Assessing the Importance of Audio/Video Synchronization for Simultaneous Translation of Video Sequences

Nicolas Staelens · Jonas De Meulenaere · Lizzy Bleumers · Glenn Van Wallendael · Jan De Cock · Koen Geeraert · Nick Vercammen · Wendy Van den Broeck · Brecht Vermeulen · Rik Van de Walle · Piet Demeester

Received: date / Accepted: date

Abstract Lip synchronization is considered a key parameter during interactive communication. In the case of video conferencing and television broadcasting, the differential delay between audio and video should remain below certain thresholds, as recommended by several standardization bodies. However, further research has also shown that these thresholds can be relaxed, depending on the targeted application and use case. In this article, we investigate the influence of lip sync on the ability to perform real-time language interpretation during video conferencing. Furthermore, we are also interested in determining proper lip sync visibility thresholds applicable to this use case. Therefore, we conducted a subjective experiment using expert interpreters, which were required to perform a simultaneous translation, and using non-experts. Our results show that significant differences are obtained when conducting subjective experiments with expert interpreters. As interpreters are primarily focused on performing the simultaneous translation, lip sync detectability thresholds are higher compared to existing recommended thresholds. As such, primary focus and the targeted application and use case are important factors to be considered when selecting proper lip sync acceptability thresholds.

Keywords Audio/video synchronization · Lip sync · Subjective quality assessment · Audiovisual quality · Language interpretation

N. Staelens · N. Vercammen · B. Vermeulen · P. Demeester
Ghent University - IBBT, Department of Information Technology, Ghent, Belgium
E-mail: {nicolas.staelens, nick.vercammen, brecht.vermeulen, piet.demeester}@intec.ugent.be

J. De Meulenaere · L. Bleumers · W. Van den Broeck
Free University of Brussels - IBBT, Studies on Media, Information and Telecommunication, Brussels, Belgium
E-mail: {jonas.de.meulenaere, lizzy.bleumers, wvdbroec}@vub.ac.be

G. Van Wallendael · J. De Cock · R. Van de Walle
Ghent University - IBBT, Department of Electronics and Information Systems, Ghent, Belgium
E-mail: {glenn.vanwallendael, jan.decock, rik.vandewalle}@ugent.be

K. Geeraert
Televic N.V., Izegem, Belgium
E-mail: k.geeraert@televic.com

1 Introduction

Perceived quality of audiovisual sequences can be influenced by the quality of the video stream, the quality of the audio stream and the differential delay between the audio and video (A/V synchronization) [14]. In the case of interactive communication, such as video conferencing, A/V synchronization is considered a key parameter [32] and is more commonly referred to as lip synchronization (lip sync) [6]. According to International Telecommunication Union (ITU)-T Recommendation P.10 [15], the goal of lip sync is to '*provide the feeling that the speaking motion of the displayed person is synchronized with that person's voice*'.

Several standard bodies such as the ITU, the European Broadcast Union (EBU) and the Advanced Television Systems Committee (ATSC) formulated a series of recommendations [1], [5], [11], [13] concerning the maximum allowed differential delay between audio and video in order to maintain satisfactory perceived quality. However, further research [3], [7], [28] has already pointed out that these recommendations can be relaxed in some cases, depending on the targeted use case and application.

Similar to video conferencing, simultaneous translation or language interpretation is also an example of in-

teractive video communication. In professional environments, such as the European Parliament, interpreters usually reside in specially equipped interpreter booths (see Figure 2) during the debates. Furthermore, these debates are recorded and broadcasted to the booths and also made available as live video streams broadcasted over the Internet. The content of such a live video stream typically consists of close-up views of the current active speaker and provides the interpreters with additional non-verbal information (gestures, facial expressions) which can facilitate the simultaneous translation.

In general, the existing recommended A/V synchronization thresholds are determined based on subjective experiments conducted using non-expert users [12]. However, interpreters can be regarded as expert users since they actively use the video stream while performing the simultaneous translation and also process the non-verbal information from the video.

Recent studies have showed that non-experts are more tolerable compared to experts during audiovisual quality assessment [26]. Furthermore, context and primary focus are also important factors to consider during quality assessment [27]. Therefore, additional research is needed to investigate whether the existing thresholds are also valid in the expert use-case of language interpretation.

In this article, we are particularly interested in investigating how delay between audio and video is perceived by real interpreters and how this delay affects their ability to perform simultaneous translations. Face-to-face interviews were organized with interpreters in order to talk about the relative importance of audio/video synchronization, the added value of having visual feedback (next to the audio signal) and which kind of (additional) information interpreters usually use or require for performing simultaneous translation. Furthermore, we also conducted a subjective audiovisual quality experiment during which the interpreters were asked to perform simultaneous translation of a number of video sequences as they would do in real-life. After each sequence, the interpreters were questioned about the audio/video delay and the overall audiovisual quality. The results of the subjective test are then compared with the results obtained during the face-to-face interviews. As a last step, we also conducted the same subjective experiment using non-expert users in order to compare the results concerning audio/video delay visibility and annoyance with the results of the expert users. In contrast with the interpreters, the non-expert users were not asked to perform a simultaneous translation of the video sequences.

The remainder of this article is structured as follows. In section 2, we start by describing different techniques for monitoring and measuring the differential delay between audio and video. Furthermore, we also provide an overview of already conducted research and existing standards defining a wide range of acceptability thresholds related to A/V synchronization. Based on this study, we highlight the importance of our study presented in this article. For obtaining ground-truth data, a subjective experiment has been set up and conducted. This will be explained in more details in section 3. Section 4 presents the results of this subjective experiment which we conducted using both experts and non-experts. The differences in the results obtained using these two targeted user groups are also discussed in more details in the same section. Finally, we conclude the article in section 5.

2 Monitoring and measuring audio/video synchronization

In order to ensure and maintain synchronized audio and video, several measurement and monitoring techniques have been proposed in literature. Furthermore, research has already been conducted in order to determine A/V synchronization acceptability thresholds for several applications such as video broadcasting and video conferencing. However, as will be explained in more details in the next sections, a wide range of different thresholds have been identified each of which are dependent on the application.

