Quantifying Emotional Similarity in Speech

John Harvill, *Student Member, IEEE*, Seong-Gyun Leem, *Student Member, IEEE*, Mohammed AbdelWahab, *Student Member, IEEE*, Reza Lotfian, *Student Member, IEEE*, and Carlos Busso, *Senior Member, IEEE*

Abstract—This study proposes the novel formulation of measuring emotional similarity between speech recordings. This formulation explores the ordinal nature of emotions by comparing emotional similarities instead of predicting an emotional attribute, or recognizing an emotional category. The proposed task determines which of two alternative samples has the most similar emotional content to the emotion of a given anchor. This task raises some interesting questions. Which is the emotional descriptor that provide the most suitable space to assess emotional similarities? Can *deep neural networks* (DNNs) learn representations to robustly quantify emotional similarities? We address these questions by exploring alternative emotional spaces created with attribute-based descriptors and categorical emotions. We create the representation using a DNN trained with the triplet loss function, which relies on triplets formed with an anchor, a positive example, and a negative example. We select a positive sample that has similar emotion content to the anchor, and a negative sample that has dissimilar emotion to the anchor. The task of our DNN is to identify the positive sample. The experimental evaluations demonstrate that we can learn a meaningful embedding to assess emotional similarities, achieving higher performance than human evaluators asked to complete the same task.

Index Terms—Speech emotion recognition, ordinal affective computing, representation learning of emotion similarity, triplet loss function, speech emotion retrieval

1 Introduction

Emotion recognition is important for a variety of problems in health, psychology, education and engineering. Automatic emotion recognition can be used to identify depression, schizophrenia, and other forms of mental conditions [1], [2], [3], improve realism in *human-robot interaction* (HRI) [4], predict learning metrics in *intelligent tutoring systems* (ITS) [5], [6], and monitor service quality in call centers [7], [8]. Creating systems that can automatically understand emotion makes it possible to deliver these services at a much higher scale, allowing the benefits of such systems to reach many users. Common formulations for *speech emotion recognition* (SER) are regression problems [9], [10], [11] and classification tasks [12], [13]. An alternative formulation is preference learning where the task is to compare the emotional content between two or more samples. Methods based on preference learning offer promising solutions for understanding emotional content in speech. These methods are rooted on the ordinal nature of emotions [14], building on the undenied evidences that relative emotional comparisons are better than absolute assessments of emotions [14], [15] (e.g., is sentence A *happier* than sentence B?). Preference learning has been explored when applied to emotional attributes [16], [17], [18], [19], [20] and categorical emotions [21], [22]. This paper proposes a novel formulation in preference learning, where the task is to identify speech samples with emotional content as close as possible to an anchor sentence.

Our proposed formulation in affective computing is related to retrieval tasks, where the goal is to extract the most emotionally-similar recording compared to an anchor speech. This problem is important in different domains. An algorithm that finds speech samples with similar emotions can help in identifying related events for surveillance applications. This formulation can also be used to provide a better emotional characterization of an input speech. For example, the anchor speech can be used to retrieve labeled samples in an emotional corpus. Then, the labels of the closest samples in the corpus can be collectively assigned to the anchor speech. Quantifying emotional similarity can also be useful in tasks such as the detection of emotional changes during a recording [23], [24] or detection of emotional hotspots [25], [26]. As a preference learning task, finding speech samples with similar emotions leverages the ordinal nature of emotions, providing direct comparisons between samples instead of absolute emotional assessments. As far as we are aware, the only study addressing this problem is our preliminary study [27], which is extended in this paper. There are important opportunities in affective computing if we have robust algorithms that can quantify emotional similarity between speech recordings.

We formulate the problem of choosing the most emotionally-similar sample compared to an anchor as a preference learning problem by building upon our preliminary work [27]. We aim to find a function that maps samples from an acoustic feature space into an emotional representation space from where we can estimate distance between emotional

• *The authors are with the Department of Electrical and Computer Engineering, The University of Texas at Dallas, Richardson, TX 75080 USA. E-mail: {jbh150030, Mohammed.Abdel-Wahab, reza.lotfian, busso}@utdallas.edu, SeongGyun.Leem@UTDallas.edu.*

Manuscript received 25 Jan. 2021; revised 31 Aug. 2021; accepted 1 Nov. 2021. Date of publication 0 . 0000; date of current version 0 . 0000.
This work was supported by the National Science Foundation (NSF) under Grants CNS-2016719 and CAREER IIS-1453781.
(Corresponding author: Carlos Busso.)
Recommended for acceptance by K. P. Truong.
Digital Object Identifier no. 10.1109/TAFFC.2021.3127390

contents. In this emotional representation, we expect that emotionally-similar samples are closer to one another than emotionally-dissimilar samples. We choose to represent this function as a *deep neural network* (DNN), relying on the triplet loss function [28]. The triplet loss function optimizes preferences by using an anchor, a positive sample, and a negative sample. In this formulation, the anchor and positive sample have very similar emotions, while the anchor and negative sample have dissimilar emotions. The loss function aims to create a representation where the anchor and the positive sample are as close as possible, while increasing the distance between the negative sample and the anchor. The key challenge is how to determine emotional similarity such that the preferences we train our network with are meaningful with respect to emotional content. For this study, we determine emotional similarity based on either emotional attributes or emotional categories. For each anchor, we estimate a ranked-order list by estimating the emotional distance using the annotations by individual evaluators. We choose sentences from the top of the list as positive samples, and sentences from lower down the list as negative samples, creating triplets to train our models.

We train and evaluate our approach using the MSP-Podcast corpus [29], showing that the proposed triplet loss approach to quantify similarity between speech recordings is a valuable tool for understanding and characterizing emotions. Given an anchor speech, the task is to choose the most emotionally-similar speech sample between two competing recordings. A comparison is successful if the positive sample is selected. The results demonstrate that the emotional attribute space provides a more reliable space to create the triplets than spaces with categorical emotional descriptors. The similarity task is easier when the anchor has more extremes values of arousal and valence, achieving accuracies up to 93%. The accuracy decreases for anchors with more neutral attribute scores. The results also show better performance as we increase the distance between the positive and negative examples. For the categorical descriptor space, the use of secondary emotions (i.e., all the emotional traits conveyed in a sentence in addition to the dominant emotion) is useful to increase the accuracy of the models. The proposed triplet loss model performs better than competitive baselines built for this task, where the differences across conditions are often statistically significant. The models are also compared with human performance by using perceptual evaluations to assess how well humans can determine emotional similarity between speech samples. On average, our triplet loss model leads to better accuracies than human performance. The evaluation demonstrates how difficult this task is even for human, and the potential of the proposed approach. The triplet loss formulation is useful for accurately retrieving emotionally-similar speech sentences. The main contributions of this study are:

- A novel ordinal formulation in affective computing, where the task is to quantify emotional similarity between speech recordings.
- A novel triplet loss approach that creates a feature embedding to quantify emotion similarity.
- An exhaustive evaluation to determine the emotional space that is more appropriate for this task, comparing the accuracy of our triplet loss approach with human performance.

The paper is organized as follows. Section 2 presents the related work, emphasizing studies using the triplet loss function and preference learning. Section 3 introduces the resources used in this study, including the database and acoustic features used to train our models. Section 4 presents our proposed approach, describing our triplet loss formulation. Section 5 describes the experimental setting for this study, including the baselines used to evaluate our models. Section 6 presents the results, describing the accuracy of our approach to discriminate emotional similarity. Section 7 presents perceptual evaluations, comparing the accuracies of our approach to human performance. Finally, Section 8 summarizes the paper, describing its implications and providing research direction for future work.

2 Related Work

To the best of our knowledge, our preliminary study [27] is the only paper formulating emotion recognition as a retrieval problem, where the aim is to identify recordings with similar emotion to a given anchor. This section discusses studies that are related to our work, focusing on contrastive learning framework such as the triplet loss function, especially applied to speech tasks, and ordinal formulation for speech emotion recognition.

2.1 Contrastive Learning Framework

Learning a similarity space has been actively researched in the field of contrastive learning [30]. Contrastive learning is designed to learn a discriminative representation by comparing the representations extracted from different samples. The main idea of this approach is to make the representations from similar samples to be closer to each other, while the representations from dissimilar samples to be far away. Such comparison-based learning allows the classifier to create more discriminative features, yielding improvements in the classification accuracy. Siamese network is an example of contrastive learning, which was first introduced by Bromley *et al.* [31] to improve the accuracy in signature verification tasks. In their study, they trained two separate *time-delay neural networks* (TDNNs) that share the weight parameters. Each TDNN extracts the hidden representations from signature images, and the distance metric is calculated by comparing the two hidden representations. If two signature images are written by the same person, the model is trained to minimize the distance metric. The trained representations can be directly used to verify the similarity between two different images [31]. The representation can also be used as an input for the output layer [32]. This learning scheme also showed good performance in face verification [33] and gait recognition [34] tasks.

Instead of using the representations from two different samples, comparison-based learning can be performed with a single sample by using data augmentation. Chen *et al.* [35] proposed a *simple framework for contrastive learning of visual representations* (SimCLR). In this approach, the system first applies simple transformations to the input images, including color distortion, cropping, and blurring. Since those transformations do not affect the class information of input images,

the feature encoder is trained to minimize the distance between the representations from two distorted image created by applying different transformation to one original image. The feature encoder can generate more discriminative input for the following classification network by using this contrastive learning method, leading to the improvement of classification performance. This approach also yields a performance improvement in speech recognition and speech emotion recognition [36], implying that the contrastive learning framework can also be successfully applied to the speech processing field.

Studies in this field have used various functions to measure the distance between representations, including cosine distance [31], *normalized temperature-scaled cross entropy loss* (NT-Xent loss) [35], and *noise contrastive estimation* (NCE) [37]. In our study, we use the triplet loss function to compare the distance between different representations. We will describe the details of the triplet loss function in Section 2.2

2.2 Triplet Loss

The triplet loss function was introduced in Schroff *et al.* [28], where the authors showed its potential for face recognition. The triplet loss function takes three samples that are processed by a function that is often implemented with DNN. One of the samples is the anchor sample. The other two samples are the positive example, which is supposed to be similar to the anchor, and a negative example, which is supposed to be different from the anchor. Schroff *et al.* [28] noted that the embeddings created by the triplet loss network were useful to determine the similarity between faces. Since then, the triplet loss function has been successfully used in several tasks including object tracking [38], person re-identification [39], face verification [40], intention detection in a dialogue system [41] and anomaly driving detection [42].

