Modeling phonetic pattern variability in favor of the creation of robust emotion classifiers for real-life applications

Bogdan Vlasenko *, Dmytro Prylipko, Ronald Böck, Andreas Wendemuth

Cognitive Systems, IESK, Department of Electrical Engineering and Information Technology, & Center for Behavioral Brain Sciences, Otto von Guericke University, D-39106 Magdeburg, Germany

Received 5 March 2012; received in revised form 27 November 2012; accepted 30 November 2012
Available online 20 December 2012

Abstract

The role of automatic emotion recognition from speech is growing continuously because of the accepted importance of reacting to the emotional state of the user in human–computer interaction. Most state-of-the-art emotion recognition methods are based on turn- and frame-level analysis independent from phonetic transcription. Here, we are interested in a phoneme-based classification of the level of arousal in acted and spontaneous emotions. To start, we show that our previously published classification technique which showed high-level results in the Interspeech 2009 Emotion Challenge cannot provide sufficiently good classification in cross-corpora evaluation (a condition close to real-life applications). To prove the robustness of our emotion classification techniques we use cross-corpora evaluation for a simplified two-class problem; namely high and low arousal emotions. We use emotion classes on a phoneme-level for classification. We build our speaker-independent emotion classifier with HMMs, using GMMs-based production probabilities and MFCC features. This classifier performs equally well when using a complete phoneme set, as it does in the case of a reduced set of indicative vowels (7 out of 39 phonemes in the German SAM-PA list). Afterwards we compare emotion classification performance of the technique used in the Emotion Challenge with phoneme-based classification within the same experimental setup. With phoneme-level emotion classes we increase cross-corpora classification performance by about 3.15% absolute (4.69% relative) for models trained on acted emotions (EMO-DB dataset) and evaluated on spontaneous emotions (VAM dataset); within vice versa experimental conditions (trained on VAM, tested on EMO-DB) we obtain 15.43% absolute (23.20% relative) improvement. We show that using phoneme-level emotion classes can improve classification performance even with comparably low speech recognition performance obtained with scant a priori knowledge about the language, implemented as a zero-gram for word-level modeling and a bi-gram for phoneme-level modeling. Finally we compare our results with the state-of-the-art cross-corpora evaluations on the VAM database. For training our models, we use an almost 15 times smaller training set, consisting of 456 utterances (210 low and 246 high arousal emotions) instead of 6820 utterances (4685 high and 2135 low arousal emotions). We are yet able to increase cross-corpora classification performance by about 2.25% absolute (3.22% relative) from UA = 69.7% obtained by Zhang et al. to UA = 71.95%.

Crown Copyright © 2012 Published by Elsevier Ltd. All rights reserved.

Keywords: Emotion phonetic pattern; Emotion phoneme classes; Cross-corpora evaluation; Affective speech; Emotion classification; Level of arousal classification

* This paper has been recommended for acceptance by Dr. Bjorn Schuller.
* Corresponding author. Tel.: +49 3916712739.
E-mail address: bogdan.vlasenko@ovgu.de (B. Vlasenko).

0885-2308/$ – see front matter. Crown Copyright © 2012 Published by Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.csl.2012.11.003
1. Introduction

Recently, attention has focused on speech as a modality from which to deduce information on speaker’s emotional state. Emotion detection and classification are playing an increasingly important role in the user-friendly human–computer interaction. It has been shown in Picard et al. (2001) that recognizing the user’s affective state is an important issue for intelligent human–computer interaction. The last decade has seen a significant number of articles on the acoustic-based emotion recognition. However, until 2009, in comparison to other standard speech processing tasks (like automatic speech recognition (ASR) and speaker recognition or verification) no standardized speech corpora and test conditions were introduced to compare performances under exactly the same conditions. Finally, Schuller et al. (2009) organized the first comparative challenge on emotion recognition from speech – the INTERSPEECH 2009 Emotion Challenge. An impressive set of emotion classification techniques has been evaluated within this challenge (Schuller et al., 2011). The organizers provided an opportunity for 33 registered research groups to evaluate their emotion classification techniques on realistic emotions in speech.

Considering the prevalence of the phoneme-level acoustic models (PLAM) within a developed ASR system, one should likewise aim at finding standardized units for emotion classification which perform well in a broad range of real-world tasks. Existence of the orthoepy, the study of correct pronunciation prescribed for a standard language (German in our case) make possible text-to-phonemes and phonemes-to-text transformation processes. Using the smallest unique phonetic unit, namely the phoneme makes it possible to use trained PLAM for speech recognition from a specified list of words. The quality of phonetic transcriptions and sufficient amount of training material are major requirements within training robust PLAM. In order to obtain phonetic annotations with high “ground truth” measures, involvement of listeners with good language skills (knowledge of the phonetic pattern and phonetic rules) into the transcription process is required. To be able to implement an ASR system for real-life applications a list of words known by the system, together with corresponding phonetic transcriptions, should be specified by a system developer. One could resume, that the existence of the standard phonetic unit (phoneme) determined and classified by “advanced” and “non-advanced” listeners is a key issue of a successful real-world application for ASR systems.

Now, we could draw an analogy between ASR with PLAM and emotion classification systems. The emotion processing community could not yet specify emotional standard units which can be easily determined and classified by any “advanced” and “non-advanced” listener. Determining and annotating of spontaneous (realistic) emotional events with reliable “ground truth” measures is one of the most important questions which is still under research. As a result, there is no methodology which specifies professional skills of an “advanced” emotion annotator. Hence we could argue that in order “to face more realistic scenarios” (Schuller et al., 2009) within future emotion challenges the organizers could provide training and test sets which are at least annotated by different groups of labelers.

The first key point of this article is not related to the improvement of the existing emotion annotation strategies, but a proposal of an experimental setup concept which can face real-life conditions. We propose to train emotional models on two different corpora and test them within cross-corpora evaluation. In the case of reliable emotional annotation and robust emotion modeling one should obtain applicable classification performance within a proposed cross-corpora experimental setup. As a first step of our experiment we evaluate the phonetic-pattern-independent emotion classifier (Vlasenko and Wendemuth, 2009), which got the second rank for the two emotion classes task on the open performance sub-challenge within the INTERSPEECH 2009 Emotion Challenge (Schuller et al., 2011). This experiment showed that the “leading” realistic emotion classification technique could not provide applicable performance within cross-corpora evaluation. In order to improve emotion classification performance within real-world conditions we propose to use phonetic pattern modeling during emotion classification, which is the second key point of this article. In our current experimental results we prove that the most important issue of simulating real-life conditions for emotion recognition evaluations is the usage of training and test sets annotated with different groups of labelers. Also, we assume that the usage of training and test sets annotated with different annotation techniques, namely dimensional (the VAM (Grimm et al., 2008) database) and categorical (the EMO-DB (Burkhardt et al., 2005) database) could make real-world application setups more realistic.

