Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 74 (2015) 37–51

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

The effect of seeing the interlocutor on auditory and visual speech production in noise ☆

Michael Fitzpatrick *, Jeesun Kim, Chris Davis

The MARCS Institute, University of Western Sydney, Australia

Received 17 March 2015; received in revised form 25 June 2015; accepted 17 August 2015
Available online 28 August 2015

Abstract

Talkers modify their speech production in noisy environments partly as a reflex but also as an intentional communicative strategy to facilitate the transmission of the speech signal to the interlocutor. Previous studies have shown that talkers can adapt both auditory and visual elements of speech produced in noise. The current study examined whether talkers adapt their auditory and visual speech production strategy as a function of their communicative setting. Participants completed an interactive communication game in various quiet and in noise conditions with/without being able to see their interlocutor. We found that the energy of talkers' speech modifications was significantly lower in conditions when interlocutors could see each other relative to conditions where they could not. Further, talkers increased the saliency of their visual speech production (measured as lip-area and lip-width) in noisy conditions for face-to-face communication. In a set of perception studies, using the speech materials from the production study as stimuli, we also demonstrated that the shift in speech production strategy across the face-to-face and non-visual communicative conditions corresponded with changes in the auditory and auditory–visual intelligibility of the speech signal produced. The results suggest that talkers actively monitor their environment and are able to adopt appropriate speech production strategies for efficient and effective communication in adverse conditions.

Keywords: Lombard speech; Speech production; Speech perception; Auditory–visual speech; Face-to-face communication

1. Introduction

One of the key factors underlying our ability to communicate in noise is that we adapt our speech production in noisy environments. When faced with noise, talkers adopt a speech production style known as Lombard speech (following Lombard, 1911). Lombard speech is characterised by increases in loudness, vowel duration and F0, as well as a shifting of energy in the speech spectrum from lower to mid/high frequencies, i.e., reduced spectral tilt (e.g., Cooke and Lu, 2010; Junqua, 1993; Patel and Schell, 2008). Some studies also suggest that the dispersion of vowel properties is altered for Lombard speech (e.g., Bond et al., 1989; Cooke and Lu, 2010). These modifications make the auditory signal significantly more intelligible for the listener in noisy environments (e.g., see Lu and Cooke, 2008; Pittman and Wiley, 2001; Van Summers et al., 1988).

Initial research examining the effects of noise on speech production conceptualised Lombard speech as a reflex, triggered by a reduction in the audibility of one's own voice (e.g., see Egan, 1972; Pick et al., 1989). Many studies have demonstrated that talkers will adapt their speech production when auditory feedback is attenuated by the presence of noise or by conditions that simulate hearing loss (e.g., Chang-Yit et al., 1975; Chen et al., 2007; MacDonald et al., 2010). More recently, however, research examining

☆ A preliminary version of this work was presented at Interspeech, 2011; and AVSP, 2011.

* Corresponding author. Tel.: +61 417406843.
E-mail address: m.fitzpatrick@uws.edu.au (M. Fitzpatrick).

http://dx.doi.org/10.1016/j.specom.2015.08.001
0167-6393/

Lombard speech under task conditions where there is interaction and feedback between interlocutors has provided evidence that Lombard speech production changes are also tailored to benefit the listener in noise. For example, the speech production changes made when interacting with an interlocutor in noise (i.e., where there is a "premium on intelligible communication", Lane and Tranel, 1971, p. 262) tend to be larger and more pronounced compared to those where communication is either absent or minimal, such as for reading lists or simply repeating words aloud(e.g., Cooke and Lu, 2010; Garnier et al., 2010; Junqua et al., 1999). Additionally, talkers also adjust their speech production in response to the properties of the background noise for the benefit of the interlocutor (e.g., see Aubanel and Cooke, 2013a, 2013b; Cooke and Lu, 2010; Lu and Cooke, 2008).

