Available online at www.sciencedirect.com
ScienceDirect
Speech Communication xxx (2014) xxx–xxx
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

Medium term speaker state detection by perceptually masked spectral features

Cenk Sezgin [a,*], Bilge Gunsel [a], Jarek Krajewski [b,c]

[a] Multimedia Signal Processing and Pattern Recognition Group, Istanbul Technical Univ., Turkey
[b] Experimental Industrial Psychology, Univ. of Wuppertal, Germany
[c] Industrial Psychology, Rhenish Univ. of Applied Sciences Cologne, Germany

Received 11 June 2013; received in revised form 12 August 2014; accepted 9 September 2014

Abstract

We propose a method based on perceptual prosodic features for medium term speaker state classification, particularly sleepiness detection. Unlike existing methods, our features represent spectral characteristics of speech in perceptual bands and also track temporal content omitting any linguistic segmentation. Despite conventional methods, we aim to model transitions between non-sleepy and sleepy modes rather than emotional states. Along with the proposed compact feature set, the developed system enable discrimination of medium term speaker states with a lower complexity compared to existing systems. This is achieved by constructing a dictionary for speech data based on bag-of-words concept. It has been identified that a training setup which is based on learned codewords, yields a robust classifier for sleepy speech. The speaker state classification has been performed by applying a two-class classification scheme on the observed test data. The numerical results, obtained on the Sleepy Language Corpus (SLC) by using Support Vector Machines (SVM) classifier, demonstrate a 10% improvement on average on unweighted recall rates compared to the benchmarking results. The introduced method is promising for online applications because of its frame based feature extraction scheme which differs from conventional segmental descriptor extraction techniques.

Keywords: Sleepiness detection; Speaker emotion recognition; Perceptual audio features; Bag-of-words; Medium term speaker state

1. Introduction

Sleepiness is an important medium term quasi-emotional state which affects safety, performance, comfort, and joy-of-use in many fields of human–machine interaction (HMI). Therefore, warning drivers against impending critical sleepiness plays an important role in preventing accidents, which induce human and financial costs. Moreover, detecting sleepiness can enhance comprehensiveness and comfort of HMI if the system output is adapted to the user's actual sleepiness-impaired attentional and cognitive resources. Furthermore, reacting to users' sleepiness state contributes to a more human-like, empathic communication, enhancing naturalness and acceptance of HMI (Picard, 1995; Tao and Tan, 2005).

There are notable efforts in literature for defining sleepiness measures. These approaches have focused mainly on measures of physiological criteria; pupil size, eye blinking, heart rate, EEG, behavioral expressions, tracking tasks, gross body movement to characterize sleepiness state (Kaida et al., 2006). However, there are challenges in using alternative criteria such as vocal expression and acoustic analysis in sleepiness detection, since these criteria require

* Corresponding author.
E-mail addresses: csezgin@itu.edu.tr (C. Sezgin), gunselb@itu.edu.tr (B. Gunsel), krajewsk@uni-wuppertal.de (J. Krajewski).

http://dx.doi.org/10.1016/j.specom.2014.09.002
0167-6393/

simplification of measurement system setup and more robustness against environmental conditions (Krajewski et al., 2008). Another difficulty in detecting medium term states, such as sleepiness, stems from the fact that the required monitoring time is longer in comparison with short-term speaker emotional states.

Some of the related work focuses on feature extraction to model impact of sleepiness on acoustic voice characteristics, while others propose novel classification schemes to improve detection performance. Nwe et al. evaluated pitch and harmonic patterns of speech to analyze flatness of voice on the DCIEM Map Task Corpus (Nwe et al., 2006; Bard et al., 1996) using statistical modeling based on Hidden Markov Models (HMM). This study revealed that sleepy speech has less variation on pitch and harmonic pattern with regard to non-sleepy speech. Tao et al. proposed a general framework to model speech characteristics by prosody, articulation and speech quality related features (Tao and Tan, 2005). A total of 8,500 features per speech sample are calculated for detecting accident-prone fatigue state classification. The class-wised averaged classification rate achieved on a small database by the 1-nearest neighbor, SVM and multi-layer perceptron classifiers are reported as over 80%.

