A SALIENCY-BASED APPROACH TO AUDIO EVENT DETECTION AND SUMMARIZATION

A. Zlatintsi[1], *P. Maragos*[1], *A. Potamianos*[2], *G. Evangelopoulos*[3]

[1] School of ECE, National Technical University of Athens, 15773 Athens, Greece
[2] Dept. of ECE, Technical University of Crete, 73100 Chania, Greece
[3] University of Houston, Dept. of CS, Houston, USA
[nzlat,maragos]@cs.ntua.gr potam@telecom.tuc.gr

ABSTRACT

In this paper, we approach the problem of audio summarization by saliency computation of audio streams, exploring the potential of a modulation model for the detection of perceptually important audio events based on saliency models, along with various fusion schemes for their combination. The fusion schemes include linear, adaptive and nonlinear methods. A machine learning approach, where training of the features is performed, was also applied for the purpose of comparison with the proposed technique. For the evaluation of the algorithm we use audio data taken from movies and we show that nonlinear fusion schemes perform best. The results are reported on the MovSum database, using objective evaluations (against ground-truth denoting the perceptually important audio events). Analysis of the selected audio segments is also performed against a labeled database in respect to audio categories, while a method for fine-tuning of the selected audio events is proposed.

Index Terms— monomodal audio saliency, modulation model, audio summarization

1. INTRODUCTION

The amount of multimedia data in the web is constantly increasing with audio/music databases, diverse recordings, lectures and presentations, TV programs archives etc. Since there is usually no labeling attached to it, there is the constant need of finding new techniques to summarize the content of such data. Current techniques propose a variety of audio features, such as short-time energy, zero crossing rate, MFCCs, chroma features and others, and the approaches include audio classification and segmentation, repetition detection, and knowledge-based rules.

This paper addresses the issue of audio event detection and summarization. We approach this issue with saliency computation which offers an abstraction of a measure of interest to audio frames. A saliency based model can be found in [1] used for the automated extraction of music snippets. After a basic feature extraction, salient segments are detected based on their occurrence frequency and their energy. Boundaries of phrases are detected so as to ensure that the final segment includes meaningful phrases. Attention models for video summarization are used in [2], where each frame of a video sequence is assigned an attention value, depending on the viewer's attention. Audio saliency, in this case, is based on energy features, since loudness attracts people's attention. In [3], video summarization is attempted, where the summarization of the audio cue is approached using segmentation and classification of audio events in order to extract the boundaries of the segments, using standard features such as MFCCs and K-means clustering for the selection of the segments included in the summary.

In this paper, we use audio data extracted from movies for audio event detection and summarization, employing a modulation model and various linear, nonlinear and adaptive fusion schemes for the construction of a saliency curve. In Section 2, we also evaluate the performance of the extracted audio events against human labeled ground truth of audio saliency. In Section 3, a machine learning approach is applied to validate the efficiency of the various saliency models. Finally, in Section 4, we examine the type of the extracted segments against ground truth of labeled audio categories on movie segments of various genre, and we propose a technique for the correction of their boundaries.

2. AUDIO ANALYSIS AND MODELING

In this paper, the issue of saliency computation in an audio stream is approached as a problem of assigning a measure of interest to audio frames, based on spectro-temporal cues. The importance of amplitude and frequency changes for audio saliency has motivated a variety of studies where subject responses are measured with respect to tones of modulated frequency or loudness [4, 5, 6]. Amplitude and frequency modulations are also important for auditory grouping [7] and recognition of audio sources and events. In the model used,

This research has been co-financed by the European Union (European Social Fund - ESF) and Greek national funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF) - Research Funding Program: Heracleitus II. Investing in knowledge society through the European Social Fund.

the saliency is quantified through a combination of modulation parameters of non-stationary components. This results in a compact representation of the audio stream by tracking the components with maximal energy contribution across frequencies and time.

