Speech Emotion Recognition – Measuring Driver States from Voice Characteristics

Jarek Krajewski, Elmar Nöth und Anton Batliner

Keywords: Speech Emotion Recognition, Emotional States, Signal Processing, Pattern Recognition

Abstract

Detecting emotions and driver states in a broad sense (e.g. annoyance, anger, and fatigue) can help to enhance acceptance and comfort of these systems themselves as well as accident related safety. Adapting the system output, e.g., in speech interfaces of Driver Assistance Systems to the actual emotional and energetic driver states might enhance the acceptance of this empathic, emotional-intelligent communication due to its improved naturalism. In addition, comfort and comprehensiveness might be improved if the system's output is adapted to the driver's actual e.g. stress-impaired attentional and cognitive resources. Furthermore, the detection of distracted driver states (e.g. pain, anger, emotional intensive conversation) could be beneficial from the viewpoint of safety concerns. Using voice communication as an indicator of emotion would have the following advantages: obtaining speech data is non-obtrusive, free from sensor application, calibration efforts, and robust against climatic environmental conditions. Several validation studies analysing e.g. basic emotions, fatigue, stress are presented in this paper reaching recognition rates for 2-class problems of about 80-90 %. Finally, current limitations and future demands (as e.g. freely accessible speech databases) for an efficient progress in speech emotions recognition are discussed.

Measuring Driver States

Many efforts have been reported in the literature for automatic measurement of emotional states. These systems mainly focus on (a) electropysiological data (e.g. EEG: Golz, Sommer, Holzbrecher & Schnupp, 2007; Hönig, Batliner & Nöth, 2007), and (b) behavioural expression data (gross body movement, head movement, mannerism, and facial expression; Horlings, Datcu & Rothkrantz, 2008) in order to characterize the user state. But these electrode- (EOG/EEG reaching 15% error rate for fatigue detection; Sommer & Golz, 2005) or video-based instruments still do not fulfil the demands of an everyday life measurement system. The major drawbacks are (a) a lack of robustness against environmental and individual-specific variations (e.g. bright light, wearing correction glasses, angle of face or being of Asian race) and (b) a lack of comfort and longevity due to electrode sensor application.

In contrast to these electrode- or video-based instruments, the utilization of voice communication as an indicator for emotional states could match the demands of everyday life measurement. Contact free measurements as voice analysis are non-obtrusive (not interfering with the primary driving task) and favourable for emotion detection, since an application of sensors would cause annoyance, additional stress and often impairs working capabilities and mobility demands. In addition, speech is easy to record even under extreme environmental conditions (bright light, high humidity and temperature), requires merely cheap, durable, and maintenance free sensors and most importantly, it utilizes already existing communication system hardware. Furthermore, speech data is omnipresent in many professional driver settings. Given these obvious advantages, the renewed interest in computational demanding analyses of vocal expressions has been enabled just recently by the advances in computer processing speed (Batliner, Steidl, Schuller et al, 2006; Juslin & Scherer, 2005; Owren & Bachorowski, 2007; Scherer, Johnstone & Klasmeyer, 2003; Schuller, Batliner, Seppi et al., 2007).

Acoustic Features

The acoustic speech emotion recognition research is mainly based on phonetics, general signal processing and computational intelligence research (e.g. Batliner, Steidl, Schuller et al., 2006). How emotions are expressed in the voice can be analyzed acoustically by measuring the characteristics of the speech wave form radiating from nostrils and mouth. Accordingly, acoustic features can be divided referring to auditive-perceptual concepts into prosody (pitch, intensity, rhythm, pause pattern, and speech rate), articulation (slurred speech, reduction and elision phenomena), and speech quality (timbre: breathy, whispery, tense, sharp, hoarse, or modal voice). A popular approach prefers the fusion of purely signal processing based features without any known auditive-perceptual correlates and perceptual-acoustic features as e.g. the speech intensity (loudness), fundamental frequency (pitch), voiced/unvoiced duration (rhythm) and several voice quality measures. Suitable measures for the voice quality category whisperiness are level of the first harmonic, bandwidth of the first formant, large amount of interharmonic noise, and steep spectral tilt. The spectrum of tense voices is flatter than that of lax voices. Creaky voice is characterized by a relatively flat spectrum, i.e. relatively strong upper harmonics. Finally, harsh and rough voices are characterized by a large amount of jitter, shimmer, and interharmonic noise.

