Emotion recognition in the noise applying large acoustic feature sets

Björn Schuller, Dejan Arsic, Frank Wallhoff, Gerhard Rigoll

Angaben zur Veröffentlichung / Publication details:

Schuller, Björn, Dejan Arsic, Frank Wallhoff, and Gerhard Rigoll. 2006. "Emotion recognition in the noise applying large acoustic feature sets." In *Speech Prosody: 3rd International Conference, Dresden, May 2 - 5, 2006; abstract book*, edited by Rüdiger Hoffmann and Hansjörg Mixdorff, 128. Dresden: TUDpress. https://www.isca-archive.org/speechprosody_2006/schuller06_speechprosody.html.

Emotion Recognition in the Noise Applying Large Acoustic Feature Sets

Björn Schuller, Dejan Arsić, Frank Wallhoff & Gerhard Rigoll

Institute for Human-Machine Communication
Technische Universität München, Germany
{Schuller; Arsic; Wallhoff; Rigoll}@tum.de

Abstract

Speech emotion recognition is considered mostly under ideal acoustic conditions: acted and elicited samples in studio quality are used besides sparse works on spontaneous field-data. However, specific analysis of noise influence plays an important factor in speech processing and is practically not considered hereon, yet. We therefore discuss affect estimation under noise conditions herein. On 3 well-known public databases - DES, EMO-DB, and SUSAS - effects of post-recording noise addition in diverse dB levels, and performance under noise conditions during signal capturing, are shown. To cope with this new challenge we extend generation of functionals by extraction of a large 4k hi-level feature set out of more than 60 partially novel base contours. Such comprise among others intonation, intensity, formants, HNR, MFCC, and VOC19. Fast Information-Gain-Ratio filter-selection picks attributes according to noise conditions. Results are presented using Support Vector Machines as classifier.

1. Introduction

Affective computing is broadly expected to form one of the most important factors in future man-machine-interaction and multi-media processing [1]. First products arise at the time and a variety of commercially promising application scenarios exists reaching from call-center supervision to public transport surveillance. Speech is considered one of the most reliable and further more comfortable modalities to automatically estimate a person's emotion [2], especially as no wiring is needed, and a person may control the amount of emotion shown. However, most use-cases are quite demanding in view of independence of the person, the spoken content, and especially of signal capturing and transmission conditions with respect to noise. Yet, experiments on acoustic affect recognition are mostly carried out on clean speech studio recordings of acted emotions. Only sparse works use spontaneous or elicited samples [3] out of the field. As the community is about to face real life conditions more studies with respect to the alluded influences therefore need to be carried out. In particular no analysis of noise effects has been fulfilled at the time [4]. We therefore aim at investigation of such herein, which is generally an important factor in speech processing tasks.

As we want to provide a broad basis of results, we chose three public databases. On two of them, originally recorded in studio quality, we carry out experiments in diverse noise levels. The third one serves as reference, as it was originally recorded in heavily noisy environment. We analyze general performance influence as well as emotion confusion in noise conditions.

As a basic approach to cope with noise we extend the feature basis of comparable works to 4k acoustic features. This high number of attributes serves as a starting point for subsequent feature selection in diverse noise levels. Likewise we can find suited attributes according to the noise situation.

The paper is structured as follows: firstly, databases are introduced in section 2; afterwards artificial addition of noise is described in section 3. Next, we discuss systematic generation of large feature sets followed by reduction to relevant ones and classification methods in sections 4 and 5. The contribution ends with results, discussion and conclusion in sections 6 and 7.

2. Databases

In order to provide results on public corpora in view of comparability we firstly decided for the popular Danish Emotional Speech Corpus (DES) [5]. In this database the four emotions *anger, joy, sadness,* and *surprise* of the MPEG-4 set plus *neutrality* are contained. Four professional Danish actors, two of them female, simulated the word *yes* and *no*, 9 sentences and two text passages in each emotion. We split the text passages into single sentences and thereby obtain 414 phrases in total. The set was recorded in 16 bit, 20 kHz PCM-coding in a sound studio. 20 test-persons, 10 of them female, reclassified the samples in a perception test. Their recognition rate was between 59% and 80% with an average resembling 67.3%.

As second dataset to observe inter-set behavior we chose the Berlin Emotional Speech Database (EMO-DB) [6], which consists of 816 phrases in total. The emotion set resembles the *"big six"* set of the MPEG-4 standard consisting of *anger, disgust, fear, joy, sadness* and *surprise*, besides an exchange of surprise in favor of *boredom, happiness* instead of *joy,* and added *neutrality*. 10 German sentences of emotionally undefined content have been acted in these emotions by 10 professional actors, 5 of them female. Throughout perception tests by 20 probands 488 phrases have been chosen that were classified as more than 60% natural and at least 80% clearly assignable. The database is recorded in 16 bit, 16 kHz under studio noise conditions. 84.3% recognition rate is reported for a human perception test.