The analysis and saliency-modeling of the audio stream is based on the AM-FM model for audio signals (speech, music, environmental sounds): $s[n] = \sum_{k=1}^{K} A_k[n] \cos\left(\int_0^n \Omega_k[n] dn\right)$. The instantaneous amplitude $A_k[n]$ and frequency $\Omega_k[n]$ are estimated by multi-band filtering $s[n]$ with Gabor filters h_k, and then, by applying the Teager energy operator Ψ and the energy separation algorithm to each filter output. We then, compute the mean instantaneous amplitude $\mathrm{MIA}[m] = (\overline{|A_j[n]|})$ and frequency $\mathrm{MIF}[m] = (\overline{\Omega_j[n]})$, from the energy-dominant modulation component along multiple frequency bands [8], i.e., the component $j = j[m]$ which maximizes the average Teager energy $\mathrm{MTE}[m] = \arg\max_k(\overline{\Psi(s * h_k)})$, where m is the frame index and $(\overline{\cdots})$ denotes time averaging. Details about the feature extraction and implementations can be found in [8, 9].

The audio stream is thus described by the 3D feature vector $\vec{F}_{\mathrm{a}}[m] = [\mathrm{MTE}, \mathrm{MIA}, \mathrm{MIF}]\,[m]$, which conveys information on excitation level, frequency content and source energy, related to the presence and evolution of audio events. Various fusion and normalization schemes are investigated for the combination of MTE, MIA and MIF, resulting in a single audio saliency curve.

2.1. Intramodal Fusion of audio features

A variety of fusion schemes were experimentally evaluated to obtain a saliency curve which forms the basis for the selection of perceptually important audio events for the creation of meaningful audio summaries. The problem examined in this paper is the low level *intramodal fusion* where the features are normalized and fused to produce a monomodal saliency curve, where each value corresponds to a measure of perceptual importance of the individual feature streams. Individual features are normalized with respect to their value range prior to fusion in order to ensure a mapping to [0, 1] and compensate for the difference in their dynamic range by least squares fitting of their values. The developed saliency curve has a number of attractive properties. It is a continuous valued function of time, constrained by appropriately designing the fusion norm to reside in $[0, 1]$ and it is formed through an unsupervised, bottom-up approach, approximating the sensory-level attention invoked by the audio stream to a listener.

The normalized feature vectors are combined using frame-level fusion of their values: $S_{\mathrm{A}} = \mathrm{fusion}(S_1, S_2, S_3)$. The fusion frameworks examined are: 1) Weighted linear combinations with equal or unequal, fixed weights. 2) Variance-based weights, inversely proportional to each feature saliency's uncertainty. 3) Nonlinear norms, e.g., max, min and weighted min. 4) Finally, time-adaptive, dynamic weights, using the syntactic structure of the video (e.g., scene and shot changes in movies) in order to find the optimum scheme which is going to be used for the final summary.

Linear Fusion: The most intuitive option is a weighted average of normalized saliency values which is based on a weighted linear combination of the audio features:

$$S_{\mathrm{lin}} = w_1 S_1 + w_2 S_2 + w_3 S_3. \quad (1)$$

The simplest such scheme is to equally weight the three features vectors (LE).

Adaptive (Variance-based) Fusion: Each feature stream is weighted inversely proportional to its variance:

$$S_{\mathrm{var}} = \sum_i (S_i / \mathrm{var}(S_i)) / \sum_i (1/\mathrm{var}(S_i)). \quad (2)$$

Nonlinear Fusion: (i) min (MI) and (ii) max (MA) fusion, i.e., taking the minimum or maximum value of the three examined audio cues at each frame respectively,

$$S_{\mathrm{min}} = \min\{S_1, S_2, S_3\}, \;\; S_{\mathrm{max}} = \max\{S_1, S_2, S_3\}. \quad (3)$$

Additionally, a new min-fusion scheme was examined, the weighted min fusion (MIVA) which can be applied globally and adaptively. In this case, each feature stream is additively weighed inversely proportional to its log variance: $w_i = \log(1/\mathrm{var}(S_i))$

$$\begin{aligned} S_{\mathrm{miva}} = &\min(S_1 - w_1, S_2 - w_2, S_3 - w_3) \\ &+ \max(w_1, w_2, w_3). \end{aligned} \quad (4)$$

This scheme is algebraically homomorphic to the linear variance-weighted scheme of (2).

The normalization intervals that were investigated are: (i) global linear normalization (GL) where scaling is performed for the whole audio stream, (ii) scene-based linear normalization (SC) where each scene is separately normalized, and (iii) shot-based linear normalization (SH) where each shot is separately normalized. Dynamic adaptation, i.e., weight updating is also considered with respect to global or local windows. For instance, for the inverse variance weighting and the min variance fusion schemes the variance of each stream can be computed at the global (VA-GL), shot (VA-SH) or scene (VA-SC) level.

