AUDIO SALIENT EVENT DETECTION AND SUMMARIZATION USING AUDIO AND TEXT MODALITIES

Athanasia Zlatintsi, Elias Iosif, Petros Maragos and Alexandros Potamianos

School of ECE, National Technical University of Athens, Greece

Email: {nzlat, maragos}@cs.ntua.gr, {iosife,potam}@central.ntua.gr

ABSTRACT

This paper investigates the problem of audio event detection and summarization, building on previous work [1, 2] on the detection of perceptually important audio events based on saliency models. We take a synergistic approach to audio summarization where saliency computation of audio streams is assisted by using the text modality as well. Auditory saliency is assessed by auditory and perceptual cues such as Teager energy, loudness and roughness; all known to correlate with attention and human hearing. Text analysis incorporates part-of-speech tagging and affective modeling. A computational method for the automatic correction of the boundaries of the selected audio events is applied creating summaries that consist not only of salient but also meaningful and semantically coherent events. A non-parametric classification technique is employed and results are reported on the MovSum movie database using objective evaluations against ground-truth designating the auditory and semantically salient events.

Index Terms— monomodal auditory saliency, affective text analysis, audio-text salient events, audio summarization

1. INTRODUCTION

Information retrieval and specifically automatic content summarization has attracted much research interest in the last few years. Because of the vast amount of the existing multimedia data in the web and our personal databases, summarization plays a key role in various domains. Areas of interest where summarization can be applicable include audio, music and video databases, lectures and presentations, sharing sites [3], TV programs [4], as well as, acoustic or video monitoring/surveillance [5] and others. In the cases where video summarization is under investigation, the audio modality is regarded to bear much of the important information, and it is used to add robustness to the overall performance of a system. However, the use of advanced and more specific information, e.g., from the text modality, is often required, in order to obtain meaningful summaries where the selected segments are of high interest. In this paper, we investigate the problem of audio salient event detection and summarization, extending previous work [1, 2], by also incorporating the text modality using features based on part-of-speech tagging and affective word continuous ratings. Our intention is to assist the auditory saliency, to measure the significance of the words and hence develop better techniques for audio summarization.

A general audio summarization system, in order to be robust has to include conspicuous and relevant acoustic events that attract human attention and outline the basic concepts and main ideas of the whole audio stream. Features like energy and loudness have been broadly used since they were reported to attract human attention [4,6–8]. Speech and audio processing techniques such as voice activity detection (VAD) and keyword spotting are usually forming a part of summarization systems for audio segmentation and detection of important words [5]. Human attention is actually targeted towards speech since it bears the semantics, which are important for understanding. For the selection of the salient events to be included in a summary, methods that are mainly used are the computation of similarity matrices or clustering [5, 9].

Analysis of text to estimate affect or sentiment is a relatively recent research topic that has attracted great interest with application to numerous domains such as tweet analysis [10], product reviews [11], or dialogue systems [12]. Text can be analyzed at different levels of granularity: from single words to entire sentences. In [13], affective ratings of unknown words were predicted using the affective ratings for a small set of words (seeds) and the semantic relatedness between the unknown and the seed words. An example of sentence-level approach was proposed in [14] applying techniques from n-gram language modeling.

In this paper, we use audio data extracted from movies for audio salient event detection and summarization. In Sec. 2, a description of the audio features, which are based on energy tracking and other perceptual features that correlate to human perception of sound, can be found. In Sec. 3, we present the text analysis, which is based on part-of-speech tagging and affective modeling of single words extracted from the subtitles of the movies. In Sec. 4, a machine learning technique is

This research work was supported by the project "COGNIMUSE" which is implemented under the "ARISTEIA" Action of the Operational Program Education and Lifelong Learning and is co-funded by the European Social Fund and Greek National Resources.

978-0-9928626-3-3/15/$31.00 ©2015 IEEE

applied to validate the efficacy of the proposed methods. Additionally, the summarization algorithm is described where automatic correction of the boundaries of the salient events (based on speech and specifically the single word level boundaries) is taking place. In Sec. 5 results are reported on the MovSum database [15].

2. AUDIO ANALYSIS AND MODELING

Spectro-temporal cues are employed to tackle the issue of auditory saliency estimation and a measure of interest is assigned to the audio frames. The importance of amplitude and frequency changes for audio saliency has motivated various studies where subject responses were measured with respect to tones of modulated frequency or loudness [16–18].

