Volume 90
Issue 5
May 2010
ISSN 0165-1684
ELSEVIER
SIGNAL
PROCESSING
Special Section on Statistical Signal & Array Processing
Guest Editors: A.M. Zoubir, M. Viberg and B. Yang
An International Journal
A publication of the European Association for Signal Processing (EURASIP)
EURASIP

Signal Processing 90 (2010) 1415–1423

Contents lists available at ScienceDirect

Signal Processing

journal homepage: www.elsevier.com/locate/sigpro

Emotion recognition from speech signals using new harmony features

B. Yang [*], M. Lugger

Chair of System Theory and Signal Processing, Universität Stuttgart, Pfaffenwaldring 47, 70550 Stuttgart, Germany

ARTICLE INFO

Article history:
Received 28 May 2009
Received in revised form
9 September 2009
Accepted 10 September 2009
Dedicated to J.F. Böhme on the occasion of his 70th birthday
Available online 23 September 2009

Keywords:
Emotion recognition
Feature extraction
Harmony features
Pitch interval

ABSTRACT

In this paper we propose a new set of harmony features for automatic emotion recognition from speech signals. They are based on the psychoacoustic harmony perception known from music theory. Starting from the estimated pitch contour of an utterance, we calculate the circular autocorrelation of the pitch histogram on the logarithmic semitone scale. It measures the occurrence of different two-pitch intervals which cause a consonant or dissonant impression. Experiments of emotion recognition using these harmony parameters in addition to state of the art features show an improved recognition performance.

1. Introduction

The human speech communication consists of two channels, the explicit channel carrying the linguistic content of the conversation (“What was said”) and the implicit channel containing the so-called paralinguistic information about the speaker (“How was it said”) [1–4]. While enormous efforts have been invested in automatic speech recognition (ASR) to extract the linguistic information from the speech samples, still much research is needed to reliably decode the implicit channel.

The list of paralinguistic properties is long: gender, age, emotion, voice quality, stress and nervousness, dialect, pathological state, alcohol or drug consumption, charisma, just to mention a few. Among these properties, the emotion plays a key role in many applications like in call centers to detect angry customers [5–8], in entertainment electronics to gather emotional user feedbacks [9], in ASR to resolve linguistic ambiguities [10–12], and in text-to-speech systems to synthesize emotionally more natural speech [13,14].

Generally, the term emotion describes the subjective feelings in short periods of time which are related to events, persons, or objects [1,15]. Since the emotional state of humans is a highly subjective experience, it is hard to find objective and universal definitions. This is the reason why there are different approaches to model emotions in the psychological literature. One approach is the definition of discrete emotion classes, the so-called basic emotions. Ekman defined seven basic emotions the humans are well familiar with: happiness, sadness, anger, anxiety, boredom, disgust, and neutral [16]. More emotions can be defined by mixtures of the basic emotions. Another approach is the utilization of continuous emotion dimensions. Schlosberg proposed a three-dimensional emotion space: activation (arousal), potency (power), and valence (pleasure, evaluation) [17]. Simpler two-dimensional emotion wheels are also known [18,19]. Both approaches can be combined to locate discrete emotions

[*] Corresponding author.
E-mail addresses: bin.yang@LSS.uni-stuttgart.de (B. Yang), marko.lugger@LSS.uni-stuttgart.de (M. Lugger).

0165-1684/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.sigpro.2009.09.009

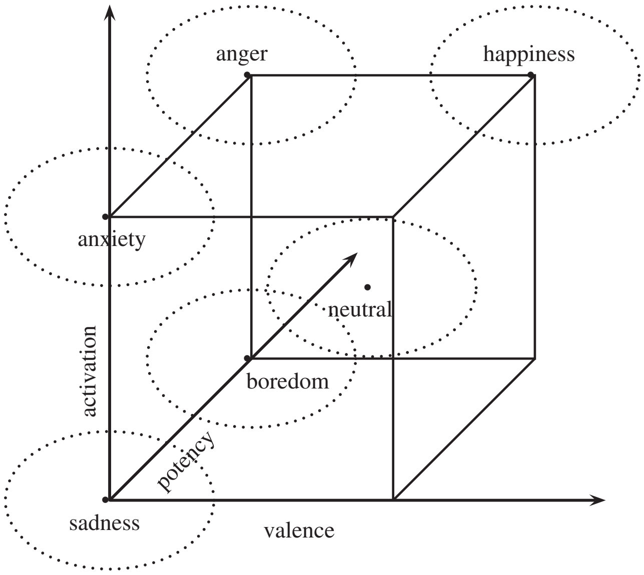

Fig. 1. Three-dimensional emotion space and six basic emotions.

in a continuous emotion space. Fig. 1 shows the locations of six common basic emotions in the three-dimensional emotion space.

Emotion recognition is a statistical pattern classification problem. It consists of two major steps, feature extraction and classification. While the theory of classification is pretty well developed [20], the extraction of distinctive features from patterns is a highly empirical issue and depends strongly on the application and database at hand.

We focus in this paper on the feature issue. One observation in emotion recognition is the particular difficulty to distinguish between, surprisingly to humans, anger and happiness [14,21–23]. A look at Fig. 1 provides an explanation: This pair of emotions differ only in the valence dimension. While it is relatively simple to discriminate different levels of activation by using so-called prosodic features (see Section 3), no sufficiently distinctive acoustic correlates for valence have been found up to now.

Our idea to combat this problem is simple. In music theory, different pitch intervals (consonant, dissonant) and chords are believed to invoke different feelings in listeners. The question is whether there is a similar mechanism between the perception of music and the perception of emotional speech? There are some works studying the relationship between emotion, voice, and music [24], but to our knowledge, there has been no attempt yet to apply the harmony perception known from music to emotion recognition. In this paper, we make a first attempt and propose a new set of harmony features for emotion recognition. We discuss their properties and demonstrate their potential in emotion recognition.

