# Building Naturalistic Emotionally Balanced Speech Corpus by Retrieving Emotional Speech From Existing Podcast Recordings

Reza Lotfian, *Student Member, IEEE,* and Carlos Busso, *Senior Member, IEEE*

**Abstract**—The lack of a large, natural emotional database is one of the key barriers to translate results on speech emotion recognition in controlled conditions into real-life applications. Collecting emotional databases is expensive and time demanding, which limits the size of existing corpora. Current approaches used to collect spontaneous databases tend to provide unbalanced emotional content, which is dictated by the given recording protocol (e.g., positive for colloquial conversations, negative for discussion or debates). The size and speaker diversity are also limited. This paper proposes a novel approach to effectively build a large, naturalistic emotional database with balanced emotional content, reduced cost and reduced manual labor. It relies on existing spontaneous recordings obtained from audio-sharing websites. The proposed approach combines machine learning algorithms to retrieve recordings conveying balanced emotional content with a cost effective annotation process using crowdsourcing, which make it possible to build a large scale speech emotional database. This approach provides natural emotional renditions from multiple speakers, with different channel conditions and conveying balanced emotional content that are difficult to obtain with alternative data collection protocols.

**Index Terms**—Affective corpus, emotion recognition, expressive speech, information retrieval, emotion ranking

## 1 Introduction

Affective computing is an important research area aiming to understand, analyze, recognize, and synthesize human emotions. Providing emotion capabilities to current interfaces can facilitate transformative applications in areas related to *human computer interaction* (HCI), healthcare, security and defense, education and entertainment. Speech provides an accessible modality for current interfaces, carrying important information beyond the verbal message. However, automatic emotion recognition from speech in realistic domains is a challenging task given the subtle expressive behaviors that occur during human interactions [1]. Current speech emotional databases are limited in size, number of speakers, inadequate/inconsistent emotional descriptors, lack of naturalistic behaviors, and unbalanced emotional content [2], [3], [4]. New advances in machine learning applied to speech processing tasks such as deep learning in *automatic speech recognition* (ASR) have relied on many hours of speech data. The research community does not have the resources to leverage powerful learning algorithms to create robust emotion models. It is important to create a large speech emotional database with naturalistic recordings which can enable transformative advances in the field of affective computing, speech processing and HCI.

Collecting data in real-life condition is a challenging task involving ethical, legal and financial considerations. A popular approach used in early studies relied on actors reading predefined sentences portraying target emotions [5], [6]. However, studies have shown that this approach results in over-emphasized expressions which differ from subtle behaviors observed during daily interactions [4], [7], [8], [9]. An alternative solution is to simulate conversations between two or more speakers collecting spontaneous, rather than read speech. Variations of this technique include creating hypothetical situations (IEMOCAP database [10]), conversation over video conference while completing a collaborative task (RECOLA database [11]) or eliciting emotions with *sensitive artificial listener* (SAL) (SEMAINE database [12], [13]). Although these methods result in more naturalistic speech corpora, they are still costly and time consuming. Some researchers rely on data recorded in uncontrolled setting during natural conversations. Examples of these databases include conversational speech recorded in call centers [7], [8], interaction of kids with robots (FAU-AIBO database [14]), TV talk-shows (VAM database [15]) and media over Internet such as interviews or video blogs [16], [17]. In most of these naturalistic recordings, however, the emotional content tends to be biased by the context and nature of the interaction, reducing the range of emotional behaviors in the corpora. It is important to create a naturalistic database with balanced emotional content.

This paper presents the data collection approach that we are using to create the MSP-PODCAST database. The approach relies on existing naturalistic recordings available on audio-sharing websites. The recent popularity of multimedia content on Internet provides unlimited resources for videos (e.g., YouTube, Vimeo), images (e.g., Flickr, Picasa, Facebook, Instagram), and audio clips (e.g., iTunes, Soundcloud). In particular, we focus on podcasts, which are prerecorded audio programs that can be downloaded or streamed. The key challenge in building an emotional speech corpus from podcasts is to select audio segments with emotionally balanced content, covering the wide spectrum of human emotions. First, we select and download

- *R. Lotfian and C. Busso are with the Erik Jonsson School of Engineering & Computer Science, The University of Texas at Dallas, TX 75080. E-mail: rxl099220@utdallas.edu, busso@utdallas.edu*

*Manuscript received December 30, 2016; revised xxx*

podcasts conveying balanced and rich emotional content. The selected recordings contain natural conversations between many different people over various topics, both positive and negative (e.g., political debates, movie reviews, sport discussions). An important criterion is to download recordings under Creative Commons licenses with less restrictive requirements, so we can share the database to the broader community. The podcasts are segmented into clean, single speaker segments, removing silent segments, background music, noisy segments, or overlapped speech. This process is automatized with algorithms for *voice activity detection* (VAD), speaker diarization, music/speech recognition, and noise level estimation. After selecting candidate speaking turns, we retrieve a set of segments conveying balanced and rich emotional content. We rely on machine learning models trained with existing corpora to retrieve samples with target emotional behaviors as described by arousal (calm versus active) and valence (negative versus positive) attributes. This approach provides control over the emotional content, increases the speaker diversity, and maintains (non-acted) spontaneous nature of the recordings. These segments are emotionally annotated with perceptual evaluations conducted on a crowdsourcing platform. We implement a novel evaluation that tracks the performance of the workers in real-time, stopping the evaluation when their performance drops below an acceptable threshold, as proposed in our previous study [18].

This study conducts proof-of-concept experiments that demonstrate that cross-corpus emotion classification along with crowdsource-based annotations can be effectively used to build naturalistic emotional database with balanced emotional content, reduced cost and reduced manual labor. Although the approach of building affective databases using media content has been previously explored [17], the contributions of this study is the use of machine learning algorithms to retrieve audio clips with balanced emotional content, providing natural stimuli with wider spectrum of emotions. We study different information retrieval methods in the context of emotion detection and compare their performances. The proposed approach relies on automatic algorithms to post-process podcasts and a cost effective annotation process, which make it possible to scale the approach to build a large scale speech emotional database.

The rest of this paper is organized as follows. Section 2 briefly summarizes some of the existing emotional databases. Section 3 explains the method used for the data collection including the selection of the podcasts, segmentation of the podcasts into short turns, post-processing steps, and emotional annotation. Section 4 describes the machine learning frameworks used to retrieve segments to be emotionally annotated. Section 5 reports the quantitative analysis on the emotion distribution of the retrieved database. Section 6 concludes the paper with a summery, final remarks and future directions.

## 2 Related Work

While there are many emotional databases in the community [3], they have clear drawbacks that limit their use to address open research challenges. Table 1 lists some of the representative corpora. The limitations include lack of naturalness, unbalanced emotional content, limited size and limited number of speakers.

TABLE 1
Summary of some of the existing emotion corpora.

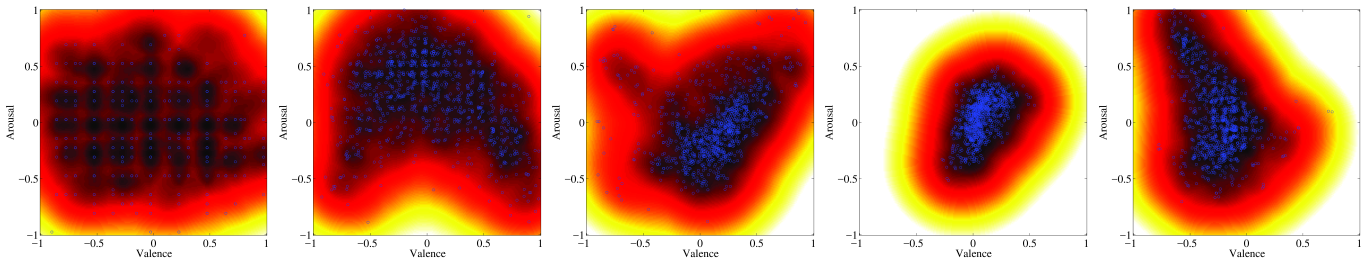
| Corpus | Size | # Spkr | Type | Lang. |
|---|---|---|---|---|
| IEMOCAP [10] | 12h26m | 10 | acted | English |
| MSP-IMPROV [19] | 9h35m | 12 | acted | English |
| CREMA-D [2] | 7,442 samples | 91 | acted | English |
| Chen Bimodal [20] | 9,900 samples | 100 | acted | English |
| Emo-DB [6] | 22m | 10 | acted | German |
| GEMEP [21] | 1,260 samples | 10 | acted | - |
| VAM-Audio [15] | 48m | 47 | spont. | German |
| TUM AVIC [22] | 10h23m | 21 | spont. | English |
| SEMAINE [13] | 6h21m | 20 | spont. | English |
| FAU-AIBO [14] | 9h12m | 51 | spont. | German |
| RECOLA [11] | 2h50m | 46 | spont. | French |

### 2.1 Lack of Naturalness

A common approach in recording emotional databases is the use of actors who are asked to portray emotions while being recorded. In most cases, the actors read a sentence portraying a target emotion. This approach was used in the Emo-DB, CREMA-D, and Chen Bimodal database. A criticism of this approach is the lack of naturalness, as the acted renditions resemble more prototypical behaviors rather than the ambiguous emotional displays observed during daily interactions [23], [24], [25], [26]. Studies have argued that better elicitation schemes can attenuate the problem of using acted renditions for this task [27], [28], [29]. The IEMOCAP and MSP-IMPROV databases are two examples, where emotions were elicited using (1) conversational settings in dyadic interactions, instead of read renditions by a single speaker, and (2) emotion-dependent contextual information which naturally triggers emotions. However, these recordings are still from actors.

