Voice Processors Based on the Human Hearing System

A VOICE PROCESSOR IS A DEDICATED AUDIO SIGNAL-PROCESSING CHIP FOR MOBILE PHONES AND PERSONAL COMPUTERS THAT ENABLES HIGH-PERFORMANCE NOISE REDUCTION AND ACOUSTIC ECHO CANCELLATION FOR HIGHLY NOISY ENVIRONMENTS. THE ALGORITHMS AND HARDWARE UNDERLYING THESE CHIPS' PERFORMANCE ARE BASED ON THE SOPHISTICATED OPERATION OF THE HUMAN HEARING SYSTEM.

Lloyd Watts
Dana Massie
Allen Sansano
Jim Huey
Audience

A new generation of mobile phones and PCs is being deployed with dedicated voice processor chips. Advanced multimicrophone noise-suppression algorithms, based on the operation of the human hearing system, are fueling this trend. These algorithms run on highly optimized signal-processing hardware.

Voice processors reside between a set of microphones and the baseband processor. They provide advanced noise reduction and acoustic echo cancellation, making it easier for mobile users to hear and be heard in highly noisy environments.

Figure 1 shows the basic architecture of a mobile phone containing a voice processor. In the *transmit* direction, two microphones pick up the voice and background sounds. The voice processor digitizes the microphone signals, performs a sophisticated binaural sound-separation algorithm, and provides cleaned-up voice to the baseband processor chip. This chip encodes the voice signal for transmission and passes it to the RF chip to transmit the signal to the nearest base station. In the *receive* direction, the RF chip demodulates the radio signal and passes it to the baseband processor chip, which decodes it. The voice processor further processes the signal, then sends it to the handset speaker.

These voice processors result in improved user experience, customer satisfaction, and network efficiency; and reduced customer churn.

The human hearing system

Figure 2 shows a simplified block diagram of the human hearing system.[1] We use two ears to help localize the spatial position of sound sources in our environment. The *cochlea*—the sensory organ of hearing—performs a sophisticated spectro-temporal analysis of the sound at each ear, decomposing the sound into its time-varying frequency components. The auditory nerve carries this spectro-temporal representation of sound to the five major cell types in the cochlear nucleus: the dorsal cochlear nucleus cells, octopus cells, multipolar cells, spherical bushy cells, and globular bushy cells. Each of these cell types computes a different specialized representation of sound. For example, the octopus cells respond strongly in the presence of strong transients, such as a drumbeat, whereas multipolar cells respond to the energy envelope of the sound in a given frequency region, discarding detailed phase information.

The lateral and medial superior olivary complexes compare the signals from the two cochlear nuclei. These first binaural regions of the auditory pathway compute the location

Published by the IEEE Computer Society

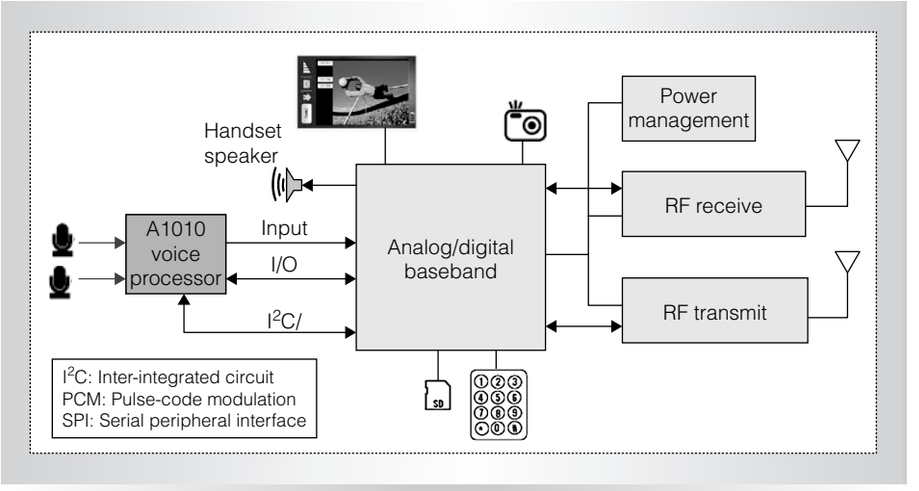

Figure 1. Mobile phone architecture. The Audience A1010 voice processor resides between the microphones and baseband processor, providing advanced noise reduction.

