Formulating Emotion Perception as a Probabilistic Model with Application to Categorical Emotion Classification

Reza Lotfian
The University of Texas at Dallas
Email: rxl099220@utdallas.edu

Carlos Busso
The University of Texas at Dallas
Email: busso@utdallas.edu

***Abstract*—Automatic recognition of emotions is an important part of affect-sensitive *human-computer interaction* (HCI). Expressive behaviors tend to be ambiguous with blended emotions during natural spontaneous conversations. Therefore, evaluators disagree on the perceived emotion, assigning multiple emotional classes to the same stimuli (e.g., sadness, anger, surprise). These observations have clear implications on emotion classification, where assigning a single descriptor per stimuli oversimplifies the intrinsic subjectivity in emotion perception. This study proposes a new formulation, where the emotional perception of a stimuli is a multidimensional Gaussian random variable with an unobserved distribution. Each dimension corresponds to an emotion characterized by a numerical scale. The covariance matrix of this distribution captures the intrinsic dependencies between different emotional categories. The process where an evaluator judges the stimuli is equivalent to sampling a point from this distribution, reporting the class with the highest value. The proposed approach recursively estimates this multimodal distribution using numerical methods. The mean of the Gaussian distribution is used as a soft label to train a *deep neural network* (DNN). Our experimental results show that the proposed training method leads to improvements in F-score over training with (1) hard-labels based on majority vote, and (2) soft-label framework proposed by other studies.**

1. Introduction

Recognizing categorical emotions such as happiness, sadness or anger from speech can have many practical applications [1], [2], [3]. Previous studies on categorical emotion recognition rely on the assumption that (1) each speech segment is often assigned to one emotional class, and (2) samples belonging to the same class share similar acoustic features. However, the boundaries between emotion classes during human interaction are ambiguous, where people rarely express extreme emotions. Samples labeled with a given emotional class may have different acoustic characteristics (e.g., shades of happiness). These observations have direct implication in machine learning frameworks for emotion recognition, where conventional approaches do not generalize when evaluated in realistic scenarios [4].

The emotional annotation process of spontaneous databases often include perceptual evaluations, where several evaluators are asked to assign an emotional class to each sample. Then, a consensus labels, such as majority vote, is calculated and assigned as ground truth. This approach aims to identify popular trends where outlier assessments are ignored. While annotators may report conflicting categories, all of them might be right, given the differences in perception and ambiguity in the expressed emotions. Emotions are driven by appraisal of a given situation [5]. Depending on the evaluator's perspective, multiple answers may be appropriate, especially if many related emotional categories are available (e.g., surprise, anger, fear, disgust). We hypothesize that individual evaluations provide richer emotional descriptions than simple consensus labels, which should be effectively leveraged while training classifiers [6], [7], [8]. The formulation of the classifier should understand the relationship between emotional classes, prioritizing the separation of unrelated categories (e.g., anger versus happiness) over related emotions (anger versus disgust) [9].

This study proposes an innovative formulation to leverage individual evaluations for emotion recognition, as opposed to consensus labels, considering the intrinsic relationship between emotional categories. We formulate emotion perception as a probabilistic model, where each speech segment has a non-observable multivariate Gaussian distribution. Each dimension is associated with the intensity measure of one emotion category (i.e., dimension equals the number of emotions). Emotional perception is equivalent to sampling a point according to this multivariate distribution. After listening to the stimuli, the evaluator draws a point in the distribution, reporting the emotional category that has the largest intensity measure. Each individual evaluation corresponds to a realization drawn from this distribution, and the task is to approximate this multivariate Gaussian distribution using numerical methods. We demonstrate that this formulation can be used in emotion recognition by (1) training the classifiers using the expected value of each emotional dimension, derived from random Gaussian distribution of individual samples (e.g., soft labels), and (2) modifying the loss function to weigh errors between emotional class according to their relations. We refer to this method as *soft label from the expected intensity of emotion* (SL-EIE). The proposed method achieves better performance than classifiers trained with hard labels derived from majority vote and soft-label proposed by Fayek et al. [7].

978-1-5386-0563-9/17/$31.00

2. Related work

Unlike other speech processing problems such as speaker identification or *automatic speech recognition* (ASR), defining ground-truth labels for speech emotion recognition is not straightforward. Emotions observed during human interactions are ambiguous, which are not necessarily captured by the target emotional descriptions or the annotation protocol (e.g., forcing an evaluator to choose a class from a predefined list [10]). The standard approach is to collect multiple annotations from experts or naïve evaluators, aiming to reach a consensus label. Given subjective nature of perceptual evaluations, the resulting ground-truth labels are noisy [9], [11].

