Speech Communication 000 (2017) 1–10

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom

The impact of automatic exaggeration of the visual articulatory features of a talker on the intelligibility of spectrally distorted speech

Najwa Alghamdi*, Steve Maddock, Jon Barker, Guy J. Brown

Department of Computer Science, University of Sheffield, United Kingdom

ARTICLE INFO

Article history:
Received 14 January 2017
Revised 18 July 2017
Accepted 28 August 2017
Available online xxx

Keywords:
Audiovisual training
Cochlear-implant simulation
Visual-speech enhancement
Lombard speech

ABSTRACT

Visual speech information plays a key role in supporting speech perception, especially when acoustic features are distorted or inaccessible. Recent research suggests that for spectrally distorted speech, the use of visual speech in auditory training improves not only subjects' audiovisual speech recognition, but also their subsequent *auditory-only* speech recognition. Visual speech cues, however, can be affected by a number of facial visual signals that vary across talkers, such as lip emphasis and speaking style. In a previous study, we enhanced the visual speech videos used in perception training by automatically tracking and colouring a talker's lips. This improved the subjects' audiovisual and subsequent auditory speech recognition compared with those who were trained via unmodified videos or audio-only methods. In this paper, we report on two issues related to automatic exaggeration of the movement of the lips/ mouth area. First, we investigate subjects' ability to adapt to the conflict between the articulation energy in the visual signals and the vocal effort in the acoustic signals (since the acoustic signals remained unexaggerated). Second, we have examined whether or not this visual exaggeration can improve the subjects' performance of auditory and audiovisual speech recognition when used in perception training. To test this concept, we used spectrally distorted speech to train groups of listeners using four different training regimes: (1) audio only, (2) audiovisual, (3) audiovisual visually exaggerated, and (4) audiovisual visually exaggerated and lip-coloured. We used spectrally distorted speech (cochlear-implant-simulated speech) because the longer-term aim of our work is to employ these concepts in a training system for cochlear-implant (CI) users.

The results suggest that after exposure to visually exaggerated speech, listeners had the ability to adapt alongside the conflicting audiovisual signals. In addition, subjects trained with enhanced visual cues (regimes 3 and 4) achieved better audiovisual recognition for a number of phoneme classes than those who were trained with unmodified visual speech (regime 2). There was no evidence of an improvement in the subsequent audio-only listening skills, however. The subjects' adaptation to the conflicting audiovisual signals may have slowed down auditory perceptual learning, and impeded the ability of the visual speech to improve the training gains.

1. Introduction

The robustness of human speech perception arises from a listener's ability to integrate and evaluate information from multiple sources. 'Audiovisual integration' refers to a listener's ability to utilise auditory and visual speech information in order to interpret the perceived message from the talker (Massaro and Cohen, 1983; Rosenblum and Saldaña, 1996). The illusion of perceiving a new audio signal when listeners are presented with an incongruent audiovisual signal - known as the McGurk effect (McGurk and MacDonald, 1976) - provides compelling evidence of the synergy of audio and visual speech during perception. This complementary support is also evident in adverse listening conditions, such as when listening to speech in noise, where visual speech cues can improve the speech intelligibility by 5–22 dB (Sumby and Pollack, 1954; MacLeod and Summerfield, 1987; Erber, 1969; Middelweerd and Plomp, 1987). The external articulators (lips, teeth, and tongue) can provide a significant proportion of the overall visual speech information gathered from the face (Summerfield et al., 1989; McGrath, 1985); McGrath (1985) found that the accuracy of identifying monophthongal vowels reached 56% by lipreading the external articulators only. By visualising these external articulators' kinematic information using a point-light technique,

* Corresponding author.
E-mail addresses: amalghamdi1@sheffield.ac.uk (N. Alghamdi), s.maddock@sheffield.ac.uk (S. Maddock), j.p.barker@sheffield.ac.uk (J. Barker), g.j.brown@sheffield.ac.uk (G.J. Brown).

http://dx.doi.org/10.1016/j.specom.2017.08.010
0167-6393/

Rosenblum et al. (1996) found that such visualisation can substantially enhance the intelligibility of speech in noise. Compared with other facial movements that provide temporal cues only, kinematic information from the mouth can provide both speech information and temporal cues (Kim and Davis, 2014b).

