A speech adapted pattern recognition framework for measuring energetic states using low level descriptors and functionals

J. Krajewski[1], R.Wieland[1], D. Sommer[2], and M. Golz[2]
[1]*Work and Organizational Psychology, Univ. of Wuppertal, Wuppertal, Germany, krajewsk@uni-wupertal.de*
[2]*Neuro Computer Science, and Signal Processing, Univ. of Applied Sciences, Schmalkalden,Germany, m.golz@fh-sm.de*

Abstract

This paper describes a general framework for detecting affective and energetic states based on prosody, articulation and speech quality related speech characteristics. The advantages of this realtime approach are that obtaining speech data is non obstrusive, free from sensor application and calibration efforts. The main part of the feature computation is the combination of frame level based speech features (LLD) and high level contour descriptors (functionals) resulting in over 8,500 features per speech sample. In general the measurement process follows the speech adapted steps of pattern recognition: (a) recording speech, (b) preprocessing (segmenting speech units of interest), (c) feature computation (using perceptual and signal processing related features, as e.g. fundamental frequency, intensity, pause patterns, formants, cepstral coefficients), (d) dimensionality reduction (filter and wrapper based feature subset selection, (un-)supervised feature transformation), (e) classification (e.g. SVM, LDA, K-NN, MLP, HMM, GMM classifier; metaclassification e.g. bagging, boosting, voting, stacking), and (f) evaluation (e.g. 10-fold cross validation, leave-one-sample-out). The validity of this approach is briefly discussed by summarizing the empirical results of a sleep deprivation study.

Keywords

Computational Intelligence, Acoustic Features, Pattern Recogntion, Sleepiness Detection, Speech Emotion Recognition

Introduction

Many efforts have been reported in the literature for measuring biosignal based energetic states [1-3]. The real-time detection systems mainly focus on (a) oculomotoric data (eye blinking, eyelid movement, and saccade eye movement) [4], (b) EEG data [5] and (c) behavioral expression data (gross body movement, head movement, mannerism, and facial expression) [6] in order to characterize the energetic state. Apart from these promising advances in analysing facial and gestural expressivity, there has been recently renewed interest in vocal expression and speech analysis. Mainly this fact is promoted by the progress in speech science. Using voice communication as an indicator of sleepiness would have the following advantages: obtaining speech data is non-invasive, non obstrusive, free from sensor application and calibration efforts, "hands- and eyes-free", and most important speech data is omnipresent in many daily life situations.

In this paper we describe a speech adapted pattern recognition framework in order to measure energetic states. Our attention is focused particularly on the processing step of feature computation. The rest of this paper is organized as follows: In Section 2 computing high level contour descriptor features is explained. The general speech adapted pattern recogntion framework is provided in Section 3, a brief summary of sleepiness detection results is given in Section 4.

High level contour descriptors as acoustic features

Frame level features (low-level descriptors)

Acoustic features can be divided according to auditive-perceptual concepts in prosody (pitch, intensity, rhythm, pause pattern, speech rate), articulation (slurred speech, reduction and elision phenomenons), and speech quality (breathy, tense, sharp, hoarse, modal voice) related features. Another distinction can be drawn from using signal processing categories as time, frequency or phase space domain features. Our approach prefers the fusion of perceptual features and purely signal processing and speech recognition based features without any known auditiv-perceptual pendants. Typical frame level based acoustic features used in emotion speech recognition and audio processing [7-9] are fundamental frequency (acoustic pendant to pitch; maximum of the autocorrelation function), intensity, duration of voiced/unvoiced segments, harmonics-to-noise ratio, position and bandwidth of 6 formants (resonance frequencies of the vocal tract depending strongly on its actual shape), 16 linear predictive coding coefficients, 12 mel frequency cepstrum coefficients ("spectrum of the spectrum"), and 12 linear frequency cepstrum coefficients (without the perceptual oriented transformation into the mel frequency scale).

