Published in final edited form as:

Comput Speech Lang. 2015 January ; 28(1): 186–202. doi:10.1016/j.csl.2014.01.003.

Speaker-sensitive emotion recognition via ranking: Studies on acted and spontaneous speech[☆]

Houwei Cao^{a,*}, Ragini Verma^a, and Ani Nenkova^b

Houwei Cao: Houwei.Cao@uphs.upenn.edu

^aDepartment of Radiology, Section of Biomedical Image Analysis, University of Pennsylvania, 3600 Market Street, Suite 380, Philadelphia, PA 19104, United States

^bDepartment of Computer and Information Science, University of Pennsylvania, 3330 Walnut Street, Philadelphia, PA 19104, United States

Abstract

We introduce a ranking approach for emotion recognition which naturally incorporates information about the general expressivity of speakers. We demonstrate that our approach leads to substantial gains in accuracy compared to conventional approaches. We train ranking SVMs for individual emotions, treating the data from each speaker as a separate query, and combine the predictions from all rankers to perform multi-class prediction. The ranking method provides two natural benefits. It captures speaker specific information even in speaker-independent training/testing conditions. It also incorporates the intuition that each utterance can express a mix of possible emotion and that considering the degree to which each emotion is expressed can be productively exploited to identify the dominant emotion. We compare the performance of the rankers and their combination to standard SVM classification approaches on two publicly available datasets of acted emotional speech, Berlin and LDC, as well as on spontaneous emotional data from the FAU Aibo dataset. On acted data, ranking approaches exhibit significantly better performance compared to SVM classification both in distinguishing a specific emotion from all others and in multi-class prediction. On the spontaneous data, which contains mostly neutral utterances with a relatively small portion of less intense emotional utterances, ranking-based classifiers again achieve much higher precision in identifying emotional utterances than conventional SVM classifiers. In addition, we discuss the complementarity of conventional SVM and ranking-based classifiers. On all three datasets we find dramatically higher accuracy for the test items on whose prediction the two methods agree compared to the accuracy of individual methods. Furthermore on the spontaneous data the ranking and standard classification are complementary and we obtain marked improvement when we combine the two classifiers by late-stage fusion.

Keywords

Emotion classification; Ranking models; Spontaneous speech; Acted speech; Speaker-sensitive

1. Introduction

Research on emotion recognition from cues expressed in human voice has a long-standing tradition (Cowie et al., 2000; Ververidis and Kotropoulos, 2006). The urgency for developing accurate methods for emotion recognition has become even greater with the wide-spread use of interactive voice systems in call centers (Petrushin, 1999; Lee et al., 2002; Yu et al., 2004), car navigation systems (Fernandez and Picard, 2003), education (Litman and Forbes-Riley, 2004) and human-robot interaction (Steidl, 2009). There is also increasing interest in incorporating emotion in search over audio content, which has motivated work on emotion prediction in talk shows (Grimm et al., 2007b) and movies (Giannakopoulos et al., 2009).

The traditional paradigm of emotion recognition in speech is to extract acoustic features from the speech signal, then train classifiers on these representations, which when applied to a new utterance are able to determine its emotion content. A variety of pattern recognition methods have been explored for automatic emotion recognition such as gaussian mixture models (Luengo et al., 2005; Vlasenko et al., 2007; Vondra and Vich, 2009), hidden Markov models (Nwe et al., 2003; Shafran et al., 2003; Meng et al., 2007), neural network (Nicholson et al., 2000) and support vector machines (Kwon et al., 2003; Tabatabaei et al., 2007; Bitouk et al., 2010; Lee et al., 2011), regression (Grimm et al., 2007b). All of these seemingly diverse methods are designed to predict the emotion of a single test utterance in isolation. In many practical applications, however, emotion analysis is to be performed on a recording of complete conversations. In recordings of meetings, a user who was not present at a meeting may want to view only parts of the discussion in which participants expressed emotions. Similarly in telephone and broadcast conversations or political debates and speeches a user may want to identify parts where emotion was expressed. In all these scenarios multiple utterances from the same speaker are available and an emotion detection system could make use of this information. In such case, the state-of-the-art classification methods produce a classification score for each test utterance that does not take into account any potentially beneficial information from other utterances present in the test set. Deciding for example if an utterance expresses anger may be easier if the decision is made with respect to a larger set of utterances conveying a variety of emotions.

Ranking approaches to the task of emotion recognition offer a way of sorting all utterances in a given sample of speech from the same speaker with respect to the degree with which they convey a particular emotion. The benefit from modeling the extent to which all possible emotions are expressed in the utterance has been documented on work on emotion profiles (Mower et al., 2011), where each utterance is characterized by the distance from the hyperplane for several binary emotion classifiers. The benefit from incorporating speaker information has been confirmed by a number of studies which show that emotion recognition is higher when the same speaker is present in both training and testing. No prior work however has shown how to exploit user-specific information when prediction is done on speakers not previously seen in training. Ranking approaches seamlessly incorporate both of these desirable properties.

In this paper we show that ranking approaches lead to considerable and consistent improvement of prediction accuracy compared to conventional classification on both acted and natural spontaneous emotional speech. We use ranking SVM for our analysis (Joachims, 2002) and rely on a standard large set of acoustic features which we describe in Section 4. First we carry out experiments on two publicly available datasets of acted emotional speech which we introduce in Section 2, and report results from speaker independent analysis using leave-one-subject-out evaluation paradigm in Section 5. In Section 6 we further evaluated the proposed ranking models on the more challenging task of spontaneous emotion recognition. We also discuss the complementarity of conventional SVM classifiers and ranking-based models and further investigate the combination of the two classifiers in Section 7.

2. Corpora of emotional speech

In this study we experiment with ranking approaches on three publicly available datasets: the Berlin emotional speech database of German emotional speech (Burkhardt et al., 2005), the LDC emotional speech database of English emotional utterances (Linguistic Data Consortium, 2002), and the FAU Aibo emotional database (Steidl, 2009). The Berlin and LDC datasets consist of acted utterances, rendered by actors to convey some target emotions. The FAU Aibo database contains realistic spontaneous emotional speech from recording of children interacting with an Aibo robot.

2.1. Berlin emotional speech database

The Berlin dataset contains Recordings of 10 native German actors (5 female/5 male), expressing in German each of the following seven emotions: *anger*, *disgust*, *fear*, *happy*, *neutral*, *sadness*, *boredom*. Each actor was asked to speak one of the 10 pre-selected sentences which were chosen to maximize the number of vowels. In our experiments, we used 454 emotional utterances corresponding to the six basic emotions (Cowie et al., 2000), which are represented in both the Berlin and LDC datasets *anger*, *disgust*, *fear*, *happy*, *sadness*, *neutral*, and excluded utterances corresponding to boredom.

Emotional labels of utterances in the Berlin dataset were validated by external subject assessment. Each utterance was rated by 20 human subjects with respect to their perceived naturalness and emotional content. Utterances for which the target emotion was not easily recognized, as well as utterances with low perceived naturalness, were removed from the dataset. Given the acquisition protocol for the dataset, samples in it can be considered unambiguous, prototypical expressions of the target emotions.

