HHS Public Access

Author manuscript

Proc ACM Interact Mob Wearable Ubiquitous Technol. Author manuscript; available in PMC 2019 June 11.

Published in final edited form as:

Proc ACM Interact Mob Wearable Ubiquitous Technol. 2018 June ; 2(2): . doi:10.1145/3214284.

A Weakly Supervised Learning Framework for Detecting Social Anxiety and Depression

ASIF SALEKIN,
Department of Computer Science, University of Virginia, Charlottesville, VA, 22903, USA

JEREMY W. EBERLE,
Department of Psychology, University of Virginia, Charlottesville, VA, 22903, USA

JEFFREY J. GLENN,
Department of Psychology, University of Virginia, Charlottesville, VA, 22903, USA

BETHANY A. TEACHMAN, and
Department of Psychology, University of Virginia, Charlottesville, VA, 22903, USA

JOHN A. STANKOVIC
Department of Computer Science, University of Virginia, Charlottesville, VA, 22903, USA

Abstract

Although social anxiety and depression are common, they are often underdiagnosed and undertreated, in part due to difficulties identifying and accessing individuals in need of services. Current assessments rely on client self-report and clinician judgment, which are vulnerable to social desirability and other subjective biases. Identifying objective, nonburdensome markers of these mental health problems, such as features of speech, could help advance assessment, prevention, and treatment approaches. Prior research examining speech detection methods has focused on fully supervised learning approaches employing strongly labeled data. However, strong labeling of individuals high in symptoms or state affect in speech audio data is impractical, in part because it is not possible to identify with high confidence which regions of a long speech indicate the person's symptoms or affective state. We propose a weakly supervised learning framework for detecting social anxiety and depression from long audio clips. Specifically, we present a novel feature modeling technique named NN2Vec that identifies and exploits the inherent relationship between speakers' vocal states and symptoms/affective states. Detecting speakers high in social anxiety or depression symptoms using NN2Vec features achieves F-1 scores 17% and 13% higher than those of the best available baselines. In addition, we present a new multiple instance learning adaptation of a BLSTM classifier, named BLSTM-MIL. Our novel framework of using NN2Vec features with the BLSTM-MIL classifier achieves F-1 scores of 90.1% and 85.44% in detecting speakers high in social anxiety and depression symptoms.

as3df@virginia.edu.

Keywords

Social anxiety; depression; anxiety; multiple instance learning; weakly supervised learning; mental disorder; audio word; embedding; weakly labeled; speech; feature modeling; assessment

1 INTRODUCTION

Social anxiety and depression are common mental health problems. Social anxiety disorder involves an intense fear of negative evaluation or rejection in social or performance situations. People with this disorder persistently avoid such situations, or endure them with significant distress, and in some cases experience strong physical symptoms such as rapid heart rate, nausea, sweating, and even full-blown panic attacks [4, 64]. The disorder occurs in about 7% of U.S. adults in a given year and in about 11% at some point in their lives [43]. Major depressive disorder is marked by persistent sadness, loss of interest or pleasure in activities, and other symptoms such as appetite and psychomotor changes, sleep disturbance, loss of energy, feelings of guilt or worthlessness, and reduced concentration [4, 64]. It is the leading cause of disability worldwide [17] and increases an individual's risk of suicide [32, 58]. About 7% of U.S. adults have the disorder in a given year, and about 14% have it at some point in their lives [43].

Because there is no objective measure or reliable biomarker of the severity of social anxiety or depression [61], gold-standard assessments for these disorders remain rooted in client self-report and clinician judgment, risking a range of subjective biases. Clinician rating scales (e.g., Hamilton Rating Scale for Depression [30]) require training, practice, and certification for inter-rater reliability [61], and client self-reports (e.g., Social Interaction Anxiety Scale ,SIAS [57]) rely on clients' ability and willingness to communicate their thoughts, feelings, and behaviors when distressed or impaired, which can alter their ability and motivation to self-report [4]. Further, distress from these disorders is often difficult for others to detect. For example, socially anxious people rate their own social performance more critically than non-anxious people, even though their actual performance is not necessarily poorer [1, 71, 87]. This suggests that social anxiety can be salient to the person but not evident to others. Socially anxious people's social avoidance and safety behavior to reduce or hide their anxiety [100] also can limit others' knowledge of their distress. Thus, relying only on subjective approaches for assessment is inadequate for reliable diagnosis, which is problematic given the high prevalence of social anxiety and depression and the vast numbers who receive no help [13]. In the United States, 50% of people with social anxiety and 22% of people with depression never talk with a provider about their symptoms [95]. Moreover, general practitioners correctly identify social anxiety and depression in only 24% and 50% of true cases [60, 99].

