

# Detection of Lip Synchronization Artifacts

Ignacio Blanco Fernández<sup>1</sup>(✉) and Mikołaj Leszczuk<sup>2</sup>

<sup>1</sup> Department of Telecommunications, Polytechnic School of Engineering of Gijón,  
Gijón, Spain

[gncblncfrnndz@gmail.com](mailto:gncblncfrnndz@gmail.com)

<sup>2</sup> Department of Telecommunications, AGH University of Science and Technology,  
Al. Mickiewicza 30, 30059 Kraków, Poland

[vq@kt.agh.edu.pl](mailto:vq@kt.agh.edu.pl)

<http://vq.kt.agh.edu.pl/>

**Abstract.** Over 10 billion hours of video are watched each month on the Internet, what, together with high definition television broadcasting and the rise in high quality video on demand makes the task of quality assessment a key one in the global multimedia market nowadays. Automating quality checking is currently based on finding major audiovisual artifacts. The Monitoring Of Audio Visual quality by key Indicators (MOAVI) subgroup of the Video Quality Experts Group (VQEG) is an open collaborative project for developing No-Reference models for monitoring audiovisual service quality. The purpose of this paper is to report the development of the audiovisual part of this project, which includes the detection of lip synchronization (also known as lip sync) artifacts.

**Keywords:** MOAVI · VQEG · Lip sync

## 1 Introduction

Automating quality checking is currently based on finding major video and audio artifacts. The Monitoring Of Audio Visual quality by key Indicators (MOAVI) subgroup of the Video Quality Experts Group (VQEG) is an open collaborative project for developing No-Reference (NR) models for monitoring audiovisual service quality. MOAVI is a complementary, industry-driven alternative to Quality of Experience (QoE), used as a subjective measure of a viewer's experiences.

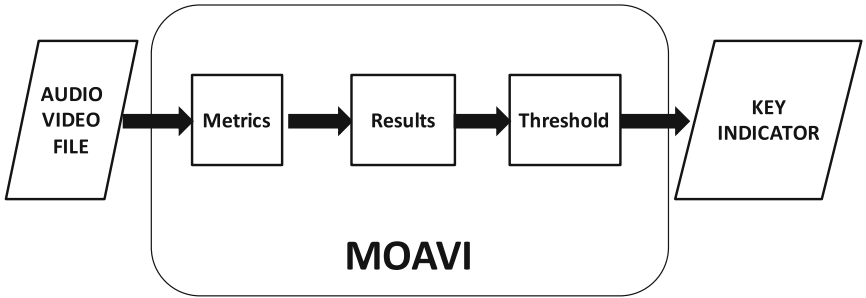
Current NR QoE models, like the reported in related research work [11], followed the less useful Full-Reference (FR) models (e.g. [4]), address measuring quality of networked multimedia, using objective parametric models. Still, these models might have slight problems in predicting overall audiovisual QoE. MOAVI can be used alternatively to automatically measure audiovisual quality by using simple indicators of perceived degradation.

The goal of the project is to develop a set of key indicators (including blockiness, blur, freeze/jerkiness effects, block missing errors, slice video stripe errors, aspect ratio problems, field order problems, interlace, lip synchronization, also known as lip sync, mute, and clipping [2]) describing service quality in general

(the list is not closed, but the major artifacts are presented), and to select subsets for each potential application. Therefore, the MOAVI project concentrates on models based on key indicators contrary to models predicting overall quality.

The video signal needs some signal processing to be performed on. Quality checking can be conducted before, during, and/or after the encoding process. However, in MOAVI, no Mean Opinion Score (MOS) is provided. A binary indicator for each artifact is provided instead showing its presence or absence.

Figure 1 shows the concept of MOAVI. The audio or video stream (only video for video artifacts, only audio for audio artifacts, and both of them together if the artifact is an audiovisual one) is the input to the system. The metric of each artifact is used to determine the level of impairment that the media to be analyzed suffers. These results are converted into binary indicators using a threshold that would determine if the artifact is in a noticeable level in the video or if it is not. In that way, MOAVI obtains a key indicator for each artifact.



**Fig. 1.** Concept of monitoring of audiovisual quality

This paper is organized as follows. Section 2 describes measuring the key audiovisual indicator – presence of lip sync. Section 3 presents the video database for the assessment of the metrics. Sections 4–6 describe the algorithms and the obtained results. Section 7 concludes the paper and summarizes the results.

## 2 Measuring Lip Sync Artifact

In the present research the process followed for the detection of audiovisual artifacts is exposed. Therefore, this paper includes the description of both the algorithm, implementation and results of three different metrics. These metrics were developed to indicate the presence or the absence of the most frequent audiovisual problem affecting an audiovisual signal, which is lip sync problem.

Lip sync is a key parameter in interactive communication. In the case of video conferencing, streaming and television broadcasting, the uneven delay between audio and video should remain below certain thresholds, recommended by several standardization bodies. However, further research has shown that the thresholds can be relaxed, depending on the targeted application and use case [9].

