Buenos Aires – 5 to 9 September, 2016
Acoustics for the 21st Century…

PROCEEDINGS of the 22nd International Congress on Acoustics

Communication Acoustics: Paper 299

Spatial modulation: Hearing the environment

Pierre Divenyi[a)]

[(a)] Center for Computer Research in Music and Acoustics, Stanford University, United States,
pdivenyi@ccrma.stanford.edu

Abstract

Auditory processing of complex sources, after an initial peripheral spectro-temporal stage, is thought to have a more central stage identify in the output time segments and frequency regions of higher activity by way of a temporal and spectral modulation analysis. Such analysis broadens the view on perception, both that of complex signals and of auditory scene analysis (ASA). When resolution of temporal and spectral modulations is adequate, the auditory system can decode complex signals and separate simultaneous sources in a scene. Although research in the modulation domain has uncovered important properties of the central (cortical) mechanism active in such analysis, so far it has bypassed the spatial dimension. The present study proposes to include spatial modulation in the horizontal plane into this mechanism. The signal emanating from multiple and diverse sources at different azimuths will first undergo peripheral binaural processing using known methods, consisting of frequency analysis, phase-compensated rectification, left-right cross-correlation, straightening, and weighted frequency integration. The output will represent azimuthal activity between -π and +π radians as a function of time. This analysis stage will be followed by the modulation analysis stage: convolution of the magnitudes, across the azimuth activity axis, with a kernel function that signifies resolution of nearby simultaneous sources. Results of the spatial modulation analysis will be shown as a function of the same input frequency analyzed and put through a stage of temporal modulation processing. Spatial and temporal modulation analysis results viewed side-by side will predict the temporal fluctuation rate and spatial source density at which perception of multiple sources should be optimal.

Keywords: Spatial hearing, multiple sound sources, auditory modeling

Spatial modulation: Hearing the environment

1 Introduction

Our awareness of the world entails our noting the presence, the properties, and the location of objects around us. According to Wundt [1], such awareness is essential for identifying these objects, and for their *aperception*, that is, on the role the objects have on us and how we should react to our having become aware of their presence. Properties of the objects are numerous but they can be classified with regard to the three essential dimensions of their occurrence: the *what*, the *when*, and the *where*. Although these three descriptors are easily explained when thinking of real-world objects and creatures most easily described in the visual domain, life forces us to characterize along the same three dimension acoustic objects, too. The interesting peculiarity of sound objects is that the "what" in them is essentially temporal – waves, vibrations. However, as these objects enter our awareness through the auditory system, temporal units are created to carry the "when" dimension, units the duration range of which is species-dependent (about syllabic for humans). The spectral dimension emerges as the main carrier of the "what," together with temporal structuration over micro-temporal intervals of durations below that of the "where" units. And, of course, there is the "where" dimension: the location of the source of the sound objects.

The situation, however, becomes more complicated when, in our complex world, we are made aware of the simultaneous presence of multiple objects, requiring us to be capable of separating them, in order to recognize at least those in the package that are crucial for making decisions. As far as sound objects are concerned, we are painfully aware of the difficulty to experience the "cocktail-party" situation: understanding the speech of a target talker amidst the brouhaha of other people talking simultaneously. To find a solution to this problem, at least in the "what" and "when" dimensions – on the spectro-temporal plane – it was found easier to deal with after transferring it into the modulation domain. Working in this domain seemed to be advantageous, first because this offered a way to gain better understanding of the central physiology of audition [2, 3], and then because the spectro-temporal receptive fields (STRFs) uncovered by spectral-temporal modulation analysis allowed modeling perception beyond the initial sensory processes [4]. However, what so far does not seem to have been explored is auditory spatial sensitivity to multisource acoustic displays analyzed in the domain of spatial modulation. In this first attempt to fill this gap, space will be represented only in the horizontal plane and expressed in terms of interaural time delays.

