Frame vs. turn-level: emotion recognition from speech considering static and dynamic processing

Bogdan Vlasenko, Björn Schuller, Andreas Wendemuth, Gerhard Rigoll

Angaben zur Veröffentlichung / Publication details:

Vlasenko, Bogdan, Björn Schuller, Andreas Wendemuth, and Gerhard Rigoll. 2007. "Frame vs. turn-level: emotion recognition from speech considering static and dynamic processing." *Lecture Notes in Computer Science* 4738: 139–47. https://doi.org/10.1007/978-3-540-74889-2_13.

Frame vs. Turn-Level: Emotion Recognition from Speech Considering Static and Dynamic Processing

Bogdan Vlasenko[1], Björn Schuller[2], Andreas Wendemuth[1], and Gerhard Rigoll[2]

[1] Cognitive Systems, IESK, Otto-von-Guericke University, Magdeburg, Germany
[2] Institute for Human-Machine Communication, Technische Universität München, Germany
Bogdan.Vlasenko@e-technik.uni-magdeburg.de, Schuller@tum.de

Abstract. Opposing the pre-dominant turn-wise statistics of acoustic Low-Level-Descriptors followed by static classification we re-investigate dynamic modeling directly on the frame-level in speech-based emotion recognition. This seems beneficial, as it is well known that important information on temporal sub-turn-layers exists. And, most promisingly, we integrate this frame-level information within a state-of-the-art large-feature-space emotion recognition engine. In order to investigate frame-level processing we employ a typical speaker-recognition set-up tailored for the use of emotion classification. That is a GMM for classification and MFCC plus speed and acceleration coefficients as features. We thereby also consider use of multiple states, respectively an HMM. In order to fuse this information with turn-based modeling, output scores are added to a super-vector combined with static acoustic features. Thereby a variety of Low-Level-Descriptors and functionals to cover prosodic, speech quality, and articulatory aspects are considered. Starting from 1.4k features we select optimal configurations including and excluding GMM information. The final decision task is realized by use of SVM. Extensive test-runs are carried out on two popular public databases, namely EMO-DB and SUSAS, to investigate acted and spontaneous data. As we face the current challenge of speaker-independent analysis we also discuss benefits arising from speaker normalization. The results obtained clearly emphasize the superior power of integrated diverse time-levels.

1 Introduction

Apart from a few attempts to classify emotions within speech dynamically [1,2], current approaches usually employ static feature vectors derived on a turn or frame level. In [2] the latter has also been shown superior to dynamic modeling. This derives mostly from the fact, that by (usually statistical) functional application to the Low-Level-Descriptors (LLD) as e.g. pitch, energy, or spectral coefficients an important information reduction takes place, which avoids phonetic (respectively spoken-content) over-modeling. Yet, it is also considered received knowledge that thereby important temporal information is lost due to a high degree of abstraction. This led to first successful attempts to integrate information on diverse time levels [3-6].

Apart from this several works point at the high influence of emotional variability within speech on the recognition of speakers [7,8]. We therefore investigate how

reliably a state-of-the art speaker recognition engine using MFCC, Cepstral Mean Substraction (CMS), and Gaussian Mixture Models (GMM) can recognize emotions instead of speakers. As such processing operates on a per-frame basis, we finally use this to accomplish the initially introduced thought of combining different temporal layers for emotion recognition within speech.

For testing we will use two public databases providing acted and spontaneous samples of emotional speech.

The paper is structured as follows: Section 2 and 3 deal with frame- and turn-level analysis of speech with respect to emotion. In section 4 two optimization strategies, namely speaker normalization and feature space optimization, are discussed. Section 5 introduces the fusion of the two approaches. Finally, in the sections 6-8 data, results and conclusions are presented.

