Multimed Tools Appl (2015) 74:345–365
DOI 10.1007/s11042-014-2136-6

Audiovisual robustness: exploring perceptual tolerance to asynchrony and quality distortion

Ragnhild Eg · Carsten Griwodz · Pål Halvorsen · Dawn Behne

Received: 2 April 2013 / Revised: 2 May 2014 / Accepted: 29 May 2014 / Published online: 22 August 2014

Abstract Rules-of-thumb for noticeable and detrimental asynchrony between audio and video streams have long since been established from the contributions of several studies. Although these studies share similar findings, none have made any discernible assumptions regarding audio and video quality. Considering the use of active adaptation in present and upcoming streaming systems, audio and video will continue to be delivered in separate streams; consequently, the assumption that the rules-of-thumb hold independent of quality needs to be challenged. To put this assumption to the test, we focus on the detection, not the appraisal, of asynchrony at different levels of distortion. Cognitive psychologists use the term temporal integration to describe the failure to detect asynchrony. The term refers to a perceptual process with an inherent buffer for short asynchronies, where corresponding auditory and visual signals are merged into one experience. Accordingly, this paper discusses relevant causes and concerns with regards to asynchrony, it introduces research on audiovisual perception, and it moves on to explore the impact of audio and video quality on the temporal integration of different audiovisual events. Three content types are explored, speech from a news broadcast, music presented by a drummer, and physical action in the

R. Eg (✉)
Department of Psychology, University of Oslo, Norway and Simula Research Laboratory, Oslo, Norway
e-mail: rage@simula.no

C. Griwodz · P. Halvorsen
Department of Informatics, University of Oslo, Norway and Simula Research Laboratory, Oslo, Norway

C. Griwodz
e-mail: griff@simula.no

P. Halvorsen
e-mail: paalh@simula.no

D. Behne
Department of Psychology, Norwegian University of Science and Technology, Oslo, Norway
e-mail: dawn.behne@svt.ntnu.no

form of a chess game. Within these contexts, we found temporal integration to be very robust to quality discrepancies between the two modalities. In fact, asynchrony detection thresholds varied considerably more between the different content than they did between distortion levels. Nevertheless, our findings indicate that the assumption concerning the independence of asynchrony and audiovisual quality may have to be reconsidered.

Keywords Audiovisual asynchrony · Temporal integration · Perceived quality · Multimedia streaming · Active adaptation

1 Introduction

A considerable body of work on synchronisation in audiovisual streaming originated several years ago [22], and rules-of-thumb for non-noticable and acceptable asynchronies were established at the time [34]. As multiplexing became the predominant solution for audiovisual streaming based on MPEG-1 and MPEG-2, asynchrony became less relevant as a research topic, and as an industry challenge. However, recent streaming approaches have turned away from multiplexing, mainly due to the introduction of multiple audio streams.

For example, video-on-demand systems that use adaptive HTTP streaming (“DASH-like” systems) can save both storage and bandwidth resources by storing and delivering separate movie soundtracks for different languages. RTP-based conferencing and telephony systems keep audio and video streams apart to simplify the independent adaptation decisions relating to codec, compression strength or resilience. Since audio and video streams in immersive systems are likely to originate from different devices, multiplexing would only complicate the process. Consequently, the current consensus in multimedia streaming is to implement congestion control and active adaptation, and to avoid packet drop, thereby staying within acceptable resource limits [7, 20, 27, 35].

While the flexibility of active adaptation alleviates problems related to buffering, the flipside is that on-time delivery will be prioritised over high quality. By reducing the frame-rate, the spatial resolution, or the quantisation parameters, lower quality streams can be transmitted in a timely manner. Active adaptation will therefore directly influence the objective auditory and visual quality, which brings forward related concerns for the subjective experience of quality. Human audiovisual integration may not be independent of quality, and neither may the existing rules-of-thumb for acceptable audiovisual asynchrony. An example is used to illustrate the motivation for this investigation:

Long router queues can contribute to noticeable asynchrony due to sporadic and severe application-layer delays (such as “stalls” and ”hiccups” in the media playout) that arise when audio is streamed over TCP [30]; adaptive video, on the other hand, is less affected. Buffering can compensate for some of the retransmission delays and recipient systems are in turn able to adjust the resulting asynchrony. Alternatively, quality reductions may reduce congestion in itself and thus better maintain synchrony. The latter approach results in loss of audiovisual quality and possibly perceptual information, it is therefore important to understand whether synchronisation demands could be less stringent when audiovisual quality is reduced.

Such an investigation should be pursued in two steps: The first step should assess whether reduced audiovisual quality affects the subjective detection of temporal asynchrony. The second step should follow only if the first step shows that signal quality affects asynchrony detection; this step should focus on the potential interaction between audiovisual asynchrony and quality distortion and assess how it may affect the quality of experience.

By applying global distortions to either the auditory or the visual modality, this paper explores the first step for three content types, news, drums, and chess. The presented results demonstrate that the perception of synchrony for these events is not greatly affected by audiovisual quality, not until the amount of impairment renders the signals almost unintelligible. We build the investigation on a body of work from both computer science and cognitive psychology. The aim is to establish acceptable audiovisual asynchrony levels and determine how they may vary depending on conditions. The investigation builds on an understanding of how temporal offsets affect perception and the integration of the human senses.

Although the perception of synchrony did not always vary consistently across quality distortions, the results still suggest that the assumption concerning the independence of asynchrony and audiovisual quality does not hold. Nevertheless, temporal perception is very robust and the nature of the audiovisual content appears to contribute to more variation than the quality. While our findings highlight the tolerance of the perceptual system, they are also specific to the selected content and applied distortions. Further explorations into temporal integration across a wider selection of audiovisual events and relevant streaming artifacts may be required before proceeding to the second step.

2 Temporal integration and quality distortion

When some event evokes simultaneous signals from two or more modalities, attention is more quickly and more accurately directed towards the source [33], and the experience of the event itself is enhanced [32]. The perceptual process that combines and creates coherence out of different sensory inputs is referred to as multisensory integration. This investigation will limit itself to audiovisual integration as they are the two modalities relevant to the majority of today's multimedia systems. Audiovisual integration aids our understanding of speech [9], assists us in temporal and spatial judgements [3], and contributes to richer experiences [12]. The integration of our senses is a complex perceptual process that relies on the convergence of information in several dimensions. In the spatial dimension, audiovisual integration is demonstrated through the ventriloquist effect - the tendency for a visual source to capture the auditory signal so that they appear to originate from the same location [18]. In the temporal dimension, the perception of synchrony between a visual and an auditory event [2] demonstrates integration between the modalities. To establish how much separation the senses can endure while remaining integrated, researchers typically attempt to establish at which thresholds the bond will break [14]. By separating an auditory and a visual event in time or space, at varying increments, there will eventually come a point where the two are no longer perceived as an entity [8].

Temporal offsets of different magnitudes are introduced as part of a methodology used to establish the perception of audiovisual synchrony. In this way, temporal integration can be explored for simple stimuli such as lights and tones [6], but also for more complex scenarios such as speech [42]. Several studies have looked at the variations in temporal integration for different events, supplying new stock to an ongoing debate on the special nature of speech. A majority of the studies has found the perception of synchrony in speech to be more tolerant to temporal offsets compared to isolated actions that typically have anticipatory moments of impact, such as a hammer hitting a peg [8], or other encounters between tools and objects [28, 36]. The robustness of intelligible speech has also been demonstrated in comparison to monkey calls [39] and non-native languages [29]. However, the temporal tolerance for speech is not unique; when compared to musical notes played on guitar or piano, asynchrony is detected sooner in speech stimuli [37]. Research on temporal integration requires audiovisual events where the sound clearly corresponds to a visible action. Most standard content used in QoE experiments are therefore inappropriate for the evaluation of perceived synchrony. Even for a sports segment or a movie scene, the movements may be too rapid or the action set too far away for the audiovisual relation to be discerned. Instead of using standard content, the current study includes three audiovisual sequences that correspond to three different types of events, speech, action, and music. To make sure the sequences represent frequently encountered media content, we selected each content from a popular multimedia arena.

