RETRIEVING SPEECH SAMPLES WITH SIMILAR EMOTIONAL CONTENT USING A TRIPLET LOSS FUNCTION

John Harvill, Mohammed AbdelWahab, Reza Lotfian and Carlos Busso

Multimodal Signal Processing (MSP) Laboratory, Department of Electrical and Computer Engineering
The University of Texas at Dallas, Richardson TX 75080, USA

jbh150030@utdallas.edu, mxa129730@utdallas.edu, reza.lotfian@utdallas.edu, busso@utdallas.edu

ABSTRACT

The ability to identify speech with similar emotional content is valuable to many applications, including speech retrieval, surveillance, and emotional speech synthesis. While current formulations in speech emotion recognition based on classification or regression are not appropriate for this task, solutions based on preference learning offer appealing approaches for this task. This paper aims to find speech samples that are emotionally similar to an anchor speech sample provided as a query. This novel formulation opens interesting research questions. How well can a machine complete this task? How does the accuracy of automatic algorithms compare to the performance of a human performing this task? This study addresses these questions by training a deep learning model using a triplet loss function, mapping the acoustic features into an embedding that is discriminative for this task. The network receives an anchor speech sample and two competing speech samples, and the task is to determine which of the candidate speech sample conveys the closest emotional content to the emotion conveyed by the anchor. By comparing the results from our model with human perceptual evaluations, this study demonstrates that the proposed approach has performance very close to human performance in retrieving samples with similar emotional content.

Index Terms— emotion retrieval, triplet loss, ranking, perception, preference learning

1. INTRODUCTION

This paper presents a ranking formulation to retrieve speech samples with similar emotional content to that of a given anchor. Emotion is an important part of everyday life. Creating machines that understand emotion is valuable to society for a variety of reasons such as detecting depression or schizophrenia [1, 2], improving *human robot interaction* (HRI) [3], monitoring quality of service in call centers [4,5], and supporting *intelligent tutoring systems* (ITSs) [6, 7]. Common formulations for *speech emotion recognition* include regression [8–10] and classification [11, 12]. An alternative formulation is preference learning where emotional behaviors are ranked according to emotional attributes [13–16], or the categorical emotions [17, 18]. Preference learning also provides suitable tools to address a novel problem in affective computing: retrieving speech samples that have similar emotional content to that of a target anchor. The ability to compare the emotional content of two or more speech samples is important for several reasons. From a theoretical perspective, this formulation takes into account the relative nature of emotions, where relative comparisons are better than absolute judgements in terms of reliability and validity [19, 20]. From an application perspective, this formulation allows machines to distinguish differences in emotional content between speech samples based on learned preferences. The ability to identify speech with similar emotional content is valuable to many applications such as speech retrieval, surveillance, and emotional speech synthesis. This is a new formulation that opens research questions. How well can a machine complete this task? How does the accuracy of automatic algorithms compare to the performance of a human performing this task? This study addresses these questions.

This study considers an anchor speech sample that is used to retrieve speech samples with similar emotional content from an audio repository. We accomplish this goal with a triplet-loss neural network that is trained on triplets consisting of an anchor, a positive sample (a recording with similar emotion) and a negative sample (a recording with different emotion). The deep learning network creates an embedding space for samples depending on acoustic features. After training, pairs of samples with similar emotional content are closer in the embedding space than dissimilar sample pairs. Using our ranking formulation, we compare two annotation methods for representing emotional content. The first representation considers a three dimensional vector for the emotional attributes *valence, arousal, and dominance* (VAD). In this representation, the samples in the database are sorted according to the deviations from the valence, arousal and dominance scores assigned to the anchor speech. The second representation considers a nine dimensional vector for categorical emotions, describing the distribution of emotions provided by different evaluators to a given speech recording.

