PRACTICAL CONSIDERATIONS ON THE USE OF PREFERENCE LEARNING FOR RANKING EMOTIONAL SPEECH

Reza Lotfian and Carlos Busso

Multimodal Signal Processing (MSP) Laboratory, Department of Electrical Engineering
The University of Texas at Dallas, Richardson TX 75080, USA
reza.lotfian@utdallas.edu, busso@utdallas.edu

ABSTRACT

A speech emotion retrieval system aims to detect a subset of data with specific expressive content. Preference learning represents an appealing framework to rank speech samples in terms of continuous attributes such as arousal and valence. The training of ranking classifiers usually requires pairwise samples where one is preferred over the other according to a specific criterion. For emotional databases, these relative labels are not available and are very difficult to collect. As an alternative, they can be derived from existing absolute emotional labels. For continuous attributes, we can create relative rankings by forming pairs with high and low values of a specific attribute which are separated by a predefined margin. This approach raises questions about efficient approaches for building such a training set, which is important to improve the performance of the emotional retrieval system. This paper analyzes practical considerations in training ranking classifiers including optimum number of pairs used during training, and the margin used to define the relative labels. We compare the preference learning approach to binary classifier and regression models. The experimental results on a spontaneous emotional database indicate that a rank-based classifier with fine-tuned parameters outperforms the other two approaches in both arousal and valence dimensions.

Index Terms: emotion recognition, preference learning, information retrieval, Rank SVM

1. INTRODUCTION

Emotions play a crucial role in social interactions, influencing rational decision making and perception [1]. An important aim in *human computer interaction* (HCI) is the design of emotion aware interfaces that are more attentive and responsive to the user's needs. While the community has made important advances in binary and multi-class speech emotion classification [2], only few studies have explored the use of preference learning [3, 4]. Preference learning is playing an important role in retrieving emotional content such as images [5, 6], videos [7], musics [8, 9], and texts [10, 11]. Emotion retrieval from speech can facilitate better solutions for call centers [12], and health care domains [13, 14]. It can also facilitate the collection of natural emotional speech databases [15]. This paper studies practical considerations in employing preference learning for emotional speech retrieval, providing comparison to other alternative machine learning solutions.

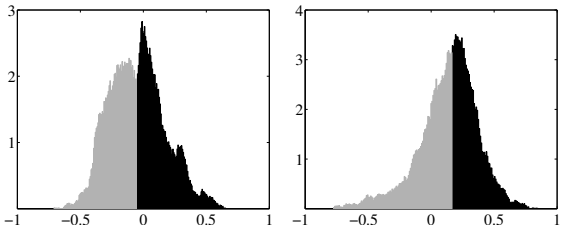
Preference learning selects between alternative samples the ones with the highest relevance according to a given criterion. It is usually implemented with pairwise comparisons, which are used to create relative rankings between the samples. Preference learning is an ideal framework for information retrieval, since it provides a ranked order sequence to select the relevant samples. Few studies have explored this framework for speech emotion recognition. The key limitations is the requirement of relative labels indicating preference between samples. An exhaustive set of labels for N samples requires $N(N-1)/2$ pairwise comparisons, which is not practical when N is large. An appealing approach is to extract these relative labels from existing emotional annotations. Cao et al. [3] used categorical labels, creating rank-based classifiers for each emotion (e.g., happy ranker). The relative rankings were created by pairing samples with different emotions, where the preferred sample conveys the emotion of the ranker. Martínez et al. [4] derived relative labels from continuous emotional descriptors for valence (negative versus positive) and arousal (calm versus active). Relative labels were created by selecting pairs of samples in which their scores were separated by a given margin. Our study explores practical implementations of this framework, comparing the results with other conventional machine learning methods.

This paper systematically analyzes different tradeoffs in building the training set for preference learning for valence and arousal. The study considers different margin values to define the pairs of samples, and the size of the training set. These parameters are clearly connected since increasing the margin will inevitably reduce the size of the training set. The analysis provides an optimal configuration that enhances the performance of the preference learning system. This framework is compared with other alternative methods including binary classification and linear regression. All these machine learning algorithms are implemented under the *support vector machine* (SVM) framework: rank-SVM, SVM and *support vector regression* (SVR). Our results indicate that preference learning outperforms SVR and binary SVM by 7% and 4% in detecting low and high level of arousal and valence, respectively.