Finally we selected the Speech Under Simulated and Actual Stress (SUSAS) database [7] as a reference for original recording in noisy field. It consists of five domains, encompassing a wide variety of stresses and emotions. We decided for the 3,949 actual stress speech samples recorded in dual-tracking workload or subject motion fear tasks, as acted samples are already covered by DES and EMO-DB in this work. 4 male speakers in an US Apache helicopter cockpit and 7 speakers, 3 of them female, in roller coaster and free fall actual stress situations are contained in this set. Two different stress conditions have been collected within the helicopter situation: *medium stress* during warm-up, where the helicopter is on the ground but running, and *high stress* during flight, where pilots are flying hover, turn and other maneuvers while speaking. Within the further samples also

Proc. Speech Prosody
ISCA

neutral samples, *fear during freefall* and *screaming* are contained as classes. Likewise a total of five emotions, respectively speaking styles, are covered. SUSAS samples are constrained to a 35 words vocabulary of short aircraft communication commands. All files are sampled in 8 kHz, 16 bit. The recordings are partly overlaid with heavy noise and background ground controller over-talk. However, this resembles realistic acoustic recording conditions, as also given in many related scenarios of interest as automotive speech interfaces or the mentioned public transport surveillance.

3. Noise Addition

In order to provide results in diverse noise levels we decided for controlled white noise addition to the samples of the DES and EMO-DB. Such additive noise overlay is a common practice in general speech processing tasks, especially in speech and speaker recognition. While this approach does not take noise influences on the speaking style as Lombard effect into account, it already forms a reasonable basis and partly covers scenarios as microphone mismatch, cellular/phone channels or voice coding effects [7]. The SNR level is chosen in relative terms with respect to the level of each individual affective speech signal. Likewise an SNR of ∞ dB resembles clean speech, while 0 dB represents signal and noise mixed at even level. We investigate the effect of noise addition in 5 dB steps starting from clean speech, moving on to slightly noise overlaid 25dB SNR and terminating at heavily overlaid -10 dB, where the original sample is hardly understandable for a human listener. However, we aim only at investigation of acoustic feature analysis in search for emotional cues. Linguistic analysis [8] is left aside in this work, as automatic speech recognition is heavily affected by noise, as well [7], and an extra study will be needed hereon.

As mentioned the SUSAS database was already recorded in heavy noise, and serves as a reference for actual noise condition and its influence on speaking style herein. Additional noise overlay is therefore spared.

4. Large Feature Set Construction

In former works we showed the higher performance of derived functionals - a common approach in speech emotion recognition ever since [10] - instead of full-blown contour classification by dynamic classification as Hidden-Markov-Models [9]. Likewise we use systematic generation of functionals f out of time-series F by means of descriptive statistics:

$$f : F \rightarrow \mathbb{R} \qquad (1)$$

Firstly, selected base-contours, respectively Low-Level-Descriptors (LLD), are calculated well known to carry information about the emotional state of a speaker. The original sampling frequency and quantization of the databases is kept, and each 10 ms a 20 ms frame is extracted by weighting with a Hamming window-function. Aiming at coverage of prosodic, articulatory and voice quality aspects, estimated feature contours contain log frame energy, pitch based on autocorrelation (ACF) in the time-domain and Dynamic Programming (DP) to minimize deviations on a global level, pitch epochs, harmonics-to-noise ratio based on ACF, formant bandwidth, position and amplitude of the first 5 formants based on LPC, polynomial roots and DP. Further more jitter and shimmer size of larynx excitation points is calculated. Thereby jitter is a measure of pitch- and shimmer one of amplitude-perturbation on a cycle to cycle basis. For spectral analysis 16 MFCCs, and spectral flux, spectral centroid, as well as spectral roll-off based on 8,192 linear DFT-spectral coefficients and polynomial dB-correction in accordance to human perception, is extracted. Dominant harmonics in the spectrum are tracked in 47 chromatic semitone intervals within human voice range by summing over three successive partials. Finally, 19 Voc19 coefficients are obtained by JSRU-style 19-channel filter-bank analysis using two second-order section Butterworth band-pass filters spaced as in [11]. Energy smoothing is done at 50Hz.

