Annotators' agreement and spontaneous emotion classification performance

Bogdan Vlasenko[1]*, Andreas Wendemuth*[2]

[1]Idiap Research Institute, Martigny, Switzerland
[2]Cognitive Systems, IIKT, Otto von Guericke University Magdeburg and
Center for Behavioral Brain Sciences Magdeburg

bogdan.vlasenko@gmail.com

Abstract

The combination of various types of data can significantly increase the amount of emotional material for training of more reliable real-life emotion classifiers. There are two well-known schemes of annotation utilized for emotional speech: multi-dimensional and categories-based. Multi-dimensional annotation is usually applied for labeling spontaneous emotional events, and categorial-based annotation is used for specification of the acted "full blown" emotional chunks. In order to simulate real-life conditions we used a cross-corpora evaluation strategy for datasets with different schemes of emotional annotation. Emotional models were trained on acted material from the EMO-DB (categories based annotation) dataset and evaluated on spontaneous data from the VAM dataset (multi-dimensional annotation). The best emotion classification performance was obtained on real-life emotional instances with the most intense arousal labels provided by a majority voting strategy (out of 17 annotators). We find that the corresponding spontaneous speech samples containing the most intensive emotional content are comparable with acted instances. The importance of employing a larger number of emotional annotators was finally addressed in our article.

Index Terms: emotion recognition, cross-corpora evaluation, phoneme-level emotional models, turn-level emotional models, emotional intensity

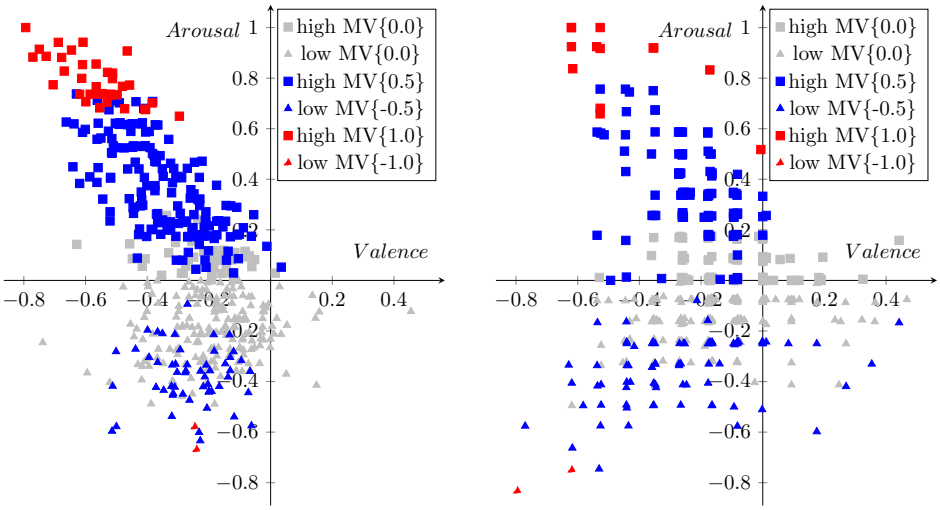
1. Introduction

It has been shown in [1, 2, 3] that recognizing the user's affective state is an important issue for developing intelligent human-computer interaction systems [4, 5, 6]. Most of these, however, require sufficient reliability, which may not be achieved yet. When evaluating the performance of emotion recognition techniques, obtainable accuracies are often overestimated. The main simplification that characterizes almost all emotion classifiers performance evaluations is that systems are usually trained and tested using the same dataset [7]. Within speaker-independent evaluations all kinds of potential mismatches between training and test data, such as different languages, acoustic channels, noises, or types of observed emotions, are usually not considered [8]. Addressing such typical sources of mismatch all at once is hardly possible, however, we believe that a first impression of the generalization ability of today's emotion classification engines can be obtained by cross-corpora evaluations. The research community could not yet specify emotional standard units which can be easily classified and determined by *any* "non-advanced" and "advanced" annotator of emotional content [9]. As a consequence, there is no unique methodology which defines the required professional skills of an "advanced" emotion annotator. Hence one can argue that using training and test sets which are at least annotated by different groups of annotators and types of annotation techniques (multi-dimensional, categorial) is an important issue of realistic scenarios.