### 2.2 Unbalanced Emotional Content

Studies have proposed several approaches to collect emotional databases with more natural interactions (SAL, TV shows, call center). Examples of these approaches include the VAM, TUM-AVIC, SEMAINE, FAU-AIBO and RECOLA databases. In spontaneous scenarios, the recordings do not follow a script and the participants are free to follow the flow of the conversation. However, controlling the emotion content during the recording is not easy. Common daily conversations do not include clear or extreme emotional manifestations, and the vast majority of the recorded samples are emotionally neutral. Furthermore, the recording protocol dictates the emotional behaviors conveyed in the corpus. If the corpus contains discussions between couples, the emotions will be biased toward negative behaviors. If the corpus contains colloquial discussions, the emotions will be biased toward positive behaviors. If the topic of discussion is noncontroversial, the recordings will convey mostly neutral behaviors. As a result, current emotional databases tend to have unbalanced emotional distribution dictated by the contextual scenarios where the corpora were recorded. This is a problem for emotion classification since the training set does not provide representative examples of certain emotional behaviors observed in daily interactions.




(a) IEMOCAP (b) MSP-IMPROV (c) SEMAINE (d) RECOLA (e) VAM

Fig. 1. Dispersion of emotional labels of five emotional speech databases. The color represents the average distance to the 20 nearest neighbor samples from a given point. Dark color represents dense number of samples.

Figure 1 depicts the distribution for the emotional speech samples for five different emotional corpora in the arousal-valence space: IEMOCAP, MSP-IMPROV, SEMAINE, RECOLA and VAM databases (the arousal-valence scores are normalized between -1 and 1). These figures are created as follows. First, we plot each speaking turn in the corpora as a dark point. Then, we color the arousal-valence space by estimating the average distance to the 20 nearest neighbor samples from a given point. Darker colors indicate higher density of samples for that region. To compensate for the differences in the size of the databases, we randomly select 1000 samples from each of the database. Since the VAM database has 947 samples, we randomly repeat some sentences until reaching 1,000 sentences. A well distributed database should cover the entire arousal-valence space with dark colors, indicating the exhaustive coverage of the emotional space. The most balanced databases with this criterion are the IEMOCAP and MSP-IMPROV corpora. As discussed in Section 2.1, these databases were recorded from actors, where the scenarios were carefully selected to elicit target emotions. With corpora recorded without actors, achieving this balance is not easy.

For the SEMAINE, RECOLA and VAM corpora, the figures show large areas with few samples, where the distribution is determined by the scenarios used to record the databases. The SEMAINE database was collected from interactions between a *user* and an *operator*. The operator portrayed a given personality, and his/her goal was to induce emotions on the user. Figure 1(c) shows the distribution in the arousal-valence space for the SEMAINE database. There are very few sentences with negative valence. Most of the sentences are neutral or slightly positive, since inducing stronger emotions in controlled recordings is nontrivial. The RECOLA database includes spontaneous interactions where participants resolved a collaborative task remotely through video conference. Despite the procedure for mood induction, Figure 1(d) shows that most of the sentences have emotionally neutral behaviors as we expect from remote collaborative dyadic interactions. The emotions are mainly positive due to the colloquial interaction between the participants. Figure 1(e) shows that the VAM database mostly covers sentences with negative valence since the recordings come from the TV talk show *Vera am Mittag*, where the participants discuss relationship issues (i.e., fatherhood, affairs, and friendship) [15]. Therefore, even with large number of samples in the corpus, it might not include enough samples for some emotional behaviors.

### 2.3 Limited Size of the Corpora

A key limitation of current databases is the size of the recordings, which prevents using complex machine learning structures (e.g., DNN with multiple layers and nodes). Most emotional databases have few hours of recordings. To the best of our knowledge, the largest emotional corpora are included in Table 1. The IEMOCAP, MSP-IMPROV, TUM AVIC and FAU-AIBO are the only corpora with over nine hours of data [1]. Without overfitting the models, the optimal machine learning structures are likely to produce performances that are not robust enough to transfer these algorithms into real applications. The size of current emotional corpora also limits the advances in ASR to create recognition systems that are robust to expressive speech.

### 2.4 Limited Number of Speakers

Current databases have limited number of speakers. This is an important problem that limits the generalization of emotion classifiers. We express emotions differently, so it is important to include multiple speakers in the corpus to capture the intrinsic inter-speaker variability associated with the expression of emotion. Most emotional corpora have less than 20 speakers. The CREMA-D, FAU-AIBO and Chen Bimodal databases are the only corpora with over 50 speakers [2]. Even for these cases, the size of the corpus per speaker is limited, which prevents reliable studies on the effect of expressive speech in speaker verification tasks. Speaker verification systems require registration data in addition to testing data, which implies that around five minutes per speaker are needed [30]. These corpora are not appropriate for these tasks.

## 3 The MSP-PODCAST Corpus

The limitations of current resources motivate us to propose a framework to record a large, emotionally balanced corpus with natural interactions from many speakers. In our previous work, we introduced the idea of building a naturalistic database from existing speech corpora providing the proof-of-concepts behind this framework [31]. That study relied on

1. We do not know the total duration of the CREMA-D and Chen Bimodal database. The papers only report the number of sentences.

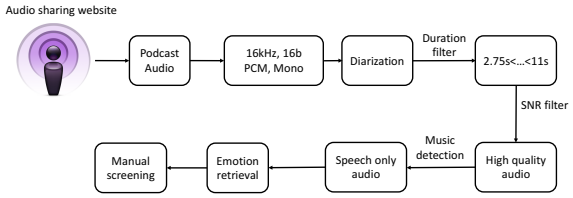



Fig. 2. Block diagram of emotional audio speech collection

the Fisher corpus [32], which consists of telephone conversation between two participants. However, the Fisher corpus is not freely available to the broader community and the interactions were mainly colloquial with very few sentences with negative emotions. Therefore, this study relies on podcasts recordings that are freely available on audio sharing websites. The emotional diversity of the corpus is achieved by selecting podcasts from a variety of topics, increasing the range of emotional behaviors. Although podcasts are not intentionally recorded to convey emotions, like any natural human interaction, the flow of the conversation between the speakers often leads to emotional speech segments. Another positive aspect of many podcasts is the lack of structure which is a characteristic of other radio-like programs. In many cases, the programs lack fixed transcriptions, where common individuals provide almost unlimited, genuine, unscripted recordings. Unlike professional radio hosts, these individuals record more lifelike conversation spanning a broader range of emotions which is appealing for this study. This section describes the proposed approach to collect the MSP-PODCAST corpus, which is summarized in Figure 2.

### 3.1 Selection of Podcasts

The MSP-PODCAST database includes collection of wide range of podcasts recordings downloaded from audio sharing websites. While the proposed data collection protocol is general and can be applied to recordings of any language, all the selected podcasts are in English. Our first criterion to select the podcasts is to include only episodes that can be shared to the broader community. We only choose podcasts that are freely available to the public under Creative Commons under CC-BY or CC-BY-SA licenses. These options are the least restrictive license clauses, allowing us to modify and redistribute the data for either commercial or non-commercial purpose. Podcasts with these licenses usually contain conversations from common individuals, without music segments, since the producers do not have the right to digitally distribute the broadcasted music. This is another advantage for this study, since we consider speech segments without music.

The second criterion to select podcasts is to maintain the naturalness and diversity of the emotional content. We are manually downloading podcasts covering different topics to increase the variety and diversity of the expressive behaviors in the corpus. Their topics include science, technology, politics, economics, business, arts, culture, medicine, lifestyle and sports. Our goal is to include interactions containing broad range of emotions. We also avoid acted recordings, which will affect the naturalness of the corpus. Therefore, we only download non-acted conversations by searching keywords related to conversations, interviews, talk shows, news, discussions, education, storytelling and debates. These interactions tend to elicit natural emotions. This is currently an ongoing effort. For this study, we include 403 podcasts.

### 3.2 Segmentation Process

After selecting the podcasts, the first step is to convert them into a consistent format. All the downloaded podcasts are converted with the software *Sound eXchange* (SoX) to be mono channel, having a sampling rate of 16 kHz, and 16 bit PCM. The podcast recordings are full length programs ranging from three to 190 minutes, including speaking segments and music segments. The recordings have one or multiple speakers. Therefore, it is important to segment the podcasts into short segments using a diarization tool. We have identified an online cloud application called *Speaker Attribution Intelligent Service* [33], which is suitable to identify and track speakers. The output of the system has information about the starting time of the segments, duration of the segments and the speaker ID number associated with the segments. We also manually segmented 105 podcasts. These podcasts are later used to train a classifier that can detect segments with background music, as discussed in Section 3.3.

