TENDENCIES REGARDING THE EFFECT OF EMOTIONAL INTENSITY IN INTER CORPUS PHONEME-LEVEL SPEECH EMOTION MODELLING

Bogdan Vlasenko

University of Passau
Chair of Complex & Intelligent Systems
Passau, Germany
bogdan.vlasenko@uni-passau.de

Björn Schuller[*]

University of Passau
Chair of Complex & Intelligent Systems
Passau, Germany
schuller@ieee.org

Andreas Wendemuth [†]

Otto von Guericke University Magdeburg
Cognitive Systems, IIKT
Magdeburg, Germany
andreas.wendemuth@ovgu.de

ABSTRACT

As emotion recognition from speech has matured to a degree where it becomes suitable for real-life applications, it is time for developing techniques for matching different types of emotional data with multi-dimensional and categories-based annotations. The categorical approach is usually applied for acted 'full blown' emotions and multi-dimensional annotation is often preferred for spontaneous real life emotions. A particularly realistic task we consider in this contribution is cross-corpus emotion recognition and its evaluation. General and phoneme-level emotional models on acted and spontaneous emotions ('very intense' and 'intense') are used in our experimental study. The emotional models were trained on spontaneous emotions from the complete VAM dataset and subsets with variable emotional intensities and evaluated on acted emotions from the Berlin EMO-DB dataset. We observe a significant classification performance gap for general models trained on very intense spontaneous emotions. As a consequence, we address the importance of collecting large corpora with very intense emotional content for training more reliable phoneme-level emotional models.

Index Terms— emotion recognition, cross-corpus evaluation, phoneme-level emotional models, turn-level emotional models, emotional intensity

1. INTRODUCTION

Emotion classification and detection are playing an increasingly important role in user-friendly human-machine interaction. It has been shown in [1, 2] that recognising the user's affective state is an important issue for development of intelligent human-computer interaction systems. Most of these, however, require sufficient reliability, which may not be achieved, yet. When evaluating performance of emotion classification techniques, obtainable accuracies are often overestimated. The main simplification that characterises almost all emotion recognition performance evaluations is that systems are usually trained and tested using the same dataset. 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. Addressing such typical sources of mismatch all at once is hardly possible; however, we believe that a good impression of the generalisation ability of today's emotion classification engines can be obtained by cross-corpora evaluations. Further, the affective computing 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 [3]. As a consequence, there is no unique methodology which defies 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 labellers and types of annotation technique (multi-dimensional, categorical) is an important issue of realistic scenarios. In our present research we are using turn-level and phoneme-level emotional models trained on spontaneous emotions as presented in the VAM [4] emotional speech samples and evaluated on acted emotions presented in the EMO-DB [5] database.

A comparably 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. In the case of acted data as presented in the EMO-DB database, one could use acoustic perception measures (naturalness and recognisability) for selection of the most reliable data. By implementing robust classification techniques for emotional speech samples with reliable

[*] The author is further affiliated with Department of Computing, Imperial College London, London, UK.

[†] The author is further affiliated with Center for Behavioral Brain Sciences Magdeburg, Germany cbbs.eu

annotation, one should obtain applicable classification performance within a cross-corpus experimental setup. In [3], it is assumed that the usage of training and test sets annotated with different annotation techniques, namely dimensional (the VAM database) and categorical (the EMO-DB database) could make real-world application setups more realistic.

