IMPROVING FASTER-THAN-REAL-TIME HUMAN ACOUSTIC EVENT DETECTION BY SALIENCY-MAXIMIZED AUDIO VISUALIZATION

Kai-Hsiang Lin, Xiaodan Zhuang, Camille Goudeseune, Sarah King, Mark Hasegawa-Johnson, and Thomas S. Huang*

University of Illinois at Urbana-Champaign**

{klin21, xzhuang2, cog, sborys, jhasegaw, t-huang1}@illinois.edu

ABSTRACT

We propose a saliency-maximized audio spectrogram as a representation that lets human analysts quickly search for and detect events in audio recordings. By rendering target events as visually salient patterns, this representation minimizes the time and effort needed to examine a recording. In particular, we propose a transformation of a conventional spectrogram that maximizes the mutual information between the spectrograms of isolated target events and the estimated saliency of the overall visual representation. When subjects are shown spectrograms that are saliency-maximized, they perform significantly better in a 1/10-real-time acoustic event detection task.

Index Terms— acoustic event detection, visual saliency, audio visualization

1. INTRODUCTION

Acoustic event detection (AED) is the detection of non-speech events in long audio recordings. Automatic AED is difficult: in the 2007 Classification of Events, Activities and Relationships (CLEAR) evaluations, namely detecting predefined acoustic events in continuous real seminar audio recordings, no system's accuracy greatly exceeded 30% [1].

In many detection tasks human perception outperforms machine perception, as it better handles the semantic gap between noisy observations and target events. For example, rifle magazine insertion clicks are detected with 100% accuracy at 0 dB SNR in white noise, babble, or jungle noise [2]. Humans can detect an anomalous sound even on its first audition [3].

However, for long audio recordings, shortcuts like high-speed playback are limited: most people cannot comprehend continuous speech faster than twice normal speed [4]. Worse yet, even after detecting an event in a relatively long segment, pinpointing the event's timestamp usually requires rewinding and replaying. Our preliminary experiments show that pure-listening AED is considerably slower than real-time playback.

However, this real-time barrier to human AED can be broken by enlisting human vision. To this end, we propose a visualization, a saliency-optimized audio spectrogram that can be examined at different temporal scales to efficiently eliminate uninteresting regions. This visualization is synchronized with the source audio recording, so an analyst searching for target events typically listens to only brief excerpts of visually interesting segments. This is so because the target events' information is embedded into visually salient patterns, which are processed by human vision with priority [5].

We formulate this visualization problem as maximizing the mutual information (MI) between the spectrogram of the target events Y and the estimated visual saliency of the examined spectrogram $\varphi(f)$ (Fig. 1). The input information Y is the spectrogram of the isolated target events in isolation (without the background noise N); the transmitted information is the visual perception by the observer. 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 results from the bottom-up attention of the human visual system. After the saliency model, (information in) the salient regions, e.g., a target event, will be recognized as Z.

2. PROPOSED METHOD

We approximate the human visual system with a communication channel that selectively attends to visual patterns in decreasing order of perceptual saliency. Therefore, when quickly examining a display, it perceives at most a few highly salient objects. The rate of information transmission is limited by finite span of attention (about six objects at a glance), and by immediate memory (about seven items) [6]. Salient patterns are processed first and are therefore more likely to be transmitted through this channel. Our algorithm uses a computational saliency model to simulate this process and generate the saliency distribution of an image.

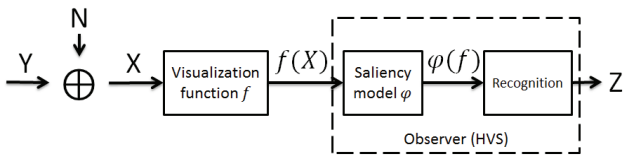


Fig. 1. Flowchart of human AED from a visual display, $f(X)$.

* Currently with Raytheon BBN Technologies.

** This work is funded by National Science Foundation grant 0807329.

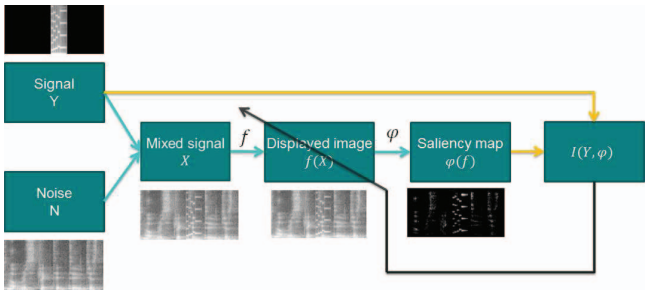


Fig. 2. Flowchart of the proposed algorithm.