2 Frame-Level Analysis (FL)

We consider using a speaker recognition system to recognize emotion from speech in the first place. Likewise, instead of the usual task to deduce the most likely speaker (from a known speaker set) ω_k form a given sequence X of M acoustic observations x [9], we will recognize the current emotion. This is usually solved by a stochastic approach following eq. 1,

$$\omega_k = \underset{\Omega}{\operatorname{argmax}} P(\omega | X) = \underset{\Omega}{\operatorname{argmax}} \frac{P(X | \omega) P(\omega)}{P(X)}. \tag{1}$$

where P(X $|\omega$) is called the speaker acoustic model, P(ω) is the prior speaker information and ω speaker model given the set of reference models $\Omega=\{\omega_1, \ldots, \omega_N\}$.

The vectors in a sequence, X, are independent and identically distributed random variables. This allows to express P(X $|\omega$) as

$$P(X | \omega) = \prod_{t=1}^{M} p(x_t | \omega). \tag{2}$$

where P($x_t|\omega$) is the likelihood of single frame x_t given model ω. This is a fundamental equation of statistical theory and is widely used in speech and speaker recognition systems using frame level analysis.

A typical state-of-the-art system uses single state HMMs as speaker acoustic model, also known as GMMs. This state is associated with an emission-probability P(X|s) which for continuous variables x is replaced with its probability density function (PDF). These PDFs are realized using weighted sums of elementary Gaussian PDFs (Gaussian Mixtures, which leads to the name GMM).

A GMM is a weighted sum of N component densities and is given by the form

$$P(x | \Omega) = \sum_{i=1}^{N} c_i b_i(x). \tag{3}$$

where x is a M-dimensional random vector, $b_i(x)$, i=1,...,N, is the component density and c_i, i = 1,… , N, is the mixture weight. Each component density is a M-variate Gaussian function of the form

$$b_i(x) = \frac{1}{(2\pi)^{M/2} |\Sigma_i|^{1/2}} \cdot \exp\left(-\frac{1}{2} (x-\mu_i)^T \Sigma_i^{-1} (x-\mu_i) \right). \tag{4}$$

with mean vector μ_i and covariance matrix Σ_i. The mixture weights satisfy the stochastic constraint that

$$\sum_{i=1}^{N} c_i = 1. \tag{5}$$

The complete Gaussian mixture model is parameterized by the mean vectors, covariance matrices and mixture weights from all component densities. These parameters are collectively represented by the notation

$$\Omega = \{c_i, \mu_i, \Sigma_i\}, i = 1,..., N. \tag{6}$$

In our emotion recognition system, each emotion is represented by such a GMM and is referred to by its model Ω.

For a sequence of T test vectors $X = x_1, x_2, \cdots, x_T$, the standard approach is to calculate the GMM likelihood as in Eq. (2) which can be written in the log domain as

$$L(X \mid \Omega) = \log p(X \mid \Omega) = \sum_{i=1}^{T} \log p(x_t \mid \Omega). \tag{7}$$

The GMM parameters are estimated by the EM-algorithm using training material for other speakers and a number of 1 to 120 Gaussian mixtures to approximate the original PDFs [10]. However, we also consider multiple states, herein, as in some speaker recognition systems, to better model dynamics. These are trained accordingly.

Speech input is processed using a 25ms Hamming window, with a frame rate of 10ms. As in typical speaker recognition we employ a 39 dimensional feature vector per each frame consisting of 12 MFCC and log frame energy plus speed and acceleration coefficients. Cepstral Mean Substraction (CMS) and variance normalization are applied to better cope with channel characteristics. The priors are chosen as an equal distribution among emotion classes.

3 Turn-Level Analysis (TL)

In order to represent a typical state-of-the-art emotion recognition engine operating on a turn-level, we use a set of 1,406 acoustic features basing on 37 Low-Level-Descriptors (LLD) as seen in table 1 and their first order delta coefficients [11]. These 37x2 LLDs are next smoothened by Low-pass filtering with an SMA-filter.

Opposing the formerly introduced dynamic modeling, such systems derive statistics per speaker turn by a projection of each uni-variate time series, respectively LLD, X onto a scalar feature x independent of the length of the turn. This is realized by use of a functional F, as depicted in eq. 8.

$$F : X \rightarrow x \in \mathbb{R}^1. \tag{8}$$

19 functionals are applied to each contour on the turn-level covering extremes, ranges, positions, first four moments and quartiles as seen in table 1. Note, that three functionals are related to position, known as duration in traditional phonetic terminology, as their physical unit is msec.

