Available online at www.sciencedirect.com
ScienceDirect
Speech Communication 53 (2011) 1162–1171
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

Emotion recognition using a hierarchical binary decision tree approach

Chi-Chun Lee [a,*], Emily Mower [a], Carlos Busso [b], Sungbok Lee [a], Shrikanth Narayanan [a]

[a] Signal Analysis and Interpretation Laboratory (SAIL), Electrical Engineering Department, University of Southern California, Los Angeles, CA 90089, USA
[b] Electrical Engineering Department, University of Texas at Dallas, Dallas, TX 75080, USA

Available online 3 July 2011

Abstract

Automated emotion state tracking is a crucial element in the computational study of human communication behaviors. It is important to design robust and reliable emotion recognition systems that are suitable for real-world applications both to enhance analytical abilities to support human decision making and to design human–machine interfaces that facilitate efficient communication. We introduce a hierarchical computational structure to recognize emotions. The proposed structure maps an input speech utterance into one of the multiple emotion classes through subsequent layers of binary classifications. The key idea is that the levels in the tree are designed to solve the easiest classification tasks first, allowing us to mitigate error propagation. We evaluated the classification framework on two different emotional databases using acoustic features, the AIBO database and the USC IEMOCAP database. In the case of the AIBO database, we obtain a balanced recall on each of the individual emotion classes using this hierarchical structure. The performance measure of the average unweighted recall on the evaluation data set improves by 3.37% absolute (8.82% relative) over a Support Vector Machine baseline model. In the USC IEMOCAP database, we obtain an absolute improvement of 7.44% (14.58%) over a baseline Support Vector Machine modeling. The results demonstrate that the presented hierarchical approach is effective for classifying emotional utterances in multiple database contexts.

Keywords: Emotion recognition; Hierarchical structure; Support Vector Machine; Bayesian Logistic Regression

1. Introduction

Emotion recognition is an integral part of quantitative studies of human behavior. The emerging areas of human behavioral signal processing and behavioral informatics offer new analytical tools to support a variety of applications, including the design of natural human–machine interfaces (HMI). Emotionally-cognizant human–computer and human–robot interfaces promise a more responsive and adaptive user experience. In real life settings, behavioral computing must reconcile information in the context of a situated interaction (Brody and Hall, 2008). This is also true of human–machine interactions where the ability to sustain interactions may be hampered by an interacting agent's inability to recognize, track and respond appropriately to the interacting partners (Pantic et al., 2005).

Many applications can benefit from an accurate emotion recognizer. For example, customer care interactions (with a human or an automated agent) can use emotion recognition systems to assess customer satisfaction and quality of service (e.g., lack of frustration) (Herm et al., 2008; Lee and Narayanan, 2005). Other tasks that rely on observational coding of human interaction, such as in therapeutic settings (Amir et al., 2010; Black et al., 2010; Lee et al., 2010) can also benefit from robust emotion recognition. Increasingly, interactive educational systems are becoming commercially available (Kanda et al., 2004; Kapoor and Picard, 2005). These systems must be able to accurately identify a child's emotional state to foster interactions and positive evaluations (Brave et al., 2005; Prendinger et al., 2005; Yildirim et al., 2005, 2011). Understanding a child's certainty in a problem solving and learning task

* Corresponding author.
E-mail address: chiclee@usc.edu (C.-C. Lee).

0167-6393/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2011.06.004

can help scaffold the interaction in a context appropriate way (Black et al., 2008). All these applications can benefit from the design of a robust emotion recognition scheme, which should also be easily adaptable to different interaction scenarios.

The computational emotion recognition framework we describe in this paper is loosely motivated by the Appraisal Theory (Lazarus, 2001) of emotions. Appraisal Theory states that emotion perception is a multi-stage conscious and unconscious process. The appraisal process can be thought of as a series of decisions (e.g., how positive is the stimulus, how novel is the stimulus, what is the cause of the stimulus, etc.). At each stage, an individual appraises the situation, reacts, and reappraises, inducing different emotions in the process (e.g., fear, surprise, and then joy). The proposed framework is inspired by the Appraisal Theory in its approximation of the appraisal and reappraisal processes. We do not, however, propose a direct interpretation or implementation of this theory; rather, we propose a simplified computational model in the form of a hierarchical binary decision tree. The framework splits a single multi-class emotion classification problem into stages of binary emotion classification tasks capturing the idea of appraisal and reappraisal. The key idea of the proposed framework is the recognition, early in the tree, of the most distinguishable emotional classes. The ambiguous emotional classes are recognized at the bottom of the tree, mitigating error propagation.

The key idea behind this proposed emotion recognition framework is the use of binary classifiers in a hierarchical tree structure. There are many well-established state of the art classifiers that can be readily implemented to work with binary classification problems, e.g., logistic regression, Support Vector Machine, Fisher discriminant analysis, etc. The system also benefits from its unweighted recall optimization criterion. In many real life interactions, the *neutral* emotion class is both the most dominant and the most ambiguous emotion class. If the system is optimized on the measure of conventional *accuracy* (number of accurately classified samples by total number of tested samples), it will likely be biased in recognizing only the dominant state accurately (Wanger, 1993). The bias is not desirable in many applications. The average unweighted recall (average percentage of number of accurately recalled utterances for each emotion class) measure can provide a way to assess the performance of our proposed classifier in emotionally biased datasets. The hierarchical structure along with the optimized decision threshold can effectively mitigate the inherent problem of class imbalance and achieve good average unweighted recall percentage.