1.1. Background

After we obtained remarkably good results on realistic emotions (Schuller et al., 2011) we decided to evaluate our method within our first cross-corpora evaluation framework (Schuller et al., 2010). Using cross-corpora
evaluation for experiments one could estimate emotion-recognition performance in conditions which are close to real-life development tasks. Several studies already provided accuracies on multiple corpora – however, only a very few consider training on one and testing on a different one (e.g., Shami and Verhelst (2006, 2007), where two and four corpora are employed, respectively). In Zhang et al. (2011) the authors used cross-corpora evaluation, to prove the suitability of unsupervised learning in a large-scale study for speech-based emotion recognition. In Schuller et al. (2010) the authors presented results employing six standard databases in a cross-corpora evaluation experiment. All earlier mentioned cross-corpora evaluations (Schuller et al., 2010; Shami and Verhelst, 2006, 2007; Zhang et al., 2011) provided emotion classification, independent of phonetic patterns, on turn or segment (frame) level. In addition, in Shami and Verhelst (2007) the authors provide a cross-corpora evaluation on the same language and different languages; in Schuller et al. (2010) and Zhang et al. (2011) the authors evaluated their classification techniques on combined and individual training sets with different recording conditions (including different room acoustics, microphone types and positions, signal-to-noise ratios, etc.), languages, and types of observed emotions. Within our first cross-corpora evaluation we found out that techniques which showed outstanding emotion classification, with speaker independent evaluation on the same dataset, are not suitable for real-world applications. After analysis of our experimental results presented in Schuller et al. (2010) we set up a hypothesis that modeling phonetic pattern variability could improve emotion classification within cross-corpora evaluation. For our current evaluation we decided to use emotional speech corpora with the same language (equal phonetic pattern) of recordings. We selected emotional speech corpora which contain recordings of adult German speakers.

Selection of corpora which have reliable emotional annotation is an important issue of our experiments, in general as well as in our setting. There are two main emotion annotation techniques: categorical and dimensional. The categorical technique is based on specification of affective speech samples with a fixed list of emotional states, so called basic emotions (Ortony and Turner, 1990); the dimensional technique specifies an emotion as a point with corresponding coordinates in a multi-dimensional emotion space. The most commonly used dimensions are valence, arousal, and dominance (Davidson, 1999; Russell and Barrett, 1999). Taking into account that in Mauss and Robinson (2009) the authors suggest that there is no “gold standard” measure of assessment of emotions, we decided to use emotional datasets annotated with both emotion labeling techniques. To be able to measure the quality of the emotional annotation, inter-rater reliability measures as an alternative to the “ground truth” have been introduced. To estimate the inter-rater agreement, it is common to use the Kappa coefficient (Carletta, 1996). Still without knowing how good each individual and all annotators together differentiate emotional information in acoustic signals this coefficient cannot tell us how reliable the emotional annotation is in total. It has been shown in Siegert et al. (2011) that emotion labels derived from Geneva Emotion Wheel or Self Assessment Manikins (SAM) provide better emotion coverage and usability within spontaneous emotional labeling. The first database, which we selected for our current experiment was the VAM (recordings from the talk show titled “Vera am Mittag”). This database provides emotion labels annotated with SAM and further processed with evaluator weighted estimators (EWE). The EWE average the individual evaluator (rater) annotation, and take into account that each rater is subject to an individual amount of disturbance during emotion annotation (Grimm et al., 2007). This type of dimensional annotation weighting can be seen as a good alternative of the “ground truth” measurement for realistic emotional corpora. The second database, which we selected, is the EMO-DB (Burkhardt et al., 2005). This database contains acted emotions annotated with a categorical approach. To perform natural emotional expressions, database developers provided two important issues: the database includes naturalness and recognizability measures for each speech sample obtained within perception test; on the first stage of the data collection three expert listeners selected 10 actors out of 40 potential speakers based on the earlier mentioned measures.

In order to assess coincidences of emotional annotations we map various emotion categories, presented in the EMO-DB database, to binary arousal dimensions (positive vs. negative +neutral). Taking into account that in Mauss and Robinson (2009) is claimed that acoustic emotion characteristics are mostly sensitive to arousal, we used only an arousal binary task. At the same time, it has not been shown yet how good acoustic features correlate with the valence dimension (El Ayadi et al., 2011). For example, one could see that in Zhang et al. (2011) the authors obtained results just slightly higher than “selection by chance” for the VAM database in a valence classification task. In Karadoğan and Larsen (2012) the authors showed that the valence dimension is more related to what speakers say, while the arousal dimension is more correlated how speakers say it.
1.2. State-of-the-art and related work

Since the beginning of emotional speech processing (for further references see Scripture, 1921; Scherer, 1986; Whissell, 1989), the usefulness of an automatic recognition of emotions in speech seems increasingly being recognized. This is reflected by the large amount of applications using user-centered human–computer interfaces. Most of these demand sufficient robustness, which may not be given yet, as Shriberg (2005), Schröder et al. (4738), Zeng et al. (2009) and others report. When evaluating the accuracy of emotion-recognition engines, attainable performances are often overrated since usually acted, prompted, or elicited emotions are considered instead of spontaneous, real-world case emotions, which are harder to recognize (cf. Schuller et al., 2010).

Speech-based emotion classifiers used in research apply a broad variety of approaches for feature extraction (Ververidis and Kotropoulos, 2006). Either dynamic analysis (Fernandez and Picard, 2003) for processing on a frame-level or static analysis for higher-level statistical functionals (Ververidis and Kotropoulos, 2004) are established. Among dynamic analyses, HMMs are dominant (Schuller et al., 2003; Lee et al., 2004; Vlasenko et al., 2007). Also, a “bag-of-frames” approach for multi-instance learning is used within dynamic analysis (Shami and Verhelst, 2006). A rarely used alternative is dynamic time warping, supporting easy adaptation. Also, dynamic Bayesian network architectures (Lee et al., 2009) could help to combine features on different time levels. On frame-level those are spectral features, whereas prosodic features are used on supra-segmental level. Related to static analysis, the list of possible classification techniques seems endless: Bayes classifier (Ververidis and Kotropoulos, 2004), multi-layer perceptrons or other types of neural networks (Schuller et al., 2004; Glüge et al., 2011), Bayesian networks (Fernandez and Picard, 2003; Cohen et al., 2003), Gaussian mixture models (Slaney and McRoberts, 1998; Lugger and Yang, 2007), random forests (Liou and Anagnostopoulo, 2009), decision trees (Lee et al., 2009), k-nearest neighbor distance classifiers (Dellaert et al., 1996), and support vector machines (SVM) (Fernandez and Picard, 2003; Batliner et al., 2006; Montacié and Caraty, 2011) are applied most often. Also, a selection of ensemble techniques (Schuller et al., 2005; Morrison et al., 2007) has been used, like bagging, boosting, multi-boosting, and stacking with and without confidence scores. Recently developed techniques such as hidden conditional random fields (Wöllmer et al., 2008), long-short-term-memory recurrent neural networks (Wöllmer et al., 2008) and support vector machines (Kockmann et al., 2009) could further be seen more frequently in near future.