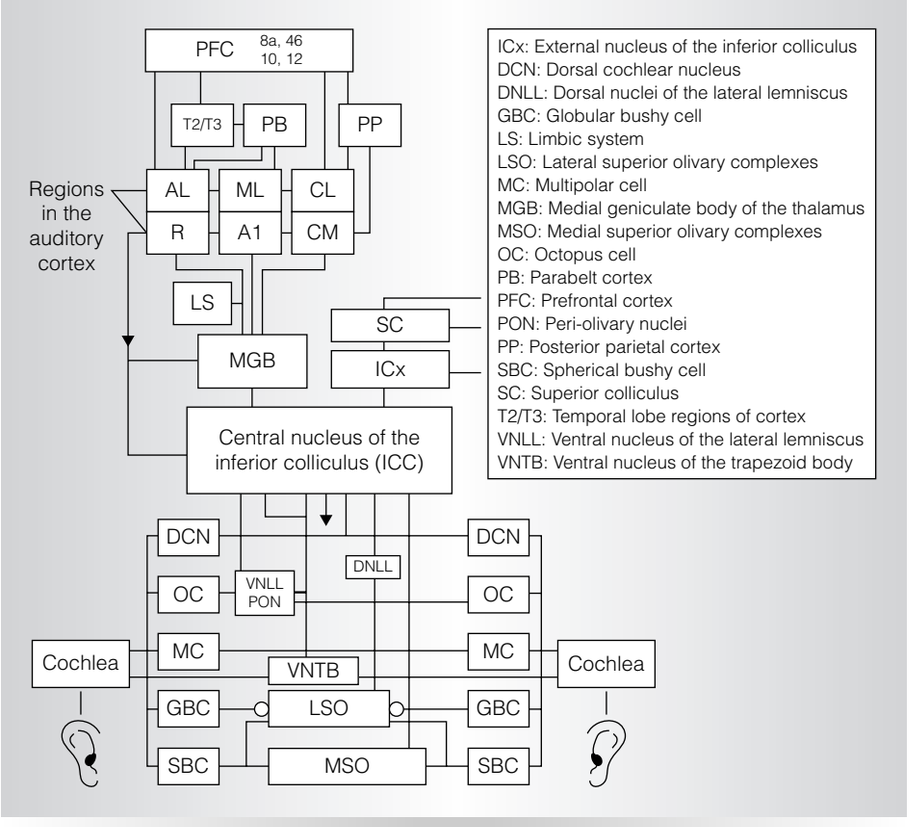

Figure 2. Simplified block diagram of the human auditory pathway. This figure shows the many different cell types and brain regions that perform the hearing function in the human brain.

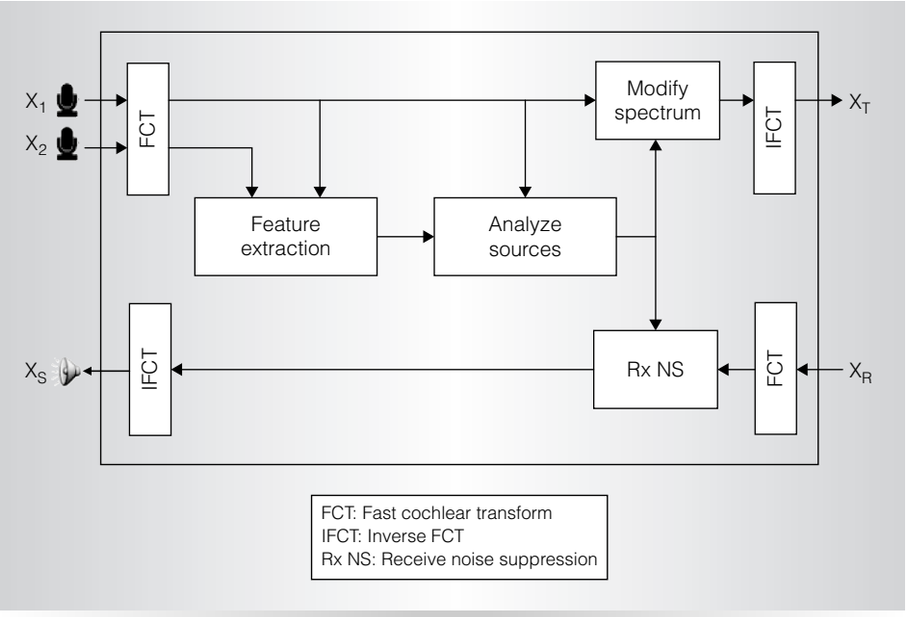

Figure 3. Block diagram of the computations performed in the Audience voice processor. The transmit noise reduction functions flow from left to right; the receive noise reduction functions flow from right to left.

of sound sources using timing and level differences of the sounds arriving at the ears.