Visual speech also significantly contributes to speech intelligibility when listeners undergo 'internal' adverse conditions (i.e. when listeners suffer limitations in perceptual skills), such as in the case of cochlear implant (CI) users. CI users experience the noise inherited from CI-processed speech, since such speech lacks important spectral and temporal cues necessary for pitch perception (Erber, 1982; Kaiser et al., 2003; Desai et al., 2008; Li et al., 2013). In addition, the amount of acoustic information CI users can receive is a function of various physical and physiological factors, including the number of implanted and activated electrodes and the severity of damage to the hearing nerve (Nie et al., 2006). Fortunately, a talker's face provides a medium in which CI users can speech-read visibly distinguished phoneme categories that are difficult to hear, such as labial (anterior consonants like bilabials) or non-labial (posterior consonants) phonemes.

Attending to visual speech in face-to-face communication is not the only strategy that CI users utilise as a way of overcoming degraded acoustic cues. Auditory perception training (or *auditory training*), for example, can significantly improve CI-listening abilities in audio-only modalities. Such training aims to enhance the central auditory system (CAS) responses to sound (i.e. CAS plasticity) by using auditory perceptual learning (Lazard et al., 2013). Recent studies have found a link between induced CAS plasticity and the introduction of visual speech in auditory training (i.e. *audiovisual training*) in which auditory skills are improved. Such studies found that the perceptual learning gained from audiovisual training was more effective in enhancing CI-simulated speech perception by normal-hearing listeners than was the case with auditory training (Bernstein et al., 2013; Pilling and Thomas, 2011; Kawase et al., 2009; Alghamdi et al., 2015a). In auditory training, acoustic signals guide 'top-down' perceptual learning. When acoustic signals are compromised by noise, however (such as in the case of CI users), the available visual signals that the audiovisual training offers can provide external support to help develop auditory perceptual learning (Bernstein et al., 2013). This can consequently shape a perceptual experience that listeners can later utilise to comprehend novel stimuli, even in auditory-only situations.

The benefits of visual speech may be affected by a range of talker-dependent factors, including lip emphasis (Lander and Capek, 2013; Campbell and Mohammed, 2010; Scott, 1987), teeth and tongue visibility (McGrath, 1985; Preminger et al., 1998; Rosenblum and Saldaña, 1996), facial hair (Kitano et al., 1985), speaking style (Kaplan et al., 1985; Scott, 1987), and talker's gender (Daly et al., 1997). This raises the question of whether or not enhancing visual speech quality can increase the benefit of such speech. In a previous study (Alghamdi et al., 2015b), we found that enhancing the appearance of a talker's lips within audiovisual training stimuli can help to improve the audiovisual and subsequent auditory recognition of CI-simulated speech by non-native, normal-hearing subjects. Using image processing and computer graphics techniques, we simulated the talker wearing lipstick in the training stimuli. According to Lander and Capek's (2013) observations of talkers who wore real lipstick, this is an effect that can improve lip-reading. We found improved training gains in audiovisual and audio-only sentence-recognition rates when the subjects were trained by audiovisual speech produced by a talker wearing simulated lipstick compared with the original unmodified audiovisual speech and the audio-only speech of the same talker (Alghamdi et al., 2015b).