In multimedia systems, synchronization is needed to ensure a temporal ordering of events. For single data streams, a stream consists of consecutive Logical Data Units (LDU). In the case of an audio stream, LDU are individual samples or blocks of samples transferred together from a source to one or more sinks. Similarly with video, one LDU may typically correspond to a single video frame and consecutive LDU – to a series of frames. These have to be presented at the sink with the same temporal relationship as they were captured giving so called “intra-stream”. The temporal ordering must also be applied to related data streams, where one of the more common relationships is the simultaneous playback of audio and video with lip sync. Both media must be “in sync”, otherwise the result will not be adjudged as satisfactory.

In general, “inter-stream” synchronization involves relationships between all kinds of media including pointers, graphics, images, animation, text, audio and video. In the following discussion, “synchronization” always refers to “inter-stream” synchronization between video and audio.

Some facts about the problem of lack of lip sync:

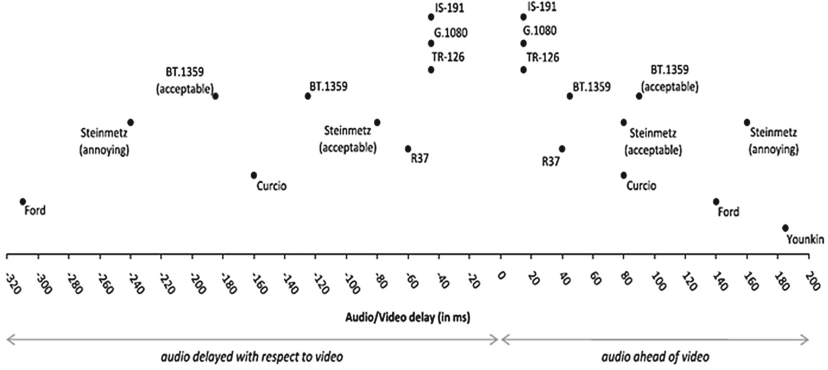
- The most common origin for lip sync artifact is the jitter produced in the transmission stage.
- Different languages make no big difference in the task of synchronizing media.
- Different languages make no big difference in the task of detecting lip sync artifact, both for human perception and for automatic detection.
- In [10] it is also stated that professional video editors and TV-related technical personnel showed a smaller level of skew tolerance. When they detected an error they could correctly state if audio is ahead of or behind video.
- Watermarks or fingerprints embedded in audio signal have been used in broadcasts to avoid this problem. However this fingerprints are not suitable for multimedia streaming through the internet.

Regarding detection thresholds, [9] refers to the large amount of different ones that authors have determined. Some authors and research groups have concluded that audio may be played up to 305 ms ahead of video and conversely video displayed up to 190 ms ahead of the audio. Both temporal skews are noticed, but can be accepted by the user without any significant loss of effect. Some authors however report a tolerance of only 4–16 ms to be acceptable.

Figure 2 shows a graphical representation of the different audio/video delay and lip sync thresholds of detectability as identified by several standard bodies and already conducted research by independent studies. The thresholds used for the metric of lip sync artifact in MOAVI are set to 100 ms when audio is delayed with respect to video and 140 ms when video is delayed versus audio. These thresholds are based on the research work we refer here [10].

### 3 Video Database for Assessment of Metrics

The development of experiments with the objective of analyzing the behavior and measuring the accuracy of the different metrics in this section requires the



**Fig. 2.** Different audio/video delay and lip sync thresholds of detectability

storage of a small database of videos and some key information about them. It is a set of 15 video sequences with lengths between 13 and 37 s, coming from all kinds of media. The videos are all taken from frontal view, although some of them include several frames in which there is a profile view. Usually only the face and the shoulders are visible. Finally, only one person is seen and listened to in each video.

Some of the sources of videos are TV news shows, others come from interviews. A small group of them are videos uploaded directly to the internet.

**Table 1.** Characteristics of the video database for the assessment of the metrics

Video	Length (s)	View	Visible	Movement
ABERCROMBIE	19,8	FRONTAL	HALF BODY	MEDIUM
ANGIE	21,6	FRONTAL	SHOULDERS	LOW
AYALA	13,9	FRONTAL	SHOULDERS	LOW
BECKHAM	18,2	FRONTAL	SHOULDERS	LOW
DICAPRIO	18,3	FRONTAL	HALF BODY	HIGH
FOXNEWS	14,3	FRONTAL	SHOULDERS	LOW
GOOGLE	27,7	FRONTAL	SHOULDERS	LOW
HAYS	25,4	FRONTAL	SHOULDERS	MEDIUM
LARRYPAGE	24,4	FRONTAL	HEAD	LOW
LISA	26,2	FRONTAL	HEAD	MEDIUM
MORRIS	24,1	FRONTAL	SHOULDERS	LOW
RESUME	25,3	FRONTAL	SHOULDERS	MEDIUM
STOSSEL	22,2	FRONTAL	HALF BODY	LOW
USAJOBS	17,9	FRONTAL	SHOULDERS	LOW
USAJOBS2	19,9	FRONTAL	SHOULDERS	LOW

The most important characteristics of each of the videos are shown in Table 1. The audio files extracted from the videos have been stored and analyzed too, in order to use them for the tests of Voice Activity Detection (VAD).