Several other emotion research works (Xiao et al., 2007; Mao and Zhan, 2010; Hassan and Damper, 2010; Albornoz et al., 2011) have also utilized hierarchical tree structure in performing emotion recognition tasks. The two most similar approaches are the DDAGSVM proposed by Mao and Zhan (2010), and the hierarchical structure proposed by Xiao et al. (2007). In both papers, the hierarchical structures are designed to operate on easier binary classification tasks in their first layer and relatively ambiguous tasks in the last layer of the tree. Our classification framework, proposed independently, shares the same design principle. However, in our framework, we do not restrict each node to classify between pairs of emotion classes. In our framework, each node is flexible in classifying between mixtures of emotion classes. The design framework proposed in this paper can be easily extended to additional emotional corpora even when the emotion class distributions differ. Our proposed framework can effectively cope with class bias.

The presented emotion recognition framework was first evaluated in the Interspeech 2009 Emotion Challenge using the AIBO database. The evaluation metric was the average unweighted recall percentage per emotion class using the AIBO database, which has two different splits: training and evaluation datasets. The database consists of affective speech collected from fifty-one children interacting with an AIBO robot dog (Schuller et al., 2009). The five emotion classes of interest are: *angry*, *emphatic*, *neutral*, *positive*, and *rest*. The class of neutrality is over-represented in this database. We demonstrated the flexibility of this classification framework by applying it to a different emotion database, the USC IEMOCAP database. This database consists of natural dyadic affective spoken interactions of professional actors. It includes both scripted plays and spontaneous dialogs. The four emotion classes of interest are: *angry*, *happy*, *sad*, and *neutral*. Both databases contain natural affective interactions, instead of utterance-by-utterance acted emotional speech.

In the AIBO database experiment, we achieve an average unweighted recall of 48.37% using leave-one speaker out (26-fold) cross validation on the training dataset. We obtain a 41.57% unweighted recall on the evaluation dataset, which is 3.37% absolute (8.82% relative) over the best baseline results presented in the Emotion Challenge baseline summarized in (Schuller et al., 2009). In the USC IEMOCAP database experiment, we achieve an average unweighted recall of 58.46% using 10-folds (leave-one-speaker out cross validation), which is a 7.44% absolute (14.58% relative) improvement over the Support Vector Machine (SVM) based baseline.

The paper is organized as follows. The two emotional databases used in our study are described in Section 2. The hierarchical classifier framework is presented in Section 3. The experimental results and discussion are provided in Section 4. Conclusions and future work are given in Section 5.

2. Emotional databases

2.1. AIBO database

The AIBO database (Steidl, 2009) consists of 51 children interacting with a Sony toy robot, AIBO, using a Wizard-of-Oz technique. The data collection was designed to

provoke emotional reactions from the children. The robot dog was programmed a priori and did not respond to the children's commands. The children were led to believe that the dog would respond; thereby making the Sony dog seem disobedient and inducing emotional speech. The database was collected at two schools (26 and 25 subjects, respectively). The data from one of the schools is used for training while the other is used for testing. The audio was recorded wirelessly with 16 bits at 48 KHz, which was further downsampled to 16 KHz. The database was segmented into *turns* by splitting the audio with a silence threshold of 1 s. Five advanced linguistics students labeled the emotional content of the database at the word level, and the following is the list of emotion classes that the annotators were asked to rate: joyful, surprised, emphatic, helpless, irritated, angry, motherese, bored, reprimanding, rest, and neutral. The weighted *Kappa* for the five annotators is 0.56, indicating a fair agreement among evaluators. The database description (Steidl, 2009) includes other metrics for computing inter-evaluator agreement that all indicate not perfect, but fair, agreement due to the nature of spontaneous dialogs. The words were combined into longer length chunks, manually defined using syntactic-prosodic criteria (Steidl, 2009). The labels of these chunks were based on majority vote over the merged words. In this study, we were provided with the five emotion classes (a subset of the whole AIBO database): Angry (includes angry, irritated, reprimanding), Emphatic, Neutral, Positive (includes motherese and joyful), and Rest. The detailed descriptions of the AIBO database collection, annotation process, and of the merging of emotion classes can be found in the cited references (Schuller et al., 2009; Steidl, 2009). A summary of the emotion class distribution used in the work is listed in Table 1. The class of *Neutral* represents about 80% of the database. In the testing split of the database, the class distribution is different from the training split as evident in the class of *Neutral* and *Positive*.

2.2. USC IEMOCAP database

The USC IEMOCAP database (Busso et al., 2008) was collected for studying multimodal expressive dyadic interactions. The design of the database assumed that by exploiting the context of dyadic interactions between actors, a more natural and richer emotional display would be elicited than in speech read by a single subject. Furthermore, the use of scripted and emotionally targeted improvisational scenarios allowed us to collect an affectively varied and balanced database. The database was collected using motion capture and audio/video recording (approximately a total of 12 h) over five dyadic sessions with 10 subjects.

Each session consists of a different dyad of male-female actors performing scripted plays and engaging in spontaneous improvised dialogs elicited through affective scenario prompts. At least three Naïve humans annotated each utterance in the database with the categorical emotion labels chosen from the set: happy, sad, neutral, angry, surprised, excited, frustration, disgust, fear and other. In this work, we consider only the utterances with majority agreement (i.e., at least two out of three annotators labeled the same emotion) over the emotion classes of: Angry, Happy, Sad, and Neutral. These classes represent the majority of the emotion categories in this database. This annotation scheme had an inter-evaluator agreement of 0.40 (Fleiss' *Kappa*), which can be considered as fair agreement between evaluators. The detailed description of the USC IEMOCAP database is in the reference (Busso et al., 2008). A summary of the emotion class distribution can be found in Table 2.