Extensive experimentation with different configurations, for the analysis of the audio stream, led to an energy-based feature set for the auditory saliency modeling. This was approached using the nonlinear differential energy operator proposed by Teager [19] and further investigated by Kaiser [20] and in [21]. Even though the Teager-Kaiser energy operator has been mainly used for energy estimation and AM-FM demodulation of the audio signal [21], it has been proven more robust to noise in comparison to the squared energy operator [22]. Besides, the multiband Teager energy cepstrum coefficients (TECCs) have proven successful in speech recognition tasks [23].

The Teager-Kaiser Energy Operator (TEO), which can track the instantaneous energy of a source, is given by

$$\Psi[x] = \dot{x}^2 - x\ddot{x}, \text{ where } \dot{x} = dx/dt. \quad (1)$$

When the Teager operator is applied to AM-FM signals of the form $x(t) = \alpha(t)cos(\phi(t))$ then Ψ yields [21]

$$\Psi[x(t)] \approx \alpha^2(t)\dot{\phi}^2(t). \quad (2)$$

Thus, it captures amplitude and frequency variation information; which has been proven to help and improve the accuracy in speech and music recognition tasks [23, 24]. Additionally, due to its sharp time resolution and lowpass behaviour Ψ can detect robustly and discriminate various acoustic events.

Since Teager energy is only meaningful in narrowband signals [21], the application of the operator is preceded by multiband filtering of the signal with Gabor filters; chosen mainly because they exhibit good joint time-frequency resolution [21]. Thus, the features used in this paper are derived using multiband analysis, where the audio signal is processed in 30 ms frames. A filterbank, of 25 linearly spaced Gabor filters, filter the signal to isolate narrowband components. Then the energy operator is estimated at the outputs and the average for the frame duration gives a measure of each channel activity; the mean instantaneous energies.

Moreover, we computed two additional perceptual features which were found to correlate to the functioning of the human auditory system. The first one is roughness proposed

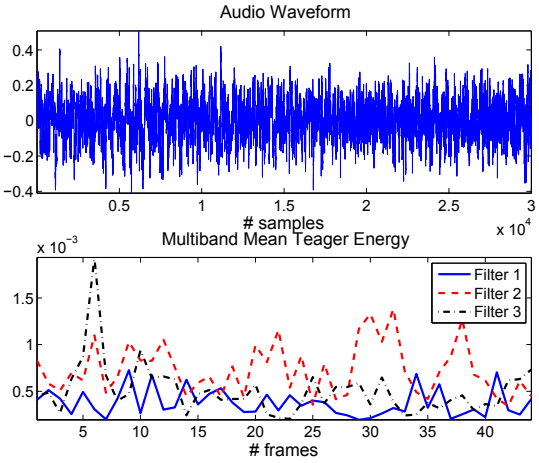


Fig. 1: Short-time features for signal analysis using 30-ms Hamming frames, updated every 1/2 of frame duration, at a 44.1-kHz sampling rate. Signal waveform (top), and the multiband Teager energy features for the first three Gabor filters (bottom figure).

in [25] and reported to be associated with human attention [26]; which is an estimation of the sensory dissonance of a sound. It expresses a sense of "stridency" of a sound due to rapid fluctuations in its amplitude and it is related to the beating phenomenon whenever pairs of sinusoids are close in frequency. An estimation of roughness can be given by computation of the peaks of the spectrum followed by averaging among all possible pairwise combinations of peaks [27]. In this work, a variant model is applied that uses more complex weighting [28]. The second perceptual feature used in this work is loudness, associated to attention as well, which corresponds to the perceived sound pressure level. For the computation of loudness the model proposed in [26] was employed, which is based on the calculation of the excitation on the basilar membrane taking into account phenomena such as the temporal frequency masking.

3. AFFECTIVE TEXT ANALYSIS

In this work, we extend the text analysis of [2, 29] and we include affective word-level modeling, using the text information available in the subtitles of each movie. High arousal and high absolute valence are actually expected to be good indicators for words related with salient events, e.g., [30]. Humans tend to pick content (movies, music) based on its affective characteristics, hence affective features is of particular interest to content delivery systems that provide personalized multimedia content, automatically extract highlights and create automatic summaries or skims. Our baseline text analysis [2, 29] can be summarized in the following steps: (i) extraction of the movie transcript from the subtitle files, (ii) audio segmentation using forced word-level alignment, (iii) part-of-speech tagging, where the different words are assigned a value of $\{0.2, 0.5, 0.7, 1\}$, and (iv) text saliency computation (assignment of a text saliency value to each frame) based on the

parser tag assigned to the corresponding word.