The triplet loss function has also been used in speech tasks including speaker verification [43], [44], speaker identification [45], speaker diarization [46], sleepiness detection [47] and noise classification [48]. Novoselov *et al.* [45] used the *cosine-similarity metric learning* (CSML) to identify speakers, which was trained with the triplet loss function. They showed that the approach was accurate and robust compared to alternative methods. Li *et al.* [43] used the triplet loss function to create embeddings for speaker verification and speaker identification tasks. The authors used *convolutional neural networks* (CNN) and *recurrent neural networks* (RNN) to extract features at the frame level, which were pooled to form an utterance-level vector used to predict the speaker identity. Bredin *et al.* [46] used a triplet loss function to train a deep neural network for speaker change detection. *Mel-frequency cepstral coefficients* (MFCCs), energy features, and their derivatives are used as input of a DNN implemented with *bidirectional long short-term memory* (BLSTM) cells. These examples show that the triplet loss function is effective for speech tasks.

The triplet loss function has been used before in other SER studies. However, the goals, and the problem formulations in previous work are radically different from the ones used in our study. Huang *et al.* [49] used the triplet loss function to create a discriminative embedding for a SER task. The embedding was then used to train a *support vector machine* (SVM). The triplet loss function aims to reduce intra-class variability, and increase inter-class variability between emotional classes. The loss function also considered a supervised term to discriminate between emotional classes. Kumar *et al.* [50] presented a similar approach using the *residual neural network* (ResNet) architecture that combines the triplet and cross entropy losses. The triplet loss function increases the distance between different emotions and reduces the distance between similar emotions. Their approach can be trained end-to-end without the need for a SVM. Han *et al.* [51] also followed a similar approach. Han *et al.* [52] used a triplet loss function incorporating both audio and video features to improve emotion prediction. Feng and Chaspari [53] applied the triplet loss to the problem of transfer learning (fine-tuning) in emotion recognition with limited data. Notice that the conventional approach to use the triplet loss function in SER is to improve the feature embedding such that sentences with similar emotions are close and sentences with different emotions are far. Our formulation does not aim to recognize a given emotional class or predict an emotional attribute. Instead, we aim to quantify emotional similarity, representing a novel contribution in affective computing.

2.3 Ordinal Nature of Emotions

There are strong evidences that emotions are better computationally represented with ordinal methods. Methods that provide comparative assessments are often more robust and more reliable than methods that assign an absolute score or an emotional category (e.g., is sentence one *more aroused* or *happier* than sentence two?). Yannakakis *et al.* [14], [15] presented a complete study with evidences across domains about the ordinal nature of emotions. In this section, we briefly describe some of the ordinal formulations that have been considered by previous studies.

An emerging formulation in affective computing is preference learning, where the task is to establish preferences between samples with respect to a given dimension. After establishing preferences, it is straightforward to rank samples according to the given criterion. If needed, rankings can be later transformed into ratings [54]. For example, Cao *et al.* [22] proposed to rank emotions with respect to emotional categories (e.g., happiness, anger, sadness). A similar formulation was proposed by Lotfian and Busso [21], where the individual evaluations between multiple annotators were used to establish preference between sentences. Preference learning has also been used to rank samples with respect to emotional attributes (e.g., valence, arousal, dominance) [16], [17], [18], [19], [20]. Martinez *et al.* [17] showed that transformations of ratings into rankings resulted in a better approach than transformations of ratings into discrete classes. Parthasarathy *et al.* [20] established preferences between samples using time-continuous annotations. They relied on the *qualitative agreement* (QA), which looks for trends across multiple evaluations. This approach was later extended for sentence level annotations of emotional attributes [16]. Many other studies have also capitalized on the ordinal nature of emotions and use a ranking approach to represent emotion. Lopes *et al.* [55] explored a ranking formulation for determining affect in horror soundscapes. Yang and Chen [56], [57] used a ranking formulation to retrieve music pieces with similar emotions.

Soleymani *et al.* [58] demonstrated a relationship between emotion rankings of movie scenes predicted from physiological signals and user self-assessment valence and arousal rankings. Mariooryad *et al.* [59] demonstrated that preference learning is an appealing approach to retrieve sentences with target emotional content, using this approach to build a database. This approach was a building block used to create the MSP-Podcast corpus [29], described in Section 3.1. Liang *et al.* [60] relied on relative and absolute models to create a multimodal emotion recognition system. The approach compared pairs of frames (audio and visual modalities) creating local rankings, which were combined to create a global ranking. The final prediction combines the absolute and rank-based models. A key reason for the success of rank-based models is that the ground truth labels are more reliable. Yannakakis *et al.* [61] used a rank-based annotation formulation for modeling affect. They found that an ordinal labeling system led to greater inter-rater agreement compared to an absolute annotation formulation. Yannakakis and Hallam [62] compared ranking and rating self-reporting methodologies for determining affect, demonstrating higher inter-evaluation agreement when using ranking methods. Holmgard *et al.* [63] demonstrated superiority of ranking-based approaches compared to class-based approaches of stress annotation for the purpose of determining stress in PTSD patients.

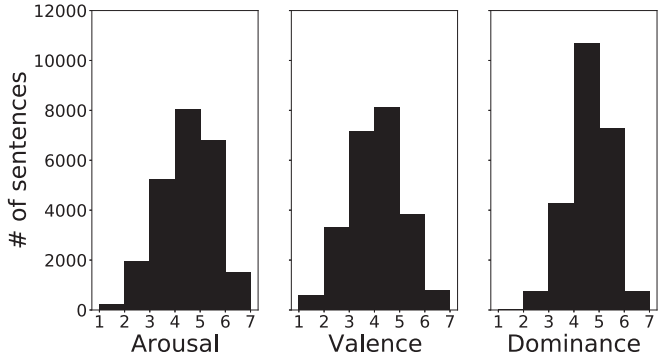
There are other alternative ordinal formulations investigated in previous studies. For example, Huang and Epps [23], [24] investigated the problem of detecting changes of emotions within a conversation. Another related formulation is to detect emotionally-salient regions in speech over time [25], [26] (e.g., emotional hotspots). The task in this study is related, and complementary, to these formulations. Our approach aims to quantify emotional similarities, which is a powerful and novel research direction in affective computing.

3 Resources

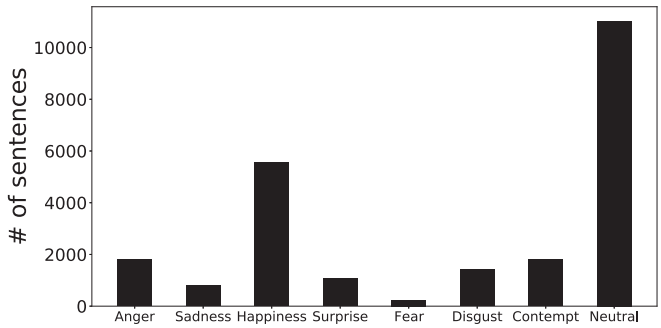
3.1 MSP-Podcast Corpus

We train and evaluate our approach with the MSP-Podcast corpus [29]. This corpus is a collection of speech sentences taken from publicly-available podcasts with Creative Commons licenses. The protocol follows the ideas presented by Mariooryad *et al.* [59], which uses *speech emotion recognition* (SER) systems to retrieve target segments to be annotated with emotional labels. The approach prioritizes samples that are more likely to be emotional, offering the tools to balance the emotional content in the corpus (e.g., finding samples with positive emotions). The sentences range in length from 2.75-11 seconds and cover a wide range of emotions. The sentences in the corpus are annotated using an improved version of the crowdsourcing protocol presented in Burmania *et al.* [64]. The approach tracks in real time the quality of the workers, stopping the evaluation when the performance drops below an acceptable level. The corpus is evaluated with interval (e.g., attributes) or nominal (e.g., categorical) emotional descriptors, where each speaking turn was annotated by five or more workers.

Valence, Arousal, and Dominance (VAD). We consider the emotional attributes valence (negative versus positive), arousal (calm versus active), and dominance (weak versus strong). Each worker assigned a score in the range of 1 to 7.



(a) Histograms of emotional attribute labels



(b) Histogram of primary categorical emotion labels

Fig. 1. Histograms describing the emotional content of the MSP-Podcast corpus (release 1.2), for (a) emotional attributes, and (b) emotional categories.

The consensus labels are the average scores provided to a speaking turn across workers. Fig. 1a illustrates the histograms of the consensus dimensional emotion attribute labels, which show the expected distributions centered around the center of the axes (value 4). While less samples are included in the extremes, the figures show that samples in the corpus are indeed labeled with extreme values for the emotional attributes. In this study, we consider the annotations in the VAD space as a vector in a three-dimensional space, where each axis represents an attribute.

Primary Emotion (PE). We also consider categorical emotions. Each worker was asked to select the most prominent emotion perceived in the speaking turn. The list includes nine possible options: anger, contempt, disgust, fear, happiness, sadness, surprise, neutral state and other. Fig. 1b illustrates the histogram of the consensus primary emotion labels, where the class neutral is the prominent emotion. Most of the emotional classes have at least 798 samples. The only exception is fear, which only has 239 samples in this version of the corpus. In this study, we ignore "other," creating an eight-dimensional space where each axis corresponds to one emotion. The consensus label for a speaking turn is the emotional category selected by most of the workers (i.e., plurality rule).

Secondary Emotion (SE). Spontaneous interactions often include ambiguous emotional content, where more than one emotion may be perceived from the speech (e.g., anger and frustration). The data collection also includes secondary emotions to capture these emotional trails. The workers were able to select multiple emotional categories. In addition to the classes included for the primary emotions, the

list was extended to include more subtle emotions: amused, frustrated, depressed, concerned, disappointed, excited, confused, and annoyed. The primary emotion is always considered as one of the secondary emotions. The SE creates a 16-dimensional space, where each axis represents an emotion.

One important goal of this study is to analyze and compare the effectiveness of the emotional attributes and the emotional categories (e.g., PE and SE) in being able to rank samples in a meaningful way such that triplets created from these rankings can train reliable DNNs to identify sentences with similar emotions.

