When the computer teaches us how we listen: Finding higher order neurons of the auditory cortex

Philippe Esling, Stephen McAdams, Léopold Crestel and Carlos Agon

***Abstract*—Concomittent advances in biophysical observations and models of artificial intelligence seem to converge towards a deeper understanding of the human brain. The A1 neurons of the primary auditory cortex are now well defined and known to process Spectro-Temporal Receptive Fields (STRF). We postulated that the learning mechanisms inside our brain might follow an identical procedure to develop all its levels of processing. Hence, our working hypothesis was that if we could develop a computational model able to learn the STRF by itself in an unsupervised manner, then what we truly learned is not only the already known STRF, but most importantly how the brain itself learns. Therefore by re-applying this learning model, we could discover the functions of neurons in more advanced (secondary) processing areas of the brain (A2, pSTG and even up to STG).**

I. Introduction

Need to find the latest article on "topology of the auditory cortex" (and topotony)

Neural memory cell = each neuron is 1 template of the 1st layer activation

Lateral inhibition is already neural ? Where is the limite between innate and learning

[4]

auditory cortex, is able to parse complex acoustic scenes into meaningful auditory objects and streams under adverse conditions. In the current work, a biologically plausible model of this process is presented. two stages: i a feature analysis stage that maps the acoustic input into a multidimensional cortical representation and ii an integrative stage that recursively builds up expectations of how streams evolve over time and reconciles its predictions with the incoming sensory input by sorting it into different clusters. This approach yields a robust computational scheme for speaker separation under conditions of speech or music interference. The model can also emulate the archetypal streaming percepts of tonal stimuli that have long been tested in human subjects.

the development of efficient and robust mathematical models which can match up to the biological performance of auditory scene analysis tasks. a extract relevant cues from the acoustic mixture in both monaural and binaural pathways, b organize the available sensory information into perceptual streams, c efficiently manage the biological constraints and computational resources of the system to perform this task in real time, and d dynamically adapt the processing parameters to successfully keep up with continuously changing environmental conditions.

biologically inspired systems that can perform intelligent processing of complex sound mixtures.

"behaviorally" realistic, requiring no prior training on specific sounds voices, languages, or other databases, no assumptions of statistical independence or multiple microphones, and can segregate sounds ranging from the simple tones and noise to the complex speech and music.

i a feature analysis stage that explicitly represents known perceptual features in a multidimensional space, such as tonotopic frequency, harmonicity pitch, and spectral shape and dynamics timbre and ii an integrative and clustering stage that reconciles incoming perceptual features with expectations derived from a recursive estimation of the state of the streams present in the environment.

abstracts and incorporates three critical properties of A1 physiology.

(i) Multidimensional feature representation: Auditory cortical neurons are selectively sensitive to a host of acoustic features sometimes in an ordered manner manifested as "response maps" such as the tonotopic axis, tuning bandwidth and asymmetry maps, thresholds, fast temporal AM and FM modulations, and interaural cues [9].

(ii) Multiscale dynamics: A1 responses exhibit temporal dynamics [6] that are commensurate with time scales of stream formation and auditory grouping as well as speech syllabic rates and intonation contours, musical melodies, and many other sensory percepts [1]. Furthermore, numerous studies have corroborated the involvement of the auditory cortex in the temporal organization of acoustic sequences

(iii) Rapidly adapting receptive fields: Physiological and imaging studies suggest that cortical receptive fields are rapidly modulated by auditory experience, behavioral context, attentional and cognitive state, expectations and memories, as well as the statistics of the stimulus (Fritz et al., 2005; Hughes et al., 2001).

adaptive cortex that optimizes its representation of incoming sounds according to the objectives of the auditory task.

This stage parallels the feature selectivity in neurons along the auditory pathway up to the auditory cortex, whereby a host of cells is tuned or best driven by different attributes of the acoustic signal along several dimensions tonotopic frequency, spectral shape, etc.

By having a rich enough representation, different sound objects occupy different regions of the feature space, hence the emergence of "clean looks" of each individual stream.

the initial analysis stage consists of the following operations.

i A frequency analysis that captures early auditory processing in the cochlea and midbrain nuclei Shamma, 1998; Wang and Shamma, 1994.

ii A harmonicity analysis which groups harmonically related components into different channels in a process consistent with the perception of pitch Goldstein, 1973; Oxenham et al., 2004; Wightman, 1973.

iii A multiscale spectral analysis of the auditory spectrogram, as presumed to occur in primary auditory cortex Schreiner, 1998.

[7]

However, relatively little is known about what specific features of natural speech are represented in intermediate and higher order human auditory cortex. In particular, the posterior superior temporal gyrus (pSTG), part of classical Wernicke's area [8], is thought to play a critical role in the transformation of acoustic information into phonetic and pre-lexical representations [4,5,9,10]. PSTG is believed to participate in an "intermediate" stage of processing that extracts spectro-temporal features essential for auditory object recognition and discards nonessential acoustic features [4,5,9–11]. To investigate the nature of this auditory representation, we directly quantified how well different stimulus representations account for observed neural responses in nonprimary human auditory cortex, including areas along the lateral surface of STG.

Good lead for evaluation : One approach, referred to as stimulus reconstruction [12–15], is to measure population neural responses to various stimuli and then evaluate how accurately the original stimulus can be reconstructed from the measured responses. Furthermore, different stimulus representations, referred to as encoding models, can be directly

compared to test hypotheses about how the neural population represents auditory function [16]

an exact reconstruction of the physical stimulus is not expected. However, analysis of stimulus reconstruction can reveal the key auditory features that are preserved in the temporal cortex representation of speech.

