Real-Time, High-Resolution Simulation of the Auditory Pathway, with Application to Cell-Phone Noise Reduction

Lloyd Watts
Audience, Inc.
440 Clyde Ave.
Mountain View, CA 94043
lwatts@audience.com

***Abstract*—Recent advances in computing and auditory neuroscience have now made it possible to produce high-resolution, real-time simulations of major portions of the human auditory pathway, including high resolution, real-time models of the cochlea, major cell-types of the cochlear nucleus, lateral and medial superior olivary complex, as well as polyphonic pitch perception based on combination-sensitive cells measured in inferior colliculus and auditory cortex, all running live on a notebook computer. These technologies are the foundation of Audience's Voice Processor chip, which provides advanced two-microphone noise reduction for cell-phones. The design of the simulated system required many careful decisions about what biological details to include, in order to preserve functional biological realism, while meeting the constraints of real-time implementation and commercial product realization.**

I. INTRODUCTION

The field of Machine Hearing is concerned with building artificial systems that can hear, often drawing on our understanding of biological hearing systems for both inspiration and guidance as to how to achieve high performance on any particular hearing task. The human auditory pathway consists of many organs and different brain regions [1], as shown in Fig. 1. It has taken about a century, from the 1890's to 1990's, to develop a thorough understanding of the cochlea alone, including the mechanisms underlying its traveling-wave behavior, compressive nonlinearity, afferent and efferent neural signals, etc. This understanding has led to a number of efficient computational models capable of matching biological measurements. Similarly, the cochlear nucleus (containing the Dorsal Cochlear Nucleus (DCN), Octopus Cells (OC), Multipolar Cells (MC), Globular Bushy Cells (GBC), and Spherical Bushy Cells (SBC) as shown in Fig. 1) has been described by Gordon Shepherd as "one of the best understood regions of the brain" [2]. The basic circuit of the cochlear nucleus is given in [3]. We have indeed been able to build detailed models of each of these cell assemblies capable of matching biological measurements. And finally, the Lateral and Medial Olivary

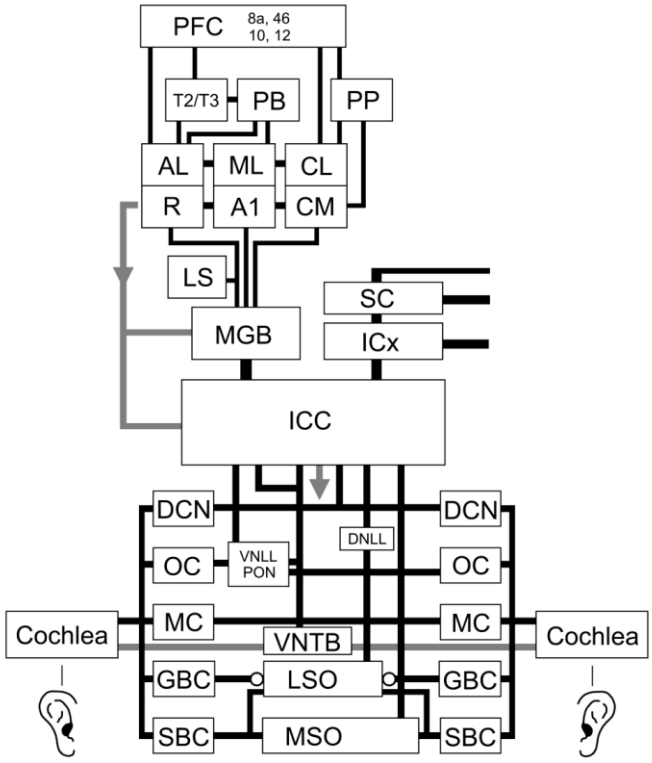

Figure 1. Human Auditory Pathway, highly simplified. [1]

Complexes (LSO and MSO) and related parts of the Central Nucleus of the Inferior Colliculus (ICC), which perform the binaural spatial localization computations, had become sufficiently well-understood by the late 1990's [4][5] that we can build efficient models of them that match the biological data.

The advent of 1GHz processors in about the year 2000 made it possible to run these biologically accurate models in real-time on inexpensive personal computers. And with

978-1-4244-5309-2/10/$26.00 ©2010 IEEE ISCAS 2010, pp. 3821-3824

hardware acceleration on Audience's dedicated low-power Voice Processor chips, it became possible in 2007 to run these sophisticated signal processing algorithms in real-time in a cell-phone, consuming less than 30mW of power. At each of these stages of development, it was necessary to make careful choices about what level of detail and abstraction to maintain, to ensure sufficient biological realism to get the desired hearing performance on the chosen task, while meeting the constraints of the implementation medium and commercial application.

