INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Speech Emotion Recognition with Emotion-Pair based Framework Considering Emotion Distribution Information in Dimensional Emotion Space

Xi Ma[1,3], *Zhiyong Wu*[1,2,3], *Jia Jia*[1,3], *Mingxing Xu*[1,3], *Helen Meng*[1,2], *Lianhong Cai*[1,3]

[1]Tsinghua-CUHK Joint Research Center for Media Sciences, Technologies and Systems, Graduate School at Shenzhen, Tsinghua University, Shenzhen 518055, China
[2]Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong SAR, China
[3]Tsinghua National Laboratory for Information Science and Technology (TNList), Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China

max15@mails.tsinghua.edu.cn, {zywu,hmmeng}@se.cuhk.edu.hk, {jjia,xumx,clh-dcs}@tsinghua.edu.cn

Abstract

In this work, an emotion-pair based framework is proposed for speech emotion recognition, which constructs more discriminative feature subspaces for every two different emotions (emotion-pair) to generate more precise emotion bi-classification results. Furthermore, it is found that in the dimensional emotion space, the distances between some of the archetypal emotions are closer than the others. Motivated by this, a Naive Bayes classifier based decision fusion strategy is proposed, which aims at capturing such useful emotion distribution information in deciding the final emotion category for emotion recognition. We evaluated the classification framework on the USC IEMOCAP database. Experimental results demonstrate that the proposed method outperforms the hierarchical binary decision tree approach on both weighted accuracy (WA) and unweighted accuracy (UA). Moreover, our framework possesses the advantages that it can be fully automatically generated without empirical guidance and is easier to be parallelized.

Index Terms: speech emotion recognition, emotion-pair, dimensional emotion space, Naive Bayes classifier

1. Introduction

Emotion recognition plays an important role in many applications, especially in human-computer interaction systems that are increasingly common today. As one of the main communication media between human beings, voice has attracted wide attentions from researchers [1]. Speech contains a wealth of emotional information. How to extract such information from speech signal is of great importance for automatic speech emotion recognition.

As an important part of speech emotion recognition, the selection of feature subspace has attracted lot of research interests. Most of these researches are devoted to finding a common and global feature subspace that is most distinctive for all kinds of emotions. However, studies have already indicated that the features associated with different emotions are not exactly the same [2]. In other words, if we can divide the whole emotion space into several subspaces and find the features that are most distinguishable for each subspace separately, the emotion recognition performance on the whole space might be boosted. Motivated by this, we propose the emotion-pair method for emotion recognition by leveraging feature subspaces. The feature subspaces are first constructed for every two different emotions (emotion-pair); bi-classifiers are then used to distinguish the emotions for each emotion-pair from the feature subspaces; the final emotion recognition result is derived by the Naive Bayes classifier based decision fusion strategy. This decision strategy is motivated by the finding that, in the dimensional emotion space, the distances between some of the archetypal emotions are closer than the others. The proposed Naive Bayes classifier aims at capturing such useful information in deciding the final emotion category for emotion recognition.

The idea of this work is similar to the previous hierarchical binary decision tree approach [3]. However, our framework possesses the advantages of being able to be fully automatically generated without empirical guidance and to be easily parallelized. In the USC IEMOCAP database [4], we can achieve an unweighted accuracy (UA) of 62.54% using 10-folds (leave-one-speaker out) cross validation, which is a 4.08% absolute (6.98% relative) improvement over the hierarchical binary decision tree approach. The weighted accuracy (WA) on the same database is 57.85%, which also outperforms the hierarchical binary decision tree approach with 1.47% absolute (2.61% relative) improvement.

The reset of the paper is organized as follows. Section 2 summarizes the previous related work. The emotion-pair based speech emotion recognition framework is then detailed in Section 3. The proposed Naive Bayes classifier based decision fusion strategy is described in Section 4. Experiments and results are presented in Section 5. Section 6 concludes the paper.

2. Related Work

As a common issue for many classification problems [5], feature selection aims to pick a subset of features that are most relevant to the target concept [6] or to reduce the dimension of features for decreasing computational time as well as improving the performance [7]. There have been many studies on feature selection for speech emotion recognition. In [2, 8, 9], prosody-based acoustic features, including pitch-related, energy-related and timing features have been widely used for recognizing speech emotion. Spectral-based acoustic features also play important role in emotion recognition, such as Linear Prediction Coefficients (LPC) [10], Linear Prediction Cepstral Coefficients (LPCC) [11] and Mel-frequency Cepstral Coefficients (MFCC) [12]. In [13], voice quality features have also been shown to be related to emotions.

