Available online at www.sciencedirect.com
ScienceDirect
Speech Communication 49 (2007) 787–800
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

Primitives-based evaluation and estimation of emotions in speech

Michael Grimm [a,*], Kristian Kroschel [a], Emily Mower [b], Shrikanth Narayanan [b]

[a] Universität Karlsruhe (TH), Institut für Nachrichtentechnik (INT), Kaiserstraße 12, 76128 Karlsruhe, Germany
[b] University of Southern California (USC), Speech Analysis and Interpretation Laboratory (SAIL), 3740 McClintock Avenue, Los Angeles, CA 90089, USA

Received 30 March 2006; received in revised form 21 December 2006; accepted 11 January 2007

Abstract

Emotion primitive descriptions are an important alternative to classical emotion categories for describing a human's affective expressions. We build a multi-dimensional emotion space composed of the emotion primitives of valence, activation, and dominance. In this study, an image-based, text-free evaluation system is presented that provides intuitive assessment of these emotion primitives, and yields high inter-evaluator agreement.

An automatic system for estimating the emotion primitives is introduced. We use a fuzzy logic estimator and a rule base derived from acoustic features in speech such as pitch, energy, speaking rate and spectral characteristics. The approach is tested on two databases. The first database consists of 680 sentences of 3 speakers containing acted emotions in the categories happy, angry, neutral, and sad. The second database contains more than 1000 utterances of 47 speakers with authentic emotion expressions recorded from a television talk show. The estimation results are compared to the human evaluation as a reference, and are moderately to highly correlated ($0.42 < r < 0.85$). Different scenarios are tested: acted vs. authentic emotions, speaker-dependent vs. speaker-independent emotion estimation, and gender-dependent vs. gender-independent emotion estimation.

Finally, continuous-valued estimates of the emotion primitives are mapped into the given emotion categories using a k-nearest neighbor classifier. An overall recognition rate of up to 83.5% is accomplished. The errors of the direct emotion estimation are compared to the confusion matrices of the classification from primitives. As a conclusion to this continuous-valued emotion primitives framework, speaker-dependent modeling of emotion expression is proposed since the emotion primitives are particularly suited for capturing dynamics and intrinsic variations in emotion expression.

Keywords: Emotion estimation; Emotion expression variations; Emotion recognition; Emotion space concept; Fuzzy logic; Man–machine interaction; Natural speech understanding; Speech analysis

1. Introduction

In recent years, automatic recognition of emotions from speech and other modalities has achieved growing interest within the human–machine interaction research community. This interest has merit, since emotion recognition is an essential part of the road map to make communication between humans and computers more human-like. Moreover, automatic assessment of affective speech continues to gain importance in the context of speech data mining. A search query on a speech archive may be located by the affective state of the target speaker in addition to, or instead of, just the semantic content.

There is a large body of literature on the "classical" approach to emotion recognition. Cowie et al. (2001) give an excellent comprehensive review. Other examples of relevant work include (Dellaert et al., 1996; Batliner et al., 2000; Oudeyer, 2003; Nwe et al., 2003; Ververidis et al., 2004). They all treat the emotion recognition problem as a multiple classification task of several emotional categories such as *angry*, *happy*, and *sad*; or simply, *negative* and *non-negative*. However, emotion psychology research has shown that, as an alternative to categories,

* Corresponding author. Tel.: +49 721 608 3790; fax: +49 721 608 3799.
E-mail address: grimm@int.uni-karlsruhe.de (M. Grimm).

0167-6393/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2007.01.010

emotions can also be described as points in a multidimensional emotion space. Cowie and Cornelius (2003) give a review of the different concepts. The multi-dimensional description benefits from a greater level of generality. Additionally, it allows for describing the intensity of emotions. These properties are necessary for an analysis of the inter- and intra-speaker emotion expression variability. In this paper, we take one step beyond current emotion recognition algorithms and propose a method for evaluating and automatically estimating these emotion primitives that determine the location of an emotion in the multi-dimensional emotion space from the speech signal.

Our approach contributes to an important challenge in automatic emotion recognition, namely recognizing emotions not only from acted speech of professional speakers but also from spontaneous speech of non-professional speakers. An increasing number of recent studies are based on spontaneous speech of naïve subjects (Douglas-Cowie et al., 2003; Yu et al., 2004; Vidrascu and Devillers, 2005; Schuller et al., 2006). For these natural emotions, a description using just one category label is not sufficient. In fact, the emotion space concept allows for a more adequate description of these emotions. In particular, gradual emotion transitions, and changes in the intensity of an emotion can easily be described. Furthermore, speaker-dependent variability in the expression of emotions, i.e., the spectrum of actually communicated emotions and the similarity of opposite emotions within this range, can be characterized. These properties are crucial for analyzing emotions in spontaneous, natural speech.

Describing emotions by attributes along bipolar axes was originally proposed by Wundt (1896). Although a general emotion description framework itself is still under discussion in the emotion psychology community (see Scherer, 2005 for instance), the concept of description by attributes has been since pursued in various forms. However, there has been only very limited research on automatic emotion recognition within the multi-dimensional emotion space framework.

Yu et al. (2004) divide the 2D emotion space of valence and arousal into three and five levels, respectively. They thereby transform the task of determining the continuous values of the emotion attributes to a more convenient multiple classification task. For the LDC CallFriend corpus, they achieve recognition rates between 54% and 67%, depending on the number of classes used. Valence and arousal are classified separately. Vidrascu and Devillers (2005) report a recognition accuracy of 82% on the two-level classification of valence into positive and negative values. Their study is based on a large corpus of a medical call center.

Fragopanagos and Taylor (2005) also motivate their choice of the activation–evaluation space by emotion psychology. They divide the emotion space into four regions for classification based on activation (positive/negative) and evaluation (positive/negative), respectively. Tested on their own database, generated through a Wizard-of-Oz experiment, they report an average recognition rate of 48.5% if only acoustic features are used as input to an artificial neural net (ANN). Combining these features with facial expression analysis or emotional salience analysis of the words or both improved the results by 0.3%, 2.5%, and 2.0%, respectively. In the case of separate classification, they report an average of 73.5% for activation and 64% for evaluation. The results are improved by up to 6% by using additional information channels.

Thus it can be summarized that using emotion dimensions as motivated by emotion psychology is a promising step toward improving the state-of-the-art in emotion recognition. However, to our knowledge there is no previous study on directly estimating the continuous-valued emotion primitives. We address this problem in this paper.

In general, there are several ways to represent emotions in a multi-dimensional emotion space. They can be distinguished by the number and meaning of their basic entities (Cowie and Cornelius, 2003; Kehrein, 2002; Schröder et al., 2001). The so-called "dimensions" are actually descriptive, generic attributes of an emotion that function as constituents. These constituents will be referred to as *primitives* in this paper. Note that these primitives are not regarded as meta-features of emotion categories but as a fully complementary description of emotions.

Two-dimensional representations include one primitive that describes the *appraisal* (or *valence, evaluation*) taking values from positive to negative. The other emotion primitive describes the *activation* (or *arousal, excitation*), and is sometimes motivated by the action tendencies of emotions. Three-dimensional representations additionally include a primitive defining the apparent strength of the person, which is referred to as *dominance* (or *power*). This third dimension is necessary to distinguish anger from fear for instance, since the dominance (or the ability to handle a situation) is the only discriminating element in this case. We chose the combination of the following three emotion primitives (Kehrein, 2002):

- *Valence (V)* – positive vs. negative,
- *Activation (A)* – excitation level high vs. low, and
- *Dominance (D)* – apparent strength of the speaker, weak vs. strong.

Our study consists of the following main parts. (1) We introduce a robust and efficient human emotion assessment method to produce the three-dimensional emotion references, which provides quick-and-easy assessment of authentic emotions in natural speech. (2) We propose a rule-based fuzzy logic method to estimate the continuous values of the emotion primitives from acoustic features derived from the speech signal. (3) We assess our emotion recognition method based on primitives by comparing the results with conventional categorical classification. (4) We finally show how emotion primitives are well suited for capturing the speaker-dependent variability in emotion expression. The rest of the paper is organized as follows.

Table 1
Databases used for this study

Description	Language	Emotion type	No. speakers	No. sentences	Avg. no. sen./speaker	No. evaluators
EMA	Am. English	Acted	3	680	227	18
VAM I	German	Authentic	19	499	26	17
VAM II	German	Authentic	28	503	28	6

Section 2 introduces the data we use. Section 3 describes the human evaluation of emotional speech in terms of the three emotion primitives. Section 4 presents details of estimating the three-dimensional emotion primitives from speech using a rule-based fuzzy logic classifier. Section 5 shows the results and provides a comparison between the results of real-valued primitives estimation and discrete emotion classification. Section 6 details how the speaker dependent variability present in expressed emotions can be described in terms of the emotion primitives. Section 7 provides conclusions and outlines future work.