The saliency model has been used to analyze the effectiveness of a visual representation. For example, Jänicke and Chen proposed a salience-based quality metric for visualization using the correspondence between the data relevance mask and the saliency map [7]. Although there have been some works on analyzing the quality of a visual representation based on saliency, there is still no good way to automatically generate the visual representation of data which is saliency-maximized.

We propose to measure the efficiency of information transmission from Y to φ through their MI. Choosing the visualization (encoding) function f that maximizes $I(Y;\varphi)$ then lets the saliency-optimized spectrogram represent the target events optimally for fast human visual examination.

Our task can be formulated as:

$$f^* = \underset{f}{\operatorname{argmax}}\, E_{X,Y}\{I(Y;\varphi(f(X)))\} \tag{1}$$

where X is the input spectrogram, Y is the ground truth (the spectrogram of the isolated acoustic event), and $f(X)$ is the displayed spectrogram, a transformation of X. The MI between the saliency map and the ground truth is thus $E\{I(Y;\varphi)\}$. 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 (Fig. 2).

2.1. Computing and transforming the spectrogram

We based our visualization on the humble spectrogram because it is familiar to audio experts, and because we found that even naïve subjects could successfully interpret details in such a time-frequency plot. Our grayscale spectrogram resolves 128 frequency bands down to 5 msec.

Our goal is to find a saliency-maximized transformation function f. This ensures that the displayed signal renders target events so that φ extracts them as salient patterns. In other words, $f(X)$ is displayed, but $\varphi(f)$ is perceived. For simplicity we use 2D linear filters: $f(X) = h[n_1, n_2] **X[n_1, n_2]$, where Eq. (1) optimizes h.

2.2. Computing the saliency map

Transforming the spectrogram of the mixed signal X yields the displayed image $f(X)$. We use an image saliency algorithm to generate the saliency map, which approximates the bottom-up attention of the human visual system. Following the framework of [8] and [9], this algorithm has three steps: extracting image features; building feature pyramids and computing each feature's center-surround difference; and combining all features' saliency maps into a single map.

The two features used by the algorithm, orientation (from Gabor-filtering different scales) and image intensity, are similar to [8].

Because a salient region must somehow differ from its neighborhood, the algorithm detects saliency with a center-surround difference (CSD) implemented by a difference of Gaussians:

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

The DoG is parameterized by two σ's, the center layer and surround layer. For computational efficiency, a Gaussian pyramid generates images filtered by Gaussians of different σ's. We use the pyramid's first and fourth layers as center and surround.

Computation of the CSD's can be formulated as:

$$\mathrm{CSD}_k = \max\{0, F_{c,k} \ominus F_{s,k}\}, \quad k \in \{I,\ O\}$$

where $F_{c,k}$ and $F_{s,k}$ are the center and surround layers of the Gaussian pyramid for feature k, and $\ominus$ denotes across-scale subtraction. Intensity is denoted by I, the four orientations by O. We combine the CSD's of different features into saliency maps, normalizing with $N(\cdot)$ before every summation [9]:

$$\begin{aligned} F_k &= N(\mathrm{CSD}_k), && k \in \{0°, 45°, 90°, 135°, I\} \\ F_O &= N\left(\textstyle\sum_k F_k\right), && k \in \{0°, 45°, 90°, 135°\} \\ S &= (F_I + F_O)/2 \end{aligned}$$

Thus, the final saliency map S is the mean of the maps for intensity and for the combined orientations.

2.3. Maximizing mutual information

We evaluate how well human visual perception captures the information in the visualization associated with the target events. This is estimated with the MI between the ground truth Y (the spectrogram of the isolated target event obtained according to Sec. 2.1), and the saliency map X of the mixed spectrogram:

$$I(X;Y) = \sum_{x\in X,\ y\in Y} p(x,y)\log\frac{p(x,y)}{p(x)p(y)}$$

Here $p(x,y)$, $p(x)$ and $p(y)$ are the joint and marginal distributions of the pixel intensity of these two images.