2 Methods

Spatial displays were created by taking four recorded or synthesized complex sounds, the waveforms and the auditory spectrograms ("cochleagrams," [5] of which are shown in Figure 1. The four sounds were placed at one of four arbitrarily selected azimuths on the 360° horizontal plane. Placement of the sounds on the azimuth was accomplished using the CIPIC database's head-related impulse responses (HRIR) measured on the KEMAR mannequin in the large-pinna configuration [6]. The KEMAR dataset includes azimuthal measurements at 10° intervals around the head. Four combinations of the four sounds-four azimuths were generated for use in the subsequent analyses. The KEMAR measurement layout and the four selected azimuths are shown in Figure 2.

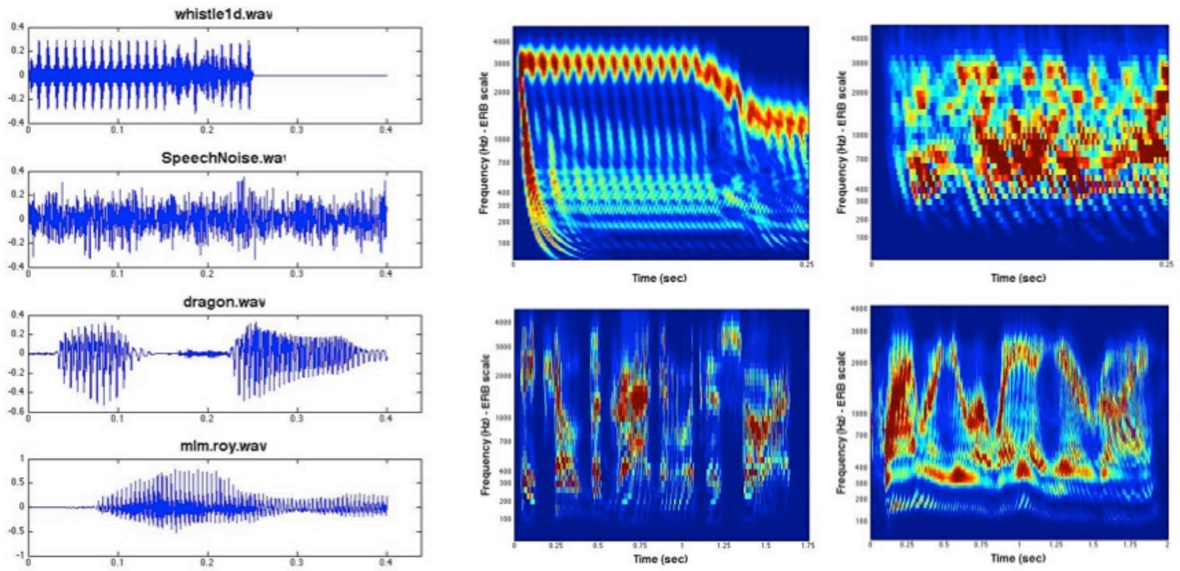

Figure 1: Time waveforms (WF, left panel) and cochleagrams (CG, right panel) of the four signals to be placed in the four azimuthal locations shown in Figure 2. Signal 1 (top WF and top right CG): a synthesized male /ui/ diphthong; Signal 2 (second WF and top right CG): synthesized speech-spectrum random noise; Signal 3 (third WF and bottom left CG): sentence "He killed the dragon with his sword." spoken by a male talker; Signal 4 (bottom WF and bottom right CG): sentence "Why were you away a year ago, Roy?" spoken by a female talker.

For the computational work two models were used. First, lateralization of the sounds in each ensemble was established by a version of the weighted-image model proposed by Trahiotis and Stern [7]. This model, just like the peripheral auditory system, first performs a frequency analysis and then finds the interaural time differences (ITDs) across frequencies using the cross-correlation method. Because the ITDs are a function of the frequency (shorter at high than at low frequencies), the ITD-frequency function is curved. A further stage of the model straightens these functions and a final stage, after assigning a higher weight to locations at the center where localization is the most accurate [8], centralizes the ITD. Because the input in the present study is a mixture emanating from four locations, the unweighted ITD-frequency functions obtained from this model were retained. Although ITD information is of little use beyond 1.5 kHz, which is also the lower limit at which interaural level difference (ILD) gains increasing importance, in the present study no ILD was added to complement ITD localization because of the well-known nonmonotonic relation of ILD and azimuth due to head diffraction and refraction [9].