Besides manual selection, many automatic feature

http://dx.doi.org/10.21437/Interspeech.2017-619

selection algorithms have also been proposed. For example, Sequential Floating Forward Selection (SFFS) [14] is an iterative method that can find a subset of features near to the optimal one. Some evolutionary algorithms such as Genetic Algorithm (GA) [15] are often used in feature selection. Feature space transformation is another type of method, including Principal Component Analysis (PCA) [7], Neural Network (NN) [16] and so on.

To describe emotions, some studies have used a psychological dimensional space such as the 3-dimensional valence-activation-dominance model and the 2-dimensional valence-activation model [17]. Besides, discrete emotion labels, the so-called archetypal emotions [18], are commonly used in speech emotion recognition. Different archetypal emotions are located at different locations in the dimensional space. Some hierarchical decision frameworks [3, 19, 20] have been proposed to classify the speech emotions, whose ideas are based on the dimensional space.

Speech emotion recognition with multiple archetypal emotions is a multi-class problem. To derive the final result, [21] has mentioned several ways to reformulate the multi-class problem to multiple binary classification problems with the decision fusion strategy [22, 23]. The common way for decision fusion is the majority voting method. However, the voting method may encounter the equal voting problem and completely ignore the relationship between emotions. Our method proposes to use the Naive Bayes classifier that aims at capturing the distance information of different emotions in the dimensional emotion space for decision fusion.

3. Emotion-Pair based Speech Emotion Recognition

Our study is based on archetypal emotions. The emotion-pair is composed of two different kinds of archetypal emotions, such as Angry and Happy. For all possible combinations of emotion-pairs, the bi-classification is used to distinguish the two emotions in each emotion-pairs. Naive Bayes classifier based decision fusion is then adopted to derive the final emotion recognition result. As shown in Figure 1, the whole method involves four steps: feature extraction, feature selection, emotion bi-classification for each emotion-pair and Naive Bayes classifier based decision fusion. This section introduces the first three steps and the Naive Bayes classifier based decision fusion will be described in the next section. For comparison, we use the same feature set, feature selection algorithm and bi-classifiers as those used in [3].

3.1. Feature Extraction

The INTERSPEECH 2009 Emotion Challenge feature set is used in our experiment. We extracted these features using the OpenSmile toolbox [24]. The feature set includes 16 low level descriptors consisting of prosodic, spectral envelope, and voice quality features. 12 statistical functionals are then computed for every low level descriptor per utterance in the USC IEMOCAP database, including mean, standard deviation, kurtosis, skewness, minimum, maximum, relative position, range, two linear regression coefficients, and their respective mean square error. This results in a collection of 384 acoustic features.

3.2. Feature Selection

We normalized features using z-normalization with respect to the neutral utterances in the training dataset. The process has underlying assumption that the average characteristics of neutral utterances across speakers do not vary extensively. Therefore, the testing examples' features are z-normalized with respect to the mean and variance of neutral utterances from training data. The normalization allows us to use acoustic features across multiple different speakers and to eliminate the effect of variations in individual speakers' speaking characteristics.

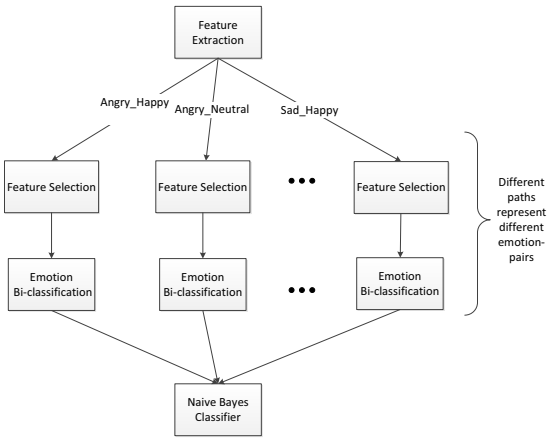

Figure 1: *Flow chart of emotion-pair based speech emotion recognition with Naive Bayes classifier based decision fusion.*

We used binary logistic regression with step-wise forward selection. The stopping criterion was based on a conditional likelihood. This feature selection process resulted in a range of 40-60 features for each binary classifier per cross validation fold.