The dorsal and ventral nuclei of the lateral lemniscus apply advanced signal conditioning, including reverberation removal. Next, the central nucleus of the inferior colliculus combines, aligns, and normalizes all of the signal representations.

The thalamus's medial geniculate body acts as an attentional gateway and relay to communicate signals from the lower brainstem regions to the limbic system regions and the auditory cortex. The limbic system regions control emotional and fear responses to sound; the auditory cortex (regions R, A1, CM, AL, ML, CL) performs high-level cognitive auditory functions such as sound source separation, speech recognition, and pitch perception.

Voice processor algorithms

Figure 3 shows the algorithmic block diagram of the Audience A1010 voice processor. In the transmit direction, the system inputs are the two microphones, X_1 and X_2, which are analogous to the two ears. Audience's proprietary fast cochlea transform (FCT), which we designed to model the detailed properties of the human cochlea, transforms the microphone signals into the spectral domain. The *feature extraction* block analyzes and compares the two FCT outputs. This block 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 using a process we modeled after the attention mechanisms in the auditory cortex. Ultimately, this process produces 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 *inverse FCT* (IFCT) converts the modified spectral representation back into a time-domain audio waveform X_T. The baseband processor then encodes this waveform for radio transmission.

The IFCT is the only computational element in the transmit pathway that isn't modeled after a real brain region. Converting the spectrum back into a sound waveform to create a functional telecommunications product is an engineering requirement.

In the receive direction, another FCT transforms the decoded signal from the baseband processor X_T into the spectral domain.

It then applies monaural noise reduction 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.

Competing two-microphone approaches include beamforming[2] (offered by ForteMedia, GTronix, National Semiconductor, and others) and blind source separation[3,4] (offered by SoftMax/Qualcomm).

Beamforming uses a time-domain delay-and-add network to form a cardioid directional pickup pattern that points toward the handset user's mouth, suppressing noise sources outside the beam. This approach's strengths include simplicity and a low computational footprint. In some cases, the algorithm is simple enough that it can be implemented directly in analog, as in the GTronix and National Semiconductor offerings. Weaknesses include poor or no suppression of noise sources in the beam, and high sensitivity to microphone placement and matching. In digital implementations, beamforming is often augmented by stationary noise suppression based on spectral subtraction, and noise gating based on voice activity detection. These augmentations improve performance at the expense of complexity, latency, suppression consistency, and loss of robustness at high noise levels.

Blind source separation is an adaptive filtering technique, usually implemented in the time domain, that uses a linear unmixing matrix to separate the two microphone signals into two uncorrelated or independent signals. Like beamforming, this technique's strengths include simplicity and a low computational footprint. Its weaknesses include poor performance in the presence of reverberation and moving or multiple noise sources, because the linear adaptive filter technique assumes an equal number of microphones and sound sources. Commercial blind source separation offerings are usually augmented by adaptation control based on voice activity detection, which reduces robustness at high noise levels.

Software versus hardware

Both technical and business considerations influence the choice of a software or hardware implementation. In the mobile phone environment, third-party developers usually can't access the digital signal processing portion of the baseband processor for programming. Therefore, only the baseband processor manufacturer can implement a software solution running on the baseband processor—for example, the manufacturer might develop, license, or acquire an in-house noise-reduction solution (as in the case of Qualcomm/SoftMax).

Third-party developers of hardware solutions include Audience and Fortemedia (digital hardware) and GTronix and National Semiconductor (analog hardware). Third-party hardware solutions add the cost of an additional chip in return for the overall solution's performance and flexibility that their noise-reduction algorithms allow.