As a side comment it may be mentioned that within the MPEG-7 standard partly similar LLDs are defined. The in the ongoing suggested methodology of time-series analysis may therefore applied on these, as well, to estimate emotion based hereon.

The contours are subsequently smoothed by symmetrical moving average low-pass filtering with a window size of three. Likewise we are less prone to noise throughout the calculation, as most feature contours as pitch or formants are prone to errors, already. Successively, speed (∂) and acceleration (∂^2) are derived as further LLDs for each basic contour in order to model temporal behavior.

Afterwards a total of 20 phrase-wise derived hi-level functionals by means of descriptive statistics per contour is computed. These are linear momentums of the first four orders, namely mean, centroid, standard deviation, Skewness and Kurtosis, as well as quartiles, ranges, extrema, extrema positions, zero-crossing-rates, and roll-off-points. Likewise roughly 4k acoustic features are obtained in total. The aim here is too build a broad feature basis for the subsequent feature selection process, throughout which is learned which attributes to prefer in which noise condition. Thereby almost redundant features are justified at this stage. However, a pre-selection process by expert knowledge of unsuited combinations is fulfilled to keep complexity within reasonable limits already prior to automatic selection.

5. Feature Selection and Classification

Besides lower extraction time-effort, reduction of features also often leads to higher classification performance, as the classifier is confronted with less complexity, if only redundant information is spared. In former works [8] we demonstrated the high effectiveness of wrapper-based search which aims at optimization of a set as a whole. However, due to the unusually high dimensionality in this domain of 4k entries in the original feature vector we apply fast Information Gain Ratio based feature selection (IGR-FS) herein. In this filter-reduction single highly relevant attributes are found by their entropy [11]. Likewise, ranking of attributes is independent of the classifier. However, we use a closed feed-back loop in order to find the optimal number of the ranked features in accordance with the target classifier.

Dealing with classification, the optimal learning method is broadly discussed [2, 4], similar to the optimal features. In [8] we made an extensive comparison on the EMO-DB database including besides Support Vector Machines (SVM) Naïve Bayes, k-Nearest Neighbors, Decision Trees, and Neural Nets. Further more we investigated construction of more powerful classifiers by means of meta-classification as MultiBoosting or Stacking. However, in our experiments both on DES and

EMO-DB SVM prevailed. We therefore concentrate on these herein.

SVM - kernel machines - are well known in the machine learning community and highly popular at the time due to their remarkable performance and generalization capabilities. The latter result from the applied structural risk minimization oriented training. Generally speaking, SVM base on a linear distance-function classification of a two-class problem. However, multi-class strategies as one-vs.-one, layer-wise decision or one-vs.-all exist. Discriminative training is achieved by optimal placement of a separation hyperplane under the precondition of linear separability. As a consequence, a dual optimization problem has to be solved throughout training process. Likewise, SVM can be seen as an analogon to electrostatics: Thereby a training sample corresponds to a charged conductor at a certain space, the decision function an electrostatic potential function and the learning target function the Coulomb energy. The precondition of linear separability is approached by a transformation of the original feature space via a kernel function that has to be found empirically.

In this evaluation we use a couple-wise one-vs.-one decision for multi-class discrimination and a polynomial kernel found optimal throughout test cycles. For more details on classifiers refer to [11].

To conclude the feature extraction, selection and classification process so far, figure 1 provides a general overview.

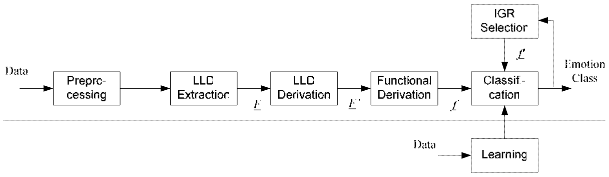

Figure 1: *Overview speech emotion recognition.*

6. Results and Discussion

Since datasets are sparse in the field of speech emotion recognition, an evaluation method which allows for training disjunctive test on all samples seems favorable. As a general evaluation mean we therefore choose the popular *j*-fold stratified cross validation (SCV).

In the first table we show the effect of white noise addition in various dB levels for the databases DES and EMO-DB (EMO). All tests have been carried out using the full 4k identical feature set, in order to focus on the direct effect of noise addition.

Table 1: *Accuracies at selected SNR levels, databases DES and EMO-DB using SVM in a 10-fold SCV and 4k features.*

Acc. [%]	∞ **dB**	**20 dB**	**10 dB**	**0 dB**	**-5 dB**	**-10 dB**
DES	68.7	61.4	54.2	53.1	51.7	49.2
EMO	86.7	83.4	83.3	78.6	72.3	67.2

A significant decrease in accuracy can be observed for each 5dB step considering a significance level of α = 0.05 and applying a Student's-t test. However, only selected steps are shown in the table due to space limitations.

