Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 66 (2015) 91–106

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Audiovisual temporal integration in reverberant environments

Ragnhild Eg [a,b,*], Dawn Behne [c], Carsten Griwodz [a,d]

[a] Media Performance Group, Simula Research Laboratory, P.O. Box 134, 1325 Lysaker, Norway
[b] Department of Psychology, University of Oslo, P.O. Box 1094 Blindern, 0317 Oslo, Norway
[c] Department of Psychology, Norwegian University of Science and Technology, NTNU, 7491 Trondheim, Norway
[d] Department of Informatics, University of Oslo, P.O. Box 1080 Blindern, 0316 Oslo, Norway

Received 12 February 2014; received in revised form 20 August 2014; accepted 8 October 2014
Available online 13 October 2014

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 continuous speech and speech syllables, along with a non-speech event, 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 affect the temporal integration of continuous speech. However, reverberation influences the temporal integration of the isolated speech syllables and the action-oriented event, 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 for the longer sequences when the temporal offsets get longer and when reverberation is introduced, suggesting that both asynchrony and reverberation add to the demands of the task.

Keywords: Audiovisual speech; Audiovisual asynchrony; Temporal integration; Reverberation

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. Unlike an echo, which returns one distinct acoustical response, reverberation arises as a mix of acoustical responses from the multiple surfaces of the enclosed space (de Lima et al., 2009). Thus, the sound that finally reaches the ear is a combination of the acoustic waves that have been conveyed directly, and the reflected ones that have been delayed in time (Assmann and Summerfield, 2004). Both the strength and

* Corresponding author at: Media Performance Group, Simula Research Laboratory, P.O. Box 134, 1325 Lysaker, Norway. Tel.: +47 414 81 909.
E-mail addresses: rage@simula.no (R. Eg), dawn.behne@svt.ntnu.nu (D. Behne), griff@simula.no (C. Griwodz).

http://dx.doi.org/10.1016/j.specom.2014.10.001
0167-6393/

the length of reverberations contribute to influence the experience of audiovisual (AV) quality (Jumisko-Pyykkö et al., 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 et al., 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 and Summerfield, 2004). Compared to quiet conditions, reverberation makes it difficult to discriminate pitch (Sayles and 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 and Silman, 1979). In line with Kurtovic's model (1975, described in Gelfand and 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.

In addition to altering speech sound intelligibility, reverberation 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 et al., 1994; Darwin and Hukin, 2000). Relatedly, when tone series are presented in simulated reverberation, as opposed to quiet conditions, it is harder to keep in synchrony with the presented tempo (Naylor, 1992). According to Naylor, the tone envelopes become smoothed to the extent that the tail of one could overlap the onset of the next. This implies that reverberation not only alters the acoustical properties of speech sounds, but acts also on the auditory perception of less complex signals. Moreover, a reverberant environment can hinder sound localisation processes. 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.

The current study considers simulated reverberation and reverberation recorded from two distinct teleconference rooms. However, instead of looking into the established effect on auditory speech intelligibility, we here explore the potential influence of auditory smearing on temporal perception. Whereas many earlier works have been restricted to auditory perception (Culling et al., 1994; Darwin and Hukin, 2000; de Lima et al., 2008), the current investigation extends this line of research to include not only auditory perception, but also the visual modality. While background noise typically will increase perceptual dependence on visual input (Alm et al., 2009; Sumby and 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 the visual signals for perceived subjective synchrony (Alais and Carlile, 2005). In other words, when the auditory and visual signals happened at the exact same moment in time, participants would not perceive that the two happened simultaneously. Another study found less accurate temporal order judgements for spatially and temporally separated AV signals in reverberant conditions, compared to anechoic conditions (Sankaran et al., 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 perceived synchrony between auditory and visual speech signals.

Synchrony remains a highly relevant challenge in teleconferencing. Due to the encoding, compression and transmission of audio and video, a temporal misalignment can take place and the two streams will be separated in time (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 and Greenberg, 2001). Nevertheless, no fundamental thresholds mark the transition from perceived synchrony to perceived asynchrony (Roseboom et al., 2009); instead, they vary with the measure and the nature of the AV event (van Eijk et al., 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 and Pisoni, 2006; Dixon and Spitz, 1980). Moreover, the perceptual tolerance to temporal offsets is inherently asymmetric (Maier et al., 2011). Thus, the points of detection 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 and Spitz, 1980). These points are typically represented as thresholds, and are defined by the temporal offset required for synchrony to be perceived at a given rate, for instance 50% of the time (Conrey and Pisoni, 2006). The thresholds also define the window of temporal integration (Keetels and Vroomen, 2012), within which sensory inputs from two modalities are considered to be aligned in time.

Perceived synchrony in speech varies depending on the sound, with asynchrony noticed at shorter offsets for bilabial stops than for the less visibly articulated velar and