

# EMOTION CLASSIFICATION FROM SPEECH USING EVALUATOR RELIABILITY-WEIGHTED COMBINATION OF RANKED LISTS

*Kartik Audhkhasi, Shrikanth S. Narayanan*

Signal Analysis and Interpretation Lab (SAIL)

Electrical Engineering Department, University of Southern California, Los Angeles, CA, U.S.A.

Email: audhkhas@usc.edu, shri@sipi.usc.edu

## ABSTRACT

In emotion recognition, a widely-used method to reconcile disagreement between multiple human evaluators is to perform majority-voting on their assigned class labels. Instead, we propose asking evaluators to rank emotional categories given an audio clip, followed by a combination of these ranked lists. We compare two well-known ranked list voting methods - Borda count and Schulze's method, with majority-voting and an evaluator model-based combination of the top ranked-labels. When tested on an emotional speech database with ground truth labels available, two interesting observations emerge. First, majority-voting performs significantly worse than the other three methods in the estimation of the given ground truth labels. Second, when performing classification using the combined labels, the two ranked list voting methods perform the best. We then propose evaluator reliability-weighted versions of these two methods, which improve the classification accuracy even further.

*Index Terms*— Emotion recognition, voting methods, evaluator reliability

## 1. INTRODUCTION

In a majority of supervised pattern recognition problems involving human behavioral data, the class labels or target variables of interest are highly subjective and inherently difficult to define concretely. For example, in emotion recognition from speech, the typical emotional categories considered are {angry, happy, sad, neutral}. However, it is well-known that the expression and experience of natural emotions are often ambiguous, and described as blended and non-prototypical. Consequently, human emotions are tough to be discretized into clearly defined disjoint classes [1]. Hence, defining classes of interest, and labeling human behavioral databases with predefined categories often is a challenging problem. Furthermore, the ambiguity in the data and its labeling severely hinders the design of pattern classification systems, which conventionally exploit high intra-class and low inter-class similarity between features. A popular approach to obtain labels which are representative of the true hidden classes is to ask multiple human evaluators to label the data. Then, a majority vote is applied on all evaluator labels associated with a given sample in the database, and it is hoped that the resulting resolved label is closer to true label of the sample than each of the individual evaluator labels. In [2], the authors demonstrate that these majority vote labels improve classification performance over each of the individual evaluator labels.

It must be noted that majority vote combination of evaluators is not optimal, since it gives equal weight to the opinion of each human evaluator. Every evaluator has varying skills and his/her own perception of the various emotion categories are in general derived differently from a multitude of cues (speech, facial and body gestures etc.) and disparate personal experiences. Many works in the past, particularly in the machine learning community, have looked at the problem of estimating hidden ground truth labels from multiple noisy evaluator labels, and modeling evaluators in general. One of the earliest works in this direction was by Smyth et al. [3], where their task was estimating the ground truth label for the presence of volcanoes in radar images. The authors pose this problem in an Expectation-Maximization (EM) framework (to be discussed in more detail in section 2). In [4], the authors go a step further, and jointly learn a model for estimating the hidden ground truth label, and a logistic regression classifier between this true hidden label and the feature vector. A more realistic model of a human evaluation process has been presented in [5], where the authors model the fact that the reliability of human evaluators varies from one sample to the next. They apply this model to emotional valence classification from speech, and show an improvement over majority logic combination of labels and previous data-independent evaluator models.

This paper approaches the problem of deriving more representative labels from human evaluators in a different paradigm. Instead of asking evaluators to give only the best matching emotional label to an audio clip, they are asked to rank the entire (finite) set of labels in decreasing order of preference. This is based on the intuition that when evaluators are restricted to give only the best label, they are forced to make a hard decision about the type of emotion present in the utterance. However, as previously noted, an utterance may contain ambiguous emotions or even a mixture (blend) of several ones. Giving evaluators the flexibility of generating a ranked list, thus, guards against the possibility of missing the true representative emotion(s) present in the utterance. Next, we use two standard voting methods for combination of multiple ranked lists, and use the estimated labels for training an emotion classifier. We note that neither of these voting methods takes different evaluator reliabilities into account. Hence, we propose a heuristic modification of these methods based on the evaluator model presented in [3]. It is shown that the new reliability-weighted voting schemes on multiple ranked lists out-perform the standard variants, and also the majority voting and EM-based maximum-a-posteriori (MAP) estimation that use only the top ranked label.