This paper is organized as follows: First we give a quick overview about the major steps of emotion recognition in Section 2 to better understand the overall system. Then we briefly introduce the state of the art features in Section 3. The new set of harmony features is described in details in Section 4. Experimental results of emotion recognition using different feature groups are reported in Section 5.

2. Emotion recognition

Different emotional states experienced by a speaker are reflected in specific patterns of acoustic features in the speech. This means, information containing the emotional state of a speaker is encoded in the acoustic signal and decoded later by the receiving listeners. For automatic emotion recognition, the first step is to find the manner how speakers encode their emotional state in the speech. This is the task of extracting distinctive features from the speech samples. After that, a classification problem to decode the emotional state from the extracted features has to be solved.

2.1. Feature extraction

Assume we have a total number of N_p patterns (utterances of a few seconds in our case). Each of them has to be assigned to one of the N_c emotion classes $\omega_1, \ldots, \omega_{N_\mathrm{c}}$. For each pattern, we generate a fixed number of N_f scalar real valued features. The result of the feature extraction process is a $N_\mathrm{f} \times N_\mathrm{p}$ feature matrix $\mathbf{X}$ whose (i,j)- th element x_{ij} denotes the feature i for pattern j. Each column of $\mathbf{X}$ is the feature vector $\underline{x}$ for one pattern. In general, the more distinctive the features are with respect to the classes to be distinguished, the better the classification performance will be.

Unfortunately, there are no general rules on how to extract good features from raw signals. This is, to a great extent, an empirical step. In emotion recognition, the encoding of the emotional state in speech is highly complex and only partially understood. This motivates our work in this paper to look for novel distinctive features.

2.2. Classification

Emotion recognition is a supervised learning problem. This means, each pattern used for the training of the classifier carries the correct emotion class label. There is a large number of classifiers for supervised learning. The most popular approaches are Bayesian learning, the linear discriminant analysis (LDA), the support vector machine (SVM) as an extension of LDA with a high-dimensional feature space, the multi-layer neural network (NN), and the hidden Markov model (HMM) to capture temporal state transitions [20,25–29].

In this paper, we use the Bayesian learning framework. Based on a model for the class-conditional likelihood $p(\underline{x}|\omega_c)$ of the random feature vector $\underline{x}$ for a given class ω_c, the Bayesian decision theory is used to derive the optimum decision boundary to minimize the overall risk [20,30]. Frequently, $p(\underline{x}|\omega_c)$ is modeled by a Gaussian or a Gaussian mixture model (GMM). The parameters of GMM, i.e. the means, covariance matrices, and mixing proportions of the Gaussians, are estimated from the feature vectors of the training patterns by the iterative expectation-maximization (EM) algorithm. The GMM model order, i.e. the number of Gaussians in $p(\underline{x}|\omega_c)$, can be estimated by information theoretic criterion like the Akaike information criterion (AIC) [31].

Unfortunately, there is no general statement about which classifier is the best one for all applications, see the "No free lunch theorem" in [20]. Currently, we observe at least two trends in emotion recognition: (a) use a very high dimensional feature space in combination with a complex classifier like SVM [25] (b) use a suitable (hierarchical, serial, parallel) combination of pretty simple classifiers like Bayesian–GMM and a moderate number of meaningful features [21,32–34].

One important issue in many applications like call centers is a speaker independent emotion classification. The speakers of the utterances to be classified are not included in the training patterns. They are unknown for the classifier and the deduced learning rules in the training phase. This means, we need one set of training patterns to train the classifier and one set of test patterns from unknown speakers to test the generalization capability of the classifier. An overfitting of the classifier is bad since a too complex classifier and a too complex decision boundary may allow perfect classification of the training patterns, but unlikely perform well on new test patterns [35].

2.3. Feature selection

Beside the previous two major steps of emotion recognition, sometimes we need an additional step called feature selection. It aims at the reduction of the feature vector length N_f. This is necessary to reduce the computational complexity in real-time applications or if we do not have sufficient training patterns. Normally, one would expect an increase in classification performance when more features are used. Nevertheless, the performance can decrease for an increasing N_f if the number of training patterns is too small. This phenomenon is known as the curse of dimensionality [20] and is one of the reasons for overfitting.

In feature selection, a subset of $\tilde{N}_f$ features are selected from the original N_f features in $\underline{x}$ without modifications. This is a combinatoric problem. It is quite similar to the situation when a rich man plans to form a new football team by buying top football players together. The strategy to buy the worldwide best 11 players does not guarantee a successful football team. It is rather important whether the players understand and complete each other well. The same argument applies to the feature selection. The safest strategy to get the best feature set is an exhaustive search which is, however, computationally impractical.

Among different suboptimum approaches for feature selection, we found the sequential floating forward selection (SFFS) algorithm promising. It is an iterative method to find a subset of features that is near the optimal one [36]. At each iteration, a new feature is added to the previous feature subset (forward step). Afterwards, the least significant features are excluded as long as the resulting subset is better in terms of the recognition rate than the previous subset with the same number of features (backward step). This conditional exclusion step is motivated by the fact that a new included feature may carry information that was already present in other features of the previous subset. Thus, the old features can be removed without losing too much discrimination performance. This process is repeated until the desired feature vector length is reached.

3. State of the art features

In speech recognition, the mel frequency ceptral coefficients (MFCC) are the state of the art feature set. They are segmental features since they are generated for each short speech frame to decode the phonemes. According to [21,37], however, MFCC features are less successful for emotion recognition since now we are looking for paralinguistic than linguistic information. In addition, the emotional state of the speaker is unlikely to change as fast as phonemes. Typically, we assign one emotion to one short utterance of a few seconds. Hence we need suprasegmental features with one feature value per utterance.