2. SEMAINE DATABASE

This study relies on the *sustained emotionally coloured machine-human interaction using nonverbal expression* (SEMAINE) database [16]. The corpus was collected using the *sensitive artificial listener* (SAL) framework, where audiovisual data of natural interactions between a user and an operator are recorded. The operator portrayed carefully selected personalities to elicit different emotional reactions in the users. This study only considers the interactions where the operator was played by another human (solid SAL). While there are 140 conversation sessions within this setting, the study only considers 91 sessions recorded by 18 users, since they have emotional evaluations that are available. The labels include evaluation of time-continuous dimensional emotions captured by FEELTRACE [17]. The raters annotated

This work was funded by NSF (IIS-1217104 and IIS-1453781).



(a) Arousal (b) Valence

Fig. 1. Distribution of turn-based labels obtained by averaging the subjective evaluations across time and raters. The figure shows the binary classification problem for testing the models.

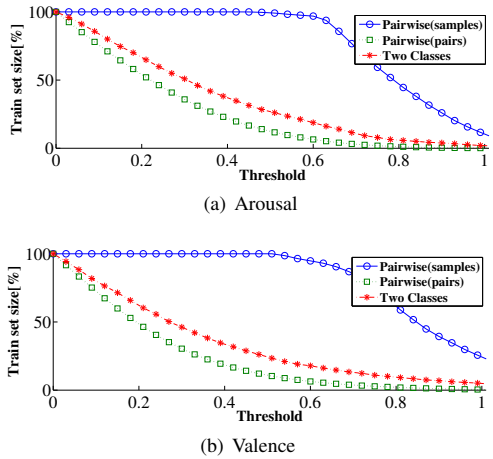
arousal, valence, expectation and power, among other emotional descriptors. This study only uses arousal and valence which are the most prominent emotional attributes. In each session, the number of evaluations per dimension varies between two and eight. With the exception of the inter-evaluator agreement analysis in Section 3.1, the annotations are combined by averaging the values across raters over the duration of the speaking turn. This approach assigns a turn level value per dimension, describing the perceived emotional content of the segment. Figure 1 shows the distributions of the labels for arousal and valence.

3. DEFINITION OF THE PROBLEM

The goal of this work is to analyze different tradeoffs in building the training set for preference learning to retrieve natural emotional speech. Since the database provides absolute ratings for arousal and valence, we derive relative labels using the same approach described in Martínez et al. [4]. Let's assume that the speaking turns s_1 and s_2 have arousal scores $e^{s_1}_{arousal}$ and $e^{s_2}_{arousal}$ (similar approach is used for valence). We estimate the absolute value of their difference $m = |e^{s_1}_{arousal} - e^{s_2}_{arousal}|$, which we refer to as *margin*. If the margin m is greater than a given threshold, we consider these speaking turns in the training set as a pair. We consider the speaking turn with the highest score as the preferred sample. We will train the preference learning classifiers with multiple examples to rank speaking turns in the test set according to their arousal or valence scores.

A second goal in this paper is to compare preference learning with other common machine learning frameworks. While the procedure for training is different across classifiers, we define a common problem to test and compare the models. We define separate binary problems consisting of detecting sentences with high and low values of arousal and valence. The binary classes are defined using median split, keeping balanced classes (see Fig. 1). For preference learning, we will rank the samples according to their relevance. We will consider a success if the samples ranked at the top of the list belong to the positive class (i.e., positive valence or high arousal – black area in Fig. 1), and the samples ranked at the bottom of the list belong to the negative class (i.e., negative valence or low arousal – gray area in Fig. 1).