When the labels are inherently noisy, studies have proposed using ensembles [12] and soft-labels [13], describing the probability or intensity associated with each class. In speech emotion recognition, studies have proposed different variants of this approach, leveraging the individual evaluations provided to each speech sample (i.e., not just the consensus label). Mower et al. [14], [15] suggested emotional profiles to describe the confidence that an emotional label is assigned to an utterance. Audhkhasi and Narayanan [8] proposed fuzzy logic to deal with ambiguous emotional utterances by generalizing class sets with partial membership, rather than single class membership. The study suggested that more information can be extracted from individual annotations assigned to a given speech sample than a consensus label. Other studies have also explored the use of individual annotations to create relative labels of categorical emotions for preference learning [6]. The study derived a probabilistic relevant score from individual annotations, indicating the intensity of a given emotional class (e.g., intensity of happiness in a speech segment). The approach was successfully used to train emotional rankers.

Fayek et al. [7] used individual evaluations to create ensembles with *deep neural network* (DNN). Each DNN is trained with the annotations from one annotator, providing the final score after fusing the ensembles. Alternatively, the study trained a single DNN with soft labels reflecting the proportion of evaluators selecting each class. Their experimental results show the benefit of considering individual evaluations in building emotion classifiers. This study is related to our paper, so we use this approach as one of the baselines. Our proposed method has two main contributions: (1) we estimate soft labels using the expected intensity of emotion after estimating a multivariate distribution representing the perceptual evaluation process, (2) we consider the intrinsic relationship between different emotional categories.

3. Methodology

During perceptual evaluations, raters are asked to listen to the stimuli and annotate their perceived emotion. In practice, listeners often recognize more than one emotion, especially if the number of choices is long including closely related emotions (e.g., excitement and happiness). We hypothesize that there is an unobservable intensity level associated with each emotional category for every speech sample. The evaluator selects the class with higher intensity according to his/her judgement. Cases when two annotators provide conflicting emotions for a speech segment may indicate that they perceive emotions with different intensity level. Therefore, the selected class does not imply complete disagreement. We propose to estimate *soft label from the expected intensity of emotion* (SL-EIE), which we estimate from the individual evaluations.

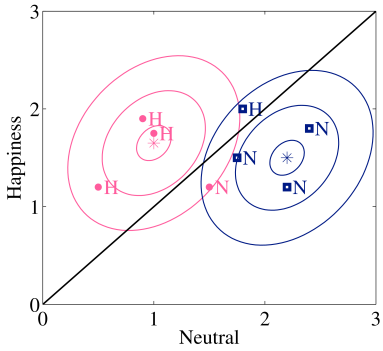

Figure 1. Illustration of approach to formulate emotion perception as a probabilistic process. Unobservable distributions of the intensity of happiness and neutral for two sentences, evaluated by four evaluators.

We formulate the problem by modeling perceptual evaluation as a random vector, where each dimension corresponds to one emotional class. The multivariate distribution of this process is hidden and the task is to estimate it from the individual evaluations. Figure 1 illustrates the key ideas when we consider only two emotions (*happiness* and *neutral*) for two speech segments (red and blue). The x-axis corresponds to the perceived intensity for *neutral*, and the y-axis corresponds to the perceived intensity for *happiness*. Each segment is evaluated by four annotators. This process is formulated as drawing four points from each of these distributions (red points for sentence 1, blue squares for sentence 2). Based on the values for x and y, the evaluator chooses *neutral* if $x > y$ or *happiness* if $y > x$. The resulting labels for these eight evaluations are denoted with *H* for *happiness* and *N* for *neutral*. Notice that the line $y = x$ defines the boundaries between the emotions. Majority vote will suggest that one sample is happy and the other is neutral. Instead, we aim to estimate the unobservable distribution from the annotations and use the expected value of the random vector to define soft-labels to train our classifiers. We assume that the covariance matrix of the distributions is independent of the individual speech samples and only depends on the relation between emotions. This formulation explains why some classes are regularly interchanged (e.g., disgust and anger) and others are clearly separated (e.g., happiness and anger).