3.3. Emotion Bi-classification for Emotion-Pair

By using the feature subset obtained in the previous step, a particular classifier can be trained for a specific emotion-pair and be designated to distinguish the emotions in that emotion-pair. As each classifier is only related to a specific emotion-pair, we call it bi-classifier. Bayesian Logistic Regression (BLR) is used as the bi-classifier in our work.

It should be noted, unlike the hierarchical binary decision tree framework, our framework can be fully automatically generated without empirical guidance. When new emotion categories are introduced to the task for recognition, only a few new emotion-pairs related to the new emotions are needed to be added to our framework. The other parts of our framework remain unchanged. However, for the hierarchical binary decision tree approach, the whole structure of decision tree needs to be reconstructed with empirical guidance in this situation. Furthermore, different bi-classifiers are independent in our method, so the framework has the advantage of being able to work in parallel.

4. Naive Bayes Classifier based Decision Fusion

After getting the emotion distinguishing result for each emotion-pair in the previous emotion classification step, a Naive Bayes classifier based decision fusion strategy is finally used to integrate the emotion bi-classification results for all emotion-pairs to derive the final emotion recognition result.

As it is well known, different archetypal emotions are located at different positions in the dimensional emotional

space. Figure 2 [25] shows the distribution of the four archetypal emotions in the three dimensional emotional space. The distance of different emotions in the emotion space, to some extent, implies the similarity between them. For example, the distance between Happy and Angry is closer than that between Happy and Sad, which indicates that Happy is more similar to Angry than to Sad. Such information might be useful for decision fusion and needs to be investigated. Take the previous example, if the target emotion is Happy, the classification result of Angry-Sad pair is more likely Angry than Sad. Motivated by this, we assume that if the emotion position distribution in the emotional space can be properly incorporated, the emotion classification results from all emotion-pairs are helpful in deriving the final emotion recognition result, even though the target emotion is not in some of the emotion-pairs.

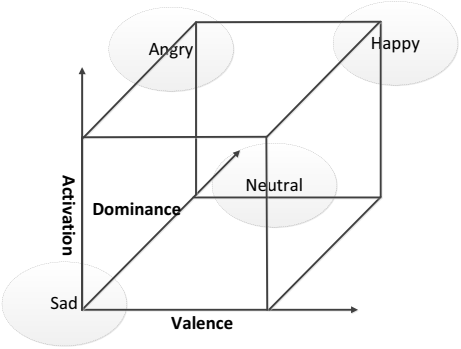

Figure 2: *The distribution of the 4 archetypal emotions in the 3-dimensional valence-activation-dominance emotion space.*

Let $E = \{e_i | i = 1, 2, ..., M\}$ be the set of emotion labels and $R = \{r_{e_j e_k} | e_j \neq e_k; r_{e_j e_k}, e_j, e_k \in E\}$ be the classification results of bi-classifiers. The relationship of the above example can be expressed in the probability form as follows (H:Happy, A:Angry, S:Sad):

$$P(r_{A\_S} = A|H) > P(r_{A\_S} = S|H) \tag{1}$$

Based on Bayes theorem, the relationship between Happy and Angry-Sad pair can be derived as follows:

$$P(H|r_{A\_S}) = \frac{P(r_{A\_S}|H)P(H)}{P(r_{A\_S})} \tag{2}$$

$$P(H|r_{A\_S}) \propto P(r_{A\_S}|H) \tag{3}$$

From equation (1) and equation (3), we can derive equation (4) when $P(H)$ and $P(r_{A\_S})$ are constant.

$$P(H|r_{A\_S} = A) > P(H|r_{A\_S} = S) \tag{4}$$

Similarly, for the decision process of the emotion-pair method, Naive Bayes classifier can be used to capture the above relationship. The posterior probability of target emotion e_i is represented as follows:

$$P(e_i|R) = \frac{P(R|e_i)P(e_i)}{P(R)} \tag{5}$$

Because the bi-classifiers are trained over different feature subspaces, we can assume that all bi-classifiers are conditional independent. Equation (5) can be represented as (6). This assumption may not be very rigorous, but it has no great effect on the final derivation result. Because the final derivation result is just a proportional relation.