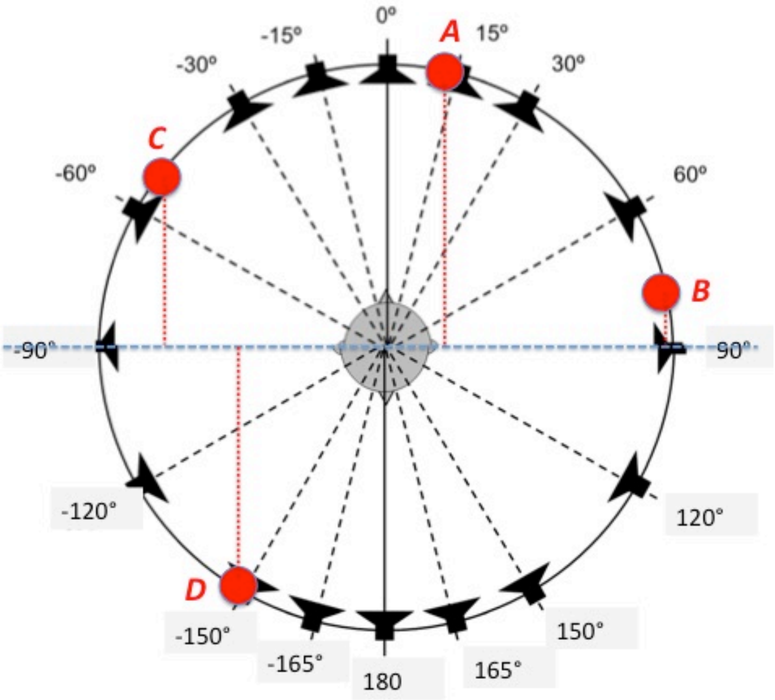

Figure 2: Schematic diagram of the recording of the KEMAR head HRIR dataset. The head was placed on a controllable turntable the center of which was exactly under the midpoint of the line between the two microphones. The four red filled circles represent the azimuthal locations used in the present study. The blue dashed line represents the interaural time difference (ITD) axis onto which those corresponding to the four selected azimuths project. Note that although the ITD scale (in µs or ms) is a monotonic function of the distance away from the ear, the function is not linear.

To compute spatial modulation, a version of the STRF model of Shamma's and his colleagues was adopted [2]. The input to this model is also the peripheral time-frequency analysis illustrated in the cochleagram display in Figure 1. However, the model addresses the question of just how well more central auditory stages can resolve the time-frequency information it receives from the peripheral stage. To answer this question, the model performs a short-term Fourier analysis using a time-frequency window of a given size and slides it across the cochleagram in a slanted direction, generating either upward or downward moving ripples. The operation is akin to dynamically modulating both the spectral and temporal layouts of the original time-frequency display, at a resolution that changes as the window size is changed by factors of 2, along the time or the frequency dimension, or both, as originally proposed by Gabor [10, 11]. This model will be used to estimate sensitivity to spectro-spatial modulation by having it process the output of the first, the weighted-image model.