Health-care providers would benefit from objective indicators of social anxiety and depression symptoms that require no extensive equipment and are readily accessible and not intrusive or burdensome to complement their self-report, interview, and other assessment modalities. Indicators of social anxiety and depression symptoms could improve diagnostic clarity and treatment planning, thereby helping ensure that people receive the most

appropriate interventions. Moreover, symptom indicators that providers can assess remotely could help close the treatment gap [41] by identifying individuals who may be in need of prevention, assessment, or treatment resources, which could be delivered in person or via eHealth modalities [90]. People with social anxiety may otherwise not seek treatment because, for example, they do not know where to find it, or fear discussing their symptoms with providers [63], and people with depression may not seek treatment in part because they think they can handle or treat their symptoms on their own or do not view their symptoms as pathological [20]. Furthermore, the ability to remotely detect affective states would help providers monitor instances of high affect both between sessions and after the end of treatment. The latter is especially important given the high relapse rates for formerly depressed individuals. Such passive outcome monitoring (e.g., [35]), requiring minimal effort from the client, could help providers identify when the client is distressed and might benefit from a just-in-time intervention or prompt providers and clients to consider scheduling a booster session.

Studies have shown that prosodic, articulatory, and acoustic features of speech can be indicative of disorders such as depression and social anxiety [14, 23, 54, 55, 80, 81, 83, 94], and research on the objective detection and monitoring of mental disorders based on measurable behavioral signals such as speech audio is proliferating [15, 16, 24, 26, 86]. State-of-the-art works on detecting mental disorders or emotional states (e.g., anxious vs. calm) from audio data use supervised learning approaches, which must be “trained” from examples of the sound to be detected. In general, learning such classifiers requires annotated data, where the segments of audio containing the desired vocal event and the segments not containing that event are clearly indicated. We refer to such data as “strongly labeled.” However, diagnosing mental disorders is a complicated and time consuming procedure that requires an annotator with a high degree of clinical training. In addition, strong labeling of mental disorders in speech audio clips is impractical because it is impossible to identify with high confidence which regions of a conversation or long speech are indicative of disorder. Supervised learning, hence, is a difficult task.

A solution is to collect long speech audio samples from individuals already diagnosed with or high in symptoms of specific mental disorders from situations that may heighten expression of the symptoms of respective disorders. This type of data is considered “weakly labeled,” meaning that although they provide information about the presence or absence of disorder symptoms, they do not provide additional details such as the precise times in the recording that indicate the disorder, or the duration of those identifying regions. We strove to use weakly labeled data to: (a) identify speakers high in social anxiety or depression symptoms, and (b) identify speakers’ state affect (anxious vs. calm). We expected that recordings from individuals high in symptoms would have regions indicative of those symptoms (i.e., regions present for persons high, but not for persons low, in symptoms). We also expected that recordings from individuals who reported a peak state of anxiety would have regions indicative of that state (i.e., regions present for persons with, but not for persons without, peak anxiety).

This approach falls under the general rubric of multiple instance learning (MIL). MIL is a weakly supervised learning approach in which labels for individual instances are unknown;

instead, labels are available for a collection of instances, usually called “bag.” A positive bag has at least one positive instance (indicating high symptoms or peak anxiety) and may contain negative instances (label noise), whereas a negative bag contains negative instances only. In this paper, we break weakly labeled audio clips into several small, contiguous segments, where the segments are the instances and the audio clip is the bag.

