Saliency-Maximized Audio Visualization and Efficient Audio-Visual Browsing for Faster-Than-Real-Time Human Acoustic Event Detection

KAI-HSIANG LIN, XIAODAN ZHUANG*, CAMILLE GOUDESEUNE, SARAH KING,
MARK HASEGAWA-JOHNSON, and THOMAS S. HUANG, University of Illinois at Urbana-Champaign

Browsing large audio archives is challenging because of the limitations of human audition and attention. However, this task becomes easier with a suitable visualization of the audio signal, such as a spectrogram transformed to make unusual audio events salient. This transformation maximizes the mutual information between an isolated event's spectrogram and an estimate of how salient the event appears in its surrounding context. When such spectrograms are computed and displayed with fluid zooming over many temporal orders of magnitude, sparse events in long audio recordings can be detected more quickly and more easily. In particular, in a 1/10-real-time acoustic event detection task, subjects who were shown saliency-maximized rather than conventional spectrograms performed significantly better. Saliency maximization also improves the mutual information between the ground truth of nonbackground sounds and visual saliency, more than other common enhancements to visualization.

Categories and Subject Descriptors: H.5.2 [**Information Interfaces and Presentation**]: User Interfaces—*Theory and methods, Evaluation / methodology*; H.1.2 [**Models and Principles**]: User/Machine Systems—*Human Information Processing*; H.5.1 [**Information Interfaces and Presentation**]: Multimedia Information Systems—*Audio input / output*; I.5.4 [**Pattern Recognition**]: Applications—*Computer vision*

General Terms: Human Factors, Algorithms, Experimentation
Additional Key Words and Phrases: Visual salience/saliency, audio visualization, acoustic event detection

ACM Reference Format:

Lin, K.-H., Zhuang, X., Goudeseune, C., King, S., Hasegawa-Johnson, M., and Huang, T. S. 2013. Saliency-maximized audio visualization and efficient audio-visual browsing for faster-than-real-time human acoustic event detection. ACM Trans. Appl. Percept. 10, 4, Article 26 (October 2013), 16 pages.
DOI: http://dx.doi.org/10.1145/2536764.2536773

This work is funded by the US National Science Foundation grant 0807329. All results and opinions are those of the authors and are not endorsed by the US National Science Foundation.
Section 6 of this article, and parts of Section 2, were previously published in Lin et al. [2012]. Section 5 provides new detail about the experimental tools used in Section 6. Sections 3 and 4 describe new experiments, never previously submitted for publication.
Authors' addresses: K.-H. Lin, Beckman Institute, 405 North Mathews Avenue, Urbana, IL 61801; email: klin21@illinois.edu; X. Zhuang; email: xzhuang@bbn.com; C. Goudeseune; email: cog@illinois.edu; S. King; email: sborys@illinois.edu; M. Hasegawa-Johnson; email: jhasegaw@illinois.edu; T. Huang; email: t-huang1@illinois.edu.
*X. Zhuang is currently affiliated with the Speech, Language and Multimedia Business Unit, Raytheon BBN Technologies, 10 Moulton St, Cambridge, MA 02138.
Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies show this notice on the first page or initial screen of a display along with the full citation. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers, to redistribute to lists, or to use any component of this work in other works requires prior specific perimssion and/or a fee. Permissions may be requested from Publications Dept., ACM, Inc., 2 Penn Plaza, Suite 701, New York, NY 10121-0701, USA, fax: +1 (212) 869-0481, or permissions@acm.org.
© 2013 ACM 1544-3558/2013/10-ART26 $15.00
DOI: http://dx.doi.org/10.1145/2536764.2536773

1. INTRODUCTION

Computers with gigabytes of storage have recently become inexpensive enough to be devoted to the humble task of recording audio, producing recordings far longer than those from the entertainment industry. Purely acoustic applications of long recordings include intelligence, surveillance, and reconnaissance (ISR) and tracking wildlife such as songbirds and whales. Equally well, such computers can record nonacoustic time series scaled into the human-audible frequence range, from sensors as diverse as accelerometers, EEGs, voltage spike detectors, and seismometers.

The deriving of insights from such recordings cannot be entirely formalized. Purely automatic analysis is inaccurate enough to justify putting a human investigator in the loop. Thus, we describe the initial stage of investigation as *browsing:* getting a rough feel for the data, for distinguishing its conventional background from surprising outliers. However, online publication of such long recordings has been discouraged by a lack of software for browsing them. Conversely, the lack of published recordings has similarly discouraged the development of audio browsers. This vicious circle is broken by the visualizations described in this article, optimized for the human investigator's senses, which enable rapid browsing of even multiday recordings.

These visualizations are implementable as lightweight browser plugins or WebGL applications. These would allow the uploading of unannotated long recordings with sparse moments of interest, such as recordings of a whale migration or the 18-day demonstration in Tahrir Square, with confidence that those moments can be found rapidly by others.

Machine perception is often outperformed by human perception, as the latter better handles the semantic gap between noisy observations and target events. In the particular field of acoustic event detection (AED), machines poorly detect and label nonspeech events in long recordings. For example, none of the systems competing in the CLEAR 2007 AED competition exceeded 30% accuracy in labeling events like quiet chair squeaks in seminar-room recordings [Temko et al. 2006; Temko 2007; Zhou et al. 2007]. It is similarly easy to devise AED tasks where trained human listeners strongly outperform machines. Here are three examples: humans can detect rifle magazine insertion clicks with 100% accuracy at 0dB SNR in both white noise and jungle noise [Abouchacra et al. 2007]; they can count cough events from audio with an intertranscriber RMS error of less than 4% [Smith et al. 2006]; and they can detect anomalous events in musical recordings in a single real-time audition [Hasegawa-Johnson et al. 2011].

Unfortunately, the power of human audition is constrained by time. For example, most people cannot comprehend continuous speech faster than twice normal speed [Arons 1997]. At high speed, nonspeech acoustic events are even harder to perceive than speech, because most interesting nonspeech events are transient and thus disproportionately masked. Worse yet, even after detecting an event in a long segment, pinpointing the event's timestamp usually requires rewinding and replaying. Our experiments show that AED by pure listening is considerably slower than real-time playback.

This real-time barrier to human AED can be broken by enlisting human vision, which efficiently interprets complex scenes at a glance [Anderson 2009; Belopolsky et al. 2008; Goldstein 2010]. To this end, we propose a visualization, a saliency-optimized audio spectrogram in which background audio with unchanging texture is dimmed, to enhance the at-a-glance salience of anomalous audio. This modified spectrogram can be rapidly skimmed using our high-speed zooming software [Goudeseune 2012] to rapidly identify sparse interesting intervals. This visualization is synchronized with the source recording, so an analyst searching for target events typically listens to only brief excerpts of visually interesting segments. The target events' information is embedded into visually salient patterns, which are processed by human vision with priority [Goldstein 2010]. In other words, this visualization suppresses uninteresting background noise.

