Efficient Video Coding in H.264/AVC by using Audio-Visual Information

Jong-Seok Lee [#1], Touradj Ebrahimi [#2]

[#] *Multimedia Signal Processing Group, Ecole Polytechnique Fédérale de Lausanne (EPFL)*
CH-1015 Lausanne, Switzerland
[1] jong-seok.lee@epfl.ch
[2] touradj.ebrahimi@epfl.ch

***Abstract*—This paper proposes an efficient video coding method which utilizes audio-visual information, based on the observation that sound-emitting regions in a video sequence attract observer's attention. The regions responsible for the sound are identified by an audio-visual source localization algorithm. Then, the result is used for encoding different regions in the scene with different quality in such a way that a region far from the sound source is coded with a lesser quality than the sound-emitting regions. This is implemented by assigning different quantization parameter values for different regions in H.264/AVC. Experimental results demonstrate the effectiveness of the proposed approach.**

I. Introduction

In the field of video coding, the characteristics of the human visual system are often exploited in order to achieve coding efficiency. When humans watch multimedia content, regions around the focus of attention are perceived with a higher scrutiny. Therefore, efficient coding can be performed in a way that regions of lesser interest where the attention is not expected to be located are encoded with lower quality without much degradation of perceived overall quality.

The region of interest (ROI) for efficient video coding can be defined in various ways. Itti [1] considered the conspicuity in terms of color, intensity, motion, etc., as the element attracting the focus of attention in his neurobiologically motivated bottom-up model. The work in [2] combined this bottom-up cues and the top-down information (i.e. human faces) in the Bayesian framework to improve the fixation model. Cavallaro *et al*. [3] assumed that moving objects in the scene are likely to be the focus of interest. Tang [4] combined different visual factors of attention such as a motion attention model, a spatiovelocity visual sensitivity model, and a visual masking model, in order to define ROIs.

Although these works have been shown to be effective through subjective and objective experiments, they are still far from the actual mechanism of attention in the human visual system. One of the interesting and important aspects of the human visual system is the visual attention guided by the acoustic modality, which has been rarely addressed previously. We can imagine that some audio events around us often attract our visual attention, e.g. making us turn our head or eyes toward the sound source. Psychological studies have demonstrated the importance of the acoustic signal in the visual attention: An auditory stimulus in a particular location draws visual attention occurring at the same spatial location [5]. Moreover, such orientation of attention can improve perception of the subsequent visual stimulus [6].

In this paper, we propose an efficient video coding method based on audio-visual information, which is implemented in the framework of H.264/AVC. We assume that the sound source and its neighboring region in the multimedia content play the role of an ROI and attract humans' visual attention, which can be used for efficient video coding. First, the sound-emitting regions are identified by an audio-visual source localization algorithm, which is applied to conventional video sequences containing only one audio channel. Then, the image frame is divided into several regions according to their spatial distance from the sound source. By utilizing the flexible macroblock ordering (FMO) scheme in H.264/AVC, each region is mapped into a slice which is encoded with a different quantization parameter (QP); for a slice far from the sound source, a large QP value is assigned so that a small number of bits is spent for encoding the slice in comparison to a slice near the sound source. The experimental results show that the proposed method can reduce the amount of necessary bits for encoding without a perceived quality degradation.

The organization of the rest of the paper is as follows: In Section II, the audio-visual source localization algorithm is described. Section III explains the way of performing efficient video coding by using the result of the source localization. Section IV demonstrates the effectiveness of the proposed method. Finally, the conclusion is made in Section V.

II. Audio-Visual Source Localization

Audio-visual source localization is to identify the spatial location of the sound source in a scene. It is useful when the scene contains multiple moving objects so that a conventional motion detection approach with only the visual information may not work satisfactorily. In this paper, it is used for finding the sound-emitting region which is expected to be attended by human observers and thus needs to be encoded in higher quality when compared to other regions.

The audio-visual source localization method used in this paper is based on the canonical correlation analysis (CCA) and finds the pixel location which shows the maximum correlation with the acoustic signal [7]. The first step of the method is to extract features for the acoustic and the visual modalities. In this paper, the difference of the luminance

MMSP'09, October 5-7, 2009, Rio de Janeiro, Brazil.
978-1-4244-4464-9/09/$25.00 ©2009 IEEE.

component between two consecutive frames is used for the visual features. The acoustic energy calculated over a moving window is extracted for the acoustic features. The window moves at the rate of the visual frame rate so that the acoustic and the visual features are temporally synchronous.