Fragopanagos et al. [6] state that, most research efforts investigated the affective speech processing on the level of complete utterances, words, or phonetic transcription independent chunks. A comparably smaller number of methods are based on phonetic pattern modelling within emotion classification [7, 8, 9, 10, 11, 12, 13, 14, 15]. Still, most of the aforementioned phoneme-level modelling emotion classification techniques used forced alignment or manual annotation for the extraction of the phoneme borders. Just some studies faced real-life conditions by using automatic speech recognition (ASR) engines for generating the phoneme alignment [16, 15, 17]. Current ASR techniques are not able to provide as good phoneme alignment on emotional speech samples as manual annotation or forced alignment do. To properly address *real-world conditions*, a phoneme-level emotion processing method presented in this contribution relies on the phoneme alignment obtained by using an ASR system which applies acoustic models adapted on affective speech samples. Several studies already reported accuracies on multiple corpora – however, only very few consider training on one and testing on a different one (e. g., [18] and [19], where two and four corpora are employed, respectively). The experimental results reported in [3, 20] showed that, a phonetic pattern dependent modelling technique provides significantly better classification performance within cross-corpus evaluations. Bone et al. [21] present a multi-corpora study on a robust, unsupervised (rule-based) method for providing a scale-continuous, bounded arousal rating operating on the vocal signal.

2. SELECTED DATABASES

For our experiments we used two emotional databases which contain recordings of adult German native speakers: the Berlin EMO-DB [5] set and the also German VAM [4] corpus.

The EMO-DB database provides speech samples with *anger, boredom, disgust, fear, joy, sadness,* and *neutral* as speaker emotions. Ten (in gender balance) professional actors speak ten German sentences with emotionally neutral linguistic meaning. For our experiment, we selected utterances which have a level of recognisability not less than 80 % and a level of naturalness not less than 60 %. For specification of the emotional categories which can be modelled 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. We consider two plots which reflect the possible location of some emotion categories in valence-arousal (VA) space [22, 23]. The first plot [23] was created by the mapping of the terms Russell [24] uses as markers for his claim of an emotion circumplex. On the second plot [22], the authors presented sympathetic forms of activation for the 24 emotion terms in the valence-arousal space. Due to the various possible locations of disgust (in positive arousal sub-space for the first plot (see [23]) and negative arousal sub-space for the second plot (see [22])), we decided to eject disgust instances from our experimental subset.

Hence, for experiments on the EMO-DB dataset we used *neutral (78 utterances)*, low arousal emotions (*boredom (79), sadness (53)*), and high arousal emotions (*anger (127), fear (55),* and *joy (64)).* Within our emotion classification experiments, we combined neutral with low-arousal emotional speech samples and later obtained a combined cover class referred to as low-arousal emotion.

Table 1. Overview of the emotional instances in the evaluated datasets

Dataset	**low arousal**	**high arousal**	**type of emotion**
EMO-DB	210	246	*acted*
VAM	502	445	*mixed spontaneous*
VAM I	244	234	*very intense spont.*
VAM II	258	211	*intense spont.*

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) [4]. In our research we use the terms *"very intense"* instead of **very good** and *"intense"* instead of **"good"**. The speech extracted from the dialogues contains a large number of colloquial expressions as well as non-linguistic vocalisations 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 labellers was employed for annotation (17 labellers for VAM I, six for VAM II). The labelling bases on a discrete five point scale for three dimensions (valence, arousal, dominance) mapped onto the interval of [-1,1]. Emotional labels generalised with the *evaluator weighted estimator* (EWE) [25] were used for our experiments. An overview of emotional instances in the utilised speech datasets is presented in Table 1. The Kiel Corpus of Read Speech [26] was used for training basic ASR acoustic models (for more details see [3]).

3. EMOTION RECOGNITION METHODS

In our research, we applied general (turn-level) phonetic pattern independent and phoneme level emotion classifiers. For our experiments, we used equal priors for the emotion categories: $P(\text{``}high\,arousal'') = P(\text{``}low\,arousal'') = 1/2$.

3.1. Acoustic feature extraction

The two considered classification techniques base on similar acoustic features. The speech signal is processed using a 25 ms Hamming window, with a 10 ms shift. A 39 dimensional feature vector consisting of 12 MFCC and the zero-order cepstral coefficient plus delta (speed) and delta-delta (acceleration) coefficients is employed. Cepstral Mean Subtraction (CMS) is applied to better cope with channel characteristics.

3.2. General emotion classification: turn-level analysis

We consider using a statistical analysis applied to ASR to recognise emotions from speech introduced in [27]. Instead of the common task to deduce the most likely word sequence hypothesis Ω_k from a given vector sequence $\mathbf{O}$ of acoustic observations $\mathbf{o}$, the task is to recognise the current speaker's emotional state.

