Medium term speaker state detection by perceptually masked spectral features

Cenk Sezgin\textsuperscript{a,}*, Bilge Gunsel\textsuperscript{a}, Jarek Krajewski\textsuperscript{b,c}

\textsuperscript{a} Multimedia Signal Processing and Pattern Recognition Group, Istanbul Technical Univ., Turkey
\textsuperscript{b} Experimental Industrial Psychology, Univ. of Wuppertal, Germany
\textsuperscript{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
Available online 18 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.

© 2014 Elsevier B.V. All rights reserved.

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 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 8500 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 6552 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 h 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 4368 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, 2012; Eyben et al., 2009) 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 (Fousec 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 (Fousec 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 data (Schuller et al., 2011; Krajewski et al., 2012). In this study, without dealing with linguistic models, we aimed to improve sleepiness detection accuracy while reducing computational complexity. The perceptual feature set
proposed in Sezgin et al. (2012) has been adapted to model audio content of sleepy data in both spectral and temporal domains.

The introduced framework intends to overcome the general absence of a valid measure in emotion recognition. In order to attain features that efficiently represent variations in sleepiness levels, we formulate acoustic features by measuring differences in loudness at different levels rather than measuring merely loudness. Furthermore, our features are capable of monitoring loudness differences not only through critical bands but also in temporal domain.

Moreover, the characteristic of a speaker state may have to be monitored for several minutes in case of medium term sleepiness state classification. To fulfill this requirement, we introduce a learning plan that utilizes a bag-of-words (BoW) scheme which generates a vocabulary for sleepiness level differences. The extracted codewords are collected in a training set as representative features for sleepy and non-sleepy speech. Hence, based on the learned dictionary, sleepiness level changes are formulated as a binary classification problem by conventional SVM. Extensive tests on the SLC data reveal discernible improvement compared to the existing methods (Schuller et al., 2011; Krajewski and Kröger, 2007; Huang et al., 2011; Krajewski et al., 2012). The preliminary results of this study are presented in a previous work (Gunsel et al., 2013). As a consequence of the frame based feature extraction scheme, rather than segment based techniques, our method does not require a pre-segmentation stage and can be conveniently adapted to online processing.

The rest of the paper is organized as follows: Section 2 describes the preprocessing steps applied on the speech data; the perceptual features are formulated in Section 3; Section 4 presents the introduced learning scheme that generates a dictionary for sleepy speech; the soft-majority voting scheme employed for speaker state recognition is described in Section 5; The discussions about the efficiency of the features and the test results as well as the computational complexity are reported in Section 6; and Section 7 summarizes conclusions and future work directions.

2. Perceptual enhancement

A number of preprocessing stages are employed in the system in order to model perceptual and physiological effects of the human ear. The perceptual masking steps, described by ITU (Thiede et al., 2000), are mainly applied to the frequency spectrum in order to enhance perceptual content of speech. The preprocessing is followed by feature extraction. The low level features used in sleepiness detection are computed at the perceptual spectrum on the Bark scale and the acoustic spectrum in Hz (Fig. 1). For the sake of improving learning capability, a chunking operation is applied to the extracted feature set and chunked features are fed into the training and the test modules. The training set is refined by a BoW (Csurka et al., 2004) scheme that provides a vocabulary for sleepiness level differences in quasi-continuum KSS. The BoW scheme integrated with the dictionary learning module enables the generation of a number of codewords representing different sleepiness levels. The resulting codewords are used for classifier training. The classification module depicted in Fig. 1 labels the categories of speech chunks based on the decision rule provided by the classifier. The sleepiness recognition rates are reported per utterance after the soft-majority voting procedure is applied to the labeled chunks. The notation and the formulation of the basic preprocessing steps are presented in the following figure.

Let \( s_n[k, n] \) denote a discrete audio signal where \( n \) and \( k \) represent index time-frame and time-sample, respectively. Each windowed frame is transformed to frequency domain by using STFT. The transformed audio frame can be expressed as:

\[
F[k_f, n] = \frac{1}{N_{FT}} \sum_{k=0}^{N_{FT}-1} h_n[k] s_n[k, n] e^{-j2\pi k_f k},
\]

where \( k_f \) is the frequency bin index and \( h_n[\cdot] \) refers to the Hann windowing function. \( N_{FT} \) denotes the size of the STFT which is set to 2048 in our system.

Successive frames of speech signal are processed based on the ITU Perceptual Evaluation of Audio Quality (PEAQ) psycho-acoustic model (Thiede et al., 2000) in order to enhance perceptual components in frequency spectrum. To achieve this goal, a spectral weighting reflecting outer and middle ear frequency response is applied to spectral components. Eq. (2) formulates the outer ear weighted spectrum

\[
F_w[k_f, n] = F[k_f, n] \cdot 10^{0.02(\mu)}.
\]

\[
W[k_f] = \left[ -0.6 \cdot 3.64 \cdot k_f^{-0.8} + \left( 6.5 \cdot e^{-0.6(k_f^{-3.5})^2} \right) - \left( 10^{-3} \cdot k_f^{3.6} \right), \right.
\]

where \( W[k_f] \) denotes the outer middle ear frequency response at the frequency bin \( k_f \) (Thiede et al., 2000). In this study, three of the features proposed for sleepiness detection are calculated using this weighted spectrum in Hz. In addition to these features, contrary to the conventional feature extraction schemes mostly operating in Mel scale, we propose an additional six low level descriptors calculated in perceptual spectrum, thus derived in Bark scale. Hence, we map spectrally weighted energy values from Hertz domain to the Bark scale by using Eq. (4)

\[
k = \text{Bark}(k_f) = 7 \ \text{arcsinh} \left( \frac{k_f}{650} \right).
\]

where \( \text{arcsinh} \ (k_f) = \ln \left( k_f + \sqrt{k_f^2 + 1} \right) \). In order to simulate the dispersion of energy along the basilar membrane and to model spectral masking effects in the Bark domain, energy components are convolved with a spreading function \( S[i, k, P] \). \( F_w[k_f, n] \) is the outer ear weighted spectral energy and let \( P[k_f, n] \) be its Bark representation, where \( k_f \) in Hz is replaced by \( k \) after it
has been mapped onto the Bark scale. The outer ear spectrally weighted energy is smeared out over frequency bands according to the level dependent spreading function of ear. Conventionally, \( S[i, k, P] \), the spreading function of the band \( i \) for an energy component at the band \( k \) is defined as a two sided exponential

\[
S[i, k, P] = \begin{cases} 
27(i - k)\Delta i; & i \leq k \\
[-24 - \frac{230}{f_c[i]} + 2\log_{10}[P[i, n]]](i - k)\Delta i; & i > k
\end{cases}
\]

where \( f_c[i] \) is the center of the Bark scale and \( \Delta i \) is the Bark scale band. The spreading function is triangular in shape, where the first part of the spreading function in Eq. (5) identifies the lower slope and the second line shows the upper slope of the triangle. The lower slope is 27 dB/Bark through all frequency and energy levels, whereas the upper slope is updated depending on frequency and energy level.