MOAVI's indicator for lip sync is based on the lip sync metric that is explained in the next sections. In the first of the sections, the audio part of the metric is explained. Signal processing used to implement a VAD algorithm is described in this first section. In the second section, the part of the metric regarding video is exposed. The combination of techniques used to detect the movement of the lips are explained. In the third and last section, the algorithm that compares the audio and visual information between each other is described. Every section includes a results subsection and a further research subsection too that will complete the approach to the method developed to detect the delay between a visual media and an audio media.

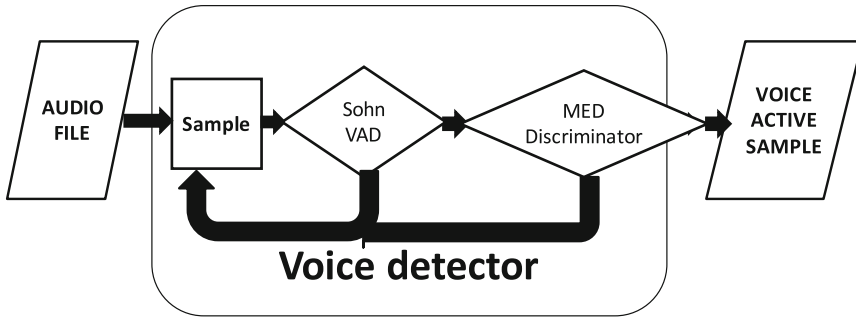
## 4 Voice Activity Detector

Developing an indicator that analyzes if the audio and the video is synchronized is a challenging goal. Nevertheless, if the process is divided into small parts, it is simplified. Therefore, the first algorithm to develop is a VAD.

### 4.1 Algorithm

In lip sync, processing the signal in utterances consisting of speech, silence, and other background noise is needed. The detection of the presence of speech embedded in various types of non-speech events and background noise is called endpoint detection, voice detection, or VAD.

The VAD algorithm consists on basically two steps. The algorithm for the detection of voice is represented in Fig. 3. The two detectors are used complementarily to obtain better results than applying just one of them.



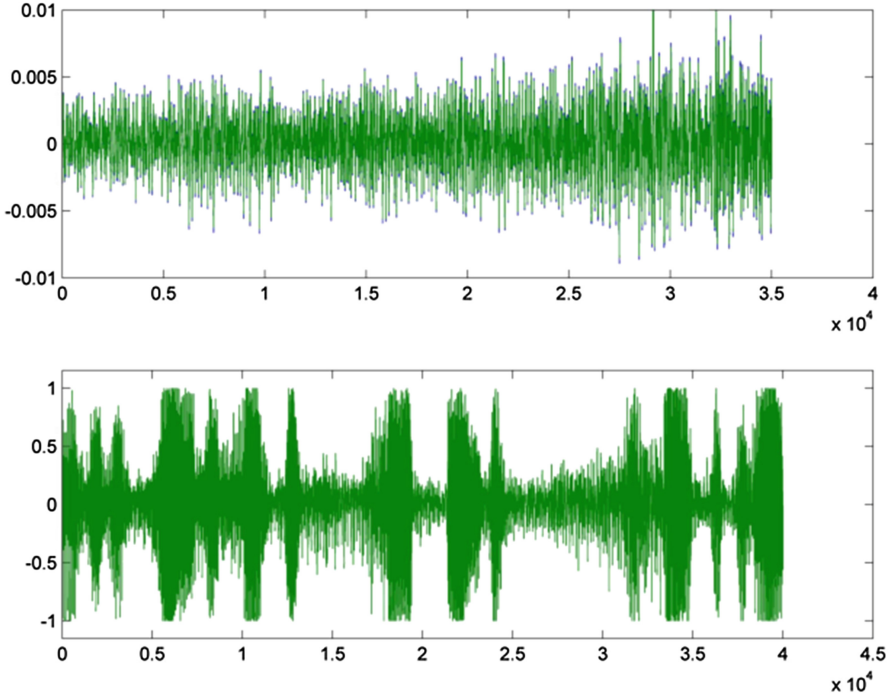
**Fig. 3.** Algorithm for the detection of speech instants artifact

The first step is the signal processing leading to detect the endpoints of the voice in the audio. An algorithm based on [8] was developed in MATLAB.

The second step is the analysis of the Minimum Energy Density (MED) feature which is a key distinction between music and similar waveforms and speech waveforms. The algorithm is described in the related research work [5] and based on that algorithm, the MATLAB code has been completed.

In [8], a VAD for variable rate speech coding is decomposed into two parts, a decision rule and a background noise statistic estimator, which are analyzed separately by applying a statistical model. A robust decision rule is derived from the generalized likelihood ratio test by assuming that the noise statistics are known a priori. To estimate the time-varying noise statistics, allowing for the occasional presence of the speech signal, a novel noise spectrum adaptation algorithm using the soft decision information of the proposed decision rule is developed. The algorithm is robust, especially for the time-varying noise.

