Commercializing Auditory Neuroscience

Lloyd Watts
Audience, Inc.
Mountain View, California

In a previous paper (Watts, 2003), I have argued that there is sufficient knowledge of auditory brain function, and sufficiently great available compute power, to begin building a realistic real-time model of the human auditory pathway, i.e., to build a machine that could hear like a human being. Full completion of a realistic model could be expected to occur in the 2015-2020 time-frame, based on reasonable extrapolations of computational capacity and advancements in neuroscience and psychoacoustics. This ambitious endeavor will require a large team of specialists, and a network of highly skilled collaborators, operating over another decade to reach its full potential. Attracting and holding such a team requires a substantial depth of financial resources. Over the period 2002-2006, the problem has been one of building the core technology, determining a viable market direction, securing financing, assembling a team, and building and executing a viable and sustainable business model that provides sufficient incentive (expected return on investment) for all participants (investors, customers, employees), in both the short term and the long term. The success of the venture has depended on showing simultaneous short-term progress on all of those synergistic and interdependent fronts, in a way that would plausibly lead to long-term success.

Scientific Foundation

The scientific foundation for the company is a detailed study of the mammalian auditory pathway, shown in Figure 1, which has been undertaken with the active assistance of eight of the world's leading auditory neuroscientists. The early philosophy was to build working, high-resolution real-time models of the various system components, and validate those models with the neuroscientists who had performed the primary research. The basic model-building began in 1998 and continued through 2002, just at the time that personal computers were crossing the 1 GHz mark, which meant that, for the first time in history, it was possible to build working, real-time models of real brain system components, *in software,* on consumer computer platforms. Early demonstrations in 2001-2002 included high-resolution real-time cochlea displays, binaural spatial representations such as Interaural Time Difference (ITD) and Interaural Level Difference (ILD), high-resolution event-based correlograms, and a powerful demonstration of real-time Polyphonic Pitch detection, all based on well-established neuroscience and psychoacoustic findings.

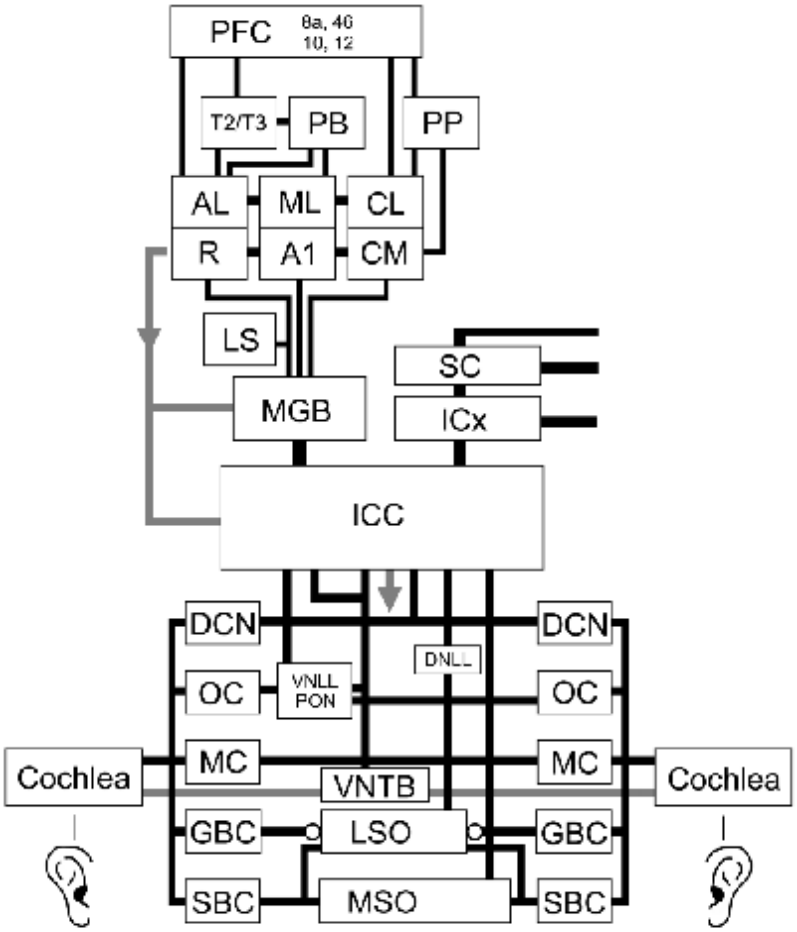

Figure 1: Auditory Pathway (highly simplified). Adapted from Young 1998, Oertel 2002, Casseday et al. 2002, LeDoux 1997, and Rauschecker and Tian 2000.