2.2. LDC emotional speech database

The LDC dataset contains emotional utterances of seven native English actors (3 female/4 male). It comprises the following 15 emotional states: *hot anger*, *cold anger*, *disgust*, *fear*, *happy*, *sadness*, *neutral*, *panic*, *despair*, *elation*, *interest*, *shame*, *boredom*, *pride*, *contempt*. Almost all of the recorded utterances are 4-syllable sentences which contain dates and numbers. In our experiments, we only focused on the six basic emotions (*hot anger*, *disgust*, *fear*, *happy*, *sadness*, *neutral*) and used 470 corresponding emotional utterances.

In comparison to the Berlin dataset, utterances in the LDC dataset were not validated by raters after the recording. Utterances in LDC dataset were simply labelled as the intended emotion given to the speakers during recording. Labeling intended emotion makes the corpus more similar to spontaneous utterances, in which emotions can be mixed and for which recognition by a listener may not be so good.

2.3. FAU Aibo emotional speech database

The FAU Aibo emotion corpus contains natural elicited emotional speech. It consists of recordings of 51 German children (30 female/21 male), interacting with the AIBO robot controlled by a human invisible to them. The speech data is annotated for the following 11 emotional states: *anger, bored, emphatic, helpless, joyful, motherese, neutral, reprimanding, rest, surprised, touchy*. The whole corpus contains 8.9 h of speech recording in total and was automatically segmented into turns using a long pause (>1 s). The corpus was originally labelled at word-level. Each word was annotated as one of the above eleven states by five listeners via majority voting. After that the turn-level labels were derived from word-level labels with confidence scores (Steidl, 2009).

We focused on the five-class (*Angry, Emphatic, Neutral, Positive, Rest*) classification problem as described in the Interspeech 2009 Challenge (Planet et al., 2009). The class **Rest** contains emotional utterances that do not belong to any of the other classes. There are 18,216 emotional turns corresponding to these five big classes onto which the original 11 emotional states were mapped. They were divided into training and testing sets as shown in Table 1.

3. Ranking SVM for emotion recognition

In this study, we make use of the SVM^{rank} toolkit to train and test our approach (Joachims, 2006).

Ranking support vector machines (SVM) are a typical pairwise method for designing ranking models. The basic idea behind them is to formalize learning to rank as a problem of binary classification on pairs that define a partial ordering and then to solve the problem using SVM classification (Joachims, 2002).

In emotion recognition specifically, instances are the feature representation of utterances. The ranking problem is to sort the utterances with respect to how much they convey a particular emotion. To train a ranker for a target emotion, we need to specify a set of pairs of instances for which one instance conveys the target emotion better than the other; the binary classification problem that the ranker will optimize is to minimize the number of pairs for which it predicts the order of the instances incorrectly. There are several alternatives for defining the partial ordering for ranking. In our initial experiments, we choose to form pairs only from utterances from the same speaker and consider all utterances that convey the target emotion to have higher scores than utterances that convey any other emotion. Stated more formally, consider training data consisting of samples from speakers $s_1, s_2, \dots, s_k$ that convey emotions $e_1, e_2, \dots, e_r$. We denote by $U_{m,i,q}$ the q th utterance expressing emotion m

and spoken by speaker i . The ranker for emotion m is learned from pairs $U_{m,i}^* > U_{n,i}^*$ such that m and n are different.¹

In testing, all utterances from a speaker whose data was not used in training is given to the ranker for a target emotion. The ranker produces a ranking score for each test utterance, allowing us to sort the utterances by decreasing score. Utterances with higher rank are considered to express the target emotion more clearly than utterances with lower rank. The output from an individual ranker is analogous to a one-versus-all binary classifier that attempts to distinguish the target emotion from all others.

The motivation for our approach is the same as that for using ranking SVMs for ranking in information retrieval. There the task is to sort webpages returned by a search engine by relevance to the query. In our task, a query is defined by each speaker in the dataset. When training a ranker for a target emotion, utterances by the same speaker that convey this emotion are more relevant than any other utterance. In testing, the ranker output gives a way of sorting all utterances by the user in terms of their relevance to the target emotion.

Fig. 1 depicts the overall training and testing set-up we adopted to build and evaluate six rankers, one for each of the basic emotions. Each line in the boxes representing data corresponds to utterances $U_{m,i,q}$ as we defined them above.

In many cases the ultimate goal is to perform multi-class emotion classification and determine what emotion is expressed by a given utterance. To perform the six-way classification problem, we need to combine the output of individual rankers into a single prediction about which is the most likely expressed emotion. However, such decisions cannot be made directly on the basis of the prediction scores given by the rankers because these scores can only be used for ordering. They do not have a meaning in an absolute sense and scores predicted from different rankers cannot be compared directly in a meaningful way.

For each test utterance U we define a normalized ranking score for each of the emotions we want to analyze:

$$1 - \frac{\text{rank}_m(u)}{N} \quad (1)$$

The score combines information about the number of test utterances from the same speaker in the test set N and the rank of that utterance given by the ranker for emotion m , $\text{rank}_m(u)$. These scores fall in the range $[0, 1)$. An utterance will have score 0 if it was the last in the sorted list defined by the SVM ranking (least likely to express the target emotion). The score will be very close to 1 for utterances at the top of the list (which resemble the most the target emotion among all utterances spoken by the speaker).

¹Specifically, all utterances of the target emotion are given score 1 and all utterances conveying other emotions a score of 0.

There are several ways in which these six scores, one from each basic emotion ranker, can be used to decide which single emotion is most likely conveyed by the utterance. For example we can predict that the emotion conveyed by the utterance is the one for which $rank_m(u)$ is the smallest. Alternatively, the scores can be viewed as a six dimensional feature set with which we train a conventional SVM classifier to decide the most likely emotion. We evaluate both alternatives in Section 5.²

4. Features

While a number of previous studies on emotion recognition focused on feature engineering and analysis, the primary goal of this paper is to introduce the novel learning methodology of ranking SVM to emotion recognition tasks. Because of this, we make use of a comprehensive set of standard acoustic features.

The common acoustic features for speech-based analysis of emotion include prosodic features and spectral features. Prosodic features are global features and they are usually calculated for the entire utterance (Dellaert et al., 1996). Typically used global prosodic features include mean pitch values for an utterance, pitch variability in given parts of the utterance, intensity and changes in intensity, voice quality percepts such as raspiness and tenseness measured using the relative spread of energy within specific energy bands, speech rate and pause duration (Fernandez and Picard, 2003; Huang and Ma, 2006; Busso et al., 2011). Spectral feature on the other hand characterize short segments of the utterance and are typically employed in automatic speech recognition. The most commonly used spectral features include Mel-frequency cepstral (MFCC) coefficients (Kwon et al., 2003), wavelet features (Neiberg et al., 2006), Linear Prediction Code (LPC) coefficients (Nicholson et al., 2000), and Perceptual Linear Prediction (PLP) coefficients (Ye et al., 2008).