2.2. Summarization Algorithm for Audio Event Detection

Since the audio streams, used in this paper, are extracted from movies, the summarization algorithm is based on [9] and follows these steps: (i) Filtering of the audio saliency curve with a median filter of length $2M + 1$ frames. (ii) Saliency threshold selection S_c dictated by the *percent of summarization* c required, where frames m with value $S_{\mathrm{A}}[m] > S_c$ are selected. For example, for 20% summarization ($c = 0.2$), S_c is selected so that the cardinality of the set of selected frames $D = \{m : S_{\mathrm{A}}[m] > S_c\}$ is 20% of the total number of frames. The result is a frame indicator function I_c for the desired level of summarization c. (iii) Combination of frames into segments. Segments that are shorter than N frames are deleted from the summary. (iv) Neighboring segments that are selected for the summary are merged if they are less than K frames apart. (v) Rendering of the selected segments into

Features		Audio Feature Fusion		
Evaluated on		Audio (A) Labeling		
		Summarization Percent		
Algorithm		20%	33%	50%
Norm	Fusion	Precision Scores		
GL-N	LE-F	68.8	66.1	61.9
GL-N	MA-F	48.8	51.2	52.6
GL-N	MI-F	**92.6**	**83.6**	**73.8**
GL-N	MIVA-GL-F	**92.6**	**83.6**	**73.8**
GL-N	MIVA-SC-F	91.1	81.9	72.8
GL-N	MIVA-SH-F	91.9	83.4	73.7
GL-N	VA-GL-F	91.6	81.0	70.5
GL-N	VA-SC-F	85.3	75.8	68.2
GL-N	VA-SH-F	90.0	82.8	72.6
SC-N	LE-F	66.1	64.3	62.0
SC-N	MI-F	77.8	73.2	69.1
SC-N	MIVA-GL-F	78.0	73.3	68.9
SC-N	MIVA-SC-F	77.6	72.3	67.6
SC-N	VA-GL-F	72.6	68.3	63.7
SC-N	VA-SC-F	72.6	65.4	61.6
SH-N	LE-F	73.2	68.8	64.2
SH-N	MI-F	68.9	67.6	64.7
SH-N	MIVA-GL-F	66.9	66.2	63.5
SH-N	MIVA-SC-F	68.4	66.9	64.4
SH-N	MIVA-SH-F	66.9	66.0	63.6
SH-N	VA-GL-F	73.2	68.9	64.2
SH-N	VA-SC-F	73.4	69.3	64.7

Table 1*:* Frame-level summarization precision scores for audio feature fusion. Audio features are evaluated on audio annotation.

a summary by using simple overlap-add on L video frames to tailor together neighboring segments. The evaluated version of the algorithm for this case operates with $M = N = 30$ frames, $K = L = 15$ frames for videos at 25 fps.

2.3. Objective Evaluation of Fusion Schemes

We evaluate three different normalization schemes: global (GL), scene-level (SC) and shot-level (SH), as well as nine fusion schemes: linear (LE), min (MI), weighted min at different levels (MIVA-GL, MIVA-SC, MIVA-SH), max (MA), and inverse variance at different levels (VA-GL, VA-SC, VA-SH). For this purpose, we used the audio stream extracted from six 30-minute movie clips from the MovSum database, a joint work of the NTUA and TUC labs, including movie clips from the following Oscar-winning movies of various genres: "Chicago", "Crash", "Departed", "Gladiator", "Lord of the Rings III" and "Finding Nemo". Each clip on the database is perceptually and cognitively labeled, regarding the salient events, forming ground-truth data for objective evaluation purposes. The evaluation was performed against the audio (A) layer, consisting of segments that are acoustically interesting. The segments that are considered acoustically salient formed a binary saliency indicator function, consisting of frames labeled as salient by at least two of the three expert labelers.