3 Results

3.1 ITDs of the four-component signals

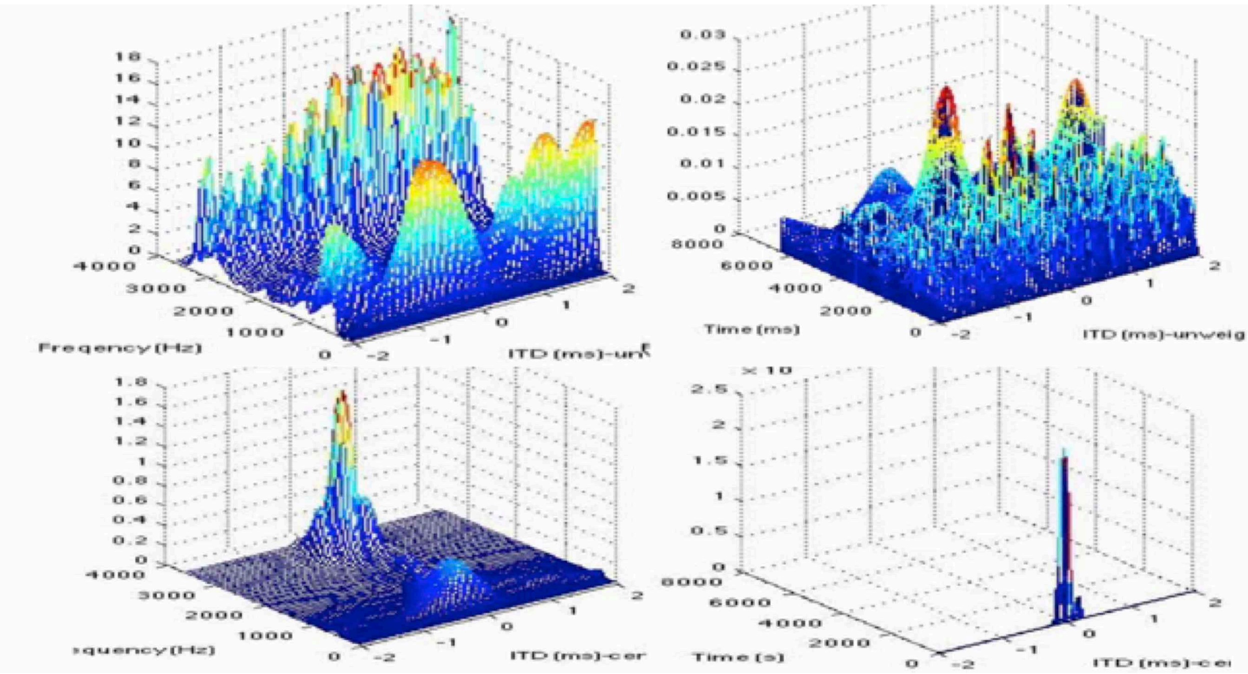

Figure 3: In the left column of the 3D plots ITD weighted images are displayed for one of the four-sound ensembles, with the x-axis indicating estimated ITD (negative means left), the y-axis frequency, and the z-axis relative magnitude. On top, the unweighted and on the bottom the straightened-centered estimates are plotted. The right 3D column shows the same weighted and centered ITD estimates for over 400 ms of the signal, the y-axis indicating the time line of the composite sound. The centered estimate on the bottom right figure shows that for the temporal mean. On the right, unweighted (left) and straightened-centered ITD estimates of a 500- and a 600-Hz warbling sinusoidal pair is shown as calculated by the Trahiotis-Stern model.

First, ITDs for the four-component four-azimuth sounds were obtained using the Trahiotis-Stern model. The result of this operation is shown in Figure 3 for two unweighted ITD estimates showing, as it should, a number of ITDs for the four-ITD complex. For a comparison, straightened-centered ITD estimates of the same configurations are also included. Both the frequency-domain estimate and the time-domain estimate suggests that several lateral location candidates (three for sure) exist for this composite multi-azimuth signal. The model was also tested on a two-component random-phase sinusoid, with the results also included in Figure 5.

3.2 Traditional STRF analysis of the four-component signals

In order to have a traditional spectral-temporal modulation look at the four-component sound, STRFs were computed for one of these ensembles. Their spectro-temporal modulation profiles are shown in Figure 4's two leftmost columns for the 4-Hz and in the two rightmost columns for the 8-Hz temporal modulations, displaying profiles for four spectral modulations (1, ½, ¼, and

1/8 octave). The figures suggest the presence of multiple sound objects, both as temporal and as spectral entities.