Fig. 2 shows all feature groups used in this paper and the corresponding number of features in each group. The five feature groups energy, pitch statistics, duration, formant, and zero-crossing rate (ZCR) are most common in emotion recognition. We call them the standard features in this paper. Recently, the group of voice quality (VQ) parameters was extensively studied in our works [21,38,39] and shows a superior performance. All these feature groups are briefly described below. The new set of harmony features will be presented in details in Section 4.

3.1. Standard features

In our implementation, each utterance is divided into 25 ms speech frames with an overlap of 15 ms between successive frames. Each frame is classified to one of three possible voicing types: voiced, unvoiced, and pause. For each voiced frame, we use the RAPT algorithm from [40] to estimate the pitch f_0, the fundamental frequency of the periodic glottal excitation. The temporal evolution of the pitch inside one utterance is called the pitch contour. Fig. 3 shows a speech signal and its corresponding pitch contour. From that, we calculate 55 pitch features by

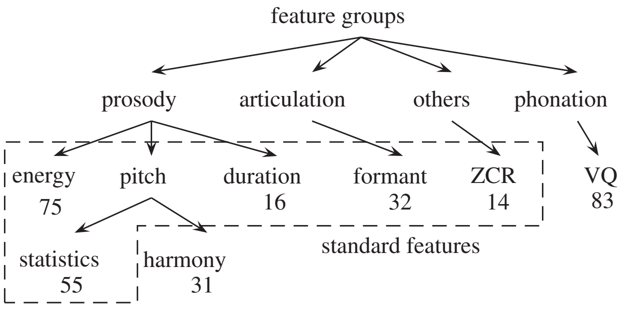

Fig. 2. Feature groups and number of features.

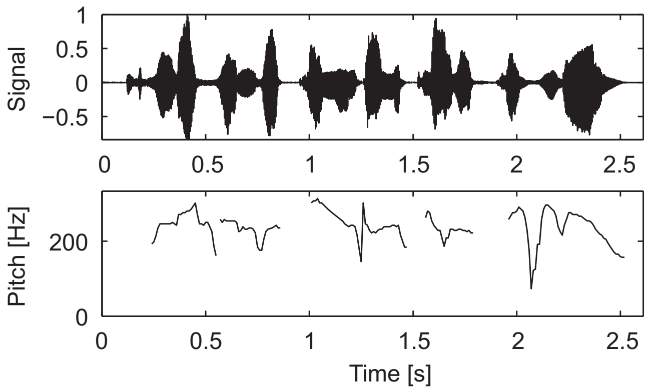

Fig. 3. An utterance and its pitch contour.

Table 1
12 intervals (f_0,f_i) of one octave in Western music.

i	Name	f_i/f_0	$N{:}D$
1	Minor second	1.0595	18:17
2	Major second	1.1225	9:8
3	Minor third	1.1892	6:5
4	Major third	1.2599	5:4
5	Perfect fourth	1.3348	4:3
6	Tritone	1.4142	7:5
7	Perfect fifth	1.4983	3:2
8	Minor sixth	1.5874	8:5
9	Major sixth	1.6818	5:3
10	Minor seventh	1.7818	9:5
11	Major seventh	1.8877	17:9
12	Octave	2.0000	2:1

measuring different statistical values like mean, median, minimum, maximum, range, variance, interquartile range over both the whole contour and segments of the contour containing only raising slopes, plateaus, and falling slopes. They describe both the mean behavior and the variability of the pitch over time and characterize the intonation (speech melody) of the utterance.

In a similar way, we calculate the other four standard feature groups from corresponding contours. The energy contour is derived from frame-based energy estimates and describes the intensity of the uttered words. Duration counts the number of uninterrupted speech frames of the same voicing type. The articulation of an utterance is measured by frequency and bandwidth of the formants of the vocal tract by a linear prediction analysis. The zero-crossing rate (ZCR) contour counts the number of zero crossings of the speech signal within each frame. More details about these feature groups can be found in [33,37,41,42].

3.2. Voice quality parameters

Beside the standard features described above, we also include voice quality (VQ) features that characterize the phonation process. In the theory of the source filter model of speech production [43], the speech signal is assumed to be the result of a filtering of the glottal excitation by the vocal tract. The process at the glottis is called phonation. It characterizes mainly the speaker specific properties, i.e. the paralinguistic information.

We calculate the VQ features by first inverse filtering the speech signal. The influence of the vocal tract is compensated to a great extent and we obtain an estimate of the glottal excitation signal. Then we extend an idea from [44] and calculate various spectral gradients of the glottal signal in the frequency domain. More details about the calculation of these VQ features can be found in [21,29].

4. New harmony features

4.1. Motivation from music theory

The term harmony in music describes the simultaneous use of different pitches. Examples are two-pitch intervals and chords involving more than two pitches. According to music theory, the relative position of these pitches to each other, i.e. the harmony structure of an interval or chord, is mainly responsible for producing a positive or negative impression on the listener [45–48].

In the standard equal temperament of Western music, one octave (frequency ratio 2) is divided into 12 logarithmically equal intervals called semitones. This means, an interval of i semitones corresponds to a frequency pair (f_0,f_i) with $f_i=f_0 2^{i/12}$. Table 1 presents the name, the frequency ratio f_i/f_0 and its rational approximation $N{:}D$ of these 12 intervals. The rational approximation is done by the MATLAB command `rat` with the tolerance value 0.02. An interval is considered to be consonant if $N{:}D$ is a ratio of small integer numbers. This is due to a greater coincidence of harmonics [45]. After only D and N wavelengthes, two sinusoidal signals with the frequency f_0 and f_i are in phase again. Hence perfect fourth $(i=5)$, perfect fifth $(i=7)$, and octave $(i=12)$ are perfect consonances, while minor/major thirds and minor/major sixths are called imperfect consonances. The remaining 5 intervals with large values for N and D are said to be dissonant.