Our intention is to examine whether the intramodal fusion of the audio cues evaluated against the ground-truth of the audio annotation (A) can form summaries that consist of segments that are both meaningful and chosen by the users as

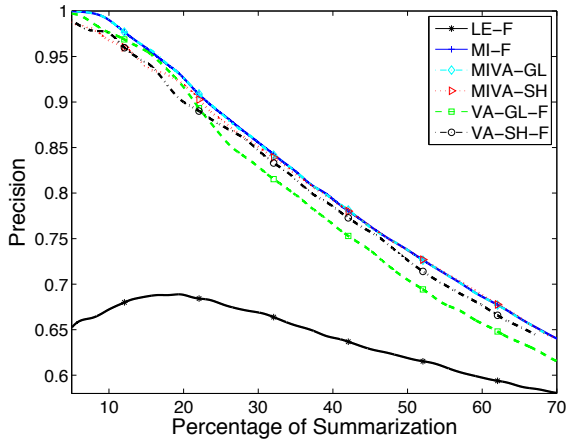

Fig. 1*:* Frame-level summarization precision scores for the five best performing fusion schemes and the baseline method LE-F. (Please see color version for better visibility.)

salient. For experimentation purposes, we altered the parameters of the summarization algorithm and specifically, the size of the minimum segment that could be selected. The results that are presented next, consider a minimum segment of 30 frames which was empirically found to be a good choise for this task.

Table 1 presents the results in terms of frame-level precision scores, since it best characterizes the frame-level performance on *event detection tasks*. The scores are presented for audio feature fusion for summaries that include 20%, 33% and 50% of the original number of frames for all possible combinations among the normalization and fusion schemes (best scores are shown in bold). We observe that for all tasks and evaluation settings global normalization significantly outperforms shot- and scene-level normalization schemes, and for the GL normalization cases (a) nonlinear MI-F and MIVA-F fusion schemes outperform linear fusions and MA-F, while (b) the inverse variance schemes (VA-GL, VA-SC, VA-SH) outperform LE-F and MA-F fusion.

Figure 1, shows frame precision results as a function of summarization level (ranging from 5% to 70%), for global normalization and the five best performing fusion schemes plus the baseline LE for audio feature fusion tasks. We observe that MI and MIVA-GL perform equally good. MIVA-SH also performs as good for summaries including over 40% of frames, followed by VA-SH and VA-GL. LE-F which is regarded as the baseline method performs significantly worse compared to the rest of the fusion schemes.

3. MACHINE LEARNING APPROACH

Next, we investigate a machine learning approach to audio summarization where classifiers are trained using the frame-based audio features presented in Sec. 2. This method is applied in order to validate the efficiency of the proposed saliency models. Specifically, we use the $\vec{F}_{\mathrm{a}}[m] =$

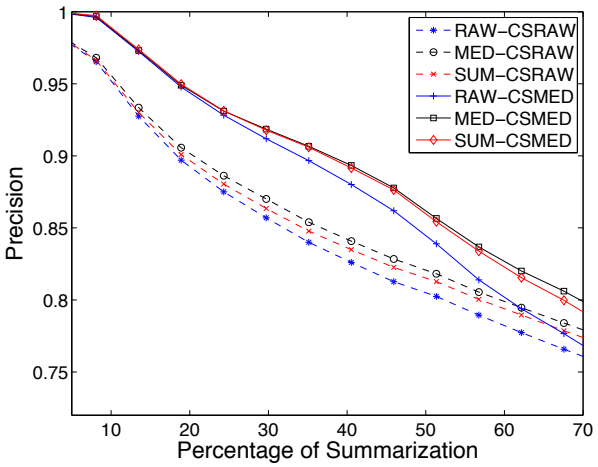

Fig. 2*:* Frame Precision using NNR classifiers.

$[\mathrm{MTE}, \mathrm{MIA}, \mathrm{MIF}]\,[m]$ feature vector along with its first and second time derivatives computed over three and five frames respectively. We report (frame precision) results using a nearest neighbor classifier[1] (NNR-k) that is trained on the manually annotated movie corpus using the binary markings (1 for frames where audio events are present, 0 otherwise). Thus the classifier output is a binary indicator function of 0's (no event) and 1's (audio event). Six-fold cross-validation is used, i.e., NNR-k models are trained on five movies and tested on the sixth. For the purposes of selecting the frames that are more likely to correspond to audio events we report results using the following (smoothing) heuristics on the classifier output: (i) No smoothing, i.e., raw frame-based results (RAW), (ii) Median-filtering on the raw classifier output with window of length $2M + 1$ (MED), (iii) Application of the summarization algorithm of Sec. 2, as if the classifier output was the thresholded saliency curve (SUM).