To our knowledge, no previous research has identified individuals high in symptoms of a mental disorder or detected state anxiety from weakly labeled audio data. The contributions of this paper are:

- We present a novel weakly supervised learning framework for detecting individuals high in symptoms of two mental disorders (social anxiety and depression) from weakly labeled audio data, adding a practical complement to health-care providers’ assessment modalities.
- We also use our approach to detect state anxiety, given the importance of determining when a person is especially anxious, which can help determine when a person would benefit from just-in-time interventions.
- We propose a novel feature modeling technique named NN2Vec (section 3.3) to generate low-dimensional, continuous, and meaningful representation of speech from long weakly labeled audio data. All existing techniques (e.g., I-vector, audio words, Emo2Vec) are designed for strongly labeled data; hence, they fail to meaningfully represent speech due to the significant label noise in weakly labeled audio. NN2Vec identifies and exploits the inherent relationship between audio states and targeted vocal events. Identifying individuals high in social anxiety and depression symptoms using NN2Vec achieved on average F-1 scores 17% and 13% higher, respectively, than those of the other techniques (sections 6.1.2, 6.2, and 7).
- MIL adaptation performs significantly better than supervised learning classifiers (where the individual instance labels are ambiguous), which fall short of generating an optimal solution due to label noise in weakly labeled data [8]. Studies have shown that emotion or mental disorder can be perceived by the temporal dynamics across speech states [50, 56, 76, 103]. To generate a sequential deep neural network solution that comprehends the temporal properties in speech while also being adaptive to noise in weakly labeled long audio data, we developed a novel MIL adaptation of bidirectional long short term memory classifier, named BLSTM-MIL (section 4.1).
- Because no existing dataset contained spontaneous speech labeled with speakers high in social anxiety, we built a dataset consisting of 3-minute samples of weakly labeled spontaneous speech from 105 participants. Our approach achieves an F-1 score of 90.1% in detecting speakers high in social anxiety symptoms. Additionally, our approach achieves an F-1 score of 93.4% in detecting anxious versus calm states based on participants’ self-reported levels of peak emotion during the speech.

- We analyzed data from a publicly available Distress Analysis Interview Corpus (DAIC-WOZ) database [92] that contains weak labels of participants' mental disorder (depressed vs. non-depressed) on 10–15 minute interviews. Our approach achieves an F-1 score of 85.44% in detecting speakers with depression, which is 33% higher than that of the best state-of-the-art work evaluated on this dataset (section 7).

2 RELATED WORK

Human event detection systems from speech can be split into three parts: feature extraction, modeling, and classification. Several combinations of features have been investigated for vocal event detection. These features can be divided into two groups according to their time span: low-level descriptors (LLDs) are extracted for each small time frame (16–45 ms is typical), such as Mel-frequency cepstral coefficients, energy, zero crossing rate, pitch, spectral centroid, reduced speech, and reduce vowel space [14, 21, 29, 77, 79, 85]. By contrast, high-level descriptors (HLDs), such as the mean, standard deviation, quartile, flatness, or skewness, are computed using the LLDs extracted for the whole audio signal or for an audio segment covering several frames [70, 77, 78, 93].

The modeling stage of an audio analytic system obtains a representation of the speech that reflects the target event information. It is expected that task performance will improve using input representation that better understands the speech-to-task relation. Different modeling approaches in the literature use different features. When dealing with LLDs, different techniques have been borrowed from other speech recognition tasks, such as supervised and unsupervised subspace learning techniques. Many of these modeling techniques apply windowing to the speech.

Recent studies on speech signal processing have achieved improved accuracy using the I-vector representation of speech [25, 39]. The I-vector extraction, originally developed for speaker recognition, consists of two separate stages: UBM state alignment and I-vector computation. UBM state alignment identifies and clusters the similar acoustic content (e.g., frames belonging to a phoneme) to allow the following I-vector computation to be less affected by the phonetic variations between features. However, noise and channel variation could substantially affect the alignment quality and, thus, the purity of extracted I-vectors. Recent studies on detecting emotion [53] and depression [69] have used I-vector modeling from strongly labeled speech data. The I-vector technique estimates the difference between real data and average data. Because majority portions of positive samples in weakly labeled audio data are actually average data, I-vector computation performs poorly.

Recently the audio-codebook model [65, 72] has been used to represent the audio signal in windows with "audio words" for vocal event detection. Several studies on text (e.g., word2vec) and language model representations [5, 27, 59, 74] have used various structures of shallow neural networks (one or two hidden layers) to model features. A recent study introduced an adaptation of the word2vec [74] approach, named Emo2Vec [76] for vocal emotion detection, which generates similar feature representations for small frames that appear with similar context (neighbor frames) for a particular targeted emotion. This paper

proposes a new and simple shallow neural network based feature modeling technique, named NN2Vec, to generate meaningful feature representation for audio event detection from long weakly labeled audio data. We consider the I-vector, audio-codebook, and Emo2vec feature modeling techniques as baselines for comparison (sections 6.1.2, 6.2, and 7).