Speech sounds are characterized by both slow and fast temporal modulations (e.g., syllable rate versus onsets) as well as narrow and broad spectral modulations (e.g., harmonics versus formants) [7]. Reconstructing the modulation representation proceeds similarly to the spectrogram, except that individual reconstructed stimulus components now correspond to modulation energy at different rates and scales instead of spectral energy at different acoustic frequencies

we estimated modulation rate tuning curves at individual STG electrode sites (n = 195) using linear and nonlinear STRFs, which are based on the spectrogram and modulation representations, respectively (Figure S4). Consistent with prior recordings from lateral temporal human cortex [31], average envelope-locked responses exhibit prominent tuning to low rates (1–8 Hz) with a gradual loss of sensitivity at higher rates

illustrates the ability of the modulation model to account for a rapid decrease in the spectrogram envelope without a corresponding decrease in the neural response.

These combined results support the idea of an emergent population-level representation of temporal modulation energy in primate auditory cortex [37].

These findings demonstrate that key features in continuous and novel speech signals can be accurately reconstructed from STG neural responses using both spectrogram and modulation-based auditory representations

The modulation representation itself is a nonlinear transformation of the spectrogram and is based on emergent tuning properties that have been identified in the auditory cortex [18].

spectrogram reconstruction has been demonstrated using neural responses from mammalian primary auditory cortex [14] or the avian midbrain [15]. Beyond primary auditory areas, further processing in intermediate and higher-order auditory cortex likely results in additional stimulus transformations [5]. In this study, we examined human STG, a nonprimary auditory area, and found that a nonlinear modulation representation yielded the best overall reconstruction accuracy

Alternatively, the true features represented by STG may not be readily inverted back to an intelligible acoustic waveform. For speech comprehension, it is hypothesized that intermediate and higher-order auditory areas extract or construct information-rich features of speech, while discarding nonessential low-level acoustic information [4,5,9, 10,43]. In the case that STG applies a highly nonlinear stimulus transformation, an exact reconstruction of the acoustic signal from STG responses would not be possible.

Although more work is needed to characterize the neural representation in the STG, this suggests that such key features are preserved at this stage in auditory processing. Our results are therefore consistent with the idea of pSTG as an intermediate stage in a hierarchy of auditory object processing [5,9,10,44]. Hierarchical auditory object processing has been hypothesized to follow a ventral “what” pathway, with an antero-lateral gradient along the superior temporal region [5,9,10,11] where stimulus selectivity increases from pure tones in primary auditory cortex to words and sentences in anterior STG [5].

At a more abstract level of representation, a recent functional imaging study also demonstrated that the semantic content of nouns could be used as an effective encoding model across multiple cortical regions [48].

Our results suggest that a similar approach may be usefully applied to the auditory cortex, where structural auditory models may partially account for responses in primary and intermediate areas (e.g., A1 and pSTG), but development of

higher level encoding models could be required to describe more anterior areas in the ventral auditory pathway.

development of neural interfaces for communication, for example by revealing the content of inner speech imagery.

[3]

combined transfer function for two directions is not symmetric, and hence units in AI are not, in general, fully separable.

lack of full separability stems from differences between the upward and downward spectral cross-sections but not from the temporal cross-sections; this places strong constraints on the neural inputs of these AI units

Neuronal responses are vigorous and well phase-locked to these spectral and temporal envelope modulations over a range of ripple velocities and densities. Measuring the amplitude and phase of the locked component of the response enables one to construct transfer functions. A transfer function can be inverse-Fourier transformed to obtain the STRF that characterizes a unit's dynamics and selectivity along the tonotopic axis.

The second important assumption deals with the separability of the temporal and spectral aspects of the responses.

separability was validated for ripples moving only in one direction (spectral envelope moving downward in frequency), a notion also known as "quadrant separability." In this report, we compare the separable functions (spectral and temporal) across upward and downward quadrants. If the functions are the same across quadrants, the responses are "fully separable" (i.e., they are separable); otherwise they are quadrant separable, which is a (specialized) form of inseparability.

Theoretically, fully separable responses imply an STRF that is fully decomposable into the product of a purely temporal impulse response and a purely spectral response field. It also implies a unit that responds equally well to upward and downward moving ripples and hence has necessarily a symmetric transfer function magnitude with respect to direction

We show that there is a directional sensitivity in the response to the upward versus downward moving components of a sound's spectral envelope. This breaks the symmetry of full spectro-temporal separability and produces quadrant separability. We propose measures to quantify quadrant and full separability. Finally, we discuss the significance of the results and their relationship to results from similar auditory and analogous visual experimental paradigms.

The STRF is measured through its two-dimensional Fourier transform, or transfer function T(w, V) 5 ^wV[STRF(t 2 x)], and then inverse transformed to compute the STRF, where the coordinates dual to t and x are w and V, respectively (see Fig. 3). By measuring the sinusoidal component with temporal frequency w of the response ywV(t) of a cell to a ripple of specific ripple velocity w and ripple density V, we can obtain the transfer function T(w, V) at one point in w 2 V space This way, we derive the amplitude uT(w, V)u and phase F(w, V) of the complex transfer function T(w, V) by measuring the amplitude and phase of the (real) response of the cell. Note that the use of complex numbers is not theoretically necessary, but it does simplify the calculations in the transfer function space considerably

Separability is an important property of the transfer functions. A fully separable transfer function is one that factorizes into a function of w and a function of V over all quadrants

A transfer function may also be only partially separable in that it is separable only for ripples moving in a given direction (upward vs. downward). In this case, the transfer function is called quadrant separable and can be expressed as the product of two independent functions

AI units respond in a phase-locked fashion to the moving ripples over a range of velocities and directions that depend on the ripple density of the spectrum. In particular, responses are usually tuned around a specific ripple velocity and density.

rich variety of shapes and cover a wide range of stimulus

parameters. The STRF describes the way AI units integrate stimulus power along the spectro-temporal dimensions. 4) We illustrate a variety of STRFs with a broad range of BFs, bandwidths, asymmetrical inhibition, temporal dynamics, and direction selectivity.