Hardware acceleration

We optimized the Audience A1010 voice processor to efficiently run the most common and computationally intensive algorithms. The FCT, for example, is an important and computationally intensive operation, so it justifies hardware acceleration. As Figure 4 shows, the FCT is a cascade of biquad filters.[5,6]

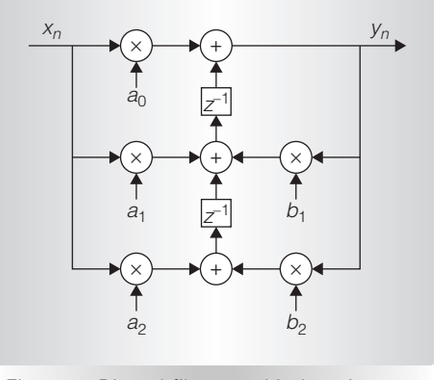

Figure 4. Biquad filter used in Lyon's cochlea model, in which $y_n = a_0x_n + a_1x_{n-1} + a_2x_{n-2} + b_1y_{n-1} + b_2y_{n-2}$. The Audience A1010 voice processor uses a proprietary variation of this method.

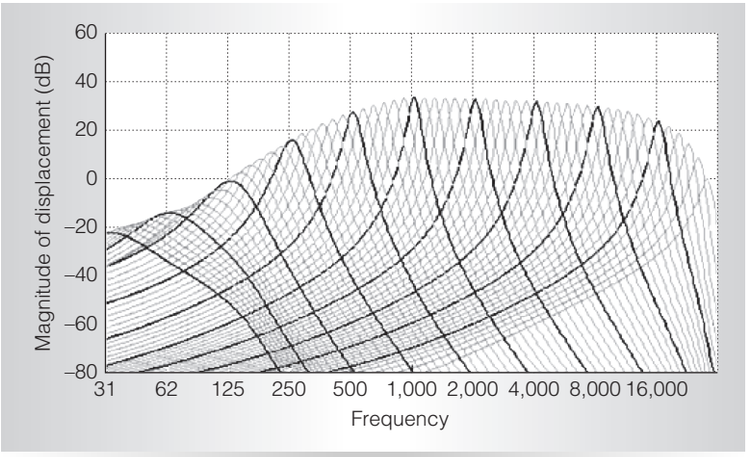

Figure 5. Transfer functions of the cochlea model. These filter responses were carefully designed to match the bandwidths and peak amplitudes of the human hearing responses.

Figure 5 shows the filter characteristics of the Audience A1010 voice processor's proprietary implementation of the FCT method, which we tuned to match human hearing frequency responses.

Audience's proprietary FCT implementation executes parallel multiplies and adds in dedicated hardware, providing accelerated performance at reduced power consumption for these critical operations. We also included dedicated instructions for other common nonlinear operations such as `sqrt`, `log`, and `exp`.

Voice processor hardware

Figure 6 shows the hardware block diagram of the Audience A1010 voice processor. The processor runs on 1.8-3.3-V power supplies, with an optional 1.2-V core power input.