### 3.3 Selecting Candidate Speaking Turns

Our goal is to consider single speaker segments, without noise or music. Overlapped speech and noisy recordings introduce additional challenges so we avoid these segments. We implement an automatic pipeline to process the turns, selecting only candidate segments that meet our criteria.

The first step consists of selecting segments that are not too long or too short. We emotionally annotate the corpus at the segment level, where each rater assigns an emotional label after listening to a segment (Sec. 3.4). There is a tradeoff in the ideal duration of the segments. If the segments are too long, the emotional content may fluctuate within the segment, so a single label may not be accurate. If the segments are too short, the listeners will have limited information to evaluate the emotional content, producing unreliable labels. Also, selecting short turns will create unreliable features for speech emotion classifications. We balance this tradeoff by considering segments with duration between 2.75s and 11s, discarding turns that are not in this range. While the speech diarization tool should remove non-speech segments, we can still have 2.75s segments dominated by silence. Therefore, we use a separate *speech activity detection* (SAD) algorithm [34], removing samples with short speech activity.

We also remove segments recorded with poor quality, or contaminated by noise. We find the *signal-to-noise ratio* (SNR) of the segments, using the *waveform amplitude distribution analysis* (WADASNR) method [35], where we discard segments with SNR values less than 20dB. We also remove phone quality speech. Audio recorded over the phone has lower sampling rate (8KHz) which affects the acoustic features used to train and test speech emotion classifiers. Therefore, this step also removes segments that do not have significant energy above 4kHz.

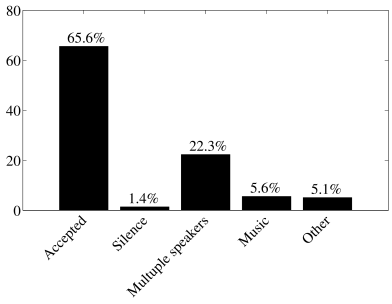



Fig. 3. Histogram of speech turns evaluated manually

We remove segments with music or speech with background music. Since the diarization tool detects any segment that has an identifiable speech pattern, it is not possible to use this information to distinguish speech from speech with background music. Instead, we built a *support vector machine* (SVM) classifier to detect music by analyzing the spectrum of the audio signal represented by pyknogram patterns [36]. This classifier was trained using the samples from the 105 podcasts that are manually segmented. After listening to 500 randomly selected samples detected as music, 63.8% of them included only music, or music in the background. There is room for improvement.

After implementing these steps, we obtain 84,125 speaking turns from the 403 podcasts. These segments form the pool of recordings from which we select the sentences to be annotated. These speech turns are expected to be emotionally unbalanced with many samples that are neutral. We propose to be more selective by automatically evaluating the emotional content of candidate speaking turns with machine learning algorithms, retrieving segments conveying target emotional behaviors. Section 4 describes this step.

The retrieved speech segments are then emotionally annotated (Sec. 3.4). This process is the most expensive step in the process, so it is important that the retrieved samples satisfy our criteria (i.e., single speaker, speech-only turns with duration between 2.75 and 11 seconds). The diarization and music removal tools are not perfect and sometime fail to detect some of the undesired samples. Therefore, we manually check the retrieved samples before uploading them for subjective evaluation. This is the only step in the pipeline that is not automatic. However, the task involves reviewing only the retrieved samples, not the tens of thousands of segments in the pool. Figure 3 shows the percentage of the retrieved samples that were accepted (65.6%). It also shows the reasons for rejection including segment with silence, multiple speakers, music and other reasons (e.g., offensive language, explicit sexual references, use of languages other than English). The most important problem is segments with more than one speaker (22.3%). We are planning to detect overlapped speech to reduce segments with this problem [37].

### 3.4 Perceptual Evaluations using Crowdsourcing

The segments that are approved during the final screening are emotionally evaluated using *Amazon mechanical Turk* (AMT). In our previous work [10], [19], we have annotated sentences with emotional attributes (arousal, valence, dominance) and categorical emotions (i.e., anger, happiness, sadness). Attribute and categorical based descriptors provide complementary information, increasing the potential use of the corpus. We follow a similar approach for this database. Figure 4 shows the questionnaire which has two parts. The first part evaluates the segments with attribute based annotations (Figs. 4(a)-4(c)). We use a seven-point Likert scale to evaluate valence (very negative versus very positive), arousal (very calm versus very active), and dominance (very weak versus very strong). We use *self-assessment manikins* (SAMs) [38], [39] to visually guide the evaluators in annotating these dimensional attributes. SAM provides a lexicon-free assessment tool which simplifies the understanding of the emotional attributes, improving their inter-evaluator agreement. The second part evaluates the segments with categorical labels. First, the evaluators need to choose the primary emotion that best describes the perceived emotion using the following options: anger, sadness, happiness, surprise, fear, disgust, contempt and neutral state. They can also choose *other* when none of the previous options are appropriate to describe the emotional content perceived by the evaluators. The evaluators can only select one option. Second, we annotate secondary emotions similar to Busso et al. [19], where the evaluators can choose all the classes that they perceive in the segment (e.g., sadness + frustration). Speech in naturalistic interactions often convey ambiguous emotions that cannot be described by a single emotion [40]. Therefore, secondary emotions provide complementary information to better describe the emotional content of the segments. We extend the list of emotion including amused, frustrated, depressed, concerned, disappointed, excited, confused, and annoyed. Similar emotional classes are grouped together to reduce cognitive load (Fig. 4(e)).

We rely on a modified version of the crowdsourcing approach proposed in Burmania et al. [18]. We noticed that the inter-evaluator agreement in annotating emotional labels increases when the evaluators, who we refer to as workers, evaluate more than one segment per *human intelligence task* (HIT). The workers can calibrate their assessment by evaluating multiple segments per task [18]. The performance drops when the worker tires or loses interest in the task. In Burmania et al. [18], we proposed to create HITs with multiple segments, where we track in real-time the quality of the workers, stopping the evaluation when the quality drops due to fatigue or lack of interest. In this method, a set of reference sentences which are already evaluated are interlaced with new sentences allowing us to continually assess the quality of the workers, stopping the evaluation when the agreement drop below an acceptable threshold. The approach significantly improves the quality of the annotations [18].

Guided by the lessons learnt in our previous perceptual evaluation, we implement three main changes on the approach. First, we increase the frequency that we include reference sentences. In Burmania et al. [18], we included five reference sentences every 20 new sentences following the pattern [5, 20, 5, 20, 5, 20, 5, 20]. The problem with this approach is that the quality was evaluated on intervals as long as 20 minutes, affecting the temporal resolution to make a decision. Instead, we add one reference sentence

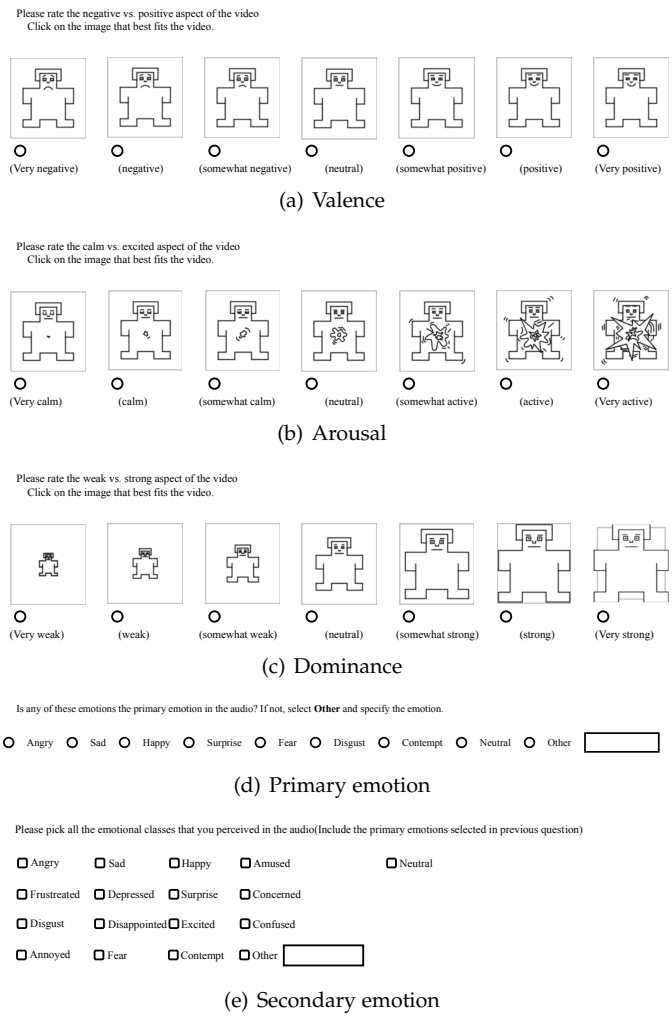



Fig. 4. Questionnaire to annotate the emotional content of the corpus. The evaluators annotate valence, arousal and dominance using SAM. The evaluation also includes primary and secondary emotions.

every four new sentences so we can detect faster when a worker drops his/her performance. In Burmania et al. [18], the overhead due to extra annotations for the reference sentences was 28%. This change reduces the overhead to 20%. Second, the presentation pattern for the four new sentences and the reference sentence is not fixed. Instead, we randomize the placement of these five sentences, making it impossible for workers to guess when they are being evaluated. This approach is resilient to most common attacks in crowdsourcing HITs. Third, the stopping criterion includes not only primary emotions, as in our previous work, but also attribute-based annotations. For primary emotions, we use the angular similarity metric, which transforms evaluations into vectors and estimates whether the angular distance between annotations increases (less agreement) or decreases (more agreement) when an extra annotation is added (see details in Burmania et al. [18]). For attribute-based annotation, we only consider arousal and valence. We estimate the absolute distance between the average score assigned to a reference sentence and the score provided by the worker.