We use the release 1.2 of the corpus, containing a total of 29,440 sentences. We have manually identified the speaker identity of 21,489 sentences (346 speakers). We use speaker information to set partitions that aims to have speaker independent sets. The test set has 7,341 sentences from 50 speakers. The development set has 2,861 sentences from 20 speakers. The train set has the remaining samples (19,238 speaking turns). To report the inter-evaluator agreement, we use the Fleiss' Kappa for primary emotion. The level of agreement for primary emotions is $\kappa = 0.234$. We use the Krippendorff's alpha for emotional attributes. The level of agreement for valence is $\alpha_{val} = 0.484$, for arousal is $\alpha_{aro} = 0.411$, and for dominance is $\alpha_{dom} = 0.322$. These levels of agreements are similar to the ones observed in other emotional databases with spontaneous speech and with a similar number of emotional classes.

3.2 Acoustic Features

The set of acoustic features used for this study comes from the *Interspeech 2013 Computational Paralinguistic Challenge* (ComParE-13) [65]. First, this feature set extracts a set of *low-level descriptors* (LLDs) such as energy, fundamental frequency and several spectral features. Second, *high-level descriptors* (HLDs) or functional are extracted from the LLDs for each sentence (e.g., mean of energy). The set defines a 6,373-dimensional vector for each speech sentence, regardless of its duration. This vector is the feature input used to train and evaluate our DNN. We use the OpenSMILE toolkit [66] to extract this feature set.

4 Proposed Approach

The purpose of this study is to develop a system that retrieves emotionally-similar speech samples to an anchor recording. For a given anchor, we formulate this problem as a pairwise comparison between two samples, where the task is to identify which of them has emotional content that is the closest to the emotional content of the anchor. The most similar sample is chosen by comparing distances in the emotional label space between the anchor and each respective sample, and choosing the sample that has the shortest distance between itself and the anchor. With this approach, it is straightforward to use a sort algorithm to rank-order recordings as a function of the distance in emotional content from the anchor. The emotional label space is assumed to be unknown during inference, and the goal in this formulation is to automatically learn representations that capture this space. For this purpose, we train a deep neural network with a triplet loss function. The purpose of this function is to make embeddings of emotionally-similar samples close together and embeddings of emotionally-dissimilar samples far apart.

4.1 Triplet Loss Function

A DNN trained with the triplet loss function creates a nonlinear function f that maps acoustic features to a d-dimensional feature space. Therefore, $f(x) \in \mathbb{R}^d$. After applying this function, we want samples that are emotionally similar to the anchor (i.e., positive samples) to be mapped close to one another, and samples that are emotionally dissimilar (i.e., negative samples) to be mapped farther apart in this d-dimensional feature space. Equation (1) shows the ideal scenario,

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

where x_i^a is the acoustic features of the anchor, x_i^p is the acoustic features of the positive sample, and x_i^n is the acoustic features of the negative sample. The parameter β is a margin to push apart positive and negative samples, and Γ is the set of all possible triplets in the training set. Equation (2) shows the triplet loss function.

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

Our results demonstrate that the use of the triplet loss function creates a representation that is effective to represent emotional content for this formulation. While the triplet loss is not novel, the formulation, and the use of this loss to solve this problem are important contributions of this study.

4.2 Emotional Label Space

We must first quantify the concept of emotional similarity to make comparisons between emotional contents in speech sentences. Our approach relies on comparing sentences using the emotional label space, obtained from the annotations. Since it is not clear which is the best emotional descriptor to achieve our goal of retrieving sentences with similar emotions, we consider categorical and attribute-based annotations.

For attribute-based descriptors, we use the three dimensional space defined by the VAD scores (Section 3.1). The score for each emotional attribute is the average value assigned for the sentence across evaluators.

For categorical descriptors, we use the *primary* and *secondary* emotions defined in Section 3.1. Instead of using a one hot vector with the consensus label, we use a soft label including the annotations from all the evaluators. We use labels from multiple annotators to create normalized histograms, providing a richer and dense representation of the emotional content in each sentence. This study evaluates three different approaches to consider and combine primary and secondary emotional labels, aiming to identify the most discriminative space for our formulation.

- The first method only uses the primary emotions for our histogram. Therefore, the histogram only has

eight dimensions. We refer to this method as the *PE* representation.

- The second method weighs the primary and secondary emotions, creating a combined representation. We only use secondary emotions that overlap with the eight emotions considered for the primary emotion set. Therefore, the histogram also has eight dimensions. We refer to this method as the *PSE(8)* representation. We weigh the primary emotions twice as much as the secondary emotions. Equation (3) illustrates the formulation used for the *PSE (8)* space, where $PSE(8)_i$ denotes the ith dimension of the $PSE(8)$ representation, P_i denotes the ith dimension of the primary emotion distribution, S_i denotes the ith dimension of the secondary emotion distribution, and α denotes the weight assigned to the primary emotion, which is set to $\alpha = 2$.

$$PSE(8)_i = \frac{\alpha P_i + S_i}{\sum_{j=0}^{7}(\alpha P_j + S_j)}. \tag{3}$$

- The third method uses all of the secondary emotions in addition to the primary emotions, creating a 16 dimensional histogram. We also weigh the primary emotions twice as much as the secondary emotions ($\alpha = 2$). We refer to this method as the *PSE(16)* representation. Equation (4) illustrates the formulation used for the *PSE(16)* representation, where $PSE(16)_i$ denotes the ith dimension of the $PSE(16)$ representation.

$$PSE(16)_i = concat\left(\frac{\alpha P_i}{\sum_{j=0}^{7}(\alpha P_j + S_j)}, \frac{S_i}{\sum_{j=0}^{7}(\alpha P_j + S_j)}\right). \tag{4}$$

All the histograms are normalized so the sum of their values is one. Before the normalization, we add a small offset to have non-zero values for all the dimensions. This step stabilizes the use of *Kullbach-Liebler Divergence* (KLD) (Section 4.3)

4.3 Triplet Generation

We need to provide meaningful preferences as input in the form of triplets to train and evaluate our models. A triplet consists of an anchor, positive, and negative samples. The anchor and positive sample should have very similar emotions. The anchor and negative sample should have very dissimilar emotions. We sort samples based on the annotation data with respect to an anchor to determine similarity. For the VAD representation, we sort with respect to the euclidean distance in the three dimensional VAD space. For the categorical emotional label representations (PE, PSE(8), PSE (16)), we sort using the KLD between the histograms of the respective sentences. We create a sorted list for *every sample* in the respective partition for which we are generating triplets. For each representation, we create 19,238 lists for the train set, 7,341 for the test set, and 2,861 lists for the development set, since we wish to use each sentence as an anchor.

Fig. 2 illustrates the method to generate the triplets from the sorted list. We randomly draw one of the top 20 sentences on the list to choose the positive sentence for a given anchor. We randomly choose from 20 sentences centered around a certain percentile in the list to choose the negative sentence. The percentile is varied in our evaluation to assess how different the emotion should be for the models to distinguish between the positive and negative samples. We consider drawing negative examples from the 10th, 20th, 25th, 40th, 60th, and 90th percentiles in the list. Intuitively, negative sentences chosen from the 90th percentile of the list should be easier to properly discriminate than those from the 10th percentile of the list, since their emotional content is more dissimilar from the anchor.

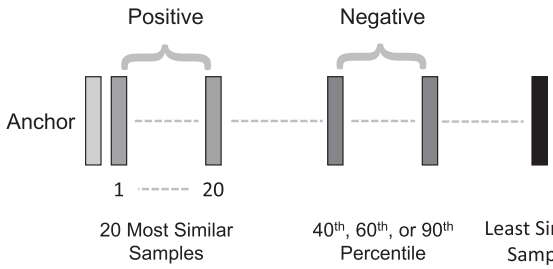


Fig. 2. Process to generate the triplet consisting of an anchor, the positive (similar emotion) and the negative (dissimilar emotion) examples. The process uses a sorted list with the distance from the anchor, where the percentile to select the negative example is varied during the evaluation.

We did not make an effort to ensure that all the positive, negative and anchor samples in a triplet come from different speakers. However, we checked the triplets used in our study, finding that less than 8.8% of the triplets for each of the four spaces include positive and anchor samples coming from the same speaker (VAD, PE, PSE(8) and PSE(16)). Furthermore, we found that there are no statistically significant differences in the performance when these triplets are removed from the test set, indicating that our experimental results are not affected by the speaker information.

4.4 Ensemble Using Multiple Models During Training

Due to local optimization during training, the performance over batches in the development set is noisy. There are points throughout the training when the model can perform well on the development set, while being stuck in a local optimal. Using a single model can be unreliable, especially using an early stopping criterion. Instead, we build an ensemble with the top models in the development set during training. We save the model weights for every 500 batches. After the training is complete, we choose the five best performing models. To increase the diversity in the ensemble, we do not consider models from nearby batches.

During inferences, we find the euclidean distance in the triplet loss embedding between the anchor and one of the competing samples, and between the anchor and the other competing sample for each of the five models. The corresponding distances are added, choosing the sample with the smallest distance. In our preliminary analysis, we find that the ensemble with the five best models has slightly better performance than simply using the best model.

5 Experimental Settings

5.1 Network Structure and Training Settings

We model function f in Equation (1) with a fully connected DNN implemented with seven layers. The model is implemented with batch normalization for the first three layers. The input of the function is the 6,373 dimensional

vector with acoustic features from the speech recording (Section 3.2). The first three layers contain 2,048 nodes, and the next three layers contain 1,024 nodes. The activation function is implemented with *rectified linear unit* (ReLU). The first five layers are implemented with dropout with a rate of p=0.2. The seventh layer is the output embedding, which has 512 nodes, mapping the 6,373 acoustic features into a 512-dimensional emotional representation space. We use this embedding as the output of f to estimate the emotional distance between recordings. The distance in the embedding is measure with the euclidean distance. The value for the variable β in Equations (1) and (2) was empirically set to 0.6. The models are trained in Keras with ADAM optimizer and a learning rate of 0.001. We use Glorot uniform initializer.

To train our networks, we create test sets for VAD, PE, PSE(8) and PSE(16) at the 40th, 60th, and 90th percentiles with 10 triplets per anchor, resulting in 192,380 training triplets per condition. We train the models using each training triplet twice (i.e., two epochs). We noticed that the validation loss flattened after two epochs so we did not add more epochs. The training batch size is an important hyperparameter for maximizing the performance on the validation set. Based on the findings of Wilson and Martinez [67], we experimented with small batch sizes setting the batch size to 10. A small batch size adds a regularization due to the noise in the estimation of the gradient. This approach creates 38,476 batches. We stop training at this point and choose the models that produced the five best validation performances.