The applied classification criteria can be expressed as:

$$\begin{aligned}\Omega_k &= \arg\max_{\Omega} \log \frac{P(\mathbf{O}|\Omega)P(\Omega)}{P(\mathbf{O})} \\ &= \arg\max_{\Omega} \log P(\mathbf{O}|\Omega, \mathcal{M}) \qquad (1)\end{aligned}$$

where: Ω is one of all system known emotions (*"low arousal"*, *"high arousal"*, "neutral" emotional state); $P(\mathbf{O}|\Omega)$ is the emotion acoustic model; $P(\Omega)$ is the prior user-behaviour information. $\mathcal{M}$ is a HMMs' parameter set [3]. For estimation parameters $\mathcal{M}$ we used forward-backward and Baum-Welch re-estimation algorithms included in the HTK toolkit [28]. For evaluations, we use configurations with different numbers of Gaussian mixtures (from two to 120).

3.3. Phonetic pattern dependent emotion classification: emotional phoneme classes

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:

$$\begin{aligned}\mathcal{W}_{\Omega_k} &= \arg\max_{\mathcal{W}_\Omega} \log \frac{P(\mathbf{O}|\mathcal{W}_\Omega)P(\mathcal{W}_\Omega)}{P(\mathbf{O})} \\ &= \arg\max_{\mathcal{W}_\Omega} \log P(\mathbf{O}|\mathcal{W}_\Omega, \mathcal{M}_{pho}) \\ &= \arg\max_{\mathcal{W}_\Omega} \log \sum_{\forall \mathbf{s}} p(\mathbf{O}, \mathbf{s}|\mathcal{W}_\Omega, \mathcal{M}_{pho}), \qquad (2)\end{aligned}$$

where $\mathcal{W}_{\Omega_k}$ is an emotional phoneme sequence build from phonemes of Ω_k – the 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"* as emotional state in our case; $P(\mathbf{O}|\mathcal{W}_\Omega)$ is an emotion acoustic model for the emotion phoneme states sequence $\mathcal{W}_\Omega$; and $P(\mathcal{W}_\Omega)$ is a priori knowledge about the affective state frequency of occurrence for the phonetic units sequence $\mathcal{W}_\Omega$.

The HMM's parameter set $\mathcal{M}_{pho}$ consists of parameters which specify *"low-arousal"* and *"high-arousal"* emotion phonemes. Namely, the full lists of phonemes are modelled 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 a 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 recognise phoneme sequences. A more detailed specification of the applied ASR engine can be found in the following section. Taking into account that the implemented ASR system and the phoneme level emotion recognition system use an identical HMM architecture (*left-to-right* monophone models with three emitting states), we could use an optimal state sequence for our phoneme level emotion modelling. With a defined optimal state sequence, we could simplify the maximisation task represented in equation 2 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 present research, the classification criteria can be expressed as:

$$\begin{aligned}\Omega_k &= \arg\max_{\Omega} \log \Big\{ p(\boldsymbol{\omega}|\hat{\mathcal{W}}_\Omega, \mathcal{M}_{pho}) p(\mathbf{O}|\boldsymbol{\omega}, \mathcal{M}_{pho}) \Big\} \\ &= \arg\max_{\Omega} \left\{ \log \pi_{\omega_1} + \sum_{t=1}^{T} \log b_{\omega_t}(\mathbf{o}_t) + \sum_{t=1}^{T} \log a_{\omega_{t-1}\omega_t} \right\}, \qquad (3)\end{aligned}$$

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 HMM's parameters is implemented in two steps. In the first step, we estimate a basic HMM's parameter set $\mathcal{M}^{ubm}_{pho}$ on emotionally neutral speech samples from the Kiel Corpus of Read Speech [26]. In the second step, we adapt $\mathcal{M}^{ubm}_{pho}$ with combined *Maximum Likelihood Linear Regression (MLLR)* (32 regression class trees) + *Maximum a Posteriori (MAP)* adaptation (hyper-parameter $\tau = 2$), (a similar setup provided an optimal recognition performance in [3, 29]). The acoustic models adapted with corresponding adaptation parameters' configuration showed the best spontaneous emotional speech recognition performance. For evaluations, we use configurations with a different number of mixtures (from 2 to 32). We evaluated these classifiers and present classification performance as a function of the number of mixtures (denoted as GMMs) in Section 4.