An important property of the responses is that for ripples moving in only one direction, the spectral and temporal functions are separable: within each quadrant they can be measured independently of each other. The property of quadrant separability makes it possible to measure the overall spectro-temporal transfer function in reasonable times using only single ripples

Inseparability is a necessary condition for the formation of more complex STRFs; direction selectivity is one possible consequence of inseparability

Direction selectivity implies that a unit is differentially responsive to one direction of ripple movement and hence must have a significant nonzero directionality index. Therefore direction selectivity necessarily implies an inseparable STRF. The opposite is not true: an inseparable STRF might reflect other factors such as asymmetric temporal and/or spectral transfer functions

Separability also places strong constraints on the underlying biological processes that give rise to the STRF shapes. For example, full separability suggests that the STRF is constituted of independent temporal and spectral processing stages. By contrast, inseparability (or just quadrant separability) implies spectrally and temporally intertwined stages of processing with the specific form of the model being entirely dependent on the details of the transfer functions. Quadrant separability in particular is a very strong constraint on both the neural inputs and the processing of the unit: almost all neural networks (whether linear or nonlinear) with multiple fully separable STRFs as inputs will in general produce a totally inseparable STRF.

cortical neuron can be easily constructed by taking inputs from (potentially) many units with (potentially) different spectral response fields and even with (potentially) different temporal impulse response properties as long as the temporal dynamics of the inputs to the cortical cell are fast compared with the temporal dynamics of the cortical cell itself

Significantly, the property of quadrant separability with temporal symmetry does not allow for any cortical inputs unless those inputs have the same temporal behavior as the neuron studied. If, for instance, all neurons in the same cortical column have similar temporal properties, including similar neural delays, this would be consistent with quadrant separability. Otherwise, cortical inputs would break quadrant separability and create a totally inseparable neuron. Total inseparability would be expected for cortical neurons in layers that receive significant input from other cortical columns or from any other neural source with significantly different temporal processing, including (but not limited to) any significant delays.

[8]

able to reproduce perceptual distance judgments between timbres as perceived by human listeners. The study demonstrates that joint spectro-temporal features, such as those observed in the mammalian primary auditory cortex, are critical to provide the rich-enough representation necessary to account for perceptual judgments of timbre by human listeners, as well as recognition of musical instruments.

by looking for brain areas that would be selective to specific sound categories, such as voice-specific regions in secondary cortical areas [22,23] and other sound categories such as tools [24] or musical instruments [25]. A hierarchical model consistent with these findings has been proposed in which selectivity to different sound categories is refined as one climbs the processing chain [26].

dimensions found will be the most salient within the sound set; but they may not capture other dimensions which could

nevertheless be crucial for the recognition of sounds outside the set. For engineering studies, dimensions may be designed arbitrarily as long as they afford good performance in a specific task. For the imaging studies, there is no suggestion yet as to which low-level acoustic features may be used to construct the various selectivity for high-level categories while preserving invariance within a category.

Responses in primary auditory cortex (A1) exhibit rich selectivity that extends beyond the tonotopy observed in the auditory nerve. A1 neurons are not only tuned to the spectral energy at a given frequency, but also to the specifics of the local spectral shape such as its bandwidth [31], spectral symmetry [32], and temporal dynamics [33] (Figure 1). Put together, one can view the resulting representation of sound in A1 as a multidimensional mapping that spans at least three dimensions: (1) Best frequencies that span the entire auditory range; (2) Spectral shapes (including bandwidth and symmetry) that span a wide range from very broad (2–3 octaves) to narrowly tuned (,0.25 octaves); and (3) Dynamics that range from very slow to relatively fast (1–30 Hz).

To circumvent these biases, we employed a model that mimics the basic transformations along the auditory pathway up to the level of A1. Effectively, the model mapped the one-dimensional acoustic waveform onto a multidimensional feature space. Importantly, the model allowed us to sample the cortical space more uniformly than physiological data available to us, in line with findings in the literature [29,30,40].

The model projects the auditory spectrogram onto a 4-dimensional space, representing time, tonotopic frequency, spectral modulations (or scales) and temporal modulations (or rates). The four dimensions of the cortical output can be interpreted in various ways. In one view, the cortical model output is a parallel repeated representation of the auditory spectrogram viewed at different resolutions. A different view is one of a bank of spectral and temporal modulation filters with different tuning

~30 K neurons

a large and uniform sampling of the space seemed desirable.

perception of musical timbre could be effectively based on neural activations patterns that sounds evoke at the level of primary auditory cortex.

at the level of primary auditory cortex does not imply that all neural correlates of sound identification will be found at this level.

cortical analysis provides a dynamic view of the spectro-temporal modulations in the signal as they vary over time.

Second, when considering more elaborate spectro-temporal cortical representations, it appears that the full representation accounts best for human performance. The match worsens if instead marginals are used by collapsing the cortical representation onto one or more dimensions to extract the purely spectral or temporal axes or scale-rate map (Figure 3, Tables 1 and 2). This is the case even if all dimensions are used separately, suggesting that there are joint spectro-temporal features that are key to a full accounting of timbre.

it is necessary to postulate the existence of nonlinearities such as divisive normalization or synaptic depression that follows a linear spectro-temporal analysis so as to account fully for the observed responses. In the current study, the exact nature of the nonlinearity remains unclear as it is implicitly subsumed in the Gaussian kernels and subsequent decisions.

unlike the small number of spectral or temporal dimensions that have been traditionally considered in the timbre literature, we cannot highlight a simple set of neural dimensions subserving timbre perception. Instead, the model suggests that subtle perceptual distinctions exhibited by human listeners are based on ‘opportunistic’ acoustic dimensions [56] that are selected and enhanced, when required, on the rich baseline provided by the cortical spectro-temporal representation.