While the binary problems in Figure 1 have clear limitations (e.g., similar samples close to the boundary belonging to different classes – see Marioorayd and Busso [18]), it provides the following advantages for this study: (a) we can measure precision rates when we retrieve different number of samples, (b) it defines a common problem where we can directly compare the benefits of increasing the margin without introducing bias in the evaluation (i.e., testing the results with samples with different margins across conditions will favor settings with larger margins as the problems become easier), and (c) it allows us to directly compare preference learning algorithms with other common machine learning approaches (see Sec. 5.2).



(a) Arousal

(b) Valence

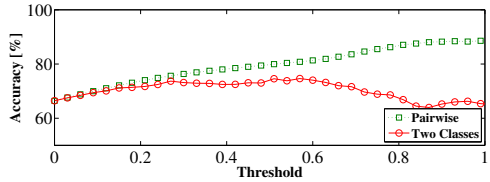
Fig. 2. The relative size of the training set for different thresholds.

3.1. Tradeoff Between Margin and Size of the Training Set

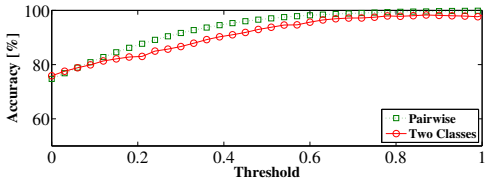
It is important to select training pairs with high level of confidence to ensure the consistency of the models. Human emotional evaluations are highly noisy since they depend on the perception of emotions, which varies not only across raters, but also within a rater during the course of the evaluation [19]. Multiple factors can affect the reliability of the relative labels defined in this study, including the intrinsic consensus between evaluators [20], the proximity between the speaking turns during the subjective evaluations, and the reaction lag of the evaluator while annotating the corpus [21, 22]. While these factors are important, we simplify the analysis in this study by considering fixed thresholds for the margins. Increasing the threshold for the margin reduces the uncertainty in the labels. However, larger thresholds reduce the number of pairs of samples that satisfy the inclusion criterion, leading to smaller training set. Therefore, the available training set size and the threshold for the margin are intrinsically related.

We study the tradeoff between the margin's threshold, and size of the training set using the labels of the SEMAINE database. Figure 2 shows the percentage of the data that meets the criterion as function of the margin's threshold for three conditions. The blue line, *Pairwise (samples)*, describes the percentage of the samples included in the training set as the threshold increases (i.e., at least one pair in the training set includes a given sample). Even with a threshold equals to 0.5, most of the samples are included in the training set. The green line, *Pairwise (pairs)*, represents the percentage of the pairs across all possible pairwise comparisons between the samples satisfying the threshold. The number of possible pairs dramatically drops as the threshold increases. For comparison, the figure also includes the red line, *Two Classes*, showing the number of individual samples with high or low ratings that are separated by a threshold around the median. This approach is commonly used in binary classification problems, where the margin aims to reduce ambiguity in the labels (i.e., detecting extreme cases). The number of samples available for training exponentially decreases as the margin increases.

We also analyze the performance of raters to identify the preferred sample using a leave-one-rater-out cross validation frame-



(a) Arousal



(b) Valence

Fig. 3. The precision of subjective evaluation for pairwise and binary classification in leave-one-out fashion for different thresholds of margin for (a) arousal and (b) valence.

work. The labels from the pairwise comparisons (ground truth) are estimated using all the emotional annotations except one. Then, we evaluate the accuracy of that rater by counting how many of the comparisons are accurately observed in his/her evaluation. We consider a success if the preferred sample has higher arousal/valence score than the other sample. We repeat this process for all the raters. Figure 3 shows the agreement between evaluators as function of the threshold (green line – *Pairwise*). The accuracies of the raters increase as we increase the margin's thresholds for activation and valence. The higher the threshold, the higher the reliability of the relative labels. For comparison, we repeat the evaluation using binary labels separated by the margin around the median (similar to the analysis in Fig. 2). The binary labels are created with all the evaluations, except one. For a sample in the lower extreme, we consider a success if the score for arousal/valence of the rater is below the median. Similarly, for a sample in the higher extreme, we consider a success if the score is above the median. The red line (*Two Classes*) shows the results as a function of the margin, which are consistent with the trends of relative labels. For similar thresholds, it is interesting that the relative labels are more reliable than the binary labels. For arousal, the performance drops for threshold higher that 0.6 since few samples remain in the set, becoming more vulnerable to outliers and unreliable evaluations. It is also interesting that the agreement is higher for valence than for arousal (the perceptual evaluation included video in addition to audio).