Most of these features can be extracted by the openSMILE feature extraction library (Eyben et al., 2010). We use the default setting and obtain a feature vector consisting of 988 spectral and prosodic features to characterize each utterance. We use these features for all experiments reported in later section.³

5. Emotion recognition experiments on acted speech

In this section we evaluate the effectiveness of the ranking SVM approach to emotion recognition and contrast their performance with conventional SVM classification on two publicly available datasets of acted speech, Berlin and LDC. In order to confirm the stability and speaker independence of the obtained classifiers, we performed all experiments using leave-one-subject-out (LOSO) paradigm. In this form of cross-validation, all samples from a given speaker are used as a test set for a model trained on the data from all other speakers and the process is repeated for all speakers. The overall performance of the classifier is

²The six dimensional vector can also be viewed as a characterization of the utterance as expressing mixed emotions, similar to the emotion profiles approach (Mower et al., 2011). We do not examine this interpretation further in this paper but it may prove useful in later work.

³Recent work has shown that feature selection approaches aimed to further narrow down this set does not lead to increased performance of emotion recognition classifiers (Planet et al., 2009).

computed by combining the predictions from all test folds and computing the overall accuracy for the entire dataset.

First, we analyze the performance of individual emotion rankers in Section 5.1. Six speaker-independent emotional rankers were constructed, one for each basic emotion. The individual rankers correspond to the “one versus all” task in the traditional literature, in which models are trained to distinguish a particular emotion from all other emotions in the dataset (Lee et al., 2004; McGilloway et al., 2000). We also explore, in Section 5.2, several different ways for defining a partial ordering on pairs for the learning stage of the ranker, inspired by dimensional theories of emotion. These however turn out to be inferior to the formulation already described.

Next, we examine the accuracy of different approaches that incorporate the results from individual rankers to perform multi-class prediction for each utterance. We report these results, along with the performance of conventional SVM classifiers, in Section 5.3.

5.1. Ranking SVM vs. binary classification

We now turn to examine the performance of rankers for each of the six basic emotions. Each of the rankers is optimized to predict the most intuitive ordering on pairs of utterances from the same speaker, which is that an utterance conveying the target emotion is scored higher than any utterance from the same speaker that conveys another emotion. For example the constraints for training an **Anger** ranker are that utterances expressing *anger* score higher than any other utterance by the same speaker.

In testing, all the utterances from the held out test speaker will be ranked. For a good ranker, examples of the target emotion should be ranked higher than examples of any of the other emotions. For each of the six rankers for the basic emotions, we report their pairwise test accuracy and the precision at k .

Pairwise accuracy is the number of pairs of utterances in the test set that are accurately ordered according to the scores predicted by the ranker. For example in a pair of instances that consists of a *anger* utterance and a *happy* utterance, an **Anger** ranker makes an accurate prediction for the pair if its ranking score for the *anger* utterance is higher than the score for the other utterance. Table 2 shows pairwise accuracy for the six rankers on the Berlin and LDC datasets.

The pairwise accuracy is very high, especially for the Berlin dataset of prototypical emotional speech. There, the rankers for all emotions have pairwise accuracy above 94% and the **Neutral** and **Sadness** rankers give pairwise accuracy of over 99%. In the LDC dataset pairwise accuracies are much lower, with **Anger** and **Neutral** rankers doing markedly better than the rankers for the other emotions.

We now turn to look at Precision at k , which is widely used to evaluate the performance of ranking models. It is defined as the number of utterances in the top k utterances in the list ordered by decreasing ranking score that express the target emotion. This metric indicates

the percentage of the top ranked k utterances according to the **Anger** ranker were indeed anger utterances.

To demonstrate the ranker's performance at various levels, Fig. 2 shows the precision at k for different k and for all rankers. Results for the Berlin dataset are shown on the left and for LDC on the right. Note that the number of utterances for a given emotion varied from speaker to speaker, especially in the Berlin dataset where utterances were excluded from the collection in a post-recording assessment. The average number of utterances for a particular emotion across speakers is given in brackets after the emotion name in the legend of the figure. A perfect ranker will attain 100% precision rate for k smaller than that average for each emotion, then drop steadily.

As the graphs in Fig. 2 indicate, the precision is high for the top results and then drops quickly, on both datasets. The precision for the **Disgust** ranker on the Berlin dataset is an outlier in this respect and it gives low precisions regardless of the choice of k . This can be explained by the fact that the number of *disgust* utterances per speaker varies widely: from 0 to 11 on the ten LOSO folds in the Berlin dataset. Some speakers do not have any utterances that express disgust so all predictions are in fact wrong—for such speakers the ranker identifies which ones sound most like disgust but none of them are actually disgust utterances.

For LDC, as we may expect given the results from pairwise accuracy, the precision for smaller k is high for the *anger* and *neural* emotions and not as impressive for the rest.

In an idealized situation where we know the number of utterances that convey the target emotion for each speaker (n), we can measure the precision at different k for different speakers, with $k = n$ for that speaker. In other words if we knew that a speaker said 7 utterances in an *angry* manner, we can look at the precision at 7 and see how many of the top ranked seven utterances were indeed *anger* utterances. This would be equivalent to viewing the ranker as a binary classifier, where we assumed that the top n utterances are the ones expressing the target emotion and the corresponding precision at k ($p @ k$) can be referred to as the one-versus-all classification accuracy with the prior knowledge of the distribution of emotions. Table 3 lists our oracle results in terms of $p @ k$ on the Berlin and LDC datasets.

For comparison, we also report analogous results for the conventional one-versus-all binary classification tasks. We list the precision at k of these one-versus-all classifiers, where the top k utterances were selected based on the output probabilities of the binary classification. Here the confidence of the classifier defines the ranking of utterances, with more confident predictions ranked higher. This setting is equivalent to using the emotion profiles approach (Mower et al., 2011). As we see next, it performs worse on the ranking task than our approach which directly learns to rank the instances.

The results reveal that we can achieve higher precision at k rate for most of the emotions with emotion rankers. The results hold for both the Berlin and the LDC dataset. On the Berlin dataset precision increases by a few percentage points for all emotions, while for the LDC dataset the precision drops slightly for *disgust* but increases for all other emotions.

In the last column of Table 3 we also list the *true* precision of the one-versus-all SVM classifiers on the Berlin and LDC datasets. To compute these numbers we do not use the oracle information about the number of utterances from the target emotion for each speaker. As expected, results are uniformly worse for all emotions.

5.2. Selection of pairwise constraints

Ranking SVM is a pairwise method and the learned model can be considerably affected by the partial ordering used in the training set. Moreover some influential theories of emotion (Grimm et al., 2007a; Giannakopoulos et al., 2009; Truong et al., 2009) analyze emotion in terms of continuous dimensions such as *VALENCE* and *AROUSAL* rather than categorical classes. *VALENCE* indicates the degree to which the emotion is positive or negative and *AROUSAL* indicates the degree of activation in emotion expression. For example *anger* is characterized by high negative valence and high arousal while typical *sadness* has high negative valence but low arousal.