For classification we use Support Vector Machines (SVM) with linear Kernel and 1-vs.-1 multi-class discrimination. One could consider the use of GMM here, as well. Yet, SVM have proven the preferred choice in many works to best model static acoustic feature vectors [11].

Table 1. Overview of applied Low-Level-Descriptors and functionals for turn-wise analysis

Low-Level-Descriptors (2x37)	**Functionals (19)**
Pitch	Mean
Energy	Standard Deviation
Envelope	Zero-Crossing-Rate
Formant 1-5 Amplitude	Quartile 1
Formant 1-5 Bandwidth	Quartile 2
Formant 1-5 Frequency	Quartile 3
MFCC Coefficient 1-16	Quartile 1 - Minimum
Harmonics-to-Noise-Ratio HNR	Quartile 2 - Quartile 1
Shimmer	Quartile 3 - Quartile 2
Jitter	Maximum - Quartile 3
Delta Pitch	Centroid
Delta Energy	Skewness
Delta Envelope	Kurtosis
Delta Formant 1-5 Amplitude	Maximum Value
Delta Formant 1-5 Bandwidth	Relative Maximum Position
Delta Formant 1-5 Frequency	Minimum Value
Delta MFCC Coefficient 1-16	Relative Minimum Position
Delta Harmonics-to-Noise-Ratio	Maximum Minimum Range
Delta Shimmer	Position of 95% Roll-Off-Point
Delta Jitter	

4 Optimization Strategies

Next, two optimization strategies are considered: First, speaker normalization (SN) by feature normalization with the whole individual speaker context. Second, feature-space optimization by correlation-based exclusion of highly correlated features (FS).

We investigate the benefits of speaker normalization, as we intend to analyze emotion independent of the speaker, herein. SN is thereby realized by a normalization of each feature x by its mean and standard deviation for each speaker individually. Thereby the whole speaker context is used. This has to be seen as an upper benchmark for ideal situations, where a speaker could be observed with a variety of emotions. Yet, it is not necessary to know the actual emotional state of observed utterances at this point.

As a high number of features is used throughout static modeling, feature space optimization seems a must in view of performance and real-time-capability. In order to optimize a set of features rather than combining attributes of single high relevance, we use a correlation-based analysis, herein. Thereby features of high class-correlation and low inter-feature correlation are kept [12]. This does not employ the target-classifier in the loop. Likewise it mostly reduces correlation within the feature space rather than evaluation of the benefit of single attributes. Still, this leads to a very compact representation of the feature space, which usually leads to an improvement of accuracy while reducing feature extraction effort at the same time.

5 Time-Level Combination

So far the two individual approaches to emotion recognition based on information processing directly on the frame-level, or on a higher turn-level, have been introduced. In order to fuse these two approaches it seems beneficial to keep utmost amounts of information for the final decision process. However, an early feature fusion is not feasible, as frame-level processing results in a dynamic number of frames. We therefore decided to include final GMM scores within the static acoustic feature vector. The process of speaker normalization and feature space optimization is extended to the likewise obtained new feature vector x'. Fig. 1 depicts the overall processing flow from an input audio file via the two streams to the final result. Overall feature selection having the GMM scores within the space reveals their high importance, as they are kept among high ranks.