2.1 Audio/video synchronization measurement techniques

In many broadcast systems, 'off-line' measurement techniques are used to maintain audio/video synchronization. Presentation time stamps (PTS), for example, can be embedded in MPEG transport streams to avoid A/V synchronization drift. Similarly, comparison of SMPTE time codes in audio and video signals can be used to synchronize the audio and video signals. These time stamps or time codes are often added after the video undergoes frame synchronization, format conversion and pre-processing. As a result, delays or misalignment in these stages remain uncompensated. Also, as time codes have no actual relation to the signal, mistimed or misaligned information can lead to a loss of A/V synchronization.

A number of solutions have been proposed that can overcome these limitations. In order for these techniques

to be useful in conferencing and broadcast environments, a number of requirements should be met. As the synchronization errors can vary in time, it is important that the measurement method responds to the A/V synchronization in a dynamic way and in real-time (in-service). Preferably, the techniques should work for all types of audio and video content, independent of the used format. Also, they should be robust to modifications of the audio and video signals that can occur during content distribution.

Roughly, three classes of methods can be distinguished for dynamically measuring the A/V synchronization based on the correspondence between both signals.

A first class exploits the relationship between acoustic speech and the corresponding lip features (such as width and height) and lip movements. In Li *et al.* [19], a high correlation between the estimated and measured visual lip features was found. Evidently, such methods are constrained to video content where lip motion is present.

Secondly, watermarking solutions have been investigated for A/V synchronization. Watermarking can, e.g., embed information about the audio signal into the video stream. The envelope of the audio signal is analyzed, from which a watermark is generated. This watermark can be embedded in the corresponding video stream. At a receiver point, the video and audio streams and the watermark can be observed to obtain a measure of the A/V synchronization. One issue with this technique is that the watermark is not necessarily robust to adaptation of the video and/or audio signal, for example, when transrating, aspect ratio conversion, or audio downmixing are applied.

In a third class of techniques, an A/V synchronization fingerprint (also referred to as 'signature' or 'DNA') is added to the audio and video signals. Features from both signals are extracted and combined into an independent data stream at a point where both signals are known to be in-sync. Later on, this data stream can be used to measure and maintain the A/V synchronization. Fingerprinting exploits characteristical features of the video or audio (such as luminance, transitions, edges, motion etc.) and uses a formula to condense the data into a small representation [18], e.g., based on robust hash codes [8]. These hash codes are sent in the data stream, and ensure that small perturbations in the audio and video features caused by signal processing operations will not change the hash bits drastically. At the detection point, signatures are again extracted based on the received signals, and a comparison is made between the generated and transmitted signatures within a short time window. The output of the correlator between both signatures will result in an estimated delay. Real-time systems based on these techniques have been described in [30,25].

To secure interoperability of A/V synchronization techniques, standardization initiatives have been started. Recently, the SMPTE 22TV Lip Sync Ad Hoc Group (AHG) has been studying the problem. The goal of this AHG is the creation of a standard for audio and video fingerprinting algorithms, transport mechanisms, and associated recommended practices. An overview of their activities is given in [29].

2.2 Audio/video synchronization perceptibility thresholds

As mentioned in the introduction, several standard bodies have already established a set of performance objectives for audio/video synchronization which has resulted in different detectability and acceptability thresholds. According to ITU-R Recommendation BT.1359, the thresholds for detecting A/V synchronization errors are at +45 ms and -125 ms [11], where a negative number corresponds with audio delayed with respect to the video. The standard also specifies that synchronization errors become unacceptable in case the delay exceeds +90 ms or -185 ms. Recommendation R37 of the EBU [5] defines that the end-to-end delay between audio and video in the case of television programs should lie between +40 ms and -60 ms. These thresholds are lower compared to the detectability thresholds as specified in ITU-R Rec. BT.1359. The ATSC Implementation Subcommittee (IS) 191 [1] argues that the recommendations from ITU-R Rec. BT.1359 are inadequate for digital TV broadcasting and state that the differential audio/video delay should remain between +15 ms and -45 ms in order to deliver tightly synchronized programs. The same thresholds are also recommended by the DSL Forum [4] and ITU-T Recommendation G.1080 [13].

Due to the fact that these international standards propose different audio/video synchronization thresholds, a lot of research has been performed and is still ongoing in order to evaluate and identify lip sync thresholds for different applications and use cases.

Steinmetz [28] performed an in-depth analysis of the influence of jitter and media synchronization on perceived quality. The goal of his study was to identify the thresholds at which lip sync becomes noticeable and/or annoying. The test sequences consisted of simulated news broadcasts, with a resolution of 240x256 pixels, in which delay up to 320 ms between audio and video was inserted. The majority of the test subjects did not detect audio/video delays up to 80 ms whereas

delays of more than 160 ms are detected by nearly all subjects. Furthermore, these thresholds are both valid for audio leading the video and video leading the audio. Results concerning the annoyance of the perceived lip sync indicate that delays up to 80 ms are acceptable for most of the subjects. When audio lags the video with more than 240 ms or audio leads the video more than 160 ms, lip sync is perceived as distracting.

The interaction effect on perceived quality of providing high quality video of Quarter Common Intermediate Format (QCIF) resolution (176x144 pixels) with accompanying low quality audio and vice versa has been studied in [21] in the case of both interactive and passive communication. The authors conclude that video has a beneficial influence on overall multimedia quality, which corresponds with the findings from Garcia *et al.* [9]. Part of the study also involved investigating the effect of lip sync on overall multimedia quality. For the lip sync experiment, audio and video were delayed up to 440 ms. Almost half of the test subjects (45%) did not detect synchronization errors when the video stream was delayed with respect to the audio stream. In the case the audio stream was delayed, only 24% of the subjects indicated that no synchronization error occurred. These results suggest that subjects are more tolerable towards audio leading the video. Further research [20] has also pointed out that more attention to lip sync is given during passive communication compared to active communication. During the latter, subjects are more concentrated on the conversation itself.

During another multimedia synchronization study, several CIF resolution (352x288 pixels) video sequences were presented to the test subjects in order to quantify the effect of A/V delay [7]. The quality of the audio and the video stream remained constant during the experiment, only the differential delay varied between -405 ms and +405 ms. Subjects were only required to evaluate the audiovisual quality of the presented sequences using a 5-grade scale. Results show that, even in the case no delay is present in the sequence, subjects never rated the sequences to be excellent quality. Furthermore, sequences with an audio offset of -40 ms were rated slightly better quality compared to the case of no delay. Audio offsets between -310 ms and +140 ms are all rated as good quality. Overall, audio lagging the video was perceived as less annoying compared to audio leading the video, which is in slight contrast with the results of Mued *et al.* [21] as discussed above.

