Available online at www.sciencedirect.com
SciVerse ScienceDirect
Computer Speech and Language xxx (2013) xxx–xxx
COMPUTER SPEECH AND LANGUAGE
www.elsevier.com/locate/csl

Comparing the consistency and distinctiveness of speech produced in quiet and in noise☆

Jeesun Kim, Chris Davis *

MARCS Institute, University of Western Sydney, Australia

Received 17 October 2012; received in revised form 29 January 2013; accepted 8 February 2013

Abstract

The study investigated whether properties of speech produced in noise (Lombard speech) were more distributed (thus potentially more distinct) and/or more consistent than those from speech produced in quiet. This was examined for auditory tokens by measuring vowel space dispersion and by determining the consistency of formant production across repeated instances. Vowel space was not expanded for speech produced in noise; there was a tendency for formants to be produced more consistently in noise (with less variation in formant frequency across repeated instances) but this was not a secure effect. The distinctiveness and consistency of Lombard visual speech were also examined using motion capture data. Relative distinctiveness was determined by comparing the amount of mouth and jaw motion for speech produced in noise and quiet; relative consistency by comparing the size of correlations for motion produced across repeated instances in the noise or in quiet conditions. Mouth, and jaw motion was larger for speech in noise, however there was no greater association between the movement measures for repeated instances of speech in noise compared to in quiet. We also examined whether the correlation between auditory and motion properties was greater for speech produced in noise than in quiet. It was found that the association between speech RMS energy and jaw motion was greater for speech in noise. The results show that although Lombard speech affects both auditory and visible articulatory properties in ways likely to enhance speech perception it does not increase production consistency.

Keywords: Vowel dispersion; Speech consistency; Lombard speech; Visual speech; Auditory visual correlation; Speech in noise; Vowel space

1. Introduction

Noise is commonly a part of the speaking and listening environment. When speaking in noise, talkers change the way they articulate and the way that speech sounds (the so called Lombard effect). In terms of acoustic properties, Lombard speech typically has greater duration, an increased F0 and increased energy at higher frequencies (Lombard, 1911; Hansen, 1996; Junqua, 1993) and such speech is more intelligible than speech produced in quiet (even when noise is mixed with both at the same signal to noise ratio, see Junqua, 1993). Much of the research on the Lombard effect has been on auditory speech, e.g., determining the acoustic changes that occur.

☆ This paper has been recommended for acceptance by 'Dr. Martin Cooke'.

* Corresponding author at: MARCS Institute, University of Western Sydney, Locked Bag 1797, Penrith, NSW 2751, Australia. Tel.: +61 2 9772 6855; fax: +61 2 9772 6040.
E-mail address: chris.davis@uws.edu.au (C. Davis).

0885-2308/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.csl.2013.02.002

Although articulatory variations in Lombard speech have been inferred from the observed changes in the acoustic signal (e.g., features related to the configuration of vocal tract, formant locations, mouth opening and sound radiation influenced by lips, see Bond et al., 1989; Summers et al., 1988), relatively little research has directly measured changes in articulations that would be clearly visible from talkers' face/head motion (i.e., the visual speech correlates of Lombard speech). It has, however, been reported that visual Lombard speech is produced with increased motion (both for movements directly related to articulation, Garnier, 2007, as well as those for that typically accompany speech, e.g., head motion, Kim et al., 2005). Moreover, there is some evidence that the visual Lombard effect is influenced by the communication factors (e.g., interactive versus non-interactive) that have been shown to affect auditory Lombard speech production (Lane and Tranel, 1971; Cooke and Lu, 2010). For example, it has been shown that in a face-to-face setting, talkers increase the saliency of their visual speech production (measured as lip-area) in noisy conditions (Fitzpatrick et al., 2011) and Garnier (2007) has pointed out the importance of understanding how a talker reorganizes her/his communication strategies in different noise environments. Further, it has been shown that when a listener can see the talker, articulatory movements produced in noise provide a more effective boost to the intelligibility of auditory speech than those produced in quiet (Fitzpatrick et al., 2012; Kim et al., 2011).

The increase in intelligibility for speech produced in noise poses the question of which speech signal properties may be responsible. Not all noise-induced changes in the speech signal increase intelligibility, for example, it has been shown that the increased F0 has little or no effect on intelligibility (Lu and Cooke, 2009). Here, we consider two production strategies that might facilitate intelligibility in noise: vowel space expansion and consistency of speech production. A talker might obtain increased intelligibility in noise if they expand the space of their vowel productions (allowing for potentially better vowel discrimination). For example, in read speech produced in quiet, talkers with larger vowel spaces were generally more intelligible than those with reduced ones (Bradlow et al., 1996). Likewise, change in vowel space arguably plays a role in infants' phonetic discrimination abilities (Liu et al., 2003). However, evidence on vowel space expansion in Lombard speech is mixed. For instance, Cooke and Lu (2010) found that in noise there were no significant changes in between-category vowel dispersion. On the other hand, Bond et al. (1989) reported that speech in noise had a more compact vowel space than that produced in quiet. Whereas others have suggested that there was a tendency for the vowel space to be expanded in noise (Mixdorff et al., 2007). Another strategy that a talker could adopt to increase intelligibility in noise is to be very consistent in the manner in which speech is produced. It has been suggested that speech perception is easier for less varied utterances (see Newman et al., 2001; Uchanski et al., 1992). This consistency could be expressed either by maintaining the constancy of spectral properties and/or the constancy in the produced duration of segments.

The current study sought evidence for use of the above two strategies in Lombard speech production (i.e., more distributed vowels and/or more consistent speech) in both auditory and visual domains. In auditory domain, vowel space expansion was examined with speech produced in quiet and in multi-talker babble noise (presented over earphones). Several measures were employed to index the production consistency of the quiet and the noise conditions: (i) the size of the correlation of first and second formant values across repeated instance of the same sentence; (ii) the mean absolute difference for first and second formant values across repeated instances. In visual domain, the distinctiveness of visual speech was determined by simply examining the size of motion and consistency by comparing the size of the correlation of jaw and mouth movements across repeated instances of the same sentence. Further, we examined whether there was a greater correlation between auditory and speech motion for speech produced in noise by measuring the association between time-aligned guided principle motion components (that relate to speech articulation) and RMS energy.

2. Methods

2.1. Participants

Four speakers, one female and three male speakers (M age = 40 years) participated in the data capture sessions. To have varied speech renditions, we selected English speakers that had different English accent: the female speaker was a native speaker of American English; two males were native speakers of Australian English and the other male a speaker of British English. All participants reported having normal vision and hearing.

Please cite this article in press as: Kim, J., Davis, C., Comparing the consistency and distinctiveness of speech produced in quiet and in noise. Comput. Speech Lang. (2013), http://dx.doi.org/10.1016/j.csl.2013.02.002