In [5], MED is used in discrimination of audio signals between speech and music. This method is based on the analysis of local energy for local subsequences of audio signals. The subsequences in the proposed method will be the ones in which voice activity has been detected in the first detector. An elementary analysis of the probability density for the power distribution in these subsequences is an effective tool supporting the decision making. It is very intuitive to try to discriminate speech and music based on shape of signal's energy envelope. As Fig. 4 shows, speech signal has characteristic high and low amplitude parts,



**Fig. 4.** Comparison between a music waveform (up) and a speech waveform (down)

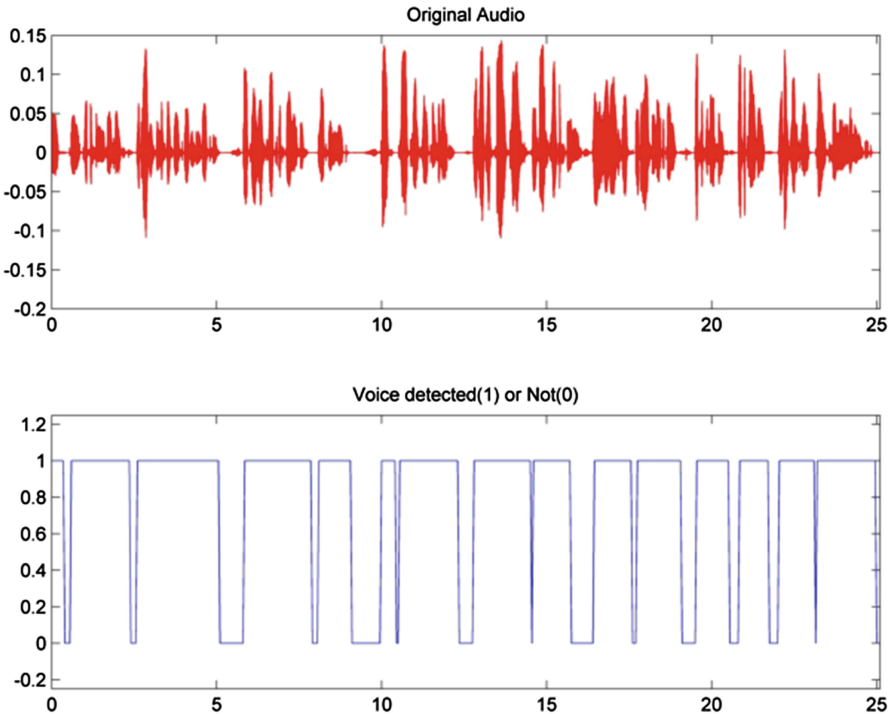
which represent voiced and unvoiced speech, respectively. On the other hand, the envelope of music signal is more steady. Moreover, it's known that speech has a characteristic 4 Hz energy modulation, which matches the syllabic rate.

Considering these characteristics, a decision is taken to discriminate between speech and music subsequences using for that issue the probability density function of short time frame energy inside some time window, which we refer to as a normalization window. The window has to be long enough to capture the nature of the signal. The value of the length of the window chosen is 200 ms, when the subsequence of speech after the first discriminator is longer than that value.

As it has been explained, these two algorithms work together to make the resulting combination more robust and to improve the accuracy of the metric in order to provide a better information which will be later compared with information coming from video, and finally will provide a lip sync artifact indicator.

## 4.2 Results

Regarding the results of the developed VAD for MOAVI, the output of the metric would be like the one presented in Fig. 5. As it can be seen, the metric provides an accurate classification of samples. Every subsequence of 50 ms is



**Fig. 5.** Example of detection of voice

classified into two different values: voiced (1) or unvoiced (0). Thus, a binary vector is constructed to be compared with the information coming from video about endpoints of speech. The final goal would be the calculation of the delay that one of the signals may have with respect to the other. The binary vector coming from the VAD metric described above is stored.

These results have been compared with the ground truth prepared by listening to the 15 audio files and developing a small database for each sound in which every instant is classified between voiced or unvoiced with a precision of 50 ms. Table 2 shows the Hamming distance, the precision, the accuracy and the F1 metric for each of the video files stored.

**Table 2.** Accuracy results of the VAD algorithm in each video from the database

Audio	Hamming Distance	Precision	Accuracy	F1 Metric
ABERCROMBIE	4	0.98	0.98	0.99
ANGIE	33	0.82	0.85	0.90
AYALA	14	0.96	0.90	0.94
BECKHAM	28	0.91	0.85	0.91
DICAPRIO	24	0.96	0.87	0.92
FOXNEWS	6	0.96	0.96	0.98
GOOGLE	32	1.00	0.88	0.93
HAYS	14	0.97	0.94	0.97
LARRYPAGE	22	0.95	0.91	0.95
LISA	15	0.94	0.94	0.97
MORRIS	4	0.98	0.98	0.99
RESUME	22	0.94	0.91	0.95
STOSSEL	8	0.99	0.96	0.98
USAJOBS	16	0.96	0.91	0.94
USAJOBS2	7	0.97	0.97	0.98

Table 3 shows the same parameters describing the performance of the metric as the Table 2, but this time the data shows the results for all the videos together. It has to be highlighted that the VAD algorithm has an accuracy of 92.17% and an F1 metric of 95.47% regarding the measurements made based on the database.