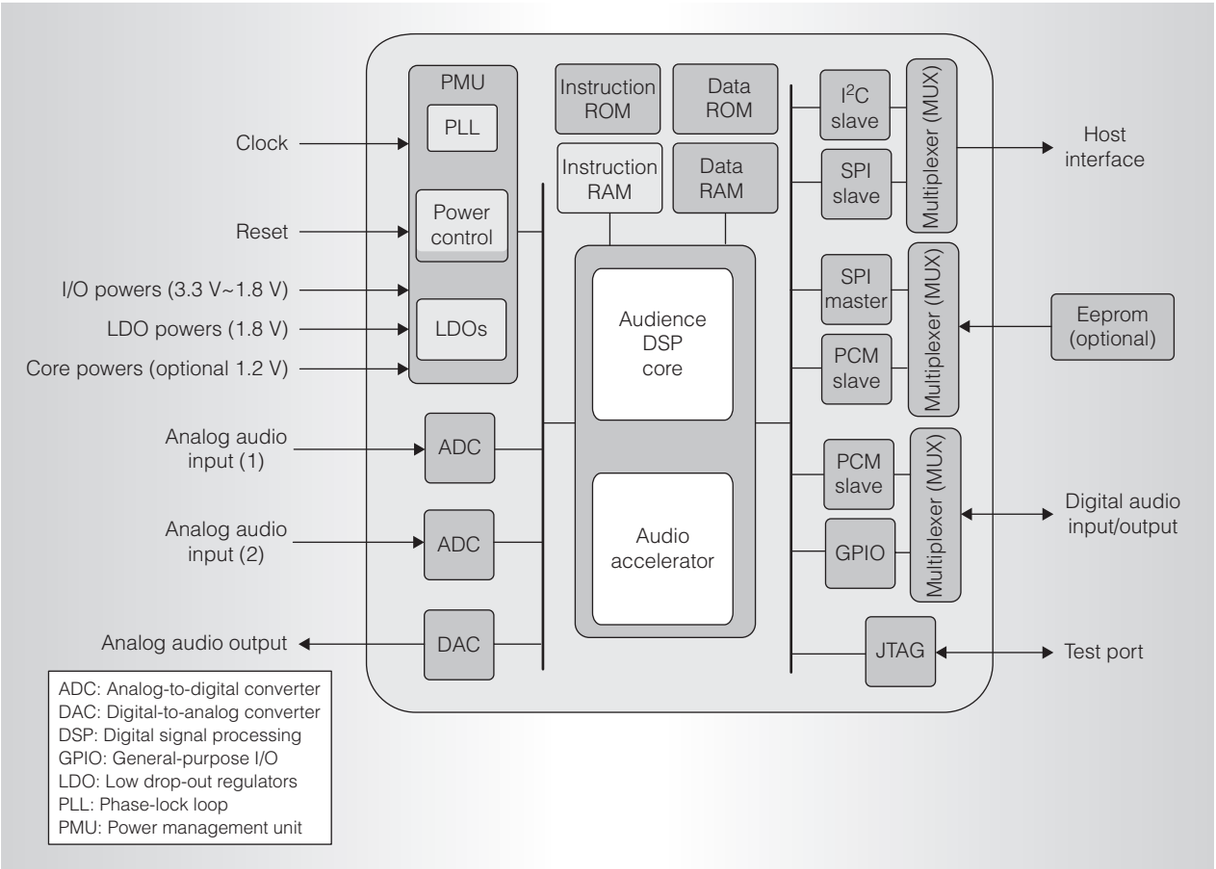

Figure 6. Block diagram of Audience's A1010 voice processor chip. The power management unit includes the clock phase-lock loop and low drop-out regulators. Analog-to-digital and digital-to-analog converters provide the analog interfaces. The voice processor supports several industry standard interfaces, including I^2C (inter-IC communication), SPI (serial peripheral interface), PCM (pulse-code modulation), and GPIO (general-purpose I/O).

It's packaged in a 48-pin chip-scale package with 0.4-mm ball spacing. Active current consumption is 15-32 mA, depending on which features are enabled. It consumes 30 μA current in sleep mode.

Figure 7 shows the die plot of the A1010 voice processor. The die size is 2.7 × 3.5 mm. We fabricated the chip using the TSMC 130-nm process.

Noise reduction performance

Figure 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 original signal's spectrum at the primary microphone on a candy-bar format phone. The bottom panel shows the noise-reduced signal's spectrum. Notice the dramatic reduction in the energy of the background noise sources, including nonstationary noise sources such as the unwanted background voice and the cell phone ringtone.

Figure 9 shows the measured mean-opinion-score (MOS) improvement when using the Audience A1010 voice processor

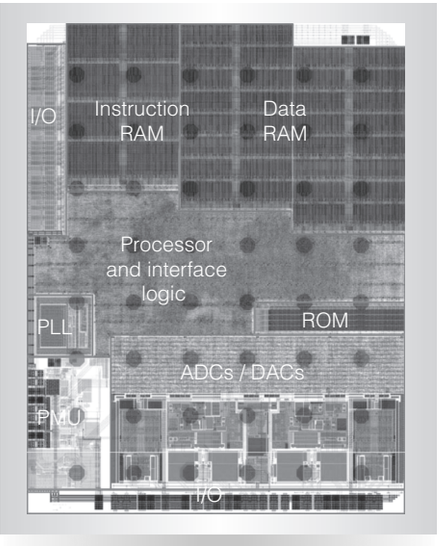

Figure 7. Die plot of the A1010 voice processor. Memory (RAM and ROM) occupies about a third of the area; mixed-signal functions (analog-to-digital and digital-to-analog converters) occupy about a third of the area; and processor logic and power management functions occupy the rest of the area.