So far, in the research community the focus was laid on prosodic features extracted on turn-level. In particular, these feature sets (from 10 to 100 features) include, among others, durations, intensity and pitch (Lee et al., 2004; Cairns and Hansen, 1994; Banse and Scherer, 1996; Li and Zhao, 1998). Only a few studies applied low-level feature modeling on a frame-level as an alternative, usually by hidden Markov models or Gaussian mixture models (Schuller et al., 2003; Vlasenko and Wendemuth, 2007). The higher success of static feature vectors is probably proved by the supra-segmental nature of the phenomena which is related to the emotional content within a speech signal (Schuller et al., 2009, 2009). Those static features are derived from low-level contours (like energy or pitch), by descriptive statistical functional applications like lower order moments (mean, standard deviation), or by specifying extremal values (Ververidis and Kotropoulos, 2004). Currently, voice quality features such as shimmer, jitter, or harmonics-to-noise ratio (HNR), and spectral and cepstral features such as formants and MFCCs have become the “new standard” feature sets as discussed for instance in Lugger and Yang (2007), Barra et al. (2006), Schuller et al. (2009). Traditionally, prosodic acoustic features, which can be classified in different ways, have been applied for affective speech processing as categorized by Batliner et al. in Batliner et al. (2011): Selection of feature sets. The ‘selective’ approach is based on phonetic and linguistic knowledge (Kießling, 1996) which is also known as ‘knowledge-based’. It has a strict systematic strategy for generating the features using a constant set of functions, which are applied to time series of different acoustic markers. This approach normally results in more than 1000 features per set. Another approach is based on brute-forcing of features (1000 up to 50,000) by analytical feature generation, partly also in combination with evolutionary generation (Schuller et al., 2008). The difference between the two approaches lies in the feature selection step: in the selective approach, the selection takes place on an empirical level before putting the features into the classification process; in the brute-force approach an automatic feature selection is required.

Staging of feature extraction. There is a “two-layered” approach, where at first features are computed on the words level; secondly, functionals such as mean values and the average values are computed for all words within one utterance. An alternative is a “single-layered” approach, where features are computed for the complete utterance.
Most of the participants of the open performance sub-challenge within the Interspeech 2009 Emotion Challenge used quite small feature sets. Surprisingly, most of them were mainly based on MFCC features, which are well-known from speech recognition and have been proven to be very useful in emotion recognition as well (cf. e.g. Schuller et al., 2011). In our current research, we decided to use MFCC features as well, modeling on frame-level for emotion recognition from speech which provides us with a flexible modeling of emotional intensity variability. Moreover, we used for this purpose continuous-density HMMs architecture.

Selection of classification paradigm. Further, in general, there are two pre-dominant emotion classification paradigms: dynamic modeling on a frame-level by means of hidden Markov models (HMMs) and suprasegmental (static) modeling by systematic feature brute-forcing (Schuller et al., 2009). Taking into account that most emotional corpora contain emotional annotation on the turn level, it is not a surprise that the latter approach shows quite good emotion classification performance. At the same time most developers of emotional corpora with utterance-level annotation cannot prove that the emotional intensity is equally distributed. With static turn-level analysis we get an acceptable performance for near-real-case emotion classification which will decrease within cross-corpora evaluation (Schuller et al., 2010). Suprasegmental modeling in comparison with frame-level modeling can provide sufficient flexibility for modeling emotional intensity variability within an utterance, hence we are using such modeling for our research as well. Still, when using static analysis, it has been shown in Schuller and Rigoll (1821) that additional sub-phrase level information can improve performance of the acoustic emotion classification. In our earlier publication (Schuller et al., 2007) we presented an emotion classification method on sub-turn entry level. We applied two different types of entries: syllables and quasi-stationary segments extracted by one-pass Viterbi beam search with token parsing based on MFCCs. Within experiments on the EMO-DB database we found out that using only segment-level information resulted in lower emotion classification performance of the general, phonetic pattern independent, suprasegmental modeling. Later we showed in Vlasenko et al. (2008), Schuller et al. (2008) that unit-specific emotion models clearly outperformed common general models, provided enough training material per unit.

Most state-of-the-art emotion recognition methods are based on turn- and frame-level analysis, independent of the phonetic transcription. Fragopanagos and Taylor (2005) state that most research efforts investigated the affective speech processing on the level of complete utterances, words, or phonetic transcription independent chunks. According to Batliner et al. (2010) words can be seen as the smallest possible chunk for analysis. A comparably smaller number of methods are based on phonetic pattern modeling within emotion classification. Goudbeek et al. presented their research on the effect of emotion dimensions on formant placement in individual vowels (Goudbeek et al., 2009). In affective speech synthesis, Inanoglu et al. developed a set of fundamental frequency (F0) conversion methods at syllable level which were evaluated for three target emotions: anger, surprise, and sadness (Inanoglu and Young, 2009). Furthermore, an emotion classification test showed that utterances converted with either F0 generation techniques could convey the desired emotion classes with accuracies above chance level. Bulut et al. studied phoneme-level signal property manipulation in transforming the emotional information conveyed in a speech utterance (Bulut et al., 2005). Busso et al. showed that the mean and the variance of the likelihood score for emotional speech differ from the results observed in neutral speech, especially for emotions with a high level of arousal and observed in some broad phonetic classes (front vowels and mid/back vowels) which present stronger differences than others (Busso et al., 2007). Lee et al. showed an acceptable speech-based emotion recognition performance using phoneme-class dependent HMM classifiers with short-term spectral features (Lee et al., 2004). In Montacié and Caraty (2011) the authors investigated phonetic variations of speech produced by intoxicated speakers, and they presented gender-dependent phoneme-based SVM classifiers. In Gajšek et al. (2012) an efficient approach to modeling the acoustic features for the emotion and intoxication recognition is presented. This approach used methods from the two pre-dominant emotion classification paradigms and combined features extracted on complete turn and monophone-based segments. Still, most of the aforementioned phoneme-level modeling emotion classification techniques used forced alignment or manual annotation for the extraction of the phoneme borders. Just some methods faced real-life conditions by using ASR engines for generating the phoneme alignment (Gajšek et al., 2012). Current ASR techniques are not able to provide as good phoneme alignment on affective speech samples as manual annotation or forced alignment. To properly address real-world conditions, a phoneme-level emotion processing method presented in this article relies on the phoneme alignment obtained by using an ASR system which applies acoustic models adapted on affective speech samples.
2. Selected databases

In order to provide a comparison of the phonetic pattern dependent and independent emotion classification performances we need to induce specific requirements to the emotional corpora. The requirements for the datasets are the following:

- **Speakers** should be in the same age group.
- The same **language** is to be used for both datasets in cross-corpora analysis.
- **Transcriptions** should be available in sufficiently high quality.
- Both datasets should have similar emotional **categories** ("basic" emotions or predominant types of general emotion categories, namely, high/low arousal)
- Datasets should contain a reliable emotional **annotation**.

