

# Monitoring of audio visual quality by key indicators

## Detection of selected audio and audiovisual artefacts

Ignacio Blanco Fernández<sup>1</sup> · Mikołaj Leszczuk<sup>2</sup>

Received: 29 July 2016 / Revised: 25 November 2016 / Accepted: 30 December 2016 /  
Published online: 15 February 2017  
© The Author(s) 2017. This article is published with open access at Springerlink.com

**Abstract** Over 10 billion hours of video are watched online every month. Together with high definition television broadcasting and the rise in high quality video on demand, this makes quality assessment a key task in the global multimedia market. Automating quality checking is currently based on finding major audiovisual artefacts. 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 on the development of the audiovisual part of this project, which includes the detection of muting, clipping and lip synchronization (also known as lip sync) artefacts.

**Keywords** MOAVI · VQEG · Mute · Clipping · Lip sync

## 1 Introduction

Automating quality checking is currently based on finding major video and audio artefacts. The Monitoring Of Audio Visual quality by key Indicators (MOAVI) subgroup of

---

✉ Mikołaj Leszczuk  
leszczuk@agh.edu.pl

Ignacio Blanco Fernández  
gncblncfrndz@gmail.com

<sup>1</sup> Polytechnic School of Engineering of Gijón, Plaza Campus Universitario 92A, 33394 Asturias, Spain

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

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.

Existing NR QoE models, such as those reported in related research work [7, 26], follow the less useful Full-Reference (FR) models (e.g. [8]), which measure the quality of networked multimedia using objective parametric models. These models have slight problems in predicting the overall audiovisual QoE. MOAVI can be used 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), muting (signal losses), and clipping [2]; the list is not complete although it does include the major artefacts) describing service quality in general, and to select subsets for each potential application. Therefore, the MOAVI project concentrates on models based on key indicators, unlike models predicting overall quality.

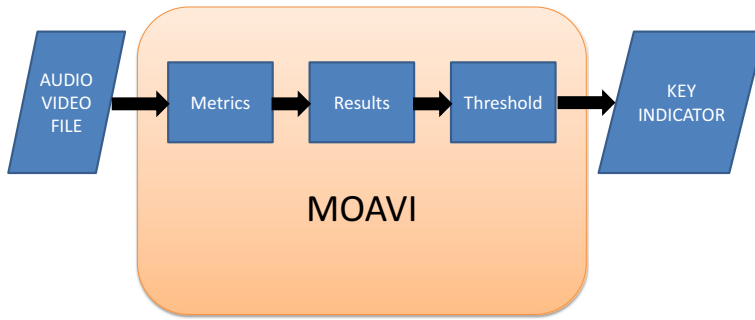
The MOAVI project focuses on indicators which are yet to be addressed by other VQEG projects. Audio quality of low bit-rate signal may be poor due to artefacts such as compression artefacts in signal coding/transmission/encoding, limited sampling rate, limited dynamics, etc.; however, these aspects have already been studied and evaluated in numerous previous VQEG works. Artefacts which are yet to be addressed are muting, clipping and lip sync. While clipping and muting detection algorithms are rather rudimentary, the main contribution of this paper is measuring the lip sync artefact.

The classic quality metric approach cannot provide pertinent predictive scores with a quantitative description of specific (new) audiovisual artefacts, such as stripe error or exposure distortions. MOAVI is an interesting approach because it can detect artefacts present in videos, as well as predicting the quality as described by consumers. In realistic situations, when video quality decreases in audiovisual services, customers can call a helpline to describe the problem and visibility of the defects or degradations in order to describe the outage. In general, they are not required to provide a Mean Opinion Score (MOS). As such, the concept used in MOAVI is completely in phase with user experience. There are many reasons for video disturbance, and they can arise at any point along the video chain transmission (filming stage to end-user stage) [13].

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 MOS is provided. A binary indicator for each artefact is provided instead showing its presence or absence.

Figure 1 shows the concept of MOAVI. The audio or video stream (video only for video artefacts, audio only for audio artefacts, and both together for audiovisual artefacts) is the input to the system. The metric of each artefact is used to determine the level of impairment of the media to be analysed. These results are converted into binary indicators using a threshold which determines whether the artefact is noticeable in the video. This way MOAVI obtains a key indicator for each artefact.

This paper is organized as follows. Section 2 describes the measurements of key audio indicators – presence of muting and clipping. Section 3 describes the measurements of key audiovisual indicator – presence of lip sync, including the video database for the assessment of the metrics, the algorithms and the results obtained. Section 4 concludes the paper and summarizes the results.



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

## 2 Measuring mute and clipping artefacts

In recent years, interest has been growing in real-time audio services over packet networks. For quality evaluation, it is essential to quantify user perception of the audio sequence. Signal loss is one of the most common degradations in audio streaming at low bit rates. The end-user perceives a silence followed by an abrupt clipping. Cell loss in packet networks or a restitution strategy could be the origin of this perceived temporal audio discontinuity.

It is important to detect and prevent or correct the clipping problem caused by digital capture, conversion and downscaling processing. The audio signal is always stored digitally in order to improve the quality of audio. In certain situations, the original audio signal may be clipped during the recording due to the impact of environmental noise or recording equipment. This means clipping can originate at the capture stage. The maximum amplitude of the clipped signal is frequently limited to a constant. This clipping distortion leads to a harsh noise. It significantly affects the subjective listening quality if the clipping intensity is strong or the clipping density is high.

Muting and clipping are the most frequent impairments present in audio streaming and audio files in general. Therefore, a key indicator for each of these artefacts needs to be added to the most suitable subset of metrics when audio is present in the file/stream being evaluated. These indicators are based on metrics developed for this project and a threshold optimized with preliminary tests carried out during the implementation and improvement phases of the development.

### 2.1 Mute

The advent of protocols for quasi real-time communications and rapidly increasing computing power are driving an increasing interest in real-time audio services over packet networks. Audio streaming is used in real-time applications since the data needs to be transmitted as soon as it is generated in order to deliver continuous media play out. These applications can only tolerate a short delay in signal restitution. However, packets of data are transmitted over unreliable, lossy networks.

Packet loss produces significant temporal impairments in the received audio. When considering quality, it is essential to quantify user perception of the played-out audio sequence. Muting caused by signal losses is one of the most common degradations in audio streaming at low bit rates. The end-user perceives a silence followed by an abrupt clipping. Cell loss in packet networks or a restitution strategy could be the origin of this perceived temporal

audio discontinuity. Packet loss or jitter could cause a sporadic or non-uniform signal loss at the decoding level because of the play-out buffer time limit.

The muting artefact presents as an absence of any kind of sound during a period of time detectable by the human ear. A typical waveform of a muted sound file is shown in Fig. 2 It is usually generated during the transmission stage where the majority of losses occur. This is why this detector should be applied near the far-end to check the correct transmission of the audio file.

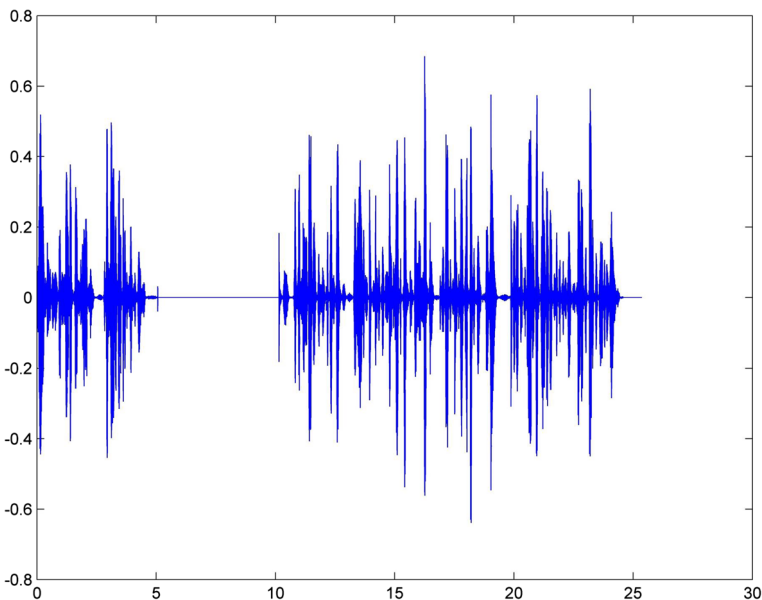
Some approaches to muting detection have been already proposed, usually in the context of automatic audio classification and segmentation. A notable example of such investigation is presented in the paper by Lu and Hankinson [14], where the concept of silence ratio has been introduced, being variation of zero-crossing rate.

### 2.1.1 Algorithm

The algorithm for the detection of the muting artefact involves the establishment of a certain threshold or set of thresholds to determine whether the audio samples analysed are suffering from sporadic audio signal loss. This way, the related research work [16] describes how different lengths, contents and local activity levels affect the quality perceived. It should be noted that the goal of the MOAVI project is to develop a set of metrics that will work without analysing the content.

Two thresholds are needed to determine whether the muting artefact is present in an audio stream: one for the duration of the silence and the other for the amplitude of local activity, which describes the greatest amplitude of the audio wave for it to be considered silenced.

