Preserving temporal integration in multimedia: Perceived synchrony across audiovisual content and quality distortions

Ragnhild Eg

Department of Psychology

University of Oslo

Media Performance Group

[ simula . research laboratory ]

Doctoral dissertation submitted to the Faculty of Social Sciences at the University of Oslo in partial fulfillment of the requirements for the degree Philosophiae Doctor

2014
All that we see or seem
Is but a dream within a dream

Edgar Allan Poe (1809-1849)
Abstract

With the introduction of modern media, our senses are facing a new reality. The human perceptual system has adapted to the physical world through millennia of evolution; on the other hand, multimedia technology has existed less than a century. The presented project set out to explore how the human senses bind auditory and visual information in various multimedia settings. Specifically, the project aimed to establish whether the loss of auditory or visual quality could have adverse effects on this perceptual binding process. Quality loss is common to multimedia content and can arise at any point during preparation, transmission, or presentation. While audio and video distortions come with a variety of audible and visible effects, often in the form of artifacts, this work takes basis in the perceptual consequences rather than the origin of the distortions. In order to assess how quality distortions affect audiovisual integration, asynchrony was introduced as an experimental tool to establish thresholds for temporal integration that can be compared across conditions. Using this as a methodology throughout, several experiments on perceived audiovisual synchrony were conducted for different content, different distortions, different scenarios, and different experimental approaches.

Instead of focusing on specific auditory or visual artifacts that can arise from the compression, encoding or transmission of multimedia content, the first set of experiments took a more generic approach. Quality distortions were introduced uniformly and consistently across the auditory and visual signals, ensuring that all sensory information would be equally affected by the masking effect. Yet, the findings revealed no significant effect of either noise or blur on perceived synchrony of audiovisual events. However, we found temporal integration to vary significantly between short and long speech excerpts, and between speech, music, and isolated physical actions. Compared to music and speech, which both are rapid and dynamic in content, greater tolerance to asynchrony was observed for the single spoken syllable and the isolated action event. This finding is possibly related to the relatively few temporal cues shared between the modalities; the temporal alignment of auditory and visual signals is an on-going process and it seems likely that it depends on continuous and consistent reference points.
The subsequent experiments looked into asynchrony applied to teleconference scenarios. In the first study, asynchrony detection for spontaneous speech was compared to perceived synchrony in recorded speech, revealing that temporal integration is more robust for the first. Likely, this is due to two related factors, the measure and the nature of the task. When detecting gradually increasing asynchrony, the slow change may make perception less sensitive to the temporal offset. Similarly, in a live conversation, numerous distractions may take attention away from the task at hand, again making perception less sensitive to the misalignment.

A common challenge to teleconference systems is reverberating acoustics, which can affect the temporal signature of auditory signals. Accordingly, a final study explored reverberation as a distortion to auditory information. Despite the lack of observed impact from other quality distortions, reverberation might still influence perceived audiovisual synchrony due to the shared temporal dimension. Results demonstrated again the robustness of temporal integrity in audiovisual speech. However, reverberation that follows isolated events can have severe impact on the subjective perception of synchrony. Nevertheless, the main lesson learnt over the course of this work relates to the remarkability of the perceptual system. Similar to earlier studies that have demonstrated the perceptual system’s capacity to compensate for substantial separations between the modalities, be they temporal, spatial, or articulatory, this work presents examples of perception’s capacity to compensate for quality discrepancies.
Acknowledgements

Ever since I started this crazy adventure, everyone in my life has been truly supportive and understanding. First and foremost, I want to thank my marvellous friends, colleagues, and leaders at Simula, without them I would surely have lost my sanity a long time ago. I am so grateful for their continuous encouragement that I would like to highlight “a few” of the memorable people and experiences that have guided me through the PhD maze: Aiko and the bunny dance, Andreas, Håkon, Håvard, Magne, Paul and Tomas, who welcomed me from the start, Badr, Elisabeth, Simone, Sunil and Teodora, all fabulous, Beata and Pengpeng, my Perceval pals, Carsten who always takes an interest, Kristian, the best Coke, walk and talk companion, Preben – kjør på!, Pål who can resolve any LaTex obstacle, and Vamsidhar with plentiful sugar supplies.

I am forever indebted to my three supervisors, Dawn Behne, Alexander Eichhorn, and Knut Inge Fostervold. Thank you for all your advice and feedback and for enduring my periods of constant questioning, followed by periods of utter silence. Dawn was also the one to inspire me and lead me down the road to multisensory perception. Alexander, Carsten and Pål deserve an additional mention for taking a leap of faith and welcoming a stranger from the land of cognitive psychology into their world of computer science. I am truly appreciative of the opportunity to join the Perceval project and the Media Performance Group.

My family and friends outside of work have also provided invaluable motivation and they have fortunately remained in my life despite endless ups and downs, among these are: Britt, both Liten and stor, Lene and Maria, the two best cloverleaves, and Marianne, may you never lose your giggles. Certainly not forgotten is my kin on the countryside, always there and always prepared.

Last, I would to thank everyone who has ever laughed at one of my silly puns. This one is for you:

“Knock, knock.” – “Who’s there?”
“Claire.” – “Claire who?”
“Clear your schedule, it’s thesis time!”
List of Publications

The thesis builds on work addressed in the following papers:

**Paper 1**
Eg, R., & Behne, D. (Submitted). Special or not, speech is tolerant: Perceived audiovisual synchrony in distorted speech.

**Paper 2**

**Paper 3**

**Paper 4**
# Table of Contents

1. **MOTIVATION** ........................................................................................................... 1

2. **MULTISENSORY PERCEPTION** ............................................................................... 3

   2.1. **AUDIOVISUAL INTEGRATION** ......................................................................... 5

3. **AUDIOVISUAL SPEECH PERCEPTION** ................................................................. 7

   3.1. **AUDITORY AND VISUAL SPEECH CUES** ..................................................... 7

   3.2. **LOSS OF SPEECH INFORMATION** ................................................................. 9

4. **AUDIOVISUAL TEMPORAL INTEGRATION** ......................................................... 12

   4.1. **AUDIOVISUAL ASYMMETRY IN TEMPORAL INTEGRATION** ....................... 13

   4.2. **TEMPORAL INTEGRATION FOR SPEECH AND DYNAMIC EVENTS** .............. 15

   4.3. **THE ASSUMPTION OF UNITY** ....................................................................... 17

   4.4. **SPEECH PERCEPTION - SPECIAL NATURE, UNITY, OR SIMPLY COMPLEXITY?** 18

5. **MULTIMEDIA AND PERCEPTION** .......................................................................... 21

   5.1. **MULTIMEDIA QUALITY** ................................................................................ 21

   5.2. **PERCEPTION OF MULTIMEDIA** ................................................................. 25

6. **LOSS OF INFORMATION IN TEMPORAL INTEGRATION** .................................... 27

   6.1. **TEMPORAL INTEGRATION IN TELECONFERENCE SCENARIOS** .................. 28

7. **METHODOLOGICAL CONSIDERATIONS** ............................................................. 30

   7.1. **BEHAVIOURAL ASSESSMENTS OF TEMPORAL INTEGRATION** ................. 30

   7.2. **STATISTICAL MEASURES OF TEMPORAL INTEGRATION** ......................... 32

8. **RESEARCH AIMS AND OUTLINES** ....................................................................... 34

   8.1. **PAPER 1 OUTLINE: PERCEIVED SYNCHRONY IN DISTORTED SPEECH** .... 34

   8.2. **PAPER 2 OUTLINE: PERCEPTUAL TOLERANCE TO ASYNCHRONY AND QUALITY LOSS** 35

   8.3. **PAPER 3 OUTLINE: TEMPORAL INTEGRATION OF LIVE SPEECH** ............ 36

   8.4. **PAPER 4 OUTLINE: TEMPORAL INTEGRATION IN REVERBERATION** ......... 37
9. GENERAL DISCUSSION AND CONCLUSIONS

9.1. TEMPORAL INTEGRATION, PAST AND PRESENT

9.2. TEMPORAL INTEGRATION ACROSS CONTENTS AND DYNAMICS

9.3. TEMPORAL INTEGRATION IS NOT EASILY DISTURBED

9.4. FINAL THOUGHTS

10. REFERENCES

11. COLLECTION OF PAPERS

11.1. PAPER 1: SPECIAL OR NOT, SPEECH IS TOLERANT: PERCEIVED AUDIOVISUAL SYNCHRONY IN DISTORTED SPEECH

11.2. PAPER 2: AUDIOVISUAL ROBUSTNESS: EXPLORING PERCEPTUAL TOLERANCE TO ASYNCHRONY AND QUALITY DISTORTION

11.3. PAPER 3: TEMPORAL INTEGRATION FOR LIVE CONVERSATIONAL SPEECH

11.4. PAPER 4: AUDIOVISUAL TEMPORAL INTEGRATION IN REVERBERANT ENVIRONMENTS
1. Motivation

From the appearance of the first telephones, radios and televisions at the end of the 19th century and the beginning of the 20th century, the advancement of media and communication technology has progressed at express rate. With the rapid speed that new multimedia systems, platforms and devices have been developed and introduced on the market, products should be available to satisfy every need and desire. From its most dehumanised view, the industry regards individuals as consumers. With this perspective, the human experience is only valuable as a means to evaluate and continue the success of a product. Thus, surveys on user satisfaction and experienced quality contribute insights on the positive and negative features with respect to personal use, and new implementations can be made to improve subsequent versions of the technology. However, surveys can only provide information on subjective experiences and opinions, they do not take into account the underlying processes responsible for these experiences. Considering that the human perceptual system has evolved to accommodate events in the physical world, and to compensate for distances and delays as they occur naturally, it is noteworthy that so little attention has been devoted to understanding how the human senses cope with digital information.

Although multimedia technology continues to evolve, certain technological challenges are not yet resolved. Download speeds are limited by the bandwidth capacity, which is a particular problem for data transmission over mobile networks. People on the go demand high quality content and they demand it instantly, but quality and speed must be weighed up against each other. Compression and encoding of audio and video increase transmission speeds at the sacrifice of fine-grained details; moreover, this type of bit-rate reduction can result in a number of different artifacts that may affect the user experience, such as blurred video, noisy audio, or temporal asynchrony between audio and video. Some auditory and visual details are not necessary for comprehension and some may not even be seen or heard, hence multimedia researchers study users’ experiences and evaluations of quality to establish how much compression people will not only tolerate, but also fail to notice.
However, people are not likely to be aware of the amount of details their vision or hearing depend on, and much less likely to provide the specifics in a user study. Thus, the motivation of this thesis was to go beyond user experience and evaluation and introduce methodologies from cognitive psychology to study the perception of multimedia quality. The thesis aims to explore how humans perceive events that have been removed from their natural settings and that may have lost essential sensory cues through transmission and the presenting device.

A general summary of multisensory perception introduces the theoretical framework, then narrows down the focus to audiovisual temporal integration, before outlining relevant problems pertaining to multimedia systems. Within this context, the project presents an alternative approach to studying multimedia quality by considering the perceptual consequences.
2. Multisensory perception

Humans perceive a limited span of the greater world, most notably with respect to the narrow frequency ranges of audible sound and visible light. Still our senses are showered with sounds, images, tastes, smells, and touches. All these inputs have to be filtered, processed, and eventually aligned, to create a coherent perceptual representation of the surroundings. Considering how few experiences are specific to one sense modality, the amount of research with focus on only one sensory system seems disproportional. While multisensory processes would be all but impossible to understand without a solid grasp of the individual perceptual systems, the human experience of the world does not proceed one sense at a time. Multisensory perception is certainly not a novel research topic; it has intrigued scientists and philosophers throughout centuries. From Aristotle’s philosophical ponderings on sensory co-ordinates (Aristotle, 1931) to Victor Urbantschitsch’s experimental explorations of the senses’ effects upon each other (cited in: Ryan, 1940). Moreover, the fact that so many human experiences are multimodal by nature is used as a recurring criticism to research areas that seem to ignore the interaction between senses by focusing on single sensory processes (Ryan, 1940).

Indeed, most of our everyday experiences involve at least two of our senses. For instance, when drinking a cup of coffee, the olfactory sense registers the smell of the ground and heated beans, the gustatory sense conveys the taste of the black brew, and the tongue and fingers feel the heat of the liquid. From a phenomenological viewpoint, much attention has been devoted to less typical crossmodal associations, such as perceived correspondence in brightness between odours, tones, and colour (Cohen, 1934). These and other atypical correspondences between two or more senses are commonly grouped together under the term synaesthesia (Laeng, Svartdal, & Oelmann, 2004; Simner, 2012). Although the definition of the term varies and the associations can be manifested in a multitude of ways, the term normally refers to a sensory experience triggered by the stimulation of another sensory modality (Simner, 2012). This could be a colour seen when hearing a word (Goller, Otten, & Ward, 2009), a taste experience induced by a tone (Beeli, Esslen, & Jäncke, 2005), or even feeling an observed touch (Blakemore, Bristow, Bird, Frith, & Ward, 2005). Of course,
multisensory perception extends far beyond curious connections between crossmodal events; similar to the described coffee experience, most humans learn associations between common sensory events through repeated exposure. Thus, understanding the perceptual mechanisms that bind both usual and unusual associations will contribute to a broader knowledge on how humans perceive and experience any number of situations encountered.

Multisensory researchers study the cognitive mechanisms that unite signals from the different modalities, and they work to localise the associated neurological activity. In recent years, significant progress has been made with increasingly sophisticated technology. This technology spans from single-neuron studies that demonstrate the importance of the superior colliculi in the integration of sensory signals (Stein & Meredith, 1993; Stein, Stanford, Ramachandran, Perrault, & Rowland, 2009), to functional and structural imaging techniques that contribute to map out the neuronal networks involved (Calvert, 2001). While studies on neuronal centres and pathways offer unique insights into the complexity of both unisensory and multisensory processes, their data are only meaningful in the context of behavioural and perceptual experiments. The firing rate and interconnections of neurons are in themselves highly interesting biological mechanisms, but to understand the conscious outcome of multisensory perception, the cause of the activity is of equal importance. Different experimental methods have been implemented in the search for multisensory pathways, where imaging technology measures brain activity during stimuli presentations. Spatial and temporal consistence between sensory signals tend to elicit enhanced neuronal activity, whereas a separation in time or space can suppress activity (Fetsch, DeAngelis, & Angelaki, 2013). For instance, simple auditory and visual stimuli presented in synchrony have been found to cause superadditive BOLD (blood oxygenation level dependent) responses in the superior colliculi, and other regions, while asynchronous presentations cause response depressions (Calvert, Hansen, Iversen, & Brammer, 2001). Similarly, spatially coincident signals from different sensory modalities may also increase the firing rate of multisensory neurons (Wallace, Wilkinson, & Stein, 1996). Continued progress in the combination of imaging technology and perceptual paradigms will surely bring forward new insights on both the structure and
function of multisensory pathways; in the meantime, multisensory percepts should be accepted as outcomes that are greater than the sum of their parts (as indicated by superadditive neural responses (Stanford & Stein, 2007)).

With imaging technology at the forefront of the most recent advances, multisensory research is still evolving as a field. However, findings from these studies rest on the application of experimental tasks, that again rest on established theories about perceptual integration. Consequently, in order to explore the neurological basis of an integrative process, first an understanding of the perceptual mechanism is required. Furthermore, to fully appreciate the complexity of the human perceptual system, it must be put in a natural context where it has to tackle a multitude of concurrent sensations. This work takes two steps in that direction. Human perceptual experience is considered across the auditory and the visual sensory systems, although the investigation is limited to the perceptual integration of auditory and visual information in the temporal domain. In order to move the research towards a more familiar environment, the conducted experiments explore the temporal integration of audiovisual information in different multimedia scenarios. Because all the experiments involve digital representations of auditory and visual content, our explorations are still removed from the physical world. However, the experiments are set in digital environments that share characteristics with so many common multimedia platforms; moreover, they address current challenges in multimedia technology.

2.1. Audiovisual integration

Among the multitude of multisensory experiences, audiovisual integration is perhaps the perceptual process that has received the most attention in research. The heavy focus on audiovisual processes is likely due to the strong correspondences between auditory and visual events (Spence, 2011). In some research areas, such as computer science, the interest is also spawned by the continuous demand for audiovisual content; relatedly, audio and video are by and large the only two sense domains that can be digitalised, transmitted, and recreated in a new location (Halsall, 2001). In the physical world, both sight and hearing provide information about the spatial location and the nature of an event;
it could be a saxophone player standing on the other side of the street, or a seagull screeching while flying past. Audiovisual events are integrated by both shared and associated cues. Shared cues are typically amodal, such as temporal and spatial cues, whereas associated cues are often learnt through experience, particularly with regards to semantics and context. Some cues are more difficult to characterise, for instance intensity, which can take the form of a perceived association between visual brightness and auditory loudness. Even pitch and contour associations have been uncovered, frequently referred to as the bouba-kiki effect (Ramachandran & Hubbard, 2001). While the majority of people will exhibit the bouba-kiki effect, it is commonly grouped together with the less frequent crossmodal correspondences typical of synaesthesia. A more common audiovisual process, familiar to most people, involves the production and perception of speech information. The diverse interest and applicability related to speech is another prominent reason for the attention devoted to research on audiovisual perception.
3. Audiovisual speech perception

Not only is speech a familiar process to the vast majority of people, it is an essential process that enables humans to communicate. Spoken language can be fully comprehended from audio alone, at least under normal circumstances. Yet, humans benefit from the additional information supplied by vision (Schwartz, Berthommier, & Savariaux, 2004). While audiovisual speech perception is far from being a recent area of study (e.g., Sumby & Pollack, 1954), researchers are still working to understand the mechanisms that are at work when integrating auditory and visual speech information. Furthermore, with the continuous advances in multimedia technology, speech has also become a source of audio and video data that must be kept intact and intelligible throughout several stages of recording, coding, transmission, and presentation. In everyday speech, auditory and visual information are integrated by the many shared characteristics that can be attributed to the same source (Welch & Warren, 1980). According to the unity assumption, described in §4.3, these characteristics are both redundant and amodal (not specific to a single modality); the most prominent examples include temporal, spatial, articulatory, and semantic cues (Vatakis, Ghazanfar, & Spence, 2008; Welch & Warren, 1980). Together, these cues strengthen the bond between auditory and visual speech, to the point where they can even overcome small discrepancies (Bertelson & Aschersleben, 2003; Vatakis & Spence, 2007). This is certainly true for natural events, where the different speeds of light and sound cause a delay in the time of arrival between the signals. However, the digital world can introduce far larger discrepancies than any natural event. Within the context of this work, we therefore focus on how the perceptual bond contributes to maintain coherence between auditory and visual speech across the spatial and temporal separations that can follow multimedia content.

3.1. Auditory and visual speech cues

For the auditory modality, speech cues are in many respects synonymous with phonemes. These speech units form the bases of syllables and words and are categorised according to their acoustical properties (Ganong & Zatorre, 1980). On the other hand, how humans learn to distinguish between visual speech cues remains a debated topic and several models have been put forward, such as the
striatal pattern classifier (Ashby & Waldron, 1999). So far, one of the better categorical accounts of basic visual speech units is probably that of visemes (Fisher, 1968). Because the lips, jaw and tongue are the only speech-producing movements that can be seen by the casual observer, the visual modality does not cover the same range of articulatory nuances as the auditory modality. Acoustically different speech sounds may appear visually similar (Hilder, Theobald, & Harvey, 2010); moreover, the same consonants can take on vastly different visual shapes depending on the context in which they are presented (Taylor, Mahler, Theobald, & Matthews, 2012). In other words, there is no one-to-one correspondence between phonemes and visemes (Chen & Rao, 1998), and visemes have likely more overlap and are thus more easily confused (Cathiard, Schwartz, & Abry, 2001). For instance, with respect to acoustical characteristics, the difference between bilabial and velar stop consonants is easily discernable from the mouth movements. Bilabial sounds are produced with the lips initially contracting then opening, while velar sounds are articulated further back in the mouth, with the back part of the tongue touching the soft palate (Kent, 1997). Conversely, other speech sounds are far more difficult to distinguish visually, even though they differ in manner or place of articulation. A few examples include labiodental nasals and labiodental stops, alveolar nasals and alveolar approximants, as well as dental fricatives and alveolar stops (Dodd, 1977).

Considering that the acoustical misidentification of place of articulation can lead to confusion among consonants (Miller & Nicely, 1955), the mouth movements of the speaker can help resolve that confusion (Green & Kuhl, 1989; Rosenblum & Saldàña, 1996). Furthermore, visemes are not the only visible speech features, the visual modality also contributes with additional information on word separation, syllable stress, intonation, and vowel identity (Burnham, Ciocca, Lauw, Lau, & Stokes, 2000; Keating et al., 2003; Navarra & Soto-Faraco, 2007). Prosody is additionally accentuated by corresponding head and facial movements (Krahmer & Swerts, 2007), and other forms of mimicry may also enhance the verbal message. Articulatory disparities between the auditory and visual modalities can actually be overcome in the perceptual integration of speech. In a landmark study on the importance of visual information in speech perception, McGurk and MacDonald (1976) demonstrated how the perceptual system can merge
conflicting information from vision and hearing. By pairing the video of a speaker articulating a syllable with the audio of another syllable produced by the same speaker, they found that many participants perceived neither of the two. In the original experiment, the syllables differed in place of articulation, and the combination of an auditory labial syllable with a visual velar syllable led to, more often than not, a perceived alveolar syllable (McGurk & MacDonald, 1976). This fusion of incongruent speech sounds is now typically referred to as the McGurk effect. Newer findings have revealed that the visual influence in the McGurk effect is enhanced with the introduction of noise (Alm, Behne, Wang, & Eg, 2009; Fixmer & Hawkins, 1998), the audiovisual fusion can also endure obscurities created by visual distortion (Eg & Behne, 2009; Fixmer & Hawkins, 1998). Furthermore, the McGurk effect has been demonstrated to be resilient to temporal offsets, at least below 80 to 170 ms, depending on the direction of asynchrony (van Wassenhove, Grant, & Poeppel, 2007). While the fusion of speech sounds is a remarkable perceptual illusion, the McGurk effect also serves as a strong example of the importance of the visual modality in speech perception.