Our results show that this proposed network achieves better performance using the VAD representation than the categorical emotional representation. We compare the performance of our network on discriminating triplets with anchors from different regions in the VAD-space. The performance varies with the location of the anchors in the VAD-space, obtaining better accuracies for anchors with extreme emotions. The evaluation also considers perceptual evaluations to compare the performance of our network with the performance of humans completing this task. We find similar performance between humans and our models when evaluating samples from certain regions of the VAD-space.

This is the first approach that combines ranking methods with triplet loss networks to retrieve samples with similar emotion. The results show that comparing the emotional content between speech samples is a difficult task for both machines and humans.

2. RELATED WORK

This study builds on work using preference learning for emotion recognition. Previous work has shown that ranking systems can be effectively used to sort sentences according to emotional attributes (e.g., which recording is more aroused?) [13–16], or emotional categories (e.g., which recording is more happy?) [17, 18]. For example, Lotfian et al. [14] explored the practical use of preference learning to rank recordings according to their either arousal or valence scores. The study analyzed the optimal margin to establish preference be-

This work was funded by NSF CAREER award IIS-1453781.

tween two recordings and optimal size of the training set to obtain good performance. Other relevant formulations in *speech emotion recognition* (SER) are to detect emotionally salient regions [21, 22], and detecting changes in emotion during a conversation [23, 24]. While these formulations are related to our task, to the best of our knowledge, retrieving sentences with similar emotion to an anchor speech sample is a new problem.

This study relies on the use of triplet-loss networks for emotion retrieval. Triplet-loss neural networks have traditionally been used for face verification [25], but this study applies these networks to discriminate between different emotional content. Huang et al. [26] has recently demonstrated the ability of triplet loss networks to train categorical emotional speech recognition systems. The authors used a triplet loss function and a supervised loss function to train a *long short-term memory* (LSTM) network. They used hard positive mining techniques to choose hard triplets. While Huang et al. [26] used variable-length processing to make use of temporal information to train the LSTM network, this study relies on *high-level descriptors* (HLDs) from the speech extracted at the sentence level. In Huang et al. [26], the embeddings were used to classify samples using a *support vector machine* (SVM). The novelty of our approach relies on choosing triplets to determine which of two competing speech samples has similar emotional content to the anchor speech sentence.

3. RESOURCES

3.1. MSP-Podcast corpus

The study relies on the MSP-Podcast corpus, which is a database of naturalistic emotional speech [27]. The samples in the database are drawn from publicly-available podcasts, after segmenting the recordings into speaking turns. This study uses version 1.2 of the corpus, which contains 29,440 audio samples (50 hours and 5 minutes of audio). The test set has 7,341 speaking turns recorded by 50 speakers. The validation set has 2,861 speaking turns recorded by 20 speakers. The remaining 19,238 speaking turns of the corpus are included in the training set. The large amount of data in the corpus creates a rich variety of emotional content on which to train our models. Lotfian and Busso [27] gives more details about this corpus.

Each sample in the corpus was annotated by at least five annotators using a crowdsourcing protocol inspired by Burmania et al. [28]. The evaluation considered the emotional attributes valence (negative versus positive), arousal (calm versus active), and dominance (weak versus strong). Each evaluator provided a score between one and seven. The evaluation also considered categorical emotions, where the raters were asked to identify the dominant emotion perceived in the recording, which we refer to as *primary emotion*. The list of emotions are anger, sadness, happiness, surprise, fear, disgust, contempt, neutral state and other. The evaluators were also asked to annotate all the emotional classes perceived in the recordings, which we refer to as *secondary emotions*. The options for the secondary emotions were extended by adding other categories [27].

3.2. Acoustic Features

The acoustic features used to train our model are the feature *computational paralinguistics challenge* (ComParE) set [29] extracted using the OpenSMILE toolkit [30]. The features consist of HLDs calculated from *low-level descriptors* (LLDs) such as *the mean of the fundamental frequency*. In total, 6,373 HLD features are extracted from each audio segment.

4. PROPOSED APPROACH

The goal of this paper is to retrieve speech samples that have similar emotional content to that of an anchor speech sample. The formulation of this problem consists of comparison between three samples: an anchor provided to query the database, and two competing samples. Our network has to decide which of these two samples have emotional content that is closer to the emotional content of the anchor speech.