3.1. Metrics of Word Semantic Similarity

Here, we provide a brief overview of two widely-used types of corpus-based metrics of word semantic similarity.

Co-occurrence-based. The underlying hypothesis is that the co-occurrence of words in a specified context (e.g., sentence, paragraph) correlates with their semantic similarity. Metrics of this type include Dice and Jaccard coefficients, point-wise mutual information, etc.

Context-based. This type relies on the distributional hypothesis of meaning suggesting that semantically similar words share similar context [31]. Given a target word w_i, a window of size $2H+1$ words is centered on every instance of w_i in the corpus, and the contextual (lexical) features are extracted. The extracted features are represented as a vector $\mathbf{x}_i$. The semantic similarity between two words, w_i and w_j, is computed by taking the cosine of their respective feature vectors:

$$Q^H(w_i, w_j) = \frac{\mathbf{x}_i \cdot \mathbf{x}_j}{|| \mathbf{x}_i || \quad || \mathbf{x}_j ||}. \tag{3}$$

Various schemes can be used for weighting the elements of feature vectors. In this work, a binary scheme is used.

3.2. Affective Rating of Words

The affective content of a word w can be rated in a continuous space ($[-1, 1]$), which includes the following dimensions: valence (v), arousal (a), and dominance (d). For any dimension, the affective content of w can be estimated via a linear combination of its semantic similarities to a set of K seed words and the corresponding affective ratings of seeds as follows [14]:

$$\hat{u}(w) = \lambda_0 + \sum_{i=1}^{K} \lambda_i \; u(t_i) \; S(t_i, w), \tag{4}$$

where $t_1 \ldots t_K$ are the seed words, $u(t_i)$ stands for the affective rating of seed t_i, while u denotes an affective dimension (v, a, or d). λ_i is a trainable weight that corresponds to seed t_i. $S(t_i, w)$ is a metric for computing the semantic similarity between t_i and w.

Regarding $S(.)$ of (4), the context-based Q^H similarity metric with $H=1$ was applied, using a text corpus consisting of more than 116 million sentences. The affective ratings of words were estimated by exploiting 600 entries of the ANEW lexicon [32] as seeds. In [14], more details are provided about the corpus, seed selection, and the training of the λ weights.

4. EXPERIMENTAL EVALUATION

4.1. MovSum Database

We have evaluated our proposed framework on seven movies from the *Movie Summarization (MovSum) database* [15], that consists of half-hour continuous segments (three and a half hours in total), which are the following: "Beautiful Mind", "Chicago", "Crash", "The Departed", "Gladiator", "Lord of the Rings – The Return of the King" and the animation movie "Finding Nemo". The specific movies were perceptually and cognitively annotated by three expert viewers considering a) monomodal auditory saliency (thus segments that were acoustically interesting; depending on the importance and the invoked attention they create to the annotator) and b) semantic saliency (e.g., phrases, actions, symbolic information, sounds), hence sequences of conceptual events not necessarily important just for the examined movie/audio stream but generally, as an objective, direct or indirect meaning. The annotated events were used as ground-truth for objective evaluation purposes. The ground-truth framewise saliency consists of frames that have been labeled salient by at least two labelers. The evaluation of the automatically selected auditory and textual events is hence performed against the audio (A) and the semantics (S) layer respectively, while the audio/semantic (AS) annotation layer is used for the evaluation of the fused audio-textual events.

