SLEEPINESS DETECTION FROM SPEECH BY PERCEPTUAL FEATURES

Bilge Günsel, Cenk Sezgin*, Jarek Krajewski***

*Multimedia Signal Processing and Pattern Recognition Group, Istanbul Technical Univ., Turkey
** Experimental Industrial Psychology, Univ. of Wuppertal, Germany
{csezgin, gunselb}@itu.edu.tr, krajewsk@uni-wuppertal.de

ABSTRACT

We propose a two-class classification scheme with a small number of features for sleepiness detection. Unlike the conventional methods that rely on the linguistics content of speech, we work with prosodic features extracted by psychoacoustic masking in spectral and temporal domain. Our features also model the variations between non-sleepy and sleepy modes in a quasi-continuum space with the help of code words learned by a bag-of-features scheme. These improve the unweighted recall rates for unseen people and minimize the language dependence. Recall rates reported based on Karolinska Sleepiness Scale (KSS) for Support Vector Machine and Learning Vector Quantization classifiers show that the developed system enable us monitoring sleepiness efficiently with a lower complexity compared to the reported benchmarking results for Sleepy Language Corpus.

Index Terms— sleepiness detection, human–machine interaction, audio emotion detection.

1. INTRODUCTION

Sleepiness is an important quasi-emotional state which affects safety, performance, comfort and joy-of-use in many fields of human–machine interaction [1]. Using speech for sleepiness detection is one of the challenging topics in the literature, because it is a more robust configuration against environmental conditions [2, 3]. Some of the related work in the literature deals with the feature extraction while others focus on classification methods to improve the detection performance. In [3] total of 8500 prosody, articulation and speech quality related features are calculated for detecting accident-prone fatigue state classification. The highest class-wised averaged rate achieved is reported as over 80%. The openEAR emotional search engine is adopted to the sleepiness detection problem in most of the recent studies. openEAR is a generic emotion detection tool, which extracts more than 6.552 features by 39 functional of 56 acoustic low-level descriptors [4]. Recently the sleepiness sub-challenge in INTERSPEECH 2011 addressed the sleepy-non-sleepy classification problem from speech [5]. Test results were reported on Sleepy Language Corpus (SLC) [6] based on 10 different levels of the KSS [7]. In [5] an extended subset of openEAR features, a total of 4368 features including spectral, energy and voice related low level descriptors and their statistical variants, is used for the sleepiness detection. The highest recognition rate achieved by SVM is reported as 70.3%. The system proposed in [8] provides 71.6% detection accuracy achieved by AdaBoost fusion of SVM and a new classifier referred as Asymmetric Simple Partial Least Squares (SIMPLS). In [9] a novel feature set is selected by applying a correlation-filter subset selection on Non Linear Dynamics (NLD) and openEAR features that yielded 565 descriptors including 395 non-linear dynamics and 170 phonetic features. A subset of the SLC data set that consists of 372 utterances collected from 77 speakers is employed for experiments. The highest recognition rates are respectively reported as 79.6% (Bayes Net) and 77.1% (AdaBoost Nearest Neighbor) for male and female speakers.

Conventional systems make use of acoustic features which are originally proposed for speech recognition hence they may not fully model the sleepiness perception because a vast majority of them, such as MFCC, are generated for short speech frames to decode the phonemes. Consequently, a high performance sleepiness detector could only be achieved by using very large feature sets or considerably small feature sets in combination with highly complex classifiers [8, 9, 10].

In this work, we attempt to improve the sleepiness detection rates while reducing the computational complexity. The sleepiness detection is formulated as a binary classification problem based on KSS (sleepy (SL) for a level exceeding 7.5, and nonsleepy (NSL) for a level equal or below 7.5). The perceptual feature set proposed in [11] is adopted to model the audio content of sleepy data. Unlike the existing features that rely on the linguistic content of speech, we aimed to learn a vocabulary for the sleepiness level differences in both perceptually masked spectral and the temporal domain. The recognition rates achieved by SVM and LVQ obtained on SLC data show recognizable improvement compared to existing methods.