The absolute perceptual threshold for detecting audio/video synchronization errors when audio is leading the video is at 185,19 ms according to the results of Younkin *et al.* [32]. This experiment did not include sequences in which the audio was lagging, but the authors assume that the detection threshold of audio lagging the video should be higher.

A similar experiment as the one conducted by Steinmetz [28] has been repeated in [3] where the focus was specifically on mobile environments. The authors argue that different detection and annoyance thresholds may apply in mobile environments due to the change in screen size, viewing distance and frame rate compared to the TV viewing environment. As such, small resolution (QCIF and Sub-QCIF) low frame rate test sequences were used during the experiment. The lip sync detection threshold, in the case of audio leading the video, is at 80 ms. It must be noted that a more strict evaluation method was used to determine this threshold compared to the results in [28]. In the case of audio lagging the video, the detection threshold appears to be content and frame rate dependent and varies between -160 ms and -280 ms.

Figure 1 provides a graphical overview of the different thresholds as identified by the international standards and research findings described above. It is clear that each application and use case scenario is characterized by different detectability thresholds. Furthermore, as the figure also shows, the acceptability thresholds span a wide range of allowable differential delay between the audio and the corresponding video stream. Therefore, additional research is needed in order to identify proper lip sync detectability thresholds in the case of simultaneous translation of video sequences and to investigate the relative importance of providing visual feedback to the interpreters.

3 Subjective quality assessment of audio/video delay during simultaneous translation

In order to collect ground-truth data concerning the visibility, annoyance and influence of A/V delay in the case of simultaneous translation, a subjective audiovisual quality experiment has been set up and conducted using expert interpreters. Furthermore, the experiment has also been conducted with non-expert users in order to investigate whether there are significant differences with the results obtained using the interpreters as both user groups have a different primary focus and expertise.

3.1 Experimental setup

Internationally standardized subjective audiovisual quality assessment methodologies, such as the ones described in ITU-T Recommendation P.911 [16] and ITU-T Rec. P.920 [17], include detailed guidelines on how to set up

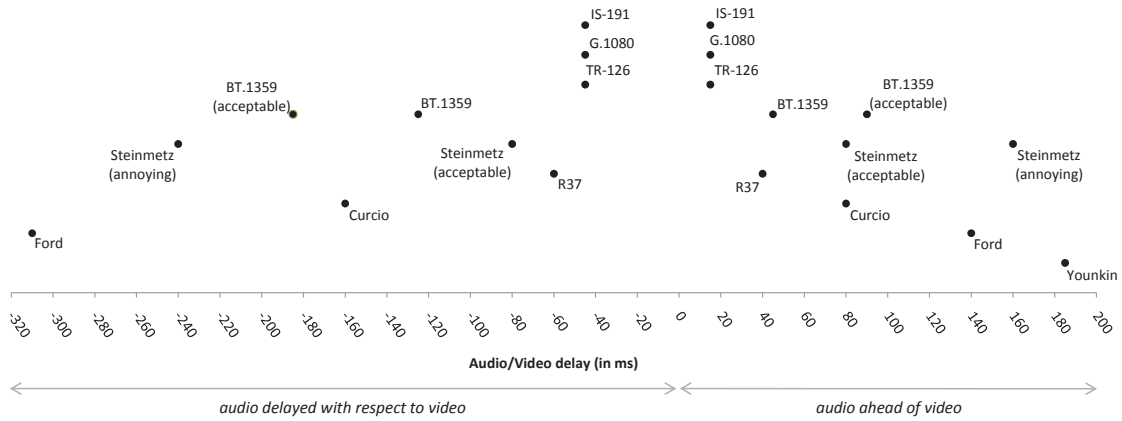

Fig. 1 Graphical representation of the different audio/video delay and lip sync detectability thresholds as identified by several standard bodies and already conducted research

and conduct such quality experiments. For the evaluation of audiovisual sequences, these methodologies describe the order in which the sequences must be presented to the test subjects and propose different rating scales which can be used by the subjects to assign a quality score to the corresponding sequence. Furthermore, the standards also pose some stringent demands related to the viewing and listening conditions by specifying, amongst others, the viewing distance between the test subject and the screen, the luminance level of the screen, the overall room illumination and the allowed amount of background noise. As such, subjective quality experiments are usually conducted in controlled environments.

Preliminary results in [24] show that subjects' audiovisual quality ratings are not significantly influenced when conducting subjective experiments in pristine lab environments, compliant with the ITU Recommendations, or on location (e.g. in a company's cafeteria with background noise and different lighting conditions). This indicates that the overall test room conditions, as specified in [16] and [17], can be relaxed to some extent.

In previous research [27], we also investigated the influence of conducting subjective quality assessment experiments in real-life environments, where subjects are not primarily focused on (audio)visual quality evaluation. Our results show that impairment visibility and annoyance are significantly influenced by subjects' primary focus and that measuring Quality of Experience (QoE) should ideally be performed in the most natural environment corresponding to the video service under test. The latter also complies with the definition of QoE which states that the quality, as perceived subjectively by the end-user, can be influenced by user expectations and context [15].

In the case of performing simultaneous translations, interpreters usually reside in special designated interpreter booths as depicted in Figure 2.

Fig. 2 Typical interpreters' booth for performing simultaneous translations

Based on the research findings mentioned above, we also opted to conduct our subjective experiments in the interpreter's most natural environment by mimicking a typical interpreter's booth as much as possible. As such, our assessment environment illustrated in Figure 3 consists of similar hardware as used in a professional environment in order to ensure that our test subjects have a similar experience compared to the real-life scenario.

As can be seen from Figure 2 and Figure 3, a display which shows a live video stream with a close-up of the person currently talking is also at the interpreter's disposal.

3.2 Audiovisual subjective assessment methodology

During subjective audiovisual quality assessment, test subjects watch and evaluate the perceived quality of a number of video sequences. In general, two different types of methodologies can be used for displaying the different test sequences to the subjects.