$$P(e_i|R) = \frac{\prod_{r_{e_j e_k} \in R} P(r_{e_j e_k}|e_i)P(e_i)}{P(R)} \tag{6}$$

From (6), the relationship between target emotion and emotion-pairs is similar to (3) and can be represented as follows:

$$P(e_i|R) \propto \prod_{r_{e_j e_k} \in R} P(r_{e_j e_k}|e_i) \tag{7}$$

After reasonable interpretations of probability, we come to the following conjecture: by using Naive Bayes classifier for decision fusion, we can introduce the distance information between different emotions in the dimensional emotion space. This information is helpful to improve the performance of emotion recognition because all emotion-pairs will provide complementary information related to target emotion in emotion space that is useful for result decision.

5. Experiments

5.1. Experimental Setup

In this work, the USC IEMOCAP database [4] is used for conducting the experiments. The database was designed for studying multimodal expressive dyadic interactions. It was collected using motion capture and audio/video recording (approximately a total of 12h) over 5 dyadic sessions with 10 subjects. Each session consists of a different dyad where one male and one female actor perform scripted plays and engage in spontaneous improvised dialogs elicited through affective scenario prompts. At least three evaluators annotated each utterance in the database with the categorical emotion labels chosen from the set: happy, sad, neutral, angry, surprised, exited, frustration, disgust, fear and other. We consider only the utterances with majority agreement (at least two out of three evaluators gave the same emotion label) over the emotion classes of: Angry, Happy, Sad and Neutral. Such configuration is the same as [3], which makes the experimental results comparable between our work and [3]. A summary of emotion class distribution can be found in Table 1.

Table 1: *Number of utterances per emotion category in the USC IEMOCAP database [4]*

Neutral	Angry	Happy	Sad	Total
1683	1083	1630	1083	5479

Our work focuses on speaker independent emotion recognition, hence the 10-folds leave-one-speaker-out cross-validation method is used to conduct the experiments. For each fold, the utterances from one speaker are used as the testing set, and the utterances from the other speakers are used as the training set.

The experimental results can be divided into two parts. In the first part, we compare and analyze the performance of the emotion-pair based framework and the hierarchical binary decision tree based framework. In the second part, we show the distribution histogram over all emotion-pair recognition results when target emotion is Happy to verify the conjecture about the distance of different emotions implying the similarity between them.

5.2. Experimental Results

We conduct emotion recognition experiments by reporting the weighed accuracies (WA) and the unweighted accuracy (UA) of different methods, where WA is the accuracy of all samples in the test set and UA is the average value of the accuracy values of all emotions. Both metrics are standard measurements used in several previous emotion recognition challenge and is adopted in [3].

In Table 2, we present weighted accuracy (WA) and unweighted accuracy (UA) in the leave-one-speaker-out setup for USC IEMOCAP database, where "Baseline" represents the hierarchical binary decision tree with Bayesian Logistic Regression (BLR), "Emotion-pair" represents the emotion-pair method with BLR. As can be seen, the UA of "Emotion-pair" is 62.54%, which is a 4.08% absolute (6.98% relative) improvement over "Baseline"; And the WA also reaches 57.85%, which is a 1.47% absolute (2.61% relative) improvement compared to "Baseline".

Through examination of confusion matrices of both hierarchical binary decision tree method and our emotion-pair method in Table 3 and Table 4, we can find that the recognition accuracies of Angry, Happy and Sad for our emotion-pair method are improved. This is because the confusion among different emotions can be alleviated by introducing the distance information of different emotions at the final decision fusion stage. The accuracy of Neutral does not show great variation, which is probably because the distances between Neutral and other emotions do not differ so much. These results indicate that the prior information of dimensional emotion space really provides helpful information to the decision fusion of the emotion-pair method. Compared to the the hierarchical binary decision tree, our framework can provide improvement in the recognition accuracy of non-neutral emotions.