Research in the area of temporal integration is plagued with a variety of approximations to temporal thresholds of perceived synchrony. This makes it all but impossible to conclude whether the variations are due to actual differences between event types or simply reflect the different measures and statistical approaches. While some experiments implement a temporal order judgement task, asking participants to determine which signal precedes the other [39], others rely on a simultaneity judgement task that requires assessments on perceived synchrony [5]. Still others ask for the detection of gradually introduced asynchrony [8], or discrimination between presentations with different temporal offsets [26]. Figure 1 provides a summary of previously published thresholds corresponding to perceived synchrony for different event types, but also established using different measures. These measures may tap into slightly different perceptual processes [40], yet the method of choice for this investigation fell on the simultaneity judgement task. This choice was based on the relative ease of synchrony/asynchrony evaluations compared to temporal order judgements. Moreover, between the two methodologies, simultaneity judgements tend to yield more stable and ecologically valid points of subjective simultaneity [11].

As mentioned, synchrony between adaptive streams can be better maintained by lowering the streaming quality. However, the perceptual consequences of such a trade-off are largely unexplored. One exception is frame rate, for which studies have found adverse effects on the temporal integration of audio and video [19, 38]. Since reduced frame rates affect the temporal nature of the video, they also affect the temporal integration of the audiovisual signals. Specifically, the subjective point of perceived synchrony is shifted with low frame rates, demanding that audio must precede video with up to 42 ms to achieve experienced synchrony [38]. Although the chance of congestion is reduced by lower frame rates, there are further reasons why this approach is seldom an attractive alternative. For instance, low frame rates have a negative impact on users' subjective enjoyment [15], and they affect users' measured heart rate [43], as well as their blood flow and skin responses [44]. Thus, active adaptation schemes should avoid implementations that affect the temporality of the transmitted signals. The commonality between the two remaining alternatives is that reductions to both the resolution and to the quantisation parameters will result in loss of spatial detail. While the implications of losing fine-grained details have not been explored for temporal integration, the effects are well-established for other perceptual processes. Not surprisingly, severe loss of visual details can lead to difficulties in recognising both people [4] and text [25]. Moreover, the auditory modality will eventually dominate speech perception when visual information is distorted [10, 24]. Conversely, noise and auditory compression artifacts can mask essential spectral frequencies and lead to distracting audio effects [23]; in speech perception, reduced audio quality typically leads to an increased dependency on the visual modality [13]. Clearly, severe quality drops will make information from either modality ambiguous; however, within the boundaries of what is intelligible, temporal integration may not be equally affected.

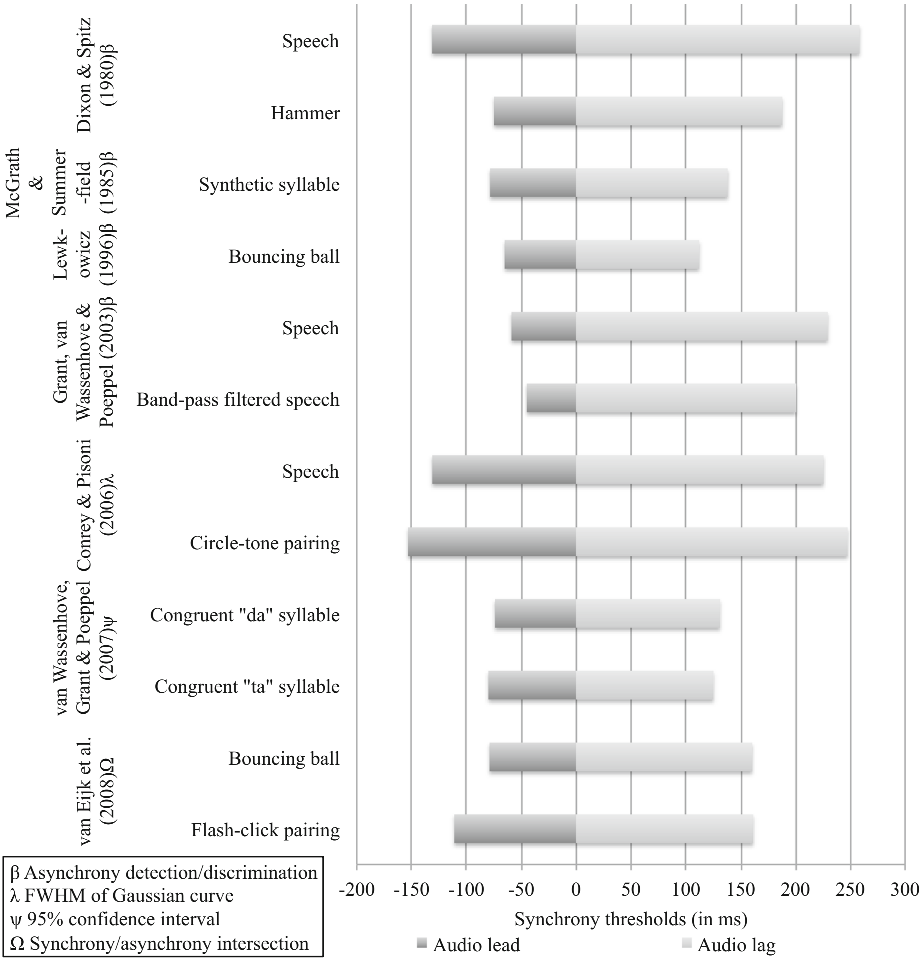


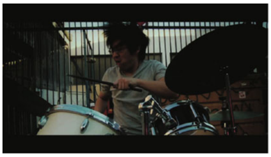
Fig. 1 Overview of previously published thresholds of perceived synchrony, illustrating the different measures and event types applied to the study of temporal integration. Where some ask participants to indicate the point at which an auditory and a visual stream are no longer perceived as synchronous [8, 14, 21, 26], others use simultaneity judgements and curve fittings to calculate the full-width at half-maximum (FWHM) [5] or the 95 % confidence interval [42]. Others still use temporal order judgements and calculate the slopes of cumulative distribution functions [11]

With new multimedia platforms come new approaches to efficient streaming, and with them come old concerns. Multiplexing is no longer always the best approach [27]; some situations will require the audio and video streams to remain separate, for instance in two-way communication systems, or when there are several audio tracks to consider [35]. While active adaptation allows for multiple streams, it introduces the potential of co-occurring asynchrony and quality distortions. Considering the detrimental effects of quality distortions on several perceptual processes, along with the limited tolerance of temporal integration, we set out to explore the possibility that media quality could influence the experience of synchrony. Although video quality drops are far more likely than audio distortions, we considered both scenarios in order to assess whether the two modalities are equally important in temporal integration. We also included three different content types to gain a broader understanding of the temporal perception of audiovisual events.