[5]

each acoustic task may have its own “signature” STRF changes, dependent on the salient cues

we found a distinct pattern of STRF change, characterized by an expected selective enhancement at target tone frequency but also by an equally selective depression at reference tone frequency

adaptive neuronal responses in A1 that can swiftly change to reflect both sensory content and the changing behavioral meaning of incoming acoustic stimuli.

highly specific taskdependent rules of plasticity in A1 neurons and the ability of single A1 neurons to change response fields in different behavioral contexts.

an important mechanism in A1 underlying active listening and a general principle of cortical processing during attentive behavior in other sensory systems

Details of the STRF changes in the trained animals depended on the specific frequencies of the reference and target tones and on the shape of the initial STRF. In general, the reference and target tones had opposite effects on the STRF during the discrimination task

When the reference frequency was placed just above the inhibitory area (fR 1 kHz) during behavior, the inhibitory sideband enlarged and almost doubled in strength, and its high frequency border moved upward, displacing the excitatory field.

two of the STRFs (Fig. 2c,d) reverted to their original prebehavior STRF shapes once the behavior ceased. However, the induced changes in the third cell (Fig. 2e) persisted in an intermediate form after behavior, which increased the net contrast between responses at the target and reference frequencies

. If all STRF changes were as described in Figure 2 (i.e., depression at the reference and potentiation at the target),

a minority of neurons showed no significant changes, either at target or reference (31 of 127, 24%), of all STRFs that did change significantly (96 of 127, 76%), 60% (58 of 96) satisfied this prediction.

very small degree of depression and potentiation of the STRF at the reference and target frequencies, respectively, even in the naive animal

, the behavioral effects (an overall depression) at the reference frequency shown in Figure 3a cannot be fully explained by sensory adaptation because exactly the same stimuli were presented to the naive anima

o sort one to four single-unit spike waveforms from the multiunit recording trace. Alternatively, after off-line sorting, we could construct a multiunit STRF by pooling spikes from all of the sorted unit clusters. This multiunit STRF sometimes looked very different spectrally and temporally from any of the constituent single-unit STRFs. However, because STRF changes (if they occur) are generally consistent regardless of the shape of the STRFs (i.e., follow a pattern of overall facilitation at target frequency and depression

plasticity in multiunit STRFs would be comparable with that measured in single-unit STRFs

. It is therefore evident that not all single units in a cluster necessarily show exactly the same pattern of adaptive plasticity, although, in this case, all changes that did occur were in the same direction

The specific type of change was influenced by the initial shape of the receptive field, the behavioral task, attention to the salient acoustic cues, and was also likely to be modulated by general influences reflecting the animal’s state of arousal, motor preparation, and reward expectation

(1) identifying the salient task cues, (2) linking them to behavior to achieve the goals of the ongoing task.

majority of STRFs (75%) of individual neurons and multiunit clusters changed significantly during performance of either the frequency discrimination

This suggests a widespread process of adaptive modulation

that affects a broad variety of STRFs throughout A1. However, there was also a stable group of STRFs that did not apparently change during behavior (25% of recorded cells), and it is interesting that, even in behaviorally labile STRFs, the plastic changes modulated the strength of preexisting inhibitory or excitatory STRF fields but seldom caused an outright change of synaptic sign at best frequency

predict that there would be virtually no lasting changes in the A1 tonotopic map, because there was no predominant behavioral focus on any single frequency throughout training.

fascinating finding: that difficult tone discrimination training (i.e., differential classical conditioning with closely adjacent CS and CS frequencies) that did not lead to successful behavioral learning still resulted in consistent receptive field changes in A1. In other words, cortical adaptive responses can occur before, or even without, associated behavioral changes.

A1 stimulus-specific adaptation

A speculative explanation for the dominant pattern of behavioral plasticity reported (overall suppression at reference and facilitation for target) is that the auditory system, for voluntary, attentive behavioral tasks, has built on a preexisting set of automatic, preattentive neural mechanisms that are normally used to detect acoustic novelty and show, in miniature, the same response pattern as seen in frequency discrimination behavior.

receptive field properties are continuously being modified

in identifying salient features of the acoustic or the visual scene, regulating adaptive plasticity, enhancing responses, and reshaping neuronal receptive field properties, enabling A1 neurons to multiplex acoustic inputs for different acoustic tasks.

[2]

a unified multiresolution representation of the spectral and temporal features

cochlear analysis of sound and the extraction of the acoustic spectrum in the cochlear nucleus are only the earliest stages in a sequence of substantial transformations of the neural representation of sound as it journeys up to the auditory cortex via the midbrain and thalamus.

organization in the more central structures of the inferior colliculus, medial geniculate body, and the cortex have only begun to be uncovered relatively recently

The model we describe is not biophysical in spirit, but rather it abstracts from the physiological data an interpretation

apparent progressive loss of temporal dynamics from the periphery to the cortex. Thus, on the auditory nerve, rapid phase locking to individual spectral components of the stimulus survives up to 4 – 9 kHz. It diminishes to moderate rates of synchrony in the midbrain under 1 kHz, and to the much lower rates of modulations in the cortex less than 30 Hz

Another important change in the nature of the neural responses is the emergence of elaborate selectivity to combined spectral and temporal features, selectivity that is typically much more complex than the relatively simple tuning curves and dynamics of auditory-nerve fiber responses

An early stage captures monaural processing from the cochlea to the midbrain. It transforms the acoustic stimulus to an auditory time-frequency spectrogramlike

. The second is called the cortical stage because it reflects the more complex spectrotemporal analysis presumed to take place in mammalian AI.

A STRF summarizes the way a cell responds to the stimulus. Along its ordinate—“frequency axis”—

. Thus, some STRFs are responsive excited or suppressed over a broad range of frequencies, exceeding an octave ii, while others are quite narrowly tuned iv.

each STRF acts as a modulation selective filter of its input spectrogram, specifically tuned to a particular range of spectral resolutions

The collection of all such STRFs then would constitute

a filterbank spanning the broad range of psychoacoustically observed scale and rate sensitivity

The model consists of two major transformations of the acoustic signal:

1 A frequency analysis stage associated with the cochlea, cochlear nucleus, and response features observed in the midbrain: This stage effectively computes an affine wavelet transform of the acoustic signal with a spectral resolution of about 10% Lyon and Shamma, 1996.