2.3. Acoustic feature extraction

Table 3 presents the acoustic features used in this work. We used the same features for the experiments on both databases to provide a common setting in which to evaluate the effectiveness of the proposed classification framework. This acoustic feature set is largely based on the findings by Schuller et al. (2007). We extracted these features using the OpenSmile toolbox (Eyben et al., 2009). The feature set includes 16 low level descriptors consisting of prosodic, spectral envelope, and voice quality features listed in Table 3. These low level descriptors are zero crossing rate, root mean square energy, pitch, harmonics-to-noise ratio, and 12 mel-frequency cepstral coefficients and their deltas. Then 12 statistical functionals were computed for every low level descriptor per utterance in the USC IEMOCAP database and per chunk in the AIBO database: mean, standard deviation, kurtosis, skewness, minimum, maximum, relative position, range, two linear regression coefficients, and their respective mean square error. This results in a collection of 384 acoustic features.

2.4. Feature selection and normalization

We normalized features using *z-normalization* with respect to the neutral utterances in the training dataset for both databases. The process has the underlying assumption that the average characteristics of neutral utterances across speakers do not vary extensively; therefore, the testing examples' features are *z-normalized* with

Table 1
AIBO database: table of emotion utterances.

	Angry	Emphatic	Neutral	Positive	Rest	Total
Train	881	2093	5590	674	721	9959
Test	611	1508	5377	215	546	8257

Table 2
USC IEMOCAP database: number of emotion utterances per category.

Angry	Happy	Sad	Neutral	Total
1083	1630	1083	1683	5480

Table 3
Acoustic features extracted ($16 \times 2 \times 12 = 384$).

Raw acoustic features + deltas	statistical functionals
Pitch (f0)	Mean, standard deviation, kurtosis
Root mean square energy (rms)	Skewness, minimum, maximum
Zero crossing rate (zcr)	Relative position, range
Harmonic to noise ratio (hnr)	Two linear regression coefficients
Mel-frequency cepstral coefficients (1–12 mfcc)	Mean square error of linear regression

respect to the mean, μ, and variance, σ^2, of neutral utterances from the training data. The normalization allows us to use acoustic features across multiple different speakers and to eliminate the effect of variations in individual speakers' speaking characteristics.

We perform feature selection on the 384 features using the standard statistics software SPSS to obtain a reduced feature set. We used binary logistic regression in SPSS with step-wise forward selection. The stopping criterion was based on a conditional likelihood criterion. Forward selection was terminated when the inclusion of an additional feature no longer increased the condition likelihood of the model statistically significantly. This feature selection process resulted in a range of 40–60 features for each binary classifier per cross validation fold. While there exist many other feature selection algorithms, we utilized binary logistic regression because it is standard and is proven effective as a feature selection method (Hosmer and Lemeshow, 2000). This feature selection algorithm was used in each experimental setup in this work to show the effectiveness of the proposed framework for the multi-class classification task. The purpose of this work is not to demonstrate the efficacy of the specific feature selection method, but instead to show how it can be incorporated in the proposed system.

3. Emotion classification framework

3.1. Building the hierarchical decision tree

Our goal is to optimize the unweighted recall percentage (average of per-class accuracies) in the classification framework. This metric is arguably a more useful metric in assessing emotional content in natural interactions when the distribution of classes is non-uniform or dominantly non-emotional. The two essential key points in our design of a emotion classification framework are listed below:

- The use of a combination of binary classifiers instead of a single multi-class classifier.
- The use of a hierarchical tree, where the top level classification is performed on the *easiest* emotion recognition task.

The structure of the framework is shown in Fig. 1. The proposed classification scheme splits the multi-class problem into a series of two-class problems, starting with the relatively *easy* classification task at the top level and leaving the harder tasks for the end. The order of the classification is important and essential in this framework. The goal is to ensure a maximum separation between any two chosen classes at each level. As depicted in Fig. 1, *Classifier 1* operates on the easiest binary classification task and classifiers in the final stage (*Classifier Stage M*) operate on the sets of binary classifications that are most ambiguous for the given acoustic features.

A key aspect in the proposed framework is to investigate the separability of the emotional classes given feature streams. This information impacts the order of the tree. We propose the following two criteria:

- *Prior knowledge:* several previous emotion recognition studies have shown the effectiveness of different feature streams in discriminating between specific emotion classes. For example, we know that acoustic features can accurately discriminate between high-activated emotion classes and low-activated emotion classes (Busso et al., 2009). Therefore, the first level classification task on any two sets of emotion classes that have distinct activation levels cab provide an initial split.
- *Empirical testing:* each emotional dataset may include different definitions and categories of emotion classes. Due to the complex combinatorial nature of finding the most distinguishable pair of emotion classes, we can rely on results obtained from a series of simple empirical studies. For example, classification based on Gaussian Mixture Models, Linear Discriminant Analysis, multi-class Support Vector Machine, and/or any other schemes can be easily trained as a preliminary step. While each classifier may obtain different accuracies, by observing the resulting confusion matrix, the discriminability between each emotion classes can be observed. The hierarchical structure can then be determined.

The approach of designing the classification tree has the potential to propagate fewer classification errors down the tree when compared to directly applying the conventional intuitive approach of classifying *non-emotional* classes vs. *emotional* classes as the first step, and splitting the broad *emotional* classes. Further, we can obtain a balanced recall percentage per emotion class by conveniently optimizing the decision threshold while performing each binary classification task.

Each classifier box shown in Fig. 1 is a binary classifier. At each level, the hard output label of the test sample is fed into the next level of classifiers to perform another set of binary classifications. This sequence of binary classification allows us to take advantage of the variability inherent in the data by creating initial classifications with high recall and identifying classification tasks with a high level of discriminability.