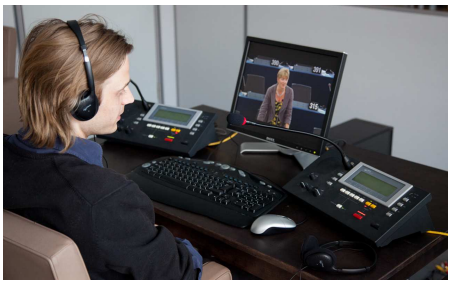

Fig. 3 Environmental setup as used during our subjective quality assessment experiment in order to mimic a realistic environment (cfr. Figure 2)

First of all, sequences can be shown pairwise using a Double Stimulus (DS) methodology. In this case, two sequences (usually the original version and an impaired or degraded version of it) are first presented to the test subjects after which they need to evaluate the quality differences between both sequences. As such, each test sequence is always presented in relation with a reference sequence. These methodologies are commonly used for evaluating the performance of video codecs [12].

A second type of methodologies, called Single Stimulus (SS), present the test sequences one at a time to the subjects. Immediately after watching the video sequence, subjects have to provide a quality rating. This means that the quality of each sequence must be evaluated without the use of an explicit reference sequence representing optimal quality. A typical trail structure of an SS methodology is depicted in Figure 4.

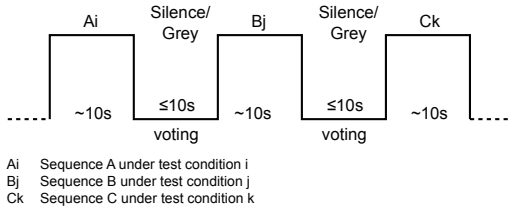

Fig. 4 Typical trail structure for an SS methodology [16], during which sequences are presented one at a time and immediately evaluated after watching

It is clear that SS methodologies correspond more with the way people watch video on their computer or on their television [10], [31]. This is also the case for the video streamed to the interpreter booths. As such, we also used the SS methodology to show the test sequences one after another to the different subjects.

After watching each video sequence, subjects were required to answer the following three questions:

1. Did you perceive any audio/video synchronization issues?
2. Do you think audio was ahead with respect to the video or vice versa?
3. How annoying does the audio/video synchronization problem appear to you, on a scale from 1 to 5?

For the last question, subjects were presented with the five-level impairment scale as depicted in Figure 5.

5 – Imperceptible
4 – Perceptible but not annoying
3 – Slightly annoying
2 – Annoying
1 – Very annoying

Fig. 5 Five-level impairment scale [16] used for collecting subjects' responses concerning audio/video delay annoyance

In case the user did not perceive any audio/video synchronization problem in the presented video sequence (thus answering no on the first question), questions 2 and 3 were automatically skipped.

As specified in ITU-T Rec. P.911, subjects also received specific instructions on how to evaluate the different video sequences. Furthermore, before the start of the real subjective experiment, two training sequences were presented to the subjects in order to get them familiarized with the subjective experiment and the range of audio/video synchronization issues they could expect. The audiovisual quality ratings given to these two training sequences are not taken into account when processing the results. A standard headset was used for playback of the audio stream. During the training sequences, the test subjects were allowed to regulate the volume of the headset.

As we are interested in assessing the influence of lip synchronization errors on the ability to perform simultaneous translation of video sequences, the interpreters participating in our subjective experiment were also required to perform this task during sequence playout. As such, the interpreters were mainly focused on the simultaneous translation of the video sequences. It must be noted that they were still aware of the possibility of audio/video synchronization errors as this was stated at the beginning of the trail. As already mentioned, the experiment was also conducted using non-expert users. These were not required to simultaneously translate the sequences and were therefore mainly focused on detecting audio/video delays.

As recommended in [12], the preferred viewing distance between the screen and the test subjects should be around seven times the screen height (H). However, as can be seen from Figure 2, interpreters are sitting closer to the screen as compared to the preferred viewing distance. Since we are targeting a more realistic

setup, we therefore did not force our test subjects to remain seated at a fixed viewing distance.

The screen used for playback of the video sequences was a standard 17 inch LCD panel with a resolution of 1024x768 pixels.

3.3 Selection, creation and impairing of video sequences

From Figures 2 and 3, it can be seen that the content shown on the displays in the interpreter booths typically consists of so called 'talking head' or 'news' sequences. These sequences are characterized by a close-up of one or more persons talking in front of the camera. Talking head sequences don't usually contain a lot of background motion except for the person who is in front of the camera. Examples of talking head MPEG-4 test sequences [23] include 'Akiyo', 'News', 'Mother&Daughter' and 'Silent'.

The source content we used for conducting our subjective experiment consisted of a joint debate during a plenary session of the European Parliament. During the debate, the camera always took a close-up of the active speaker. From that video content, of which we obtained the original recordings, we then selected one speaker whose native spoken language was English and who delivered a continuous speech of about 5 minutes long.

ITU-R Recommendation BT.1359 [11] specifies that the overall delay between audio and the corresponding video track should fall within the range [-185 ms, +90 ms] and that the detectability thresholds are at -125 ms and +45 ms. In this study, we want to evaluate how audio/video delay is perceived by interpreters, which are experts towards performing simultaneous translation of video sequences but not concerning video quality. As such, their detectability and acceptability thresholds may be different from the ones recommended. Therefore, we inserted delay between the audio and the video in the range of [-240 ms, +120 ms]. The source video content was captured at 25 frames per second at a resolution of 720x406 pixels. For the experiment, the delay step size was chosen to match the video frame rate which implies that the delay varied in steps of 40 ms.

For inserting delay between the audio and the video, the selected video sequence was first split into 10 shorter clips, each about 30 seconds long. This duration is slightly longer compared to the sequence duration as recommended by the ITU methodologies [16]. However, according to the results in [28], using clips with 30 seconds duration is needed for getting the subjects' impression on audio/video synchronization. We made sure that no cutting occurred in the middle of a sentence. Then, the audio and the video track were demuxed and additional delay was inserted in the audio track. Finally, the audio and the video track were remuxed back together. In this article, we are only investigating the influence of audio/video delay. Therefore, we neither changed the quality of the video nor the audio stream. As a result, the quality of the different processed video sequences matched the quality of the original source content. During the subjective experiment, the video sequences were played back in the original order, one after another. This way, we ensured that the natural flow of the speech was not broken and that the conversation remained logical to the interpreters.