.

Market Focus and Product Direction

Many avenues for commercialization were explored in the early years of the company. After a two-year sojourn into noise-robust speech recognition from 2002-2004, the market was re-assessed, and it was determined that the company's greatest commercial value was in the extraction and reconstruction of the human voice, and that the technology could be applied to improving the quality of telephone calls made from noisy environments. This insight was driven by the enormous sales volume of the cell-phone market, and the urgent need for cell-phone users to place calls from noisy locations and still be heard clearly ("Can you hear me now? Good."). The speech recognition work was de-emphasized and the company began its focus in earnest on commercializing a two-microphone non-stationary noise suppressor for the mobile telephone market.

Technology

Figure 2 shows a block diagram of Audience's Cognitive Audio System, designed to extract a single voice out of a complex auditory scene.

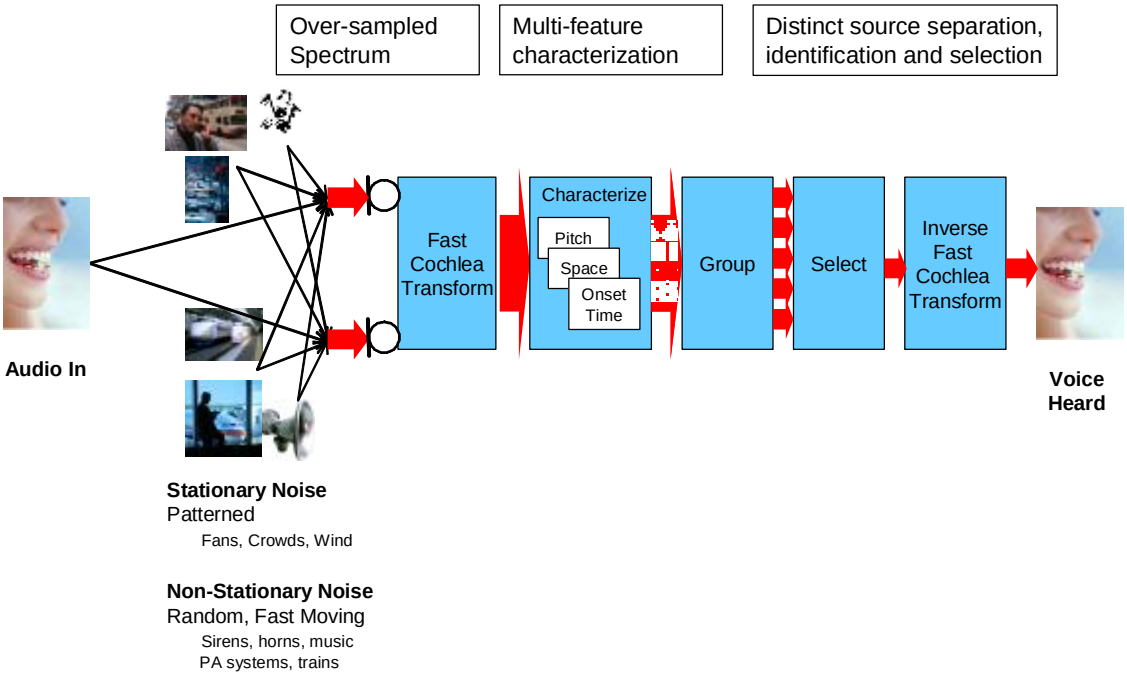

Figure 2: Architecture of the Cognitive Audio System.

The major elements in the system include:

- **Fast Cochlea Transform™ (FCT):** The FCT provides a high-quality spectral representation of the sound mixture, with sufficient resolution and without introducing frame artifacts, to allow the various components of the multiple sound sources to be characterized.
- **Characterization:** In the Characterization block, the attributes of sound components that are used by human beings for grouping and stream separation are computed. These attributes include the pitches of the constituent non-stationary sounds, the spatial location cues (used when multiple microphones are available), onset timing and other transient characteristics, estimation and characterization of quasi-stationary background noise levels, etc. These attributes are then associated with the raw FCT data as acoustic tags which are used in the subsequent Grouping process.
- **Grouping:** The Grouping block performs a type of clustering operation in various low-dimensionality spaces such that sound components with common or similar attributes may be mutually associated into a single auditory stream, and sound components with sufficiently dissimilar attributes are associated with different auditory streams. Ultimately, the streams are tracked through time and associated with persistent or recurring sound sources in the auditory environment.