Our postulate is that similar mechanisms could also affect the production and perception of emotional speech. It might be that certain frequency pairs cause a more pleasant impression on the listener than others. Our idea is to take a more careful look at the pitch contour of an utterance. In Section 3.1, we only calculated some statistical values from the pitch contour without a psychoacoustic motivation. We called them the statistical features of the pitch group in Fig. 2. Below we want to derive additional harmony features from the same pitch contour to characterize the relationship between different pitches.

Note that, in contrast to intervals and chords in music, different pitches of speech are not produced at the same time. But due to the human memory, they can be treated in the same way as simultaneous pitches as long as they occur in a short time period like in a short utterance.

4.2. Interval features

First we study features describing pitch intervals. They form a subset of harmony features. We call them the interval (INT) features. Let f denote the pitch estimates from the voiced speech frames. We transform them from the linear scale to the logarithmic semitone scale

$$s(f) = L\log_2(f/f_{\text{ref}}) + s_{\text{ref}}. \tag{1}$$

$L = 12$ semitones per octave are used in Western music. The MIDI standard proposes $f_{\text{ref}} = 440\,\text{Hz}$ for the standard pitch A4 and $s_{\text{ref}} = s(f_{\text{ref}}) = 69$ to count pitches. In our case, the choice of f_{ref} and s_{ref} is arbitrary because we are interested in the relative position of the frequencies rather than their absolute values.

Let S be a random variable whose realization is s. We denote its probability density function (PDF) by $p(s)$. We propose to use the second-order autocorrelation of $p(s)$ to describe the occurrence of pitch pairs with a given logarithmic distance s:

$$r(s) = \int_{-\infty}^{\infty} p(s+\lambda)p(\lambda)\,\mathrm{d}\lambda. \tag{2}$$

The following properties of $r(s)$ hold:

P1 $r(s)$ can be interpreted as the PDF of the difference $I = S_1 - S_2$ of two independent random variables S_1 and S_2 with the same PDF $p(s)$. This implies $\int_{-\infty}^{\infty} r(s)\,\mathrm{d}s = 1$.

P2 If the pitch contour contains several discrete pitches s_i with the probability P_i, i.e. $p(s) = \sum_i P_i \delta(s - s_i)$ with $\delta(s)$ being the Dirac function, then

$$r(s) = \left(\sum_i P_i^2\right)\delta(s) + \sum_{i\neq j} P_i P_j[\delta(s - s_{ij}) + \delta(s - s_{ji})],$$

with $s_{ij} = s_i - s_j$. The first term with $\delta(s)$ describes the correlation of each pitch with itself (interval prime) and is irrelevant for the harmony study. The remaining mixture terms reflect pitch distances at s_{ij} and s_{ji}.

The above introduced measures $p(s)$ and $r(s)$ have two drawbacks: First they treat harmonics of a fundamental frequency as independent pitches. This does not coincide with the observation in music that the perception of a pitch interval is largely invariant with respect to modifications of pitch frequencies by powers of 2 (octaves). Second, $p(s)$ is defined for $s \in \mathbb{R}$ and we do not have enough pitch samples from a short utterance to estimate $p(s)$ reliably.

Therefore, we introduce below a circular pitch on the semitone scale

$$s_\circ = \text{mod}_L(s), \quad 0 \le s_\circ < L. \tag{3}$$

$\text{mod}_L(s)$ is the modulo of the division s/L. It maps all octaves into a single one. In music information retrieval, similar concepts as pitch class profiles (PCP) [49] and chroma vectors [50] are known for chord estimation. But these concepts use directly the short-time spectrum of music signals with a modulo logarithmic frequency scale as features, while we perform an explicit pitch estimation and consider the correlation of the pitch PDF.

Let $p_\circ(s)$ be the PDF of the random variable $S_\circ$ whose realization is $s_\circ$. Its relationship to the PDF of the non-circular pitch S is

$$p_\circ(s) = \begin{cases} \sum_{k=-\infty}^{\infty} p(s+kL), & 0 \le s < L, \\ 0 & \text{otherwise}. \end{cases} \tag{4}$$

As a measure for the occurrence of pitch pairs without taking the height of their octaves into account, we introduce the circular correlation of $p_\circ(s)$

$$r_\circ(s) = \int_0^L p_\circ(\text{mod}_L(s+\lambda))p_\circ(\lambda)\,\mathrm{d}\lambda \tag{5}$$

for $0 \le s < L$. We can show (without proof):

P3 $r_\circ(s) = \sum_{k=-\infty}^{\infty} r(s+kL)$ for $0 \le s < L$.

P4 $r_\circ(s)$ also has the interpretation of the PDF of a random variable. Let $S_{\circ,i} = \text{mod}_L(S_i)$ $(i = 1, 2)$ be two independent random variables with the same PDF $p_\circ(s)$. Then $r_\circ(s)$ is the PDF of the circular pitch distance $I_\circ = \text{mod}_L(S_{\circ,1} - S_{\circ,2}) = \text{mod}_L(S_1 - S_2)$.

P5 $r_\circ(s) = r_\circ(L - s)$. This symmetry property illustrates why two complementary intervals with a sum equal to one octave (e.g. perfect fourth and perfect fifth) provide the same consonant or dissonant impression, see Table 1.

In addition to $r_\circ(s)$, we introduce a single dissonance parameter DIS to summarize the consonance and dissonance effects of all occurring pitch intervals. Let $d(s)$ be a suitable dissonance function for the logarithmic pitch distance s or equivalently the frequency ratio $2^{s/L}$. $d(s)$ is large if this pitch interval sounds dissonant and $d(s)$ is small if it sounds consonant. The mean dissonance DIS is defined by

$$\text{DIS} = \int_0^L d(s)r_\circ(s)\,\mathrm{d}s. \tag{6}$$

It is the expectation of the random variable $d(I_\circ)$ where $I_\circ = \text{mod}_L(S_1 - S_2)$ is the circular pitch distance with the PDF $r_\circ(s)$.