Substantial parts of emotional datasets are annotated with a categorial approach. Several human labelers are usually employed for emotion annotation. The final emotional label is then selected with a "majority voting" approach. For our experiment we selected the VAM (VAM I + VAM II) [10] database with spontaneous emotions. During the annotation process several labelers were able to select one from five (-1, -0.5, 0.0, 0.5, 1) numerical values for each emotional dimension. The corresponding numerical values represent the level of emotional intensity for each modality. Afterwards, the obtained numerical values were processed using an *evaluator weighted estimator* (EWE) [11]. We decided to use the "majority voting" approach for parsing on different experimental datasets. During experiments we used EWE emotional labels mapped on a two-class problem. We determined three different subsets with defined "majority voting" winners. Afterwards we organized a cross-corpora evaluation test for each subset.

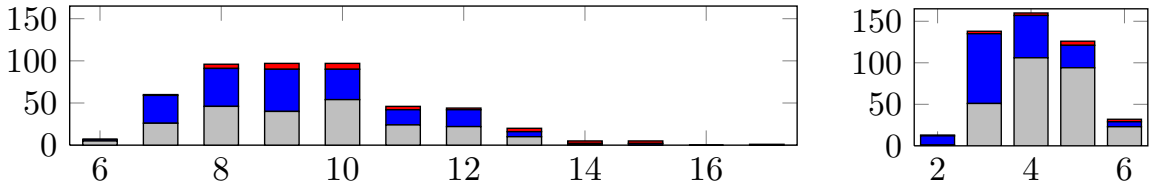
Two dimensional plots in Figure 1 display a distribution of the emotional instances presented in the VAM datasets. EWE estimations were mapped into valence-arousal (VA) space. The major part of emotional instances is located in the negative valence subspace, and just few samples in negative arousal subspace correspond to the positive emotions (positive valence). A comparable small number of the training samples for positive valence was the main reason, why we have trained our classifier just for the arousal discrimination task. In order to train reliable emotional classification techniques one should have sufficient amount of training data with reliable emotional annotation.

Fragopanagos et al. [12] state that most research efforts investigated the affective speech processing on the level of complete utterances, words, or phonetic transcription independent chunks [13, 14]. A comparably smaller number of methods are based on phonetic pattern modeling within emotion classification [15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26]. Several studies already reported accuracies on multiple corpora – however, only a very few consider training on one and testing on a different one (e.g., [27] and [28], where two and four corpora are employed, respectively). The experimental results reported in [9] showed that the phonetic pattern dependent modeling technique provides significantly better classification performance within the cross-corpora evaluations. In our research we trained phoneme-level emotional models on acted data from the EMO-DB [29] dataset and evaluate the obtained models on spontaneous emotions from the different subsets of the VAM dataset.



(a) VAM I (b) VAM II

Figure 1: Distribution of high (square) and low (triangle) arousal instances in VAM I and VAM II subsets. Gray - MV{0.0}, blue - MV{-0.5, 0.5}, red - MV {-1,1}.



(a) VAM I (b) VAM II

Figure 2: Distribution of maximal number of agreeing annotators for arousal dimension. Datasets VAM I and VAM II. Gray (bottom) - MV{0.0}, blue (midle) - MV{-0.5, 0.5}, red (top) - MV {-1,1}.

2. Selected databases

For training emotional models we selected the EMO-DB database which covers *anger, boredom, disgust, fear, joy, neutral,* and *sadness* speaker emotions. Ten (5 male, 5 female) professional actors speak ten German sentences with emotionally neutral linguistic meaning. For our experiment we selected utterances which have a level of naturalness not less than 60% and a level of recognizability not less than 80%. For specification of the emotional categories which can be modeled on the speech material presented in both datasets, we investigated possibilities to map the emotional states to the predominant type of general emotion categories, namely, high- and low- arousal [9, 30].