3.3.1. Automatic speech recognition engine

For our ASR engine we applied a continuous density HMM 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 transformations: a 'basic' *Maximum Likelihood Linear Regression (MLLR)* with 32 regression 'classes' and *Maximum a Posteriori (MAP)* with $\tau = 2$. Phoneme level bi-gram language models are applied in the ASR engine for specification of the optimal state sequences $\boldsymbol{\omega}$ as in equation 3.

4. EXPERIMENTAL RESULTS

As the numbers of emotional instances in the selected speech corpora (see Table 1) are unbalanced, we selected *unweighted average recall (UA)* as suited measure for emotion-recognition performance. UA is calculated as the sum of all class accuracies, divided by the number of classes, thus (intentionally) ignoring the number of instances per class.

During the first experimental phase, we employed general emotion classification techniques as described in Section 3.2. Two sets of emotional models have been trained on the VAM I and VAM II dataset, accordingly. These sets of emotional models have been evaluated within intra-corpus speaker independent evaluations on the same datasets (VAM I, VAM II). For our first experiment we used the Leave-One-Speaker-Group-Out (LOSGO) strategy.

Figure 1 displays recognition rates for phonetic-pattern independent non-optimised (arbitrary number of GMM mixtures in a range from 2 to 120) emotion classifiers trained on different VAM subsets and evaluated on the complementary subsets of the VAM database (e. g., trained on VAM I, evaluated on VAM II). The solid black line depicts the baseline results corresponding to Leave-One-Speaker-Group-Out (LOSGO) evaluation with optimal mixture (GMMs) number [30].

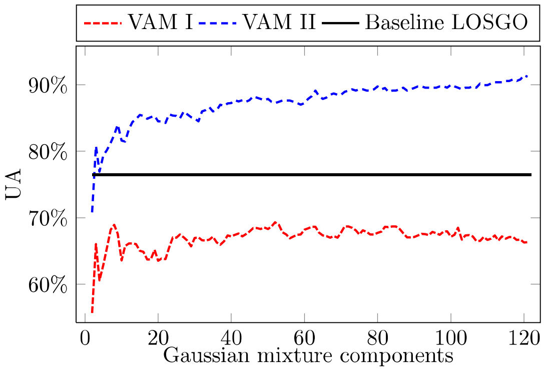

Fig. 1. *UA rates for intra-corpus evaluation. Phonetic-pattern independent (general) models trained on VAM II and evaluated on VAM I (blue dashed line); and trained on VAM I and evaluated on VAM II (red dashed line). Baseline results correspond to a LOSGO evaluation with the optimal mixture (GMMs) number.*

During the second experimental phase, we also employed general emotional models. The different sets of emotional models have been trained on the complete VAM dataset, and the VAM I and VAM II subsets. The obtained sets of emotional models have been evaluated within cross-corpus evaluations on the EMO-DB set.

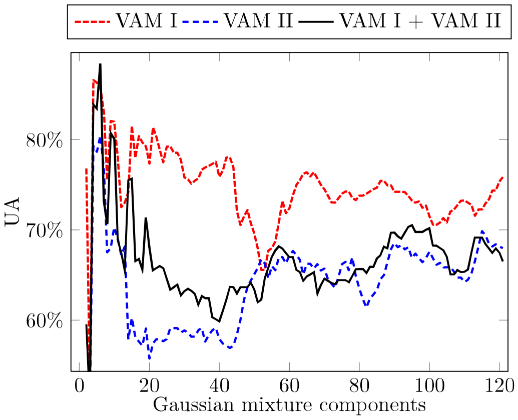

Fig. 2. *UA rates for cross-corpus evaluations. Phonetic-pattern independent (general) models trained on a subset or the complete VAM dataset and evaluated on the EMO-DB dataset.*

Figure 2 displays recognition rates for phonetic-pattern independent non-optimised (an arbitrary number of mixtures in a range from 2 to 120) emotion classifiers trained on emotional samples from the complete VAM dataset (VAM I + VAM II) and the VAM I and VAM II subsets as evaluated on the EMO-DB dataset.

