Adaptive cortical model for auditory streaming and monaural speaker separation

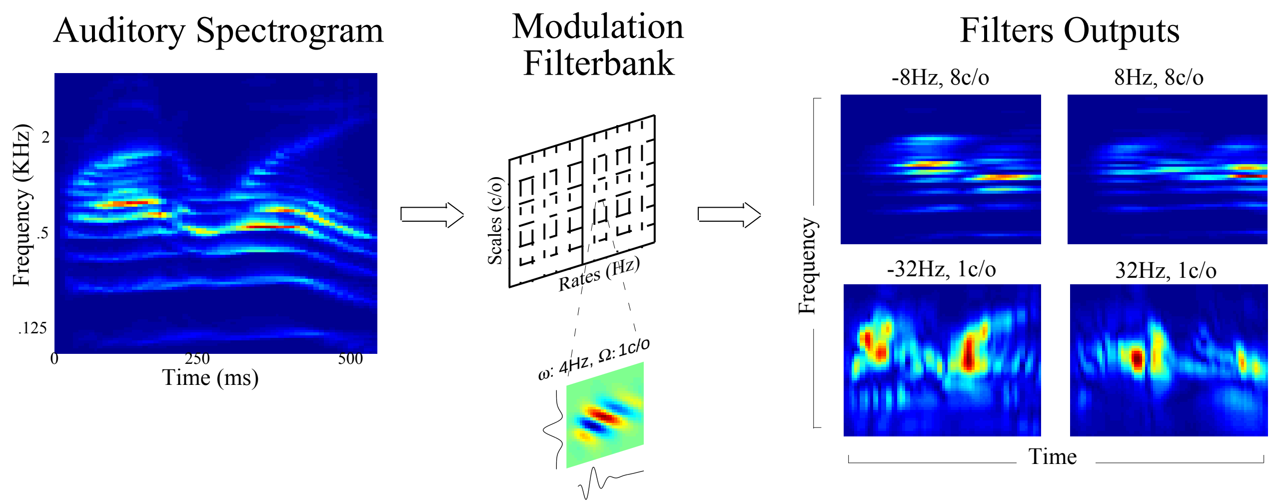
Mounya Elhilali & Shihab Shamma

Neural Systems laboratory
Institute for Systems Research
Department of Electrical and Computer Engineering
University of Maryland, College Park

Motivation & Framework

We present a biologically-inspired model of dynamic recognition and learning of auditory objects in the auditory cortex, based on unsupervised learning and the statistical theory of Kalman prediction.

This sound organization scheme uses an underlying model of cortical processing, where neural receptive fields (STRFs) are modelled by a two-dimensional multi-resolution filter-bank.

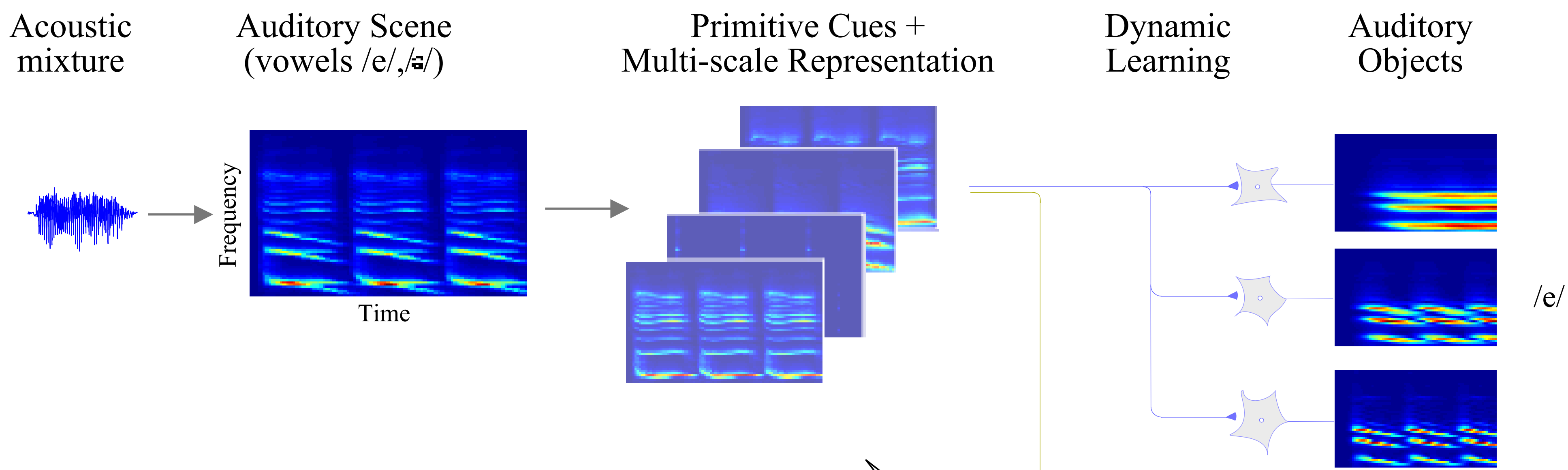


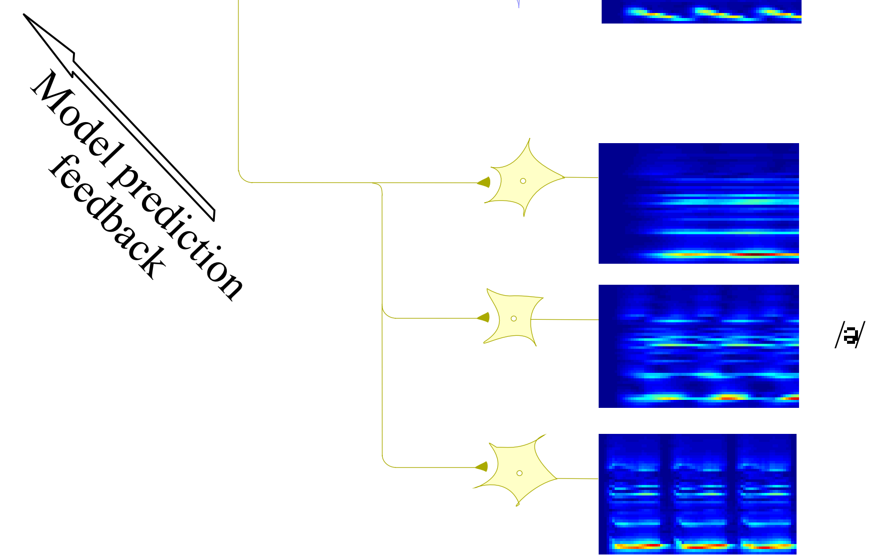
This cortical representation sets the framework for unsupervised organization of sound elements into perceptual streams, using predictions of an internal representation of the environment.

Center for Auditory and Acoustic Research

Institute for Systems Research
University of Maryland

Model Schematic





<u>Primitive cues:</u>
* Pitch extraction (based on harmonic templates)
* Onset/Offset (common start and end of events)
* Spatial location (possibly ...)

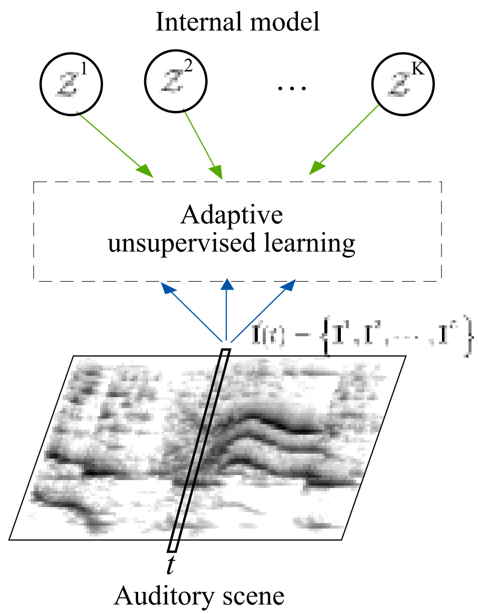
<u>Multi-scale Representation:</u>
* Timbre cues

<u>Dynamic Learning:</u>
* Multi-resolution cortical representation
* Unsupervised learning
* Kalman filtering

Center for Auditory and Acoustic Research

Institute for Systems Research
University of Maryland

Learning Algorithm



Optimization function:

$$\boxed{\mathrm{J} = max\ \mathrm{P}(model \mid \mathrm{Input})}$$

⇨ maximizing the model representation of the streams given the input data

⇨ Learning clusters which maximize a temporal continuity constraint at the output of cortical dynamic filters.

Let
- Z(t): model internal representation
- I(t): Input data of primitive cues (in multi-scale representation)

Y(t): Output through cortical filters

A,B: Cortical filter parameters

Maximizing J ⇨ Minimizing $-\log \mathrm{P}(\mathrm{I} \mid \mathrm{Z}) - \log \mathrm{P}(\mathrm{Z})$ i.e $\Delta \mathrm{J}/\Delta \mathrm{Z} = 0$

⇨ Kalman filter formulation

$$\mathrm{Z}(t) = \mathrm{A}.\ \mathrm{I}(t) + \mathcal{N}$$
$$\mathrm{Z}(t) = \mathrm{B}.\ \mathrm{Z}(t\text{-}1) + \mathrm{C}.\ \mathrm{Y}(t) + \mathcal{N}$$