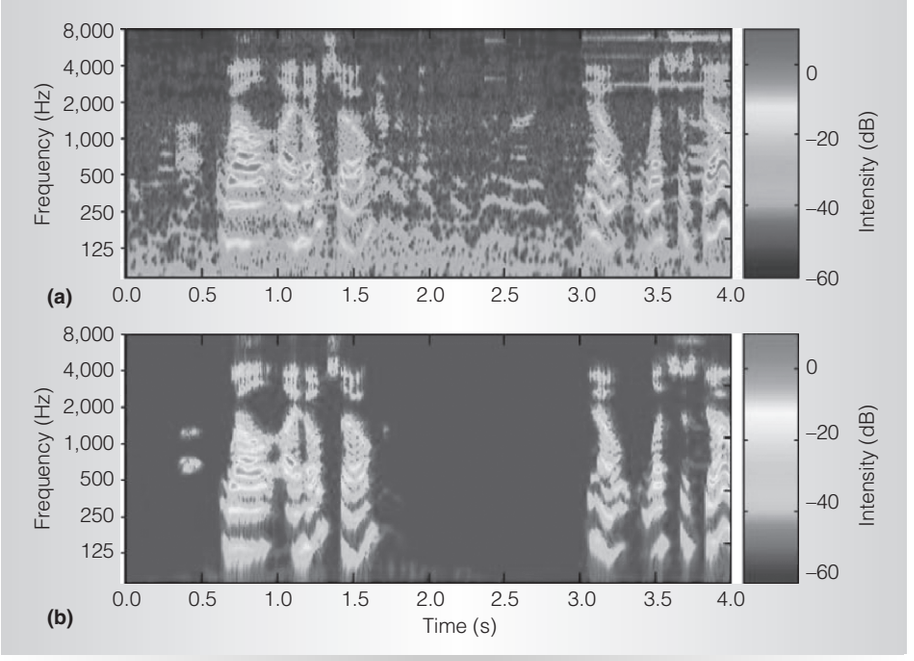

Figure 8. The fast cochlea transform (FCT) representation of a microphone signal before (a) and after (b) noise reduction. The logarithmic frequency scale shows a doubling in frequency (each tic mark represents doubling). The signal is a voice recorded on a busy street with traffic noise, a nearby talker, and a cell phone ringtone.

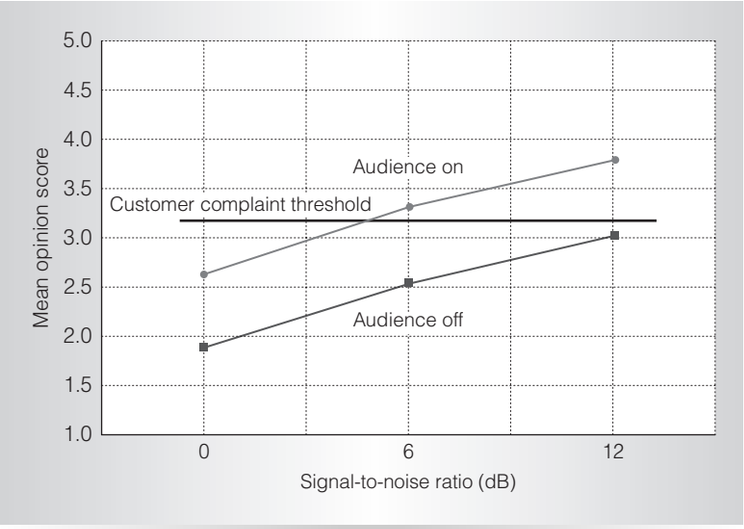

Figure 9. Subjective performance measurement using ITU-T P.835 Amendment 1, Appendix III methodology. The Audience A1010 voice processor improves performance an average of 0.77 MOS.

Table 1. Power consumption and noise suppression performance of the A1010 voice processor.

Measure	Power consumption (mA)	Total noise level reduction (dB)	Efficiency (TNLR/power consumption) (dB/mA)
Transmit noise suppression (Tx NS)	14	25	1.8
+ Receive noise suppression (Rx NS)	7	15	2.1
+ AEC	1	35	n/a
+ VE	2	n/a	n/a
Chip circuitry	8	n/a	n/a
Total	32	n/a	n/a

as a function of input signal-to-noise ratio. We used the methodology described in the International Telecommunications Union's Telecommunication Standardization Sector (ITU-T) standard P.835 Amendment I, Appendix III.[7] The average improvement is 0.77 MOS.