The output of the Grouping block is the raw FCT data associated with each stream, and the corresponding acoustic tags.

- **Selector:** The Selector block allows the separated auditory sound sources to be prioritized and selected as appropriate for the given application.
- **Inverse Fast Cochlea Transform:** In the telephony applications, the primary output of the system is reconstructed, cleaned-up, high-quality voice. The Inverse Fast Cochlea Transform block converts the FCT data back into digital audio for subsequent processing, including encoding for transmission across a cell-phone channel.

Details of Technical Approach

- **Fast Cochlea Transform™ (FCT):** The first stage of processing must have adequate resolution to support high-quality stream separation:

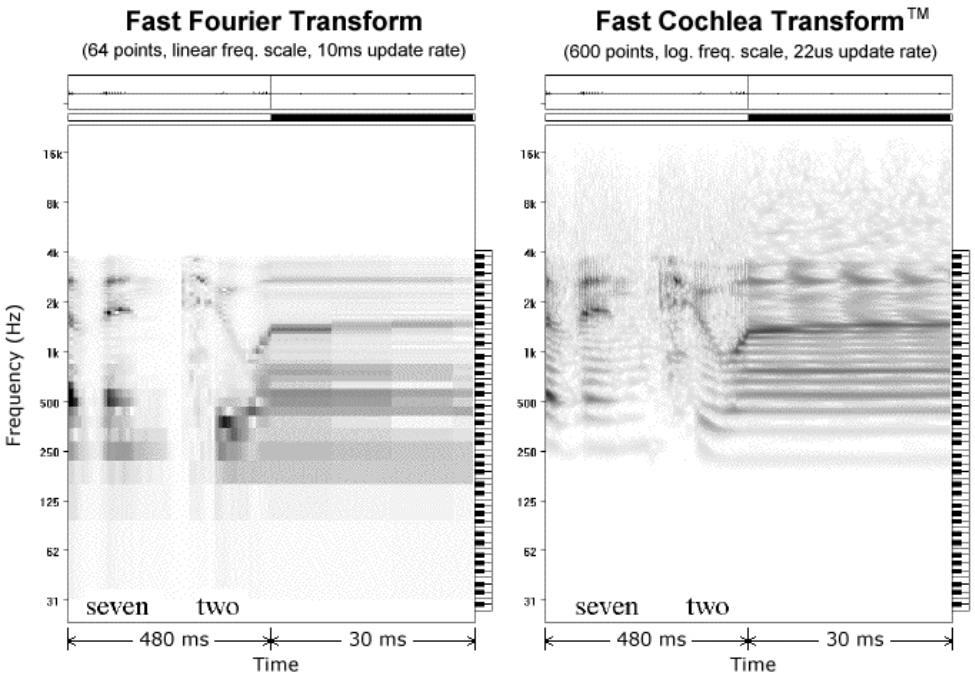

Figure 3. Comparison of Fast Fourier Transform and Audience's Fast Cochlea Transform™.

In Figure 3, a comparison is made between the conventional Fast Fourier Transform (FFT) and Audience's Fast Cochlea Transform™ (FCT). In many applications, the FFT is updated every 10 ms, giving it coarse temporal resolution as seen in the right half of the FFT panel. The FCT is updated every audio sample, which allows resolution of glottal pulses, as necessary to compute periodicity measures on a per-formant basis, as a cue for grouping voice components. Similarly, the FFT is often configured to give poor spectral resolution at low frequencies, since often the following processor (such as a speech recognizer back-end) is only interested in a smooth estimate of spectral envelope. The FCT is designed to give high resolution

information about individual resolved harmonics, so that they may be tracked and used as grouping cues in the lower formants.

The importance of high resolution is even greater in a multi-source example, as shown in Figure 4. In this example, the speech is corrupted by a loud siren. The low spectro-temporal resolution of the frame-based FFT makes it more difficult to resolve and track the siren, and therefore more difficult to remove it from the speech. The high spectro-temporal resolution of the FCT makes it much easier to resolve and track the siren, as distinct from the harmonics of the speech signal, and the boundaries between the two signals are much better defined, to allow high performance in the subsequent grouping and separation steps.