Table 1: Overview of emotional instances in VAM

	VAM I		VAM II	
Subset	**low**	**high**	**low**	**high**
MV{0.0}	188	40	178	97
MV{-0.5,0.5}	54	159	78	101
MV{-1.0,1.0}	2	35	2	13
Total	244	234	258	211

The VAM database consists of 12 hours of audio-visual recordings taken from a German TV talk show. The corpus contains 947 utterances with spontaneous emotions from 47 guests of the talk show, recorded from unscripted, authentic discussions. The VAM database contains two parts VAM I (19 speakers who had been roughly classified as "very good" with respect to the emotions conveyed) and VAM II (28 speakers who had been roughly classified as "good" with respect to the emotions conveyed). The speech extracted from the dialogs contains a large number of colloquial expressions as well as non-linguistic vocalizations and partly covers different German dialects. For annotation of the speech data, the audio recordings were manually segmented to the utterance level, where each utterance contained at least one phrase. A large number of human labelers were employed for annotation (17 annotators for VAM I, 6 annotators for VAM II) [10]. The labeling bases on a discrete five point scale for three dimensions (valence, arousal, dominance) mapped onto the interval of [-1,1]. For our evaluations, we used only arousal measures extracted from the annotation processed with *evaluator weighted estimators* (EWE). The original dimensional annotations were mapped into a two class problem (high-arousal > 0 vs. low-arousal ≤ 0). For our experiments we split VAM I and VAM II into three subsets with majority winner 0.0; -0.5 and 0.5; -1.0 and 1.0 (later specified as MV{0.0}, MV{-0.5,0.5}, MV{-1.0,1.0}). In Table 1 one can find the numbers of emotional utterances present in the evaluated 6 subsets of

emotional speech data. For example for the VAM I dataset and MV{-0.5,0.5} subset we have 54 low arousal instances with "majority voting" winner -0.5 and 159 high arousal instances with "majority voting" winner 0.5. Figure 2 presents information about the distribution of the maximum number of agreeing annotators for each subset of emotional speech data. VAM II contains more data with more than half of the annotators agreeing on the same emotional label in the MV{-1.0, 1.0} subset. On the other hand VAM II contains a comparable number of emotional instances with determined majority voting winner: MV{0.0} and MV{-0.5, 0.5}. The Kiel Corpus of Read Speech [31] was used for training basic ASR acoustic models (for more details see [9]).

3. Acoustic feature extraction

In our research we applied low-level feature modeling on frame-level for emotion recognition from speech. The speech signal is processed using a 25 ms Hamming window, with a 10 ms shifting step. A 39 dimensional feature vector consisting of 12 MFCC and zero-order Cepstral coefficient plus delta and delta-delta (acceleration) coefficients is employed. Cepstral Mean Subtraction (CMS) is applied to better cope with channel characteristics.

4. Emotion classifier

The implemented classification technique is based on a two-stages classification process. On the first stage German phonetic transcriptions with corresponding phoneme alignments are generated for each test utterance. During the second stage we use the corresponding phoneme alignment for phoneme-level emotion classification. The applied classification criteria can be expressed as:

$$\begin{aligned} \mathcal{W}_{\Omega_k} &= \arg\max_{\mathcal{W}_\Omega} \log P(\mathbf{O}|\mathcal{W}_\Omega, \mathcal{M}_{pho}) \\ &= \arg\max_{\mathcal{W}_\Omega} \log \sum_{\forall \mathbf{s}} p(\mathbf{O}, \mathbf{s}|\mathcal{W}_\Omega, \mathcal{M}_{pho}) \end{aligned} \quad (1)$$

where $\mathcal{W}_{\Omega_k}$ is an emotional phoneme sequence built from phonemes of Ω_k emotional class, $\mathcal{M}_{pho}$ is a phoneme level **HMM/GMM's** parameter set, $\mathbf{s} = [s_1, s_2, \ldots, s_T]$ is a state sequence associated with the observation vector sequence $\mathbf{O} = [o_1, o_2, \ldots, o_T]$, $\mathcal{W}_\Omega$ is a possible phoneme emotion sequence for Ω_1 ="*low arousal*" or Ω_2 ="*high arousal*" emotional state in our case; $P(\mathbf{O}|\mathcal{W}_\Omega)$ is an emotion acoustic model for the emotion phoneme states sequence $\mathcal{W}_\Omega$; $P(\mathcal{W}_\Omega)$ is a priori knowledge about the affective state frequency of occurrence for the phonetic units sequence $\mathcal{W}_\Omega$.