**Table 3.** Accuracy results of the VAD algorithm in the whole video database

Total Frames	Hamming Distance	Precision	Accuracy	F1 Metric
3182	249	0.95	0.92	0.95



## 5 Lip Activity Detector

This section exposes the sub-metric of the lip sync metric based on video analysis. The combination of techniques detecting the frames with lip motion is explained.

### 5.1 Algorithm

In this research the video metrics are developed in OpenCV, a cross-platform library of programming functions mainly aimed at real-time computer vision.

The reason to use an OpenCV implementation is the easy and fast implementation, the fast execution of high level metrics based on optimization for multi-core systems and the advance vision research by providing not only open but also optimized code for basic vision infrastructure.

The algorithm to track and detect lips activity in this environment is explained in Fig. 6. It can be observed that for every frame, the algorithm classifies it into two different groups, e.g. frames in which the lips are moving and frames in which they are not. The block diagram represents the following algorithm:

- From the video file to analyze, the next frame is read. In case it is the first one, two frames have to be read.
- In that frame, a Haar cascade is used for the detection of the mouth region based on OpenCV implementation of Viola and Jones algorithm for face detection. The Viola and Jones object detection framework is the first object detection framework to provide competitive object detection rates in real-time. It was proposed in 2001 by Viola and Jones [12]. Although it can be trained to detect a variety of object classes [1,6], like for the mouth region in this algorithm, it was motivated primarily by the problem of face detection. The mouth region will be our Region Of Interest (ROI).
- In the ROI of the frame, we measure the motion that has appeared between the previous frame and the new one. The algorithm for estimating the amount of motion will be explained in detail in the next figure.

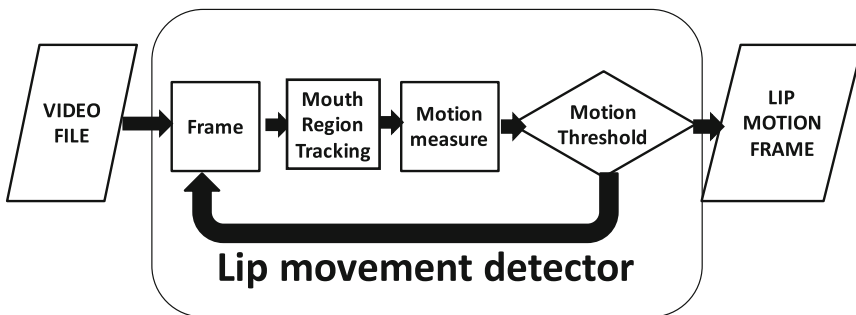


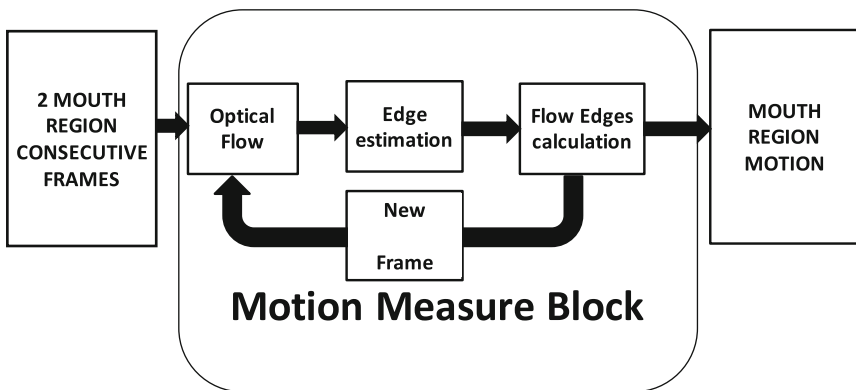
Fig. 6. Algorithm for the detection of lip movement

- A motion threshold will be compared with the calculated motion to determine if the output of the metric is lip-active. This threshold has been optimized for the final output of the metric, which is the audiovisual delay.
- The first of the two frames is released and the last frame read is used to compare with the next one, until we reach the end of the video file.

Figure 6 describes the algorithm in general. However, the key block for the detection of lip movement is the one named “motion measure”. Figure 7 explain in more detail the process carried out to determine the amount of movement between two frames in the mouth ROI. The algorithm is described here:

- The inputs of the block are two consecutive frames in which the mouth region has been located.
- The optical flow between them is calculated. The implementation is based on the algorithm exposed in the related research work carried out by Farneback [3]. Optical flow estimates the quantity and direction of the motion in every corresponding point of the two consecutive frames the algorithm receives
- Once the direction and intensity of motion is estimated, the next step is to discriminate between movement of all the face and movement of the lip region independently. This has been achieved by the calculation of the edges of the optical flow output. This stands for knowing the laplacian of the motion field, and analyzing the borders. If the border is in the mouth ROI, we consider it as an indicator for the independent movement of the lips.
- The last step is to “count” how much edges of the optical flow have been discovered in the mouth region. The number of these edges is strongly correlated with the amount of lip motion in the frame.