2. Data

For this study we use two databases. The first corpus, called the *EMA Corpus*,[1] contains speech with acted emotions in American English. The second corpus, called the *VAM Corpus*,[2] contains spontaneous speech with authentic emotions that was recorded from guests in a German TV talk-show. Table 1 summarizes the key facts about both databases: language, emotion elicitation type (acted or natural), number of speakers and sentences, average number of sentences per speaker, and number of evaluators.

The two databases are deliberately chosen to contain two different emotion production styles. While the spontaneous speech database is used to push the application oriented research on authentic emotions, the acted speech database is used to provide a comparison with state-of-the-art emotion categorization.

The use of these two different databases was also partly motivated by our goal to explore if the proposed methods hold good for two different languages and across natural and acted emotional speech. We however test the emotion primitives estimator only with speech of the same language that has been used for training. Nevertheless, similar recognition results for both languages, English and German, may imply cross-cultural robustness of the proposed method.

2.1. Acted speech corpus

The *EMA Corpus* (Lee et al., 2005) contains 680 sentences of emotional speech, produced by one professional (f) and two non-professional (1f/1m) speakers. The female speakers read 200 sentences, and the male speaker read 280 sentences. These recordings consist of 10 (14) sentences, each of them repeated 5 times in 4 different emotions. Each block consisted of 14 sentences that were randomized within the block. Each repetition for a given emotion was block-wise; the subjects produced all sentences within a given block in the same emotion. This was repeated for each of the four emotions, in a random order of emotions (Lee et al., 2005). All sentences are in English, spoken by native speakers of American English. As described in (Grimm et al., 2006a), the EMA corpus was evaluated by four native speakers of American English. For each sentence, the evaluators assigned one of the category labels from among *happy, angry, sad, neutral,* and *other* to the utterance. The average human recognition rate of the acted emotions was 81.8%. Happy emotion was most poorly recognized (76.6%). This was due to the fact that several sentences that were intended to be happy were perceived as neutral emotions. See Table 2 for the confusion matrix, given as an average of all three speakers in the database. Similar results were reported by Bulut et al. (2002).

To assess the inter-evaluator agreement, we used the parameter κ derived from the *Kappa statistics* (Carletta, 1996),

$$\kappa = \frac{P_A - P_0}{1 - P_0}. \quad (1)$$

This parameter describes the level of inter-evaluator agreement $\kappa \in [0, 1]$. P_A represents the proportion of the evaluators that assigned the same class label, and P_0 corrects for their agreement by chance. We found a moderate inter-evaluator agreement of $\kappa = 0.48$ between the four evaluators, which is a typical value for such categorical emotion assessment by humans (cf. Vidrascu and Devillers, 2005).

2.2. Natural speech corpus

The second database, the *VAM Corpus*, consists of recordings of invited guests in a German TV show called

Table 2
Confusion matrix of emotion class labeling of EMA corpus, in percent, by four human listeners ($\kappa = 0.48$)

	Angry	Happy	Neutral	Sad	Other
Angry	**80.3**	2.2	4.1	0.7	12.7
Happy	3.2	**75.6**	11.8	1.3	8.1
Neutral	1.2	0.4	**84.0**	11.8	2.6
Sad	0.3	0.6	6.3	**87.5**	5.3

[1] The acronym *EMA* stands for electromagnetographic articulatory study. However, the articulatory data were analyzed in a different study.
[2] The acronym *VAM* is the abbreviation of the talk-show title *Vera am Mittag*.

Vera am Mittag.[3] This show is broadcasted Monday through Friday on Free-TV with a regular duration of one hour. Each show contains five dialogues between two or three guests, moderated by an anchorwoman. The speakers mostly discuss personal problems or family issues in a spontaneous unscripted fashion. The first part of this corpus, *VAM I*, was first used in (Grimm and Kroschel, 2005a). The second part, *VAM II*, contains sentences from additional speakers in the talk-show that were evaluated after the initial experiment was reported.

In total, the VAM database contains 1002 emotional utterances from 47 speakers (11m/36f). All signals were recorded using a sampling frequency of 16 kHz and 16 bit resolution.

The dialogues were manually segmented at the utterance level. Each utterance contained at least one intermediate phrase. The video stream was not analyzed in this study. The speakers were selected by a preliminary evaluation during the data segmentation and selection step to guarantee that each speaker showed both neutral expressions and at least some emotional deviation from the neutral state.

The emotions covered in the spontaneous speech corpus are summarized in Section 3.2, after introducing the evaluation method.

3. Primitives-based emotion evaluation

Evaluation of the emotions contained in the speech data was done through human listener tests. A popular, and widely used tool for the human evaluation of emotions in a multi-dimensional emotion space is the *Feeltrace* tool developed by Cowie et al. (2000). This instrument allows for time-continuous and value-continuous assessment of emotions in the activation–evaluation space. The method is based on Plutchik's concept of defining emotions as positions within a circle, wherein the angle determines the character of the emotion, and the distance from the origin determines the intensity of the emotion. We did not use this instrument since (1) it is restricted to a two-dimensional emotion space that has been shown not to be adequate for distinguishing certain emotions such as fear and anger, for instance, see (Cowie et al., 2000), (2) a square space (or a cube in 3D) is more appropriate for our chosen primitives, since valence is a bipolar rather than an angular-type periodic entity (Russell and Mehrabian, 1977), and (3) time-continuous evaluation was not well suited for our utterance-level units.

The evaluation method described below builds upon our preliminary work reported in (Grimm and Kroschel, 2005b,c). The novel aspects reported here include a more intuitive scaling and orientation of the axes. Additionally, the evaluation tool was extended to include elicitation of the evaluator's background such as language comprehension capabilities, and self-evaluation of his/her personality with respect to handling emotions. However, we could not observe any statistically significant difference in the evaluation of emotions by humans of different cultural background or different self-evaluation.

Section 3.1 describes the utterance-based assessment method. Section 3.2 contains the primitives-based evaluation results on the acted and spontaneous speech databases, respectively.

3.1. Evaluation method

For the evaluation of emotions in the 3D emotion space of *valence*, *activation*, and *dominance*, we propose to use the self assessment manikins (SAMs) proposed originally by Lang (1980). This instrument consists of an array of five images per primitive (see Fig. 1). These images allow us to avoid the use of categorical labels for emotions. Evaluating emotions using SAMs is fast and very intuitive. Note that the SAMs originate from self-assessment, however in our case, the speech was not evaluated by the speakers themselves.

For each utterance n in the database, $1 \leqslant n \leqslant N$, the evaluator k, $1 \leqslant k \leqslant K$, chooses 3 values $\hat{x}_{n,k}^{(i)}$ – one for each emotion primitive $i \in \{valence, activation, dominance\}$. The selection of the icons is mapped to integer values $\{1,2,3,4,5\}$ and then transformed to unity space $[-1,+1]$. For intuitive comprehension of the primitives, the axes are oriented from *negative* to *positive* (*valence*), *calm* to *excited* (*activation*), and *weak* to *strong* (*dominance*).

Although it can be assumed that each evaluator assesses the emotional content of an utterance to the best of his/her knowledge, the assessment does not necessarily reflect the emotion truly felt by the speaker. There is a number of "input- and output-specific issues", as Fragopanagos and Taylor call it (Fragopanagos and Taylor, 2005). Both the expression and perception of emotions are subject to several influences, such as display rules and cognitive effects. From a signal processing viewpoint, these influences can be modeled as signals with superimposed noise on top of the hidden "true" emotion. Assuming an unbiased ensemble of evaluators, the hidden emotion can best be deter-

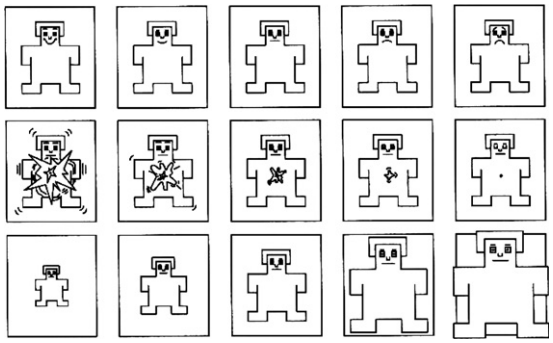

Fig. 1. Self assessment manikins (Fischer et al., 2002). This evaluation tool is used for a text-free, three-dimensional assessment of emotion in speech.

[3] English: *Vera at noon*; *Vera* is the name of the talk-show host.

mined by estimating it from the combined assessment results of several evaluators.