As the metrics for the MOAVI project are NR, we cannot compare the file with the original. An NR audio metric explores the audio file at the sample level in order to detect and measure the distortions which may have been generated.



**Fig. 2** Example of sound waveform with mute artefact

When the characteristics of the artefact are known, the detection algorithm is simple. Figure 3 shows a schematic view of the process which determines whether the mute artefact is present in an audio file. Each sample is compared to the amplitude threshold. If its value is lower, we check whether the number of successive low amplitude samples is sufficient to become noticeable. If the silence is sufficiently long, the key indicator for the muting artefact is positive, indicating the presence of the artefact in the analysed sound.

The paper [16] provides experiments for setting the duration threshold. It has been shown that, for most types of content, a signal loss of 10ms is detectable (with the exception of news or speech-based content). An unequivocal detection, close to the probability of 1, is attained for a discontinuity of 30ms. This result is also valid for preliminary tests carried out in this research. Thus, the duration threshold is 30 milliseconds or its equivalent in samples.

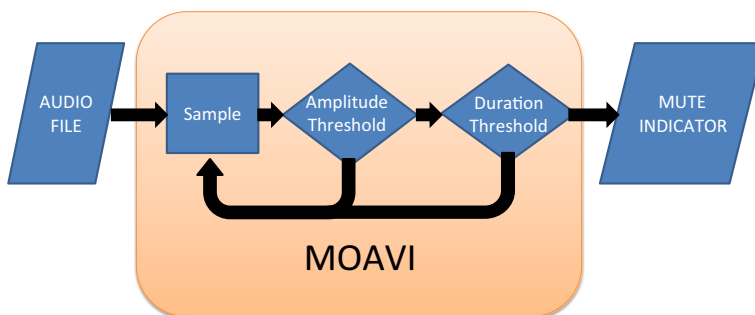
Regarding the amplitude, the threshold for the minimum amplitude in the digital signal detectable by a listener depends on the player configuration characteristics such as volume settings or distance between the listener and the speaker. However, if muting is considered to be an artefact which occurs when the signal is completely lost, the amplitude threshold has to be the minimum amplitude in absolute value different from zero that the codification can admit. Therefore, the assumption made here is that muting is only present when the audio signal is a sequence of zeros or the complete absence of audio signal.

A sound file can carry information for two channels. In fact since the majority of streaming, broadcasts and music are produced, transmitted and displayed using stereo digital equipment, it is common for the mute detection algorithm to analyse and synchronize both channels. Therefore, the solution is simple: since the human ear can only declare as mute a file with both channels silenced, the logical operation to be introduced between the two channels is the ‘AND’ operation. This means that the key indicator is active only if both of the channels are detected as mute.

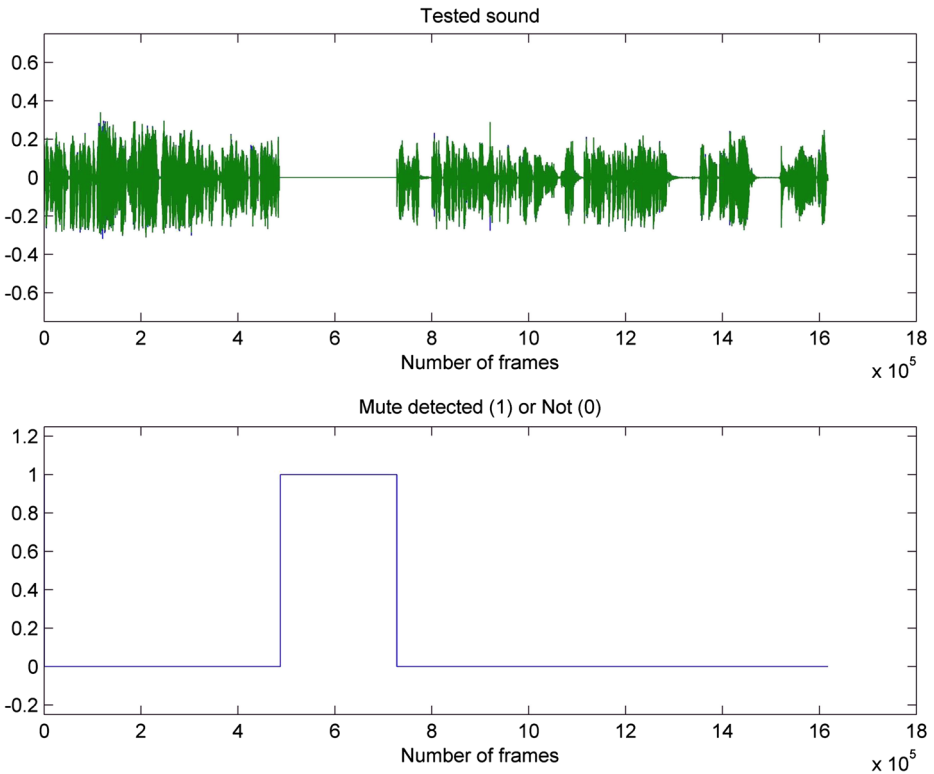
As the metric takes into account every sample it is extremely accurate while indicating the start and end of the muted subsequence, which can be helpful in the detection of the data packet which has been lost. This data packet could even be requested to be sent again from the production/distribution centre, which solves the mute artefact problem in this scenario.

### 2.1.2 Results

Regarding the results obtained, the detection of the mute artefact in a simulated sequence impaired by a signal loss is shown in Figs. 4 and 5. In the first figure, the silence was artificially introduced between samples number  $5 \cdot 10^5$  and  $7 \cdot 10^5$  approximately.



**Fig. 3** Algorithm for the detection of mute artefact



**Fig. 4** Example of detection of the mute artefact in an audio sequence 1

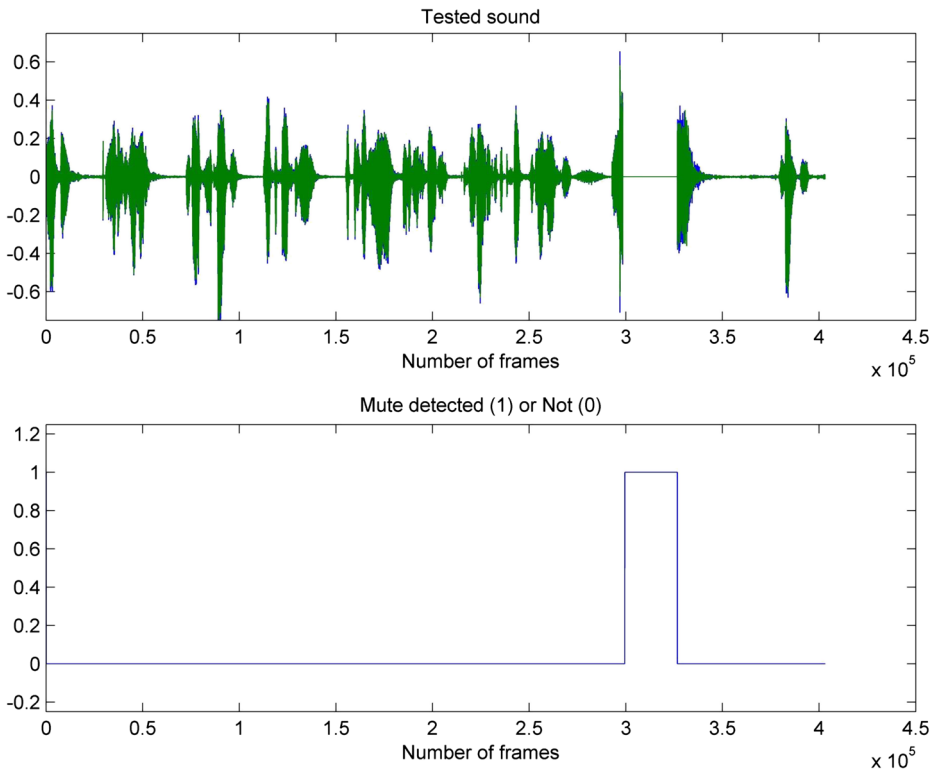
In the second figure, the silence was introduced between samples number  $3 \cdot 10^5$  and  $3.2 \cdot 10^5$  approximately. In both the example sequences it can be observed that the algorithm works accurately and it detects the artefact at the time positions when it was introduced. Additionally, the metric detects muting discriminates the silent moments during speech (pauses when only background noise is heard) from artificial silence, or loss of the audio signal which is actually the mute impairment that the metric was developed to detect.

Experiments were conducted to evaluate the accuracy of this detector. The set of ten audio files used as an input for the experiments was similar to that shown in Fig. 4 in the sense that an artificial mute artefact was introduced to them. In this regard, the mute artefact was present in the input audio files as silent samples of different lengths.

An accuracy rate of 95 % was found for this metric under these conditions. Most of the samples that were erroneously marked as "non-muted" (false negatives) were the first muted samples which the detector encountered from the muted section.

One of the limitations of this algorithm are the potential false negatives when a signal bias (i.e. DC offset) is introduced in the audio wave. Under these circumstances, a muted signal does not imply small values of samples and thus it would not be detected.

Although psychoacoustic experiments are not the object of this research, we use the available publications to determine the optimal thresholds for the minimum duration of the silence and the minimum noticeable amplitude of the waveform [16].