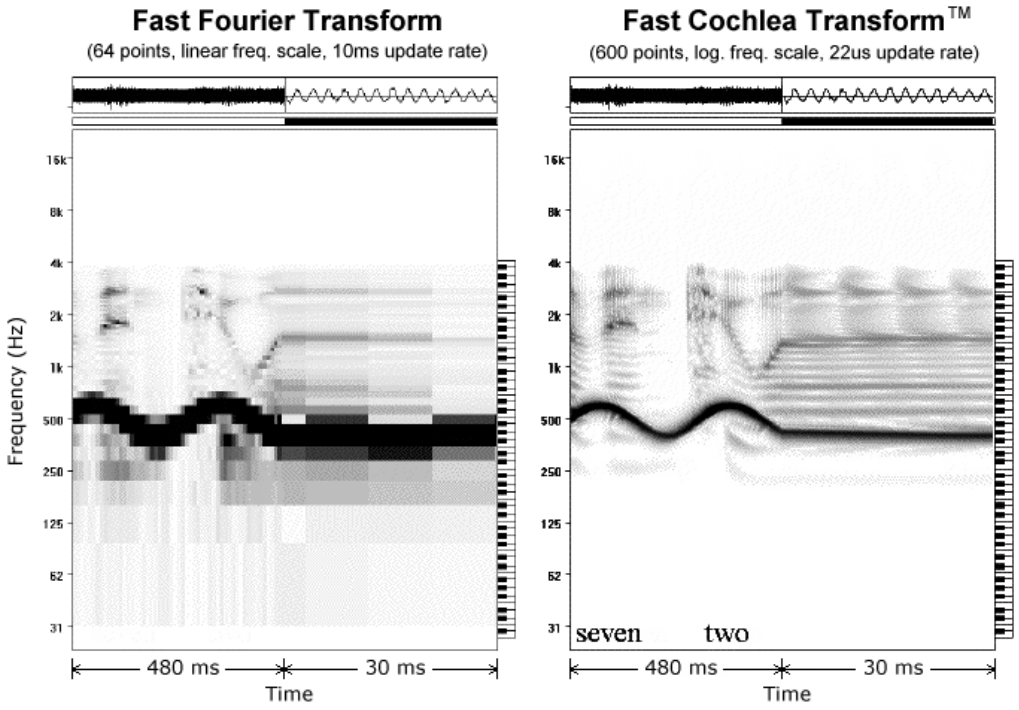

Figure 4. Multi-stream separation demonstration (speech + siren).

Note that the Fast Cochlea Transform creates a redundant, oversampled representation of the time-varying auditory spectrum. We have found this to be necessary to meet the joint requirements of perfect signal reconstruction, with no aliasing artifacts, at low latency, with a high degree of modifiability in both the spectral and temporal domains.

- **Characterization Block – Pitch Extraction.** Audience's Polyphonic Pitch algorithm is capable of resolving the pitch of multiple speakers simultaneously, and detecting multiple musical instruments simultaneously. An example is shown in Figure 5, extracting the simultaneous pitches of a male and female speaker.

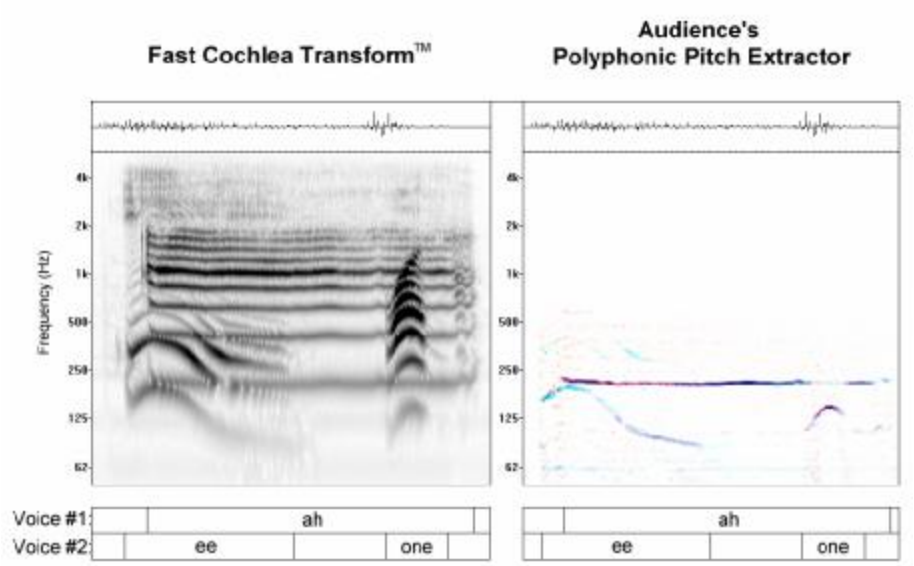

Figure 5. Polyphonic pitch for separating multiple simultaneous voices.