To be able to obtain reliable phoneme alignments we decided to select corpora which contain speech material recorded from speakers of a common age group (adult speakers). Taking into account that German language is one of the most popular language of recordings of publicly available well annotated datasets, we decided to select German phonetic patterns for our experiment. For training of the ASR engine, which we used for phoneme alignment, the Kiel Corpus of Read Speech (Kohler, 1996) was selected. For evaluation of our emotional models we selected two databases that matched our requirements: VAM (Grimm et al., 2008) and EMO-DB (Burkhardt et al., 2005).

The VAM database consists of 12 h 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 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 used for annotation (17 labelers for one half of the data, six for the other). 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 EWE. During mapping of the original dimensional annotations into a two class problem (level of arousal $>0$ vs. $\leq 0$) we obtained 502 low and 445 high arousal emotional utterances.

To show a distribution of the emotional instances presented in the VAM dataset we mapped the non-binarized labels processed with EWE into valence-arousal (VA) space. As one can see from Fig. 1, the major part of emotional instances is located in the negative valence subspace and just few samples, mostly in the negative arousal subspace, correspond to the positive emotions (have positive valence). A comparably small number of the training samples for positive valence was the additional reason why we have trained our classifiers just for arousal discrimination task.

The second database, which we selected for our experiment is the EMO-DB (Burkhardt et al., 2005) which covers anger, boredom, disgust, fear, joy, neutral, and sadness speaker emotions. Ten (five female) professional actors speak ten German sentences with emotionally neutral linguistic meaning. In order to sustain our requirements for the reliability of the emotional content, we selected utterances which have a level of naturalness not less than 60% and a level of recognizability not less than 80%. In order to specify emotional categories which can be modeled on the speech material presented in both datasets, we investigated possibilities to map the aforementioned emotional states to the predominant type of general emotion categories, namely, high/low arousal. We found two graphs which reflect possible location of some emotion categories on VA space (Fontaine et al., 2007; Scherer, 2005). The first graph (Scherer, 2005) was created by the mapping of the terms which (Russell, 1983) uses as markers for his claim of an emotion circumplex. On the second graph (Fontaine et al., 2007) 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 graph (see Scherer, 2005) and negative arousal sub-space for the second graph (see Fontaine et al., 2007), we decided to eject disgust instances from our experimental dataset. As a result, for our current 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 class referred as low-arousal emotion. In total, we used 456 utterances (210 low and 246 high arousal emotions) with defined measures received within perception tests.
To train basic HMMs with Gaussian mixture models (GMMs) for the ASR engines we used the Kiel Corpus of Read Speech (Kohler, 1996). Later, the basic ASR models have been adapted on the affective speech samples in order to provide better emotion speech processing. 1041 utterances for 6 female speakers and 1033 utterances for 6 male speakers were used for our experiments. Breathing and mechanical noises were manually deleted from the selected acoustic material to reach a qualitative estimation of acoustic parameters. By such data quality improvements, we compensated for a comparably small amount of training samples for the training of a robust monophone-based ASR system. The data quality improvement can be automatized, if reliable noise detection techniques are made available. Given those techniques, models for different types of noises can be built, which allow the HMMs’ architecture to identify noise segments. In the same way as described above, these noisy segments can be removed from the data, leading to improvement of the trained classifiers. This method will be applicable on real-life applications where annotations are not available.

3. German phonetic pattern and numbers of evaluated phonetic instances

A phoneme is the smallest acoustic component of speech and it is widely used as the sub-word unit for automatic speech recognition. The German language contains 13 unreduced vowels, 2 reduced vowels, 3 diphthongs, 6 plosives, 9 fricatives, 3 nasals, and 2 liquids. The acoustic segments of the vowels are usually long in duration and are spectrally well represented. They are generally reliably and easily recognized by human beings and by ASR systems (Rabiner and Juang, 1993).

Table 1 shows the total number of German vowel (unreduced, reduced, diphthongs) instances presented in the selected speech datasets and their corresponding SAM-PA and IPA symbols.

4. Experimental setup

Within our experiments we decided to use two different evaluation measures to present emotion-recognition performances: unweighted average recall (UA) and weighted average recall (WA). Unweighted average recall is the sum of all class accuracies, divided by the number of classes, without considering the number of instances per class. Weighted average recall, also known as accuracy, is the accuracy per class, including consideration of the number of instances per class. In other words WA (accuracy) is the number of instances with correctly classified classes, divided by the total number of classified instances. As the numbers of emotional instances in the selected speech corpora (see Section 2)
are unbalanced, the primary measure which should be taken into account is unweighted average recall. In a case of equal classification scores for both emotional states we mark a test utterance as “unrecognized”.

During the emotion recognition phase we applied straightforward recognition on turn level or sequential emotion processing on each phonetic unit. In order to map emotion classification results obtained on a phonetic unit (phoneme or word in our case) onto the turn-level, we consider three strategies known from multi-instance learning for each matching segment and phonetic unit at the same time:

- **An un-weighted majority vote (MV)**: we counted the number of units assigned for each emotional state. The emotional state which had a maximum number of units was selected.
- **A maximum length vote (MLV)**: we counted the total durations of units assigned for each emotional state. The emotional state which had the largest total duration was selected.
- **Maximum classifier prediction score multiplied with the length vote (MSL)**: we summarize the classifier prediction score multiplied with the length vote for each emotional unit presented in the turn. The emotional state which had the largest estimated sum was selected.

Within our current experiments we used speaker independent (SI) evaluations for tuning parameters and cross-corpora evaluation for the final experiments on the classifiers with the obtained optimal parameters. Also we used a SI evaluation strategy to show discriminative characteristics of the indicative vowels (cf. Section 5.3.2) within emotion classification with emotional phoneme classes. To address speaker independence within our evaluations we applied leave-one-speaker-out (LOSO) or leave-one-speakers-group-out (LOSGO) strategies. In the case of 10 or less speakers in a corpus, namely EMO-DB, we apply the LOSO strategy; otherwise, namely VAM, we select 5 speaker groups with nearly equal amount of male as well as female speakers and samples per group for LOSGO evaluation.

5. Emotion recognition methods

In our research we applied low-level feature modeling on frame-level for emotion recognition from speech. The continuous-density HMMs have been used for this purpose. We implemented a multivariate Gaussian mixture model (GMM) for modeling production probabilities. Two different units of analysis have been investigated for dynamic analysis: turn and phoneme. Turn-level modeling or so-called general emotion models do not take into account the phonetic content of the analyzed speech sample. Phoneme-level emotion modeling relies on the phonetic characteristics of the analyzed speech sample.