4.3. Practical implementation

For the practical implementation, we approximate the PDF by a histogram and the expectation by a sample average. Plot (a) in Fig. 4 shows the histogram $h_S(k)$ of the pitch S using a histogram bin width of $\frac{1}{2}$ semitone. We used the pitch contour from Fig. 3 and chose $f_{\text{ref}} = 440\,\text{Hz}$ and $s_{\text{ref}} = 69$ in (1). Plots (b) and (c) show the histogram $h_{S_\circ}(k)$ of the circular pitch $S_\circ$ and the histogram $h_{I_\circ}(k)$ of the circular pitch distance $I_\circ$. In both cases, each semitone is divided into $q = 5$ equal steps resulting in a total number of $Lq = 60$ histogram bins. The histograms differ from the PDFs $p(s), p_\circ(s), r_\circ(s)$ in scaling. This, however, does not matter since the Bayesian–GMM classifier we used is invariant with respect to a scaling of the features. Similar to property P5 in the previous subsection, the histogram $h_{I_\circ}(k)$ $(0 \le k < Lq)$ is symmetric in the sense of $h_{I_\circ}(k) = h_{I_\circ}(Lq - k)$. Hence we only use $Lq/2 = 30$ values $h_{I_\circ}(k)$ $(1 \le k \le Lq/2)$ as our new interval features for emotion recognition. The value $h_{I_\circ}(0)$ counts only the

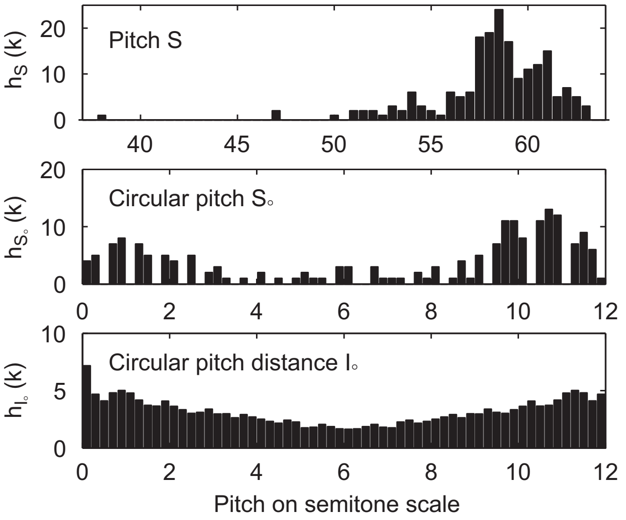

Fig. 4. Histogram of pitch and pitch difference.

correlation of each pitch with itself (interval prime) and is irrelevant for our purpose, see property P2.

The expectation in (6) is approximated by

$$\mathrm{DIS} = \sum_{k=1}^{Lq/2} d(s_k) h_{I_\circ}(k). \tag{7}$$

$s_k = (k+0.5)/q$ is the center of the k-th histogram bin. The dissonance function $d(s)$ is implemented as the geometric mean $\sqrt{N(s)D(s)}$ where $N(s)/D(s)$ is the rational approximation of the frequency ratio $2^{s/L}$ with a tolerance value of 0.02. Using DIS and 30 histogram values of $h_{I_\circ}(k)$, we have a total number of 31 interval (INT) features.

4.4. Triad features

In addition to the previous interval features, the group of harmony features can also contain measures for chords. In contrast to intervals, the perception of three-pitch triads like major, minor, diminished, and augmented is not completely understood yet [51]. Nevertheless, our concept of autocorrelation of pitch PDF in Section 4.2 can be easily extended to triads. In this case, we use the third-order circular autocorrelation of $p_\circ(s)$

$$r_\circ(s_1, s_2) = \int_0^L p_\circ(\mathrm{mod}_L(s_1+\lambda)) p_\circ(\mathrm{mod}_L(s_2+\lambda)) p_\circ(\lambda)\, \mathrm{d}\lambda. \tag{8}$$

It has the interpretation of the bivariate PDF of two modulo pitch differences $I_{\circ,i} = \mathrm{mod}_L(S_i - S_3)$ $(i=1,2)$ of a triad (S_1, S_2, S_3) where S_i are three independent random variables with the same PDF $p(s)$. This third-order correlation will be implemented in the future.

5. Experiments

5.1. Emotion database

In emotion recognition, the used database containing the training and test patterns plays a crucial role. A list of the most important databases is given in [52]. One distinguishes between acted and natural speech databases. Clearly, emotion recognition from natural speech, which is the goal in practice, is much more difficult than emotion recognition from acted speech due to the much larger variation of emotional expressions in natural conversation. Unfortunately, natural speech databases for emotion recognition (e.g. from call centers) are seldom public available due to the privacy of speakers. In addition, the acquisition and labeling of a large size database is very expensive.

For these reasons, we use the public Berlin emotion database of acted speech [53] in this paper to classify six basic emotions: happiness, sadness, anger, anxiety, boredom, and neutral. It is a pretty small database containing 694 utterances for these six emotions. Ten actors have been invited to speak short utterances between two and five seconds in German. The signals are sampled at 16 kHz.

5.2. Procedure

We perform a speaker independent emotion recognition by using the "leaving-one-speaker-out" cross validation. This means, 10 speakers are partitioned into one group of nine speakers for training and one last speaker for testing. We use the Bayesian classifier with a Gaussian class-conditional likelihood. Experiments with a GMM class-conditional likelihood and order estimation by AIC have not shown remarkable performance improvements for this database. The reason is the small diversity of the acted emotions in the Berlin database. For each utterance, we calculate a total number of 306 features, see Fig. 2. In order to evaluate the benefit of new feature sets, we study the recognition rate using an increasing number of feature groups [54]. The baseline is to select the best $\tilde{N}_\mathrm{f}$ features by the SFFS algorithm from the 192 standard features (energy, pitch statistics, duration, formant, ZCR). Then we repeat the same experiment by selecting the best features from 275 standard-plus-VQ features and from 306 standard-plus-VQ-plus-INT features. The recognition rate shown in the plots below is averaged over all 10 speaker partitions and all emotions to be distinguished.