II. BIOLOGICALLY ACCURATE COMPUTATIONAL MODELS

A. Cochlea and Multipolar Cells

The cochlea model used in Audience's high-resolution biological simulations is based on the Pole-Zero Filter Cascade (PZFC) model developed by Lyon [6][7]. With careful tuning, we were able to match the critical bandwidths and latencies of the human hearing system, providing the optimal time-frequency resolution trade-off as a foundation for building a machine hearing system with human-level performance. The system shown in Fig. 2 and Fig. 3 uses 600 filter stages, spanning 20Hz-20KHz, or 10 octaves, with 60 filter stages per octave (5 filters per musical semitone), and runs at CD quality sampling rate of 44.1 kSamples/second.

B. Binaural Hearing System

The model of the binaural system is an efficient cochlea-domain event-based binaural comparator, which captures the essential computations performed by the SBC, GBC, LSO and MSO cell assemblies [8]. The output of this binaural system is shown in Fig. 3.

For both the cochlea/multipolar system and the binaural system, it was necessary to make choices about the level of abstraction of the model. For example, the human cochlea has 3000 inner hair cells, whereas our cochlea model had only 600 filter stages. The choice of 600 filters gave us frequency resolution of 5 points/musical semitone, sufficient to tell if a musical note was in tune within 20 cents – not as high-resolution as an expert listener, but sufficiently accurate that we felt we weren't missing any essential details of the computation, and at 600 filters, we could display the output on commercial computer displays, and run the algorithm in real-time on the notebook computers of 2001.

Also, the cochlea/multipolar system did not explicitly model transduction to spikes by the inner hair cells – several neuroscientists specializing in the cochlea and cochlear nucleus advised us that going to this level of detail would simply slow down the computation and not provide any additional essential insights or performance improvement to our model. The compressive nonlinearity in our cochlea model is a simple square-root compression, without the explicit time-constants and amplitude dependent frequency shifts as observed in the real cochlea, and preserved in more sophisticated cochlear models [7]. We felt that our choice of abstraction level was sufficiently accurate to provide a foundation for the higher-level processes of binaural localization, stream separation, and speech recognition, and in fact, including those additional complexities would have made it harder to build those higher-level processes. We speculate that the biological system compensates for those effects, and that we would not lose much, if anything, at the system level, by simply leaving them out.

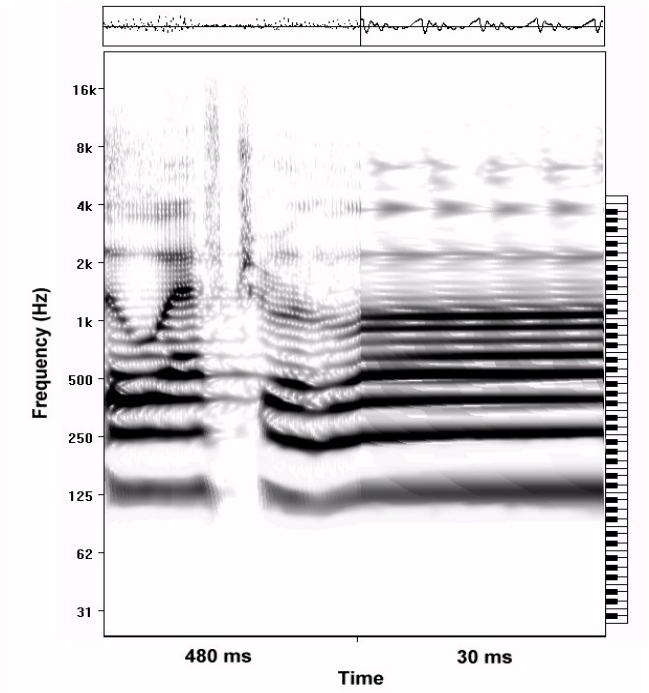

Figure 2. Output of model of Cochlea and Multipolar Cells for the utterance "so after a lot of thought".