The frequency spreading is computed for each critical band \( k \) of each audio frame \( n \) by smearing the spectral energy over the frequency as in Eq. (6) Thiede et al., 2000,

\[
E_i[k, n] = \frac{\left(\sum_{i=0}^{N_s-1} P[i, n]S[i, k, P]^{0.4}\right)^{2/3}}{B_i[k, n]}
\]

where \( N_s \) is the number of the Bark scales and \( B_i[k, n] \) is a normalizing factor calculated for a reference energy level of 0 dB for each band of frame \( n \). The main purpose of applying the frequency spreading is to model the masking aspect of the ear, which perceptually smoothenes frequency transitions.

We also perform time domain spreading that corresponds to forward masking. Among conventional forward masking functions, which are commonly used in audio compression, we prefer to use the model introduced in PEAQ which enables the tracking of sleepiness level variations of successive frames over time. To design forward masking, the energy levels in each critical band are smeared out over time according to Eq. (7)

\[
E[k, n] = a \cdot E[k, n - 1] + (1 - a) \cdot E_i[k, n]^{0.3},
\]

where \( E[k, 0] = 0 \) and \( 0 < a < 1 \) controls the time constant for decaying energies. \( E[k, n] \) is compliant with the initial time spreading effect of \( E_i[k, n] \). Until now, the proposed formulations have been entirely based on a single frame, including the introduced time spreading function. Current values of \( E[k, n] \) depend on the past frames as in Eq. (7) and \( E[k, n] \) behaves like a frequency dependent, filtering over time. \( E[k, n] \) is the final energy level reached on the Bark scale after the frequency and time masking operations and is identified as

\[
E[k, n] = \max_{i} \left( E[k, n], E_i[k, n]^{0.3} \right),
\]

where \( n \) is the actual frame number, and \( k \) the band index. It is an important to note that six of nine perceptual features are extracted using the frequency and time masking on the Bark scale. The remaining three features are extracted in frequency domain after outer ear spectral weighting.

Fig. 2 illustrates outputs of the preprocessing modules. Outputs are plotted respectively for sample utterances collected from NSL and SL modes of the SLC database. In the first row, time domain representations of the audio files are given. The second row illustrates the corresponding spectrograms obtained by STFT. In the third row, the spectrograms are weighted by an auditory filter bank, which resembles the hearing model of outer ear. The last row shows the spectrograms after spectral and forward masking on the Bark scale.

3. Extraction of low level speech features

The preprocessing stages detailed in Section 2 are employed to model perceptual and physiological effects of the human ear. The preprocessing is followed by the feature extraction. Unlike existing studies, we hypothesize that variances between different sleepiness levels are more discriminative than emotional data itself. Thus, detection of loudness difference between sleepiness levels is proposed to be more efficient in comparison to loudness detection at a single sleepiness level. In order to demonstrate this
approach practically, we make use of a reference concept to distinguish sleepiness levels with respect to one another. Hence, six out of nine of our features reflect content variations of the observed speech frames from a reference. The reference is specified to highlight the difference between class variations.

In the context of this study, we have used speech frames recorded at different KSS levels as the reference. The effects of the reference selection on speaker state recognition performance also appear in the test results. We should emphasize that the subjective nature of sleepiness detection needs to employ a reference criterion, which scales the effect of the subjectivity.

The low level features used in sleepiness detection are computed on the Bark scale and the acoustic spectrum in Hz (Fig. 1). The descriptors included in the first group are: The normalized sleepiness level difference (NSD), the normalized spectral envelope difference (NSED1), the average of NSED1 over critical bands (NSED2), the temporal average of NSED1 through successive Y audio frames (NSED3), the average number of non-sleepy blocks (ANSB) and the overall loudness of the frame (OLOF). The perceptual 10 dB bandwidth ($W_{10}$) and 5 dB bandwidth ($W_{5}$) of SL/NSL blocks and the average harmonic structure magnitude (AHSM) are calculated at the weighted spectrum in Hz. Table 1 lists these features and gives a brief description of each one where a detailed formulation are presented in the following subsections.

3.1. Normalized sleepiness level difference

We use the variations observed on loudness through critical bands as a discriminative feature for medium term speaker state detection. To formulate variations at critical bands we use local loudness, which is described as the perceived loudness of an emotional signal after it has been reduced by a masker signal (Thiede et al., 2000). Masker signal causes loudness to be perceived at different frequency bands and is effective, particularly at low frequencies. Therefore, locality of loudness is established by adaptively masking low frequency components. Generally, masking describes an effect by which an audible signal becomes inaudible when a louder signal masks it. We refer the reference audio signal as the masker in our system and...
compute a local loudness rather than conventional loudness. In conclusion, we evaluate a localized loudness with respect to a reference set. The Bark energy of the sleepiness level difference is normalized with the masker in this feature.

If \( P[k, n] \) is the outer ear weighted spectral energy of sleepy speech on the Bark scale, \( P_{\text{PerDi}}[k, n] \) is the variation between the observed \( (P_{\text{Obs}}[k, n]) \) and the reference \( (P_{\text{Ref}}[k, n]) \) weighted energies. The \( \text{NSD} \) is formulated as the ratio of the emotional difference \( P_{\text{PerDi}}[k, n] \) given to the masking threshold \( M[k, n] \) formulated below

\[
\text{NSD} = 10 \log_{10} \left( \frac{1}{Y} \sum_{n=1}^{Y} \left( \frac{1}{N_c} \sum_{k=0}^{N_c-1} \frac{P_{\text{PerDi}}[k, n]}{M[k, n]} \right) \right),
\]

where \( N_e = 109, k \) denotes the number of critical bands, \( n \) refers to the frame number and \( Y \) is the chunk size. The masking threshold \( M[k, n] \) is calculated by

\[
M[k, n] = \frac{E[k, n]}{10^m} \quad (10)
\]

can be used to represent feature vectors corresponding to chunks, knowing that feature vector of a chunk is averaged over \( Y \) successive frames.