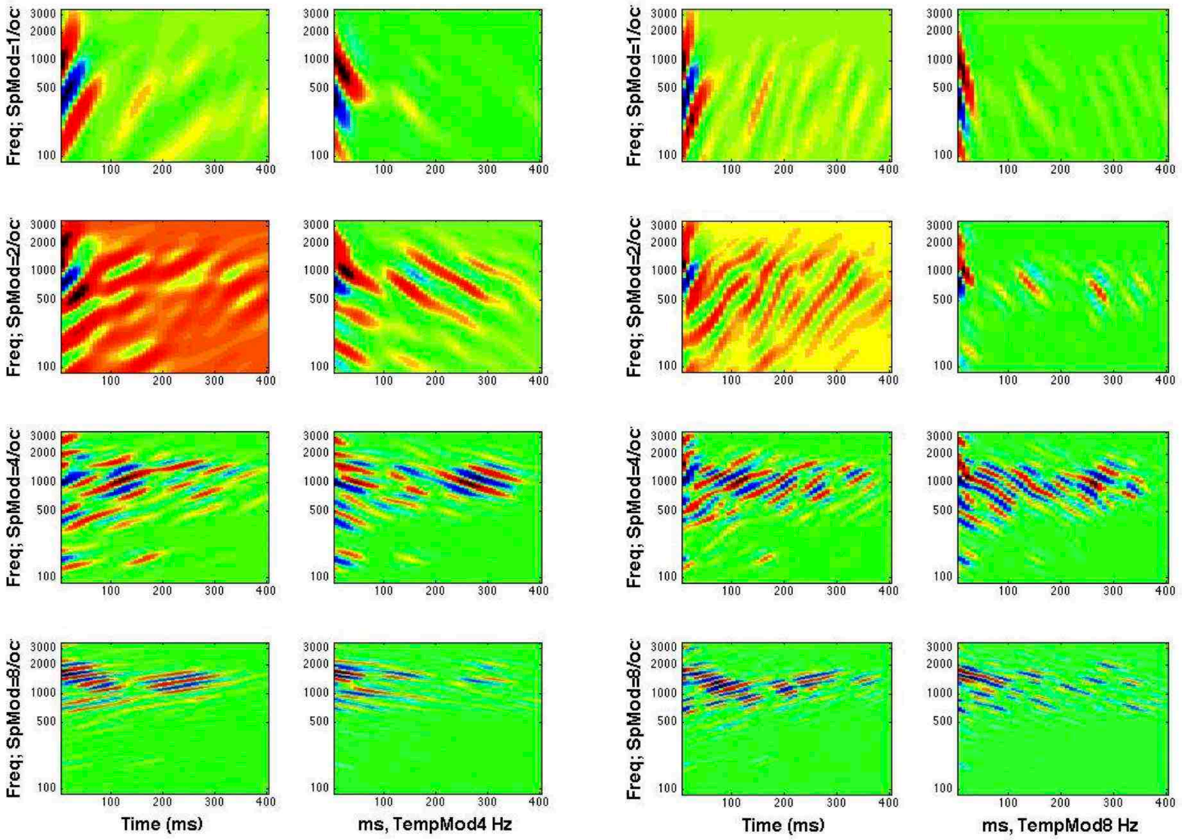

Figure 4: Spectro-temporal receptive field (STRF) analysis of one four-sound ensemble for four spectral modulations (on the ordinate) and two temporal modulations, the two left columns showing upward and downward ripples of a 4-Hz, and the two right most columns of an 8-Hz modulation.

3.3 Spatial STRF analysis

When the time axis of a cochleagram is replaced by the ITDs calculated by the weighted-image model, spatial STRF spectro-temporal fields are obtained. Although ITD-frequency maps are also obtained by the Trahiotis-Stern model – as shown in the 3D plots of Figure 3 and the 2D plot below – the STRF analysis offers the additional feature of being able to observe the degree to which multiple objects are distinct either in their spectra, their location, or both. Figure 5a illustrates the way the weighted-image model deals with the simple two-object ensemble of a 500- and a 600-Hz tone with random phase fluctuations passed through a 400-Hz wide bandpass filter. Figure 5b shows the unweighted image of the same two-object input processed by the STRF model (shown only at two temporal modulation rates for sake of economy).