### 3.2. Loss of speech information

Due to the complex and dynamic combination of the acoustical characteristics inherent in auditory speech, some nuances are easily lost and the meaning may not be successfully conveyed to the listener. Surrounding sounds can become background noise or audio can be corrupted by the transmitting medium, these in turn mask or degrade the frequencies of human speech (Kim & Stern, 2011; Miller & Nicely, 1955; Petrovsky, Azarov, & Petrovsky, 2011; Rogers, Lister, Febo, Besing, & Abrams, 2006). Mercifully, the ambiguity caused by degraded speech signals can be relieved by the added contribution of visual speech information (Girin, Schwartz, & Feng, 2001; Sumby & Pollack, 1954). Seeing the face of someone talking improves the ability to detect speech (MacLeod & Summerfield, 1987), to comprehend speech (Schwartz et al., 2004), and to correctly identify phonetic attributes (Alm et al., 2009). However, as illustrated by the McGurk effect (McGurk & MacDonald, 1976), the visual contribution can alter the audiovisual percept if the information from the two modalities is not
congruent. This observation was extended in an investigation of speech detection in sentences presented with white noise; response accuracy was predictably better for the noisy audio tracks that were paired with matching video recordings compared to those paired with mismatching video recordings (Grant & Seitz, 2000).

On the other hand, when visual speech information is somehow obscured, its impact is also reduced. For instance, the rate of responses matching the auditory signal increases when participants are presented with incongruent audiovisual syllables where the video has been degraded using mosaic transformations (MacDonald, Andersen, & Bachmann, 2000). Participants’ ability to correctly identify syllables articulated visually by a digital talking head also decreases when the stimulus is masked by the same type of spatial quantisation (Campbell & Massaro, 1997). Moreover, increased reliance on the auditory modality has been observed for McGurk stimuli with the video blurred by Gaussian filters (Thomas & Jordan, 2002) or decreased pixel resolution (Eg & Behne, 2009), or even visual obscuring using paper affixed to the monitor (Fixmer & Hawkins, 1998). Visual information is undeniably beneficial to speech perception, but humans do not rely on all the information that is available under normal conditions. In a series of experiments on spatial frequency and speech comprehension, the authors noted that performance was better for the high-frequency than for the low-frequency band-pass filters, but with a peak performance at the mid-range frequencies (Munhall, Kroos, Jozan, & Vatikiotis-Bateson, 2004). Although the visual enhancement of the filtered images never reached the level of the unfiltered control condition, the number of correctly identified key words was still far greater than the auditory-only condition.

Results from studies on speech perception in noisy and visually degraded environments demonstrate the importance of seeing the facial movements of a speaker. More than that, they highlight the remarkable resilience of the perceptual system. People are able to detect and understand spoken words and syllables in noise that is louder than the speech itself, in other words, at negative signal-to-noise ratios (Alm et al., 2009; Girin et al., 2001; MacLeod & Summerfield, 1987). Furthermore, at severe levels of visual distortions, such as mosaic
transformation, blurring, or filtering, the influence of the visual modality persists in the form of infrequent McGurk fusion responses (Eg & Behne, 2009; MacDonald et al., 2000) and improved word identification (Munhall et al., 2004). Similar observations have been made using very sparse visual speech representations in the form of illuminated dots (Rosenblum, Johnson, & Saldaña, 1996). Although participants typically do not perceive these point-light displays as speech movements, their responses still reflect influence from the visual recordings (Rosenblum, 2008). Even dynamically expanding rectangles and ovals can improve speech detection in noisy conditions (Bernstein, Auer, & Takayanagi, 2004). Nevertheless, the need for an intact auditory signal arguably becomes more important when the visual speech signal is degraded, thus diminishing the added benefit this modality would normally contribute to speech comprehension.
4. Audiovisual temporal integration

As seen from studies on speech recognition (Grant & Seitz, 2000) and fusion of speech syllables (McGurk & MacDonald, 1976), the integration of audiovisual speech information is facilitated by the auditory and visual convergence of articulatory cues. As mentioned, the integration of the two modalities depends on the correspondence between several amodal cues (Welch & Warren, 1980), for speech and for other audiovisual events. For two signals to be attributed to the same source, they must occur at the same time. Thus, temporal coincidence is a required perceptual cue, although it need not imply absolute simultaneity. A metaphorical perceptual buffer compensates for short temporal offsets between sensory signals, and in the case of auditory and visual signals, these offsets can extend to a few hundred milliseconds (Conrey & Pisoni, 2006; Grant, van Wassenhove, & Poeppel, 2003; Maier, Di Luca, & Noppeney, 2011; Vatakis & Spence, 2006a). In general, the term temporal integration describes the process that is responsible for perceiving synchrony between audiovisual signals (Arrighi, Alais, & Burr, 2006). The temporal integration between an auditory and a visual event is typically assessed using subjective judgements of synchrony, or temporal order, across different temporal offsets; the different methodologies common to this field are described in §7. Despite the remarkable tolerance of the perceptual system, the bond between two modalities will eventually break when the separation between them is pushed too far. In the case of temporal misalignment, the breaking points typically correspond to the temporal offsets where asynchrony is perceived more frequently than synchrony; thus, they define the thresholds at which the senses are no longer integrated (Grant et al., 2003). The thresholds, in turn, establish the window of temporal integration (Vroomen & Keetels, 2010). With the massive body of research into audiovisual synchrony perception that has accumulated over the decades, it is evident that there are no universal thresholds for the temporal integration of the two modalities (e.g., (Maier et al., 2011; Spence & Squire, 2003; Zampini, Guest, Shore, & Spence, 2005)). Instead, these thresholds are statistically derived values that represent a dynamic range of tolerable asynchronies (van Eijk, Kohlrausch, Juola, & van de Par, 2009), and these vary depending on the nature of the events and other conditions (Dixon & Spitz, 1980; Vatakis & Spence, 2006a), as well as between individuals (e.g.
Fouriezos et al., 2007). As outlined in §11.2, our studies using dynamic audiovisual stimuli uncovered patterns consistent with these earlier findings, with temporal tolerance depending on the distinct events.

### 4.1. Audiovisual asymmetry in temporal integration

Past studies on the temporal integration of audiovisual content have established that the perceptual system is inherently asymmetric in its capacity to compensate for asynchrony. When an auditory signal arrives before a visual signal (audio lead asynchrony), the perceptual system is fairly sensitive to the discrepancy; however, when the visual signal arrives first (audio lag asynchrony), rather large temporal offsets go unnoticed (Lewkowicz, 1996). The robustness of the temporal tolerance to audio lag asynchrony is especially prominent for speech (Grant et al., 2003). Figure 1 presents key findings from select works on temporal integration for speech and other audiovisual events. As seen, the windows of temporal integration vary greatly between studies, emphasising the nonexistence of universal thresholds. Moreover, the asymmetry between audio lead and audio lag asynchrony is demonstrated in all the presented studies, for the vast majority of conditions. For instance, several studies have presented windows of temporal integration with audio lag thresholds more than 100 ms longer relative to objective synchrony, as compared to audio lead thresholds (Dixon & Spitz, 1980; Grant et al., 2003; Maier et al., 2011; Petrini, Dahl, et al., 2009; Vatakis & Spence, 2006a). Some have speculated that the origin for this asymmetry must be the faster processing of auditory signals, which requires an earlier processing of visual signals to make up for the lag (Vroomen & Keetels, 2010). Others attribute the asymmetry to the different speeds of light and sound (≈300,000 km/s and ≈345 m/s\(^1\), respectively (Lide & Haynes, 2009)), with the perceptual processes compensating for the delayed arrival of the auditory signal relative to the visual signal (Alais & Carlile, 2005; King, 2005; van Eijk, Kohlrausch, Juola, & van de Par, 2008). In other words, audio lead asynchrony cannot occur in nature and is contingent on neural processing time, while audio lag asynchrony varies naturally according to distance, explaining why the perceptual system has a greater capacity to compensate for misalignments in this direction (Alm & Behne, 2013).

\(^1\) Speed of sound in dry air at ≈23°C.
Figure 1. Windows of temporal integration established for a selection of audiovisual stimuli and experimental methods. Audio lead and lag thresholds are defined by the implemented method, the labels in parentheses are explained in §7.2. Stimuli are colour-coded according to the nature of the audiovisual events, from top to bottom: dark green represents longer excerpts of continuous speech, while light green is one-word or one-syllable stimuli, blue corresponds to simple and repetitive stimuli, yellow indicates music-related stimuli, and red is associated with physical action events. Temporal thresholds are not published in all of the illustrated works; those missing were calculated from the data provided.
A review of recent advances in multisensory integration remarks that neurons that receive cues from more than one sense modality tend to show enhanced responses to multisensory stimuli, provided that they take place in close spatial and temporal proximity (Fetsch et al., 2013). From this probabilistic perspective, the eventual percept rests on the neuronal activation following exposure to multisensory stimuli; the closer together two sensory cues are in space and time, the more likely is their integration (Fetsch et al., 2013). Extending on the speculations of others, it follows that the temporal processing of corresponding auditory and visual cues can tolerate some delays in either direction, with the rate of neuronal activity increasing as the events come closer together in time and space. Moreover, this response enhancement may be stronger for sensory cues that are separated in the arguably more natural direction of light before sound.

4.2. Temporal integration for speech and dynamic events

As mentioned, the perceptual tolerance to audiovisual asynchrony varies between individuals (Fouriezos et al., 2007) and according to signal precedence (Maier et al., 2011; Petrini, Dahl, et al., 2009). Furthermore, the nature (van Eijk, Kohlrausch, Juola, & van de Par, 2008) and complexity (Arrighi, Alais, & Burr, 2006; Fujisaki & Nishida, 2005) of the audiovisual event will define the window of temporal integration. In Figure 1, the distinction in temporal integration for different categories of stimuli is highlighted using colours. Most notable is the great variation across all studies. Also noteworthy is the tendency for perceived synchrony for simple and short stimuli, such as light and tones or spoken syllables, to break apart at shorter intervals compared to longer and more complex stimuli, such as music pieces or spoken sentences. Some relate the strong audiovisual bond for spoken language to the special nature of speech (Liberman & Whalen, 2000; Tuomainen, Andersen, Tiippana, & Sams, 2005), and the existence of neurological processes that are specialised for phonological information, auditory and visual (Liberman, 1982). Indeed, studies comparing temporal integration for speech and more physical events have found that the perceptual tolerance to asynchrony is stronger for dynamic speech than for action sequences with predictable moments of impact (Dixon & Spitz, 1980; Miner & Caudell, 1998; Moore, 2000; Vatakis & Spence, 2006a). Of course, the special
nature of speech remains a debated topic (Klatt, 1979; Massaro & Chen, 2008; Nearey, 1997; Vroomen & Stekelenburg, 2011). Yet, if siding with the former, then the strong temporal tolerance associated with speech should be attributed to specialised processes that bind phonological information across modalities.

Consistent with a strong perceptual integration for audiovisual speech, Dixon and Spitz (1980) found that the detection of gradually introduced asynchrony occurred sooner for their sequence with a hammer hitting a nail than for their prose sequence. Corresponding results have been found for audiovisual presentations of other hitting actions, with greater perceptual sensitivity to asynchrony for action sequences than for spoken sentences and musical content (Vatakis & Spence, 2006a). Considering the ability of the perceptual system to maintain coherence even with inconsistencies between modalities, as demonstrated by the McGurk effect (McGurk & MacDonald, 1976), one account would argue that the audiovisual bond is strengthened by common semantic cues (van Wassenhove et al., 2007). By extending this perceptual paradigm into the temporal dimension, van Wassenhove, Grant and Poeppel (2007) set out to explore whether the absence of semantic cues could affect the temporal integration of spoken syllables. As predicted, they found larger windows of temporal integration for audiovisual syllables that were consistent across modalities, as compared to the incongruent McGurk stimuli. Hence, in the presence of several consistent amodal cues, the perceptual integration of multisensory signals is more likely to be preserved. Furthermore, the robustness of the temporal integration of intelligible audiovisual speech is also demonstrated in contrast to monkey calls (Vatakis, Ghazanfar, et al., 2008) and unfamiliar languages (Navarra, Alsius, Velasco, Soto-Faraco, & Spence, 2010).

Not all findings support the proposition of a stronger temporal bond for speech-related content. In an investigation of perceived synchrony in audiovisual speech, Conrey and Pisoni (2006) compared recordings of spoken words with simple flash and tone presentations. Here, the authors found a smaller window of temporal integration for the speech stimuli, resulting in the proposal that the multisensory processing of audiovisual synchrony is common across speech and other events (Conrey & Pisoni, 2006). Moreover, the strong temporal bond
observed for speech tokens is not unique. In comparison to musical notes played on a guitar or a piano, synchrony is maintained longer for music than for spoken sentences (Vatakis & Spence, 2006a). Similarly, in our own experiments (§11.2), we found that asynchrony was detected at shorter temporal offsets in the news broadcast than in the drumming sequence. Seeing how no constant thresholds can be established for temporal integration, neither within nor between event types or individuals, comparison studies like these are still too few to form confident conclusions. As a preliminary note, it is surmised that the temporal binding of speech does not always distinguish itself from other audiovisual events.

### 4.3. The assumption of unity

If speech is no more special than other audiovisual events, an alternative explanation is required to account for the strong temporal tolerance reported in the majority of studies (as outlined in Figure 1). Indeed, speech is not the only process to exhibit consistency between auditory and visual cues; through experience, we learn to associate a number of different sounds to their visual sources. According to the assumption of unity, sensory integration is strengthened by congruent and redundant characteristics, or cues, that are associated with a common source across modalities (Vatakis, Ghazanfar, et al., 2008; Vatakis & Spence, 2007, 2008; Welch & Warren, 1980). Consequently, the level of consistency and the number of common characteristics define the strength of the perceptual integration. With respect to the strong bond between auditory and visual speech, this could be explained by the number and consistency of the cues common to both modalities, such as temporal and spatial co-occurrence, semantic congruence, and signal salience (Vatakis, Ghazanfar, et al., 2008). Together, these cues create a united percept that persists even across small separations, such as temporal offsets (Bertelson & Aschersleben, 2003; Vatakis & Spence, 2007) or spatial dislocation (Bertelson, Vroomen, de Gelder, & Driver, 2000; Jack & Thurlow, 1973). Consistent with this proposition, for incongruent audiovisual speech syllables presented at different levels of asynchrony, the rate of responses matching the visual modality is higher within, than outside, the range of perceived synchrony (Soto-Faraco & Alsius, 2009). As for events that humans typically have less experience with, such as paired presentations of circles and
tones (Conrey & Pisoni, 2006), they would share less characteristics and should therefore be less resilient to separations.

Support of the unity assumption comes from studies that compare perceptual integration of audiovisual stimuli across manipulations to the congruency of amodal cues. For instance, in the temporal integration of spoken words or syllables where the gender of the voice does not match the talking face, asynchrony tends to be noticed at shorter temporal offsets compared to presentations where the face and voice belong together (Vatakis & Spence, 2007). Furthermore, considering perceptual integration from McGurk fusions of speech syllables, Tuomainen, Andersen, Tiippana, and Sams (2005) applied sine-wave representations of spoken syllables to manipulate the experimental context without altering the stimuli. In general, participants will not perceive the sine waves as speech sounds unless their significance is pointed out, which allows for a comparison of audiovisual integration across speech and non-speech conditions. Indeed, the authors found that the McGurk effect was almost absent when participants were unaware of the speech representation, while the rate of fusion responses approached the level of natural speech when the sine waves were initially perceived as speech syllables (Tuomainen et al., 2005). These results provide a clear example of the profound difference in the perceptual processing and integration of auditory and visual information when one modality does not convey semantic information. Combined, these studies demonstrate the contribution of semantic or phonological cues in the perceptual integration of audiovisual speech. Moreover, they highlight that the strong perceptual bond associated with audiovisual speech does not extend to all vocal productions. Instead, this bond is likely limited to intelligible speech sounds.

4.4. **Speech perception - Special nature, unity, or simply complexity?**
An argument against the proposed specialised processes for human speech ascribes any differences in perceptual integration to the complex dynamics of speech production. When comparing spoken language to other audiovisual events, the comparison extends beyond a mere comparison between speech-related and speech-unrelated events. More commonly, the comparisons are made between
dynamic and continuous speech excerpts and isolated and predictable action events that culminate in a moment of impact. Hence, instead of specialised speech processes or amodal associative cues, the observed differences could simply be linked to the frequency, predictability, complexity, or visibility of the experimental stimuli.

To minimise the natural complexity of spoken language, sine-wave speech representations have also been used in a study on temporal integration (Vroomen & Stekelenburg, 2011). Here, the sine waves yielded almost identical thresholds to natural speech, whether they were perceived as speech tokens or not. Although the speech manipulation had an effect on McGurk fusion responses, similar to that presented by Tuomainen and colleagues (2005), the temporal thresholds remained consistent across speech presentations. Considering that temporal integration of sine waves is not altered by the awareness of the speech representation, the authors propose that judging the order of an auditory and a visual stream does not require a perceptual pairing (Vroomen & Stekelenburg, 2011). In other words, the lack of semantic and articulatory cues, consistent across modalities, did not affect the subjective perception of temporality, which could be viewed as a disputation against both the unity assumption and the special nature of speech. In an alternative approach to reduce the natural complexity of speech, one study used band-pass filtered spoken sentences at four different spectral bands (Grant et al., 2003). This technique maintains the intelligibility of the speech signals by preserving different acoustic speech envelopes, but the level of perceived correlation with visually presented mouth movements changes across the different bands (Grant & Seitz, 2000; Grant, 2001). Nevertheless, the perception of audiovisual synchrony did not vary significantly across the spectral bands, nor did they differ from the natural speech condition (Grant et al., 2003). Again, the manipulation to the auditory speech signal had no effect on the temporal integration of the two modalities, suggesting that the bond between auditory and visual speech is no more special than other audiovisual events.

On the other hand, the two studies by Grant and colleagues (2003) and Vroomen and Stekelenburg (2011) are based on the premise that altering the auditory signal
could in some way diminish the dynamics typical of human speech. Although sine-wave speech can strengthen the link between the auditory and visual modality when using McGurk stimuli (Tuomainen et al., 2005; Vroomen & Stekelenburg, 2011) and filtered speech can vary in association with visual speech movements depending on the spectral band (Grant & Seitz, 2000; Grant, 2001), these findings reflect phonological processes that may not be relevant to the temporal integration of audiovisual speech. When resolving conflicts between contradictory articulatory cues or when relying on both modalities for semantic understanding, the perceptual system is engaging a number of interacting activities. Temporal alignment between auditory and visual signals reflects only one of many activities that must converge in order to achieve speech comprehension. Accordingly, audiovisual temporal integration may not depend on all the factors that must be in place for overall speech perception. Moreover, it could be that auditory speech signals are particularly resilient to some degradations, such as loss of frequencies, so that even severe manipulations can be insufficient to remove the temporal cues required to align the auditory events with the visual events. Possibly, the temporal integration between the visual and the degraded auditory modality will be maintained as long as the overall temporality can be discerned from the most salient auditory and visual cues. This proposition was further explored for both the auditory and the visual modality in §11.1 and §11.2, where global distortions were applied to either modality at levels ranging from barely noticeable to severely disturbing.
5. Multimedia and perception

The term multimedia is one that seems straightforward at first glance. Through the past decades, we have come to associate multimedia with any entertaining or informative content broadcasted to television sets, computers, mobile devices, and even more specialised devices such as teleconference platforms and game consoles. Definitions of the term refer to text, images, audio or video transferred over the network between people or between a person and a system (Halsall, 2001), where information is integrated, manipulated, presented, stored, and communicated in at least one continuous and one discrete medium (Steinmetz & Nahrstedt, 1995). For the sake of brevity and familiarity, we use the term to describe digitally presented audiovisual content.

5.1. Multimedia quality

In the field of psychology, noise (Alm et al., 2009; Sumby & Pollack, 1954), distortion (Eg & Behne, 2009; MacDonald et al., 2000), and other disturbances (Fixmer & Hawkins, 1998), are used to assess the reliance on a sensory modality that is rendered less intelligible. In multimedia research, similar disturbances are considered artifacts that result from encoding, decoding, compression, or transmission (Liu, Hsu, & Lee, 2008; Punchihewa, Bailey, & Hodgson, 2003; Yuen & Wu, 1998). Figure 2 illustrates the chain of events from the capture of audiovisual content, through compression and encoding, to the streaming of multimedia files. Compression speeds up transmission, therefore the network bandwidth sets the criteria for the multimedia quality. However, high-resolution monitors and televisions require better quality than small devices such as tablets and mobile phones. Thus, the screen resolutions must be considered along with the bandwidth capacity. For instance, by implementing active adaptation, the streaming of multimedia content will adjust according to the available resources; this kind of adaptation typically involves downscaling the quality when bandwidth is limited or upscaling when the capacity improves (Creusere & Hardin, 2011; Michalos, Kessanidis, & Nalmpantis, 2012; Su, Yang, Lu, & Chen, 2009). Quality scaling is an efficient way to reduce the bitrate of a video and it enables content to be delivered in a timely manner without packet drops (loss of data) or stalling (small delays or freezing) (Wu, Claypool, & Kinicki, 2006).
Figure 2. A step-by-step multimedia streaming example. Audio and video are captured by a microphone and a camera and then transferred to a server. Media files are compressed and encoded at the server, before they are streamed to any user device that sends a request for the content.

However, downscaling of audio and video streams will eventually result in perceptible artifacts and overall loss of quality (de Koning, Veldhoven, Knoche, & Kooij, 2007; Goldmann et al., 2010). For instance, scalable video coding
(SVC) is a common technique used to encode and compress H.264 videos, thus enabling bandwidth adaptation (Schwarz, Marpe, & Wiegand, 2007). This technique downscales video streams by manipulating the temporal resolution (frame rate), the spatial resolution (pixel rate), or the quantisation parameter (fidelity, granularity of encoded units) (Ni, Eg, Eichhorn, Griwodz, & Halvorsen, 2011; Segall & Zhao, 2008). Table 1 presents a simplified summary of the most common artifacts associated with this downscaling technique, and Figure 3 provides a visual illustration for two of these. Additionally, artifacts, or quality degradations, may arise from the technology in use, be they malfunctions in the equipment used for capturing, shortcomings in transmission, or distortion originating in the presentation set-up. Reverberation is an example of the latter, although it can also arise from equipment malfunctions. Reverberation is a frequent problem in teleconference systems and is further illuminated in §6.1.

Table 1. Overview of SVC techniques with definitions of the most commonly related artifacts (Ni et al., 2011; Schwarz et al., 2007; Segall & Zhao, 2008; Yuen & Wu, 1998).