Collectively, these findings have led to the proposal that Lombard speech is the result of both adaptation to noise and communicative strategies (Garnier et al., 2010; Hazan and Baker, 2011; Lane and Tranel, 1971). This proposal points to the need to study Lombard speech in communicative situations, where there is feedback and interaction between interlocutors, in order to explore the modifications that are made as a function of adaptation to noise, and those that are motivated by improving the intelligibility of speech for the listener.

The hyper- and hypo-articulation (H&H) model described by Lindblom (1990) provides a useful theoretical framework that can accommodate the reflexive and communicative adaptations of Lombard speech. This model proposes that variability in produced speech reflects a dynamic adjustment that attempts to satisfy the competing goals of minimising articulatory effort and maximising intelligibility. When demands for intelligibility are low, talkers are more likely to revert to a reduced style of speech production (i.e., 'hypo-speech') which minimises the effort of production. However, when the demands for intelligibility are increased (such as when talking in noise), talkers are more likely to increase the clarity of their produced speech in order to effectively communicate with their conversational partner. One of the implications of the H&H model is that talkers are sensitive not only to the constraints placed on them by the environment (i.e., noise), but also to the specific needs of the listener. Through ongoing feedback and interaction between speech partners, talkers are able to converge upon production strategies that optimally balance efficiency of output with effectiveness of communication (Lindblom, 1990).

Based on the H&H model of production, changes in the communicative environment may be expected to have a significant impact on speech production in noise. However, previous research has primarily focused on the effect of the auditory environment (i.e., the level of noise presented to the talkers) on the auditory signal properties of Lombard speech. Relatively few studies have examined how differences in communicative setting might influence speech production in noise (although see Aubanel et al., 2012; Cooke and Lu, 2010; Garnier et al., 2010; Hazan and Baker, 2011). In the current study, therefore, we examined whether talkers make different auditory and visual speech adaptations in noise depending on whether they are communicating with their interlocutor face-to-face (FTF), or whether they have to rely only on auditory cues to communicate (i.e., in a non-visual (NV) condition).

It is currently unclear whether Lombard speech production varies between FTF and NV conditions. Previous studies examining Lombard speech in communicative settings have either restricted communication to NV conditions (e.g., Cooke and Lu, 2010), or to FTF conditions (e.g., Garnier et al., 2010). However, there are several reasons to expect that this variation in communicative setting (i.e., FTF vs. NV) will matter for Lombard speech production. First, it is well established that seeing the talker's moving face (visual speech) enhances speech perception in noise (e.g., Robert-Ribes et al., 1998; Ross et al., 2007; Sumby and Pollack, 1954; Summerfield, 1987). Second, it has been demonstrated that speech produced in noise is also accompanied by changes to the visual speech signal ("visual Lombard speech", e.g., see Garnier et al., 2010; Kim et al., 2011). While larger lip/jaw motion is generally required to produce louder speech (e.g., see Schulman, 1989), these visible Lombard speech changes have been demonstrated to lead to greater enhancements in the auditory–visual (AV) intelligibility of the speech signal the interlocutor receives (Kim et al., 2011), suggesting that talkers might have some flexibility to adapt the intelligibility of their visual Lombard speech for their interlocutor. Plausibly then, when the interlocutors can see each other, compared to when they cannot, there will be less need for the talker to strive to make the auditory signal intelligible. Equally, for both talker and listener, there will be an increased premium on producing visually intelligible speech. Therefore, it can be predicted that talkers will make more pronounced visual but less pronounced auditory modifications in FTF communication, relative to non-visual (NV) conditions.

The aim of the current study was to test this prediction by addressing two specific questions: (1) whether talkers in noise adopt different speech production strategies across communicative conditions where they either can or cannot see their interlocutor; and (2) if so, whether the modifications talkers make to their auditory and visual Lombard speech across the different communicative settings lead to changes in intelligibility.

To examine the first question we conducted a speech production study where we analysed talkers' auditory and visual Lombard speech in FTF and NV conditions. An important consideration in the study was to examine speech production in settings where there was an emphasis on the need for talkers to communicate to complete the task. That is, we wanted to examine speech produced in conditions where talkers were free to interact with one another, received feedback as to the success or failure of their communication, and were able to adapt their speech