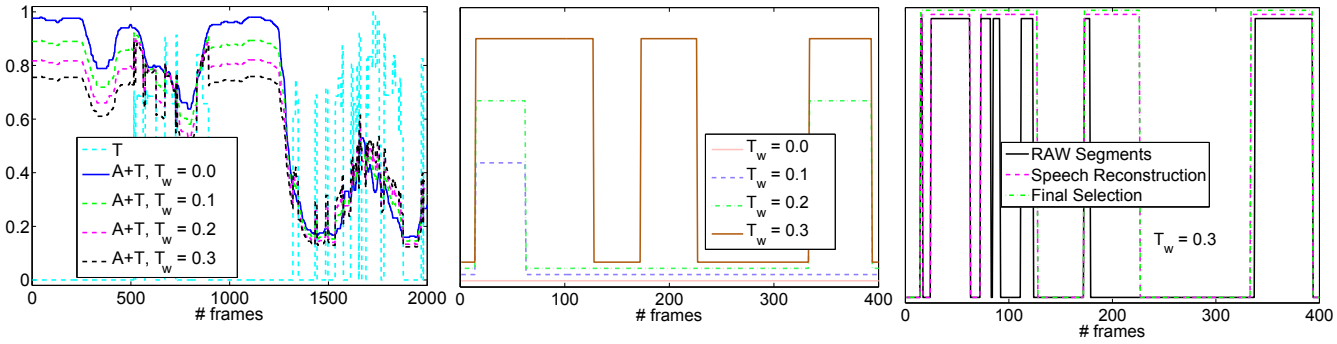
4.2. Machine Learning Approach

For the audio-text salient event detection and audio summarization we follow a non-parametric data-driven classification approach, following the same framework as in [1, 2]. The resulting temporal sequence of the 27 audio features along with its first and second temporal derivatives (computed over 3 and 5 frames respectively) and the 4 text features, compose the set of features used for the classification, where a K-Nearest Neighbor Classifier (KNN) is employed. Particularly, framewise saliency is considered as a two-class classification problem, and a seven-fold cross-validation is used, hence KNN models are trained on six movies and tested on the seventh. In order to obtain various compression rates results and be able to create summaries of variable rates, a confidence score is defined for each classification result, i.e., each frame.

4.3. Summarization Algorithm for Audio Event Detection

In this paper, a new summarization procedure is adopted, extending our previous summarization algorithm [2], consisting of technical characteristics considered imperative so as to make the summaries smoother, concerning audio transitions, while also improving the comprehension of the semantics.

For the audio summary production we are using the salient frames' confidence scores, which also allow the creation of variable rates summaries. Specifically, we use frames or segments with high confidence scores, as an indicator function curve that marks the conspicuous audio and text events. The newly adopted method that we follow for the creation of a "saliency curve" and the selection of the events to be included in the summary, include the following steps: **(1)** median filtering of the audio confidence scores C_A, so as to obtain a smoother and coarse audio attention curve. **(2)** The text confidence scores C_T that were trained only on the



(a) Late fusion of A and T *(b) Effect of text weight in the selected segments* *(c) Segment selection after step 5 and 6*

Fig. 2*:* (a) Late fusion of the audio and text modalities using different text weight, where $w = \{0.1, 0.2, 0.3\}$. (b) Effect of the used text weight w on the selected segments. (c) Effect of reconstruction opening and the movie summarization algorithm for $w = 0.3$. All figures concern summaries at ($\times$5) rate. (Note that y-axis on *(b)* and *(c)* does not imply a scale measure.)

speech segments are used; while frames with no speech are set to zero. **(3)** Late fusion of the audio (A) and text (T) modalities is performed, where a fixed weight w for the text stream is chosen: $C_{AT} = C_A + w \cdot C_T$. In this paper we present results for weights $w = 0.10, 0.20, 0.30$. **(4)** Sorting of the confidence scores so as to define the high confidence frames/segments that will be included in the summary, according to the number of frames needed. **(5)** In order to create summaries that do not include only salient but also semantically coherent events, we perform correction of the boundaries of the selected segments, taking in this case into consideration the boundaries of the single words. This procedure which we call *"speech reconstruction"* is achieved using ideas from mathematical morphology and specifically, the reconstruction opening [33]:

$$\rho^{-}(M|X) \triangleq \text{connected components of } X \text{ intersecting } M. \quad (5)$$

If we consider as reference X the single words after the alignment and as marker M the raw salient events that were selected to be included in the summary (Step **4**), then we are able to redefine the boundaries of those segments, and thus extract large-scale components, containing exactly the input components X that intersect the marker. This boundary correction performed is regarded significant for the comprehension of the semantics, since it ensures that no words will be clipped, and the creation of smoother transitions [1].

After speech reconstruction the final step of the summarization algorithm is taking place for the combination of the selected segments into a summary, as described in [2, 29]. Thus, in Step **(6a)** segments that are shorter than N frames are deleted from the summary, while neighboring segments selected for the summary are merged if they are less than K frames apart, where $K = 25$ and $N = K/2$, while **(6b)** the final rendering of the selected segments into a summary is performed by using simple overlap-add. Figures 2a and 2b show the effect of text saliency (using different text weights w) on the fusion and the selection of segments, respectively, and Fig. 2c shows the initial selected raw segments, the same segments after speech reconstruction and after the final step (Step **6**) of the movie summarization algorithm for $w = 0.3$.