#### 3.4.1 Details of the Implementation

This framework relies on a reference set that is emotionally annotated before the evaluation. We selected segments from our pool, excluding the retrieved sentences. We collected five evaluations per sentence using the questionnaire in Figure 4. This phase was completed by workers who passed a qualification test consisting of annotations of 10 sentences from the IEMOCAP database. We considered annotations of arousal, valence and primary emotions creating thresholds so that about 60% of workers who attempted the qualification test obtained the qualification.

A problem with HITs with qualification tests is the reduced number of workers interested in the tasks, since they are less willing to complete entry tests. Therefore, we only rely on the real-time quality control framework in the evaluation of the retrieved samples. We estimate the average performance metric over the three most recent reference sentences to determine if the evaluation continues or stops for the metrics for arousal, valence and primary emotions. Since we consider three reference sentences, the worker has to complete at least 12 sentences. The decision to stop the evaluation is determined with two thresholds per metric: *average* and *low* performance thresholds. The *average* performance thresholds are set such that the new annotations are as good as the ones in the reference set (inter-evaluator agreement does not change as we add this extra label). The *low* performance thresholds are set so that the new labels are only better than 10% of the annotations in the reference set in term of inter-evaluator agreement. Considering the average metrics for arousal, valence and primary emotion over the three most recent reference sentences, the evaluation stops if (1) any of the metrics is below the *low* performance threshold, (2) two of the metrics are below the *average* performance threshold, and (3) the workers choose to quit anytime after answering the first 12 evaluations. The workers are able to evaluate up to 100 sentences per HIT if they provide high quality annotations.

#### 3.4.2 Training and Payment

After accepting our HIT, workers are redirected to the instruction page where they can learn about the task step-by-step. We created a training video showing how to complete the HIT. This video introduces the concepts of emotional attributes arousal, valence and dominance, which are less familiar to naïve evaluators. The payment of the workers includes a fix rate for the first 12 sentences. Afterward, the workers are paid with bonus for each annotated sentence. The payment for the extra sentences is twice as much as the first 12 sentences to encourage the workers to maintain high performance in their current HIT instead of answering a new HIT. In total, 278 different workers evaluated the set of 2,317 utterances (Table 2), providing at least five evaluations per sentence.

## 4 Retrieving Expressive Behaviors

Since subjective evaluations to annotate the emotional content of sentences requires resources, we need to give priority to a subset of sentences that are more likely to be emotional. By retrieving sentences with target emotions we can control, up to some extend, the emotional content of the corpus,

creating a balanced corpus. This study focuses on balancing the emotional content in terms of attribute-based labels. The attribute dominance is highly correlated with arousal so we do not include this dimension, focusing only on arousal and valence. We prefer attribute-based annotations over categorical annotations in this study due to (1) the diversity on the emotional classes considered across studies [3], and (2) the existence of annotations for arousal and valence in many corpora.

We propose to build emotion detectors by training different models using available emotional corpora to find the most expressive samples among the unlabeled pool of speech samples. We formulate the problem as a retrieval task where the objective is to identify sentences with high and low arousal, and high and low valence. Using these methods, we expect a better balance in the emotion distribution over the arousal-valence space by separately targeting these four problems. Other formulations are possible, but are left as future work (e.g., retrieving categorical emotions [41], finding hotspots [42], [43]).

We evaluate solutions under three machine learning problems: classification, preference learning and regression. Using classification, we use the confidence level of the classifiers to retrieve sentences from the pool that are most likely to be in the target arousal-valence region. Using preference learning, we rank unlabeled samples according to their valence or arousal scores, selecting the sentences at the top (high) and bottom (low) of the sorted list. Using regression, we directly estimate the arousal and valence scores of the sentences, selecting the ones with higher and lower values for each attribute. This section describes different solutions under these three machine learning problems, evaluating their performance with existing emotional corpora. The best algorithms are then used to retrieve the sentences for the MSP-PODCAST corpus.

## 4.1 Classification Based Approaches

### 4.1.1 Support Vector Machine

SVM has been successfully used in emotion classification [44], [45]. Given $n$ training samples $(x_1, y_1), \ldots, (x_n, y_n)$, where $x_i$ is the feature vector and $y_i \in \{-1, 1\}$ is the class label, the canonical formulation of SVM is:

$$\begin{aligned} &\min_{w,\zeta} \frac{1}{n} \sum\nolimits_{i=1}^{n} \zeta_i + \lambda ||w||^2 \\ &\text{subject to } y_i(x_i \cdot w + b) \geq 1 - \zeta_i, \qquad \zeta_i \geq 0 \; \forall i \end{aligned} \tag{1}$$

where $\zeta_i$ is the nonzero slack variable. The algorithm fits a hyperplane in the feature space that maximizes the margin between positive and negative training samples. This approach can be used to address retrieval problems by considering the margin to the hyperplane as a measure of confidence (e.g., $w \cdot x_i - b$). We build separate binary classifiers to detect sentences with low and high level of arousal and valence, retrieving the samples with the largest distance to the hyperplane. The SVM classifiers are built with linear kernel trained with *sequential minimal optimization* (SMO). We use the implementation provided by the LibSVM toolkit [46]. The SVM complexity parameter is set to $c$ = 0.1, following the setting used in previous studies [47]. For each dimension, the positive and negative classes are separated by defining a margin that balances the size of classes. The margins are 0.5 for arousal and 0.4 for valence following the results from Lotfian and Busso [48].

### 4.1.2 Bayesian-Optimal Classifier for Dichotomized Labels

The formulation in Section 4.1.1 requires to dichotomize the numerical labels for arousal and valence into two categorical classes. Binary labels $Q_Y$ are derived from a continuous attribute $\mathbf{Y}$ by dichotomizing it to upper half $\mathcal{U}$ (positive class) and lower half $\mathcal{L}$ (negative class). This common practice in emotion recognition has limitations which we have discussed in Mariooryad and Busso [49]. Under Gaussian assumptions, we proposed a *Bayesian-optimal classifier for dichotomized labels* (BOC-DL) by considering the original continuous scores $\mathbf{Y}$ [49]. If we denote the random variable for the feature vector as $\mathbf{X}$ (i.e., $x_i$ is one realization of $\mathbf{X}$), the classification task is reduced to

$$Label(X) = \begin{cases} \mathcal{L}, & \text{if } \mathbf{\Sigma_{YX}\Sigma_{XX}}^{-1}\mathbf{X} < 0. \\ \mathcal{U}, & \text{if } \mathbf{\Sigma_{YX}\Sigma_{XX}}^{-1}\mathbf{X} \geq 0. \end{cases} \tag{2}$$

where $\mathbf{\Sigma_{YX}}$ is the cross-covariance matrix of the feature vector $\mathbf{X}$ and continuous labels $\mathbf{Y}$, and $\mathbf{\Sigma_{XX}}$ is the covariance matrix of $\mathbf{X}$. The confidence for this classifier is given by $\mathbf{\Sigma_{YX}\Sigma_{XX}}^{-1}\mathbf{X}$, which we use to sort the unlabeled set selecting the sentences with the highest and lowest values.

## 4.2 Ranking Based Approaches

Preference learning [50], [51] provides ideal frameworks to rank sentences according to a given emotional dimension [52]. While preference learning has been used in multiple applications in information retrieval [53], its application in affective computing is limited [41], [43], [48], [52], [54], [55], [56].

Preference learning algorithms are trained with pair of samples with preference relationship with respect to a given metric (i.e., sentence A has higher arousal than sentence B). These relative labels can be derived from existing attribute-based annotations by selecting a safe margin needed to establish a distinctive preference between two sentences. We follow the practical considerations for emotion ranking discussed in Lotfian and Busso [48]. After scaling the attribute values in the range $[-1, 1]$, we select a threshold of 0.4 for valence and 0.5 for arousal. Pair of samples separated by these thresholds are used to train the proposed preference learning algorithms.