**Fig. 5** Example of detection of the mute artefact in an audio sequence 2

## 2.2 Clipping

As noted in [4, 29] on a restoration method of clipped audio signals based on MDCT, the audio signal is always stored digitally in order to improve audio quality. In certain situations, the original audio signal may be clipped during the recording due to environmental noise or recording equipment. The maximum amplitude of the clipped signal is often limited to a constant. This clipping distortion leads to a harsh noise. It significantly affects subjective listening quality if the clipping intensity or density is high.

Clipping can be divided into two classes: digital clipping and analogue clipping. For digital clipping, when the signal amplitude exceeds the upper limit of the recording equipment during the transcription, the signal amplitude will be a constant in the peak region. In analogue recording systems, the signal can be clipped by impedance mismatch or the overflow of the input electrical level. Analogue clipping shows a small deviation in amplitude, and the sample values in the clipped region are not exactly equal to each other. In both digital and analogue clipping, the front-end of the clipped signal is always in the peak regions.

While analogue and digital input clipping can occur in the observed streams, they need to be distinguished. Although input analogue signals can be over-amplified, in fact artificial amplification is not common in real equipment. On the other hand, digital over-amplification is introduced when certain parts of the digital processing chain are not connected

correctly – digital signal is equalized without signal compression/limitation – by the digital compressor/limiter algorithm.

A typical waveform of a clipped signal tends to be similar to the one showed in Fig. 6. The waveform in the clipped areas is a constant or semi constant value, which is usually the highest value that the amplitude of the audio signal can have.

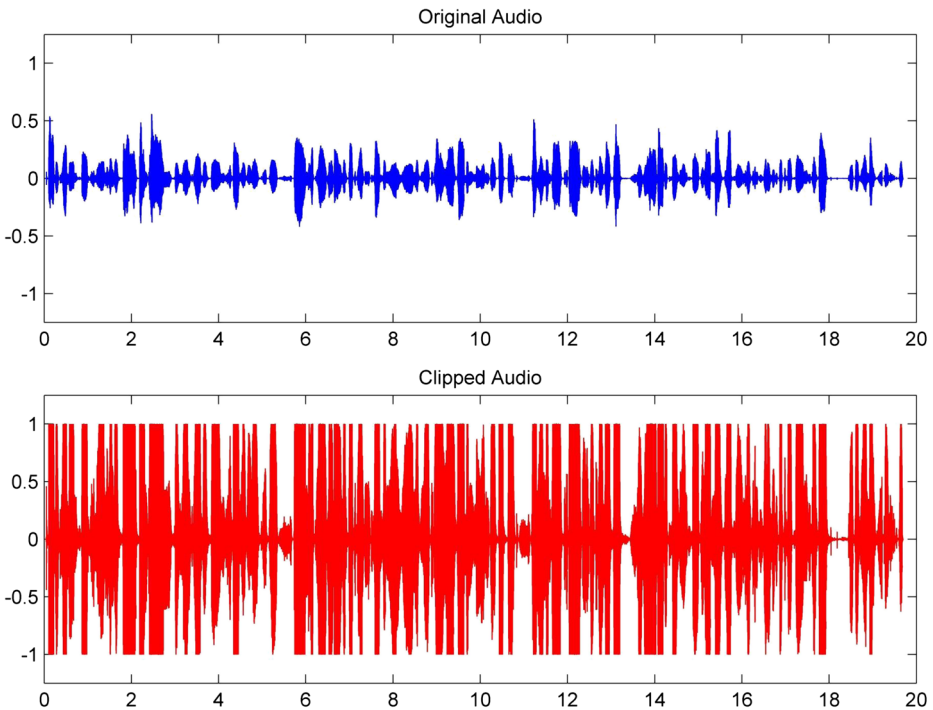
There is also another type of clipping in which the artefact is produced during the stage before the audio signal level is reduced or converted. In this case, the constant or semi constant amplitude can be any value. In this type of clipping, none of the signal samples are higher than the constant. Thus, the waveform appears to be cut off at the mid value.

Whereas clip detection has been already investigated for a quite long time, most of the proposed solutions (like the one by Person and Muccioli [17]) was related to analogue signals. Nevertheless, recently, solutions for digital signals (like the one by Skoglund and Linden [19]) started to emerge as well.

The following section explains the algorithm we used to detect of clipping (both types).

### 2.2.1 Algorithm

The algorithm for the detection of the clipping artefact involves setting a certain threshold or set of thresholds to determine whether each of the analysed audio samples is limited to a constant amplitude. This method has been used to study how different lengths and contents affect the perceived quality. As the goal of the MOAVI project is to develop a set of metrics that work without analysing the content, this is not taken into account in the clipping metric.



**Fig. 6** Example of waveform of an audio signal suffering clipping



This means that two thresholds are needed to determine whether the clipping artefact is present in an audio stream: one for the number of samples following each other restricted to a constant, and one for the maximum variation of the amplitude value in two consecutive samples to be considered *constant*; this represents the amplitude gap between two consecutive audio samples which are candidates to be clipped.

As the metrics for the MOAVI project are NR, we cannot compare the file with the original. An NR audio metric explores the audio file at the sample level in order to detect and measure the distortions which may have been generated, so the NR clipping metric cannot compare the analysed signal with the original.

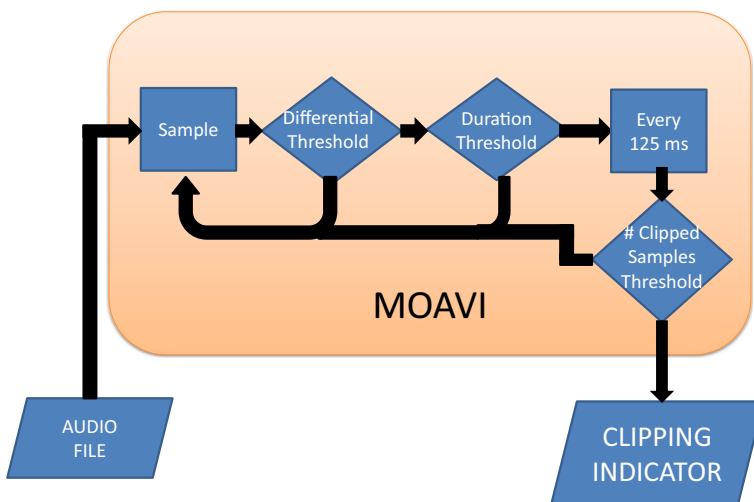
Figure 7 shows a schematic view of the process used to determine whether the clipping artefact is present in an audio file. Each sample is compared to the previous sample to determine whether the gap between their amplitudes is greater than the differential threshold. If the gap is lower, and thus two or more samples have a very similar amplitude, we check whether the number of consecutive low-amplitude samples is sufficient to be noticeable by a human listener as clipping (harsh noise).

If the length of the constant or semi-constant values is sufficient, the sample becomes a candidate to be clipped. Every 125 milliseconds, the number of candidate samples is compared with the total number of samples analysed in those 125 milliseconds. Therefore, the key indicator for the clipping artefact is positive if this ratio is higher than 30 percent. If the key indicator is positive, it indicates the presence of the clipping artefact in the analysed sound.

The percentage of candidate samples to be clipped (30 percent) and the length of the audio sub-sequence (125 milliseconds) over which the clipped/not clipped decision is made is based on preliminary tests, which show that the best behaviour occurs when applying the pertinent threshold to this length of sequence.

### 2.2.2 Results

Regarding the results, the algorithm detecting this artefact is simulated over a sequence impaired by generated clipping. This process involves two steps. In the first step, the audio



**Fig. 7** Block diagram describing the algorithm for the detection of the clipping artefact

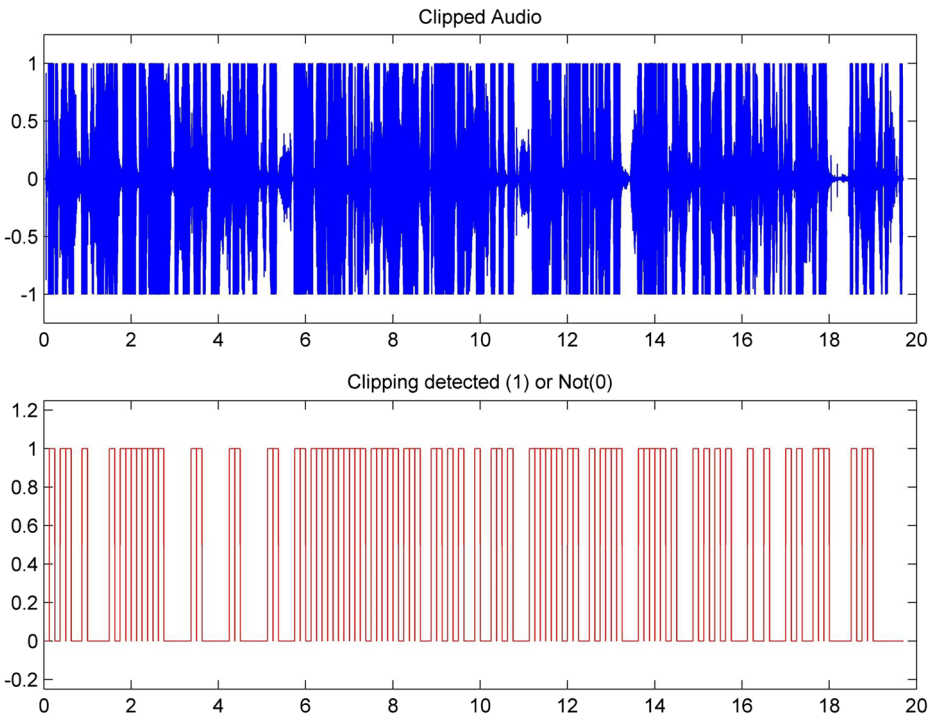
signal is amplified until some of its samples reach the top amplitude (over-amplification). In the second step, the amplitudes are cut above the maximum value which can be reached by a sound file with a given bit depth. This generates a waveform similar to an audio signal affected by the impairment *naturally*, during the capture or processing stage (see Fig. 6).