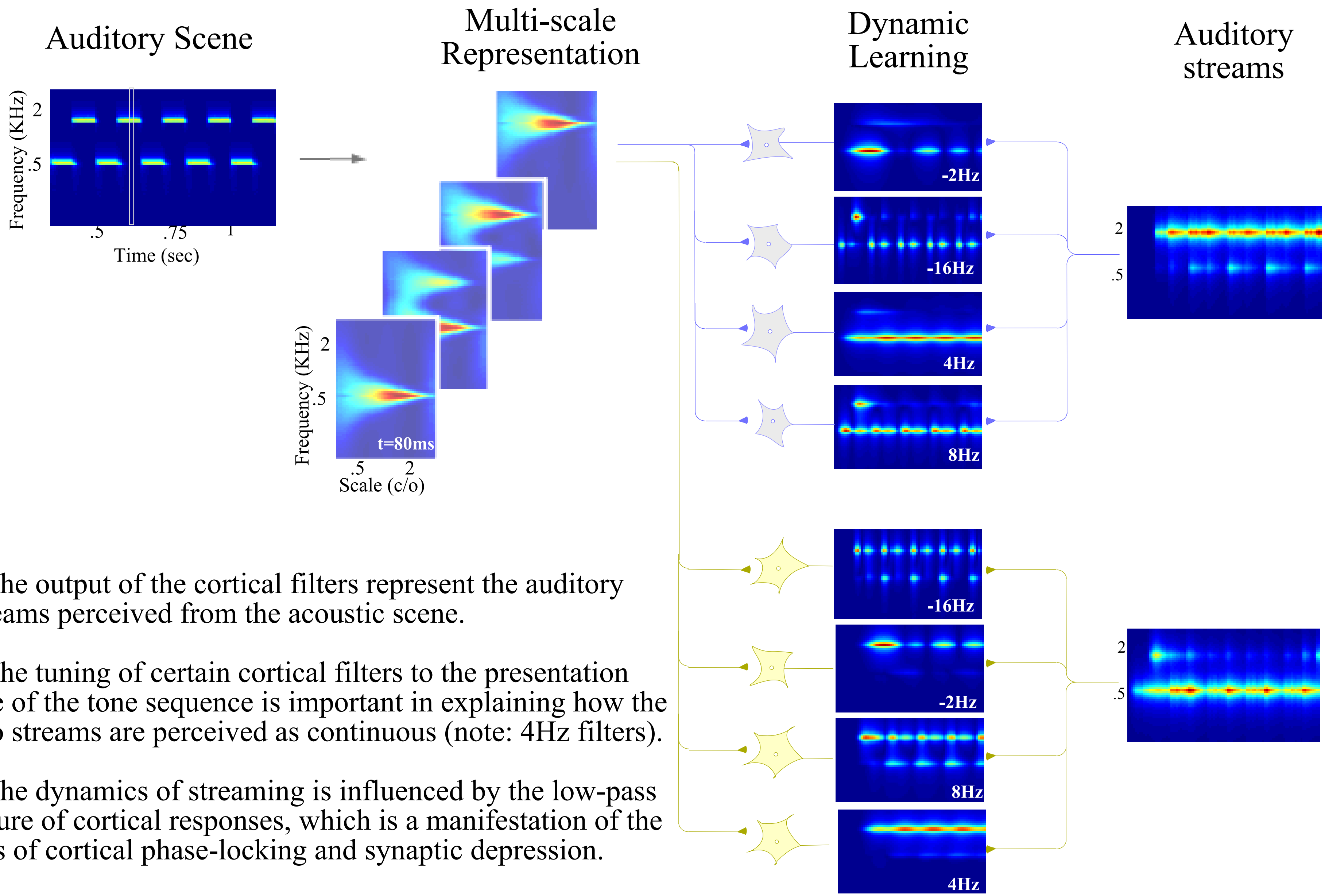
⇨ Learning function (following Kalman theory)

$$\mathrm{Z}(t+1) = \mathrm{Z}^*(t) + \mathrm{G}(t)\ (\mathrm{I}(t) - \mathrm{A}.\mathrm{Z}^*(t))$$

⇨ Competitive learning step

$$min\ (\mathrm{I}(t) - \mathrm{A}.\mathrm{Z}^*(t))$$

Alternating Tone Sequence



* The output of the cortical filters represent the auditory streams perceived from the acoustic scene.

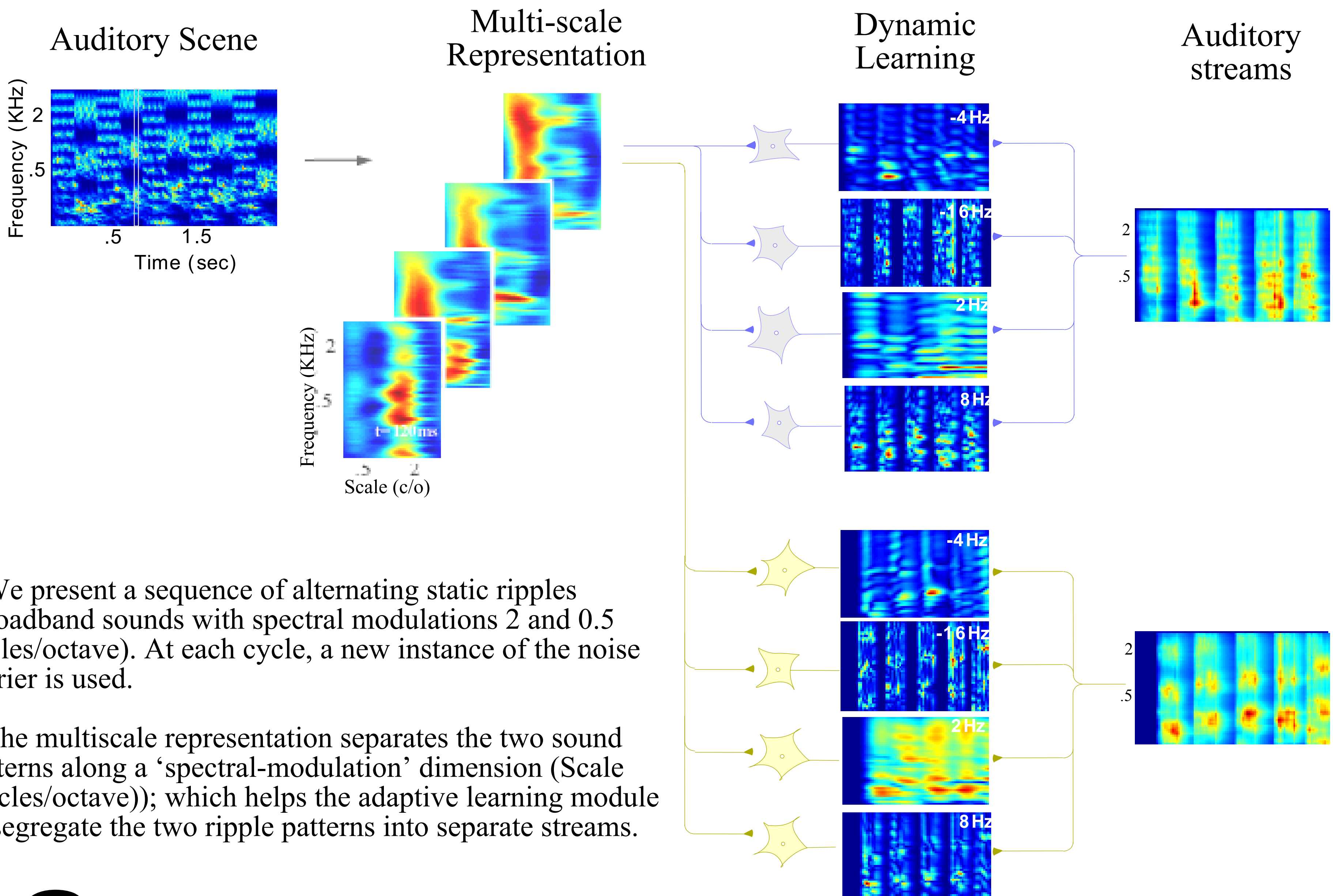
* The tuning of certain cortical filters to the presentation rate of the tone sequence is important in explaining how the two streams are perceived as continuous (note: 4Hz filters).

* The dynamics of streaming is influenced by the low-pass nature of cortical responses, which is a manifestation of the loss of cortical phase-locking and synaptic depression.

Center for Auditory
and Acoustic Research

Institute for Systems Research
University of Maryland

Alternating Ripple Sequence



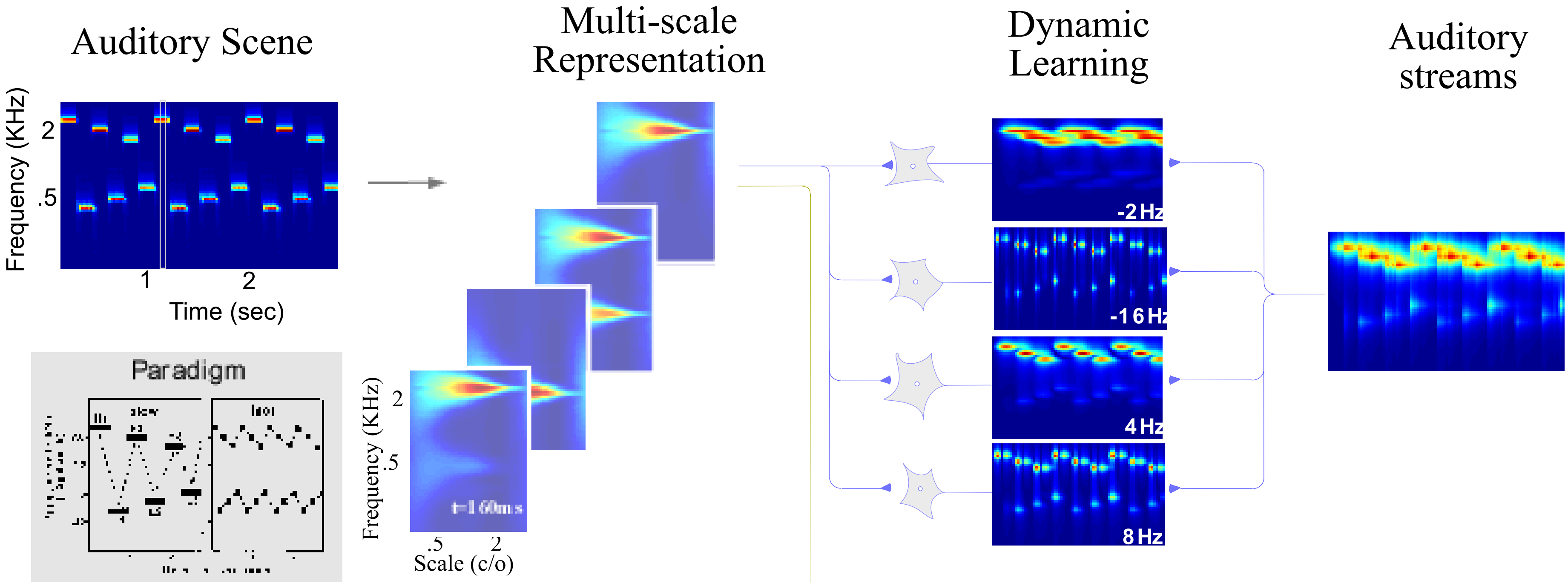
* We present a sequence of alternating static ripples (broadband sounds with spectral modulations 2 and 0.5 cycles/octave). At each cycle, a new instance of the noise carrier is used.

* The multiscale representation separates the two sound patterns along a ‘spectral-modulation’ dimension (Scale (cycles/octave)); which helps the adaptive learning module to segregate the two ripple patterns into separate streams.