3.1. Database

This study relies on the MSP-PODCAST emotional speech corpus collected at the University of Texas at Dallas [16]. The database includes a large set of speech segments from podcast recordings available in audio sharing websites. The podcasts are selected from various topics including

politics, sports, talk shows, and movies, therefore, conveying a broader range of emotions. Podcasts are segmented into speech turns using speaker diarization tools. We implement an automatic process that selects only speech segments with one speaker, without noise, background music, or phone quality audio. The duration of the segments are between 2.75s and 11s. Following the ideas described in Mariooryad et al. [17], we use different machine learning formulations to retrieve segments from the pool of available segments conveying target emotion behaviors. These segments are then emotionally annotated. The data collection process is an ongoing effort, where the current study uses 13,432 speech segments (21 hrs, 15 min). We have manually annotated the speaker identity of 147 speakers (9,670 segments). We use segments from 50 speakers as our test set (4,283 segments), and data from 10 speakers as our development set (1,860 segments). The train set includes the rest of the corpus (7,289 segments). This partition attempts to create speaker independent datasets for train, test and development sets.

The speech segments are emotionally annotated using an improved version of the crowdsourcing method proposed by Burmania et al. [18] (see details on Lotfian and Busso [16]). Within the perceptual evaluation, the raters are asked to choose the primary emotion from anger, sadness, happiness, surprised, fear, disgust, contempt, and neutral. They can also choose other if none of the previous labels are suitable. The current version of the corpus is not balanced across emotional classes, with many happy and neutral sentences and very few fear sentences. Therefore, we remove the label fear for this study (we replace individual annotations for "fear" by "other"). While not used in this study, the annotation also includes secondary emotions and emotional attributes (arousal, valence and dominance). Each speech segment is annotated by at least five annotators.

3.2. Theoretical Framework

The following derivation is done for each sample. We model the perceived intensity for a given stimuli with vector $\boldsymbol{x} = [x_1, x_2, ..., x_D]^T$, where each dimension corresponds to one of the D emotions. Vector $\boldsymbol{x}$ is a realization of a random vector $\boldsymbol{X}$ with multivariate Gaussian distribution:

$$p(x) = \frac{1}{\sqrt{(2\pi)^D|\boldsymbol{\Sigma}|}} \exp[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})^T] \doteq \mathcal{N}(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma}) \quad (1)$$

where $\boldsymbol{\mu}$ is the mean vector and $\boldsymbol{\Sigma}$ is the covariance matrix. We estimate the probability of an annotator selecting each class with the vector $\boldsymbol{p} = [p_1, p_2, ..., p_D]^T$, where p_j is the proportion of evaluators selecting emotion j (this is the soft label vector proposed by Fayek et al. [7]).

An annotator selecting class j is equivalent to the event $x_j > x_i, \quad \forall i \neq j$. This probability can be estimated from the multivariate Gaussian distribution:

$$p_j = \boldsymbol{P}(X_1 \leq X_j, ..., X_{j-1} \leq X_j, X_{j+1} \leq X_j, ..., X_D \leq X_j) \quad (2)$$

where X_j is the jth element of the random vector $\boldsymbol{X}$. Given the distribution of $\boldsymbol{X}$ (Eq. 1), p_j can be estimated as

$$p_j = \int_{-\infty}^{\infty}\int_{-\infty}^{x_j}...\int_{-\infty}^{x_j} \frac{1}{\sqrt{(2\pi)^D|\boldsymbol{\Sigma}|}} \exp[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})]dx_1...dx_{j-1}dx_{j+1}...dx_D dx_j \quad (3)$$

To take advantage of numerical methods available to calculate the *cumulative distribution function* (CDF) of a Gaussian distribution, we convert Equation 3 by defining $\boldsymbol{Y}$ such that,

$$\begin{aligned} Y_i &= X_i - X_j, \qquad \forall i \neq j \\ Y_j &= X_j \end{aligned} \quad (4)$$

Therefore, $\boldsymbol{Y} = \boldsymbol{H}_j\boldsymbol{X}$,

$$\boldsymbol{Y} = \begin{array}{c} \\ j \end{array} \overset{\begin{array}{cccccc} & & & j & & \end{array}}{\begin{pmatrix} 1 & 0 & \dots & -1 & \dots & 0 \\ 0 & 1 & \dots & -1 & \dots & 0 \\ \vdots & \vdots & & \ddots & & \vdots \\ 0 & 0 & \dots & 1 & \dots & 0 \\ \vdots & \vdots & & \ddots & & \vdots \\ 0 & 0 & \dots & -1 & \dots & 1 \end{pmatrix}} \boldsymbol{X} \quad (5)$$