2 A spectrotemporal multiresolution analysis stage postulated to conclude in the primary auditory cortex: This stage effectively computes a two-dimensional affine wavelet transform with a Gabor-like spectrotemporal mother-wavelet

As with the early auditory stage, the multiresolution cortical model is highly schematic and lacks realistic biophysical mechanisms and parameters. Nevertheless, the model aims to capture perceptually significant features in the auditory spectrogram, and hence justify its relevance through its successful application in accounting for a variety of perceptual thresholds and tasks

real cortical STRFs Fig. 1 are far more complex than the simple Gabor-like shapes we have employed in the model.

A. Peripheral auditory processing

The initial stage of the model starts with a transformation of the signal from a pressure time waveform to a spatiotemporal activation pattern. It captures the basic processing taking place at the level of the auditory periphery Pickles, 1988, including cochlear filtering, hair-cell transduction, and auditory-nerve and cochlear-nucleus spectrotemporal sharpening. Briefly, it consists of a cochlear-filter bank of 128 constant-Q highly asymmetric bandpass filters Q= 4 equally spaced on a logarithmic frequency axis x with center frequencies spanning a range of 5.3 octaves i.e., with a resolution of 24 channels per octave. Next, a hair-cell stage transduces the cochlear-filter outputs into auditorynerve patterns via a three-step process consisting of highpass filtering, a nonlinear compression, and low-pass leakage, effectively limiting the temporal fluctuations below 5 kHz. Finally, a lateral inhibitory network performs a sharpening of the filter-bank frequency selectivity mimicking the role of cochlear-nucleus neurons Sacks and Blackburn, 1991; Shamma, 1998. It is modeled as a first difference operation across the frequency channels, followed by a halfwave rectifier, and then a short-term integrator. Extensive details of the biophysical grounds, computational implementation, and perceptual relevance of this model can be found in Wang and Shamma 1994 and Yang et al. 1992. We complement the model above with an additional onset sharpening stage to emphasize the presence of transient events in this spectrographic representation. We apply a high-pass filter filter cutoff about 400 Hz to the output of each frequency channel to boost the transient energy in the signal.

A time-frequency activity pattern Pt,x that represents a noise-robust Wang and Shamma, 1994 equivalent of an acoustic spectrum, called a sharpened-onset auditory spectrogram see Fig. 2. It not only encodes the instantaneous power in each frequency band but also captures the temporal dynamics of the spectral components falling within the bandwidth of each band, giving rise to fast “envelope modulations” of the signal.

The processing of the acoustic signal in the cochlea is modeled as a bank of 128 constant-Q asymmetric bandpass filters equally spaced on the logarithmic frequency scale spanning 5.3 octaves. The cochlear output is then transduced into inner hair cells potentials via a high pass and low pass operation. The resulting auditory nerve signals undergo further spectral sharpening via a lateral inhibitory network. Finally, a midbrain model resulting in additional loss in phase locking is performed using short term integration with time constant 4 ms resulting in a time frequency representation called as the

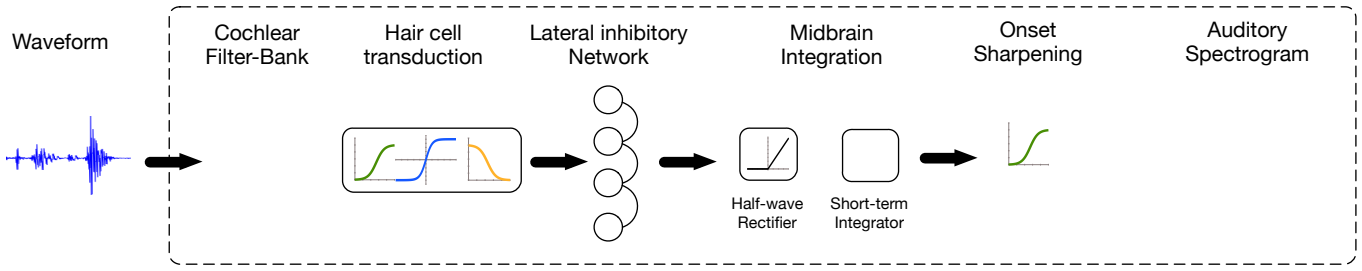

Figure 1. Peripheral auditory processing workflow

auditory spectrogram.

The stages of the early auditory model are illustrated in Fig. 2. In brief, the first operation is an affine wavelet transform of the acoustic signal $s(t)$. It represents the spectral analysis performed by the cochlear filter bank. This analysis stage is implemented by a bank of 128 overlapping constant Q ($Q_{10dB} \approx 3$) bandpass filters with center frequencies CFs that are uniformly distributed along a logarithmic frequency axis x, over 5.3 oct 24 filters/ octave. The impulse response of each filter2 is denoted by $h(t;x)$. These cochlear filter outputs $y_{coch}(t,x)$ are transduced into auditory-nerve patterns $y_{AN}(t,x)$ by a hair cell stage consisting of a highpass filter, a nonlinear compression $g(\cdot)$, and a membrane leakage low-pass filter $w(t)$ accounting for decrease of phase-locking on the auditory nerve beyond 2 kHz. The final transformation simulates the action of a lateral inhibitory network LIN postulated to exist in the cochlear nucleus Shamma, 1989, which effectively enhances the frequency selectivity of the cochlear filter bank Lyon and Shamma, 1996; Shamma, 1985b. The LIN is simply approximated by a first-order derivative with respect to the tonotopic axis and followed by a half-wave rectifier to produce $y_{LIN}(t,x)$. The final output of this stage is obtained by integrating $y_{LIN}(t,x)$ over a short window, $\mu(t,\tau) = e^{-t/\tau}u(t)$, with time constant τ = 8 ms mimicking the further loss of phase locking observed in the midbrain. The mathematical formulation for this model can be summarized as follows: ycocht,x = st

$$y_{coch}(t,x) = s(t) \otimes_t h(t;x)$$

$$y_{AN}(t,x) = g\left(\delta_t y_{coch}(t,x)\right) \otimes_t w(t)$$

$$y_{LIN}(t,x) = max\left(\delta_x y_{AN}(t,x), 0\right)$$

$$y_{final}(t,x) = y_{LIN}(t,x) \otimes_t \mu(t;\tau)$$

cochlear filtering is essentially linear, lacking such phenomena as two-tone suppression and level-dependent tuning, which are critical in some applications Carney, 1993. The lateral inhibition model is very schematic and lacks details of single neurons Shamma, 1989. We also have no explicit adaptive properties in our current model