This study addresses this problem using a triplet-loss neural network. While triplet-loss networks have shown impressive performance in other fields, this structure has not been fully explored in speech emotion recognition. The goal of a triplet-loss network is to pull embeddings of emotionally similar samples together and to push embeddings of emotionally dissimilar samples apart. This is accomplished by applying the triplet loss function to a neural network during training.

4.1. Triplet Loss Function

The network maps the acoustic features extracted for each sample into a d-dimensional embedding space represented by $f(x) \in \mathbb{R}^d$. The goal is to make sample x_i^a (anchor) closer to all other similar samples x_i^p (positive), and farther from dissimilar samples x_i^n (negative). In this study, similarity is determined by distances in the VAD-space or P+S-space (see Sec. 4.2). The corresponding formulation is:

$$||f(x_i^a) - f(x_i^p)||_2^2 + \alpha < ||f(x_i^a) - f(x_i^n)||_2^2 \qquad \forall f(x_i^a), f(x_i^p), f(x_i^n) \in \Gamma \tag{1}$$

where α is a margin to help separate positive and negative samples, and Γ is the set of all possible triplets in the training set. The function $f(x)$ is learned using a fully connected *deep neural network* (DNN). Through training, the triplet loss is minimized:

$$L = max[0, \sum_i^N (||f(x_i^a) - f(x_i^p)||_2^2 - ||f(x_i^a) - f(x_i^n)||_2^2 + \alpha)] \tag{2}$$

While we do not assign classes to the samples, we use the triplet loss function to map samples with similar emotional content to similar locations in the embedding space.

4.2. Emotion Representation

An important aspect of the proposed approach is to identify an appropriate representation to describe emotions. Our approach uses two alternative representations relying on either the *core affect* theory or *basic emotion* theory.

The first representation uses emotional attributes. For each dimension, we estimate the average score provided by the evaluators to the speaking turn. Then, we create a three dimensional vector representing the speaking turn as one point in the VAD-space. We refer to this approach as VAD representation.

The second representation relies on categorical emotions. We create a nine dimensional vector where each dimension represents an emotion (Sec. 3.1). This vector provides the distribution of emotions assigned to the speaking turn by the evaluators. The vector considers the primary and secondary emotions. This study only uses secondary emotions from the original nine possible primary emotions, discarding the additional categories added to the list. We weight the primary emotion twice to emphasize that the emotion was dominant. We aggregate the selection across evaluators, normalizing the vector to obtain the distribution. We refer to this approach as P+S representation (i.e., primary plus secondary emotions).

4.3. Triplet Generation

The success of a triplet-loss network is largely dependent upon the quality of the triplets that it receives for training [25]. This study

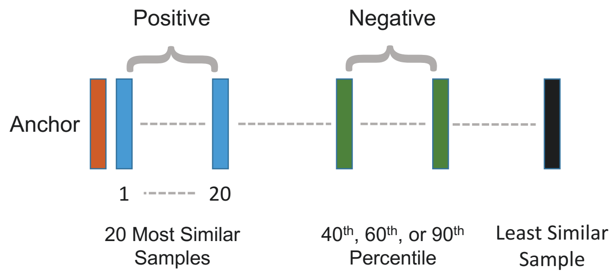

Fig. 1. Choosing positive and negative samples with respect to an anchor to generate a triplet. After sorting the samples according to their distance from an anchor sample, we randomly select a positive sample from the 20 closest samples, and a negative sample from 20 samples located at the Xth percentile in the list.