Two examples of clipping being detected are presented in Figs. 8 and 9. In both figures, clipping was artificially introduced over the entire file, since in most cases the clipping artefact affects the entire file.

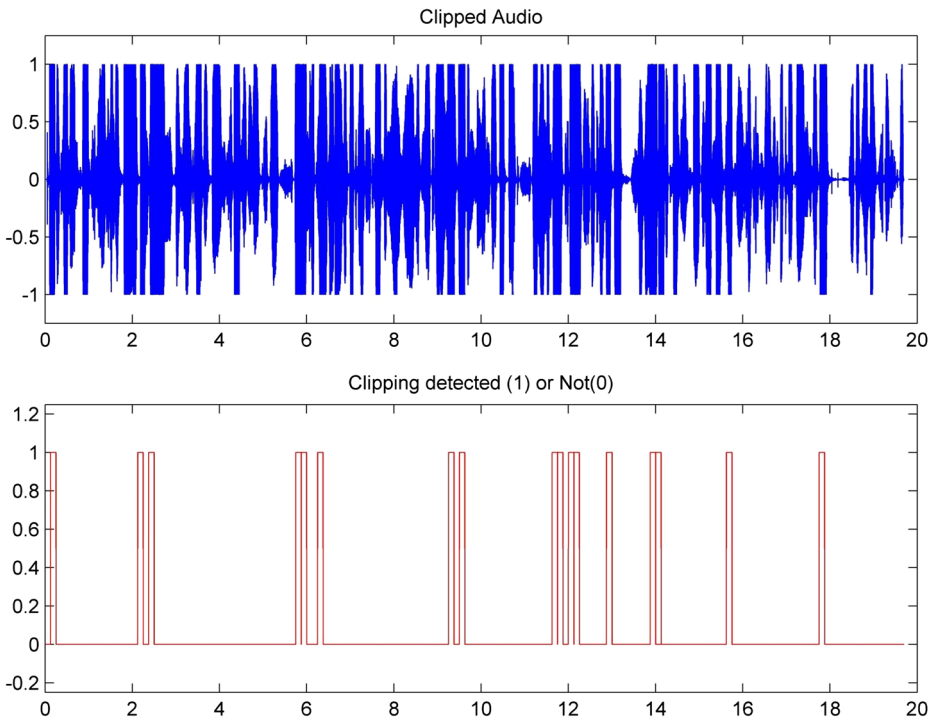
In Fig. 8 the amplification is 24 dB. This makes the clipping more noticeable and the signal cuts are greater. This produces a harsh noise when the sound is played, becoming more noticeable as the cuts become greater. In Fig. 9, the amplification is 15 dB. This means that the number of sub-sequences detected as clipped is lower; however, the indicator remains positive since the artefact is detected.

This shows that the detection occurs in the instants when the waveform is cut or limited by a constant, which corresponds to the instants when the sound is impaired when the file is played. Thus, the MOAVI indicator for clipping increases when clipping appears in the entire file, although the metric is able to determine accurately which samples are clipped in case this information is needed.

We conducted experiments to evaluate the accuracy of this detector using a set of ten audio files. The files were similar to the file shown in Fig. 8 in that they included an artificial clipping artefact. The clipping artefact was present in the input audio files as a set of



**Fig. 8** Example of detection of the clipping artefact in an audio sequence



**Fig. 9** Example of detection of the clipping artefact in an audio sequence

samples with the maximum possible amplitude. Different values and lengths were used for this evaluation.

An accuracy rate of 90 % was found for this metric under these conditions. Most of the samples erroneously marked as “non-clipped” (false negatives) were the first clipped samples found by the detector.

Although psychoacoustic experiments are not the object of this research, we use the available publications [29] to determine the optimal thresholds for the minimum duration of the silence and the minimum noticeable amplitude of the waveform.

### 2.3 Limitations and further research

There are three main limitations to further research:

- The results could be enhanced by applying adaptive thresholds depending on the content.
- Being a NR metric, it is impossible to discriminate a silence introduced by the loss of a sound file packet and a normal silence which would not be an artefact. Therefore, the false alarm ratio can be high and content-dependent.
- Being a NR metric, it is impossible to discriminate a clipping introduced while the file undergoing capturing, processing, transmitting and displaying from deliberately-introduced clipping which would not be an artefact. However, deliberately-introduced clipping is less frequent than in the case of silence, and it is not significant.

### 3 Measuring the lip sync artefact

This paper examines the process of detection of audiovisual artefacts. We describe the algorithm, implementation and results of three different metrics developed to indicate the presence or absence of the lip sync artefact, which is the most common problem affecting audiovisual signals.

Lip syncing is a key parameter in interactive communication. In video conferencing, streaming and television broadcasting, the uneven delay between audio and video should remain below certain thresholds, recommended by several standardization bodies. However, research shows that the thresholds can be relaxed, depending on the targeted application and use case [21].

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). For audio streams, LDUs are individual samples or blocks of samples transferred together from a source to one or more sinks. Similarly with video, one LDU typically corresponds to a single video frame, and consecutive LDUs to a series of frames. They have to be presented at the sink with the same temporal relationship as they are captured, giving an 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 satisfactory.

In general, inter-stream synchronization involves relationships between many types 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.

Until recently, lip syncing was impossible to detect automatically by state-of-the-art solutions. This is due to the difficulty in obtaining the correct algorithm (technique) to detect this artefact and the high cost of equipment required for processing video and audio. Additionally, analysis of literature and patents covering the lip sync detection problem shows that several solutions use this formulation [3, 9, 11, 22, 24, 28]; however, none of them are innovative scalable solutions and offer potential commercial applications, unlike the results of the research presented in this paper. The majority of existing solutions (including that patented by LG Electronics [9, 11]) attempt to circumvent the difficulties in detecting this artefact by introducing external timestamps to audio and video signals. Another approach represents a solution known as QuMax2000 (patented by the KWILL Corporation) [24]; this requires no external marks, but instead it requires simultaneous access to audiovisual streams with and without the lip sync artefact, which makes the solution unsuitable for non-laboratory conditions. Similarly, LipTracker (patented by the Pixel Instruments Corporation) [3] is not a suitable solution. While the general concept of detecting the lip sync artefact carries certain similarities with the the solution proposed in this paper, an analysis of the patent indicates the existence of significant algorithmic differences. In addition, it should be noted that LipTracker, originally developed in 2005, is simply a closed-mounted rack 19" laboratory solution for analysing analogue signals and the detection of the lip sync artefact in limited cases, such as television news programmes or talk shows [22].

Recently, some more related approaches to developing methods for bi-modal (audio-video) lip speech detection have been proposed, for example in the paper by Czyżewski et al. [5]. These methods can be potentially combined with the method proposed in this paper, in order to achieve higher accuracy.

Some more facts about the lip sync problem:

- The most common origin for the lip sync artefact is jitter produced in the transmission stage.
- Different languages make no significant difference in synchronizing media.
- Different languages make no significant difference in the detection of the lip sync artefact, both for human perception and for automatic detection.
- In [23] it is also stated that professional video editors and TV-related technical personnel show a lower level of skew tolerance. When they detect an error, they are able to correctly state whether audio is ahead of or behind video.
- Watermarks or fingerprints embedded in an audio signal are used in broadcasts to avoid this problem. However, this method is not suitable for on-line multimedia streaming.

Regarding detection thresholds, [21] describes the high number of thresholds determined by the authors. Some authors and research groups have concluded that audio may be played up to 305 ms ahead of video and conversely video can be displayed up to 190 ms ahead of the audio. Both temporal skews are noticed, but they can be accepted by the user without any significant loss of effect. However, some authors report a tolerance of only 4-16 ms.

Figure 10 is a graphical representation of different audio/video delay and lip sync thresholds of detectability as identified by several standard bodies and independent studies. The thresholds used for the lip sync artefact in MOAVI are set to 100 ms when the audio is delayed with respect to video and 140 ms when video is delayed versus the audio. These thresholds are based on research work by Steinmetz on human perception of jitter and media synchronization, referred to here [23].

### 3.1 Video database for the assessment of metrics

The development of experiments analysing the behaviour and measuring the accuracy of different metrics in this section requires a small database of videos and key information about them. It is a set of 15 video sequences between 13 and 37 seconds longs, originating from various types of media. The videos are all taken from a forward-facing camera, although some include several frames with a profile view. Usually only the face and the shoulders are visible. Only one person is seen and heard in each video.