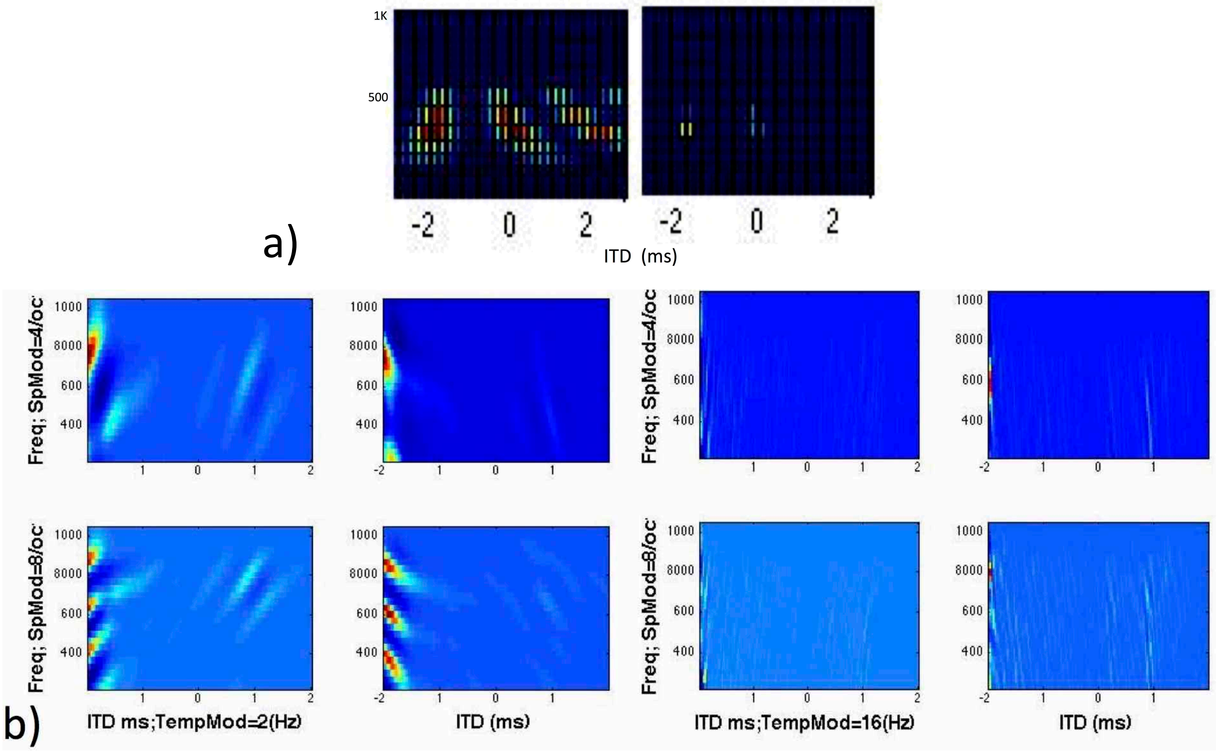

Figure 5: Panel a): Frequency (ordinate) and ITD (abscissa) plot of the estimated ITDs of a 500- and 600-Hz tone pair; unweighted estimates on the left and straightened/centered estimates on the right. **Panel b):** STRF analysis of the unweighted weighted-image model frequency-ITD estimate results. Among the 4 temporal modulation rate – spectral modulation upward and downward moving ripple pairs only two are shown here: ¼th and 1/8th octave ripples at the slowest (2-Hz) and the fastest (16-Hz) modulation rates, chosen because many of them clearly indicate the presence of more than one object.

Of more interest is the STRF analysis of the weighted-image output of the four-sound ensembles. Results of that analysis are shown in Figure 6, again incompletely displayed for reasons of space economy. In that figure upward-downward ripples of the same spectral width and same temporal modulation rate were computed with Gabor kernels of two time-frequency sizes: on the left a kernel considered “normal” by Shamma and his group, and a kernel twice that size. It was expected that the latter would produce poorer spectral and temporal resolution than the former, and therefore would not be as efficient to identify distinct auditory objects. In fact, an earlier piece of research demonstrated that such broadening of the Gabor kernel can account for the decreased ability of elderly listeners to separate frequency-modulated targets from distractors [12].

