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

In this work, we break the real-time barrier of human audition by producing rapidly searchable visualizations of the audio signal. We propose a saliency-maximized audio spectrogram as a visual representation that enables fast detection of audio events by a human analyst. This representation minimizes the time needed to examine a particular audio segment by embedding the information of the target events into visually salient patterns. In particular, we find a visualization function that transforms the original mixed spectrogram to maximize the mutual information between the label sequence of target events and the estimated visual saliency of the spectrogram features. Subject experiments using our human acoustic event detection software show that the saliency-maximized spectrogram significantly outperforms the original spectrogram 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 acoustic event detection is difficult. For example, in the 2007 Classification of Events, Activities and Relationships (CLEAR) Evaluations, even the top rated AED system achieved only around 30% accuracy for detection of predefined acoustic events in continuous real seminar audio recordings [1].

Human perception outperforms machine perception in many detection tasks as it is better at dealing with the semantic gap between the noisy acoustic observations and the 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]. Further, humans recognize some audio anomalies even if the sound has not been heard before [3].

However, the human auditory system has two intrinsic constraints. First, most people are only capable of listening to one sound at a time, making the rapid browsing of large audio databases much harder than the parallel browsing of images or video. Second, playing back audio with higher speed is not a very effective option. For example, comprehension of continuous speech is feasible for most people only up to about 2X normal speed [4]. For AED, even after hearing an acoustic event in a relatively long audio segment, the timestamp of the target event is not easy to find out without replaying the audio. According to our preliminary experiments, detecting acoustic events by listening is much slower than real time.

In this work, we break the real-time barrier of human audition by engaging both human audition and vision. The audio signal is visualized using an innovative saliency-optimized audio spectrogram that can be visually examined at different temporal scales to efficiently eliminate uninteresting areas. The human analysts also have access to the original audio signal, i.e., they can play back the visually interesting segments, usually only necessary for a short duration to find the target events.

We propose the saliency-optimized audio spectrogram as the key to effective audio visualization that enables fast browsing and detection of audio events by human analysts. This representation minimizes the time needed to examine a particular audio segment by embedding the information of the target events into visually salient patterns. As salient patterns are processed by human vision with priority [5], the human analyst only needs to spend the time sufficient to examine these patterns in order to detect the events.

As illustrated in Fig. 1, we formulate the problem as maximizing the mutual information (MI) between the spectrogram of the target events Y and the estimated visual saliency of the spectrogram $\varphi(f)$ to be examined. The input information Y is the spectrogram of the clean audio for target events and the transmitted information is the visual perception by the observer. N is the background noise and X is the spectrogram of the mixed signal. f is the visualization function which will convert the spectrogram to the saliency-optimized audio spectrogram. $\varphi(f)$ is the saliency map, i.e., 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.

The human visual system (HVS) is approximated as a communication channel that selectively attends to visual patterns in decreasing order of perceptual salience, and is therefore limited to the perception of only a few highly salient objects when the time available for examining a display is limited. The maximum rate of information transmission is limited due to the finite span of attention and immediate memory. (Span of attention will encompass about six objects at a glance and span

of immediate memory is about seven items in length [6].) The salient patterns are processed first and therefore are more likely to be transmitted through this communication channel. In our algorithm a computational saliency model is used to simulate this process and generate the saliency distribution of an image.

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 measuring the efficiency of information transmission from Y to φ using MI between them. By optimizing the visualization (encoding) function $f(x)$ to maximize $I(Y;\varphi)$, the saliency-optimized spectrogram embeds the target event information into a representation that is optimal for fast human visual examination.