5.2 Baselines

The goal of our proposed approach is to take a set of acoustic features from speech samples and map them into an emotional representation space that can be used to determine emotional similarity. To demonstrate the benefits of our method, we compare the results with different baselines. There is no direct method proposed by other studies that we can compare our models given that this is a novel formulation. Instead, we use three general approaches to indirectly solve our novel formulation.

The first general approach consists of predicting or classifying the emotional content, using the SER output to estimate emotion similarity. For the PE, PSE(8), and PSE(16) spaces (i.e., categorical emotions), we first train a simple classifier to predict which of the primary emotions the speech sentences belong to. Then, we take our trained classifier and use the softmax outputs to predict which of two sentences is most similar to an anchor sentence. The classifier is a DNN with three hidden layers. The input of the network is also the 6,373 dimensional feature vector described in Section 3.2. Each layer has 1,024 nodes and uses batch normalization. The input layer uses dropout with a rate of p=0.2, whereas the hidden layers use a dropout rate equal to p=0.5. The hidden layers use the *exponential linear unit* (ELU) as the activation function. The output layer has eight nodes corresponding to the eight primary emotions it predicts, using a softmax activation function. We train our classifier for 50 epochs. To predict the emotional similarity using our classifier, we take the KLD between the softmax outputs of the anchor and the first sample, and between the anchor and the second sample. The smaller the KLD, the more similar the distributions. Therefore, the sentence with the smallest KLD is chosen as the most emotionally-similar sample to the anchor. For the VAD space (emotional attributes), we use the ladder networks proposed by Parthasarathy and Busso [11], [68] to predict valence, arousal, and dominance (i.e., regression models). The model was trained with the train set, optimizing performance on the development set of the MSP-Podcast corpus (Section 3.1). This network also uses the 6,373 dimensional feature vector as input. Then, we find the euclidean distance in the predicted VAD space between the anchor and the first sample, and between the anchor and the second sample. We choose the sample with the smallest distance as the most emotionally-similar sentence to the anchor.

The second general approach estimates feature-level representations, estimating emotional distance in the feature space. We tested the extended version of the *Geneva minimalistic acoustic parameter set* (eGeMAPS) [69] and the ComParE-13 set [65] (Section 3.2). These features sets are used for SER tasks, so their feature representations are considered as appropriate spaces to assess emotional similarities. All of the feature-level representations are extracted using the OpenSMILE toolkit [66]. We used the HLDs to represent the audio so that one speech recording has a single representation regardless of its duration. We calculated the L2 distance of the feature representation to determine the distance between the anchor and the positive and negative samples. We select the sample with the closest distance to the anchor. We use this approach for the VAD, PE, PSE(8), and PSE(16) spaces.

The third general approach consists of estimating model-level representations, quantifying emotional similarity in general audio representations. We tested four additional models designed to extract general audio representations: YAMNet [70] built upon the MobileNetV1 architecture [71], VGGish [72], TRILL [73], and Wav2Vec 2.0 [74]. To train these models, we used the same training set used to create our proposed triplet loss model. All of these models can be trained in an unsupervised manner. We followed the same architecture and training procedure described in the references of these four methods [70], [72], [73], [74]. For the VGGish, YAMNet, and TRILL networks, the input is resampled to 16 khz mono. The input is the log mel-spectrogram computed using the *short-time Fourier transform* (STFT) with a window of 25 ms and a hop of 10 ms. The STFT coefficients are mapped into a normalized 64 Mel bin. According to their original implementations, we cropped the feature into 960 ms segments, with the step size of 480 ms, 960 ms, 170 ms for YAMNet, VGGish, TRILL, respectively. If the last segment has a duration less than 960 ms, we drop the segment. Wav2vec 2.0 takes the raw wav input for its feature encoder, which is normalized to have zero mean and unit variance. The encoder output has a receptive field of 25 ms and a stride of 20 ms. For the VGGish, YAMNet, TRILL, and Wav2Vec 2.0 models, we perform average pooling on the features generated by the models to have a fixed length output. This vector is used as a sentence-level representation of the sample. Using the same approach used for the feature-level representation experiment, we calculated the L2 distance of the sentence-level representation to quantify emotional similarity between the anchor and the positive and negative samples.

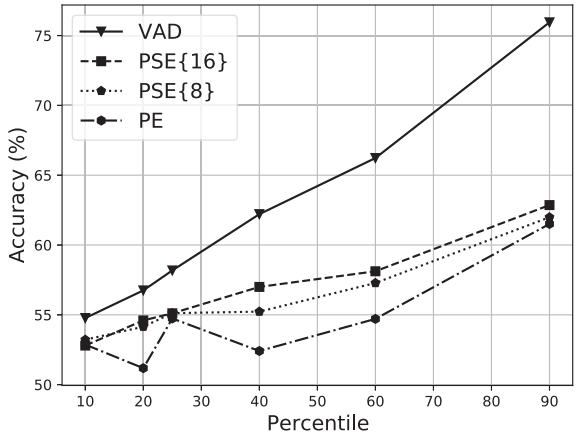


Fig. 3. Global model accuracy when the models are built with emotional attribute space (VAD) and categorical descriptor spaces (PE, PSE(8), and PSE(16)). The results are presented when the negative examples are drawn from the 10th, 20th, 25th, 40th, 60th or 90th percentile conditions.

6 Experimental Evaluation

This section compares the performance of the proposed model using an emotional attributes representation (VAD) and the categorical emotion representations (PE, PSE(8), PSE(16)). The results are evaluated over triplets created on the test sets that are used neither to train nor optimize the performance of the models. For a given triplet in the test set, the binary task in our formulation is to select which of the two recordings (i.e., positive and negative samples) is more emotionally similar to the anchor. We measure performance in terms of accuracy (i.e., percentage of positive samples correctly determined across all test triplets). Performance at chance is 50% given this formulation. We establish significance between the models and their respective baselines by splitting the test set into 100 sets of 734 triplets, estimating the average performance across the 100 sets. Then, we evaluate the differences using the one-tailed two sample proportion t-test, asserting significance at p-value < 0.025.

6.1 Global Performance

Our first evaluation is to analyze the global performance of the proposed models as we vary the selection of the negative samples by drawing negative examples from different percentile in the ranked list.

Fig. 3 shows the performance of our models on the entire test set. This figure shows that representing the emotional content using the emotional attribute space (VAD) leads to better performance for our formulation than using categorical emotion spaces (PE, PSE(8), PSE(16)). The best accuracy is 75.9% , which is obtained using the VAD space in the 90th percentile condition. It is interesting that the concept of emotion similarity between speech sentences is better established for machine-learning using acoustic features with emotional attributes than emotional categories. As expected, the performance increases as the emotional distance between the positive and negative examples increases. Selecting the negative examples from the 90th percentile in the ranked list leads to better accuracy in our models. Our result shows that selecting the negative samples from the 10th and 20th percentiles shows worse performance than selecting from higher percentiles in every emotional label space, indicating that the difficulty of the task increases as the distance between the positive sample and the negative sample gets closer. In the following experiments, we only selected the negative samples from the 40th, 60th and 90th percentile conditions.

TABLE 1
Global Accuracy Using the VAD Space to Create the Triplets

Model	VAD – 90th [%]	VAD – 60th [%]	VAD – 40th [%]
Triplet loss	**75.9***	**66.2***	**62.2***
Ladder	73.7	63.7	60.0
YAMN	56.7	55.0	54.8
VGGish	58.4	55.0	54.9
TRILL	55.4	55.3	55.1
Wav2Vec2	57.6	53.7	52.7
eGeMAPS	56.7	53.8	53.7
ComParE13	51.2	50.6	50.9

The results are presented when the negative examples are drawn from the 40th, 60th or 90th percentile conditions. An asterisk (*) indicates that the performance of the triplet loss approach is significantly better than all of the baselines (one-tailed two sample proportion t-test with p-value < 0.025).

Fig. 3 also shows interesting results when we use categorical descriptors. Adding secondary emotions is particularly useful, leading to consistent improvements in accuracy. There are clear improvements when we compare the PE and PSE(8) spaces, indicating that adding secondary emotions help in finding better representations to assess emotional similarity. We also observe consistent improvements by increasing the emotional space from the 8 dimensional space in PSE(8) to the 16 dimensional space in PSE(16). As more information is added, the representation to assess emotional similarity improves, helping our models to become progressively more accurate. This finding demonstrates that secondary emotions are useful for providing a nuanced understanding of emotional content in speech sentences, providing necessary information in addition to the one provided by primary emotions.

Tables 1 and 2 shows the results when we compare the models with the baseline methods. We highlight higher accuracies in bold, adding an asterisk (*) when the differences in performance achieved by the proposed triplet loss method and the baseline methods are statistically significance. For the VAD space, Table 1 shows significant improvements by using our proposed method over the baselines for all the percentile conditions. The ladder network baseline is a state-of-the-art framework for predicting emotional attributes, leading to a competitive baseline for this novel task. It is interesting that our models are able to improve performance over the baselines, where our models can more robustly discriminate easier triplets. For harder triplets (i.e., 40th percentile condition), the proposed approach achieves 62.2%, which is significantly better than chance.

Table 2 shows the accuracy of the triplet loss methods and baseline approach for the categorical emotional spaces. The overall performances are not as good as the ones reported for the VAD space. Our result shows that the triplet loss model shows the best performance among the baselines for the PSE(8), and PSE(16) conditions. For the PE space, our model shows the best performance for the 90th percentile condition. The exception in these results is for the 40th and 60th percentile conditions, where our model achieves lower performance. The evaluation with the

TABLE 2
Global Accuracy Using the Categorical Description Spaces to Create the Triplets

	PE			PSE(8)			PSE(16)		
Model	90th [%]	60th [%]	40th [%]	90th [%]	60th [%]	40th [%]	90th [%]	60th [%]	40th [%]
Triplet loss	**61.5***	54.7	52.4	**61.9***	**57.2***	**55.2***	**62.8***	**58.1***	**57.0***
Classifier	49.3	49.2	50.7	51.2	50.8	51.0	51.8	51.0	51.3
YAMN	55.7	53.5	53.0	56.1	53.8	52.8	55.8	54.2	54.0
VGGish	55.5	**55.7***	54.5	54.9	53.6	52.8	54.5	53.1	52.6
Trill	53.6	54.3	53.0	53.8	53.5	52.9	54.0	53.6	53.1
Wav2Vec2	55.4	54.2	**58.2***	54.4	52.7	52.2	55.5	52.8	52.8
eGeMAPS	54.0	51.2	51.8	53.8	52.1	52.1	53.7	52.9	52.0
ComParE13	49.3	49.8	50.8	51.4	50.6	51.1	52.3	51.1	51.4

The results are presented when the negative examples are drawn from the 40th, 60th or 90th percentile conditions. An asterisk (*) indicates that one approach is significantly better than all of the other models (one-tailed two sample proportion t-test with p-value < 0.025).

similarity metrics in the feature/model representation spaces demonstrate the benefits of the proposed triplet model approach. The performances of the baseline are slightly above chance, indicating the difficulty of this task. The proposed triplet loss framework is able to reach up to 62.8% accuracy (PSE(16); 90th percentile condition).