<table>
<thead>
<tr>
<th>Temporal resolution</th>
<th>Spatial resolution</th>
<th>Quantisation parameter</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Downscaled frame rate:</strong></td>
<td><strong>Downscaled pixel rate:</strong></td>
<td><strong>Downscaled fidelity:</strong></td>
</tr>
<tr>
<td>Fewer video frames per second</td>
<td>Fewer pixels per video frame</td>
<td>More granular encoding</td>
</tr>
<tr>
<td>- Jerkiness</td>
<td>- Loss of sharpness</td>
<td>- Blockiness</td>
</tr>
<tr>
<td>Discontinuity in a moving sequence</td>
<td>Unsharpening of edges and contrasts</td>
<td>Visible boundaries between quantised blocks (encoded units) in a video frame</td>
</tr>
<tr>
<td>- Judder</td>
<td>- Loss of details</td>
<td>- Ringing</td>
</tr>
<tr>
<td>Flipping motion of a still-frame in the direction of movement</td>
<td>Disappearance of fine-grained details</td>
<td>Shimmering echoes around hard edges</td>
</tr>
<tr>
<td>- Jitter</td>
<td>- Bluriness</td>
<td>- Noise</td>
</tr>
<tr>
<td>Variations in motion speed</td>
<td>Global haziness that is particularly noticeable for textures and features</td>
<td>Random variations between pixels</td>
</tr>
</tbody>
</table>
Visible and auditory artifacts are not the only potential quality degradations that result from encoding, compression and transmission. Since these techniques are frequently performed separately for audio and video, the streams can become temporally misaligned. As discussed, audiovisual asynchrony becomes noticeable once the temporal offset exceeds a certain threshold; this is detrimental not only to sensory integration (Zampini et al., 2005), speech comprehension (Grant & Greenberg, 2001), and other perceptual processes (Arrighi et al., 2006; Fujisaki, Shimojo, Kashino, & Nishida, 2004), but also to the subjective experience of quality (Steinmetz, 1996). With the consumer demands for high-speed, high quality content on the one hand, and the practical limitations on the other hand, the challenge is to find the best trade-off. Since lower bitrates improve delivery time, but also reduce the subjective and perceptual quality, we set out to explore
the relationship between single-modality artifacts and asynchrony as an audiovisual artifact and as a perceptual integration process (§11.1 and §11.2).

5.2. Perception of multimedia

Studies on Quality of Experience were originally designed to establish the levels of audio/video quality that will satisfy the expectations of consumers while at the same time adhering to the available resources (de Lima et al., 2009; Goldmann et al., 2010; Ni, Eichhorn, Griwodz, & Halvorsen, 2009; Wu et al., 2009). This line of research focuses on the subjective experience of users, typically from quality ratings or preference selections (ITU-T, 1998). In recent years, methodologies from cognitive psychology have been adopted, and adapted, to a similar line of research. With the newly coined term, Quality of Perception, researchers are now exploring perceptual measures for different quality impairments (Ghinea & Thomas, 2005); these approaches include eye-tracking (Gulliver & Ghinea, 2004; Ninassi, Le Meur, Le Callet, & Barba, 2007) and perceptual thresholds (Ni et al., 2011). Our investigations have focused on a single perceptual process, namely temporal integration (§11.1, §11.2, §11.3, and §11.4).

As outlined, by lowering the bitrate of an audio or video stream, the file size decreases and the download speed increases. As a consequence, the media quality goes down; moreover, the synchrony between streams can be affected. The impact such a trade-off will have on perception, and specifically on sensory integration, remains a largely unexplored research area. Yet there are always exceptions, in this case for frame rate. With lower frame rates, the temporal acuity of motion and other visual information is reduced, with adverse effects demonstrated for attention and content comprehension (Gulliver & Ghinea, 2006), speech comprehension (Knoche, Meer, & Kirsh, 2005), and lipreading (Vitkovitch & Barber, 1996). Seeing how frame rate and audiovisual asynchrony are artifacts that both affect the temporal nature of the content, the potential relationship between the two seems natural to explore. Indeed, a relationship was established in a study on temporal integration for audiovisual speech, with increasingly larger audio lags required to maintain subjective perception of synchrony (Vatakis & Spence, 2006c). However, as these and other studies
(Wilson & Sasse, 2000a, 2000b) demonstrate, reducing the frame rate of a video stream is rarely a desirable option for enhancing the speed of transmission.

With respect to SVC and active adaptation schemes, the two remaining video downscale options are then reductions to the spatial resolution or the quantisation parameter. The perceptual commonality between the two alternatives is the loss of spatial detail. Although the potential impact of losing detailed spatial information has previously not been studied for the perception of audiovisual asynchrony, clear detrimental effects have been established for several other perceptual processes. Loss of fine-grained visual details makes it more difficult to identify people and faces, familiar or unfamiliar (Bachmann, 1991; Burton, Wilson, Cowan, & Bruce, 1999; Gold, Bennett, & Sekuler, 1999), as well as text and letters (Gold et al., 1999; Marcel, 1983). For the auditory modality, excessive compression may serve to mask important spectral frequencies, which in turn may lead to distracting audio artifacts (Liu et al., 2008). As outlined in §3.2, audiovisual speech perception also suffers in the presence of auditory and visual disturbances. In short, loss of visual information eventually renders the auditory modality as the only reliable sensory source (Eg & Behne, 2009; MacDonald et al., 2000), whereas audio quality reductions lead to greater dependence on visual speech information (Grant & Seitz, 2000). Clearly, loss of detail or quality in either modality will make the sensory information ambiguous or even unreliable, and it puts strain on several perceptual processes. Yet, within the limits of deprived sensory signals that remain intelligible, the implications for audiovisual temporal integration are still unknown.
6. Loss of information in temporal integration

On the one hand, results from speech perception studies would indicate that the integration of audiovisual events depends on the quality, or salience, of the sensory signals. On the other hand, temporal integration is a separate perceptual process that may not require the same spectra of auditory and visual details than speech-related processes. As mentioned, research in this area is lacking, but some studies have looked at the role of stimulus intensity or saliency for more basic perceptual processes. From a set of experiments on the temporal perception of visual stimuli, it was found that in rapid serial presentations of two simultaneous light-flashes, the dimmest flash is more often than not perceived as coming before the brightest flash (Roufs, 1963; Bachmann, Põder, & Luiga, 2004). Moreover, the subjective rating of a light’s brightness becomes brighter when the light is accompanied by an auditory cue (Stein, London, Wilkinson, & Price, 1995). Granted, these ratings may reflect late-stage decisional processes rather than early sensory integration (Odgaard, Arieh, & Marks, 2003), yet the associated enhancement of the visual percept remains relevant in the current context. In another study on basic processes in the perception of auditory and visual information, Colavita (1974) discovered that presentations with both light and tone stimuli almost exclusively yielded responses to the light. This dominance of the visual modality endured even when the subjective ratings of corresponding brightness and loudness would indicate far higher intensity for the tone. Furthermore, with instructions to respond only to the tone during audiovisual presentations, participants persisted with pressing the light key for more than half of the trials. Accordingly, during simultaneous presentations of a light and tone, it seems that the visual signal is perceived before the auditory, to the extent that the tone is sometimes not registered at all (Colavita, 1974). While the isolation and simplicity of these auditory and visual stimuli make the findings difficult to generalise to more dynamic audiovisual events, the Colavita effect serves as a demonstration of an imbalance between the two modalities, at least for temporal perception. Combined with findings established from speech perception, these studies lay the foundation for a predicted influence of signal quality on audiovisual temporal integration.
6.1. Temporal integration in teleconference scenarios

With the many potential quality degradations that may arise from multimedia streaming, deprived sensory signals are certainly not experiences limited to laboratory settings. In most everyday settings, people are surrounded by wanted and unwanted media outlets with great varieties in content, intensity, and quality. Even in more specialised settings can the technology disturb the signal and the available perceptual information. One such setting is teleconferencing, which has become commonplace in both business and personal life. Due to the physical separation between two people communicating over a teleconference system, not only the lack of distortion, but also the coherence between the transmitted auditory and visual speech signal, are crucial to comprehension. This is particularly relevant to the larger teleconference platforms that are in use today, where groups of people can interact in an environment that approximates an extended meeting room.

The expansive arenas of the larger platforms and the enclosed spaces of the smaller platforms each bring with them distinct technological challenges to tackle, but reverberation is a problem common to many of these systems. Reverberation arises due to the acoustic response of an enclosure (ITU, 2009), which can be described as a temporal smearing of the audio. In short, two sound signals are combined, the signal that is conveyed directly from the main sound source and the signal’s echo that has been reflected and delayed by the enclosure (Assmann & Summerfield, 2004). Speech signals that are subjected to this temporal smearing are generally rated lower on experienced quality (de Lima et al., 2008), but reverberation affects more than the quality of an auditory signal. The signatures of the speech sounds themselves are altered (Assmann & Summerfield, 2004), for instance, through the distortion of the amplitude spectra of vowels (Culling, Summerfield, & Marshall, 1994). This loss of information can cause confusion regarding the arrival time of a speech signal, which in turn makes it more difficult for the perceptual system to discriminate the difference in arrival times between the two ears (Hartmann, 1983). This lateral precedence of one signal before the other, called the interaural time difference, serves as a localisation cue; the disturbance from reverberating sound components therefore makes it difficult to find the origin of a sound, and thereafter retain attention to it
(Culling et al., 1994; Darwin & Hukin, 2000). Hence, reverberation can be very detrimental to the natural ability to locate a speaker in a room full of people.

Of particular concern is the influence of reverberating acoustics on human speech, considering that teleconference systems are designed to facilitate spoken communication. Related to the mentioned distortion of vowel spectra (Culling et al., 1994), reverberation leads to more frequent vowel confusion compared to non-reverberant environments (Cox, Alexander, & Gilmore, 1987), for instance with combined vowels, or diphthongs, perceived as single vowels, or monophthongs (Nábělek, 1988). Confusion among consonants of different voicing and place of articulation can also occur because of reverberation (Cox et al., 1987), prominently so for consonants at the end of a word (Gelfand & Silman, 1979). Furthermore, irrelevant noises and voices in the background can be echoed and mixed in with the transmitted audio, providing additional disturbance and exaggerating the problem of reverberation in teleconferences. In §11.4 we disregard speech comprehension and consider instead the temporal disruption of reverberating acoustics on the perceptual integration of audiovisual speech, again within the temporal dimension.
7. Methodological considerations

In research on temporal integration, no clearly defined methodology has arisen as the set standard. Instead, new experiments are frequently based on approaches from similar studies, often combining features from different studies according to what can best be adapted to the problem at hand. In our investigations we have followed the examples of others and chosen to adapt the methodology to the relevant questions. Because our research questions target separate challenges, the experimental designs and methodologies do differ between some of the experiments, but they are all described in detail in §11.1, §11.2, §11.3, and §11.4.

7.1. Behavioural assessments of temporal integration

Studies on audiovisual asynchrony typically implement a behavioural task to assess the temporal integration between the modalities. At least four distinct experimental tasks have been implemented in past research, these include synchrony discrimination, asynchrony detection, simultaneity judgements (SJ), and temporal order judgements (TOJ) (Dixon & Spitz, 1980; Grant et al., 2003; Zampini et al., 2005; Zampini, Shore, & Spence, 2003). As seen from the outline in Table 2, all of these approaches require participants to make judgements on the temporality of audiovisual sequences. However, the great variations in windows of temporal integration for different events and experiments, highlighted in Figure 1, emphasise how much the task, setting, and content, influence the measure of temporal integration. Thus, the results from these experiments reflect not only the variables of interest, but also experimental implementations that differ between the majority of studies. Consequently, any attempt at cross-comparing research findings will be flawed.
Table 2. Short descriptions of four different experimental tasks designed to assess audiovisual temporal integration.

<table>
<thead>
<tr>
<th>Task</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Discrimination of synchrony</td>
<td>Two audiovisual stimuli presented in sequence, followed by a forced-choice selection of the stimulus that was perceived to be the most, or least, synchronous (Grant et al., 2003).</td>
</tr>
<tr>
<td>Detection of asynchrony</td>
<td>Audiovisual sequences deployed with gradual increases or decreases in asynchrony; participants reply when they perceive the change (Dixon &amp; Spitz, 1980).</td>
</tr>
<tr>
<td>Simultaneity judgement (SJ)</td>
<td>Audiovisual stimuli presented individually, with subsequent response selections to specify perceived synchrony or asynchrony between the modalities (Zampini et al., 2005).</td>
</tr>
<tr>
<td>Temporal order judgement (TOJ)</td>
<td>Audiovisual presentations with focus on precedence between modalities; participants indicate if the auditory or visual signal is perceived to come first (Zampini et al., 2003).</td>
</tr>
</tbody>
</table>

Considering that the wording and presentation of tasks and responses can influence subjective strategies (van de Par, Kohlrausch, & Juola, 2002), it follows that the experimental approaches may be tapping into slightly different perceptual mechanisms (Vatakis, Navarra, Soto-Faraco, & Spence, 2008). For instance, the most important distinction between the SJ and the TOJ task lies in the judgement of simultaneity versus successiveness (van Eijk et al., 2008). Because the focus on simultaneity should be a less demanding task, SJ will arguably provide a more stable measure of subjective synchrony, also referred to as the point of subjective simultaneity (PSS) (van Eijk et al., 2008). Conversely, the TOJ measure is assumed to be less biased towards a perception of synchrony since this neutral response option is not offered, and this could possibly make it a more sensitive temporal measure (Vatakis, Navarra, Soto-Faraco, & Spence, 2008). In other words, perfect temporal alignment may not be necessary for perceiving simultaneity between modalities, but the attention required to discern temporal order may make the perceptual system more sensitive to these misalignments (Grant et al., 2003). Yet, for the TOJ task, PSS is established as the point of
maximum uncertainty in temporal order and can therefore reflect an unconscious choice in response strategy, possibly skewing the subjective midpoint (van Eijk et al., 2008). With respect to external validity, which demands that the visual signal precedes the auditory, the literature on temporal integration would suggest that the SJ task obtains the most appropriate PSS measure (see summary table in Van Eijk et al., 2008).

Based on the outlined strengths and weaknesses of the SJ and TOJ tasks, we chose to use SJ in our investigations. This way, we aimed to make the numerous experimental trials a little less demanding for participants and hoped at the same time to obtain more externally valid measures of subjective simultaneity. Furthermore, in one experiment (§11.3), we compared the SJ task with the more uncommon approach of asynchrony detection.

### 7.2. Statistical measures of temporal integration

As mentioned, the experimental designs are not identical across the studies we have conducted. However, they all reflect three essential measures, the PSS, the temporal thresholds for audio lead and audio lag asynchrony, and the window of temporal integration, which is defined by the temporal thresholds. SJ thresholds can be derived from the 50% point on a Gaussian curve (the full-width at half-maximum) (Conrey & Pisoni, 2006; Petrini, Russell, & Pollick, 2009; Zampini et al., 2005), although some only report a more conservative sensitivity measure that corresponds to the standard deviation of the curve (Vatakis, Navarra, et al., 2008; Vroomen & Stekelenburg, 2011). In addition, some use other curve functions altogether (van Wassenhove et al., 2007). On the other hand, TOJ thresholds are normally calculated as the 25% and 75% points on a cumulative normal distribution (Vatakis, Maragos, Rodomagoulakis, & Spence, 2012). In discrimination studies, thresholds are determined by correctly identified stimuli, for instance at a rate of 70.7% (McGrath & Summerfield, 1985), and the PSS is calculated as the mid-point between thresholds (Fouriezos et al., 2007). Asynchrony detection thresholds are simply averages of the temporal offset presented at the time of detection (Dixon & Spitz, 1980); in our detection experiment (§11.3), we followed suit and set the PSS as the mid-point between
thresholds. For the SJ task, the PSS corresponds to the mean of a fitted Gaussian curve (Vatakis, Navarra, et al., 2008; Zampini et al., 2005), which is similar to the TOJ task where the mid-point of the cumulative distribution represents the PSS (van Eijk et al., 2008; Vatakis, Navarra, et al., 2008).

When calculating the PSS, we consistently used the means of fitted Gaussian curves for the SJ task. Yet, in our first line of investigation, we wished to separate our results from the PSS typically derived from TOJ measures. Because of this, we applied the label MPS (mean point of perceived synchrony) in §11.1 and §11.2, but switched to the more familiar PSS terminology in §11.3 and §11.4. The thresholds themselves are based on the two points of a Gaussian curve that cross at the 50% perceived synchrony, in §11.1, §11.3 and §11.4. In §11.2, we wanted to give recommendations on synchrony thresholds that could be applied to multimedia streaming and we found it prudent to use a more conservative measure; thus the thresholds reflect one standard deviation around the PSS. The absolute sum of the thresholds define the window of temporal integration, regardless of the statistical approach. As mentioned, the thresholds from the detection task in §11.3 are consistent with the temporal offset presented at the time of detection, averaged across participants, and the PSS is calculated as the average between the audio lead and the audio lag threshold. Finally, to evaluate statistically significant differences we used the repeated-measures analysis of variance (ANOVA) F-test, followed by paired comparison t-tests (Howell, 2002).
8. Research aims and outlines

The presented work is set mainly in the cross-section between cognitive psychology and computer science. We make use of basic research methodology and behavioural studies to consider potent challenges in contemporary multimedia technology. Our overall goal was to explore how the human perceptual system integrates auditory and visual information in scenarios that deviate from the natural settings the human senses have evolved to accommodate. By itself, any multimedia platform will remove natural cues that humans typically take for granted, but that the perceptual system may rely upon, such as spatial co-location. Moreover, the artificiality of multimedia settings can be exaggerated by quality distortions and other artifacts, as well as asynchrony introduced from the misalignment of auditory and visual streams. In our investigations of such applied scenarios, we narrowed down the focus to perceived audiovisual synchrony, thereby evaluating relevant research problems according to thresholds that define temporal integration. This approach allowed us to explore whether the temporal bond would vary depending on the nature of the audiovisual event and on external factors such as audio and video quality. Accordingly, we planned to generalise results from more stringent experiments to applied contexts and to supply the multimedia community with recommendations for acceptable temporal offsets. In the end, the conducted studies aimed to shed light on the perceptual ability to compensate for audiovisual separations in time and quality.

8.1. Paper 1 outline: Perceived synchrony in distorted speech

Paper 1 (§11.1) assesses the perception of synchrony for a long and a short speech episode in order to explore the effects of audiovisual quality on temporal integration, and the potential difference caused by the complexities in the spoken episodes. Speech stimuli were thus exposed to separations between the modalities both in the temporal domain and in the domain of signal saliency, in the form of quality loss. In theory, if the bond between auditory and visual speech is particularly robust due to the special nature of speech (Liberman & Whalen, 2000; Tuomainen et al., 2005), then the thresholds for temporal integration should be comparable across the two speech scenarios. Conversely, if the complexity of naturally produced speech takes attention away from the temporal
misalignment (Grant et al., 2003; Vroomen & Stekelenburg, 2011), then asynchrony should be detected sooner for the isolated speech syllable. Moreover, the inherent complexity of speech may be masked by quality distortions, which in turn may increase perceptual tolerance to asynchrony. Hence, another goal of this study was to investigate whether a concurrent separation in two dimensions, temporal and saliency, could have a more adverse effect on perceptual integration than temporal misalignments have on their own.

Audio and video quality reductions were considered in separate experiments to ensure that no confusion would arise from alternations between background noise and visual blur, and to avoid an unmanageable number of trials. A third experiment considered the combined effect of auditory and visual distortions. More noticeable than any effect of audiovisual quality was the remarkable perceptual resilience to distortions that at times approached the severe. With regards to the complexity of speech, the results consistently showed a more resilient temporal integration for the short, compared to the long speech episode. While this finding did not coincide with predictions, the explanation may simply be related to the predictability of a single spoken syllable. Furthermore, the vastly greater number of audiovisual events contained within the long speech episode provides numerous sensory cues to judge synchrony against. One speculation that arose from this study was the prominence of articulatory cues, and the possibility that the perception of temporal synchrony rests primarily on salient phonemes and visemes.

8.2. Paper 2 outline: Perceptual tolerance to asynchrony and quality loss

Rooted in the multimedia context, paper 2 (§11.2) focuses on two challenges in recent streaming approaches. With adaptive streaming over HTTP (Michalos et al., 2012), content providers have to consider the maximum quality downscaling that consumers will tolerate, but an old concern regarding the synchrony between audio and video streams has resurfaced (Petlund, Evensen, Griwodz, & Halvorsen, 2008). Considering the disruptions caused by quality degradations in speech perception and other perceptual processes (MacDonald et al., 2000), it
follows that the temporal integration of audiovisual events may be similarly affected. In a series of experiments, this study explores a potential interaction between asynchrony and audiovisual quality on temporal integration. To retain some of the typical characteristics of current media content for our experimental stimuli, we selected audiovisual sequences from three popular entertainment services. These represent three distinct audiovisual events, speech, action, and music. In multimedia streaming, downscaled quality is far more likely to be noticeable for the video than for the audio. Still, we included distortion in both domains in order to assess the importance of each modality in the perception of audiovisual synchrony.

The most prominent finding is the consistency in the windows of temporal integration, from no quality loss to severe quality loss. This previously observed temporal resilience to loss of perceptual information is a finding that repeats itself across our studies. Furthermore, the results revealed marked differences in the temporal perception of the three content types, with asynchrony in the speech sequence detected at shorter temporal offsets than for the action or music sequence. Although these results go against the works of those who have found comparatively robust temporal integration for audiovisual speech, they also reflect the complexities that are typical of modern multimedia content. A jam session on the drums or a chess piece moved across a board, these are not events presented in isolation; they bring with them a lot of irrelevant, and often distracting, information, which could possibly contribute to greater perceptual tolerance than that observed for more controlled experiments.

8.3. Paper 3 outline: Temporal integration of live speech

In paper 3 (§11.3), the aim was to compare temporal integration of audiovisual speech across two experimental scenarios. One scenario involves live conversations and makes use of the asynchrony detection task; the other scenario is set up in more controlled surroundings and approaches temporal integration through the SJ task. Accordingly, the results shed a hint of light on the generalisability and ecological validity of temporal integration windows derived from controlled experimental studies to those that should reflect the
unpredictability and disturbances of real-life settings. Because life is full of distractions that divert attention away from on-going conversations, any disparity in the speech conveyed is more likely to be overlooked. Thus, temporal offsets between the audio and video during a live teleconference could be harder to perceive than corresponding asynchrony presented in a behavioural task. Furthermore, the distinctions between the experimental approaches may also speak for a more conservative window of temporal integration using the SJ task, compared to asynchrony detection where the gradual change in synchrony could contribute to desensitisation.