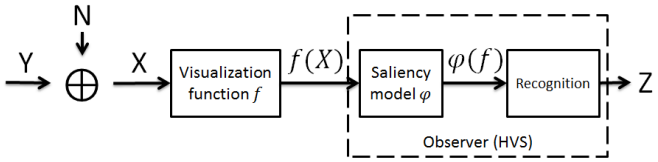


Fig. 1. Flowchart of human AED from a visual display, $f(X)$

2. PROPOSED METHOD

Our task can be formulated as:

$$f^* = arg \max_f E_{X,Y}\{I(Y,\ \varphi(f(X)))\} \quad (1)$$

where X is the input audio spectrogram, Y is the ground truth (the spectrogram of the acoustic event), $f(X)$ is the displayed image, which is a transformed version of X. $\varphi(f)$ is the saliency map and $E\{I(Y,\varphi)\}$ is the MI between the saliency map and the ground truth. As in Fig 2, five modules are involved to solve for the optimized transformation function f in Eq (1): Spectrogram computation, visualization transformation, saliency map computation, MI computation and MI maximization.

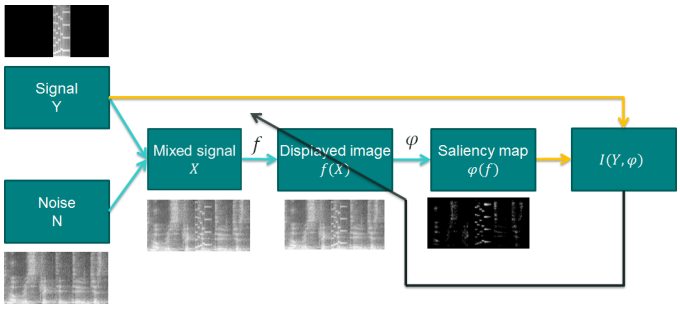


Fig. 2. Flowchart of the proposed algorithm

2.1. Spectrogram computation

We choose spectrogram as our basic visualization for the following reasons: Audio signal experts know how to get information from spectrograms, and in our preliminary experiments even naive human subjects prefer a time-frequency plot to any other display. In our experiments, the spectrogram is generated with sampling rate of 5 msec and the frequency resolution is 128 frequency bands. Spectrograms are saved as grayscale images.

2.2. Visualization transformation

The goal of this work is to find a saliency-maximized transformation function f. f is the key to ensure that the displayed signal embed target event info in a way that will be extracted by φ as salient patterns ($f(X)$ is what is actually displayed, but $\varphi(f)$ is what is perceived). For simplicity, we use 2D linear filters in this work, i.e., $f(X) = h[n_1, n_2] * *X[n_1, n_2]$, and Eq. (1) optimizes $h[n_1, n_2]$.

2.3. Saliency map computation

After transforming the spectrogram of the mixed signal X, we get the displayed image $f(X)$. We use an image saliency algorithm to generate the saliency map, which approximates the perception of the human vision system. The image saliency algorithm follows the saliency framework developed by Itti et al. [8] and Walther and Koch [9]. It consists of three steps: extracting image features, building a feature pyramid and computing center-surround difference of each feature, and combining the saliency maps of all features into one saliency map.

We use two features similar to [8]: intensity of the grayscale spectrogram and orientation obtained by applying Gabor filters to different scales of the spectrogram. Color information is not used in this work.

A salient region has some property different from its neighborhood. To detect saliency, the algorithm uses center-surround difference (CSD), which is implemented by difference of Gaussian (DoG).

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

The DoG is parameterized by two σ's which correspond to the center layer and surround layer. For computational efficiency, a Gaussian pyramid is used to generate images filtered by Gaussians of different σ's. In our experiments, the center and surround layers are the first and fourth layers of the Gaussian pyramid.

The computation of CSD can be formulated as:

$$CSD_k = max\,(0, F_{c,k} \ominus F_{s,k}), \quad k \in \{I,\ O\} \quad (3)$$

where $F_{c,k}$ and $F_{s,k}$ are center and surround layers of Gaussian pyramid of feature k. I is intensity and O is orientations including 0°,45°,90°, and 135°. $\ominus$ stands for across-scale subtraction.

CSDs of different features are combined as:

$$F_k = N\left(CSD_k\right), \quad k \in \left\{I,\ 0^\circ, 45^\circ, 90^\circ, 135^\circ\right\}$$
$$F_o = N\left(\sum_j F_j\right), \quad j \in \{0^\circ, 45^\circ, 90^\circ, 135^\circ\} \tag{4}$$
$$S = \frac{1}{2}(F_I + F_o)$$

where F_I is intensity saliency map, F_o is orientation saliency map after combining saliency maps of four orientations, and $N(\cdot)$ is the normalization operation. The final saliency map S is the average of the intensity and orientation saliency maps. We apply a normalization operation before summing maps of different features [9].