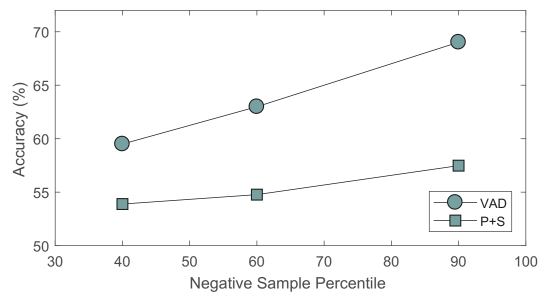

Fig. 2. Performance of the proposed triplet-loss network using sorted lists created over the VAD-space and P+S-space. The negative samples are drawn from the 40th, 60th or 90th percentiles.

creates the triplets by ordering samples with respect to their distance to the anchor sample in the emotional space. For the VAD-space, we consider the Euclidean distance in the three dimensional space between an anchor sample and each of the samples in the set. For the P+S-space, we rely on the *Kullbach-Liebler divergence* (KLD), comparing the distribution of the anchor sample with the distribution of all other samples in the set. Once the samples are ordered from the lowest to the largest distances from a given anchor, we select the positive and negative examples. Figure 1 illustrates this process. The positive sample is chosen randomly as one of the twenty closest samples to the anchor. The negative sample is randomly chosen from a set of twenty samples located at the Xth percentile of the list. We consider the 40th, 60th, or 90th percentiles. Notice that the distance between the positive and negative example is higher when we use the negative sample from the 90th percentile in the list.

5. EXPERIMENTAL EVALUATION

The evaluation considers the train, test and validation partitions in the MSP-Podcast corpus. Although we can generate multiple triplets per sample, this study only creates one triplet per speech sample by separately considering the train, test and validation partitions (i.e., the anchor, positive and negative samples belong to the same partition). We create 19,238 training triplets, 2,861 validation triplets, and 7,341 test triplets. We evaluate the triplet-loss network by estimating its accuracy in detecting positive and negative examples. The network is correct when it chooses the positive sample over the negative one. Thus the accuracy reported for the experiments is the percentage of times the embeddings of the anchor and positive sample are closer to one another than the embeddings of the anchor and the negative sample. The distance between two embeddings is estimated with Euclidean distance.

The input of the network is a 6,373 feature vector. The network has three intermediate hidden layers with 1,024 nodes per layer. The output layer is implemented with 512 nodes. Therefore, the network maps the 6,373 acoustic features to a 512 dimensional embedding. We use dropout (p=0.2) between layers and batch normalization. The hidden layers use the *rectified linear unit* (ReLU) activation function. For training, we use the Adam optimizer and a batch size of 512. The models are trained for 15 epochs on a total of 19,238 triplets. The results reported throughout the paper are the averages of the performance of 10 models with random weight initializations.

5.1. Global Performance

We evaluate the proposed approach by training and testing networks on triplets created with either the VAD-space or the P+S-space. The negative samples are selected from the 40th, 60th, or 90th percentile of the list (i.e., six conditions). Figure 2 shows the accuracy of the models. The best performance that we obtain for this task is 69%, when we use the VAD-space drawing negative samples from the 90th percentile. The results show that the triplets created based on the ordering of VAD vectors are better than those created from the P+S vectors. We compare the average accuracies over the 10 models for each condition using the one-tailed two sample population mean t-test, asserting significance at p-value$<$0.05. The differences are statistically significant at each percentile, showing that using the valence, arousal and dominance space is more appropriate for this task than using categorical emotions. Another trend to notice is that as the percentile increased, the performance improved for both emotional spaces. This result is intuitive, because a higher percentile implies the negative samples are drawn from farther down the list, increasing the difference between positive and negative samples. While the VAD-space is superior to the P+S-space based on the results of this experiment, the fact that the performance increased for both methods as the negative sample percentile increased shows that both methods are valuable in being able to rank samples in a meaningful way with respect to emotional content. We focus the rest of the discussion on results obtained using the VAD-space, due to its superior performance over the P+S-space.

5.2. Performance per Region in the VAD-Space

In order to gain a better understanding of the performance of the model with respect to the expressiveness of the anchor samples, we divide the VAD-space into 27 cubes by splitting the values for each attribute into low (1-3), medium (3-5) and high (5-7). The triplets are placed into each cube according to their anchors' VAD value. Figure 3 illustrates this process, listing the number of triplets in the testing set per cube. Cubes marked with "None" did not contain any triplets. The central cube has 3,243 test samples, which convey fairly neutral emotions. Samples in the cubes toward the edges present more expressive behaviors.