The \( \text{NSD} \) formulated in Eq. (9) can be computed at any frame \( n \) of the reference speech. There is no need to process SL, NSL and reference frames at the same order and also no need to use just one reference frame for each SL or NSL speech frame. As it is presented in Section 5, the reference is applied at frame level, but training and classification is applied at chunk level. It is shown in Section 6 that this results in an increased representation capability.

### 3.2. Normalized spectral envelope difference

We examine frequency spreading energy level variations in critical bands through successive frames, in order to track temporal content of speech. Normalized loudness of these variations in spectral domain gives us information about sleepiness levels of speech.

Let \( E_{\text{Sp}}[k, n] \) denote the loudness changes in time by exploiting the frequency spreading excitation pattern, \( E_{\text{Sp}}[k, n] \), of the \( r \)th frame at the \( k \)th critical band (see Eq. (6)). The formulation given in Eq. (12) is used to model the loudness changes within critical bands.

#### Table 1

| Low level descriptors used for sleepiness detection. |
|-----------------------------------|-----------------------------------|
| 1. Normalized sleepiness level difference (\( \text{NSD} \)) | Average of the masked variations between the pitch patterns of SL/NSL audio and the reference audio computed over the Bark scales of an audio frame |
| 2. Normalized spectral envelope difference (\( \text{NSED1} \)) | Normalized envelope variations of the unsmeared SL/NSL pitch patterns from the reference within the successive frames for each critical band |
| 3. \( \text{NSED2} \) | Average of \( \text{NSED1} \) over all critical bands |
| 4. \( \text{NSD} \) | The temporal average of \( \text{NSED1} \) through successive \( Y \) audio frames |
| 5. Overall loudness of the frame (\( \text{OLOF} \)) | Sum across all critical bands of all outer ear weighted loudness values of an audio frame |
| 6. Average number of non-sleepy blocks (\( \text{ANSB} \)) | Expected number of non-sleepy blocks within a time interval |
| 7. Average harmonica structure magnitude (\( \text{AHSM} \)) | Average of the fundamental frequencies estimated from the log spectrum of the correlations of sleepiness differences through critical bands for successive \( Y \) audio frames |
| 8. 10 dB perceptual band with \( (W_{E1}) \) | The highest frequency component which exceeds the noise floor by at least 10 dB |
| 9. 5 dB perceptual band with \( (W_{E2}) \) | The highest frequency component which exceeds the noise floor by at least 5 dB |
The temporal average of the
average number of non-sleepy blocks (ANSB) provides a measure for occurrence of high excitation levels through successive frames on the Bark scale. Since the nature of NSL speech tends to have higher excitation pattern peaks than SL speech, ANSB improves accuracy of NSL–SL discrimination.

A probabilistic approach that estimates the number of excitation patterns remaining above a loudness threshold is applied (Thiede et al., 2000) hereby to calculate expected number of NSL blocks within a time interval. Let \( e[k, n] \) denote the difference between the excitation levels of the reference and observed (SL, NSL) speech computed on the Bark scale \( k \) of the audio frame \( n \) in \( \text{dB} \). \( E_o[k, n] \) refers to the frequency spreading excitation pattern where \( O \) and \( R \) sub-indices refer to the observed (SL, NSL) and the reference signals

\[
e[k, n] = \frac{e_o[k, n]}{e_r[k, n]} = \frac{10 \log_{10}(E_o[k, n])}{10 \log_{10}(E_r[k, n])},
\]

Our aim is to specify speech frames in which excitation level difference is above a certain threshold. The probability of an excitation pattern remaining above a sleepiness threshold can be modeled by Thiede et al. (2000)

\[
p[k, n] = 1 - \left( \frac{d_1}{(EsO[k, n])^b} \right),
\]

where \( b \) is a constant and normalizing coefficient. Assuming that the observed frames are uncorrelated, the total probability of declaring the speech frame \( n \) as NSL can be calculated by

\[
s[k, n] = \left( \frac{d_1}{E_{oO}[k, n]} \right)^{1.7} + c_1 E_{oO}[k, n] + c_2 E_{oO}[k, n]^2 + c_3 E_{oO}[k, n]^3 + c_4 E_{oO}[k, n]^4
\]

where the parameters shown in Eq. (17) are respectively set to; \( c_0 = -0.2, \ c_1 = 0.05, \ c_2 = -0.001, \ c_3 = 5 \cdot 10^{-6}, \ c_4 = 9 \cdot 10^{-11} \) and \( d_1 = 5.9 \) (Thiede et al., 2000).

Basicallly, the ANSB is computed as the average number of blocks declared as NSL within one second. We expect
the excitation pattern of the sleepiness difference of the NSL speech to have higher peaks, which means relatively greater probability of the detection, $p[k,n]$. Fig. 5 illustrates the extracted ANSB features, which support this expectation. The excitation level differences for SL speech are lower and less noticeable on the higher Bark scales. As can be seen in Fig. 5, the SL–NSL speaker state discrimination capability of the ANSB is promising.

3.4. Overall loudness of the frame

The overall loudness of the frame (OLOF) measures average loudness through Bark scales. The loudness at the kth Bark scale of the nth frame can be formulated as

$$L[k,n] = \left( \frac{EIN[k]}{sThr[k] \cdot 10^4} \right)^{0.23} \left[ 1 - \left( 1 - \frac{sThr[k]}{EIN[k]} \right)^{0.23} \right] - 1,$$

(18)

where $EIN[k]$ is the internal noise of the ear. The threshold index $sThr[k]$ is calculated according to

$$sThr[k] = 10^{\frac{Bark(sThr[k]) - 1}{2}} \cdot \left( -2 - 0.05 \cdot \text{atan} \left( \frac{k}{\text{Bark}} \right) - 0.75 \cdot \text{atan} \left( \frac{k}{\text{Bark}} \right)^2 \right),$$

(19)

$$sThr[k] = \text{Bark} \left( sThr[k_f] \right).$$

(20)

The OLOF over critical bands, $L_{total}[n]$, is calculated as the sum across all filter channels of all loudness values above zero, as

$$L_{total}[n] = \frac{24}{N_c} \sum_{k=0}^{N_c} \max \{L[k,n], 0\}. \quad (21)$$

It is important that we incorporate the human auditory system (HAS) into the loudness formulation that yields another perceptual feature for sleepiness detection. The discrimination capability of OLOF can be observed from Fig. 6 in which loudness is plotted over the critical bands of NSL and SL speech. As expected, the overall loudness of NSL audio is much more dominant through the critical bands.

