Available online at www.sciencedirect.com

SciVerse ScienceDirect

Computer Speech and Language 28 (2014) 607–618

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Animated Lombard speech: Motion capture, facial animation and visual intelligibility of speech produced in adverse conditions

Simon Alexanderson, Jonas Beskow *

KTH Speech, Music and Hearing, 100 44 Stockholm, Sweden

Received 24 October 2012; received in revised form 7 February 2013; accepted 25 February 2013
Available online 5 March 2013

Abstract

In this paper we study the production and perception of speech in diverse conditions for the purposes of accurate, flexible and highly intelligible talking face animation. We recorded audio, video and facial motion capture data of a talker uttering a set of 180 short sentences, under three conditions: normal speech (in quiet), Lombard speech (in noise), and whispering. We then produced an animated 3D avatar with similar shape and appearance as the original talker and used an error minimization procedure to drive the animated version of the talker in a way that matched the original performance as closely as possible. In a perceptual intelligibility study with degraded audio we then compared the animated talker against the real talker and the audio alone, in terms of audio-visual word recognition rate across the three different production conditions. We found that the visual intelligibility of the animated talker was on par with the real talker for the Lombard and whisper conditions. In addition we created two incongruent conditions where normal speech audio was paired with animated Lombard speech or whispering. When compared to the congruent normal speech condition, Lombard animation yields a significant increase in intelligibility, despite the AV-incongruence. In a separate evaluation, we gathered subjective opinions on the different animations, and found that some degree of incongruence was generally accepted.

Keywords: Lombard effect; Motion capture; Speech-reading; Lip-reading; Facial animation; Audio-visual intelligibility

1. Introduction

Humans are extremely versatile in their way of compensating for external factors in spoken communication. In the presence of acoustic noise, humans seamlessly adapt their speech production and perception strategies in order to compensate for the acoustic external conditions and maintain communication. According to the theory of Hyper–Hypo articulation (Lindblom, 1990), speakers tend to economize their speech production with the goal to make themselves understood in a particular communicative situation. Speech produced in noise exhibits not only increased loudness, but also larger articulatory movements (Fitzpatrick et al., 2011).

Adaptation occurs not only in the production system but also in speech perception. When the acoustic channel is subject to disturbances, humans tend to rely more on other information sources, such as vision. It is well known that the visual modality has a strong influence on speech perception, and can even alter the perception in its entirety (McGurk and MacDonald, 1976). In acoustically adverse conditions, visible speech enhances speech comprehension.

* Corresponding author. Tel.: +46 8 790 8965.
E-mail addresses: simonal@kth.se (S. Alexanderson), beskow@kth.se, jbeskow@gmail.com (J. Beskow).

0885-2308
http://dx.doi.org/10.1016/j.csl.2013.02.005

This audiovisual speech enhancement effect is particularly important for people with hearing impairments, but is equally present in normal hearing persons in the presence of external noise (Summerfield, 1992).

While both of these effects have been extensively studied in isolation, the combined effect – i.e. how audiovisual speech perception is affected by the presence of noise during speech production, has been considerably less investigated. According to Campbell and Mohammed (2010), speakers speaking using normal voice, as opposed to whispering or shouting, are usually judged easiest to speech-read, and Lombard speech can arguably be considered a form of shouting. On the other hand, Kim et al. (2011) found that the audiovisual speech enhancement in a speech in noise audiovisual sentence intelligibility task, was greater for speech recorded in noisy conditions than for speech recorded in quiet conditions. Similar results were reported by Vatikiotis-Bateson et al. (2007), in a study following up on Vatikiotis-Bateson et al. (2006), who found no support for larger AV enhancement for Lombard speech.

In our work we are concerned with talking characters that can be automatically animated from either text or speech with high enough accuracy that they provide substantial visual intelligibility enhancement. While several such systems have been presented (see Bailly et al. (2003) for an overview), none of these have the capability of modelling the types of articulatory variations associated with Lombard speech. Given the reported increases in AV intelligibility for Lombard speech, we want to investigate whether these enhancements carry over to our talking head animation, when we drive the animation from motion capture. More interestingly, we want to see if animation modelled after Lombard speech can be used to enhance the visual intelligibility in a *general* sense, i.e. even for a non-Lombard voice, in spite of the potential voice/face mismatch that this would entail.

The ability of a virtual character to adapt articulatory effort can be important for other reasons as well. Recent work in statistically driven speech synthesis (c.f. Raitio et al., 2011) include the capability of adapting the voice quality to Lombard speech. In the interest of coherence, we would expect the same capability to be present in a talking head animation, in the interest of an audio-visually coherent user experience. Similarly, if the animated face is being driven from real speech, as in Salvi et al. (2009), it would be desirable to have the ability to adapt to the speaking style of the input voice.

In the on-going Lipread project, we are designing a talking face to be used in an e-learning environment for training of speech-reading. In a training situation such as this one, it is important to be able to modify the animation style to give the learner a more diverse training situation. In summary, having the ability to model speaking style in terms of articulatory effort would potentially give us two things: facial animation that can conform to the style of a given real or synthesized voice, and a visual speech synthesis system where the degree of articulatory effort (and potentially the intelligibility), can be explicitly controlled.

In the current study we used speech in quiet, speech in noise and whispering as a way of eliciting different degrees of articulatory effort from our talker while he was recorded using motion capture equipment. Then we mapped the motion capture data onto an animated face rig and produced animated versions of the recorded sentences. Finally we conducted a perceptual evaluation study, where we compared the intelligibility of the generated facial animation against the natural video as well as audio-only versions.

2. Audio, video and motion corpus

We have recorded a multimodal (audio, video, motion capture) corpus covering different speaking styles. The purpose of the corpus is twofold. Firstly, we want to investigate the visual intelligibility of an animated talker with different speaking styles as outlined in Section 1, which is the main focus of this article. To this end we need suitable material for intelligibility testing. Secondly, we want the corpus to function as training data for a statistically based visual speech synthesizer, so we need a phonetically balanced data set. To satisfy these goals we have chosen to record a Swedish sentence set previously developed for the purpose of audiovisual intelligibility testing (see Öhman, 1998) as well as a set of nonsense VCV words.

2.1. Data recording

The speaker was a male Swedish actor who was seated face to face with a listener (approx. 2 m away), and was instructed to read short sentences and words from a monitor, and make sure that the listener understood what was being said. Both listener and speaker wore headphones (Sennheiser HD 600), where they could hear their own speech as picked up by a common omni-directional microphone, at a level that was pre-adjusted to roughly compensate for the