This paper is organized as follows. Section 2 discusses the emotional speech database used in this paper. An oracle experiment is also presented as a motivation for using the complete ranked list of labels as opposed to just the top label. Next, in section 3, we discuss

---

This work was supported by the NSF, DARPA and Army.

the label fusion approach presented in [3], which we use as our baseline along with majority voting. We then discuss the two ranked list voting methods - Borda count and Schulze’s method. This section concludes by defining a global evaluator reliability metric based on the model in [3], and by proposing simple heuristic modifications to the two ranked list voting methods based on this metric. Section 4 presents the experiments and results, and the paper concludes in section 5 with a discussion and few areas of future work.

## 2. RANKED-LIST EMOTIONAL SPEECH DATABASE

An emotional speech database was collected from one male speaker speaking five repetitions of seven sentences, in each of the five types of emotions (neutral, hot anger, cold anger, happiness and sadness) and three speaking styles (normal, fast and intense). Thus, the total number of files in the database was 525. It may be noted that for the purpose of this paper, the number and type of emotion categories used is not an important factor, as is the acoustic variability, since we are mainly interested in the evaluations of the audio data by multiple evaluators, and the benefits obtained by different algorithms of combining them.

For evaluation, three evaluators (all native speakers of American English) were asked to listen to these audio clips, and rank the five emotions in decreasing order of preference. In addition, for the top-ranked choice, they were also asked to rate their confidence and the strength of the emotion (both on a scale of 1 to 5). In ranking the emotions, the evaluators were given a sixth choice, “other”, which signified that the audio clip contained an emotion which was not in the given set of five emotions.

This database contains all acted emotions, since the speaker was given the emotion type with which a particular sentence was to be read. Hence, for the purpose of an initial analysis, we assume that the identity of the emotion to be acted is representative of emotional content of the utterance. Next, we compute the oracle accuracy of the top  $N$  ranks assigned by each evaluator, where  $N \in \{1, 2\}$ . The accuracy of the top  $N$  ranks for a given evaluator is defined to be the percentage of utterances in the database when the top  $N$  ranks assigned by the evaluator contained the true emotional label. Obviously, the oracle accuracy of the top 6 ranks for each evaluator will be 100%. Figure 1 gives the top-1 and top-2 oracle accuracies of the three evaluators. As can be observed, the oracle accuracy rises by roughly 15% absolute for all three evaluators, upon incorporating the second ranked emotion in addition to the top ranked one. This motivates the idea that the high ranked emotions other than the top ranked one also contain significant information about the true emotional content of the utterance.

## 3. ALGORITHMS FOR COMBINATION OF RATINGS BY MULTIPLE EVALUATORS

We first discuss the approach presented in [3], which performs estimation of the true label using just the top ranked label assigned by each evaluator. Figure 2 shows the Bayesian network structure used for learning the evaluator model and inference of the ground truth labels in [3]. It is hypothetically assumed that each evaluator can observe the true hidden label for each sample in the database. The evaluator then generates his own label by sampling from a  $K$ -valued categorical distribution conditioned on the value of the hidden label. This distribution is denoted by  $\{A^r(k_1, k)\}_{k_1=1}^K$ . Thus, the  $r^{th}$  evaluator has a  $K \times K$  reliability matrix,  $A^r$ . Large diagonal entries of the matrix signify a more reliable evaluator, since he is very likely

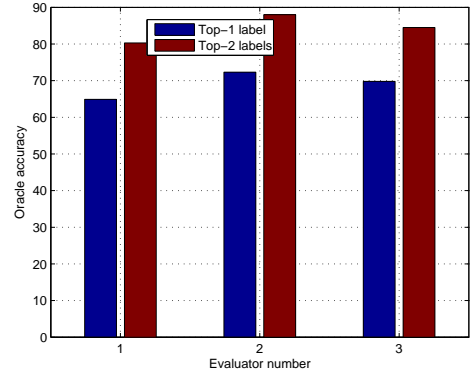


Fig. 1. Top-1 and top-2 oracle accuracies of the three evaluators.

to label the sample with the true hidden label. The hidden labels are assumed to be generated from a prior categorical distribution,  $P(y = k) = p_k$ .

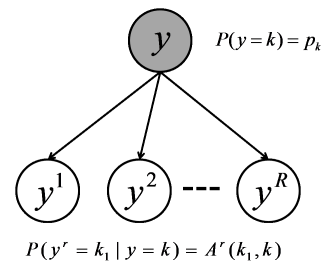


Fig. 2. Bayesian network structure used in [3] for  $R$  evaluators.  $y$  is the true hidden label and  $y^r$  is the label assigned by the  $r^{th}$  evaluator.