We investigated several orderings of emotions, inspired by the literature that adopts *VALENCE* and *AROUSAL* as the basis for their analysis. Examples of investigated pairwise constraints for **Anger** rankers in our studies are listed as follows.

- a. Anger > Others
- b. Anger > Disgust & Fear & Sadness > Neutral & Happy
- c. Anger > Disgust & Fear & Sadness > Neutral > Happy
- d. Anger > Disgust & Fear > Sadness > Neutral > Happy
- e. Anger > Disgust > Fear > Sadness > Neutral > Happy
- f. Anger > Happy > Neutral > Sadness > Fear > Disgust
- g. Anger > Happy > Fear > Disgust > Sadness > Neutral

The results reported in the previous section correspond to formulation (a). Now we evaluate precision at k , where k was total number of utterances for the target emotion, for the alternative formulations as well.

The corresponding precision for the **Anger** classifier is listed in the Table 4.⁴ From the experimental results we noticed significant difference among different pairwise constraints in terms of precision, and the simplest one-versus-all based rankers trained with setting (a) performed better than any other rankers constructed with complicated pairwise constraints.

5.3. Multi-class classification of emotion

Given a sample of speech, the emotion rankers indicate how relevant each of the utterances is to a particular emotion. However, the rankers do not directly give a way to decide which particular emotion is expressed by a given utterance. In order to classify the unknown test utterance as expressing one of the basic emotions, we need to combine ranking scores from

⁴Similar conclusions can be drawn from the results for other emotions as well. **Anger** only results are shown as a representative example because it has the most stable recognition rates for both datasets and is representative of the general trends.

different rankers. We implemented a rule-based and a supervised learning approach to combine the ranker scores into a final prediction. Next we give details on the implementation.

5.3.1. Rule-based combination—In the rule-based approach, the emotion of an utterance is decided by directly comparing the ranks assigned by the rankers for the six basic emotions. The utterance is classified as conveying the emotion for which it achieved highest rank. If an utterance had the same rank assigned by more than one ranker, a decision about which emotion to pick was made randomly. The accuracy of speaker-independent, multi-class classification on the Berlin and LDC datasets is shown in the column (1) of Table 5.

5.3.2. SVM classifier combination—In the supervised learning approach, we do two passes of training and testing. First, leave-one-subject-out paradigm is used to train emotion rankers. The test predictions from each fold are used to generate the six dimensional feature vector of normalized ranking scores described in Section 3. Then a standard SVM classifier is trained for the six-class classification task, again using leave one subject out paradigm.

The multi-class classification accuracy of this model on the Berlin and LDC datasets are shown in column (2) of Table 5. Overall for the six emotions the supervised combination slightly outperforms the rule-based combination on both datasets. The absolute performance gain are 2.4% and 1.7% for the Berlin and LDC datasets respectively.

It is interesting to note however that the rule-based combination leads to higher accuracy of predicting *neutral* utterances on the Berlin dataset and *sadness* instances on the LDC dataset. For all other emotions, the SVM classifier trained with normalized ranking scores yield notably higher emotion recognition accuracy. For instance, the absolute improvement in recognition accuracy of *disgust* is 13% for Berlin and 12.5% for LDC datasets.

5.3.3. Comparison with conventional SVM—Finally, in order to investigate the usefulness of ranking SVM in emotion recognition, we compare the performance of ranking SVM based multi-class classifiers to the results of conventional SVM classifiers. We conducted two sets of experiments.

In our first set of experiments, we performed the standard task of multi-class classification of the six basic emotions. We trained baseline SVM classifiers with radial basis kernels constructed with the LIBSVM toolkit (Chang and Lin, 2001). These multi-class classifiers were directly trained with the large standard set of 988 acoustic feature vectors, unlike the ranking SVM based classifiers which uses only six features derived from the rankers for specific emotions. The accuracy of speaker-independent, multi-class classification on the LDC and Berlin datasets is given in column (3) of Table 5.

The overall performance for the standard multi-class classifiers is much lower than that of the ranking SVM based one. The absolute degradation is as high as 7.3% and 8.3%, for the Berlin and LDC datasets respectively. The standard SVM classifiers also performed lower emotion recognition consistently on most emotions except for *disgust* in the Berlin dataset

and *sadness* in the LDC dataset, where the conventional SVM classifier performed much better.

In the next set of experiments we pinpoint exactly which aspects of the model contribute to the better performance of the ranking-based classifier compared to its conventional SVM classifier counterpart. In the previous section we presented strong evidence that the ranker scores rank utterances much more accurately than that can be done using the class probabilities from one-versus-all standard emotion classifiers. To confirm that the ranking scores lead to improvements in multi-class prediction, we trained top-level SVM classifiers with a new 6-dimensional feature set. By analogy to the approach we took in Section 5.3.2, the features are the target class probabilities predicted by the six conventional one-versus-all binary classifiers. Another reason for the improvement in multi-class prediction is that we do not use the raw scores from the ranker but instead define a normalized ranking score which is used as a feature in the second level classifier. We test how much this modeling decision affects performance, we perform a similar score normalization of the SVM class probabilities. Since the predicted probabilities also provide a natural ranking of the test instances, we can generate ranking scores, exactly as we defined them in Section 3, based on the output probabilities from the one-versus-all classifiers and use them as features for multi-class classification. Finally, improvements may be due to the natural way in which the ranker approach incorporates speaker information. To tease apart this contribution, we perform speaker-independent (SI) ranking in which we sort the target class probabilities from one-versus-all classifiers across all speakers or speaker-relative (SR) ranking which considered within-speaker sorting. The speaker-relative approach correspond exactly to the approach we took for the ranking-based multi-classification, the only difference is the type of learning model: classifier or ranker.

The goal of these experiments is to better tease apart the reasons for the considerable improvement in accuracy for the ranking-based multi-class classifier. Improvement could come from the fact that a second layer of learning was applied. Comparing the results with two-level learning with features derived from conventional classifier will show if the benefit comes from the use of ranker specifically. Further, the improvement could be due to the fact that in our approach we generate features from the ranking of utterance only for a particular speaker. The speaker independent and speaker relative experiments will allow us to analyze if this choice made a difference.

As the detailed results indicate, we reap benefits both from using the rank SVM and from the speaker relative generation of features for the second level of learning.

The multi-classification accuracies of all these classifiers are given in the column (4), (5), and (6) of Table 5 for the Berlin and LDC datasets. Column (4) lists the results based on the actual one-versus-all output probabilities, while column (5) and (6) show results based on speaker-independent (SI) and speaker-relative (SR) rank-related systems, where ranking is performed based on the class probability from the standard classifier.

Clearly the ranking SVM based multi-class classifiers outperform all versions of those based on one-versus-all SVMs. In contrast to what we observed when we combined the scores

from the ranking SVMs, combining the actual prediction probabilities of binary classifiers did not improve classification performance. On the other hand, when applying the one-versus-all based ranking scores instead of the true probabilities, speaker independent ranking based classifier performed worst among all classifiers, while significant improvement can be achieved on speaker-relative ranking based classifiers by considering the speaker differences in emotion expressivity.