The HMMs parameter set $\mathcal{M}_{pho}$ consists of parameters which specify *"low-arousal"* and *"high-arousal"* emotion phonemes. Namely, the full lists of phonemes are modeled for *"low-arousal"* and *"high-arousal"* emotions, independently. Hence, $2 \times 36 = 72$ emotional phoneme models are implemented for the EMO-DB database. In order to simplify the emotion classification process we decided to use a fixed phoneme sequence $\hat{\mathcal{W}}$ with corresponding optimal state sequence $\omega = [s_1^{opt}, s_2^{opt}, \ldots, s_T^{opt}] = [\omega_1, \omega_2, \ldots, \omega_T]$. To specify a fixed phoneme sequence we used an ASR engine to recognize phoneme sequences. With a defined optimal state sequence we could simplify the maximization task represented in equation 1 by estimating $p(\mathbf{O}, \mathbf{s}|\mathcal{W}_\Omega, \mathcal{M}_{pho})$ just for the optimal state sequence. In this case, implemented in our current research, the classification criteria can be expressed as:

$$\begin{aligned} \Omega_k &= \arg\max_{\Omega} \log \left\{ p(\boldsymbol{\omega}|\hat{\mathcal{W}}_\Omega, \mathcal{M}_{pho}) p(\mathbf{O}|\boldsymbol{\omega}, \mathcal{M}_{pho}) \right\} \\ &= \arg\max_{\Omega} \left\{ \log \pi_{\omega_1} + \sum_{t=1}^{T} \log b_{\omega_t}(\mathbf{o}_t) + \sum_{t=1}^{T} \log a_{\omega_{t-1}\omega_t}, \right\} \end{aligned} \quad (2)$$

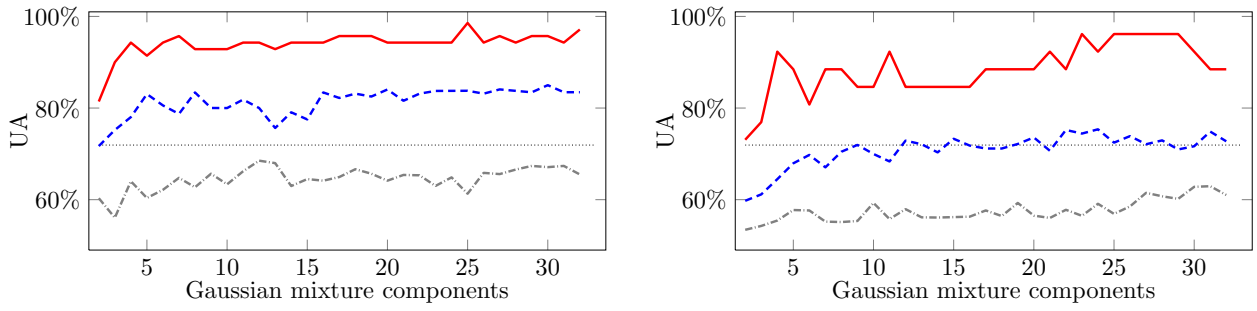
where $\hat{\mathcal{W}}_\Omega$ is an optimal phoneme sequence $\hat{\mathcal{W}}$ build from emotional phonemes from an emotional class Ω.

Considering an initial state distribution π_i, state transaction probabilities a_{ij} and observation generation probability distributions $b_i(\mathbf{o}_t)$, we estimate two main multipliers $p(\boldsymbol{\omega}|\hat{\mathcal{W}}_\Omega, \mathcal{M}_{pho})$ and $p(\mathbf{O}|\boldsymbol{\omega}, \mathcal{M}_{pho})$. The first one is the probability of passing through the optimal state sequence $\boldsymbol{\omega}$, the second one is the probability of observing the acoustic feature vector sequence $\mathbf{O}$ given the state sequence $\boldsymbol{\omega}$. These multipliers will be estimated for both emotional phoneme classes. The estimation of the HMMs parameters is implemented in two steps. In the first step, we estimate a basic HMMs parameter set $\mathcal{M}_{pho}^{ubm}$ on emotionally neutral speech samples from the Kiel Corpus of Read Speech [31]. In the second step, we adapt $\mathcal{M}_{pho}^{ubm}$ with combined *Maximum Likelihood Linear Regression (MLLR)* (32 regression class trees) + *Maximum a Posteriori (MAP)* adaptation (hyper-parameter $\tau = 2$), (a similar approach provided an optimal recognition performance in [9, 30]). A corresponding adaptation parameter setup is used for the employed ASR engine. We evaluate these classifiers and present the classification performance as a function of the number of GMMs (from 2 to 32) in Section 5.