2. PROPOSED SYSTEM

Fig. 1 illustrates the main blocks of the overall system. The low level features used in sleepiness detection are computed at perceptual spectrum in Bark scale as well as the acoustic spectrum in Hz. The feature set is referred as *perceptual* because in order to model physiological and the perceptual effects of the human ear we apply the outer ear masking for both domains and an additional psychoacoustic masking in Bark [12]. Basically we have a training scheme in which the SL-NSL classifier is designed based on a training feature set refined by a Bag-of-Features (BoF) [13] scheme. The idea behind performing BoF is to learn the vocabulary for the sleepiness level differences in the quasi-continuum Karolinska Sleepiness Scale. Thus a descriptor (codeword) is assigned to each sub-cluster independently and the resulting set of codewords are used for model learning. In our work, the codewords are specified by vector quantization applied on the features whitened by Principle Component Analysis (PCA). Classification scheme illustrated in Fig.1 performs labeling the received SL-NSL samples based on the decision rule provided by the model learning block. Sleepiness recognition rates are reported per utterance after majority voting of sample labels. The system is detailed in the following section in which formulation of feature extraction is also given.

3. PERCEPTUAL FEATURES

We use a compact feature set that includes 9 descriptors. 3 out of 9 are calculated in Hz and 6 are computed in Bark. Table 1 lists these features and gives a brief description of each where more explanation is presented in the following.

Let $F[k_f, n]$ denote the Short Time Fourier Transform (STFT) of the audio sample where n is the index of time-frames and k_f is the frequency bin index. Corresponding masked spectral component is given as

$$F_e[k_f, n] = \left| F[k_f, n] \right| \; 10^{\frac{W[k_f]}{20}}. \qquad (1)$$

where the weighting function $W[k_f]$ denotes the outer middle ear frequency response at frequency bin k_f [12].

We can simply monitor sleepiness levels based on the variations in signal bandwidth, because perceived timbre, dullness and muffling effects in speech change according to the sleepiness level resulting in different perceptual bandwidths for NSL and SL audio. Therefore, we define the first feature, "10dB perceptual bandwidth (*BW1*)", as the frequency corresponding to the spectral component which exceeds the noise floor at least by 10 dB. The feature "5dB perceptual bandwidth (*BW2*)" is specified in a similar way where the spectral components exceeds the noise floor at least 5dB.

As a third feature, Average Harmonic Structure Magnitude

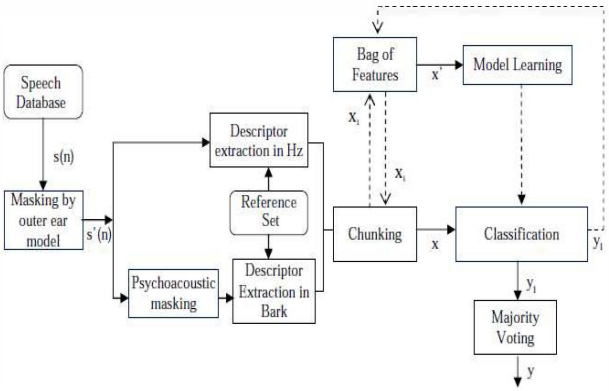

Fig. 1 The overall system block scheme. Dashed lines indicate steps carried out only during classifier training.

(*AHSM*), is defined in order to model the monotonous nature of NSL audio that is much more similar to a periodical signal with stable harmonics with respect to SL audio. On the other hand due to the intonation fluctuations of the speech in sleepy mode, the SL signal should not have a periodic structure as clear as the NSL signals. Conventionally fundamental frequency thus harmonic structure is estimated from the log spectrum of correlation function of audio signal [4, 14]. Unlike these methods we use the correlation of sleepiness differences through critical bands instead of time domain audio signal itself. Furthermore, in sleepiness detection, the general absence of a single valid audio measure for each person makes it necessary to acquire a wide variety of features including subjective self-assessment measures of sleepiness state. To overcome this difficulty, we aimed to learn the sleepiness level differences with respect to a reference in both perceptual spectral bands as well as the temporal domain. Hence first the outer ear weighted energy differences between SL/ NSL audio and the reference set are computed through the critical bands. Then the correlation of the energy differences through the critical bands is obtained. Fundamental frequency is estimated from the log spectrum of the correlation function. Average value of the fundamental frequencies estimated for successive *Y* audio frames is reported as *AHSM*. This is referred as chunking in Fig.1. The idea behind performing chunking is to make the sleepiness levels of speech more tractable. Hence *Y* is specified as long enough where the sleepy audio signal can be considered stationary. We set the length of audio frames to 43ms with 50% overlapping and *Y* is set to 70 frames.