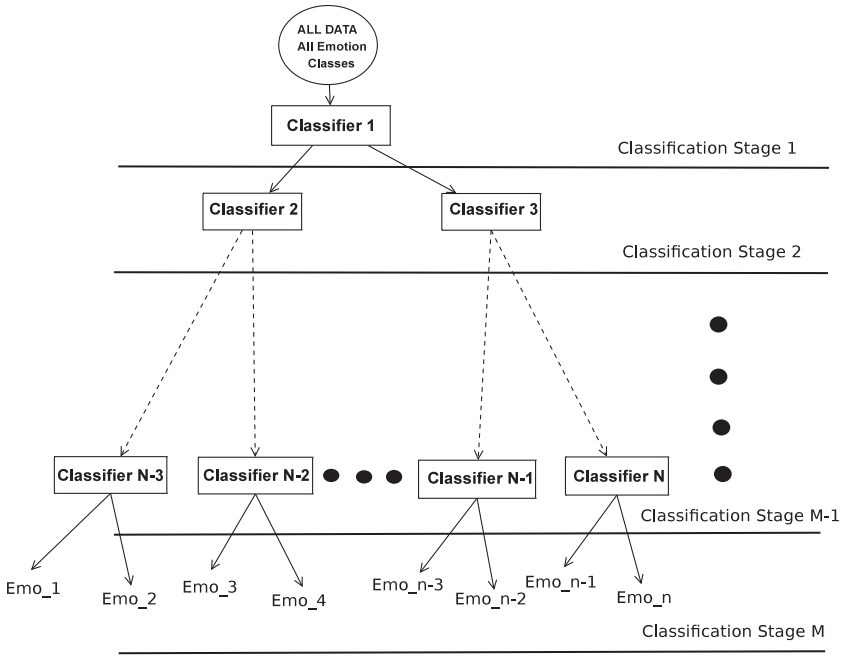

Fig. 1. Proposed classification framework: a hierarchical binary decision tree with the easiest task as the first stage and the most ambiguous task as the last stage.

3.2. Building the hierarchical decision tree for the AIBO database and the IEMOCAP database

Fig. 2 presents the proposed trees for the AIBO (left) and IEMOCAP (right) databases. The realization for each database differs but follows the structure illustrated in Fig. 1. Both frameworks are determined through a combination of criteria mentioned above. For the AIBO database, the classes considered are: Angry, Emphatic, Positive, Neutral, and Rest. We placed A/E vs. P at the first classification stage because multiple iterations of preliminary classification tasks using acoustic features demonstrated that a high-level of discrimination between these two groups of classes. We delay the decision between N and R until the end, and again based on the empirical observation regarding the high level of similarity and ambiguity between N and R classes of this database. We trained a total of six classifiers listed as follows (the classifiers were trained using all the data from the training set with class labels relevant to the task):

- Angry/Emphatic vs. Positive (A&E vs. P)
- Angry vs. Emphatic (A vs. E)
- Angry vs. Neutral/Rest (A vs. N&R)
- Emphatic vs. Neutral/Rest (E vs. N&R)
- Positive vs. Neutral/Rest (P vs. N&R)
- Neutral vs. Rest (N vs. R)

The same design process was applied to the USC IEMOCAP database. The right panel in Fig. 2 shows the decision sequence order. The emotion classes of interest in this task are: Angry, Happy, Sad, and Neutral. We placed A/H vs. S as the first classification step.

In this experiment, we use the same set of acoustic features, hypothesizing that they can accurately discriminate between these two groups of emotion classes. The neutral class is delayed until the last stage due to the difficulties in recognizing the *neutral* class (Metallinou et al., 2010). A total of five binary classifiers for the USC IEMOCAP database were trained and are listed below:

- Angry/Happy vs. Sad (A&H vs. S)
- Angry vs. Happy (A vs. H)
- Angry vs. Neutral (A vs. N)
- Happy vs. Neutral (H vs. N)
- Sad vs. Neutral (S vs. N)

3.3. Classifier for binary classification tasks

Under this hierarchical framework, the specific binary classifier can be determined tailored to the problem domain. Many different binary classifiers have shown promising results in performing classification. For example, both Bayesian Logistic Regression (BLR) (Genkin et al., 2007) and Support Vector Machine (SVM) (Vapnik, 1995) have been shown to be effective in classification tasks. Logistic regression provides a discriminative model to be used as a classifier (Agresti, 1990) and the Bayesian version is a method to prevent data overfitting by placing a prior centered at zero on the weights of the models. SVM is a maximum margin classifier that finds the largest separation between two classes.

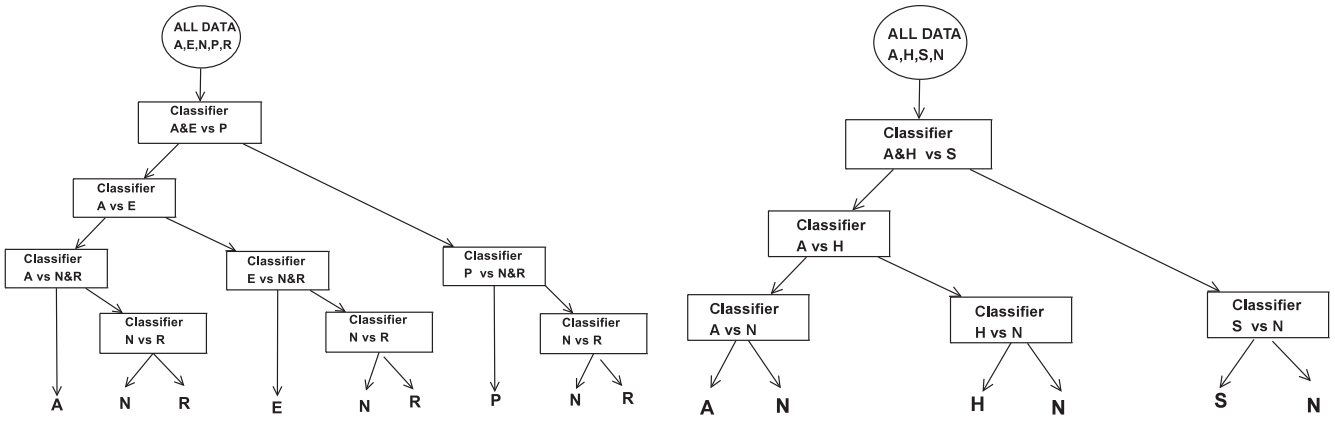

Fig. 2. *Left*: proposed hierarchical structure for the AIBO database. *Right*: hierarchical structure for the USC IEMOCAP database.