3.5. Average harmonic structure magnitude

Average harmonic structure magnitude (AHSM) emphasizes harmonic structure of speech and reflects variations in fundamental frequency. It is observed that NSL speech is more similar to a periodical signal having stable harmonics with respect to SL speech including intonation variations. Since non-linear frequency transformation would smear harmonic structure and prevents detection of sleepiness level variances, we prefer to assess harmonic
structures of sleepy speech in linear frequency spectrum using the AHSM.

The extraction of the AHSM can be briefly summarized as follows: First the outer ear weighted spectral energy differences between SL and NSL speech and the reference are computed through critical bands. Subsequently, correlation of sleepiness variances through critical bands is obtained. Fundamental frequency is estimated based on the log spectrum of the calculated correlation function. The average value of the fundamental frequencies, which are estimated for successive speech frames, is reported as the AHSM. The details of the AHSM feature can be attained from Sezgin et al. (2012). We use the correlation of sleepiness variances through critical bands instead of time domain components.

Fig. 7 illustrates the distribution of the correlation for a NSL–SL speech pair through critical bands over time. We aim to highlight the monotonous nature of NSL speech that is much more similar to a periodical signal with stable harmonics by using the AHSM as a representative feature. On the other hand, due to the intonation fluctuations of speech in sleepy mode, SL speech should not have a periodic structure as clear as NSL audio. The autocorrelation values computed for the NSL mode clearly indicate the periodic monotonous nature of the NSL speech with respect to the SL speech in Fig. 7.

3.6. Perceptual bandwidth

Perceived timbre, dullness and muffling effects in audio change according to sleepiness levels that yield different perceptual bandwidths. The perceptual bandwidth feature aims to distinguish sleepiness levels based on the variations in signal bandwidth.

In order to compute the perceptual bandwidth, the highest frequency corresponding to maxima of spectrum is obtained. Loudness level at this frequency is used as an estimate of a noise floor. Beginning with higher frequencies, a scanning process is applied to figure out the highest frequency component exceeding the noise floor by at least 10 dB. This frequency is defined as the estimated perceptual bandwidth. Hence, a rough estimate of the observed emotional speech bandwidth is computed for each frame and the mean bandwidth value calculated over $Y$ successive frames is used in sleepiness detection.

We also use another version of the perceptual bandwidth which exceeds the noise floor by at least 5 dB towards lower frequencies. Again the feature is extracted from the frame base and then elaborated on chunk base by averaging the feature over $Y$ successive frames. Fig. 8 plots the 5 dB bandwidth estimates through the samples taken from NSL and SL speech. The NSL speech owns a greater noise floor that yields a lower perceptual bandwidth.

![Fig. 6. Overall loudness calculated through the Bark scales of (a) NSL and (b) SL speech.](image)

![Fig. 7. Perceptual harmonic structure of (a) NSL and (b) SL speech.](image)
with regard to SL in Fig. 8. On the basis of this finding, this feature can be effective for detection of SL speech, since the variance of the bandwidth for the SL utterances is very small.

4. Dictionary learning

We model speaker state detection as a two-class classification problem (SL and NSL states). Medium term speaker state recognition requires longer monitoring intervals, which leads to difficulties in content representation and therefore classifier design. We use the BoW scheme to generate a vocabulary for the scattered content (to various KKS levels) of speech. The BoW model is commonly used for different problems to determine whether natural groupings are present in data (Webb, 2002). In order to detect compact clusters, we simply perform $k$-means clustering on all the perceptual feature vectors. The codewords are then identified as the centers of the learned clusters, thus they can be considered as representative of similar feature vectors. The number of the clusters $K$ is referred to as the codebook size with an analogy to the size of the word dictionary. The extracted codewords establish the training set, as the initial feature set is used for verification. Thus, dictionary codewords constitute a whole training set, as the initial feature set is used for verification.

Another issue that needs to be clarified is the size of the initial large training set. We perform frame based feature extraction. However frame based speaker state reporting results in computational complexity and problems of accuracy. The conclusion we have drawn from the evidence in our work is that the consolidation of successive $Y$ frames into a representative mean vector which is referred to as a chunk. The idea behind chunking is to make the speaker state transitions more tractable in medium term. $Y$ denotes a group of frames in which speaker state can be considered stationary. In this work $Y$ is empirically set to 70 and feature extraction has been performed on 43 ms frame length with 50% overlap corresponding to approximately 1 s chunk length. The frame-wise computed $M$-dimensional initial features are fed into the dictionary learning module as chunks after being averaged over $Y$ frames (Fig. 1).

5. Speaker state recognition by soft majority voting

We designed a nonlinear SVM classifier by using a Gaussian Radial Basis Function (GRBF) kernel formulated in Eq. (23) in order to address the nonlinear nature of the speaker sleepiness recognition problem,

$$K(x_{iR}, x_{jR}) = e^{-\frac{(x_{iR} - x_{jR})^2}{2\sigma^2}}$$

where $\sigma$ is the width of the kernel.
In Eq. (23) (according to the notation in Section 4) \( x_{iR} \) denotes the \( M \)-dimensional perceptual feature vector which is extracted from the \( i \)th audio chunk with respect to \( R \)th corresponding reference audio chunk, where \( R = 1, 2, \ldots, N \). Similarly, \( x_{iR} \) denotes the \( M \)-dimensional feature vector extracted from the \( i \)th audio chunk with \( R \)th corresponding reference. \( N \) is chosen as 1 or greater in such a way that guarantees a satisfactory recognition rate at the dictionary learning stage. It is important to note that the number of the representative feature vectors extracted from one speech utterance is equal to the number of the reference speech chunks \( N \) that is different from conventional methods. Furthermore, as it is reported in Section 6, the higher the value of \( N \), the higher the recognition accuracy. For simplicity, the same notation can be used to represent feature vectors corresponding to chunks, knowing that the feature vector of a chunk is averaged over \( Y \) successive frames.