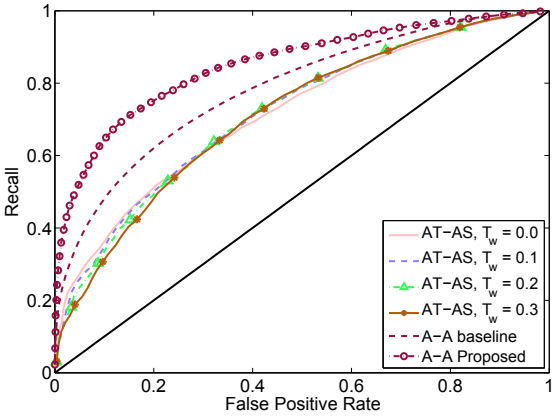


Fig. 3*:* Saliency classification ROC curves while changing the percentage of frames in summary, for audio-text (AT) on audio-semantics (AS) annotation, and for the baseline and the proposed audio on audio (A) annotation.

5. RESULTS AND DISCUSSION

Figure 3 shows Receiver Operating Characteristic (ROC) curves for saliency classification, while changing the percentage of frames in summary (between 1–100%), for audio on audio (A-A) and audio-text on audio-semantics (AT-AS) annotation. The results for the proposed method (A-A and AT-AS) are produced using the new summarization algorithm, and specifically for A on A we use the sorted median filtered confidence. For the baseline method, results are shown for the sorted confidence scores as presented in [2]. An important observation is that the proposed audio configuration outperforms the baseline accomplishing a really good performance. Regarding the fusion of text with audio (AT-AS) the improvement is not as noticeable; probably only a bit for longer summaries. However, we have to emphasize here the positive outcome of the text modality when fused with audio for the creation of the summaries. a) Segments or frames that

were not chosen as salient by the audio modality will be most probably emphasized by the text. b) We manage to produce summaries including events selected in a more uniform and consistent manner from the whole audio stream, which otherwise would be discarded (see Fig. 2b). Finally, c) text assists the summarization through the speech reconstruction process which is imperative for meaningful summaries. Concluding, our preliminary subjective evaluation, by a few expert users, confirmed our observations and results, since the summaries were judged to have good quality. Our intentions however are to plan extensive human evaluation in the future.

6. CONCLUSIONS

A new and improved synergistic approach was adopted for perceptually salient event detection with application in audio summarization, where saliency computation of audio streams was assisted by the text modality. Our experimental evaluation using a simple classifier confirms the adequacy of the proposed algorithms; specifically, we showed that our proposed audio frontend, based on perceptually inspired features, for auditory saliency estimation outperforms the baseline system over the saliency annotated MovSum database. Moreover, the audio summary can be further improved when incorporating the text modality affecting both the selection of events and the correction of their boundaries. For future work, we aim to further refine our methods, possibly use adaptive weights for the text modality, and plan a quality of experience (QoE) evaluation where the quality of the summaries will be measured in a systematic manner by humans.

REFERENCES

[1] A. Zlatintsi, P. Maragos, A. Potamianos, and G. Evangelopoulos, "A saliency-based approach to audio event detection and summarization," in *Proc. European Signal Process. Conf.*, 2012.

[2] G. Evangelopoulos, A. Zlatintsi, A. Potamianos, P. Maragos, K. Rapantzikos, G. Skoumas, and Y. Avrithis, "Multimodal saliency and fusion for movie summarization based on aural, visual, textual attention," *IEEE Trans. on Multimedia*, vol. 15(7), pp. 1553–1568, 2013.

[3] A. Pikrakis, "Audio thumbnailing in video sharing sites," in *Proc. European Signal Process. Conf.*, 2012.

[4] H. Duxans, X. Anguera, and D. Conejero, "Audio based soccer game summarization," in *IEEE BMSB*, 2009.

[5] D. Damm, D. von Zelledmann, M. Oispuu, M. Häge, and F. Kurth, "A system for audio summarization in acoustic monitoring scenarios," in *Proc. European Signal Process. Conf.*, 2012.

[6] Y.-F. Ma, L. Lu, H.-J. Zhang, and M. Li, "A user attention model for video summarization," in *Proc. ACM Multimedia*, 2003.

[7] L. Lue and H.-J. Zhang, "Automated extraction of music snippets," in *Proc. Int'l Conf ACM*, 2003.

[8] Y. Ma, X.S. Hua, L. Lu, and H. Zhang, "A generic framework of user attention model and its application in video summarization," *IEEE Trans. on Multimedia*, vol. 7, no. 5, pp. 907–919, Oct 2005.

[9] W. Jiang, C. Cotton, and A.C. Loui, "Automatic consumer video summarization by audio and visual analysis," in *Proc. ICME*, 2011.