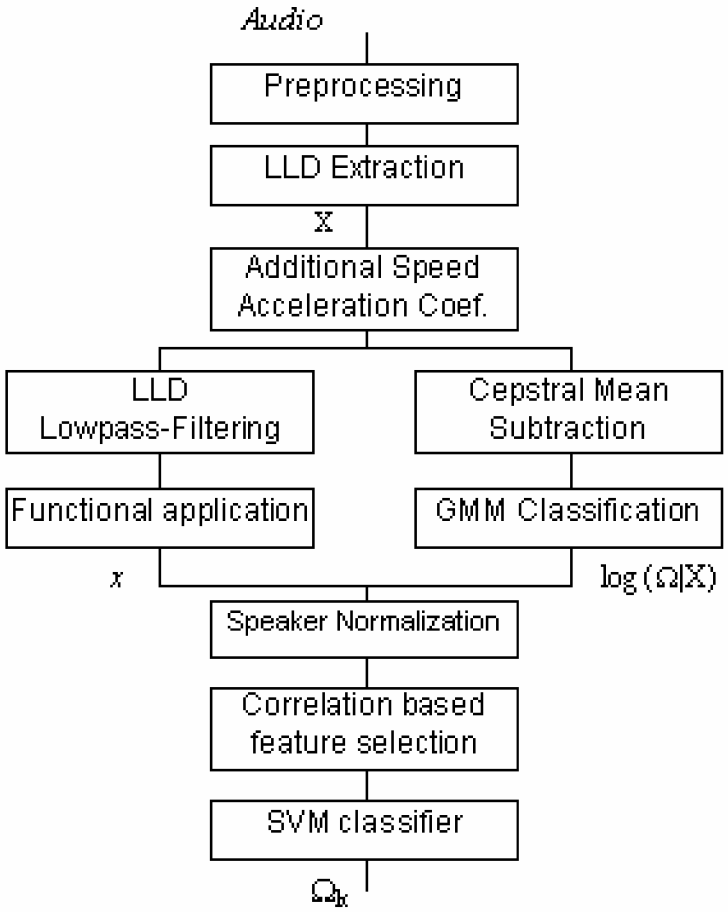

Fig. 1. Processing flow for the combined frame- and turn-level analysis

6 Acted and Spontaneous Data

To demonstrate effectiveness of each single approach and the fusion on acted and spontaneous data, we decided first for the popular studio recorded Berlin Emotional Speech Database (EMODB) [13], which covers the 'big six' emotion set (MPEG-4) besides boredom instead of surprise, and added neutrality. 10 (5f) professional actors speak 10 German emotionally undefined sentences. 494 phrases are marked as min.

60% natural and min. 80% assignable by 20 subjects. 84.3% accuracy are reported for a human perception test.

Second, we selected the Speech Under Simulated and Actual Stress (SUSAS) database [14] as a reference for spontaneous recordings. As additional challenge speech is partly masked by field noise. It consists of five domains, encompassing a wide variety of stresses and emotions. We decided for the 3,663 actual stress speech samples recorded in subject motion fear and stress tasks, as acted samples are already covered by EMODB in this work. 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: medium stress, and high stress. Within the further samples also 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 over-talk. However, this resembles realistic acoustic recording conditions, as also given in many related scenarios of interest such as automotive speech interfaces or public transport surveillance.

7 Experimental Results

Results are presented for each modeling technique individually (TL and FL), and for the combination of these two. Thereby the effects of speaker normalization SN and feature space optimization FS as described are shown, too.