Moreover, it is shown that the emotional differences (variances) in audio are more discriminative than the data itself thus can be used to enrich the discrimination capability of extracted speech features [11]. To lay over this approach on a practical basis we make use of a *reference* concept to distinguish sleepiness levels with respect to another. Hence 6 out of 9 of our features, including *AHSM*, reflect the

content variations of observed audio samples from the reference audio set. Also the subjective nature of the SL-NSL discrimination forces us to employ a reference which alleviates the effect of this subjectivity. As it is detailed in Section 4, the reference audio set is chosen in such a way that to highlight between class variations. Data belonging to test speaker is not included in the reference set, thus there is no need for a priori information about the test speaker.

Table 1 List of features used for sleepiness detection.

Low level descriptors calculated in Hz	
Average harmonics structure magnitude(*AHSM*)	Average of the fundamental frequencies estimated from the log spectrum of the correlations of sleepiness differences
10dB perceptual bandwidth (*BW1*)	The highest frequency component which exceeds the noise floor by at least 10 dB.
5dB perceptual bandwidth (*BW2*)	The highest frequency component which exceeds the noise floor by at least 5 dB.
Low level descriptors calculated in Bark	
Average number of non-sleepy blocks (*ANSB*)	Expected number of non-sleepy blocks within a time interval.
Normalized sleepiness level difference (*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.
Normalized Spectral Envelope Difference(*NSED1*)	Normalized envelope variations of the un-smeared SL/NSL pitch patterns from the reference within the successive frames for each critical band.
NSED2	Average of *NSED1* over all critical bands
NSED3	The temporal average of *NSED1* through successive *Y* audio frames.
Overall loudness of the frames (*OLF*)	Sum across all critical bands of outer ear weighted loudness values of an audio frame.

The un-smeared excitation pattern $E_s[k,n]$ is computed for each critical band of each audio frame by smearing the spectral energy over the frequency as in Eq.(2) [12],

$$E_s[k,n] = \left(\sum_{k=0}^{N_c - 1} P_e[k,n] S_{dB}(i,k,n,P_e[k,n])^{0.4} \right)^{\frac{1}{0.4}}, \quad (2)$$

where N_c denotes the number of critical bands and is set to 109 according to PEAQ [12]. In Eq.(2) $P_e[k,n]$ be the Bark representation of the outer ear weighted energy given by Eq.(1) and is referred conventionally as the loudness (pitch pattern) computed at critical band k, where k_f in Hz is replaced by k. $S_{dB}(i,k,n,P_e)$ denotes the spreading function of the band i for an energy component at the band k. In order to track temporal changes encountered in SL-NSL speech, we also perform time domain spreading to monitor the pitch variations over the frames, again in each critical band. Let $\overline{E}_{der}[k,n]$ denote the envelope changes of the $E_S[k,n]$ described by Eq(2). Eq.(3) model the envelope changes within a critical band over successive frames.

$$\overline{E}_{der}[k,n] = a \quad \overline{E}_{der}[k,n \quad 1] + (1 \quad a) \left| E_S[k,n]^{0.3} \quad E_S[k,n \quad 1]^{0.3} \right| \quad (3)$$

We normalize the differences between the envelope changes of SL-NSL data as in Eq.(4) to calculate a new feature $NSED1$ where the parameter β controls the minimum difference thus normalization.

$$NSED1[k,n] = \frac{\left| NSE_{SL/NSL}[k,n] \quad NSE_R[k,n] \right|}{+ \; NSE_{SL/NSL}[k,n]} \quad (4)$$

We derive two statistical descriptors from the normalized spectral envelope difference described in Eq.(4). Basically, the first descriptor ($NSED2$) is calculated by taking the average of normalized differences over all z=109 Bark scales. The temporal average of the normalized differences through successive Y audio frames yields the second statistical descriptor ($NSED3$).

We compute a new feature, normalized sleepiness level difference (*NSD*) as in Eq.(5), where Z = 109 denotes the number of critical bands, n refers to the audio frame number, and $M[k,n]$ adaptively masks the low frequency components to highlight the high frequency bands of SL-NSL audio.

$$NSD = 10 \log_{10} \frac{1}{Y} \sum_{n=1}^{Y} \left(\frac{1}{Z} \sum_{k=0}^{Z-1} \frac{P_{e_{SL/NSL}}[k,n] \quad P_{e_R}[k,n]}{M[k,n]} \right). \quad (5)$$

NSD enable us monitoring variations between the pitch patterns over the critical bands.