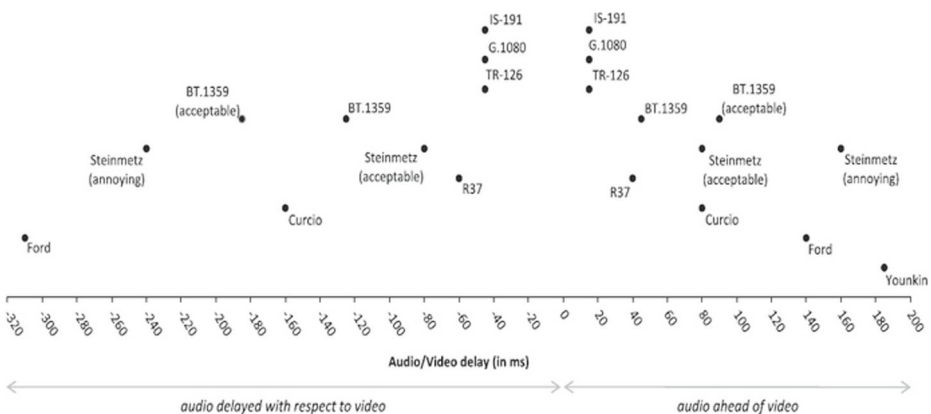


Fig. 10 Different audio/video delay and lip sync thresholds of detectability

Some of the videos originate from TV news shows or interviews; a few are videos uploaded directly to the internet.

The most important characteristics of each video are shown in Table 1. The audio files extracted from the videos have been stored and analysed, so they can be used for tests of Voice Activity Detection (VAD).

The MOAVI indicator for lip sync is based on the lip sync metric explained below. The audio part of the metric is described first, followed by the signal processing used to implement a VAD algorithm. The video part of the metric described in the second section, explaining the combination of techniques used to detect the lip movement. In the third and final section, the algorithm comparing the audio and visual information is described. Each section includes a results subsection and a further research subsection describing the method developed to detect the delay between the visual and audio and audio media.

### 3.2 Voice activity detector

VAD, also known as speech activity detection or speech detection, is a technique used in speech processing in which the presence or absence of human speech is detected [18]. The main applications of VAD are in speech coding, speech recognition and speech searching [25].

Developing an indicator analysing whether audio and video are synchronized is a challenging goal. The process is simplified if the task is divided into smaller parts, therefore the first algorithm to develop is a voice activity detector.

#### 3.2.1 Algorithm

In lip syncing, it is necessary to process the signal in utterances including speech, silence and background noise. The detection of speech embedded in various types of non-speech events and background noise is known as endpoint detection, voice detection, or VAD.

**Table 1** Characteristics of the video database for the assessment of 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 VAD algorithm includes two steps. The algorithm for the detection of voice is represented in Fig. 11. The two detectors are used together to obtain better results.

The first step is signal processing leading to the detection of the endpoints of voice in the audio. An algorithm based on [20] was developed in MATLAB.

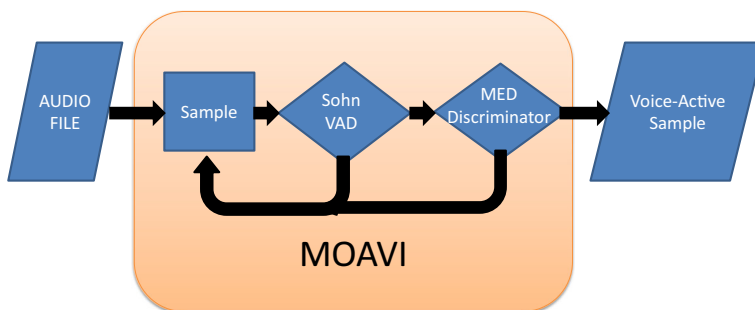
The second step is the analysis of the Minimum Energy Density (MED) feature which is a key distinction between music and similar waveforms and speech waveforms. The algorithm is described in [10]; the MATLAB code was completed based on this algorithm.

In [20], a VAD for variable rate speech coding is decomposed into two parts - a decision rule and a background noise statistic estimator - which are analysed 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 a speech signal, a noise spectrum adaptation algorithm using soft decision information of the proposed decision rule was developed. The algorithm is robust, especially for time-varying noise.

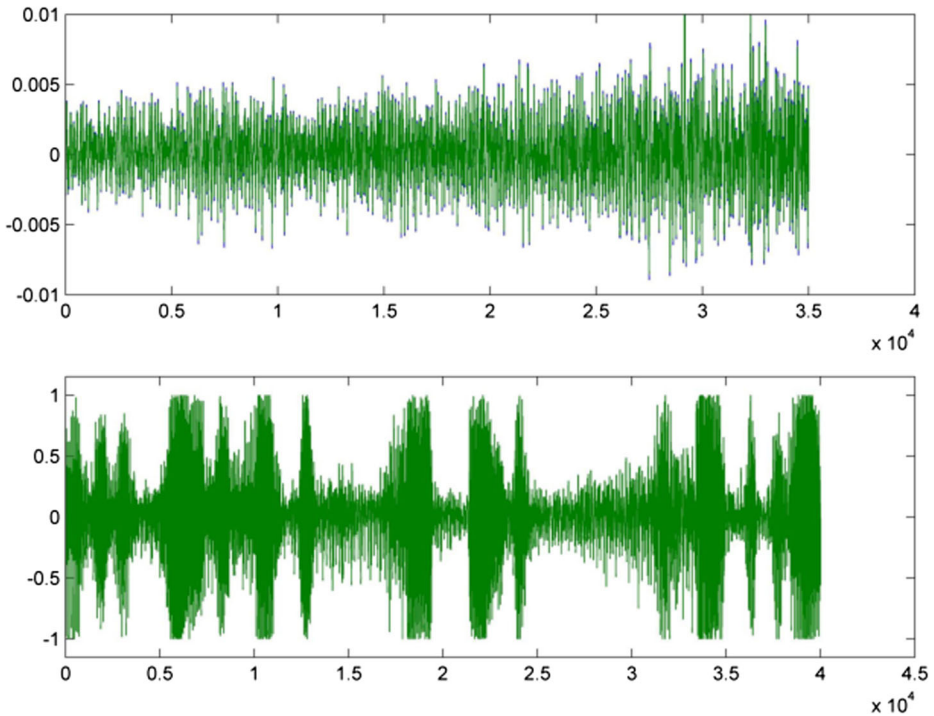
In [10], MED is used to discriminate between speech and music audio signals. This method is based on the analysis of local energy for local sub-sequences of audio signals. The sub-sequences in this method will be those in which voice activity has been detected in the first detector. An elementary analysis of the probability density for the power distribution in these sub-sequences is an effective tool supporting the decision-making. Distinguishing between speech and music is intuitive, based on shape of the signal's energy envelope. As Fig. 12 shows, speech signals have distinctive high and low amplitude parts, which represent voiced and unvoiced speech, respectively. In turn, the music signal envelope is more steady. Moreover, it is known that speech has a distinctive 4 Hz energy modulation, which matches the syllabic rate.

Considering these characteristics, a decision is made to discriminate between speech and music sub-sequences using the probability density function of short timeframe energy inside a time window known as the normalization window. The window has to be long enough to capture the nature of the signal. This value is 200 ms, when the sub-sequence of speech after the first discriminator is longer than this value.

As explained above, these two algorithms work together to make the resulting combination more robust and to improve the accuracy of the metric in order to provide better information to be compared with information coming from video; this provides a lip sync artefact indicator.



**Fig. 11** Algorithm for the detection of the speech instants artefact



**Fig. 12** Comparison between a music waveform (*up*) and a speech waveform (*down*)

### 3.2.2 Results

Regarding the results of the VAD developed for MOAVI, the output of the metric resembles the one presented in Fig. 13. The metric provides an accurate classification of samples. Every subsequence of 50 ms is classified into two different values: voiced (1) or unvoiced (0). Thus, a binary vector is constructed to be compared with information originating from the video concerning endpoints of speech. The final goal is calculating the delay between the signals. The binary vector originating from the VAD metric described above is stored.