Table 2: *Comparison of weighted accuracy (WA) and unweighted accuracy (UA) between the hierarchical binary decision tree (Baseline) and the emotion-pair method (Emotion-pair) on USC IEMOCAP database.*

	WA	UA
Baseline	56.38%	58.46%
Emotion-pair	57.85%	62.54%

Table 3: *Confusion matrix by using the hierarchical binary decision tree with Bayesian Logistic Regression (BLR).*

Actual \ Predict	Neutral	Angry	Happy	Sad
Neutral	**54.51%**	6.89%	15.20%	23.40%
Angry	16.62%	**65.40%**	15.26%	2.72%
Happy	26.13%	19.57%	**41.72%**	12.58%
Sad	21.70%	2.22%	3.88%	**72.21%**

Table 4: *Confusion matrix by using the emotion-pair method with Bayesian Logistic Regression (BLR).*

Actual \ Predict	Neutral	Angry	Happy	Sad
Neutral	**53.98%**	6.19%	12.39%	27.43%
Angry	15.46%	**68.04%**	12.37%	4.12%
Happy	21.94%	15.82%	**50.51%**	11.73%
Sad	14.93%	1.49%	5.97%	**77.61%**

To further verify that the distance information of different emotions in the dimensional emotion space can really contribute to the final decision fusion with Naive Bayes classifier, we performed statistical analysis on the emotion distinguishing results of the bi-classifiers for all emotion-pairs (i.e. the output of "Emotion Bi-classification" modules in Figure 1). For each target emotion, all its corresponding utterances in the test set are recognized by the bi-classifiers of all emotion-pairs; the histogram of the identified emotion categories (i.e. the output of the bi-classifiers) can then be calculated and plotted. Take Happy as the target emotion as example, Figure 3 depicts the histogram of the identified emotion categories, where vertical axis represents the proportion of the emotion-pair recognition results of the testing utterances that fall into each of the emotion categories (corresponding to horizontal axis) over all emotion-pair recognition results. The higher the proportion of an emotion category is, the closer that specific emotion should be to Happy in the dimensional emotion space. From the figure, we can see that the proportion really reflects such expectations. For example, the distance between Happy and Angry is smaller than Happy and Sad in the emotion space, so the proportion of Angry is higher than Sad in the histogram.

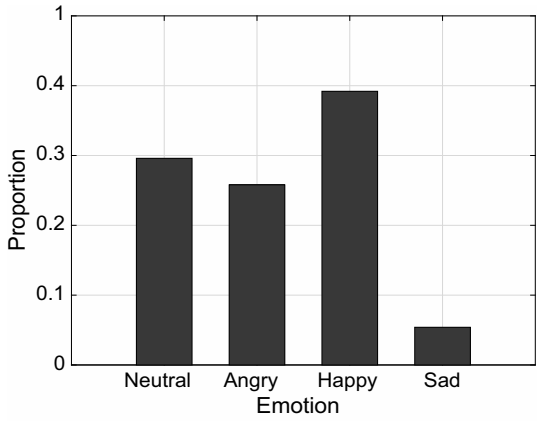

Figure 3: *Histogram of the identified emotions recognized by the bi-classifiers of all emotion-pairs, with Happiness as the target emotion.*

6. Conclusion

In this paper, we propose a speech emotion recognition framework by distinguishing different emotion-pairs in different feature subspaces, and use the Naive Bayes classifier to make the final decision by considering the relationship between different emotions in the dimensional emotion space. Experimental results have proved that our approach can achieve better results compared to the hierarchical binary decision tree method. Furthermore, our framework can be fully automatically generated without empirical guidance and is easier to be parallelized. Considering the promotion space is relatively large in feature selection, emotion bi-classification and decision fusion, our future work will be devoted to optimize these three parts.

7. Acknowledgement

This work is supported by National High Technology Research and Development Program of China (2015AA016305), National Natural Science Foundation of China (NSFC) (61375027, 61433018, 61370023 and 61171116), joint fund of NSFC-RGC (Research Grant Council of Hong Kong) (61531166002, N CUHK404/15) and Major Program for National Social Science Foundation of China (13&ZD189).

8. References

[1] M. Ayadi, M. Kamel, and F. Karray, "Survey on speech emotion recognition: features, classification schemes and database," *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011.

[2] C. Busso, S. Lee, and S. Narayanan, "Analysis of emotionally salient aspects of fundamental frequency for emotion detection," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 4, pp. 582–596, 2009.