5.3. Two-class results

First we investigate for which of the three emotion dimensions in Fig. 1 the new interval features are distinctive. For this purpose, we compare all five pairs of emotions where both emotions in each pair differ only in one dimension. Fig. 5 shows the recognition rates for an increasing number $\tilde{N}_\mathrm{f}$ of selected features. As known from literature, the recognition rate is not monotonically increasing in $\tilde{N}_\mathrm{f}$ due to the curse of dimensionality and the suboptimality of the feature selection algorithm SFFS.

In plot (a), we compare anger vs. happiness which differ only in the valence dimension. The use of the voice quality (VQ) features in addition to the standard features improves the recognition rate by roughly 4%. The new

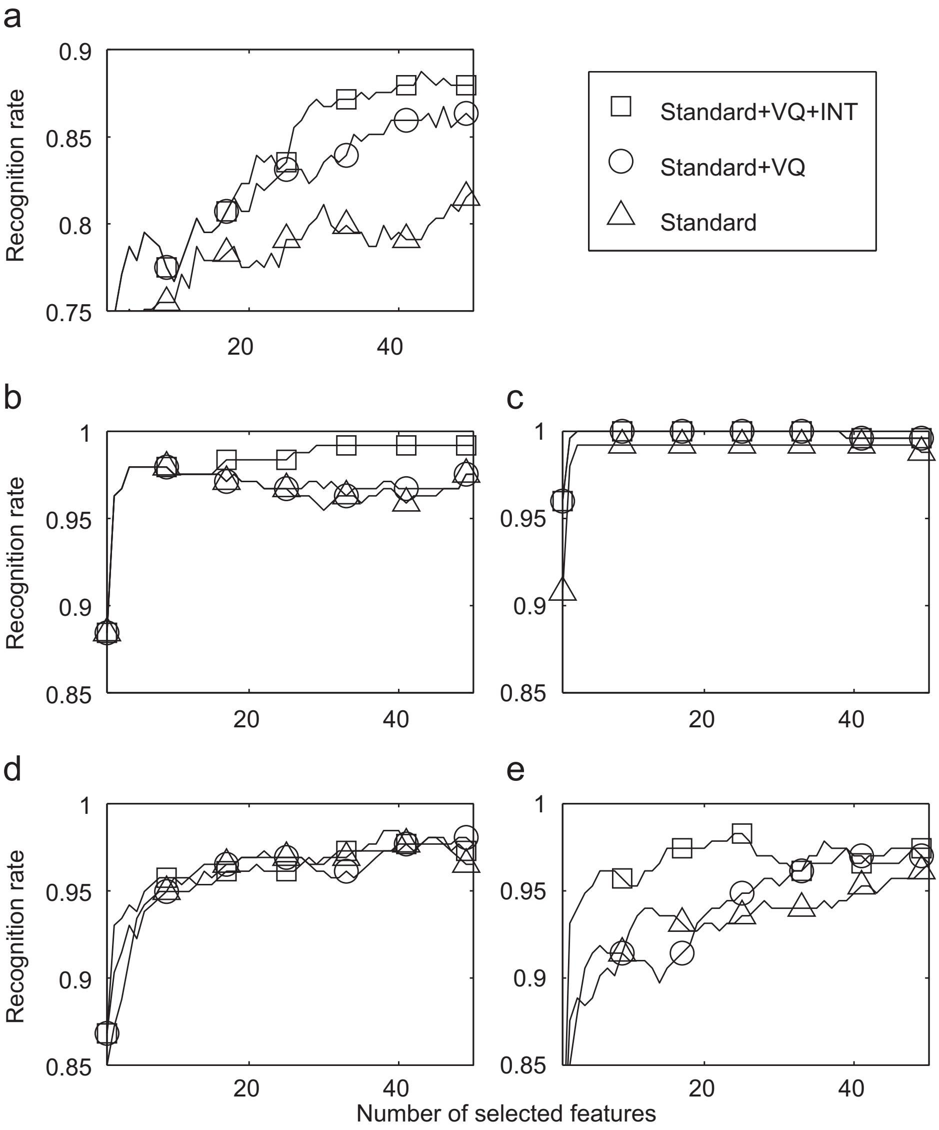

Fig. 5. Results of two-class emotion recognition: (a) Anger vs. happiness, (b) Sadness vs. anxiety, (c) Boredom vs. anger, (d) Anxiety vs. anger, (e) Sadness vs. boredom.

group of interval (INT) features further improves the classification by 2%. In plot (b) and (c), we compare sadness vs. anxiety and boredom vs. anger. The discriminative dimension in this case is activation. In the former case, INT features improve the discrimination while VQ features do not. In the latter case, only a few standard features are sufficient to distinguish between boredom and anger. In plot (d) and (e), we study the potency dimension which distinguishes anxiety from anger and sadness from boredom. Along this dimension, neither the VQ nor the INT features contribute remarkably to the classification for a large number $\tilde{N}_f$ of selected features.

The conclusion from these experiments is clear: The new set of harmony features is useful for emotion recognition, but not equal for all emotion dimensions. It seems to be distinctive for the valence (plot a) and, in part, for the activation (plot b) dimension. But the deep relationship between pitch intervals and emotion perception is still an open question.

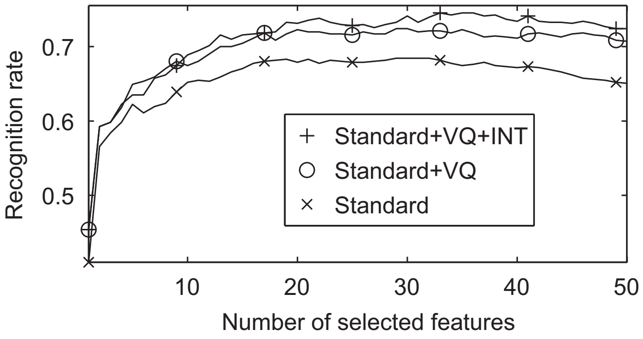

Fig. 6. Result of six-class emotion recognition.

