Available online at www.sciencedirect.com

ScienceDirect

Computer Speech and Language 28 (2014) 629–647

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Approaching speech intelligibility enhancement with inspiration from Lombard and Clear speaking styles☆

Elizabeth Godoy[a,*], Maria Koutsogiannaki[a,b], Yannis Stylianou[a,b]

[a] *Institute of Computer Science, Foundation for Research and Technology Hellas, Crete, Greece*
[b] *Multimedia Informatics Lab, Computer Science Department, University of Crete, Greece*

Received 15 October 2012; received in revised form 5 August 2013; accepted 11 September 2013
Available online 4 October 2013

Abstract

Lombard and Clear speech represent two acoustically and perceptually distinct speaking styles that humans employ to increase intelligibility. For Lombard speech, increased spectral energy in a band spanning the range of formants is consistent, effectively augmenting loudness, while vowel space expansion is exhibited in Clear speech, indicating greater articulation. On the other hand, analyses in the first part of this work illustrate that Clear speech does not exhibit significant spectral energy boosting, nor does the Lombard effect invoke an expansion of vowel space. Accordingly, though these two acoustic phenomena are largely attributed with the respective intelligibility gains of the styles, present analyses would suggest that they are mutually exclusive in human speech production. However, these phenomena can be used to inspire signal processing algorithms that seek to exploit and ultimately compound their respective intelligibility gains, as is explored in the second part of this work. While Lombard-inspired spectral shaping has been shown to successfully increase intelligibility, Clear speech-inspired modifications to expand vowel space are rarely explored. With this in mind, the latter part of this work focuses mainly on a novel frequency warping technique that is shown to achieve vowel space expansion. The frequency warping is then incorporated into an established Lombard-inspired Spectral Shaping method that pairs with dynamic range compression to maximize speech audibility (SSDRC). Finally, objective and subjective evaluations are presented in order to assess and compare the intelligibility gains of the different styles and their inspired modifications.

Keywords: Lombard effect; Clear speech; Intelligibility enhancement

1. Introduction

In real-world communications, speakers and listeners are often immersed in noisy environments that influence both speech production and intelligibility. When faced with adverse communication conditions, human beings physically alter their manner of speaking in order to make their speech more intelligible. For example, humans adopt "Lombard" or "Clear" speaking styles, depending respectively on whether or not they are immersed in a noisy environment. Just as humans adopt their speech production to increase intelligibility, speech enhancement techniques seek to modify the speech signal in order to make it more intelligible for human listeners. With growing numbers of applications using

☆ This paper has been recommended for acceptance by S. King.

* Corresponding author. Tel.: +30 6943851605.
E-mail address: godoyec@gmail.com (E. Godoy).

0885-2308/$ – see front matter
http://dx.doi.org/10.1016/j.csl.2013.09.007

speech technologies in commercial (e.g., mobile telephone, GPS, customer service systems), military (e.g., Air Force, Ground troop relays) and medical (e.g., assisted-speech) contexts, modifications that help "speaking-devices" to be more intelligible (and consequently, relevant) are currently in high demand. Considering the intelligibility gains of the human speaking styles, acoustic phenomena observed in Lombard and Clear speech can be used to inspire speech signal modifications for intelligibility enhancement. Indeed, this is the approach adopted in the present work.

Though Lombard and Clear speech are both highly intelligible, the styles are perceptually and acoustically distinct. To begin with, the "Lombard" effect refers to the ways in which humans reflexively modify their speech when speaking in a noisy environment (Lombard, 1911). Perceptually, Lombard speech can be described as "tense" and "loud," with increased vocal effort. Compared its "normal" counterpart, Lombard speech incorporates many prosodic and segmental acoustic-phonetic modifications (Summers et al., 1988; Hanley and Steer, 1949; Dreher and O'Neill, 1957; Junqua, 1993; Womack and Hansen, 1996; Garnier et al., 2006; Lu and Cooke, 2008, 2009; Lu, 2009; Davis and Kim, 2012). Observations from these Lombard-related works include: decreased speaking rate, increased pitch, higher energy, decreased spectral tilt or spectral "flattening," as well as formant shifts particularly for F1 and F2, formant bandwidth reduction and vowel-to-consonant energy re-distribution. Among these observations, the Lombard increase in intelligibility has been shown to be largely attributed to spectral modifications (Lu and Cooke, 2009), particularly increased spectral energy in an inclusive formant band or, otherwise stated, a decreased spectral "tilt."