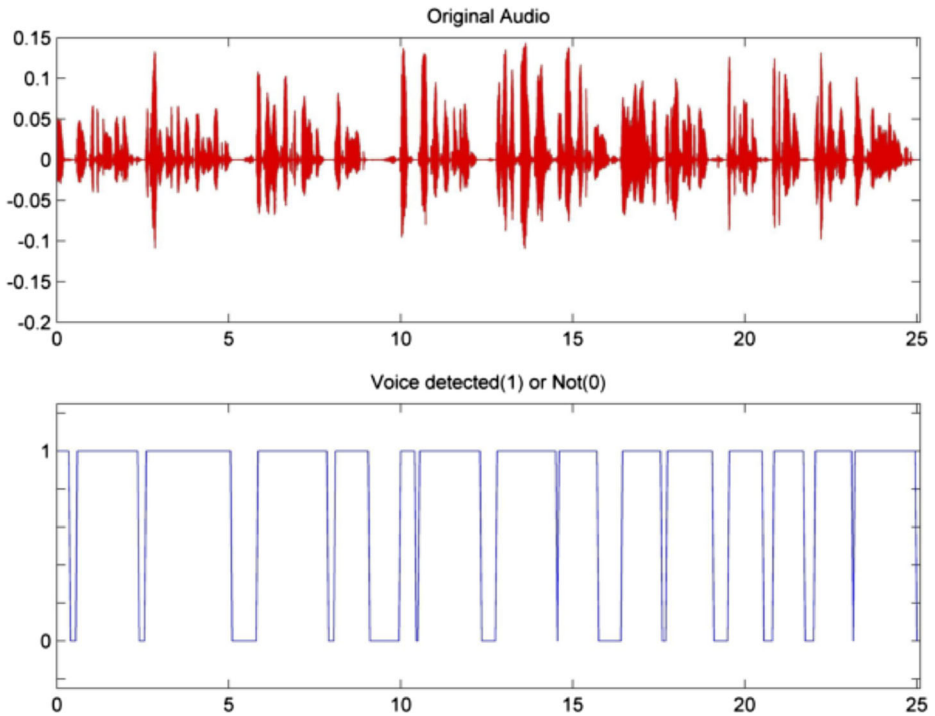
These results were compared with the ground truth prepared by listening to the 15 audio files and developing a small database for each sound in which every instant is classified between voiced or unvoiced with a precision of 50 ms. The selected audios were selected based on two characteristics: they mainly comprised human voice, and they featured different environments/sources, such as old radio, recent interviews or noisy conferences.

Table 2 shows the Hamming distance, precision, accuracy and the F1 metric for each of the video files stored.

Table 3 shows the same parameters describing the performance of the metric as Table 2, although this time the data shows the results for all the videos together. In this regard, the total Hamming distance column shows the sum of all the Hamming distances calculated for each audio file, and the precision, accuracy and F1 metric are the mean of the corresponding statistical indicator for each audio file.

It should be noted 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.





**Fig. 13** Example of voice detection

### 3.3 Lip activity detector

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

#### 3.3.1 Algorithm

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

OpenCV is fast and easy to use; it provides fast execution of high level metrics based on the optimization of multi-core systems and advance research by providing open and optimized code for basic vision infrastructure.

The algorithm tracking and detecting lip activity in this environment is explained in Fig. 14. The algorithm classifies each frame 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:

- The next frame is read in the video being analysed. If it is the first frame, two frames have to be read.
- In this frame, a Haar cascade is used for the detection of the mouth region based on an OpenCV implementation of the Viola and Jones algorithm for face detection. The Viola and Jones object detection framework is the first such framework to provide competitive rates in real-time. It was proposed in 2001 by Viola and Jones [27]. Although it can be

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

Audio	Frames	Hamming Distance	Precision	Accuracy	F1 Metric
ABERCROMBIE	200	4	0.98	0.98	0.99
ANGIE	220	33	0.82	0.85	0.90
AYALA	350	14	0.96	0.90	0.94
BECKHAM	185	28	0.91	0.85	0.91
DICAPRIO	185	24	0.96	0.87	0.92
FOXNEWS	150	6	0.96	0.96	0.98
GOOGLE	265	32	1.00	0.88	0.93
HAYS	235	14	0.97	0.94	0.97
LARRYPAGE	245	22	0.95	0.91	0.95
LISA	250	15	0.94	0.94	0.97
MORRIS	200	4	0.98	0.98	0.99
RESUME	245	22	0.94	0.91	0.95
STOSSEL	400	8	0.99	0.96	0.98
USAJOBS	265	16	0.96	0.91	0.94
USAJOBS2	350	7	0.97	0.97	0.98

trained to detect a variety of object classes [1, 12], for example the mouth region as in this algorithm, its development was motivated by the problem of face detection. The mouth region will be our Region Of Interest (ROI).

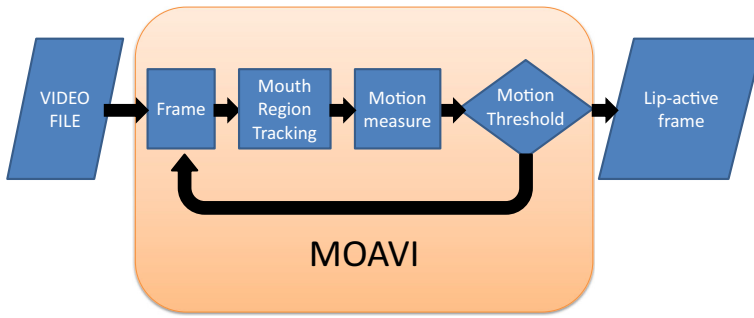
- In the ROI of the frame, we measure the motion that appears between the previous and current frame. The algorithm for estimating the amount of motion is explained in detail in the next figure.
- A motion threshold is compared with the calculated motion to determine if the output of the metric is lip-active. This threshold was 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 14 describes the algorithm in general. The key block for the detection of lip movement is known as motion measure. Figure 15 explains 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.

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

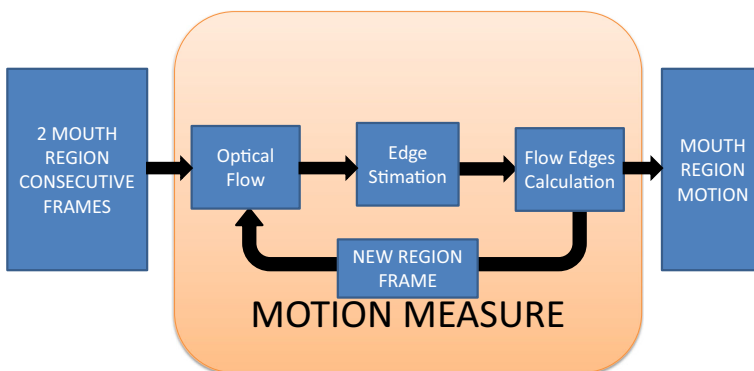
Total Frames	Total Hamming Distance	Precision	Accuracy	F1 Metric
3182	249	0.95	0.92	0.95



**Fig. 14** Algorithm for the detection of lip movement

- The optical flow between them is calculated. The implementation is based on the algorithm described in research carried out by Farneback [6]. 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 the movement of the entire face and the movement of the lip region independently. This was achieved by calculating the edges of the optical flow output. This involves knowing the Laplacian of the motion field, and analysing the borders. If the border is in the mouth ROI, we consider it as an indicator of independent movement of the lips.
- The final step is to count how much of the edge region of the optical flow was discovered in the mouth region. The number of these edges is strongly correlated with the amount of lip motion in the frame.

The total information from the OpenCV metric is loaded into MATLAB to be processed and to continue with the comparison with information coming from the audio part. This means that 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 study into C++ and OpenCV.



**Fig. 15** Detailed block diagram for motion measure

### 3.3.2 Results

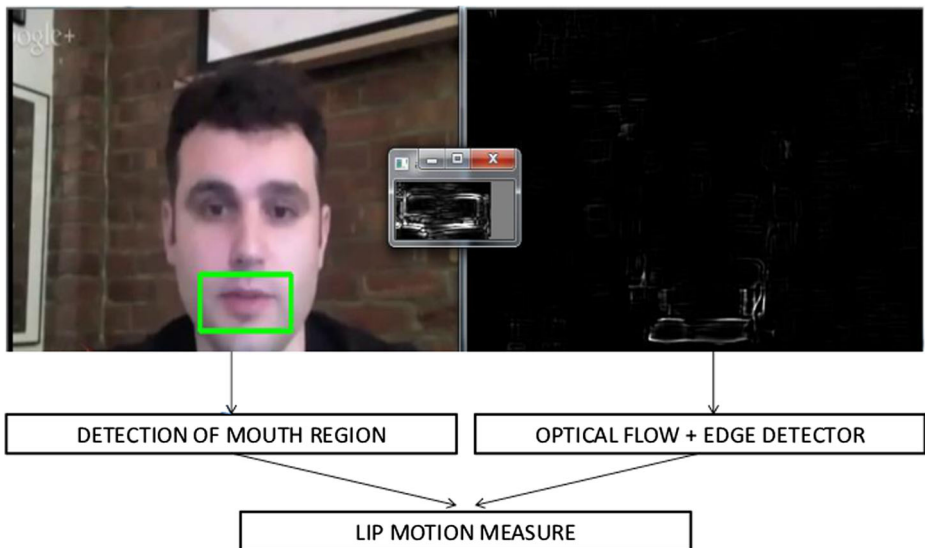
The output of the algorithm for Lip Activity Detection (LAD) is a binary vector showing the instants in which the video information analysis provides evidence of lip movement. This binary vector is compared with the binary vector obtained with the VAD algorithm. The comparison is carried out using the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

Being a video metric has the advantage of showing its behaviour in an image, which is not possible for audio metrics. Figure 16 shows the graphical output for a frame of the LAD metric for MOAVI. The frame originates from one of the audiovisual sequences, named STOSSEL, which 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. The graphical representation of the output of the metric is shown in the middle of the figure.

