Available online at www.sciencedirect.com

ScienceDirect

Computer Speech and Language 36 (2016) 72–92

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

# Detecting paralinguistic events in audio stream using context in features and probabilistic decisions[☆]

Rahul Gupta[a,*], Kartik Audhkhasi[b], Sungbok Lee[a], Shrikanth Narayanan[a]

[a] *Signal Analysis and Interpretation Laboratory (SAIL), University of Southern California, 3710 McClintock Avenue, Los Angeles, CA 90089, USA*

[b] *IBM Thomas J Watson Research Center, 1101 Kitchawan Rd, Yorktown Heights, NY 10598, USA*

Received 28 November 2014; received in revised form 24 June 2015; accepted 7 August 2015
Available online 11 September 2015

## Abstract

Non-verbal communication involves encoding, transmission and decoding of non-lexical cues and is realized using vocal (e.g. prosody) or visual (e.g. gaze, body language) channels during conversation. These cues perform the function of maintaining conversational flow, expressing emotions, and marking personality and interpersonal attitude. In particular, non-verbal cues in speech such as paralanguage and non-verbal vocal events (e.g. laughters, sighs, cries) are used to nuance meaning and convey emotions, mood and attitude. For instance, laughters are associated with affective expressions while fillers (e.g. um, ah, um) are used to hold floor during a conversation. In this paper we present an automatic non-verbal vocal events detection system focusing on the detect of laughter and fillers. We extend our system presented during Interspeech 2013 Social Signals Sub-challenge (that was the winning entry in the challenge) for frame-wise event detection and test several schemes for incorporating local context during detection. Specifically, we incorporate context at two separate levels in our system: (i) the raw frame-wise features and, (ii) the output decisions. Furthermore, our system processes the output probabilities based on a few heuristic rules in order to reduce erroneous frame-based predictions. Our overall system achieves an Area Under the Receiver Operating Characteristics curve of 95.3% for detecting laughters and 90.4% for fillers on the test set drawn from the data specifications of the Interspeech 2013 Social Signals Sub-challenge. We perform further analysis to understand the interrelation between the features and obtained results. Specifically, we conduct a feature sensitivity analysis and correlate it with each feature's stand alone performance. The observations suggest that the trained system is more sensitive to a feature carrying higher discriminability with implications towards a better system design.

*Keywords:* Paralinguistic event; Laughter; Filler; Probability smoothing; Probability masking

## 1. Introduction

Non-verbal communication involves sending and receiving non-lexical cues amongst people. Modalities for transmitting non-verbal cues include body language, eye gaze and non-verbal vocalizations. Non-verbal communication is hypothesized to represent two-thirds of all communication (Hogan and Stubbs, 2003) and its primary functions

[☆] This paper has been recommended for acceptance by Björn Wolfgang Schuller.

[*] Corresponding author. Tel.: +1 2135871103.
*E-mail address:* guptarah@usc.edu (R. Gupta).

http://dx.doi.org/10.1016/j.csl.2015.08.003

Table 1
Statistics of laughter and filler annotations in the SVC corpus.

| Event | Total number of segments | Statistics over the segment lengths (in milliseconds) | | |
|---|---|---|---|---|
| | | Mean | Standard deviation | Range |
| Laughter | 1158 | 943 | 703 | 2–5080 |
| Filler | 2988 | 502 | 262 | 1–5570 |

include reflecting attitude and emotions (Argyle et al., 1970; Mehrabian and Ferris, 1967; Halberstadt, 1986), assisting dialog process (Bavelas and Chovil, in press; Johannesen, 1971) as well as expressing personality (Isbister and Nass, 2000; Cunningham, 1977). Studies suggest that non-verbal communication is a complex encoding-decoding process (Zuckerman et al., 1975; Lanzetta and Kleck, 1970). Encoding relates to the generation of non-verbal cues, usually in parallel with verbal communication and decoding involves interpretation of these cues (Argyle, 1972; O'sullivan et al., 1994). Studies broadly classify non-verbal communication into two categories, visual and vocal (Streeck and Knapp, 1992; Poyatos, 1992). Visual cues include communication through body gestures, touch and body distance (Ruesch and Kees, 1956) and vocal cues comprise paralanguage (e.g., voice quality, loudness) and non-verbal vocalizations (e.g., laughters, sighs, fillers) (Schuller et al., 2008; Bowers et al., 1993). Both these channels of non-verbal communication have been extensively studied and the literature suggests their relationship to varied phenomena and constructs including language development (Harris et al., 1986), child growth (Mundy et al., 1986; Curcio, 1978), relationship satisfaction (Kahn, 1970; Boland and Follingstad, 1987) and psychotherapy process (Gupta et al., 2014). This extension of non-verbal communications research beyond understanding their primary functions reflects their significance in interaction.

Our focus in this work is on non-verbal vocalizations (NVVs) in speech. Previous research links various forms of non-verbal vocalizations such as laughters, sighs and cries to emotion (Goodwin et al., 2009; Gupta et al., 2012), relief (Soltysik and Jelen, 2005; Vlemincx et al., 2010) and evolution (Furlow, 1997). The importance of each of these non-verbal vocalizations is highlighted by the role they play in human expression. Therefore a quantitative understanding of their production and perception can have a significant impact on both behavioral analysis and behavioral technology development. In this paper, we aim to contribute to the analysis of these non-verbal vocalizations by developing a system for detection of non-verbal events in spontaneous speech.

Several previous works have proposed detection methods for NVVs. Kennedy and Ellis (2004) demonstrated the efficacy of using window-wise low level descriptors from speech (Cortes and Vapnik, 1995) in detecting laughters in meetings. Truong and Van Leeuwen (2005) investigated perceptual linear prediction (PLP) and acoustic prosodic features for NVV detection using Gaussian mixture models. Várallyay et al. (2004) performed acoustic analysis of infant cries for early detection of hearing disorders. Schuller et al. (2008) presented static and dynamic modeling approach for recognition of non-verbal events such as breathing and laughter in conversational speech. In particular, the Interspeech 2013 Social Signals Sub-challenge (Schuller et al., 2013) led to several investigations (Kaya et al., 2013; Pammi and Chetouani, 2013; Krikke and Truong, 2013; Brueckner and Schulter, 2014; An et al., 2013) on frame-wise detection of two specific non-verbal events: laughters and fillers. Building upon on our efforts (Gupta et al., 2013) on the same challenge dataset (Salamin et al., 2013) (that was the winning entry in the challenge), in this paper we perform further analysis and experiments. Previous works in this research field have primarily focused on local characteristics and our approach investigates the benefits of considering context during the frame-wise prediction. Our methods are inspired from the fact that the non-verbal events occur over longer segments (and hence analysis frames). The temporal characteristics of these events has been investigated in studies such as (Mowrer et al., 1987; Bachorowski et al., 2001; Candea et al., 2005). These studies reveal interesting patterns such as a positive correlation between duration of laughter (Mowrer et al., 1987) and number of intensity peaks and similarity in duration of fillers across languages (Candea et al., 2005). Bachorowski et al. (2001) went further into the details of laughter types (e.g. voiced vs unvoiced) and their relation to laughter durations. More studies on laughter and filler duration and its relation to their acoustic structures can be found in (Vettin and Todt, 2004; Sundaram and Narayanan, 2007; Vasilescu et al., 2005). As statistics (presented later in Table 1) on our database of interest also show that laughters and fillers exist over multiple analysis frames, we hypothesize that information from neighboring frames can be utilized to reduce the uncertainty associated with the current frame.