Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 49 (2007) 144–158

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Automatic discrimination between laughter and speech

Khiet P. Truong *, David A. van Leeuwen

TNO Human Factors, Department of Human Interfaces, P.O. Box 23, 3769 ZG Soesterberg, The Netherlands

Received 18 October 2005; received in revised form 17 November 2006; accepted 4 January 2007

Abstract

Emotions can be recognized by audible paralinguistic cues in speech. By detecting these paralinguistic cues that can consist of laughter, a trembling voice, coughs, changes in the intonation contour etc., information about the speaker's state and emotion can be revealed. This paper describes the development of a gender-independent laugh detector with the aim to enable automatic emotion recognition. Different types of features (spectral, prosodic) for laughter detection were investigated using different classification techniques (Gaussian Mixture Models, Support Vector Machines, Multi Layer Perceptron) often used in language and speaker recognition. Classification experiments were carried out with short pre-segmented speech and laughter segments extracted from the ICSI Meeting Recorder Corpus (with a mean duration of approximately 2 s). Equal error rates of around 3% were obtained when tested on speaker-independent speech data. We found that a fusion between classifiers based on Gaussian Mixture Models and classifiers based on Support Vector Machines increases discriminative power. We also found that a fusion between classifiers that use spectral features and classifiers that use prosodic information usually increases the performance for discrimination between laughter and speech. Our acoustic measurements showed differences between laughter and speech in mean pitch and in the ratio of the durations of unvoiced to voiced portions, which indicate that these prosodic features are indeed useful for discrimination between laughter and speech.

Keywords: Automatic detection laughter; Automatic detection emotion

1. Introduction

Researchers have become more and more interested in automatic recognition of human emotion. Nowadays, different types of useful applications employ emotion recognition for various purposes. For instance, knowing the speaker's emotional state can contribute to the naturalness of human–machine communication in spoken dialogue systems. It can be useful for an automated Interactive Voice Response (IVR) system to recognize impatient, angry or frustrated customers who require a more appropriate dialogue handling and to route them to human operators if necessary (see e.g., Yacoub et al., 2003). In retrieval applications, automatic detection of emotional acoustic events can be used to segment video material and to browse through video recordings, e.g., Cai et al. (2003) developed a 'hotspotter' that automatically localizes applause and cheer events to enable video summarization. Furthermore, a meeting browser that also provides information on the emotional state of the speaker was developed by Bett et al. (2000). Note that the word 'emotion' is a rather vague term susceptible to discussion. Often the terms 'expressive' or 'affective' are also used to refer to emotional speech. We will continue using the term 'emotion' in its broad sense.

The speaker's emotional and physical state expresses itself in speech through paralinguistic features such as pitch, speaking rate, voice quality, energy etc. In the literature, pitch is indicated as being one of the most relevant paralinguistic features for the detection of emotion, followed by energy, duration and speaking rate (see Bosch ten, 2003). In general, speech shows an increased pitch variability or range and an increased intensity of effort when people are in a heightened aroused emotional state (Williams and Stevens, 1972; Scherer, 1982; Rothganger et al., 1998; Mowrer et al., 1987). In a paper by Nwe

* Corresponding author. Tel.: +31 346 356 339; fax: +31 346 353 977.
E-mail address: khiet.truong@tno.nl (K.P. Truong).

0167-6393/$ - see front matter © 2007 Published by Elsevier B.V.
doi:10.1016/j.specom.2007.01.001

et al. (2003), an overview of paralinguistic characteristics of more specific emotions is given. Thus, it is generally known that paralinguistic information plays a key role in emotion recognition in speech. In this research, we concentrate on audible, identifiable paralinguistic cues in the audio signal that are characteristic for a particular emotional state or mood. For instance, a person who speaks with a trembling voice is probably nervous and a person who is laughing is most probably in a positive mood (but bear in mind that other moods are also possible). We will refer to such identifiable paralinguistic cues in speech as ''paralinguistic events.'' Our goal is to detect these ''paralinguistic events'' in speech with the aim to make classification of the speaker's emotional state or mood possible.