Note that in the three experiments where we employed target class probabilities to rank utterances, only the speaker sensitive experiment, in which the utterances from the same speaker are considered as a group, leads to significant improvement over the standard multi-class SVM classification. The choice of model—ranker rather than classifier—leads to further improvement. Based on these results we can conclude that ranking and speaker sensitivity contribute roughly equally to the improved performance.

6. Emotion recognition experiments on spontaneous speech

In this section we investigate the performance of ranking models on the more challenging tasks of classification of spontaneous emotional speech. We perform experiments on the FAU Aibo dataset which contains mostly neutral utterances with a relatively small portion of less expressive emotional utterances.⁵

The goal of these experiments is to demonstrate that the benefits from ranking approaches are not constrained to acted speech and that they can boost the recognition of emotion in spontaneous interaction. We first evaluate, in Section 6.1, the performance of individual emotion rankers, as we did in our work in acted emotion datasets. Next, we examine the performance of a ranking-based classification system that incorporates the results from individual rankers to perform multi-class prediction for each utterance. We report these results, along with the performance of conventional SVM classifiers, in Section 6.2.

6.1. Evaluation of individual emotional rankers

In Table 6 we list the pairwise accuracy for the five rankers on the FAU Aibo dataset. Not surprisingly, the performance is much lower on this dataset than on the acted Berlin and LDC datasets (cf. Table 2). There is a striking difference in performance on the three coherent emotion classes *anger*, *emphatic*, and *positive*, on which the ranker performs well with pairwise accuracies close to 75%, and the mixed emotion class *rest* for which pairwise accuracy drops below 60%. The pairwise accuracy on *neutral* is also lower compared to the coherent emotions.

Furthermore, Fig. 3 shows the precision at k for different k for all rankers for the FAU dataset. Unlike in the acted datasets we discussed before, the average number of utterances per speaker (given in brackets after the emotion name in the legend of the figure) vary considerably for different emotions. For instance, the average number of utterances is more than 200 for *neutral*, but less than 10 for *positive*. As the graph indicates, the precision is high for the top results and then drops steadily for all emotions except of *positive*. This is

⁵Some of the results from this section have been previously presented at Interspeech 2012 (Cao et al., 2012).

because the number of *positive* instances are very different for different speakers. For some speakers there is no positive utterances available while for some other speakers more than 30 examples can be found.

Similarly to the results we reported in Table 3 for acted emotional speech, we also analyze the precision at k ($p @ k$) for spontaneous speech. Here results are computed for a fixed k equal to the total number of target emotions for each emotion rankers. The results on spontaneous speech are summarized in Table 7. We also give the respective results for one-versus-all binary classification in the table, for the sake of comparison.

Consistent with what we found on the acted Berlin and LDC datasets, the results on the spontaneous FAU database confirm that we can achieve higher precision at k with emotion rankers than with the conventional binary classifiers. Remarkably we also observe that the performance gain is bigger for the coherent emotion classes *anger* and *emphatic* than for *neutral* and *rest*.

6.2. Multi-class classification results on spontaneous speech

Now we turn to examine the performance of ranking-based approaches for the task of multi-class emotion classification on the spontaneous data. As we described in Section 5.3, in order to classify a test utterances as expressing one of the emotion classes, we need to combine ranking scores from different rankers. On the acted speech datasets we found that the supervised combination outperforms the rule-based combination. For this reason, here we study the performance of the two-pass system for emotion classification on spontaneous speech by training a standard SVM classifier to combine the ranking scores from the individual rankers.

We use a leave-one-subject-out paradigm to train emotion rankers on the training set. The test predictions from each emotion ranker in each fold are normalized to assign rank scores comparable across rankers, as described in Section 3. These rank scores (one for each emotion) are used to represent the utterance and a conventional multi-class SVM classifier is trained, again using leave-one-subject-out paradigm on the training set. We compare the performance of this system to that of a conventional 5-class SVM classifier.

Since the number of instances per emotion class varies widely in the FAU Aibo database as shown in Table 1, we used the unweighted average (UA) recall, also known as balanced accuracy, as performance metric for the emotion recognition experiments presented below. The results for the conventional SVM classifier and the ranking-based system are presented in Table 8.

The UA recall for the conventional and ranking-derived classifiers were 41.5% and 39.4% respectively. The ranking-derived classifier has 10% better accuracy in identifying *Anger* and 4% better for *Emphatic* utterances. It does, however, make more mistakes on the *Neutral* class compared to the conventional classifier. Given that the two classifiers perform well on different classes, we expect that the combination of the two approaches will further improve performance. Both classifiers have very low accuracy for *Rest*. This is not

surprising because *Rest* is a catch-all class combining different infrequent emotions, unlike the other four classes which contain utterances expressing the same emotion.

7. Discussion

7.1. Complementarity of SVM and ranking SVM

In order to better understand how the conventional and ranking-derived classifiers differ from each other, we carried out further analysis to investigate the complementarity of the two classifiers. We first divided the testing data into two groups according to their prediction results by the two classifiers.

Group I (SVM agree with Ranking SVM): contains instances classified as conveying the same emotion by the ranking SVM and the conventional SVM classifiers.

Group II (SVM differ from Ranking SVM): contains instance classified as conveying different emotions by the two classifiers.

We first list the proportion of the entire dataset that falls in the two groups on the Berlin, LDC, and FAU Aibo databases respectively in Table 9. There is high agreement between the conventional SVM and ranking SVM on the acted Berlin dataset of unambiguous, prototypical emotion expressions; there the two method give the same prediction on more than 75% of the testing data. However, the agreement between the two classifiers decreases for the less prototypical speech; the two classifiers disagree with each other for more than half of the test instances on the LDC and FAU datasets.

Table 10 summarizes the performance on all three datasets of the classifiers on subsets of the test data where their predictions agree and disagree (Groups I and II as we defined them above). The difference in performance on the two groups immediately stands out. For instance, the UA recall for Group I is 88.4% and 56.6% on Berlin and LDC respectively, more than double the UA recall of 37.1% and 26.8% achieved by the conventional SVM classifiers on the Group II partition of the test data. Consistent with this behavior, on the spontaneous FAU data we obtain UA recall of 50.1% on the Group I part of the test data, while the results for Group II of the same SVM system is only 32.2%. These results suggest that the ranking system can be used in tandem with the conventional SVM classifier in order to find a smaller subset for which prediction accuracy is much higher than if prediction is performed on the entire dataset. Agreement between the two classifiers gives a strong indication that the prediction is correct.

In addition, the ranking-based classifiers perform significantly better than conventional SVM, for both acted and spontaneous emotional datasets, on the Group II partition where the two classifiers give different predictions. The corresponding UA recall is 57.0% and 44.8% for the acted Berlin and LDC respectively, which show absolute improvement of 19.9% and 18.0% compared with the results for conventional SVM classifiers.