Because the objective function $E\{I(Y;\varphi)\}$ in Eq. (1) is non-convex and non-differentiable, we merely 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 (using the conventional spectrogram). Only the training data is used to optimize f.

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

3. EXPERIMENTAL DESIGN

To evaluate the proposed algorithm, we compared the conventional and the saliency-maximized spectrograms both objectively and subjectively. Objective comparison measured the $I(Y;\varphi)$ of target events. Subjective comparison was the F-score of humans using either spectrogram to detect acoustic events.

We simulated data for this task using sound effects as target events superimposed on ("mixed with") the realistically noisy background of a seminar room [10]. All 62 sound effects were obviously foreign to a seminar room. Both the target events and the background audio were split into disjoint sets for training and for evaluation. For objective comparison, we made training and evaluation samples by mixing each target event with a temporally center-aligned background of four times the event's duration. Section 4.3 describes the data for subjective evaluation.

4. EXPERIMENTAL RESULTS

4.1. Saliency-maximized spectrogram

The saliency-maximizing transformation learned by the proposed algorithm attenuates background speech and emphasizes non-speech events (Fig. 3). The three target events are obscured in the conventional spectrogram, but instantly visible in the saliency-maximized spectrogram. Conventional image enhancements such as edge detection and Wiener filters are much less effective.

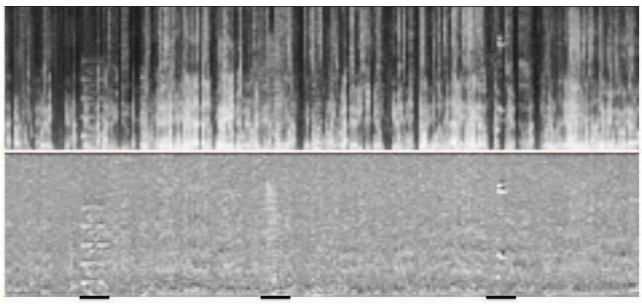


Fig. 3. Qualitative improvement of a spectrogram: conventional (top), saliency-maximized (bottom). Three target events are marked with black underlines.

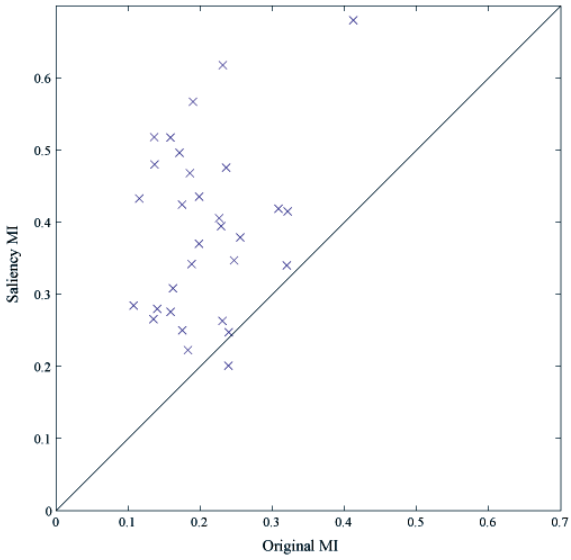


Fig. 4. Comparison of MI for 31 evaluation samples. Almost all samples yield a larger MI when the spectrogram is saliency-maximized.

4.2. Objective measures

Our objective measure is the empirical MI between the spectrogram and the ground truth. Fig. 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 the background audio used in this evaluation were used in training.)

4.3. Subjective experiments

We measured human subjects' AED performance with both kinds of spectrogram, using an otherwise identical computer interface.

To let a human subject conveniently browse audio, we developed an audio visualization interface called Timeliner [3]. This shows a multi-hour recording that subjects can smoothly and rapidly zoom temporally, over a range from 10 μs to tens of minutes per horizontal pixel. Subjects can also listen to any part of the recording (Fig. 5). For these experiments the display device was a 17-inch CRT screen, while audio was presented with ear buds (Fig. 6).

We asked twelve subjects, unfamiliar with spectrograms, to detect anomalous target events in 80-minute recordings of seminar room background noise. Into each recording we mixed 40 sound effects randomly chosen from the testing set, 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 tenth of the targets. We therefore instructed subjects to first scan for a visually suspicious pattern and then verify it by listening, before annotating the temporal position of that target.