3 Experimental set-up

Due to the lack of standardised content suitable for synchrony evaluations, we selected experimental stimuli from three popular entertainment services, TV, movie, and YouTube. Selection criteria included clear correspondence between auditory and visual events and good representation with respect to content type. Detailed descriptions of the speech, action, and music content are included in Table 1. The duration of all sequences was set to 13 seconds, determined by the time required to keep the coherence of the speech excerpt. Furthermore, care was taken to find sequences that kept the event of interest in focus, so that no distractions would turn attention away from the speaker, the chess piece, or the drummer. Because audio intensity varied between the videos, all were normalised to an average intensity of 70 dB. The video resolution was established from the sequence with the lowest

Table 1 Content description of the sequences included in the experiment

Content 1: Speech	Content 2: Chess	Content 3: Drums
		
The Norwegian news broadcast shows a seated anchor-woman filmed from the chest up. She presents a piece of news on the return of two football players who have been out of the game due to injuries. The scene composition remains the same throughout the 13 seconds. The broadcast was provided by the National Library of Norway, with permission to use for research purposes.	The video portrays a game of chess played by two young men in a Renaissance setting. The scene opens with only the chessboard in view and zooms out to gradually include the two players and some of the surroundings. Five pieces are moved during the selected 13 seconds. The sequence was retrieved from the movie Assassin's Creed: Lineage (Part 1), with permission from Ubisoft.	A young man introduces the music video by hitting his drumsticks together three times before commencing to play the drums. The camera zooms out to include the alley where he sits. The 13-second-long sequence concludes with the appearance of the drummer's clones playing bass-guitars. The video was produced by Freddie Wong and Brandon Laatsch for the freddiew channel.

resolution, 1024 × 576 pixels. By applying distortions that were both global and regular, we eliminated uncontrolled variations due to irregularly occurring artifacts. Pink noise added to the auditory signals and gaussian blur filters applied to the video signals met our criteria. While these distortions may be less relevant from an encoding perspective, they provide uniform masking of the signals, ensuring that all auditory and visual events are equally affected. Provided that all artifacts can be said to remove or mask sensory information, the perceptual effects of noise and blur should be generalisable to more common encoding and compression artifacts.

A range of blur levels were tried out and narrowed down in a pilot study, resulting in the final range used in all experiments, with blur filters set to 2 × 2, 4 × 4 or 6 × 6 pixels. In addition, an undistorted stimuli level was kept at the 1024 × 576 pixel resolution. Audio distortion levels were determined based on results from the quality relation experiment (Section 4) and a subsequent pilot study. In the end, these included the no-noise condition, and noise levels at 60 dB, 70 dB, and 80 dB. The videos were displayed on iMac 11.3 computers running the Superlab software with audio presented through AKG K271 circumaural headphones. The displays were set to iMac standards with full brightness and output audio intensity was kept at an average of 70 dB.

4 Quality relation experiment

The main purpose of the first experiment was to establish levels of auditory and visual distortions to implement in the temporal integration experiments. The hypotheses are therefore addressed by the two following experiments, while this first correspondence experiment builds the foundation for the investigation. Seeing how different severities of pink noise and gaussian blur provide no common ground for comparisons, this experiment was designed to explore subjective relations between distortions for the two modalities. By obtaining a measure on the perceived correspondence in severity of the auditory and visual distortions, we also wanted to explore whether the loss of information could be more detrimental for one modality over the other. The incentive was to establish three corresponding distortion levels for each modality to be carried through to the two subsequent experiments. Only the audio distortion levels were selected in this manner due to the experimental set-up; video distortion levels were chosen as best-fits to previous research [10, 24], with their appropriateness tested in comparison to audio distortion.

4.1 Method

4.1.1 Stimuli and procedure

The video sequences were presented in four simultaneous video quality versions on-screen (blur levels described in Section 3, randomly designated to one of four screen locations. Figure 2 shows an example of such a presentation for the drums video. The corresponding audio track was presented through headphones, with a random auditory distortion level. This included audio presented without noise, or with pink noise at 55 dB, 60dB, 65 dB, 70 dB, 75 dB, or 80 dB. Each blur level was presented twice in the same location for each noise level, so that the same condition was repeated eight times for every content. In total, this resulted in 168 trials, with 56 repetitions of each video content. Participants were asked to consider the perceived quality of the audio and select the video quality they deemed to be the best subjective match in distortion severity. The randomised quality versions were

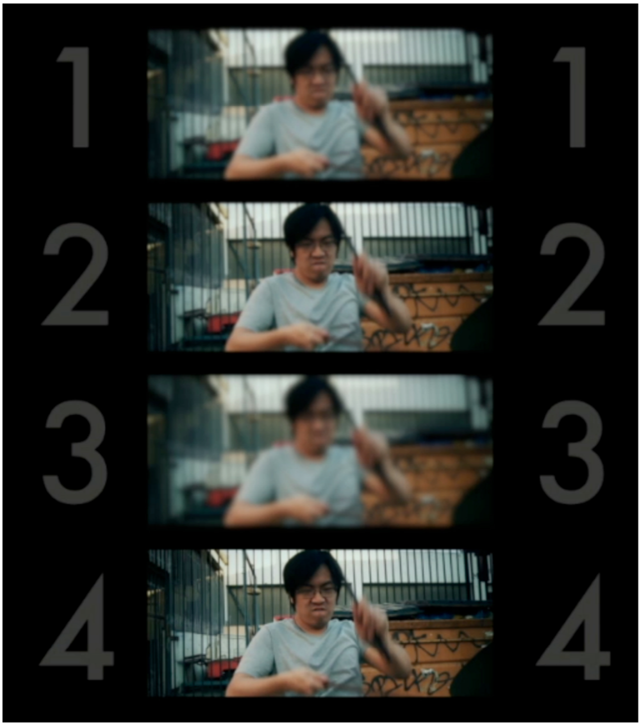


Fig. 2 Illustration of the drums sequence presented with all four blur distortion levels simultaneously

labelled with numbers from 1 to 4 and responses were made by selecting the same number on the keyboard. Responses could be given at any time during the stimulus presentation, thus the total duration of the experiment varied between 40 and 70 minutes.

4.1.2 Participants

With ages spanning from 19 to 29 (*mean* = 24), a total of 21 participants were recruited from the Norwegian University of Science and Technology. All were informed of their rights as volunteers and gave their consent, and all received a gift card upon completion of the test.

4.2 Results and discussion

The perceived correspondence between the presented auditory distortion level with the chosen visual distortion level was initially evaluated from the mean distribution of responses, averaged across participants. Considering the greatly similar distributions for the three video sequences, responses were later collapsed across content. The resulting averages were compared using matched-sample t-tests, which pair participants' scores across conditions and indicate whether the variation is large enough to be statistically significant [16]. Taking basis in the noise distortion level with the highest rate of responses corresponding to a specific blur distortion level, the t-tests compared this best match to all other noise distortion

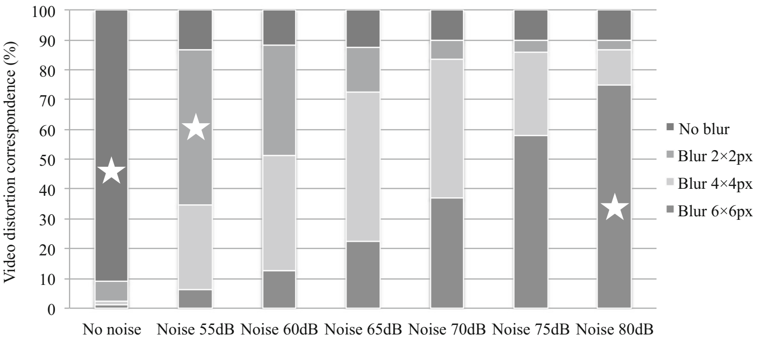


Fig. 3 Response distribution on video distortion levels perceived to best correspond to the presented audio distortion level. *Stars* indicate significantly higher rates of blur level responses for the indicated noise level, as compared to all other noise levels

levels. The results, presented in Fig. 3, thus give an indication on the best perceptual correspondence in distortion severity between presented pink noise levels and selected gaussian blur levels.