A commonly used methodology for determining detectability thresholds is the staircase method [2] which would adaptively adjust (increase or decrease) the delay between the audio and the video in consecutive video sequences, depending on the subject's responses. However, using such methodology, subjects can pick up the delay behavior in the different sequences and anticipate their responses [32]. Therefore, we randomly inserted the delay in each video sequence. Furthermore, as we have a fixed playout order, no adaptive re-ordering of the sequences is possible. An overview of the delay inserted in each video sequence is listed in Table 1.

Table 1 Inserted delay between audio and video in each video sequence. Negative numbers imply that the audio is delayed with respect to the video.

Sequence	A/V delay (in ms)
01	0
02	-120
03	80
04	-80
05	-200
06	40
07	-160
08	120
09	-240
10	-40

4 Results

Using the subjective video quality assessment methodology, as explained in section 3.2, the expert users were presented with the 10 different audiovisual sequences which they were asked to simultaneously translate / interpret, just as they would do in a normal real-life situation. Afterwards, we repeated exactly the same experiment using non-expert users which were only required

to evaluate the audio/video synchronization of the sequences.

In this section, we first present the results obtained using our interpreter test subjects. Then, we compare these results with the findings from the non-experts.

4.1 Interpreters' evaluation

Fifteen expert users, of which ten female and five male, participated in this experiment. The average age was 25, with a minimum age of 20 and a maximum age of 41. As recommended by ITU Recommendation P.911, at least 15 subjects should participate in the experiment. In the case of expert users, Nezveda *et al.* [22] even showed that a significant lower number of subjects can be used.

In order to contextualize these participants and to elaborate the quantitative data, both interviews and observational research has been conducted. Before the experiment was due, a short interview took place, questioning the participants about their experiences in interpreting, the use of video conferencing tools, what they usually focus on while they interpret, the importance of visual cues, and how they normally prepare an interpretation session.

4.1.1 Interview to contextualize the interpreters

Of the test subjects, 10 had at least one year experience in interpreting English to Dutch (and vice versa) and experience with performing simultaneous translations during video conferencing. Their practical knowledge ranged from exercises in class to actual interpreting at conferences.

In general, real-life interpreting is preferred to the use of video conferencing tools as the latter may conceal considerable contextual information. It is believed that limited information about the speaker and the public impedes a proper translation. In this respect, anticipating unexpected events were recorded as well. In addition, it was also felt that one is more dependent on the technological functioning.

It was repeatedly indicated throughout the interviews that the primary focus in (real-time) interpreting is directed to the spoken word. As such, visual cues are only of secondary importance. Still, the majority of the expert users consider it helpful to have additional non-verbal information provided in visual cues such as gestures, facial expressions and lip movements. It serves as a comfort during their translation and it creates the setting in which the speaker talks. On the other hand, actively avoiding visual cues was also often cited, especially when difficulties with translating are encountered (e.g. high speech rate).

In the beginning of the experiment, the participants were informed about the nature of the sequences they were about to see. Consequently, no preparations on the subject could be made. Normally, the specific vocabulary inherent to the sector they are about to work for is thoroughly studied as well as related documents and information about the speakers. Lacking this information makes translating a more demanding task. This could affect the translation performance of the subjects, but because the main focus of the research is lip synchronization, the effect on the results is considered small.

4.1.2 Visibility and annoyance of audio/video synchronization issues

After watching each individual video sequence, subjects were required to indicate whether they perceived any audio/video synchronization issues, rate the audiovisual quality and identify whether audio was leading the video or vice versa.

In Figure 6, the percentage of the expert subjects who actually perceived the corresponding delay between the audio and the video is depicted.

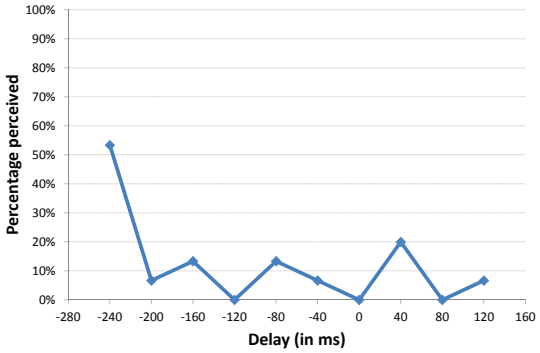

Fig. 6 Percentage of expert users who did perceive lip sync issues compared to the actual inserted delay.

In general, almost none of the expert subjects detected the desynchronization between audio and video (at most one or two subjects), even in the case where the delay is up to -240 ms. This can be explained by the fact that the expert users are primarily focused on the simultaneous translation of the audio track. As indicated during the pre-interview, visual cues are only of secondary importance and by some even actively avoided

in order to focus solely on the spoken content. The latter is especially the case when parts of the conversation become more difficult to translate. During the simultaneous translation of the different video sequences, the interpreters are also actively communicating. Results in [20] indicate that less attention is given to lip sync during active communication. According to the results presented in Figure 6, the delay between audio and video may exceed the 160 ms threshold recommended by Steinmetz [28].

Due to the low detection thresholds, there is no clear difference concerning visibility of lip sync when audio is delayed or ahead of the video signal. The graph shows that the delay between the audio and the video can be more than -240 ms or 120 ms before reaching a detection threshold of 100%.

During the subjective experiment, we observed that the participants mainly focused on the screen. Exceptionally, some of them closed their eyes, looked away or even sat back for a while. Afterwards, they explained sometimes having problems interpreting and translating the sequences, caused by the high speech rate, the dense information, uncertainty about a translation or in some cases the asynchronicity between the audio and the video. The latter is remarkable as the above graph shows that only a small percentage of the experts actually perceived this asynchronicity.

The overall average quality ratings given to the different sequences, as shown in Figure 7, remain high as only a small percentage of the experts detect the A/V synchronization issues.

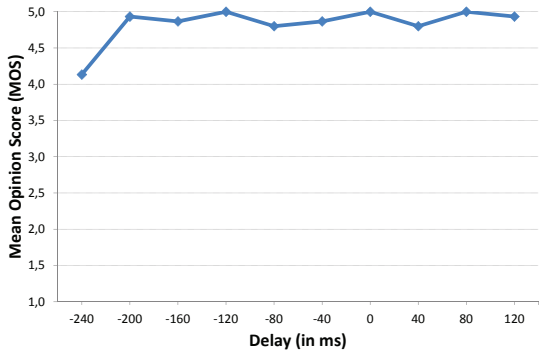

Fig. 7 MOS scores given by the experts to the sequences with inserted delay between audio and video.