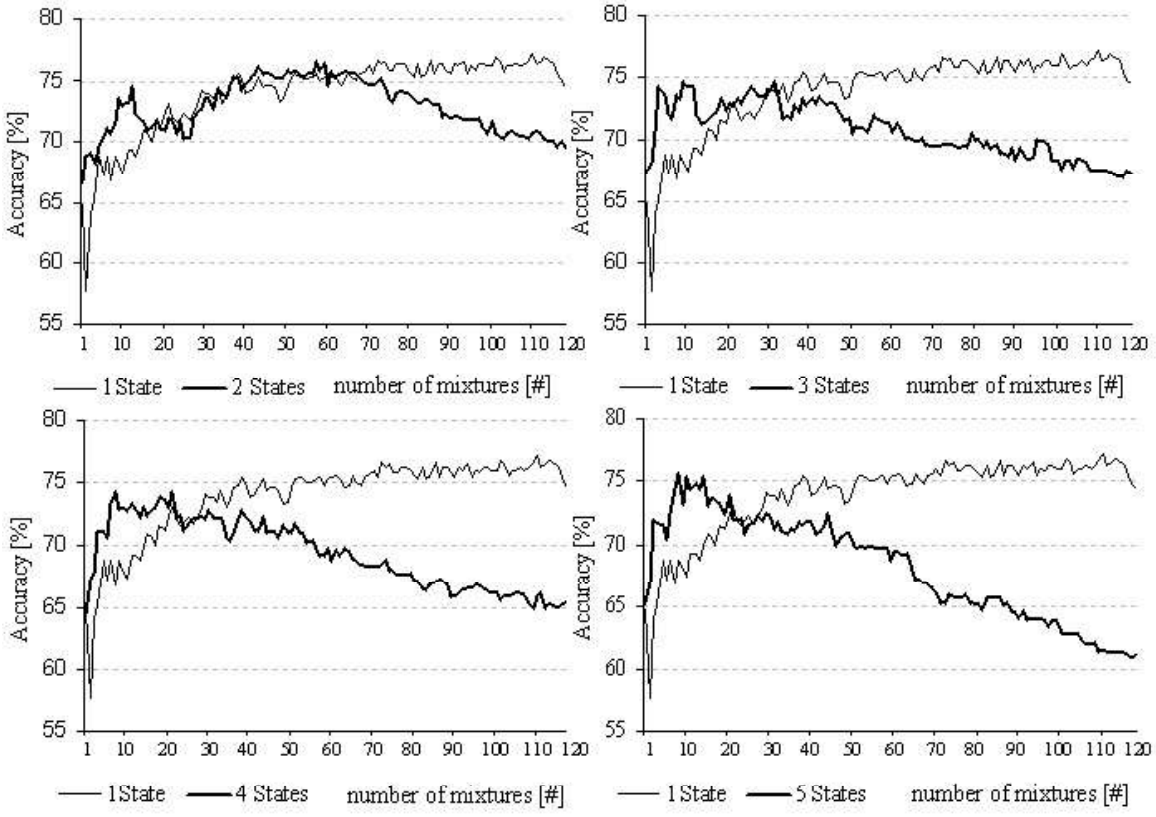

Fig. 2. Accuracy depending on the number of mixtures and number of HMM states, LOSO evaluation, database EMODB

As can be seen in fig. 2. single state HMM show the most stable and robust results.

For EMODB, we provide results of a leave-one-speaker-out (LOSO) evaluation to face the challenge of speaker independence. For SUSAS we decided for 10-fold stratified cross validation (SCV), as only 7 speakers are contained in the chosen spontaneous subset. On the other hand this is possible, as roughly 500 phrases are available per speaker.

Table 2. Combination of turn level and frame level analysis, databases EMODB with LOSO evaluation and speaker dependent 10-fold SCV for SUSAS. TL and FL abbreviate turn and frame levels. SN and FS represent speaker adaptation and feature selection. (√) indicates that the technique has been applied.

Accuracy [%]	SN	FS	EMODB	SUSAS
TL	-	-	74.9	80.8
TL	√	-	79.6	80.8
TL	√	√	83.2	83.3
FL	-	-	77.1	67.1
TL+FL	√	-	81.6	81.3
TL+FL	√	√	89.9	83.8

During feature selection the original 1,406 features have been reduced to 76 for EMODB. For SUSAS 71 features have been selected on the whole dataset, and 33-107 features were observed as optimum for the individual speakers. This underlines the brute-force nature of the creation of a >1k feature space in order to find a very compact robust final set. Tab. 1 shows the summarized results.