When a pair of the feature vectors is given, CCA finds the projection vectors by which the correlation of the projected data becomes the maximum. If we denote $\mathbf{x} \in \mathrm{R}^m$ and $\mathbf{y} \in \mathrm{R}^n$ the acoustic and the visual feature vectors, respectively, the projection vectors found by CCA, $\mathbf{w}_x$ and $\mathbf{w}_y$, are obtained by

$$\mathbf{w}_x, \mathbf{w}_y = \arg\max_{\mathbf{w}_x, \mathbf{w}_y} \frac{\mathbf{w}_x^T (\mathbf{Y}^T \mathbf{X}) \mathbf{w}_y^T}{\sqrt{\mathbf{w}_x^T (\mathbf{X}^T \mathbf{X}) \mathbf{w}_x \mathbf{w}_y^T (\mathbf{Y}^T \mathbf{Y}) \mathbf{w}_y}}, \quad (1)$$

where $\mathbf{X} = \{\mathbf{x}_i\}_{i=t}^{t+N-1}$ and $\mathbf{Y} = \{\mathbf{y}_i\}_{i=t}^{t+N-1}$ are the collections of the acoustic and the visual feature vectors over N frames, respectively, and which contain a feature vector for each time point i in their rows. In our case, the value of m is 1 because we use only the energy component of the acoustic signal. Consequently, $\mathbf{w}_x$ becomes a scalar value which can be considered in $\mathbf{w}_y$, and thus can be regarded as unity. Since we use the differential pixel values for the visual features, the value of n is basically equal to the number of pixels in images. However, in order to reduce the computational complexity, we ignore the pixel locations where there is no pixel value change and, thus, n becomes much less than the number of pixels. This selection is updated at every time step t.

Alternatively, solving the above problem is equivalent to solving the following problem [8]:

$$\mathbf{X} = \mathbf{Y}\mathbf{w}_y. \quad (2)$$

Since the dimension of $\mathbf{w}_y$ is usually much larger than N, if $\mathbf{Y}$ is full rank, there are an infinite number of solutions for the above equation. Imposing a spatial sparsity criterion leads to a unique solution:

$$\min \| \mathbf{w}_y \|_1 \quad \text{subject to} \quad \mathbf{X} = \mathbf{Y}\mathbf{w}_y, \quad (3)$$

where $\|\cdot\|_1$ is the l^1-norm. The above constrained optimization problem can be solved by the linear programming technique. The solution $\mathbf{w}_y$ can be interpreted as "cross-modal energy" concentrated on the pixel location responsible for the sound signal.

In order to consider the spatio-temporal consistency of the solutions over multiple frames, a weighting scheme is incorporated in the above formulation. Then, the problem (3) is modified as

$$\min \sum_{i=1}^{n} | f_i w_{yi} | \quad \text{subject to} \quad \mathbf{X} = \mathbf{Y}\mathbf{w}_y, \quad (4)$$

where w_{yi} is the i-th component of $\mathbf{w}_y$ and f_i the weighting factor given by

$$f_i = \max_{1 \le j \le n} w_{yj}^{old} - w_{yi}^{old} + 1, \quad (5)$$

where w_{yi}^{old} is the i-th component of the spatially smoothed version of the solution for the previous frame. In other words, if a pixel location has received a large energy value in the solution for the previous frame, the weight values for the location and its neighbouring pixels are set to be small so that large values are obtained for these locations in the solution for the current frame. Adding 1 in (5) ensures that all weights are greater than zero. Smoothing of the solution for the previous frame is done by applying a Gaussian filter to the solution. The problem (4) is also solved by the linear programming.

Our source localization method has advantages over existing methods in the following senses: It does not require a special setup with multiple microphones such as microphone arrays in previous approaches [9], but only one channel acoustic signal is used. Moreover, unlike previous approaches [10], it does not have any assumption on the region or object to be localized. Also, it does not require a training phase which may need manually processed training data.

III. Coding based on Audio-Visual Information

The result of the audio-visual source localization is used to determine which region is to be encoded in higher and which in lower quality. The quality of a region is controlled by varying the QP in H.264/AVC. By assuming that the sound-emitting region is attended more than the other regions, a large QP is assigned to regions far from the sound source, thereby coding efficiency can be obtained without much degradation of perceived quality.