The parameters of this model ( $\{p_k\}_{k=1}^K$  and  $\{A^r\}_{r=1}^R$ ) can be learned using the EM algorithm. Given optimal estimates of these parameters, one can find the maximum-a-posteriori (MAP) estimate of the hidden ground truth label  $y$  when evaluator labels  $\{y^1, \dots, y^R\}$  are provided, as follows:

$$\begin{aligned}
 y_{MAP} &= \arg \max_{k \in \{1, \dots, K\}} P(y = k | y^1, \dots, y^R) \\
 &= \arg \max_{k \in \{1, \dots, K\}} p_k \prod_{r=1}^R \prod_{k_1=1}^K A^r(k_1, k)^{y_{k_1}^r}
 \end{aligned} \tag{1}$$

where  $y_{k_1}^r = 1$  iff  $y^r = k_1$ , and 0 otherwise.

However, we note that both majority-voting and the above method utilize only the top-ranked label from each evaluator. As was empirically indicated in the previous section, labels at other ranks also potentially possess significant information about the true hidden label. We next discuss two standard voting methods, which utilize the entire ranked list while estimating the best representative label.

### 3.1. Borda count and Schulze’s method

The Borda count is a voting method which requires each voter (or evaluator in our case) to give a complete ranked list of all possible categories [6]. The best candidate or option is given rank 1, the second best option rank 2 and so on. Next, the  $l^{th}$  ranked option is given

$M-l+1$  votes (where  $M$  is the size of the ranked list). For example, in a ranked list of 6 options, the first option would get 6 votes, the second one 5, and the last one 1. Next, the total votes of a given category are computed by adding up the votes which it received from all the evaluators. The category having the most number of votes is declared the winner. The Borda count method is widely used in various places, including certain political elections in Slovenia and selecting sports award winners in the United States (e.g. in Major League Baseball and NCAA). It must be noted that Borda count is susceptible to “strategic voting”, where the voters are divided into camps favoring particular candidates. In such situations, there is huge variability in the number of votes received by these favored candidates from one ranked list to another, and as a consequence, the mediocre candidates emerge with higher vote counts.

Schulze’s method [7] works by considering pairwise preferences among voters for various candidates. The first step is the computation of the pairwise preference matrix,  $D$ . The  $(k_1, k_2)$  entry of this matrix denotes the number of voters which prefer candidate  $k_1$  to  $k_2$ . The entries of this matrix can be visualized as edge weights of a fully connected, directed graph between all candidate nodes. A chain from candidate  $k_1$  to  $k_2$  is defined as an ordered set of candidate nodes  $(c_1 = k_1, \dots, c_n = k_2)$  with the property that  $\forall i \in \{1, \dots, n-1\}, D(c_i, c_{i+1}) > D(c_{i+1}, c_i)$ . The “strength” of the path  $(c_1, \dots, c_n)$  is defined to be the weight of the weakest link in the path. That is, it is defined as under:

$$p(c_1 = k_1, c_2, \dots, c_n = k_2) = \min_{i \in \{1, \dots, n-1\}} D(c_i, c_{i+1}) \quad (2)$$

The “strength of the strongest path” between two candidate nodes  $(k_1, k_2)$  is defined to be the maximum of the strength of all paths connecting these two nodes.

$$p_S(k_1, k_2) = \max_{(c_2, \dots, c_{n-1})} p(c_1 = k_1, c_2, \dots, c_n = k_2) \quad (3)$$

The candidate  $k_1$  is said to have beaten  $k_2$  if  $p_S(k_1, k_2) > p_S(k_2, k_1)$ . Thus, a candidate  $k$  is declared to be the overall winner if  $p_S(k, k_j) > p_S(k_j, k) \forall k_j \neq k$ . Schulze’s method can be used to generate a ranked list of winners based on the number of candidates each candidate beats.

As we can observe from the above two voting methods, no notion of reliability of voters or evaluators is used. Hence, even though we are able to utilize the entire ranked list, equal emphasis is being given to the rankings by all evaluators. We hypothesize that presence of such reliability information while combining ranked lists can generate more representative labels. In the next subsection, we discuss a global evaluator reliability metric based on the model in [3], and use it to modify Borda count and Schulze’s method.