6.2 Performance per Region With VAD Space

We examine the results of our models on triplets with anchors from different regions in the VAD space to demonstrate that the performance is highly contingent upon the *difficulty* of the triplets. The regions we evaluate for the emotional attributes are specific volumes in the VAD space. As mentioned in Section 3.1, each emotional dimension was rated in a scale from one to seven. We split each dimension into low (1-3), medium (3-5) and high (5-7) scores creating a $3 \times 3 \times 3$ cube. We only consider five regions within this cube, which we visualize in Fig. 4. Regions 1, 2, 3 and 4 are the corners of this cube in the arousal-valence space, regardless of the dominance value. We consider these regions since arousal and valence are the most common dimensions used in previous studies [75]. The fifth region is the center of the space, where most of the sentences with neutral emotions

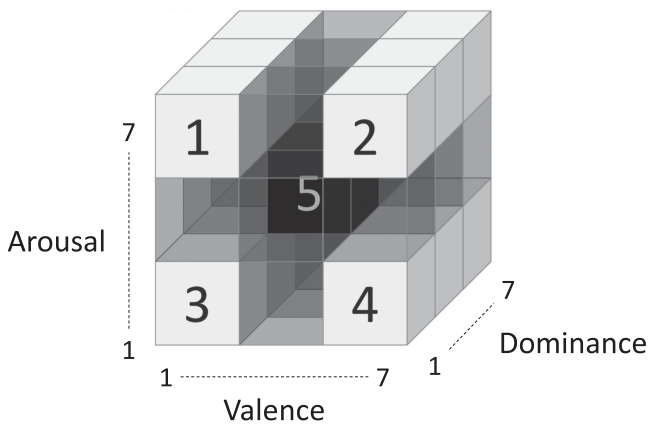


Fig. 4. Regions in the VAD space considered to analyze the performance of the proposed model as a function of the location of the anchors. Regions 1 to 4 include the extreme values for valence and arousal. Region 5 include samples in the middle of the VAD space.

TABLE 3
Accuracy of the Triplet Loss Model When the Anchor Belongs to the Regions in the VAD Space Defined in Fig. 4

	Performance by region in VAD space		
Region	40th [%]	60th [%]	90th [%]
1	68.0	71.5	81.6
2	68.1	68.4	81.0
3	77.7	85.9	93.0
4	79.0	85.0	89.5
5	57.9	62.1	71.4

The results are presented when the negative examples are drawn from the 40th, 60th or 90th percentile conditions.

are located (e.g., all the sentences with all the attributes in the range 3-5).

Table 3 shows the performance of our models for each of the regions in the VAD space. We notice that the performance is around 90% for regions 3 and 4 in the 90th percentile condition (e.g., low arousal). We also observe high accuracy when the anchor is located in regions 1 and 2 (high arousal). These results show that our models are fairly accurate at discriminating different emotional content when the anchor conveys clear emotional information. While triplets with anchors from more extreme regions in the arousal and valence space are fairly easy to discriminate, the triplets from region 5 achieve the worst performance for each percentile condition. Region 5 represents the most difficult triplets, because the anchors come from the center of the VAD space. This location holds triplets with anchors that are fairly neutral in all three dimensions. Since the distribution of the corpus is centered in this region, the actual separation between the positive and negative samples in the emotional space is less than the distance observed when the anchor is in other regions. In fact, the distance in the VAD space between the anchor and a negative sample in an extreme region can theoretically be twice as much as the distance between the anchor and negative sample in the center. The global averages in Table 1 are closer to the performance on region 5 than the performances of regions 1-4, since the data concentrates in the center of the VAD space.

6.3 Performance per Emotion With Categorical Spaces

This section analyzes the performance of our models in further details by grouping the sentences, as a function of their consensus emotional categories (primary emotion). Table 4 shows the results for each categorical emotion by percentile condition. The PSE(16) space is generally the best space for this retrieval task, although the trend is not always consistent across all the emotions and conditions. In the 90th percentile condition, we observe performance over 65% for sadness, surprise and neutral speech. Fear is the emotional class with the worst performance, which is the emotion with the least samples in the corpus (Fig. 1b). The lack of representation in the corpus might explain its poor performance. Across emotional classes, we observe worse performance than models trained in the VAD space, which provides evidences of the superiority of emotional attributes over categorical descriptors for ordinal tasks.

TABLE 4
Accuracy of the Triplet Loss Model When the Anchor Belongs to Each of the Primary Emotions

	PE			PSE(8)			PSE(16)		
Emo	40th [%]	60th [%]	90th [%]	40th [%]	60th [%]	90th [%]	40th [%]	60th [%]	90th [%]
Ang	55.1	59.7	61.4	61.4	63.9	63.2	61.2	65.1	61.4
Hap	52.3	55.1	61.4	57.0	59.4	60.6	60.3	60.2	62.1
Sad	61.4	64.1	69.9	61.2	54.6	61.6	61.9	56.7	69.8
Con	53.8	54.7	59.8	53.9	59.4	58.2	57.4	59.3	59.7
Dis	51.1	54.2	58.2	56.5	56.2	57.4	56.5	58.9	58.9
Fear	51.3	52.3	57.0	53.9	52.3	56.4	50.2	50.2	53.6
Sur	53.3	57.9	63.4	58.4	57.9	59.1	57.7	61.0	65.7
Neu	49.6	52.9	62.7	53.1	55.8	66.8	54.4	55.8	65.3
Avg	53.5	56.4	61.7	56.9	57.4	60.4	57.5	58.4	62.1

The results are presented when the negative examples are drawn from the 40th, 60th or 90th percentile conditions.

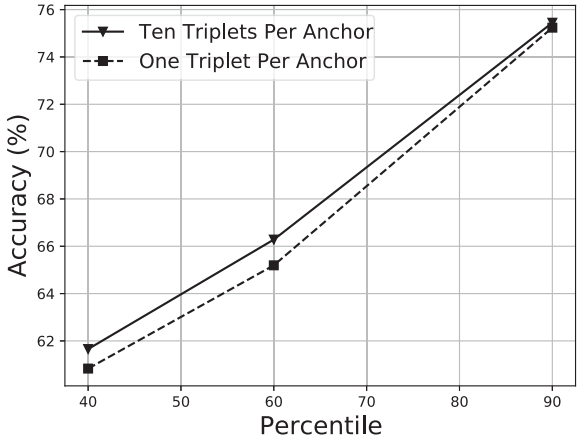


Fig. 5. Evaluation of the benefits in using multiple triplets per anchor. The results are presented when the negative examples are drawn from the 40th, 60th or 90th percentile conditions.

6.4 Benefits of Using Multiple Triplets per Anchor

We use a train set consisting of 10 triplets per anchor in the corpus. For a triplet in the train set, the anchor, the positive sample and the negative sample all belong to the train set. Different triplets per anchor can be easily constructed by using different sentences as positive and negative samples, providing more training examples to build our model. This is one of the benefits of using ordinal formulations in affective computing [53]. This section compares the benefits observed by adding multiple triplets per sentence. The evaluation only considers the VAD space, since it is the emotional space with the best performance in previous evaluations.

Fig. 5 compares the results of models trained with either one or ten triplets per anchor using the VAD space. When training the model with one anchor per sample, we increase the number of epochs to match the total number of batches used for the models trained with ten triplets per anchor. Fig. 5 shows consistent improvements across percentile conditions. The differences between both models are statistically significant for the 40th and 60th percentile conditions. This result indicates that adding multiple triplets per anchor leads to improvements in the accuracy of the models.

6.5 Selection of Positive Samples

We want positive samples that are close to the anchor. We also want some randomness so that, if we choose an anchor, the positive sample is not deterministically defined. We achieve this goal by randomly drawing the positive sample from the top of list (Section 4.3). Notice that we can eventually select more than one triplet per sample if we want to extend the triplets used to train and evaluate the results. A larger number of samples considered in the top of the list increases the probability of selecting triplets with different anchor-positive examples. In this paper, we consider 20 samples in the top of the lists. Given the size of the corpus, this threshold is not very important. We expect that even the top 30 samples will be very close to the anchor given the density of the emotional content in the corpus. To demonstrate this point, we compare the performance of the proposed method by using different thresholds to choose the positive samples for the VAD space. We draw the positive sample from the list with the top 10, 15, 20, 25, and 30 similar samples to the anchor. We estimate the accuracies on the development set to avoid using the test set to pick this threshold. Table 5 shows the results in the VAD space. In most of the conditions, changing the threshold for selecting positive samples does not make significant differences in the accuracy of our proposed approach.