First, a priority map is generated based on the localization result, which represents the weighted distance between each pixel and the nearest localized energy location (Fig. 2(b)). When there are multiple energy locations, a pixel near a smaller energy is assigned with a larger distance than one near a larger energy. We use the Euclidean distance to measure the distance between pixels. Then, the image frame is divided into L partitions (slices) according to the priority. The linearly spaced values within the range of the priority values are assigned as the boundaries of the partitions.

The FMO scheme in H.264/AVC is used for assigning different QPs for different slices. A slice is a group of macroblocks to be encoded together. Each slice can be decoded independently. In H.264/AVC, there are six different pre-defined types of grouping patterns (Types 0 to 5). However, in our case, a slice can have an arbitrary shape depending on the priority map, which does not correspond to pre-defined Types 0 to 5 but Type 6 in FMO. Fig. 2(c)-(d) shows examples of slice grouping with different values of L.

The QP for slice j, QP(j), is determined by

$$\mathrm{QP}(j) = \min\left[\mathrm{QP}_0 + j \cdot \Delta\mathrm{QP},\ 51\right], \quad j = 0, \ldots, L-1, \quad (6)$$

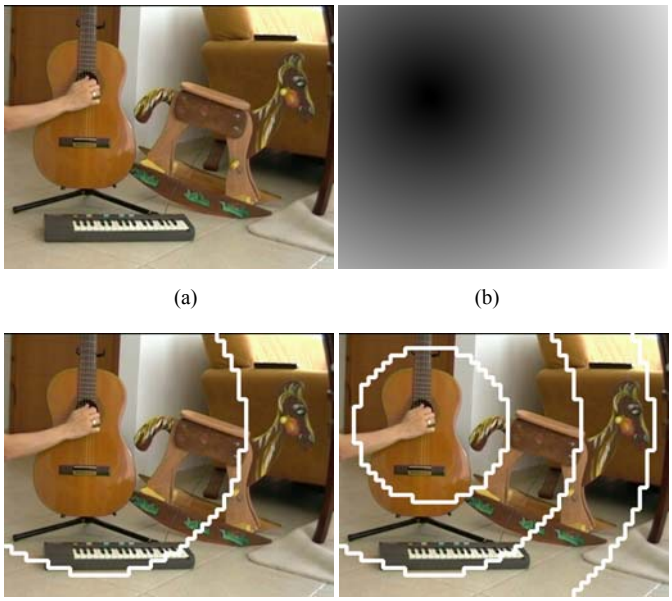
where QP_0 is the QP value for the highest priority (i.e. the sound-emitting region) and ΔQP the incremental value of QP between each slice.

In order to reduce the overhead for sending additional bits containing the slice information, we update the slice groups only at a pre-defined rate (every fifth frame in our case).

IV. Experiments

A. Setup

Two test video sequences are used in our experiments. In Data #1, a hand plays a guitar and then moves to play a synthesizer, while a wooden horse is rocking throughout the whole sequence [8]. Data #2 is selected from the "groups" section of the CUAVE database [11], where three people



(a) (b)

(c) (d)

Fig. 1. Example image frame where the guitar sound is produced by the hand while a wooden horse is rocking at the same time. (a) Original frame. (b) Priority map. (c) Slice boundary for L=2. (d) Slice boundaries for L=4.

pronounce continuous English digits in turn; while a person is speaking, other persons also move their heads and mouths. The two video sequences are about 10 seconds long. Their frame rates are 25 fps and 29.97 fps, respectively.

As mentioned in Section II, the difference of the luminance component between two consecutive frames is used for the visual features and the acoustic energy calculated over a moving window is used for the acoustic features. The value of N was set to 32 for Data #1 and 16 for Data #2, where we consider different rates of sound and motion in the two data.

We used the JM Reference Software version 15.1 for H.264/AVC coding [12]. The constant QP mode, the rate control (adaptive QP) scheme, and the proposed method are compared. The encoder uses the baseline profile in order to use the FMO scheme. The RDO is enabled. The reference frame number is set to 5. The search range of full search motion estimation is 32. The context adaptive variable length coding (CAVLC) is used.