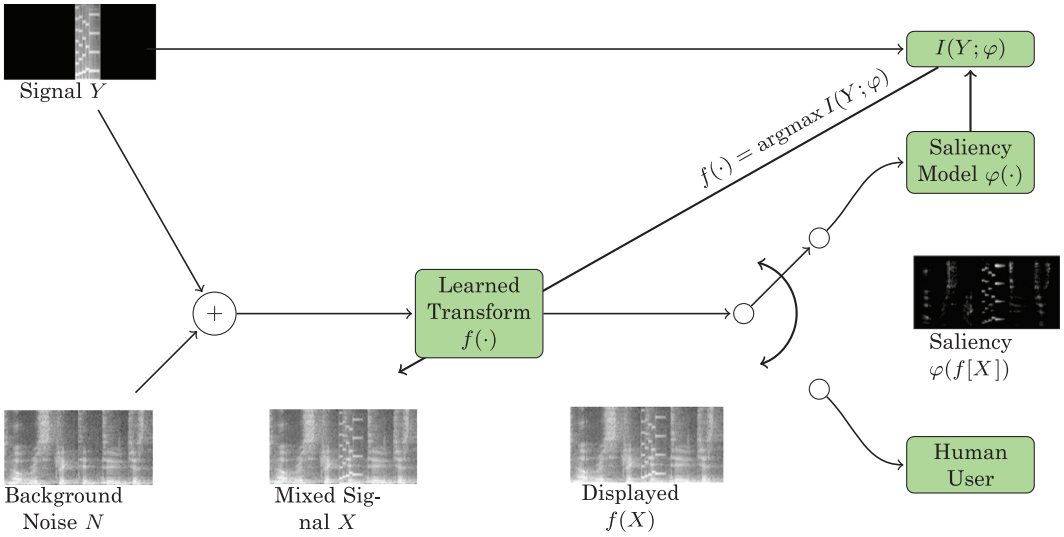


Fig. 1. Flowchart of human acoustic event detection from a visual display. X is a spectrogram of the input mixed audio signalsumming the audio waveform of Y and N, from which the system computes the displayed image $f(X)$. The transformation from spectrogram to displayed image is learned in order to optimize the mutual information $I(Y;\varphi(f(X)))$, where $\varphi(\cdot)$ is a model of human bottom-up attention allocation (saliency). After learning, the transformed image $f(X)$ is displayed to human users to speed up their search for anomalies.

We formulate this visualization problem as maximizing the mutual information (MI) [Cover and Thomas 2006; Shannon and Weaver 1949] between the spectrogram of a target event Y and the estimated visual saliency of the examined spectrogram $\varphi(f)$ (Figure 1). The input information Y is the spectrogram of the target event in isolation (without the background noise N); the transmitted information is the observer's visual percept. The visualization function f converts the mixed-signal spectrogram X to the saliency-optimized spectrogram. The saliency map $\varphi(f)$ is the output of the saliency model, which models the human visual system's signal-driven (bottom-up) allocation of attention. The visualization $f(X)$ maximizes the MI $I(Y;\varphi(f(X)))$ between noise-free events and the saliency map of training examples. To evaluate $f(X)$, we use it to generate images from another set of target events, disjoint from those used to train $f(X)$. These images are presented to human subjects, whose task performance we then measure.

2. SALIENCY-MAXIMIZED AUDIO VISUALIZATION

Let the spectrogram of the target acoustic event be $Y[n_1, n_2]$, an RGB matrix indexed by row index n_1 and column index n_2, where the time scale of the row and column index depend on the zoom of the display (Figure 1). In an AED task, users do not observe $Y[n_1, n_2]$ directly; instead, they observe $X[n_1, n_2]$, the spectrogram of the signal mixed with background noise. The background noise, with spectrogram $N[n_1, n_2]$, is irrelevant to the task (e.g., orchestral music [Hasegawa-Johnson et al. 2011] or speech [Lin et al. 2012]). To help users correctly identify where $Y[n_1, n_2]$ is nonzero, we propose to transform the image prior to display, using a learned image transformation $f[n_1, n_2] = f(X[n_1, n_2])$.

The parameters of the learned transformation f maximize how much information about the target event's location is communicated to the user. The information communicated through a noisy channel

can be measured as

$$I(Y;\Phi) = E_{Y,\Phi}\left[\log_2\left(\frac{p_{Y,\Phi}(y,\varphi)}{p_Y(y)p_\Phi(\varphi)}\right)\right], \tag{1}$$

where Y is the input of the channel, Φ is the output of the channel, and $p_{Y,\Phi}(y,\varphi)$ is their joint probability density with marginals $p_Y(y)$ and $p_\Phi(\varphi)$ [Shannon and Weaver 1949]. Notice that $I(Y;\Phi)$ is an expectation. Because the signals used to train the system differ from those shown to the user during evaluation, our algorithms are therefore trained by maximizing the stochastic approximation

$$\hat{I}(Y;\Phi) = \sum_{t=1}^{T}\log_2\left(\frac{\hat{p}_{Y,\Phi}(y_t,\varphi_t)}{\hat{p}_Y(y_t)\hat{p}_\Phi(\varphi_t)}\right), \tag{2}$$

where (y_t,φ_t) are input-output pairs observed in the training corpus, and $\hat{p}_{Y,\Phi}(y_t,\varphi_t)$ is the binned empirical probability mass function of the training data.

The channel output Φ deserves further comment. Our visualization application communicates with listeners through a channel that includes four types of distortion: (1) the addition of background noise; (2) the scaling of the spectrogram, modeled by choosing the appropriate time scale for indices n_1 and n_2; (3) the intentional distortion caused by the learned mapping $f(X[n_1,n_2])$; and (4) the limited processing power of human visual attention. We approximate the human visual system with a communication channel that attends to visual patterns selectively, in decreasing order of saliency. Therefore, when quickly examining a display, it perceives at most a few highly salient objects. The rate of information transmission is limited by the finite span of attention (about six objects at a glance), and by immediate memory (about seven items) [Miller 1956]. Our model of the communication channel is a sort of homunculus model, according to which a high-level cognitive process detects exactly those acoustic events whose visible evidence is attended to by the low-level visual attention system. The image received by the high-level cognitive processes is therefore $\varphi[n_1,n_2] = \varphi(f[n_1,n_2])$, where $\varphi(f[n_1,n_2])$ is a nonlinear multiscale saliency transform, based on a saliency model that zeros out all pixels of $f[n_1,n_2]$ except for those that will likely attract attention (the "salient" pixels) [Itti et al. 1998].

Saliency of information is important in the field of human factors engineering [Wickens and Hollands 1999]. Examples of this include a quality metric for visualization that uses the correspondence between a data relevance mask and a saliency map [Jänicke and Chen 2010], and a saliency-based perceptual tool for visual design [Rosenholtz et al. 2011]. But although saliency has been used to *analyze* the quality of a visual representation, it has not yet been effectively used to automatically *generate* saliency-maximized visual representations.