During the third experimental phase we employed phonetic pattern dependent emotional models. Three different sets of phoneme-level emotional models have been trained on the complete VAM dataset and the VAM I and VAM II subsets. Obtained sets of phoneme-level emotional models have been evaluated within cross-corpus analyses on the EMO-DB dataset.

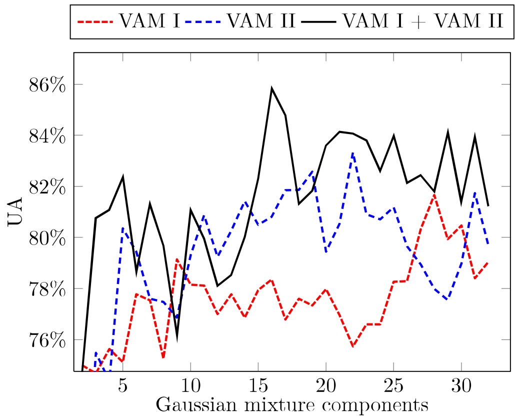

Fig. 3. *UA rates for cross-corpus evaluations. Phonetic-pattern dependent models trained on a subset or the complete VAM dataset and evaluated on the EMO-DB dataset.*

Figure 3 displays recognition rates for phonetic-pattern dependent non-optimised (an arbitrary number of mixtures in a range from 2 to 32) emotion classifiers trained on emotional samples from the VAM dataset (VAM I + VAM II) and VAM I and VAM II subsets and evaluated on the EMO-DB set.

5. CONCLUSIONS

Results presented in Figure 1 show a strong classification gap between emotional models trained on the 'very intense' (VAM I) and 'intense' (VAM II) subsets. Unexpectedly, the general emotional models trained on the VAM II dataset material provide outstanding results on the VAM I dataset material. Corresponding results are considerably better than *baseline* results obtained during LOSGO evaluation. This raises the interesting question whether it is less important to include emotional instances with 'very intense' emotional content to classify such instances correctly.

In order to find an answer to this question one should look at Figure 2. In the case of the second evaluation phase, we used the same models trained on 'very intense' (VAM I) and 'intense' (VAM II) subsets. The corresponding emotional models have been evaluated on acted emotions from the EMO-DB database. Results obtained during cross-corpus evaluation show a steep classification performance gap for emotional models trained on VAM I and VAM II samples. Emotional models trained on 'very intense' (VAM I samples) spontaneous emotions provide outstanding classification performance on acted emotional instances.

Finally, during the third evaluation phase we evaluated phoneme-level emotional models. Phonetic-pattern dependent emotional models trained on 'very intense' and 'intense' emotional samples provide comparable classification performance within cross-corpus evaluation. At the same time, the VAM I dataset ('very intense' emotions) contains a comparable smaller number of instances for each phonetic unit in comparison with the VAM II dataset. Emotional models trained on the complete VAM dataset provide the best (in comparison with VAM I and VAM II) performance. By increasing the number of phoneme instances with very intense emotional instances, one could train more reliable emotion classifiers.

The main outcome of our evaluation experiments (see Figure 1 and Figure 2) is that there is a big classification performances gap for emotional models trained on spontaneous data with different emotional intensity. The second important outcome was obtained during the analysis of cross-corpora evaluation on the EMO-DB dataset. In order to train reliable phoneme-level emotion classifiers one should have sufficient speech samples with very intense emotional content.