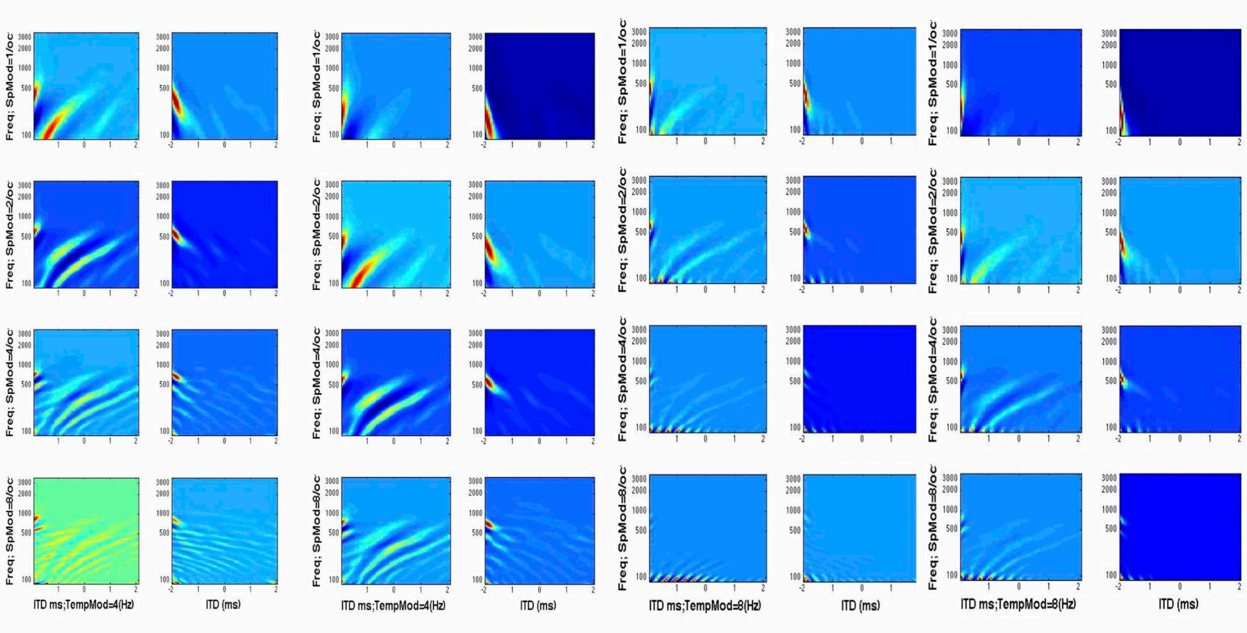

Figure 6. STRF receptive field functions of one four-sound ensemble first processed by the weighted image model to obtain auditory-analogue estimates of ITDs across the frequency range. Each row of figures presents different spectral modulations, 1-, ½-, ¼-,and 1/8-octave from the top to the bottom. Upward and downward sweeping modulation ripples alternate across the eight columns. The temporal modulation rate is 4 Hz in the left four figure columns,\ and 8 Hz in the right four figure columns. The second pair of columns in both modulation rates was processed with a Gabor kernel (both time and frequency) twice as broad as that with which the first pair was processed; the finer resolution of the ripples on the left pairs of the columns than those on the right pairs is clearly visible.

4 Discussion and Conclusions

The work presented embodies the first attempt to express resolution of auditory objects in space by way of a transform into the modulation domain. As the last figure shows, this resolution is controlled by a kernel function performing the transform and will affect the degree to which auditory objects can be recognized as distinct entities. Future research should extend spatial resolution on the horizontal plane to also include elevation, should integrate spectral, temporal, and spatial resolution of auditory objects, and should compare theoretical resolution functions with perceptual data collected using listeners.

Acknowledgments

The author wishes to acknowledge the gracious assistance offered by Drs. Richard M. Stern and Clayton D. Rothwell on the use of the weighted-image model. Support for the presentation of this paper was provided by the Center for Computer Research in Music and Acoustics of Stanford University.

References

[1] Wundt, W.M., *Grundzüge der physiologischen Psychologie*. 1874, Leipzig: W. Engelmann.

[2] Chi, T., et al., *Spectro-temporal modulation transfer functions and speech intelligibility.* J Acoust Soc Amer, 1999. **106**(5): p. 2719-2732.

[3] Depireux, D.A., et al., *Spectro-Temporal Response Field Characterization With Dynamic Ripples in Ferret Primary Auditory Cortex.* J. Neurophysiol., 2001. **85**(3): p. 1220-1234.