In our participation of the 2009 Emotion Challenge (Lee et al., 2009), two different classifier types were used, Bayesian Logistic Regression and Support Vector Machines. Bayesian Logistic Regression obtained the best accuracy though it was not statistically significantly better than Support Vector Machines. As a result, we decided to employ only the Bayesian Logistic Regression as the choice for each of the binary classifier boxes. The feature selection algorithm presented in Section 2.4 is based on logistic regression. It is well suited to Bayesian Logistic Regression, since they share many properties in common. However, the specific choice on the binary classifier can be made along with feature selection to obtain performance optimized for the specific task.

Single class bias is an issue in this emotion recognition task as it may bias the results towards the over-represented class, in this case – *neutral*. Prior work has shown the effectiveness of using Synthetic Minority Oversampling Technique (SMOTE) (Chawla et al., 2002) to deal with the over-representation of a single class. However, in this paper, instead of generating artificial data samples to balance the classes, we exploit our prior knowledge about the class distribution of the training splits in the two databases to adjust the decision threshold on the Bayesian Logistic Regression to obtain a balanced recall across the emotion classes of interest.

3.3.1. Bayesian Logistic Regression

A general binary logistic regression model is a discriminative model of the form shown in Eq. (1).

$$p(y = 1|\beta, x) = \psi(B^T x) \tag{1}$$

where y is the class label $(+1, -1)$, x is the input feature vector, β's are the model parameters, and ψ is the logistic function defined in Eq. (2)

$$\psi(z) = \frac{\exp(z)}{1 + \exp(z)} \tag{2}$$

In Bayesian Logistic Regression (BLR), we place a Gaussian prior with $\mu = 0$ and covariance $\sigma^2 I$ on the model parameters β's shown in Eq. (3) and perform a *maximum a posteriori estimation* of the model parameters to prevent overfitting of the parameters on the training data. This prior on the model parameters has the same effect as the ridge logistic regression where the model parameters' $\|L_2\|$ norms are constrained. Another possible prior is a Laplacian prior, which has the same effect as lasso logistic regression. In this work, Gaussian prior is used since it offered better accuracy than a Laplacian prior in our empirical testing.

$$p(\beta_j|\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{\beta_j^2}{2\sigma^2}\right) \tag{3}$$

The BBR software (Genkin et al., 2007) was used for Bayesian Logistic Regression model training and threshold tuning.

4. Experiment setup and results

The effectiveness of the proposed hierarchical classification method was evaluated on the two different databases introduced in Section 2: the AIBO and the USC IEMOCAP databases. The first set of experiments utilizes the AIBO database and follows the guidelines used in the 2009 Interspeech Emotion Challenge (Lee et al., 2009). A training dataset with emotion labels was used to develop our algorithm. The labels for the testing database were unknown (the performance metrics were given through a website interface). To show that the algorithm can be easily applied in another database, we applied the proposed classification framework to the USC IEMOCAP database.

4.1. AIBO database

Two predefined subsets of the AIBO database were available for this task, a training dataset and an unlabeled evaluation dataset. Two different experiments were designed based on this structure. In Experiment I, we analyzed our hierarchical structures using only the training data subset. We used leave one speaker out (26-fold) cross-validation was used to estimate the classification

performance (average unweighted recall). This cross-validation method was used to simulate the scenario in which the unlabeled evaluation dataset consists of a disjoint speaker set. Experiment I serves as the development phase. In Experiment II, the framework was trained on the entirety of the training dataset and tested on the evaluation dataset.

- *Experiment I* : Leave one speaker out (26-fold) cross-validation on the training dataset (AIBO database)
- *Experiment II* : Evaluate performance on the unlabeled evaluation dataset (AIBO database)

4.1.1. Results of Experiment I on the AIBO database

The unweighted recall for Bayesian Logistic Regression was 48.27% (Table 4). The columns of the confusion matrix found in Table 4 represent our hypothesized class labels and the rows are the annotated ground truth class labels. The *conventional* framework presented was based on classifying non-emotional vs. emotional classes as the first step; Fig. 3 shows the conventional structure for the AIBO database experiment I.

Several observations can be made from examining the results. While the conventional hierarchical structure obtains approximately the same weighted accuracy as the proposed framework, the proposed method outperforms the conventional method in unweighted accuracy. The confusion matrices show that the largest improvement is in the emotion class of *Positive*, which is confused mostly with *Neutral* and *Rest*. The effect of tackling the classification problem involving the *Positive* emotion as the first step is essential as evidenced by the increase in the recognition accuracy of this class. The recall for A/E vs. P for the proposed method at the first step is at 94.82%. The first stage binary classifier is able to separate these two emotion groups with high accuracy. Therefore, we are able to retain the majority of the members of the two groups of emotion classes by placing this classification task as the first step in the proposed structure as compared to the conventional method.

Table 4
Experiment I: summary of result.

Unweighted recall (UA)					Weighted recall (WA)
Bayesian Logistic Regression (BLR): proposed					
48.27%					**48.82%**
	Angry	Emphatic	Neutral	Positive	Rest
Angry	**504**	145	126	53	53
Emphatic	395	**1078**	412	101	107
Neutral	506	1020	**2703**	776	585
Positive	21	31	121	**439**	62
Rest	97	130	185	171	**138**
Bayesian Logistic Regression (BLR): conventional					
38.42%					**48.66%**
	Angry	Emphatic	Neutral	Positive	Rest
Angry	**420**	129	193	30	109
Emphatic	342	**916**	596	58	181
Neutral	395	923	**3207**	317	748
Positive	18	30	248	**132**	246
Rest	106	133	217	94	**171**

We are classifying the emotion class, *Rest*, at about the chance level. This is expected because this class is not as strictly defined as the emotions in the other classes. *Rest* is misclassified more often as either *Neutral* or *Positive* compared with *Angry* or *Emphatic* (Table 4). This indicates that the *Rest* is acoustically similar to the *Positive* and the *Neutral* in this database.