In order to obtain results for variable compression rates, we define a confidence score for each classification result, i.e., each frame. We choose as confidence score the portion of the k nearest neighbors that are marked as 1 (audio events); this roughly corresponds to the posterior probability of the audio event class for that frame. Results are presented both for the raw confidence scores (CSRAW), as well as, the median filtered confidence curve with window of length $2K + 1$ (CSMED). Results are shown in Fig. 2 for all combinations of post processing of classifier output and confidence scores, e.g., "SUM-CSMED" refers to using the summarization algorithm for post-filtering classifier output and median filtering for smoothing the confidence score curves. The parameters are optimized to achieve the best possible classification accuracy scores to $k = 250$ neighbors for NNR, and $M = K = 50$ frames for median filtering.

Overall, precision scores achieved using the NNR classifier are better than those achieved using the saliency curve approach with the exception of the $5 - 20\%$ summarization region, where only the median filtered confidence scores (CSMED) achieve better performance. Post-filtering of the classifier output (MED, SUM) improves on the baseline precision somewhat (RAW) in the 30-70% summarization region. Finally, post-filtering of the confidence scores (CSMED) significantly improves precision in the 5-50% region over the raw estimates (CSRAW).

[1]Similar results can be obtained using Gaussian mixture models or support vector machine classifiers.

4. EVENT STRUCTURE ANALYSIS AND SUMMARIZATION

In this section, we analyze and evaluate the type of segments chosen by the summarization algorithm, while we propose a technique for correction of the boundaries of the selected segments included in the final audio summary.

The evaluation is performed on a different database of movie segments of various genres, consisting of four 3-4 min long clips from a documentary, a music documentary, and two different movies, including a variety of audio classes. The segments of this database are labeled in respect to various audio categories such as: speech, music, background music, song, environmental sounds, e.g., wind, waves etc, machine sounds, "foley" sounds, e.g., laughter, applause, footstep, knocking, and impact sounds, e.g., "bang" (gunshot), "boom" (explosion) and smash sounds.

The following discussion concerns the type of the automatic extracted audio events. We evaluate the best fusion scheme from the previous analysis MI-F (GL-N) and compare it to the baseline LE-F (GL-N). We note that MI-F fusion includes almost all speech segments for longer summaries, while only the most prominent, high intensity speech segments for smaller summaries. It also includes intense and loud music segments (which did not function as background music), all impact sounds, e.g. gunshots, machine sounds and sounds that stood out in silence. For the music documentary, we observe that the music segments are favored, most probably because of their high intensity in comparison to speech (interview segments). Table 2 shows the percentage of frames extracted by the summarization algorithm belonging to a specific audio category. The total percentage of frames of each type of audio is also presented for reference reasons, (shown in "% of frames for each audio category").

Audio Summarizer: In this final step, we describe an algorithm for the adjustment of the automatic extracted audio events. As already discussed, we use the audio saliency as an indicator function curve that marks the most prominent audio segments. This is automatically performed by the summarization algorithm, Sec. 2.2, depending on the required threshold set by the user. In order to make the system more robust and be able to choose segments that are not only salient but also form meaningful phrases, we perform correction of the boundaries of the extracted events, using the boundaries of the manually segmented audio categories. This is achieved

Type of Audio	Movie (3 min)					Music Documentary (4 min)				
Algorithm	MI-F		LE-F		% of frames	MI-F		LE-F		% of frames
Summarization Percent	20%	50%	20%	50%	for each audio	20%	50%	20%	50%	for each audio
	% of detected frames				category	% of detected frames				category
Audio Category										
Speech	28.0	82.0	18.8	70.9	36.6	0	25.9	1.5	29.2	61.0
Music	20.4	55.5	23.3	43.0	17.5	58.0	88.1	50.5	80.9	35.0
Background Music	36.8	68.6	27.2	51.2	35.9	0	25.1	1.3	28.8	52.1
Song	-	-	-	-	-	83.1	100	76.87	98.6	19.8
Environmental	38.6	52.5	71.9	95.0	8.3	-	-	-	-	-
Machine	33.5	81.2	23.4	62.9	4.6	50	100	59.5	59.5	0.7
Foley	18.0	47.6	17.55	41.8	10.0	0	10.6	19.5	100	1.9
Impact	100	100	86.1	100	2.3	0	100	0	38.5	0.2

Table 2*:* Frame-level summarization percentage of correct frames belonging to a specific audio category for audio feature fusion for the best fusion scheme MI-F and the baseline LE-F with GL normalization.