Results from an explorative analysis of the temporal thresholds suggested that the window of temporal integration was wider for the detection task, but not significantly so. However, the distributions were not consistent in the directions they extended. Asynchrony in the live teleconference was detected at significantly longer audio lead thresholds compared to the two recorded speech sequences, which was also reflected in the PSS. From this, we surmised that the detection task could possibly increase the tolerance to audiovisual asynchrony, at least in the direction in which the perceptual system is particularly sensitive. Furthermore, it is equally feasible that the live setting, which encouraged spontaneous statements, contributed with random distractions that also increased the perceptual tolerance to temporal offsets.

8.4. Paper 4 outline: Temporal integration in reverberation

In order to address two relevant technical challenges within teleconferencing, paper 4 (§11.4) considers the potential interaction between reverberation and asynchrony. Detrimental effects of reverberation have already been established for auditory speech intelligibility (Cox et al., 1987; Nábělek, 1988; Sayles & Winter, 2008); instead, the presented study focuses on the temporal integration of visual signals with reverberant auditory signals. Based on findings that background noise will lead to a stronger visual dependence (Alm et al., 2009; Sumby & Pollack, 1954), it follows that the auditory disturbance from reverberation could introduce a similar visual reliance. Furthermore, reverberation smears the temporal signature of a sound (Assmann &
Summerfield, 2004), which leads to the expectation that it will create a disturbance to temporal perception and consequently affect the temporal integration of the auditory and visual modalities. These predictions are explored for two speech sequences and an action-oriented scenario, with reverberating audio tracks to simulate a small and a large teleconference room. The comparison between audiovisual content sheds light on the distinct influence of reverberation on synchrony perception according to the nature of the event, and the two sequences portraying separate speakers allow for a broader generalisability that is not limited to an individual’s speaking style. To understand how the cognitive load might increase with the auditory uncertainty created by reverberation, we included an additional measure on perceptual evaluation times; thereby comparing the time spent by participants on judging the synchrony of stimuli.

Similar to our previous experiments on temporal integration, we found that the perception of audiovisual synchrony in speech is very resilient to signal quality deteriorations. Neither the temporal thresholds nor the evaluation times reflected any considerable negative impact from the reverberant conditions. In contrast, the perception of synchrony for the action event was severely influenced by both types of reverberation, with the PSS, and window of temporal integration, shifting by more than 150 ms in the audio lead direction. The same skew is reflected in evaluation times, which are significantly longer for audio lead asynchrony, as well as shorter for audio lag asynchrony in reverberant conditions. Following the assumption that the upcoming speech sound serves to mask the reverberating components of the preceding sound, the robustness of temporal integration for audiovisual speech may again be attributed to its dynamic nature. Events that occur in isolation seem more likely to be affected by this temporal smearing, implicating a magnification of small disturbances. For a live teleconference, this could imply that irrelevant background noise may become more distracting if it is accentuated by reverberating echoes.
9. General discussion and conclusions

The presented work originated from one broad question: How can we evaluate the effect of media quality on audiovisual perception? Knowing that the perceptual integration of auditory and visual inputs rests on the convergence of several sensory cues, the first logical step appeared as an exploration of what happens when we introduce divergence between the cues. Going from the past to the present approaches to perceptual integration, described in §2, §3, and §4, we found it best to focus the planned work on one perceptual mechanism. Because the work would be set in the context of multimedia streaming and signal quality, we did not want to focus exclusively on speech perception. Other audiovisual events are equally prevalent in modern media, among them music and physical actions. Audiovisual events converge in at least two dimensions they all have in common, the spatial and the temporal dimension. Seeing that the spatial dimension would be less relevant within the context we had set, the choice fell on an investigation of the temporal integration of media content. Thus ensued our work on the perception of audiovisual synchrony across variations of media content and media quality.

9.1. Temporal integration, past and present

One consideration that has been continuously emphasised throughout this work is the subjective nature of temporal integration tasks and the great versatility observed across experiments and measures. Figure 1 provides a visualisation of the variability in windows of temporal integration from publications within this field; furthermore, it presents findings from two of our studies in the context of related research. Within the arguably generous parameters set by these experiments, it is clear that our thresholds of temporal integration are consistent with the previous findings of others. While this consistency will not bring the field any closer to narrow down the range of temporal thresholds, it does support the validity of our results. At the most extreme, the thresholds that define audiovisual temporal integration can extend beyond 300 ms in the audio lag direction (Maier et al., 2011; Vroomen & Stekelenburg, 2011) and 150 ms in the audio lead direction (Conrey & Pisoni, 2006; Eg & Behne, 2013; Fouriezos et al., 2007; Maier et al., 2011; Zampini et al., 2005). At the other extreme, temporal
integration thresholds have been established at offsets as short as 100 ms for audio lag (Lewkowicz, 1996; Vatakis et al., 2012; Vroomen & Stekelenburg, 2011) and 50 ms for audio lead asynchrony (Grant et al., 2003; Lewkowicz, 1996; Vatakis et al., 2012). In the end, the thresholds themselves may be less important than their applicability. Most researchers use asynchrony as a tool to explore the strength of the perceptual bond between modalities, hence the main interest lies in the difference between experimental conditions, not between past and present studies. In our own explorations we followed this exact approach and compared temporal thresholds across our variables of interest, thereby gaining an understanding of what integrates and separates the senses.

Moreover, temporal thresholds vary between measures. Some have speculated that the perceptual processes involved in judging simultaneity are not the same as those involved in judging temporal order (van Eijk et al., 2008; Vatakis, Navarra, et al., 2008). Indeed, comparisons of the two approaches show greater windows of temporal integration for SJ, along with PSS further in the audio lag direction (Maier et al., 2011). According to our findings (§11.3), thresholds derived from asynchrony detection are also distinct from the SJ task; the first provided a wider window of temporal integration than the latter. The greater temporal tolerance of the detection approach is possibly linked to a perceptual desensitisation that comes from the gradual increase in temporal misalignment. One consideration that is important to bear in mind is that temporal thresholds are determined from different measures. This means that temporal thresholds can vary both due to cognitive factors and due to statistical approximations. One study tackles this problem by including non-parametric statistics, in addition to the more common parametric analyses (Maier et al., 2011). Circumventing the forced adaptation to distribution functions and focusing solely on response proportions, Maier and colleagues (2012) present three assumption-free indices that can be compared across tasks, peak location, width, and asymmetry. Their findings point to subjective synchrony that is shifted more towards audio lag for SJ than for TOJ, consistent with the literature. Furthermore, interaction effects with larger increases in width and reductions in asymmetry indicate that SJ is a more sensitive measure than TOJ, at least for the temporal integration of speech. Future
endeavours of this kind could contribute to shed more light on the cognitive mechanisms that are responsible for perceiving synchrony and temporal order.

9.2. Temporal integration across contents and dynamics

Although it is not feasible to generalise across the variance seen in Figure 1, a few trends do emerge. For instance, in accordance with the notion that speech classifies as a special audiovisual process, the thresholds observed for audio lag asynchrony tend to extend further than for action-related and simple stimuli. In speech, the natural delay in arrival between a visual and an auditory signal may be further extended due to certain speech movements that precede voice onset (Grant et al., 2003). This learnt association might contribute to a particularly tolerant temporal integration for auditory speech signals that lag behind the visual speech movements. However, a similar perceptual tolerance to audio lag asynchrony is also seen for music-related stimuli, which brings back the notion of stimuli complexity. As with speech, the sounds and movements evoked when playing an instrument are rapid and dynamic. Hence, the commonality between speech and music lies in the fluctuating nature of the audiovisual events, which might divert attention away from any temporal offsets. On the other hand, when comparing perceived synchrony for a 13 second-long news broadcast with a single spoken syllable in §11.1, we found no indications that the isolated event of speaking a syllable would lead to a smaller window of temporal integration. Instead, the perceptual tolerance to asynchrony was greater for the short speech excerpt. Neither did we find a stronger temporal integration when comparing a long speech excerpt with a chess and a drumming sequence in §11.2.

We can only speculate on the cause, but based solely on these findings, it is not likely that the robust temporal integration typical of audiovisual speech is attributable only to the complexity of spoken language. One possibility is that the temporal perception of crossmodal signals aligns according to prominent events, such as spoken labial syllables or action events with sharp moments of impact. In longer speech excerpts, a number of prominent events are available for the continuous alignment of the modalities. Conversely, for music and action-oriented sequences, there may be only a few prominent events; furthermore, the
musical tempo may be too fast for the perceptual system to separate between events. A less causational viewpoint might assign the apparent differences in perceived synchrony between audiovisual events to mere coincidence. Yet, a few studies, including our own (§11.2), have run direct comparisons in within-subjects designs and found significant variations between speech, action, and even musical stimuli (Dixon & Spitz, 1980; Vatakis & Spence, 2006a, 2006b). Combined, the literature seems to leave us with more questions than answers.

To follow up these questions, future topics of investigation include a continuation of Vatakis and colleagues’ (2012) investigation of temporal integration across consonant-vowel combinations, to explore whether the syllabic differences remain present in continuous speech. Also, the complexity of spoken language can be assessed according to other features, such as speaking rate, word structure, or language familiarity (the latter has been explored for Spanish and English (Navarra et al., 2010)). Furthermore, to improve the ecological validity of experiments on temporal integration, more versatile stimuli should be included. For example, the action stimuli presented in past research are isolated events, typically with predictable moments of impact (Dixon & Spitz, 1980; Vatakis & Spence, 2006a); still, the physical actions encountered in real life, or presented as multimedia, do not necessarily culminate in a bang, knock or crash. Stimuli portraying knocking hammers, beating bats, and smashing blocks are likely selected due to the clear correspondence between the auditory and visual events. This opens the door for a new methodological challenge, which rests on the premise that speech and music are not the only continuous physical events that can be assessed according to their audiovisual synchrony.

9.3. **Temporal integration is not easily disturbed**

When low bandwidth necessitates severe compression, or multimedia quality deteriorates during another stage of streaming, the perceptual system loses sensory information. The detrimental effects of image and video distortions have been demonstrated for several perceptual processes, such as face and symbol recognition (Bachmann, 1991; Burton et al., 1999; Gold et al., 1999). Similarly, with noise or distortions in the auditory modality, speech comprehension becomes
more difficult and the influence of visual speech movements increases (Eg & Behne, 2009; Grant & Seitz, 2000; MacDonald et al., 2000). While these findings may not necessarily be relevant in the temporal dimension, others have found that the intensity of basic auditory and visual signals can influence the detection or perceived temporal order of events (Bachmann et al., 2004; Colavita, 1974; Roufs, 1963). Both the assumption of unity and the theory on the special nature of speech, along with learnt associations, implicate a more robust perceptual binding with greater coherence among sensory cues. If the convergence of coherent cues strengthens sensory integration, then it follows that divergence between modalities should weaken the bond, also for temporal integration. Working from this assumption, we ran a series of experiments on perceived synchrony for different audiovisual events with distorted quality of information in one or both modalities.

However, we found limited support for our prediction. Instead, our investigations in §11.1 and §11.2 showed repeatedly that the quality or saliency of a sensory signal has little impact on temporal integration. Indeed, synchrony responses remain fairly consistent throughout, even when the auditory or the visual event is close to being fully masked by noise or blur. Furthermore, the variations that were uncovered were not sufficiently systematic to propose any deductions. One trend that is hinted at from the wider response distributions, is a slight increase in the windows of temporal integration with the introduction of white noise. If this trend had been more pronounced, we could assume that the disturbance from the noise led to a more tolerant temporal bond. With auditory signals masked by the noisy frequencies, they become less salient, providing less information on any crossmodal events. Thus, it is possible that the temporal alignment of the modalities is burdened by the lack of clearly corresponding audiovisual events. Yet, such an inference cannot be made from the current set of results.

Only in our study on reverberation, §11.4, did we observe a marked influence of disturbing elements on temporal integration, although this influence was absent for the speech stimuli. However, the impact of reverberation on the perception of synchrony in the chess sequence was significant. Little difference is observed when comparing the effects of reverberation from a small and a large
teleconference room, but the contrast with the non-reverberant condition is remarkable. In the presence of reverberating acoustical components, participants judge the chess sequence to be in synchrony when the audio leads the video by more than 100 ms. Compared to the non-reverberant condition, and to findings from related research (Dixon & Spitz, 1980; Vatakis & Spence, 2006a), this is a sizeable skew away from the more natural audio lag direction. Audiovisual presentations with predictable moments of impact provide perception with time to anticipate the events, particularly with respect to the visual movement. In theory, this anticipation could make the perceptual system extra sensitive to any temporal misalignment; however, the nature and the portrayal of an event could be equally influential. In our selected chess sequence, the camera zooms slowly out and pans from an overhead view to a front view, gradually changing the angle of perspective. Consequently, the three first moves are observed at the closest distance, and these are all viewed from above. For this particular stimulus, it is possible that the depth estimation required for these moves makes it difficult to predict the moment of impact. Indeed, the time spent by participants when evaluating the chess sequence points to greater uncertainty in synchrony judgements when the auditory leads the visual signal; on average, participants spent more than 2 seconds longer on audio lead than on audio lag presentations. Possibly, the visual movement could permit an earlier capture of the auditory event, which would allow the remainder of the movement to capture the reverberating components that extend the event. This would explain both the shift in subjective synchrony, and the direction of the shift. Of course, this explanation is limited to very similar scenarios, but it is likely that the perception of other isolated events can be disturbed in the presence of reverberation. While the dynamics of speech seem to counteract reverberation, thus preserving temporal integrity, an isolated sound may not necessarily be followed by a new auditory signal that masks the temporal smear. In this way, reverberation could serve to accentuate the event.

Due to the lack of a pronounced effect of stimuli quality on the perception of synchrony, we make no assumptions regarding the related impact of signal saliency on the strength of the perceptual bond. Based on the numerous examples of the adverse effect of auditory/visual masking on speech perception, it seems
unlikely that our findings are related to an absence of perceptual impact. A more likely explanation attributes this near-absence to a resilient temporal perception that does not depend on fine-grained auditory and visual details to temporally align, and maintain, an auditory and a visual event. Additionally, the lack of an effect of audiovisual quality on temporal perception makes it difficult to put our findings in the context of a theoretical assumption. Based on the assumption of unity, we predicted that the temporal bond would weaken with additional separation between the modalities. This could either mean that the saliency of a signal cannot be considered along with other sensory cues that contribute to unite the modalities, or it could mean that the premise of the unity assumption does not hold. Unfortunately, neither explanation receives support from our presented results. We aim to extend our work beyond the temporal dimension to further investigate whether the perceptual integration of audiovisual content is hampered by the digital presentation of multimedia, where bandwidth restrictions can cause audiovisual quality to go up and down. In connection to this, we also aim to extend the applicability of the multimedia context by introducing more common compression artifacts, such as blockiness (Schwarz et al., 2007). Furthermore, we wish to apply our investigations to audiovisual events that are not related to speech. By exploring non-speech scenarios for other multisensory processes, such as spatial integration, future research could help clarify whether the detrimental effects of masked information are limited to the lack of comprehension and the articulatory confusion that typically occurs in speech perception.

9.4. Final thoughts
The presented work originated from a question on how multimedia quality affects audiovisual perception, and a related question on how to assess this. In a set of studies on perceived synchrony in multimedia, we have contributed with findings on how the human perceptual system handles loss of sensory information, while maintaining temporal integrity. In the end, we learnt that the perceptual system does this well. Temporal integration seems to be a very robust mechanism that can compensate for divergence in several dimensions, in this case, concurrent inconsistency in temporal alignment and signal saliency. Considering the extent of the distortions we applied, we find the perceptual system to be remarkably
resilient. When maintaining the subjective experience of audiovisual synchrony, perception can often make do with what little information is made available.

With regards to the context of this work, the minimal influence of audiovisual quality on perceived synchrony speaks in favour of the current streaming approaches. According to our observations, the temporal integrity of an auditory and a visual stream does not rely significantly on the quality of either modality. Despite the temporal dimension shared by asynchrony and reverberation, no interaction could be found between the two, at least not for speech content. If a relationship had been established between the quality of audiovisual information and asynchrony, it would be advisable to incorporate this consideration into existing compression, encoding and transmission solutions. However, the perception of synchrony is largely independent of the signal quality, which means that content providers need not add this as a concern. Of course, this does not remove loss of sensory information as a concern for other perceptual processes, with speech perception as a prominent example. Nor does it remove asynchrony as a concern, the temporal thresholds presented by us and others (see Figure 1) provide guidelines for the maximum offsets content providers should allow between audio and video streams.

Finally, on the contribution to research on audiovisual perception, we have observed windows of temporal integration that conform to earlier findings on related topics. While the perception of synchrony in speech is typically very resilient to temporal offsets, we found the same to apply to long-running action sequences and musical content. Moreover, the asymmetry commonly observed for the perceptual tolerance to temporal offsets is equally present in our results. These may not be major contributions on their own, but this work has taken another step towards more realistic scenarios by applying stimuli derived from broadcasted multimedia content. We aimed to improve the ecological validity of findings by taking some of the control away from the experiment, and from the consistencies in our results, we seem to have succeeded in this small endeavour.
10. References


11. Collection of papers

11.1. Paper 1: Special or not, speech is tolerant: Perceived audiovisual synchrony in distorted speech

Authors: Ragnhild Eg and Dawn Behne
Venue: *Perception*
Status: Manuscript submitted on February 12, 2014.
Special or not, speech is tolerant:
Perceived audiovisual synchrony in distorted speech

Ragnhild Eg & Dawn Behne

Abstract

Several studies have found the temporal integration of audiovisual speech to be more robust compared to other audiovisual events. This robustness could be attributed to many shared characteristics between auditory and visual speech, such as semantic, spatial, and articulatory cues, that bind the modalities. Alternatively, the temporal tolerance observed for speech could be related to the complexity of spoken language. To explore whether the robustness of temporal speech perception is dependent on corresponding cues to unite the modalities, this study explored the consequences of removing the congruence between auditory and visual signal saliency. By distorting the quality of the auditory or the visual signal, the resulting disparity in saliency aimed to separate the modalities. Simultaneity judgements were used to evaluate the perception of synchrony for two different speech episodes, a long, dynamic news broadcast and a short, controlled speech syllable. We found significant interactions for both auditory and visual distortion, but our quality manipulations contributed only to small variations in simultaneity judgements. On the other hand, the difference between speech episodes remained consistent, with asynchrony detected at longer offsets for the syllable than for the news broadcast. Our findings provide further support for the robustness of temporal integration in speech.
1 Introduction

The past two decades have seen a resurgence of research into multimodal perception of temporal synchrony, expanding the area with diverse topics such as synaesthetic congruency (Keetels & Vroomen, 2011), temporal frequency (Fujisaki & Nishida, 2005), and predictability of events (Cook, Van Valkenburg, & Badcock, 2011). This renewed interest has contributed to a growing body of research into the perception of audiovisual synchrony, and speech, with its bimodal nature, remains a prominent research topic (e.g., Maier, Di Luca, & Noppeney, 2011; Vatakis, Maragos, Rodomagoulakis, & Spence, 2012; Vroomen & Stekelenburg, 2011). Furthermore, asynchrony is not only a useful experimental tool, it is also an obstacle that is prevalent for many multimedia systems. Teleconferences are becoming commonplace in many work arenas and the use of such systems helps maintain communication and collaboration across distances, while cutting down on resources spent on travels. However, the incentive to use teleconference platforms will surely be tied to the user experience (Strohmeier, Jumisko-Pyykkö, & Kunze, 2010; Watson & Sasse, 1996), which again depends on the intelligibility and coherence of the audio and video (McCotter & Jordan, 2003; Reiter & Weitzel, 2007; Wijesekera, Srivastava, Nerode, & Foresti, 1999; Yap & Balota, 2007). The intelligibility of speech signals communicated through a teleconference platform is dependent on at least two key challenges, the synchrony between the auditory and visual signals (Grant & Greenberg, 2001; Staelens et al., 2012) and the quality of the signals (Fixmer & Hawkins, 1998; MacDonald et al., 2000; Ross, Saint-Amour, Leavitt, Javitt, & Foxe, 2007). This paper focuses on both challenges by establishing temporal ranges for non-noticeable audiovisual asynchrony in speech and exploring how these ranges vary as audio and video quality go down.

The detection of stimulus onset asynchrony (SOA) is not based on absolute thresholds, it takes the form of a dynamic range of intervals that are shaped by interacting events (Roseboom, Nishida, & Arnold, 2009). Prior research has established that the perceptual system is more sensitive to asynchrony when the auditory signal precedes the visual signal (Lewkowicz, 1996). Conversely, the perception of synchrony remains fairly robust when sound lags behind visual
events (Grant et al., 2003). Sensitivity to temporal offsets varies to some degree between individuals (e.g. Fouriezos et al., 2007), content type (van Eijk, Kohlrausch, Juola, & van de Par, 2008), content complexity (Arrighi, Alais, & Burr, 2006; Fujisaki & Nishida, 2005), and experimental set-ups (van de Par et al., 2002; van Eijk et al., 2008; Vatakis, Navarra, et al., 2008). The perceptual mechanisms at work in integrating audiovisual speech are also of great interest in the temporal domain (Massaro, Cohen, & Smeele, 1996; Navarra et al., 2010). The landmark work of Dixon and Spitz (1980) showed that the detection of gradually introduced asynchrony occurred at smaller displacements for an action-oriented video than for a speech video. In fact, their video portraying a hammer hitting a peg resulted in a temporal window where asynchrony was not detected between about 75 ms auditory lead and 188 ms auditory lag. When the video showed a narrator, the temporal window expanded to about 131 ms auditory lead and 258 ms auditory lag. Greater sensitivity to asynchrony for action events was also demonstrated in a study that compared the recordings of a male speaker with different segments portraying someone in the act of hitting an object with a tool (Vatakis & Spence, 2006a). Another study on the synchrony aspect of speech perception considered that the stronger integration observed for speech may be related to semantic or other associative cues between the auditory and visual modalities (van Wassenhove et al., 2007). Participants were presented with ordinary speech syllables and McGurk stimuli, which are incongruently matched auditory and visual speech syllables that differ in place of articulation and typically yield a fused percept of an intermediary syllable (McGurk & MacDonald, 1976). Van Wassenhove, Grant and Poeppel (2007) found a significantly larger temporal window of synchrony for congruent syllables than for incongruent syllables, proposing that the disparity between modalities for McGurk stimuli leads to a lower tolerance to asynchrony. With inconsistencies in both temporality and semantics, the study on perceived synchrony for McGurk stimuli reveals how the perceptual bond between modalities can be weakened by multiple disparities.