Each subject annotated four different recordings, either

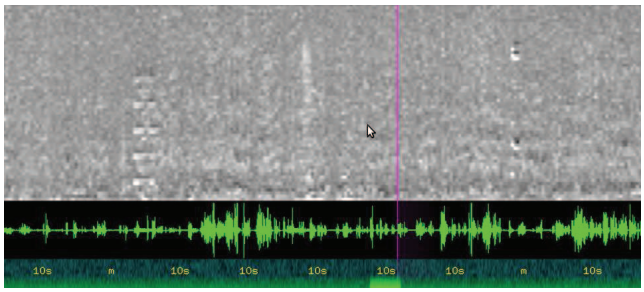


Fig. 5. Components of Timeliner's interface: spectrogram, waveform, and time-axis.

Fig. 6. Configuration of the human subject experiment.

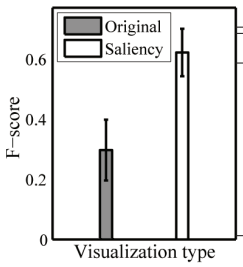
two saliency-maximized followed by two conventional spectrograms, or in the reverse order. This was balanced across subjects. Afterwards, subjects were asked which spectrogram was more helpful (we explained nothing to them about spectrograms or saliency). All preferred the saliency-maximized one.

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 [11].

The F-scores' analysis of variance used three factors: spectrogram *type* (conventional or saliency-maximized), *order* of presentation of spectrogram type, and which *recordings* were used (Fig. 7b). The saliency-maximized spectrogram significantly outperformed the conventional one, with no significant interaction between these three factors.

5. CONCLUSION

The proposed saliency-maximized audio spectrogram enables much-faster-than-real-time audio browsing by rendering target events as salient patterns processed with priority by the human visual system. In a 1/10-real-time AED task, human subjects achieved 100% relative improvement in event detection F-score with the saliency-maximized spectrogram, as compared to the conventional spectrogram (Fig. 7a).



(a)

Source	d.f.	F	P
Type	1	142.65	<0.001
Order	1	0.0	0.978
Recording	3	0.3	0.823
Type × Order	1	0.4	0.533
Type × Rec.	3	1.86	0.154
Order × Rec.	3	0.29	0.836

(b)

Fig. 7. (a) F-scores for the AED task (error bars indicate standard deviation); (b) ANOVA of the F-scores.

6. REFERENCES

[1] X Zhuang, X Zhou, M A Hasegawa-Johnson, and T S Huang, "Real-world acoustic event detection," *Pattern Recognition Letters*, vol. 31, no. 12, pp. 1543–1551, 2010.

[2] K S Abouchacra, T Letowski, and T Mermagen, "Detection and localization of magazine insertion clicks in various environmental noises," *Military Psychology*, vol. 19, no. 3, pp. 197–216, 2007.

[3] M Hasegawa-Johnson, C Goudeseune, J Cole, H Kaczmarski, H Kim, S King, T Mahrt, J Huang, X Zhuang, K Lin, H Sharma, Z Li, and T Huang, "Multimodal Speech and Audio User Interfaces for K-12 Outreach," in *APSIPA ASC*, 2011.

[4] B Arons, "SpeechSkimmer: a system for interactively skimming recorded speech," *ACM Transactions on Computer-Human Interaction (TOCHI)*, vol. 4, no. 1, pp. 3–38, 1997.

[5] E B Goldstein, *Sensation and perception*, Wadsworth, 8th edition, 2009.

[6] G A Miller, "The magical number seven, plus or minus two: some limits on our capacity for processing information," *Psychological Review*, vol. 63, no. 2, pp. 81, 1956.

[7] H Jänicke and M Chen, "A salience-based quality metric for visualization," *Computer Graphics Forum*, vol. 29, no. 3, pp. 1183–1192, 2010.

[8] L Itti, C Koch, and E Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998.

[9] D Walther and C Koch, "Modeling attention to salient proto-objects," *Neural Networks*, vol. 19, no. 9, pp. 1395–1407, 2006.

[10] J Carletta, "Unleashing the killer corpus: experiences in creating the multi-everything AMI Meeting Corpus," *Language Resources and Evaluation Journal*, vol. 41, no. 2, pp. 181–190, 2007.

[11] C J van Rijsbergen, *Information Retrieval*, Butterworth-Heinemann, 2nd edition, 1979.