In (Grimm and Kroschel, 2005c), two different methods to merge the evaluation results of several evaluators were discussed. We choose the *evaluator weighted estimator (EWE)*,

$$x_n^{\mathrm{EWE},(i)} = \frac{1}{\sum_{k=1}^{K} r_k^{(i)}} \sum_{k=1}^{K} r_k^{(i)} \hat{x}_{n,k}^{(i)}. \tag{2}$$

This estimator averages the individual evaluators' responses, and takes into account that each evaluator is subject to an individual amount of disturbance during evaluation. This is done by introducing evaluator-dependent weights $r_k^{(i)}$,

$$r_k^{(i)} = \frac{\sum_{n=1}^{N} \left(\hat{x}_{n,k}^{(i)} - \frac{1}{N} \sum_{n'=1}^{N} \hat{x}_{n',k}^{(i)} \right) \left(\bar{x}_n^{(i)} - \frac{1}{N} \sum_{n'=1}^{N} \bar{x}_{n'}^{(i)} \right)}{\sqrt{\sum_{n=1}^{N} \left(\hat{x}_{n,k}^{(i)} - \frac{1}{N} \sum_{n'=1}^{N} \hat{x}_{n',k}^{(i)} \right)^2} \sqrt{\sum_{n=1}^{N} \left(\bar{x}_n^{(i)} - \frac{1}{N} \sum_{n'=1}^{N} \bar{x}_{n'}^{(i)} \right)^2}}. \tag{3}$$

These evaluator-dependent weights measure the correlation between the listener's responses, $\{\hat{x}_{n,k}^{(i)}\}_{n=1,\ldots,N}$, and the average ratings of all evaluators, $\{\bar{x}_n^{(i)}\}_{n=1,\ldots,N}$, where

$$\bar{x}_n^{(i)} = \frac{1}{K} \sum_{k=1}^{K} \hat{x}_{n,k}^{(i)}. \tag{4}$$

The assessment quality is determined by calculating the standard deviation $\sigma_n^{(i)}$ of the evaluations,

$$\sigma_n^{(i)} = \sqrt{\frac{1}{K-1} \sum_{k=1}^{K} \left(\hat{x}_{n,k}^{(i)} - x_n^{\mathrm{EWE},(i)} \right)^2}. \tag{5}$$

A comparison of the EWE and a maximum likelihood estimator, as well as a discussion of intrinsic evaluation errors due to emotion space quantization by the SAMs can be found in (Grimm and Kroschel, 2005c). It was shown that the EWE yields up to 20% better results than the maximum likelihood estimator. It converges into the maximum likelihood estimator in the case of equal weights for all estimators. Corrected by the emotion space quantization error, the actual evaluation error was found to be

$$\epsilon_n^{(i)} = \sigma_n^{(i)} - b(K), \tag{6}$$

where $b(K) \in (\frac{1}{8}, \frac{1}{8}\sqrt{2}]$ is a constant bias, and depends on the number of evaluators K.

3.2. Evaluation results

The described method for evaluating emotion primitives was applied to both the EMA and the VAM database. For each sentence, the EWE estimate was calculated according to (2). With $K = 18$, 17, and 6 evaluators for the 3 corpora, respectively, we use a comparatively high number of evaluators for this task. Most other studies involve 2 evaluators (Vidrascu and Devillers, 2005; Vidrascu and Devillers, 2005; Schuller et al., 2005) or, at most, 4–5 evaluators (Yu et al., 2004; Fragopanagos and Taylor, 2005; Lee and Narayanan, 2003).

The inter-evaluator agreement was measured by determining the standard deviations $\sigma_n^{(i)}$ of the assessment and the correlation coefficients $r^{(i)}$ using Eqs. (5) and (3), respectively.

The standard deviation, on the one hand, measures the suitability of a particular sentence for our task. A low standard deviation indicates that the emotional expression is perceived by all human listeners similarly. The inter-evaluator correlation, on the other hand, measures the agreement among the individual evaluators and thus focuses on the more general evaluation performance.

The average results for each database are reported in Table 3. On average, the standard deviation was between 0.28 and 0.38 for each primitive. Thus, the standard deviation was slightly above 0.25, i.e. half the distance between two SAMs, indicating good evaluation results. There was no significant difference between the database containing acted emotions and the databases containing authentic emotions. Note that the standard deviation includes the quantization error due to the discretization of the SAMs in the emotion primitive space.

The inter-evaluator correlation was moderate to high with values in the range of 0.48–0.79. The correlation was in general greater for the EMA database than for the VAM database. This result is probably due to the more stereotypical nature of the emotions portrayed by the actors. Furthermore it could be observed that the *valence* primitive yields a smaller inter-evaluator correlation than *activation* or *dominance*. In particular, for the VAM database containing authentic emotions from talk show dialogues this result might be due to the fact that the distribution of *valence* values was narrower than the distributions of *activation* or *dominance*, and thus evaluators' deviations by the same amount resulted in a smaller correlation coefficient.

Table 3
Average standard deviation $\bar{\sigma}$ and correlation coefficient r for the emotion primitives evaluation of the EMA corpus and the VAM I/II corpus by human listeners, averaged over all speakers and all sentences

	Standard deviation $\bar{\sigma}$			Correlation coefficient r		
	Valence	Activation	Dominance	Valence	Activation	Dominance
EMA	0.35	0.36	0.35	0.63	0.79	0.75
VAM I	0.30	0.38	0.33	0.49	0.78	0.68
VAM II	0.28	0.30	0.29	0.48	0.66	0.54

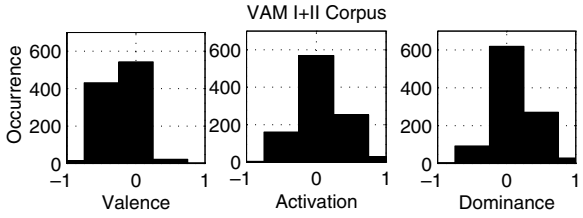

Fig. 2. Histogram of emotions in VAM corpus.

All correlation coefficients were statistically significant ($p < 0.0001$).

Fig. 2 shows the histogram of the emotions in the VAM talk show database. It has to be noted that a large percentage of the utterances in this database contains neutral or negative speech with high *activation* and *dominance* values. This distribution is probably due to the nature of the topics discussed in the talk show, which include family problems, paternity questions and friendship issues. The restrictiveness of recording a wide spectrum of emotions is an intrinsic problem in spontaneous speech processing. Moreover, averaging assessment results naturally tends to result in a more Gaussian-like distribution. We addressed the problem of unequally distributed emotions in the database by using a rule-based emotion primitives estimator that is not influenced by *a priori* probabilities, cf. Section 4.

We calculated the standard deviations for each sentence to discard a few outliers: All utterances that had been evaluated with a standard deviation $\sigma_n^{(i)} > 0.5$ for any of the emotion primitives $i \in \{V, A, D\}$ were not used for the further study. In many of these cases the utterances were too long, and contained more than one, conflicting emotions. The remaining utterances were all evaluated with a deviation of one SAM or less. Thus the VAM I database was reduced to 490 utterances (98.2%), the VAM II database was reduced to 489 utterances (94.2%), and the EMA database was reduced to 614 sentences (90.3%), respectively. The resulting new average standard deviations were marginally smaller than the ones reported above.

For comparison: Cowie et al. report similar standard deviations using the Feeltrace tool on a different evaluation task and three evaluators (Cowie et al., 2000). From Cowie et al. (2000, Fig. 4), it can be inferred that the standard deviation of their chosen primitives, *evaluation* and *activation*, was in the range of 0.2–0.3.

It can be summarized that (1) the SAMs are well suited for evaluating emotions in speech, (2) the inter-evaluator correlation on *activation* and *dominance* is higher than on *valence*, and (3) the inter-evaluator correlation on acted emotions is slightly higher than the one on authentic emotions.

4. Primitives-based emotion estimation

In this section, we focus on automated emotion estimation from speech. Specifically, we describe a *fuzzy logic* inference system for primitives-based automated emotion estimation. Fuzzy logic lends itself to continuous-valued estimates of emotions in spontaneous natural speech. Such continuous-valued emotion estimates are necessary to automatically assess temporal dynamics in emotion, or to tackle the problem of a speaker-dependent variability in emotion expression.

The emotion estimator described below builds upon our previous work (Grimm and Kroschel, 2005a; Hernandez, 2005). The preliminary results reported were based on a fraction of our database, and a smaller number of evaluators than in the present study.

Fuzzy logic was chosen because the nature of linguistic emotion class labels is inherently fuzzy and vague. Fuzzy logic transforms crisp values into fuzzy values using membership grades. The crisp values that are extracted from the acoustic speech signal are processed as linguistic variables. For instance, the mean value of the pitch is processed as a *high*, *medium*, or *low* mean pitch value. While the idea of applying fuzzy logic to the problem of emotion recognition has been previously discussed with other objectives (Lee and Narayanan, 2003; Huang and Akagi, 2005), fuzzy logic has not been used to estimate continuous values of emotion primitives yet. We consider this approach in this paper. An alternative would be multidimensional, kernel-based regression methods (Schölkopf and Smola, 2002), which we will analyze in the future. Section 4.1 describes the pre-processing and the acoustic feature extraction. Section 4.2 details the proposed estimation method, and it describes how the rule system is derived from acoustic features.

4.1. Pre-processing and feature extraction

All signals were sampled at 16 kHz sampling rate and a resolution of 16 bit. They were processed at the utterance-level.