We obtain a good recall for four of the emotion classes (Angry: 57.2%, Emphatic: 51.5%, Positive: 65.1%, Neutral: 48.4%) excepting the class, *Rest* (19.1%). This indicates that the structure of our framework is able to handle the highly skewed database and is able to obtain a more balanced retrieval rate on the emotion classes. This is essential in emotion recognition since in natural human interaction, *Neutral* is often be the majority of expressed emotions. The balancing of the recognition accuracy using the proposed structure is advantageous because it is able to

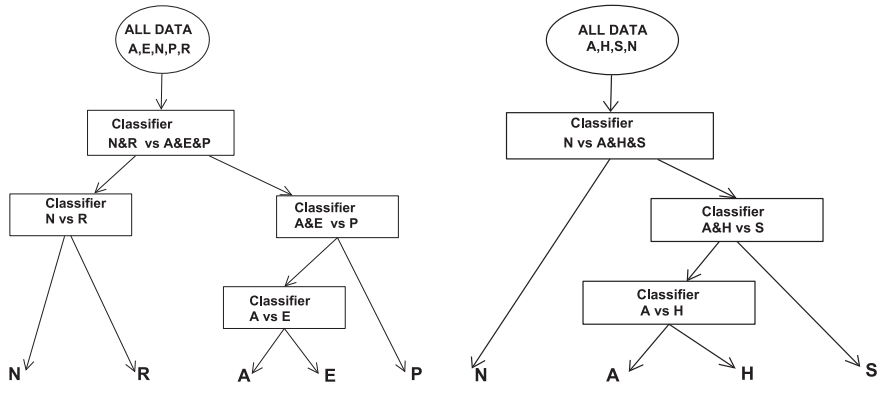

Fig. 3. *Left*: conventional hierarchical structure for the AIBO database. *Right:* conventional hierarchical structure for the USC IEMOCAP database.

identify several other less frequently expressed but informative emotion classes.

4.1.2. Results of Experiment II on the AIBO database

In Experiment II, we evaluated our framework on the evaluation dataset, which was the actual task for the 2009 Emotion Challenge. The six classifiers were trained on the entirety of the training dataset. The unweighted recall using Bayesian Logistic Regression was 41.57%. The summary of the results is shown in Table 5. An HMM baseline is also presented (Schuller et al., 2009) because HMMs are generally effective for mitigating the effect of class bias. SVM based baseline is used as our baseline since it obtains the highest accuracy.

Our proposed framework using Bayesian Logistic Regression achieved the highest average unweighted recall. It improves the accuracy measure of the baseline model (multi-class SVM with a SMOTE class balancing technique) presented (Schuller et al., 2009) by 3.37% absolute (8.82% relative). The average unweighted recall rate on the three *emotional* classes (*Angry, Emphatic*, and *Positive*) is about 52% where the average unweighted recall rate on *non-emotional* (*Neutral* and *Rest*) classes is only about 25%. This result demonstrates that our proposed framework is capable of retrieving the emotional utterances even given that some of these emotion classes are only a small portion of the database. The characteristics of the proposed framework is advantageous in real world applications where the majority of expressions are often *Neutral*. Furthermore, we also notice a large discrepancy between Experiments I and II for this database. We speculate that except for the fact that the two datasets were recorded at different places with different subjects, the primary reason for this discrepancy is that the evaluation dataset may be more unbalanced (the class of *Neutral* concentrates almost 65% of the dataset). It would be interesting to investigate whether an improved classification accuracy could be obtained by incorporating the knowledge of emotion class distribution on this portion of the database.

In summary, our proposed framework for the five-class emotion recognition as a sequence of binary classification tasks is able to improve the unweighted recall by 3.37% absolute (8.82% relative) compared with using Support Vector Machine with SMOTE baseline on the unlabeled evaluation dataset, which is the baseline provided by the 2009 Emotion Challenge. Since the AIBO database contains realistic and spontaneous interactions, it is encouraging to see that the framework has the potential to overcome the class imbalance problem in the database and to achieve a good recall percentage especially on the *emotional* classes.

Table 5
Experiment II: summary of result.

	Unweighted recall (UA)	Weighted recall (WA)
Bayesian Logistic Regression (BLR)		
SVM baseline	38.2%	39.2%
HMM baseline	35.9%	37.2%
BLR	**41.57%**	**39.87%**

	Angry	Emphatic	Neutral	Positive	Rest
Angry	**290**	171	65	63	22
Emphatic	210	**752**	325	136	85
Neutral	748	1094	**2057**	1109	369
Positive	23	13	39	**131**	9
Rest	95	58	134	197	**62**

4.2. USC IEMOCAP database

In order to show such framework can be easily applied in another emotional database, the USC IEMOCAP database was used with leave-one-speaker out cross validation evaluation scheme. The leave-one-speaker out cross validation setup was used to emulate the AIBO evaluation condition in which the testing data consists speaker set disjoint from that of the training set (the USC IEMOCAP database does not specify the two splits to be used for training and testing. Therefore, we utilize a different experimental setup for the USC IEMOCAP database). We used this evaluation scheme to make this experiment comparable to the AIBO database experimental setup. In each fold, we used nine speakers as the training dataset and one speaker as the testing dataset.