A. Results

Figs. 2 and 3 show the relative reduction of bitrate of the sequences produced by the proposed approach when compared to those produced by JM with fixed QP. When QP_0 is small, the advantage of the proposed method in terms of coding gain is clearly seen. On the other hand, for large QP_0 values, the gain is small and even becomes negative because the amount of the bits for sending the slice group information is not negligible any more. Also, the effects of L and ΔQP can be observed; using large values of L or ΔQP produces large gains. The gain for Data #1 is larger than that for Data #2 because the background region which is encoded with large QP values in the proposed method contains larger motion by

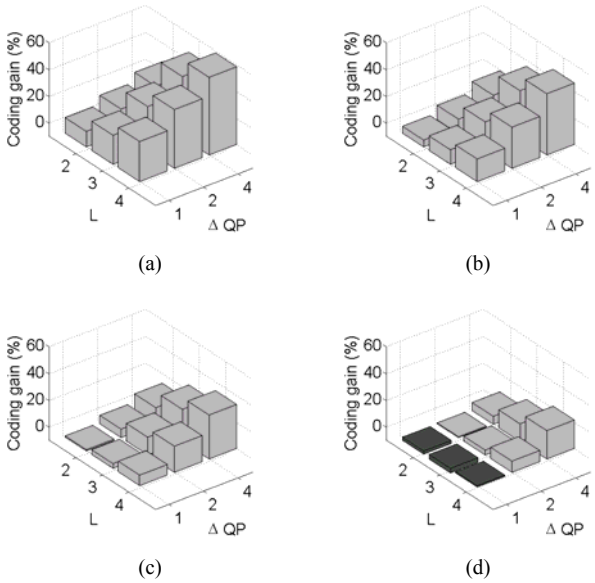


Fig. 2. Coding gain (%) by the proposed method in comparison to JM (fixed QP) for Data #1 when (a) QP_0=22, (b) QP_0=26, (c) QP_0=30 and (d) QP_0=34. The dark bars indicate negative values.

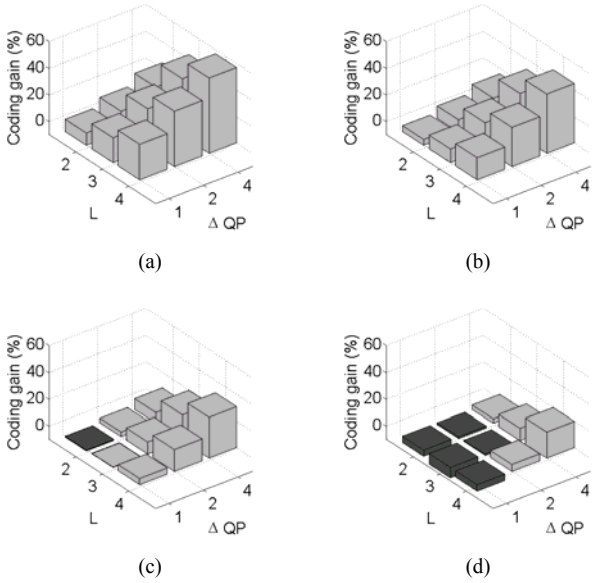


Fig. 3. Coding gain (%) by the proposed method in comparison to JM (fixed QP) for Data #2 when (a) QP_0=22, (b) QP_0=26, (c) QP_0=30 and (d) QP_0=34. The dark bars indicate negative values.

the rocking horse in Data #1 than that by the silent people in Data #2.

In order to investigate the perceived quality of the encoded video sequences, we conducted a simple subjective quality evaluation. The following four conditions are compared:

- JM with constant QP=26;
- Proposed method with QP_0=26, L=4, ΔQP=1;
- Proposed method with QP_0=26, L=4, ΔQP=2;
- Proposed method with QP_0=26, L=4, ΔQP=4.

The test has been performed according to the guidelines provided by standards [13]. The double stimulus continuous

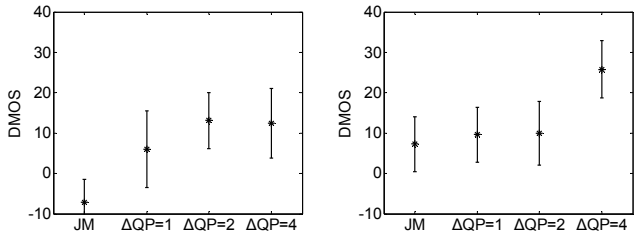


Fig. 4. Results of the subjective test comparing JM and the proposed method with three different values of ΔQP. The differential mean opinion score (DMOS) values and confidence intervals are shown for (a) Data #1 and (b) Data #2.