Center for Auditory and Acoustic Research

Institute for Systems Research
University of Maryland

Alternating Tone Cycle

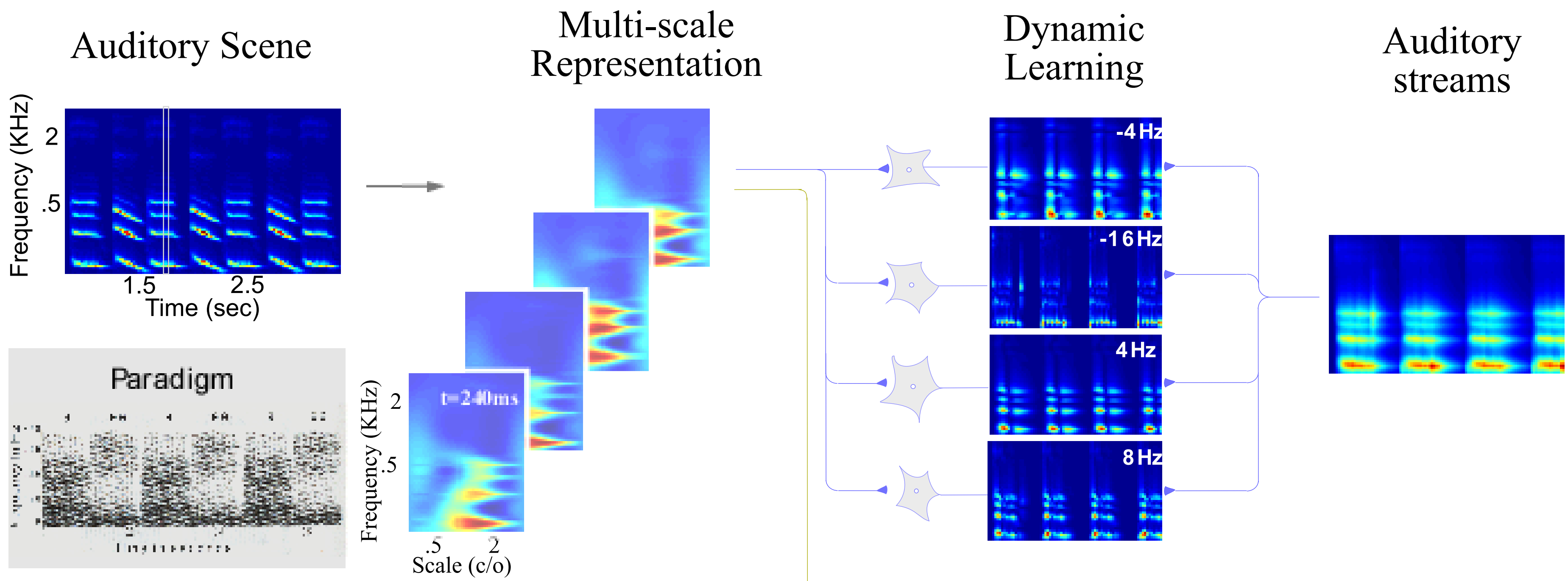


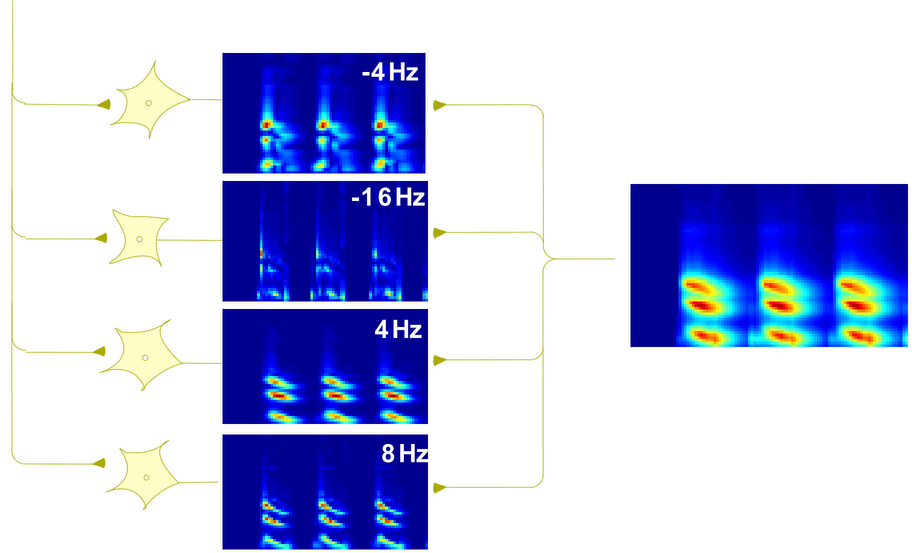
* This simulation is a direct test of the principle of *sequential integration*. The tones in the low frequency region fall in the same cluster, because the acoustic features from patterns L1, L2, L3 appear to be similar (by virtue of frequency proximity), and dissociated from the other "competing" patterns H1, H2, H3. This perception is only maintained as long as the sequences are repeated at a relatively fast rate, guaranteeing that the dynamics of the cortical model are commensurate with the presentation rate.

Center for Auditory and Acoustic Research

Institute for Systems Research
University of Maryland

Alternating Vowels





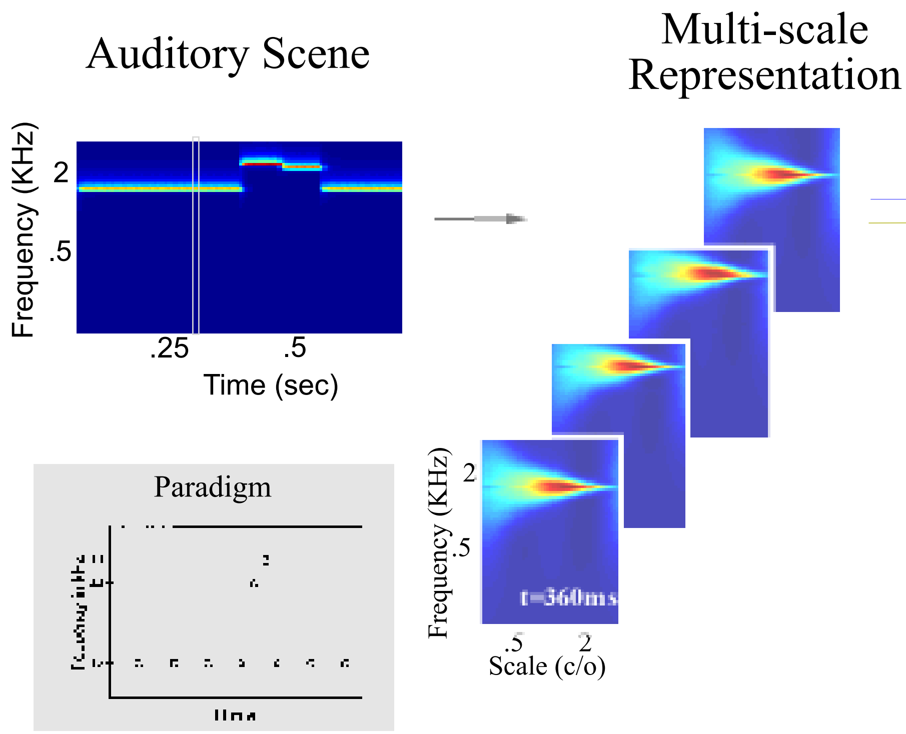
* The signals in this simulation are natural vowels /e/ and /ə/ from a same male speaker.

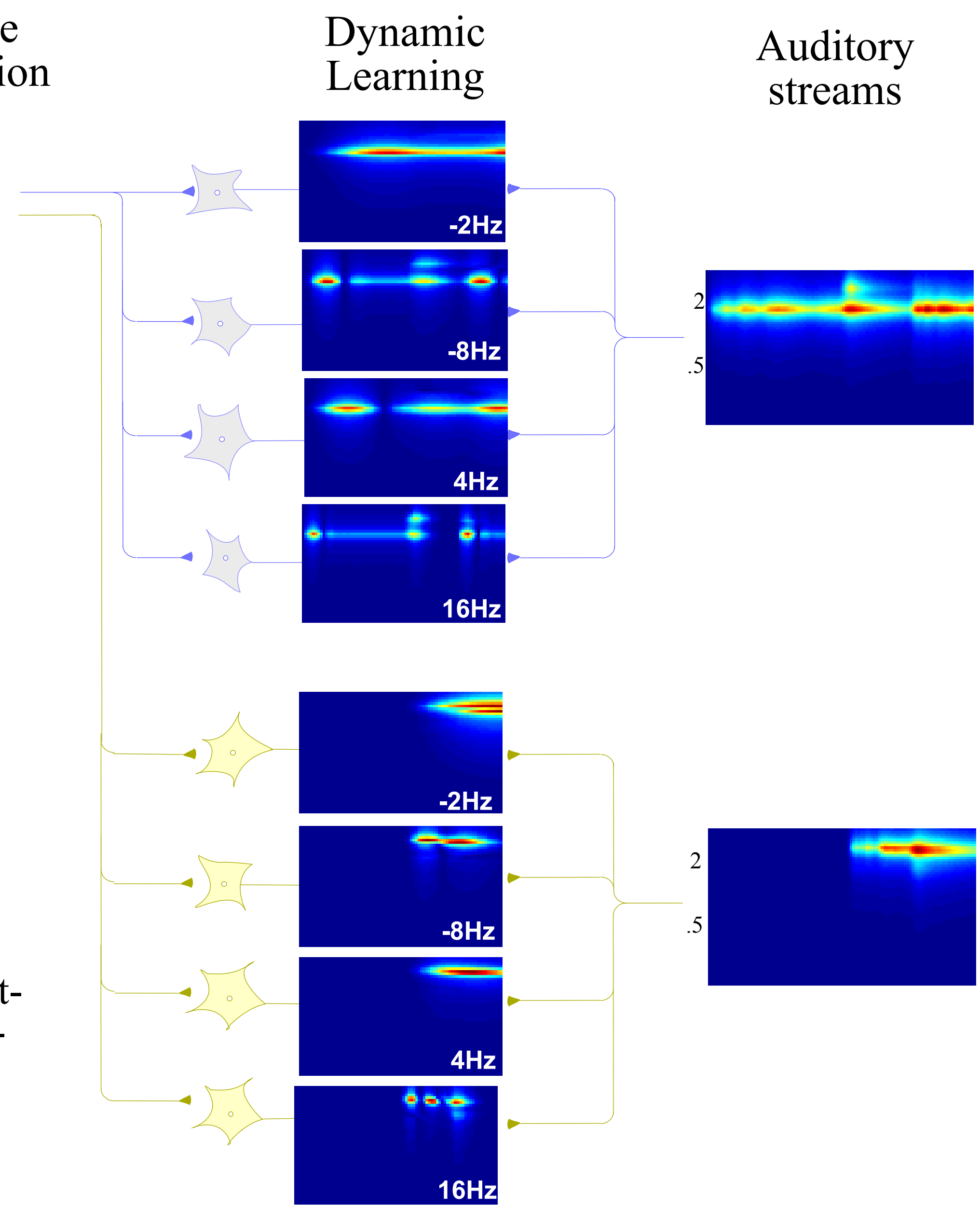
* Even though they are produced at the same pitch (same speaker), the two vowels exhibit very distinct spectral shapes (different formant positions and relative intensities). The divergence between the two patterns promote streaming effects, hence contributing to their segregation into two separate streams.