Since $\boldsymbol{Y}$ is a linear transformation of a multivariate Gaussian random vector, $\boldsymbol{y}$ is also a multivariate Gaussian random vector: $\mathcal{N}(\boldsymbol{y}; \boldsymbol{H}_j\boldsymbol{\mu}, \boldsymbol{H}_j\boldsymbol{\Sigma}\boldsymbol{H}_j^T)$. Therefore, p_j in equation 2 can be estimated with

$$p_j = P(Y_1 \leq 0, \dots, Y_{j-1} \leq 0, Y_j < \infty, Y_{j+1} \leq 0, \dots, Y_D \leq 0) \quad (6)$$

The next step is to find an estimate for the parameter $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ that satisfies the vector $\boldsymbol{p}$ estimated from the annotations for all probabilities p_j greater that zero:

$$\begin{aligned} &\forall j; \quad p_j > 0 \\ &P(Y_1 \leq 0, \dots Y_j < \infty, \dots, Y_D \leq 0) = p_j \end{aligned} \quad (7)$$

Notice that adding a constant to $\boldsymbol{x}$ that satisfies Equation 2 is also an acceptable solution. For consistency, the intensity vectors $\boldsymbol{x}$ across sentences should be in the same range. Therefore, we add an extra constraint where the intensity of neutral class is always equal to 1, serving as a reference.

First, we estimate the covariance matrix $\boldsymbol{\Sigma} = E[(\boldsymbol{X}-\boldsymbol{\mu})(\boldsymbol{X}-\boldsymbol{\mu})^T]$. Since the intensity vectors $\boldsymbol{x}$ is an unobserved variable, we approximate $\boldsymbol{\Sigma}$ with $\boldsymbol{p}$ using the subjective evaluations. Estimating $\boldsymbol{\Sigma}$ from $\boldsymbol{p}$ will produce a covariance matrix that is different from the covariance matrix of $\boldsymbol{x}$. However, this approach captures the dependencies between different emotional categories, which is the key information provided by $\boldsymbol{\Sigma}$. Before estimating $\boldsymbol{\Sigma}$, we normalize the vector $\boldsymbol{p}$ using a two-step approach. First, we multiply $\boldsymbol{p}$ by a constant such that the mean for *neutral* is equal to 1. Then, we subtract the mean vector of $\boldsymbol{p}$ to create a zero-mean variable $\hat{\boldsymbol{p}}$. If we have n sentences, $\boldsymbol{\Sigma}$ can be estimated with:

TABLE 1. COVARIANCE MATRIX BETWEEN DISTRIBUTION OF EMOTIONS ($\widetilde{\boldsymbol{\Sigma}}$). THE MATRIX IS ESTIMATED WITH $\boldsymbol{p}$ IN EQUATION 8.

	ANG	SAD	HAP	SUR	DIS	CON	NEU
ANG	0.24	-0.02	-0.11	-0.02	0.04	0.03	-0.15
SAD	-0.02	0.13	-0.06	-0.01	-0.01	-0.01	-0.01
HAP	-0.11	-0.06	0.68	-0.02	-0.10	-0.11	-0.25
SUR	-0.02	-0.01	-0.02	0.16	-0.01	-0.02	-0.09
DIS	0.04	-0.01	-0.10	-0.01	0.18	0.03	-0.12
CON	0.03	-0.01	-0.11	-0.02	0.03	0.20	-0.10
NEU	-0.15	-0.00	-0.25	-0.09	-0.12	-0.10	0.79

ANG: Anger, SAD: Sadness, HAP: Happiness, SUR: Surprised, DIS: Disgust, CON: Content, NE: Neutral

$$\widetilde{\boldsymbol{\Sigma}} = \begin{bmatrix} \hat{\boldsymbol{p}}_{(1)}, & \hat{\boldsymbol{p}}_{(2)}, & \ldots, & \hat{\boldsymbol{p}}_{(n)} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{p}}_{(1)}^T \\ \hat{\boldsymbol{p}}_{(2)}^T \\ \vdots \\ \hat{\boldsymbol{p}}_{(n)}^T \end{bmatrix} \tag{8}$$

where $\hat{\boldsymbol{p}}_{(i)}$ is the normalized version of $\boldsymbol{p}$ for the training sample i. We assume $\widetilde{\boldsymbol{\Sigma}}$ to be independent of the specific speech sentence, so we can use the same covariance matrix across all the samples in the train and test sets. Table 1 shows the estimated values for the training set.