B. Cortical processing

Our current understanding of cortical processing reveals that cortical units exhibit a wide variety of receptive field profiles Kowalski et al., 1996; Miller et al., 2002; Elhilali et al., 2007. These response fields, also called spectrotemporal receptive fields STRFs, represent a time-frequency transfer function of each neuron, hence capturing the specific sound features that selectively drive the cell best. Functionally, such rich variety implies that each STRF acts as a selective filter specific to a range of spectral resolutions or scales and tuned to a limited range of temporal modulations or rates, covering the broad

span of psychoacoustically observed modulation sensitivities in humans and animals Eddins and Bero, 2007; Green, 1986; Viemeister, 1979.

The central stage further analyzes the spectro-temporal content of the auditory spectrogram using a bank of modulation selective filters centered at each frequency along the tonotopic axis, modeling neurophysiological receptive fields. This step corresponds to a 2D affine wavelet transform, with a spectro-temporal mother wavelet, define as Gabor-shaped in frequency and exponential in time. Each filter is tuned (Q = 1) to a specific rate (v in Hz) of temporal modulations and a specific scale of spectral modulations (V in cycles/octave), and a directional orientation (+ for upward and 2 for downward).

For input spectrogram z(t,f), the response of each STRF in the model is given by:

$$r_{\pm}\left(t, f; \omega, \Omega, \theta, \phi\right) = z\left(t, f\right) *_{t,f} STRF_{\pm}\left(t, f; \omega, \Omega, \theta, \phi\right)$$

where t,f denotes convolution in time and frequency and h and w are the characteristic phases of the STRF's which determine the degree of asymmetry in the time and frequency axes respectively. The model filters STRFz(t,f; v,V; h,w) filters can be decomposed in each quadrant (upward + or downward 2) into RF(t; v; h) into SF(f; V; w) corresponding to rate and scale filters respectively. Details of the design of the filter functions STRFz can be found in [58]. The present study uses 11 spectral filters with characteristic scales [0.25, 0.35, 0.50, 0.71, 1.00, 1.41, 2.00, 2.83, 4.00, 5.66, 8.00] (cycles/octave) and 11 temporal filters with characteristic rates [4.0, 5.7, 8.0, 11.3, 16.0, 22.6, 32.0, 45.3, 64.0, 90.5, 128.0] (Hz), each with upward and downward directionality. All outputs are integrated over the time duration of each note. In order to simplify the analysis, we limit our computations to the magnitude of the cortical output rz(t,f; v,V; h,w) (i.e. responses corresponding to zero-phase filters).

The second analysis stage mimics aspects of the responses of higher central auditory stages especially the primary auditory cortex.

bank of filters that are selective to different spectrotemporal modulation parameters that range from slow to fast rates temporally, and from narrow to broad scales spectrally.

Three features are of particular interest: i it is centered on a particular center frequency CF. The location of the excitatory white and inhibitory black stripes

ii the modulation rate along the time axis is about 16 Hz; and iii the excitatory portions are separated on the vertical axis by about 1 oct, giving rise to a spectral "scale" sensitivity to peaks separated by 1 oct,

The filter output is computed by a convolution of its STRF with the input auditory spectrogram yfinalt,x, i

responses across the filter bank, with different stimuli being differentiated by which filters they activate best. The response map provides a unique characterization of the spectrogram, one that is sensitive to the spectral shape and dynamics of the entire stimulus.

We assume a bank of "idealized" STRFs as depicted in Fig. 5a. Each STRF is selective to a narrow range of temporal and spectral modulations and is also directionally sensitive to either upward or downward drifting modulations. A complete set of such STRFs with a range of temporal and spectral selectivity e.g., 1 – 300 Hz, and 0.25–8 peaks or cycles/octave would be sufficient to decompose and characterize the modulations in the auditory spectrogram. More realistic complex STRFs can be readily formed by superposition of these basic STRFs. We define the STRF as a real function that is formed by combining two complex functions in a manner consistent with extensive physiological data. Specifically, experimental STRFs are not necessarily time-frequency separable. Instead, we have found that they are almost always so-called "quadrant separable."3 This requires that the STRF be represented as the real of the product of a complex temporal and a complex

spectral "impulse response" function, $h_{IRT}(t)$ and $h_{IRS}(x)$, as follows: $STRF = \mathcal{R}\{h_{IRT}(t) \cdot h_{IRS}(x)\}$, where

$$h_{IRS}(x:\Omega,\phi) = h_{irs}(x;\Omega,\phi) + j\hat{h}_{irs}(x;\Omega,\phi)$$

$$h_{IRT}(t;\omega,\theta) = h_{irt}(t;\omega,\theta) + j\hat{h}_{irt}(t;\omega,\theta)$$

$\mathcal{R}\{\cdot\}$ denotes the real part, and $h(\cdot)$ and $\hat{h}(\cdot)$ denote Hilbert transform pairs. The real functions $h_{irs}(\cdot)$ and $h_{irt}(\cdot)$ are defined by sinusoidally interpolating seed functions $h_s(\cdot)$, $h_t(\cdot)$ and their Hilbert transforms Wang and Shamma, 1995

$$h_{irs}(x:\Omega,\phi) = h_s(x;\Omega)\,cos\;\phi + \hat{h}_s(x;\Omega)\,sin\,\phi$$

$$h_{irt}(t;\omega,\theta) = h_t(t;\omega)\,cos\,\theta + \hat{h}_t(t;\omega)\,sin\,\theta$$

where Ω and ω are the spectral density and velocity parameters of the filters; ϕ and θ are characteristic phases; $h_s(\cdot)$ and $h_t(\cdot)$ are the spectral and temporal functions that determine the modulation selectivity of the STRF, and $\hat{h_s}(\cdot)$ and $\hat{h}_t(\cdot)$ are their Hilbert transforms. In addition, the directional sensitivity of the STRF is modeled as