We propose to measure the efficiency of information transmission from Y to φ through their MI. The visualization (encoding) function f is chosen to maximize $I(Y;\varphi)$, to represent the target events optimally for fast human visual examination. Hence,

$$f^* = \operatorname*{argmax}_f \hat{I}(Y;\varphi(f(X))), \tag{3}$$

where X is the input spectrogram; Y is the ground truth (the spectrogram of the isolated event); and $f(X)$ is the displayed spectrogram, a transformation of X. Five modules solve for the optimized transformation function f: computing the spectrogram, transforming the visualization, computing the saliency map, computing the MI, and maximizing the MI (Figure 1). Optimization of f uses only the training data; entirely separate acoustic events and background noise are used to evaluate the derived f^*.

2.1 Computing and Transforming the Spectrogram

We base our visualization on the humble spectrogram because it is familiar to audio experts, and because even naïve subjects can successfully interpret its details. Our grayscale spectrogram resolves

128 frequency bands down to 5msec, in linear rather than logarithmic frequency scale so as to preserve the high-frequency information that is useful for distinguishing target events.

Our goal is to find a transformation function f that is saliency maximized, rendering target events so that φ extracts them as salient patterns. For simplicity, we use linear filters: $f(X) = h[n_1, n_2] ** X[n_1, n_2]$, where $**$ denotes 2D convolution, and Equation (3) optimizes h.

2.2 Computing the Visual Saliency Map

The saliency map is generated by an image saliency algorithm based on prior work [Itti et al. 1998; Walther and Koch 2006], estimating how the human visual system allocates attention to any particular pixel [Frintrop et al. 2010; Borji and Itti 2013]. Our saliency algorithm is independent of acoustic events, with parameters chosen empirically from the literature. During training, we used this algorithm to learn a spectrogram transformation function, the linear filter h in Section 2.1. It was used as well in objective evaluation, to evaluate competing visualizations. Of course, it was not needed in subjective evaluation, which used actual humans instead of algorithmic estimation. Our algorithm has three steps: building feature pyramids, computing each feature's center-surround difference (CSD), and combining all features' saliency maps into a single map (Figure 2).

An important step in any biologically plausible model of bottom-up attention is the parallel computation of early visual features such as intensity, orientation, and color. Because we use a grayscale spectrogram, we omit color, leaving orientation and intensity [Itti et al. 1998]. These features are computed by filtering the displayed image: $I_k[n_1, n_2] = B_k[n_1, n_2] ** f[n_1, n_2]$, where $k \in \{I, 0^\circ, 45^\circ, 90^\circ, 135^\circ\}$. The filter $B_I[n_1, n_2] = \delta[n_1, n_2]$ is the delta function (so $I_I = f$). The other four filters are Gabor filters with orientations of $\{0^\circ, 45^\circ, 90^\circ, 135^\circ\}$. A strong response of $[n_1, n_2]$ in I_k indicates that f has property k in a neighborhood of $[n_1, n_2]$. For example, $I_{45^\circ}[n_i, n_j]$ responds strongly to a 45° bar at $[n_i, n_j]$ in f.

Because a salient region differs from its neighborhood, the algorithm detects saliency with a CSD, implemented by convolving the input image with a difference of Gaussians (DoG) filter. Thus, $\mathrm{CSD}_k[n_1, n_2] = I_k[n_1, n_2] ** \mathrm{DoG}[n_1, n_2]$, where

$$\mathrm{DoG}[n_1, n_2] = \frac{1}{2\pi}\left(\frac{e^{-(n_1^2+n_2^2)/2\sigma_c^2}}{\sigma_c^2} - \frac{e^{-(n_1^2+n_2^2)/2\sigma_s^2}}{\sigma_s^2}\right). \tag{4}$$

This DoG function is parameterized by two σ's. The first Gaussian has the smaller σ_c; the second has the larger σ_s. (Subtracting these Gaussians approximates a Mexican Hat function.) Filtering an image with the central Gaussian, the first term in Equation (4), averages the features within approximately σ_c pixels of the output pixel. The surround Gaussian, the second term, computes a similar average for σ_s. Thus, $\mathrm{CSD}_k[n_1, n_2]$ estimates how much the pixels near $[n_1, n_2]$ stand out, relative to those in the surrounding disc of radius σ_s.

Because Gaussian filtering with large σ's is computationally expensive, we approximate it by filtering recursively with a small Gaussian kernel and downsampling. A dyadic Gaussian pyramid generates images filtered by Gaussians of different σ's. The displayed input image is filtered by a 2D separable Gaussian kernel $[1\,5\,10\,10\,5\,1]/32$ and downsampled twofold. Repeating this builds the layers of the pyramid [Walther and Koch 2006]. The filters B_k are applied to the Gaussian pyramid to extract features from different layers:

$$\mathrm{CSD}_k = \max\{0, F_{k,c} \ominus F_{k,s}\}, \qquad k \in \{I, 0^\circ, 45^\circ, 90^\circ, 135^\circ\} \tag{5}$$

where $F_{k,c}$ and $F_{k,s}$ are the center and surround layers of the pyramid for feature k, and $\ominus$ denotes across-scale subtraction. We use the pyramid's first and fourth layers as center and surround. Fullwave rectification commonly computes the magnitude of the response of weak centers on strong surrounds