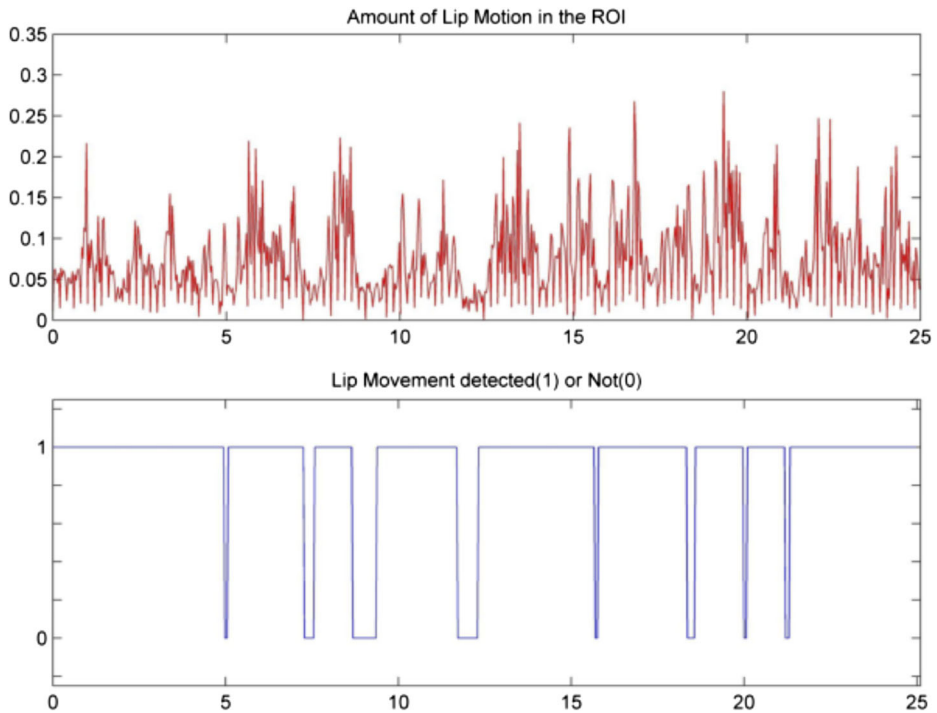
The results subsection of the LAD shows graphs of the outputs of the metrics described above. A typical output of the motion measure block is represented in the upper graph of Fig. 17. The binary vector determined from this information is shown in the graph below. This binary vector, based on the threshold of the amount of motion, indicates which of the frames are considered active in terms of lip movement.

### 3.4 Delay calculation

The goal of the previous algorithms, VAD and LAD, was to provide a binary vector originating from the audio information and another from the video information. In the second step, they are compared with each other to obtain the delay between them. This section explains the algorithm used in this comparison and shows the results.



**Fig. 16** Graphical output of the LAD algorithm



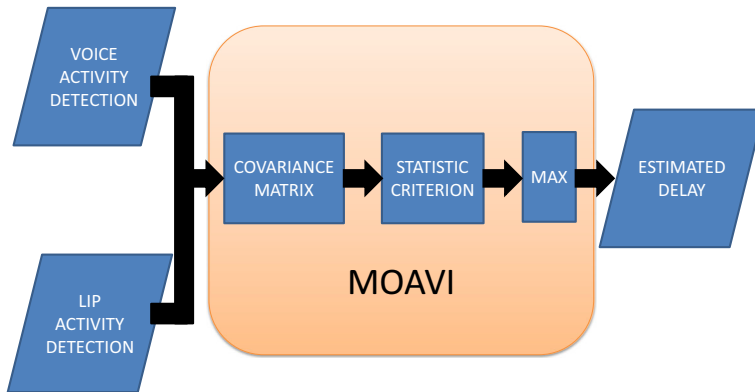
**Fig. 17** Example of detection of lip activity

### 3.4.1 Algorithm

Some delay estimation algorithms were implemented in the time-domain. For example, the basic but well-known delay estimation based on cross-correlation was used in this application, without good results. Most advanced time delay estimation algorithms are implemented in the frequency-domain, such as the generalized cross-correlation method. The problem with using the frequency-domain is the lack of accuracy in the spectral estimation for short signal segments. The delay algorithm needed in this synchronization stage aims to estimate the time shift of the audio with respect to video, and it needs to be used in short audiovisual sequences such as those stored in the database described above.

For this reason, the estimation algorithm found in [15] is a time-domain implementation that satisfies the needs of this application. The proposed information delay criterion is used. The basis of the algorithm is a time-domain implementation of the maximum likelihood method. Although numerically motivated convergence criteria are commonly used, our method uses statistically motivated convergence criteria.

The delay algorithm is outlined in the block diagram (Fig. 18). The implementation was done in MATLAB. The first input of the delay estimator is the binary vector from the VAD, while the second input is the binary vector from the LAD. Both vectors have the same length. The delay algorithm introduces different delays between the two signals, and calculates the likelihood of the pair of signals for each delay introduced. The delay that maximizes the likelihood value is the estimated delay of the two signals, and thus the output of the delay algorithm. The algorithm process is as follows:



**Fig. 18** Block diagram for delay estimation

- First, a covariance matrix is constructed based on the possible delays. In this metric, the possible delays were set to  $\pm 2$  s.
- The criterion is built up next. 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 with this method is that it is assumed that the audio and video activity are perfectly synchronized, meaning that when a person is talking and the lips are visible, the viewer can see the lips moving only when a sound can be heard.

This is clearly not accurate. One example of an absence of audiovisual speech correlation is *noisy, unvoiced motion* of the lips, such as smiling or licking of the lips. They are impossible to discriminate using this algorithm, although some differences are accepted and the estimated delay remains accurate. An example of a problem which can be corrected easily is the absence of complete synchronization between lip activity and voice activity even when the lip sync artefact has not occurred. It can be observed that lip activity starts around 300 ms before the voice can be perceived. This is a stationary delay which can be corrected simply by taking into account the 300 ms in the estimated delay. The results shown below include this artificially added gap.

### 3.4.2 Results

Section 3.2.2 shows that the accuracy of the Voice Activity Detector is 92.7 %. It has been noted that in certain situations the VAD method is not able to perfectly discriminate between human speech and other sounds. In addition, the Lip Activity Detector experiences difficulties in certain situation, such as discriminating lip motion while speaking and other types of lip movement.

In these circumstances, the two binary vectors used as inputs for the Delay Estimation Algorithm are not going to be active ( $value = 1$ ) at the same instants, even if no delay is introduced. This is why detecting the Lip Sync artefact is challenging. It is also the reason why an advanced delay estimation algorithm is used. The results of estimating the delay using this algorithm are presented in this subsection.

Since the Delay Estimation Block is the final stage of the Lip Sync Artefact Key Indicator Determination, the output of this block is a key indicator. Therefore, if the estimated delay is above the thresholds determined in previous sections (140 ms), the determined Lip Sync Artefact Key Indicator is active.

Delays of 0, 300, 500 and 800 ms are artificially introduced to analyse the delays determined by the metric. The absolute error is also calculated. An average gap of 34.28 ms (standard deviation gap: 32.92 ms) is calculated for the 60 estimations carried out during the experiment. Moreover, only 12 % possible cases failed the test by detecting a delay when none was present. This is a satisfactory result, since in 88 % of the test audiovisual sequences the binary key indicator is correct. Thus, in 88 % of cases, the key indicator determines correctly whether the lip sync artefact is present and the threshold is exceeded and whether the audio is delayed with respect to the video or vice versa.

### 3.5 Limitations and future research

As limitations, we list a few main aspects which should be improved during further research.

With respect to VAD, certain sounds that should not be detected as speech because they appear without any correlation with video information are actually detected as voice activity. Examples could be speakers which are not visible in the scene (common in films) or other background music. Further research should include audio signal processing in terms of speaker recognition to discriminate between different speakers.

With respect to LAD, certain *noisy* lip movements which should not be detected as speech because they appear without any correlation with audio information are actually detected as lip activity. Examples could be people smiling or licking their lips, which are impossible to discriminate using this algorithm. Further research should include video signal processing in terms of speaker recognition to discriminate between different people in the scene.

With respect to the Delay Estimator, further research should be capable of detecting both types of delays rather than just audio delayed with respect to video.

## 4 Conclusions

The purpose of this paper was to report the development of the audiovisual part of the MOAVI project, which includes the detection of mute, clipping and lip synchronization (also known as lip sync) artefacts.

Regarding the results obtained for the mute artefact, the algorithm works accurately and detects the artefact at the time positions when it was introduced. We suggest that two further phenomena are evaluated in future research which, if detected, should improve the mute detection accuracy. First of all, muting may be detected if there is no audio and lip movement is recognized, which is done with respect to lip sync detection. Muting may be detected if the first sample of a sequence with a value of 0 is preceded by a high value (this often produces an annoying effect).

Regarding the clipping results obtained, the detection occurs at the instants in which the waveform is cut or limited by a constant, which is exactly the instants that sound annoying when the file is played.

Regarding the results of the lip sync indicator, in 88 % of the test audiovisual sequences, the binary key indicator is correct.

**Acknowledgments** Research work co-funded by the National Centre for Research and Development, Poland, conferred on the basis of the decision number EUREKA C 2013/1-5/MITSU/2/2014.

**Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

## References

1. Baran R, Glowacz A, Matiolanski A (2015) The efficient real- and non-real-time make and model recognition of cars. *Multimed Tools Appl* 74(12):4269–4288. doi:[10.1007/s11042-013-1545-2](https://doi.org/10.1007/s11042-013-1545-2)
2. Cerqueira E, Janowski L, Leszczuk M, Papir Z, Romaniak P (2009) Video artifacts assessment for live mobile streaming applications. In: Mauthe A, Zeadally S, Cerqueira E, Curado M (eds) *Future multimedia networking, lecture notes in computer science*, vol 5630. Springer, Berlin Heidelberg, pp 242–247. doi:[10.1007/978-3-642-02472-6\\_26](https://doi.org/10.1007/978-3-642-02472-6_26)
3. Cooper J (2014) System and method for av sync correction by remote sensing. <https://www.google.ch/patents/US20140354829>. US Patent App. 14/460,305
4. Czyzewski A, Ciarkowski A, Kostek B, Cichowski J (2013) Online sound restoration system for digital library applications. *Proc Meet Acous* 20(1):055004. doi:[10.1121/1.4863268](https://doi.org/10.1121/1.4863268). URL <http://scitation.aip.org/content/asa/journal/poma/20/1/10.1121/1.4863268>
5. Czyzewski A, Kostek B, Szykalski M, Ciszewski TE (2017) Building knowledge for the purpose of lip speech identification. Springer International Publishing, Cham, pp 3–14. doi:[10.1007/978-3-319-43982-2\\_1](https://doi.org/10.1007/978-3-319-43982-2_1)
6. Farneback G (2001) Very high accuracy velocity estimation using orientation tensors, parametric motion and simultaneous segmentation of motion field
7. Garella JP, Grampín E, Sotelo R, Baliosian J, Joskowicz J, Guimerans G, Simon M (2016) Monitoring QoE on digital terrestrial TV: a comprehensive approach. In: 2016 IEEE International symposium on broadband multimedia systems and broadcasting (BMSB), pp 1–6. doi:[10.1109/BMSB.2016.7522008](https://doi.org/10.1109/BMSB.2016.7522008)
8. Glowacz A, Grega M, Gwiazda P, Janowski L, Leszczuk M, Romaniak P, Romano SP (2010) Automated qualitative assessment of multi-modal distortions in digital images based on GLZ. *Ann Telecommun Annales des Télécommun* 65(1):3–17. doi:[10.1007/s12243-009-0146-6](https://doi.org/10.1007/s12243-009-0146-6)
9. Han C, Kim J (2009) Method and apparatus for testing lip-sync of digital television receiver. <https://www.google.ch/patents/US7586544>. US Patent 7,586,544
10. Kacprzak S, Ziółko M (2013) Speech/music discrimination via energy density analysis. In: Dediu AH, Martín-Vide C, Mitkov R, Truthe B (eds) *Statistical language and speech processing, lecture notes in computer science*, vol 7978. Springer, Berlin Heidelberg, pp 135–142. doi:[10.1007/978-3-642-39593-2\\_12](https://doi.org/10.1007/978-3-642-39593-2_12)
11. Kim J, Han C (2005) Method and apparatus for testing lip-sync of digital television receiver. <http://www.google.com.gt/patents/WO2005004470A1?cl=zh>. WO Patent App. PCT/KR2004/001,616
12. Leszczuk M, Baran R, Skoczylas L, Rychlik M, Slusarczyk P (2014) Public transport vehicle detection based on visual information. In: Dziech A, Czyzewski A (eds) *Multimedia communications, services and security, communications in computer and information science*, vol 429. Springer International Publishing, pp 16–28. doi:[10.1007/978-3-319-07569-3\\_2](https://doi.org/10.1007/978-3-319-07569-3_2)
13. Leszczuk M, Hanusiak M, Farias MCQ, Wyckens E, Heston G (2016) Recent developments in visual quality monitoring by key performance indicators. *Multimed Tools Appl* 75(17):10,745–10,767. doi:[10.1007/s11042-014-2229-2](https://doi.org/10.1007/s11042-014-2229-2)
14. Lu G, Hankinson T (2000) An investigation of automatic audio classification and segmentation. In: WCC 2000 - ICSP 2000. 2000 5th international conference on signal processing proceedings. 16th World computer congress 2000, vol 2, pp 776–781. doi:[10.1109/ICOSP.2000.891627](https://doi.org/10.1109/ICOSP.2000.891627)
15. Moddemeijer R (1999) On the convergence of the iterative solution of the likelihood equations
16. Pastrana R, Gicquel J, Colomes C, Cherifi H (2004) Sporadic signal loss impact on auditory quality perception



17. Person A, Muccioli J (1995) Adjustable clip detection system. <https://www.google.com/patents/US5453716>. US Patent 5,453,716
18. Ramirez J, Segura J, Gorriz J (2007) Voice activity detection. Fundamentals and speech recognition system robustness. INTECH Open Access Publisher. <https://books.google.pl/books?id=AB3qoAEACAAJ>
19. Skoglund J, Linden J (2014) Audio clipping detection. <https://www.google.com/patents/US20140226829>. US Patent App. 13/767,387
20. Sohn J, Sung W (1998) 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. doi:[10.1109/ICASSP.1998.674443](https://doi.org/10.1109/ICASSP.1998.674443)
21. 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 (2012) Assessing the importance of audio/video synchronization for simultaneous translation of video sequences. *Multimed Syst* 18(6):445–457. doi:[10.1007/s00530-012-0262-4](https://doi.org/10.1007/s00530-012-0262-4)
22. Stanger L (2007) Method and apparatus for lipsync measurement and correction. <https://www.google.ch/patents/US7212248>. US Patent 7,212,248
23. Steinmetz R (1996) Human perception of jitter and media synchronization. *IEEE J Select Areas Commun* 14(1):61–72. doi:[10.1109/49.481694](https://doi.org/10.1109/49.481694)
24. Vanderhoff W, Laparidis A, Halstead R, Downey W, Chen L, Parrino R (2013) System for testing set-top boxes and content distribution networks and associated methods. <http://www.google.com/patents/US8595784>. US Patent 8,595,784
25. Vavrek J, Pleva M, Lojka M, Vizslay P, Kiktová E, Hládek D, Juhár J, Pleva M, Kiktova E, Hladek D et al (2013) Tuke at mediaeval 2013 spoken web search task. In: *MediaEval*
26. Venkatesh R, Ajit B, Bopardikar S, Perkis A, Hillestad OI (2002) No-reference metrics for video streaming applications
27. Viola P, Jones M (2001) Robust real-time object detection. In: *International journal of computer vision*
28. Yamasaki H, Furuya O, Mitsui A (2012) Image synthesizing device, coding device, program, and recording medium. <https://www.google.com/patents/US20120120312>. US Patent App. 13/357,862
29. Zhang D, Bao C, Deng F, Xia B, Chen H (2011) A restoration method of the clipped audio signals based on MDCT. In: 2011 IEEE International symposium on signal processing and information technology (ISSPIT), pp 253–257. doi:[10.1109/ISSPIT.2011.6151569](https://doi.org/10.1109/ISSPIT.2011.6151569)



**Ignacio Blanco Fernández** is a researcher and professional analyst in the fields of web and video quality and performance. Ignacio completed his Msc in Telecommunications Engineering at University of Oviedo, Spain on 2013. He became interested in video and software quality during the investigation he conducted for his Master Thesis at AGH University of Science and Technology in Cracow, Poland. He collaborated then with INDECT European project. After finishing his studies, he joined Hewlett-Packard at 2013 to become a performance testing engineer. He has recently joined Experis IT at web optimization projects related to banking industry. His research interests lie in the area of audiovisual signal treatment and web/app software quality, ranging from theory to design to implementation.



**Mikołaj Leszczuk** PhD. He started his professional career in 1996 at COMARCH SA as manager of the Multimedia Technology Department, and then at COMARCH Multimedia as the CEO. Since 1999 has been employed at the AGH Department of Telecommunications. In 2000 he moved to Spain for a fourmonth scholarship at the Universidad Carlos III de Madrid. After returning to Poland, he was employed at the Department of Telecommunications as a research and teaching assistant, and in 2006, he successfully defended his doctoral dissertation as an assistant professor. His current research interests are focused on multimedia data analysis and processing systems, with particular emphasis on Quality of Experience. He (co-)authored approximately 130 scientific publications of which 23 are publications in journals of the JCR database. He has been teaching at undergraduate and graduate levels. He has cosupervised 1 PhD student and supervised (promoted) approximately 40 MSc students of various nationalities. He has participated more than 20 major research projects, including FP4, FP5, FP6, FP7, Horizon 2020, OPIE, Culture 2000, PHARE, eContent+, and Eureka!. Between 2009 and 2014, he was the administrator of the major international INDECT research project, dealing with solutions for intelligent surveillance and automatic detection of suspicious behaviour and violence in urban environments. He is a member of VQEG (Video Quality Experts Group, board member), IEEE (Institute of Electrical and Electronics Engineers), and GAMA (Gateway to Archives of Media Art). The latter organization collaborates with the VQiPS (Video Quality in Public Safety) working group. More information:<http://www.linkedin.com/in/miklesz>.