Speech Communication 100 (2018) 18–29

Contents lists available at ScienceDirect

Speech Communication

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

Measuring communication difficulty through effortful speech production during conversation

Timothy Beechey[a,b,c,*], Jörg M. Buchholz[a,c], Gitte Keidser[a,b,d]

[a] HEARing Cooperative Research Centre, Melbourne, Australia
[b] National Acoustic Laboratories, Sydney, Australia
[c] Department of Linguistics, Audiology section, Faculty of Human Sciences, Macquarie University, Sydney, Australia
[d] School of Health and Rehabilitation Sciences, University of Queensland, Brisbane, Australia

ARTICLE INFO

Keywords:
Lombard speech
Conversation
Communication
Effort
Accommodation

ABSTRACT

This study describes the use of a novel conversation elicitation framework to collect fluent, dynamic conversational speech in simulated realistic acoustic environments of varying complexities. Our aim is to quantify speech modifications during conversation, which characterize effortful speech, as a function of the difficulty of the acoustic environment. We report speech production data at the acoustic-phonetic level (vocal level, mid-frequency emphasis, formant frequencies and formant bandwidths), as well as at higher levels of analysis including utterance duration and turn overlap durations. The sensitivity and test-retest reliability of different speech production measures to changes in acoustic environment are reported. We propose a multi-dimensional view of effortful speech modifications. Considering speech modifications across different linguistic levels provides a richer view of the effects of the acoustic environment on communication as compared with consideration of low-level acoustic-phonetic markers alone. Finally, we describe how consideration of speech modification data may form the basis of a measure of communication effort with scope for the assessment of the impacts of hearing impairment and amplification upon ease of spoken communication.

1. Introduction

Speech produced in noise (Lombard speech) is characterized by increased vocal effort which is manifested in acoustic changes such as increased intensity, mid-frequency emphasis, higher first formant (F1) frequencies and fundamental frequency (F0). Many studies of Lombard speech have argued that these speech modifications have a communicative basis (Cooke and Lu, 2010; Garnier et al., 2010; Hazan and Baker, 2011; Junqua et al., 1999; Lane and Tranel, 1971). The communicative view of Lombard speech attributes the speech modifications listed above to talkers' intention to increase the intelligibility of their speech for the hearer, relative to neutral speech, in difficult listening conditions. This is consistent with the fact that communication is an inherently interactive behavior which is shaped by dynamic feedback between interlocutors, and accommodation in response to that feedback (Schober and Clark, 1989). Dynamic feedback and accommodation distinguish conversation from passive listening. These two strategies can help to improve communication by providing interlocutors with opportunities to signal comprehension difficulties and therefore to influence the speech production of their communication partner. For example, Branigan et al. (2011) demonstrated that hearers' comprehension when listening to dialogs was better than when listening to monologues and was maximized when the hearer participated in a dialog. When overhearing a dialog, the hearer benefited from the feedback and accommodation that occurred between talkers. When participating in the dialog the hearer was able to elicit accommodation tailored to their own comprehension difficulties. However, studies of Lombard speech have generally not considered holistic communicative contexts and therefore may not adequately reflect many aspects of realistic communicative interactions. Many Lombard speech studies have not considered speech produced during conversations and therefore have not captured the effects of dynamic feedback and accommodation between interlocutors. For example, Lu and Cooke (2009) measured changes in speech when talkers read sentences to a listener and Junqua et al. (1999) considered speech directed towards an automated voice recognition system. While Cooke and Lu (2010) and Hazan and Baker (2011) measured acoustic phonetic changes in speech during conversational interactions, both studies separated talkers into different booths, or with an acoustically transparent screen. Such separation removes important aspects of natural interaction, including visual cues and a more general sense of co-location. Relatively few studies have considered speech modifications that occur within

---

[*] Corresponding author.
E-mail address: tim.beechey@nal.gov.au (T. Beechey).

https://doi.org/10.1016/j.specom.2018.04.007
Received 28 June 2017; Received in revised form 28 March 2018; Accepted 23 April 2018

conversations between co-located talkers. Notable exceptions include Aubanel et al. (2011) where talkers sat across a table without any visual obstruction as well as studies of visual analogues of Lombard speech (Davis et al., 2006; Kim et al., 2005).

In addition, the types of maskers generally employed in Lombard speech studies have either been stationary noise or constructed babble noise. Live competing speech has been employed as a masker (Aubanel and Cooke, 2013; Aubanel et al., 2011, 2012) which introduces informational masking and allows for the study of temporal strategies whereby talkers attempt to exploit predictable gaps in competing speech to maximize the intelligibility of their own speech. Lombard speech studies have not, to-date, considered the effects of realistic background noise representing real-world locations where conversations are likely to take place.