An initial large training set is formed by assigning the class label \( y_i \in \{-1, +1\} \) to each training vector, which is represented as chunks. A dictionary for each class has been learned based on the learning scheme described in Section 4. Codewords, which represent the variations of the initial training set, form the final training vectors for each class. The final training set, consisting of \( K \) codevectors, is used to design SVM by maximizing the Wolfe dual of the Lagrange function given in Eq. (24) Vapnik, 1998

\[
\max_x W(x) = \max_x \left( \sum_{i=1}^{K} z_i - \frac{1}{2} \sum_{i,j=1}^{K} z_i z_j y_i y_j K(x_{iR}, x_{jR}) \right)
\]

subject to constraints

\[
\sum_{i=1}^{K} z_i y_i = 0, \quad 0 \leq z_i \leq C, \quad i = 1, \ldots, K
\]

where \( z_i \) is the \( i \)th Lagrange multiplier corresponding to the \( i \)th training vector, and \( K \) is the number of Lagrange parameters which is equal to the number of codewords. If the training set is not linearly separable, deviations from the decision boundary is controlled by the misclassification cost parameter \( C \), which defines an upper bound for the Lagrange multipliers. Consequently, a subset of the training samples is specified as support vectors that correspond to \( s \leq K \).

In a similar manner, a test set is constructed with \( x_{iR} \) which is an \( M \)-dimensional feature vector representing a chunk of speech in test data. It should be noted that the same reference set has been used for feature extraction in both test and training stages. The number of feature vectors extracted for each chunk is equal to the number of reference speech frames \( N \) as it is in the training session. The SL/NSL classification of each test feature vector is performed according to Eq. (26)

\[
V(x_i) = \text{sgn}\left( \sum_{s \in SV} z_s y_s K(x_i, x_{sR}) + b \right)
\]

where \( V(\cdot) \) describes the decision rule of the binary SVM classifier. \( x_{sR}, s \in SV, \) represents the support vectors (SVs) obtained at SVM training stage and \( b \) is the bias term (Vapnik, 1998). Following the classification phase, the final decision is given per each speech chunk using the soft-majority voting rule (Sezgin et al., 2012; Gunsel et al., 2013).

We apply a soft-majority voting (SMV) decision rule that assigns the class label \( j \) to a chunk of test speech based on the final probability computed for each chunk by Eq. (27), where the class label \( j \) can be either sleepy (SL) or non-sleepy (NSL).

\[
P_{j, \text{Final}} = P_{j, P} + P_{j, V} \cdot V(\cdot).
\]

The final probability computed for the \( i \)th test chunk given in Eq. (27) consists of two terms. \( P_{j, P} \) denotes the posterior probability assigned to the test sample \( x_{iR} \) by the classifier. Therefore, the first term gives the average of the \( N \) decision probabilities assigned by SVM. The term increases the final probability of class \( j \), when the decision probabilities over \( N \) samples are high. SMV denotes the posterior probability of SL/NSL class when \( N \) is set to 1 in which SVM labels only one chunk of the representative feature vector. The second term aims to increase the precision on the majority of decisions made by SVM. The second term of the final probability is formulated by Eq. (28)

\[
P_{j, V} = \frac{1}{N} \sum_{i=1}^{N} g(p(w_i|x_{iR})).
\]

where \( g(p(w_i|x_{iR})) = \begin{cases} 1, & \arg\max_{x_i}(p(w_i|x_{iR})) \\ 0, & \text{otherwise} \end{cases} \) and \( \sum P_{j, V} = 1 \) which makes use of the relative probability values of the categories by a quantification linear function \( g(p(w_i|x_{iR})) \). The overall approach of SMV takes greater amount of statistical data into consideration.

6. Experimental results

We have evaluated the speaker state recognition performance using the proposed and the existing systems on the Sleepy Language Corpus (SLC). Section 6.1 describes the training and the test data. Impact of the dictionary learning in terms of computational complexity and recognition accuracy has been discussed in Section 6.2. Section 6.3 presents the sleepiness detection recall rates achieved by the proposed features and their comparison with state-of-the-art methods. Section 6.3 also covers test results that evaluate the diversity measures of the proposed features.

6.1. SLC data and experimental scenarios

We have performed the tests on the SLC data (Krajewski and Kröger, 2007), which is used in the Speaker State Challenge (Schuller et al., 2011), to compare our performance with existing systems. The annotated SLC database provides 10 KSS labels where only 2 main classes are defined: sleepy (SL) for a level exceeding 7.5 on the
KSS, and non-sleepy (NSL) for a level equal or below 7.5. The scores range from 1 to 10: extremely alert (1), very alert (2), alert (3), rather alert (4), neither alert nor sleepy (5), some signs of sleepiness (6), sleepy, but no effort to stay awake (7), sleepy, some effort to stay awake (8), very sleepy, great effort to stay awake, struggling against sleep (9), extremely sleepy, cannot stay awake (10) Krajewski et al., 2010.

Automatic classification of the sleepiness state of a speaker is confounded by at least four factors in the SLC database. Firstly, there are a number of speakers, in total 99, in the database. The speakers have diverse vocal tracts, glottis physiologies and differing ranges of vocal expressivity such as activation level during communication. Secondly, the database consists of three styles of speech: read speech, spontaneous speech and command/control. Vocal characteristics of each style differ greatly. For instance, speech rate and intonation are expected to be less variable in read or command speech, as compared to spontaneous speech for a given speaker. Thirdly, the utterance durations range from 0.5 to over 60 s. This indicates drastically different amounts and types of information. Different speakers, speech styles and utterance durations provide a complex dataset that makes feature generalization challenging. Fourthly, the binary two-class problem gets more complicated because of the 10-level scattered sleepiness level structure of the SLC. Learning the sub-classes requires attention during the training phase (Krajewski and Kröger, 2007).

In the related literature two test scenarios are defined on the SLC database, where the test case 1 (TC1) covers the “TRAIN vs DEVEL” and the test case 2 (TC2) comprises “TRAIN + DEVEL vs TEST” test cases. The content of the test data employed in both scenarios, TC1 and TC2, are listed in Table 2. The training set (TRAIN) in TC1 consists of 3366 utterances collected from 36 subjects and the test set (DEVEL) includes 2915 utterances collected from 30 subjects. Second test scenario (TC2) employs an enlarged training set that is constituted by including the set DEVEL into the TRAIN (TRAIN + DEVEL). The test data (TEST) of TC2 consists of 2808 utterances collected from 33 subjects.

### 6.2. Impact of dictionary learning

We introduce the dictionary learning scheme to reduce computational complexity while preserving speaker state classification accuracy. Gain of dictionary learning can be explained based on the number of feature vectors included in training set (\(L\)) for both TC1 and TC2 (Table 2).