2.4. Mutual information maximization

With the ground truth, i.e., the spectrogram of the clean target event, obtained according to 2.1 and the saliency map of the mixed spectrogram generated in 2.3, we can evaluate, for each transformed image, the extent to which human subjects will pay attention to the information associated with the target acoustic event. In this work, we use mutual information between the saliency map and the ground truth as such a measure:

$$I\left(X,Y\right) = \sum_{y\in Y}\sum_{x\in X} p(x,y)\log(\frac{p(x,y)}{p(x)p(y)}) \tag{5}$$

where X is the saliency map and Y is the spectrogram of the clean target event. $p(x,y)$, $p(x)$ and $p(y)$ are the joint distributions and the marginal distributions of the intensity of these two images.

The objective function in Eq. (1) is non-convex and non-differentiable. We use simulated annealing to estimate f in order to approximate the global optimum. The initial transformation is set to $h[n_1,n_2] = \delta[n_1,n_2]$, which is also the baseline transformation.

In our experiments, linear filters of different sizes from 5x5 to 15x15 were evaluated. The optimized average MIs for these transformations are similar. The filter used in the subject experiments is set to size 5x5, chosen from several optimization trials by inspecting the visualizations generated from the training data.

3. EXPERIMENT

We evaluate the proposed algorithm with objective and subjective comparison between the original and the saliency-maximized spectrograms. For objective comparison, the $I(Y;\varphi)$ with the target event is compared between the original and the saliency-maximized spectrograms. In subjective comparison, human subjects were asked to detect acoustic events using either type of spectrograms separately and the result is measured by the F-score.

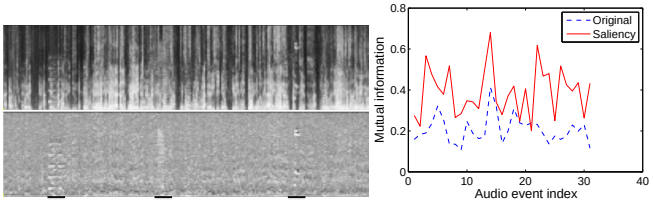
3.1. Evaluation data set

We simulate data for this task using artificial sound effects as the target acoustic event, and seminar room recordings as the background. There are 62 different sound effects, clearly not belonging to the seminar room scenario, split into two non-overlapping sets for training and evaluation. The background in this data set is selected from the AMI Corpus [1], and is quite realistic and noisy. Background data used in training and evaluation are also non-overlapping. Within the training and evaluation sets respectively, the target events are mixed with the background audio. For objective comparison, the training and evaluation samples are obtained by mixing each target event, temporally center-aligned, into background audio of four times the length. The data for subjective evaluation is detailed in Sec 3.2.3.

3.2. Experimental results

3.2.1. Saliency-maximized spectrogram

The saliency-maximizing transformation learned by the proposed algorithm attenuates background speech spectrograms, and visually emphasizes non-speech events. Fig. 3 (a)is used to demonstrate the qualitative improvement between the original and the saliency-maximized spectrograms. There are three acoustic events in this signal (labeled with black underline at the bottom), which are almost completely hidden in the original spectrogram. In contrast, the background audio is suppressed and the target events become more visible even at a quick glance in the saliency-maximized spectrogram.