The next step is to estimate the mean vector $\boldsymbol{\mu}$ using numerical methods. We iteratively adjust the entries of $\boldsymbol{\mu}$, one at a time, to satisfy Equation 2 for each emotional category. Since the soft-labels p_i are derived from a limited number of annotators (e.g., five), some of the emotions may not be selected producing zero entries in $\boldsymbol{p}$. If $p_i = 0$ (e.g., class i was not selected by the annotators), then Equation 3 cannot be satisfied. To address this problem, we adjust the entries of vector $\boldsymbol{p}$ by estimating the probability of classes not selected due to limited evaluators, which we denoted λ. Intuitively, this probability decreases as we increase the number of evaluators. We experimentally estimate λ as a function of the number of annotations available for an individual sample. This is an approximation since λ can also depend on other parameters (e.g., emotion, agreement level). As part of the protocol to increase quality used in the crowdsourcing evaluation, some sentences in the MSP-PODCAST corpus were evaluated by more than five annotators. To estimate $\lambda(n)$ (i.e., λ when a sample is evaluated by n raters), we consider all the sentences with more than n annotations. At random, we select n of them. Then, we select at random another sample, evaluating whether the emotional label of this additional sample was already selected by the first n evaluators. Figure 2 shows the value for $\lambda(n)$, which is the proportion of cases when the label of the additional evaluator is new. The dashed line shows the probability of an unseen label if the labels are drawn randomly proportional to the population of each label in the entire database. This line shows the subjective nature of the emotion evaluations. We adjust $\boldsymbol{p}$ such that $\sum_j \tilde{p}_j = 1 - \lambda(n)$.

Algorithm 1 shows the procedure to estimate the expected intensity vector $\boldsymbol{\mu}$ for an individual speech sample. We start with an initial vector $\boldsymbol{\mu}_{initial}$. For emotional class j with $p_j > 0$, we adjust μ_j to satisfy Equation 2 for $\tilde{p}_j$ (i.e.,p_j after adjusting for $\lambda(n)$). We repeat this process for all the emotional classes selected by the evaluators. We iteratively repeat this process k times since the dimensions are not independent. After the iterations, we normalize the estimated values such that the intensity for *neutral* is one (i.e., $\tilde{\mu}_{neutral} = 1$). This step brings the intensity of different speech sample in the same scale.

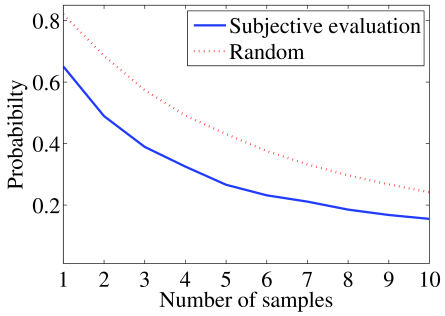

Figure 2. Estimation of $\lambda(n)$, the probability that an extra evaluator will select an emotion not selected by the previous n evaluators.

TABLE 2. THE EXPECTED VALUE OF A LABEL $\boldsymbol{p}$ AND EXPECTED VALUE OF THE INTENSITY OF CATEGORIES $\boldsymbol{w}$ FOR THE MSP-PODCAST DATABASE

	ANG	SAD	HAP	SUR	DIS	CON	NEU
$\boldsymbol{p}$	0.08	0.05	0.20	0.07	0.08	0.11	0.33
$\boldsymbol{\mu}_{initial}$	-0.07	0.30	-0.75	0.26	0.39	0.41	1.00

ANG: Anger, SAD: Sadness, HAP: Happiness, SUR: Surprised, DIS: Disgust, CON: Content, NE: Neutral

The initial vector $\boldsymbol{\mu}_{initial}$ is set using algorithm 1, where the input vector for the probabilities consists of the prior of each class in the MSP-PODCAST corpus. Since all the emotional classes are included, $\lambda = 0$. Table 2 shows the prior probability of each emotion and the corresponding initial intensity values $\boldsymbol{\mu}_{initial}$.