Training features are extracted from the training data provided by the Speaker State Challenge organizers. In order to compare our performance with the existing methods, we did not make any optimization on the training data. \(L\) is the size of the initial training set on which dictionary learning has been applied. Using BoW on the training set, \(K\) codewords are provided which we use for classifier training. It is shown in Table 2 that as the described learning scheme is performed, the number of feature vectors can be reduced about 150 times. Note that the amounts of the codewords assigned to each sleepiness level do not have to be equal.

Regarding the reference set concept, to learn smooth transition patterns between sleepiness levels, we have used a large reference set at the training stage. Initially, 7 reference subjects, selected from the NSL (KSS = 2) and the SL (KSS = 8) speech, are used as the reference for TC1. The size of reference set has increased to 15 in TC2. Since, the introduced scheme is designed to model loudness variabilities rather than loudness itself, the reference utterances are specified both from the SL and the NSL content to highlight the variations. At the test stage, to reduce computational load, we have used 7 and 4 of these reference utterances at the extraction of test features in TC1 and TC2, respectively. It should be noted that none of the training and the test utterances are used in the reference. The last column of Table 2 gives the total number of test feature vectors for each test case. Final scores for the speaker state classification accuracy are reported after the soft majority voting described in Section 5. Since the TRAIN and DEVEL data subsets are biased roughly (70% – 30%) towards the NSL utterances, using weighted average recall would be misleading (Kaida et al., 2006; Schuller et al., 2011). Hence, unweighted average recall rate (UA) is used as the generic performance metric in this work. UA is a commonly used measure that does not weigh class accuracies by size of each class, but gives each class equal weights.

In Section 6.3, it is shown that our classification performance outperforms existing systems that have been employed for comparison. Since we use SVM as it is in the comparable tasks, we can quantify the computational gain of dictionary learning and the used compact feature set based on the complexity of SVM. Burges formulated the complexity of SVM as (Burges, 1999)

\[
O(S^3 + (S^2 \cdot L) + (S \cdot M \cdot L))
\]

for the training stage, where \(S\) is the number of support vectors, \(M\) is the dimension of feature vectors and \(L\) is the number of training samples, if most support vectors are not upper bound and \(S/L \ll 1\). In test phase the

<table>
<thead>
<tr>
<th>Test scenario</th>
<th>Train data</th>
<th># original utterances</th>
<th>Test data</th>
<th># original utterances</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>TC1</td>
<td>TRAIN</td>
<td>3366</td>
<td>DEVEL</td>
<td>2915</td>
<td>77,022</td>
</tr>
<tr>
<td>TC2</td>
<td>TRAIN + DEVEL</td>
<td>6281</td>
<td>TEST</td>
<td>2808</td>
<td>13,181</td>
</tr>
</tbody>
</table>
complexity of SVM is stated as $O(CS)$, where $C$ is the number of operations required to evaluate kernel. Since we use RBF kernels, $C$ is $O(M)$, in order of the dimension of data vectors. It is obvious that our computational complexity will be much lower at the classification stage, knowing $M = 9$ in our study, as it is in the order of thousands in comparable systems. Furthermore, we can remark that the number of support vectors $S$ can be equal to, at most, $K$, the number of codewords in the proposed scheme. Since $K$ is 150 times lower than $L$, the proposed dictionary learning scheme highly reduces the computational complexity at the training stage.

### 6.3. Speaker state recognition performance

We employ UA recall rates as an evaluation measure and SVM as the classifier for medium term speaker state recognition. Conventionally, SVM is commonly used in the related literature, because it is well fitted to large acoustic feature sets and its training complexity does not depend on feature set size (Weninger et al., 2012). Hence, in our work SVM has been selected to compare our performance against well-known openSMILE features (Eyben et al., 2009; Schuller et al., 2011; Krajewski and Kroger, 2007). LibSVM tool of WEKA toolkit (Witten and Frank, 2000) is used with the RBF kernel parameters $C$: 100, $Nu$: 0.5.

The unique impact of the features extracted in Hz and Bark scale is tested separately. Table 3 respectively reports recall rates for SL data ($RR_{SL}$), NSL data ($RR_{NSL}$) and unweighted average accuracy (UA) achieved by the six Bark features, the three Hertz features and all of the nine features for the scenario TC1. It is shown from Table 3 that Bark features provide 17% and 30% higher accuracy in the detection of speaker states from NSL and SL speech, respectively. The Bark features also have more impact on NSL distinction with regard to Hz features. As a result of psychoacoustic frequency and time masking applied to the Bark scale, smooth changes of non-sleepy speech become tractable with the perceptual features. On the other hand, inclusion of the outer ear model into feature extraction in Hz makes us capable of detecting abrupt sleepiness differences from sleepy speech. According to these findings using all of the nine features have more impact on detection of non-sleepy speaker states (NSL) instead of only using the six Bark features.

Based on the fact that all of the features are necessary to improve speaker state recognition, we have compared the proposed scheme to the state-of-the art systems. Table 4 reports the comparative recall rates. IS2011 Winner refers to the highest scores reported by the Interspeech 2011 Speaker States Challenge participants (Schuller et al., 2011; Huang et al., 2011) where the acoustic features of openSMILE (Eyben et al., 2009) are used. IS2011 SSC denotes the highest baseline performance declared by the organizers of the Challenge (Schuller et al., 2011). The results are obtained by openSMILE that extracts more than 6552 features by 39 function of 56 acoustic low-level descriptors (LLD) and corresponding first and second order delta regression coefficients. Table 4 reports the system performance on the SLC for TC1 and TC2 scenarios in terms of $RR_{SL}$, $RR_{NSL}$ and unweighted average accuracy per class. The conclusion drawn from Table 4 is that the SVM with the perceptual features achieves the highest detection rates for both TC1 and TC2 scenarios, compared to the performance by the IS2011 Winner and IS2011 SCC. We observe that the conventional features provide higher accuracy on the non-sleepy speech data (NSL). Furthermore, improvement achieved by the proposed scheme becomes more tractable on the sleepy speech (SL) with the improvement ranging between 15% and 29%. Thus, the introduced scheme enables us to monitor spectral and temporal changes of sleepy speech efficiently.