The analysis in this section shows that the labels are more reliable as the threshold increases. However, the size of the training set decreases. Furthermore, classifiers that are trained based on large margins may not be able to perform well in the testing set for pairs of samples with minimal differences (assuming mismatched train and test conditions). Section 5.1 analyzes the tradeoff in terms of classification performance in retrieving emotional speech.

4. PREFERENCE LEARNING (RANK SVM)

This study uses Rank-SVM as our preference learning algorithm. Rank-SVM is an extension of SVM to rank elements instead of classifying them into categorical classes [23]. The problem consists in determining the order of pairs of samples according to a given dimension. Rank-SVM can be formulated as the following optimization problem:

$$\min_{w,\zeta} \quad \frac{1}{2}\|w\|^2 + C\sum\nolimits_i \zeta_i$$
$$\text{subject to } \langle w, (x_i^{(1)} - x_i^{(2)})\rangle \geq 1 - \zeta_i, \zeta_i \geq 0 \text{ for } i \in [l] \quad (1)$$

The samples s_1 and s_2 with feature vectors $x_i^{(1)}$ and $x_i^{(2)}$, respectively, form the pair set i belonging to the training set l. We assume that s_1 is preferred over s_2. ζ_i represents the nonzero slack variable, and C is the soft margin variable. The weight vector w is determined by maximizing the margin of the support vectors [24]. For testing the model, the hyperplane defined by $\hat{w}$ can be used to estimate the ranking between samples s_1 and s_2. s_1 would be preferred over s_2 if $\langle w, (x_i^{(1)} - x_i^{(2)})\rangle \geq 0$. Otherwise s_2 is preferred over s_1. We use the implementation provided by the SVM-rank toolkit [25].

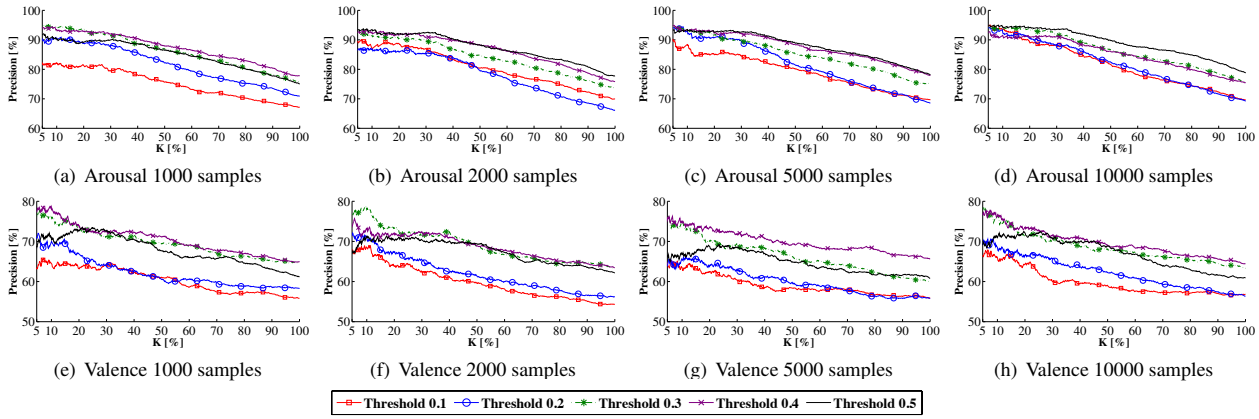
We train the Rank-SVM with a subset of the pairs from the SEMAINE database that satisfy the margin's threshold (the threshold and the size of the training set are parameters analyzed in Sec. 5.1). All the experiments are evaluated with speaker independent partitions where all the data from one speaker is only used for either development, training or testing. In particular, the recordings from eight randomly selected speakers are used as a development set (feature selection and setting the parameters of the SVMs). The recordings from the other ten speakers are divided into two folds for training and testing the models using cross-validation.