Center for Auditory and Acoustic Research

Institute for Systems Research
University of Maryland

Capturing Interference Tones



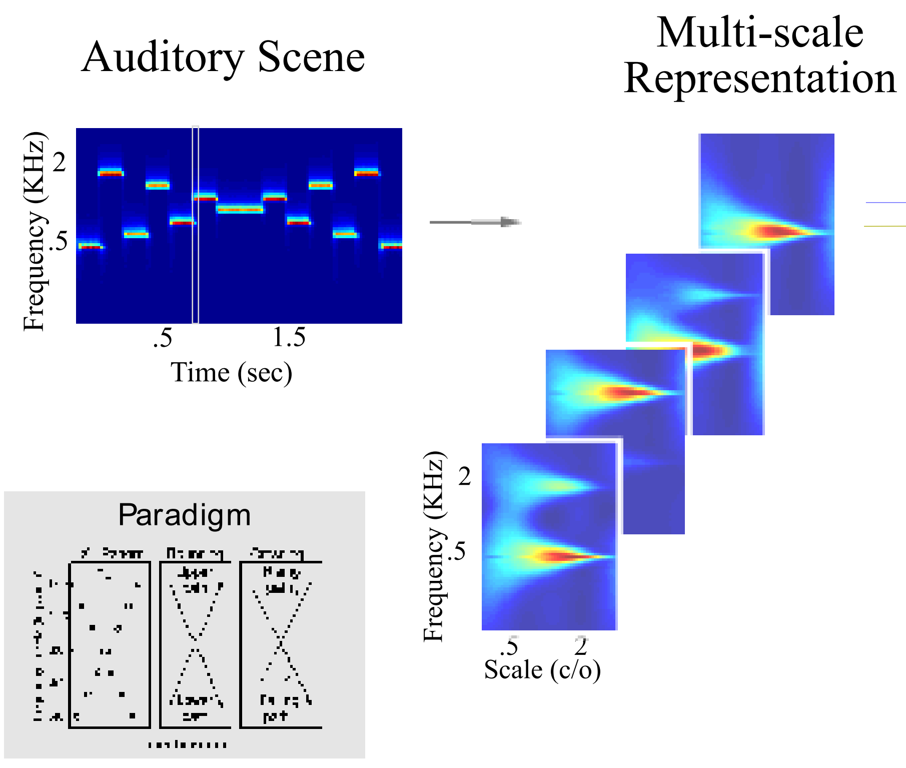


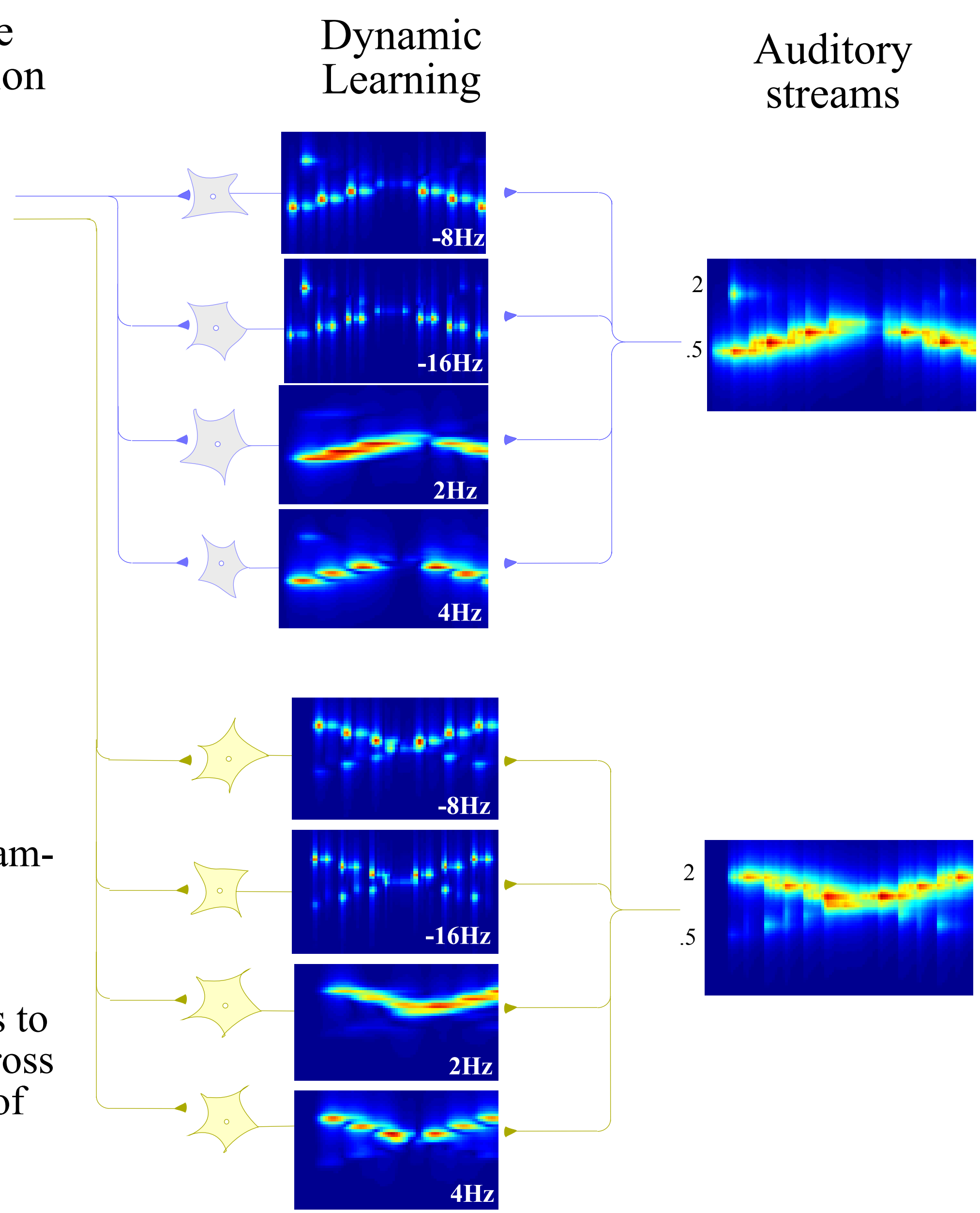
* By presenting a repeated tone sequence (C-C-C preceding A-B in the paradigm panel), we can capture the effect of interfering tones X, hence segregating the sequence of interest (AB) from its surrounding interference.

Center for Auditory and Acoustic Research

Institute for Systems Research
University of Maryland

Crossing Trajectories





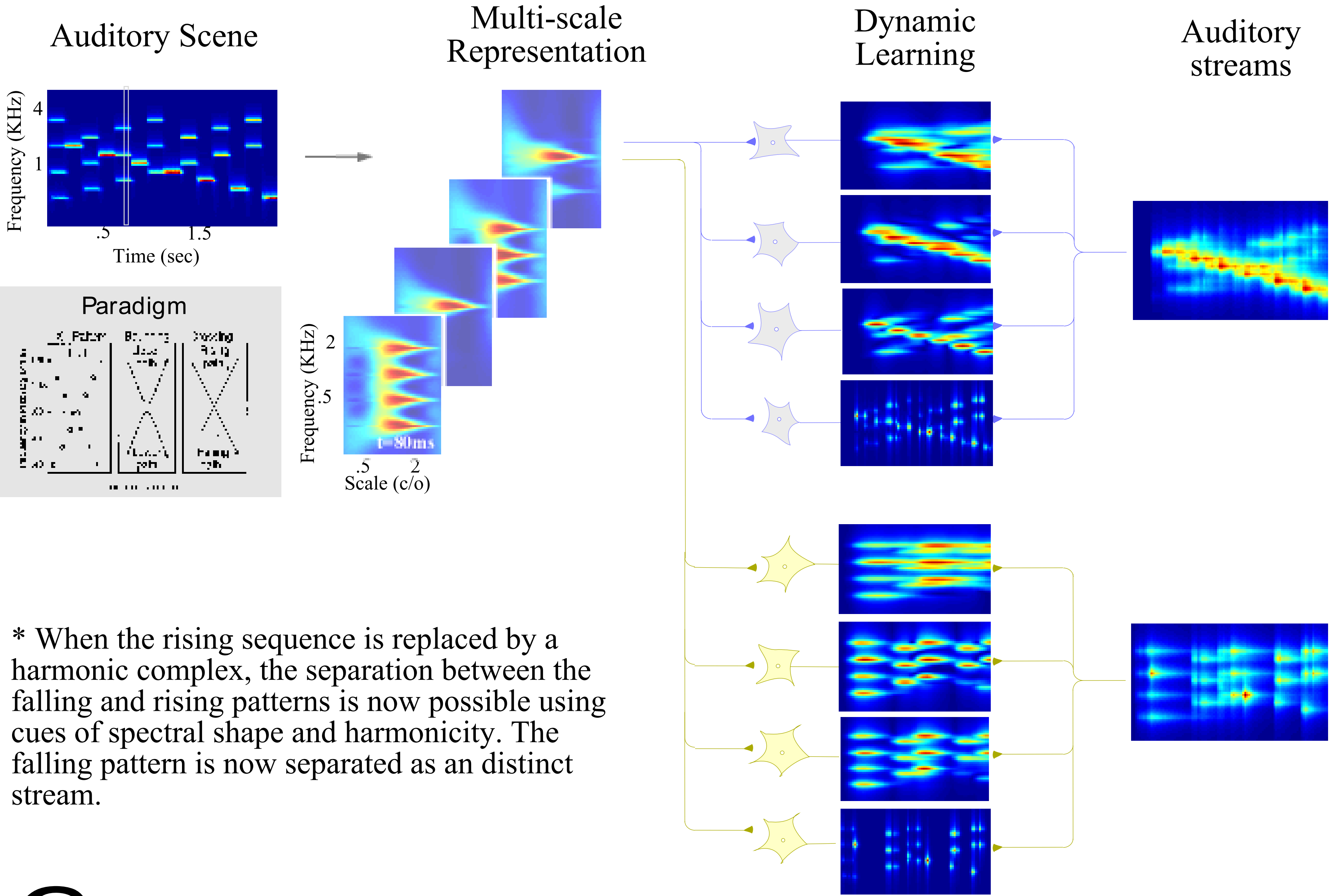
* The theory of crossing trajectories tests the streaming ability of a rising vs. falling tone sequence.

* In the case of pure tones, the streamed sounds exhibit a ‘bouncing’ effect, where the system fails to follow the rising and falling trajectories as they cross each other. This effect is illustrated in the output of the learning model.