$$STRF_{\Downarrow} = \mathcal{R}\{h_{IRT}(t) \cdot h_{IRS}(x)\}$$

$$STRF_{\Uparrow} = \mathcal{R}\{h^{*}_{IRT}(t) \cdot h_{IRS}(x)\}$$

, where $*$ denotes the complex conjugate; $\Downarrow$ and $\Uparrow$ denote downward and upward moving direction respectively. Note, the downward STRF shown in Fig. 5a is a special case of $\theta = \phi = 0$. We choose $h_s(\cdot)$ to be a Gabor-like function commonly used in the vision literature to describe the analogous spatial aspect of a receptive field Jones and Palmer, 1987. It is defined as the second derivative of a Gaussian function; $h_t(\cdot)$ is assumed to be a gamma function e.g., as in Slaney 1998.

$$h_s(x) = \left(1 - x^2\right) e^{-x^2/2}$$

$$h_t(t) = t^2 e^{-3.5t} sin\,(2\pi t)$$

And for different scales and rates

$$h_s(x;\Omega) = \Omega h_s(\Omega x)$$

$$h_t(t;\omega) = \omega h_t(\omega t)$$

Therefore, the STRF in general is an inseparable spectrotemporal function of $h_s(\cdot)$ and $h_t(\cdot)$, with a specific highly constrained spectrotemporal structure known as "quadrant separable." The spectrotemporal response of a downward upward cell c for an input spectrogram $y(t,s)$ is then given by

$$r_{c\Downarrow(\Uparrow)}(t,x:\omega_c,\Omega_c,\theta_c,\phi_c) = y(t,x) \otimes_{tx} \mathcal{R}\left\{h^{(*)}_{IRT}(t;\omega_c,\theta_c) \cdot h^{(*)}_{IRS}(x;\Omega_c\right.$$

where $\otimes$ denotes convolution with respect to both t and x. This multiscale multirate or multiresolution spectrotemporal response is called "cortical representation." Substituting Eqs. 5–8 into Eq. 9, the cortical representation at downward or upward cell c can be rewritten as

$$r_{c\Downarrow}(t,x:\omega_c,\Omega_c,\theta_c,\phi_c) = y(t,x) \otimes_{tx} \left[\left(h_t h_s - \hat{h}_t \hat{h}_s\right) cos\,(\theta_c + \phi_c) + \left(\hat{h}\right.\right.$$

and

$$r_{c\Uparrow}(t,x:\omega_c,\Omega_c,\theta_c,\phi_c) = y(t,x) \otimes_{tx} \left[\left(h_t h_s + \hat{h}_t \hat{h}_s\right) cos\,(\theta_c - \phi_c) + \left(\hat{h}\right.\right.$$

where $h_t \equiv h_t(t;\omega_c)$ and $h_s \equiv h_s(x;\Omega_c)$ to simplify notation.

C. Integrative and clustering stage

The second integrative and clustering stage induces stream segregation by reconciling incoming sensory information with gradually formed expectations. integrate these incoming looks and set up computational rules to cluster them by labeling them according to the different streams present in the scene.

This integration process postulates that clusters of A1 neurons with typical multiscale dynamics of 2 – 30 Hz Miller et al., 2002 integrate their sensory inputs to maintain a form of a working memory representation. This memory trace is used to build expectations of how a stream evolves over time and makes predictions about what is expected at the next time instant. By reconciling these expectations with the actual incoming sensory cues, the system is able to assign incoming features to the perceptual group that matches them best Nix and Hohmann, 2007. Specifically, the integrative stage consists of different cortical clusters two clusters in the current model, both governed by a recursive Markovian process which i integrates the input of each cortical array with dynamics typical of time constants of A1, ii uses a Kalman-filter-based estimation to track the evolution of each array/stream over time, and iii utilizes the recent auditory experience to infer what each cluster expects to "hear" next. FIG.

II. Results

We explore the model's ability to mimic human perception as we vary these two critical parameters frequency separation between the two notes FAB and tone repetition time.

In many circumstances, it is reported that buildup of streaming can range from a few hundreds of milliseconds to several seconds Anstis and Saida, 1985; Bregman, 1978. To reproduce these time scales, it is necessary to incorporate more biological realism to the cortical processing stage via simple adaptation processes or synaptic depression mechanisms known to operate at the level of thalamocortical projections see Elhilali et al. 2004 for details. Habituation of A1 responses over time has been postulated as a possible neural mechanism responsible for the observed perceptual buildup of streaming Micheyl et al., 2005.

Timbre-based segregation Stream segregation in general can be induced between sound sequences that differ sufficiently along any feature dimension in the cortical representation, including different distributions along the spectral analysis axis timbre, harmonicity axis pitch, and other axes not included

Implicit in the notion of a stream is a sequence of sounds that share consistent or smoothly varying properties

The processes of particular interest here are the multiscale analysis, the cortical dynamics, and the adaptive nature of cortical processing

We remove the multiscale analysis stage and basically perform the clustering operation directly on the harmonic patterns extracted from the auditory spectrogram

We argued earlier that the contribution of the cortical spectral analysis is to map different spectral patterns into different regions of the multiscale axis

Using the tonotopic axis alone, the two vowels overlap greatly. In contrast, the multiscale analysis reveals the different timbre structures arising from the two vowels.

strengthens the claim that the topographic organization of mammalian auditory cortex with neurons of different spectral resolutions and sensitivities orthogonal to its tonotopic organization does indeed underlie the system's ability to distinguish between natural sounds of distinct timbres e.g., speech Sutter, 2005.