### 4.2.1 Rank-SVM

Ranking SVM is an extension to rank samples instead of assigning categorical classes [57]. The formulation is very similar to regular SVM. Let's assume that sample $i$ is preferred over sample $j$ (e.g., $x_i \succ x_j$). The task is to find a hyperplane such that $w \cdot x_i > w \cdot x_j$. This is equivalent to solving the binary classification problem $w \cdot (x_i - x_j) > 0$, where $x_i - x_j$ is just the difference between the feature vectors of the samples. If we have $M$ pairwise comparisons as labels, where $x_i$ and $x_j$ are the $k$ pair in the training set, the formulation for Rank-SVM is:

$$\min_{w,\zeta} \quad \frac{1}{2}\|w\|^2 + C\sum\nolimits_k \zeta_k$$
$$\text{subject to } z_k\langle w,(x_i - x_j)\rangle \geq 1-\zeta_k, \quad \zeta_k \geq 0 \text{ for } k \in 1\ldots M \tag{3}$$

where $z_k = +1$ or $z_k = -1$ depending on the preference relation between $x_i$ and $x_j$, $\zeta_k$ represents the nonzero slack variable, and $C$ is the trade-off parameter. Similar to the approach used in SVM, the weight vector $w$ is determined to maximize the margin between samples [58]. Since this function is linear, the value of $\langle \hat{w}, x\rangle$ can be used to estimate the ranking of unlabeled samples. We employ the RankSVM toolkit described by Joachims [59].

#### 4.2.2 Gaussian Process Ranking

*Gaussian process ranking* (RP-rank) is a probabilistic kernel approach for preference learning [60]. It has been shown that this method has advantages over Bayesian methods for model selection and probabilistic prediction. The approach uses an unobservable latent function $f(x_i)$ that respects preference relations. If $x_i$ is preferred over $x_j$, then $f(x_i) > f(x_j)$. The approach assumes that this latent function is a realization of a Gaussian process with zero mean and covariance matrix $\Sigma$, where its $ij$-th entry is

$$K(x_i, x_j) = \exp(-\frac{\kappa}{2}\|x_i - x_j\|^2), \tag{4}$$

where $\kappa > 0$. Under this formulation, the prior probability of the latent function is given by:

$$P(\boldsymbol{f}) = \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}}\exp(-\frac{1}{2}\boldsymbol{f}^T\Sigma^{-1}\boldsymbol{f}), \tag{5}$$

where $\boldsymbol{f} = [f(x_1), f(x_2), ..., f(x_n)]^T$. Chu and Ghahramani [60] proposed to estimate a likelihood function $P(\mathcal{D}|\boldsymbol{f})$, where $\mathcal{D}$ is the pairwise preference relation between the samples in the training set. The model assumes that the latent function is not perfect, adding a Gaussian noise with zero mean and unknown variance. Using Bayes' theorem, they derived an expression for $P(\boldsymbol{f}|\mathcal{D})$, which is used to estimate the latent function $f$, using *maximum a posteriori estimation* (MAP). The estimated function is used to define preference between samples in the testing set. The study of Chu and Ghahramani [60] provides further details on this method.

### 4.3 Regression Based Approaches

#### 4.3.1 Support Vector Regression

SVR performs linear regression in a high-dimension feature space using a similar formulation to SVM, where the key difference is the constraints in the optimization (Eq. 6),

$$\min_{w,\zeta} \quad \frac{1}{2}\|w\|^2 + C\sum\nolimits_i \zeta_i + \zeta_i^*$$
$$\text{s.t: } y_i - f(x_i, w) \leq \zeta_i^* + \epsilon, f(x_i, w) - y_i \leq \zeta_i + \epsilon \tag{6}$$

where $\zeta_i$ and $\zeta_i^*$ are (non-negative) slack variables that allow the model to converge to a feasible solution [61]. For retrieval, the unlabeled samples are sorted according to the predicted attribute values, selecting the top and bottom samples in the sorted list.

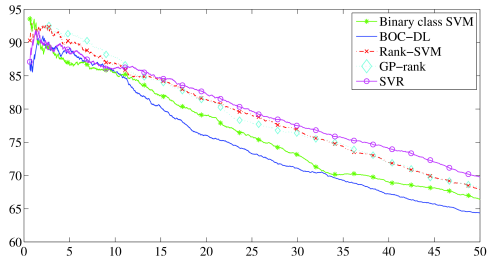
### 4.4 Performance in Existing Corpora

We evaluate the performance of the five different machine learning frameworks using existing emotional corpora: the *interactive emotional dyadic motion capture* (IEMOCAP) database and the MSP-IMPROV database.

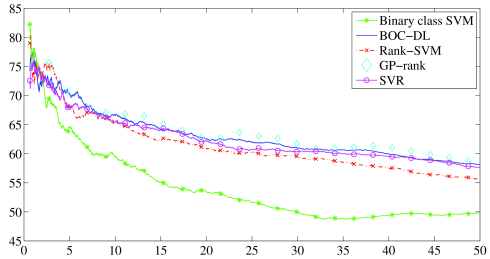
IEMOCAP Database [10]: This corpus was collected at the *University of Southern California* (USC). Ten actors participated in five dyadic interactions using scripted plays and spontaneous improvisations that were carefully selected to elicit happiness, anger, sadness and frustration. Other emotions were also elicited as dictated by the dialogs between the actors. These two techniques are rooted in well-established theories and methods of theater, providing emotional manifestations closer to natural interactions [27]. The corpus contains approximately twelve hours of data, which was manually segmented, transcribed and emotionally annotated with categorical (3 raters) and attribute-based (2 raters) labels at the turn level. For attribute-based labels, the corpus have annotations for arousal (calm versus active), valence (negative versus positive) and dominance (weak versus strong), using a five-point Likert scale. This study considers turns in which three independent evaluators reached majority vote agreement on categorical labels. Turns with overlapped speech were excluded from the experiments resulting in 4,784 speaking turns. Further information about the database is provided in Busso et al. [10].

The MSP-IMPROV Database [19]: This corpus is an acted corpus of dyadic interactions collected at *The University of Texas at Dallas* (UTDallas). The corpus was collected to explore emotion perception with audiovisual stimuli with congruent and conflicting emotional content (e.g., happy voice and angry face). To achieve this goal, hypothetical scenarios were created for a set of target sentences. Twelve actors improvised these scenarios in six dyadic sessions with the goal of uttering the target sentence. By adding emotion-dependent contextual information, the corpus provides conversational renditions of these target sentences conveying different emotions (happiness, anger, sadness and neutral speech). An interesting aspect of the corpus is the inclusion of all the speaking turns that led one of the actor to utter the target sentence. Furthermore, the corpus contains all the natural interactions between the actors during the breaks. For comparison, eight actors also recorded read renditions of the target sentences portraying target emotions. This study uses the entire corpus which comprises 8,438 turns (over 9 hours) of emotional sentences. The emotional labels of the corpus were collected with perceptual evaluations using crowdsourcing [18]. We discuss the annotation process in more details in Section 3.4. The annotations included values for arousal, valence and dominance using a five-point Likert scale. More details on this corpus are provided in Busso et al. [19].

For this analysis, we combine the IEMOCAP [10] and MSP-IMPROV [19] databases, creating two partitions, one for training (50%) and one for testing (50%) the models. The partitions are speaker independent where all the data from a given speaker belongs exclusively to one of the sets. The total number of sentences is 13,222. The features from both databases are normalized subtracting the mean and




(a) Arousal




(b) Valence

Fig. 5. Precision at $K$ of different retrieval methods: BOC-DL, SVM-rank, Gaussian-process rank (GP-rank), binary class SVM, support vector regression (SVR).

dividing by the standard deviation.

The goal is to retrieve samples with high and low arousal, and samples with high and low valence. The feature set for this experiment is the extended version of the *Geneva minimalistic acoustic parameter set* (eGeMAPS) [62], extracted with OpenSMILE [63]. The set contains prosodic, spectral and voice quality features that were carefully selected based on their potential as speech markers for affective changes, successful use in previous studies, and their theoretical significance. The reduced size of the set removes the need for feature selection, facilitating the reproduction of our results by other research groups.

We follow the approach used in Lotfian and Busso [48] to consistently assess performance across the five machine learning algorithms. For each emotional attribute, the testing set is split into two classes with the same size (e.g., low and high arousal; low and high valence). We consider a success if the retrieved sample is in the correct side of the split (e.g., a retrieved sample for low valence is correct if its value is below the median value). We use the precision at $K$ (P@K) metric, which is widely used in information retrieval. It measures the precision rate when we retrieve $k\%$ of the samples. For this problem, we measure P@K by considering the first $k$ percent of the utterances listed either on the top $k\%$ of the list (high arousal/valence) or on the bottom $k\%$ of the list (low arousal/valence). For example, a P@10 equals to 87% for arousal indicates that when you retrieve 20% of the samples (10% for low and 10% for high arousal), 87% of these samples belong to the correct side of the split. Notice that P@50 includes all the testing data, so its value is equivalent to accuracy in binary classification problems.

Figure 5 shows P@K for arousal and valence for the five machine learning algorithms considered in this study. For arousal, Figure 5(a) shows that SVR, Rank-SVM and GP-rank are very competitive with P@20 above 80%. For valence, Figure 5(b) shows that the precision rates are lower. GP-rank and Rank-SVM, which are ranking-based methods, have higher performance. We also observe good performance with the BOC-DL method. Considering these results, we consider BOC-DL, GP-rank, and SVR, which are the best methods for classification, preference learning and regression, respectively. We employ these classifiers to retrieve emotional samples from the unlabeled segments.