Center for Auditory and Acoustic Research

Institute for Systems Research
University of Maryland

Crossing Trajectories (cont'd)

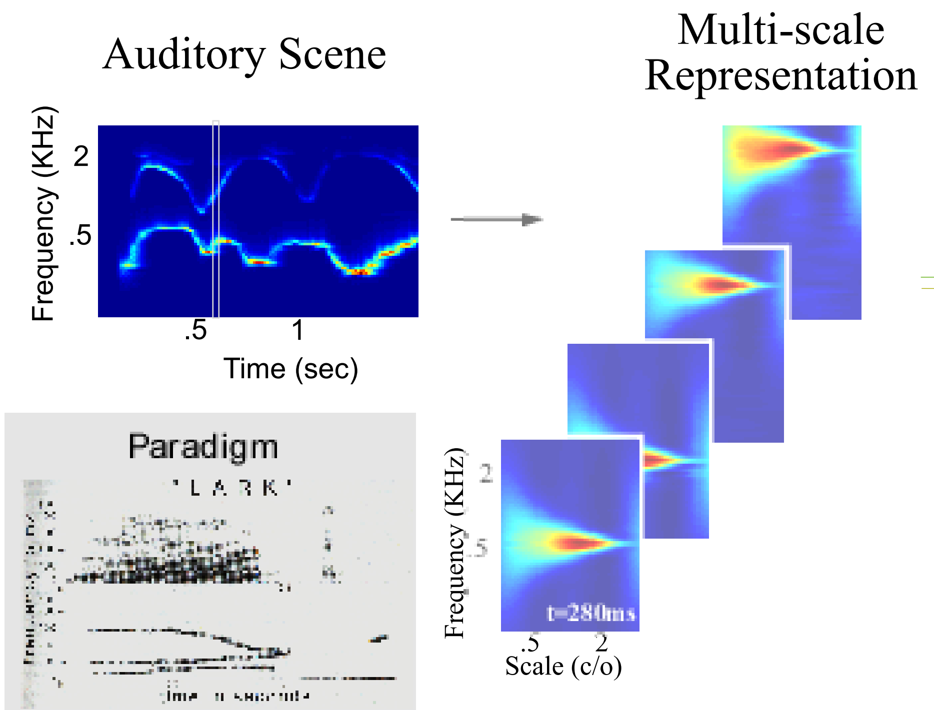


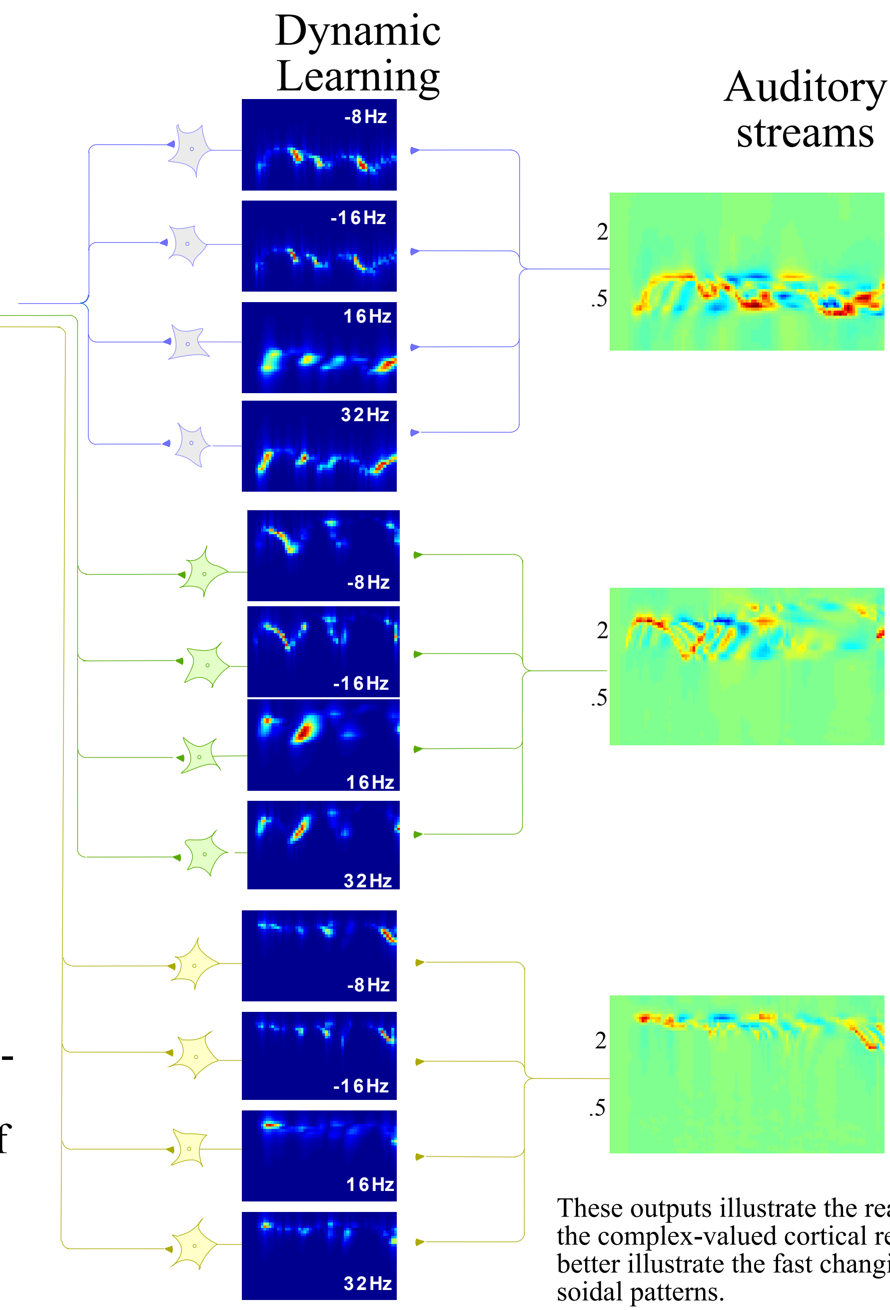
* When the rising sequence is replaced by a harmonic complex, the separation between the falling and rising patterns is now possible using cues of spectral shape and harmonicity. The falling pattern is now separated as an distinct stream.

Center for Auditory and Acoustic Research

Institute for Systems Research
University of Maryland

Sinewave Speech





These outputs illustrate the real-part of the complex-valued cortical response to better illustrate the fast changing sinusoidal patterns.

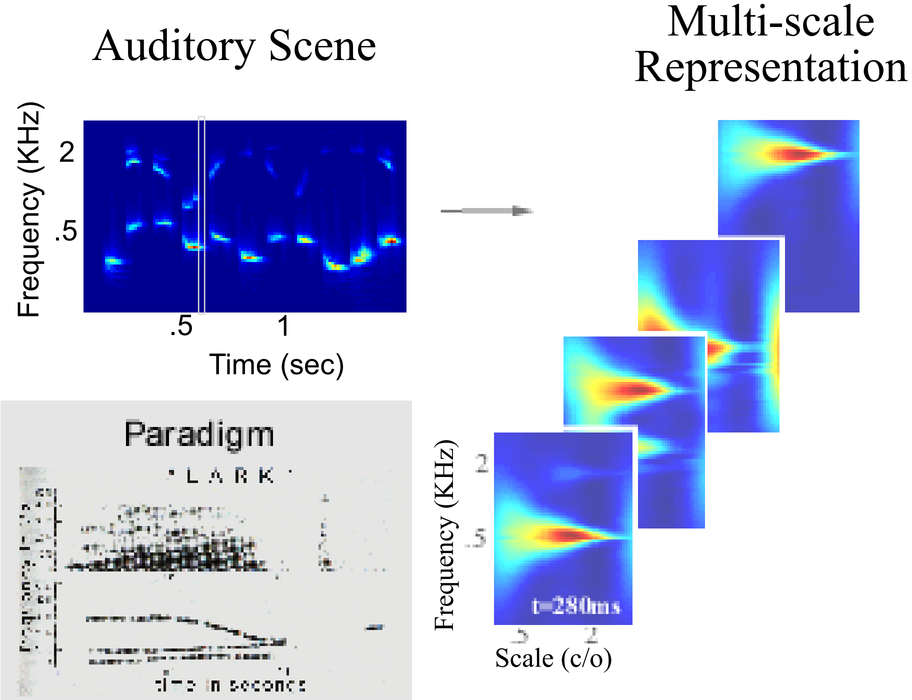
* Sinewave speech is constructed by sinusoidal waves tracking the formant trajectories of speech. It initially sounds like a collection of tones and beeps.

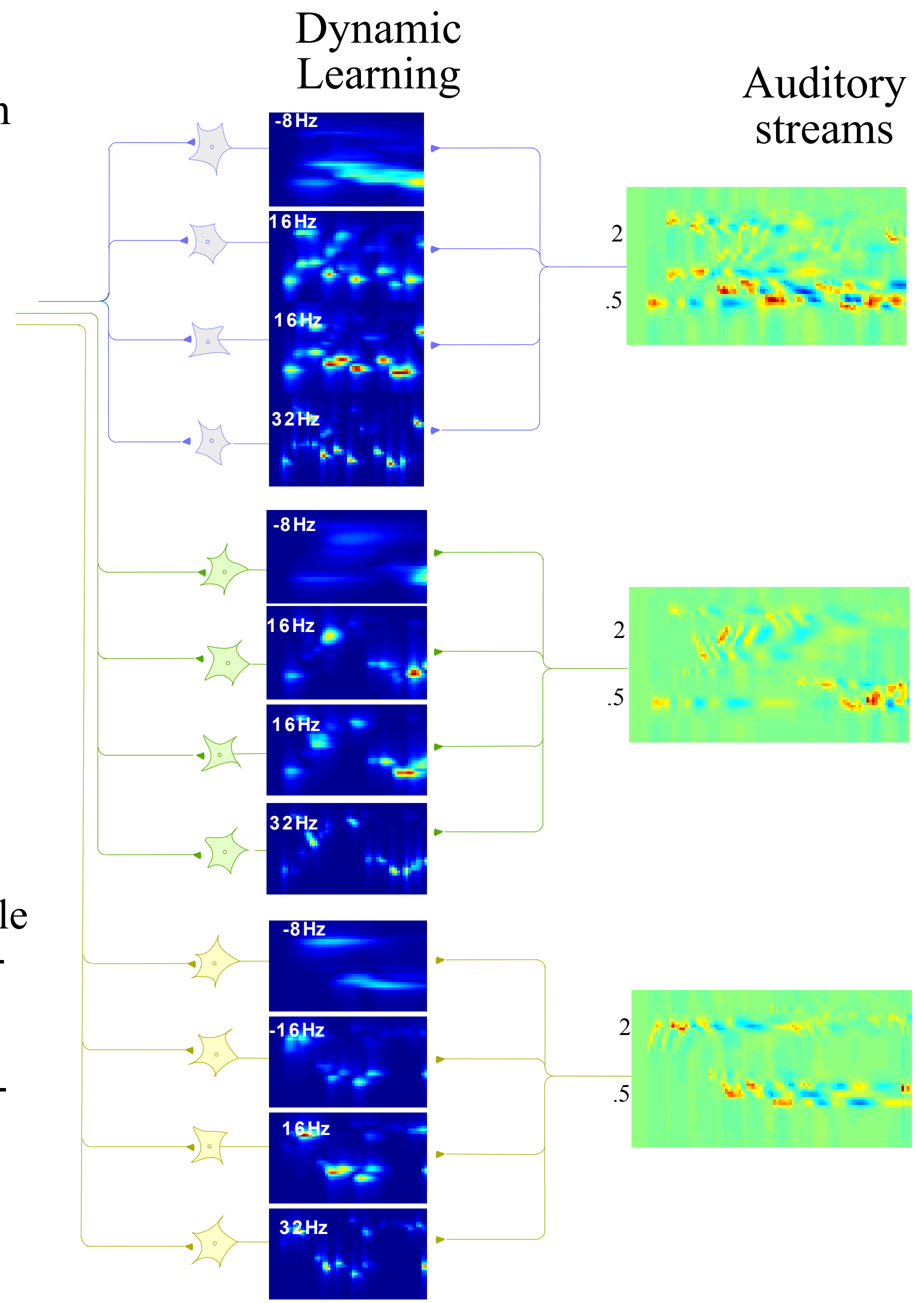
*We hypothesize that each sinewave is clustered into a separate stream, hence making it difficult for the auditory system to integrate information across perceptual objects to be able to recognize the linguistic meaning of the sentence.