Typical acoustic features which implicitly cover the categories mentioned above and are used frequently in Speech Emotion Recognition (SER) (see Table 1) are (a) fundamental frequency, (b) intensity, (c) duration of voiced/unvoiced speech segments, (d) harmonics-to-noise ratio (HNR), (e) formant positions (F1-F7), (f) formant bandwidths (Fbw1-Fbw7), (g) mel frequency cepstrum coefficients (MFCCs), (h) linear frequency cepstrum coefficients (LFCCs), (i) linear predictive coding coefficients, and (j) spectral features derived from the long term average spectrum (LTAS).

Table 1. Basic acoustic feature contours (frame-level descriptors).

Frame level based feature	Description
Fundamental frequency (F0)	acoustic equivalent to pitch; rate of vocal fold vibration; maximum of the autocorrelation function; models prosodic structure; speech melody indicator
Energy	models intensity, based on the amplitude in different intervals; average squared amplitude within a predefined time segment; stressing structure
Harmonics-to-noise Ratio (HNR)	spectral energy in voiced vs. unvoiced segments; ratio between harmonic and aperiodic signal energy; breathiness indicator
Formant position (F1-F7)	resonance frequencies of the vocal tract (VT) depending strongly on its actual shape; represent spectral maxima, and are known to model spoken content and speaker characteristics; influenced by lower jaw angle, tongue body angle, tongue body horizontal location, tongue tip angle, tongue tip horizontal location, relative lip height, lip protrusion, velum height
Formants bandwidth (Fbw1-Fbw7)	model VT shape and energy loss of speech signal due to VT elasticity (yielding wall effect), viscoeleasticity of VT tissue or heat conduction induced changes of air flow (jet streams, turbulences); width of the spectral band containing significant formant energy (- 3 dB threshold)
Duration of voiced-unvoiced segments	models temporal speech rhythm aspects as speech rate and pause structure
Mel frequency cepstrum coefficients (MFCCs)	"spectrum of the spectrum"; have been proven beneficial in speech emotion recognition, and speech recognition tasks; homomorphic transform with equidistant band-pass-filters on the Mel-scale; Holistic and decorrelated representation of spectrum
Linear frequency cepstrum coefficients (LFCCs)	similar to MFCC but without the perceptually oriented transformation into the mel frequency scale; emphasize changes or periodicity in the spectrum, while being relatively robust against noise

Linear Predictive Coding (LPC)	Provides an accurate and economical representation of the envelope of the short-time power spectrum. One of the most powerful speech coding analysis techniques providing very accurate estimates of speech parameters and is known as being relatively efficient for computation at the same time
Long Term Average Spectrum (LTAS)	averages out formant information; giving general spectral trends; relative amount of energy within predefined frequency bands; speech quality

The speaker's emotional state expresses itself through prosodic-phonetic information as pitch, intensity, speaking rate, duration, articulation and voice quality (see Figure 2). Furthermore, this prosody related information conveyed by speech can be enriched by paralinguistic information, which contains short, emotional non-speech expressions, comprising both clear non-speech sounds and interjections with a phonemic structure (cf. affective bursts, Schroeder, 2001). Laughing, sniffs, sobs, or morphing speech into a cry ("noooo …"), swearing, yawning, interjections (e.g. "wow!"), deep breathing as sound of relief are typical examples of these paralinguistic events, which might be used for speech emotion recognition (Truong & van Leeuwen, 2007). Furthermore, emotional information is carried in the human speech over the information channel of syntax and semantics (lexical content). Syntactic channels include information as part-of-speech categories (POS; e.g. number of nouns, non-inflected adjectives), other syntax-related information (n-grams, bag of words, elaborated vs. simple construction, many vs. few word per sentence; many vs. few negations; 1st person singular use, tentative words, present vs. past tense), Linguistic Inquiry and Word Count (LIWC; e.g. categories as anger words, inclusive words, social processes), and MRC psycholinguistic features (concreteness of words, imageability; cf. Mairesse, Walker, Mehl & Moore, 2007). Moreover, the analysis of semantic channels might include: topic selection (e.g. self-focused, pleasure vs. problem talk, single vs. multiple topics, strict selection vs. think out loud ; bag of words features), and conversational behaviour characteristics (listen vs. initiate conversation, back-channel behaviour, formal style). The major drawbacks of SER using syntactic or semantic information are the difficulties related to an automatic identification of these categories. In addition, a word tagger could help to count semantic and syntactic categories which might provide information about sleepiness as well (e.g. part-of-speech classes; Batliner Nutt, Warnke et al., 1999).