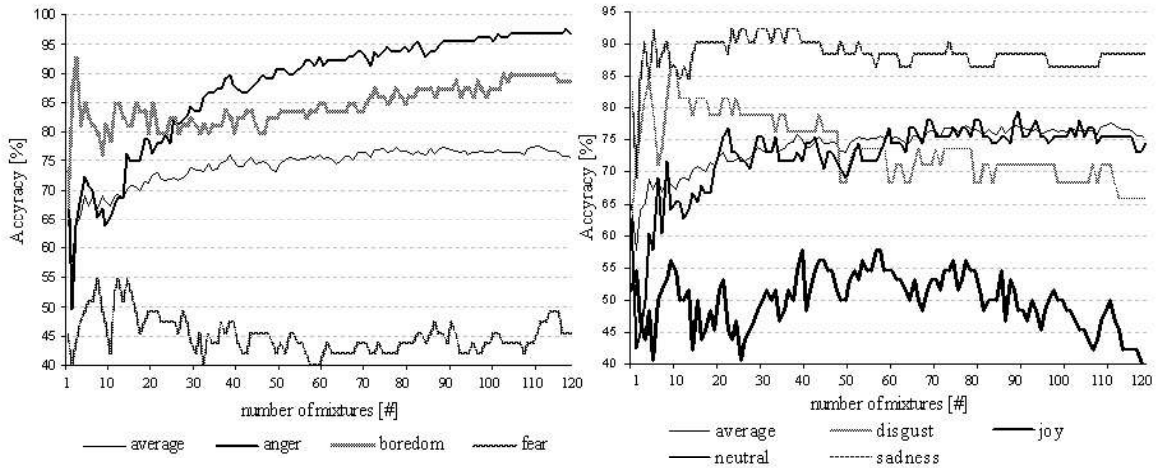

Fig. 3. Accuracy depending on the number of mixtures, LOSO evaluation, database EMODB

In the following the influence of the chosen number of mixtures for frame-level analysis is illustrated in detail for EMODB and SUSAS overall results and for each emotional state independently. As can be seen in fig. 3, a surprisingly high number of mixtures (>60) compared to the size of the database seems beneficial. Yet, not all emotions benefit from increase of mixtures, as e.g. fear. A similar behavior is observed for the SUSAS database (fig. 4)

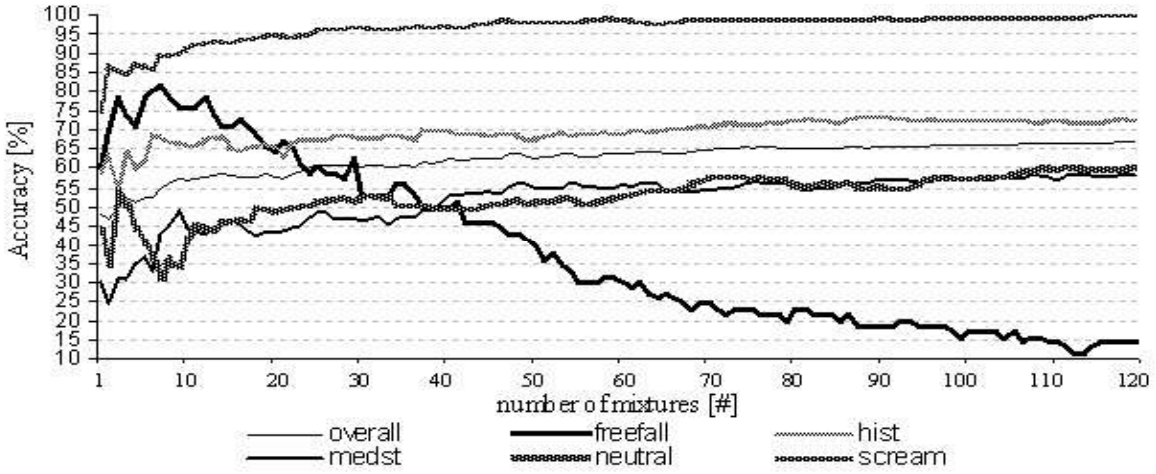

Fig. 4. Accuracy depending on the number of mixtures, mean of speaker-dependent 10-fold SCV evaluation, database SUSAS

where freefall is represented by very few samples and therefore is highly negatively influenced by the increase of mixtures. This comes, as classes with sparse data suffer from over-adaptation with respect to high accuracy. Also, this derives from the fact that no prior information was used in eq. 1. This however allows for more comparability between FL and TL modeling. In general ~50 mixtures seem optimal.