2. Focus on automatic laughter detection

We have decided first to concentrate on laughter detection, due to the facts that laughter is one of the most frequently annotated ''paralinguistic events'' in recorded natural speech databases, it occurs relatively frequently in conversational, spontaneous speech and it is an emotional outburst and acoustic event that is easily identified by humans. Laughter detection can be meaningful in many ways. The main purpose of laughter detection in this research is to use laughter as an important cue to the identification of the emotional state of the speaker(s). Furthermore, detecting laughter in, e.g., meetings can provide cues to semantically meaningful events such as topic changes. The results of this research can also be used to increase the robustness of non-speech detection in automatic speech recognition. And finally, the techniques used in this study for discrimination between laughter and speech can also be used for similar discrimination tasks between other speech/non-speech sounds such as speech/music discrimination (see e.g., Carey et al., 1999).

Several studies have investigated the acoustic characteristics of laughter (e.g., Bachorowski et al., 2001; Trouvain, 2003; Bickley and Hunnicutt, 1992; Rothganger et al., 1998; Nwokah et al., 1993) and compared these characteristics to speech. Of these studies, the study by Bachorowski et al. (2001) is probably the most extensive one using 97 speakers who produce laugh sounds, while the other studies mentioned here use 2–40 speakers. Although studies by Bachorowski et al. (2001) and Rothganger et al. (1998) conclude that F_0 is much higher in laughter than in speech and that speech is rather monotonic, lacking a strongly varying melodic contour that is present in laughter, there are other studies that report on mean F_0 measures of laughter that are rather speech-like (Bickley and Hunnicutt, 1992). There are also mixed findings on intensity measures of laughter: while Rothganger et al. (1998) report on higher intensity values for laughter that even resemble screaming sounds, Bickley and Hunnicutt (1992) did not find large differences in amplitude between laughter and speech. Researchers did agree on the fact that the measures were strongly influenced by the gender of the speaker (Bachorowski et al., 2001; Rothganger et al., 1998) and that laughter is a highly complex vocal signal, notable for its acoustic variability (Bachorowski et al., 2001; Trouvain, 2003). Although there exists high acoustic variability in laughter, both between and within speakers, Bachorowski et al. (2001) noted that some cues of the individual identity of the laughing person are conveyed in laughter acoustics (i.e., speaker dependent cues). Furthermore, culture specific laughs may also exist: although no significant differences were found between laughter from Italian and German students (Rothganger et al., 1998), laughter transcriptions by Campbell et al. (2005) show that Japanese laughter can be somewhat different from the more typical ''haha'' laughs that are commonly produced in Western culture. A similarity between laughter and speech was found by Bickley and Hunnicutt (1992): according to their research, the average number of laugh syllables per second is similar to syllable rates found for read sentences in English. However, they (Bickley and Hunnicutt, 1992) also identified an important difference between laughter and speech in the durations of the voiced portions: a typical laugh reveals an alternating voiced-unvoiced pattern in which the ratio of the durations of unvoiced to voiced portions is greater for laughter than for speech. This is one of the features that can be used for the development of a laughter detector.

Automatically separating laughter from speech is not as straightforward as one may think since both sounds are created by the vocal tract and therefore share characteristics. For example, laughter usually consists of vowel-like laugh syllables that can be easily mistaken for speech syllables by an automatic speech recognizer. Additionally, there are different vocal-production modes that produce different types of laughter (e.g., voiced, unvoiced) which causes laughter to be a very variable and complex signal. Furthermore, laughter events are typically short acoustic events of approximately 2 s (according to our selected laughter segments taken from the ICSI database, see Section 4.1). Several researchers have already focused on automatic laughter detection; usually these studies employed spectral/cepstral features to train their models. Cai et al. (2003) tried to locate laughter events in entertainment and sports videos: they modeled laughter with Hidden Markov Models (HMM) using Mel-Frequency Cepstral Coefficients (MFCCs) and perceptual features such as short-time energy and zero crossing rate. They achieved average recall and precision percentages of 92.95% and 86.88% respectively. In the LAFCam project (Lockerd and Mueller, 2002), a system was developed for recording and editing home videos. The system included laughter detection using Hidden Markov Models trained with spectral coefficients. They classified presegmented laughter and speech segments correctly in 88% of the test segments. For automatic segmentation and classification of laughter, the system identified segments as laughter correctly 65% of the time. Kennedy and Ellis (2004) developed their laugh detector by training a Support Vector Machine (SVM) with Mel-Frequency Cepstral Coefficients, their deltas, spatial