5.1. Acoustic feature extraction

The speech signal is processed using a 25 ms Hamming window, with a 10 ms shifting step. As in typical speech recognition system we employ a 39-dimensional feature vector per each frame consisting of 12 MFCC and zero-order cepstral coefficient plus delta and delta-delta (acceleration) coefficients. Cepstral mean subtraction (CMS) is applied to better cope with channel characteristics.
5.2. General phonetic pattern independent emotion classification: turn-level analysis

We consider using a statistical analysis applied to ASR to recognize emotions from speech (Vlasenko and Wendemuth, 2009). Instead of the common task to deduce the most likely word sequence hypothesis \( \Omega_k \) from a given vector sequence \( \Omega \) of acoustic observations \( o \), the task is to recognize the current speaker’s emotional state. This is solved using a stochastic approach applied to ASR, namely Bayesian decision rule, with a different argument interpretation:

\[
\Omega_k = \arg \max_{\Omega} \log P(\Omega | \Omega) = \arg \max_{\Omega} \log \frac{P(\Omega | \Omega) P(\Omega)}{P(\Omega)}
\]

where \( \Omega \) is one of all system known emotions (“low arousal” and “high arousal” emotional states in our case); \( P(\Omega | \Omega) \) is the emotion acoustic model; \( P(\Omega) \) is the prior user-behavior information.

In the case of turn-level analysis, the emotion acoustic model consists a fixed number of HMMs’ states \( n \). Each state is modeled with acoustic driven analysis which does not consider phonetic units alignment. Taking into account that the selected corpora contain emotion annotation on turn level, one emotion is assigned for a complete utterance. In other words, within the training and testing observation the feature vector sequence \( \Omega \) contains all feature vectors extracted from a complete utterance. In the more general case, our HMMs-based dynamic classifier can detect several emotional chunks within the same utterance.

In simple cases the priors in the user-behavior model \( P(\Omega) \) have an equal distribution among emotion classes (in our case \( P(\text{“high arousal”}) = P(\text{“low arousal”}) = 0.5 \). It is also possible to implement phonetic transcription and a user-behavior model dependent on the history of emotional states (in our case \( P(\text{“high arousal”}) = \alpha \), \( P(\text{“low arousal”}) = 1 - \alpha \), where \( 0 < \alpha < 1 \). The weighting coefficient \( \alpha \) can be estimated using semantic analysis presented in Karadogan and Larsen (2012). This semantic analysis could be designed by combination of the Latent Semantic Analysis (LSA; Landauer et al., 1998) and Affective Norms for English Words (ANEW; Bradley and Lang, 2012). With ANEW and LSA techniques we could measure the arousal for each word in the test utterance. Afterwards this measurement could be used for estimation of the weighting coefficient \( \alpha \). Taking into account that we selected emotional corpora with German speech material (we could not implement ANEW for German words), we decided to use the simple behavior model for our experiments. Keeping in mind that we use equal priors for both emotional states and assuming that \( P(\Omega) \) is not correlated with the user’s emotional state, Eq. (1) can be simplified:

\[
\Omega_k = \arg \max_{\Omega} \log P(\Omega | \Omega) P(\Omega) = \arg \max_{\Omega} \log P(\Omega | \Omega, \mathcal{M}) P(\Omega)
\]

where \( \mathcal{M} \) is an HMMs’ parameter set. The HMMs’ parameter set \( \mathcal{M} = (\pi, A, B) \) consists of the following parameters (Rabiner, 1989): \( \pi \) – initial state distribution; \( A \) – state transition probability matrix; \( B \) – observation generation probability distribution. Corresponding parameters specify acoustic characteristics of emotion phonemes pronounced with high and low arousal.

The HTK toolkit (Young et al., 2009) was used to build these models, using standard techniques such as forward-backward and Baum–Welch re-estimation algorithms. In Section 6, we determine the optimal classifier architecture (state number \( n \) and number of Gaussian mixture components).

5.3. Phonetic pattern dependent emotion classification: emotional phoneme classes

In the case of phonetic pattern dependent emotion modeling, Eq. (1) is slightly modified:

\[
\Omega_k = \arg \max_{\Omega_{\text{pho}}} \log P(\Omega | \Omega_{\text{pho}}) P(\Omega_{\text{pho}})
\]

The arguments of Eq. (3) are interpreted in the following way:

\( \Omega_{\text{pho}} \) is a possible phoneme emotion sequence (fixed phoneme states sequence, obtained with ASR system) for \( \text{pho} = \text{“low arousal”} \) or \( \text{pho} = \text{“high arousal”} \) emotional state in our case; \( P(\Omega | \Omega_{\text{pho}}) \) is an emotion acoustic model for the emotion phoneme states sequence \( \Omega_{\text{pho}} \); \( P(\Omega_{\text{pho}}) \) is a priori knowledge about the affective state frequency of occurrence for the phonetic units sequence \( \Omega_{\text{pho}} \).
In simple cases the priors in the phonetic-unit-specific emotion model $P(\Omega_{\text{pho}})$ have an equal distribution among emotion classes (in our case $\forall 2^\Omega$: $P(\Omega_{\text{high arousal}})=P(\Omega_{\text{low arousal}})=0.5$). It is also possible to provide phonetic transcription and a prior information modeling dependent on the history of emotional states (see previous section). In our current research we applied a simple prior modeling technique and HMMs with left-to-right topology and three emitting states for emotion acoustic modeling. In this case Eq. (3) can be simplified:

$$
\Omega_k = \arg\max_{\Omega_{\text{pho}}} \log P(O|\Omega_{\text{pho}}) = \arg\max_{\Omega_{\text{pho}}} \log P(O|\Omega_{\text{pho}}, M_{\text{pho}})
$$

$$
= \arg\max_{\Omega_{\text{pho}}} \log \sum_{\text{States sequence } s} P(O, s|\Omega_{\text{pho}}, M_{\text{pho}})
$$

where $M_{\text{pho}}$ is a phoneme-level HMMs’ parameter set, $s = [s_1, s_2, \ldots, s_T]$ is a state sequence associated with on observation vector sequence $O = [o_1, o_2, \ldots, o_T]$.

In the case of phonetic level modeling, instead of processing on acoustic driven states, we use phoneme states. The parameter set $M_{\text{pho}}$ consists of parameters which specify “low-arousal” and “high-arousal” emotion phonemes. Namely, the full lists of phonemes (36 phonemes for the EMO-DB database and 39 phonemes for the VAM database) 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 and $2 \times 39 = 78$ emotional phoneme models on the VAM.