- **Characterization Block – Spatial Localization.** These representations are valuable for stream separation and sound source location, when stereo microphones are available. Figure 6 shows the response of the binaural representations to a sound source positioned to the right of the stereo microphone pair.

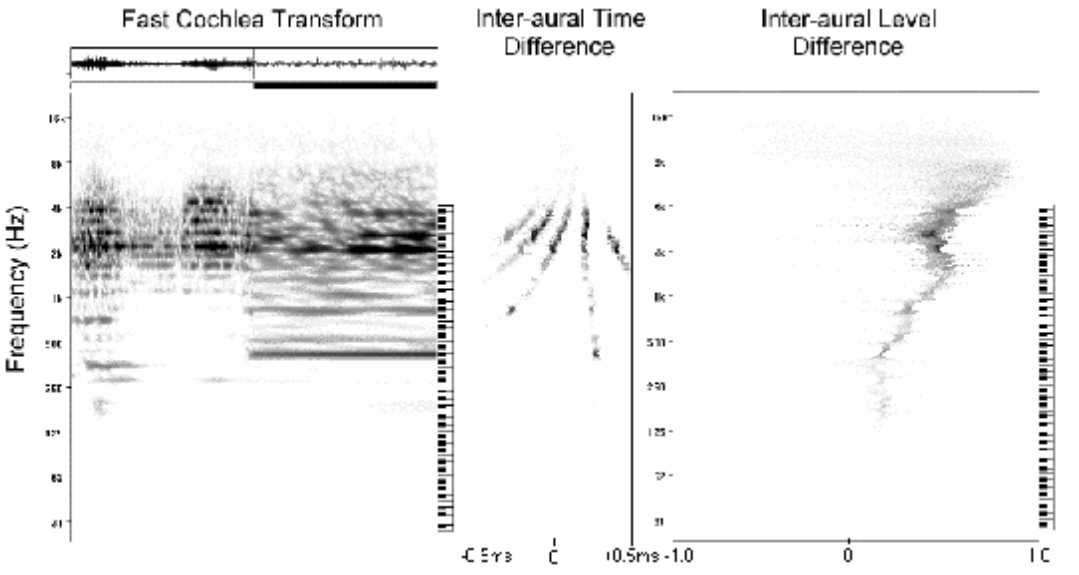

Figure 6. Response of cochlea model, and inter-aural time and level difference computations for spatial localization.

- **Stream Separation**. Figure 7 shows an example of a complex audio mixture (voice recorded on street corner with nearby conversation, passing car noise, and cell-phone

ringing) in the cochlea representation, and then after sound separation, in which only the voice has been preserved.

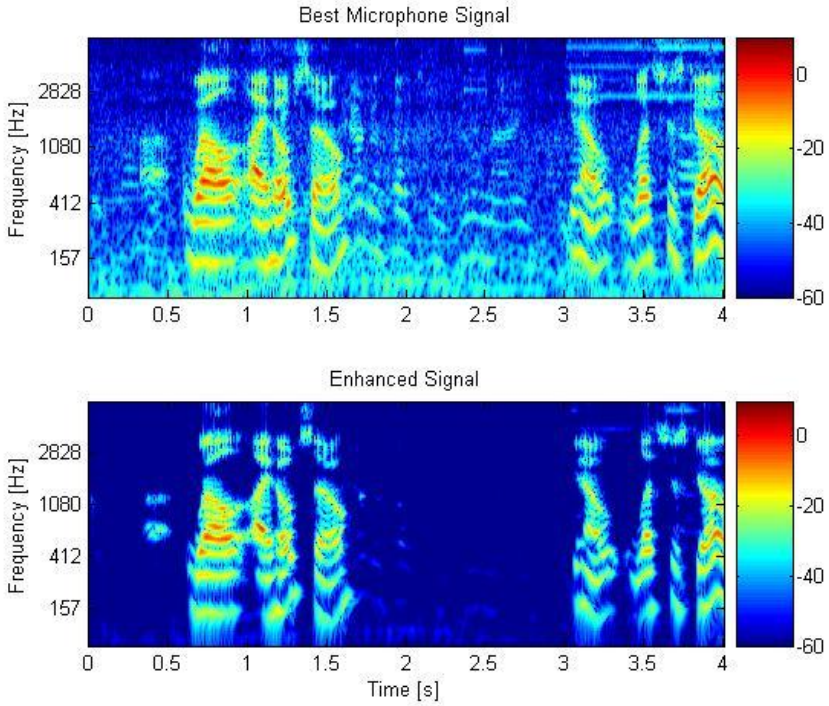

Figure 7. Separation of a voice from a street corner mixture, using Audience's Real-Time Embedded Software Handset Implementation. (Top Panel) Mixture of voice with car noise, other voice, and cell-phone ring-tone. (Bottom Panel) Isolated voice.