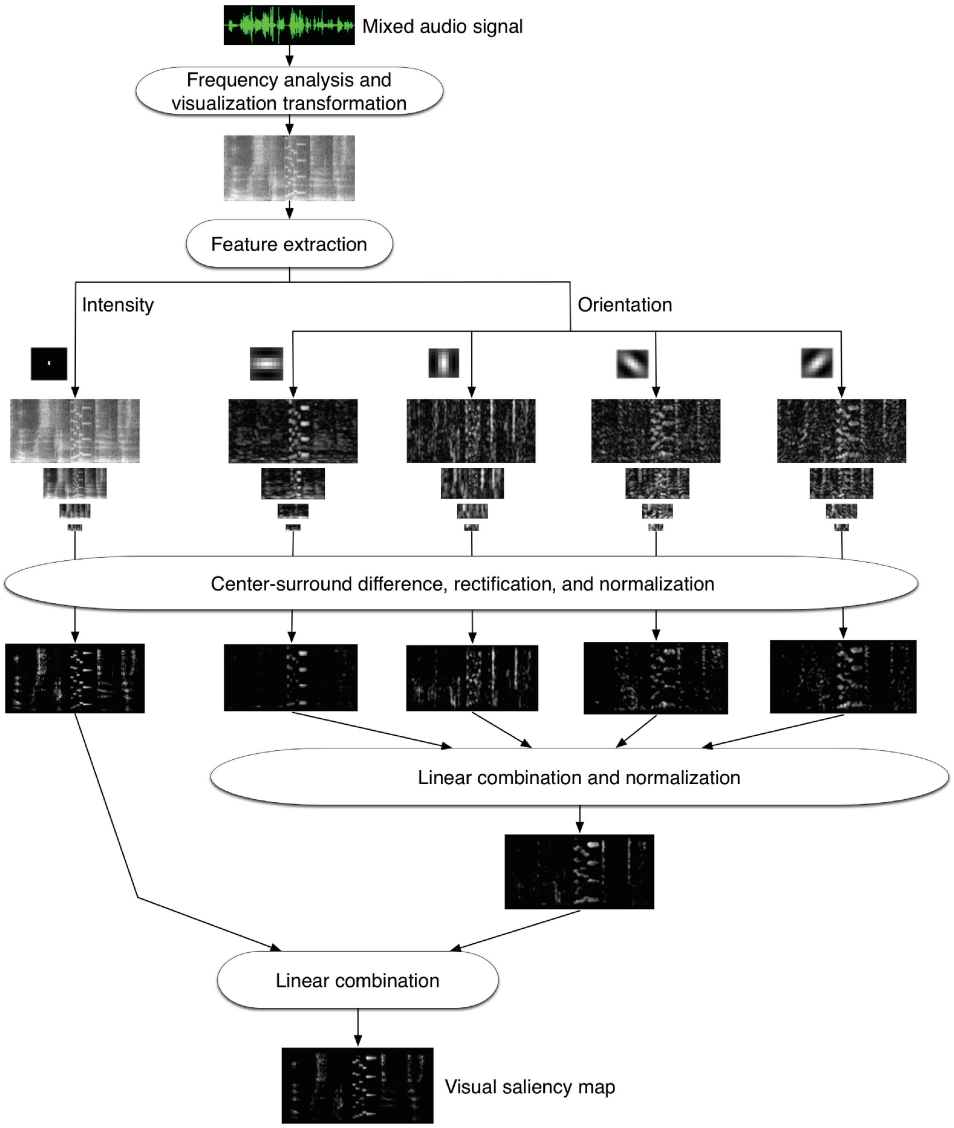


Fig. 2. Computation of visual saliency. The mixed audio signal is converted to a spectrogram and then transformed to a displayed image. (Here, the transformation function is the initial transformation of $\delta[n_1, n_2]$ in Section 2.3, so the image is the spectrogram itself.) From this, a delta-function filter and four truncated 9×9 Gabor filters (enlarged here for clarity) extract intensity and orientation features. CSD is implemented by across-scale subtraction between layers of a Gaussian pyramid. To detect strong centers on weak surrounds, each CSD is halfwave rectified. The CSDs are then normalized and combined, yielding a single saliency map.

(or vice versa). We instead use (positive) halfwave rectification in Equation (5) because, in an AED task, target events almost always have more energy than their background.

We combine the CSDs of different features into a single saliency map, normalizing before every summation with a nonlinear operator $N(\cdot)$. This operator restores similar dynamic range to the CSDs

of different visual modalities and also enhances strong local peak response. It is implemented with a large 2D DoG filter. This is followed by positive rectification [Itti and Koch 2001].

The final saliency map $S[n_1, n_2]$ is the mean of the normalized maps for intensity and for the combined orientations:

$$\begin{aligned} \overline{F_I} &= N\left(\mathrm{CSD}_I\right), \\ \overline{F_O} &= N\left(\sum_k N\left(\mathrm{CSD}_k\right)\right), \qquad k \in \{0^\circ, 45^\circ, 90^\circ, 135^\circ\} \\ S &= (\overline{F_I} + \overline{F_O})/2. \end{aligned}$$

2.3 Maximizing Mutual Information

To evaluate how well human visual perception captures the information in the visualization associated with the target events, we estimate the MI between the ground truth Y (the spectrogram of the isolated target event, obtained according to Section 2.1) and the saliency map Φ of the transformed spectrogram of the mixed signals (Equation (2)).

Because the objective function $\hat{I}(Y; \varphi)$ (Equation (2)) is nonconvex and nondifferentiable, we can only approximate the global maximum. Simulated annealing estimates f from an initial transformation of $h[n_1, n_2] = \delta[n_1, n_2]$. This transformation is also the baseline one and corresponds to the conventional spectrogram $f[n_1, n_2] = X[n_1, n_2]$.

The audio targets and background noises (y_t, φ_t) used to maximize MI are similar to, but disjoint from, those used in evaluation.

We evaluated linear filters with sizes from 5×5 to 15×15, all with similar optimized mean MIs. For human-subject experiments we chose a 5×5 filter, after inspecting the visualizations generated from the training data.

3. ALTERNATIVE ENHANCEMENTS OF VISUALIZATION

Several traditional algorithms exist for enhancing visualizations. Figure 6 shows the average MI between saliency distribution and the ground truth of various enhancements, whereas Figure 8 shows the corresponding spectrograms. We tested these enhancements:

(a) *Energy thresholding with MI maximization.*
Thresholding, ubiquitous in human perception [Goldstein 2010], suppresses any signal below a threshold. It has been applied in computer vision algorithms such as foreground detection [Bovik 2009; Gonzalez and Woods 2001]. As an alternative to our proposed saliency map generation, for learning the 5×5 linear visualization transformation filter we replaced saliency map generation with energy thresholding [Otsu 1979], zeroing any pixels with intensity below the threshold.

(b) *Event-specific Wiener filter.*
The Wiener filter suppresses noise using a filter that minimizes mean-squared error using the known autocorrelations and cross-correlations of the input signals [Lim 1990]. We gave each evaluation event its own Wiener filter, unrealistically using knowledge of the autocorrelations and cross-correlations of the noisy signal and target event of each testing input spectrogram. This oracle let us investigate the performance upper bound for audio visualization using the Wiener filter.

(c) *Event-independent Wiener filter.*
Realistic audio browsing lacks event-specific autocorrelations and cross-correlations. Here, we estimated one filter for all target events, using the average autocorrelation and cross-correlation of spectrograms in the training corpus. We trained a single 25×25 Wiener filter based on the

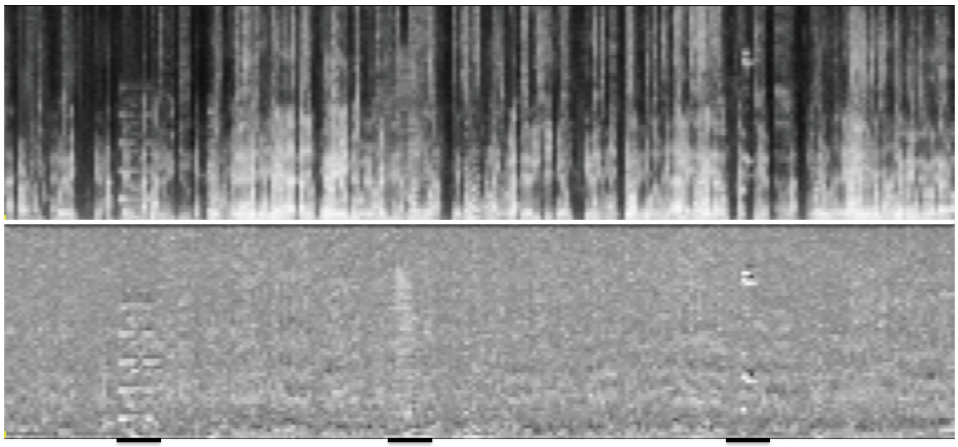


Fig. 3. Qualitative enhancement of a spectrogram: conventional (top), saliency-maximized (bottom). Three target events are marked with black underlines.

statistical property of the patches of the training spectrograms. This filter was then applied to the evaluation spectrograms, to estimate how effective Wiener filtering is under real-world browsing conditions.