Speaking style is a determining factor in the quality of visual speech (Kaplan et al., 1985; Scott, 1987). According to Lindblom's

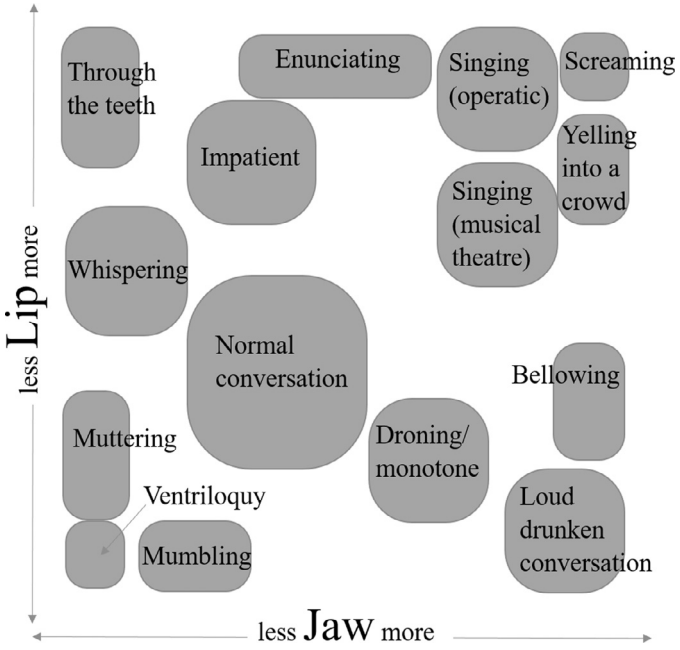

Fig. 1. The transition from hypo- to hyper-articulation in the Jaw Lip viseme model (JALI), which simulates visual speech by aggregating functions related to jaw motion and lip motion (Edwards et al., 2016): This shows that distinct speaking styles exert different articulatory-energy modifications. Permission to use this image has been granted by Edwards et al. (2016).

hypo-hyper (H&H) theory of speech production (Lindblom, 2012), talkers make articulatory energy modifications from hypo- to hyper-articulated speech in order to adapt to the demands of the listening situation. This may create a variety of speaking styles that exert different energy magnitudes in order to move the external articulators (Fig. 1) (Edwards et al., 2016). This concept led us to investigate the transition from hypo- to hyper-articulated speech as a modification method for enhancing visual speech. Theobald et al. (2006) addressed speaking-style exaggeration in 2D videos in order to support the forensic lip-reading of surveillance videos; they exaggerated the lip movements of a talker in a video by amplifying the principle components (PCs) of the talker's mouth shapes and appearance that were elicited from the video. They found improved lip-reading performance among inexperienced lip-readers.

The study we present in this paper investigates the impact of exaggerating the visual articulatory features (in particular the mouth kinematics) on the visual benefit of audiovisual speech. The main challenge of such an enhancement is that the combination of exaggerated visual signals and the original unexaggerated acoustic signals could create conflicting, incongruent inputs. The impact of exposure to conflicting inputs has been widely investigated in the behavioural-studies literature (McGurk and MacDonald, 1976; Bertelson et al., 2003; Bermant and Welch, 1976; Saldaña and Rosenblum, 1993; De Gelder and Vroomen, 2000). These impacts can be classified as immediate biases or after-effects (Bertelson et al., 2003). An example of immediate biases may be observed in spatial conflict (such as the ventriloquism effect), where visual stimuli can influence sound localisation (Bermant and Welch, 1976), and in identity conflict (i.e. the McGurk effect) McGurk and MacDonald (1976). Another study also found that exposure to mismatched inputs can create an after-effect on perceptual modalities used for adapting to this conflict. For example, visual speech has the ability to recalibrate auditory perception after exposure to the conflicting audiovisual stimuli observed in the McGurk effect (Bertelson et al., 2003). Audiovisual recalibration has also been observed in tempo-

Please cite this article as: N. Alghamdi et al., The impact of automatic exaggeration of the visual articulatory features of a talker on the intelligibility of spectrally distorted speech, Speech Communication (2017), http://dx.doi.org/10.1016/j.specom.2017.08.010