In the next table we show effects of feature selection by IGR-FS on the accuracy for the databases DES and EMO-DB. *Best N* thereby stands for the reduced feature set at the optimum size as described in section 5 in view of accuracy. Reduction always helps to increase performance, but feature sets differ largely at the various noise levels and for the diverse databases. Here again, we show only the extrema of clean speech and highly noise overlaid -10 dB SNR samples.

Table 2: *Accuracies at selected SNR levels, database DES using SVM in a 10-fold SCV and best N features by IGR-FS.*

Acc. [%]	**DES**	**DES best N**	**EMO**	**EMO best N**
∞ **dB**	68.7	**74.5**	86.7	**86.9**
-10 dB	49.2	**54.9**	67.2	**71.1**

Next, confusions are presented in total sample number for the clean speech case (table 3) and for the worst investigated case of -10 dB SNR (table 4). Due to space limitations we decided only for the EMO-DB corpus. However, similar behavior can be reported for DES.

Table 3: *Confusions clean speech, database EMO-DB using SVM in a 10-fold SCV and best N features by IGR-FS.*

Classified True [#]	**A**	**D**	**F**	**H**	**N**	**S**	**B**
Anger	**119**	0	2	6	0	0	0
Disgust	2	**32**	1	0	3	0	0
Fear	4	1	**47**	2	1	0	0
Happiness	14	0	2	**40**	2	0	0
Neutral	0	0	1	0	**69**	2	7
Sadness	0	0	0	0	1	**50**	2
Boredom	0	0	1	2	8	1	**67**

Table 4: *Confusions at -10 dB SNR level, database EMO-DB using SVM in a 10-fold SCV and best N features by IGR-FS.*

Classified True [#]	**A**	**D**	**F**	**H**	**N**	**S**	**B**
Anger	**110**	4	0	13	0	0	0
Disgust	7	**21**	3	3	1	0	3
Fear	2	2	**43**	5	1	0	2
Happiness	15	5	5	**31**	2	0	0
Neutral	0	3	2	0	**48**	2	23
Sadness	0	0	0	0	3	**44**	6
Boredom	0	2	0	1	19	7	**50**

The typical confusion pairs within similar emotional activity level are found in the clean speech case: anger is likely confused with happiness, even though of opposite valence, and neutrality with boredom (marked grey in the tables). Besides these two couples statistically mostly uninteresting confusions are found. Generally, these confusions seem due to the fact that activity level is more easily discriminated than valence. In order to manifest this we clustered the emotions of EMO-DB into active/passive and positive/negative. Thereby neutral, sadness and boredom are labeled passive, opposing all remaining as active. Respectively happiness is labeled positive opposing the remaining as negative. Within a 10-fold SCV with SVM and best N IGR-FS features 97.5% discrimination accuracy was observed in the first case, but only 92.4% in the second case.

This trend comes out even clearer for DES where 96.1% accuracy for active/passive opposes 82.6% for positive/negative discrimination.

Interestingly, confusion of these couples is raised in number of occurrence under severe noise conditions, as can be seen in table 4. Likewise, noise addition seemingly does not lead to random confusions.

If we take a closer look at single emotions, it can be seen that the highest absolute losses in accuracy on EMO-DB occur for neutrality and boredom, but also for disgust. Table 5 therefore shows label-wise accuracies for clean speech and the absolute loss if compared to -10 dB SNR level. Additionally, these rates are also shown for DES. Emotions not contained in the datasets are marked by "-". On both databases anger suffers least from noise addition, while sadness is recognized most easily in total. For DES this is true with a significant gap to the remaining emotions.

Table 5: *Emotion-wise accuracies for clean speech and absolute loss at -10 dB SNR levels, databases DES and EMO-DB using SVM in a 10-fold SCV and best N features by IGR-FS.*

Acc. [%]	DES ∞ dB	DES loss	EMO ∞ dB	EMO loss
Anger	77.6	8.2	93.7	7.1
Disgust	-	-	84.2	29.0
Fear	-	-	85.4	7.3
Happiness/Joy	58.1	19.8	69.0	15.5
Neutral	77.7	27.1	88.5	26.9
Sadness	86.9	21.4	94.3	11.3
Boredom	-	-	84.8	21.5
Surprise	72.2	21.5	-	-

On the real samples of the SUSAS database we finally achieve 77.8% correct recognition rate within 10-fold SCV and using SVM applying the full feature set. By IGR-FS reduction accuracy is boosted to impressive 84.9% in average. Neutrality is recognized with 76.0%, fear during freefall with 88.6%, medium stress with 82.2%, high stress with 90.6%, screaming with 97.9% accuracy. Neutral samples are exclusively confused with stress, and mostly with medium stress.