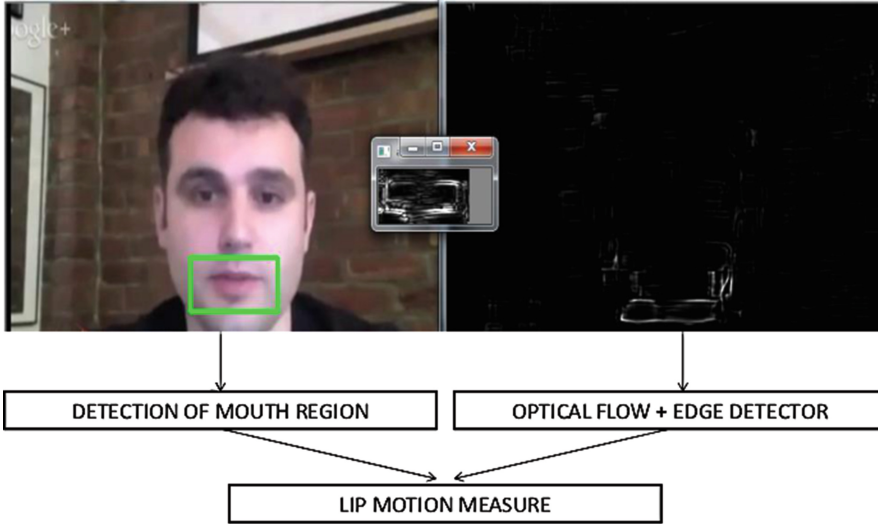
The whole information coming from the OpenCV metric is loaded into MATLAB in order to process it and continue with the comparison with the information coming from the audio part. This means only the video part of the lip sync algorithm is implemented in OpenCV. Future plans include the full implementation of the metrics included in this research into C++ and OpenCV.



**Fig. 7.** Detailed block diagram for motion measure

## 5.2 Results

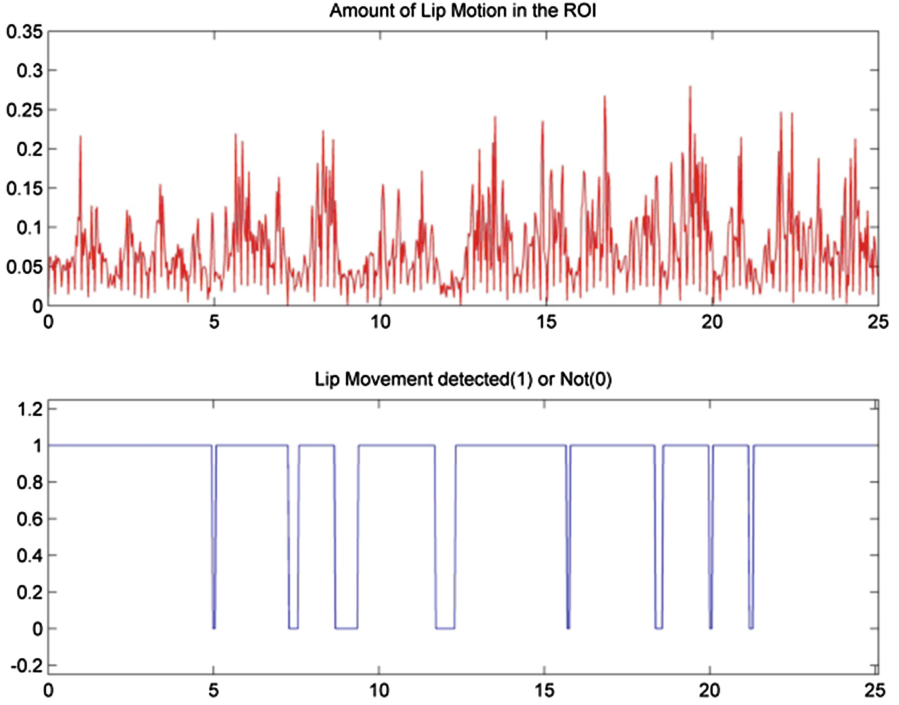
The output of the algorithm for Lip Activity Detection (LAD) should be a binary vector showing the instants in which the video information analysis provides evidence of lip movement. This binary vector should be compared with the binary vector obtained with the VAD algorithm. The comparison will be carried out using the delay calculation algorithm which will be explained in next section.



**Fig. 8.** Graphical output of the LAD algorithm

Being a video metric has the advantage of having the possibility of showing its behavior in an image, something not possible for audio metrics. Figure 8 shows the graphical output for a frame of the LAD metric for MOAVI. It is a frame coming from one of the audiovisual sequence, named “STOSSEL”, that is included in the MOAVI database. All elements presented by OpenCV can be seen in this capture. The green rectangle shows the position of the mouth and defines the ROI of the frame. The optical flow is calculated and the edges of its output are drawn in the black and white square on the right. In the middle of the figure, it can be seen the graphical representation of the output of the metric.

In this results subsection of the LAD, some graphs of the outputs that the metrics described above provide are shown. The typical output of the motion measure block is represented in the upper graph of the Fig. 9. The binary vector determined from that information is shown in the graph situated under it. This binary vector, based on the threshold for the amount of motion, indicates which of the frames are considered active in terms of lip movement.