[3] C. Lee, E. Mower, C. Busso, S. Lee, and S. Narayanan, "Emotion recognition using a hierarchical binary decision tree approach," *Speech Communication*, vol. 53, no. 9-10, pp. 1162–1171, 2011.

[4] C. Busso, M. Bulut, C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. Chang, S. Lee, and S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Journal of Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[5] M. Dash and H. Liu, "Feature selection for classification," *Intelligent Data Analysis*, vol. 1, no. 3, pp. 131–156, 1997.

[6] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of machine learning research*, no. 3, pp. 1157–1182, 2003.

[7] P. Belhumeur, J. Hespanha, and D. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 711–720, 1997.

[8] R. Cowie, E. Douglas-Cowie, and N. Tsapatsoulis, "Emotion recognition in human-computer interaction," *IEEE Signal Processing Magazine*, vol. 18, no. 1, pp. 32–80, 2001.

[9] E. Vayrynen, J. Kortelainen, and T. Seppanen, "Classifier-based learning of nonlinear feature manifold for visualization of emotional speech prosody," *IEEE Transactions on Affective Computing*, vol. 4, no. 1, pp. 47–56, 2013.

[10] L. Rabiner and R. Schafer, *Digital Processing of Speech Signals*. Prentice Hall, Upper Saddle River, New Jersey 07458, USA, 1978.

[11] B. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *the Journal of the Acoustical Society of America*, vol. 55, no. 6, pp. 1304–1312, 1974.

[12] S. Davis and P. Mermelstein, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[13] C. Gobl and A. N. Chasaide, "The role of voice quality in communicating emotion, mood and attitude," *Speech Communication*, vol. 40, no. 1-2, pp. 189–212, 2003.

[14] D. Ververidis and C. Kotropoulos, "Fast and accurate sequential floating forward feature selection with the bayes classifier applied to speech emotion recognition," *Signal Processing*, vol. 88, no. 12, pp. 2956–2970, 2008.

[15] F. J. Ferri, V. Kadirkamanathan, and J. Kittler, "Feature subset search using genetic algorithms," in *in: IEE/IEEE Workshop on Natural Algorithms in Signal Processing, IEE*. Press, 1993.

[16] D. Yu, M. Seltzer, and J. Li, "Feature learning in deep neural networks - studies on speech recognition tasks," *arXiv:1301.3605*, 2013.

[17] H. Schlosberg, "Three dimensions of emotion," *Psychological Review*, vol. 61, no. 2, pp. 81–88, 1954.

[18] A. Ortony and T. Turner, "What's basic about basic emotions," *Psychological Review*, vol. 97, no. 3, pp. 315–331, 1990.

[19] Z. Xiao, E. Dellandrea, W. Dou, and L. Chen, "Automatic hierarchical classification of emotional speech," in *Proceedings of the Ninth IEEE International Symposium on Multimedia Workshops*. IEEE Computer Society Washington, DC, USA, 2007, pp. 291–296.

[20] Q. Mao and Z. Zhan, "A novel hierarchical speech emotion recognition method based on improved ddagsvm," *Computer Science and Information Systems*, vol. 7, no. 1, pp. 211–222, 2010.

[21] E. Allwein, R. Schapire, and Y. Singer, "Reducing multiclass to binary: a unifying approach for margin classifiers," *The Journal of Machine Learning Research*, vol. 1, pp. 113–141, 2001.

[22] T. Ho, J. Hull, and S. Srihari, "Decision combination in multiple classifier systems," *IEEE Trans. Pattern Anal. Machine Intell*, vol. 16, pp. 66–75, 1994.

[23] J. Kittler, M. Hatef, R. Duin, and J. Matas, "On combining classifiers," *IEEE Trans. Pattern Anal. Machine Intell*, vol. 20, no. 3, pp. 226–239, 1998.

[24] F. Eyben, F. Weninger, and F. Gross, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proceedings of the 21st ACM international conference on Multimedia*. ACM New York, ISBN: 978-1-4503-2404-5 DOI:10.1145/2502081.2502224, 2013, pp. 835–838.

[25] M. Lugger and B. Yang, *Psychological Motivated Multi-Stage Emotion Classification Exploiting Voice Quality Features*. Speech Recognition, France Mihelic and Janez Zibert (Ed.), InTech, DOI: 10.5772/6383., 2008.