For our ASR engine we applied a continuous density HMMs technique based on multivariate GMMs with 32 mixture components. In order to compensate the mismatch of acoustic characteristics between neutral speech samples and affective speech material we applied two model-based transforms: a basic *MLLR* with 32 regression classes and *MAP* with $\tau = 2$. Phoneme level bi-gram language models are applied in the ASR engine for specification of the optimal state sequences $\boldsymbol{\omega}$ in equation 2. Acoustic models adapted with corresponding adaptation parameters configuration showed the best spontaneous emotional speech recognition performance. We trained our phoneme-level emotional models on speech samples from the EMO-DB database. For our experiments we used equal priors: $P(\text{"}high\, arousal\text{"}) = P(\text{"}low\, arousal\text{"}) = 1/2$.

5. Experimental results

As the numbers of emotional instances in the selected speech corpora (see Table 1) are unbalanced, we selected *unweighted average recall (UA)* for specification of emotion-recognition performances. Unweighted average recall is the sum of all class accuracies, divided by the number of classes, without considering the number of instances per class. Figure 3 displays recognition rates for phonetic-pattern dependent non-optimized (an arbitrary number of GMMs) emotion classifiers trained on acted emotional instances from EMO-DB database and evaluated on the subsets of the VAM database. Baseline results obtained on the complete VAM datasets with optimal classifier configurations ($UA = 71.92\%$, 31 GMMs) are illustrated with dotted-black line. The baseline result was obtained during cross-corpora presented in [9]. The experimental results on the VAM I dataset (see Figure 3(a)) show that the classification



(a) VAM I

(b) VAM II

Figure 3: Unweighted average recall rates as a function of the number of Gaussian mixture models. Emotional models are trained on the EMO-DB dataset and evaluated on 6 VAM data subsets. Solid-red: MV {-1,1}, dashed-blue: MV{-0.5, 0.5}, dotdashed-gray: MV{0.0}, doted-black: baseline result for optimized configuration evaluated on the complete VAM database.

performances for the MV {-1,1} and MV{-0.5, 0.5} subsets are better in comparison with baseline results. For VAM II dataset material we obtained outperforming classification performance for the MV {-1,1} subset. The best classification performance was obtained on the VAM I, MV {-1,1} subset ($UA = 98.57\%$, 25 GMMs) and the VAM II, MV subset {-1,1} ($UA = 96.15\%$, 25..29 GMMs).

6. Conclusions

The main outcome of our evaluation experiment (see Figure 3) is that phoneme-level emotion modeling provides outstanding classification performance on preselected spontaneous emotional samples. The selection criteria are based on a majority voting strategy. "Full blown" acted emotions annotated with a categorial approach could be associated with spontaneous emotions with majority winner for the highest possible arousal value (-1 and 1 in our case). Our phoneme-level models outperformed baseline emotion classification performances for the earlier mentioned three subsets VAM I, MV{-0.5, 0.5} - 203 samples; VAM I, MV {-1,1} - 37 samples; VAM II, MV {-1,1} -13 samples. We also showed that the phoneme can be seen as the smallest possible acoustic unit for cross-corpora classification of emotional arousal in two classes: low, high. Emotional models trained on acted emotional speech samples from the EMO-DB database could provide outstanding classification performance for the most expressive spontaneous emotional speech samples from the VAM dataset.

The second important outcome was obtained during analyzing the distribution of maximal numbers of agreeing annotators presented in Figure 2. By defining a threshold for the number of agreeing annotators for the VAM I dataset (about 8) and the VAM II dataset (about 3) we could specify requirements for selection of the most expressive samples. All three subsets with outperforming classification performance contained emotional instances with level of agreement larger than the proposed threshold.

7. Discussion and Outlook

We obtained unexpected and notable results within cross-corpora evaluation on the most expressive spontaneous emotional speech samples. Experimental results presented in Figure 3(a) and Figure 3(b) show that emotional instances annotated with a larger number of annotators have better classification performance. In order to improve the reliability of the multi-dimensional emotional annotation technique one should apply EWE measures for multi-labeler annotations (at least 17 human labelers). Creation of new well annotated (by using an approved annotation approach like in [32]) emotional corpora [33] will help us to make a more detailed emotional speech analysis.