Figure 3 shows the accuracy per cube, when we draw negative samples from the 90th percentile of the sorted list. In general, we observe better performance towards the outer areas of the VAD-space. We believe that this result is due to the distribution of the data. For anchors in the central cube, negative samples in the 90th percentile in the list may not be far enough from the positive samples, as there are many samples in the center. For anchors in the extreme of the space, the distance in the VAD-space between positive and negative samples is larger, simplifying this task. The best performance is associated with regions where the arousal value is either low or high, suggesting that this may be the most discriminative dimension for this task using acoustic features. We also notice similar performance across the valence dimension when arousal and dominance are held

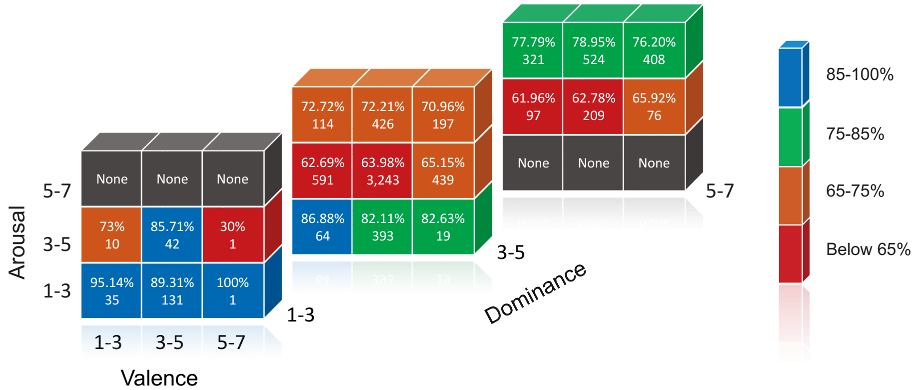

Fig. 3. Analysis of the accuracy of the triplet-loss network as a function of the anchors VAD score. The results are clustered into 27 cubes, when we draw negative samples from the 90th percentile of the sorted list. Extreme values of the attributes lead to better performance.

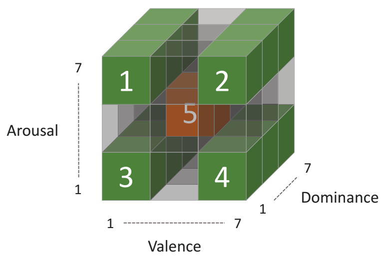

Fig. 4. Perceptual evaluation regions.

Table 1. Comparing human performance and our proposed triplet-loss network.

	Triplet network		Human performance
Region	Entire Test Set	60 Triplets	60 Triplets
	90th Percentile	90th Percentile	90th Percentile
1	76.46%	82%	**86.67%**
2	74.50%	**96%***	73.33%
3	89.80%	**98%***	82.22%
4	83.50%	**74%**	66.67%
5	63.98%	65%	**75.31%**
	40th Percentile	40th Percentile	40th Percentile
1	66.69%	64%	**75.56%**
2	65.98%	64%	**80%***
3	78.79%	**78%**	65.56%
4	65.50%	**66%**	57.78%
5	56.59%	49%	**60%***

constant. This result indicates that valence may be less discriminant for comparing emotional content using acoustic features. This result agrees with previous studies showing the changelles in detecting valence from speech [31, 32].

5.3. Human Performance per Region in the VAD-Space

We use triplets from the test set to evaluate the ability of humans to detect which candidate speech is more similar to the emotion of the anchor. To simplify the evaluation, we only consider the five regions shown in Figure 4. These regions mostly consider arousal and valence scores, discarding dominance. We create two separate sets of 30 triplets each, one of which has the negative samples drawn from the 90th percentile and one which has the negative samples drawn from the 40th percentile. For each set of 30 triplets, five triplets belong to each of the regions one through four, and ten triplets belong to region five. The perceptual evaluations are performed by 9 subjects in our laboratory, who evaluated all 60 triplets.