4.2.1. Experiment result of the USC IEMOCAP database

A summary of the classification accuracy is shown in Table 6. The average unweighted recall is 58.46%, which is a 7.77% absolute improvement (15.16% relative) compared to a recently published result (Metallinou et al., 2010) on the same four emotion classes using Hidden Markov Models trained with acoustic features. In the current work, a mutliclass SVM was presented as additional baseline classification. Our proposed method obtains a 7.44% absolute (14.58%) improvement over the SVM baseline. The conventional hierarchical structure classifies *Neutral* vs. others as the first step (Fig. 3).

Several observations can be made by looking at the results. Through examination of the confusion matrices of both the conventional structure and proposed structure, we observe the same trend as seen in the AIBO database. The recognition accuracy of *Sad* and *Happy* increased significantly. These two classes are mostly confused with the class of *Neutral*. This confusion is alleviated by making this assessment at the first stage leaving the assignment of *Neutral* to a later step. This structure improves the recognition accuracy (Table 6).

There is very little confusion noticed between the combined angry/happy and sad class; the recall percentage at the first stage classification (A/H vs. S) is 85% and 87.5%, respectively. Most of the angry/happy vs. sad emotional utterances were successfully recalled and split at the first binary classification stage. The recognition accuracy of the emotion class, *Happy*, is the lower (41.7%) compared to the emotion classes of *Angry* (65.4%) and *sad* (47.64%). This is in accordance with the trend found in previous work

Table 6
Experiment: summary of the USC IEMOCAP database classification result.

	Unweighted recall (UA)	Weighted recall (WA)
HMM baseline	50.69%	N/A
SVM baseline	51.02%	42.41%
BLR conventional	**53.55%**	**53.47%**
BLR proposed	**58.46%**	**56.38%**

Bayesian Logistic Regression (BLR): proposed

	Angry	Happy	Sad	Neutral
Angry	**720**	168	30	183
Happy	319	**680**	205	426
Sad	24	42	**782**	235
Neutral	116	256	394	**918**

Bayesian Logistic Regression (BLR): conventional

	Angry	Happy	Sad	Neutral
Angry	**683**	155	32	231
Happy	290	**585**	183	572
Sad	36	42	**516**	489
Neutral	103	190	245	**1146**

on this database (Metallinou et al., 2010). This likely results from the reliance on only acoustic features (the emotional evaluation included audiovisual stimuli). Previous work has demonstrated that happiness can be more accurately modeled by incorporating facial expression features (Metallinou et al., 2010). One of the most noticeable results in this experiment is that the recall percentage for the neutral class is 54.54%. This is an encouraging outcome considering the highly ambiguous nature of the neutral class, which is evident in previous results (35.23%) on the same database (Metallinou et al., 2010). While the conventional hierarchical approach obtains a higher recognition accuracy on the emotion class of *neutral*, the proposed framework improves the recognition rates on the other three emotional classes without losing much of the recognition rate on the *neutral*. This is likely because the neutral assignment is made at the last step. We have multiple binary classifiers to separate the different emotional classes from neutral instead of having one multi-class classifier to identify neutral class. This approach takes into account the fuzziness in the definition of the *neutral* emotion class; it can be an emotion class itself or can be used a way to describe a user state that is not emotional. Overall, the result improved 7.77% absolute compared to the recently published results on the same set of emotion classes on the same database (Metallinou et al., 2010). The experimental conditions differed between these two works with respect to the features used. While not directly comparable, it is still encouraging to see that without exhaustive tuning and optimization, the proposed framework can provide significant improvement in the overall emotion recognition accuracy.

5. Conclusions

Accurate emotion recognition systems are essential for the advancement of human behavioral informatics and in the design of effective human–machine interaction systems. Such systems can help promote the efficient and robust processing of human behavioral data as well as in the facilitation of natural communication. In this work, a multilevel binary decision tree structure was proposed to perform multi-class emotion classification. The framework was designed by empirical guidance and experimentation. The easiest subset of classification problems were placed at the top level to reduce the accumulation of error. This classification framework was introduced first in the Interspeech 2009 Emotion Challenge (where it placed first on the classifier sub-challenge task) and has been since tested on another emotional database and reported in this paper. The results show encouraging recognition rates that are competitive with the state of the art.

Many future modifications can be integrated within this framework. Instead of outputting hard labels at every level, a soft label, such as a measure of probability or even profile based representation (Mower et al., 2011), can be used to enhance the modeling power of the proposed framework. Also, since the choice of binary classifier is flexible and largely dependent on the feature selection technique, the framework can be further improved by optimizing the choice of binary classifier along with the appropriate feature selection method at each classification stage. The major limitation of the approach described here is the empirical nature of the proposed hierarchical structure. While the proposed method has the advantage of being intuitive and efficient to design, it does not ensure an optimal solution. Our future work plans to investigate an automatic procedure to generate the hierarchical structure. This can minimize the need for several iterations of empirical testing. A specific related question for future work surrounds the derivation of a hierarchical structure that will not only optimally balance performance accuracy and combinatorial complexity but also yield results that are intuitively interpretable in light of psychological theories of emotions.

References

Agresti, A., 1990. Categorical Data Analysis. In: Wiley Series in Probability and Mathematical Statistics. Wiley, New York.

Albornoz, E.M., Milone, D.H., Rufiner, H.L., 2011. Spoken emotion recognition using hierarchical classifiers. Comput. Speech Lang. 25 (3), in press.

Amir, N., Mixdorff, H., Ofer Amir, D.R., Diamond, G.M., Pfitzinger, H.R., Levi-Isserlish, T., Abramson, S., 2010. Unresolved anger: Prosodic analysis and classification of speech from a therapeutic setting. In: Speech Prosody.

Black, M., Chang, J., Narayanan, S., 2008. An empirical analysis of user uncertainty in problem-solving child-machine interactions: Linguistic analysis of spontaneous children speech. In: Proc. Workshop on Child, Computer and Interaction, Chania, Greece.