The work presented by Krajewski, in Bard et al. (1996), examines the effect of microsleep endangered sleepiness states on acoustic voice characteristics. A total of 45,088 features are calculated per speech window where the speech samples are generated by a car simulator emulating sleep deprivation. The highest detection rate is reported as 85.1% for SVM by a reduced dimensional 130-D feature set.

In (Krajewski et al., 2010), nonlinear dynamic (NLD) features are proposed in order to improve prediction of fatigue from speech. The NLD features consist of 375 state space features that capture temporal information, 110 fractal features that quantify self-affinity and 5 entropy features that measure regularity of speech signal fluctuations. It is shown that the NLD features provide additional information regarding dynamics and structure of sleepy speech compared to commonly applied speech emotion recognition features as the Stanford Sleepiness Scale (SSS) is used.

In most recent studies, openSMILE emotional feature extractor has been adapted to the sleepiness detection problem. openSMILE is a generic short time emotional state detection tool which extracts more than 6,552 features by 39 functionals of 56 acoustic low-level descriptors (Eyben et al., 2009). The sleepiness sub-challenge in INTERSPEECH 2011 addressed the sleepiness classification problem from speech by using openSMILE features (Eyben et al., 2009; Schuller et al., 2011). Test results have been reported on the Sleepy Language Corpus (SLC) Krajewski and Kröger, 2007 featuring 21 hours of speech recordings of 99 subjects given in 10 different levels on the Karolinska Sleepiness Scale (KSS) Kaida et al., 2006. The KSS is a common, well-established and standardized subjective sleepiness questionnaire measure. In this work we also use the KSS to evaluate the level of sleepiness state.

Kaida et al. and Krajewski et al. used an extended subset of openSMILE features; a total of 4,368 features including energy related, spectral and voice related low level descriptors and their statistical variants for sleepiness detection. The highest recognition rate achieved by SVM is reported to be 70.3% (Kaida et al., 2006; Krajewski and Kröger, 2007). The system proposed by the winner of the challenge provides 71.6% detection accuracy achieved by employing AdaBoost with SVM and the Asymmetric Simple Partial Least Squares (SIMPLS) classifiers (Huang et al., 2011).

In (Krajewski et al., 2012), an optimized feature set is specified by applying a correlation-filter subset selection on NLD and openSMILE features that yielded 565 descriptors, including 395 non-linear dynamics and 170 phonetic features. The performance has been reported on a subset of the SLC data that includes 372 utterances of 77 speakers. The highest recognition rates are respectively reported as 79.6% (Bayes Net) and 77.1% (AdaBoost Nearest Neighbor) for male and female speakers.

The aforementioned features in the related works establish a broad space with numerous and redundant features. This is mainly because the existing features are primarily proposed for speech recognition rather than emotion or sleepiness detection (Lugger and Yang, 2008; Yang and Lugger, 2010). Therefore, these features may not fully model sleepiness because a vast majority of them, such as MFCC, are generated for short speech frames to decode phonemes. Consequently, a high performance sleepiness detector could only be achieved by using very large feature sets (Krajewski et al., 2010; Eyben et al., 2009; Krajewski et al., 2012) or considerably small feature sets in combination with highly complex classifiers (Huang et al., 2011; Krajewski et al., 2012). Recently conducted studies show that deep neural networks (DNNs) can effectively generate discriminative features that approximate complex nonlinear dependencies between features, and therefore improve recognition performance (Fousek et al., 2013; Heigold et al., 2013). Stuhlsatz et al. proposed Generalized Discriminant Analysis (GerDA) based on DNNs for learning low dimensional discriminative features from a large set (Stuhlsatz et al., 2011). This approach allows a fast and simple linear classification. Currently, DNNs have a disadvantage compared with GMMs that requires training on massive data sets in favor of making good use of large cluster machines. This issue may be offset by tuning DNNs more efficiently, so they do not require as much data to achieve the same performance. However, finding alternatives for parallelizing and fine-tuning DNNs is still a major concern (Fousek et al., 2013).

Another important issue is that these methods perform segmental feature extraction either using linguistic models or preprocessing schemes in order to handle medium term

Please cite this article in press as: Sezgin, C. et al., Medium term speaker state detection by perceptually masked spectral features, Speech Comm. (2014), http://dx.doi.org/10.1016/j.specom.2014.09.002