The robustness of the integration between sensory modalities is possibly related to the characteristics they share. For instance, the assumption of unity postulates that the presence of several coherent and redundant properties, consequently
associated with one source across modalities, strengthens sensory integration (Welch & Warren, 1980). In this way, human speech is united across sensory modalities by certain amodal factors, such as temporal and spatial co-occurrence, as well as semantic consistency (Vatakis, Ghazanfar, et al., 2008). With sufficiently consistent or salient cues, the unity may even overcome minor separations such as temporal offsets (Bertelson & Aschersleben, 2003; Vatakis & Spence, 2007). In support of the unity assumption, Tuomainen, Andersen, Tiippana, and Sams (2005) found that sine-wave representations of spoken syllables do not contribute to perceptual fusion when combined with video recordings of incongruent syllables. Only when participants are told that the sounds they hear represent speech do they perceive the syllables, and proceed to exhibit the McGurk effect. A given sound has a profoundly different contribution to sensory integration depending on how it is initially perceived; when understood as speech, the association between the sine-wave syllable and the visual syllable comes to modulate their integration. Furthermore, when comparing temporal order judgements for human speech sounds with that of two monkeys producing a grunt or a “coo” sound, performance was found to be significantly worse for the first (Vatakis, Ghazanfar, et al., 2008). Not only does this point to a stronger temporal integration for the human speech, it indicates that the effect is related to not merely any vocal production.

The tendency for perceptual integration to overcome inconsistencies between modalities is not exclusive to the semantic and temporal domains; it also influences perception when the auditory and visual signals are separated in other dimensions. The ventriloquist effect is an example of how the predicament of spatial separation is resolved by attributing the sound source to a different object than the actual origin (e.g., Jack & Thurlow, 1973). The same tendency is observed for more qualitative dimensions, such as intensity or saliency. Audiovisual speech presented in noise, or with other auditory distortions, is still perceived with contributions from both modalities (Sumby & Pollack, 1954). The association between visual and auditory speech cues is so strong that it extends to the point where visual contributions can enhance the auditory percept (Kim & Davis, 2004), as long as they are semantically consistent (Grant & Seitz, 2000). For the special case of the McGurk effect, noise tends to exaggerate the
perceptual illusion (Alm et al., 2009). Even at quite severe levels of auditory degradation is perception able to fuse the different syllables, thereby overcoming the incongruence by creating a percept that is consistent with both (Grauwinkel & Fagel, 2006). While auditory distortion may facilitate integration for incongruent audiovisual stimuli, degradations to the visual modality seem to enhance the inconsistency between the senses. In an earlier study on audiovisual speech integration with blurred faces, we found that visual speech cues became less salient with reduced pixel resolution, thus increasing the rate of responses matching the auditory stimulus (Eg & Behne, 2009). At 75x60 pixels, a small share of McGurk fusions was still reported, while at 45x36 pixels these were almost absent. The results are consistent with earlier findings by MacDonald, Andersen and Bachmann (2000). Conversely, the visual contribution to word identification has been found to persist for a range of spatial frequencies when combining band-pass filtered video sequences with clear audio signals (Munhall et al., 2004), although the most severe levels resulted in performance worse than for audio presented alone.

Auditory and visual distortions can clearly affect the integration of audiovisual speech cues, but that does not necessarily imply that the temporal integration of the two modalities is affected by the loss of sensory information. This is by and large an area yet to be explored. For the visual modality, research suggests that with rapid serial presentations of two simultaneous light-flashes, the dimmer of the two is likely to be perceived as preceding the other (Roufs, 1963; Bachmann, Põder, & Luiga, 2004). Furthermore, the intensity of a light is perceived to be brighter when it is presented with an accompanying auditory cue (Stein et al., 1995). Although this effect may be more related to late-stage decisional processes than early sensory integration (Odgaard et al., 2003), the implicated perceptual enhancement remains relevant. When simple auditory and visual signals are presented simultaneously without the possibility of enhancing, but rather impeding, the detection of the other, vision tends to win over audition; this occurs even when the subjective loudness of a tone is double that of the perceived brightness of a light (Colavita, 1974). Although the brevity and lack of dynamics in these simple stimuli make it difficult to generalise to longer and more complex stimuli, the study demonstrates a basic perceptual imbalance in the temporal
domain. Be the separation temporal, semantic or qualitative, the perceptual unity between the auditory and visual modalities should only endure up to a certain point. Beyond that, the two should be perceived as originating from independent sources.

While the integration of auditory and visual speech signals appears to be especially robust, the bond between modalities is not unquestionably related to the characteristics they share. The greater temporal tolerance associated with speech could also be connected to the complexity of spoken language. Where the production of words is a dynamic process, action events usually occur in isolation, with ample time to anticipate the moment of impact. Thus the difference may be attributed to the greater predictability of audiovisual action events, rather than intrinsic unity factors that are stronger for speech. Vroomen and Stekelenburg (2011) applied the same type of sine-wave speech as Tuomainen and colleagues (2005) to minimise the natural complexity of spoken language, and compared temporal order and simultaneity judgements for participants presented with natural speech, sine-waves not introduced as speech, or sine-waves introduced as speech. Response distributions were almost identical across conditions, even for those that were unaware of the speech representations. Unlike previous studies that have carried out comparisons between action and speech events, these results did not show a difference between speech and non-speech. The authors conclude that temporal order judgements do not depend upon the match between auditory and visual speech cues. This is consistent with the similarities in asynchrony detection between speech stimuli and simple circle-tone pairings presented by Conrey and Pisoni (2006). Also speaking against a special audiovisual binding for speech is a study that compared synchrony perception for natural and artificial speech (Grant et al., 2003). The absence of a mediating effect of the applied bandpass-filters could indicate that the associations between auditory and visual speech cues are no more special than other such pairings; however, the reverse can also be said. The bandpass-filters may not have sufficiently degraded auditory intelligibility and the cues that remained could have been sufficient to bind the two modalities together.

The extent to which the human perceptual system preserves synchrony in speech
may be attributed to spatial, semantic and other binding cues, or it may be related to difficulties in discerning the temporal complexity of speech. In order to explore this issue further, we conducted a study with two types of speech stimuli. A news broadcast served as a representation of the dynamic speech typical of everyday language, while the recording of the spoken syllable /ba/ offered a controlled, singular speech token. By choosing a bilabial stop consonant, we ensured that both the auditory and visual speech tokens would have high perceptual salience (Kent, 1997); moreover, a recent study has found greater temporal sensitivity for these salient speech sounds, likely due to shorter processing latencies (Vatakis et al., 2012). Consequently, the dynamics, and likely the saliency of speech tokens, will differ between the isolated speech syllable and the continuous news reading. Working from the assumption that discrepancy in an additional dimension should further separate the auditory and the visual speech signals, the subjective perception of synchrony is in turn expected to become less resilient to temporal offsets. We therefore introduced separation in the quality domain by manipulating the saliency of modality-specific signals using two distortion techniques, background noise and video blur. Auditory and visual saliency reductions were explored in separate experiments due to the potential perceptual confusion arising from alternations between modality distortions, and to keep the number of conditions at a manageable level. The two experiments applied reduced signal quality at different intensities to the visual modality and the auditory modality, respectively. A third experiment was run simultaneously with the second, this explored the combined effect of auditory and visual distortions. Through the design of the experiments, the study aimed to investigate whether separation in two dimensions (ie., temporality and saliency) could have adverse effects on the temporal integration of speech signals, hence reducing tolerance to temporal offsets. Conversely, if the temporal tolerance can mainly be attributed to the complexity of the dynamic nature of speech, then the added distortions may serve to mask some temporal events in the auditory or visual signal and in turn enhance the perceptual tolerance to asynchrony.
2 Method

Three experiments were designed to assess the potential interaction between temporal offset and signal salience disparity on audiovisual temporal integration in speech. Experiment 1 manipulated a discrepancy in signal salience by reducing the video quality by means of Gaussian blurring. Experiment 2 took a similar approach to reduce audio quality, with pink noise distortion applied to the auditory modality. Experiment 3 aimed to explore a potential combined effect of auditory and visual saliency reductions to see whether the applied distortions might cancel each other out.

2.1 Experiment 1

2.1.1 Stimuli and material
The speech stimuli outlined in this paper were presented together with other audiovisual sequences in a larger study on temporal integration for different contents. We used a short and a long speech episode to represent the types of speech stimuli used in similar studies (e.g., Conrey & Pisoni, 2006; Van Wassenhove et al., 2007), both are described in detail in Table 1. Unfortunately, the technology used to assess the perception of synchrony can itself introduce asynchrony. Misalignments of the audio and video tracks are almost unavoidable when the streams are encoded and compressed separately; however, the introduced variations can be controlled in retrospect with the correct equipment (Maier et al., 2011). We checked the accuracy of synchrony manipulations by considering the alignment between lip movements and the corresponding spectrograms, we did this for the release burst of the syllable and the frication of a v-initial word spoken at 7.7 seconds into the news sequence. To keep the conditions that could be controlled as equal as possible, the average audio intensity was set to 70 dB and resolution to 1024x576 pixels. Asynchronous audio tracks were adjusted in Audacity version 2.0.1 by editing out the selected asynchrony duration either at the beginning or at the end of the track. These durations corresponded to the lead or lag-time, so that the sound would start playing slightly sooner or slightly later with respect to the video track. The edited audio files were imported in the audio interchange file format (AIFF) to Final Cut Pro X and exported with the video track using the accompanying Compressor.
software and the H.264 encoder to convert files to QuickTime movies. Fade-ins and fade-outs were used to avoid giving away temporal cues; video onsets and offsets remained the same throughout for the same reason. Initially, we selected temporal offsets from prior research (Conrey & Pisoni, 2006; van Wassenhove et al., 2007; Vatakis & Spence, 2006a; Vroomen & Stekelenburg, 2011) and ran a pilot study with asynchrony levels increasing by 50 ms up to 500 ms audiovisual separation. Results from the pilot demonstrated the characteristic asymmetry in the detection and sensitivity to asynchrony where the auditory signal comes before the visual signal (audio lead) and asynchrony where the auditory signal comes after the visual signal (audio lag). Thus, the final selection of asynchrony levels were based on the distribution of synchrony responses and included the objectively synchronous condition, as well as SOAs of 50 ms, 100 ms, 150 ms, and 200 ms, for audio lead, and 100 ms, 200 ms, 300 ms, and 400 ms, for audio lag. Gaussian blur was chosen to distort the visual signal due to its ability to filter out high spatial frequencies and fine details while preserving global outlines. Following another pilot, blur filters were applied at the three levels included in Figure 1, 2x2 pixels, 4x4 pixels, and 6x6 pixels, in addition to the undistorted condition. This made a total of 72 stimuli, all repeated twice. Videos were presented using Superlab 4.5 running on 24” iMac 7.1 computers with monitor resolution of 1920x1200 pixels. Audio was presented through AKG K271 circumaural headphones and responses were recorded by Cedrus RB-530 response pads.
Table 1. Brief stimuli descriptions of the long and short speech episodes included in the study.

<table>
<thead>
<tr>
<th>Long speech episode: News</th>
<th>Short speech episode: Syllable</th>
</tr>
</thead>
<tbody>
<tr>
<td>The Norwegian news</td>
<td>The syllable /ba/ is spoken by</td>
</tr>
<tr>
<td>broadcast shows an anchor-</td>
<td>a Norwegian female, filmed so</td>
</tr>
<tr>
<td>woman shot from the chest</td>
<td>that her head, neck and</td>
</tr>
<tr>
<td>up, with the same scene</td>
<td>shoulders fits within the</td>
</tr>
<tr>
<td>composition throughout.</td>
<td>frame. The total duration of</td>
</tr>
<tr>
<td>During the 13-second-long</td>
<td>the clip is set to 1 second.</td>
</tr>
<tr>
<td>segment, she presents a</td>
<td>The vocal articulation starts</td>
</tr>
<tr>
<td>piece of news on the</td>
<td>after 200 ms and lasts</td>
</tr>
<tr>
<td>return of some football</td>
<td>approximately 420 ms. The</td>
</tr>
<tr>
<td>players. The broadcast</td>
<td>clip is part of a series of</td>
</tr>
<tr>
<td>was provided by the</td>
<td>spoken syllables recorded at</td>
</tr>
<tr>
<td>National Library of Norway,</td>
<td>the Speech Lab, NTNU.</td>
</tr>
<tr>
<td>for research purposes.</td>
<td></td>
</tr>
</tbody>
</table>

Figure 1. Photo representations of the syllable stimuli /BA/, articulated by a female Norwegian speaker. The photos demonstrate the visual effect of applying Gaussian blur at 2x2 pixels, 4x4 pixels, and 6x6 pixels, along with the original resolution.

2.1.2 Participants
The experiment was completed by 19 participants (5 males, 14 females) aged between 19 and 41 (M=22.63, SD=4.79). All were native Norwegian speakers with self-reported normal hearing and normal or corrected vision.
2.1.3 Procedure

The current study is part of a larger project on temporal integration that was carried out in the Speech Lab at the Norwegian University of Science and Technology. Participants’ task was to evaluate audiovisual presentations on perceived synchrony. The experiment ran over two sessions, set one week apart. Each session included one repetition of every stimulus; these were randomised separately for all participants and between the individual sessions. At the beginning of the first session, participants were requested to read through an information sheet, give written consent, and provide some background information. They then received instructions on how the experiment would progress and how to handle the headphones and the response pad. The latter included two labelled buttons, “SYNC” for presentations perceived to be synchronous and “ASYNC” for presentations perceived to be asynchronous. Two practice trials commenced the experiment and a break was included halfway through. Participants progressed at their own pace, thus the total duration of each session varied between 50 and 70 minutes.

2.2 Experiment 2

2.2.1 Stimuli and material

The stimulus material consisted of the same speech stimuli described for Experiment 1, this time with distortion introduced to the auditory modality. Although white noise is common in speech research, for this study, the decision fell on pink noise to mask the speech signal. The choice was motivated by the nature of the frequency distributions for the two types of noise. Where white noise covers all frequencies with equal probability, pink noise \((1/f)\) noise gives higher weighting to low frequencies, making the pattern more dynamic and yielding a spectrum more similar to speech (Voss & Clarke, 1978). The levels of audio distortion were again narrowed down following a pilot study, with final pink noise levels set to 60 dB, 70 dB and 80 dB, as well as the original undistorted audio track. Audio signals with noise were normalised to 70 dB, so that only the signal-to-noise ratio varied between conditions. Figure 2 shows the sound waves of the speech syllable, presented in quiet and in noise.
2.2.2 Participants
Ten males and 15 females aged between 18 and 33 years volunteered for the two experiments ($M=24.36$, $SD=3.85$), all had Norwegian as their native language and reported normal hearing and normal or corrected vision.

2.2.3 Procedure
The procedure was the same as described for Experiment 1. Experiments 2 and 3 were run one after the other, with the order counterbalanced across participants.

2.3 Experiment 3
2.3.1 Stimuli and material
Experiment 3 combined the distortion manipulations described for Experiments 1 and 2. Gaussian blur and pink noise were combined using the two highest levels of each to create four new distortion levels. This excluded 2x2 pixel blur and 60 dB noise, along with the two smallest temporal offsets, 50 ms audio lead and 100 ms audio lag.

Figure 2. Sound wave representations of the syllable /BA/, articulated by a female Norwegian speaker. After the introduction of noise the auditory signals were normalised back to 70dB, thus the sound waves illustrate the masking effect of pink noise at 60dB, 70dB, and 80dB, in addition to the quiet condition.
2.3.2 Participants
The experiment was run with the same 10 male and 15 female participants described for Experiment 2.

2.3.3 Procedure
The experiment proceeded in the same manner as described for Experiment 1. Experiments 2 and 3 were run one after the other, with the order counterbalanced across participants.

3 Results

Responses were treated as synchrony match or synchrony non-match and thereafter converted to percentages. Repeated-measures ANOVA F-tests were run for the initial analyses, with effect sizes calculated using the partial eta-squared statistic ($\eta^2$). Differences between factor levels were deduced from graphs plotted with standard errors of the mean. The interactions between asynchrony and distortion were further explored by calculating theoretical distributions using Gaussian curve fittings. Curves were plotted with SOA times along the x-axis and mean perceived synchrony along the y-axis, maximum amplitude was constrained to 100% perceived synchrony. The distributions were fitted across participants due to the insufficient number of factor level repetitions, thus making the commonly used point of subjective simultaneity (PSS) inappropriate. The approach reported here, using mean point of synchrony (MPS), is adapted from Conrey and Pisoni (2006). Reported results include standard deviations (SD), full-width at half-maximum (FWHM), and goodness of fit ($R^2$), for the individual curves. Extra sum-of-squares F-tests indicated whether the models of best fit were best described with a single curve, or with separate curves for each distortion level.

3.1 Experiment 1
3.1.1 Speech type
Perception of simultaneity was found to differ between the short and the long speech episodes, with an overall main effect [$F(1,18) = 41.38, p<.001, \eta^2 = .70$] and a significant interaction with SOA [$F(8,144) = 8.84, p<.001, \eta^2 = .33$].
Figure 3 illustrates the wider distribution observed for the spoken syllable as compared to the news broadcast; the relationship suggests an audiovisual integration more tolerant to temporal offsets for the isolated speech sound. The most extreme difference between the two distributions became apparent when the auditory signal led before the visual by 100 ms, at which approximately 20% of responses reported the news broadcast to be synchronous, as opposed to nearly 65% for the syllable.

![Figure 3](image)

*Figure 3.* Mean percent synchrony perceived at the presented SOAs, averaged across auditory distortion levels and plotted separately for news and syllable stimuli. Error bars represent standard errors of the mean.

3.1.2 Asynchrony and distortion

With synchrony responses distributed nearly across the entire scale and corresponding well with previous research, the main effect for asynchrony was as expected \( F(8,144) = 186.44, p<.001, \eta^2 = .91 \). Blur distortion did not contribute with an overall effect \( F(3,54) = 1.03, \text{ns} \), nor did it vary significantly with speech type \( F(3,54) = 2.09, \text{ns} \). However, it did vary with SOA \( F(24,432) = 1.76, p<.015, \eta^2 = .09 \). Considering the main effect of speech type, along with the significant interaction between blur and SOA, the absence of a significant three-way interaction was a surprise \( F(24,432) = 0.95, \text{ns} \). Still, with the
different response distributions yielded by the short and long speech episodes, the relationship between asynchrony and distortion was considered separately for the two stimuli types. As seen in Table 2, the Gaussian curve fittings further established that this was not a straightforward relationship. Both extra sum-of-squares F-tests revealed that the models for the theoretical distributions were best explained with a single curve for both the news \([F(9,24) = 0.18, \text{ ns}]\) and the syllable \([F(9,24) = 0.25, \text{ ns}]\) episodes. Figure 4 presents the temporal thresholds at which 50 % of responses correspond to perceived synchrony, showing the slight effect of visual signal saliency on temporal integration.

Table 2. Statistical results from Gaussian curve fittings with mean percent perceived synchrony plotted across SOA, run separately for news and syllable stimuli and for all levels of visual distortion. Results include mean point of synchrony (MPS), standard deviation (SD), full-width at half-maximum (FWHM), and goodness of fit \(R^2\).

<table>
<thead>
<tr>
<th></th>
<th>Blur</th>
<th>MPS</th>
<th>SD</th>
<th>FWHM</th>
<th>(R^2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>News</td>
<td>None</td>
<td>72</td>
<td>112</td>
<td>263</td>
<td>.93</td>
</tr>
<tr>
<td></td>
<td>2x2 px</td>
<td>73</td>
<td>115</td>
<td>272</td>
<td>.92</td>
</tr>
<tr>
<td></td>
<td>4x4 px</td>
<td>79</td>
<td>118</td>
<td>278</td>
<td>.93</td>
</tr>
<tr>
<td></td>
<td>6x6 px</td>
<td>89</td>
<td>116</td>
<td>272</td>
<td>.94</td>
</tr>
<tr>
<td>Syllable</td>
<td>None</td>
<td>74</td>
<td>172</td>
<td>406</td>
<td>.90</td>
</tr>
<tr>
<td></td>
<td>2x2 px</td>
<td>69</td>
<td>152</td>
<td>357</td>
<td>.93</td>
</tr>
<tr>
<td></td>
<td>4x4 px</td>
<td>68</td>
<td>170</td>
<td>399</td>
<td>.92</td>
</tr>
<tr>
<td></td>
<td>6x6 px</td>
<td>72</td>
<td>168</td>
<td>396</td>
<td>.93</td>
</tr>
</tbody>
</table>
Figure 4. 50% perceived synchrony thresholds derived from Gaussian curve fittings run separately for auditory distortion levels. Thresholds represent the temporal offset (in ms) required for synchrony to be reported at chance level, on average across participants. Thresholds are grouped according to stimuli, (a) presents results for the news broadcast and (b) presents results for the spoken syllable.

3.2 Experiment 2

3.2.1 Speech type

Perception of synchrony was found to differ between the short and the long speech episodes \( F(1,24) = 64.01, p<.001, \eta^2 = .73 \). Moreover, the response distributions seen in Figure 5 for the long news clip and the short speech syllable are visibly distinct from each other, also supported by the significant interaction between speech type and SOA \( F(8,192) = 19.61, p<.001, \eta^2 = .45 \). This is predominantly observed for auditory lead stimuli, where more than a third of the responses reflect an incapacity to detect asynchrony for the speech syllable at the most extreme temporal offset of 200 ms. More than 50% of responses report that the short speech episode is perceived as synchronous when presented with 150 ms auditory lead, while a corresponding percentage is found between 50 ms and 100 ms for the long speech episode. Overall, the temporal perception of the short syllable is seemingly more tolerant to asynchrony than that of the news broadcast.
Figure 5. Mean percent synchrony perceived at the presented SOAs, averaged across visual distortion levels and plotted separately for news and syllable stimuli. Error bars represent standard errors of the mean.