Even when the delay goes up to -240 ms, the quality of that particular video sequence is still not perceived as being annoying (MOS $>$ 4), similar to the results obtained in [7]. Analyzing the individual quality ratings given by the test subjects to each video sequence showed that the quality score drops on average by 1.3, with a standard deviation of 0.4, in case a lip sync problem is detected.

Finally, the interpreters were also asked to indicate, in the case of an A/V synchronization issue, whether they perceived the audio to be delayed with respect to the video or vice versa. As the graph in Figure 8 shows, very few experts are able to correctly classify the relationship between the video and the audio track.

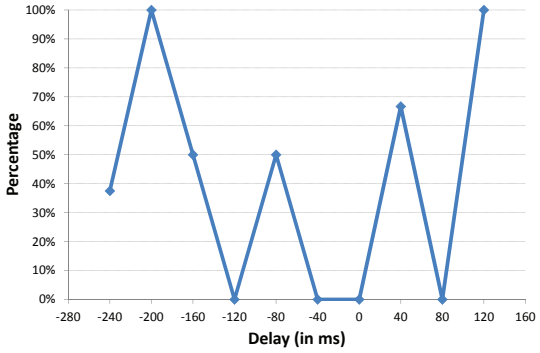

Fig. 8 Percentage of the experts who correctly determined whether audio was leading video or vice versa, in case they perceived A/V synchronization issues.

It must be noted that the graph only takes into account the subjects who actually detected the A/V synchronization problem. As such, this graph should be closely inspected in relation to the graph from Figure 6 when interpreting the results. For example, even though the classification accuracy is 100% in the case of a delay of -200 ms, only one of the test subjects actually detected this synchronization issue.

In the case of a delay of -240 ms, 53% of the subjects detected the synchronization problem. However, only 38% of them is able to correctly detect that the audio was indeed delayed with respect to the video. Further analysis of the individual responses showed that subjects fail to identify whether audio is ahead or delayed compared to the video. Even when a particular subject identifies different sync problems, he/she is not able to differentiate delayed sound from delayed video. As such, similar to the question whether they perceived a synchronization issue, subjects are again trying to guess the answer.

Our results show a high correlation between the different test subjects. It also clearly shows that, when interpreters are mainly focused on performing the simultaneous translation, audio/video delay is not a pri-

mary concern to them. Furthermore, the test subjects fail to combine real-time interpretation with assessing the audiovisual quality of the presented sequences. Even in the case of a severe differential delay ($\geq$ 240 ms) between audio and video, synchronization issues become only slightly detectable.

4.1.3 Post-experimental interview: extending the quantitative data

Throughout the interviews it was recurrently indicated that the used audiovisual sequences were demanding and required high concentration. Interestingly, the provided reasons included mainly factors associated to the content of the sequences (e.g. high speech rate, vocabulary, or diction) or themselves (lack of preparation) and only to some the detected desynchronizations. Furthermore, the participants assessed their performance worse than what they normally achieve. The discrepancy between the low detection rates and the encountered difficulties suggest that the participants were highly involved in completing the test, leaving little to no capacity to assess the (de-)synchronization. This is supported by the expressed uncertainty regarding their detections and whether audio or video was leading. Furthermore, as the contextualization interviews indicated, visual cues are secondary to auditory cues, meaning that less attention is paid to the video in the first place. Only when the delay was up to -240 ms, the desynchronization was substantially more detected. A modest part of the participants expressed during the interviews that, when the desynchronization was perceived, it did disturb them in completing their translation. The desynchronization amplified the difficulties one already had, manifesting itself primarily as a loss of concentration. Yet, the MOS scores indicate that none of the sequences were considered annoying.

Despite the low detection rate, audio/video synchronization is often considered important. A correlation seems to exist between the experienced difficulties and the allocated weight to audio/video synchronization. The data suggests that the more difficulties were encountered while translating, the more the importance of synchronization is emphasized. Quoting the participants, the maximal allowed delay varies from none or milliseconds to not more than a few words.

Nevertheless, an impaired audio-visual stream was recurrently preferred to a single audio track. As long as the delay is not too high, nor too long, video is considered a valuable asset as it provides the interpreter with a certain comfort. Even in the case of this experiment, in which the speaker showed little expressions or gestures, the video was considered helpful to more than one participant.

4.2 Comparison with non-experts users

In this section, we investigate how the average end-users perceive audio/video synchronization in order to see whether there is a significant difference with respect to the interpreters. Test subjects were asked to watch the same audiovisual sequences as the interpreters and evaluate whether they perceived any audio/video synchronization issues. In contrast to the expert interpreters, the non-expert users were not asked to perform a simultaneous translation of the speech. As a result, the non-experts are primarily focused on detecting A/V synchronization issues.

A total number of 24 non-expert users, aged between 24 and 34 years old participated with the subjective experiment.

4.2.1 Detecting audio/video delay

Figure 9 shows the percentage of viewers who perceived any kind of A/V synchronization problem, compared to the actual delay inserted between the audio and the video signal.

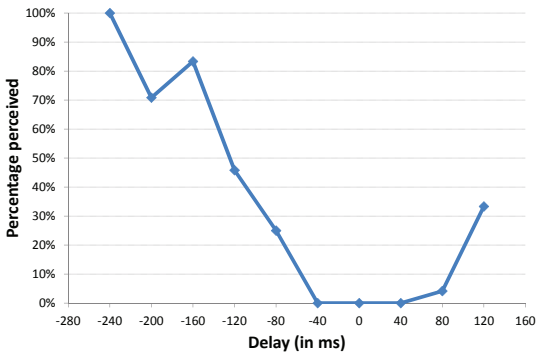

Fig. 9 Percentage of non-expert viewers who perceived lip sync issues compared to the actual inserted delay.

The graph clearly shows that delays up to one video frame (= [-40 ms , 40 ms]) are not detected at all. This also corresponds with the A/V synchronization thresholds recommended by the ITU [13], the ATSC [1] and the DSL Forum [4]. Furthermore, when the audio is delayed by 240 ms compared to the video signal, all subjects also detected the desynchronization. The detection threshold shows more or less a linear behavior

with respect to the actual inserted delay. As can be seen in Figure 9, a delay of -160 ms is slightly more detected than a delay of -200 ms. However, based on the statistical Z-test, we found that there is no statistical difference between the percentages of the subjects who perceived the delays of -160 ms and -200 ms. In case the audio is 120 ms ahead of the video signal, only 33% of the subjects detect that lip sync is out of sync. This implies that the audio can lead the video with more than 120 ms difference. Corresponding with the results in [28], delays up to two video frames (= [-80 ms, 80 ms]) are only detected by a small amount of subjects. An interesting remark is that audio/video desynchronization is apparently less detected when the audio is ahead of the video which was also concluded by Mued *et al.* [21].