6.6 Computational and Memory Requirements

We denote by M the number of model ensembles (Section 4.4). Our proposed model requires approximately M times more space to store the parameters and approximately M times more computations during inference than the regression and classification baselines. Without using ensembles, our model can have a similar complexity and memory requirement than the baselines. The only difference is in the number of nodes in the output layer, where our model has 512 nodes, while the baseline for the categorical emotion has eight nodes (primary emotions), and the baseline for emotional attributes (ladder network) has three nodes (arousal, valence and dominance). To confirm our analysis, we checked the actual inference time and number of parameters of our model and the classification baseline. Since there are differences in the number of nodes and layers between the triplet loss models and the classification baseline, we set them to the same value for this analysis. The baseline implemented with this modified architecture led to very similar performance than the ones reported in Table 2. With this setting, our classification baseline model has 25,683,410 parameters. The triplet loss model has 131,000,050 parameters with the ensembles, and 26,200,010

TABLE 5
Global Accuracies of the Triplet Loss Model in the VAD Space as a Function of the Number of Samples in the Top of the List to Select the Positive Sample

Threshold	VAD – 90th [%]	VAD – 60th [%]	VAD – 40th [%]
10	74.1	66.1	62.6
15	75.1	66.5	61.2
20	75.9	66.2	62.2
25	75.1	65.9	61.4
30	75.5	66.6	62.3

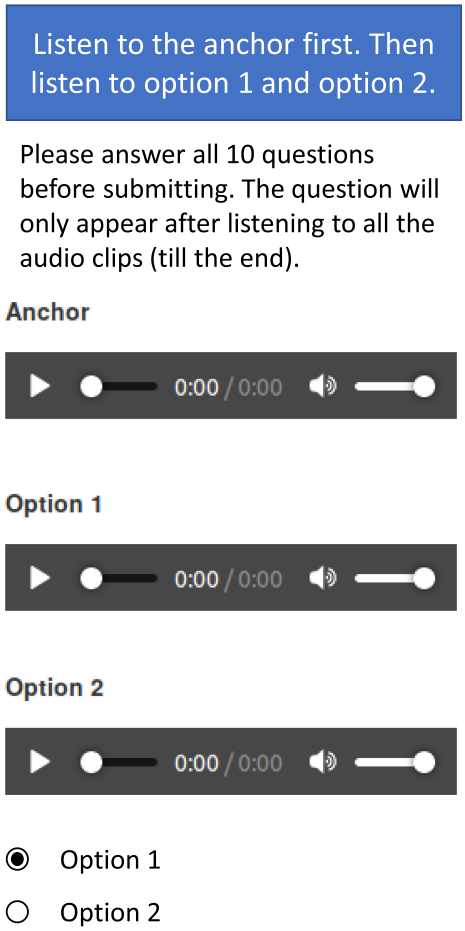


Fig. 6. Interface for the perceptual evaluation using AMT. The worker listen to the anchor and two competing recordings, selecting the file that is more emotionally similar to the anchor.

parameters without the ensembles. For inference time, the classification baseline model takes 0.14ms to compare the similarity for one triplet, and the triplet loss model takes 0.50 ms with ensembles, and 0.11 ms without ensembles. This analysis confirms that our method can achieve similar complexity and memory requirements if we do not include the ensembles. With the ensembles, these values are approximately M times higher.

7 Comparison With Human Performance

As a new formulation in affective computing, we ask how hard is this task for human evaluators? Can people reliably predict which sample is more emotionally similar to a given anchor? How does human performance compare with the accuracy obtained by our models? This section evaluates these questions by using perceptual evaluations, where we ask annotators to perform the task of selecting the more emotionally-similar sentence to an anchor.

The task in the perceptual evaluation is to listen to an anchor audio and two competing audios. They are asked to choose the competing audio that is the most emotionally-similar audio to the anchor. We use *Amazon Mechanical Turk* (AMT) for the evaluation. Fig. 6 shows the interface used for the evaluation. Each *human intelligent task* (HIT) includes ten triplets, which are presented one after the other. To avoid unreliable answers, the option to submit the survey is only activated after a worker has listened to the three sentences (anchor and two competing sentences). For workers to qualify for this task, they need to live in the United States and have a history of acceptance rate above 95% on AMT.

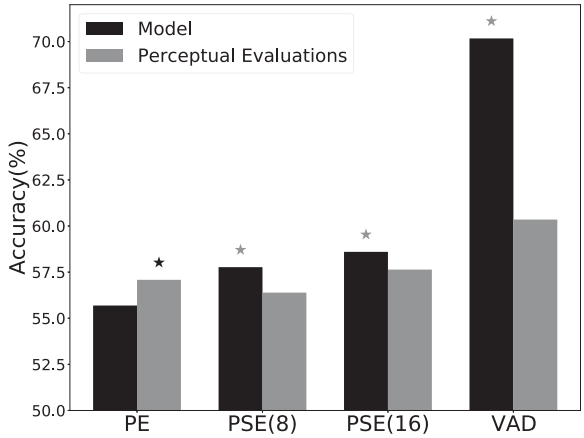


Fig. 7. Global comparison of model and human performance for each of the emotional spaces. We denote with an asterisk (*) on top of the bar when the difference between conditions is statistically significant.

The data considered for this evaluation uses the VAD space and the PE, PSE(8), and PSE(16) spaces. For emotional attributes, we select 10 triplets for each of 19 different cubes that form the VAD space (see Fig. 8), and for each of the 40th, 60th, and 90th percentile conditions. These triplets are chosen randomly from the subset of triplets belonging to each cube. In eight of the 27 cubes, we do not have enough sentences to form the triplets, so we exclude them from our study (i.e., black cubes without numbers in Fig. 8). Therefore, we evaluate 570 triplets for the emotional attributes (10 triplets $\times$ 19 cubes $\times$ 3 percentile conditions). For the categorical emotions, we randomly choose 20 triplets per each of the primary emotions (i.e., the consensus label of the anchor belongs to the primary emotions). We select 20 triplets at the 40th, 60th, and 90th percentile conditions for the PE, PSE(8), and PSE(16) spaces. In total, we evaluate 1,440 triplets for categorical emotions (20 sentences $\times$ 8 emotions $\times$ 3 percentile conditions $\times$ 3 emotional spaces). Our perceptual evaluation is conducted by 262 workers in total. We asked three independent workers to evaluate each triplet. We measured inter-evaluator agreement between workers by using the Fleiss' Kappa metric. The agreement is $\kappa = 0.07$. We also observed low agreement in the perceptual evaluation reported in our preliminary study [27], which was conducted in controlled laboratory environment ($\kappa = 0.15$). The low agreement shows the difficulty of this task for humans. We evaluate the differences between human and model performance using a one-tailed two sample proportion t-test, asserting significance at p-value < 0.025.

7.1 Global Performance of Perceptual Evaluations

We first evaluate the global comparison between human and model performance when using triplets created in the VAD, PE, PSE(8) and PSE(16) spaces. Fig. 7 shows the perceptual evaluation results, which aggregates the results across the 40th, 60th, and 90th percentile conditions. We also include side-by-side the global performance from the models for comparison. Similar to the results observed from our models, creating the triplets in the VAD space leads to better human

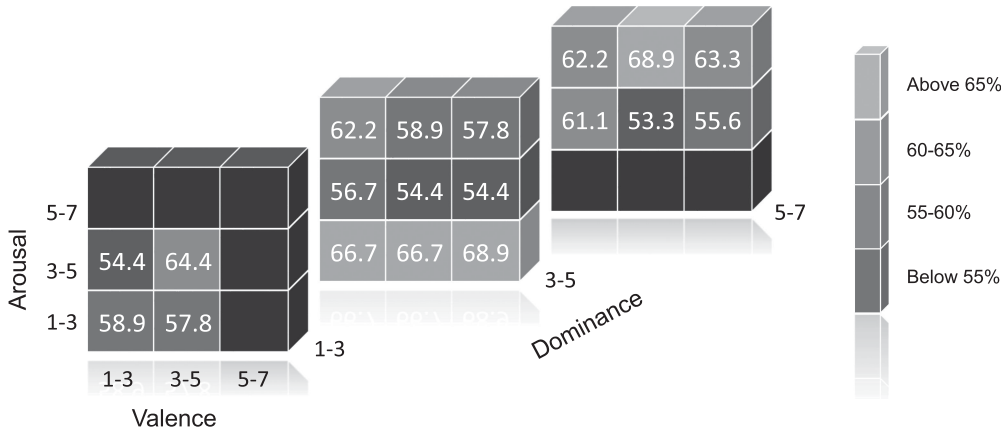


Fig. 8. Detailed results for the accuracy of perceptual evaluations in the VAD space. The number in each cube gives the average accuracy for triplets with anchors in the respective cube. Black cubes without numbers are areas in the VAD space without enough samples in release 1.2 of the MSP-Podcast corpus.

performance. The performance is higher than the accuracy achieved by the triplets created with categorical emotions. Contrary to our models, we do not observe clear improvements by adding secondary emotions. The accuracies using the PE, PSE(8) and PSE(16) spaces are very similar.

The perceptual evaluation results show that finding emotional similarly is a hard task for human workers. In general, we observe that human performance is not as good as the model performance for this task. The accuracies of our models are significantly better than human performance for the VAD, PSE(8) and PSE(16) spaces. Fig. 7 shows that human performance is only better in the PE space.

7.2 Human Performance per Region in VAD Space

This section analyzes human performance for anchors belonging to different regions in the VAD space. Fig. 8 shows the results of the perceptual evaluations on an expanded cube. The numbers on each cube are the average human accuracies for all the anchors belonging to the cubes in the VAD space. Lighter areas indicate higher performance. The general trend is that areas farther from the center of the cube have higher human accuracies. It also appears that dominance is the attribute with the least influence in the results. In contrast, cubes with extreme arousal and valence scores generally have better performance. The best accuracy is 68.9% for the cube with high valence, low arousal, medium dominance. These results match the trends observed in Table 3, confirming that triplets with anchors from expressive regions are easier to discriminate than those from neutral regions, even for human.

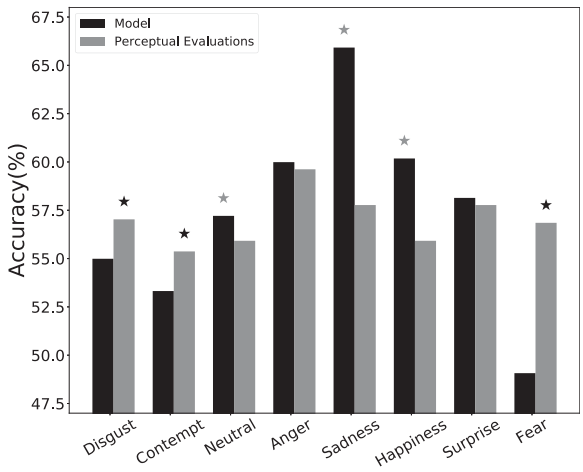


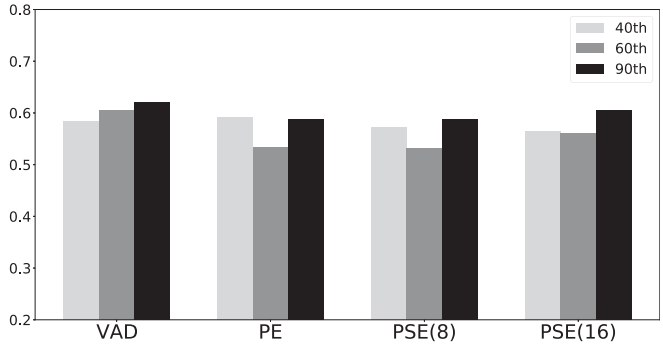
Fig. 9. Global comparison between model and human performances for triplets where the anchor belong to each of the primary emotions. We denote with an asterisk (*) on top of the bar when the difference between conditions is statistically significant.