TABLE 2
Number of sentences retrieved under different settings. Due to overlapped sets, the total number of distinct sentences is 2317.

| Set | # Turns | Retrieval approach |
|---|---|---|
| High Arousal | 200 | GP-rank |
| Low Arousal | 200 | GP-rank |
| High Valence | 200 | GP-rank |
| Low Valence | 200 | GP-rank |
| High Arousal | 200 | SVM regression |
| Low Arousal | 200 | SVM regression |
| High Valence | 200 | SVM regression |
| Low Valence | 200 | SVM regression |
| High Arousal | 200 | BOC-DL |
| Low Arousal | 200 | BOC-DL |
| High Valence | 200 | BOC-DL |
| Low Valence | 200 | BOC-DL |
| Random | 100 | Random |
| Total | 2317 | |

## 5 Analysis of the Emotional Content

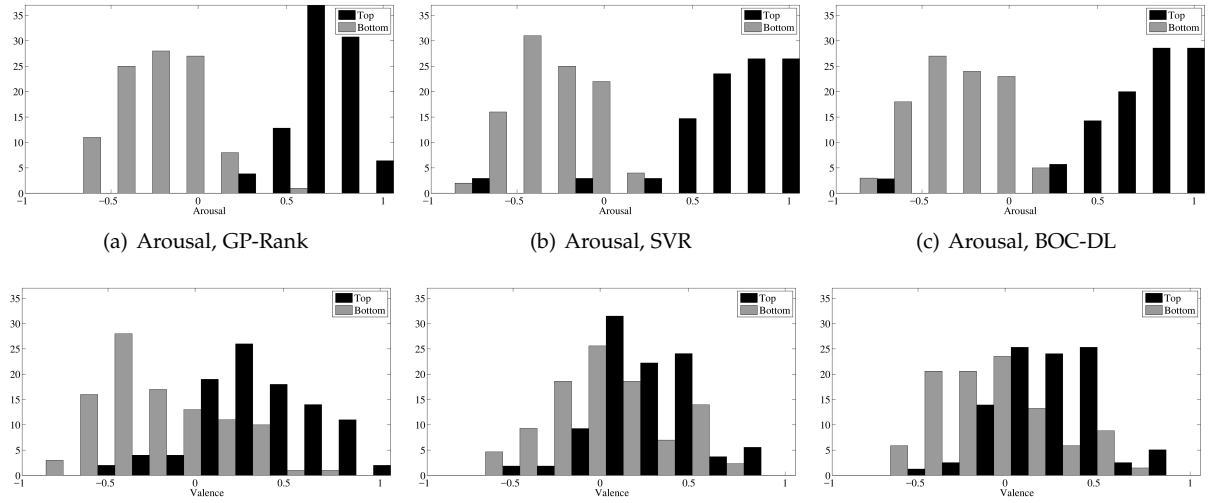
### 5.1 Retrieving Emotional Segments

We evaluate the BOC-DL, GP-rank, and SVR methods on the 84,125 unlabeled sentences extracted from the podcasts. We train these framework with all the data using the eGeMAPS feature set. For each machine learning method, we select 200 sentences for each of the four problems (low and high arousal, low and high valence). As explained in Section 3.3, these retrieved sentences are manually evaluated to correct errors in our pipeline to create the segments (e.g., music, noise, more than one speaker). Rejected segments are replaced by the next segments in the sorted lists until reaching 200 sentences per condition. We also retrieve 100 sentences at random to demonstrate the benefits of using machine learning to build the corpus. Table 2 summarizes the different settings and the corresponding number of turns to create the corpus. Some of the segments were retrieved by more than one method, so the total number of sentences is 2,317. The retrieved samples are then emotionally evaluated using the crowdsourcing-based perceptual evaluation described in Section 3.4.

### 5.2 Emotional Content of Retrieved Sentences

The samples are retrieved to span the arousal-valence space. Therefore, the analysis mainly focuses on the scores provided for arousal and valence. Each segment was evaluated by five workers, where the global score corresponds to the average value assigned to the segment. We linearly map the scores in the range -1 to 1.

Figure 6 shows the histograms for arousal and valence for the retrieved samples for the three machine learning methods. Gray bars provide the histograms of the retrieved samples for either low arousal or low valence. The black




(a) Arousal, GP-Rank (b) Arousal, SVR (c) Arousal, BOC-DL

(d) Valence, GP-Rank (e) Valence, SVR (f) Valence, BOC-DL

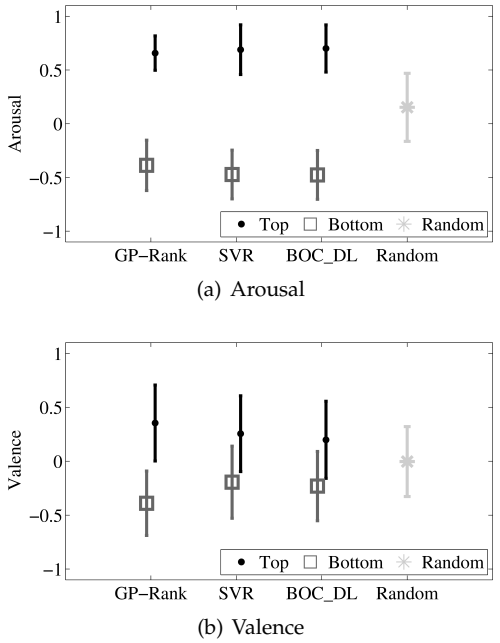
Fig. 6. Histogram of arousal and valence scores assigned to the samples retrieved by the machine learning algorithms. Gray bars correspond to samples in the bottom of the sorted lists, and black bars correspond to samples in the top of the sorted lists.

bars provide the histogram of the retrieved samples for either high arousal or high valence. The figure shows that the approaches were very effective for arousal, creating clear separation between sentences with low and high arousal. For valence, there is a difference between low and high valence, but the separation between these classes is not as clear as the ones for arousal. Predicting valence from speech is a hard problem since there are few acoustic parameters that are discriminative for this attribute [64]. Figure 7 shows error plots with the average and standard deviation of the arousal and valence scores assigned to the retrieved sentences. This figure confirms the results from the histogram, where we clearly observe the effectiveness of the proposed methods in retrieving emotional samples with low and high values of arousal and valence.

Figure 7 also shows the mean and standard deviation of the scores assigned to the sentence retrieved at random. The figure shows that the mean value is close to zero for arousal and valence. Most of the sentences in the podcasts are emotionally neutral which confirms our initial observations. The figure shows that when machine learning algorithms are used, however, we are able to retrieve emotional sentences validating our approach.

Figures 6 and 7 show that GP-Rank is the method that creates the best separation between the retrieved sentences for low and high valence. The approach also has competitive performance for arousal, showing the potential of preference learning for this task. Figure 8 presents the scattering plot for the sentences retrieved with the GP-rank method. The distribution of the samples indicates that the arousal ranker selects high and low arousal samples that are symmetrically distributed in terms of valence. For valence, however, the retrieved segments tend to have high arousal scores. An area for future research is to retrieve sentences with arousal and valence scores contained in a target area of the arousal-valence space (e.g., low valence + low arousal).

Figure 9 depicts the dispersion map of the emotional speech segments retrieved by the machine learning algorithms. We follow the same approach used in Figure 1

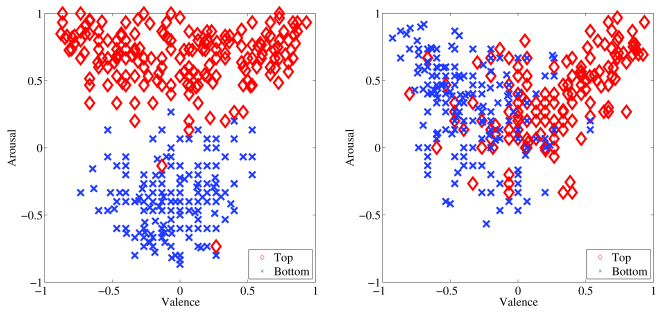



(a) Arousal

(b) Valence

Fig. 7. Mean and the standard deviation of samples retrieved by the machine learning algorithms. The figure shows the results from the samples in the top and bottom of the sorted lists for arousal and valence.

(Sec. 2.2), where we randomly select 1,000 sentences, so the results are comparable. The MSP-PODCAST corpus has better diversity than databases collected without actors (SEMAINE, RECOLA, and VAM corpora). The corpus has smaller areas with lighter color which indicates that we have samples across most of the arousal-valence space. The proposed approach can select naturalistic recordings with balanced emotional content spanning broader range of emotions.




(a) Arousal (b) Valence

Fig. 8. Distribution of retrieved samples for arousal and valence in valence-arousal plane retrieved with GP-rank method. For each dimension, different colors indicate the samples are intended to be perceived as high or low.