Comparing the visibility of lip sync between the interpreters (Figure 6) and the average end-users (Figure 9) highlights the importance of the primary focus, similar to the results obtained in [27]. Despite the fact that the interpreters were also asked to evaluate the A/V synchronization, performing the simultaneous translation requires all their attention.

In general, our results obtained using non-experts correspond much more with results from already conducted research.

4.2.2 Audiovisual quality ratings for sequences with audio/video synchronization delays

When inspecting the MOS scores given to the different video sequences, as depicted in Figure 10, we notice that delays up to two video frames are still rated perfect quality which corresponds with their corresponding visibility thresholds (see Figure 9). Furthermore, delays up to 120 ms are perceivable but not rated annoying (MOS > 4). These results are similar the different A/V synchronization thresholds proposed by ITU-R Rec. BT.1359 [11] and Steinmetz [28]. Test subjects also perceive delays of -240 ms as annoying.

In accordance with our findings in the previous section, audiovisual quality is rated slightly higher when the audio is ahead of the video signal. However, this is not a significant difference. Therefore, it cannot be assumed that sequences with audio ahead of video are indeed less annoying compared to the sequences in which the audio is delayed with respect to the video.

On average, individual quality ratings drop by 1.5, with a standard deviation of 0.3, in case a non-expert detects an A/V synchronization problem. This is a slightly higher drop compared to the interpreters because the non-experts are primarily focused on audiovisual quality evaluation.

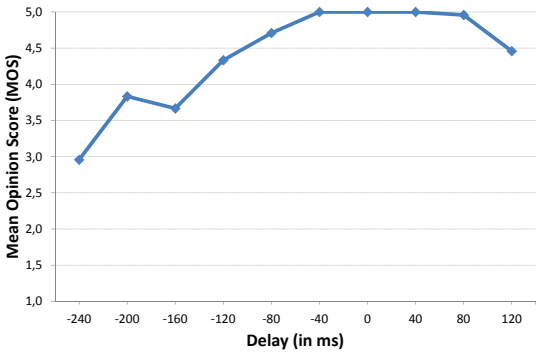

Fig. 10 MOS scores given by the non-experts to the sequences with inserted delay between audio and video.

4.2.3 Identifying whether the audio stream is delayed or ahead with respect to the corresponding video track

In case the test subjects perceived an A/V synchronization issue, they were also asked to indicate whether they perceived the audio was delayed with respect to the video track or vice versa. Figure 11 depicts the percentage of the subjects who correctly determined whether the audio was delayed or ahead of the video. Remark that only the results of subject who really perceived an A/V sync issue are taken into account.

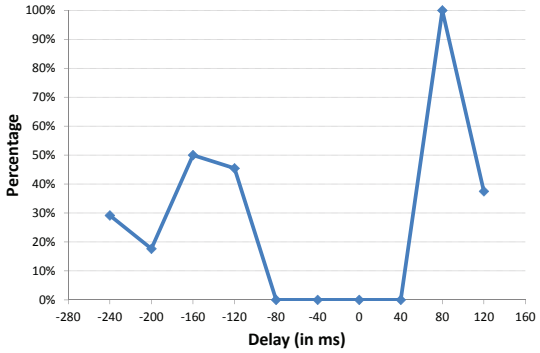

Fig. 11 Percentage of the non-experts who correctly determined whether audio was leading video or vice versa, in case they perceived A/V synchronization issues.

As the graph shows, it is difficult for the test subjects to determine the exact relationship between the audio and the video. Only a limited number of subjects are capable of correctly detecting whether the audio leads the video or vice versa. Even in the case of a delay of -240 ms, which is detected by 100% percent of the

test subjects (cfr. Figure 9), only 29% of the subjects correctly identified that the audio was delayed with respect to the video track. In case of a delay of 80 ms, the plot shows that 100% of the subjects correctly classified the relationship between the video and the audio track. However, it must be noted that only 1 subject detected the A/V sync issue in this case. In general, A/V sync becomes noticeable when the delay is more than 120 ms (in both directions).

As such, similar to the evaluations done by the interpreters, the non-experts also fail in identifying the direction of the differential delay between audio and video.

5 Conclusions

As indicated during the pre-experimental interviews, visual cues are only of secondary importance to the interpreters. Having a challenging task to complete, as experienced by several of the expert users, interpreters are primarily focused on performing the simultaneous translation. As such, detecting A/V desynchronization while interpreting a conversation poses a great challenge to most of our test subjects. Consequently, the majority of the interpreters do not perceive lip sync problems when the differential delay between audio and video remains below 240 ms. This detection threshold is significantly higher compared to the thresholds recommended by the different standard bodies and already conducted research (see Figure 1).

Both the experimental data and the post-experimental interviews suggest a low importance of desynchronized audio/video during simultaneous translation. Desynchronization seems to amplify existing difficulties, rather than causing difficulties by itself.

Despite the low detection rate and the high MOS scores, only a minority considers A/V synchronization important. Underlying this contradictory finding is the expectation that desynchronized audio and video will hamper the task of the interpreter eventually.

Conducting the same subjective experiment using non-experts highlights the importance of the primary focus. It is clear that lip sync is much easier detected when subjects are actively evaluating the audiovisual quality of the video sequences.

In contrast with the research findings from the interpreters, the results concerning lip sync visibility and acceptability obtained from our non-experts correspond with the results from already conducted subjective studies and with the recommendations from different standard bodies.

These differences, both in visibility and acceptability thresholds between the interpreters (experts) and non-experts, highlight the importance of considering the targeted application and use case when determining and investigating appropriate A/V synchronization thresholds.

Acknowledgements The research activities that have been described in this paper were funded by Ghent University, the Interdisciplinary Institute for Broadband Technology (IBBT) and the Institute for the Promotion of Innovation by Science and Technology in Flanders (IWT). This paper is the result of research carried out as part of the OMUS project funded by the IBBT. OMUS is being carried out by a consortium of the industrial partners: Technicolor, Televic, Streamovations and Excentis in cooperation with the IBBT research groups: IBCN & MultimediaLab & WiCa (UGent), SMIT (VUB), PATS (UA) and COSIC (KUL).