(d) *Nonnegative matrix factor deconvolution.*
Besides noise filtering, we also tested an example of blind source separation, namely nonnegative matrix factor deconvolution (NMFD), an extension of nonnegative matrix factorization (NMF) [Lee and Seung 1999; Berry et al. 2007]. NMF has been applied to magnitude spectrograms of polyphonic music, modeling each instrument with an instantaneous frequency signature [Smaragdis and Brown 2003]. Extending this, NMFD models each instrument with a time-frequency signature by considering temporal structure [Smaragdis 2004]. NMFD has two important parameters: the number of basis functions R and the temporal length of the factors T. We tuned these parameters for the training set and applied them to the evaluation set (we used $R = 17$ and $T = 31$). Some of the basis functions were chosen to reconstruct the target events. There were multiple combinations of basis functions for partial reconstruction of the spectrogram. We chose the reconstructed spectrogram with the highest correlation coefficient to the ground truth event, again as an oracle for an upper bound on NMFD's performance.

4. OBJECTIVE EVALUATION

Data for training and evaluating the algorithm used electronic sound effects, such as those common to 1980's video games, as target events, superimposed on the realistically noisy background of an ongoing seminar [Carletta 2007]. All 62 sound effects were obviously foreign to a seminar room. The lengths of the sound effects are shorter than 5 seconds, except for one lasting for 9 seconds. Both the target events and the background audio were split into disjoint subsets—half for training and half for evaluation. Samples for training and evaluation were made by adding each target event to a temporally center-aligned background four times longer than the event itself.

The saliency-maximizing transformation learned by our proposed algorithm emphasizes nonspeech events and attenuates the background, which is strongly heterogeneous and has significant speech presence (Figure 3). The three target events in the figure are obscured in the conventional spectrogram, but instantly visible in the saliency-maximized spectrogram.

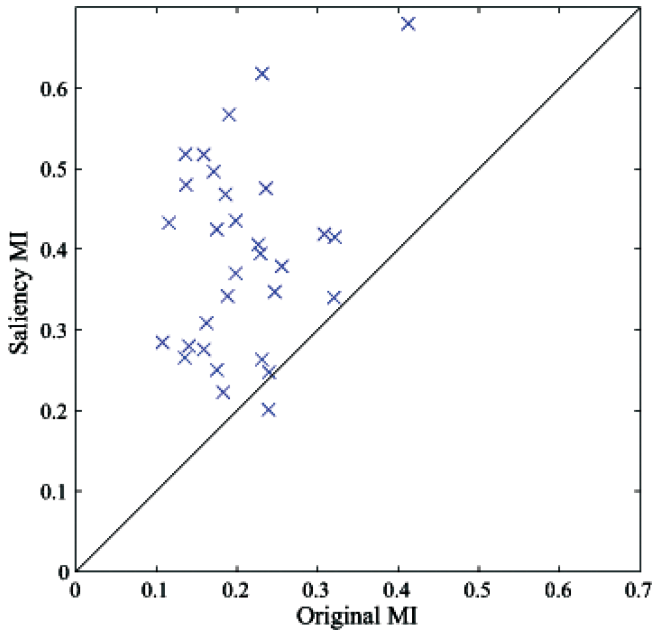


Fig. 4. Comparison of MI for 31 evaluation samples. Almost all samples yield a larger MI when the spectrogram is saliency maximized.

Our objective measure is the empirical MI between the saliency map of the spectrogram and the ground truth. (Each spectrogram has its own saliency map, generated by the same algorithm.) Figure 4 shows the quantitative improvement due to maximizing saliency. Both axes measure the $I(Y;\varphi)$ of evaluation samples. (Recall that neither these samples nor these backgrounds were used in training.)

The proposed algorithm most improves low-SNR target events that are still barely visible in the unenhanced spectrogram. The net improvement of the MI between conventional and saliency-maximized spectrograms of evaluation samples is maximized around –20dB SNR (Figure 5). At very low SNR, the target's visual pattern is too buried to be enhanced. Conversely, at very high SNR, the visual pattern is already so salient that little improvement is possible.

We compare the performance of the proposed algorithm with the alternatives discussed in Section 3. Figure 6 shows that the proposed method had the largest average MI improvement. NMFD and the event-specific Wiener filter were second best, the latter with smaller standard deviation. Compared to baseline, energy thresholding and event-independent Wiener filtering actually degraded performance. Because the proposed method actually uses MI to improve the spectrogram, it may be unfairly favored in the comparison of Figure 6. Nevertheless, these results are corroborated by another measure of performance—the correlation coefficient. Figure 7 shows that NMFD had the largest average correlation improvement. The proposed method and event-specific Wiener filtering performed similarly, while energy thresholding and event-independent Wiener filtering again perform worse than baseline. This order may be explained because the oracular event-specific Wiener filter uses the most information about the evaluation spectrogram; NMFD also uses correlation with ground truth to choose the spectrogram's best partial reconstruction. This provides crucial prior knowledge about the target event's temporal location and visual pattern, for choosing the most related components for partial reconstruction. Of the three enhancements trained without such "cheating," energy thresholding

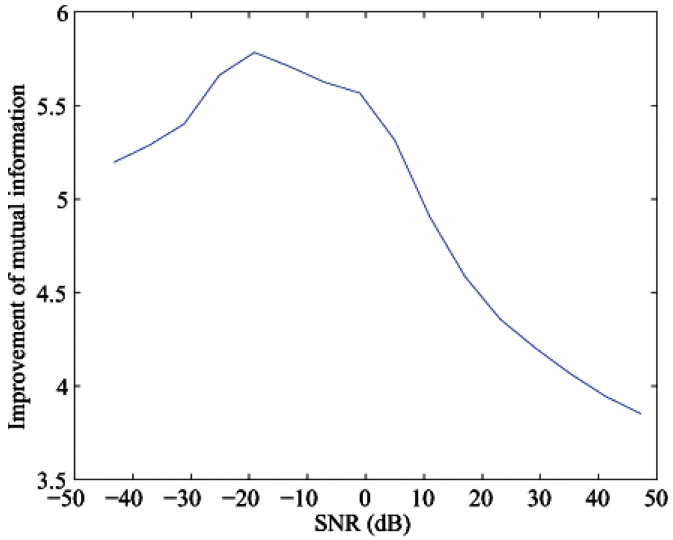


Fig. 5. Sensitivity to SNR of MI improvement due to saliency maximization.