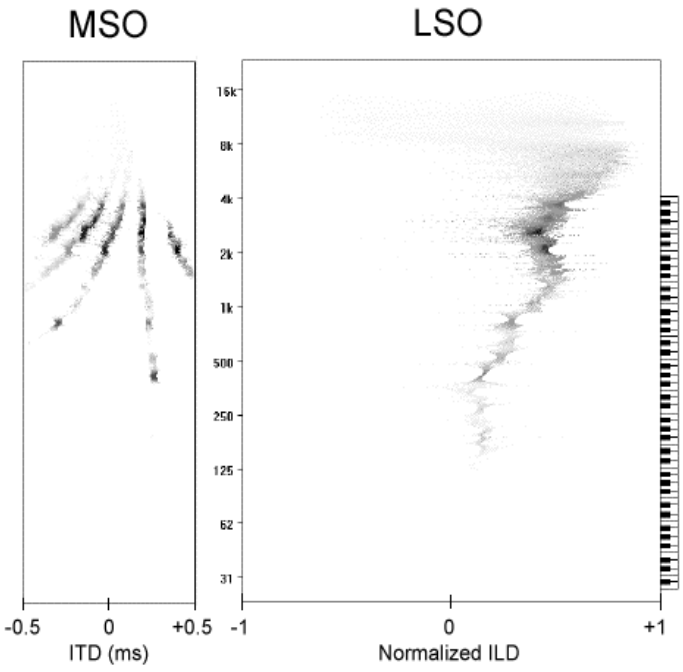

Figure 3. Output of binaural hearing model, showing the Inter-Aural Time Delay (ITD) and Normalized Inter-Aural Level Difference (ILD) computed in the MSO and LSO, respectively, for a sound source coming from the right.

III. COMMERCIAL PRODUCT DEVELOPMENT

A. Algorithms

Fig. 4 shows the algorithmic block diagram of the Audience A1010 Voice Processor. In the transmit direction, the inputs to the system are the two microphones X_1 and X_2, which are analogous to the two ears. The microphone signals are transformed into the spectral domain using Audience's proprietary Fast Cochlea Transform (FCT), which was carefully designed to model the detailed properties of the human cochlea. The two outputs of the FCT are analyzed and compared in the Feature Extraction block, which performs the essential computations that are computed in the cochlear nucleus, olivary complex, and inferior colliculus. The Analyze Sources block determines the allocation of spectral energy to foreground and background sources in a way that is modeled after the attention mechanisms in auditory cortex, ultimately producing a fine-grained mask to apply to the microphone signals in the Modify Spectrum block, such that the foreground speech is preserved, and the background noise is attenuated. The modified spectrum is then passed to the Inverse Fast Cochlea Transform (IFCT), which converts the spectral representation back into a time-domain audio waveform X_T, where it can be encoded for radio transmission by the baseband processor.

Note that the Inverse Fast Cochlea Transform is the only computational element in the transmit pathway that is not modeled after a real brain region – it is an engineering requirement to convert the spectrum back into a sound waveform to create a functional telecommunications product.

In the receive direction, the decoded signal from the baseband processor X_R is transformed into the spectral domain using another FCT, and monaural noise reduction is applied in the Receive Noise Suppression (Rx NS) block. Note that the Rx NS block takes an input from the transmit pathway – the Voice Processor is already analyzing the local noise environment in which the device is being used, so it can adjust the incoming signal for optimal volume and spectral shape relative to the local noise. Finally, another IFCT converts the noise-reduced receive-side spectrum back into sound signal X_S to be played through the handset speaker.

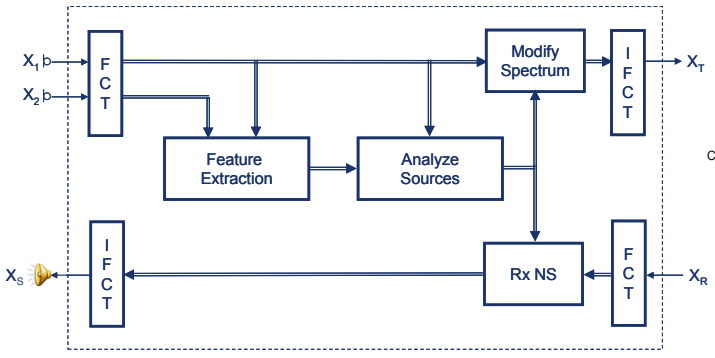

Figure 4. Algorithmic block diagram of Audience's A1010 Voice Processor.

B. Hardware Acceleration

The Audience A1010 Voice Processor has been optimized to efficiently run the most common and computationally intensive algorithms. The Fast Cochlea Transform (FCT), is a good example of an important operation which is computationally intensive, so that hardware acceleration is justified. The FCT is a cascade of biquad filters (Pole-Zero Filter Cascade) in a proprietary implementation, carefully tuned to match human hearing frequency responses, as shown in Fig. 5.