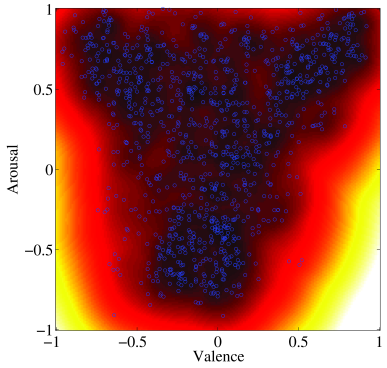



Fig. 9. Dispersion of emotional content in the retrieved speaking turns.

### 5.3 Reliability of Emotional Annotations

Table 3 shows the reliability of emotional annotations obtained with the crowdsourcing based evaluation. For attribute based annotations, we assess inter-evaluator agreement using Krippendorff's alpha coefficient. The values are all above 0.4. Valence scores have the highest inter-evaluator agreement, reaching $\alpha = 0.459$. For primary emotions, we measure inter-evaluator agreement using Fleiss' kappa. The agreement is lower than the ones reported in our previous work [18]. The key difference is the extended emotional classes considered in this evaluation. We previously considered anger, sadness, happiness, neutral state, and other. In this study, we consider these classes in addition to surprise, fear, disgust, and contempt. Adding more options reduces the inter-evaluators agreement, as the separation between the classes decreases.

### 5.4 Analysis of Primary Emotions

Figure 10 shows the distribution of the primary emotions assigned to the retrieved samples. For this purpose, we aggregate the labels assigned by the workers using majority vote rule. While the distribution for arousal and valence

TABLE 3
Reliability of annotations

| | Primary emotion | Arousal | Valence | Dominance |
|---|---|---|---|---|
| $\alpha$ | 0.229 | 0.426 | 0.459 | 0.402 |

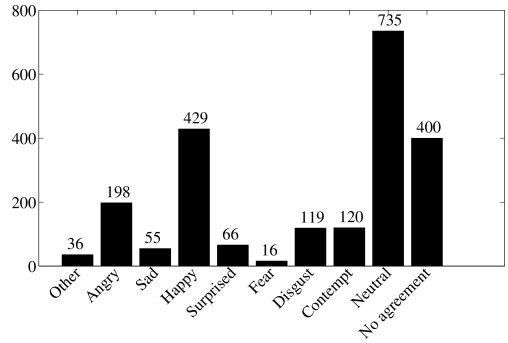



Fig. 10. Distribution of primary discrete emotions of retrieved samples by all three methods. Labels are assigned based on the majority vote consensus. Label *No Agreement* indicates that majority agreement does not exist for those samples.

are balanced, the distribution for emotional categories is less balanced. The categories with more samples are neutral state, happiness and anger. We have fewer samples labeled for fear and sadness. There are also many samples without agreement. We consider the unbalance distribution of primary emotions as an opportunity to investigate retrieval systems to identify speech sentences with a given emotional class (e.g., angry ranker) [41].

### 5.5 Scalability of the Approach

The proposed approach provides a systematic framework to build a large-scale database for speech emotion research. Using the proposed methodology, we have emotionally annotated 18,238 sentences (27h,42m) of emotional speech by increasing the number of podcasts (920), increasing the size of the reference set, and using multiple machine learning algorithms to retrieve target emotional behaviors [41], [42], [43], [48], [56], [65].

The identities of the speakers are not directly available, which is important for emotion recognition (e.g., to keep speaker independent partitions to train and test emotional models), and for evaluating the effect of emotional speech on other speech tasks such as speaker verification [30]. To address this issue, we are manually annotating the identity of the speakers on the corpus. We have identified the speaker information in 9,670 sentences, which belong to 151 speakers. From the 2,317 sentences retrieved in this study, we have speaker information for 1,583 sentences recorded by 83 speakers. The database includes recordings from multiple speakers, which is an important feature of the corpus. To the best knowledge of the authors, this corpus is the largest speech emotional corpus with natural interactions reported in the community. This is an ongoing project in our research group, where the goal is to extend the size of the corpus to enable us to train deep learning structures with millions of parameters toward robust speech emotion classifiers.

## 6 Conclusions

This work proposed a new method to build naturalistic emotional databases from publicly available speech samples. The recordings are selected from podcasts with spontaneous conversations providing almost unlimited resources

of naturalistic expressive interactions. While other non-acted databases convey unbalanced emotional content as dictated by the recording protocols, the emotional content is carefully selected by training machine learning algorithms with existing emotional corpora which retrieve samples with target emotions. As a proof-of-concept, we demonstrated that this technique can be used to create a corpus with samples across the arousal-valence space. For this purpose, we trained and evaluated machine learning alternatives for classification, preference learning and regression, where the task was to identify sentences with low and high values of arousal and valence. Relying on the three most successful algorithms on the evaluations with existing corpora, we retrieved 2,317 sentences which were emotionally annotated using crowdsourcing evaluations. The experimental evaluation demonstrated the potential of this approach, where the emotional content of the retrieved speech sentences cover almost the entire arousal and valence space. The distribution of the corpus is more balanced than the emotional content of other emotional corpora.

The key advantage of this approach is its scalability. Using this approach, we have extended the size of the corpus with over 27 hours of emotional data. As discussed in Section 5.4, the distribution of the corpus across primary emotions is not as balanced as the scores for attribute dimensions, where we have few sentences for certain emotional classes (e.g., fear, sadness). To address this problem, we are exploring preference learning solutions for categorical emotions (e.g., sad ranker). We are also increasing the number of podcasts with emotional content including unrepresentative emotional classes increasing the pool of candidate sentences.

Creating a large speech emotional corpus will open new opportunities for the community, allowing the use of the latest deep learning solutions in speech emotion recognition. We are starting to observe the benefits of this corpus, by improving emotion classification performance with more complex deep learning structures, which can be reliably trained as we increase the size of the corpus.

## ACKNOWLEDGMENTS

This work was funded by a National Science Foundation under grant IIS-1453781 (Career Award) and Microsoft Research.

## REFERENCES

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

[2] H. Cao, D. Cooper, M. Keutmann, R. Gur, A. Nenkova, and R. Verma, "CREMA-D: Crowd-sourced emotional multimodal actors dataset," *IEEE Transactions on Affective Computing*, 2014.

[3] D. Ververidis and C. Kotropoulos, "A state of the art review on emotional speech databases," in *First International Workshop on Interactive Rich Media Content Production (RichMedia-2003)*, Lausanne, Switzerland, October 2003, pp. 109–119.

[4] E. Douglas-Cowie, N. Campbell, R. Cowie, and P. Roach, "Emotional speech: Towards a new generation of databases," *Speech Communication*, vol. 40, no. 1-2, pp. 33–60, April 2003.

[5] M. Liberman, K. Davis, M. Grossman, N. Martey, and J. Bell, "Emotional prosody speech and transcripts," Philadelphia, PA, USA, 2002, Linguistic Data Consortium.

[6] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *9th European Conference on Speech Communication and Technology (Interspeech'2005 - Eurospeech)*, Lisbon, Portugal, September 2005, pp. 1517–1520.

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

[8] L. Devillers and L. Vidrascu, "Real-life emotions detection with lexical and paralinguistic cues on human-human call center dialogs," in *Interspeech - International Conference on Spoken Language (ICSLP)*, Pittsburgh, PA, USA, September 2006, pp. 801–804.

[9] E. Douglas-Cowie, R. Cowie, and M. Schröder, "A new emotion database: considerations, sources and scope," in *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, Newcastle, Northern Ireland, UK, September 2000, pp. 39–44.

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

[11] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions," in *2nd International Workshop on Emotion Representation, Analysis and Synthesis in Continuous Time and Space (EmoSPACE 2013)*, Shanghai, China, April 2013.

[12] E. Douglas-Cowie, R. Cowie, C. Cox, N. Amir, and D. Heylen, "The sensitive artificial listener: an induction technique for generating emotionally coloured conversation," in *Second International Workshop on Emotion: Corpora for Research on Emotion and Affect, International conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco, May 2008, pp. 1–8.

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

[14] S. Steidl, "Automatic classification of emotion-related user states in spontaneous children's speech," Ph.D. dissertation, Universität Erlangen-Nürnberg, Erlangen, Germany, January 2009.

[15] M. Grimm, K. Kroschel, and S. Narayanan, "The Vera am Mittag German audio-visual emotional speech database," in *IEEE International Conference on Multimedia and Expo (ICME 2008)*, Hannover, Germany, June 2008, pp. 865–868.

[16] T. Rahman and C. Busso, "A personalized emotion recognition system using an unsupervised feature adaptation scheme," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2012)*, Kyoto, Japan, March 2012, pp. 5117–5120.

[17] L. Morency, R. Mihalcea, and P. Doshi, "Towards multimodal sentiment analysis: Harvesting opinions from the web," in *Proceedings of the 13th international conference on multimodal interfaces (ICMI 2011)*, Alicante, Spain, November 2011, pp. 169–176.

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

[19] C. Busso, S. Parthasarathy, A. Burmania, M. AbdelWahab, N. Sadoughi, and E. Mower Provost, "MSP-IMPROV: An acted corpus of dyadic interactions to study emotion perception," *IEEE Transactions on Affective Computing*, vol. 8, no. 1, pp. 67–80, January-March 2017.