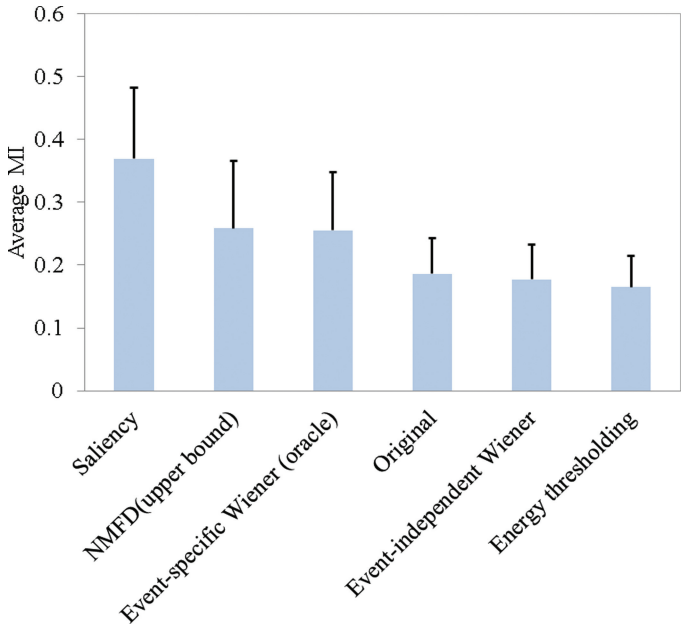


Fig. 6. Average MI using different methods (error bars indicate standard deviation).

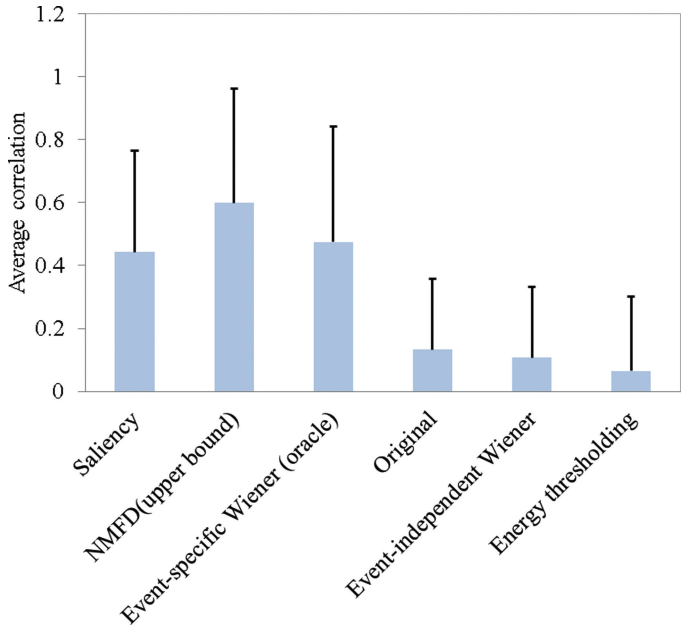


Fig. 7. Average correlation coefficients using different methods (error bars indicate standard deviation).

and event-independent Wiener filtering performed poorly, whereas the proposed algorithm performed best. The proposed method, using linear filters, is also among the computationally fastest of these visualizations. In particular, it is four orders of magnitude faster than the NMFD that we used (non-negativity constraints often slow down convergence). Note also that what the proposed method tries to increase is a target event's saliency; except for energy thresholding, the alternatives increase an event's SNR—a different (and perhaps harder) problem.

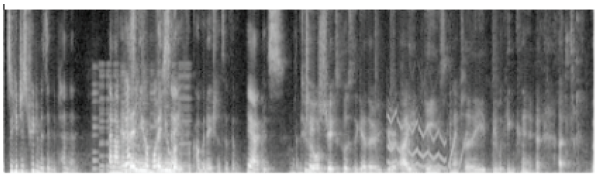
5. VISUALIZATION-GUIDED AUDIO BROWSER

The Timeliner audio browser (Figure 9) displays the waveform of a long audio recording, labeled in units ranging from weeks down to milliseconds depending on the current zoom level [Goudeseune 2012]. Timeliner also displays audio visualization features such as audio spectrograms or the outputs of event classifiers. The user can smoothly zoom in to interesting subintervals, to find even very brief anomalous segments without long periods of listening.

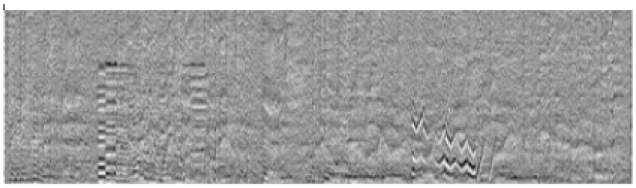
In conventional audio editors and browsers, the color of each pixel or texel is computed by undersampling data from the corresponding time interval. In Timeliner, such an interval might be tens of minutes long, and naïve undersampling flickers distractingly when panning and zooming at 60 frames per second. This flickering entirely obscures the data until the pan or zoom stops, eliminating any benefit of these continuous gestures.

To restore smoothness and to improve performance for long recordings, Timeliner first computes a multiscale cache for the recording and for the derived features. Given any subinterval of the recording, this cache yields the minimum, mean, and maximum data values found during that interval, for either scalar or vector data. (The time taken to compute this min-mean-max is logarithmic with respect to the full recording's duration, and independent of the subinterval's duration.) Final rendering maps each pixel's or texel's triplet to a hue-saturation-value color through a predefined transfer function.

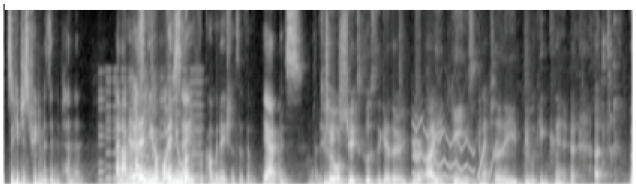
(a) Spectrogram of two target events in isolation.



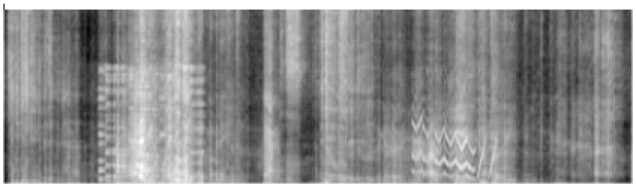
(b) Spectrogram of the same events with background noise.



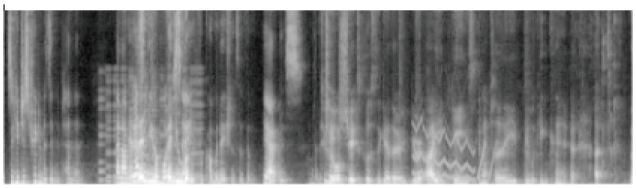
(c) Saliency-maximized spectrogram (our proposed method).