Table 1 lists the cumulative results. The table also reports the performance of our triplet-loss network over the 60 triplets evaluated by the raters, and over all the sentences in each of these five regions. Percentages in bold indicate better performance by either the model or humans, and an asterisk indicates statistical significance (one-tailed two sample proportion t-test, p-value$<$0.05). When the negative samples are drawn from the 90th percentile, we can see that the only statistically significant results are in regions two and three where the proposed model outperforms human perception. When the negative samples are drawn from the 40th percentile, we notice a decrease in performance for both humans and our model compared to triplets from the 90th percentile. We also notice that humans perform statistically significantly better than our model on regions two and five. These results indicate that our model performs better on easy triplets with expressive anchors whereas humans perform better on more difficult triplets with less expressive anchors.

6. CONCLUSIONS

This paper proposed to determine speech recordings with similar emotional content to that of an anchor speech sample. This approach is consistent with the ordinal nature of emotion [19, 20], and offers new opportunities in practical applications. We addressed this problem with a triplet-loss network that compares the embedding of the anchor speech with the ones of the positive sample (i.e., similar emotion to the anchor), and a negative sample (i.e., different emotion from the anchor). This study demonstrated that evaluating emotion similarity in the VAD-space is better than in the P+S-space, for this task. The results showed that triplets with expressive anchors are easier to discriminate than triplets with neutral anchors due to larger distances between positive and negative samples in the VAD-space. The study also showed that the performance of our model for the task of retrieving similar emotional samples is close to human performance, overall, and is even superior when the anchor speech is in certain regions in the VAD-space.

While the results are competitive, we are exploring alternative options to improve the accuracy of the models, especially for anchors in the center of the VAD-space. One limitation of the study is that the number of human evaluators is small. Also, humans only evaluate a few samples from each region, thus the samples taken from each region may not properly represent each region's distribution. We are planning to extend the perceptual evaluation to cover more examples. A future work in this area is to perform a similar study where each partition includes data from a given speaker. Due to differences in how humans express emotion, the task may become easier when one subject's emotional expression is learned in depth with preference learning through a triplet loss neural network.

7. REFERENCES

[1] L.A. Low, N.C. Maddage, M. Lech, L.B. Sheeber, and N.B. Allen, "Detection of clinical depression in adolescents' speech during family interactions," *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 3, pp. 574–586, March 2011.

[2] J. Edwards, H.J. Jackson, and P.E. Pattison, "Emotion recognition via facial expression and affective prosody in schizophrenia: a methodological review," *Clinical Psychology Review*, vol. 22, no. 6, pp. 789–832, July 2002.

[3] P. Rani, C. Liu, N. Sarkar, and E. Vanman, "An empirical study of machine learning techniques for affect recognition in human-robot interaction," *Pattern Analysis and Applications*, vol. 9, no. 1, pp. 58–69, May 2006.

[4] L. Devillers, C. Vaudable, and C. Chastagnol, "Real-life emotion-related states detection in call centers: A cross-corpora study," in *Interspeech 2010*, Makuhari, Japan, September 2010, pp. 2350–2353.

[5] C.M. Lee and S.S. Narayanan, "Toward detecting emotions in spoken dialogs," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 2, pp. 293–303, March 2005.

[6] S.K. D'Mello, S.D. Craig, B. Gholson, S. Franklin, R. Picard, and A.C. 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.

[7] A. De Vicente and H. Pain, "Informing the detection of the students' motivational state: An empirical study," in *International conference on Intelligent Tutoring Systems (ITS 2002)*, S.A. Cerri, G. Gouardères, and F. Paraguaçu, Eds., vol. 2363 of *Lecture Notes in Computer Science*, pp. 933–943. Springer-Verlag Berlin Heidelberg, Biarritz, France and San Sebastian, June 2002.