3.2.2 Asynchrony and distortion

The share of synchrony responses were again spread out across most of the scale, with a predictable main effect of asynchrony \([F(8,192) = 156.43, p<.001, \eta^2 = .87]\). Auditory distortion had no significant impact on synchrony perception on its own \([F(3,72) = 1.82, ns]\), but it did vary between the short and long speech episodes \([F(3,72) = 9.57, p<.001, \eta^2 = .29]\). While the two-way interaction between asynchrony and noise was not found to be significant \([F(24,576) = 1.30, ns]\), the three-way interaction between asynchrony, noise and speech type was \([F(24,576) = 2.93, p<.001, \eta^2 = .11]\). Despite the significant three-way interaction, the extra sum-of-squares F-tests for the curve fittings suggested that the variation was best explained by Gaussian models collapsed across auditory distortion levels for news \([F(9,24) = 0.17, ns]\), as well as for the syllable \([F(9,24) = 1.60, ns]\). Table 3 summarises the remaining results from the Gaussian curve fittings, while Figure 6 shows the 50% perceived synchrony thresholds. Going by the significant three-way interaction, some variation in response distributions can be attributed to the different levels of auditory distortion. FWHM values are greater with noise at 80dB than in the absence of noise, pointing to a stronger temporal integration as the modalities are further separated by saliency.
disparities. If this difference had been more pronounced, the most likely cause would be attributed to an auditory masking of salient speech cues. However, these are only slight tendencies and the contradiction between the three-way interaction and the best-fit models is likely related to significant, but not sufficiently systematic, variations in responses.

*Table 3.* Statistical results from Gaussian curve fittings with mean percent perceived synchrony plotted across SOA, separated by speech and noise conditions. Results include mean point of synchrony (MPS), standard deviation (SD), full-width at half-maximum (FWHM), and goodness of fit ($R^2$).

<table>
<thead>
<tr>
<th>Noise</th>
<th>MPS</th>
<th>SD</th>
<th>FWHM</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>News</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>None</td>
<td>59</td>
<td>119</td>
<td>281</td>
<td>.93</td>
</tr>
<tr>
<td>60 dB</td>
<td>61</td>
<td>117</td>
<td>275</td>
<td>.92</td>
</tr>
<tr>
<td>70 dB</td>
<td>64</td>
<td>119</td>
<td>280</td>
<td>.92</td>
</tr>
<tr>
<td>80 dB</td>
<td>62</td>
<td>133</td>
<td>313</td>
<td>.87</td>
</tr>
<tr>
<td>Syllable</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>None</td>
<td>29</td>
<td>170</td>
<td>399</td>
<td>.94</td>
</tr>
<tr>
<td>60 dB</td>
<td>34</td>
<td>191</td>
<td>450</td>
<td>.96</td>
</tr>
<tr>
<td>70 dB</td>
<td>28</td>
<td>176</td>
<td>415</td>
<td>.97</td>
</tr>
<tr>
<td>80 dB</td>
<td>43</td>
<td>178</td>
<td>418</td>
<td>.97</td>
</tr>
</tbody>
</table>
Figure 6. 50% perceived synchrony thresholds derived from Gaussian curve fittings run separately for visual distortion levels. Thresholds represent the temporal offset (in ms) required for synchrony to be reported at chance level, on average across participants. Thresholds are grouped according to stimuli, (a) presents results for the news broadcast and (b) presents results for the spoken syllable.

3.3 Experiment 3

3.3.1 Speech type
As before, the main effect of speech type \[ F(1,24) = 59.79, \ p<.001, \ \eta^2 = .71 \], and the interaction between speech type and SOA \[ F(6,144) = 21.85, \ p<.001, \ \eta^2 = .48 \], indicate a distinction between the two types of speech stimuli. As seen in Figure 7, asynchrony is detected at shorter offsets for the news broadcast than for the speech syllable, but predominantly for lead asynchrony.
3.3.2 Asynchrony and distortion

The main effect of asynchrony \([F(6,144) = 128.53, \ p<.001, \ \eta^2 = .84]\), again points to an unsurprising change in perception of synchrony with increasing temporal offsets. No main effect of the two-modality distortion was uncovered \([F(4,96) = 2.72, \ ns]\), neither was an interaction between distortion and SOA \([F(24,576) = 1.45, \ ns]\). However, an interaction between distortion and speech type was revealed \([F(4,96) = 2.96, \ p<.05, \ \eta^2 = .11]\), along with a three-way interaction between SOA, speech type and two-modality distortion \([F(24,576) = 2.23, \ p<.001, \ \eta^2 = .09]\). Again the curve fittings yielded a single Gaussian model for each speech type, as seen in the former two experiments, with non-significant extra sum-of-squares F-tests for both news \([F(12,20) = 0.42, \ ns]\) and syllable \([F(12,20) = 0.96, \ ns]\). Table 4 summarises the curve data and Figure 8 shows the temporal thresholds for all distortion level combinations. The results for the syllable show the same tendency as observed for Experiment 2. An interesting trend to note is the increasing FWHM values observed for the syllable conditions where blur distortion goes from 4x4 to 6x6 pixels, but pink noise is kept at the same level. These hint at a combined masking effect where the most detrimental visual blurring may have interacted with noise in the auditory signal to decrease

*Figure 7.* Mean percent synchrony perceived at the presented SOAs, averaged across audiovisual distortion levels and plotted separately for news and syllable stimuli. Error bars represent standard errors of the mean.
the prominence of the articulatory cue. This may also explain the absence of a similar trend for the news broadcast, since this provided several articulatory cues for perception to temporally align the auditory and visual signals. The distributions are nevertheless not sufficiently divergent to make further inferences.

*Table 4.* Statistical results from Gaussian curve fittings with mean percent perceived synchrony plotted across SOA, separated by speech condition and for all levels of audiovisual distortion. Results include mean point of synchrony (MPS), standard deviation (SD), full-width at half-maximum (FWHM), and goodness of fit ($R^2$).

<table>
<thead>
<tr>
<th>Distortion</th>
<th>MP</th>
<th>SD</th>
<th>FWHM</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>70</td>
<td>116</td>
<td>273</td>
<td>.94</td>
</tr>
<tr>
<td>News</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>70 dB, 4x4 px</td>
<td>80</td>
<td>123</td>
<td>291</td>
<td>.97</td>
</tr>
<tr>
<td>70 dB, 6x6 px</td>
<td>83</td>
<td>124</td>
<td>292</td>
<td>.97</td>
</tr>
<tr>
<td>80 dB, 4x4 px</td>
<td>68</td>
<td>135</td>
<td>317</td>
<td>.92</td>
</tr>
<tr>
<td>80 dB, 6x6 px</td>
<td>74</td>
<td>125</td>
<td>294</td>
<td>.98</td>
</tr>
<tr>
<td>Syllable</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>None</td>
<td>29</td>
<td>169</td>
<td>397</td>
<td>.91</td>
</tr>
<tr>
<td>70 dB, 4x4 px</td>
<td>38</td>
<td>180</td>
<td>424</td>
<td>.95</td>
</tr>
<tr>
<td>70 dB, 6x6 px</td>
<td>20</td>
<td>223</td>
<td>526</td>
<td>.93</td>
</tr>
<tr>
<td>80 dB, 4x4 px</td>
<td>17</td>
<td>200</td>
<td>471</td>
<td>.95</td>
</tr>
<tr>
<td>80 dB, 6x6 px</td>
<td>32</td>
<td>207</td>
<td>488</td>
<td>.96</td>
</tr>
</tbody>
</table>
Figure 8. 50% perceived synchrony thresholds derived from Gaussian curve fittings run separately for audiovisual distortion levels. Thresholds represent the temporal offset (in ms) required for synchrony to be reported at chance level, on average across participants. Thresholds are grouped according to stimuli, (a) presents results for the news broadcast and (b) presents results for the spoken syllable.

4 Discussion

Judging from the substantial body of research into the perceptual consequences of temporal misalignments between sensory modalities (e.g., Grant & Greenberg, 2001; Staelens et al., 2012), it follows that the technical challenge of asynchrony in multimedia systems can have dire outcomes for speech intelligibility. While some have found that the perception of audiovisual speech is very resilient to temporal offsets (e.g., Dixon & Spitz, 1980), others have not found speech perception to be more robust to asynchrony than other audiovisual stimuli (Conrey & Pisoni, 2006; Grant et al., 2003; Vroomen & Stekelenburg, 2011). The robust temporal integration of speech could arguably be attributed to shared amodal characteristics; speech is proposed to have especially strong audiovisual associations binding the two modalities (Vatakis & Spence, 2007). Extending beyond the temporal and spatial domain, speech is also linked by semantic and articulatory cues (Vatakis, Ghazanfar, et al., 2008); combined, they all serve to maintain a consistent and coherent perceptual experience (Welch & Warren,
1980). On the other hand, in the studies where speech stimuli were not differentiated from other stimuli, the experimental set-ups typically compared speech to audiovisual events with dynamics similar to that of speech (Grant et al., 2003; Vroomen & Stekelenburg, 2011). The basis for the current set of studies was motivated by the plausibility of an alternative account; the greater temporal tolerance generally observed for speech stimuli may stem from the complex nature of speech. Where a motion event typically has a distinct progression, often culminating in a foreseeable moment of impact, speech is constant and dynamic. So the frequency of changes in the audiovisual cues contained within speech may be too rapid for small temporal offsets to be noticed. To extend upon the body of research that has collectively set out to determine what makes speech special, the current study has taken a different approach to the matter at hand. We worked from the hypothesis that a disparity in signal saliency would contribute to separate the sensory modalities and reduce the perceptual tolerance to temporal offsets. Furthermore, by including two speech stimuli that differed, most importantly, in complexity, we could also investigate whether speech dynamics could influence the perceptual tolerance to asynchrony. A news broadcast provided a representative example of the type of dynamic speech encountered every day, while a short spoken syllable offered a controlled version of a common speech sound.

Results showed a consistent difference in distributions between the two speech types. Perception was found to be more tolerant to temporal offsets for the short than for the long speech episode, particularly with the auditory preceding the visual signal. For the syllable, the time between the commencement of the video and the onset of the auditory articulation is admittedly very limited when the auditory signal leads the visual. Yet, with 150-200 ms available for the shorter SOAs, this could not be the only contributing factor to the greater proportion of synchrony responses for the short speech episode. If previously established differences in the temporal perception of speech and other audiovisual events are mainly due to the dynamic nature of speech, then we should have found a stronger tolerance to asynchrony for the news broadcast, compared to the syllable. Conversely, if the temporal perception of speech is especially resilient due to strong associations between auditory and visual speech cues, then results
should have been similar for the two speech scenarios. We found no support for either proposition. More likely, the stronger perceptual tolerance to asynchrony observed for the syllable could be explained by the single prominent articulatory cue on which to base simultaneity judgements, which did not provide sufficient information for accurate temporal perception. Conversely, for the long speech episode there are several articulations, possibly providing more temporal references and improving perceptual sensitivity to asynchrony. While any inferences drawn from two isolated speech examples are bound to be speculative, the current findings are not consistent with an explanation model that attributes the especially strong perceptual tolerance to temporal offsets in speech to the complexity and continuity of spoken language.

Considering the level of degradation applied to the visual modality, the lack of a clear effect of blur on temporal integration bears witness to the great capacity of the perceptual system to make use of what little information remains available. Fine-grained visual details are seemingly not essential to the temporal processing of events; this raises a question on whether global outlines could be sufficient to perceive the timing of speech movements. Although this question cannot be considered with respect to the presented results, it is one that would be worth exploring. When looking at the overall effect of auditory distortion, its impact was also modest. For the long speech episode, a slight increase in the window of temporal integration followed the increase to 80 dB noise. On the other hand, the effect of pink noise on the detection of asynchrony for the short speech episode was more immediate, showing a marked increase in the window of temporal integration already at 60 dB. While more variations were observed between auditory distortion levels for the speech syllable, the windows of temporal of integration were greater for all noise levels, compared to the quiet condition. The limited effect of noise on the temporal perception of the news episode may again be attributed to the greater number of articulatory cues available. Conversely, the short available processing time and the isolation of the articulation may have contributed to enhance the auditory masking effect for the speech syllable.

As for the combination of auditory and visual distortions, tendencies toward a masking effect were again noted. The combination of auditory and visual distortions contributed to stronger temporal integration when visual blurring was
at its worst, but as before, predominantly for the speech syllable. This seemingly
greater effect of audiovisual distortion could be an indication that the two-
modality masking accomplished what noise and blur could not do alone. If so,
then the masking of both auditory and visual articulatory cues may have
strengthened the bond and the temporal integration by reducing the saliency of
both modalities. Either way, none of the plotted theoretical distributions differed
significantly from the others; in fact, the best-fitting models were always a single
fitted Gaussian curve.

Overall, the auditory and visual masking of noise and blur did not contribute to
clear separations of the auditory and visual modalities, at least not with respect to
temporal perception. This was a surprising finding, considering the impact
sensory salience has been shown to have on other perceptual processes. Although
possible, it is unlikely that this lack of unambiguous effects is related to the
selection of distortion levels, at least not for noise. The selections were based on
pilot studies and the most extreme levels were chosen on the merits that the
signals should remain intelligible, but just barely. For noise, this could be double-
checked by inspecting the spectra of the auditory signals. Alternatively, temporal
perception could be less dependent on the articulatory cues masked by Gaussian
blur and pink noise than is the case for speech comprehension. If so, it may
suggest that the perceptual binding of audiovisual speech, across temporal offsets,
depends not on the ever-changing speech cues, but more on prominent
articulatory events. The other main finding was the greater perceptual tolerance to
asynchrony for the syllable. This resilience might also be related to a perceptual
strategy in temporal perception that does not rely on the constant dynamics of
speech. Another interesting implication arises from the established detection
thresholds; across almost all conditions, the auditory signal had to lag behind the
visual signal with more than 200 ms before the asynchrony was detected at
greater than chance levels. It seems that the perceptual tolerance to temporal
offsets in speech is indeed very tolerant, and this does not change with the
introduction of disturbing elements. The production of speech involves a complex
range of facial movements, centred around the mouth, coupled with equally
complex sounds with interchanging pitch, voicing, formants, and others. All these
cues are united in the perception of speech to create coherent strings of

87
information. Yet the temporal aspect need not require all the finer details that are necessary to understand meaning. It is beyond the scope of this study to make assumptions on the nature of the suggested key events, with one exception. More than likely, these events would have to be rich in both auditory and visual information, thus providing salient cues to align the two modalities in time. The syllable /ba/ was chosen as a stimulus for those very reasons, and perhaps those cues served to better unite the auditory and visual modalities for that particular speech sound.

The current study could find no indications that the strong temporal integration observed for speech should be related to its dynamic nature, nor did the study contribute with significant support to the assumption that the integration of speech is especially strong due to shared amodal characteristics. However, the findings demonstrate the robustness of audiovisual temporal integration and they indicate that temporal alignment may primarily depend on prominent articulatory events. This could also explain the difference found between speech types, where the temporal integration was stronger for a short speech syllable than for a longer news broadcast. For users and providers of telecommunications systems, the presented results should be good news. The challenges related to providing audiovisual streams of acceptable synchrony and quality may well be complicated, and the consequences can certainly be detrimental. However, as far as temporal perception is concerned, one challenge does not depend heavily on the other.

Acknowledgements
The authors would like to acknowledge the helpful feedback from Carsten Griwodz and Knut Inge Fostervold, and the valued assistance of Camilla Hellum Foyn. This work was completed as part of the Perceval project, funded by the Research Council of Norway through the VERDIKT programme.
References


frequency requirements for audiovisual speech perception. *Perception &

Society of America, 84*(2), 476–484.

Perception of audiovisual speech synchrony for native and non-native

articulatory information enables the perception of second language sounds.
*Psychological Research, 71*(1), 4–12.

the Acoustical Society of America, 101*(6), 3241–54.

in adaptive video streaming to handheld devices. In *Proceedings of the 19th

streaming from coarse-grained videos. In *Proceedings of the 18th
International Workshop on Network and Operating Systems Support for
Digital Audio and Video* (pp. 103–108). Williamsburg.

on an image affect your perception of quality? Applying visual attention to
image quality metric. In *Proceedings of the 14th IEEE International

perceived brightness: Sensory interaction versus response bias. *Perception &
Psychophysics, 65*(1), 123–132.

for improving the user experience for time-dependent thin-stream

expertise affects perceived audiovisual asynchrony. *Experimental Brain

Petrini, K., Russell, M., & Pollick, F. (2009). When knowing can replace seeing


Vatakis, A., Maragos, P., Rodomagoulakis, I., & Spence, C. (2012). Assessing the effect of physical differences in the articulation of consonants and vowels on


11.2. Paper 2: Audiovisual robustness: Exploring perceptual tolerance to asynchrony and quality distortion

Authors: Ragnhild Eg, Carsten Griwodz, Pål Halvorsen, and Dawn Behne

Venue: Multimedia Tools and Applications

Status: Accepted for publication.
Audiovisual robustness: Exploring perceptual tolerance to asynchrony and quality distortion

Ragnhild Eg · Carsten Griwodz · Pål Halvorsen · Dawn Behne

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 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

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

C. Griwodz and P. Halvorsen
Department of Informatics, University of Oslo, Norway and Simula Research Laboratory, Norway E-mail: {griff, paalh}@simula.no

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

A considerable body of work on synchronisation in audiovisual streaming originated several years ago [21], and rules-of-thumb for non-noticeable and acceptable asynchronies were established at the time [33]. 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 [7,19,26,34].

While the flexibility of active adaptation alleviates problems related to buffering, the flip-side 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 [29]: 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 [32], and the experience of the event itself is enhanced [31]. 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 [11]. 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 [17]. 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 [13]. 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 [36]. 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 [27, 39]. The robustness of intelligible speech has also been demonstrated in comparison to monkey calls [37] and non-native languages [28]. 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 [40]. 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
Ragnhild Eg et al.

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 [37], 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 [25]. 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 [38], 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 [35].

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 [18, 41]. 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 [41]. 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 [14], 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 [24]. Moreover, the auditory modality will eventually dominate speech perception when visual information is distorted [10, 23]. Conversely, noise and auditory compression artefacts can mask essential spectral frequencies and lead to distracting audio effects [22]; in speech perception, reduced audio quality typically leads to an increased dependency on the visual modality [12]. 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.

With new multimedia platforms come new approaches to efficient streaming, and with them come old concerns. Multiplexing is no longer always the best approach [26]; some situations will require the audio and video streams to remain separate, for instance
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, 13, 20, 25], others use simultaneity judgements and curve fittings to calculate the full-width at half-maximum (FWHM) [5] or the 95% confidence interval [36]. Others still use temporal order judgements and calculate the slopes of cumulative distribution functions [35].

in two-way communication systems, or when there are several audio tracks to consider [34]. 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 resolution, 1024x576 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 2x2, 4x4 or 6x6 pixels. In addition, an undistorted stimuli level was kept at the 1024x576 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
Table 1: Content description of the sequences included in the experiment.

<table>
<thead>
<tr>
<th>Content 1: Speech</th>
<th>Content 2: Chess</th>
<th>Content 3: Drums</th>
</tr>
</thead>
<tbody>
<tr>
<td>The Norwegian news broadcast shows a seated anchorman 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.</td>
<td>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.</td>
<td>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.</td>
</tr>
</tbody>
</table>

this manner due to the experimental set-up; video distortion levels were chosen as best-fits to previous research [10, 23], 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, 60 dB, 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 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 [15]. 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 levels. The results, presented in Figure 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 correspond very well, with over 90% match. The 2x2 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 4x4 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 6x6 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 6x6 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 2x2 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 [42]. Audiovisual sequences were presented successively with a short response period in-between, consistent with the "Absolute Category Rating" method recommended by the ITU-T [16]. 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 Figure 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 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 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 %, \( p < .05 \) [15]. Effect sizes were calculated using the partial eta-squared (\( \eta^2_p \)) statistic [30]; 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 Figure 4. For the auditory distortion experiment, the results from the repeated-measures ANOVAs are presented in Table 3 and distribution curves are plotted in Figure 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 perceived synchrony are still represented by the means of the fitted curves, but the window of temporal integration is here defined

### Table 2 Results from repeated-measures ANOVAs for visual distortion.

<table>
<thead>
<tr>
<th></th>
<th>Speech</th>
<th></th>
<th>Chess</th>
<th></th>
<th>Drums</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( F )</td>
<td>( p )</td>
<td>( \eta^2_p )</td>
<td>( F )</td>
<td>( p )</td>
<td>( \eta^2_p )</td>
</tr>
<tr>
<td>Asynchrony</td>
<td>144.49</td>
<td>&lt;.001</td>
<td>.89</td>
<td>59.18</td>
<td>&lt;.001</td>
<td>.77</td>
</tr>
<tr>
<td>Blur</td>
<td>0.16</td>
<td>ns</td>
<td>.01</td>
<td>3.44</td>
<td>&lt;.05</td>
<td>.16</td>
</tr>
<tr>
<td>Asynch*Blur</td>
<td>1.49</td>
<td>ns</td>
<td>.08</td>
<td>1.45</td>
<td>ns</td>
<td>.08</td>
</tr>
</tbody>
</table>

### Table 3 Results from repeated-measures ANOVAs for auditory distortion.

<table>
<thead>
<tr>
<th></th>
<th>Speech</th>
<th></th>
<th>Chess</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( F )</td>
<td>( p )</td>
<td>( \eta^2_p )</td>
<td>( F )</td>
</tr>
<tr>
<td>Asynchrony</td>
<td>122.75</td>
<td>&lt;.001</td>
<td>.84</td>
<td>71.57</td>
</tr>
<tr>
<td>Noise</td>
<td>2.13</td>
<td>ns</td>
<td>.08</td>
<td>8.50</td>
</tr>
<tr>
<td>Asynch*Noise</td>
<td>2.75</td>
<td>&lt;.001</td>
<td>.10</td>
<td>2.50</td>
</tr>
</tbody>
</table>
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.

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.
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.

Table 4 Results from the repeated-measures ANOVA exploring content.

<table>
<thead>
<tr>
<th></th>
<th>F</th>
<th>p</th>
<th>η²</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asynchrony</td>
<td>113.69</td>
<td>&lt;.001</td>
<td>.86</td>
</tr>
<tr>
<td>Content</td>
<td>7.35</td>
<td>&lt;.005</td>
<td>.29</td>
</tr>
<tr>
<td>Asynch*Content</td>
<td>5.60</td>
<td>&lt;.001</td>
<td>.24</td>
</tr>
</tbody>
</table>

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 Figure 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, Figure 8, and Figure 9.