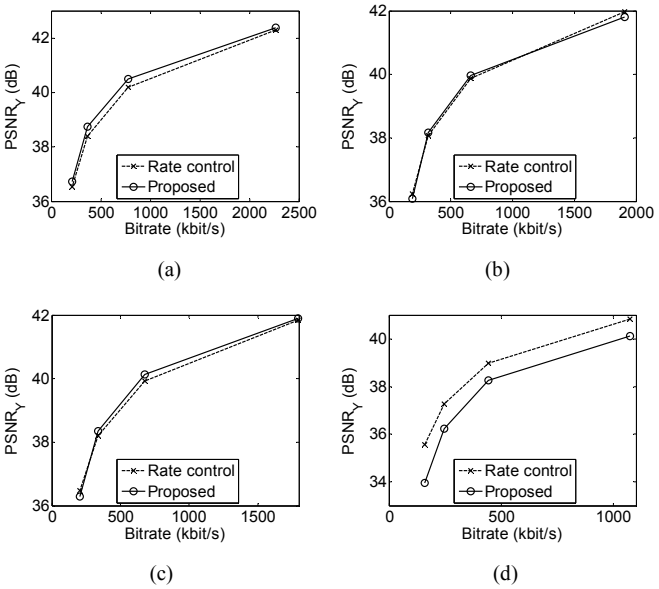


Fig. 5. Rate-distortion curves obtained from the JM rate control scheme and the proposed method for Data #1. (a) *L*=2, ΔQP=1. (b) *L*=2, ΔQP=4. (c) *L*=4, ΔQP=1. (d) *L*=4, ΔQP=4.

quality scale (DSCQS) method was used. Eleven subjects participated in the test.

Fig. 4 shows the results of the test in terms of the differential mean opinion score (DMOS) values and the 95% confidence interval for the two test sequences. It can be inferred that the difference of the perceived quality between JM and the proposed method is statistically insignificant when the value of ΔQP is small (i.e. ΔQP=1 for Data #1 and ΔQP=1 or 2 for Data #2), where the coding gain is 17% for Data #1 and 17% and 29% for Data #2, respectively. Since Data #1 contains large motion by the rocking horse in the background, the subjects could notice quality degradation in that region when ΔQP becomes larger than 1. On the other hand, the motion in the background of Data #2 is rather small and thus the quality degradation is not noticeable until ΔQP becomes larger than 2.

Figs. 5 and 6 compare the rate-distortion curves obtained by using the rate control scheme of JM and the proposed method for the two test sequences, respectively. Each curve is created by using the initial QP values of 22, 26, 30 and 34. It

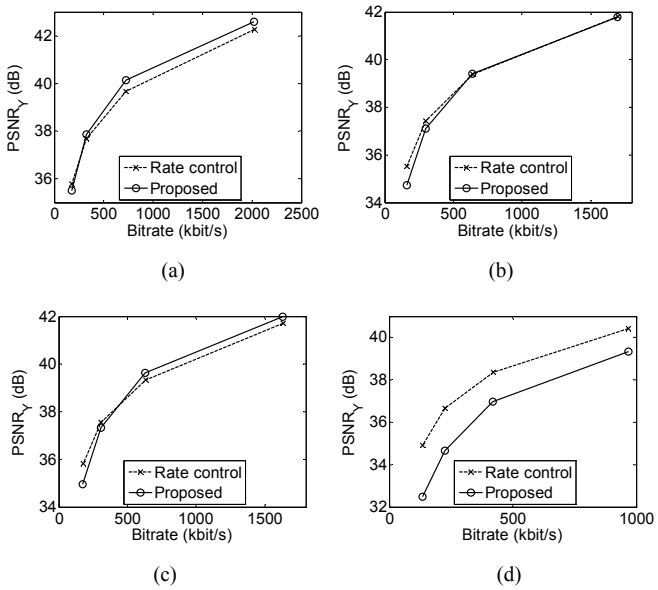


Fig. 6. Rate-distortion curves obtained from the JM rate control scheme and the proposed method for Data #2. (a) *L*=2, ΔQP=1. (b) *L*=2, ΔQP=4. (c) *L*=4, ΔQP=1. (d) *L*=4, ΔQP=4.

is observed that, for small values of *L* and ΔQP, higher PSNR values are obtained by the proposed method compared to the rate control scheme. As the values of *L* and ΔQP are larger, the quality of the sequences produced by the proposed method becomes worse than those by JM with rate control. However, it should be noted that lower PSNR values by the proposed method than those by JM with rate control do not always imply that the video sequences created by the proposed method have worse quality perceived by humans, because measuring PSNR values does not consider humans' focus of attention but computes average distortion over the whole image frames.