[10] P. Nakov, S. Rosenthal, Z. Kozareva, V. Stoyanov, A. Ritter, and T. Wilson, "Semeval 2013 task 2: Sentiment analysis in twitter," in *Proc. of 2nd Joint Conf. on Lexical and Computational Semantics (*SEM), 7th Int'l. Workshop on Semantic Evaluation*, 2013, pp. 312–320.

[11] M. Hu and B. Liu, "Mining and summarizing customer reviews," in *Proc. Conf. on Knowledge Discovery and Data Mining*, 2004.

[12] C. M. Lee and S. S. Narayanan, "Toward detecting emotions in spoken dialogs," *IEEE Trans. on Speech and Audio Process.*, 2005.

[13] P. Turney and M. L. Littman, "Unsupervised learning of semantic orientation from a hundred-billion-word corpus," Tech. Rep. ERC-1094, National Research Council of Canada, 2002.

[14] N. Malandrakis, A. Potamianos, E. Iosif, and S. Narayanan, "Distributional semantic models for affective text analysis," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 21(11), pp. 2379–92, 2013.

[15] A. Zlatintsi, P. Koutras, N. Efthymiou, P. Maragos, A. Potamianos, and K. Pastra, "Quality evaluation of computational models for movie summarization," in *Proc. QoMEX*, Costa Navarino, Greece, May 2015.

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

[17] 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.

[18] 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.

[19] H. M. Teager and S. M. Teager, "Evidence for nonlinear sound production mechanisms in the vocal tract," in *Speech Production and Speech Modelling*, W.J. Hardcastle and A. Marchal, Eds., vol. 15. NATO Advanced Study Institute, Series D, Boston, MA: Kluwer, July 1989.

[20] J.F. Kaiser, "On a simple algorithm to calculate the energy of a signal," in *Proc. IEEE Int'l. Conf. Acoust., Speech, Signal Process.*, 1990.

[21] P. Maragos, J.F. Kaiser, and T.F. Quatieri, "Energy separation in signal modulations with application to speech analysis.," *IEEE Trans. Signal Process*, vol. 41, pp. 30243051, 1993.

[22] D. Dimitriadis, A. Potamianos, and P. Maragos, "A comparison of the squared energy and Teager-Kaiser operators for short-term energy estimation in additive noise," *IEEE Trans. on Signal Process.*, 2009.

[23] D. Dimitriadis, P. Maragos, and A. Potamianos, "On the effects of filterbank design and energy computation on robust speech recognition," *IEEE Trans. on Audio, Speech and Language Process.*, Aug. 2011.

[24] A. Zlatintsi and P. Maragos, "AM-FM modulation features for music instrument signal analysis and recognition," in *Proc. 20th European Signal Processing Conference*, 2012.

[25] R. Plomp and W.J.M. Levelt, "Tonal consonance and critical bandwidth," *Jour. Acoust. Soc. of Am. (JASA)*, vol. 38, pp. 548–560, 1965.

[26] E. Zwicker and H. Fastl, *Psychoacoustics, Facts and Models*, Springer, 2nd edition, 1999.

[27] W. Sethares, *Tuning, Timbre, Spectrum, Scale*, Springer-Verlag, 1998.

[28] P.N. Vassilakis, *Perceptual and Physical Properties of Amplitude Fluctuation and their Musical Significance*, Ph.D. thesis, Univ. of California, 2001.

[29] G. Evangelopoulos, A. Zlatintsi, G. Skoumas, K. Rapantzikos, A. Potamianos, P. Maragos, and Y. Avrithis, "Video event detection and summarization using audio, visual and text saliency," in *Proc. Int'l Conf. Acoustics, Speech and Signal Processing*, 2009.

[30] N. Malandrakis, A. Potamianos, G. Evangelopoulos, and A. Zlatintsi, "A supervised approach to movie emotion tracking," in *Proc. ICASSP*, Prague, Czech Republic, May 2011.

[31] Z. Harris, "Distributional structure," *Word*, vol. 10(23), 1954.

[32] M. Bradley and P. Lang, "Affective norms for English words (ANEW): Stimuli, instruction manual and affective ratings. Tech. report C-1.," The Center for Research in Psychophysiology, Univ. of Florida, 1999.

[33] P. Maragos, *The Image and Video Processing Handbook*, chapter Morphological Filtering for Image Enhancement and Feature Detection, pp. 135–156, Elsevier Acad. Press, 2nd edition, 2005.