The core operations of parallel multiplies and adds in Audience's proprietary implementation of the FCT have been implemented in dedicated hardware to provide accelerated performance at reduced power consumption for this critical operation. In addition, there are dedicated instructions for other common nonlinear operations such as *sqrt*, *log, exp*, etc.

C. Voice Processor Hardware

Fig. 6 shows the hardware block diagram of the Audience A1010 Voice Processor. It runs on 1.8-3.3V power supplies, with optional 1.2V core power input. It is packaged in a 48-pin Chip-Scale Package (CSP), with 0.4mm ball spacing. Active current consumption is 15-32mA, depending upon which features are enabled. It consumes 30µA current when in sleep mode.

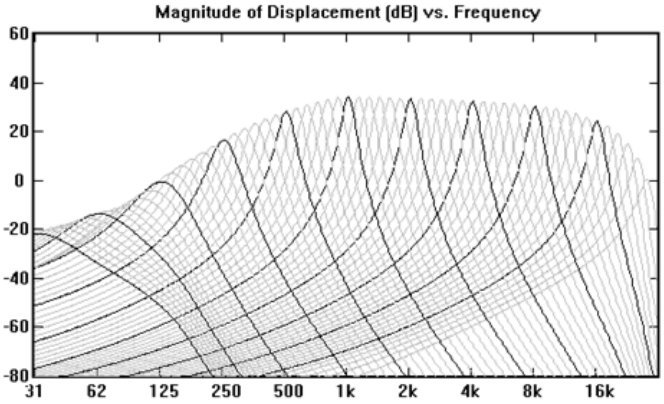

Figure 5. Transfer functions of the FCT.

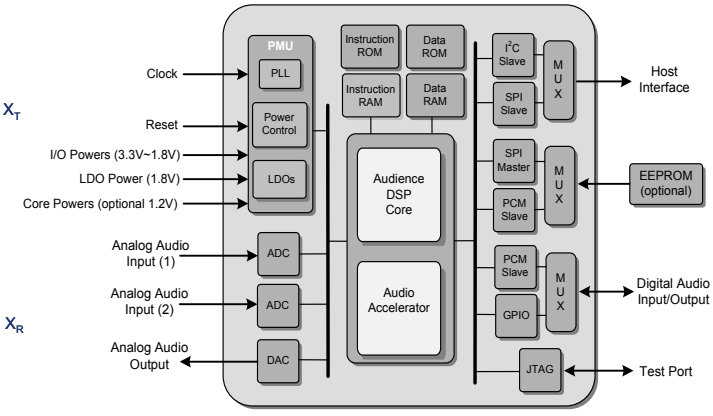

Figure 6. Block Diagram of the Audience A1010 Voice Processor.

Fig. 7 shows the die plot of the A1010 Voice Processor. The die size is 2.7 x 3.5 mm. The chip was fabricated on the TSMC 130nm process.

D. *Noise Reduction Performance*

Fig. 8 shows the effect of the Voice Processor's noise reduction on a real recording made on a busy street. The top panel shows the spectrum of the original signal at the primary microphone on a candy-bar format phone. The bottom panel shows the spectrum of the noise-reduced signal. Notice that there has been a dramatic reduction in the energy of the background noise sources, including the non-stationary noise sources such as the unwanted background voice and the cell-phone ringtone.

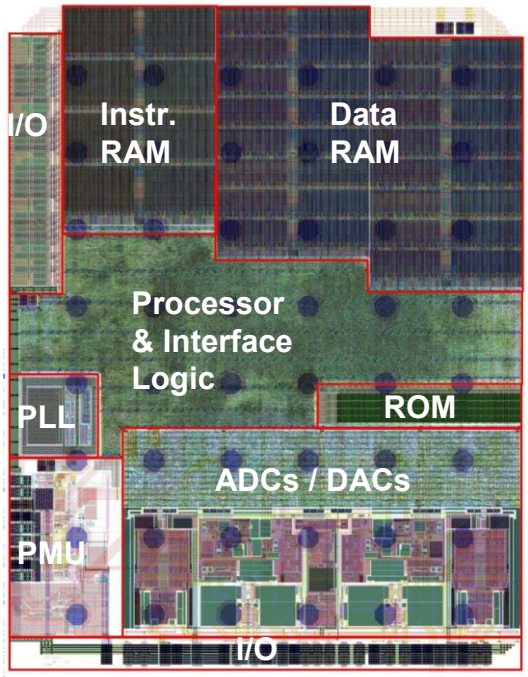

Figure 7. Die Plot of the A1010 Voice Processor.