The acoustics of emotional speech have been studied for many years. In general, the differences of the prosodic characteristics between emotionally loaded and neutral speech have been analyzed and reported (Murray and Arnott, 1993; Banse and Scherer, 1996; Cowie et al., 2001). The major acoustic speech features considered include fundamental frequency f0 (“pitch”), speaking rate, intensity, and voice quality. For example, Murray and Arnott state that angry speech is slightly faster, has a very much higher pitch average, much wider pitch range, and higher intensity (Murray and Arnott, 1993). Some of these characteristics can be related directly to physiological changes in the vibration of the vocal chords.

The number of features extracted from the speech signal varies significantly from approximately 10 basic features such as mean values and range in pitch and intensity (Lee et al., 2001) to 276 in the case of systematic application of functionals to a set of basic trajectories (Schuller et al., 2006). This spectrum results from the fact that it is still unclear which features are suited best, and that the

feature set is highly dependent on the data and the classification task. We chose $M = 46$ acoustic features that were derived from the pitch and the energy contour of the speech signal, as well as features related to the speaking rate and spectral characteristics. This is in accordance with most studies in this field. The emotionally colored prosody of the utterance is thus described in terms of statistics, such as mean value, standard deviation, and percentiles.

The following features were extracted from the speech signal:

Pitch related features: f0 mean value, standard deviation, median, minimum, and maximum, 25% and 75% quantiles, difference between f0 maximum and minimum, difference of quartiles.

These features related to the fundamental frequency f0 describe the intonation and speaking melody. They capture monotone speech or highly accented syllables, for example. The pitch was estimated using autocorrelation method since it was shown to give good results in a wide range of applications (Nagel, 2005).

Speaking rate related features: ratio between the duration of unvoiced and voiced segments, average duration of voiced segments, standard deviation of duration of voiced segments, average duration of unvoiced segments, and standard deviation of duration of voiced segments.

These features describe the temporal characteristics in the prosody. They might reveal whether the speech sounds urged or relaxed, for example.

Intensity related features: intensity mean, standard deviation, maximum, 25% and 75% quantiles, and difference of quartiles.

Intensity related features are used to capture the energy in speaking, and helps to discriminate shouting from sad or depressed speech, for example.

Spectral features: mean value and standard deviation of 13 Mel frequency cepstral coefficients (MFCC).

The MFCCs are very common in automatic speech recognition (ASR). While the short-term statistics are very useful for phoneme recognition, the long-term statistics indicate voice quality and are thus often included in the feature set for automatic emotion recognition.

A principal component analysis (PCA) was applied to the feature set to reduce the number of features using an eigenvalue threshold of 0.01. However, the estimation results were best when all features were used.

The described features form the basis of the rule system in the fuzzy inference emotion estimator. Each feature m, $1 \leqslant m \leqslant M$, is related to each of the emotion primitives $x^{(i)} := x^{\mathrm{EWE},(i)}$, $i \in \{V, A, D\}$, to be estimated.

4.2. Rule-based fuzzy logic emotion estimation

The fuzzy logic classifier consists of the three components *fuzzification*, *inference*, and *defuzzification* (Kroschel, 2004). Until the last step of defuzzification, the emotion primitives $x^{(i)}$, $i \in \{V, A, D\}$, will therefore be represented by the fuzzy, linguistic variables

$$\begin{aligned} x^{(V)} &\rightarrow B_l^{(V)} \in \mathscr{B}^{(V)} = \{negative, neutral, positive\} \\ x^{(A)} &\rightarrow B_l^{(A)} \in \mathscr{B}^{(A)} = \{calm, neutral, excited\} \\ x^{(D)} &\rightarrow B_l^{(D)} \in \mathscr{B}^{(D)} = \{weak, neutral, strong\}. \end{aligned} \tag{7}$$

The membership functions of these fuzzy variables are depicted in Fig. 3. The three emotion primitives are estimated separately. In the following, we briefly summarize the three elements of the fuzzy inference system reported in (Grimm and Kroschel, 2005a; Hernandez, 2005). Fig. 4 shows an example of the fuzzy logic inference system. It is based on an example of two features and intends to give a compact overview on the individual elements of the fuzzy logic estimator.

4.2.1. Fuzzification

In the fuzzification step, each feature m is transformed from a crisp value v_m to three fuzzy variables A_j, $j = 1, 2, 3$, where

$$A_j \in \mathscr{A} = \{low, medium, high\}. \tag{8}$$

This reflects the fact that, for example, the absolute value of the average fundamental frequency is not relevant, but it is important to distinguish between low, medium and high pitch average. These generalized terms $\mathscr{A}$ are applied to all features, although when talking about an individual feature we would rather use more specific terms for description.

The *degree of membership* $\mu_{j,m}$ of each linguistic variable A_j is determined by the value of the membership function $\mu_{A_j,m}(\alpha)$ at the point of the crisp feature value,

$$\mu_{j,m} = \mu_{A_j,m}(v_m). \tag{9}$$

The membership functions of the input features, $\mu_{A_j,m}(v_m)$, have the same piecewise linear shape as the membership functions of the emotion primitives depicted in Fig. 3. This shape is common in fuzzy logic systems (Kroschel, 2004), and it has been found to be well suited for emotion representation, too (Hernandez, 2005). The edges of the membership functions are determined by the 10% and 90% quantiles of the distributions of the feature values, $Q_{10} = Q_{10}(m)$ and $Q_{90} = Q_{90}(m)$. Thus for feature m the membership functions are defined as follows:

$$\mu_{1,m} = \begin{cases} 1, & v_m < Q_{10} \\ \frac{v_m - Q_{50}}{Q_{10} - Q_{50}}, & Q_{10} \leqslant v_m < Q_{50} \\ 0, & v_m \geqslant Q_{50} \end{cases} \tag{10}$$

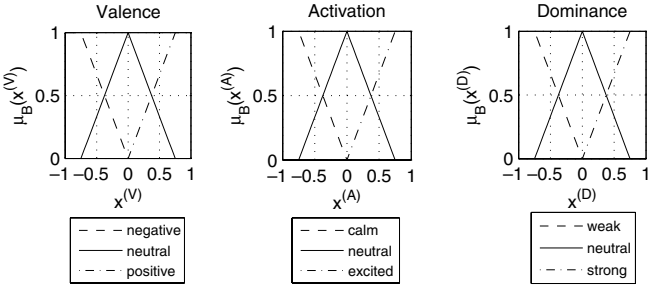

Fig. 3. Membership functions of the emotion components.

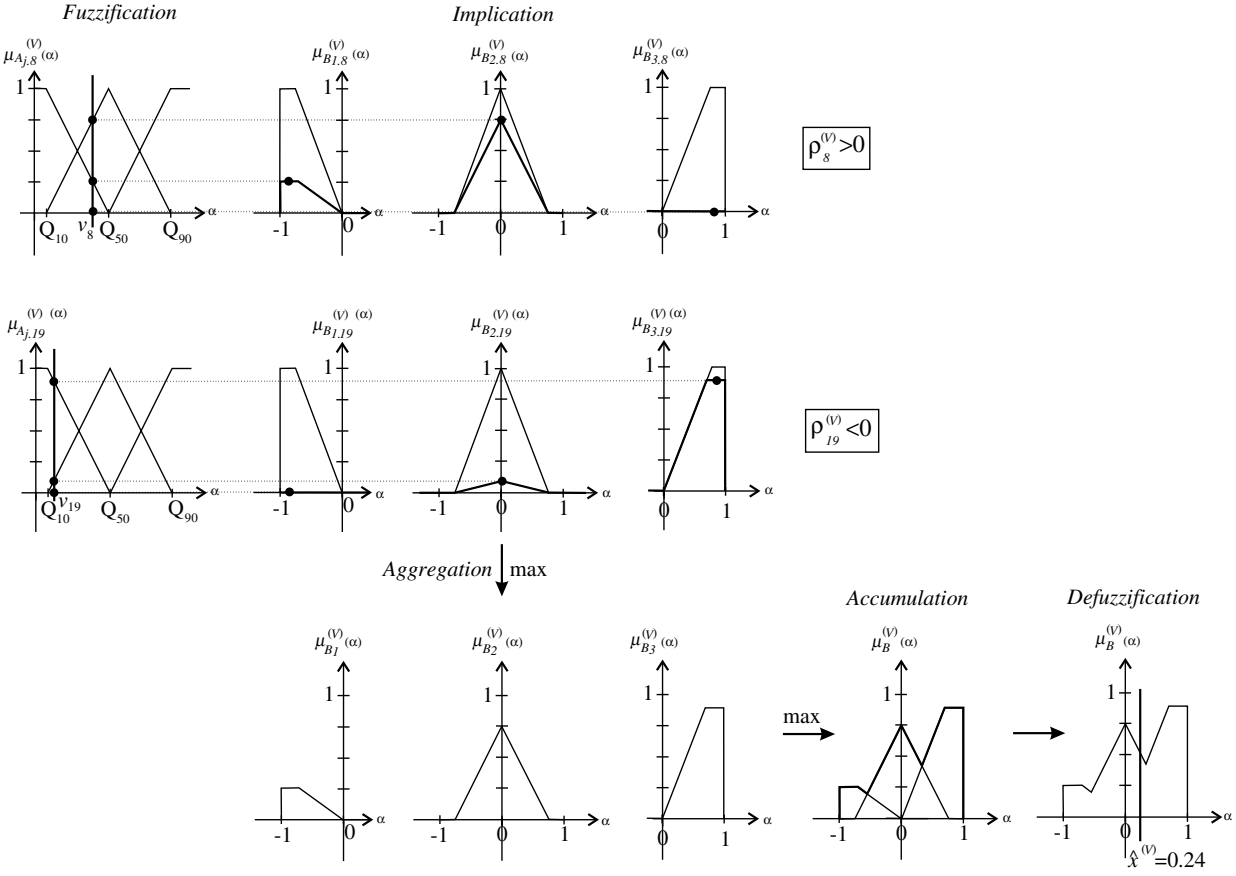

Fig. 4. Fuzzy logic for emotion primitives estimation: fuzzification, inference (implication, aggregation, accumulation) and defuzzification for *valence*, an example using two features.