We use summation of loudness through critical bands as another discriminative feature. To enclose the hearing model, we compute the loudness as the excitation pattern normalized by the internal noise of ear (E_{IN}) as in Eq.(6), which is also different from conventional formulations.

$$L_{total}[n] = \sum_{k=0}^{Z-1} L[k,n] = \sum_{k=0}^{Z-1} \frac{E[k,n]}{E_{IN}[k]} \quad (6)$$

Since the nature of NSL audio tends to have higher excitation pattern peaks in comparison to SL, the feature average number of non-sleepy blocks (*ANSB*) provides a measure for the occurrence of high excitation levels through successive frame groups analyzed in bark scale, therefore improves the accuracy of sleepiness detection. To calculate the *ANSB* within a time interval, we use a probabilistic approach that estimates the number of frames in which the excitation level difference remains over a threshold [11].

4. TEST RESULTS

We have performed the sleepiness detection tests on the SLC data corpus [6] used in the Speaker State Challenge [5] to compare our performance with the existing systems. The SLC data includes 9089 utterances, which features 21 hours

of speech recordings of 99 subjects. The sampling rate of speech is 16 kHz. According to the data used for training and test stages, test scenarios are named as *Train vs Develop* and *Train+Develop vs Test* as in [5]. Number of utterances used for the training and test are respectively 3366 and 2915 for *Train vs Develop*. For *Train+Develop vs Test,* we have 6281 and 2808 utterances, respectively. The sleepiness detection rates achieved by the SVM and LVQ classifiers are reported for SL and NSL speech. Both of these classifiers are supervised thus a training scheme has been applied before classification. We used Weka libSVM toolbox and LVQ toolbox [15].

Table 2 reports the recall rates of SL data (RR_{SL}), NSL data (RR_{NSL}), and unweighted accuracy on average (UA), for the both scenarios. It can be concluded that the SVM provides higher recognition rates compared to the LVQ. Gender based recognition rates listed at the last three columns of Table 2 reports higher detection rates for males compared to females.

In order to investigate impact of individual features, extensive tests are performed by attribute evaluation tools of Weka. It is observed that six of the features, namely *NSD*, *NSED3*, *ANSB*, *BW2*, *AHSM*, and *OLF*, have dominance on the sleepiness monitoring. Table 3 reports the performance achieved by 6 and 9 features on SLC for *Train vs Develop*. It can be concluded that none of the features are redundant. The improvement becomes recognizable on the detection of SL speech and up to 10% increase is observed on the UA detection rates when all of the features are used.

Table 4 reports the sleepiness detection performance achieved by the proposed system compared to the existing ones. IS2011 Winner refers the highest scores reported by the Interspeech 2011 Speaker States Challenge participants [8] where the features of openEAR are used. IS2011 SSC denote the highest baseline performance declared in [5] and the results are obtained by the openEAR. It can be seen from Table 4 that the SVM with perceptual features achieves the highest detection rates for both of the test cases. Even though the number of reference utterances used for training is small, the SVM provides higher accuracy compared to the IS2011 Winner and IS2011 SCC. It can be concluded that UA rates achieved by our system with SVM are increased from 80% to 90% when the number of reference utterances is increased from 4 to 57 (large-ref) at the training stage. It is concluded during our work that the samples from opposite classes have to be included in the reference set at the training stage. For the reported results, number of codewords learned by BoF are 4K (out of 77K feature vectors) and 8K (out of 77K feature vectors) for the test scenarios *Train vs Develop* and T*rain+Develop vs Test*, respectively. After learning the codewords by BoF, the ref set size at test stage is reduced. Nevertheless, at the test stage the computational complexity of SVM with RBF kernel is $O(M)$ where M is the size of feature vectors [16]. Knowing that M is equal to 9 while it is in the order of thousand in comparable systems, our computational complexity at the test stage will be much more lower even though the training complexity can be considered comparable when the size of reference set is large. Hence the performances achieved by two different classifiers confirm the perceptual feature set can be efficiently used for sleepiness detection.

Table 2 SL-NSL recognition rates for utterances.(%).

Train (3366 utter) *vs Develop* (2915 utter)						
Classifier	RR_{SL}	RR_{NSL}	UA	M	F	UA
SVM	88.5	94.6	91.6	92.8	90.9	91.9
LVQ	83.6	87.5	85.5	90.4	83.6	87.0
Train+Develop vs Test (2808 utter)						
SVM	79.9	80.1	80.0	89.40	76.40	82.90
LVQ	63.2	78.7	71.0	77.9	67.8	72.9

Table 3 Impact of features on sleepiness detection (%).

	Train vs Develop					
	6 dominant features			All Features		
Classifier	RR_{SL}	RR_{NSL}	UA	RR_{SL}	RR_{NSL}	UA
SVM	73.8	93.2	83.5	89.1	97.2	93.2
LVQ	75.7	82.9	79.3	83.6	87.6	85.6

Table 4 Overall performance obtained on the SLC data compared to the existing systems (%).