[20] L. Chen, "Joint processing of audio-visual information for the recognition of emotional expressions in human-computer interaction," Ph.D. dissertation, University of Illinois at Urbana-Champaign, Champaign, IL, USA, 2000.

[21] T. Bänziger, H. Pirker, and K. Scherer, "GEMEP - Geneva multimodal emotion portrayals: a corpus for the study of multimodal emotional expressions," in *First International Workshop on Emotion: Corpora for Research on Emotion and Affect (International conference on Language Resources and Evaluation (LREC 2006))*, Genoa,Italy, May 2006, pp. 15–19.

[22] B. Schuller, R. Müller, F. Eyben, J. Gast, B. Hörnler, M. Wöllmer, G. Rigoll, A. Höthker, and H. Konosu, "Being bored? recognising natural interest by extensive audiovisual integration for real-life application," *Image and Vision Computing*, vol. 27, no. 12, pp. 1760–1774, November 2009.

[23] R. Cowie, E. Douglas-Cowie, and C. Cox, "Beyond emotion

archetypes: Databases for emotion modelling using neural networks," *Neural Networks*, vol. 18, no. 4, pp. 371–388, May 2005.

[24] E. Douglas-Cowie, L. Devillers, J. Martin, R. Cowie, S. Savvidou, S. Abrilian, and C. Cox, "Multimodal databases of everyday emotion: Facing up to complexity," in *9th European Conference on Speech Communication and Technology (Interspeech'2005)*, Lisbon, Portugal, September 2005, pp. 813–816.

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

[26] A. Batliner, K. Fischer, R. Huber, J. Spilker, and E. Nöth, "Desperately seeking emotions or: actors, wizards and human beings," in *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, Newcastle, Northern Ireland, UK, September 2000, pp. 195–200.

[27] C. Busso and S. Narayanan, "Recording audio-visual emotional databases from actors: a closer look," in *Second International Workshop on Emotion: Corpora for Research on Emotion and Affect, International conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco, May 2008, pp. 17–22.

[28] T. Bänziger and K. Scherer, "Using actor portrayals to systematically study multimodal emotion expression: The GEMEP corpus," in *Affective Computing and Intelligent Interaction (ACII 2007), Lecture Notes in Artificial Intelligence 4738*, A. Paiva, R. Prada, and R. Picard, Eds. Berlin, Germany: Springer-Verlag Press, September 2007, pp. 476–487.

[29] F. Enos and J. Hirschberg, "A framework for eliciting emotional speech: Capitalizing on the actor's process," in *First International Workshop on Emotion: Corpora for Research on Emotion and Affect (International conference on Language Resources and Evaluation (LREC 2006))*, Genoa,Italy, May 2006, pp. 6–10.

[30] S. Parthasarathy, C. Zhang, J. Hansen, and C. Busso, "A study of speaker verification performance with expressive speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2017)*, New Orleans, LA, USA, March 2017, pp. 5540–5544.

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

[32] C. Cieri, D. Miller, and K. Walker, "The Fisher corpus: A resource for the next generations of speech-to-text," in *International conference on Language Resources and Evaluation (LREC 2004)*, Lisbon, Portugal, May 2004.

[33] "The speaker attribution intelligent service," http://diarizationservice3.cloudapp.net/, 2016, retrieved December 22, 2016.

[34] A. Ziaei, L. Kaushik, A. Sangwan, J. Hansen, and D. Oard, "Speech activity detection for NASA apollo space missions: challenges and solutions," in *Interspeech 2014*, Singapore, September 2014, pp. 1544–1548.

[35] C. Kim and R. Stern, "Robust signal-to-noise ratio estimation based on waveform amplitude distribution analysis," in *Interspeech 2008*, Brisbane, Australia, September 2008, pp. 2598–2601.

[36] A. Potamianos and P. Maragos, "Speech formant frequency and bandwidth tracking using multiband energy demodulation," *The Journal of the Acoustical Society of America*, vol. 99, no. 6, pp. 3795–3806, June 1996.

[37] N. Shokouhi, A. Ziaei, A. Sangwan, and J. Hansen, "Robust overlapped speech detection and its application in word-count estimation for prof-life-log data," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2015)*, Brisbane, Australia, April 2015, pp. 4724–4728.

[38] L. Fischer, D. Brauns, and F. Belschak, *Zur Messung von Emotionen in der angewandten Forschung*. Pabst Science Publishers, Lengerich, 2002.

[39] M. Grimm, K. Kroschel, E. Mower, and S. Narayanan, "Primitives-based evaluation and estimation of emotions in speech," *Speech Communication*, vol. 49, no. 10-11, pp. 787–800, October-November 2007.

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

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

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

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

[44] Y.-L. Lin and G. Wei, "Speech emotion recognition based on HMM and SVM," in *International Conference on Machine Learning and Cybernetics (ICMLC 2005)*, vol. 8, Guangzhou, China, August 2005, pp. 4898–4901.

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

[46] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. Witten, "The WEKA data mining software: An update," *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 1, pp. 10–18, June 2009.

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

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

[49] S. Mariooryad and C. Busso, "The cost of dichotomizing continuous labels for binary classification problems: Deriving a Bayesian-optimal classifier," *IEEE Transactions on Affective Computing*, vol. 8, no. 1, pp. 119–130, January-March 2017.

[50] J. Doyle, "Prospects for preferences," *Computational Intelligence*, vol. 20, no. 2, pp. 111–136, May 2004.

[51] J. Fürnkranz and E. Hüllermeier, *Preference learning*. Springer, November 2010.

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

[53] R. Herbrich, T. Graepel, P. Bollmann-Sdorra, and K. Obermayer, "Learning preference relations for information retrieval," in *Workshop on Learning for Text Categorization (Workshop held in conjunction with ICML/AAAI-98)*, Madison, WI, USA, July 1998, pp. 80–84.

[54] H. Cao, R. Verma, and A. Nenkova, "Combining ranking and classification to improve emotion recognition in spontaneous speech," in *Interspeech 2012*, Portland, Oregon, USA, September 2012, pp. 358–361.

[55] ——, "Speaker-sensitive emotion recognition via ranking: Studies on acted and spontaneous speech," *Computer Speech & Language*, vol. 29, no. 1, pp. 186–202, January 2015.

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

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

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

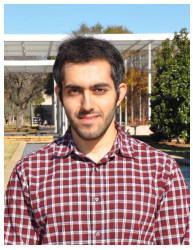
[59] T. Joachims, "Training linear SVMs in linear time," in *ACM SIGKDD international conference on Knowledge discovery and data mining*, Philadelphia, USA, August 2006, pp. 217–226.

[60] W. Chu and Z. Ghahramani, "Preference learning with Gaussian processes," in *International conference on machine learning (ICML 2005)*, Bonn, Germany, August 2005, pp. 137–144.

[61] D. Basak, S. Pal, and D. Patranabis, "Support vector regression," *Neural Information Processing-Letters and Reviews*, vol. 11, no. 10, pp. 203–224, October 2007.

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

[63] F. Eyben, M. Wöllmer, and B. Schuller, "OpenSMILE: the Munich versatile and fast open-source audio feature extractor," in *ACM International conference on Multimedia (MM 2010)*, Florence, Italy, October 2010, pp. 1459–1462.
[64] C. Busso and T. Rahman, "Unveiling the acoustic properties that describe the valence dimension," in *Interspeech 2012*, Portland, OR, USA, September 2012, pp. 1179–1182.
[65] S. Parthasarathy and C. Busso, "Jointly predicting arousal, valence and dominance with multi-task learning," in *Interspeech 2017*, Stockholm, Sweden, August 2017.



**Reza Lotfian** (SM'17) received his BS degree (2006) with high honors in Electrical Engineering from the Department of Electrical Engineering, Amirkabir University, Tehran, Iran and MS degree (2010) in Electrical Engineering from the Sharif University (SUT), Tehran, Iran. He is currently pursuing his Ph.D. degree in the Electrical and Computer Engineering at the University of Texas at Dallas (UTD), Richardson, Texas, USA. He joined the Multimodal Signal Processing (MSP) laboratory in 2013. His research interest includes the area of speech signal processing, affective computing, human machine interaction, and machine learning.

**Carlos Busso** (S'02-M'09-SM'13) 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 (2008) in electrical engineering from the University of Southern California (USC), Los Angeles, in 2008. He is an associate professor at the Electrical and Computer Engineering Department of The 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 (MSP) laboratory [http://msp.utdallas.edu]. He is a recipient of an NSF CAREER Award. In 2014, he received the ICMI Ten-Year Technical Impact Award. He also received the Hewlett Packard Best Paper Award at the IEEE ICME 2011 (with J. Jain). 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 digital signal processing, speech and video processing, and multimodal interfaces. His current research includes the broad areas of affective computing, multimodal human-machine interfaces, modeling and synthesis of verbal and nonverbal behaviors, sensing human interaction, in-vehicle active safety system, and machine learning methods for multimodal processing. He is a member of ISCA, AAAC, and ACM, and a senior member of the IEEE.