5.4. Six-class results

Fig. 6 shows the result when we classify all six basic emotions happiness, sadness, anger, anxiety, boredom, and neutral by a single stage (flat) Bayesian–GMM

Table 2
Confusion matrix of a six-class emotion recognition using Standard + VQ features.

	Happ.	Bored.	Neutr.	Sadn.	Anger	Anxi.
Happ.	**51.8**	0	1.8	0	32.1	14.3
Bored.	0	**77.7**	8.0	12.5	0	1.8
Neutr.	3.9	11.5	**67.3**	4.8	2.9	9.6
Sadn.	0	5.7	5.0	**86.0**	0	3.3
Anger	*21.9*	0.7	0	0	**71.5**	5.8
Anxi.	13.2	0.8	2.5	5.0	7.4	**71.1**

Table 3
Confusion matrix of a 6-class emotion recognition using Standard + VQ + INT features.

	Happ.	Bored.	Neutr.	Sadn.	Anger	Anxi.
Happ.	**52.7**	0.9	1.8	0	33.9	10.7
Bored.	4.5	**84.8**	0.9	3.6	0	6.3
Neutr.	8.7	13.5	**52.9**	6.7	2.9	15.4
Sadn.	0.8	5.8	1.7	**87.6**	0	4.1
Anger	*8.0*	0	0	0	**86.1**	5.8
Anxi.	6.6	1.7	5.0	2.5	7.4	**76.9**

classifier. Again, we observe a considerable improvement of the recognition rate by roughly 4% and 2%, respectively, when we use the VQ and INT features in addition to the standard ones.

Tables 2 and 3 present the confusion matrices when a total number of $\tilde{N}_{\mathrm{f}} = 50$ features were selected from the Standard+VQ feature groups (Table 2) and from the Standard + VQ + INT feature groups (Table 3). The first column indicates the given emotion. The numbers in each row represent the percentage of patterns classified to different emotions. Clearly, the use of harmony features significantly improves the classification of anger, in particular against happiness. Also the recognition of boredom is improved. As a price, the state neutral becomes more confused with other emotions. The average recognition rate over all emotions is improved by 2%.

6. Conclusion

The contributions of this paper are twofold. First we gave a quick review of automatic emotion recognition from speech signals. Then we presented a novel set of so-called harmony features based on the psychoacoustic perception of pitch intervals and chords from music theory. We discussed their properties. First experiments have shown that these new features are correlated to the valence and activation dimension of emotion and are able, if used in addition to the state of the art features, to improve the recognition performance. Future works include experiments with the triad features in (8), an optimization of the dissonance function $d(s)$ in (6), a study of these harmony features in other classifiers or combinations of classifiers, and an evaluation of these harmony features with more realistic non-acted emotional speech.

Acknowledgment

We thank Stefan Uhlich for valuable discussions.

References

[1] R. Cowie, et al., Emotion recognition in human–computer interaction, IEEE Signal Processing Magazine 18 (2001) 32–81.

[2] R. Cowie, E. Douglas-Cowie, Speakers and hearers are people: reflections on speech deterioration as a consequence of acquired deafness, in: K.E. Spens, G. Plant (Eds.), Profound Deafness and Speech Communication, Wiley, New York, 1995.

[3] H. Traunmüller, Paralinguistic phenomena, in: U. Ammon, N. Dittmar et al.(Eds.), Sociolinguistics: An International Handbook of the Science of Language and Society, Walter de Gruyter, Berlin, 2005, pp. 653–665.

[4] J. Laver, The Phonetic Description of Voice Quality, Cambridge University Press, Cambridge, 1980.

[5] F. Burkhardt, M. van Ballegooy, et al., An emotion-aware voice portal, in: Electronic Speech Signal Processing Conference, 2005.

[6] F. Burkhardt, T. Polzehl, et al., Detecting real life anger, in: Proceedings of the IEEE ICASSP, 2009, pp. 4761–4764.

[7] L. Devillers, L. Vidrascu, Real-life emotions detection with lexical and paralinguistic cues on human–human call center dialogs, in: INTERSPEECH, 2006.

[8] D. Morrison, R. Wang, et al., Ensemble methods for spoken emotion recognition in call-centres, Speech communication 49 (2007) 98–112.

[9] A. Batliner, et al., You stupid tin box—children interacting with the AIBO robot: a cross-linguistic emotional speech corpus, in: Proceedings of the Fourth International Conference of Language Resources and Evaluation, 2004, pp. 171–174.

[10] J. Hirschberg, C. Avesani, The role of prosody in disambiguating potentially ambiguous utterances in English and Italian, in: ESCA Workshop on Intonation, 1997.

[11] R. Gretter, D. Seppi, Using prosodic information for disambiguation purposes, in: INTERSPEECH, 2005, pp. 1821–1824.

[12] M. Cernak, C. Wellekens, Emotional aspects of intrinsic speech variabilities in automatic speech recognition, in: International Conference on Speech and Computer, 2006, pp. 405–408.

[13] M. Schröder, Emotional speech synthesis: a review, in: EUROSPEECH, 2001, pp. 561–564.

[14] M. Schröder, R. Cowie, et al., Acoustic correlates of emotion dimensions in view of speech synthesis, in: EUROSPEECH, 2001, pp. 87–90.

[15] M. Lewis, J. Haviland-Jones, L.F. Barrett, Handbook of Emotions, The Guilford Press, 2008.

[16] P. Ekman, An argument for basic emotions, Cognition and Emotion 6 (1992) 169–200.

[17] H. Schlosberg, Three dimensions of emotions, Psychological Review 61 (1954) 81–88.