Table 1 shows the Audience A1010 voice processor's power consumption and noise suppression performance. We use a modified version of the total noise-level reduction (TNLR) measurement specified in ITU-T standard G.160.[8]

Test standards

In 2006, it was becoming clear that the emerging noise suppression technology would impact the mobile telephone industry. At that time, common industry practice was to test noise-suppression algorithms with pink noise, car noise, street noise, and cafeteria babble noise. All of these are *quasistationary* noise sources, and thus this test methodology wouldn't be useful for evaluating the performance of an advanced *nonstationary* noise suppressor.

In October 2007, an industry consortium led by Audience, and with support from T-Mobile, AT&T, Sprint, Motorola, and Nokia, developed a new test methodology for testing nonstationary noise suppressors. Published as ITU-T P.835 Amendment I, Appendix III,[7] the new standard specifies six noise types. Two of these noise types—single-voice distractor and music containing drums—are nonstationary. The standard also requires virtual motion of the noise sources in a four-loudspeaker recording environment to simulate realistic noise environments.

In addition to subjective performance measures, wireless carriers need objective performance measures that can be automated for low-cost testing of new phone models. Existing standard methods, such as the ITU-T G.160 standard's TNLR measure,[8] don't account for the high suppression and nonstationary noise-reduction capabilities of this new generation of voice processors. Therefore, begun in 2008 and continuing through 2009, Audience is leading an effort to improve the accuracy of the TNLR and other related objective measures in the ITU-T G.160 standard.

The next generation of voice processors promises to bring exciting new functionality to mobile telephones, PCs, and converged entertainment devices. We believe that this is just the beginning of establishing this new class of co-processor device that will assume an increasingly important role in mobile telephone architecture, analogous to graphics processors in the PC and video game architectures. MICRO

References

1. L. Watts, ''Visualizing Complexity in the Brain,'' *Computational Intelligence: The Experts Speak,* D. Fogel and C. Robinson, eds., John Wiley & Sons/IEEE Press, 2003, pp. 45-56.
2. W. Kellermann, ''Beamforming for Speech and Audio Signals,'' *Handbook of Signal Processing in Acoustics,* D. Havelock, S. Kuwano, and M. Vorlander, eds., Springer, 2008, pp. 691-702.
3. A.J. Bell and T.J. Sejnowski, ''An Information Maximization Approach to Blind Separation and Blind Deconvolution,'' *Neural Computation,* vol. 7, no. 6, Nov. 1995, pp. 1129-1159.
4. T. Lee, A.J. Bell, and R. Orglmeister, ''Blind Source Separation of the Real World Signals,'' *Proc. Int'l Conf. Neural Networks* (ICNN), IEEE Press, 1997, pp. 2129-2134.
5. R.F. Lyon, *A Signal-Processing Model of Hearing,* tech. report, Xerox PARC, 1978; http://www.dicklyon.com/tech/Hearing/Sig_Proc_Model_of_Hearing-Lyon1978.pdf.
6. M. Slaney, *Lyon's Cochlear Model,* tech. report No. 13, Advanced Technology Group, Apple Computer, 1988; http://cobweb.ecn.purdue.edu/~malcolm/apple/tr13/LyonsCochlea.pdf.
7. Int'l Telecomm. Union, Telecomm. Standardization Sector (ITU-T) P.835 Amendment I, Appendix III, Subjective Test Methodology for Evaluating Speech Comm. 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.
8. Int'l Telecomm. Union, Telecomm. Standardization Sector (ITU-T) G.160, *Voice Enhancement Devices,* http://www.itu.int/rec/T-REC-G.160-200806-P/en.

Lloyd Watts is the founder, chair, and chief technology officer of Audience. His research interests include auditory signal processing and signal processor architecture. Watts has a PhD in electrical engineering from the California Institute of Technology. He is a member of the IEEE.

Dana Massie is the director of DSP architecture at Audience. His research interests include VLSI architectures and DSP algorithm development and implementation. Massie previously was director of the Creative Technology Advanced Technology Center, where he led the development of the EMU10K1, a multimedia coprocessor for PCs that accelerated environmental audio extensions.

Allen Sansano is the director of VLSI engineering at Audience. His research interests include low-power VLSI design and system-on-chip architectures. Sansano has an MS in electrical engineering from the University of Washington.

Jim Huey is the vice president of VLSI engineering at Audience. His research interests include microprocessor architecture/design, Bayesian inference, and decision theory. Huey has a BS in electrical engineering and computer science from Duke University.

Please direct any questions and comments about this article to Lloyd Watts, Audience Inc., 440 Clyde Ave., Mountain View, CA 94043; lwatts@audience.com.