$$\mu_{2,m} = \begin{cases} 0, & v_m < Q_{10} \\ \frac{v_m - Q_{10}}{Q_{50} - Q_{10}}, & Q_{10} \leqslant v_m < Q_{50} \\ \frac{v_m - Q_{90}}{Q_{50} - Q_{90}}, & Q_{50} \leqslant v_m < Q_{90} \\ 0, & v_m \geqslant Q_{90} \end{cases} \tag{11}$$

$$\mu_{3,m} = \begin{cases} 0, & v_m < Q_{50} \\ \frac{v_m - Q_{50}}{Q_{90} - Q_{50}}, & Q_{50} \leqslant v_m < Q_{90} \\ 1, & v_m \geqslant Q_{90}, \end{cases} \tag{12}$$

where Q_{50} is an abbreviation for $(Q_{10} + Q_{90})/2$.

Fig. 4 (top left) shows the fuzzification of a crisp feature value (v_8) into membership grades of the fuzzy variables *negative, neutral, positive* for *valence* ($\mu_{1,8}, \mu_{2,8}, \mu_{3,8}$).

4.2.2. Inference

The rule base is derived from the correlation $\rho_m^{(i)}$ between the acoustic features m, with $1 \leqslant m \leqslant M$, and the emotion $x_n^{(i)}$ attested by human listeners (cf. Section 3),

$$\rho_m^{(i)} = \frac{\sum_{n=1}^{N} (v_{m,n} - \bar{v}_m)\left(x_n^{(i)} - \bar{x}^{(i)}\right)}{\sqrt{\sum_{n=1}^{N} (v_{m,n} - \bar{v}_m)^2} \sqrt{\sum_{n=1}^{N} \left(x_n^{(i)} - \bar{x}^{(i)}\right)^2}}, \tag{13}$$

where $\bar{v}_m = \frac{1}{N}\sum_{n=1}^{N} v_{m,n}$, and $\bar{x}^{(i)} = \frac{1}{N}\sum_{n=1}^{N} x_n^{(i)}$.

Thus for each linguistic input variable $A_j \in \mathscr{A}$, one rule is formulated to link it to each linguistic output variable $B_l \in \mathscr{B}$,

$$\text{IF } v_m \text{ is } A_j \text{ THEN } x^{(i)} \text{ is } B_l^{(i)}. \tag{14}$$

The sign of the correlation coefficient $\rho_m^{(i)}$ thereby determines which variable pairs (j, l) are related to one another,

$$l_m^{(i)} = 2 + (j_m - 2) \cdot \text{sign}(\rho_m^{(i)}), \quad j_m = 1, 2, 3; \quad m = 1, \ldots, M. \tag{15}$$

For example, we derive the following rules from $\rho_8^{(A)} = 0.8$ and $\rho_{19}^{(V)} = -0.4$, respectively: If the pitch range ($m = 8$) is *high* ($j_8 = 3$) then the *activation* ($i = A$) is *excited* ($l_8^{(A)} = 3$). Or, if the average pause duration between consecutive words ($m = 19$) is *high* ($j_{19} = 3$) then the *valence* ($i = V$) is *negative* ($l_{19}^{(V)} = 1$). That is why in Fig. 4 a *low* feature value of feature 8 is implied to *negative*, while a *low* feature value of feature 19 is implied to *positive*.

The absolute value $|\rho_m^{(i)}|$ determines the importance of the rule and is defined as the rule weight. This way the rules are generated in an automatic way. The expert knowledge is reflected in the fact that features which are highly correlated with the emotion primitives are given large impact in the rule base.

Applying the rules to each acoustic feature, each fuzzy input yields a *degree of support* for each fuzzy output variable. This *degree of support* is the membership grade of the feature assigned to the appropriate fuzzy variables of the emotion primitives, multiplied with the rule weight.

In the *aggregation* step, the degrees of support of all acoustic features are fused using a maximum operator. This maximization has been found to be superior to sum aggregation (Hernandez, 2005). In Fig. 4 the aggregation is found in vertical direction. It can be applied before or after implication as it gives the same result for the chosen operators.

The *implication* draws the actual conclusion and scales the output membership functions by the appropriate aggregated degree of support using a multiplication (product implication). By this, the output membership functions depicted in Fig. 3 are scaled to the appropriate level determined by the rules. Still, the emotion primitives are described by the values of the fuzzy variables, cf. (7).

In the *accumulation* step, the three scaled membership functions of the fuzzy variables $B_l^{(i)}$, $l = 1, 2, 3$, are accumulated using a maximum operator. For *valence*, for example, this accumulation fuses the three fuzzy variables *negative, neutral,* and *positive* that were used to scale the three output membership functions into *one* curve for *valence*. Thus the result is one continuous membership function $\mu_B^{(i)}(\alpha)$ describing the fuzzy value of *valence*, *activation*, and *dominance* for $i \in \{V, A, D\}$, respectively. Fig. 4 shows the accumulation in the bottom row: the three output membership functions resulting from the implication are depicted in thin lines, while the accumulation result is depicted in a bold line. Another elaborate example of this inference system is described in (Grimm and Kroschel, 2005a).

4.2.3. Defuzzification

The last step in the fuzzy logic emotion estimator, *defuzzification*, transforms the fuzzy values to crisp values. We use the centroid method,

$$\hat{x}^{\mathrm{FL},(i)} = \frac{\int_{-1}^{1} \alpha \cdot \mu_B^{(i)}(\alpha)\, \mathrm{d}\alpha}{\int_{-1}^{1} \mu_B^{(i)}(\alpha)\, \mathrm{d}\alpha}. \tag{16}$$

The defuzzification is shown in Fig. 4, bottom right, for the same sample values. The result is one crisp, real-valued number per emotion primitive.

The crisp emotion estimates are normalized by a constant factor $c = 1.63$ to account for the restricted interval of possible values. This restriction results from the shape of the membership functions $\mu_{B_l}^{(i)}$. The centroid method has been shown to give better results than the mean of maximum or bisector method, respectively (Hernandez, 2005).

5. Results

The fuzzy logic emotion estimator was applied to the EMA and the VAM databases. The rule base was constructed for male and female speakers separately, and for all speakers jointly. In total, we defined 13 different scenarios, as itemized in Table 4. The number of speakers and the number of sentences used in each scenario are stated as "#Sp." and "#Sen.", respectively.

For the emotion estimation test using the described fuzzy logic method we generated the rule base from all available utterances, depending on the scenario, and then tested with each of the utterances sequentially. Due to the large database size and the nature of the rule base in the classifier, which was determined in a generic way from the correlation between the individual acoustic features and the emotion primitives, we found that there was no difference in the results if the tested utterance was excluded from the training set.

The following sections discuss these results. Section 5.1 describes several aspects of the estimation results, for example the impact of the individual features, the emotion type, and the speaker dependency. Section 5.2 compares the results of the real-valued emotion primitives estimation to the results of a classical discretized emotion classification task.

5.1. Estimation results

The automatic estimation of emotion primitives was assessed by calculating the estimation error

$$e_n^{(i)} = \left| x_n^{\mathrm{EWE},(i)} - \hat{x}_n^{\mathrm{FL},(i)} \right| \tag{17}$$

for each utterance in the database, $1 \leqslant n \leqslant N$, and for each emotion primitive $i \in \{V, A, D\}$ separately. The mean error for each scenario is reported in Table 4. On average, the estimation error was between 0.16 and 0.28 for the different scenarios. These errors are comparable to the standard deviation in the human evaluation of emotions in the emotion space, cf. Table 3 and Eq. (6). They are in the range of half the distance between two evaluation manikins and thus notably small. Since these results are based on a large number of samples, $N > 100$, they can be regarded as statistically significant.

The correlation coefficients were also used as a means for assessing the estimation results. For all scenarios, the correlation coefficient was found to be positive, and for most of the scenarios we found fairly high correlation in the range of 0.70–0.85. Again, for the EMA database the correlation coefficient was in general higher than for the VAM database. Using separate classifiers for male and female speakers, or increasing the database size by joining VAM I and II, did not improve the correlation significantly.

