ELSEVIER
Available online at www.sciencedirect.com
ScienceDirect
Computer Speech & Language xxx (2017) xxx-xxx
www.elsevier.com/locate/csl

Audio-visual word prominence detection from clean and noisy speech ☆

Martin Heckmann

Honda Research Institute Europe GmbH, D-63073 Offenbach am Main, Germany

Received 22 December 2015; received in revised form 8 May 2017; accepted 28 September 2017
Available online xxx

Abstract

In this paper we investigate the audio-visual processing of linguistic prosody, more precisely the detection of word prominence, and examine how the additional visual information can be used to increase the robustness when acoustic background noise is present. We evaluate the detection performance for each modality individually and perform experiments using feature and decision fusion. For the latter we also consider the adaptive fusion with fusion weights adjusted to the current acoustic noise level. Our experiments are based on a corpus with 11 English speakers which contains in addition to the speech signal also videos of the speakers' heads. From the acoustic signal we extract features which are well known to capture word prominence like loudness, fundamental frequency and durational features. The analysis of the visual signal is based on features derived from the speaker's rigid head movements and movements of the speaker's mouth. We capture the rigid head movements by tracking the speaker's nose. Via a two-dimensional Discrete Cosine Transform (DCT) calculated from the mouth region we represent the movements of the speaker's mouth. The results show that the rigid head movements as well as movements inside the mouth region can be used to discriminate prominent from non-prominent words. The audio-only detection yields an Equal Error Rate (EER) averaged over all speakers of 13%. Based only on the visual features we obtain 20% of EER. When we combine the visual and the acoustic features we only see a small improvement compared to the audio-only detection for clean speech. To simulate background noise we added 4 different noise types at varying SNR levels to the acoustic stream. The results indicate that word prominence detection is quite robust against additional background noise. Even at a severe Signal to Noise Ratio (SNR) of −10 dB the EER only rises to 35%. Despite this the audio-visual fusion leads to marked improvements for the detection from noisy speech. We observe relative reductions of the EER of up to 79%.

1. Introduction

When humans communicate they not only listen to what is said but also to how it is said. These prosodic variations play a vital role in human communication (Shriberg, 2005). For spoken dialog systems one situation where the prosodic information is particularly important is after a misunderstanding between the human and the machine (Litman et al., 2006; Levow, 2004). Humans use prosodic cues to highlight a correction following a misunderstanding when talking to another human but also when talking to a machine (Swerts et al., 2000). A distinguishing feature of corrections is that they are frequently hyperarticulated and hence perceived as highly prominent (Litman et al.,

☆ This paper has been recommended for acceptance by Prof. R. K. Moore.

E-mail address: martin.heckmann@honda-ri.de

http://dx.doi.org/10.1016/j.csl.2017.09.002
0885-2308/ 2017 Published by Elsevier Ltd.

2006). Acoustically prominence is mainly realized via changes in segment duration and intensity as well as fundamental frequency modifications (Streefkerk, 2002).

At least since the observations of Sumby and Pollack in 1954 that seeing a speaker's face and lip movements improves human speech intelligibility scores in noise (Sumby and Pollack, 1954) it is well acknowledged that the visual modality contains a lot of information relevant to communication. As a result, the field of audio-visual speech recognition emerged which in particular targets the robust recognition in noise (Potamianos et al., 2003; Zhou et al., 2014; Heckmann et al., 2002; Kolossa et al., 2009).

Furthermore, it has been shown that humans are able to use visual information to extract linguistic prosodic cues and in particular word prominence (Graf et al., 2002; Munhall et al., 2004; Beskow et al., 2006; Swerts and Krahmer, 2008; Al Moubayed and Beskow, 2009). Yet so far this information has not been used in systems to extract linguistic prosody. Based on the findings in audio-visual speech recognition one expects that the visual information is particularly beneficial in situations when the acoustic signal is impaired. Such impairments by additional background noise and reverberations from the room commonly occur when speech processing is applied in real world applications, in particular on hand held devices, with robots and in cars. Due to the typical large distances between the microphones and the speaker's mouth, the impact of these disturbances can be quite high. In the realm of speech recognition this is a topic which has received a lot of attention (see Li et al. (2014) for a current overview). The aforementioned speech impairments in principle also apply to the prosodic speech analysis. Yet much less effort has been spent to cope with these impairments in this context. As far as we are aware of, the impact of speech impairments and viable countermeasures have only be addressed for audio-only emotion classification (Schuller et al., 2007; Eyben et al., 2013; You et al., 2006).

We introduced audio-visual word prominence detection in Heckmann (2012) on a dataset of 3 speakers. In Heckmann (2013) we extended the dataset to 16 speakers and improved the acoustic feature extraction. Next, we included context features, i.e. features spanning across the current segment, and feature contour modeling in Schnall and Heckmann (2014) and Heckmann (2014). In this paper we use the same acoustic feature extraction as in our previous work but introduce an improved visual processing including a correction of the speaker's head tilt. Based on this we evaluated the performance of rigid head movements and movements inside the mouth region compared to the acoustic features. This evaluation showed that there is a large variation in visual detection performance for the different speakers. Nevertheless, the visual features contribute overall a lot of information. Furthermore, we investigated different audio-visual fusion schemes. Here we saw that a fusion on the decision level clearly outperforms a feature fusion. A key aspect of this paper is the evaluation of the prominence detection also from noisy audio signals. We performed this evaluation for different types of noise added to the speech signal at a wide range of Signal to Noise Ratio (SNR) levels. This evaluation also included an evaluation of the importance of two of the main cues to word prominence, fundamental frequency and intensity variations, in varying noise conditions. The results showed that the word prominence detection is quite robust against background noise and that the relative importance of the fundamental frequency and intensity derived features depends on the noise type. A further important part of this evaluation is the assessment of the audio-visual fusion schemes with varying degrees of degradations in the acoustic channel. We will demonstrate that the decision fusion is also superior when the acoustic signal is degraded by noise and that improvements of 79 % compared to an audio-only detection can be obtained from an audio-visual fusion. However, adaptively weighting the two modalities dependent on the acoustic noise level does not further improve the results.

2. Prior work

Quite a few approaches to detect prosodic word prominence have been proposed in the past (Streefkerk, 2002; Tamburini, 2003; Wang and Narayanan, 2007; Jeon and Liu, 2010; Schillingmann et al., 2011). Some authors rather focus on the detection of pitch accent, one acoustic realization leading to prosodic prominence (Shriberg and Stolcke, 2004; Levow, 2005; Rosenberg and Hirschberg, 2009). All these approaches have in common that they only consider the acoustic modality and assume that the signal is not distorted by noise.

The fusion of acoustic and visual information has received a lot of attention in the past (Atrey et al., 2010). A key question in this domain was to find models to optimally fuse the acoustic and visual modalities (Teissier et al., 1999; Potamianos et al., 2003; Yoshida et al., 2009). Particularly relevant to audio-visual speech recognition are models which are able to cope with the complex temporal relation between auditory and visual cues in speech. Depending

Please cite this article as: M. Heckmann, Audio-visual word prominence detection from clean and noisy speech, Computer Speech & Language (2017), http://dx.doi.org/10.1016/j.csl.2017.09.002