Additionally on the predictions for which the two methods disagree (Group II), on the spontaneous FAU dataset, the overall UA recall of the ranking-based system is considerably higher than that of the traditional classification approach, i.e. 36.8% compared to 32.2%. As

we already observed in our prior experiments, the ranking-based system is particularly successful at predicting the coherent emotion classes *anger*, *emphatic*, *positive*. On these classes, the UA of the ranking system is 40.4%, compared to the 28.6% for the standard SVM classifier; this difference translates to 11.8% absolute and 41.4% relative improvement for the ranking approach. This promising improvement suggests that the advantage of our proposed approach will be more significant in realistic situations, where the most typical tasks are retrieving emotional instances from large sample sets where the majority of expression are likely to be *Neutral*.

7.2. Combination of SVM and ranking SVM classifiers

Finally, we consider the direct combination of conventional SVM and ranking-based classifiers into a single prediction. As shown in López-Cózar et al. (2010), many classifier combination methods can be effective in the emotion recognition tasks. Here, we investigate the performance of two often-used late-stage fusion methods.

In the first combination approach, we combine predictions from the classifiers by multiplying the posterior probabilities for each emotion class produced by each of the classifiers. The emotion class with highest resulting probability is predicted as the class for each instance in the test data.

In the other combination approach, we resort to supervised learning for the combination. For each utterance, we generate a new feature representation that consists of the output probabilities for each emotion class, produced by the conventional and the ranking-derived multi-class SVMs. A new multi-class classifier is trained with these representations.

The performance of the two fusion systems are summarized in Table 11. For the sake of comparison, we also repeat the results of the original classifiers in the same table. The two late-stage fusion systems generally have similar performance. For the acted Berlin and LDC datasets, the two fusion systems perform much better than the conventional SVM systems but not better than the ranking SVM systems. This may be because on these datasets the conventional SVM classifiers perform significantly worse than the ranking systems for all most emotion classes and so the combination of predictions is not beneficial.

For the spontaneous FAU Aibo database, however, the performance clearly improves when the two classifiers—conventional and ranking-derived—are combined. On this dataset, the two individual classifiers achieve comparable performance in general, but each shows advantages on different emotions. As a result, the fusion system achieves the best UA recall of 44.4%, by taking advantage of information provided by both the conventional and the ranking based classifiers. Compared with the SVM classifiers, the combined system exhibits much higher recall rate on all clean emotion classes and maintained a reasonable recall rate on *Neutral*.

8. Conclusions

In this paper, we introduced a novel ranking model for emotion recognition. In contrast to the state-of-the-art emotion classification systems which rate single test utterance

independently of each other with one prediction score, the ranking approach is applicable to situations when multiple utterances from the same speaker are available. In such situations, all utterances by the same speaker are considered at prediction time, and the degree to which they resemble each of the emotions is calculated. These assessments are then converted to a final prediction of a single dominant emotion. Our experiments reveal that both the choice of model—ranking rather than classification, and the speaker-relative information contribute to the consistent improvements over the standard approaches. In many applications, speaker information is already available as part of the recording setup i.e. through the use of directional (noise canceling) microphones. Application of the approach is feasible even in general settings, where speaker recognition and clustering could be performed automatically (Kinnunen and Li, 2010; Anguera Miro et al., 2012).

We first investigated ranking models on two publicly available datasets of acted speech and evaluated model effectiveness in terms of precision rate of the individual emotional ranker, as well as the overall multi-class classification accuracy. Compared with traditional one-versus-all binary classifiers, higher precision rate was achieved with rankers. Some emotional utterances which were misclassified by traditional binary classifiers can be correctly predicted by ranking models. The advantages of the ranking view for emotion analysis were further proven in multi-class classification tasks. The recognition results indicate that emotion classification directly based on ranking SVM scores significantly outperform traditional SVM algorithm. Furthermore, creating SVM classifiers with normalized ranking SVM scores additionally improves the emotion recognition.

Similarly to previous work (Bitouk et al., 2009), we observe that the overall performance of emotion recognition on the LDC and Berlin datasets varies considerably. Performance on the LDC dataset is much lower than the that on the Berlin dataset. However, we noticed that *anger* emotion can be successfully classified by ranking models in both of these two datasets. In addition, we also find that ranking models offer significant improvement in recognition of *fear* and *happy* emotion on both datasets. We also observe that much of the improvement that we obtained between directly comparing the ranks to come up with a multi-class predicting and training a classifier that combines the normalized ranking scores was due to the refinement of *happy* models, where significant reduction were observed on *disgust-happy* confusions on both the Berlin and LDC datasets.

In addition, we further evaluated the proposed the ranking models for emotion recognition tasks on realistic spontaneous speech from the FAU Aibo. Compared with conventional SVM classifiers, ranking-based classifiers recognize many emotional instances better but have relatively lower accuracy for neutral utterances. By combining the two approaches, we achieve performance better than that of either individual classifier. The combination system achieves UA recall of 44.4%. The benchmark result of Interspeech 2009 emotion challenge, which does not use any information on speaker identity, is 38.2% reported in Planet et al. (2009). Much higher results of 44.0% can be obtained by majority voting among the best seven systems that participated in the challenge Schuller et al. (2011).

The FAU Aibo corpus contains spontaneous speech and the majority of utterances in it are neutral. Only a relatively small portion of the utterances in the corpus are expressions of

spontaneous emotional speech. Our promising results on the FAU Aibo dataset suggest that the proposed emotion recognition system should be able to retrieve emotional instances from large samples of data in many realistic situations.

Finally, we discussed the complementarity of conventional SVM classifiers and ranking-based classifiers and further observed that performing ranking-SVM based classification in parallel with conventional SVM classification and examining their predictions may help us to differentiate highly reliable predictions from relatively poor ones.

Acknowledgments

This work was supported in part by the grant NIH R01-MH073174.

References

- Anguera, Miro X.; Bozonnet, S.; Evans, N.; Fredouille, C.; Friedland, G.; Vinyals, O. Speaker diarization: a review of recent research. *Transactions on Audio, Speech & Language Processing*. 2012; 20(2):356–370.
- Bitouk D, Nenkova A, Verma R. Improving emotion recognition using class-level spectral features. *INTERSPEECH*. 2009:2023–2026.
- Bitouk D, Verma R, Nenkova A. Class-level spectral features for emotion recognition. *Speech Communication*. 2010:613–625. [PubMed: 23794771]
- Burkhardt F, Paeschke A, Rolfes M, Sendlmeier WF, Weiss B. A database of german emotional speech. *INTERSPEECH*. 2005:1517–1520.
- Busso C, Metallinou A, Narayanan SS. Iterative feature normalization for emotional speech detection. *ICASSP*. 2011:5692–5695.
- Cao H, Verma R, Nenkova A. Combining ranking and classification to improve emotion recognition in spontaneous speech. *INTERSPEECH*. 2012
- Chang, CC.; Lin, CJ. LIBSVM: a library for support vector machines. 2001. Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>
- Cowie R, Douglas-cowie E, Savvidou S, McMahon E, Sawey M, Schrder M. feeltrace: an instrument for recording perceived emotion in real time. *Proceeding of ISCA Workshop on Speech and Emotion*. 2000:19–24.
- Dellaert F, Polzin T, Waibel A. Recognizing emotion in speech. *Proceedings of the CMC*. 1996:1970–1973.
- Eyben F, Wöllmer M, Schuller B. Opensmile: the munich versatile and fast open-source audio feature extractor. *ACM Multimedia*. 2010:1459–1462.
- Fernandez R, Picard R. Modeling drivers' speech under stress. *Speech Communication*. 2003:145–159.
- Giannakopoulos T, Pikrakis A, Theodoridis S. A dimensional approach to emotion recognition of speech from movies. *ICASSP*. 2009:65–68.
- Grimm M, Kroschel K, Mower E, Narayanan S. Primitives-based evaluation and estimation of emotions in speech. *Speech Communication*. 2007a Oct 10–11.49:787–800.
- Grimm M, Kroschel K, Narayanan SS. Support vector regression for automatic recognition of spontaneous emotions in speech. *ICASSP*. 2007b:1085–1088.
- Huang R, Ma C. Toward a speaker-independent real-time affect detection system. *ICPR*. 2006; (1): 1204–1207.
- Joachims T. Optimizing search engines using clickthrough data. *Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (KDD)*. 2002:133–142.
- Joachims T. Training linear SVMs in linear time. *Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (KDD)*. 2006
- Kinnunen T, Li H. An overview of text-independent speaker recognition: from features to supervectors. *Speech Communication*. 2010; 52(1):12–40.