8 Discussion and Future Work

Within this paper we introduced speaker recognition motivated emotion recognition on a frame-level and its fusion with turn-level-based emotion recognition. The results presented do not allow for a direct comparison between these two, as a different number of LLDs has been used. Still, even using a typical speaker recognition system shows surprisingly high performance for the recognition of emotion within speech though it should be noted that FL clearly falls behind TL for the SUSAS database. When investigating the optimal number of mixtures to be used, it seems favorable to provide a minimum of 50 mixtures. However, some emotions may be negatively influenced by too high a number of mixtures. Considering dynamical modeling, no gain could be observed for use of several HMM states, as long as an adequate number of mixtures is provided.

Speaker normalization and feature space optimization both clearly help to improve overall results. Thereby it has to be noted that less than 10% of the original feature space suffices to get an optimum performance.

The highest accuracy is however obtained by the suggested fusion of both approaches. This is in particular true for EMODB. For SUSAS it is not too clear, whether the extra effort is justified.

It has to be mentioned that for both databases results in the order of human perception can be reported. This is true, even though speaker independence and spontaneous data in noisy environment have been faced, yet once at a time.

In future work we plan to investigate phonetic unit bound HMM models. Furthermore, a combination of functional application on syntactical unit motivated

chunks combined with dynamic modeling capabilities of an HMM seem a promising variant. Finally, the findings have to be verified on datasets without limited predefined spoken content.

Acknowledgements

The work has been conducted in the framework the NIMITEK project (Sachsen-Anhalt Federal State funding). Bogdan Vlasenko acknowledges support by a graduate grant of the Federal State of Sachsen-Anhalt.

References

1. Polzin, T.S., Waibel, A.: Detecting emotions in speech, Cooperative Multimodal Communication, 2nd Int. Conf. 98, CMC (1998)
2. Schuller, B., Rigoll, G., Lang, M.: Hidden Markov Model-Based Speech Emotion Recognition. In: Proc. ICASSP 2003, IEEE, Hong Kong, China, vol. II, pp. 1–4 (2003)
3. Lee, Z., Zhao, Y.: Recognition emotions in speech using short-term and long-term features. In: Proc. ICSLP, pp. 2255–2558 (1998)
4. Jiang, D.N., Cai, L.-H.: Speech emotion classification with the combination of statistic features and temporal features. In: Proc. ICME 2004, IEEE, Taipei, Taiwan, pp. 1967–1971 (2004)
5. Murray, L.R., Arnot, I.L.: Toward the simulation of emotion in synthetic speech: A review of the literature of humans vocal emotion. JASA 93(2), 1097–1108 (1993)
6. Schuller, B., Rigoll, G.: Timing Levels in Segment-Based Speech Emotion Recognition. In: Proc. INTERSPEECH 2006, ICSLP, ISCA, Pittsburgh, PA, pp. 1818–1821 (2006)
7. Klasmeyer, G., Johnstone, T., Bänziger, T., Sappok, C., Scherer, K.R.: Emotional Voice Variability in Speaker Verification. In: Proc. ITRW on Speech and Emotion, ISCA, Newcastle, UK (2000)
8. Shahin, I.: Enhancing speaker identification performance under the shouted talking condition using the second order circular Hidden Markov Models. Speech Communication 48(8), 1047–1055 (2006)
9. Reynolds, D.: Speaker identification and verification using Gaussian mixture speaker models. Speech Communication 17, 91–108 (1995)
10. Young, S., Evermann, G., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodland, P.: The HTK-Book 3. Cambridge University, Cambridge, England (2002)
11. Schuller, B., Seppi, D., Batliner, A., Maier, A., Steidl, S.: Towards More Reality in the Recognition of Emotional Speech. In: Proc. ICASSP 2007, Honolulu, Hawaii (2007)
12. Witten, I.H., Frank, E.: Data Mining: Practical machine learning tools with Java implementations, p. 133. Morgan Kaufmann, San Francisco (2000)
13. Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., Weiss, B.: A Database of German Emotional Speech. In: Proc. INTERSPEECH 2005, ISCA, Lisbon, Portugal, pp. 1517–1520 (2005)
14. Hansen, J.H.L., Bou-Ghazale, S.: Getting Started with SUSAS: A Speech Under Simulated and Actual Stress Database. In: Proc. EUROSPEECH-97, Rhodes, Greece, vol. 4, pp. 1743–1746 (1997)