Contour descriptors (functionals)

After splitting the speech signal into 10 ms frames and computing the above mentioned frame level features (Low-Level Descriptors, LLD; see [10]), the values of each frame level feature are connected to contours. This procedure results in about 57 speech feature contours (e.g. the fundamental frequency contour, the bandwidth of formant 4 contour etc.), which are joined by their first and second derivates (delta and delta-delta contours). Furthermore these 171 speech feature contours are described by elementary statisitics (linear moments, values and positions of extrema, quartiles, ranges, length of time periods beyond threshold values, regression coefficients, etc.), and spectral descriptors (spectral energy of low frequency bands vs. high frequency bands, etc.) resulting in about 8,500 high-level speech features (171 speech contours x 50 functionals).

Speech adapted pattern recognition framework

The measurement process follows the speech adapted steps of pattern recognition (see Table 1): (a) recording speech, (b) preprocessing (segmenting speech units of interest), (c) feature computation (using perceptual and signal processing related features, as e.g. fundamental frequency, intensity, pause patterns, formants, cepstral coefficients), (d) dimensionality reduction (filter and wrapper based feature subset selection, (un-)supervised feature transformation), (e) classification (e.g. SVM, LDA, K-NN, MLP, HMM, GMM classifier; metaclassification e.g. bagging, boosting, voting, stacking), and (f) evaluation (e.g. 10-fold cross validation, leave-one-sample-out).

Empirical validation results

We conducted a within-subject sleep deprivation design (N = 17; 8.00 p.m to 4.00 a.m). During the night of sleep deprivation a well proved, standardised self-report sleepiness measure, the Karolinska Sleepiness Scale (KSS) was used every hour just before the speech recordings. The verbal material consisted of a German phrase: "Ich suche die Friesenstraße" ["I´m searching for the Friesen Street"]. The sentence was taken from simulated communication with a driver assistance system. The participants recorded other verbal material at the same session, but in this article we focus on the material described above. For training and classification purposes the records were further divided in two classes: sleepy (SS) and non sleepy (NSS) with the boundary value KSS ≥ 6. (46 samples NSS, 22 samples SS). During the night, the subjects were confined to the laboratory and supervised throughout the whole period. Between sessions, they remained in a room, watched DVD, and talked. Non caffeinated beverages and snacks were available ad libitum.

Table 1. *Processing steps and alternative specifications of the pattern recognition based speech acoustic measurement. The here used realizations of the pattern recognition specifications are printed in italics.*

Pattern recognition step	Specification
Recording	
Source of verbal material	*Human to human,* human to machine; *monolgue* vs. dialogue situation; speech databases (e.g. AEC, Sympafly, IFA, EMO-DB)
Speaking style	Vowel phonation, isolated words, connected speech, *read speech,* fluent speech, spontenous speech
Speech segment	Different vowels, different consonant type (fricative, stop, glide), consonant cluster, syllables, words, intonation unit, *phrases*
Recording situtation	Noisy vs. *noise subdued enviroment* (e.g. driving with open window vs. laboratory recording); rough vs. clean speech signal quality (e.g. telephon call, radio communication vs. clean recording in 22.05 kHz, 16 bit)
Preprocessing	
Segmentation	*Manual,* (semi-)automatical segmentation (e.g. MAUS system) of speech signal in units of interest
Framing	Size of frames (10-*20 ms*), degree of overlapping, window function (*hamming*, hanning)
Feature extraction	
Low level descriptors	*Fundamenal frequency, intensity-, harmonics-to-noise ratio, formant position and bandwidth (F1-F6), LPC, MFCC, LFCC, voiced speech segments, unvoiced speech segments*
Functionals	*linear moments, extrema values and positions, quartiles, ranges, length o time periods beyond threshold values, regression coefficients); spectral descriptors (spectral energy of low frequency bands vs. high frequency bands);* state space feature (largest lyapunov coefficient); automatic feature generation (genetic algorithms)
Normalization	*Individual speaker specific baseline correction*, age/ gender specific normalization; noise filtering
Dimensionality reduction	
Feature Subset Selection	*Filter based subset selection (correlationsfilter); wrapper-based subset selection* (forward selection, backward elimination, *genetic algorithm selection*)
Feature Transformation	*Unsupervised (principle component analysis)*; supervised (linear discriminant analysis)
Classification	
Classifier choice	Classification granularity (*binary* or multiple class prediction)*; 1-nearest neighbour, multi-layer perceptron, support vector machine,* linear discriminant analysis, hidden markov model, decision tree, gaussian mixture model; parameter optimization;
Metaclassification	Bagging, boosting, *voting*, stacking
Validation	
Evaluation strategy	*10-fold cross validation,* leave-one-speaker-out (LOSO), multiple-hold-o

