Available online at www.sciencedirect.com

SciVerse ScienceDirect

Computer Speech and Language 28 (2014) 314–325

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Classification of social laughter in natural conversational speech☆

Hiroki Tanaka [a,*], Nick Campbell [a,b]

[a] *Augmented Human Communication Laboratory, Nara Institute of Science and Technology, Takayama-cho 8916-5, Ikoma-shi, Nara, Japan*
[b] *Speech Communication Lab, CLCS, Trinity College Dublin, College Green, Dublin, Ireland*

Received 13 August 2012; received in revised form 9 July 2013; accepted 23 July 2013
Available online 5 August 2013

Abstract

We report progress towards developing a sensor module that categorizes types of laughter for application in dialogue systems or social-skills training situations. The module will also function as a component to measure discourse engagement in natural conversational speech. This paper presents the results of an analysis into the sounds of human laughter in a very large corpus of naturally occurring conversational speech and our classification of the laughter types according to social function. Various types of laughter were categorized into either polite or genuinely mirthful categories and the analysis of these laughs forms the core of this report. Statistical analysis of the acoustic features of each laugh was performed and a Principal Component Analysis and Classification Tree analysis were performed to determine the main contributing factors in each case. A statistical model was then trained using a Support Vector Machine to predict the most likely category for each laugh in both speaker-specific and speaker-independent manner. Better than 70% accuracy was obtained in automatic classification tests.

Keywords: Laughter; Prosody; Paralinguistic information; Non-verbal behaviour; Classification; Support Vector Machines

1. Introduction

In human–human interaction, communication involves both verbal and nonverbal information, and the latter serves especially to express discourse engagement. One of the most common nonverbal vocalizations in social conversation is laughter (Petridis, 2011) which is also reported as the most frequently annotated acoustic nonverbal behavior in meeting corpora (Laskowski and Burger, 2007) where 8.6% of the time a person vocalizes in a meeting is spent on laughing and 0.8% is spent on laughing while talking. Laughter is a universal and prominent feature of human communication (Jung, 2003), and expressed by both vocal and facial expressions. It is a powerful affective and social signal (Vinciarellia et al., 2009). There is no culture where laughter is not found. However, current dialogue systems and computer-based social skills training (a training method for people with autism or Asperger syndrome to learn social function (Ozonoff and Miller, 1995)) do not take into account laughter (Golan and Baron-Cohen, 2006).

In a seminal study of the segmentation of laughs, Trouvain and Schroder (2004) suggest that we consider laughter as articulated speech, where at the low level there are sound segments that are either vowels or consonants. At the next higher level, there are syllables consisting of sound segments. The next higher level deals with larger units such as

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

* Corresponding author. Tel.: +81 743 72 5265; fax: +81 743 72 5269.
E-mail address: hiroki-tan@is.naist.jp (H. Tanaka).

0885-2308/$ – see front matter.
http://dx.doi.org/10.1016/j.csl.2013.07.004

phrases which are made up of several syllables. Owren and Understanding (2007) recommend the term 'bout' for the longer sequence, and 'call' for the individual syllables; we will adopt that terminology in this study.

Some earlier work on the automatic segmentation of laughter has been reported in the literature. Truong et al. (2007) reported automatic laughter segmentation in meetings. They performed laughter vs. speech discrimination experiments comparing traditional spectral features and acoustic phonetic features, and concluded that the performance of laughter segmentation can be improved by incorporating phonetic knowledge into the models. Scherer et al. (2012) reported that the total accuracy of detecting laughter from natural discourse in human–computer interaction reached over 90% in online and offline detection experiments with speech and visual information. Kennedy and Ellis (2004) focused on joint laughter in meetings, which means participants (more than just one) laugh simultaneously (Glenn, 1991; Jefferson, 1979; Kangasharju and Nikkot, 2009), and they obtained detection results with a correct accept rate of 87% and a false alarm rate of 13% by using Support Vector Machines.

Types of laughter vary in natural conversational speech, and some classifications have been reported in the literature regarding different categories of laughter. Most types of laughter were discussed in Shimizu et al. (1994), and the major work is the discrimination of laughter into two types, voiced and unvoiced, based on acoustics (Bachorowski and Owren, 2001; Hudenko et al., 2009). Laurence and Laurence (2007) deal with a study of laughs in spontaneous speech and explore the positive and negative valence of laughter towards their global aim of detecting emotional behavior in speech. The conclusion of their acoustic analysis is that unvoiced laughs are more often perceived as negative and voiced segments as positive. Previous work in the literature has also discussed whether laughter patterns can be defined through stereotypes (Bachorowski et al., 2001; Trouvain and Schroder, 2004; Sundaramb and Narayananc, 2007). However, laughter is not simply positive or negative, or even defined by stereotypes; it is quite usual for people to infer different degrees of emotion and engagement based on its perceptions, and it is common for people to make use of social laughter in sophisticated social interaction. In this study we tested perceptual types of laughter to determine the main characteristics of laughter in social interaction by reference to the above previous studies.

Automatic classification of four phonetic types of laughter in a natural-speech conversation corpus was conducted by Campbell et al. (2005), based on perceptual impressions of laughter, in which a laughter episode is considered as a sequence of speech-like phonetic segments (after Bachorowski et al., 2001). The work described 4 different laughter types: voiced, chuckle, breathy and nasal, and modeled each laugh as composed of different combinations of these segments using Hidden Markov Models (HMMs) statistical classification. The study reported an automatic discrimination using 3–15 states with mfcc-based HMMs for 4 functions of laughter (hearty, amused, satirical, and polite). In categorizing emotional classification the work achieved 76% accuracy. However because of the hidden nature of the statistical modeling the report did not provide explicit details about which specific acoustic features contributed to the various categorizations of the laughter.

We report progress towards developing a sensor module that categorizes types of laughter for application in dialogue systems or social skills training situations. In the present study we only make use of the audio information but recognize that facial expression also carries an important channel of communicative information (Carroll and Russel, 1996; De Gelder and Vroomen, 2000). This paper reports a study of laughs in a corpus of human–human dialogues recorded from Japanese telephone conversational speech (The Expressive Speech, 2013). We employed a corpus of natural spontaneous speech where laughter occurred naturally as a consequence of the dialogue interaction. We specifically avoid the use of contrived laughter or even specifically elicited laughs since they may not be representative of natural spontaneous interaction.

In the following sections we first provide details of the corpus, then introduce two Experiments. Experiment 1: a perceptual test by Japanese students to determine the number and types of easily discriminated laughter, and Experiment 2: describing the acoustic feature extraction, presenting the results of an analysis of the main acoustic features and finally reporting a classification of type of laughter using statistical methods.

2. Data: natural types of laughter

We used two types of Japanese corpora. First, the Expressive Speech Processing (ESP) corpus (The Expressive Speech, 2013) was used for this study. The speech data were recorded over a period of several months, with paid volunteers coming to an office building in a large city in Western Japan once a week to talk with specific partners in a separate part of the same building over an office telephone. While talking, they each wore a head-mounted Sennheiser HMD-410 close-talking dynamic microphone and recorded their speech directly to DAT (digital audio