In this experiment, the methodology is limited to the presented range of noise and blur distortion levels. The results therefore have to be interpreted within the confinement of the available options; with a larger selection of blur distortion levels, the subjective selections may well have been spread out more in either direction. With that note of caution, the results do serve to highlight the best match in severity between the presented distortion levels. Not surprisingly, the auditory and visual tracks presented without distortion were judged to be correspond very well, with over 90 % match. The 2 × 2 pixel blur distortion was deemed the best correspondence to 55 dB pink noise with match responses above 50 %, but reports were fairly high at nearly 40 % also for 60 dB noise. No clear trend was evident for 4 × 4 pixel blur, with match responses distributed fairly evenly across 65 dB and 70 dB noise, at 50 % and 46 % respectively. At 75 %, the subjective match between 80 dB noise and 6 × 6 pixel blur was far clearer.

Within its confinement, this experiment presents a good correspondence between auditory and visual distortion at the lower levels, as well as between the higher levels. Following these results, audio distortions for the subsequent experiments were set to 60 dB, 70 dB, and 80 dB. We first established 80 dB pink noise as the most severe noise level, since it was the significantly best match to the most severe blur level at 6 × 6 pixels; moreover, the two were also the overall best corresponding distortion levels. The two remaining auditory distortion levels were chosen at decrements from this point. While 55 dB noise was judged to be the best subjective match to 2 × 2 pixel blur, our pilot study had indicated insignificant perceptual consequences from its masking effect, so we decided on the second-best match instead. The two lower noise levels were thus set to 60 dB and 70 dB.

5 Temporal integration experiments

To investigate the combined effect of distortion and asynchrony on the temporal integration of audio and video, two experiments were planned and carried out. One explored the perceptual consequences of visual distortion, while the other focused on auditory distortion. Potential differences in temporal integration for the three content types were also considered.

5.1 Method

5.1.1 Stimuli and procedure

The temporal integration experiments followed a standard simultaneity judgement methodology from cognitive psychology, outlined in [41]. Audiovisual sequences were presented successively with a short response period in-between, consistent with the ”Absolute Category Rating” method recommended by the ITU-T [17]. Asynchrony was determined by temporal offsets that varied between presentations, and participants were asked to judge whether the audio and video appeared to be synchronous or asynchronous. Initially, the plan was to implement the content described in Section 3 for the two synchrony experiments; however, participant feedback necessitated an adjustment following the visual distortion experiment. The drums video was not included in the auditory distortion experiment due to reported difficulties in perceiving the temporal events within the clip; these difficulties were deemed to contribute to unnecessarily high cognitive loads.

Asynchrony levels were manipulated by shifting the audio track relative to the video track, including more of the original audio at either the beginning or the end of the sequence. The video onsets and offsets remained the same for all stimuli, to avoid giving away visual timing cues. Initial asynchrony levels were based on a body of research (summarised in Fig. 1) that has contributed with a range of thresholds for perceived synchrony, but also a general consensus on the greater perceptual sensitivity to auditory signals preceding visual signals. Given the wide spread of previously published thresholds, we carried out a pilot study to establish appropriate asynchrony levels for our dynamic video content. Thus, stimuli presentations included synchronous audio and video, and audio tracks leading before video tracks (audio lead) with 50 ms, 100 ms, 150 ms, and 200 ms, as well as audio tracks lagging behind video tracks (audio lag) with 100 ms, 200 ms, 300 ms, and 400 ms.

For the visual distortion experiment, the blur levels were the same as described in Sections 3 and 4. Similarly, the auditory distortion experiment implemented the noise levels established from the quality relation experiment, pink noise at 60 dB, 70 dB, and 80 dB, along with a no-noise condition. For the visual distortion experiment, the different content, asynchronies, and distortion levels added up to 108 stimulus conditions. With each condition repeated twice, the total number of presentations came to 216. For the auditory distortion experiment, the exclusion of one content resulted in 72 stimulus conditions and 144 trials. Stimuli were randomised separately for every participant.

Participants had to watch the sequences for the entire duration before being prompted for a response. They responded by pressing one of two buttons on a Cedrus RB-530 response-box, choosing either the ”synchronous” or ”asynchronous” alternative. The full experiment was typically completed in 60-70 minutes, including a break half-way through.

5.1.2 Participants

The 19 volunteers who participated in the visual distortion experiment were aged between 19 and 41 years (*mean* = 23), while the ages for the 25 participants in the auditory distortion experiment spanned from 18 to 33 years (*mean* = 24). All participants were recruited from

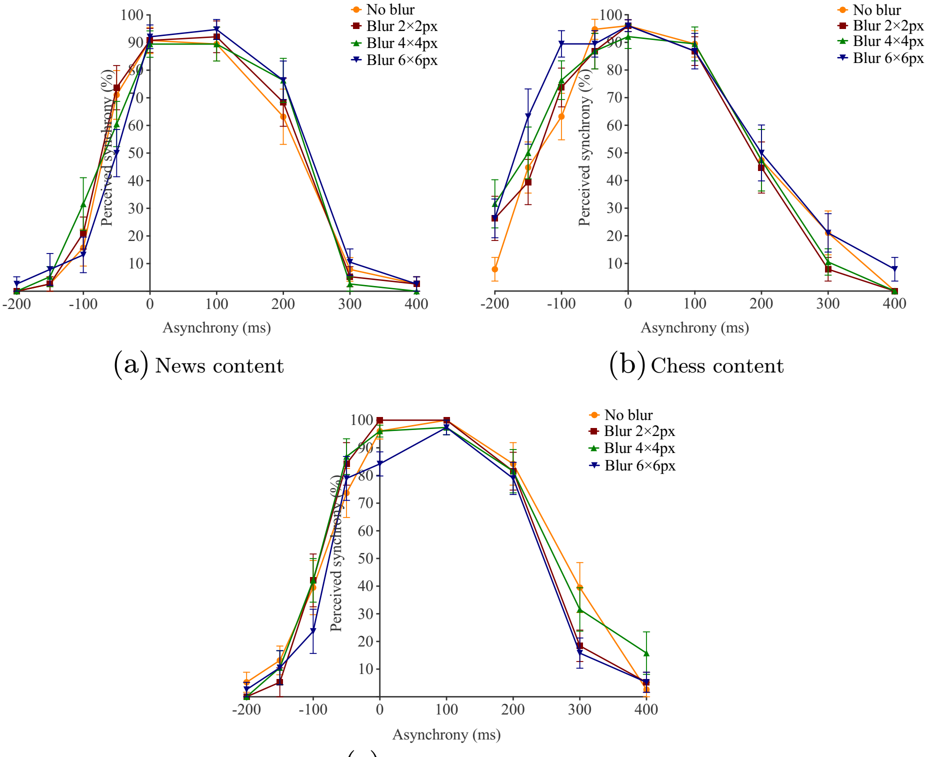
Table 2 Results from repeated-measures ANOVAs for visual distortion

	Speech			Chess			Drums		
	F	p	η^2_p	F	p	η^2_p	F	p	η^2_p
Asynchrony	144.49	< .001	.89	59.18	< .001	.77	134.49	< .001	.88
Blur	0.16	*ns*	.01	3.44	< .05	.16	3.43	< .05	.16
Asynch*Blur	1.49	*ns*	.08	1.45	*ns*	.08	1.24	*ns*	.06

the Norwegian University of Science and Technology, but none took part in both experiments. Participants were informed of their rights as volunteers and gave their consent prior to the test, and they received a gift card as compensation for their time.