Black, M., Katsamanis, N., Lee, C.-C., Lammert, A.C., Baucom, B.R., Christensen, A., Georgiou, P.G., Narayanan, S., 2010. Automatic classification of married couples' behavior using audio features. In: Proc. Interspeech, Makuhari, Japan, 2010.

Brave, S., Nass, C., Hutchinson, K., 2005. Computers that care: Investigating the effects of orientation of emotion exhibited by an embodied computer agent. Internat. J. Human–Comput. Stud. 62 (2), 161–178.

Brody, L., Hall, J., 2008. Gender and Emotion in Context. The Guilford Press.

Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J., Lee, S., Narayanan, S., 2008. IEMOCAP: Interactive emotional dyadic motion capture database. J. Language Resour. Eval. 42, 335–359.

Busso, C., Lee, S., Narayanan, S., 2009. Analysis of emotionally salient aspects of fundamental frequency for emotion detection. IEEE Trans. Audio Speech Language Process. 17 (4), 582–596.

Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002. Smote: Synthetic minority over-sampling technique. J. Artif. Intell. Res. 16, 321–357.

Eyben, F., Woellmer, M., Schuller, B., 2009. Speech and music interpretation by large-space extraction, Tech. rep., Institute for Human–Machine Communication, Technische Universitaet Muenchen. <http://www.sourceforge.net/projects/openSMILE>.

Genkin, A., Lewis, D.D., Madigan, D., 2007. Large-scale Bayesian logistic regression for text categorization. Technometrics 49 (3), 291–304.

Hassan, A., Damper, R.I., 2010. Multi-class and hierarchical svms for emotion recognition. In: Proc. Interspeech, pp. 2354–2357.

Herm, O., Schmitt, A., Liscombe, J., 2008. When calls go wrong: How to detect problematic calls based on log-files and emotions. In: Proc. Interspeech.

Hosmer, D., Lemeshow, S., 2000. Applied Logistic Regression, second ed., Wiley Series in Probability and Statistics.

Kanda, T., Hirano, T., Eaton, D., Ishiguro, H., 2004. Interactive robots as social partners and peer tutors for children: A field trial. Human–Comput. Interact. 19 (1), 61–84.

Kapoor, A., Picard, R., 2005. Multimodal affect recognition in learning environments. Proc. 13th Annual ACM Internat. Conf. on Multimedia. ACM, New York, NY, USA, pp. 677–682.

Lazarus, R., 2001. Relational meaning and discrete emotions. Appraisal Process. Emotion: Theor., Methods, Res., 37–67.

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

Lee, C.-C., Mower, E., Busso, C., Lee, S., Narayanan, S., 2009. Emotion recognition using a hierarchical binary decision tree approach. In: Proc. Interspeech, 2009.

Lee, C.-C., Black, M., Katsamanis, A., Lammert, A., Baucom, B., Christensen, A., Georgiou, P.G., Narayanan, S., 2010. Quantification of prosodic entrainment in affective spontaneous spoken interactions of married couples. In: Proc. Interspeech, Makuhari, Japan, 2010.

Mao, Q.-R., Zhan, Y.-Z., 2010. A novel hierarchical speech emotion recognition method based on improved ddagsvm. Comput. Sci. Inform. Systems 7 (1), 211–222.

Metallinou, A., Lee, S., Narayanan, S., 2010. Decision level combination of multiple modalities for recognition and analysis of emotion expression. In: Proc. Internat. Conf. on Acoustics, Speech, and Signal Processing.

Metallinou, A., Busso, C., Lee, S., Narayanan, S., 2010. Visual emotion recognition using compact facial representation and viseme information. In: Proc. Internat. Conf. on Acoustics, Speech, and Signal Processing (ICASSP).

Mower, E., Mataric, M.J., Narayanan, S.S., 2011. A framework for automatic human emotion classification using emotional profiles. IEEE Trans. Audio Speech Language Process. 19:5, 1057–1070.

Pantic, M., Sebe, N., Cohn, J., Huang, T., 2005. Affective multimodal human–computer interaction. In: Proc. 13th Annual ACM Internat. Conf. on Multimedia, ACM New York, NY, USA, 2005, pp. 669–676.

Prendinger, H., Mori, J., Ishizuka, M., 2005. Using human physiology to evaluate subtle expressivity of a virtual quizmaster in a mathematical game. Internat. J. Human–Comput. Stud. 62 (2), 231–245.

Schuller, B., Batliner, A., Seppi, D., Steidl, S., Vogt, T., Wagner, J., Devillers, L., Vidrascu, L., Amir, N., Kessous, L., Aharonson, V., 2007. The relevance of feature type for the automatic classification of emotional user states: Low level descriptors and functionals. In: Proc. Interspeech.

Schuller, B., Steidl, S., Batliner, A., 2009. The interspeech 2009 emotion challenge. In: Proc. Interspeech, Brighton, UK .

Steidl, S., 2009. Automatic classification of emotion-related user states in spontaneous children's speech, Logos, Verlag, Berlin.

Vapnik, V.N., 1995. The Nature of Statistical Learning Theory. Springer, New York.

Wanger, H.L., 1993. Measuring performance in category judgment studies on nonverbal behavior. J. Nonverbal Behav. 17 (1), 3–28.

Xiao, Z., Dellandrea, E., Dou, W., Chen, L., 2007. Automatic hierarchical classification of emotional speech. In: Proc. ISMW, pp. 291–296.

Yildirim, S., Lee, C.M., Lee, S., Potamianos, A., Narayanan, S., 2005. Detecting politeness and frustration state of a child in a detecting politeness and frustration state of a child in a conversational computer game. In: Proc. Eurospeech, Lisbon, Portugal.

Yildirim, S., Narayanan, S., Potamianos, A., 2011. Detecting emotional state of a child in a conversational computer game. Comput. Speech Language 25 (1), 29–44 (Special Issue on Affective Speech).