7.3 Human Performance in Categorical Spaces

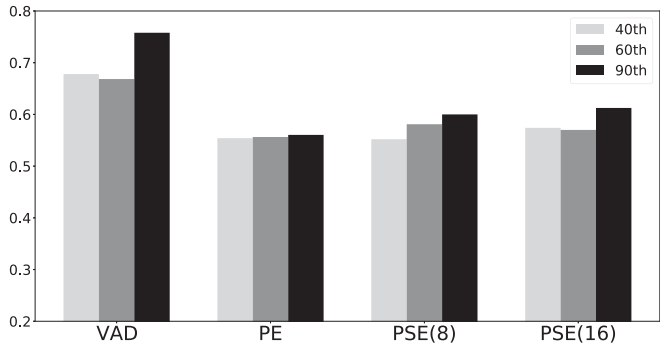
This section further analyzes human performance when the consensus label for the anchor belong to one of the eight primary emotions. Fig. 9 shows the aggregated results across all percentile conditions (40th, 60th and 90th), and emotional category spaces (PE, PSE(8), and PSE(16)). For comparison, we also include the triplet loss model performance. We add an asterisk on top of the bar when one of the conditions is significantly better than the other. For five of the eight emotions, the triplet loss function achieves better results than human performance. The exceptions are only fear, disgust, and contempt. In the human evaluations, we observe that anger has the highest performance, and neutral speech has one of the lowest performances. Anger is a more extreme emotion than neutral speech. Since this emotion is often identified with low valence, high arousal and high dominance, this result further support the consistent observations found with emotional attributes where regions in the extreme of the VAD space have higher performance than the central region.

7.4 Human Performance for Percentile Conditions

This section analyzes the human performance in our emotional similarity task on the triplets when the negative samples are drawn from the 40th, 60th or 90th percentiles in the ranked lists of the anchors. Fig. 10a shows the results from the perceptual evaluations. For comparison, we also include the equivalent results for the performance of the triplet loss model in Fig. 10b. Fig. 10a does not show a clear pattern in the accuracies by human performance observed across percentile conditions. The accuracies of the 90th percentile condition are only slightly higher for the VAD and PSE(16) spaces. However, the trend is not as clear as the one observed in Fig. 10b for the accuracies of our triplet loss models when the emotional separation between positive and negative samples increases in the triplet.



(a) Human Performance



(b) Triplet loss Models

Fig. 10. Global comparison between model and human performances when the negative examples are drawn from the 40th, 60th or 90th percentile conditions.

8 Conclusion

This paper proposed a new formulation in affective computing aiming to identify emotional similarity between speech sentences. The proposed task chooses between two competing speech samples the one that is more emotionally-similar to an anchor. The paper proposed a novel solution based on the triplet loss function that creates a feature embedding that reduces the distance between speech recordings that are emotionally similar, and increases the distance between speech with different emotions. We used an ensemble of DNNs trained with the triplet loss function. The triplets, consisting of an anchor, a positive sample and a negative sample, were created using alternative emotional spaces including emotional attributes and categorical descriptors. The results demonstrated that our proposed triplet loss model outperforms competitive baselines, obtaining improvements in accuracy that were often statistically significant. The experimental evaluation demonstrated that creating the triplets using the emotional attribute space leads to better performance than using spaces created with categorical descriptors. For categorical emotions, we evaluated the benefits of using secondary emotions, finding clear improvements when this information was considered. We also examined the performance of our triplet loss models by region in the VAD space. We observed that triplets where the anchor is more emotional are easier to discriminate than those with anchors in neutral regions of the VAD space.

We performed perceptual evaluations to evaluate and compare how well humans can perform this task. We observed similar trends to those exhibited by our triplet loss models. Humans perform better on triplets created using the emotional attribute space than those created using the categorical descriptor spaces. We also observed that humans perform better on triplets with anchors from more expressive regions on the VAD space, as well as more expressive primary emotions. Interestingly, the triplet loss models showed better accuracies than human performance for this difficult task across conditions, showing the potential of our proposed models. These observations are important as they give grounding to the fact that our models are perceiving emotional content for this task in ways similar to that of humans. It demonstrates that such a representation is meaningful and should be expanded upon in further research.

The novel formulation and proposed framework represent important contributions in the area of affective computing. The study provides further evidences of the benefits of using ordinal methods in speech emotion recognition [14], [15]. The ability to quantify emotional similarity opens important research opportunities enabling exciting applications. For example, we can describe the emotional samples by aggregating the emotional descriptors of samples that are similar to a recording. This type of characterization is richer than assigning a single emotional category or a single score per emotional attribute. While our ensemble method improves the performance reported in our preliminary study [27], there is still room for improvement in terms of accuracy. One possible research direction is to use additional acoustic features or add other modalities for training our networks.

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 Trans. Biomed. Eng.*, vol. 58, no. 3, pp. 574–586, Mar. 2011.

[2] J. Edwards, H. J. Jackson, and P. E. Pattison, "Emotion recognition via facial expression and affective prosody in Schizophrenia: A methodological review," *Clin. Psychol. Rev.*, vol. 22, no. 6, pp. 789–832, Jul. 2002.

[3] A. Rosenfeld *et al.*, "Big data analytics and AI in mental healthcare," 2019, *arXiv:1903.12071*.

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

[5] 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 *Proc. Int. Conf. Intell. User Interfaces Affect. Interact.*, 2005, pp. 7–13.

[6] A. De Vicente and H. Pain, "Informing the detection of the students' motivational state: An empirical study," in *Proc. Int. Conf. Intell. Tutoring Syst.*, S. A. Cerri, G. Gouardères, and F. Paraguaçu, Eds., 2002, pp. 933–943.

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

[8] C. M. Lee and S. S. Narayanan, "Toward detecting emotions in spoken dialogs," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 2, pp. 293–303, Mar. 2005.

[9] 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 *Proc. Int. Conf. Acoust., Speech, Signal Process.*, Kyoto, Japan, 2012, pp. 4157–4160.

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

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

[12] E. M. Albornoz, D. H. Milone, and H. L. Rufiner, "Spoken emotion recognition using hierarchical classifiers," *Comput. Speech Lang.*, vol. 25, no. 3, pp. 556–570, Jul. 2011.

[13] R. Lotfian and C. Busso, "Curriculum learning for speech emotion recognition from crowdsourced labels," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 4, pp. 815–826, Apr. 2019.

[14] G. N. Yannakakis, R. Cowie, and C. Busso, "The ordinal nature of emotions: An emerging approach," *IEEE Trans. Affective Comput.*, vol. 12, no. 1, pp. 16–35, Jan.–Mar. 2021.

[15] G. N. Yannakakis, R. Cowie, and C. Busso, "The ordinal nature of emotions," in *Proc. Int. Conf. Affective Comput. Intell. Interact.*, San Antonio, TX, USA, 2017, pp. 248–255.

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

[17] H. P. Martinez, G. N. Yannakakis, and J. Hallam, "Don't classify ratings of affect; Rank them!," *IEEE Trans. Affective Comput.*, vol. 5, no. 2, pp. 314–326, Jul.–Sept. 2014.

[18] R. Lotfian and C. Busso, "Practical considerations on the use of preference learning for ranking emotional speech," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2016, pp. 5205–5209.

[19] S. Parthasarathy, R. Lotfian, and C. Busso, "Ranking emotional attributes with deep neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process*, 2017, pp. 4995–4999.

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

[21] R. Lotfian and C. Busso, "Retrieving categorical emotions using a probabilistic framework to define preference learning samples," in *Proc. Interspeech*, 2016, pp. 490–494.

[22] H. Cao, R. Verma, and A. Nenkova, "Speaker-sensitive emotion recognition via ranking: Studies on acted and spontaneous speech," *Comput. Speech Lang.*, vol. 29, no. 1, pp. 186–202, Jan. 2015.

[23] Z. Huang and J. Epps, "Detecting the instant of emotion change from speech using a martingale framework," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process*, 2016, pp. 5195–5199.

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

[25] S. Parthasarathy and C. Busso, "Predicting emotionally salient regions using qualitative agreement of deep neural network regressors," *IEEE Trans. Affective Comput.*, vol. 12, no. 2, pp. 402–416, Apr.–Jun. 2021.

[26] S. Parthasarathy and C. Busso, "Defining emotionally salient regions using qualitative agreement method," in *Proc. Interspeech*, 2016, pp. 3598–3602.

[27] J. Harvill, M. AbdelWahab, R. Lotfian, and C. Busso, "Retrieving speech samples with similar emotional content using a triplet loss function," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Brighton, UK, 2019, pp. 7400–7404.

[28] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*, 2015, pp. 815–823.

[29] R. Lotfian and C. Busso, "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings," *IEEE Trans. Affective Comput.*, vol. 10, no. 4, pp. 471–483, Oct.–Dec. 2019.

[30] P. H. Le-Khac , G. Healy, and A. F. Smeaton, "Contrastive representation learning: A framework and review," *IEEE Access*, vol. 8, pp. 193907–193934, Oct. 2020.

[31] J. Bromley *et al.*, "Signature verification using a "siamese" time delay neural network," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 7, no. 4, pp. 669–688, Aug. 1993.

[32] G. Koch, R. Zemel, and R. Salakhutdinov, "Siamese neural networks for one-shot image recognition," in *Proc. ICML Deep Learn. Workshop*, 2015, pp. 1–8.

[33] S. Chopra, R. Hadsell, and Y. LeCun , "Learning a similarity metric discriminatively, with application to face verification," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2005, pp. 539–546.

[34] C. Zhang, W. Liu, H. Ma, and H. Fu, "Siamese neural network based gait recognition for human identification," in *IEEE Int. Conf. Acoust., Speech Signal Process.*, 2016, pp. 2832–2836.

[35] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Mach. Learn.*, H. Daumé III and A. Singh, Eds., Jul. 2020, pp. 1597–1607.

[36] D. Jiang, W. Li, M. Cao, W. Zou, and X. Li, "Speech SIMCLR: Combining contrastive and reconstruction objective for self-supervised speech representation learning," 2020, *arXiv:2010.13991*.

[37] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, "Unsupervised feature learning via non-parametric instance discrimination," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3733–3742.

[38] X. Dong and J. Shen, "Triplet loss in siamese network for object tracking," in *Proc. Eur. Conf. Comput. Vis.*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds., 2018, pp. 459–474.

[39] A. Hermans, L. Beyer, and B Leibe, "In defense of the triplet loss for person re-identification," *in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2017, pp. 354–355.