5.2 Results and discussion

The temporal integration experiments explore the effects of distortion and asynchrony, along with their interaction, on the perception of synchrony. All participants' responses were converted to percentages and averaged across stimulus repetitions. These data were



(a) News content

(b) Chess content

(c) Drums content

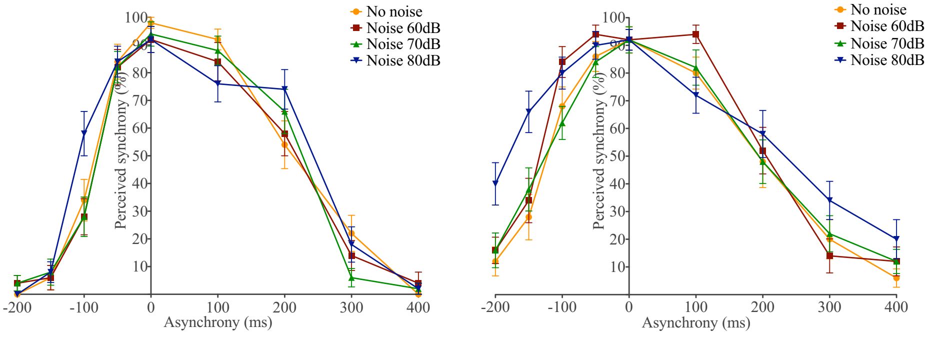
Fig. 4 Rates of perceived synchrony (in percent), averaged across participants and distributed over asynchrony levels, for the different levels of blur distortion. *Negative values* indicate audio lead asynchrony, while *positive values* represent audio lag asynchrony

Table 3 Results from repeated-measures ANOVAs for auditory distortion

	Speech			Chess		
	F	*p*	η_p^2	*F*	*p*	η_p^2
Asynchrony	122.75	< .001	.84	71.57	< .001	.75
Noise	2.13	*ns*	.08	8.50	< .001	.26
Asynch*Noise	2.75	< .001	.10	2.50	< .01	.09

then analysed separately for each experiment and each content. Repeated-measures analysis of variance (ANOVA) F-tests were calculated to contrast the variation in responses attributable to the introduced conditions with the variation attributable to randomness. This in turn determined the statistical significance of the differences between distortion and asynchrony levels, indicated by a probability less than 0.05 %, $p<$ .05 [16]. Effect sizes were calculated using the partial eta-squared (η_p^2) statistic [31]; this provided a measure, comparable across experiments, for the strength of the effect of asynchrony and/or distortion on reported synchrony. Differences between distortion levels were deduced from graphs plotted with standard errors of the mean illustrating the response variation. ANOVA results for the visual distortion experiment are presented in Table 2, and the distribution of average responses are illustrated in Fig. 4. For the auditory distortion experiment, the results from the repeated-measures ANOVAs are presented in Table 3 and distribution curves are plotted in Fig. 5.

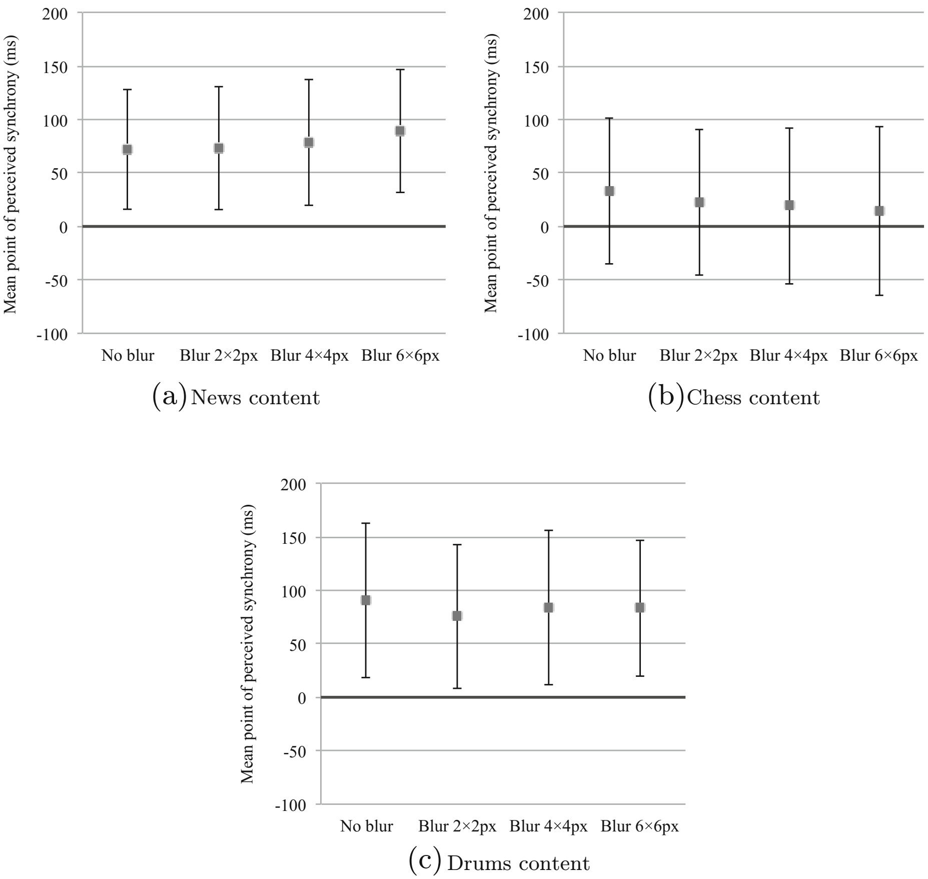
Following a common procedure for simultaneity judgement experiments [5], we performed Gaussian curve fittings to establish asynchrony detection thresholds. Curves were fitted with the overall averages for all distortion levels, distributed across asynchrony levels, which yielded four distributions for each content. Unlike earlier studies, this investigation sought to establish thresholds that can be applied to current media streaming solutions. The more traditional 50 % threshold for asynchrony detection, calculated using the full-width at half-maximum, was therefore traded with a more conservative measure. The mean points of



(a) News content

(b) Chess content

Fig. 5 Rates of perceived synchrony (in percent), averaged across participants and distributed over asynchrony levels, for the different levels of noise distortion. *Negative values* indicate audio lead asynchrony, while *positive values* represent audio lag asynchrony



(a) News content

(b) Chess content

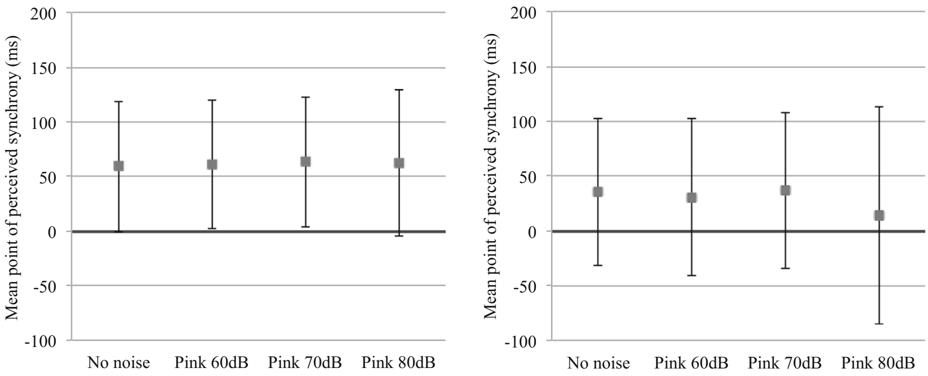
(c) Drums content

Fig. 6 Mean points of perceived synchrony (in milliseconds) for each level of blur distortion, grouped according to content. *Error bars* represent the windows of temporal integration as one standard deviation around the mean. *Negative values* indicate audio lead asynchrony, while *positive values* represent audio lag asynchrony

perceived synchrony are still represented by the means of the fitted curves, but the window of temporal integration is here defined as one standard deviation around the mean. Figure 6 presents the temporal thresholds for the visual distortion experiment, while the thresholds for the auditory distortion experiment are displayed in Fig. 7.