5.1.1. Impact of individual features

The ranking of the rules with respect to the rule weights $|\rho_m^{(i)}|$ was database and speaker-gender dependent. The highest correlation between an individual acoustic feature and the emotion was found to be the 25% quantile of the

Table 4
Mean error and correlation to reference for the automated emotion primitives estimation (“Estimation” columns) of the EMA corpus and the VAM I/II corpus, respectively

Scenario	Selection	Database	#Sp.	#Sen.	Estimation		Evaluation	
					Mean error	Mean correlation	Mean error	Mean correlation
1	All	VAM I	19	478	0.27	0.71	0.21	0.65
2	All	VAM II	28	469	0.23	0.43	0.15	0.56
3	All	VAM I + II	47	947	0.24	0.60	0.18	0.61
4	Male	VAM I	4	90	0.28	0.85	0.20	0.66
5	Female	VAM I	15	388	0.28	0.68	0.20	0.63
6	Male	VAM II	7	106	0.17	0.42	0.11	0.41
7	Female	VAM II	21	363	0.23	0.44	0.14	0.58
8	Male	VAM I + II	11	196	0.25	0.70	0.15	0.52
9	Female	VAM I + II	36	751	0.26	0.58	0.18	0.61
10	Female	EMA(1)	1	200	0.23	0.80	0.22	0.80
11	Female	EMA(2)	1	200	0.16	0.82	0.22	0.66
12	Male	EMA(3)	1	280	0.17	0.79	0.23	0.67
13	All	EMA	3	680	0.19	0.75	0.22	0.72

Manual results of the human evaluation are added for comparison (“Evaluation” columns).

pitch ($m = 6$) for the male speakers in the VAM I database with $\rho_6^{(V)} = 0.70, \rho_6^{(A)} = 0.89$, and $\rho_6^{(D)} = 0.91$. Other features of high correlation to emotion primitives included the f0 median and the standard deviation of the 3rd and 13th MFCC, respectively.

In general, it was observed that all features had a non-zero rule weight, and thus at least partly contribute some information about the emotion. Although there is some agreement with the feature ranking found in other studies (Schuller et al., 2005), it has to be suspected that the ranking strongly depends on the data used.

5.1.2. Natural emotions vs. acted emotions

In general, the error was higher for the natural speech database VAM (0.17 in scenario 6, to 0.28 in scenarios 4 and 5) than for the acted speech database EMA (0.16 in scenario 11, to 0.23 in scenario 10). The error in recognizing acted emotions (0.19 in scenario 13) was approximately 20% below the error in recognizing authentic emotions (0.24 in scenario 3), when all speakers were used. Thus, acted emotions yielded better recognition results.

The result of the human evaluation of the acted emotions (EMA) also gave higher inter-evaluator agreement (0.66–0.80) than for the spontaneous, natural emotions in the VAM corpus (0.41–0.65). For these stereotype emotions the machine recognition even outperformed the human evaluation in terms of error and correlation.

5.1.3. Impact of the database

The two modules VAM I and II are comparable in the number of speakers and sentences. The mean error of the estimation was similar for the two modules VAM I and II (0.27 and 0.23). This was not the case for the evaluation (0.21 and 0.15), which gave a smaller error for VAM II. However, this might be due to the different set of evaluators or due to more explicit emotional content which led to a smaller inter-evaluator deviation.

The correlation coefficient of the estimation was higher for VAM I than for VAM II (0.71 and 0.43 in scenarios 1 and 2, respectively). The same tendency was found for the evaluation (0.65 and 0.56). This discrepancy might be due to the different *a priori* distributions of the emotions in the two database modules. While VAM I has a very narrow distribution of emotion primitives, in particular for *valence*, VAM II has a much wider distribution. For example, the variance for *valence* in VAM I was only 67% of the variance in VAM II. Since the variance of the distribution contributes reciprocally to the correlation coefficient (cf. Eq. (3)), the correlation coefficient for VAM II is intrinsically smaller than for VAM I.

When we compared the separate modules to the joint database (scenarios 1–3) we found that the mean error for the joint database was between the results of the two modules. The same observation was made when only male (scenarios 4, 6, 8) or only female speakers (scenarios 5, 7, 9) were analyzed. Thus we could not make the observation that a larger database automatically yielded better results. However, the advantage gained from using the larger joint VAM database in terms of more thorough training of the rule base might have been over-compensated by different emotion expression styles of the different speakers.

5.1.4. Gender dependency

For the scenarios comparing gender-specific versus non-gender-specific rule bases, we could not observe consistent tendencies despite the fact that male and female speakers express their emotions differently (Schröder et al., 2001). Only for the male speakers in VAM II, a gender-dependent rule base gave remarkable improvements from 0.23 to 0.17 in average error values (scenarios 2, 6, 7). For all other scenarios of VAM I (scenarios 1, 4, 5) or VAM I + II (scenarios 3, 8, 9), the mean error was approximately the same, and independent from using separate estimators for both male and female speakers or one joint estimator.

This result might be caused by the method the rules in the rule system are derived, which are all based on the same feature set for all scenarios. Those features that might indeed have different values depending on the emotion *and* the gender were ruled out by features depending on the emotion only. This can easily happen in the case of greater rule weights.

5.1.5. Speaker dependency

For the EMA database we found that the estimators using only one speaker to build the rule system (scenarios 10, 11, 12), when tested on that particular speaker's speech, achieved better results than the estimator using all three speakers (scenario 13). This coincides with previous work indicating that speaker-dependent training of the estimator achieves the most accurate emotion classification results.

5.1.6. Comparison with respect to the emotion primitives

The mean errors and correlation coefficients mentioned in Table 4 contain the average values for each of the emotion primitives *valence*, *activation*, and *dominance*. We observed that for each of the scenarios, the error in the *valence* dimension was greater than the error in either the *activation* or *dominance* dimension. For scenario 3 (all speakers of VAM I + II) for instance, the individual mean values of the estimation error are 0.34, 0.19, and 0.20, for *valence*, *activation*, and *dominance*, respectively. When observing the correlation between the emotion estimates and the emotion reference we observed a similar discrepancy. For each scenario, the correlation coefficient was smaller for *valence* than for *activation* or *dominance*. In scenario 3 for instance, the correlation coefficient for *valence* was 0.34, while it was 0.73 and 0.71 for *activation* and *dominance*, respectively. Thus *valence* was more difficult to estimate automatically using our feature set than *activation* or *dominance*. Note that the better results obtained for *activation* and *dominance* are in accordance with the inter-evaluator correlation, cf. Table 3.

5.1.7. Comparison to the manual results of the evaluation done by listeners

The estimation results achieved with the automated method can be compared to the evaluation results of the human listeners. While for classical emotion categorization we could simply compare confusion matrices, it is less intuitive to compare emotion assessment results within the emotion space approach in terms of real-valued emotion primitives.

In Table 4, the last two columns recall the evaluation results of Table 3, corrected for the emotion space quantization (cf. Section 3.1) and more detailed for the individual scenarios. The evaluation measures only compare the evaluators amongst themselves, i.e. we do not have a "ground truth" for the emotion. Therefore, the comparison between the estimation error and the evaluation error, as well as the comparison of the respective correlation coefficients can only be a rough one.

We can see that the estimation performs in the same range as the human evaluation. In most cases the human agreement is still higher than the machine recognition. However, a comparison of the human evaluation performance and the machine recognition with respect to the database modules is difficult due to the different set of evaluators.

It can be summarized that the fuzzy logic system for emotion estimation is well suited for this task, since a small estimation error and a moderate to high correlation to the emotion reference can be observed. Estimation results for acted emotions in performed speech yield slightly higher results than for authentic emotions expressed in unrehearsed natural speech.

5.2. Comparing emotion primitives estimation and categorical classification

The proposed method to estimate emotion primitives from acoustic features derived from the speech signal was shown to yield low errors and a high correlation to the reference. Since most previous studies target discrete emotion categories instead of continuous-valued emotion primitives, it is important to compare the estimation errors from the fuzzy logic estimator to the performance achieved with categorical recognition. To facilitate this comparison, we analyzed the EMA corpus, since each sentence in this corpus has a defined emotion category label. The recognition rates for the EMA corpus may serve as a rule of thumb for the VAM corpus, for which strict emotion categorization could not be applied due to the lack of objective emotion class labels.

Toward enabling this comparison, we conducted a straightforward classification task. We used the emotion primitives estimates as an input to a k-nearest neighbor (kNN) classifier. The kNN classifier estimates the *a posteriori* probability $P(Q|\boldsymbol{x})$ of the emotion class $Q \in \{angry, happy, neutral, sad\}$ given the emotion primitives $\boldsymbol{x} = (x^{(V)}, x^{(A)}, x^{(D)})^{\mathrm{T}}$ for a local volume element in the 3D emotion space from given training data (Kroschel, 2004). Depending on the feature set and the data, this classifier achieved results comparable to support vector machines and linear discriminant classifiers (Lee and Narayanan, 2005; Hammal et al., 2005), and in some cases outperformed these other classifiers (Dellaert et al., 1996; Yu et al., 2002). The experiment was done using *leave-one-out* (LOO) cross-validation.