We highlighted the importance of using an emotional dataset with reliable emotional labels for training emotional models. Emotional models trained on acted emotions from the EMO-DB database provide more stable classification performance on spontaneous emotions. By using a majority voting strategy implemented for a large number of annotators one could preselect the most expressive spontaneous emotions samples. By using a combined training set with acted and the most expressive spontaneous emotional samples one could improve the reliability of emotion classifiers.

From our perspective, detection of the high expressive emotional events and implementation of emotion adaptive dialog management could make spoken dialog system more user friendly. We assumed that emotional instances presented in the VAM I dataset have more reliable emotional content marked as "very good" by dataset developers [10]. Developers of emotional dataset should address attention to the intensity of spontaneous emotions. Detection of non-expressive emotional instances requires more data with reliable and "delicate" emotional annotation. From the other side, the emotion research community should address the question of the applicability of the detection of non-expressive emotions.

The emotion research community should provide a better fundamental analysis of human emotion perception and production. With more detailed human emotion's perception and generation analysis the affective computing community will be able to specify emotional standard units which can be easily determined and classified by *any* "advanced" and "non-advanced" listener. This will also enable us to make our emotion classification techniques more robust.

8. Acknowledgments

This research is associated and supported by the Transregional Collaborative Research Center SFB/TRR 62 "Companion-Technology for Cognitive Technical Systems" funded by the German Research Foundation (DFG).

9. References

[1] R. W. Picard, E. Vyzas, and J. Healey, "Toward Machine Emotional Intelligence: Analysis of Affective Physiological State," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, pp. 1175–1191, 2001.

[2] B. Vlasenko and A. Wendemuth, "Heading toward to the natural way of human-machine interaction: The nimitek project," in *Proceedings - 2009 IEEE International Conference on Multimedia and Expo, ICME 2009*, 2009, pp. 950–953.

[3] K. Scherer, J. Sundberg, L. Tamarit, and G. Salomo, "Comparing the acoustic expression of emotion in the speaking and the singing voice," *Computer Speech and Language*, vol. 29, no. 1, pp. 218–235, 2015.

[4] J. Kim, D. Erickson, S. Lee, and S. Narayanan, "A study of invariant properties and variation patterns in the converter/distributor model for emotional speech," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2014, pp. 413–417.

[5] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Mller, and S. Narayanan, "Paralinguistics in speech and language - state-of-the-art and the challenge," *Computer Speech and Language*, vol. 27, no. 1, pp. 4–39, 2013.

[6] Z. Zhang, J. Deng, E. Marchi, and B. Schuller, "Active learning by label uncertainty for acoustic emotion recognition," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2013, pp. 2856–2860.

[7] B. Schuller, B. Vlasenko, F. Eyben, M. Wöllmer, A. Stuhlsatz, A. Wendemuth, and G. Rigoll, "Cross-Corpus Acoustic Emotion Recognition: Variances and Strategies," *IEEE Transactions on Affective Computing*, vol. I, pp. 119–131, 2010.

[8] B. Schuller, B. Vlasenko, F. Eyben, G. Rigoll, and A. Wendemuth, "Acoustic Emotion Recognition: A Benchmark Comparison of Performances," in *IEEE ASRU 2009*, 2009, pp. 552–557.

[9] B. Vlasenko, D. Prylipko, R. Böck, and A. Wendemuth, "Modeling phonetic pattern variability in favor of the creation of robust emotion classifiers for real-life applications," *Computer Speech and Language*, vol. 28, no. 2, pp. 483 – 500, 2014.

[10] M. Grimm, K. Kroschel, and S. Narayanan, "The Vera am Mittag German Audio-Visual Emotional Speech Database," in *ICME 2008*, 2008, pp. 865–868.

[11] M. Grimm, K. Kroschel, E. Mower, and S. Narayanan, "Primitives-based evaluation and estimation of emotions in speech," *Speech Communication*, vol. 49, no. 10-11, pp. 787–800, 2007.

[12] N. Fragopanagos and J. G. Taylor, "Emotion recognition in human-computer interaction," *Neural Networks*, vol. 18, no. 4, pp. 389–405, 2005.