Fig. 7 shows example frames of the sequences produced by the proposed method and the rate control scheme of JM to compare their subjective quality when the two sequences have the same bitrate. When one observes the region around the hand which plays the synthesizer, where more attention is expected to be placed, it can be noticed that the details in this region are clearer in the image frame produced by the proposed method when compared to that by JM, because more bits are allocated to regions producing the sound signal.

V. CONCLUSION

We have presented a method for ROI coding based on audio-visual information. After the audio-visual source localization algorithm detects the location of the sound source in the scene, the priority of allocating quality is given to the regions close to the source. We used the FMO scheme in H.264/AVC in order to assign different QP values for different regions according to their distance from the source. The experimental results showed that we could obtain encoding efficiency in terms of bitrates without perceived quality degradation.

(a)

(b)

Fig. 7. Example frames of the sequences produced by (a) JM 15.1 with the rate control scheme and (b) the proposed approach (QP_0=30, L=4, ΔQP=2), which have the same bitrate of 293 kbps.

In our further work, we will perform more thorough subjective tests to evaluate the effectiveness of the proposed method in terms of perceived quality. We also plan to conduct experiments with high definition content, in which we expect that our approach is even more effective when compared to standard definition content used in this paper.

ACKNOWLEDGMENT

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2011) under grant agreement no. 216444 (PetaMedia), and the Swiss NCCR Interactive Multimodal Information Management (IM2).

REFERENCES

[1] L. Itti, "Automatic foveation for video compression using a neurobiological model of visual attention," *IEEE Trans. Image Processing*, vol. 13, no. 10, pp. 1304-1318, 2004.

[2] G. Boccignone, A. Marcelli, P. Napoletano, G. D. Fiore, G. Iacovoni, and S. Morsa, "Bayesian integration of face and low-level cues for foveated video coding," *IEEE Trans. Circuits, Syst. Video Technol.*, vol. 18., no. 12, pp. 1727-1740, Dec. 2008.

[3] A. Cavallaro, O. Steiger, and T. Ebrahimi, "Semantic video analysis for adaptive content delivery and automatic description," *IEEE Trans. Circuits, Syst. Video Technol.*, vol. 15, no. 10, pp. 1200-1209, Oct. 2005.

[4] C.-W. Tang, "Spatiotemporal visual considerations for video coding," *IEEE Trans. Multimedia*, vol. 9, no. 2, pp. 231-238, Feb. 2007.

[5] J. Driver and C. Spence, "Attention and the crossmodal construction of space," *Trends in Cognitive Sciences*, vol. 2, no. 7, pp. 254-262, Jul. 1998.

[6] J. J. McDonald, W. A. Teder-Salejarvi, F. D. Russo, and S. A. Hillyard, "Neural substrates of perceptual enhancement by cross-modal spatial attention," *Journal of Cognitive Neuroscience*, vol. 15, no. 1, pp. 10-19, 2003.

[7] J.-S. Lee, F. De Simone, and T. Ebrahimi, "Video coding based on audio-visual attention," in *Proc. ICME'09*, 2009, pp. 57-60.

[8] E. Kidron, Y. Y. Schechner, and M. Eland, "Cross-modal localization via sparsity," *IEEE Trans. Signal Processing*, vol. 55, no. 4, pp. 1390-1404, Apr. 2007.

[9] H. Zhou, M. Taj, and A. Cavallaro, "Target detection and tracking with heterogeneous sensors," *IEEE J. Sel. Top. Sign. Proc.*, vol. 2, no. 4, pp. 503-513, Aug. 2008.

[10] P. Besson, V. Popovici, J.-M. Vesin, J.-P. Thiran, and M. Kunt, "Extraction of audio features specific to speech production for multimodal speaker detection," *IEEE Trans. Multimedia*, vol. 10, no. 1, pp. 63-73, Jan. 2008.

[11] E. Patterson, S. Gurbuz, Z. Tufekci, and J. N. Gowdy, "CUAVE: a new audio-visual database for multimodal human-computer interface research," in *Proc. ICASSP'02*, 2002, pp. 2017-2020.

[12] H.264/AVC JM Reference Software. [Online]. Available: http://iphome.hhi.de/suehring/tml

[13] Recommendation ITU-R BT.500-11, "Methodology for the subjective assessment of the quality of television pictures," Geneva, Switzerland, 2002.