As the distributions in Figure 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 signifi-
cant 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 Figure 6. Hence, 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 Figure 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 Figure 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, Figure 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 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.

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 Figure 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, 13]. 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
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.

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, 39]. 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 [24], and speech movements [10, 23]. Auditory distortions will similarly mask important acoustical information [22], making it difficult to understand speech sounds [12]. 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 [20, 25, 36], 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 H.264 or MP4 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.

Acknowledgements 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 Helium Føyn 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

10. Eg, R., and Behne, D. Distorted visual information influences audiovisual perception of voicing. In INTERSPEECH (Brighton, 2009), pp. 2903–2906.


11.3. Paper 3: Temporal integration for live conversational speech

Authors: Ragnhild Eg and Dawn Behne


Temporal integration for live conversational speech

Ragnhild Eg¹, Dawn M. Behne²

¹Simula Research Laboratory, Lysaker, Norway
²Department of Psychology, Norwegian University of Science and Technology, Norway
rage@simula.no, dawn.behne@svt.ntnu.no

Abstract

The difficulty in detecting short asynchronies between corresponding audio and video signals demonstrates the remarkable resilience of the perceptual system when integrating the senses. Thresholds for perceived synchrony vary depending on the complexity, congruency, and predictability of the audiovisual event. For instance, synchrony is typically detected sooner for simple flash and tone combinations than for speech stimuli. In applied scenarios, such as teleconference platforms, the thresholds themselves are of particular interest; since the transmission of audio and video streams can result in temporal misalignments, system providers need to establish how much delay they can allow. This study compares the perception of synchrony in speech for a live two-way teleconference scenario and a controlled experimental set-up. Although methodologies and measures differ, our explorative analysis indicates that the windows of temporal integration are similar for the two scenarios. Nevertheless, the direction of temporal tolerance differs; for the teleconference, audio lead asynchrony was more difficult to detect than for the experimental speech videos. While the windows of temporal integration are fairly independent of the context, the skew in the audio lead threshold may be a reflection of the natural diversion of attending to a conversation.

Index Terms: audiovisual speech, temporal integration, synchrony perception, teleconference

1. Introduction

Perception of synchrony is frequently applied as a tool to evaluate the temporal integration of audiovisual events. Historically, researchers have looked at very simple stimuli, such as flash and tone combinations, to study basic perceptual processes (e.g., [1,2]). In later years this methodology has been extended to more complex audiovisual events, in particular to speech (e.g., [3–5]). The growing body of research on the temporal integration of audiovisual events has established the ability of human perception to realign sensory signals so that short temporal offsets go unnoticed, with an inherent asymmetry that favours the precedence of visual over auditory signals (e.g., 6). This buffer is likely in place to avoid perceptual conflicts and to ensure coherent sensory experiences, similar to the ventriloquist effect that compensates for spatial displacements [7], and the McGurk effect [8], which illustrates perceptual strategies for incongruent speech tokens. In this respect, the detection of asynchrony is an informative measure of the limitations of perceptual integration. The window of temporal integration will vary according to both the context [9,10] and the applied experimental methodology [11,12] such that the difference between conditions is often of greater interest than the specific temporal thresholds (e.g., [13,14]). Temporal offsets between simple audiovisual events such as light and sound combinations are typically detected at fairly short offsets [12]. More complex events that involve, for instance, musical instruments, can yield wider windows of temporal integration [10]. Windows of temporal integration observed for audiovisual speech are also found to be fairly robust to asynchrony [3,15]. Figure 1 shows an overview of a range of previously published research with windows of temporal integration for different speech stimuli, derived from different measures. This selection of work illustrates the variations in perceptual tolerance to asynchrony, not only between spoken words [3, 16], sentences [9,10,15,16] and syllables [5], but also across different experimental settings and methodologies (e.g. [9,15,17]). In general, the temporal perception of syllables appears to be less tolerant than that of full words and sentences (e.g. [10]). Furthermore, the temporal thresholds derived from temporal order judgement (TOJ) and simultaneity judgement (SJ) measures could reflect different perceptual strategies [11]; that TOJ measures involve the additional task of determining the order of two signals may make them more demanding, but the focus on precedence may also make them more sensitive [12]. Thresholds for perceived synchrony are of direct relevance when it comes to multimedia platforms (e.g., [18]). For teleconference systems the delay of an audio or video stream can have severe consequences for both the quality [19] and the intelligibility [20] of the perceived speech. Windows of temporal integration can therefore serve as guidelines for system providers to indicate the maximum misalignment that can be tolerated. As mentioned, perception of synchrony depends on the nature of the audiovisual event, but in a live conversation there are bound to be other influences and disturbances that are controlled for in experimental settings. Attention is a likely source of variation; should attention be captured by one modality or engaged by another task, larger temporal offsets may go unnoticed [21]. The current study assesses the detection of asynchrony in live conversations that take place over a teleconference platform, in order to explore the generalisability of temporal integration thresholds derived from isolated audiovisual events. By comparing the teleconference with a more controlled experimental setting, the study aims to shed light on the complexities of a real-life scenario and the consequent predicted potential of increased perceptual tolerance to asynchrony. An SJ measure is used for the experimental setting, as the task of judging synchrony versus asynchrony is more similar to asynchrony detection than to TOJ. Furthermore, by comparing the SJ derivation, with perceived synchrony thresholds at 50 % [3], to the more direct measure of asynchrony detection, the study may also provide insight on the appropriateness of SJ as a broadly applicable methodological approach.
2. Method

The current study was planned and run as two experiments. The TelCo experiment was carried out as a live teleconference between two people engaged in a conversation, whereas the SJ experiment was conducted in a laboratory setting with two television broadcasts used as speech stimuli.

2.1. Participants

The TelCo experiment included 6 male and 4 female adults under the age of 50. They were all Cisco employees who had been asked and agreed to participate. Participants for the SJ experiment were recruited at the University of Oslo, and in total they included 9 females and 11 males between 20 and 38 years old (\(M=25.60, SD=4.35\)).

2.2. Stimuli and procedure

Due to the differences between the two experimental methods, the implemented stimuli and procedures are presented separately.

2.2.1. TelCo asynchrony detection

The TelCo experiment was run like a teleconference, with two volunteers participating in a question game carried out in two of Cisco’s teleconference rooms at Lysaker, Norway. One participant, the respondent, sat in the smaller of two rooms, and would draw a ticket that stated the name of a person, animal, object, or place. The other participant, the correspondent, sat in the larger and would try to guess what was printed on the ticket by asking a series of yes/no questions. In addition to the respondent’s task in the conversation, the respondent also focused on the experimental task of assessing the synchrony throughout. Halfway through, the two participants switched roles.

Two experimenters sat in the same room as the correspondent in order to continuously manipulate the temporal offset between the audio and the video. Audio lead or audio lag asynchrony was introduced in a random order determined beforehand. The offsets increased in steps of 30 ms and the experimenters took care to start each step during a pause in the conversation. The respondent was instructed to raise a hand the
Participants completed eight rounds, four audio-lead and four audio-lag repetitions, before switching roles. Control measurements with a flash- and tone-device were completed twice for every participant, once for audio lead and once for audio lag. This was done to make sure that the introduced offsets corresponded to the audiovisual asynchrony coming through to the small conference room.

2.2.2. Simultaneity judgements for speech

For the SJ experiment, two speech sequences, News and P.M., were selected from previously aired television broadcasts on the basis of their differences in shooting angle and movement of the speaker. The News sequence shows a female news anchor filmed in studio, while P.M. is an excerpt with the Norwegian Prime Minister in a current issues show. Video playback duration was set to 13 seconds, so that the coherence of both sequences was maintained. Audio editing was done with Audacity (2.0.1) [22] and Praat [23], with average audio intensity at 70 dB. Videos were edited in Final Cut Pro (10.0.8), with 1024x576 pixel resolution. Temporal offsets were based on our earlier experiments [e.g., 3] and introduced by displacing the audio track relative to the video track. Audio lead asynchrony was presented at 50 ms, 100 ms, 150 ms, and 200 ms, while audio lag asynchrony was set to 100 ms, 200 ms, 300, and 400 ms. The two ranges of asynchrony levels reflect the asymmetry in perceptual sensitivity to lead and lag asynchrony [6].

The experiment was conducted in a meeting room at the University of Oslo, with videos presented using the Superlab software running on a 2.53 GHz MacBook Pro with a 15" monitor (1440x900 pixel resolution). Two Logitech Z4i satellite speakers (8.5 watts each, >92 dB S/N, 35 Hz - 20 kHz frequency response) were placed on either side of the monitor and participants sat at a distance of approximately 70 cm.

Participants were asked to attend to both the audio and the video and make decisions on whether they perceived them to be synchronous or asynchronous. Responses were collected with a Cedrus RB-530 response-box. The experiment was divided into blocks, in which single instances of stimulus conditions were presented in a random order. As responses could be given at any time, the duration of the experiment varied between individuals, with an upper restriction of 90 minutes including breaks between blocks. Thus the total number of blocks, and thereby also repetitions, varied between 6 and 8 blocks, depending on the rate of progression.

3. Results

Detection times for the asynchrony introduced in the TelCo experiment were averaged across repetitions to establish audio lead and lag thresholds. The point of subjective simultaneity (PSS) was calculated as the mean of the lead and lag thresholds, whereas the window of temporal integration (TI) spans the two thresholds. For the SJ experiment, responses were scored as synchrony match or non-match and proportions were calculated across repetitions of each stimulus condition. Gaussian curves were then fitted individually across the range of temporal offsets. The TI is represented by the full-width at half-maximum, which was calculated from the standard deviation of each curve, while the PSS corresponds to the mean of the curve. From these statistics, the audio lead and audio lag thresholds were also established.

An initial one-way ANOVA assessed the effect of the order of participant roles for the TelCo experiment on the detection of lead and lag asynchrony, as well as the derived TI and PSS. None of the measures were found to differ significantly between the first and second respondents, indicating that the order of experimental tasks did not impact participants’ detection of asynchrony.

The PSS, TI, and lead and lag thresholds for the TelCo and SJ experiments are derived from different procedures, different measures, different participants, and different calculations; the statistical analyses are therefore carried out purely exploratively. To gain some insight into possible variations in the perception of synchrony for the different audiovisual conditions, we ran one-way ANOVAs for lead and lag thresholds, the TI, and the PSS, to compare TelCo, News and P.M. Main effects for PSS [F(2,47)=16.07, p<.001], audio lead thresholds [F(2,47)=29.68, p<.001], and audio lag thresholds [F(2,47)=4.48, p=.02], indicate that temporal perception might not be consistent across the three contexts. The differences between thresholds are plotted in Figure 2; however, Figure 2 also illustrates the similarities in the TI widths. The corresponding lack of an effect for TI [F(2,47)=1.02, ns] suggests that the overall tolerance for asynchrony is similar across scenarios, although with a directionality that may depend on characteristics of the speech events. The variations in the perception of lead and lag asynchrony are demonstrated by the different thresholds for TelCo, News and P.M, presented in Figure 2. The graph shows that the differences are particularly prominent for lead asynchrony, and that the temporal offsets required for detection are quite long for the TelCo scenario. The PSS plotted in Figure 3 also highlights how subjective synchrony is closest to objective synchrony for TelCo compared to the video sequences. The main effects were further explored with Dunnett C post-hoc analyses, with significant contrasts represented by black arrows in Figures 2 and 3. The greater variance between participants in the SJ experiment, as compared to the TelCo experiment, may explain why the audio lag threshold does not differ significantly between News and P.M. Still, the significant differences between all audio lead thresholds, and two of the PSS contrasts, emphasize the notion that the asymmetry in temporal tolerance may be the major source of variation in responses.
Figure 2: Windows of temporal integration with audio lead and audio lag thresholds. Thresholds for the TelCo scenario are established from the points of asynchrony detection, whereas thresholds for News and P.M. are calculated from the fitted Gaussian distributions and correspond to synchrony perceived at chance level. The black arrows indicate audio lead and audio lag thresholds that differ significantly from the others, while error bars represent standard deviations. Windows of temporal integration for the three conditions were not significantly different.

4. Discussion

Earlier studies into the perception of synchrony between auditory and visual events have contributed to a field of research with widely varying thresholds of temporal integration (e.g., [9,10,20]). Considering the differences found when comparing methodologies and stimulus complexities [12], and the assumed contribution of attention [21], the results from this study are surprisingly consistent across conditions. The windows of temporal integration did not differ significantly across the three experimental scenarios in the current study, implying that the perceptual tolerance to temporal offsets is more constant than we originally expected. On the other hand, significant differences between all audio lead thresholds point to a bias in the direction of temporal integration of different audiovisual events. This skew is also reflected in the difference between audio lag thresholds when comparing TelCo and P.M., and in the PSS contrasts between TelCo and P.M., as well as between News and P.M. In other words, the overall tolerance to audiovisual asynchrony is more or less the same for the teleconference and the two speech sequences, but the directionality sets them apart.

The distinctly greater tolerance to audio lead asynchrony observed for the teleconference scenario, compared to the speech sequences, could possibly reflect the predicted complexities of a real-life scenario. The impact from the added attentional demand of the question game might only be manifested for asynchrony where the margins for detection were already narrow. Given that temporal perception is especially sensitive to auditory signals that precede visual signals [6], the disturbances attributable to the teleconference could have contributed to a greater perceptual tolerance in this direction. If so, we deduce that temporal integration may be even more robust in the busy surroundings of everyday life than is generally found in experimental settings.

As for the two speech sequences from the SJ experiment, we surmise that the intentional choice of video content with different speaker characteristics and shot angles may have affected the perceptual sensitivity to lead and lag asynchrony. While one speaker is positioned face forward, the other is facing sideways and viewed at an angle. The viewing angle of a speaker can indeed influence performance on speech recognition tasks, at least in the presence of distractions [25]. Thus temporal speech cues are also likely affected by the visibility of a speaker’s face. Moreover, the clarity of the acoustical phonemes, along with the prominence of the speech movements, is also likely to contribute to the accuracy of participants’ simultaneity judgements.

Overall, the temporal integration of audiovisual speech shows a remarkable resilience to asynchrony. With the narrowest window of temporal integration approaching 350 ms, the potential leeway available for teleconference platforms is quite remarkable. Our results correspond to previously published thresholds established for continuous speech stimuli [9,15,20]. Although TOJ measures tend to yield more narrow windows of temporal integration [10], particularly for short speech segments [5], SJ and detection tasks can be argued to provide more ecologically valid measures of synchrony perception [12]. Based on the results from the current study and related works, our most conservative recommendations to developers of teleconference platforms, and similar systems, would therefore be to ensure that thresholds do not exceed 50 ms for audio lead asynchrony and 200 ms for audio lag asynchrony. Within this window of temporal integration, perception can compensate for the temporal misalignment and asynchrony is unlikely to be noticed.

Figure 3: PSS for the three speech scenarios, calculated as the mid-point between lead- and lag-thresholds for TelCo and represented by the mean of the Gaussian distribution for News and P.M. Significant differences in PSS are highlighted by the black arrows and standard deviations are shown as error bars.
5. Acknowledgements

The authors would like to thank the team at Cisco Norway for the rewarding collaboration.

6. References


11.4. Paper 4: Audiovisual temporal integration in reverberant environments

Authors: Ragnhild Eg and Dawn Behne

Venue: *Speech Communication*

Status: Submitted on February 12, 2014.
Audiovisual temporal integration in reverberant environments

Ragnhild Eg & Dawn Behne

Abstract

With teleconferencing becoming more accessible as a communication platform, researchers are working to understand the consequences of the interaction between human perception and this unfamiliar environment. Given the enclosed space of a teleconference room, along with the physical separation between the user, microphone and speakers, the transmitted audio often becomes mixed with the reverberating auditory components from the room. As a result, the audio can be perceived as smeared in time, and this can affect the user experience and perceived quality. Moreover, other challenges remain to be solved. For instance, during encoding, compression and transmission, the audio and video streams are typically treated separately. Consequently, the signals are rarely perfectly aligned and synchronous. In effect, timing affects both reverberation and audiovisual synchrony, and the two challenges may well be inter-dependent. This study explores the temporal integration of audiovisual speech and non-speech sequences across a range of asynchrony levels for different reverberation conditions. Non-reverberant stimuli are compared to stimuli with added reverberation recordings. Findings reveal that reverberation does not influence the temporal integration of speech, which has positive implications for providers of teleconference solutions. However, reverberation does have a severe impact on the temporal integration of an action-oriented sequence, with perceived subjective synchrony skewed towards audio lead asynchrony and away from the more common audio lag direction. Furthermore, less time is spent on simultaneity judgements as the temporal offsets get longer and when reverberation is introduced, suggesting that both asynchrony and reverberation add to the demands of the task.
1 Introduction

Teleconference systems have evolved from being a direct communication platform between two individuals to becoming an extended meeting arena for larger groups of people. With larger groups and larger meeting rooms come larger challenges to tackle, such as reverberating sound components that tend to extend in time as the room size and source distance increase (de Lima et al., 2009). Reverberation is the consequence of the acoustic response from an enclosure (ITU, 2009), characterised by the temporal smearing of an auditory signal. Thus, the sound that reaches the ear is a combination of the energy spectrum that has been conveyed directly, and the reflected one that has been delayed in time (Assmann & Summerfield, 2004). Both the strength and the length of reverberations contribute to influence the experience of AV\(^2\) quality (Jumisko-Pyykkö, Reiter, & Weigel, 2007). In speech, the resulting effect not only disturbs the experienced quality (de Lima et al., 2008), but also alters the signature and intelligibility of the spoken sounds (Cox, Alexander, & Gilmore, 1987).

Specifically, reverberation may transform dynamic speech phonemes into more static elements, thereby flattening formants and blurring the onset and offset of certain consonants and vowels, while extending others (Assmann & Summerfield, 2004). Compared to quiet conditions, reverberation makes it difficult to discriminate pitch (Sayles & Winter, 2008) and it can create confusion among vowels (Cox et al., 1987). For example, the perception of reverberant speech will typically merge the two-vowel sound of a diphthong into a single-vowel monophthong (Nábělek, 1988). Furthermore, confusion related to consonant place of articulation and voicing has also been established (Cox et al., 1987), especially for consonants that follow a vowel at the end of a word (Gelfand & Silman, 1979). In line with Kurtovic’s model (1975, described in Gelfand & Silman, 1979), the energy reflected from the preceding vowel is believed to mask the subsequent consonant and thereby make the articulation features less intelligible. This masking would be far less detrimental for a consonant in a word-initial position.

\(^2\) AV is used as an abbreviation throughout for any audiovisual content or process.
Reverberation also leads to confusion in the arrival of an auditory signal, hampering the perceptual capacity to discern the precedence to one ear before the other (Hartmann, 1983). Because this precedence, or interaural time difference, is an important cue for localising sound sources, reverberation contributes to difficulties in establishing the origin of a sound and even retaining attention to it (Culling, Summerfield, & Marshall, 1994; Darwin & Hukin, 2000). In a natural environment with several people engaged in a conversation, binaural cues would normally assist in locating the speaker; however, in a teleconference, the reverberation that could arise from the transmission would be detrimental to this process (Nunes et al., 2011). Moreover, the potential disturbance from background noises and voices may serve to enhance the problem of reverberation in teleconferences.

This study considers reverberation from two distinct teleconference rooms, but instead of looking into the established effect on auditory speech intelligibility, it explores the potential influence of auditory smearing on temporal perception. Many earlier works have been restricted to auditory perception (Culling et al., 1994; Darwin & Hukin, 2000; de Lima et al., 2008), but we extended the current investigation to include the visual modality. While background noise typically will increase perceptual dependence on visual input (Alm, Behne, Wang, & Eg, 2009; Sumby & Pollack, 1954), less is known about the perceived correspondence between vision and hearing in the presence of reverberation. One study used reverberant depth cues to demonstrate perceptual alignment to simulated source distances, where greater distances required auditory signals to lag further behind for perceived subjective synchrony (Alais & Carlile, 2005). Another study found less accurate temporal order judgements for spatially and temporally separated AV signals in reverberant conditions, compared to anechoic conditions (Sankaran, Carlile, & Leung, 2013). Combined, these findings point to a decreased sensitivity to temporal offsets in the presence of reverberation. Despite the relevance to teleconference systems, as far as we know, no study has been carried out to directly explore the impact of reverberation on the temporal integration of auditory and visual speech signals.
The encoding, compression and transmission of audio and video can result in a temporal misalignment between the two streams (Bang et al., 2009). Short temporal offsets are rarely noticeable, but once they exceed certain durations, they can be detrimental to both the subjective experience of quality (Steinmetz, 1996) and the intelligibility of spoken sounds (Grant & Greenberg, 2001). Nevertheless, no fundamental thresholds mark the transition from perceived synchrony to perceived asynchrony (Roseboom, Nishida, & Arnold, 2009); instead, they vary with the measure and the nature of the AV event (van Eijk, Kohlrausch, Juola, & van de Par, 2008). For instance, perceptual tolerance to asynchrony is typically greater for spoken words and sentences than for more action-oriented events, such as a hitting hammer (Conrey & Pisoni, 2006; Dixon & Spitz, 1980). Moreover, the perceptual tolerance to temporal offsets is inherently asymmetric (Maier, Di Luca, & Noppeney, 2011). Thus, detection thresholds tend to reflect a lesser tolerance to asynchrony where the auditory signal precedes the visual signal (audio lead) than to asynchrony where the visual signal arrives first (audio lag) (Dixon & Spitz, 1980).

To assess the perception of synchrony, at least four approaches have been implemented in past research: detection of gradually introduced asynchrony (Dixon & Spitz, 1980), discrimination between sequences that differ in AV asynchrony (Grant, van Wassenhove, & Poeppel, 2003; McGrath & Summerfield, 1985), simultaneity judgements (Conrey & Pisoni, 2006; Fujisaki, Shimojo, Kashino, & Nishida, 2004; Roseboom et al., 2009; Zampini, Guest, Shore, & Spence, 2005), and temporal order judgements (Bertelson & Aschersleben, 2003; Parise & Spence, 2008; Vatakis & Spence, 2006). For this study, we implemented simultaneity judgements to explore temporal integration under variable conditions. The benefits of simultaneity judgements are the relative ease of the task and their estimation of perceived synchrony, referred to as point of subjective simultaneity (PSS). Considering the respective speeds of light and sound, the perceptual adaptation to late-coming sound is deemed to have higher ecological validity, and this is reflected in the PSS established from simultaneity judgements (van Eijk et al., 2008). Furthermore, simultaneity judgements yield temporal windows of AV integration, confined by the thresholds at which the
perception of auditory lead and lag asynchrony occurs at chance (van Wassenhove, Grant, & Poeppel, 2007).