To further investigate potential differences in temporal thresholds between content, an individual analysis was carried out using only responses to the undistorted video presentations from the visual distortion experiment. Results from this analysis are presented in Table 4, Figs. 8, and 9.

As the distributions in Fig. 4, along with the main effects of asynchrony, exemplify, participants' synchrony responses spread out across the full range of perceived synchrony for all content. The effect of blur distortion was less clear-cut. No significant interaction with asynchrony was found, and only for the chess and drums content were main effects uncovered. Further inspection of the mean points of perceived synchrony and the windows of temporal integration revealed no consistent effect of blur, illustrated in Fig. 6. Hence,



(a) News content (b) Chess content

Fig. 7 Mean points of perceived synchrony (in milliseconds) for each level of noise distortion, grouped according to content. *Error bars* represent the windows of temporal integration as one standard deviation around the mean. *Negative values* indicate audio lead asynchrony, while *positive values* represent audio lag asynchrony

the results suggest that the temporal integration of our speech and action events is resilient to loss of visual details, to the point were only global motion features may be required for these visual events to be temporally aligned with the corresponding auditory signals.

Unlike visual blur, auditory distortion did contribute to interactions between asynchrony and noise. The distributions presented in Fig. 5 further support this notion, with more synchrony reported for 80 dB noise at the temporal offsets where asynchrony appeared more difficult to discern (around the 50 % mark). This tendency is especially prominent for the chess sequence. With a wider window of temporal integration, seen in Fig. 7, the most severe level of auditory distortion contributes to a more tolerant temporal integration of the chess event. By obscuring the auditory signal, the temporal events are likely to become less salient, making it more difficult to align the two modalities. Possibly, the top-down processing of speech may have served to alleviate this effect somewhat.

The difference in distributions for the three content types, Fig. 8, did yield both main effects and a significant interaction between asynchrony and content. Figure 8 portrays how the mean point of perceived synchrony is fairly close to objective synchrony at approximately 30 ms audio lag for the chess content, while for the drums content this point is extended to more than 90 ms audio lag. The skew in temporal tolerance to audio lead asynchrony observed for the chess content, as compared to news and drums, may be related to the visual angle and the slowness in the movement of the hand holding the chess piece. Thus, the bird's eye view shot at the beginning of the video sequence may have created

Table 4 Results from the repeated-measures ANOVA exploring content

	F	p	η_p^2
Asynchrony	113.69	$< .001$	.86
Content	7.35	$< .005$	.29
Asynch*Content	5.60	$< .001$	.24

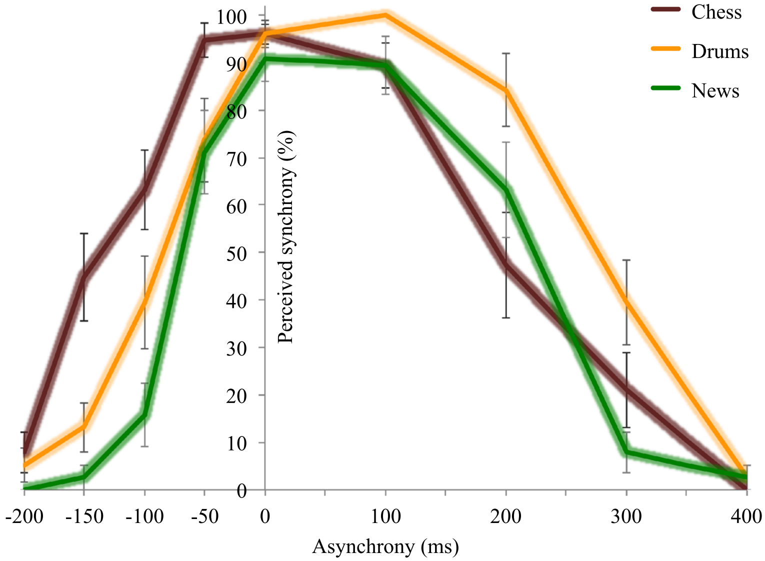
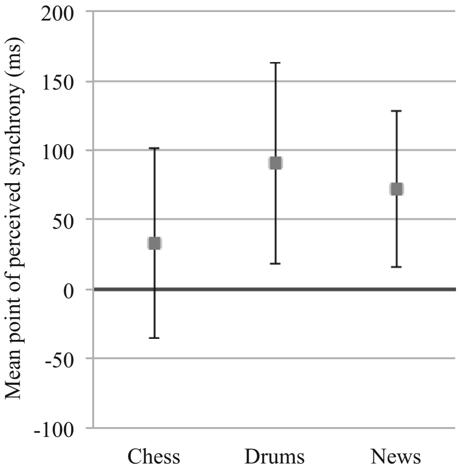


Fig. 8 Rates of perceived synchrony (in percent) derived from the zero blur distortion level, averaged across participants and distributed over asynchrony levels, for the different content. *Negative values* indicate audio lead asynchrony, while *positive values* represent audio lag asynchrony

uncertainty in the visual domain by covering up the moment of impact. The wider window of temporal integration observed for the drumming video, in comparison to the news content, suggests that the rapid movements of the drummer are harder to align across modalities than are the fluctuating articulations of the news broadcaster. Again, this may be a reflection of the top-down processes associated with speech, where experience aids the anticipation of the coming syllable.

Fig. 9 Mean points of perceived synchrony (in milliseconds) for each content, derived from the zero blur distortion level. *Error bars* represent the windows of temporal integration as one standard deviation around the mean. *Negative values* indicate audio lead asynchrony, while *positive values* represent audio lag asynchrony



The inconsistencies in temporal integration between such different content highlight the difficulties in establishing universal thresholds for acceptable asynchrony. However, a conservative approach is to take basis in the temporal thresholds closest to objective synchrony. Judging from the data collected in the two experiments, it seems prudent to avoid audio streams that lead before the video; moreover, video streams should preferably not lag by more than 100 ms. With respect to the more liberal windows of temporal integration presented in Fig. 1, this is indeed a cautious recommendation. Moreover, our findings are specific to three audiovisual sequences and future investigations may reveal different distributions for other content. Still, the presented windows of temporal integration are in agreement with past research on continuous and dynamic events [5, 8, 14]. Thus, our recommendations may err on the side of caution, but synchrony maintained between 0 ms and 100 ms audio lag should offer a safe margin for most scenarios.