- Kwon OW, Chan K, Hao J, Lee T. Emotion recognition by speech signals. *Proceedings of 8th European Conference on Speech Communication and Technology*. 2003:125–128.
- Lee C, Yildirim S, Bulut M, Kazemzadeh A, Busso C, Deng Z, Lee S, Narayanan S. Emotion recognition based on phoneme classes. *INTERSPEECH*. 2004:205–211.
- Lee CC, Mower E, Busso C, Lee S, Narayanan S. Emotion recognition using a hierarchical binary decision tree approach. *Speech Communication*. 2011; 53(9–10):1162–1171.
- Lee CM, Narayanan SS, Pieraccini R. Classifying emotions in human-machine spoken dialogs. *ICME*. 2002; (1):737–740.
- Linguistic Data Consortium. Emotional prosody speech and transcripts. University of Pennsylvania; 2002. LDC Catalog No.: LDC2002S28
- Litman DJ, Forbes-Riley K. Predicting student emotions in computer-human tutoring dialogues. *Proceedings of Association for Computational Linguistics*. 2004:352–359.
- López-Cózar, R.; Silovsky, J.; Griol, D. F2 – new technique for recognition of user emotional states in spoken dialogue systems. *Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue*; 2010. p. 281–288.
- Luengo I, Navas E, Hernez I, Snchez J. Automatic emotion recognition using prosodic parameters. *Interspeech*. 2005:493–496.
- McGilloway S, Cowie S, Douglas-Cowie E, Gielen S, Westerdijk M, Stroeve S. Approaching automatic recognition of emotion from voice: a rough benchmark. *ISCA Workshop on Speech and Emotion*. 2000:200–205.
- Meng H, Pittermann J, Pittermann A, Minker W. Combined speech-emotion recognition for spoken human-computer interfaces. *IEEE International Conference on Signal Processing and Communications*. 2007:1179–1182.
- Mower E, Mataric MJ, Narayanan SS. A framework for automatic human emotion classification using emotion profiles. *IEEE Transactions on Audio, Speech & Language Processing*. 2011; 19(5): 1057–1070.
- Neiberg D, Elenius K, Laskowski K. Emotion recognition in spontaneous speech using GMMs. *Interspeech*. 2006
- Nicholson J, Takahashi K, Nakatsu R. Emotion recognition in speech using neural networks. *Neural Computing and Applications*. 2000
- Nwe T, Foo S, Silva LD. Speech emotion recognition using hidden Markov models. *Speech Communication*. 2003; 41(4):603–623.
- Petrushin V. Emotion in speech: recognition and application to call centers. *Proceedings of Artificial Neural Networks in Engineering*. 1999:7–10.
- Planet S, Sanz II, Socoró JC, Monzo C, Adell J. GTM-URL contribution to the interspeech 2009 emotion challenge. *INTERSPEECH*. 2009:316–319.
- Schuller B, Batliner A, Steidl S, Seppi D. Recognising realistic emotions and affect in speech: state of the art and lessons learnt from the first challenge. *Speech Communication*. 2011; 53(9–10):1062–1087.
- Shafraan I, Riley M, Mohri M. Voice signatures. *Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop*. 2003:31–36.
- Steidl, S. PhD thesis. 2009. Automatic classification of emotion related user states in spontaneous children's speech.
- Tabatabaei T, Krishnan S, Guergachi A. Emotion recognition using novel speech signal features. *IEEE International Symposium on Circuits and Systems*. 2007:345–348.
- Truong KP, van Leeuwen DA, Neerincx MA, de Jong FMG. Arousal and valence prediction in spontaneous emotional speech: felt versus perceived emotion. *INTERSPEECH*. 2009:2027–2030.
- Ververidis D, Kotropoulos C. Emotional speech recognition: resources, features, and methods. *Speech Communication*. 2006; 48(9):1162–1181.
- Vlasenko B, Schuller B, Wendemuth A, Rigoll G. Frame vs. turn-level: emotion recognition from speech considering static and dynamic processing. *Affective Computing and Intelligent Interaction*. 2007:139–147.

- Vondra M, Vich R. Recognition of emotions in german speech using gaussian mixture models. *Multimodal Signals: Cognitive and Algorithmic Issues*. 2009:256–263.
- Ye C, Liu J, Chen C, Song M, Bu J. Speech emotion classification on a riemannian manifold. *Advances in Multimedia Information Processing – PCM 2008*. 2008:61–69.
- Yu C, Aoki PM, Woodruff A. Detecting user engagement in everyday conversations. *INTERSPEECH*. 2004

Fig. 1.
Ranking system for emotion recognition.

Fig. 2.

Precision at k along with different K on Berlin and LDC datasets. The numbers followed by emotion states are the average number of testing examples per emotion of all LOSO folds.

Fig. 3.

Precision at k along with different K on FAU Aibo dataset. The numbers in parenthesis after the emotion name are the average number of testing examples per emotion across all LOSO folds.

Table 1

Number of instances for the 5-class problem.

	Speaker	A	E	N	P	R
Train	26	881	2093	5590	674	721
Test	25	611	1508	5377	215	546

Table 2

Pairwise accuracy (%) for the six emotion rankers on the Berlin and LDC datasets. Here accuracies are the average of all LOSO folds.

	Berlin dataset	LDC dataset
Anger	98.17	94.92
Disgust	98.28	68.68
Fear	97.71	70.38
Happy	91.88	81.01
Neutral	99.81	96.76
Sadness	99.96	66.16

Table 3

$p @ k$ (precision at k (%); k = total number of utterances for target emotions) from emotional rankers, conventional one-versus-all classifiers, and overall precision rate of one-versus-all classification for Berlin (top) and LDC (bottom) datasets.