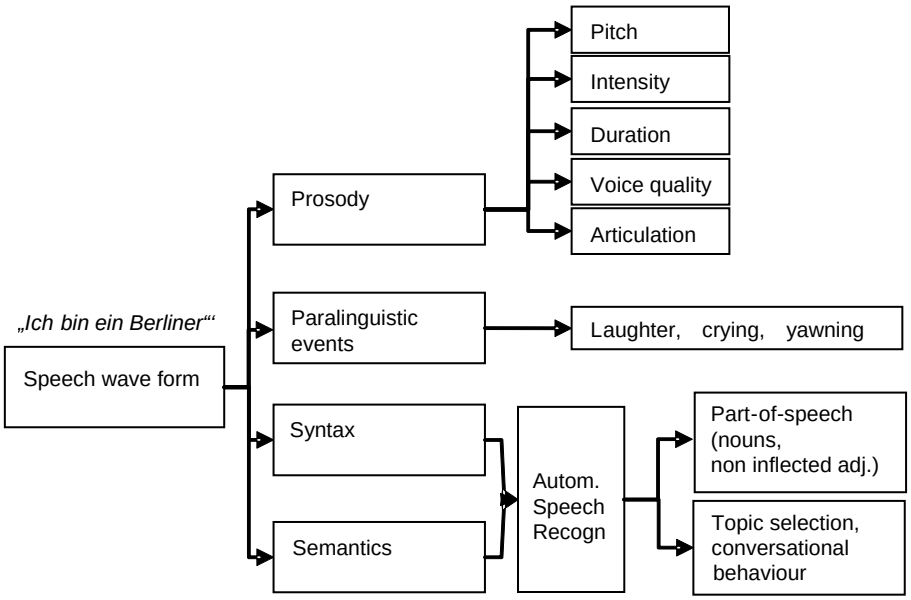

Figure 2: Multiple information channels for detecting emotional states from speech

Empirical Validation Results in Speech Emotion Recognition (SER)

SER can be useful is a broad range of application scenarios, as e.g. automatic interactive voice response system, in which the system recognizes impatient, angry or frustrated users. Appropriate validation scenarios of SER must adapt exactly to their application context to capture the speech material, context factors, and actual speaker variety of this certain application. In general, emotion research can be viewed as going from the analysis of acted speech to more realistic application scenarios. Emotions occurring in realistic, non-prompted spontaneous speech seem to be more difficult to recognize compared to acted speech (Batliner et al., 2003). Nevertheless, SER engines are able to detect emotional states in naturalistic settings (speaker-independent, spontaneous speech) within a range of about 80-90% CL for 2-class problems and 4-class problems with a performance of > 60% CL (Ang, Dhillon, Krupski, Shriberg & Stolcke, 2002; Batliner & Huber, 2007; Devillers et al., 2005; Krajewski & Kröger, 2007), which seems close to the performance of a single labeller (Steidl, Levit, Batliner, Nöth & Niemann, 2005). Due to large inter-individual differences in speaker characteristics and the way speakers employ different acoustic features in different ways, higher classification rates could be achieved by personalized, speaker-dependent classification (Batliner & Huber, 2007). Further fine-tuning of classification performances can be obtained by pre-selection of merely prototypical cases, close talk microphones in a quiet surrounding, a size-restriction of the used vocabulary (Batliner, Steidl, Hacker, Nöth & Niemann, 2005).