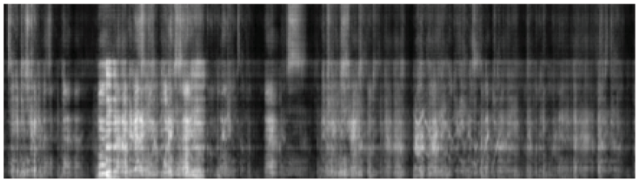
(d) Spectrogram enhanced with maximized MI, based on energy thresholding instead of visual saliency.



(e) Spectrogram enhanced by event-specific Wiener filtering (an oracle result).



(f) Spectrogram enhanced by event-independent Wiener filtering.



(g) Spectrogram enhanced by NFMD (upper bound).

Fig. 8. Various visualizations of spectrograms.

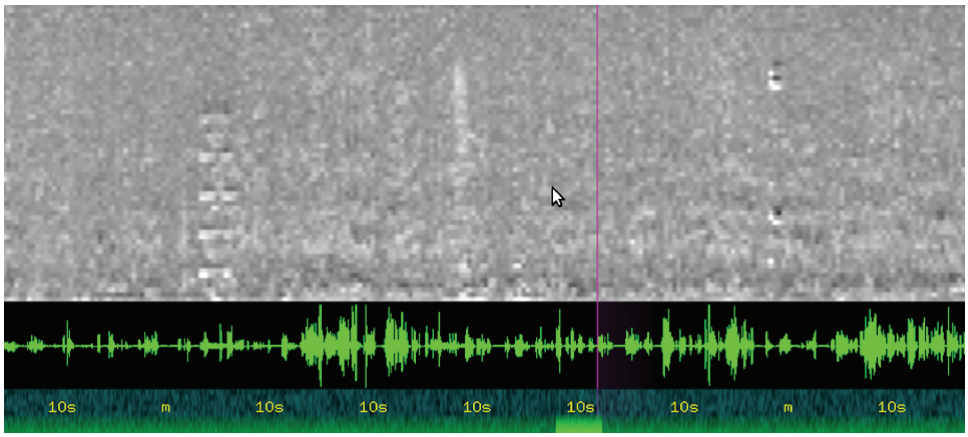


Fig. 9. Components of Timeliner's interface: saliency-maximized spectrogram, waveform, and time axis.

Fig. 10. Configuration of the human subject experiment.

Inspired by the left hand on keyboard/right hand on mouse layout that emerged in 1990s real-time games, Timeliner's two-handed input frees the user's gaze from hunting for keys. The "WASD" keys pan and zoom, and the spacebar starts and pauses audio playback. The mouse and its scrollwheel also pan and zoom, so users can choose whatever modality they find most familiar.

Timeliner's file parsing and user interface are implemented in the scripting language Ruby, whereas its heavier computation is done in C++ to reduce memory usage. Graphics are rendered with OpenGL and its utility toolkit GLUT. Timeliner runs natively on Linux and Windows, and is distributed open source.

6. SUBJECTIVE EVALUATION

We measured human subjects' AED performance with both saliency-maximized and conventional spectrograms, using an otherwise identical computer interface. Timeliner enabled convenient audio browsing. Video was presented with a 17-inch CRT and audio with ear buds (Figure 10).

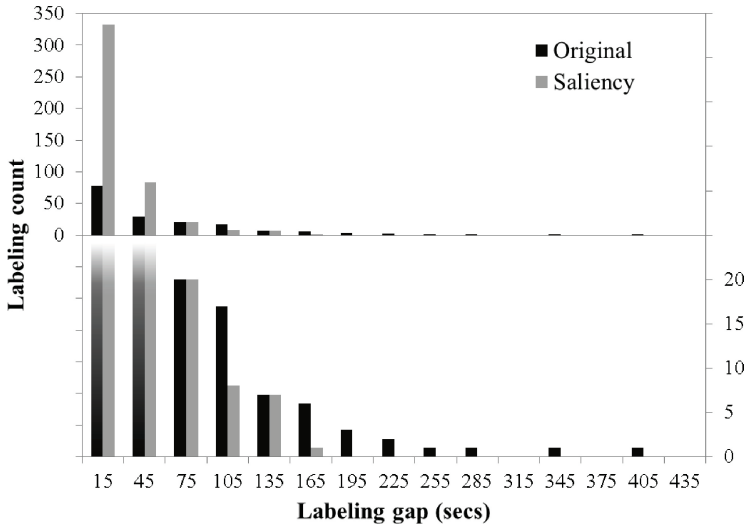


Fig. 11. Histogram of durations between consecutive correct labelings. The lower subfigure's ordinate is magnified to better display values at gaps over 75 seconds.

We asked 12 subjects, who were unfamiliar with spectrograms, to detect anomalous target events in 80-minute recordings of seminar room background noise [Carletta 2007]. Into each recording, we mixed 40 sound effects randomly chosen from the objective evaluation set of 31 different ones, uniformly distributed but without overlap, at various amplitudes. Because the task lasted only 8 minutes, naïve listening (real-time search) would expect to find only a 10th of the targets. We therefore instructed subjects to first scan for a visually suspicious pattern and then verify it by listening before annotating that target's temporal position.

Each subject annotated six different recordings, using either three saliency-maximized followed by three conventional spectrograms, or the reverse order. This ordering was balanced across subjects. The first and the fourth sessions were just for practice with each visualization; only the other sessions were evaluated for performance. The recordings used in the non–practice sessions were balanced across subjects. To restore subjects' vigilance and to reconfigure the computer between sessions, subjects rested for about 1 minute, or longer if they desired. (This is also why we did not use one very long session.) Afterward, subjects were asked which spectrogram was more helpful (we explained nothing to them about spectrograms or saliency). All preferred the saliency-maximized one.

To quantify subjects' AED performance from their annotated timestamps, we computed their recall and their precision. Recall was the fraction of targets whose durations contained a timestamp (how many were hit). Precision was the fraction of timestamps that were in some target (hits per try). A subject's F-score was the harmonic mean of their precision and recall [Rijsbergen 1979].

The experiment was a within-subject design, comparing subjects' F-scores when using either conventional or saliency-maximized spectrograms. A paired samples t-test revealed a significant difference in the F-scores for conventional ($M = 0.2752$, $SD = 0.0777$) and saliency-maximized ($M = 0.5846$, $SD = 0.1033$) spectrograms; $t(11) = -12.976$, $p < .05$. This suggests that the saliency-maximized spectrogram significantly outperformed the conventional one.

Maximizing saliency increased not only F-scores but also stability. Figure 11 shows that it increased the number of events found within 30 seconds, from 47% to 74%, and also decreased the longest time

between detections, from 5 minutes to 3 minutes. The longer tail for the histogram of the conventional spectrogram correlates with the frustration that many subjects reported while using it.