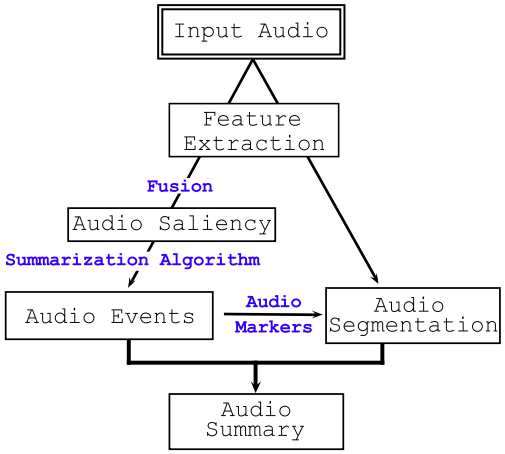

Fig. 3*:* Block diagram of the summarization system.

using ideas from mathematical morphology and specifically, the reconstruction opening: $\rho^{-}(M|X) \triangleq$ connected components of X intersecting M. In such a way, we can extract large-scale components by knowing only smaller markers inside them, i.e. the initially extracted audio events. In this case, since we do not employ an automatic audio segmentation algorithm, we perform this final step by using as marker M the thresholded saliency function, marking the segments that will be included in the final summary and as reference X the labeled audio-specific annotation. Note that for audio including speech, a VAD algorithm as in [8] could provide automatic segmentation. This action is regarded significant for the performance of the system, especially for the comprehension of speech segments. The boundary correction is expected to improve the method's accuracy, since human labelers tend to choose unified segments especially concerning speech. Figure 3 shows the process of audio summarization procedure from feature extraction to the final adjusted summary.

5. CONCLUSIONS

Linear and nonlinear fusion schemes have been proposed to integrate audio cues from movie clips in order to create a monomodal saliency curve with applications to audio event detection. The thresholded audio saliency in combination with manually segmented events constitutes the final system for audio summarization. Among the various normalization and fusion schemes investigated, global normalization (GL), min fusion (MI), weighted min (MIVA) and shot-variance (VA-SH) schemes work very well. The evaluation of MI-F against labeled database with audio categories showed that it is well suited for both generic audio streams and music-oriented too. The machine learning approach employed achieved better results than those achieved using the saliency computation approach, with the exception of the $5-20\%$ summarization region. However, we consider significant the fact that the proposed bottom-up saliency models gain almost as good scores for the smaller summaries. For future work, we intend to perform automatic segmentation and classification of the audio streams in order to create a fully automatic summarization system.

6. REFERENCES

[1] L. Lue and H.-J. Zhang, "Automated Extraction of Music Snippets", *in Proc. Int'l Conf. ACM*, 2003.

[2] Y.-F. Ma, L. Lu, H.-J. Zhang and M. Li, "A User Attention Model for Video Summarization", *in Proc. ACM Multimedia*, pp. 533-542, 2003.

[3] J. Wei, C. Courtenay and A.C. Loui, "Automatic consumer video summarization by audio and visual analysis", *Multimedia and Expo (ICME-11)*, pp. 1-6, Jul. 2011.

[4] J.B. Fritz, M. Elhilali, S.V. David and S.A. Shamma, "Auditory attention–focusing the searchlight on sound", *Current op.inion in neurobiology*, vol. 17, no. 4, pp. 437-455, Aug. 2007

[5] C. Kayser, C.I. Petkov, M. Lippert and N.K. Logothetis, "Mechanisms for allocating auditory attention: an auditory saliency map", *Current Biology*, vol. 15, no. 21, pp. 1943-1947, 2005.

[6] M. Elhilali, J. Xiang, S.A. Shamma and J.Z. Simon, "Interaction between attention and bottom-up saliency mediates the representation of foreground and background in an auditory scene", *PLoS biology*, vol. 7, no. 6, Jun. 2009.

[7] R.P. Carlyon, *How the brain separates sounds*, Trends in Cognitive Sciences, vol. 8, no. 10, pp. 465-471, Oct.2004.

[8] G. Evangelopoulos and P. Maragos, "Multiband modulation energy tracking for noisy speech detection," *IEEE Trans. Audio Speech Language Processing*, vol. 14, pp.2024-2038, Nov 2006.

[9] G. Evangelopoulos, A. Zlatintsi, G. Skoumas, K. Rapantzikos, A. Potamianos, P. Maragos and Y. Avrithis, "Video event detection & summarization using audio, visual & text saliency," *Proc. ICASSP*, Taipei, Taiwan, 2009.