Alternatively to the Lombard effect, "Clear" speech strategies are adopted when a speaker in a quiet environment addresses a listener facing a communication barrier. Perceptually, Clear speech can be described as overly or extremely articulated speech, with an increased effort to distinguish between sounds, often involving slowing down the speaking rate. Unlike the Lombard reflex, "Clear" speech is the result of an active communication strategy that can vary from speaker-to-speaker. Nonetheless, a variety of such strategies have been analysed in many works and certain common characteristics have emerged (Picheny et al., 1986, 1989; Krause and Braida, 2004a,b; Drullman et al., 1994a,b; Hazan and Simpson, 1996; Hazan and Markham, 2004; Hazan and Baker, 2010, 2011; Bradlow et al., 2003; Ferguson and Kewley-Port, 2002, 2007; Liu, 2006; Amano-Kusumoto and Hosom, 2011). Specifically, compared to "casual" or "conversational" speech, Clear speech has been shown to exhibit: decreased speaking rate (with increased vowel duration and longer, more frequent pauses), increased pitch, increased consonant energy, expanded vowel space (with corresponding F1 and F2 shifts), spectral flattening or increased energy at higher frequencies and increased modulation depth in the temporal signal envelope. Of these modifications, it appears that the spectral flattening and especially vowel space expansion are among the most influential phenomena (Amano-Kusumoto and Hosom, 2011).

While Lombard and Clear speech have individually been well-studied in the literature, the styles are rarely addressed simultaneously. One exception adopting a similar approach to this work can be found in Skowronski and Harris (2006), in which commonalities between the two styles in terms of energy re-distribution between voiced and unvoiced parts of speech are exploited in developing speech intelligibility enhancement algorithms. Unlike Skowronski and Harris (2006), the present work focuses on spectral phenomena observed in Lombard and Clear speech. First, the same acoustic analyses are applied to distinct Lombard-normal and Clear-casual corpora in an effort to highlight significant properties of the styles and compare corresponding observations. As can be discerned from the discussions above, previous studies have mentioned varying degrees of spectral flattening associated with Clear speech and formant shifts for the Lombard effect. However, the relative extent of these modifications has not been compared between the two styles. While the corpora in this work are distinct, the analyses and processing are common and key observations are drawn in comparing statistics of each intelligible style with its respective counterpart, thus remaining consistent within each corpus. Consequently, analyses in the present work take initial steps towards merging studies of Lombard and Clear speech. Nonetheless, the underlying questions ultimately concern how to exploit the spectral characteristics observed in Lombard and Clear speech in designing algorithms for speech signal intelligibility enhancement.

Generally, speech modifications for intelligibility enhancement can be classified into several groups. First, there are techniques to enhancing intelligibility that exploit audio and signal properties, such as the amplitude compression scheme in Niederjohn and Grotelueschen (1976), dynamic range compression in Blesser (1969) and a method for peak-to-rms reduction in Quatieri and McAulay (1991). Second, certain speech intelligibility enhancement methods focus on speech-in-noise and exploit knowledge of the noise masker, such as the optimizations based on a speech intelligibility index in Sauert and Vary (2006) and the glimpse proportion maximization in Tang and Cooke (2011). Third, in the context of text-to-speech synthesis, adaptation approaches have been explored to increase intelligibility, as in Langner and Black (2005) and Raitio et al. (2011). Fourth, certain techniques aim to study the impact of particular acoustic