7. Conclusions

Within this work we showed effects of noise conditions for speech emotion recognition on two public databases. On DES the best guess accuracy resembles 20.0%, always choosing the maximum a-priori class 31.0%, and human perception 67.3% in average. Herein we obtained an outstanding maximum of 74.5% for clean speech, and 54.9% at -10 dB SNR level. For EMO-DB best guess resembles 14.3%, maximum a-priori selection 26.0%, human performance 84.3%. Remarkable 87.5% could be obtained for clean speech in this work, 71.11 at -10 db SNR level as a maximum. Likewise it can be summarized that automatic speech emotion recognition suffers as expected under noise influences, but an astonishingly high performance can be observed at severe noise levels compared to other speech processing tasks as speech or speaker recognition.

Tests on the spontaneous emotion database SUSAS which is recorded in strong field noise conditions also speak for a robust recognition already without extra noise cancellation effort. Impressive 84.9% correct recognition rate for 5 emotions can be reported compared to 20% best guess.

Feature selection improves accuracy already when employing fast IGR-FS. However, feature sets strongly vary with noise conditions. Likewise, optimal sets can be chosen according to the noise level to reduce complexity for the classifier by a concentration on attributes less prone to noise.

Future works aim at investigation of noise cancellation techniques known in speech processing as use of beam-forming through microphone arrays or adaptive noise filtering in the field of speech emotion recognition. Furthermore different specific noise types as channel and coding influences, overlaid speech, and distant talk shall be analyzed.

8. Acknowledgements

This work highly benefits from the contributions of the student researcher Thomas Mikschl.

9. References

[1] Shriberg, E., 2005. Spontaneous Speech: How People Really Talk And Why Engineers Should Care, *Proc. INTERSPEECH 2005*, ISCA, Lisbon, Portugal, 1781-1784.

[2] Cowie, R.; Douglas-Cowie, E.; Tsapatsoulis, N.; Votsis, G.; Kollias, S.; Fellenz, W.; Taylor, J. G., Jan. 2001. Emotion recognition in human-computer interaction, *IEEE Signal Processing magazine*, vol. 18, no. 1, 32–80.

[3] Batliner, A.; Seidl, S.; Hacker, C.; Nöth, E.; Niemann, H., 2005. Tales of Tuning – Prototyping for Automatic Classification of Emotional User States, *Proc. INTERSPEECH 2005*, ISCA, Lisbon, Portugal, 489-492.

[4] Pantic, M; Rothkrantz, L., Sep. 2003. Toward an Affect-Sensitive Multimodal Human-Computer Interaction, *Proceedings of the IEEE*, vol. 91, 1370-1390.

[5] Engberg, I. S.; Hansen, A. V, 1996. *Documentation of the Danish Emotional Speech Database DES*, Aalborg, Denmark.

[6] Burkhardt, F.; Paeschke, A.; Rolfes, M.; Sendlmeier, W.; Weiss, B., 2005: A Database of German Emotional Speech, *Proceedings of the INTERSPEECH 2005,* ISCA, Lisbon, Portugal, 1517-1520.

[7] Hansen, J.H.L.; Bou-Ghazale, S., 1997. Getting Started with SUSAS: A Speech Under Simulated and Actual Stress Database, *Proc. EUROSPEECH-97*, Rhodes, Greece, vol. 4, 1743-1746.

[8] Schuller, B.; Müller, R.; Lang, M.; Rigoll, G., 2005. Speaker Independent Emotion Recognition by Early Fusion of Acoustic and Linguistic Features within Ensembles. *Proc. INTERSPEECH 2005*, ISCA, Lisbon, Portugal, 805-809.

[9] Schuller, B.; Rigoll, G.; Lang, M., 2003. Hidden Markov Model-Based Speech Emotion Recognition, *Proc. ICASSP 2003*, IEEE, Hong Kong, China, vol. II, 1-4.

[10] Amir, N., Ron, S, 1998. Towards an automatic classification of emotions in speech. *Proc. 5th International Conference of Spoken Language Processing*, Sydney, Australia, 555–558.

[11] Holmes, J. N., 1980. The JSRU 19-channel Vocoder, *IEE Proceedings*, vol. 1, part F, 127.

[12] Witten, I. H.; Frank, E., 2000. *Data Mining, Practical machine learning tools with Java implementations*, Morgan Kaufmann, San Francisco, 133.