A different aspect of SER is the question, which features might be sensitive to emotional speech (Juslin & Laukka, 2003; Juslin & Scherer, 2005; Owren & Bachorowski, 2007; Scherer, 2003; Scherer, Johnstone & Klasmeyer, 2003). Referring to the relevance of single acoustic features within the context of classifiers, the most relevant feature classes for SER seem to be Mel Frequency Cepstral Coefficients (MFCC), duration features, then pitch variability (jitter), and intensity features (cf. Schuller, Batliner, Seppi et al., 2007). Voice quality features and basic pitch features as F0 mean, max, or range are less important. MFCC features which are standard features in word recognition are probably successful because they are coarse but robust measures and less error prone (e.g. for octave jumps).

Limitations and Future Work

A short sketch sums up possible starting points for future research facing primarily the challenge of improving the measurement precision of SER:

(a) Recording (Corpus Engineering): Collecting emotional speech samples from different types of speakers and speaking styles would provide a broadly diversified learning data set to compute submodels for different responder groups and different confounder constellations (e.g. a submodel for emotional speakers with head cold). Thus, collecting different, naturalistic emotional speech samples (since acted data is not suitable for real life purposes) within an open source speech database would provide the infrastructural research background that enhances further progress in acoustic sleepiness analysis. Emotional speech databases as e.g. FAU-Aibo Emotion Corpus (Batliner, Steidl & Nöth, 2008) could serve as a model for this kind of open source speech corpora.

(b) Preprocessing: Finding and segmenting emotional sensitive phonetic units (phones, consonant clusters, or syllables in different word and phrasal unit positions) could improve the overall detection rates, especially in free and spontaneous speech with unrestricted vocabulary size and domain. Automatic speech recognition could serve for this purpose. In addition, a word tagger could help to count semantic and syntactic categories which might provide information about sleepiness as well (e.g. part-of-speech classes; Batliner et al., 1999).

(c) Feature extraction: The signal processing features derived from state space domains as, e.g., average angle or length of embedded space vectors, and recurrence quantification analyses

should be computed, and feature transformation applied. Moreover, evolutionary feature generation methods could be used to find further features (Pachet & Roy, 2009; Schuller, Reiter & Rigoll, 2006). In addition, different normalization procedures could be applied as, e.g., computing speaker specific baseline corrections not on high-level features, but on duration adapted low-level contours. Additionally, hierarchical functionals (Schuller, Wimmer, Mösenlechner, Kern & Rigoll, 2008) might help to identify emotional sensitive subparts within a speech segment. Moreover, automatic speech recognition output could serve as source for identifying emotion-sensitive phones.

(d) Dimensionality reduction: For finding the optimal feature subset, further supervised filter based subset selection methods (e.g. information gain ratio) or supervised wrapper-based subset selection methods should be applied (e.g. sequential forward floating search, genetic algorithm selection). Another method for reducing the dimensionality of the feature space are unsupervised feature transformation methods (e.g. PCA network, nonlinear autoassociative network, multidimensional scaling, independent component analysis, Sammon map, enhanced Lipschitz embedding, SOM) or supervised feature transformation methods (e.g. LDA).

(e) Classification: A rejection class should be added to the classification task. Furthermore, future work on SER could consider more metaclassifier methods as bagging, boosting, or stacking including exhaustive parameter optimizations. Dividing between male and female classification models might be as promising as applying maximum-likelihood Bayes classifiers or fuzzy membership indexing.

References

Ang, J., Dhillon, R., Krupski, A., Shriberg, E. & Stolcke, A. (2002). Prosody-based automatic detection of annoyance and frustration in human-computer dialog. *Proc. of ICSLP. 341-344.*

Batliner, A., Nutt, M., Warnke, V., Nöth, E., Buckow, J., Huber, R. & Niemann, H. (1999). Automatic annotation and classification of phrase accents in spontaneous speech. *Proceedings of the European Conference on Speech Communication and Technology, 6,* 519–522.

Batliner, A., Steidl, S. & Nöth, E. (2008). Releasing a thoroughly annotated and processed spontaneous emotional database: The FAU Aibo emotion corpus. *Proceedings of LREC on Corpora for Research on Emotion and Affect,* Marrakesh, 28-31.

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.

Devillers, L., Vidrascu, L. & Lamel, L. (2005). Challenges in real-life emotion annotation and machine learning based detection Source. *Neural Networks, 18,* 407-422.