- **Audio Reconstruction using the Inverse Fast Cochlea Transform**. After sound separation in the Cochlea (spectral) domain, it is possible to reconstruct the audio waveform for transmission, playback or storage, using Audience's Inverse Fast Cochlea Transform.

So far, the product direction of the company has remained true to the original goal, namely, to create a successful commercial entity by building machines that could hear like a human being. Along the way, we have found points of divergence between what the brain does (computes with spikes, uses slow wetware, does not reconstruct audio) and what our system must do to be commercially viable (computes with conventional digital representations, uses fast silicon hardware, requires an inverse spectral transformation), but in general, the insights that have come from studying the neuroscience and psychoacoustics of hearing have indeed led to insights that have translated into greater signal processing capacity and robustness.

Product Implementation

In the early days of the company, I assumed that it would be necessary to build dedicated hardware (integrated circuits, or silicon chips) to efficiently support the high

compute-load of the brain-like algorithms, and for that reason, I set investor expectations that Audience would be a fabless semiconductor company with a strong intellectual property position (my catch-phrase was "the nVidia of sound input"). In 1998, Paul Allen advised an early focus on the algorithms, while remaining flexible on the implementation technology, since the project was likely to take many years and the implementation technology changes so fast (Allen, 1999). Eight years later in 2006, that counsel continues to serve the company. As we now engage the market with a specific product, we are finding acceptance of both dedicated hardware solutions and embedded software solutions, for reasons that have less to do with computational demands and more to do with the details of integrating our solution into the existing cell-phone platform (lack of mixed-signal support for a second microphone, for example). So, the company ends up being a fabless semiconductor company after all, but for very different reasons than had been expected when the company was founded in the year 2000.

Conclusions

At the turn of the twenty-first century, there is sufficient knowledge of auditory brain function, and sufficiently great available compute power, to begin building a realistic real-time model of the human auditory pathway, i.e., to build a machine that could hear like a human being. This ambitious endeavor will require a large team of specialists, and a network of highly skilled collaborators, to reach its full potential. Attracting and holding such a team requires a substantial depth of financial resources. Over the period 2002-2006, the problem has been one of building the core technology, determining a viable market direction, securing financing, assembling a team, and building and executing a viable and sustainable business model that provides sufficient incentive (expected return on investment) for all participants (investors, customers, employees), in both the short term and the long term. The success of the venture has depended on showing simultaneous short-term progress on all of those synergistic and interdependent fronts, in a way that would plausibly lead to long-term success.

References

Allen, P. (1999), Interval Research internal project review, personal communication.

Bregman, A. (1990), *Auditory Scene Analysis*, MIT Press.

Casseday, J., Fremouw, T., Covey, E. (2002), in Oertel, D., Fay, R., and Popper, A., ed., *Integrative Functions in the Mammalian Auditory Pathway*, Springer-Verlag, New York, pp. 238-318.

Oertel, D. (2002), in Oertel, D., Fay, R., and Popper, A., ed., *Integrative Functions in the Mammalian Auditory Pathway*, Springer-Verlag, New York, pp. 1-5.

Rauschecker, J., and Tian, B. (2000), "Mechanisms and streams for processing of "what" and "where" in auditory cortex", *Proceedings of the National Academy of Sciences*, vol. 97, no. 22, 11800-11806.

Shepherd, G. (1998), *The Synaptic Organization of the Brain*, fourth edition, Oxford University Press, p. vi.

Watts, L. (2003), "Visualizing Complexity in the Brain", in *Computational Intelligence: The Experts Speak,* edited by D. Fogel and C. Robinson, IEEE Press/Wiley, 2003, pp. 45-56.

Young, E. (1998), in Shepherd, G., ed., *The Synaptic Organization of the Brain*, fourth edition, Oxford University Press, pp. 121-158.