(a) (b)

Fig. 3. (a) Original (top) v.s saliency-maximized (bottom) spectrograms; (b) Comparison of mutual information for each evaluation sample

3.2.2. Objective measures

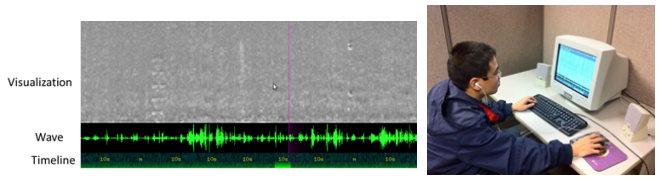
Fig. 3 (b) is used to demonstrate the quantitative improvement between the original and saliency-maximized spectrograms. The horizontal axis represents the 31 testing samples containing different sound effects and the vertical axis is their $I(Y;\varphi)$. None of these 31 test events are used in training, nor is there any train/test overlap in the background speech signals. The dashed curve is the MI between the spectrograms of the mixed signal and the ground truth, while the other curve between the saliency-maximized spectrogram and the ground truth. We can clearly observe that almost all saliency-maximized spectrograms have higher $I(Y;\varphi)$ than the corresponding original spectrograms.

3.2.3. Subjective experiments

We examined the human subjects' AED performance using an identical interface with either the original or the saliency-maximized spectrograms.

[1] http://corpus.amiproject.org

To enable a human subject to conveniently browse audio, we develop an audio visualization interface called Timeliner [3]. Timeliner (Fig 4) provides access to temporally-scaled audio visualization and audio playback at arbitrary times. The user can view a multi-hour recording in a single screen, and then smoothly and rapidly zoom temporally in to regions of interest, changing from scales coarser than ten minutes per pixel to as small as 10μs per pixel. A 17" CRT screen is used as the display and the user uses ear buds to listen to the audio (Fig 4 (b))



(a) (b)

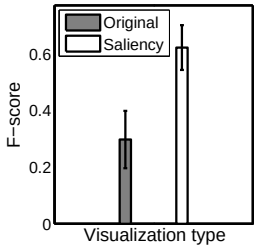
Fig. 4. (a) The interface of Timeliner; (b)The configuration of the human subject experiment

Twelve human subjects not familiar with spectrograms are asked to detect the audio events that should not appear in a meeting room. Each subject is asked to label four different files: two salience-maximized spectrograms then two original spectrograms, or in the reversed order. Such orders are balanced across all subjects. Each file is a 80-min seminar room recording, containing 40 instances of randomly seletected sound effects in the testing set. The sound effects are randomly adjusted to produce various SNR. The subject has eight minutes to detect events in each audio file. In order to do faster-than-real-time search, human subjects are instructed to leverage visually suspicious patterns on the displayed audio visualization and verify them by audio playback whenever necessary. Subjects need to label the temporal position after finding a target event.

After finish all the files, each subject is asked which of the two visualizations is more helpful. Without knowing the properties of the two types of visualization, all of them prefer the saliency-maximized spectrogram.

Human AED performance is measured by the F-score of the hypothesized event locations the subjects produce through the interface, relative to ground truth locations of all events. The F-score is the geometric mean of the precision and recall. Precision is the percentage of hypothesized timestamps falling into any ground truth event region over the total number of hypothesized timestamps, and recall is the percentage of the number of ground truth events that overlap with any hypothesized timestamp.

The F-scores of human AED using different visualization methods are illustrated in Figure 5 (a), error bars indicating standard deviation. A three-way ANOVA analysis (Fig. 5 (b)) has three factors: visualization type (original or saliency-maximized spectrogram), subject experience order (exposure to files using the original or the saliency-maximized spectrogram first) and file ID (audio content). The saliency-maximized spectrogram significantly outperforms the original spectrogram, and there is no significant interaction between the three factors.



(a)

Source	d.f.	F	P
Type	1	142.65	0
Order	1	0	0.9778
File	3	0.3	0.8234
Type*order	1	0.4	0.5334
Type*file	3	1.86	0.154
Order*file	3	0.29	0.8356

(b)

Fig. 5. (a) F-score of human AED using different audio visualization; (b) Three-way ANOVA of the F-score

4. CONCLUSION

The proposed algorithm maximizes the mutual information between target events and the estimated visual saliency of the displayed spectrogram. The saliency-maximized spectrogram is applied to Timeliner which is a flexible interface using both audition/vision. Subjects achieve significantly better performance when using the saliency-maximized spectrogram than using the original spectrogram in a 1/10-real-time AED task.

5. 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 and C Goudeseune, "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 Pub Co, 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] Chen M Jänicke H., "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.