Various types of supervised classifiers have been used for vocal event detection, including hidden markov models [51], gaussian mixture models [55, 88, 104], support vector machines (SVM) [2, 55, 77], k-nearest neighbor [73, 77], and many others [68].

With the recent success of deep learning approaches in speech recognition [31], research on audio event detection studies is shifting from conventional methods to modern deep learning techniques [7, 49]. Several studies [36, 40, 75] have used the convolutional neural network (CNN) to identify the presence of vocal events. The CNN learns filters that are shifted in both time and frequency. Using these filters, the CNN exploits spatially local correlation by enforcing a local connectivity pattern between neurons of adjacent layers: Each neuron is connected to only a small region of the input volume. However, conventional CNNs fail to capture long temporal context information. We consider the CNN as another baseline.

Several studies on detecting emotion and mental disorder from speech [50, 56, 76, 103] have shown that temporal properties in a speech signal provide important information about emotion and mental disorder and have used sequential classifiers such as the recurrent neural network (RNN) and the long short term memory classifier (LSTM). Both the RNN and the LSTM have feedback loops that let them maintain information in “memory” over time. But the LSTM outperforms the RNN [10], as it does better at avoiding the vanishing gradient problem and captures longer temporal context information. Given that LSTM performs better for long audio data, we consider the bidirectional LSTM as a baseline.

Some recent studies [52, 56] have combined a CNN feature extraction architecture with sequential LSTM, an approach named CNN-LSTM that can learn to recognize and synthesize sequential dynamics in speech. In the CNN-LSTM, the CNN acts as the trainable feature detector for the spatial signal. It learns features that operate on a static spatial input (windows) while the LSTM receives a sequence of high-level representations to generate a description of the content. We consider the CNN-LSTM as a baseline.

Due to the presence of label noise (discussed in section 4.1), conventional neural networks fall short of generating an optimal solution when trained on weakly labeled data. To generate a deep neural network solution that captures the temporal properties in speech while also being adaptive to noise, we developed a MIL adaptation of BLSTM (BLSTM-MIL). Although no study has used weakly supervised learning to identify vocal events in weakly labeled speech data, several recent studies [37, 46, 47, 89] have detected rare environmental sound events (e.g., car horn, gun shot, glass break) from weakly labeled audio clips (where the event is a small fraction of a long environmental audio clip). Two of these studies [46, 47] used MIL approaches, the mi-SVM and the deep neural network-based MIL (DNN-MIL), for environmental audio event detection. We consider the mi-SVM and the DNN-MIL as baselines (section 6.1.3 and table 7).

Some recent studies on social anxiety and depression monitoring systems [6, 9, 91] have used smartphone sensors, text information, call information, and GPS data to understand how depression or social anxiety levels are associated with an individual's mobility and communication patterns. To our knowledge, no study has identified whether a speaker belongs to a high versus low social anxiety group or is experiencing an anxious versus calm vocal state from audio data. Although two prior studies [97, 98] have found a strong correlation between vocal pitch (F0) and social anxiety and one study [48] has shown that pitch (F0) and energy are indicative of state anxiety, we used pitch and energy as two of our LLDs.

No study on the detection of depression from speech audio signals [15, 16, 24, 26, 86] has applied weakly supervised learning approaches to weakly labeled audio data. These studies have used LLDs [26, 86] such as pitch, RMS, MFCC, and HNR as features and SVMs [62, 86], hidden Markov models [16], and linear support-vector regression models [26] as supervised learning classifiers. A depression detection approach evaluated on the DAIC-WOZ database [92] used I-vector features with an SVM classifier. Recently, another depression detection study [56] named DepAudioNet evaluated on this database applied a CNN-LSTM classifier using LLD. We consider the two most recent depression detection approaches evaluated on the DAIC-WOZ dataset [56, 62] as baselines.

3 FEATURE MODELING

In the following sections, we discuss extracted LLDs from raw audio signal (section 3.1) and use an audio-codebook approach to map the audio signal to audio words (section 3.2) and our new NN2Vec feature modeling approach (section 3.3) to reflect the audio information for MIL algorithms. Our novel MIL solution uses NN2Vec with a new BLSTM-MIL (section 4.1) classifier to detect vocal events from weakly labeled data.