In order to simplify the emotion classification process we decided to use a fixed phoneme states sequence with corresponding optimal state sequence $\omega = [\omega_1, \omega_2, \ldots, \omega_T] = [s_{1\text{opt}}, s_{2\text{opt}}, \ldots, s_{T\text{opt}}]$. To specify a fixed phoneme states sequence we used an ASR engine to recognize phoneme states sequences, see Fig. 2. More detailed specification of the applied ASR engine can be found in the following section. Taking into account that the implemented ASR system architecture is identical to the phoneme-level emotion recognition system, we can use an optimal state sequence for our phoneme-level emotion modeling. Let us assume that $\omega$ is the state sequence associated with the recognized phoneme states sequence $\Omega_{\text{pho}}$. With a defined optimal state sequence we simplify the maximization task represented in Eq. (4) by estimation of $P(O, s|\Omega_{\text{pho}})$ just for optimal state sequence. In this case, implemented in our current research, the classification criteria can be expressed as:

$$
\Omega_k = \arg\max_{\Omega_{\text{pho}}} \log P(O, \omega|\Omega_{\text{pho}}, M_{\text{pho}}) = \arg\max_{\Omega_{\text{pho}}} \{ p(\omega|\Omega_{\text{pho}}, M) p(O|\omega, M_{\text{pho}}) \}
$$

$$
= \arg\max_{\Omega_{\text{pho}}} \{ \prod_{t=1}^{T} b_{\omega t}(o_t) a_{\omega t-1\omega t} \} = \arg\max_{\Omega_{\text{pho}}} \left\{ \log \pi_{\omega 1} + \sum_{t=1}^{T} \log b_{\omega t}(o_t) + \sum_{t=1}^{T} \log a_{\omega t-1\omega t} \right\}
$$

$$
= \arg\max_{\Omega_{\text{pho}}} \left\{ \sum_{t=1}^{T} \log b_{\omega t}(o_t) + \sum_{t=1}^{T} \log a_{\omega t-1\omega t} \right\}
$$

Fig. 2. Phoneme-level emotion recognition with emotional phoneme classes.
Considering an initial state distribution \( \pi \), state transaction probabilities \( a_{ij} \), observation generation probability distributions \( b_i(o) \), we estimate two main multipliers \( p(\omega|\Omega_{\text{pho}}, M) \) and \( p(O|\omega, M_{\text{pho}}) \). The first one is the probability of passing optimal state sequence \( \omega \), the second one is the probability of observing acoustic feature vector sequence \( O \) given state sequence \( \omega \). 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 \( M^{\text{ubm}}_{\text{pho}} \) on emotionally neutral speech samples from the Kiel Corpus of React Speech (Kohler, 1996). In the second step, we adapt \( M^{\text{ubm}}_{\text{pho}} \) with combined Maximum Likelihood Linear Regression (MLLR) (32 regression class trees) + Maximum a Posteriori (MAP) adaptation (hyper-parameter \( \tau = 2 \)). The chosen adaptation parameters provide the most robust emotional speech recognition, as shown in Vlasenko et al. (2012).

Taking into account that to date the emotion processing community could not specify emotional standard units which can be easily determined and classified by any “advanced” and “non-advanced” listener, we are not in the position to say if one, several or all “emotionally colored” phonemes contained in a given word “make it sound” emotional. The same argument can be projected on the turn level by measuring the emotional intensity of the words within an utterance. As a result, we decided to estimate the conditional probability in Eq. (5) on two possible segment types: word, or phoneme. We divided the optimal state sequence \( \omega \) into a set of subsequences, which corresponds to an alignment of the phonetic units (phoneme or word). The corresponding phoneme alignments have been obtained by implementation of a phoneme- or word-level language model (LM) for the ASR engine. Afterwards, we estimated the corresponding conditional probabilities on all possible state sub-sequences and provided segment-level emotion classification. Finally, we map the emotion classification results obtained using phonetic units (phonemes or words in our case) onto the turn-level, by using the three strategies presented in Section 4. Within the straightforward emotion recognition process, we select the emotional state which has a higher probability of passing the optimal phoneme states sequence \( \omega \) and of observing the acoustic feature vector sequence \( O \) as specified in Eq. (5). In Section 6, we specify the optimal number of Gaussian mixture components for the continuous-density distribution function used for modeling of the state output probability distribution.

### 5.3.1. Automatic speech recognition engine

The most robust and general acoustic technique for automatic speech recognition are HMMs. We applied a continuous-density HMMs technique based on a 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 and MAP adaptation. For the MLLR, optimal affective speech recognition performance was obtained with 32 regression classes where only means have been transformed. For the MAP, optimal performance was obtained with \( \tau = 2 \) (Vlasenko et al., 2012). All other ASR parameters (like word insertion penalty, language scale factor, etc.) have not been optimized and used as “standard” values defined in HTK examples. Within our cross-corpora evaluations, for decoding the optimal state sequence for the test dataset samples we used acoustic models adapted on emotional speech samples from the training dataset. For example, for cross-corpora evaluation on the EMO-DB speech samples we used ASR models adapted on the emotional speech samples from the VAM database.

For phonetic representation of the words in our ASR vocabularies we used a simplified version of BAS SAM-PA (Bavarian, 1996) with a set of 39 phonemes (18 vowels and 21 consonants). In order to implement ASR a phoneme-level transcription is needed, which requires a corresponding lexicon containing phonetic transcription of words presented in a corpus. Since the VAM corpus does not provide such a lexicon, we created it by ourselves using a combined approach. The major part of the word transcriptions (1216 items) was taken from other German corpora, namely Verbmobil (Hess et al., 1995) and SmartKom (Schiel et al., 2002). For the rest (688 words) we created transcriptions using grapheme-to-phoneme conversion with the Sequitur G2P converter (Bisani and Ney, 2008). Using G2P also faces real-life conditions, namely, we could not provide manual phonetic transcriptions which could occur within a possible human–computer interaction. The converter was trained on a combined lexicon based on SmartKom and Verbmobil lexicons (12,460 German words at all). Before applying the G2P software to the missing VAM lexicon, we tested it on the combined lexicon, where 1% randomly selected words were moved into the test set. The phoneme error rate was 5.33% (56 from 1050), the word error rate 29.13% (37 from 127). For language modeling (LM) we used bi-gram and zero-gram models implemented on word level. Zero-grams assumes that all the words (in the dictionary) have the same a-priori probability. In order to solve a problem with out-of-vocabulary (OOV) words we applied a bi-gram language modeling on phoneme level.
5.3.2. Using only indicative vowels for emotion classification

In the case of using only indicative vowels for phonetic pattern dependent emotion classification, the HMMs’ parameter set \( \mathcal{M}_{\text{pho}} \) specifies only \( 2 \times 7 = 14 \) emotional phoneme models for both databases. In our previous research (Vlasenko et al., 2011), we showed that a convenient Neyman–Pearson criterion with only one average F1 value extracted from a short list of vowels could provide remarkably good discriminative characteristics within straightforward detection of high arousal acted and spontaneous emotions. We called these vowels “indictive vowels” ([a, e, E, @, 6, a, u]). In order to show emotion discriminative characteristics of the indicative vowels, we decided to apply an emotional phoneme class analysis only to seven indicative vowels. In Section 7.1 we present a comparison of the emotion classification performances obtained with emotional phoneme classes modeling with a complete list of phonemes, and with indicative vowels only.