For continuous speech stimuli, the window of temporal integration tends to be fairly robust. For example, in a temporal order experiment where stimuli included a spoken sentence, the audio lead and lag thresholds were established at 118 ms and 190 ms, respectively (Vatakis & Spence, 2006). Another study assessing the perception of simultaneity for spoken sentences found thresholds at 131 ms for audio lead and 225 ms for audio lag (Conrey & Pisoni, 2006). Yet another experiment used a discrimination task where participants had to indicate which of two spoken sentences was out of synchrony, for the unfiltered speech stimuli the audio lead threshold was approximately at 60 ms and the audio lag threshold was around 230 ms (Grant et al., 2003). As these results illustrate, variations are great across studies and methodologies, but the characteristic asymmetry of temporal speech perception remains consistent, as does the overall robustness to asynchrony.

This study addresses two technical challenges relevant to teleconference systems, reverberation and asynchrony, and explores the temporal integration of visual signals with reverberant auditory signals. Knowing that reverberation alters the temporal signatures of certain speech characteristics (Cox et al., 1987), and may even mask subsequent speech sounds (Gelfand & Silman, 1979), we assume that increased reverberation will lead to increased temporal distortion. By creating a temporal ambiguity, we expect reverberation to contribute to a greater perceptual tolerance to AV asynchrony by widening the temporal window of integration. Furthermore, due to the extended masking that may result when the audio precedes the video, we believe that the increased tolerance will be most prominent for auditory lead asynchrony. To explore whether the temporal distortion expected from reverberation may differ between speech sounds and less complex auditory signals, the applied stimuli include two speech scenarios, as well as an action-oriented scenario. By including two speech sequences based on separate speakers, we will also assess potential variations in temporal integration that are not related to reverberation, but to speaking style. Finally, we want to address the time spent by participants when evaluating the simultaneity of AV
presentations at different levels of asynchrony and reverberation. Because uncertainty and cognitive load may vary across conditions, the evaluation times could contribute with insights on the effort required to make simultaneity judgements.

2 Method

Reverberation is a consequence of sound transmitted in an enclosed area and the auditory signature is therefore unique to that enclosure. Nevertheless, to keep stimuli consistent across conditions, we used the approach of simulated reverberation. In other words, auditory signals were filtered using impulse responses recorded from two teleconference rooms that differed in size and reverberation time. The experiment was set up in a university meeting room, with video presented on a computer screen and audio through loudspeakers. Although the experimental location contributed with its own reverberating components, the temporal distortion from the simulated reverberation was preserved within the audio signals. We aimed to make the setting realistic and to use stimuli with natural speech production and a familiar action event.

2.1 Stimuli and material

To explore the effect of auditory reverberations on the temporal integration of audio and video, we selected three AV sequences: News, P.M., and Chess. With teleconference systems as the applied scenario, two of the sequences contained television excerpts that represented well-known speech scenarios. One speech sequence, the News content, came from a Norwegian news broadcast, with a female anchor filmed in studio and speaking directly into the camera. A second speech sequence, P.M., was an excerpt from a current issues show, where the Norwegian Prime Minister was filmed in studio, talking to a show host. The Chess content, which contained no speech, served as a predictable object-action sequence to compare with the dynamic speech contents. We made sure that the duration was kept the same for all video sequences by selecting an excerpt from the News content that was logical and semantically representative in isolation. From this basis, we established a common duration of 13 seconds and found an equally logical excerpt from the P.M. content that fit within this constraint. For
the Chess content, the 13 seconds included five full moves on the board. All editing was performed with Audacity (2.0.1) (Audacity Team, 2012), Praat (Boersma & Weenink, 2012), and Final Cut Pro (10.0.8), with average audio intensity at 70 dB and video resolution set to 1024x576 pixels.

Before manipulating the asynchrony of the video sequences, we ensured that their synchrony was maintained throughout. Admittedly, temporal alignment cannot be guaranteed when audio and video tracks are edited and encoded separately (Bang et al., 2009), but we controlled the synchrony of the speech sequences by comparing the visual lip movements for voiced initial syllables and the corresponding spectrograms. Similarly, for the Chess sequence we verified that the moment the chess piece made contact with the board coincided with the auditory event on the spectrogram. Experimental asynchrony levels were based on an earlier study where we found the characteristic asymmetry in temporal integration, with greater sensitivity to auditory lead than to auditory lag asynchrony (Eg & Behne, 2012), consistent with prior research (Conrey & Pisoni, 2006; van Wassenhove et al., 2007; Vroomen & Stekelenburg, 2011). When introducing asynchrony, only the audio tracks were displaced, whereas the onsets and offsets of the video tracks were kept constant throughout to avoid giving away visual cues. By editing out the duration of the temporal offset at the start of the audio track, audio lead asynchrony was applied at 50 ms, 100 ms, 150 ms, and 200 ms. Meanwhile, including a small part of the audio track preceding the selected sequence served to apply audio lag asynchrony at 100 ms, 200 ms, 300, and 400 ms.

To simulate reverberation, audio tracks were manipulated using impulse responses recorded in two of Cisco’s teleconference rooms in San Jose, California. Impulse responses used to filter audio tracks for the 636 ms long reverberation condition, Reverb.636, came from recordings from a small conference room designed for 4 people, which measured \( \approx 3.5 \times 4 \) meters. Recordings for the 538 ms long reverberation condition, Reverb.538, were carried out in a large conference room that could seat up to 16 people and measured \( \approx 6.5 \times 8.5 \) meters. The non-reverberant condition is referred to as Quiet. Figure 1
includes examples of audio from the three contents presented in quiet and reverberation.

2.2 Participants
A total of 9 female and 11 male participants, aged 20 to 38 years ($M=25.60$, $SD=4.35$), were recruited from the University of Oslo. Participants received information about the experiment and the relevant ethical considerations, and also gave their consent, prior to the experimental task.

2.3 Procedure
We tested participants individually in a meeting room at the University of Oslo, with videos presented using the Superlab software running on a 2.53 GHz MacBook Pro with a 15” monitor (1440x900 pixel resolution). They sat approximately 70 cm from the monitor, with two Logitech Z4i satellite speakers (8.5 watts each, $>92$ dB S/N, 35 Hz - 20 kHz frequency response) placed immediately next to the monitor, on each side, and the subwoofer turned off. Participants were asked to attend to the audio and the video and to decide whether they perceived them to be in synchrony or not. Responses could be given at any time during the presentation by pressing one of two labelled buttons on the Cedrus RB-530 response-box. The labels read “synchronous” and “asynchronous”. Given the long duration of the presentations, the time taken to make a response is not equivalent to reaction time. Instead, the time measures are considered as representative of a conscious decision-making process and are included and analysed as evaluation times. The maximum duration of the full experiment was limited to 90 minutes. Because evaluation times varied between individuals, the number of repetitions also varied. All stimulus conditions were randomised within a block, with the same 81 trials repeated for every block. Each block was preceded by a break. Depending on the rate of progression, participants completed between 6 and 8 blocks, corresponding to the number of stimulus repetitions.
Figure 1. Spectrograms and pitch contours derived from \( \approx800 \) ms audio excerpts from the Chess, News, and P.M. contents. The top row represents a single chess piece being put down on the board in: (a) Quiet, (b) Reverb.636, and (c) Reverb.538. The middle row contains the phrase “i går kveld” (“yesterday evening”, in English) delivered by the news broadcaster in: (d) Quiet, (e) Reverb.636, and (f) Reverb.538. Finally, the bottom row depicts the words “sett bedre” (“seen better”, in English) spoken by the Prime Minster in: (g) Quiet, (h) Reverb.636, and (i) Reverb.538.
3 Results and discussion

3.1 Simultaneity judgements

Because of the unequal number of repetitions completed by participants, we converted the simultaneity judgement scores to ratios and averaged them across repetitions. Individual Gaussian curves, distributed across asynchrony levels, could then be fitted for each content and reverberation condition. With the mean of each fitted curve defining the individual PSS, the grand average determined the PSS for each stimulus condition, as illustrated in Figure 2. The full-width at half-maximum of each curve established both the temporal window of integration and the temporal thresholds for lag and lead asynchrony (Conrey & Pisoni, 2006). The temporal thresholds correspond to the points where synchrony and asynchrony are reported at the same rate, at chance levels. We assessed potential outliers from the temporal thresholds by excluding any scores that exceeded the value of the maximum temporal offset. With this criterion, scores from one female participant were excluded from the analyses for the Chess content. Figure 3 shows the 50% thresholds for audio lead and audio lag asynchrony, highlighting the different impact of reverberation on the speech and action stimuli. Repeated-measures ANOVAs explored the effect of asynchrony and reverberation, along with their interaction, on the subjective perception of synchrony. Based on the differences between distributions, we ran separate analyses for each content; all results are presented in Table 1. Effect sizes are reported with the partial-eta square statistic ($\eta_p^2$).
Figure 2. Points of Subjective Simultaneity (PSS) for all contents and reverberation conditions, averaged across participants. Error bars illustrate the standard deviations of each Gaussian distribution, centred around the mean.

Figure 3. Temporal thresholds for lead and lag asynchrony, separated by reverberation conditions and averaged across participants, for (a) Chess, (b) News, and (c) P.M. Error bars correspond to the standard deviation of each threshold.
Table 3. Results from three repeated-measures ANOVAs, run separately for Chess, News, and P.M., to explore the effect of reverberation on the perception of synchrony. The table includes degrees of freedom (df), the $F$-statistics, the significance ($p$-value), and the effect sizes ($\eta_p^2$).

<table>
<thead>
<tr>
<th></th>
<th>df</th>
<th>$F$</th>
<th>$p$-value</th>
<th>$\eta_p^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chess</td>
<td>Asynchrony</td>
<td>8, 144</td>
<td>134.26</td>
<td>&lt;.001</td>
</tr>
<tr>
<td></td>
<td>Reverberation</td>
<td>2, 36</td>
<td>15.98</td>
<td>&lt;.001</td>
</tr>
<tr>
<td></td>
<td>Reverb*Async</td>
<td>16, 288</td>
<td>33.53</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>News</td>
<td>Asynchrony</td>
<td>8, 152</td>
<td>91.02</td>
<td>&lt;.001</td>
</tr>
<tr>
<td></td>
<td>Reverberation</td>
<td>2, 38</td>
<td>2.80</td>
<td>ns</td>
</tr>
<tr>
<td></td>
<td>Reverb*Async</td>
<td>16, 304</td>
<td>1.10</td>
<td>ns</td>
</tr>
<tr>
<td>P.M.</td>
<td>Asynchrony</td>
<td>8, 152</td>
<td>99.75</td>
<td>&lt;.001</td>
</tr>
<tr>
<td></td>
<td>Reverberation</td>
<td>2, 38</td>
<td>0.24</td>
<td>ns</td>
</tr>
<tr>
<td></td>
<td>Reverb*Async</td>
<td>16, 304</td>
<td>1.40</td>
<td>ns</td>
</tr>
</tbody>
</table>

Analyses uncovered main effects of asynchrony for all contents, as seen in Table 1, but the influence from reverberation was only significant for the Chess sequence. The interaction between reverberation and asynchrony reveals a prominent shift from audio lag to audio lead asynchrony, which is reflected in the PSS, Figure 2, and the temporal thresholds, Figure 3a. However, Figures 2, 3b and 3c illustrate the consistency in the temporal perception of the speech sequences, in quiet or in any of the two reverberating scenarios. The windows of temporal integration for News and P.M. correspond well to earlier works (Conrey & Pisoni, 2006; Dixon & Spitz, 1980), with the characteristic asymmetry that points to greater perceptual tolerance when visual signals precede, rather than succeed, auditory signals (Grant et al., 2003). With the missing impact of reverberation on the temporal integration of AV speech, it is clear that participants were equally able to judge AV synchrony in quiet and reverberant environments. Considering the severely audible reverberation, the lack of an effect on the temporal perception of AV speech is surprising. Seemingly, reverberant conditions can contribute to misidentifications of vowels and
consonants (Cox et al., 1987), but will still not affect the perceived temporality of speech. Although the acoustical signal is likely distorted from the reverberation (Culling et al., 1994), the temporal distortion may affect nothing more than the intelligibility of the speech sounds. Moreover, as several studies have demonstrated, the perception of synchrony for continuous speech is very resilient to temporal offsets (Conrey & Pisoni, 2006; Dixon & Spitz, 1980; Grant et al., 2003; Vatakis & Spence, 2006), and this finding certainly appears to extend to reverberant environments.

On the other hand, the results for the Chess segment tell a different story. Judging from the PSS, longer audio lead times are required for synchrony to be perceived in reverberation, regardless of room size. Related to this result are the audio lead and lag thresholds that become significantly skewed with reverberation, to the point where participants were unable to perceive audio lead asynchrony for the temporal offsets presented in the experiment. However, the room size showed no significance on the influence of the reverberant environment. With the absent effect of reverberation upon the temporal integration of speech, the severe impact observed for the Chess sequence was also surprising. Still, when taking into account the contrast found when comparing the perception of synchrony for AV speech and more action-oriented events (Dixon & Spitz, 1980; Vatakis & Spence, 2006), it is plausible that these results can be attributed to the isolation of the AV event. Unlike the continuous speech of the news anchor and the prime minister, the acoustical signal of the chess piece touching the board was brief and not masked by preceding and succeeding events. In effect, the temporal smearing of the auditory signal was uninterrupted and may have carried greater perceptual impact. The ambiguity related to the onset of the auditory event as it is extended in time thus appears to make audio lead asynchrony close to impossible to perceive for isolated events.

### 3.2 Evaluation times

Analyses of evaluation times are included to investigate whether participants find it more challenging to make judgements of simultaneity in reverberant than in undisturbed conditions, in which case they would presumably require longer time to evaluate the two options. Any evaluation time that exceeded the duration of the
video sequences was excluded as an outlier, which was the case for 246 of the 12380 presentations (<2%). The remaining evaluation times were averaged across stimulus repetitions for each participant. Three repeated-measures ANOVAs, one for each content, yielded the results presented in Table 2 and Figure 4.

**Figure 4.** Average evaluation times for all reverberation and asynchrony conditions, with error bars representing the standard error of the mean.

**Table 4.** Results from three repeated-measures ANOVA, run separately for each content to explore participants’ evaluation times for quiet and reverberant conditions. The table includes degrees of freedom (df), the F-statistics, the significance (p-value), and the effect sizes ($\eta_p^2$).

<table>
<thead>
<tr>
<th></th>
<th>df</th>
<th>F</th>
<th>p-value</th>
<th>$\eta_p^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asynchrony</td>
<td>Chess</td>
<td>8,152</td>
<td>14.06</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Reverbation</td>
<td>Chess</td>
<td>2,38</td>
<td>21.21</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Reverb*Async</td>
<td>Chess</td>
<td>16,304</td>
<td>3.83</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Asynchrony</td>
<td>News</td>
<td>8,152</td>
<td>8.40</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Reverbation</td>
<td>News</td>
<td>2,38</td>
<td>5.25</td>
<td>&lt;.011</td>
</tr>
<tr>
<td>Reverb*Async</td>
<td>News</td>
<td>16,304</td>
<td>1.08</td>
<td>ns</td>
</tr>
<tr>
<td>Asynchrony</td>
<td>P.M.</td>
<td>8,152</td>
<td>14.67</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Reverbation</td>
<td>P.M.</td>
<td>2,38</td>
<td>1.47</td>
<td>ns</td>
</tr>
<tr>
<td>Reverb*Async</td>
<td>P.M.</td>
<td>16,304</td>
<td>0.67</td>
<td>ns</td>
</tr>
</tbody>
</table>
As seen from the analyses presented in Table 2, participants’ evaluation times reflect an effect of asynchrony for all contents, as well as main effects of reverberation for Chess and News. The distributions presented in Figure 4 illustrate how evaluation times are, for the most part, shorter with long temporal offsets. For the two speech sequences, participants required more time to make judgements on synchrony when the temporal offsets were close to objective synchrony. Arguably, these asynchrony levels should be harder to discern and would therefore require more consideration. In line with this result are the shorter evaluation times observed for speech sequences presented with audio and video in synchrony. Reductions in evaluation times also seem to follow the ease of synchrony perception for the Chess sequence. Considering the PSS estimated for Chess, presented in Figure 2 with and without reverberation, these are less skewed toward the audio lag direction, compared to News and P.M. It follows that simultaneity judgements should be less demanding with audio lag than audio lead asynchrony, consistent with the observed evaluation times.

Oddly, the main effects for reverberation show different trends for Chess and News. With average evaluation times ranging from 6.73 seconds and 6.75 seconds for Reverb.636 and Reverb.538, respectively, to 7.11 seconds in quiet, simultaneity judgements for the Chess content were made quicker in reverberation. Conversely, for the News content, the evaluation times increased from 5.58 seconds in quiet to 5.64 seconds for Reverb.636 and 5.83 seconds for Reverb.538. Thus, the time spent on simultaneity judgements for News is longer with reverberation than in quiet, but all over they are shorter than for Chess. However, the interaction with asynchrony was only significant for the Chess sequence. As seen from Figures 2 and 3, audio lead asynchrony was difficult to discern for the Chess content presented in reverberation and this result appears to be reflected in the judgement-making process. The figures convey a tendency for Reverb.538 to increase evaluation times when temporal offsets were short and simultaneity judgements were likely harder to make. In the audio lag direction, an opposite tendency can be found, with shorter time spent on reverberating sequences. As mentioned, this is likely connected to the relative ease of discerning audio lag asynchrony for the Chess sequence, and the related difficulties in discerning audio lead asynchrony. Accordingly, the shorter
evaluation times seem to correspond to the less demanding simultaneity judgements. The overall decrease in time spent on evaluating the Chess sequence as audio lag asynchrony increases seems consistent with the trend seen for the speech sequences as audio lead and lag asynchrony becomes greater. In general, more time was spent on simultaneity judgements for AV sequences when asynchrony and/or reverberation conditions should implicate higher cognitive loads.

4 Conclusions

Following the premise that reverberation leads to an alteration of the temporal signature of an acoustic signal, this study has investigated the influence of a reverberant environment on AV temporal perception. Considering the technical challenges related to teleconference platforms, and the associated perceptual consequences, the study has specifically addressed how the co-occurrence of asynchrony and reverberation affects the temporal integration of AV speech. In addition, an action-oriented sequence was included to shed light on the differences in temporal integration that are often found between continuous speech and action events. With respect to the two speech sequences, the temporal smearing resulting from reverberation had no impact on the temporal integration of the auditory and visual signals. Similar to previous findings (Conrey & Pisoni, 2006; Dixon & Spitz, 1980; Grant et al., 2003; Vatakis & Spence, 2006), the current thresholds for perceived synchrony fell in favour of a temporal robustness when it comes to continuous speech. Moreover, if these results are applicable to a broader speech context, they imply that the temporal ambiguity that arises from reverberations could be masked by subsequent speech events. In natural and dynamic speech, one syllable tends to follow the other, and this could possibly cover up or work against the temporal smearing. All in all, the established windows of temporal integration, with and without reverberation, indicate that the perceptual system operates with a temporal buffer when integrating sensory signals across modalities and this buffer is particularly resilient to auditory signals lagging behind. The implications of such a theoretical buffer offer good news to providers of teleconference systems, suggesting that perfect objective synchrony is not required for these services.
As surmised from related studies (Dixon & Spitz, 1980; Vatakis & Spence, 2006), the temporal integration does indeed differ between the AV action-oriented event and the speech sequences. In the absence of reverberation, the temporal thresholds for the Chess sequence show a more symmetrical distribution between audio lead and lag asynchrony. Likely, the distinctiveness and relative slow movement of the chess piece touching the board, as compared to the dynamics of continuous speech, create an uncertainty regarding the specific moment of impact. This uncertainty could in turn extend the window of temporal integration. With reverberation smearing the sound of the Chess sequence, the magnitude of the window of temporal integration does not change much, but it is shifted even further in the audio lead direction. Again this result may be attributed to the isolation of the AV Chess event, where the reverberation is allowed to distort the auditory signal without a subsequent event following closely and masking the temporal smearing. Although audio lead asynchrony could go almost unnoticed for an event with an ambiguous moment of impact, the reverberant environment has a clear detrimental effect on AV temporal integration. The shift away from objective synchrony and the arguably more ecologically valid audio lag direction (van Eijk et al., 2008) is bound to affect perceptual processes beyond temporal integration. In the best case, reverberation may only affect the perceived quality; in the worst case, it may also affect the intelligibility of the AV content. While teleconferencing was the main focus of the study, few environments are free from disturbing events. Noises in the background from moving objects or people could be mixed with the speech signal, along with their reverberating components. As a result, the noise signals could be more detrimental to the comprehension and integration of the transmitted speech, than the reverberations of the speech sounds themselves.

With evaluation times as an added measure, we also assessed potential variations in cognitive load that could be attributed to asynchrony and reverberation. Our analyses show that the time spent on judging simultaneity did differ according to perceived synchrony. With the exception of the Chess sequence presented with audio leading the video, time spent on judging conditions with extreme temporal offsets was shorter than for presentations closer to objective synchrony. Fully
synchronous presentations also tended to be evaluated more quickly than presentations with asynchrony levels close to the temporal thresholds, which are harder to discern. For the Chess sequence, where reverberation led to an extremely tolerant temporal perception of audio lag asynchrony, evaluation times were shorter when compared to the quiet condition. Again the presumed cognitive load of the simultaneity judgement task was associated with time spent on the task. Consequently, these results demonstrated perceptual consequences that go beyond temporal integration. AV presentations that were perceived as synchronous more than half of the time were still affecting the cognitive processing, adding load to an already hardworking system.

Acknowledgements

The authors would like to thank the team at Cisco Norway for the collaboration and their assistance with the stimuli material. Professor Carsten Griwodz at Simula Research Laboratory has also been invaluable with his contributions and feedback to this work.
References


“Knock, knock.” – “Who’s there?”


“This is the end!”