[8] M. Wöllmer, A. Metallinou, N. Katsamanis, B. Schuller, and S. Narayanan, "Analyzing the memory of BLSTM neural networks for enhanced emotion classification in dyadic spoken interactions," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2012)*, Kyoto, Japan, March 2012, pp. 4157–4160.

[9] S. Parthasarathy and C. Busso, "Jointly predicting arousal, valence and dominance with multi-task learning," in *Interspeech 2017*, Stockholm, Sweden, August 2017, pp. 1103–1107.

[10] S. Parthasarathy and C. Busso, "Ladder networks for emotion recognition: Using unsupervised auxiliary tasks to improve predictions of emotional attributes," in *Interspeech 2018*, Hyderabad, India, September 2018, pp. 3698–3702.

[11] E.M. Albornoz, D.H. Milone, and H.L. Rufiner, "Spoken emotion recognition using hierarchical classifiers," *Computer Speech & Language*, vol. 25, no. 3, pp. 556–570, July 2011.

[12] R. Lotfian and C. Busso, "Curriculum learning for speech emotion recognition from crowdsourced labels," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. To Appear, 2019.

[13] S. Parthasarathy and C. Busso, "Preference-learning with qualitative agreement for sentence level emotional annotations," in *Interspeech 2018*, Hyderabad, India, September 2018, pp. 252–256.

[14] R. Lotfian and C. Busso, "Practical considerations on the use of preference learning for ranking emotional speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2016)*, Shanghai, China, March 2016, pp. 5205–5209.

[15] S. Parthasarathy, R. Lotfian, and C. Busso, "Ranking emotional attributes with deep neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2017)*, New Orleans, LA, USA, March 2017, pp. 4995–4999.

[16] S. Parthasarathy, R. Cowie, and C. Busso, "Using agreement on direction of change to build rank-based emotion classifiers," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 11, pp. 2108–2121, November 2016.

[17] 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.

[18] 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 2015.

[19] G.N. Yannakakis, R. Cowie, and C. Busso, "The ordinal nature of emotions," in *International Conference on Affective Computing and Intelligent Interaction (ACII 2017)*, San Antonio, TX, USA, October 2017, pp. 248–255.

[20] G.N. Yannakakis, R. Cowie, and C. Busso, "The ordinal nature of emotions: An emerging approach," *IEEE Transactions on Affective Computing*, vol. To appear, 2019.

[21] S. Parthasarathy and C. Busso, "Predicting emotionally salient regions using qualitative agreement of deep neural network regressors," *IEEE Transactions on Affective Computing*, vol. To appear, 2019.

[22] S. Parthasarathy and C. Busso, "Defining emotionally salient regions using qualitative agreement method," in *Interspeech 2016*, San Francisco, CA, USA, September 2016, pp. 3598–3602.

[23] Z. Huang and J. Epps, "Detecting the instant of emotion change from speech using a martingale framework," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2016)*, Shanghai, China, March 2016, pp. 5195–5199.

[24] Z. Huang, J. Epps, and E. Ambikairajah, "An investigation of emotion change detection from speech," in *Interspeech 2015*, Dresden, Germany, September 2015, pp. 1329–1333.

[25] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2015)*, Boston, MA, June 2015, pp. 815–823.

[26] J. Huang, Y. Li, J. Tao, and Z. Lian, "Speech emotion recognition from variable-length inputs with triplet loss function," in *Interspeech 2018*, Hyderabad, India, September 2018, pp. 3673–3677.

[27] 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, 2019.

[28] 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.

[29] 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.

[30] 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.

[31] C. Busso and T. Rahman, "Unveiling the acoustic properties that describe the valence dimension," in *Interspeech 2012*, Portland, OR, USA, September 2012, pp. 1179–1182.

[32] K. Sridhar, S. Parthasarathy, and C. Busso, "Role of regularization in the prediction of valence from speech," in *Interspeech 2018*, Hyderabad, India, September 2018, pp. 941–945.