6. Parameter tuning

In order to estimate optimal parameters for the classification techniques presented in Section 5 we applied speaker-independent strategies, namely LOSO and LOSGO. The corresponding parameters are: (a) number of HMM states for the phonetic pattern independent emotion classification technique; (b) number of Gaussian mixture components. In this section we specify these parameters of the implemented emotional models which provide the best classification performance on each emotional dataset.

At first, we estimate optimal parameters for the phonetic pattern independent classifier. In our previous article (Vlasenko et al., 2007) we showed that HMMs with single-state architecture and 90 Gaussian mixture components show the most stable and robust results for the seven classes’s emotion classification task on EMO-DB’s speech samples. In order to prove robustness of a single-state architecture on a two-class problem we estimated the emotion classification performance as a function of the numbers of Gaussian mixture components (values are in range from 2 to 120) for single or multiple states (\( n = 1–5 \)) architectures. For this, we decided to use spontaneous emotions presented in the VAM database.

As can be seen in Fig. 3, single-state [\( n = 1 \)] HMMs show the most stable and robust results. The best classification performance on the VAM’s speech samples has been reached with 77 Gaussian mixture components. Within evaluation of the single state architecture on the EMO-DB’s speech material the best classification performance has been reached with 117 Gaussian mixture components. In our evaluations of the phonetic pattern independent classifiers, we followed these findings and used a single state architecture with corresponding number of Gaussian mixture components (77 for emotional models trained on VAM, and 117 for emotional models trained on EMO-DB).

Finally, we estimated optimal parameters for phonetic pattern dependent emotion classifiers. In order to find an optimal value, we evaluated classifiers which used numbers of the Gaussian mixture components in a range from 2 to 32. Within LOSO and LOSGO evaluations we found out that the optimal classification performance has been reached with 29 components for the VAM database and 31 components for the EMO-DB database. In our evaluations of the phonetic pattern dependent classifiers, we again followed these findings and implemented phoneme-level modeling with 29 Gaussian mixture components for the emotional models trained on the VAM speech samples and 31 Gaussian mixture components for the emotional models trained on the EMO-DB speech samples.
Table 2

<table>
<thead>
<tr>
<th>Database</th>
<th>Word level zero-gram</th>
<th>Word level bi-gram</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>UA</td>
<td>WA</td>
</tr>
<tr>
<td>Complete phoneme set</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EMO-DB</td>
<td>92.90</td>
<td>92.98</td>
</tr>
<tr>
<td>VAM</td>
<td>69.11</td>
<td>69.59</td>
</tr>
<tr>
<td>Only indicative vowels</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EMO-DB</td>
<td>87.94</td>
<td>88.16</td>
</tr>
<tr>
<td>VAM</td>
<td>65.85</td>
<td>66.10</td>
</tr>
</tbody>
</table>

7. Experiments

In this chapter we present a comparison of emotion recognition results obtained with two different setups: (a) LOSO evaluation using emotion phoneme classes with complete phoneme set vs. indicative vowels set and (b) cross-corpora evaluation of general (phonetic pattern independent) and emotional phoneme classes based emotion classification techniques. All experiments concerning our emotion-recognition techniques are performed with the optimal classification parameters obtained within the speaker independent experiment presented in Section 6.


For our first experiment we decided to implement the method introduced in Section 5.3 for the complete phoneme set and for indicative vowels only. Within our evaluation we used word-level language modeling and LOSO strategy.

The best emotion classification performances on the utterance level were obtained with bi-gram language models. The main outcomes of the results presented in Table 2 are that the indicative vowels analysis (UA on EMO-DB about 88.62%, UA on VAM about 67.56%) with emotional phoneme classes performed comparably close to the emotion recognition with the complete phoneme set (UA on EMO-DB about 92.86%, UA on VAM about 71.85%). Also we compared our best phoneme-level emotion classification for the VAM database result with the best result for general models, UA = 76.46% obtained with the parameters tuning in Section 6. The phonetic pattern independent classifier outperforms our phoneme-level emotion models, which is due to the comparably high level of incorrect transcriptions in part of the G2P generated lexicon and quite high levels of ASR driven transcription errors.

7.2. Cross-corpora evaluation

In this chapter we present results of cross-corpora experiments concerning our emotion-recognition techniques with optimal classification parameters (see Section 6).

7.2.1. Phonetic pattern independent models

For our first cross-corpora evaluation we employed phonetic pattern independent (general) emotion classifiers.

As can be seen in Table 3, general models could not provide sufficient emotion classification performance in real-world conditions, namely cross-corpora evaluation. Within parameters tuning and also in our previous emotion
classification benchmark evaluation (Schuller et al., 2009) we showed that an continuous-density HMMs based on MFCC features provided classification performance about UA = 76.46% in a LOSO evaluation. Using “full-blown” emotions presented in the EMO-DB database for training general emotion models resulted in 9.26% absolute (relative 12.11%) classification performance declining, in evaluation on spontaneous emotions presented in the VAM dataset. As a result, we can conclude that the phonetic content variability presented in the EMO-DB and VAM should be modeled within the emotion classification processes in order to improve recognition performance in real-life evaluations. With the experimental result presented in the next section we want to prove our hypothesis.

7.2.2. Phonetic pattern dependent models

Now we evaluate the method introduced in Section 5.3 in conditions close to real-world scenarios. To find the score unit for best recognition, we implemented phoneme- and word-level straightforward classification.

As one can see from Table 4, using phonemes as score units provided the best emotion classification performance. By using phoneme-level emotion models trained on the VAM database material we were able to recognize emotions from the EMO-DB database with superior classification performance of UA = 83.94%. Interchanging training and recognition databases we obtained classification performance of UA = 71.92%. It can be seen from Table 4 that the phonetic pattern dependent emotion models could improve classification performance in real-life conditions.

In Table 5 we present experimental results obtained with additional approaches, appending the straightforward classification technique. In order to map emotion classification results obtained on various segment types (phoneme and word) onto the turn level, we consider three strategies (MV, MLV, and MSL) known from multi-instance learning for each segment, and we apply language modeling on phoneme-level and word-level.

As can be seen in Table 5 emotion classification performances obtained with two different language models are comparable. The best emotion classification performances were obtained with maximum length vote strategy for phoneme-level matching segments (VAM: UA = 71.95%, EMO-DB: UA = 82.99%) and maximum classifier prediction score multiplied with the length vote for word-level matching segments and bi-gram language models (VAM: UA = 70.92%, EMO-DB: UA = 83.02%).