3.3. Loss Function

A key aspect of training a DNN is to define an objective function that effectively measures the disagreement between the ground truth and predicted labels. The objective is to minimize the value of this function. Previous studies have

Algorithm 1 Estimation of $\boldsymbol{\mu}$ per each sentence

Input:
$\boldsymbol{p}$: Probability of classes
$\widetilde{\boldsymbol{\Sigma}}$: Covariance matrix
k: Number of iterations
$\lambda(n)$: Probability of unseen labels
$\boldsymbol{\mu}_{initial}$: initial intensity vector
Output: Estimated emotion intensity $\tilde{\boldsymbol{\mu}}$
$\hat{\boldsymbol{p}} \leftarrow \boldsymbol{p}/(1-\lambda(n))$
$\tilde{\boldsymbol{\mu}} \leftarrow \boldsymbol{\mu}_{initial}$
for $i <= k$ **do**
  for j **where** $p_j > 0$ **do**
    find $\overline{\mu}_j$ such that $F(0,0,\ldots,\infty,\ldots,0) = \tilde{p}_j$ for $\mathcal{N}(\boldsymbol{x}; \boldsymbol{H_j}[\tilde{\mu}_0,\ldots,\tilde{\mu}_{j-1},\overline{\mu}_j,\ldots,\tilde{\mu}_D]^T, \boldsymbol{H_j}\widetilde{\boldsymbol{\Sigma}}\boldsymbol{H_j}^T)$
    $\tilde{\mu}_j \leftarrow \overline{\mu}_j$
for $j = 1 : D$ **do**
  $\tilde{\mu}_j \leftarrow \tilde{\mu}_j - \tilde{\mu}_{neutral} + 1$

used categorical cross-entropy as the loss function to train DNN with either hard labels [7], [19] or soft-labels [7]. We define our cost function based on the Mahalanobis distance to reflect a more meaningful measure of cost for disagreements between the predictions and subjective evaluations assigned to test speech. The loss function is defines as:

$$\mathcal{L}(\boldsymbol{\theta}, \tilde{\boldsymbol{\mu}}) = 1 - \exp[-\frac{1}{2}(\boldsymbol{\theta} - \tilde{\boldsymbol{\mu}})\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta} - \tilde{\boldsymbol{\mu}})^T] \quad (9)$$

where θ is the intensity value predicted by the network. The key feature of this loss function is that there is a smaller penalty for confusing related emotions, as dictated by the covariance matrix. For example, the penalty for confusing *anger* and *happiness* is larger than the penalty of confusing *angry* and *disgust*. This property is desirable from the application perspective.

4. Experimental Evaluations

The performance of the proposed SL-EIE is evaluated using DNN with two fully connected hidden layers. Each hidden layer consists of 512 nodes with *rectified linear unit* (ReLU). We use dropout with 20% drop rate for the hidden layers to avoid overfitting. The output is a softmax layer with one output per emotion, with the exception of *neutral* which is always equal to 1. The predicted class corresponds to the emotional category with the higher output value. If this value is less than 1, the network selects the class *neutral*. We use the loss function defined in Equation 9. The network is trained with 50 epochs without early stopping. The parameters of the networks are optimized maximizing the performance on the development set.

We consider two baseline systems using DNN with the same configuration used for our model (i.e., two hidden layers with 512 nodes). The first one corresponds to a DNN trained with labels derived with the majority vote rule. This method excludes samples from the training set without majority vote agreement or with label *other*. The second baseline is trained with soft-labels derived by estimating the proportion of evaluators selecting each emotional class (vector $\boldsymbol{p}$ in Sec. 3.2). Both baselines are trained with categorical cross-entropy as the loss function.

To evaluate the frameworks using a consistent and fair comparison, the ground truth labels for the test set are estimated with majority vote rule. Samples without majority vote agreement or with label *other* are removed from the test set. For each emotional class, we estimated the precision and recall rates. Then, we estimate the average precision and average recall across emotional classes (we report these metrics in Table 3). From these values, we estimate the F1-score rate. The F1-score is not affected by unbalanced classes. We consider a seven class problem (anger, sadness, happiness, surprised, disgust, contempt, and neutral).

4.1. Acoustic Features

The study relies on the extended version of the *Geneva minimalistic acoustic parameter set* (eGeMAPS) which contains 88 *high level descriptors* (HLDs) [20]. This set was carefully selected based on their performance in previous paralinguistic problems, and theoretical significance. Since the set has only 88 features, we do need feature selection, simplifying the evaluation, and facilitating the reproduction of the results by other research groups. The features are extracted with OpenSMILE [21].

TABLE 3. CLASSIFICATION PERFORMANCE OF PROPOSED SL-EIE METHOD ON THE MSP-PODCAST DATABASE.