### 3.2. A global evaluator reliability metric

As we defined in a previous subsection, the diagonal entries of the matrix  $A^r$  (for the  $r^{th}$  evaluator),  $A^r(k, k)$ , denote the probability that the evaluator generated the label  $k$ , given that the true hidden label was also  $k$ . Thus, we define the global reliability metric for each evaluator to be the following:

$$\epsilon_r = P(\text{evaluator “}r\text{” reliable}) \quad (4)$$

$$= \sum_{k=1}^K P(\text{evaluator “}r\text{” reliable} | y = k) p_k = \sum_{k=1}^K p_k A^r(k, k)$$

Note that  $\epsilon_r \in [0, 1]$ , with more reliable evaluators having higher values of  $\epsilon_r$ . This reliability metric can be computed directly from

the estimated parameters of the reliability model from [3] presented in the previous subsection.

### 3.3. Reliability-weighted Borda count and Schulze’s method

As we noted previously, both ranked list voting methods weight the labels from all evaluators equally. We make heuristic modifications to both these methods, taking the global reliability  $\epsilon_r$  into account. We first square the reliability values, and then normalize their sum to 1. The square is taken to increase the differences between numerically close values, and could be replaced by exponentiation of any order greater than 1 based on cross-validation. Thus, we define the following reliability-dependent weights for the  $R$  evaluators:

$$w^r = \frac{\epsilon_r^2}{\sum_{r=1}^R \epsilon_r^2} \quad \forall r \in \{1, \dots, R\} \quad (5)$$

In the reliability-weighted version of Borda count, instead of adding the votes given to a particular category (or label) by all evaluators, we first multiply the vote counts by the weight of the individual evaluators. Addition of these “reliability-weighted” votes gives us the final count of votes for a given category. Similarly for Schulze’s method, while computing the entry  $D(k_1, k_2)$  of the pairwise preference matrix, we weight each vote in preference of option  $k_1$  as compared to  $k_2$  by the weight of the evaluator who gave that vote.

## 4. EXPERIMENTS AND RESULTS

We first conducted experiments to evaluate the ability of each of the label combination methods to predict the label of acted emotion, assuming it to be the ground truth. Table 1 gives this accuracy, averaged over 30-fold cross validation (29 folds used for training of the method in [3]), and the p-value computed using Mann-Whitney’s signed rank test. Clearly, all methods perform significantly better than majority voting at the 5% level. In addition, all methods except majority voting perform equally well.

Method	Accuracy	p-value
Majority-voting	75.7	reference
Method in [3]	78.8 (4.1)	0.006
Borda count	77.8 (2.8)	0.02
Schulze’s method	77.6 (2.5)	0.02
Modified Borda count	77.5 (2.4)	0.03
Modified Schulze’s method	77.4 (2.2)	0.04

**Table 1.** Accuracy of various algorithms in estimating the label of the acted emotion. Figures in brackets show percentage improvement over majority-voting.

Based on Table 1, it appears that using the entire ranked list does not improve the estimation of the true label as compared to [3]. However, we realize that the ground truth label in the database is the identity of the emotion which the speaker was asked to synthesize in his speech. This fact does not make it the representative emotion present in the audio. As we noted earlier, the audio could contain a mix of several emotions. This makes the available ground truth label itself unreliable. Thus, accuracy of the estimated labels with respect to the given ground truth may not be a good measure for evaluation. Hence, as a further experiment, we perform emotion classification using the estimated label as the ground truth. The entire database

was randomly divided into two halves. One half was used for training the algorithm in [3], and the estimated labels from all the algorithms were computed on the other half.

We extracted 13 Mel Filter Bank (MFB) coefficients from each audio clip over 25 ms frames with a 10 ms shift. The component-wise mean and standard deviation of this vector were computed over the entire utterance, resulting in two 13-dimensional feature vectors. Discrete cosine transform (DCT) was then applied independently on each of these vectors, and the top 2 dimensions were selected. Concatenation of these vectors resulted in a 4-dimensional feature vector for each audio clip.

Next, we trained a multinomial logistic regression classifier in Weka [8] using the labels generated by each of the label combination methods (on second half of the database), with 10-fold cross validation. The average classification accuracies are reported in Table 2. Three observations can be made immediately. First, the ranked list combination methods heavily out-perform majority voting and the method presented in [3]. Second, among the ranked list methods, the reliability-weighted versions proposed in this paper perform better than the standard variants. Both these observations suggest that using the entire ranked list, and introducing the notion of evaluator reliability while combining them, gives emotional labels that are more representative of the audio, leading to an improvement in the classification accuracy. The final observation we make is the slightly worse performance of the method in [3] as compared to the majority voting based combination. We recall from Table 1 that majority voting performed worse in estimating the true label given in the database. This performance reversal suggests that accuracy in estimating the true labels present in the database might not be a robust method to evaluate the relative merits of the various approaches.