### Table 4


<table>
<thead>
<tr>
<th>Database</th>
<th>Train</th>
<th>Test</th>
<th>Phoneme level bi-gram</th>
<th>Word level zero-gram</th>
<th>Word level bi-gram</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>UA</td>
<td>WA</td>
<td>Unr.</td>
</tr>
<tr>
<td>VAM</td>
<td>EMO-DB</td>
<td>83.94</td>
<td>84.21</td>
<td>0.0</td>
<td>81.98</td>
</tr>
<tr>
<td>EMO-DB</td>
<td>VAM</td>
<td>71.92</td>
<td>72.22</td>
<td>0.0</td>
<td>69.09</td>
</tr>
</tbody>
</table>

### Table 5

Classification performance [%] of the emotional phoneme-classes analysis with word or phoneme emotion matching segments mapped onto turn level with (MV, MLV, MSL). Evaluated on the EMO-DB and VAM corpora with cross-corpora strategy. Abbreviation: Unr. – unrecognized.

<table>
<thead>
<tr>
<th>Database</th>
<th>Train</th>
<th>Test</th>
<th>LM</th>
<th>MV</th>
<th>MLV</th>
<th>MSL</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>UA</td>
<td>WA</td>
<td>Unr.</td>
<td>UA</td>
</tr>
<tr>
<td>Phoneme-level language modeling</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VAM</td>
<td>EMO-DB</td>
<td>Bi-gram</td>
<td>79.62</td>
<td>80.26</td>
<td>3.07</td>
<td>82.99</td>
</tr>
<tr>
<td>EMO-DB</td>
<td>VAM</td>
<td>Bi-gram</td>
<td>69.53</td>
<td>69.69</td>
<td>2.96</td>
<td>71.95</td>
</tr>
<tr>
<td>Word-level language modeling</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VAM</td>
<td>EMO-DB</td>
<td>Zero-gram</td>
<td>77.13</td>
<td>76.97</td>
<td>7.89</td>
<td>81.57</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Bi-gram</td>
<td>80.48</td>
<td>80.70</td>
<td>7.24</td>
<td>82.35</td>
</tr>
<tr>
<td>EMO-DB</td>
<td>VAM</td>
<td>Zero-gram</td>
<td>66.73</td>
<td>67.27</td>
<td>3.27</td>
<td>69.28</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Bi-gram</td>
<td>66.74</td>
<td>67.16</td>
<td>6.76</td>
<td>70.88</td>
</tr>
</tbody>
</table>
8. Conclusion

The main outcomes of the first evaluation experiment (see Table 2) are that the indicative vowel analysis with a simple straightforward detection technique, and the more elaborate emotional phoneme classes approach, perform comparably close to baseline emotion recognition performance. Results presented in Table 3 proved that the “leading” realistic emotion classification technique could not provide applicable performance within cross-corpora evaluation.

The experimental results presented in Tables 3–5 proved that the phonetic pattern dependent modeling technique presented in Section 7.2.2 improves the classification performance within cross-corpora evaluation. The best emotion classification performances within cross-corpora evaluation were obtained with the straightforward technique based on phoneme segments for VAM trained and EMO-DB tested emotion acoustic models (UA = 83.94%) and with maximum length vote strategy applied for the phoneme-level segments for EMO-DB trained and VAM tested models (UA = 71.95%). We could summarize that emotional phoneme-classes with phoneme-level bi-gram language models for ASR-driven phonetic transcription generation increase cross-corpora classification performance about 4.72% absolute (relative 7.02%) for emotion models trained on the EMO-DB dataset samples and evaluated on the VAM dataset. With phonetic pattern dependent emotion models trained on spontaneous emotions we obtain a considerable performance gain (16.49% absolute (relative 24.45%) improvement) for the classification of the level of arousal of acted emotions. These results also show that emotion classification techniques employing a common feature set for the phoneme-level modeling can extract more emotion discriminating information than using turn-level analysis independent of phonetic transcriptions.

One can see from columns “Umr.” (Unrecognized) in Tables 2 and 5 that a significant relative number of the utterances were classified as emotionally unrecognized with the word-level matching scores based on the un-weighted majority vote strategy. This means that these utterances obtained an equal number of words which were classified as low and high arousal. Taking into account that the EMO-DB and VAM databases contain only turn-level emotion annotation, we suppose that some words within an utterance could be theoretically assigned to the neutral speaking style or even differ from the annotated emotional state. By using high standard emotional corpora with word-level emotion annotation we will be able to provide better phoneme-level emotion analysis.

9. Discussion and future work

We showed that using phoneme-level emotion classes could improve classification performance even with comparably low speech recognition performance obtained with scant a priori knowledge about the language, implemented as a zero-gram for word-level modeling and a bi-gram for phoneme-level modeling. As an alternative to the word-level language models we implemented phoneme-level bi-gram models which can solve problems with appearances of OOV words. We proved that the best cross-corpora classification performance could be obtained with phoneme-level emotion analysis and phoneme-level matching segments. Cross-corpora evaluation results on the VAM database can be compared with state-of-the-art results presented in Zhang et al. (2011). In this article, the authors achieved the best emotion classification performance for the VAM speech samples, UA = 69.7%. In order to train phonetic pattern independent emotion models the authors used 6820 (4685 high arousal and 2135 low arousal) training samples from 5 publicly available databases: ABC (Schuller et al., 2007), AVIC (Schuller et al., 2009), DES (Engbert and Hansen, 1996), eNTERFACE (Martin et al., 2006), SAL (Wöllmer et al., 2008). By using almost 15 time less training samples (456 utterances containing 210 low and 246 high arousal emotions) we were able to increase cross-corpora classification performance by about 2.25% absolute (3.22% relative) from UA = 69.7% (Zhang et al., 2011) to UA = 71.95%, see Table 5. By using more robust affective speech recognition techniques we will be able to further improve emotion classification performance of the emotion phoneme classes technique.

As can be seen in Table 1, there is rather few material available for detailed statistical emotional speech analysis on the phoneme level. Creation of new well-annotated emotional corpora will help us to perform a more detailed emotional speech analysis. Within the SFB/TRR 62 “Companion-Technology for Cognitive Technical Systems” (http://www.sfb-trr-62.de/) project we collect a new speech corpus with spontaneous emotions. In our future research we would like to evaluate our emotion classification techniques on our collected speech material and other new publicly available emotional corpora. It is recommendable that the emotion research community should address more attention to human emotion’s perception: with more rigorous perception analysis of human emotion’s, the affective computing community will be able to define emotional standard units that can be easily determined and classified by
any “advanced” and “non-advanced” listener. This will also enable us to make our emotion processing techniques more robust.

Acknowledgements

This research is associated and supported by the Transregional Collaborative Research Center SFB/TRR 62 “Companion-Technology for Cognitive Technical Systems” (http://www.sfb-trr-62.de/) funded by the German Research Foundation (DFG). Further, the research leading to these results has received funding from the federal state Sachsen-Anhalt with the Graduiertenförderung (LGFG scholarship). We also acknowledge the DFG for financing our computing cluster used for parts of this work. Finally, this work was conducted within the framework of the Magdeburg Center for Behavioral Brain Sciences.

References