The size of the training set that directly depends on the size of the reference set in our scheme, is a crucial parameter in all pattern recognition problems. We aim to minimize the number of utterances employed as the reference by dictionary learning. In order to evaluate the SL–NSL classification accuracy with the reference set size, the reference set size has been increased 16 times. However, the SLC corpus includes 10 sub-classes (KSS levels) which makes our binary classification problem more complicated. This complication forces us to select the references among different categories instead of a unique category which differs from our previous work (Sezgin et al., 2012). Table 5 illustrates the plot of UA recall rates against the number of the reference utterances for TC2. UA rates achieved by the SVM increased from 80% to 90%, when the number of the reference utterances increased from 4 to 57. 10% higher accuracy is, however, achieved by a cost of about 16 times higher classification load. During our study we concluded that samples from different classes, SL or NSL, have to be included in the reference set, especially at the training stage. After determining the content of the reference subjects based on the silhouette measure Eq. (22) during training stage, optimum size at test stage can be determined heuristically.

### Table 3
Sleepiness detection rates achieved by the features extracted in Hz and Bark scale regarding TC1.

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Bark features: NSD, NSED1, ANSB, NSED2, NSED3, OLOF</th>
<th>Hz features: $W_{E1}, W_{E2}, AHSM$</th>
<th>All features</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$RR_{SL}$</td>
<td>$RR_{NSL}$</td>
<td>UA</td>
</tr>
<tr>
<td>SVM</td>
<td>87.1</td>
<td>88.0</td>
<td>87.5</td>
</tr>
</tbody>
</table>
which is \( \frac{Q_{11} - N_{00}}{N_{11} + N_{00}} \) when speaker state labels are as below:

\[ Q_{F1,F2} = \frac{N_{11}N_{00} - N_{01}N_{10}}{N_{11}N_{00} + N_{01}N_{10}} \]  

(30)

where \( N_{11} \) and \( N_{00} \) are the numbers of instances that the predictions of the features \( F1 \) and \( F2 \) are both correct or incorrect, respectively. \( N_{01} \) and \( N_{10} \) are the numbers of instances where only \( F1 \) or \( F2 \) commit an error. Note that \( Q_{F1,F2} \) will be equal to 1, when speaker state labels are exactly the same for both features. \( Q_{F1,F2} \) will be equal to -1 when there is no consensus between decisions made individually.

Table 7 reports the pairwise \( Q \)-statistics calculated for TC1 on where all values are lower than 0.361. This explicitly shows that there is no significant correlation between the features, hence none of them are redundant. \( Q \)-statistics exhibit solid agreement with the recall rates reported in Table 7. It can be concluded that the \( NSD \) is the most discriminative feature, since higher negative scores are calculated for it.

Since we have shown that the features are not redundant within themselves, it makes sense to observe the impact of all features in Bark and Hertz domains at various KSS levels. Table 8 reports the recognition rates achieved by SVM for TC1. It is to be seen that the features extracted from the Bark spectrum highly improve the speaker state classification accuracy for the NSL speech (KSS < 7.5) compared to the features in Hertz. However, 9% complementary positive effect on the overall performance (UA) is achieved by using all of the features.

In order to evaluate the correlation between the proposed features and KSS levels, Spearman’s rank correlation coefficient is computed. Spearman’s rank correlation coefficient is a non-parametric measure of statistical dependence between two variables and is computed for ranking two variables \( x_i, y_i \) as below:

\[ \rho = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2\sum(y_i - \bar{y})^2}} \]  

(31)

where \( x_i \) is the UA recall rate computed for each KSS level by SVM and \( y_i \) is the KSS deviation from KSS level threshold (7.5) referring to the border of SL–NSL. \( \bar{x} \) is the mean value of \( x_i \) and \( \bar{y} \) stands for the mean value of \( y_i \) which is 7.5. Eq. (32) gives the sum of correlations of all KSS levels

\[ \rho = \sum_{j=1}^{10} \rho_j \]  

(32)
Table 7
Pairwise $Q$-statistics measuring whether the features commit the same errors on the utterances.

<table>
<thead>
<tr>
<th></th>
<th>WE2</th>
<th>AHSM</th>
<th>OLOF</th>
<th>NSED1</th>
<th>ANEB</th>
<th>NSED2</th>
<th>NSED3</th>
<th>NSLD</th>
</tr>
</thead>
<tbody>
<tr>
<td>$W_{E1}$</td>
<td>-0.031</td>
<td>0.057</td>
<td>0.028</td>
<td>0.060</td>
<td>-0.010</td>
<td>0.047</td>
<td>-0.021</td>
<td>-0.048</td>
</tr>
<tr>
<td>$W_{E2}$</td>
<td>0.024</td>
<td>0.009</td>
<td>0.011</td>
<td>0.055</td>
<td>-0.034</td>
<td>-0.101</td>
<td>-0.306</td>
<td>0.016</td>
</tr>
<tr>
<td>AHSM</td>
<td>0.011</td>
<td>0.038</td>
<td>0.009</td>
<td>-0.001</td>
<td>-0.040</td>
<td>-0.166</td>
<td>0.12</td>
<td>0.133</td>
</tr>
<tr>
<td>OLOF</td>
<td>0.019</td>
<td>0.047</td>
<td>0.014</td>
<td>-0.01</td>
<td>-0.16</td>
<td>-0.14</td>
<td>0.084</td>
<td>-0.259</td>
</tr>
<tr>
<td>NSED1</td>
<td>0.006</td>
<td>0.001</td>
<td>-0.084</td>
<td>0.014</td>
<td>-0.021</td>
<td>-0.133</td>
<td>-0.266</td>
<td>0.04</td>
</tr>
<tr>
<td>ANEB</td>
<td>-0.021</td>
<td>-0.099</td>
<td>-0.099</td>
<td>-0.133</td>
<td>-0.266</td>
<td>0.016</td>
<td>-0.07</td>
<td>0.099</td>
</tr>
<tr>
<td>NSED2</td>
<td>-0.133</td>
<td>-0.266</td>
<td>-0.133</td>
<td>-0.266</td>
<td>0.016</td>
<td>-0.07</td>
<td>0.099</td>
<td>0.361</td>
</tr>
<tr>
<td>NSED3</td>
<td>-0.266</td>
<td>0.016</td>
<td>-0.07</td>
<td>0.099</td>
<td>0.361</td>
<td>0.12</td>
<td>0.099</td>
<td>0.06</td>
</tr>
</tbody>
</table>

Table 8
Distribution of correctly classified instances (%) versus sleepiness scales for NSL (KSS < 7.5) and SL (KSS > 7.5) with SVM (PCA) for TC1.