The averaged accuracy rates (ratio correctly classified samples through all samples) of three different classifiers (1-nearest neighbour, multi-layer perceptron, support vector machine) were over 80%. Due to the hypothezised sleepiness related physiological changes in cognitive speech planning, respiration, phonation, articulation, and radiation, the results for the reported classification performance were largely as could be expected. This is consistent with previous sleepiness related findings, that suggest an association of acoustic features [11,12] with sleepiness.

References

1. Kollias, S., Amir, N., Kim, J., Grandjean, D. (2004). *Description of potential exemplars: Signals and Signs of Emotion*. HUMAINE Human-Machine Interaction Network on Emotions.
2. Galley, N. (2007). On the way to a sleep warner. "Monitoring sleep and sleepiness with new sensors within medical and industrial applications", SENSATION IP 2nd International Conference, 4-5 June 2007, Chania, Greece.
3. Golz, M., Sommer, D., Mandic, D. (2006). Establishing a gold standard for drivers microsleep detection. *Proc. Int Conf Monitoring Sleep and Sleepiness (SENSATION 2006*, Basel, Switzerland), 29.

4. Caffier, P.P. (2002). The spontaneous eye-blink as sleepiness indicator in patients with obstructive sleep apnoea syndrome-a pilot study, *Sleep Medicine* **2**, 155-162.
5. Sommer, D., Chen, M., Golz, M., Trunschel, U., Mandic, D. (2005). Fusion of state space and frequency domain features for improved microsleep detection. W. Dutch et al. (Eds.): Int Conf Artifical Neural Networks (ICANN 2005), (pp. 753-759). Springer: Berlin.
6. Vöhringer-Kuhnt, T., Baumgarten, T. Karrer, K. & Briest, S. (2004). Wierwille's method of driver drowsiness evaluation revisited. *Proceeding of International Conference on Traffic & Transport Psychology*.
7. Schuller, B. (2006). *Automatische Emotionserkennung aus sprachlicher und manueller Interaktion.* [Automatic emotion recognition from verbal and manual interaction]. Dissertation, Technische Universität München.
8. Batliner, A., Steidl, S., Schuller, B, Seppi, D., Laskowski, K, Vogt, T., Devillers, L., Vidrascu, L., Amir, N., Kessous, L & Aharonson, V. (2006). Combining Efforts for Improving Automatic Classification of Emotional User States. In T. Erjavec & J.Z. Gros (Eds.), *Language Technologies, IS-LTC 2006*, (pp. 240-245). Ljubljana, Slovenia: Infornacijska Druzba.
9. Mierswa, I. , Morik, K. (2005). Automatic feature extraction for classifying audio data. Kluwe.
10. Vlasenko, B., Schuller, B., Wendemuth, A., Rigoll, G. (2007). Combining Frame and Turn-Level Information for Robust Recognition of Emotions within Speech. *Proceedings of Interspeech*, 2249-2252.
11. Nwe, T.L., Li, H**.,** Dong, M. (2006). Analysis and Detection of Speech under Sleep Deprivation. *Proceeding of Interspeech* 2006, 17-21.
12. Krajewski, J., Kröger, B. (2007). Using prosodic and spectral characteristics for sleepiness detection. *Proceedings of Interspeech,* 1841-1844.