[13] J. Arias, C. Busso, and N. Yoma, "Shape-based modeling of the fundamental frequency contour for emotion detection in speech," *Computer Speech and Language*, vol. 28, no. 1, pp. 278–294, 2014.

[14] C. Busso, A. Metallinou, and S. Narayanan, "Iterative feature normalization for emotional speech detection," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2011, pp. 5692–5695.

[15] C. M. Lee, S. Yildirim, M. Bulut, A. Kazemzadeh, C. Busso, Z. Deng, S. Lee, and S. Narayanan, "Emotion recognition based on phoneme classes," in *Interspeech 2004*, 2004, pp. 889–892.

[16] M. Bulut, C. Busso, S. Yildirim, A. Kazemzadeh, C. Lee, S. Lee, and S. Narayanan, "Investigating the role of phoneme-level modifications in emotional speech resynthesis," in *Interspeech 2005*, 2005, pp. 801–804.

[17] B. Vlasenko, D. Prylipko, D. Philippou-Hübner, and A. Wendemuth, "Vowels formants analysis allows straightforward detection of high arousal acted and spontaneous emotions," in *Interspeech 2011*, 2011, pp. 1577–1580.

[18] C. Busso, S. Lee, and S. Narayanan, "Using Neutral Speech Models for Emotional Speech Analysis," in *Interspeech 2007*, 2007, pp. 2225–2228.

[19] B. Vlasenko and A. Wendemuth, "Processing affected speech within human machine interaction," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2009, pp. 2039–2042.

[20] M. Goudbeek, J. Goldman, and K. R. Scherer, "Emotion dimensions and formant position," in *Interspeech 2009*, 2009, pp. 1575–1578.

[21] B. Vlasenko, B. Schuller, A. Wendemuth, and G. Rigoll, "On the Influence of Phonetic Content Variation for Acoustic Emotion Recognition," in *PIT 2008*, 2008, pp. 217–220.

[22] C. Montacié and M.-J. Caraty, "Combining multiple phoneme-based classifiers with audio feature-based classifier for the detection of alcohol intoxication," in *Interspeech 2011*, 2011, pp. 3205–3208.

[23] B. Vlasenko, B. Schuller, K. Mengistu, G. Rigoll, and A. Wendemuth, "Balancing spoken content adaptation and unit length in the recognition of emotion and interest," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2008, pp. 805–808.

[24] R. Gajšek, F. Mihelič, and S. Dobrišek, "Speaker state recognition using an HMM-based feature extraction method," *Computer Speech and Language*, vol. 27, no. 1, 2013.

[25] B. Vlasenko and A. Wendemuth, "Determining the smallest emotional unit for level of arousal classification," in *Proceedings - 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, ACII 2013*, 2013, pp. 734–739.

[26] S. Mariooryad and C. Busso, "Compensating for speaker or lexical variabilities in speech for emotion recognition," *Speech Communication*, vol. 57, pp. 1–12, 2014.

[27] M. Shami and W. Verhelst, "Automatic classification of emotions in speech using multi-corpora approaches," in *BENELUX/DSP (SPS-DARTS 2006)*, 2006, pp. 3–6.

[28] ——, "Automatic classification of expressiveness in speech: A multi-corpus study," ser. Computer Science / Artificial Intelligence, 2007, pp. 43–56.

[29] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A Database of German Emotional Speech," in *Interspeech 2005*, 2005, pp. 1517–1520.

[30] B. Vlasenko, D. Philippou-Hbner, and A. Wendemuth, "Parameter optimization issues for cross-corpora emotion classification," in *Proceedings - 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, ACII 2013*, 2013, pp. 454–459.

[31] K. J. Kohler, "Labelled data bank of spoken standard German - the Kiel Corpus of read and spontaneous speech," in *ICSLP 1996*, 1996, pp. 1938–1941.

[32] I. Siegert, R. Böck, D. Philippou-Hübner, B. Vlasenko, and A. Wendemuth, "Appropriate Emotional Labeling of Non-acted Speech Using Basic Emotions, Geneva Emotion Wheel and Self Assessment Manikins," in *ICME 2011*, 2011.

[33] S. Mariooryad, R. Lotfian, and C. Busso, "Building a naturalistic emotional speech corpus by retrieving expressive behaviors from existing speech corpora," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2014, pp. 238–242.