<table>
<thead>
<tr>
<th>SVM (TC1)</th>
<th>KSS Levels</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7–7.5</th>
<th>7.5–8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bark (%)</td>
<td>94.3</td>
<td>92.9</td>
<td>85.2</td>
<td>88.4</td>
<td>89.7</td>
<td>87.1</td>
<td>76.9</td>
<td>92.2</td>
<td>91.3</td>
<td>94.2</td>
<td></td>
</tr>
<tr>
<td>Hz (%)</td>
<td>34.3</td>
<td>36.5</td>
<td>52.1</td>
<td>63.4</td>
<td>40.5</td>
<td>64.4</td>
<td>43.2</td>
<td>71.7</td>
<td>69.4</td>
<td>87.5</td>
<td></td>
</tr>
<tr>
<td>All (%)</td>
<td>100</td>
<td>96.5</td>
<td>92.4</td>
<td>97.5</td>
<td>92.4</td>
<td>98.3</td>
<td>87.6</td>
<td>92.4</td>
<td>89.7</td>
<td>94.2</td>
<td></td>
</tr>
</tbody>
</table>

Table 9
Spearman’s rank correlation ($\rho_j$) coefficients computed for the proposed features against KSS deviation from KSS = 7.5 with SVM.

<table>
<thead>
<tr>
<th>SVM</th>
<th>$\rho_1$</th>
<th>$\rho_2$</th>
<th>$\rho_3$</th>
<th>$\rho_4$</th>
<th>$\rho_5$</th>
<th>$\rho_6$</th>
<th>$\rho_7$</th>
<th>$\rho_8$</th>
<th>$\rho_9$</th>
<th>$\rho_{10}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>All features</td>
<td>0.23</td>
<td>0.05</td>
<td>0.08</td>
<td>0.06</td>
<td>0.06</td>
<td>0.02</td>
<td>0</td>
<td>-0.03</td>
<td>-0.07</td>
<td>-0.02</td>
</tr>
<tr>
<td>$W_{E1}$</td>
<td>0.22</td>
<td>0.04</td>
<td>0.10</td>
<td>0.08</td>
<td>0.05</td>
<td>0.02</td>
<td>-0.01</td>
<td>-0.03</td>
<td>-0.07</td>
<td>-0.02</td>
</tr>
<tr>
<td>$W_{E2}$</td>
<td>-0.12</td>
<td>-0.08</td>
<td>-0.08</td>
<td>-0.06</td>
<td>-0.05</td>
<td>-0.02</td>
<td>-0.01</td>
<td>0.02</td>
<td>0.06</td>
<td>0.14</td>
</tr>
<tr>
<td>AHSM</td>
<td>0</td>
<td>0.08</td>
<td>0.12</td>
<td>0.12</td>
<td>0.01</td>
<td>0.06</td>
<td>-0.01</td>
<td>0</td>
<td>-0.06</td>
<td>-0.11</td>
</tr>
<tr>
<td>OLOF</td>
<td>-0.20</td>
<td>-0.04</td>
<td>-0.10</td>
<td>-0.07</td>
<td>0.01</td>
<td>0</td>
<td>-0.02</td>
<td>0.03</td>
<td>0.04</td>
<td>0.09</td>
</tr>
<tr>
<td>NSED1</td>
<td>-0.02</td>
<td>-0.12</td>
<td>-0.09</td>
<td>-0.05</td>
<td>-0.03</td>
<td>-0.02</td>
<td>-0.02</td>
<td>0.02</td>
<td>0.06</td>
<td>0.13</td>
</tr>
<tr>
<td>ANEB</td>
<td>-0.14</td>
<td>-0.08</td>
<td>-0.07</td>
<td>-0.06</td>
<td>-0.04</td>
<td>-0.03</td>
<td>-0.01</td>
<td>0.02</td>
<td>0.06</td>
<td>0.13</td>
</tr>
<tr>
<td>NSED2</td>
<td>0.07</td>
<td>-0.03</td>
<td>-0.05</td>
<td>-0.03</td>
<td>0</td>
<td>0</td>
<td>-0.03</td>
<td>0.01</td>
<td>0.02</td>
<td>0.12</td>
</tr>
<tr>
<td>NSED3</td>
<td>0.12</td>
<td>0.15</td>
<td>0.08</td>
<td>0.09</td>
<td>0.06</td>
<td>0.04</td>
<td>-0.01</td>
<td>-0.02</td>
<td>-0.06</td>
<td>-0.09</td>
</tr>
<tr>
<td>NSD</td>
<td>0.11</td>
<td>0.13</td>
<td>0.03</td>
<td>0.01</td>
<td>0.05</td>
<td>-0.01</td>
<td>-0.04</td>
<td>0.01</td>
<td>0.01</td>
<td>0</td>
</tr>
</tbody>
</table>

where $j$ contributing to 10 KSS levels and $\rho_j$ indicating one KSS level at a time.

Table 9 reports the Spearman’s rank correlation coefficients calculated for each feature against KSS levels. The absolute value of the correlation of the features with KSS deviation is relatively small ($<0.2$) which indicates that the accuracy change with the deviation is limited. The correlation exhibits a stable behavior through the KSS levels except SL–NSL border where it performs on the lowest base particularly ($\sim0.01$ to 0.03) where KSS = 7.5.

7. Conclusions

We propose a binary classification scheme with compact number of features for speaker sleepiness state recognition from speech. The main contribution of this paper is the proposed compact and sparse representation that provides around 10% improvement on the unweighted recall rates compared to the benchmarking results.

Unlike existing systems that rely on phonetic features, we propose a sleepiness detection system for speech that integrates psychoacoustic masking and temporal masking into feature extraction. The proposed perceptual feature set differs from the existing work not only because it integrates the human auditory system into the feature extraction scheme, but it also aims to model emotional differences rather than emotions themselves. This is achieved by formulating seven out of nine features based on loudness differences between the sleepy and non-sleepy modes. Inspired by the reference usage in audio quality measurements, we describe a reference set concept that enables us to model loudness differences. In addition, these features aim to model speaker state variations through individual critical bands rather than the whole spectrum, which is also different from the existing feature extraction schemes. Furthermore, our features are capable of tracking speaker state variations over time. Three out of nine features, $NSED_1$, $NSED_2$, and $NSED_3$, track spectral envelope variations in critical bands within successive speech frames. The average number of non-sleepy blocks, $ANSB$, is also formulated as a feature to track temporal changes occurring in a specific time interval.
The introduced dictionary learning scheme allows learning the sparse content of sleepy and non-sleepy speech in an efficient way. This yields a significant decrease in computational complexity at the classifier training stage. Another advantage of the developed model is that it does not apply a linguistic segmentation stage, and therefore can be easily adapted to online processing.

References