	Rec. [%]	Pre.[%]	F1-score[%]
Human performance	38.2	41.1	39.6
Majority vote	25.7	24.2	24.9
Soft-label [7]	27.2	23.7	25.3
SL-EIE [proposed]	28.1	24.5	26.2
SL-EIE [cross-entropy as loss function]	27.6	24.6	26.0

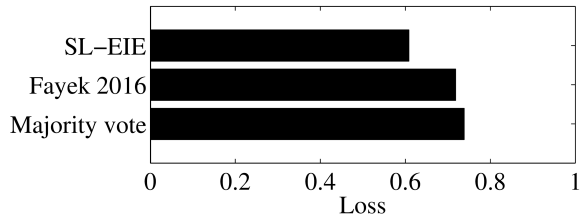

Figure 3. The error between estimated labels and ground-truth based on the loss function

4.2. Performance of Classification

Table 3 reports the results of the experimental evaluation, which demonstrates the benefits of using the proposed SL-EIE method. As a reference, the table also reports human performance on this task. We obtain this value by randomly removing one of the annotators for each sample in the test set. The annotators are compared with the consensus label obtained from the rest of the evaluators (i.e., majority vote after removing one evaluation). The F1-score is only 39.6% which reveal the complexity of this seven-class problem (chance performance is 14.3%). The recordings in the MSP-PODCAST corpus are from spontaneous emotional interactions, as opposed to prototypical behaviors. Note that the evaluators unconsciously use semantic information in the perception of emotions. Our DNN classifiers rely only on acoustic features.

The performance of the baseline systems are F1-score of 24.9% when we train with majority vote labels, and 25.3% when we train with the soft-label approach proposed by Fayek et al. [7] (i.e., using $\boldsymbol{p}$). Using the proposed SL-EIE based labels, we obtain a classification performance of 26.2% when we use cost function defined in Equation 9, and 26% when we use categorical cross-entropy as loss function. This result represents an improvement of 1.3% (majority vote) and 0.9% (soft-labels) over the baselines. The results reveal that (1) using soft-labels is an effective way of training speech emotion classifiers with DNN, and (2) the proposed SL-EIE based labels provide additional information, which results in improved performance, even when the testing labels are obtained with majority vote.

An alternative way of comparing the algorithms is by estimating the average loss $\mathcal{L}$ between the predicted and ground-truth labels for the test set using Equation 9. Notice

that the proposed model was trained to minimize this loss function. Figure 3 gives the average results for the two baseline systems and our proposed method. The SL-EIE method achieves the lowest loss, showing the advantages of the proposed soft-labels. The ANOVA test indicates that the differences are statistically significant, asserting significance at $p = 0.05$. Pairwise comparisons show that the difference between the three groups are all statistically different.

5. Conclusions

This study proposes a new framework to address the problem of classifying categorical emotions in spontaneous speech. The approach provides soft-labels inspired by the emotion perception process. Each sentence has a non-observable multivariate Gaussian distribution where the dimensions correspond to the emotional categories. In this formulation, assigning a label to the sample is equivalent to sampling a point in this distribution, where the selected category is the emotions with the highest intensity. This distribution captures the dependencies between emotional classes. The task consists of estimating this distribution from individual evaluations, reporting the expected intensities for each emotion as the soft labels. Using a spontaneous emotional database, we experimentally show the benefit of using this representation for training emotional classifiers.

There are several research directions to improve this framework. An assumption made in the derivation is the existence of a universal covariance matrix that is used for all the sentences. Likewise, we assume that the probability of unseen labels can be estimated only as a function of the number of evaluators. Future work should evaluate the benefit of alternative formulations. For example, the model can also consider the bias and reliability of individual evaluators to derive a better model for the emotion.

Acknowledgments

This work was funded by NSF CAREER award IIS-1453781. Microsoft supported the MSP-PODCAST corpus.

References

[1] C. G. Kohler, W. Bilker, M. Hagendoorn, R. E. Gur, and R. C. Gur, "Emotion recognition deficit in schizophrenia: association with symptomatology and cognition," *Biological Psychiatry*, vol. 48, no. 2, pp. 127–136, July 2000.

[2] C. A. Mazefsky and D. P. Oswald, "Emotion perception in Asperger's syndrome and high-functioning autism: the importance of diagnostic criteria and cue intensity," *Journal of Autism and Developmental Disorders*, vol. 37, no. 6, pp. 1086–1095, July 2007.

[3] S. D'Mello, S. Craig, B. Gholson, S. Franklin, R. Picard, and A. Graesser, "Integrating affect sensors in an intelligent tutoring system," in *Affective Interactions: The Computer in the Affective Loop Workshop at 2005 International Conference on Intelligent User Interfaces*, San Diego, CA, USA, January 2005, pp. 7–13.