7. CONCLUSION

Our proposed saliency-maximized spectrogram enables audio browsing that is much faster than real time. In AED, it improves the MI between the ground truth of non–background sounds and visual saliency, more than other common enhancements to visualization. In a 1/10-real-time AED task, compared to conventional spectrograms, it increased stability and improved subjects' F-score by 100%.

We plan to extend this enhancement of visualization to time series derived from nonacoustic sensors. We also wish to use more refined models of human perception, such as color saliency, although these will require more training data.

Finally, we wish to apply this enhancement to more realistic audio events. (It increased the saliency of a few background sounds, notably rapidly rising pitch in female speech and squeaky writing on a whiteboard; these were indeed sometimes mislabeled as anomalous.) As nonelectronic sounds have subtler visual patterns, they may demand a transformation more elaborate than the current 5×5 linear filter and a concomitantly larger training dataset.

REFERENCES

ABOUCHACRA, K. S., LETOWSKI, T., AND MERMAGEN, T. 2007. Detection and localization of magazine insertion clicks in various environmental noises. *Mil. Psychol. 19,* 3, 197–216.

ANDERSON, J. R. 2009. *Cognitive Psychology and Its Implications*. Worth.

ARONS, B. 1997. Speechskimmer: A system for interactively skimming recorded speech. *ACM Trans. Comput. Hum. Interact. 4,* 1, 3–38.

BELOPOLSKY, A. V., KRAMER, A. F., AND GODIJN, R. 2008. Transfer of information into working memory during attentional capture. *Vis. Cognit. 16,* 4, 409–418.

BERRY, M., BROWNE, M., LANGVILLE, A., PAUCA, V., AND PLEMMONS, R. 2007. Algorithms and applications for approximate nonnegative matrix factorization. *Comput. Stat. Data Anal. 52,* 1, 155–173.

BORJI, A. AND ITTI, L. 2013. State-of-the-art in visual attention modeling. *IEEE Trans. Pattern Anal. Mach. Intell. 35,* 1, 185–207.

BOVIK, A. 2009. *The Essential Guide to Image Processing*. Academic Press.

CARLETTA, J. 2007. Unleashing the killer corpus: Experiences in creating the multi-everything AMI meeting corpus. *Lang. Resour. Eval. 41,* 2, 181–190.

COVER, T. M. AND THOMAS, J. A. 2006. *Elements of Information Theory*. Wiley-Interscience.

FRINTROP, S., ROME, E., AND CHRISTENSEN, H. I. 2010. Computational visual attention systems and their cognitive foundations: A survey. *ACM Trans. Appl. Percept. 7,* 1, 6:1–6:39.

GOLDSTEIN, E. B. 2010. *Sensation and Perception*. Wadsworth.

GONZALEZ, R. C. AND WOODS, R. E. 2001. *Digital Image Processing* (2nd ed.). Addison-Wesley Longman, Boston, MA.

GOUDESEUNE, C. 2012. Effective browsing of long audio recordings. In *Proceedings of the 2nd ACM International Workshop on Interactive Multimedia on Mobile and Portable Devices (IMMPD'12)*. ACM, New York, 35–42.

HASEGAWA-JOHNSON, M. A., GOUDESEUNE, C., COLE, J., KACZMARSKI, H., KIM, H., KING, S., MAHRT, T., HUANG, J.-T., ZHUANG, X., LIN, K.-H., SHARMA, H. V., LI, Z., AND HUANG, T. S. 2011. Multimodal speech and audio user interfaces for K-12 outreach. In *Proceedings of the Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC'11).* 526–531.

ITTI, L. AND KOCH, C. 2001. Feature combination strategies for saliency-based visual attention systems. *J. Electron. Imaging 10,* 1, 161–169.

ITTI, L., KOCH, C., AND NIEBUR, E. 1998. A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. Pattern Anal. Mach. Intell. 20,* 11, 1254–1259.

JÄNICKE, H. AND CHEN, M. 2010. A salience-based quality metric for visualization. *Comput. Graphics Forum 29,* 3, 1183–1192.

LEE, D. D. AND SEUNG, H. S. 1999. Learning the parts of objects by non-negative matrix factorization. *Nature 401,* 6755, 788–791.

LIM, J. S. 1990. *Two-Dimensional Signal and Image Processing*. Prentice Hall.

LIN, K.-H., ZHUANG, X., GOUDESEUNE, C., KING, S., HASEGAWA-JOHNSON, M., AND HUANG, T. S. 2012. Improving faster-than-real-time human acoustic event detection by saliency-maximized audio visualization. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'12).* 2277–2280.

MILLER, G. 1956. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychol. Rev. 63,* 2, 81–97.

OTSU, N. 1979. A threshold selection method from gray-level histograms. *IEEE Trans. Syst. Man Cybern. A, Syst. Humans 9,* 1, 62–66.

RIJSBERGEN, C. J. V. 1979. *Information Retrieval* (2nd ed.). Butterworth-Heinemann, Newton, MA.

ROSENHOLTZ, R., DORAI, A., AND FREEMAN, R. 2011. Do predictions of visual perception aid design? *ACM Trans. Appl. Percept. 8,* 2, 1–20.

SHANNON, C. AND WEAVER, W. 1949. *The Mathematical Theory of Communication*. University of Illinois Press, Urbana.

SMARAGDIS, P. 2004. Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs. In C. G. Puntonet and A. Prieto, Eds., *Independent Component Analysis and Blind Signal Separation*. Springer, 494–499.

SMARAGDIS, P. AND BROWN, J. C. 2003. Non-negative matrix factorization for polyphonic music transcription. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*.

SMITH, J. A., EARIS, J. E., AND WOODCOCK, A. A. 2006. Establishing a gold standard for manual cough counting: Video versus digital audio recordings. *Cough 2*, 6:1–6.

TEMKO, A. 2007. *CLEAR 2007 AED Evaluation Plan and Workshop*.

TEMKO, A., MALKIN, R., ZIEGER, C., MACHO, D., NADEU, C., AND OMOLOGO, M. 2006. Acoustic event detection and classification in smart-room environments: Evaluation of CHIL project systems. *Cough 65,* 48, 5.

WALTHER, D. AND KOCH, C. 2006. Modeling attention to salient proto-objects. *Neural Netw. 19,* 9, 1395–1407.

WICKENS, C. D. AND HOLLANDS, J. G. 1999. *Engineering Psychology and Human Performance* (3rd ed). Prentice Hall.

ZHOU, X., ZHUANG, X., LIU, M., TANG, H., HASEGAWA-JOHNSON, M. A., AND HUANG, T. S. 2007. HMM-based acoustic event detection with AdaBoost feature selection. In *Proceedings of the Classification of Events, Activities and Relationships Evaluation and Workshop*.

Received January 2013; revised May 2013; accepted July 2013