6 Conclusion

With the use of multiple audio tracks becoming more common, an old challenge relating to the synchrony of streams is resurfacing. This investigation has revisited one of the assumptions behind the predominant rules-of-thumbs and found cause for concern. We ran two separate experiments, one with distorted audio and one with distorted video, to explore the interaction between asynchrony and media quality on the temporal integration of audio and video. Within the context of the selected content, the effects of audiovisual quality were not straightforward. Yet, the findings still demonstrate that the perception of synchrony can be influenced by the quality of both auditory and visual information.

Further conclusions cannot easily be drawn from the current dataset. This goes particularly for video quality, with isolated main effects of blur for two of the three content types, but no consistent variation in responses. Seeing how video quality affects the perception of synchrony for the chess and the drums videos in opposite directions, the only safe assumption is that there exists a co-dependence. As for audio quality, pink noise yielded both main effects and significant interactions with asynchrony, but again with slightly ambiguous response variations. Still, audio distortions appear to mask the temporal events of the auditory signal, particularly for the chess sequence, leading to a widening of the window of temporal integration. In other words, poor audio quality may make the perceptual system more tolerant to audiovisual asynchrony. In terms of multimedia streaming, the implications of an interactive effect between media quality and asynchrony may not be the greatest. Nevertheless, it remains a relevant consideration for active adaptation and other solutions that stream audio and video separately, and that adjust streaming quality according to the available bandwidth. If a downscaled and distorted video could make the perceptual system more sensitive to the delay of one stream, then it would be advisable to make sure that either quality or synchrony is maintained.

Furthermore, temporal integration varies between audiovisual events, similar to what has been noted by others [8, 36]. The three content studied have quite distinct distributions for perceived synchrony, with asynchrony detected sooner in speech than in either the chess or the drums sequence. From a cognitive perspective, this may well reflect the activation of top-down processes for speech. This again might imply that the anticipation building up when observing the culmination of a physical event cannot compete with the years of training most people possess when it comes to predicting the next syllable in a stream of speech [1]. With respect to media providers, the relative intolerance to asynchrony in our selected speech excerpt serves as a reminder for particular caution when it comes to preserving synchrony for this type of content.

The most prominent finding from this set of experiments is perhaps the robustness of the perceptual system, with the temporal integration of audio and video withstanding quite severe losses of auditory and visual information. This result suggests that neither fine-grained visual details, nor every nuance of the acoustical spectrum, is required for temporal events to align across the sensory systems. Temporal integration is certainly an important aspect to the perception of multimedia; however, perceived synchrony does not equal comprehension. As noted, the loss of visual information can put a burden on perception, making it harder to recognise people [4], text [25], and speech movements [10, 24]. Auditory distortions will similarly mask important acoustical information [23], making it difficult to understand speech sounds [13]. Consequently, this investigation demonstrates an impressive perceptual tolerance to quality loss, but only for the temporal domain. Human perception is likely less tolerant with regards to the loss of perceptual details necessary for understanding a message.

When it comes to the thresholds of perception's temporal tolerance, we recommend caution. Like previous works have shown [21, 26, 42], asynchrony can be detected at very short thresholds, likely depending on the nature of the audiovisual event. Our observations indicate that such caution would involve the avoidance of video stream delays entirely, since the windows of temporal integration for two of three content types do not extend to audio lead asynchrony, irrespective of distortion. For active adaptation and similar streaming solutions, this means that forced delays to the audio stream could be advisable for scenarios where the synchrony of streams is difficult to maintain. On the other hand, the auditory delay must remain moderate; at the most conservative, our findings suggest that audio lag asynchrony should not exceed 100 ms.

A window of temporal integration spanning only 100 ms provides a very small margin for maintaining perceived synchrony in adaptive multimedia streaming. With so little leeway, it would be desirable to improve the understanding of the factors that are at play. Given the difference between content, additional experiments are required to assess whether the detection of asynchrony will vary equally between other action, music and speech events. Further explorations should include more sequences to represent these categories, as well as additional content types to represent the most common broadcasted events, such as sports, song, and dialogue. More than that, the possibility that distortion caused by bitrate adaptation could influence the perception of synchrony should be further explored. This investigation considers global and regular distortions as approximations to more common auditory and visual artifacts. Although these distortions ensured experimental control, they may not equal the severity of the most common adaptation techniques. Future work should therefore consider the perceptual consequences of artifacts that arise due to downscaling approaches, such as MPEG-2, H.264 and HEVC encoding, or that result from transmission. For instance, excessive compression may cause blockiness or ringing, two highly visible artifacts that may be more detrimental to visual intelligibility than blurring. Moreover, data packets can be lost during transmission, which can result in a number of different artifacts; black squares may replace the missing data and jitter, judder, or jerkiness may be experienced because of missing frames. With the inclusion of applicable artifacts and a greater span of distortion levels, subsequent studies should reveal whether the temporal integration of audiovisual content is as robust as the presented results would indicate, or whether they are specific to the current context.

Finally, the effect of media quality on temporal integration may be moderate, but it may still have marked impact on the user experience. We therefore aim to follow through with the suggested second step and explore the relationship between asynchrony, distortion, and quality of experience.

Acknowledgments The authors would like to thank the editors and the anonymous reviewers for their helpful comments and suggestions. Thanks are also extended to Dr. Alexander Eichhorn and Dr. Knut Inge Fostervold, for their feedback along the way, to Camilla Hellum Foyn for her assistance, and to all the participants who volunteered their time. This work was completed as a contribution to Perceval, project number 193034, and was funded by the Research Council of Norway through the VERDIKT programme.

References

1. Alm M, Behne D (2013) Audio-visual speech experience with age influences perceived audio-visual asynchrony in speech. J Acoust Soc Am 134(4):3001–3010
2. Arrighi R, Alais D, Burr D (2006) Perceptual synchrony of audiovisual streams for natural and artificial motion sequences. J Vision 6(3):260–268
3. Bolognini N, Frassinetti F, Serino A, Làdavas E (2005) "Acoustical vision" of below threshold stimuli: interaction among spatially converging audiovisual inputs. Exp Brain Res 160(3):273–282
4. Burton AM, Wilson S, Cowan M, Bruce V (1999) Face recognition in poor-quality video: evidence from security surveillance. Psychol Sci 10(3):243–248
5. Conrey B, Pisoni DB (2006) Auditory-visual speech perception and synchrony detection for speech and nonspeech signals. J Acoust Soc Am 119(6):4065–4073
6. Cook LA, Van Valkenburg DL, Badcock DR (2011) Predictability affects the perception of audiovisual synchrony in complex sequences. Atten Percept & Psychophys 73:2286–2297
7. Davies M, Zeiss J, Gabner R (2012) Evaluating two approaches for browser-based real-time multimedia communication. In: Proceedings of the 10th international conference on advances in mobile computing & multimedia, MoMM '12, ACM, New York, pp 109–117
8. Dixon NF, Spitz L (1980) The detection of auditory visual desynchrony. Percept 9:719–721
9. Dodd B (1977) The role of vision in the perception of speech. Percept 6:31–40
10. Eg R, Behne D (2009) Distorted visual information influences audiovisual perception of voicing. In: INTERSPEECH, Brighton, pp 2903–2906
11. Van Eijk RLJ, Kohlrausch A, Juola JF, Van de Par S (2008) Audiovisual synchrony and temporal order judgments: effects of experimental method and stimulus type. Percept & Psychophys 70(6):955–968
12. Eldridge M, Saltzman E (2010) Seeing what you hear: visual feedback improves pitch recognition. Eur J Cogn Psychol 22(7):1078–1091
13. Grant KW, Seitz PF (2000) The use of visible speech cues for improving auditory detection. J Acoust Soc Am 108(3):1197–1208
14. Grant KW, Van Wassenhove V, Poeppel D (2003) Discrimination of auditory-visual synchrony. In: Proceedings of the international conference on audio-visual speech processing, St. Jorioz, pp 31–35
15. Gulliver SR, Ghinea G (2006) Defining user perception of distributed multimedia quality. ACM Trans Multimed Comput Commun Appl 2(4):241–257
16. Howell DC (2002) Statistical methods for psychology, 5th edn. Duxbury, Pacific Grove
17. ITU-T (1998) P.911. Subjective audiovisual quality assessment methods for multimedia applications. International telecommunication union, Geneva
18. Jack CE, Thurlow WR (1973) Effects of degree of visual association and angle of displacement on the "ventriloquism" effect. Percept Mot Skills 37:967–979
19. Knoche H, De Meer H, Kirsh D (2005) Compensating for low frame rates. In: Van der Veer GC, Gale C (eds) CHI extended abstracts, ACM, Portland, pp 1553–1556