This study uses the acoustic features provided for the speaker state challenge at INTERSPEECH 2013 [26]. This set includes 6308 *high level descriptors* (HLDs) extracted using the OpenSMILE toolkit [27]. The set includes prosodic, spectral and voice quality features. We extract this set for each speaking turn of the users. Given the high dimensional feature vector, we reduce the number of features using feature selection. Since applying feature selection that maximizes the performance of a classifier is computationally expensive for a large feature set, we simplify the approach by using a two-level feature selection scheme. First, we remove less-informative features using information gain, reducing the number of features to 500. The information gain is separately implemented for arousal and valence, using binary labels (i.e., low versus high classes – see Fig. 1). Then, we select 50 features by maximizing the performance of the Rank-SVM using *floating forward feature selection* (FFFS). FFFS searches for the best features starting with an empty set. It adds one by one the features that increase the objective function. At each step, the algorithm evaluates whether removing features increases the performance, minimizing the risk of potential local optima. For each method discussed in this paper (Rank-SVM, SVM and SVR), the objective function for FFFS is defined as the precision rate after retrieving the top 10% and bottom 10% of the samples. With this approach, all the classifiers use 50 features.

5. EXPERIMENTAL RESULTS

5.1. Preference Learning with Rank-SVM

We evaluate the effect of building the training set under different conditions on the performance of the Rank-SVM. The parameters are the number of pairs in the training set and the threshold on the margin to define the sample pairs. These pairs of samples are randomly selected from the training set, which, on average, contains 1343 speaking turns over the two-fold cross validation evaluation. Motivated by our focus on retrieval, we estimate performance using precision at k (P@K). This metric is widely used in information retrieval. It measures the precision rate when k samples are retrieved (e.g., success rate in retrieving the top k samples in the list). We



(a) Arousal 1000 samples (b) Arousal 2000 samples (c) Arousal 5000 samples (d) Arousal 10000 samples

(e) Valence 1000 samples (f) Valence 2000 samples (g) Valence 5000 samples (h) Valence 10000 samples

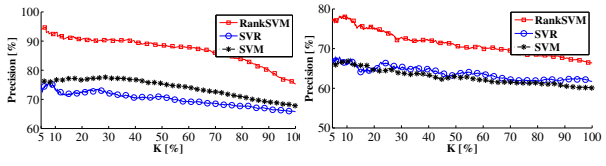
Fig. 4. Precision of differentiating between the k top and bottom ranked percentile of the samples.

measured the precision by considering the first k percent of utterances in the testing set with the highest and lowest ranked samples (i.e., $k/2$ from the top and $k/2$ from the bottom percentile of the ordered samples). As mentioned in Section 3, we consider a success if samples in the top percentile have scores above the median, and samples in the bottom percentile have scores below the median (Fig. 1). When k=100%, the evaluation considers all the samples in the testing set. Therefore, this precision rate equals the accuracy of Rank-SVM in determining samples above or below the median (binary classification using all the samples). Notice that P@K does not reflect the actual ranking within the selected samples from the top or bottom of the ranked list, but the performance in retrieving samples.

Figure 4 reports P@K results for arousal and valence when the size of the training set varies between 1000 and 10000 pairs, and the threshold for the margin varies from 0.1 to 0.5 (i.e., threshold $\in \{0.1, 0.2, 0.3, 0.4, 0.5\}$). When the threshold is small, the size of the training set affects the performance of the system. Increasing the size beyond approximately five times the size of the number of original samples does not lead to further improvement. The P@K is more sensitive to the value of the threshold. Increasing the margin between pair of samples in the training set improves the performance. From the results, we set the size of the training set to 5000. We set the threshold equals to 0.5 for arousal, and 0.4 for valence.

5.2. Comparison with Other Machine Learning Framework

Finally, we compare the performance of preference learning with two other alternative methods: binary classification and regression. The methods are implemented under the support vector framework to reflect the performance of models under similar machine learning methods. We train separate binary SVMs to recognize low versus high level of arousal and valence. For training, we use a margin around the median to define the classes similar to the analysis discussed in Section 3.1. For simplicity, we use a margin equals to 0.5 for arousal, and 0.4 for valence to be consistent with the thresholds



(a) Arousal (b) Valence

Fig. 5. Precision at K [%] of different methods.