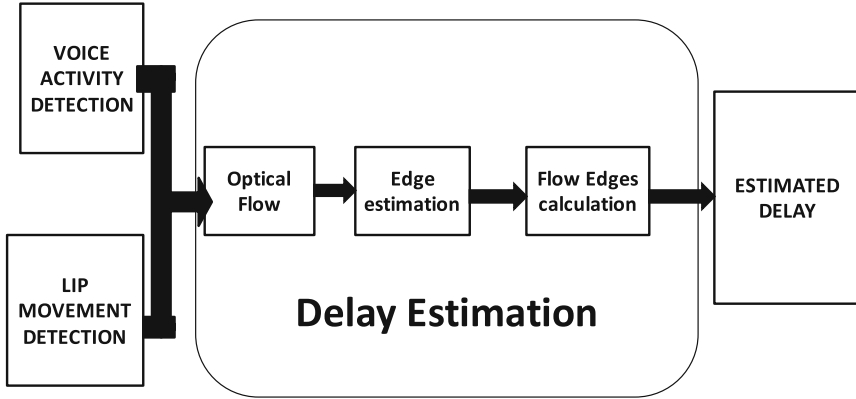
**Fig. 9.** Example of detection of lip activity

## 6 Delay Calculation

The goal of the previous algorithms, VAD and LAD, was to provide a binary vector coming from the audio information and another one from the video information. In a second step, they have to be compared with each other to obtain the delay that one of them has with respect with the other. This section explains the algorithm to carry out this comparison and shows the obtained results.

### 6.1 Algorithm

Some delay estimation algorithms have been implemented in time-domain. For example, the primitive but well-known delay estimation based on cross-correlation method has been tried in this application, without good results. Most advanced time delay estimation algorithms are implemented in frequency-domain; for example the generalized cross-correlation method. The problem that comes out when using frequency domain is the lack of accuracy in the spectral estimation in case of short signal segments. The delay algorithm needed in this synchronization stage will have the goal of estimating the time shift of audio with respect to video, and it must be possible to be used in short audiovisual sequences like the ones stored in the database described above.



**Fig. 10.** Block diagram for delay estimation

For this reason, the estimation algorithm found in [7] is a time-domain implementation that satisfies the needs of this application. The proposed information theoretic delay criterion is used. The basis of the proposed algorithm relies in a time-domain implementation of the maximum likelihood method. Usually numerically motivated convergence criteria are used but in the proposed method, statistically motivated convergence criterion was used instead.

The delay algorithm is outlined in the block diagram (Fig. 10). The implementation has been done in MATLAB. The first input of the delay estimator is the binary vector from the VAD. The second input is the binary vector from the LAD. Both vectors are going to have the same length to compare them and to adjust which of the possible delays makes maximum the likelihood between the two signals. The process followed in the present algorithm is described here:

- First, a covariance matrix is constructed based on the possible delays that are assumed. In this metric, the possible delays were set to  $\pm 2$  s.
- In a second step, the criterion is build up. The goal is to establish a statistically motivated convergence criterion to make the decision.
- Finally the maximum of the criterion is calculated. The estimated delay will be the shift that corresponds to that maximum.

One of the problems of this method is that it's assumed that the audio activity and the video activity are perfectly synchronized. It's supposed that when a person is talking and the lips are visible, the viewer can see the lips moving only when some sound can be heard.

Obviously, this is not an accurate approach. The first example of lack of audiovisual speech correlation can be the *noisy, unvoiced motion* of the lips, such as smiling, or wetting lips, which are impossible to discriminate using this algorithm, although some differences are accepted and still the estimated delay will be accurate. There is luckily another example of the problems that can be easily corrected. That is the lack of complete synchronization between lip

activity and voice activity even when the lip sync artifact has not occurred. It can be observed that lip activity usually starts around 300 ms in average before the voice activity starts to be perceived. This is a stationary delay that can be perfectly corrected just by taking into account this 300 ms in the estimated delay. In the next section, the results show this artificially added gap.

## 6.2 Results

As it can be seen in the other results sections of the lip sync indicator, the accuracy of the previous metrics is quite high. There are some specific situations in which the VAD method is not able to perfectly discriminate between human speech and other sounds, and the same happens in a few situation in the case of LAD method with lip motion for speaking and other kinds of lip motion.

In these circumstances, it is obvious that the two binary vectors used as inputs for the Delay Estimation Algorithm are not going to be active (value = 1) in the same instants, even if no delay is introduced. This is the reason why the goal of detecting Lip Sync artifact is challenging. On the other hand, this is the reason why an advanced delay estimation algorithm is used and the results of estimating the delay with this algorithm are presented in this subsection.

It is important to understand that, being the Delay Estimation Block the last of the stages for the Lip Sync Artifact Key Indicator Determination, the output of this block is going to be the Key Indicator. Therefore, if the estimated delay is over the thresholds that were determined in previous sections (140 ms), the determined Lip Sync Artifact Key Indicator will be active.