Glenn Van Wallendael and Jan De Cock would also like to thank the Institute for the Promotion of Innovation through Science and Technology in Flanders for financially supporting their Ph.D. and postdoctoral grant, respectively.

The authors would also like to thank Dr. Bart Defrancq, Lecturer and Coordinator of the PP in Conference Interpreting at University College Ghent, for his contributions to this work and support in acquiring the expert test subjects.

References

1. ATSC IS-191: Relative timing of sound and vision for broadcast operations (2003)
2. von Békésy, G.: A new audiometer. Acta Otolaryngologica **35**, 411–422 (1947)
3. Curcio, I.D., Lundan, M.: Human perception of lip synchronization in mobile environment. International Symposium on A World of Wireless, Mobile and Multimedia Networks **0**, 1–7 (2007)
4. DSL Forum Technical Report TR-126: Triple-play Services Quality of Experience (QoE) requirements. DSL Forum (2006)
5. EBU Recommendation R37: The relative timing of the sound and vision components of a television signal (2007)
6. Firestone, S., Ramalingam, T., Fry, S.: Voice and Video Conferencing Fundamentals, chap. 7, pp. 223–255. Cisco Press (2007)
7. Ford, C., McFarland, M., Ingram, W., Hanes, S., Pinson, M., Webster, A., Anderson, K.: Multimedia synchronization study. Tech. rep., National Telecommunications and Information Administration (NTIA), Institute for Telecommunication Sciences (ITS) (2009)
8. Fridrich, J., Goljan, M.: Robust hash functions for digital watermarking. In: International Conference on Information Technology: Coding and Computing (ITCC) (2000)
9. Garcia, M.N., Schleicher, R., Raake, A.: Impairment-factor-based audiovisual quality model for iptv: Influence of video resolution, degradation type, and content type. EURASIP J. Image and Video Processing **2011** (2011)
10. Huynh-Thu, Q., Garcia, M.N., Speranza, F., Corriveau, P., Raake, A.: Study of rating scales for subjective quality assessment of high-definition video. IEEE Transactions on Broadcasting **57**(1), 1–14 (2011)
11. ITU-R Recommendation BT.1359: Relative timing of sound and vision for broadcasting (1998)
12. ITU-R Recommendation BT.500: Methodology for the subjective assessment of the quality of television pictures (2009)

13. ITU-T Recommendation G.1080: Quality of Experience requirements for IPTV services. International Telecommunication Union (ITU) (2008)
14. ITU-T Recommendation J.148: Requirements for an objective perceptual multimedia quality model. International Telecommunication Union (ITU) (2003)
15. ITU-T Recommendation P.10/G.100 Amd 2: Vocabulary for performance and quality of service (2008)
16. ITU-T Recommendation P.911: Subjective audiovisual quality assessment methods for multimedia applications. International Telecommunication Union (ITU) (1998)
17. ITU-T Recommendation P.920: Interactive test methods for audiovisual communications. International Telecommunication Union (ITU) (2000)
18. Kudrle, S., Proulx, M., Carrières, P., Lopez, M.: Fingerprinting for solving A/V synchronization issues within broadcast environments. SMPTE Motion Imaging Journal pp. 47–57 (2011)
19. Mued, L., Lines, B., Furnell, S., Reynolds, P.: Acoustic speech to lip feature mapping for multimedia applications. In: 3rd International Symposium on Image and Signal Processing and Analysis (ISPA), pp. 829–832 (2003)
20. Mued, L., Lines, B., Furnell, S., Reynolds, P.: The effects of lip synchronization in ip conferencing. In: International Conference on Visual Information Engineering, pp. 210 – 213 (2003)
21. Mued, L., Lines, B., Furnell, S., Reynolds, P.: The effects of audio and video correlation and lip synchronization. Campus-Wide Information Systems **20**, 159–166 (September 2003)
22. Nezveda, M., Buchinger, S., Robitza, W., Hotop, E., Hummelbrunner, P., Hlavacs, H.: Test persons for subjective video quality testing: Experts or non-experts? In: QoEMCS workshop at the EuroITV - 8th European Conference on Interactive TV (2010)
23. Pereira, F., Alpert, T.: MPEG-4 video subjective test procedures and results. IEEE Transactions on Circuits and Systems for Video Technology **7**(1), 32–51 (1997)
24. Pinson, M.H.: Impact of lab effects and environment on audiovisual quality. VQEG_MM_2011_010_audiovisual quality repeatability, Seoul, South-Korea (2011)
25. Radhakrishnan, R., Terry, K., Bauer, C.: Audio and video signatures for synchronization. In: IEEE International Conference on Multimedia and Expo (ICME), pp. 1549–1552 (2008)
26. Speranza, F., Poulin, F., Renaud, R., Caron, M., Dupras, J.: Objective and subjective quality assessment with expert and non-expert viewers. In: Second International Workshop on Quality of Multimedia Experience (QoMEX), pp. 46–51 (2010)
27. Staelens, N., Moens, S., Van den Broeck, W., Mariën, I., Vermeulen, B., Lambert, P., Van de Walle, R., Demeester, P.: Assessing Quality of Experience of IPTV and Video on Demand Services in Real-Life Environments. IEEE Transactions on Broadcasting **56**(4), 458–466 (2010)
28. Steinmetz, R.: Human perception of jitter and media synchronization. IEEE Journal on Selected Areas in Communications **14**(1), 61 –72 (1996)
29. Stojancic, M., Eakins, D.: Interoperable AV sync systems in the SMPTE 22TV Lip Sync AHG: content-fingerprinting-based audio-video synchronization. SMPTE Motion Imaging Journal pp. 47–57 (2011)
30. Terry, K., Radhakrishnan, R.: Detection and correction of lip-sync errors using audio and video fingerprints. In: SMPTE Annual Tech Conference and Expo (2009)
31. Winkler, S.: Digital Video Quality - Vision Models and Metrics. John Wiley & Sons (2005)
32. Younkin, A., Corriveau, P.: Determining the amount of audio-video synchronization errors perceptible to the average end-user. IEEE Transactions on Broadcasting **54**(3), 623–627 (2008)