20. Krasic C, Li K, Walpole J (2001) The case for streaming multimedia with TCP. In: Proceedings of the international workshop on interactive distributed multimedia systems and telecommunication services, IDMS, Lancaster, pp 213–218
21. Lewkowicz DJ (1996) Perception of auditory-visual temporal synchrony in human infants. J Exp Psychol Hum Percept Perform 22(5):1094–1106
22. Li L, Karmouch A, Georganas ND (1994) Multimedia teleorchestra with independent sources: part 2 - synchronization algorithms. Multimed Syst J 1(4):154–165
23. Liu CM, Hsu HW, Lee WC (2008) Compression artifacts in perceptual audio coding. IEEE Trans Audio Speech Lang Process 16(4):681–695
24. MacDonald J, Andersen Sr, Bachmann T (2000) Hearing by eye: how much spatial degradation can be tolerated. Percept 29(10):1155–1168
25. Marcel AJ (1983) Conscious and unconscious perception: experiments on visual masking and word recognition. Cogn Psychol 15:197–237
26. McGrath M, Summerfield Q (1985) Intermodal timing relations and audio-visual speech recognition by normal-hearing adults. J Acoust Soc Am 77(2):678–685
27. Michalos MG, Kessanidis SP, Nalmpantis SL (2012) Dynamic adaptive streaming over HTTP. J Eng Sci Technol Rev 5(2):30–34
28. Miner N, Caudell T (1998) Computational requirements and synchronization issues for virtual acoustic displays. Presence: Teleoperators Virtual Environ 7(4):396–409
29. Navarra J, Alsius A, Velasco I, Soto-Faraco S, Spence C (2010) Perception of audiovisual speech synchrony for native and non-native language. Brain Res 1323:84–93
30. Petlund A, Evensen K, Griwodz C, Halvorsen P (2008) TCP mechanisms for improving the user experience for time-dependent thin-stream applications. In: Proceedings of the IEEE conference on local computer networks (LCN), IEEE, Montreal, pp 176–183
31. Pierce CA, Block RA, Aguinis H (2004) Cautionary note on reporting eta-squared values from multifactor ANOVA designs. Educ Psychol Meas 64(6):916–924
32. Ross LA, Saint-Amour D, Leavitt VM, Javitt DC, Foxe JJ (2007) Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy environments. Cereb Cortex 17:1147–1153
33. Santangelo V, Spence C (2007) Multisensory cues capture spatial attention regardless of perceptual load. J Exp Psychol: Hum Percept Perform 33(6):1311–1321
34. Steinmetz R (1996) Human perception of jitter and media synchronization. IEEE J Sel Areas Commun 14(1):61–72
35. Stockhammer T (2011) Dynamic adaptive streaming over HTTP: standards and design principles. In: ACM MMSys, MMSys '11, ACM, New York, pp 133–144
36. Vatakis A, Spence C (2006a) Audiovisual synchrony perception for music, speech, and object actions. Brain Res 1111(1):134–142
37. Vatakis A, Spence C (2006b) Audiovisual synchrony perception for speech and music assessed using a temporal order judgment task. Neurosci Lett 393(1):40–44
38. Vatakis A, Spence C (2006c) Evaluating the influence of frame rate on the temporal aspects of audiovisual speech perception. Neurosci Lett 405(1-2):132–136
39. Vatakis A, Ghazanfar AA, Spence C (2008a) Facilitation of multisensory integration by the 'unity effect' reveals that speech is special. J Vision 8(9):1–11
40. Vatakis A, Navarra J, Soto-Faraco S, Spence C (2008b) Audiovisual temporal adaptation of speech: temporal order versus simultaneity judgments. Exp Brain Res 185(3):521–529
41. Vroomen J, Keetels M (2010) Perception of intersensory synchrony: a tutorial review. Atten Percept & Psychophys 72(4):871–884
42. Van Wassenhove V, Grant KW, Poeppel D (2007) Temporal window of integration in auditory-visual speech perception. Neuropsychologia 45(3):598–607
43. Wilson G, Sasse M (2000a) Listen to your heart rate: counting the cost of media quality. In: Paiva A (ed) Affective interactions, lecture notes in computer science, vol 1814. Springer, Berlin, pp 9–20
44. Wilson GM, Sasse MA (2000b) Do users always know what's good for them? Utilising physiological responses to assess media quality. In: People and computers XIV - usability or else! Sunderland, pp 327–339

Ragnhild Eg is a postdoctoral candidate in the Media Department at Simula Research Laboratory. As a member of an ongoing multi-disciplinary project, she focuses on the human perception of multimedia. With main interests related to the perceptual integration of auditory and visual information, her research explores how this basic process is affected by the constraints imposed by technology. At the time of writing this biography, she had recently submitted her PhD in psychology to the University of Oslo. She also holds a Master in Cognitive and Biological Psychology from the Norwegian University of Science and Technology, and a Double Bachelor in Journalism and Psychology from the University of Queensland.

Carsten Griwodz leads the Media Department at the Norwegian research company Simula Research Laboratory AS, Norway, and is Professor at the University of Oslo. He received his Diploma in Computer Science from the University of Paderborn, Germany, in 1993. From 1993 to 1997, he worked at the IBM European Networking Center in Heidelberg, Germany. In 1997 he joined the Multimedia Communications Lab at Darmstadt University of Technology, Germany, where he obtained his doctoral degree in 2000. He joined the University of Oslo in 2000 and Simula Research Laboratory in 2005. His interests lie in the improvement of system support for interactive distributed multimedia, with heterogeneous parallel processing, operating systems and protocol support for streaming applications and multiplayer games in particular. He leads the StorIKT project Verdione that investigates system support for the World Opera, and is member of the Center for Research-based Innovation ”Information Access Disruptions” that develops next generation search technology.

Pål Halvorsen is a senior researcher at Simula Research Laboratory and a professor at the Department of Informatics, University of Oslo, Norway. He received his doctoral degree (Dr.Scient.) in 2001 from the Department of Informatics, University of Oslo, Norway. His research focuses mainly at distributed multimedia systems.

Dawn M. Behne is Associate Professor and director of the NTNU Speech Laboratory in the Department of Psychology at the Norwegian University of Science and Technology. She earned a B.A. from Ripon College, an M.A. from The Pennsylvania State University in 1985 and a PhD from the University of Wisconsin-Madison in 1989. Her research has focused on category development, audiovisual integration and crossmodal speech perception, as well as cognitive neurological organization of speech processing. She is a member of The Royal Norwegian Society of Sciences and Letters.