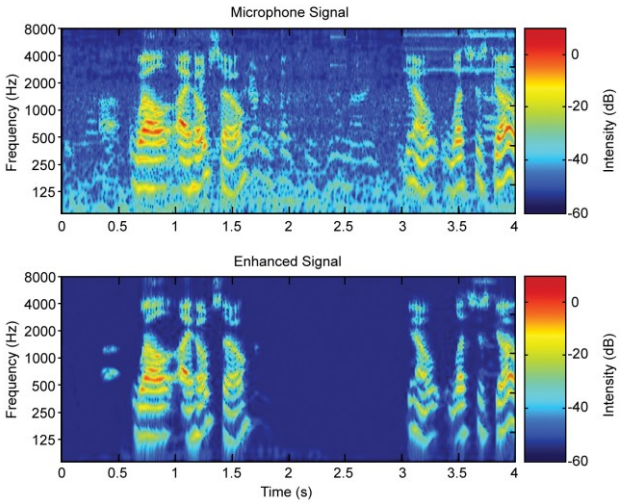

Figure 8. The Fast Cochlea Transform representation of a microphone signal, before and after noise reduction.

Fig. 9 shows the measured Mean-Opinion-Score improvement with the Audience A1010 Voice Processor, as a function of input Signal-To-Noise Ratio (SNR), using the methodology described in ITU-T standard P.835 Amendment I Appendix III [9]. The average improvement is 0.77 MOS.

IV. CONCLUSION

Audience has developed a sophisticated real-time model of the human auditory pathway, and used it as the foundation for a commercial cell-phone noise reduction product. Success in this endeavor required a detailed study of the neuroscience, and careful choices of abstraction level of the model, so as to meet the constraints of the implementation medium and commercial application.

REFERENCES

[1] L. Watts, "Visualizing Complexity in the Brain," in *Computational Intelligence: The Experts Speak,* D. Fogel and C. Robinson, Eds., IEEE Press/Wiley, pp. 45-56, 2003.

[2] G. Shepherd, *The Synaptic Organization of the Brain,* Fourth Edition, Oxford University Press, p. vi.

[3] E. Young, "The Cochlear Nucleus", in *The Synaptic Organization of the Brain,* Fourth Edition, G. Shepherd, Ed., Oxford University Press, pp. 121-158.

[4] Yin, T. (2002), in Oertel, D., Fay, R., and Popper, A., ed., *Integrative Functions in the Mammalian Auditory Pathway*, Springer-Verlag, New York, pp. 99-159.

[5] 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.

[6] R. F. Lyon, "A Signal-Processing Model of Hearing", Xerox PARC technical report, 1978. Available at http://www.dicklyon.com .

[7] M. Slaney, "Lyon's Cochlear Model," Apple Technical Report No. 13, Advanced Technology Group, Apple Computer, Inc., Cupertino, CA, 1988.

[8] L. Watts, "Computation of Multi-Sensor Time Delays", United States Patent Number 6,792,118, September 14, 2004.

[9] ITU-T P.835 Amendment I Appendix III, *Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm* – Appendix III: *Additional provisions for non-stationary noise suppressors,* http://www.itu.int/rec/T-REC-P.835-200710-I!Amd1/en .

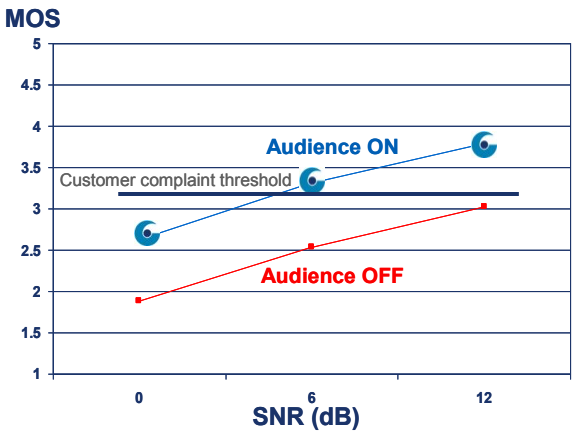

Figure 9. Subjective Performance Measurement using P.835 Test Methodology.