Available online at www.sciencedirect.com

SciVerse ScienceDirect

Computer Speech and Language 27 (2013) 288–300

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Analysis of the visual Lombard effect and automatic recognition experiments☆

Panikos Heracleous[a,*], Carlos T. Ishi[a], Miki Sato[a], Hiroshi Ishiguro[b], Norihiro Hagita[a]

[a] *ATR, Intelligent Robotics and Communication Laboratories, 2-2-2 Hikaridai Seika-cho, Soraku-gun, Kyoto-fu 619-0288, Japan*
[b] *ATR, Hiroshi Ishiguro Laboratory, 2-2-2 Hikaridai Seika-cho, Soraku-gun, Kyoto-fu 619-0288, Japan*

Received 4 April 2011; received in revised form 16 April 2012; accepted 11 June 2012
Available online 19 June 2012

Abstract

This study focuses on automatic visual speech recognition in the presence of noise. The authors show that, when speech is produced in noisy environments, articulatory changes occur because of the Lombard effect; these changes are both audible and visible. The authors analyze the visual Lombard effect and its role in automatic visual- and audiovisual speech recognition. Experimental results using both English and Japanese data demonstrate the negative effect of the Lombard effect in the visual speech domain. Without considering this factor in designing a lip-reading system, the performance of the system decreases. This is very important in audiovisual speech automatic recognition in real noisy environments. In such a case, however, the recognition rates decrease because of the presence of acoustic noise and because of the Lombard effect. The authors also show that the performance of an audiovisual speech recognizer depends also on the visual Lombard effect and can be further improved when it is considered in designing such a system.

Keywords: Lip-reading; Automatic speech recognition; Hidden Markov models (HMMs); Fusion; Noise robustness

1. Introduction

Speech is bi-modal in nature and includes the audio and visual modalities. Speech can be perceived using not only audio information but also information provided by the mouth/face movements. Automatic visual speech recognition (i.e., automatic lip-reading) attempts to automatically recognize speech provided by the mouth/lips. However, since many sounds look similar on the mouth/lips (i.e., visemes), speech cannot be totally recognized using visual information alone. In fact, automatic visual speech recognition has applications in audiovisual speech recognition, whereas visual information is used as a complement to audio information to increase the robustness against the noise.

In noisy environments, the talker increases the intelligibility of his/her speech (Lombard, 1911), and, during this process, several characteristics of speech change (the Lombard effect) (Bond and Moore, 1990; Castellanos and Casacuberta, 1996). As a result, the performance of an automatic speech recognizer operating in a noisy environment

☆ This paper has been recommended for acceptance by Philip Jackson, Ph.D.
* Corresponding author.
E-mail address: panikos@atr.jp (P. Heracleous).

0885-2308/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.csl.2012.06.003

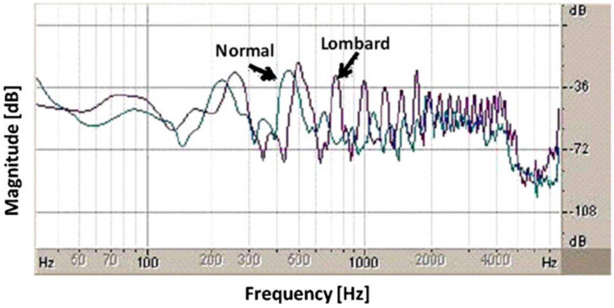

Fig. 1. Power spectrum of a normal, clean and a Lombard word.

decreases not only because of the noise contamination but also because of these modifications (Junqua, 1993; Wakao et al., 1996; Hansen, 1996).

Previously, in Heracleous et al. (2007), the role of the Lombard effect in non-audible murmur (NAM) recognition using a NAM microphone was investigated. A NAM microphone is a special acoustic sensor that is attached behind the talker's ear and can capture very quietly uttered speech (i.e., non-audible murmur). The results showed that, although a NAM microphone is very robust against noise, the recognition performance of a NAM recognizer decreases in noisy environments because of the Lombard effect.

Although many studies have addressed the problem of the Lombard effect in audio-only automatic speech recognition, only a few studies have addressed this issue with reference to automatic visual speech recognition.

In Huang and Chen (2001), audiovisual speech recognition experiments using noisy and Lombard data were presented. In this study, it was also briefly mentioned that the Lombard effect is present not only in the audio channel but also in the visual channel, and a few results were also presented.

In Davis et al. (2006) and Garnier et al. (2006), the changes that occur in the visual correlates of speech articulation when speech is produced in noisy environments were considered. In these studies, results were presented showing visual differences in the lip/mouth sector when speech was produced in a noisy environment or when Lombard speech was used. However, in these studies, analysis and experimental results related to visual speech recognition were not reported.

In this study, the authors comprehensively analyzed the visual Lombard effect phenomenon with respect to automatic visual- and audiovisual speech recognition using real noisy data and Lombard data and showed significant progress compared to the previously limited studies. Specifically, several isolated word and continuous phoneme recognition experiments were conducted in both the Japanese and the English languages, using data from several speakers. In addition, two fusion methods were used to integrate the audio and the visual streams in the audiovisual recognition experiments. The authors also showed that, when designing an audiovisual speech recognition system, further improvements in the recognition rates can be achieved by considering the visual Lombard effect in the statistical model training.

2. Acoustic Lombard effect

When speech is produced in noisy environments, the speech production process is modified, by a set of apparently preconscious behaviors called Lombard effect. Specifically, due to the reduced auditory feedback, the talker attempts to increase the intelligibility of his/her speech. During this process, several characteristics of speech change. In particular, the intensity of speech increases, the fundamental frequency (F0) and formants shift, the durations of vowels increase, and the spectral tilt changes. Because of these modifications, the performance of a speech recognizer decreases.

The analysis of the Lombard effect is not trivial because it depends on the speaker, on the noise type, and on the noise level (Wakao et al., 1996; Chi and Oh, 1996). One way to investigate the effect of the Lombard effect is to analyze clean speech uttered while the speaker is listening to noise through headphones or earphones (i.e., Lombard speech). Even though Lombard speech does not contain any noise components, modifications in speech characteristics can be realized.