We tested $k \in \{1, 3, 5, 7, 9\}$ as a parameter of the kNN classifier, and applied the classification to both the individual speakers of the EMA database (speaker-dependent task) and the combined set of all sentences across all speakers in the database (speaker-independent task). The best recognition rate was achieved using $k = 7$ for the speaker-dependent, and $k = 9$ for the speaker-independent task. For these parameters, the average recognition rate was 83.5% for the speaker-dependent, and 66.9% for the speaker-independent task. The confusion matrices are

Table 5
Confusion matrix for automatically classifying emotion categories from emotion primitives using a *k*NN approach: speaker-dependent results

	Angry	Happy	Neutral	Sad
Angry	**91.9**	2.0	4.3	1.9
Happy	18.8	**80.5**	0.7	0.0
Neutral	0.7	0.0	**85.4**	13.9
Sad	0.0	0.0	26.6	**73.4**

Table 6
Confusion matrix for automatically classifying emotion categories from emotion primitives using a *k*NN approach: speaker-independent results

	Angry	Happy	Neutral	Sad
Angry	**84.1**	13.9	2.0	0.0
Happy	14.6	**72.9**	10.4	2.1
Neutral	1.2	13.5	**49.1**	36.2
Sad	0.6	5.1	30.8	**63.5**

given in Tables 5 and 6, respectively. *Angry* achieved the best recognition results (91.9% and 84.1%, respectively). In contrast, the classification error was highest for *neutral* and *sad* for both tasks (*neutral* → *sad*: 13.9%/36.2%, *sad* → *neutral*: 26.6%/30.8%). Another major difference in the results between the speaker-dependent and the speaker-independent task was found in the mutual misclassification of *happy* and *neutral* (*happy* → *neutral*: 0.7%/10.4%, *neutral* → *happy*: 0.0%/13.5%).

It can be summarized that the emotion primitives estimation lends itself well to emotion categorization. An average estimation error of 0.17 corresponded to an average recognition rate of 83.5% for speaker-dependent emotion recognition. The emotion estimation and classification errors are mainly due to ambiguous neutral speech that actually sounded emotionally charged (happy or sad) rather than neutral. This effect is caused by speaker-dependent expression variations that cannot be captured by the emotion categories, as will be discussed in the following section.

6. Speaker-dependent variability in emotion expression

In this section, we consider speaker dependent variability in the inherent range of emotional expressions. Previous emotion recognition experiments, which target overall categorization of speech-based evidence into a few discrete emotion classes, by nature of their formulation are not suited to describing the underlying variability well. In contrast, the real-valued emotion space approach using emotion primitives considered in this paper lends itself to a conceptual description of speaker dependent variability as described below.

To start with, let us recall that there is a degradation in classification performance if we switch from speaker-dependent to speaker-independent emotion recognition. The reason for the higher amount of misclassifications is the high degree of speaker-dependency in the expression of emotions. To investigate this fact further, we analyzed the emotion primitives obtained from the EMA corpus as a function of both the emotion category and the speaker (Grimm et al., 2006a).

Based on evaluation results from 18 human listeners, we calculated the centroids and the covariances for each emotion cluster in the 3D emotion primitives space. Fig. 5 shows the results indicating the 2σ-regions for each emotion category and each speaker individually. In (Grimm et al., 2006a), these cluster centroids were found comparable to class distributions achieved with other emotion evaluation methods. The position of the individual emotion clusters and the distance of their centroids is a measure for the variability in the emotion expression behavior of a speaker (Grimm et al., 2006b).

The emotion classes clustered in individual subspaces of the emotion primitives space. However, the *sad* and *neutral* classes, as well as the *happy* class of speaker 3 were located relatively closely. The spatial ambiguity of these emotion categories led to a high number of misclassifications. The considerably greater variance of the emotion clusters for speaker 1 explains the *happy* and *angry* mismatches observed in the classification experiment. Similarly, the low confusion error of *angry* and *sad* can be explained by the large distance between the respective emotion clusters in the emotion primitive space.

Both, for *neutral* ↔ *happy* and *neutral* ↔ *sad*, we observed an increase in error rates of approximately 10% absolute, when we switched from speaker-dependent to speaker-independent emotion recognition. These results can also be explained by the speaker-dependent emotion cluster locations, as shown in Fig. 5. It can be observed that the distance between the *neutral* and the *sad* clusters of speaker 3 is relatively large when compared to the distance between the *neutral* cluster of speaker 3 and the *sad* cluster of speaker 2. Thus the confusion error was higher when the classifier was trained with sentences from both speakers than in the case of training with sentences from speaker 3 only.

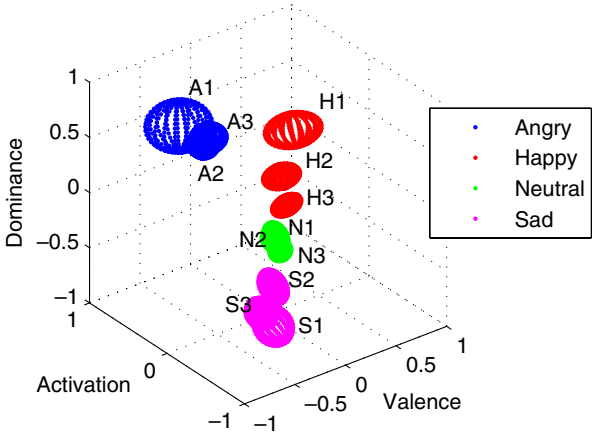

Fig. 5. Covariance plot of the emotion classes *Angry* (A), *Happy* (H), *Neutral* (N), and *Sad* (S) for speakers 1, 2, and 3 in the EMA database.

The location of the cluster for happy emotions in the 3D emotion space was significantly different for the individual speakers. While the calculated centroid was $\mathbf{x}(happy) = (0.58, 0.44, 0.27)^T$ for speaker 1, it was $\mathbf{x}(happy) = (0.23, 0.13, 0.11)^T$ and $\mathbf{x}(happy) = (0.12, -0.08, -0.01)^T$ for speakers 2 and 3, respectively. Thus, the expression of happiness in the latter ones was manifested in a manner that was closer to *neutral*. Such small emotion variability could be due to individual expression patterns, a long-term mood, or other affective influences.

These inter-speaker differences are the main reason for the difficulty in speaker-independent emotion recognition. The wide range of emotion primitive values within *one* emotion category stresses the fact that emotion recognition in terms of simple emotion categories is not sufficient.

7. Conclusions and outlook

In this study, we focused on a novel approach towards automatic estimation of emotions in speech, using emotion primitives, rather than attempting direct classification of emotion categories. Following the concept of a three-dimensional emotion space, we defined emotions to be composed by the values of the three emotion primitives, namely *valence*, *activation*, and *dominance*, each assumed to take values in the range of $[-1, +1]$. We analyzed both acted and spontaneous speech. We tested several scenarios including gender-specific classification, and speaker-dependent vs. speaker-independent emotion estimation.

We introduced a text-free, image-based evaluation method, the self assessment manikins (SAMs). This method was shown to yield low assessment deviation, with an average standard deviation $\bar{\sigma}$ in the range of 0.28–0.38. A moderate to high inter-evaluator agreement was measured. The correlation between the evaluators was between 0.48 and 0.79, depending on the database and the primitive indicating that emotions were fairly reliably assessed in the emotion primitive space by the human listeners.

We also described methods for directly estimating the emotion primitives from acoustic features derived from the speech signal. For the feature set, we extracted 46 different speech signal characteristics describing pitch, energy, speaking rate, and spectrum on an utterance-based segmentation level.

Using a rule-based fuzzy logic estimator, the emotion primitives were estimated on the basis of simple IF–THEN rules. The mean error was between 0.17 and 0.28. Speaker-dependent classification ($\bar{e}_{\min} = 0.16$) gave up to 15% better results than speaker-independent classification for the acted speech. The estimation error for the acted database ($\bar{e} = 0.19$) was approximately 20% below the error for the authentic emotion database ($\bar{e} = 0.24$), which is in accordance with other studies in this field. Moreover, it was found that gender-dependent estimators improved the results only in one case.

The correlation between the automatically-derived estimates of the emotion primitives and the human evaluation ratings was moderate to high ($0.42 \leqslant \bar{r} \leqslant 0.85$). Again, the correlation was higher for the acted database and the speaker-dependent task in general. However, the discrepancy between the experimental settings only resulted in relatively small deviations in the results. The machine recognition was compared to the manual evaluation by human listeners. It showed that the results were in the same range, especially with respect to the mean error. Apart from the acted emotions case, the human performance was always slightly superior to the automatic emotion recognition.

Classical emotion categories were analyzed in terms of the emotion primitives. It was found that the confusion rates of emotion categorization are reflected in the distances of the respective emotion cluster centroids in the 3D emotion primitives space. Moreover, speaker-dependent variations in emotion expression were found to result in significant deviations of the cluster locations in the emotion space.