[4] Elhilali, M., T. Chi, and S.A. Shamma, *A spectro-temporal modulation index (STMI) for assessment of speech intelligibility.* Speech Communication, 2003. **41**(2–3): p. 331-348.

[5] Lyon, R.F. *A computational model of binaural localization and separation*. in *Proceedings of IEEE ICASSP*. 1983.

[6] Algazi, V.R. and R.O. Duda, *The CIPIC HRTF Database*, D. University of California, Editor 2004, Center for Image Processing and Integrated Computing, University of California, Davis: Davis, CA.

[7] Trahiotis, C. and R.M. Stern, *Lateralization of bands of noise: Effects of bandwidth and differences of interaural time and phase.* J Acoust Soc Amer, 1989. **86**(4): p. 1285-1293.

[8] Mills, J.W., *On the minimum audible angle.* J Acoust Soc Amer, 1958. **30**: p. 237-246.

[9] Macaulay, E.J., W.M. Hartmann, and B. Rakerd, *The acoustical bright spot and mislocalization of tones by human listeners.* J Acoust Soc Amer, 2010. **127**(3): p. 1440-1449.

[10] Zibulski, M. and Y. Zeevi, *Discrete multiwindow Gabor-type transforms.* IEEE Trans. Sig. Proc., 1997. **45**(6): p. 1428-1442.

[11] Wolfe, P.J., S.J. Godsill, and M. Dorfler, *Multi-Gabor dictionaries for audio time-frequency analysis.* Applications of Signal Processing to Audio and Acoustics, 2001 IEEE Workshop on the, 2001: p. 43-46.

[12] Divenyi, P., *Decreased ability in the segregation of dynamically changing vowel-analog streams: a factor in the age-related cocktail-party deficit?* Front Neurosci, 2014. **8**(144).

1. Wundt, W.M., *Grundzuege der physiologischen Psychologie*. 1874, Leipzig: W. Engelmann.
2. Chi, T., et al., *Spectro-temporal modulation transfer functions and speech intelligibility.* The Journal of the Acoustical Society of America, 1999. **106**(5): p. 2719-2732.
3. Depireux, D.A., et al., *Spectro-Temporal Response Field Characterization With Dynamic Ripples in Ferret Primary Auditory Cortex.* J. Neurophysiol., 2001. **85**(3): p. 1220-1234.
4. Elhilali, M., T. Chi, and S.A. Shamma, *A spectro-temporal modulation index (STMI) for assessment of speech intelligibility.* Speech Comm., 2003. **41**(2–3): p. 331-348.
5. Lyon, R.F. *A computational model of binaural localization and separation*. in *Proceedings of IEEE ICASSP*. 1983.
6. Algazi, V.R. and R.O. Duda, *The CIPIC HRTF Database*, D. University of California, Editor 2004, Center for Image Processing and Integrated Computing, University of California, Davis: Davis, CA.
7. Trahiotis, C. and R.M. Stern, *Lateralization of bands of noise: Effects of bandwidth and differences of interaural time and phase.* J Acoust Soc Amer, 1989. **86**(4): p. 1285-1293.
8. Mills, J.W., *On the minimum audible angle.* Journal of the Acoustical Society of America, 1958. **30**: p. 237-246.
9. Macaulay, E.J., W.M. Hartmann, and B. Rakerd, *The acoustical bright spot and mislocalization of tones by human listeners.* J Acoust Soc Am, 2010. **127**(3): p. 1440-1449.
10. Zibulski, M. and Y. Zeevi, *Discrete multiwindow Gabor-type transforms.* IEEE Trans. Sig. Proc., 1997. **45**(6): p. 1428-1442.
11. Wolfe, P.J., S.J. Godsill, and M. Dorfler, *Multi-Gabor dictionaries for audio time-frequency analysis.* Applications of Signal Processing to Audio and Acoustics, 2001 IEEE Workshop on the, 2001: p. 43-46.
12. Divenyi, P., *Decreased ability in the segregation of dynamically changing vowel-analog streams: a factor in the age-related cocktail-party deficit?* Front Neurosci, 2014. **8**(144).