[40] Z. Ming, J. Chazalon, M. M. Luqman, M. Visani, and J. Burie, "Simple triplet loss based on intra/inter-class metric learning for face verification," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, 2017, pp. 1656–1664.

[41] F. Ren and S. Xue, "Intention detection based on siamese neural network with triplet loss," *IEEE Access*, vol. 8, pp. 82242–82254, 2020.

[42] Y. Qiu, T. Misu, and C. Busso, "Use of triplet loss function to improve driving anomaly detection using conditional generative adversarial network," in *Proc. Intell. Transp. Syst. Conf.*, 2020, pp. 1–7.

[43] C. Li *et al.*, "Deep speaker: An end-to-end neural speaker embedding system," 2017, *arXiv:1705.02304*.

[44] C. Zhang and K. Koishida, "End-to-end text-independent speaker verification with triplet loss on short utterances," in *Proc. Interspeech*, 2017, pp. 1487–1491.

[45] S. Novoselov, V. Shchemelinin, A. Shulipa, A. Kozlov, and I. Kremnev, "Triplet loss based cosine similarity metric learning for text-independent speaker recognition," in *Proc. Interspeech*, 2018, pp. 2242–2246.

[46] H. Bredin, "TristouNet: Triplet loss for speaker turn embedding," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2017, pp. 5430–5434.

[47] P. Wu, S. K. Rallabandi, A. W. Black, and E. Nyberg, "Ordinal triplet loss: Investigating sleepiness detection from speech," in *Proc. Interspeech*, 2019, pp. 2403–2407.

[48] S. J. Bu and S. B. Cho, "Automated learning of in-vehicle noise representation with triplet-loss embedded convolutional beamforming network," in *Proc. Int. Conf. Intell. Data Eng. Automated Learn.*, C. Analide, P. Novais, D. Camacho, and H. Yin, Eds., 2020, pp. 507–515.

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

[50] P. Kumar, S. Jain, B. Raman, P. P. Roy, and M. Iwamura, "End-to-end triplet loss based emotion embedding system for speech emotion recognition," in *Proc. Int. Conf. Pattern Recognit.*, 2021, pp. 8766–8773.

[51] J. Han, Z. Zhang, G. Keren, and B. Schuller, "Emotion recognition in speech with latent discriminative representations learning," *Acta Acustica United Acustica*, vol. 104, no. 5, pp. 737–740, Sep. 2018.

[52] J. Han, Z. Zhang, Z. Ren, and B. W. Schuller, "EmoBed: Strengthening monomodal emotion recognition via training with crossmodal emotion embeddings," *IEEE Trans. Affective Comput.*, vol. 12, no. 3, pp. 553–564, Jul.–Sep. 2021.

[53] K. Feng and T. Chaspari, "A siamese neural network with modified distance loss for transfer learning in speech emotion recognition," in *Proc. AAAI-20 Workshop Affective Content Anal.: Interactive Affect. Response*, 2020, pp. 1–7.

[54] Y. Baveye, E. Dellandrea, C. Chamaret, and L. Chen, "From crowdsourced rankings to affective ratings," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops*, 2014, pp. 1–6.

[55] P. Lopes, A. Liapis, and G. N. Yannakakis, "Modelling affect for horror soundscapes," *IEEE Trans. Affective Comput.*, vol. 10, no. 2, pp. 209–222, Apr.–Jun. 2019.

[56] Y. Yang and H. H. Chen, "Music emotion ranking," in *Proc. Int. Conf. Acoustics, Speech, Signal Process.*, 2009, pp. 1657–1660.

[57] Y.-H. Yang and H. H. Chen, "Ranking-based emotion recognition for music organization and retrieval," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 4, pp. 762–774, May 2011.

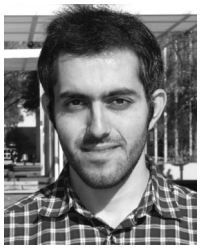
[58] M. Soleymani, G. Chanel, J. J. M. Kierkels, and T. Pun, "Affective ranking of movie scenes using physiological signals and content analysis," in *Proc. ACM Workshop Multimedia Semantics*, 2008, pp. 32–39.

[59] S. Mariooryad, R. Lotfian, and C. Busso, "Building a naturalistic emotional speech corpus by retrieving expressive behaviors from existing speech corpora," in *Proc. Interspeech*, 2014, pp. 238–242.
[60] P. P. Liang, A. Zadeh, and L.-P. Morency, "Multimodal local-global ranking fusion for emotion recognition," in *Proc. ACM Int. Conf. Multimodal Interaction*, 2018, pp. 472–476.
[61] G. N. Yannakakis and H. P. Martinez, "Grounding truth via ordinal annotation," in *Proc. Int. Conf. Affective Comput. Intell. Interact.*, 2015, pp. 574–580.
[62] G. N. Yannakakis and J. Hallam, "Ranking vs. preference: A comparative study of self-reporting," in *Proc. Affective Comput. Intell. Interact.*, S. D'Mello, A. Graesser, B. Schuller, and J.-C. Martin, Eds., 2011, pp. 437–446.
[63] C. Holmgård, G. N. Yannakakis, H. P. Martinez, and K.-I. Karstoft, "To rank or to classify? Annotating stress for reliable PTSD profiling," in *Proc. Int. Conf. Affective Comput. Intell. Interaction*, 2015, pp. 719–725.
[64] A. Burmania, S. Parthasarathy, and C. Busso, "Increasing the reliability of crowdsourcing evaluations using online quality assessment," *IEEE Trans. Affective Comput.*, vol. 7, no. 4, pp. 374–388, Oct.–Dec. 2016.
[65] B. Schuller *et al.*, "The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in *Proc. Interspeech*, 2013, pp. 148–152.
[66] F. Eyben, M. Wöllmer, and B. Schuller, "OpenSMILE: The Munich versatile and fast open-source audio feature extractor," in *Proc. ACM Int. Conf. Multimedia*, 2010, pp. 1459–1462.
[67] D. Wilson and T. Martinez, "The general inefficiency of batch training for gradient descent learning," *Neural Netw.*, vol. 16, no. 10, pp. 1429–1451, Dec. 2003.
[68] S. Parthasarathy and C. Busso, "Semi-supervised speech emotion recognition with ladder networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 2697–2709, Sep. 2020.
[69] F. Eyben *et al.*, "The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing," *IEEE Trans. Affective Comput.*, vol. 7, no. 2, pp. 190–202, Apr.–Jun. 2016.
[70] M. Plakal and D. Ellis, "YAMNet," 2020. [Online]. Available: https://github.com/tensorflow/models/tree/master/research/audioset/yamnet
[71] A. G. Howard *et al.*, "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*.
[72] S. Hershey *et al.*, "CNN architectures for large-scale audio classification," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2017, pp. 131–135.
[73] J. Shor *et al.*, "Towards learning a universal non-semantic representation of speech," in *Proc. Interspeech*, 2020, pp. 140–144.
[74] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. Adv. Neural Inf. Process. Syst.*, Virtual, 2020, pp. 12449–12460.
[75] J. A. Russell and L. F. Barrett, "Core affect, prototypical emotional episodes, and other things called emotion: Dissecting the elephant," *J. Pers. Soc. Psychol.*, vol. 76, no. 5, pp. 805–819, May 1999.

John Harvill (Student Member, IEEE) received the BS degree with high honors in biomedical engineering from the University of Texas at Dallas (UTD) in 2018. He is pursuing his PhD degree at the Electrical and Computer Engineering Department at the University of Illinois at Urbana-Champaign. His research interests include the areas of voice conversion, automatic speech recognition, and emotion recognition.

Seong-Gyun Leem (Student Member, IEEE) received the BS and MS degrees in computer science and engineering from Korea University, Seoul, South Korea, in 2018 and 2020, respectively. He is currently working toward the PhD degree in electrical engineering with the University of Texas at Dallas. His current research interests include speech emotion recognition, noisy speech processing, and machine learning.

Mohammed AbdelWahab (Student Member, IEEE) received the BSc degree in electrical and electronic engineering from Ain Shams University, Cairo, Egypt, in 2010, and the MS degree in electrical engineering from Nile university, Cairo, Egypt 2012. He is currently working toward the PhD degree in electrical engineering with the University of Texas at Dallas. His research interest includes speech signal processing, emotion recognition, artificial intelligence, and machine learning.



Reza Lotfian (Student Member, IEEE) received the BS degree with high honors in electrical engineering from the Department of Electrical Engineering, Amirkabir University, Tehran, Iran, in 2006, the MS degree in electrical engineering from the Sharif University, Tehran, Iran, in 2010, and the PhD degree in electrical engineering from the University of Texas at Dallas. He is currently a research scientist with Cogito Corp, Boston, Massachusetts, USA. His research interests include speech signal processing, affective computing, human machine interaction, and machine learning.

Carlos Busso (Senior Member, IEEE) received the BS and MS degrees with high honors in electrical engineering from the University of Chile, Santiago, Chile, in 2000 and 2003, respectively, and the PhD degree in electrical engineering from the University of Southern California (USC), Los Angeles, in 2008. He is currently an associate professor with Electrical Engineering Department, University of Texas at Dallas (UTD). He was selected by the School of Engineering of Chile as the best electrical engineer graduated in 2003 across Chilean universities. At USC, he received a provost doctoral fellowship from 2003 to 2005 and a fellowship in Digital Scholarship from 2007 to 2008. At UTD, he leads the Multimodal Signal Processing. He is the co-author of the winner paper of the Classifier Sub-Challenge event at the Interspeech 2009 emotion challenge. His research interests include human-centered multimodal machine intelligence and applications, affective computing, multimodal human-machine interfaces, nonverbal behaviors for conversational agents, in-vehicle active safety system, and machine learning methods for multimodal processing. His work has direct implication in many practical domains, including national security, health care, entertainment, transportation systems, and education. He was the general chair of ACII 2017 and ICMI 2021. He is a member of ISCA, AAAC, and a senior member of the ACM. He was the recipient of a NSF CAREER Award, the ICMI 10 Year Technical Impact Award in 2014. In 2015, his student was the recipient of the third prize IEEE ITSS Best Dissertation Award (N. Li). He was also the recipient of the Hewlett Packard Best Paper Award at the IEEE ICME 2011 (with J. Jain), the Best Paper Award at the AAAC ACII 2017 (with Yannakakis and Cowie), and the Best of IEEE Transactions on Affective Computing Paper Collection in 2021 (with R. Lotfian).