Center for Auditory and Acoustic Research

Institute for Systems Research
University of Maryland

Sinewave Speech (cont'd)



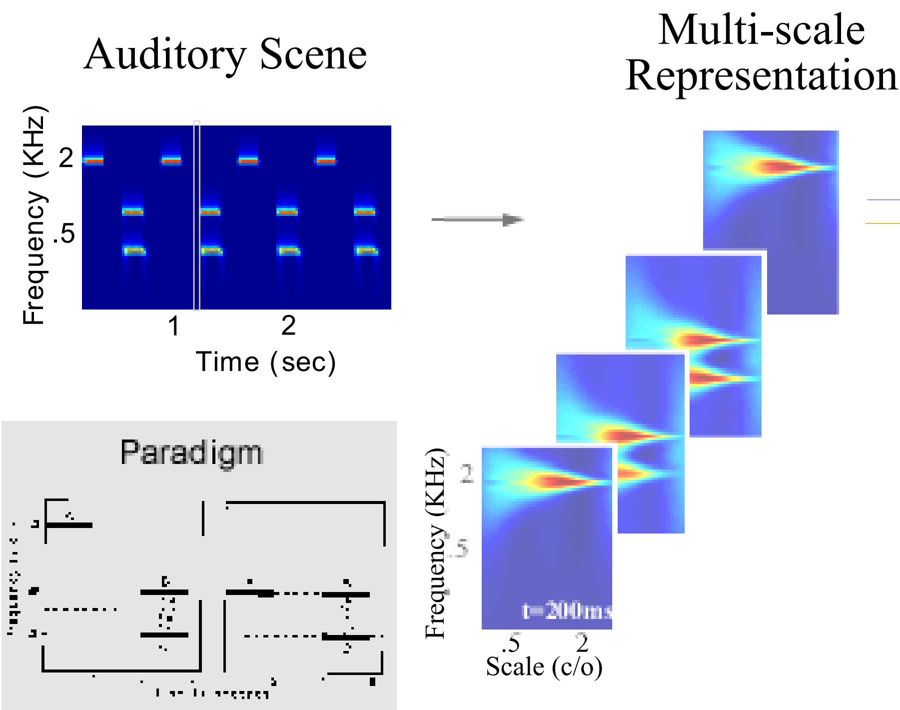


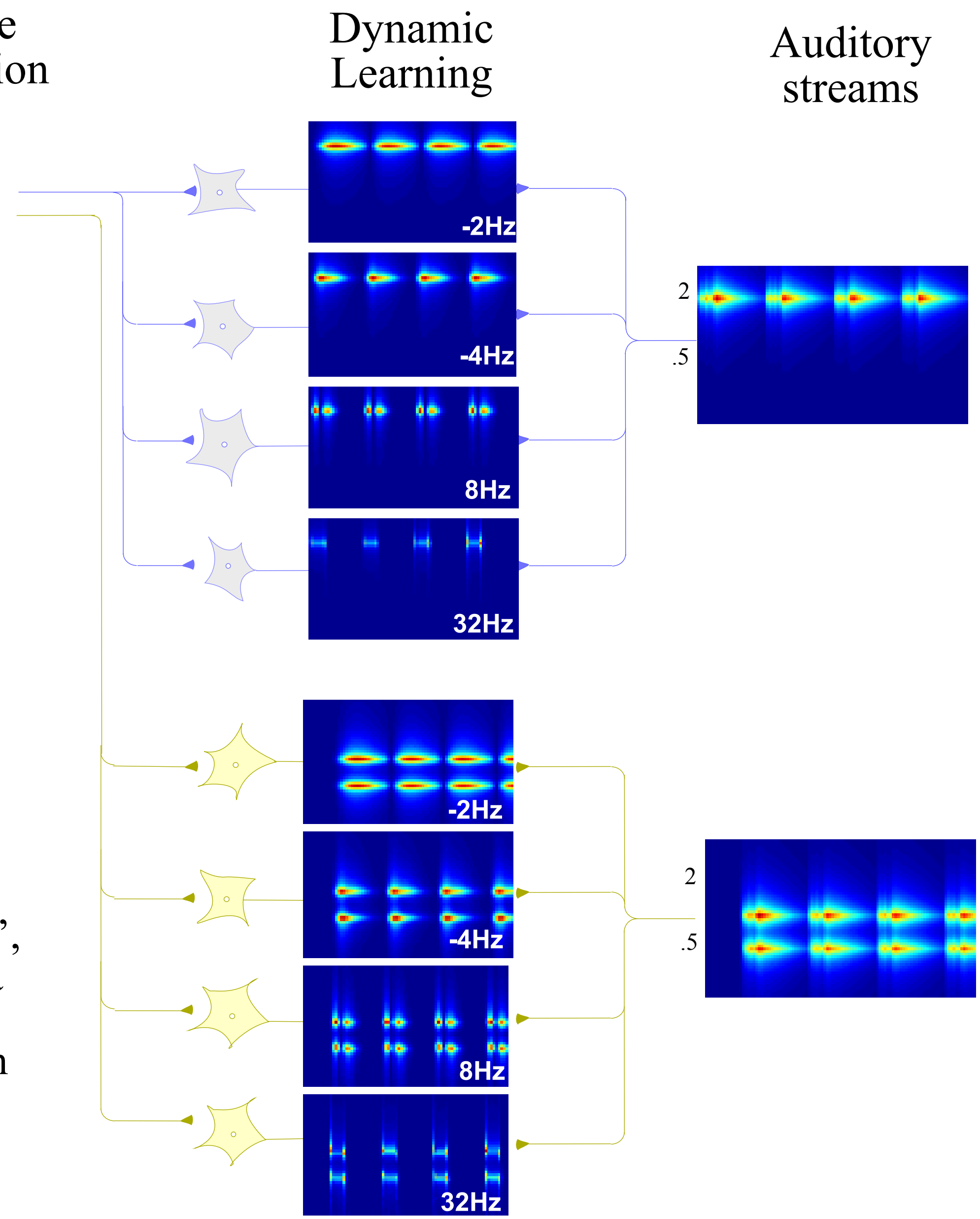
* Interestingly, sinewave speech sounds more intelligible when segments of silence are interleaved with the original signal. The silence portions introduce 'common-onset' cues giving evidence that the three sinusoidal waves should be integrated together in the same perceptual stream.