We stipulate in the current model that cortical time constants play a role in the process of auditory scene organization by facilitating the tracking of sound streams over the course of few to tens of hertz. To test this hypothesis, we modified the range of cortical time constants implemented in the model by adjusting the parameters of the temporal filters

Key to this organizational role are the multiple time scales typically observed in cortical responses, as well as an internal representation of recent memory that allows the smooth evolution of streams over time. A powerful aspect of our formulation is its real-time capability because the model forms auditory streams as it receives its input data, requiring no prior

training on a specific speech corpus or early exposure to a set of voices, sound categories, and patterns.

augmenting the cortical multidimensional representation with a spatial dimension whose responses are computed from midbrainlike processing of binaural cues

can potentially guide our understanding of the brain function in general and biological auditory scene analysis, in particular.

The organization of the auditory pathway up to the auditory cortex indicates that different auditory features are extracted from the incoming sounds at various stages and probably organized into auditory objects at the cortical level Nelken, 2004. This rich image which emerges at the level of A1 effectively projects the acoustic waveform into a higher dimensional perceptual space in a mapping reminiscent of operations taking place in classification and regression techniques such as support vector machines and kernelbased classifiers Cristianini and Shawe-Taylor, 2000; Herbrich, 2001.

allowing them to occupy nonoverlapping parts of the perceptual space

Evidence from auditory cortical physiology is consistent with this view Woolley et al., 2005 and suggests that the correspondence between cortical tuning and spectrotemporal features in natural sounds constitutes a mapping that effectively enhances discriminability among different sounds.

Information in sound occurs on multiple time scales with different temporal features having distinct acoustic manifestations, neural instantiations, and perceptual roles. At the level of the central auditory system particularly the primary auditory cortex, numerous physiological investigations have shown that cortical responses appear to be particularly tuned to relatively slow rates of the order of few to tens of hertz. The sluggishness of cortical responses has been postulated to correspond very closely to important information components in speech and music.

the appropriate choice of time constants for this process is crucial for achieving the desired performance.

. The present model is grounded in this view and actually relies on the choice of cortical dynamics to regulate the tempo of the feature integration.

Our present implementation accomplishes this process via an ongoing accumulation of expectations of each object. These expectations can be thought of as a matched filter that permits into the cluster only sensory patterns that are broadly consistent with recent history of that class.

. The physiological plausibility of this mechanism rests on the existence of feedback projections that mediate the adaptive representation of biological information under continuously changing behavioral contexts and environments. Adaptive signal processing techniques such as Kalman filtering have been successfully implemented to model many forms of dynamic neural adaptation and plasticity in hippocampal and motor circuits Eden et al., 2004; Srinivasan et al., 2006; Wu et al., 2006.

The present model is based on the premise that streaming is reflected in the response properties of neurons in the primary auditory cortex.

e identified a possible correlate of stream segregation in the primary auditory cortex A1

changes in the spectrotemporal tuning of cortical receptive fields in a direction that promotes streaming and facilitates the formation of two segregated objects Yin et al., 2007. The question remains, however, as to where and how exactly in the model does the adaptive nature of the STRFs emerge and serve functionally to promote streaming?

Specifically, if we were to imagine recording from a cortical cell represented by one of the integrators in Fig. 1 in the quiescent state i.e., without feedback, we would observe responses with spectral and temporal selectivity that mimics the STRFs typically seen in A1 Miller et al., 2002. During streaming, the top-down feedback would alter the input into

this cell at the junction labeled “unsupervised clustering” in Fig. 1, effectively changing the selectivity of the cell or its STRF.

This view of the relationship between streaming and rapid STRF plasticity makes several specific predictions that need to be explored in the future. For instance, STRF changes in the model essentially represent the neural correlate of the so-called “buildup” of streaming.

However, an alternative hypothesis is that the top-down “feedback loop” in the model is enabled only when the listener’s attention is engaged. Clearly, without feedback, clustering ceases and no streams can form. Attention, we postulate, engages the feedback loop and enables streaming.

further postulate that “selective attention” to one stream or another can modulate the gain in the appropriate feedback loop, and hence favor the formation and perception of one e.g., the foreground stream over the other the background.

III. Applications

A. Audio classification

Use it for MIREX evaluation

B. Mimicking human perception

C. Targeting auditory

IV. Conclusion and future work

References

[1] Robert P Carlyon and Shihab Shamma. An account of monaural phase sensitivity. *The Journal of the Acoustical Society of America*, 114(1):333–348, 2003.

[2] Taishih Chi, Powen Ru, and Shihab A Shamma. Multiresolution spectrotemporal analysis of complex sounds. *The Journal of the Acoustical Society of America*, 118(2):887–906, 2005.

[3] Didier A Depireux, Jonathan Z Simon, David J Klein, and Shihab A Shamma. Spectro-temporal response field characterization with dynamic ripples in ferret primary auditory cortex. *Journal of neurophysiology*, 85(3):1220–1234, 2001.

[4] Mounya Elhilali and Shihab A Shamma. A cocktail party with a cortical twist: how cortical mechanisms contribute to sound segregation. *The Journal of the Acoustical Society of America*, 124(6):3751–3771, 2008.

[5] Jonathan B Fritz, Mounya Elhilali, and Shihab A Shamma. Differential dynamic plasticity of a1 receptive fields during multiple spectral tasks. *The Journal of neuroscience*, 25(33):7623–7635, 2005.

[6] Lee M Miller, Monty A Escabí, Heather L Read, and Christoph E Schreiner. Spectrotemporal receptive fields in the lemniscal auditory thalamus and cortex. *Journal of neurophysiology*, 87(1):516–527, 2002.

[7] Brian N Pasley, Stephen V David, Nima Mesgarani, Adeen Flinker, Shihab A Shamma, Nathan E Crone, Robert T Knight, Edward F Chang, et al. Reconstructing speech from human auditory cortex. *PLoS-Biology*, 10(1):175, 2012.

[8] K Patil, D Pressnitzer, S Shamma, and M Elhilali. Music in our ears: the biological bases of musical timbre perception. *PLoS computational biology*, 8(11):e1002759, 2012.

[9] Christoph E Schreiner and John V Urbas. Representation of amplitude modulation in the auditory cortex of the cat. ii. comparison between cortical fields. *Hearing research*, 32(1):49–63, 1988.