Finally, the majority of Lombard speech studies have considered low-level acoustic-phonetic parameters such as vocal level, F0, formant frequencies, spectral tilt and vowel duration. Only relatively few studies have considered factors inherent to conversational interaction such as turn-taking and talker overlaps (Aubanel and Cooke, 2013; Aubanel et al., 2011, 2012). As a result, relatively little is known about how conversational dynamics affect Lombard speech. Consideration of speech modifications at higher linguistic levels is crucial for our understanding of the communicative nature of Lombard speech and how talkers may employ different strategies in different circumstances. For example, a talker may vary their vocal level independently of their speaking rate or they may vary their F0 independently of their turn-taking behavior. As a result, it is logically possible for talkers to modify their speech in complex and even contradictory ways. Analysis of speech modifications at different linguistic levels may therefore provide a richer understanding of communicative effort than consideration of a single level of behavior. To understand communicative effort, it is informative to consider different ways in which talkers may modify their speech in challenging communication settings. Among other strategies, a talker may modify their speech in terms of: (i) the rate of vocal fold vibration which forms the voiced sound source of speech; (ii) the articulation of speech sounds through the shape and compliance of the vocal tract; (iii) overall vocal level; (iv) rate of production; (v) length and complexity of utterances; or (vi) manner of interaction with their communication partner, such as turn-taking behavior. These modifications reflect vocal behavior at different linguistic levels from low-level acoustic-phonetic changes up to prosodic, syntactic and discourse-pragmatic changes. A comprehensive review of talker strategies is provided by Cooke et al. (2014a).

The aim of this study was to investigate how talkers modify their speech when communicating in realistic acoustic environments of differing complexity at both the acoustic-phonetic level and the interactive level. As a secondary aim, we sought to investigate the reliability of automated acoustic analyses, rather than manual annotation methods, to determine whether the rapid acquisition of speech effort data could plausibly be employed in clinical settings in the future. It was hypothesized that speech modifications at the acoustic-phonetic level, such as vocal level, F0 and formant frequencies, will follow a different pattern of change than modifications at the interactive level, such as turn-taking behavior. Consideration of such a range of speech modifications provides a richer understanding of communicative strategies employed by talkers than consideration of acoustic-phonetic factors alone. It will be argued that automatically extracted Lombard speech measures at multiple linguistic levels may be used to measure changes in communication difficulty and effort during conversation. This approach may be generalized to measure the effects of other factors such as hearing impairment or cognitive impairment on degree of communication difficulty.

2. Methods

2.1. Subjects and materials

Ten male and 10 female native Australian English-speaking adults aged between 18 and 51 years (mean = 28.7 years, standard deviation = 7.97 years) with normal pure-tone hearing thresholds (i.e. < 20 dB HL) between 250 Hz and 8 kHz were tested in pairs. Participants were recruited through advertisements on the Macquarie University campus and through word-of-mouth and received a payment to cover their travel expenses. Participants were naive to the purpose of the study. Treatment of participants was approved by the Australian Hearing Ethics Committee and conformed in all respects to the Australian Government's National Statement on Ethical Conduct in Human Research.

2.1.1. Conversation elicitation task

In order to record fluent, dynamic conversations, a puzzle task was designed to elicit realistically complex utterances and balanced contributions from participants while encouraging engagement. The purpose of the task described here is solely to facilitate fluent, balanced conversations which are as representative of everyday verbal communication as possible. Completion of the task is not a measure of interest as task completion may depend on cognitive resources that are not directly relevant for successful communication. A total of 8 puzzles were constructed on 10 × 10 grids with each square containing a tangram image and one of three colors, which were labeled in the subject instructions as “pink”, “dark blue” and “light blue” (Fig. 1a). The square colors were chosen to allow for the collection of multiple tokens of the corner vowels from the color names p[ɪ]nk, d[ɐ]rk bl[u]e and light bl[u]e, though analysis of specific vowels is not a focus of the present study.

The object of the puzzles is to find the unique path from the marked start square to the diagonally opposite end square by moving horizontally or vertically between squares containing identical colors or pictures. A single puzzle was created and then 7 additional puzzles were derived with identical structures by rotating and flipping the original puzzle and substituting different tangram images. The complexity of the puzzles ensured that participants could not detect that the puzzles had a common solution. For each puzzle, two complementary participant views were created by removing half of the information from each

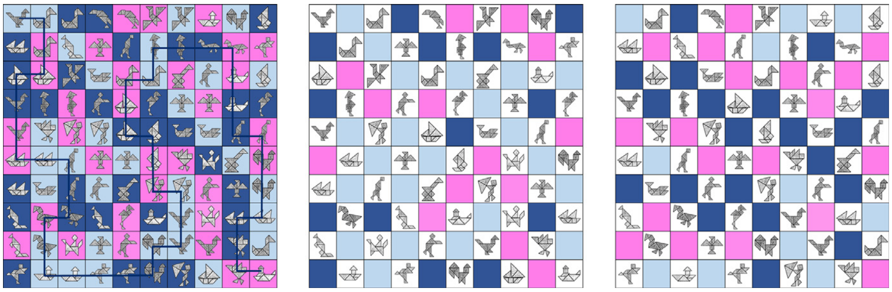

Fig. 1. Example of a complete (unseen) puzzle with solution (left panel) and complementary participant views (center and right). (For interpretation of the references to color in this figure, the reader is referred to the web version of this article.)