Center for Auditory and Acoustic Research

Institute for Systems Research
University of Maryland

Tone in a Mixture



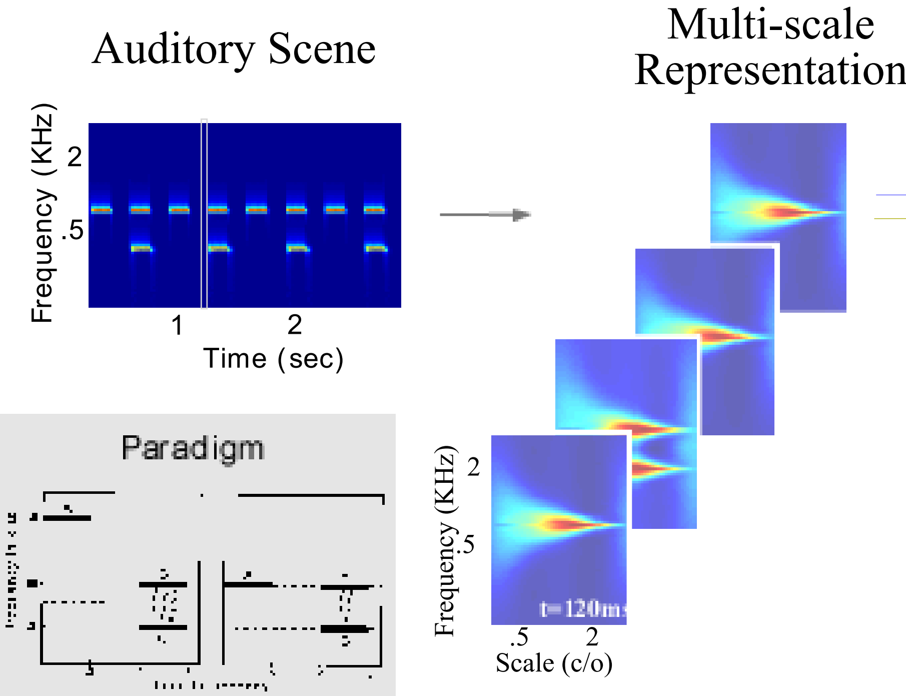


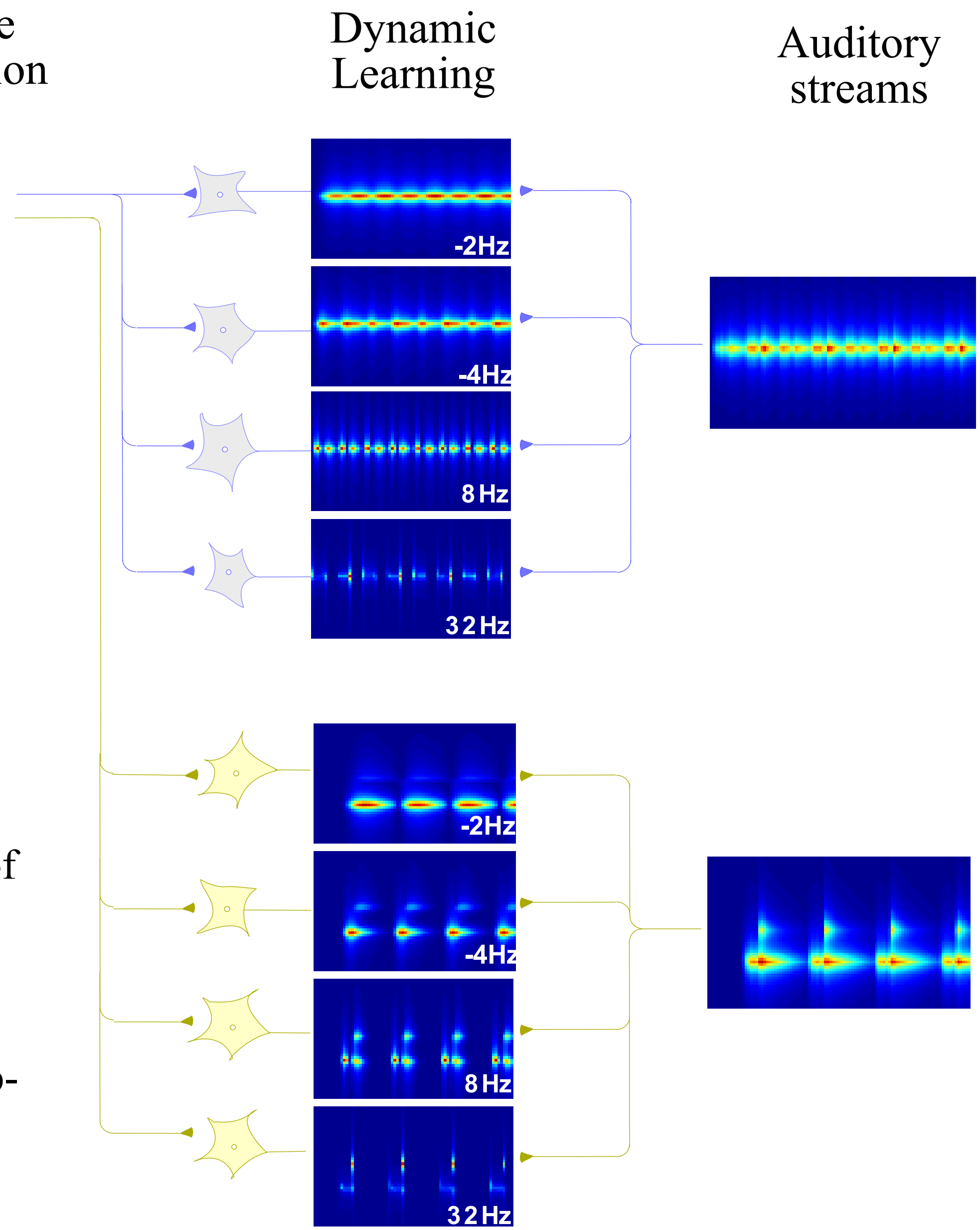
* When alternating a tone 'A' with a complex 'BC', the occurrence of tones B and C together delivers a strong onset cue, hinting that these two elements should be grouped together, separately from stream 'A'.

Center for Auditory
and Acoustic Research

Institute for Systems Research
University of Maryland

Tone in a Mixture (cont'd)



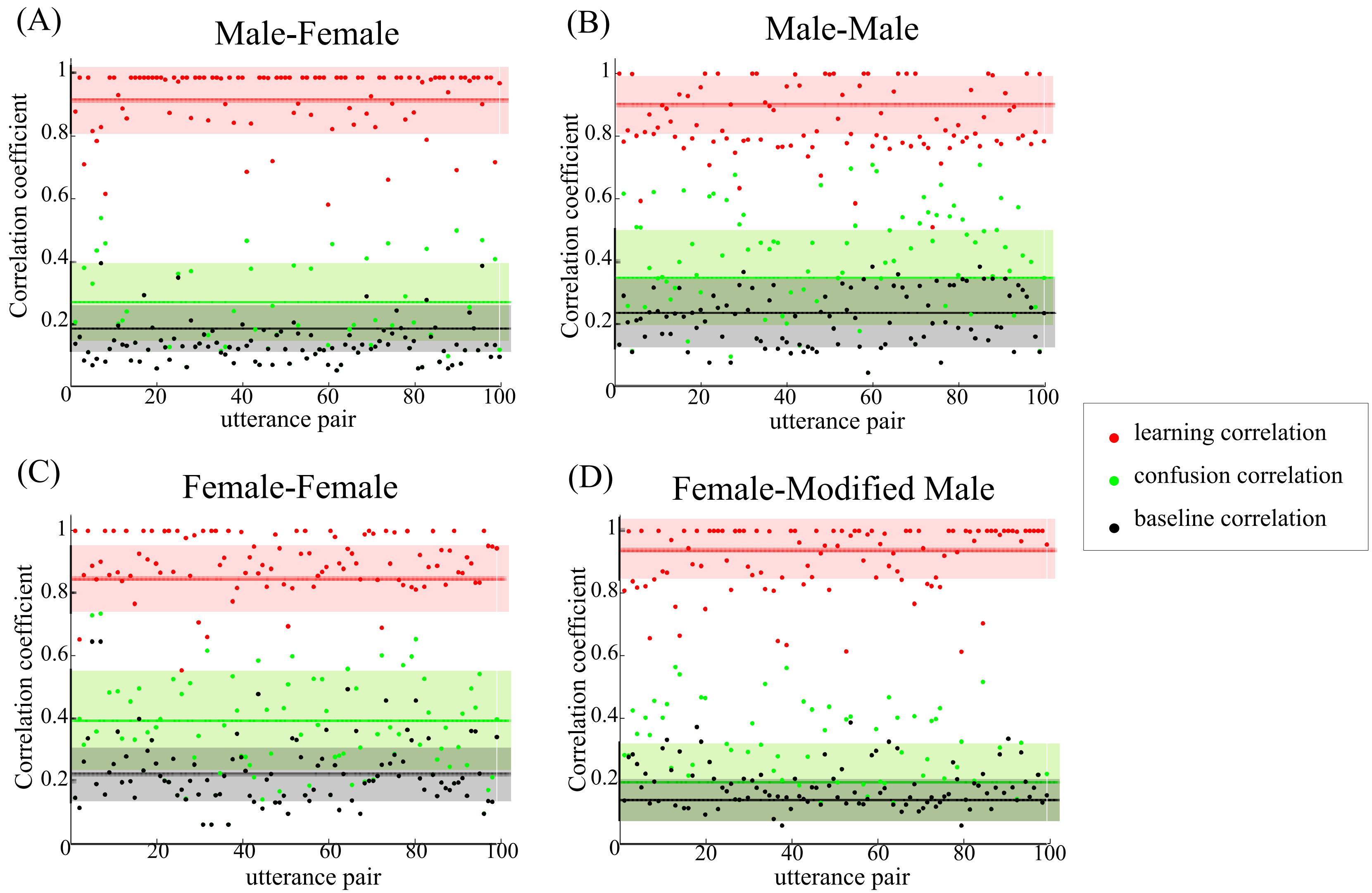


* When the frequency of tone 'A' is equal to that of tone 'B', the organization of the streams now follows Bregman's *old-plus-new heuristic* where new patterns are separated into 'old' sound elements (sound 'B' is a continuation of an old pattern 'A'), plus a new element 'C' which segregates into a separate stream.

Center for Auditory and Acoustic Research

Institute for Systems Research
University of Maryland

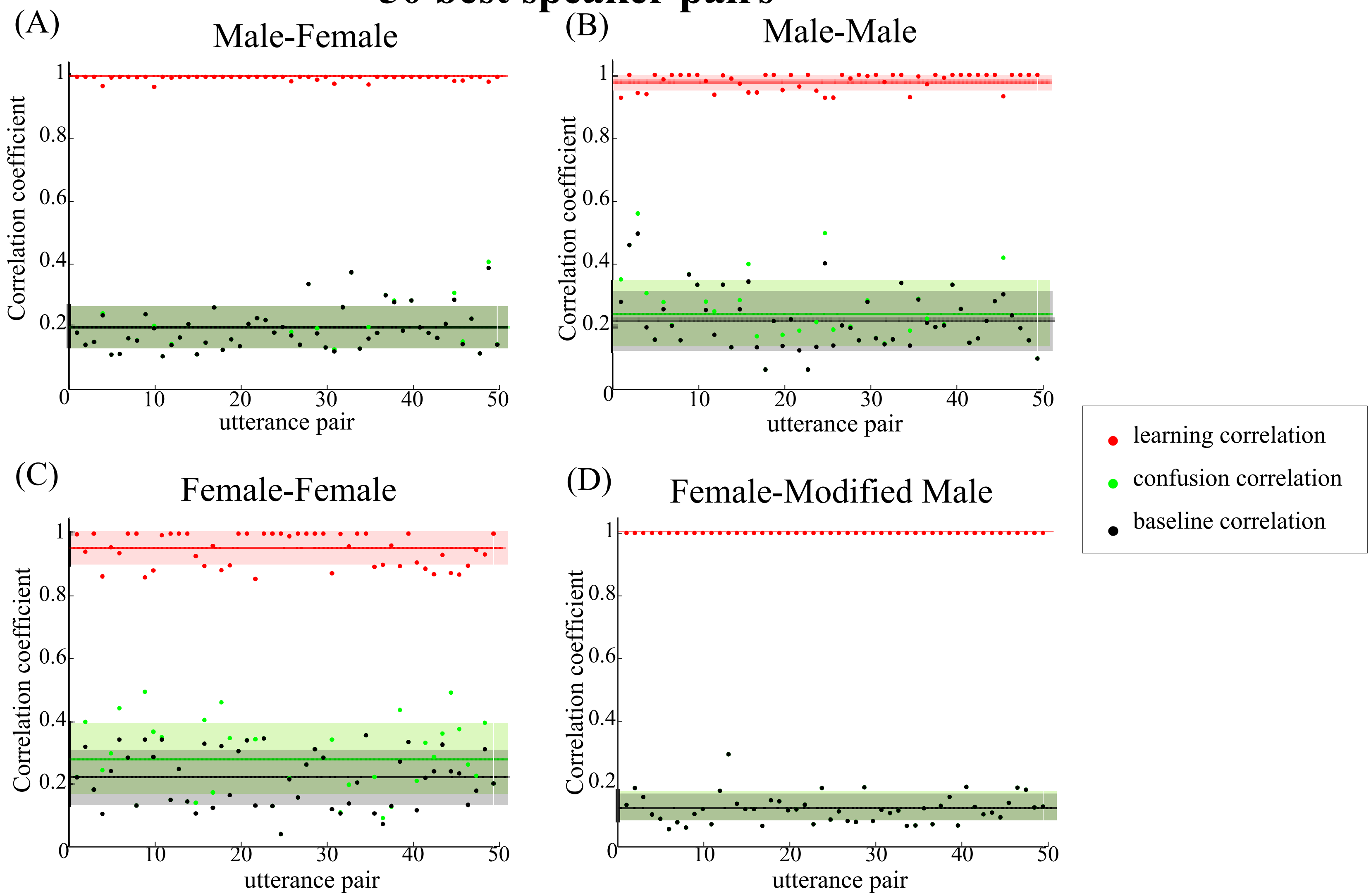
Speech Segregation (with original utterances)



* A pair of sentences (1-3 sec) from two speakers are analyzed and mapped into a multi-scale representation. The features extracted from both speaker are then combined in an array of sound patterns, with no reference to which speaker they belong to. They are then clustered using the adaptive learning model.