It can be summarized that emotion primitives provide a good means of handling emotions in both acted and spontaneous productions of natural speech. The results of automatic estimation are in the range of the human evaluation performance. Emotion primitives lend themselves naturally to capturing speaker-dependent emotion expression variations.

In our future work we plan to fuse other modalities with the acoustic analysis. The integration can be accomplished in a straightforward manner by implementing additional rules in the rule base. In this framework, speaker-dependent emotion expression variations are explicitly described by speaker models: this can in turn lead to model-based or parameter-driven emotion recognition that is adapted to individual speakers. Finally, the integration of the automatic emotion estimation into specific man–machine interaction applications, such as a humanoid robot, will indicate further needs in this field of research.

Acknowledgments

This work was supported by a grant of the German Academic Exchange Service DAAD (to M. Grimm), and partly by a grant of the Collaborative Research Center (SFB) 588 "Humanoid Robots – Learning and Cooperating Multimodal Robots" of the Deutsche Forschungsgemeinschaft (to K. Kroschel and M. Grimm).

This research was also supported in part by grants from the National Science Foundation, including a CAREER award (to S. Narayanan) and a graduate fellowship (to E. Mower), the Department of the Army and the Office of Naval Research MURI program.

References

Banse, R., Scherer, K., 1996. Acoustic profiles in vocal emotion expression. J. Personality Social Psychol. 70 (3), 614–636.

Batliner, A., Fischer, K., Huber, R., Spilker, J., Nöth, E., 2000. Desperately seeking emotions or: Actors, wizards, and human beings. In: Proc. ICASSP, pp. 195–200.

Bulut, M., Narayanan, S., Syrdal, A., 2002. Expressive speech synthesis using a concatenative synthesizer. In: Proc. ICSLP, Denver, CO.

Carletta, J., 1996. Assessing agreement on classification tasks: the kappa statistic. Comput. Linguist. 22 (2), 249–254.

Cowie, R., Cornelius, R., 2003. Describing the emotional states that are expressed in speech. Speech Communication 40, 5–32.

Cowie, R., Douglas-Cowie, E., Savvidou, S, McMahon, E., Sawey, M., Schröder, M., 2000. 'FEELTRACE': an instrument for recording perceived emotion in real time. In: Douglas-Cowie, E., Cowie, R., Schröder, M. (Eds.), Proc. ISCA Workshop on Speech and Emotion: A Conceptual Framework for Research, Textflow, Belfast, pp. 19–24.

Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., Taylor, J., 2001. Emotion recognition in human–computer interaction. IEEE Signal Process. Mag. 18 (1), 32–80.

Dellaert, F., Polzin, T., Waibel, A., 1996. Recognizing emotion in speech. In: Proc. ICSLP, Vol. 3, Philadelphia, PA, USA, pp. 1970–1973.

Douglas-Cowie, E., Cowie, R., Schröder, M., 2003. The description of naturally occurring emotional speech. In: Proc. 15th Internat. Conf. on Phonetic Sciences, Barcelona, Spain, pp. 2877–2880.

Fischer, L., Brauns, D., Belschak, F., 2002. Zur Messung von Emotionen in der angewandten Forschung. Pabst Science Publishers, Lengerich.

Fragopanagos, N.F., Taylor, J.G., 2005. Emotion recognition in human–computer interaction. Neural Networks 18 (4), 389–405.

Grimm, M., Kroschel, K., 2005a. Rule-based emotion classification using acoustic features. In: Proc. 3rd Internat. Conf. on Telemedicine and Multimedia Communication, Kajetany, Poland.

Grimm, M., Kroschel, K., 2005b. Evaluierung von natürlichen Emotionen in Sprachsignalen. In: Proceedings 31. Deutsche Jahrestagung für Akustik, DAGA'05, München, Germany, pp. 731–732.

Grimm, M., Kroschel, K., 2005c. Evaluation of natural emotions using self assessment manikins. In: Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), San Juan, Puerto Rico, pp. 381–385.

Grimm, M., Mower, E., Narayanan, S., Kroschel, K., 2006a. Combining categorical and primitives-based emotion recognition. In: Proc. 14th European Signal Processing Conference (EUSIPCO), Florence, Italy.

Grimm, M., Kroschel, K., Narayanan, S., 2006b. Modeling emotion expression and perception behavior in auditive emotion evaluation. In: Proc. ISCA 3rd Internat. Conf. on Speech Prosody, Dresden, Germany, pp. 9–12.

Hammal, Z., Bozkurt, B., Couvreur, L., Unay, D., Caplier, A., Dutoit, T., 2005. Passive versus active: Vocal classification system. In: Proc. Eusipco, Antalya, Turkey.

Hernandez, C., 2005. Einsatz von Fuzzy Logic zur Erkennung von Emotionen in der Sprache, Studienarbeit, Universität Karlsruhe (TH), Germany.

Huang, C.-F, Akagi, M., 2005. A multi-layer fuzzy logical model for emotional speech perception. In: Proc. Eurospeech, Lisbon, Portugal, pp. 417–420.

Kehrein, R., 2002. The prosody of authentic emotions. In: Proc. Speech Prosody Conference, pp. 423–426.

Kroschel, K., 2004. Statistische Informationstheorie: Signal- und Mustererkennung, Parameter- und Signalschätzung, fourth ed. Springer, Berlin.

Lang, P., 1980. Behavioral treatment and bio-behavioral assessment. In: Sidowski, J.B. et al. (Eds.), Technology in Mental Health Care Delivery Systems. Ablex Publishing, Norwood, NJ, pp. 119–137.

Lee, C., Narayanan, S., 2003. Emotion recognition using a data-driven fuzzy inference system. In: Proc. Eurospeech, Geneva, pp. 157–160.

Lee, C.M., Narayanan, S.S., 2005. Toward detecting emotions in spoken dialogs. IEEE Trans. Speech Audio Process. 13 (2), 293–303.

Lee, C., Narayanan, S., Pieraccini, R., 2001. Recognition of negative emotions from the speech signal. In: Proc. IEEE ASRU, Trento, Italy, pp. 240–243.

Lee, S., Yildirim, S., Kazemzadeh, A., Narayanan, S, 2005. An articulatory study of emotional speech production. In: Proc. Eurospeech, pp. 497–500.

Murray, I., Arnott, J., 1993. Toward the simulation of emotion in synthetic speech: a review of the literature on human vocal emotion. J. Acoust. Soc. Amer. 93 (2), 1097–1108.

Nagel, A., 2005. Robuste Pitch-Extraktion für die Erkennung von Emotionen in der Sprache. Diploma Thesis, Universität Karlsruhe (TH), Germany.

Nwe, T.L., Foo, S.W., De Silva, L.C., 2003. Speech emotion recognition using hidden markov models. Speech Communication 41 (4), 603–623.

Oudeyer, P.-Y., 2003. The production and recognition of emotions in speech: features and algorithms. Int. J. Hum. Comput. Stud. 59 (1–2), 157–183.

Russell, J., Mehrabian, A., 1977. Evidence for a three-factor theory of emotions. J. Res. Personality 11, 273–294.

Scherer, K.R., 2005. What are emotions? And how can they be measured? Social Sci. Inf. 44 (4), 693–727.

Schölkopf, B., Smola, A.J., 2002. Learning with Kernels. The MIT Press.

Schröder, M., Cowie, R., Douglas-Cowie, E., Westerdijk, M., Gielen, S., 2001. Acoustic correlates of emotion dimensions in view of speech synthesis. In: Proc. Eurospeech, Vol. 1, Aalborg, pp. 87–90.

Schuller, B., Lang, M., Rigoll, G., 2005. Speaker independent emotion recognition by early fusion of acoustic and linguistic features within ensembles. In: Proc. Interspeech, Lisbon, Portugal, pp. 805–808.

Schuller, B., Lang, M., Rigoll, G., 2006. Recognition of spontaneous emotions by speech within automotive environment. In: Proc. 32. Deutsche Jahrestagung für Akustik, DAGA'06, Braunschweig, Germany, pp. 57–58.

Ververidis, D., Kotropoulos, C., Pitas, I., 2004. Automatic emotional speech classification. In: Proc. ICASSP, Montreal, Canada, pp. 593–596.

Vidrascu, L., Devillers, L., 2005. Real-life emotion representation and detection in call centers data. In: Proc. First Internat. Conf. on Affective Computing and Intelligent Interaction (ACII), Beijing, China, pp. 739–746.

Vidrascu, L., Devillers, L., 2005. Detection of real-life emotions in call centers. In: Proc. Eurospeech, pp. 1841–1844.

Wundt, W., 1896. Grundriss der Psychologie. W. Engelmann, Leipzig.

Yu, Y., Chang, E., Li, C., 2002. Computer recognition of emotion in speech. In: Proc. Intel Internat. Science and Engineering Fair.

Yu, C., Aoki, P.M., Woodruff, A., 2004. Detecting user engagement in everyday conversations. In: Proc. 8th Internat. Conf. on Spoken Language Processing (ICSLP), Vol. 2, Jeju Island, Korea, pp. 1329–1332.