Delays of 0, 300, 500 and 800 ms are artificially introduced to analyze the delays that the metric determines. The absolute error is also calculated. An average gap of 154.8 ms is calculated for the 60 estimations carried out for the experiment. Moreover, in 80 % of the test audiovisual sequences, the binary key indicator is correct. Thus, in the 80 % of the times, this key indicator determines correctly not only if the lip sync artifact is present and the threshold is overcome if the audio is delayed with respect to the video or vice versa.

## 7 Limitations and Future Research

As limitations, we can list a few main aspects that would be important to improve as further research.

With respect to VAD, some sounds that should not be detected as speech because they appear without any correlation with video information are actually detected as voice active. Examples of this sounds could be speakers that are not visible in the scene (more and more frequent in today's films), background music with voice are not detectable. Further research will include audio signal processing in terms of speaker recognition to discriminate between different speakers.

With respect to LAD, some *noisy* lip movement that should not be detected as speech because they appear without any correlation with audio information are actually detected as lip active. Examples of this lip movements could

be people smiling or wetting lips, which are impossible to discriminate using this algorithm. Further research will include video signal processing in terms of speaker recognition to discriminate between different people in the scene.

With respect to Delay Estimator, further research will expect to be capable of detecting both senses of delays, not only audio delayed with respect to video.

**Acknowledgments.** The work was co-financed by The Polish National Centre for Research and Development (NCBR), as a part of the EUREKA Project №. C 2012/1-5 MITSU.

## References

1. Baran, R., Glowacz, A., Matiolanski, A.: The efficient real- and non-real-time make and model recognition of cars. *Multimedia Tools Appl.* **74**(12), 4269–4288 (2015). <http://dx.doi.org/10.1007/s11042-013-1545-2>
2. Cerqueira, E., Janowski, L., Leszczuk, M., Papir, Z., Romaniak, P.: Video artifacts assessment for live mobile streaming applications. In: Mauthe, A., Zeadally, S., Cerqueira, E., Curado, M. (eds.) *FMN 2009*. LNCS, vol. 5630, pp. 242–247. Springer, Heidelberg (2009)
3. Farneback, G.: Very high accuracy velocity estimation using orientation tensors, parametric motion and simultaneous segmentation of motion field. In: *Proceedings of the Eighth IEEE International Conference on Computer Vision*, Vancouver, Canada (2001)
4. Glowacz, A., Grega, M., Gwiazda, P., Janowski, L., Leszczuk, M., Romaniak, P., Romano, S.: Automated qualitative assessment of multi-modal distortions in digital images based on glz. *Ann. Telecommun. - Ann. Telecommun.* **65**(12), 3–17 (2010). <http://dx.doi.org/10.1007/s12243-009-0146-6>
5. Kacprzak, S., Ziółko, M.: Speech/music discrimination via energy density analysis. In: Dediu, A.-H., Martín-Vide, C., Mitkov, R., Truthe, B. (eds.) *SLSP 2013*. LNCS, vol. 7978, pp. 135–142. Springer, Heidelberg (2013)
6. Leszczuk, M., Baran, R., Skoczylas, L., Rychlik, M., Ślusarczyk, P.: Public transport vehicle detection based on visual information. In: Dziech, A., Czyżewski, A. (eds.) *MCSS 2014*. CCIS, vol. 429, pp. 16–28. Springer, Heidelberg (2014)
7. Moddemeijer, R.: On the convergence of the iterative solution of the likelihood equations. In: Schouwhamer Immink, K.A. (ed.) *Ninth Symposium on Information Theory in the Benelux*, May 26–27, 1988, Mierlo (NL), pp. 121–128. *Werkge-meenschap Informatie- en Communicatietheorie*, Enschede (NL) (1999). ISBN: 90-71048-04-7
8. Sohn, J., Sung, W.: A voice activity detector employing soft decision based noise spectrum adaptation. In: *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 1, pp. 365–368, May 1998
9. Staelens, N., De Meulenaere, J., Bleumers, L., Van Wallendael, G., De Cock, J., Geeraert, K., Vercammen, N., Van den Broeck, W., Vermeulen, B., Van de Walle, R., Demeester, P.: Assessing the importance of audio/video synchronization for simultaneous translation of video sequences. *Multimedia Syst.* **18**(6), 445–457 (2012). <http://dx.doi.org/10.1007/s00530-012-0262-4>

10. Steinmetz, R.: Human perception of jitter and media synchronization. *IEEE J. Sel. Areas Commun.* **14**(1), 61–72 (1996)
11. Venkatesh, R., Bopardikar, A.S., Perkis, A., Hillestad, O.I.: No-reference metrics for video streaming applications. In: *Proceedings of the 14th International Packet Video Workshop (PV 2004)*, Irvine, CA, USA, 13–14 December 2004
12. Viola, P., Jones, M.: Robust real-time object detection. *Int. J. Comput. Vis.* **57**(2), 137–154 (2001)



Multimedia Communications, Services and Security  
8th International Conference, MCSS 2015, Kraków,  
Poland, November 24, 2015. Proceedings  
Dziech, A.; Leszczuk, M.; Baran, R. (Eds.)  
2015, XII, 211 p. 91 illus. in color., Softcover  
ISBN: 978-3-319-26403-5