Speech Segregation (with original utterances)

- 50 best speaker pairs -

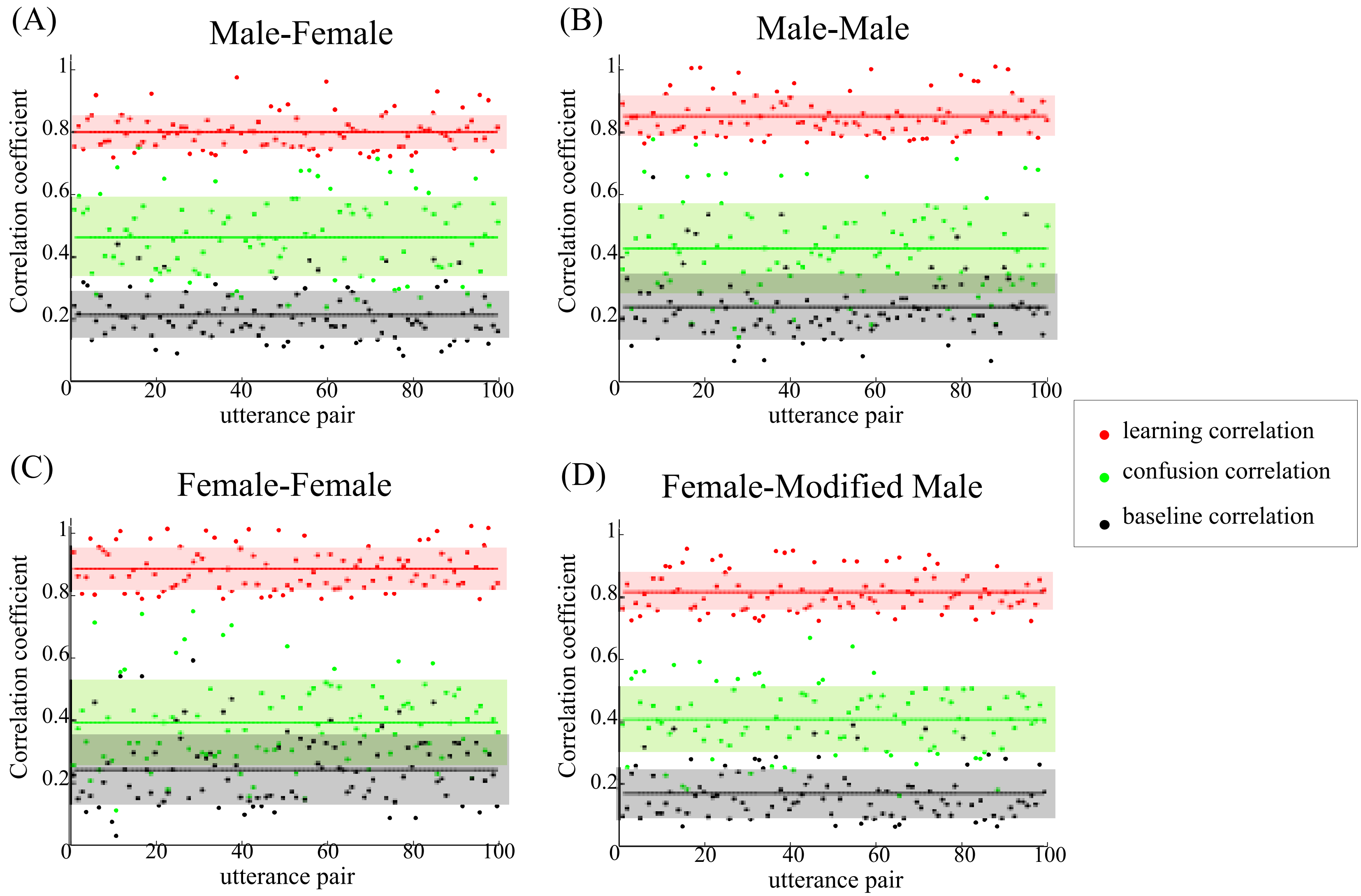


* Segregation results are quantified by a correlation coefficient between the learned (A’) and original sentence (A). The baseline correlation is computed between the original sentences (A and B), while the confusion correlation relates the learned sentence (A’) to the utterance from the other speaker (B).

Center for Auditory and Acoustic Research

Institute for Systems Research
University of Maryland

Speech Segregation (from mixture)



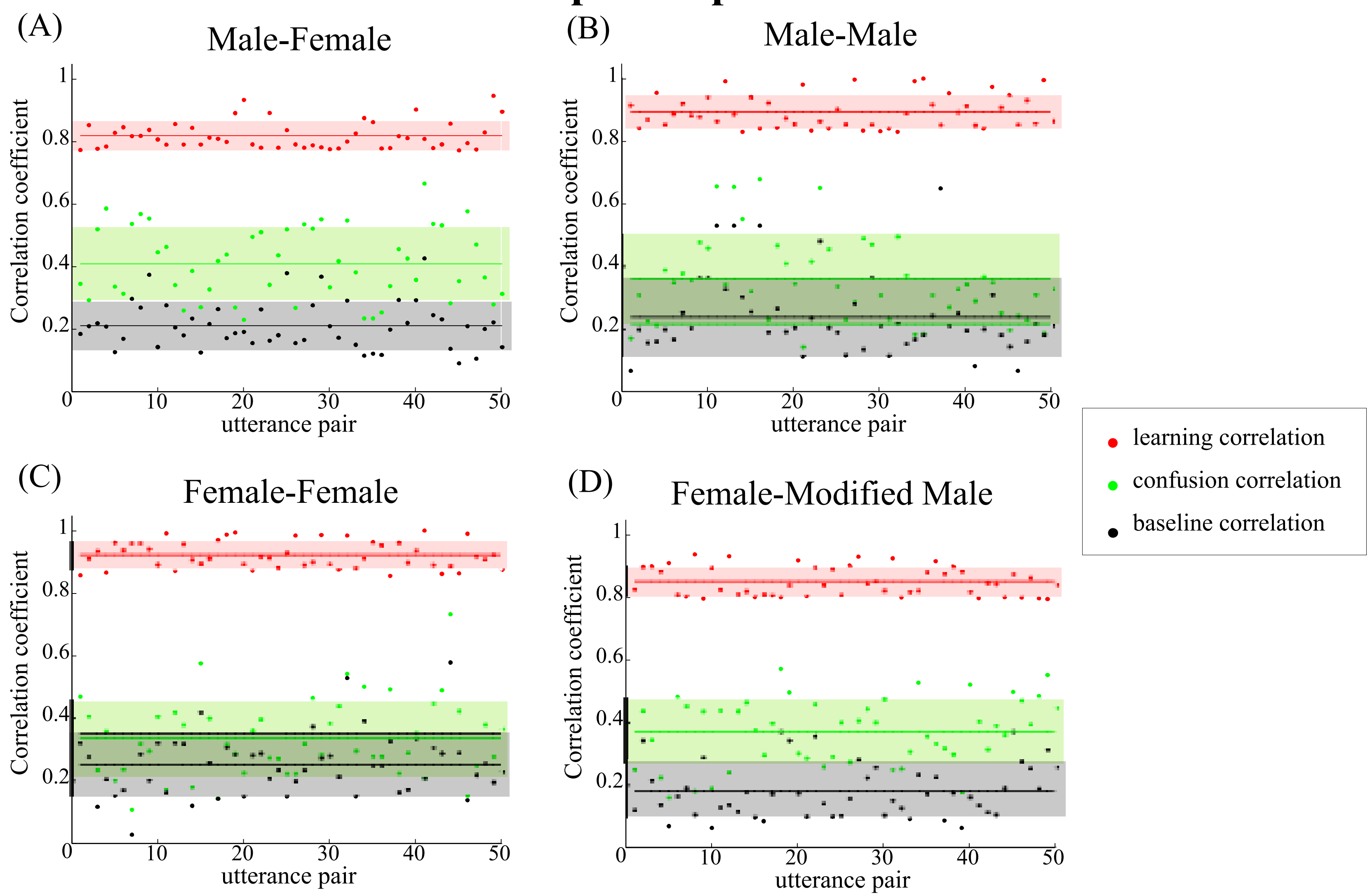
* Sound mixtures are analyzed using a set of pitch and onset cues extracted from the mixture spectrogram, and mapped onto a multi-scale representation.

Center for Auditory and Acoustic Research

Institute for Systems Research
University of Maryland

Speech Segregation (from mixture)

- 50 best speaker pairs -



* While not as high as the values obtained with the original sentences, the correlation coefficients in this test indicate a relatively successful performance of the adaptive learning model in segregating concurrent speakers.

Center for Auditory and Acoustic Research

Institute for Systems Research
University of Maryland

Conclusions

* We develop and test a cortical model for sound organization based on adaptive learning and Kalman estimation. The model is founded on perceptual principles of auditory grouping and stream formation. Such principles are translated into a computational scheme that combines aspects of bottom-up sound processing with an internal representation of the world, which adapts its intrinsic representation based on the residual error between its own predictions and the actual sensory input.

* The model is extremely valuable in exploring various aspects of sound organization in the brain, allowing us to gain interesting insight into the neural basis of auditory scene analysis.

References

* A. S. Bregman. Auditory scene analysis: The perceptual organization of sound. MIT Press, 1990.
* A. S. Bregman and P. A. Ahad. Demonstrations of auditory scene analysis: The perceptual organization of sound. Compact Disk, Department of Psychology, McGill University.
* D. S. G. Pollock. A handbook of time-series analysis, signal processing and dynamics. Academic Press, 1999.
* X. Yang, K. Wang, and S. A. Shamma. Auditory representations of acoustic signals. IEEE transactions on information theory, 38(2):824-839, 1992.