Berlin dataset	Precision at k ($p @ k$)		Precision
	Ranker	Classifier	Classifier
Anger	89.6	86.6	77.1
Disgust	82.6	76.1	77.3
Fear	81.2	76.8	83.3
Happy	69.0	54.9	55.7
Neutral	97.5	89.9	78.4
Sadness	99.4	98.4	88.8
LDC dataset	Precision at k ($p @ k$)		Precision
	Ranker	Classifier	Classifier
Anger	75.0	72.4	53.9
Disgust	41.7	45.8	33.3
Fear	37.6	34.1	29.5
Happy	47.6	42.8	31.8
Neutral	81.3	79.2	74.6
Sadness	34.2	30.3	25.6

Table 4

$p @ k$ (k = total number of utterances for the target emotion) for different **Anger** rankers which trained with different pairwise constraints on Berlin and LDC datasets.

	Berlin dataset	LDC dataset
(a)	89.6	75.0
(b)	70.1	42.1
(c)	63.8	32.9
(d)	64.6	35.5
(e)	66.1	35.5
(f)	68.5	39.5
(g)	78.7	46.1

Table 5

Speaker-independent, multi-class emotion classification accuracy for six emotion task on Berlin (top) and LDC (bottom) datasets when we (1) directly compare ranking SVM scores and (2) use normalized classification scores to train an SVM classifier. Classification accuracy for (3) conventional SVM multi-class classifier, (4) classifiers trained with one-versus-all binary classification probabilities, (5) classifiers trained with one-versus-all based speaker independent (SI) ranking and (6) classifiers trained with one-versus-all based speaker relative (SR) ranking are also given to allow comparison.

Berlin dataset	(1) Ranking compare	(2) Ranking learning	(3) Multi-class	(4) 1-Versus-all prob. based	(5) 1-Versus-all SI ranking	(6) 1-Versus-all SR ranking
Anger	88.2	89.8	81.1	77.2	82.7	85.0
Disgust	60.9	73.9	80.4	82.6	58.7	60.9
Fear	75.4	76.8	63.8	73.9	66.7	78.3
Happy	71.8	74.6	62.0	47.9	50.7	62.0
Neutral	88.6	87.3	82.3	82.3	82.3	86.1
Sadness	88.7	90.3	85.5	90.3	74.2	91.9
Average	81.1	83.5	76.2	75.3	71.6	79.0
LDC dataset	(1) Ranking compare	(2) Ranking learning	(3) Multi-class	(4) 1-Versus-all prob. based	(5) 1-Versus-all SI ranking	(6) 1-Versus-all SR ranking
Anger	76.3	76.3	72.4	63.2	57.9	68.4
Disgust	34.4	46.9	27.1	21.9	17.7	42.7
Fear	34.1	37.6	23.5	38.8	24.7	38.8
Happy	47.6	58.3	47.6	32.1	38.1	58.3
Neutral	83.0	84.9	54.7	62.3	66.0	79.2
Sadness	32.9	10.5	36.8	19.7	28.9	11.8
Average	48.7	50.4	42.1	37.7	36.4	48.1

Table 6

Pairwise accuracy (%) for five emotion rankers on FAU Aibo dataset.

	Anger	Emphatic	Neutral	Positive	Rest
Pairwise accuracy (%)	78.56	74.59	70.87	75.77	57.58

Table 7

$p @ k$ (precision at k (%); k = total number of utterances for target emotions) from emotion rankers, conventional one-versus-all classifiers, and overall precision rate of one-versus-all classification for the FAU Aibo dataset.

	Precision at k ($p @ k$)		Precision
	Ranker	Classifier	Classifier
Anger	36.5	30.8	29.3
Emphatic	46.4	37.7	35.1
Neutral	77.2	73.5	73.2
Positive	19.5	19.1	18.3
Rest	15.0	13.9	13.5

Table 8

Multi-class classification rate (%) of conventional SVM and ranking SVM on the FAU AIBO datasets for the 5-class task.

	Anger	Emphatic	Neutral	Positive	Rest	U/A
SVM	50.1	42.6	40.9	53.5	20.5	41.5
Ranking SVM	60.1	46.5	27.0	52.6	11.0	39.4

Table 9

Proportion of Group I (same predictions) and Group II (different predictions) on the Berlin, LDC and FAU Aibo datasets.

	Berlin	LDC	FAU
Group I	76.4%	46.0%	45.8%
Group II	23.6%	54.0%	54.2%

Table 10

Multi-class classification rate (%) on the Berlin, LDC, and FAU Aibo datasets. Group I: test utterance classified as the same emotion class by conventional and ranking-based classifiers; Group II: test utterance classified as different emotions by two classifiers.

Berlin	Group I	Group II	
	SVM/Ranking	SVM	Ranking
Anger	94.2	25.0	70.8
Disgust	93.5	53.3	20.0
Fear	83.3	19.0	81.0
Happy	74.0	33.3	61.9
Neutral	89.6	41.7	58.3
Sad	95.8	50.0	50.0
UA	88.4	37.1	57.0
LDC	Group I	Group II	
	SVM/Ranking	SVM	Ranking
Anger	85.7	35.0	50.0
Disgust	42.5	16.1	50.0
Fear	29.2	21.3	41.0
Happy	71.4	23.8	45.2
Neutral	96.2	14.8	74.1
Sad	14.3	50.0	8.3
UA	56.6	26.8	44.8
FAU	Group I	Group II	
	SVM/Ranking	SVM	Ranking
Anger	68.7	22.7	43.4
Emphatic	58.3	31.8	40.4
Neutral	38.2	41.6	37.1
Positive	73.1	31.3	37.5
Rest	12.4	33.8	25.8
UA	50.1	32.2	36.8

Table 11
Multi-class classification rate (%) of different individual and fusion systems on the Berlin, LDC, and FAU Aibo datasets.

Berlin	Anger	Disgust	Fear	Happy	Neutral	Sad	UA
SVM	81.1	80.4	63.8	62.0	82.3	85.5	75.8
Ranking SVM	89.8	73.9	76.8	74.6	87.3	90.3	82.1
Fusion_multiplication	88.2	76.1	78.3	66.2	88.6	91.9	81.6
Fusion_feature	89.0	73.9	75.4	66.2	91.1	93.5	81.5
LDC	Anger	Disgust	Fear	Happy	Neutral	Sad	UA
SVM	72.4	27.1	23.5	47.6	54.7	36.8	43.7
Ranking SVM	76.3	46.9	37.8	58.3	84.9	10.5	52.4
Fusion_multiplication	78.9	38.5	30.6	40.5	92.5	31.6	52.1
Fusion_feature	73.7	54.2	36.5	60.7	67.9	25.0	53.0
FAU	Anger	Emphatic	Neutral	Positive	Rest	UA	
SVM	50.1	42.6	40.9	53.5	20.5	41.5	
Ranking SVM	60.1	46.5	27.0	52.6	11.0	39.4	
Fusion_multiplication	58.1	48.5	37.6	57.2	19.4	44.2	
Fusion_feature	58.1	46.7	35.6	69.8	11.9	44.4	