[4] C. Busso, M. Bulut, and S. Narayanan, "Toward effective automatic recognition systems of emotion in speech," in *Social emotions in nature and artifact: emotions in human and human-computer interaction*, J. Gratch and S. Marsella, Eds. New York, NY, USA: Oxford University Press, November 2013, pp. 110–127.

[5] K. Scherer, "Appraisal theory," in *Handbook of cognition and emotion*, T. Dalgleish and J. Power, Eds. New York, NY, USA: John Wiley & Sons Ltd, March 1999, pp. 637–663.

[6] R. Lotfian and C. Busso, "Retrieving categorical emotions using a probabilistic framework to define preference learning samples," in *Interspeech 2016*, San Francisco, CA, USA, September 2016, pp. 490–494.

[7] H. M. Fayek, M. Lech, and L. Cavedon, "Modeling subjectiveness in emotion recognition with deep neural networks: Ensembles vs soft labels," in *International Joint Conference on Neural Networks (IJCNN 2016)*, Vancouver, BC, Canada, July 2016, pp. 566–570.

[8] K. Audhkhasi and S. S. Narayanan, "Emotion classification from speech using evaluator reliability-weighted combination of ranked lists," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2011)*, Prague, Czech Republic, May 2011, pp. 4956–4959.

[9] S. Steidl, M. Levit, A. Batliner, E. Nöth, and H. Niemann, ""Of all things the measure is man" automatic classification of emotions and inter-labeler consistency," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2005)*, vol. 1, Philadelphia, PA, USA, March 2005, pp. 317–320.

[10] J. A. Russell, "Forced-choice response format in the study of facial expression," *Motivation and Emotion*, vol. 17, no. 1, pp. 41–51, March 1993.

[11] L. Devillers, L. Vidrascu, and L. Lamel, "Challenges in real-life emotion annotation and machine learning based detection," *Neural Networks*, vol. 18, no. 4, pp. 407–422, May 2005.

[12] M. Abdelwahab and C. Busso, "Ensemble feature selection for domain adaptation in speech emotion recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2017)*, New Orleans, LA, USA, March 2017, pp. 5000–5004.

[13] C. Thiel, "Classification on soft labels is robust against label noise," in *Knowledge-Based Intelligent Information and Engineering Systems (KES 2008)*, ser. Lecture Notes in Computer Science, I. Lovrek, R. Howlett, and L. Jain, Eds. Zagreb, Croatia: Springer Berlin Heidelberg, September 2008, vol. 5177, pp. 65–73.

[14] E. Mower, A. Metallinou, C.-C. Lee, A. Kazemzadeh, C. Busso, S. Lee, and S. Narayanan, "Interpreting ambiguous emotional expressions," in *International Conference on Affective Computing and Intelligent Interaction (ACII 2009)*, Amsterdam, The Netherlands, September 2009, pp. 1–8.

[15] E. Mower, M. Mataric, and S. Narayanan, "A framework for automatic human emotion classification using emotional profiles," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 5, pp. 1057–1070, May 2011.

[16] R. Lotfian and C. Busso, "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings," *IEEE Transactions on Affective Computing*, vol. To appear, 2017.

[17] S. Mariooryad, R. Lotfian, and C. Busso, "Building a naturalistic emotional speech corpus by retrieving expressive behaviors from existing speech corpora," in *Interspeech 2014*, Singapore, September 2014, pp. 238–242.

[18] A. Burmania, S. Parthasarathy, and C. Busso, "Increasing the reliability of crowdsourcing evaluations using online quality assessment," *IEEE Transactions on Affective Computing*, vol. 7, no. 4, pp. 374–388, October-December 2016.

[19] D. Bertero and P. Fung, "A first look into a convolutional neural network for speech emotion detection," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2017)*, New Orleans, LA, USA, March 2017, pp. 5115–5119.

[20] F. Eyben, K. Scherer, B. Schuller, J. Sundberg, E. André, C. Busso, L. Devillers, J. Epps, P. Laukka, S. Narayanan, and K. Truong, "The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, April-June 2016.

[21] F. Eyben, M. Wöllmer, and B. Schuller, "OpenSMILE: the Munich versatile and fast open-source audio feature extractor," in *ACM International conference on Multimedia (MM 2010)*, Florence, Italy, October 2010, pp. 1459–1462.