[18] R. Plutchik, Emotion: A Psychoevolutionary Synthesis, Harper & Row, New York, 1980.

[19] A. Hanjalic, Extracting moods from pictures and sounds: towards truly personalized TV, IEEE Signal Processing Magazine 23 (2006) 90–100.

[20] R.O. Duda, P.E. Hart, D.G. Stork, Pattern Classification, second ed., Wiley, New York, 2001.

[21] M. Lugger, B. Yang, Psychological motivated multi-stage emotion classification exploiting voice quality features, in: F. Mihelic, J. Zibert (Eds.), Speech Recognition, In-Tech, 2008 (Chapter 22).

[22] C. Pereira, Dimensions of emotional meaning in speech, in: ITRW on Speech and Emotion, 2000, pp. 25–28.

[23] Y. Xu, S. Chuenwattanapranithi, Perceiving anger and joy in speech through the size code, in: Proceedings of the International Conference on Phonetic Sciences, 2007, pp. 2105–2108.

[24] I. Fonagy, Emotions voice and music, Royal Swedish Academy of Music, 1981, pp. 51–79.

[25] B. Schuller, D. Arsic, et al., Emotion recognition in the noise applying large acoustic feature sets, in: Speech Prosody, Dresden, 2006.

[26] B. Schölkopf, A. Smola, Learning with Kernels, MIT Press, Cambridge, MA, 2002.

[27] V.N. Vapnik, The Nature of Statistical Learning Theory, Springer, Berlin, 2000.

[28] B. Schuller, G. Rigoll, M. Lang, Hidden Markov model-based speech emotion recognition, in: Proceedings of the IEEE ICASSP, 2003.

[29] M. Lugger, B. Yang, Extracting voice quality contours using discrete hidden Markov models, in: Proceedings of the Speech Prosody, 2008.

[30] S.M. Kay, Fundamentals of Statistical Signal Processing, Detection Theory, Vol. 2, Prentice-Hall, Englewood Cliffs, NJ, 1998.

[31] H. Akaike, A new look at the statistical model identification, IEEE Transactions on Automatic Control 19 (1974) 716–723.

[32] L.I. Kuncheva, Combining Pattern Classifiers Methods and Algorithms, Wiley, New York, 2004.

[33] D. Ververidis, C. Kotropoulos, Emotional speech classification using Gaussian mixture models, in: IEEE International Symposium on Circuits and Systems, 2005, pp. 2871–2874.

[34] M. Lugger, B. Yang, Combining classifiers with diverse feature sets for robust speaker independent emotion recognition, in: Proceedings of the EUSIPCO, 2009.

[35] J. Reunanen, Overfitting in making comparisons between variable selection methods, Journal of Machine Learning Research 3 (2003) 1371–1382.

[36] P. Pudil, F.J. Ferri, et al., Floating search methods for feature selection with nonmonotonic criterion, Pattern Recognition—Conference B: Computer Vision 2 (1994) 279–283.

[37] T. Nwe, S. Foo, L.D. Silva, Speech emotion recognition using hidden Markov models, Speech communication 41 (2003) 603–623.

[38] M. Lugger, B. Yang, The relevance of voice quality features in speaker independent emotion recognition, in: Proceedings of the IEEE ICASSP, vol. 4, 2007, pp. 17–20.

[39] M. Lugger, B. Yang, Cascaded emotion classification via psychological emotion dimensions using a large set of voice quality parameters, in: Proceedings of the IEEE ICASSP, 2008, pp. 4945–4948.

[40] D. Talkin, W. Kleijn, K. Paliwal, A robust algorithm for pitch tracking (RAPT), Speech Coding and Synthesis (1995) 495–518.

[41] C.M. Lee, S.S. Narayanan, Toward detecting emotions in spoken dialogs, IEEE Transactions on Speech and Audio Processing 13 (2005) 293–303.

[42] R. Banse, K.R. Scherer, Acoustic profiles in vocal emotion expression, Journal of Personality and Social Psychology 70 (1996) 614–636.

[43] G. Fant, Acoustic Theory of Speech Production, The Hague, Mouton, 1970.

[44] K. Stevens, H. Hanson, Classification of glottal vibration from acoustic measurements, in: O. Fujimura, M. Hirano (Eds.), Vocal Fold Physiology, Hiltop University Press, 1994, pp. 147–170.

[45] H.L.F. Helmholtz, On the Sensations of Tone as a Physiological Basis for the Theory of Music, second ed., Dover Publications, New York, 1877.

[46] R. Plomp, J.M. Levelt, Tonal consonance and critical bandwidth, Journal of the Acoustical Society of America 38 (1965) 548–560.

[47] G. Schellenberg, S. Trehub, Frequency ratios and the perception of done patterns, Psychonomic Bulletin & Review (1994) 191–201.

[48] D. Schwartz, C. Howe, D. Purves, The statistical structure of human speech sounds predicts musical universals, The Journal of Neuroscience (2003) 7160–7168.

[49] T. Fujishima, Real-time chord recognition of musical sound: a system using common lisp music, in: ICMC, 1999, pp. 464–467.

[50] G.H. Wakefield, Mathematical representation of joint time-chroma distribution, in: SPIE, vol. 3807, 1999.

[51] N.D. Cook, T.X. Fujidawa, The psychophysics of harmony perception: harmony is a three-tone phenomenon, Empirical Musicology Review 1 (2) (2006) 106–126.

[52] D. Ververidis, C. Kotropoulos, A state of the art review of emotional speech databases, in: Proceedings of the First Richmedia Conference, 2003.

[53] F. Burkhardt, A. Paeschke, et al., A database of German emotional speech, in: Proceedings of the Interspeech, 2005, pp. 1517–1520.

[54] M. Lugger, B. Yang, An incremental analysis of different feature groups in speaker independent emotion recognition, in: Proceedings of the International Conference on Phonetic Sciences, 2007, pp. 2149–2152.