Table 1. Accuracy of machine learning methods in classifying the entire data into low or high arousal/valence (P@K with k=100%).

Dimension	Rank-SVM [%]	SVR [%]	SVM [%]
Arousal	75.1	65.5	68.1
Valence	66.8	62.1	61.7

used to define the training set in Rank-SVM. Therefore, the SVM classifiers are trained with the samples in the extreme of their distributions. For testing, we sort the speech samples in the testing set in descending order according to the distance from the hyperplane. Likewise, we train *support vector regressions* (SVR) mapping the acoustic features into a score that estimates the arousal or valence score associated with the speaking turn. We can achieve an ordinal regression or ranking of the testing samples by sorting the scores. We use the implementation of SVM and SVR provided in LIBSVM [28]. The approach for feature selection is consistent with the method used for Rank-SVM, selecting 50 features across conditions.

Figure 5 compares the performance for SVM, SVR and rank-SVM using P@K. For arousal, rank SVM is near 20% better than the other two alternative models for most of the values of k (i.e., percentage of the data retrieved by the system). For valence, rank-SVM also offers the best performance. An interesting case is when k=100%, which is equivalent to the accuracy of these systems in categorizing the entire data into low or high level of arousal or valence (binary problems). Table 1 lists the performance for this case. While the performance from SVM and SVR are similar, Rank-SVM provides improvements over 7% for arousal, and over 4% for valence, demonstrating the benefits of using preference learning.

6. CONCLUSIONS

The study evaluated practical considerations in training preference learning algorithms. We study the tradeoff between the margin's threshold to define the sample pairs for training, and the size of the training set. The classification results demonstrate the importance of increasing the separation between the training sample pairs, making the relative labels more reliable. The performance is less sensitive to the size of the training set, as long as the number of pairs is higher than the number of samples in the corpus. We compared the performance of preference learning with the ones achieved with binary SVM and SVR for the problem of retrieving emotional samples. The results clearly demonstrate the benefits of using preference learning, even for binary problems when all the testing samples have to be categorized into either low or high level of arousal (or valence).

7. REFERENCES

[1] R. W. Picard, "Affective computing," MIT Media Laboratory Perceptual Computing Section, Cambridge, MA,USA, Technical Report 321, November 1995.

[2] C. Busso, S. Mariooryad, A. Metallinou, and S. Narayanan, "Iterative feature normalization scheme for automatic emotion detection from speech," *IEEE Transactions on Affective Computing*, vol. 4, no. 4, pp. 386–397, October-December 2013.

[3] H. Cao, R. Verma, and A. Nenkova, "Speaker-sensitive emotion recognition via ranking: Studies on acted and spontaneous speech," *Computer Speech & Language*, vol. 29, no. 1, pp. 186–202, January 2014.

[4] H. Martinez, G. Yannakakis, and J. Hallam, "Don't classify ratings of affect; rank them!" *IEEE Transactions on Affective Computing*, vol. 5, no. 2, pp. 314–326, July-September 2014.

[5] W. Wang and Q. He, "A survey on emotional semantic image retrieval," in *IEEE International Conference on Image Processing (ICIP 2008)*, San Diego, CA, USA, October 2008, pp. 117–120.

[6] S. Schmidt and W. G. Stock, "Collective indexing of emotions in images. A study in emotional information retrieval," *Journal of the American Society for Information Science and Technology*, vol. 60, no. 5, pp. 863–876, May 2009.

[7] J. Kierkels, M. Soleymani, and T. Pun, "Queries and tags in affect-based multimedia retrieval," in *IEEE International Conference on Multimedia and Expo (ICME 2009)*, Amsterdam, The Netherlands, June-July 2009, pp. 1436–1439.

[8] K. Trohidis, G. Tsoumakas, G. Kalliris, and I. P. Vlahavas, "Multi-label classification of music into emotions," in *International Conference on Music Information Retrieval (ISMIR 2008)*, Philadelphia, PA, USA, September 2008, pp. 325–330.