	Train vs Develop			*Train + Develop vs Test*		
	RR_{SL}	RR_{NSL}	UA	RR_{SL}	RR_{NSL}	UA
SVM (large-ref)	96.7	96.0	96.3	94.2	87.6	90.9
SVM	89.1	97.2	93.2	79.9	80.1	80.0
IS2011 Winner [5]	60.3	75.7	68.0	64.2	79.1	71.6
IS 2011 SSC [8]	NA	NA	67.3	NA	NA	70.3

4. CONCLUSIONS

Unlike the existing systems that rely on phonetic speech features, we propose a sleepiness detection scheme that integrates psychoacoustic and temporal masking into feature extraction. The perceptual code words learned by BoF enable us to model the temporal and spectral content of sleepy data in a quasi-continuum space. Extensive tests on SLC data demonstrate that we recognizably improve the performance compared to the existing schemes in terms of the UA as well as the individual SL and NSL recall rates. As a result of the frame based feature extraction scheme, rather than mostly used segment based techniques, the developed method does not require a pre-segmentation stage and can be easily adopted to online processing.

5. REFERENCES

1. T. Nwe, L. Li, H., M. Dong, Analysis and detection of speech under sleep deprivation, in *Proc. of INTERSPEECH*, vol. 9, pp. 17-21, 2006.
2. J. H. Yang, Z. Mao, L. Tijerina, T. Pilutti, J. F. Coughlin, E. Feron, Detection of driver fatigue caused by sleep deprivation, *IEEE Trans. on Systems Man and Cybernetics*, vol. 39, no. 4, 2009.
3. J. Krajewski, A. Batliner, M. Golz, Acoustic sleepiness detection - Framework and validation of a speech adapted pattern recognition approach, *Behavior Research Methods*, vol. 41, pp. 795–804, 2009.
4. F. Eyben, M. Wollmer, and B. Schuller, openEAR– Introducing the Munich Open-Source Emotion and Affect Recognition Toolkit, in *Proc. of the Affective Computing and Intelligent Interaction (ACII)*, 2009.
5. B. Schuller, A. Batliner, S. Steidl, F. Schiel, and J. Krajewski., "The INTERSPEECH 2011 Speaker State Challenge," in *Proc. INTERSPEECH,* pp. 3201–3204, 2011.
6. J. Krajewski, The Center of Interdisciplinary Speech Science, Univ. of Wuppertal, Germany.
7. K. Kaida, M. Takahashi, T. Akerstedt, A. Nakata, Y. Otsuka, T. Haratani ve K. Fukasawa, Validation of the Karolinska Sleepiness Scale Against Performance and EEG Variables, *Clinical Neurophysiology,* pp. 1574-1581, 2006.
8. D. Huang, S. S. Ge, Z. Zhang,"Speaker State Classification Based on Fusion of Asymmetric SIMPLS and Support Vector Machines," in *Proc* of *INTERSPEECH,* pp.3301-3304, 2011.
9. J. Krajewski, S. Schnieder, M. Golz, A. Batliner, B. Schuller, "Applying multiple classifiers and nonlinear dynamics feature for detecting sleepiness from speech," *Journal of Neurocomputing*, vol. 84, pp. 65-75, 2012.
10. M. E. Ayadia, M. S: Kamelb, F. Karrayb, "Survey on speech emotion recognition; features, classification schemes, and databases," *Pattern Recognition*, vol. 44(3), pp. 572-587, 2011.
11. M. C. Sezgin, B. Gunsel, G. K. Kurt, "Perceptual Audio Features for Emotion Detection," *EURASIP J. on Audio, Speech, and Music Processing*, vol. 16, 2012.
12. International Telecommunications Union Recommendation BS.1387-1, *Method for objective measurements of perceived audio quality*. (2000).
13. G. Csurka, C. Dance, L.X. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," in *Proc. of ECCV Workshops*, 2004.
14. H. G. Kim, N. Moreau, T. Sikora, MPEG-7 Audio and Beyond, John Wiley and Sons Ltd., England (2005).
15. H Witten, E Frank, Data Mining: Practical Machine Learning Tools and Techniques With Java Implementations, Morgan Kaufman (2000).
16. C. J. C. Burges, "A tutorial on Support Vector Machines for Pattern Recognition," Kluwer Academic Publishers, Boston (1999).