Method	Accuracy	Weighted F-measure
Majority-voting	32.4	0.293
Method in [3]	30.1 (-7.1)	0.279 (-4.8)
Borda count	50.8 (56.8)	0.499 (70.3)
Schulze's method	50.7 (56.5)	0.505 (72.4)
Modified Borda count	52.7 (62.6)	0.515 (75.8)
Modified Schulze's method	53.8 (66.0)	0.535 (82.6)

**Table 2.** Emotion classification accuracy and F-measure using labels estimated by different algorithms. Figures in brackets show percentage improvement over majority-voting

## 5. CONCLUSION AND FUTURE WORK

This paper presented the idea that asking evaluators to rank emotions from a set of possible labels according to their applicability to a given audio clip can be beneficial to emotion recognition performance. This idea is based on the hypothesis that the various emotion description categories are ambiguous (ambiguity in perception), and a given audio clip does not typically contain a single emotion (non-prototypicality in production). Asking evaluators to rank emotions in an audio clip, followed by a combination of these ranked lists seems to be an effective way to arrive at a more representative emotional class, as compared to methods based on just the top-ranked emotions. We use two well-known ranked list voting methods - Borda count and Schulze's method, and compare them with majority voting and the method presented in [3] on the top ranked label. We also

propose evaluator reliability-weighted versions of the two ranked list combination methods, where evaluator reliability is computed using the result of [3]. While estimating the ground truth label given in the database, we find that majority voting performs the worst, and all other algorithms perform equally well. Emotion classification experiments using the estimated ground truth labels show the reliability weighted versions of the two ranked list voting methods performing the best, followed by their standard version. We interestingly observe that here, majority voting performs slightly better than the method from [3], which indicates that the ground truth labels given in the database may themselves be noisy and not representative of the emotional content of the audio.

There are several directions for future work. First, we have incorporated evaluator reliability information in the two ranked list combination methods in a heuristic fashion. It would be interesting and useful to adopt a more theoretical approach, and design an algorithm which incorporates evaluator reliability at a more fundamental level. For example, it would be interesting to investigate and design models of how evaluators generate these ranked lists. We can also investigate the applicability of more sophisticated evaluator models such as the one presented in [5]. We would like to investigate ways in which the representativeness of a given class label given the data can be quantified. Finally, it would be interesting to test the ideas presented in this paper on a larger emotion recognition database, and other subjective pattern classification problems in general.

## 6. ACKNOWLEDGEMENT

The authors would like to thank Mr. Jangwon Kim and Dr. Sungbok Lee (members of SAIL at USC) for organizing the data collection.

## 7. REFERENCES

- [1] M. Wollmer, F. Ebyen, S. Reiter, B. Schuller, C. Cox, E. Douglas-Cowie, and R. Cowie, "Abandoning emotional classes - Towards continuous emotion recognition with modeling of long-range dependencies," in *Proceedings of InterSpeech*. ISCA, 2008, pp. 597–600.
- [2] E. Mower, M. J. Mataric, and S. Narayanan, "Evaluating evaluators: A case study in understanding the benefits and pitfalls of multi-evaluator modeling," in *Proceedings of InterSpeech*. ISCA, 2009, pp. 1583–1586.
- [3] P. Smyth, U. Fayyad, M. Burl, P. Perona, and P. Baldi, "Inferring ground truth from subjective labeling of Venus images," in *Advances in Neural Information Processing Systems*, 1995, pp. 1085–1092.
- [4] V. C. Raykar, S. Yu, L. H. Zhao, A. Jerebko, C. Florin, G. H. Valadez, L. Bogoni, and L. Moy, "Supervised learning from multiple experts: Whom to trust when everyone lies a bit," in *Proceedings of ICML*, 2009.
- [5] K. Audhkhasi and S. S. Narayanan, "Data-dependent evaluator modeling and its application to emotional valence classification from speech," in *Proceedings of InterSpeech*, 2010.
- [6] D. Black, "The theory of committees and elections," *Cambridge University Press*, 1968.
- [7] M. Schulze, "A new monotonic and clone-independent single-winner election method," *Voting Matters*, vol. 17, October 2003.
- [8] I. H. Witten and E. Frank, "Data Mining: Practical machine learning tools and techniques," *San Francisco: Morgan-Kaufmann*, 2005.