6. DISCUSSION AND OUTLOOK

We obtained revealing results within cross-corpus evaluations of emotional models trained on spontaneous data with different intensity. Corresponding models have been evaluated on acted emotions from the same language. We showed that, the phoneme can be seen as possible acoustic unit for classification of the level of emotional arousal. Even with insufficient amount of training data, phoneme-level emotional models outperform general phonetic pattern independent emotional models.

We highlighted the importance of using an emotional dataset with high intensity spontaneous emotions for the training of phoneme-level emotional models. By using sufficient amount of spontaneous emotional instances with intense content one could train robust emotion recognition models with stable classification performance. The combination of 'very intense' spontaneous and acted emotional data could solve the problem of sparse emotional training material.

7. ACKNOWLEDGMENT

The research leading to these results has received funding from the European Unions Horizon 2020 Programme research and innovation programme under grant agreements No. 644632 (MixedEmotions), IMI RADAR-CNS under grant agreement No. 115902 and ERC Starting Grant iHEARu under grand agreement No. 338164.

8. 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. 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.

[3] 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.

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

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

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

[7] 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 *Proc. of Interspeech 2004*, Jeju Island, Korea, 2004, pp. 889–892.

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

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

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

[11] M. Goudbeek, J.P. Goldman, and K. R. Scherer, "Emotion dimensions and formant position," in *Proc. of Interspeech 2009*, Brighton, UK, 2009, pp. 1575–1578.

[12] B. Schuller, B. Vlasenko, D. Arsic, G. Rigoll, and A. Wendemuth, "Combining Speech Recognition and Acoustic Word Emotion Models for Robust Text-Independent Emotion Recognition," in *Proc. of ICME 2008*, Hannover, Germany, 2008, pp. 1333–1336.

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

[14] B. Schuller, B. Vlasenko, R. Minguez, G. Rigoll, and A. Wendemuth, "Comparing One and Two-Stage Acoustic Modeling in the Recognition of Emotion in Speech," in *Proc. of IEEE ASRU 2007*, Kyoto, Japan, 2007, pp. 596–600.

[15] 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.

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

[17] B. Vlasenko and A. Wendemuth, "Annotators' agreement and spontaneous emotion classification performance," in *Proc. of Interspeech 2015*, Dresden, Germany, 2015, pp. 1546–1550.

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

[19] M. Shami and W. Verhelst, "Automatic classification of expressiveness in speech: A multi-corpus study," Computer Science / Artificial Intelligence, pp. 43–56. 2007.

[20] B. Vlasenko and A. Wendemuth, "Determining the smallest emotional unit for level of arousal classification," in *Proc. of ACII 2013*, Geneva, Switzerland, 2013, pp. 734–739.

[21] Daniel Bone, Chi-Chun Lee, and Shrikanth Narayanan, "Robust unsupervised arousal rating: A rule-based framework withknowledge-inspired vocal features," *IEEE Trans. Affective Computing*, vol. 5, no. 2, pp. 201–213, 2014.

[22] J. R. J. Fontaine, K. R. Scherer, E. B. Roesch, and P. C. Ellsworth, "The World of Emotions is not Two-Dimensional," *Psychological Science*, vol. 18, no. 12, pp. 1050–1057, 2007.

[23] K. R. Scherer, "What are emotions? And how can they be measured?," *Social Science Information*, vol. 44, no. 4, pp. 693–727, 2005.

[24] J. A. Russell, "Pancultural Aspects of the Human Conceptual Organization of Emotions.," *Journal of Personality and Social Psychology*, vol. 45, no. 6, pp. 1281, 1983.

[25] M.l 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.

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

[27] B. Vlasenko and A. Wendemuth, "Processing Affected Speech Within Human Machine Interaction," in *Proc. of Interspeech 2009*, Brighton, UK, 2009, pp. 2039–2042.

[28] S. Young, G. Evermann, M. Gales, Hain T., D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book (for HTK Version 3.4)*, 2009.

[29] B. Vlasenko and A. Wendemuth, "Parameter optimization issues for cross-corpora emotion classification," in *Proc. of ACII 2013*, Geneva, Switzerland, 2013, pp. 454–459.

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