[9] Y.-H. Yang, Y.-C. Lin, Y.-F. Su, and H. Chen, "A regression approach to music emotion recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 2, pp. 448–457, February 2008.

[10] T. Danisman and A. Alpkocak, "Feeler: Emotion classification of text using vector space model," in *AISB 2008 Convention Communication, Interaction and Social Intelligence*, vol. 2, Aberdeen, Scotland, April 2008, pp. 53–59.

[11] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Foundations and trends in information retrieval*, vol. 2, no. 1-2, pp. 1–135, January 2008.

[12] V. Petrushin, "Emotion in speech: Recognition and application to call centers," in *Proceedings of the Artificial Neural Networks in Engineering (ANNIE 1999)*, St. Louis, MO, November 1999.

[13] M. Kranzfelder, A. Schneider, S. Gillen, and H. Feussner, "New technologies for information retrieval to achieve situational awareness and higher patient safety in the surgical operating room: the MRI institutional approach and review of the literature," *Surgical Endoscopy*, vol. 25, no. 3, pp. 696–705, March 2011.

[14] K. Pollak, R. Arnold, A. Jeffreys, S. Alexander, M. Olsen, A. Abernethy, C. Sugg Skinner, K. Rodriguez, and J. Tulsky, "Oncologist communication about emotion during visits with patients with advanced cancer," *Journal of Clinical Oncology*, vol. 25, no. 36, pp. 5748–5752, December 2007.

[15] 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.

[16] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schröder, "The SEMAINE database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 5–17, January-March 2012.

[17] R. Cowie, E. Douglas-Cowie, S. Savvidou, E. McMahon, M. Sawey, and M. Schröder, "'FEELTRACE': An instrument for recording perceived emotion in real time," in *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*. Newcastle, Northern Ireland, UK: ISCA, September 2000, pp. 19–24.

[18] S. Mariooryad and C. Busso, "The cost of dichotomizing continuous labels for binary classification problems: Deriving a Bayesian-optimal classifier," *IEEE Transactions on Affective Computing*, vol. To appear, 2015.

[19] A. Metallinou and S. Narayanan, "Annotation and processing of continuous emotional attributes: Challenges and opportunities," in *2nd International Workshop on Emotion Representation, Analysis and Synthesis in Continuous Time and Space (EmoSPACE 2013)*, Shanghai, China, April 2013.

[20] R. Cowie and G. McKeown, "Statistical analysis of data from initial labelled database and recommendations for an economical coding scheme," Belfast, Northern Ireland, UK, September 2010, SEMAINE Report D6b. [Online]. Available: http://semaine-project.eu

[21] S. Mariooryad and C. Busso, "Analysis and compensation of the reaction lag of evaluators in continuous emotional annotations," in *Affective Computing and Intelligent Interaction (ACII 2013)*, Geneva, Switzerland, September 2013, pp. 85–90.

[22] ——, "Correcting time-continuous emotional labels by modeling the reaction lag of evaluators," *IEEE Transactions on Affective Computing*, vol. 6, no. 2, pp. 97–108, April-June 2015, special Issue Best of ACII.

[23] R. Herbrich, T. Graepel, and K. Obermayer, "Support vector learning for ordinal regression," in *International Conference on Artificial Neural Networks (ICANN 1999)*, Edinburgh, UK, September 1999, pp. 97–102.

[24] V. Vapnik, *Statistical learning theory*. New York, NY, USA: John Wiley & Sons, September 1998.

[25] T. Joachims, "Training linear SVMs in linear time," in *ACM SIGKDD international conference on Knowledge discovery and data mining*, Philadelphia, USA, August 2006, pp. 217–226.

[26] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, M. Mortillaro, H. Salamin, A. Polychroniou, F. Valente, and S. Kim, "The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in *Interspeech 2013*, Lyon, France, August 2013, pp. 148–152.

[27] 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.

[28] C.-C. Chang and C.-J. Lin, "LIBSVM: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, pp. 27:1–27, April 2011.