Golz, M., Sommer, D., Holzbrecher, M. & Schnupp, T. (2007). Detection and Prediction of Driver's Microsleep Events. In RS4C (Eds.), *Proceedings 14th International Conference Road Safety on Four Continents.* Bangkok, Thailand.

Hönig, F., Batliner, A. & Nöth, E. (2007). Fast Recursive Data-driven Multi-resolution Feature Extraction for Physiological Signal Classification. In J. Hornegger, et al. (Eds.): *3rd Russian-Bavarian Conference on Biomedical Engineering, Erlangen, 47-52.*

Horlings, R., Datcu, D. & Rothkrantz, L. J. M. (2008). Emotion Recognition using Brain Activity. *ACM International Conference Proceedings Series, 374.*

Juslin, P. N. & Laukka, P. (2003). Communication of emotions in vocal expression and music performance: Different channels, same code? *Psychological Bulletin, 129,* 770-814.

Juslin, P. N. & Scherer, K. R. (2005). Vocal expression of affect. In J. A. Harrigan, R. Rosenthal & K. R. Scherer (Eds.), *The new handbook of methods in nonverbal behavior research* (pp. 65-135). New York: Oxford University Press.

Krajewski, J. & Kröger, B. (2007). Using prosodic and spectral characteristics for sleepiness detection. In H. van Hamme & R. van Son (Eds.), *Interspeech Proceedings* (pp. 1841-1844). Antwerp: University Antwerp.

Mairesse, F., Walker, M., Mehl, M. & Moore, R. (2007). Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of Artificial Intelligence Research, 30,* 457-500.

Owren, M. J. & Bachorowski, J.-A. (2007). Measuring emotion-related vocal acoustics. In J. Coan & J. Allen (Eds.). *Handbook of emotion elicitation and assessment, (*pp. 239-266). New York: Oxford University Press.

Pachet, F. & Roy, P. (2009). Analytical Features: a knowledge-based approach to audio feature generation. *Journal on Audio, Speech, and Music Processing*

Scherer, K. R. (2003). Vocal communication of emotion: A review of research paradigms. *Speech Communication, 40,* 227-256.

Scherer, K. R., Johnstone, T. & Klasmeyer, G. (2003). Vocal expression of emotion. In R. J. Davidson, K. R. Scherer, H. Goldsmith (Eds.). *Handbook of the Affective Sciences* (pp. 433-456). New York and Oxford: Oxford University Press.

Schuller, B., Batliner, A., Seppi, D., Steidl, S., Vogt, T., Wagner, J., Devillers, L., Vidrascu, L., Amir, N., Kessous, L. & Aharonson, V. (2007). The relevance of feature type for the automatic classification of emotional user states: Low level descriptors and functionals.*Proceedings of Interspeech, 8,* 2253-2256.

Schuller, B., Wimmer, M. Mösenlechner, L., Kern, C. & Rigoll, G. (2008). Brute-forcing hierarchical functionals for paralinguistics: A waste of feature space? *Proceedings International Conference on Acoustics, Speech, and Signal Processing, 33,* 4501–4504.

Schuller, B., Reiter, S. & Rigoll, G. (2006). Evolutionary Feature Generation in Speech Emotion Recognition. *Proceedings ICME,* 95-98.

Sommer, D., Chen, M., Golz, M., Trunschel, U. & Mandic, D. (2005). Fusion of state space and frequency domain features for improved microsleep detection. In W. Dutch et al. (Eds.), *Proceedings International Conference Artificial Neural Networks (ICANN 2005),* (pp. 753-759). Springer: Berlin.

Steidl, S., Levit, M., Batliner, A., Nöth, E. & Niemann, H. (2005).Of All Things the Measure is Man" - Classification of Emotions and Inter-Labeler Consistency. In IEEE (Eds.), *Proceedings of ICASSP - International Conference on Acoustics, Speech, and Signal Processing, 1,* (pp. 317-320). Conference Management Services: Texas.

Truong, K. P. & van Leeuwen, D.A. (2007). Automatic discrimination between laughter and speech. *Speech Communication, 49,* 144–158.

Vlasenko, B., Schuller, B., Wendemuth, A. & Rigoll, G. (2007). Combining frame and turnlevel information for robust recognition of emotions within speech. *Proceedings of Interspeech, 8,* 2249-2252.