3.1 Audio Features

Our approach segments the audio clips into overlapping windows and extracts a feature set from each window. Extracted feature sets represent the inherent state of audio from that window. Based on the previous studies on audio features associated with human vocal event detection (section 2), we considered the LLDs shown in the left column of table 1, as well as their delta and delta-delta coefficients. Each window is segmented into overlapping 25-ms frames with 10-ms overlap, from which LLDs are extracted. Next, the 8 functionals shown in the right column of table 1 are applied to extract the audio window representation. In total, 272 features are extracted from each of the overlapping windows. We evaluated window size from 500 ms to 10 seconds.

3.2 From Audio to Words

We use the audio-codebook model [65, 72] to represent the audio signal in a window with audio words. The audio words are not words in the typical, semantic meaning of words, but rather fragments of the audio signal represented by features. We need robust features to represent the audio state in a window. Inspired by [45], we use a GMM-based clustering

method to generate the audio codebook from the functional representations mentioned in section 3.1.

To generate the codebook, a GMM-based model is trained on randomly sampled data from the training set. The resulting clusters form the codebook audio words. Once the codebook is generated, acoustic HLDs within a certain range of the audio signal are assigned to the closest audio words (GMM cluster centers) in the codebook.

The discriminating power of an audio-codebook model is governed by the codebook size. The codebook size is determined by the number of clusters C generated by the GMM. In general, larger codebooks are thought to be more discriminating, whereas smaller codebooks should generalize better, especially when HLDs extracted from frames can be distorted with distance, environmental noise, and reverberation, as smaller codebooks are more robust against incorrect assignments. However, a codebook that is too small is too generic, and, hence, unable to capture the change in speech states in various small frames. Hence, through the audio-codebook approach, audio clips are converted to a sequence of audio words.

3.3 NN2Vec Approach

Human emotion or mental states are represented by a sequence of audio states [52, 56, 76], which are represented by audio words. Our assumption is that regions (subsequences of audio words) indicative of targeted vocal events (high symptom/disorder classification or state anxiety) are common (occur with high probability) across positive audio clips, and not present or rarely present (occur with low probability) in negative audio clips. MIL requires that the feature modeling learn the inherent relation between audio states and vocal events (positive class) and that the generated feature representations indicate the positive class. Conventional feature modeling (audio word, I-vector, etc.) techniques cannot learn this relation effectively from weakly labeled long audio clips (section 6.1.2).

To identify and exploit the inherent relationship between audio states and vocal events from weakly labeled data, we developed a neural network-to-vector conversion (NN2Vec) approach that generates an N dimensional dense vector representation for each of the audio words. The contributions of NN2Vec are:

- **Representational efficiency:** Audio word representation relies on the notion of one hot encoded vector, where an audio word is represented by a sparse vector with a dimension equal to the size of the vocabulary with a 1 at the index that stands for the word and 0s everywhere else. Hence, the feature representation dimension is significantly high, which is difficult for a classifier to optimize using limited weakly labeled data. NN2vec is a shallow neural network model that generates a fixed-length dense vector for each of the audio words. This means that the model learns to map each discrete audio word representation (0 through the number of words in the vocabulary) into a low-dimensional continuous vector space from their distributional properties observed in training. This is done by a relatively straightforward optimization that starts with a more or less random assignment and then progressively reduces the overall error with a gradient descent method. We evaluated with codebook sizes V from 500 to 5000

and found that the best NN2Vec dimension N is between 20 and 50, based on V. Hence, compared to high-dimensional sparse audio word features, NN2Vec features represent audio states with significantly low-dimensional distributed representation.

- **Mapping efficiency:** An interesting property of NN2vec vectors is that they not only map the states of audio (audio words) in a smaller space, but also encode the syntactic relationships between audio states. NN2Vec vectors are similar for audio states with similar probability of occurring in positive audio clips. Neural networks typically respond in a similar manner to similar inputs. Generated distributed representations are designed to take advantage of this; audio states that should result in similar responses are represented by similar NN2Vec vectors, and audio states that should result in different responses are represented by quite different NN2Vec vectors. Hence, identification of sequences of states indicative of a mental disorder should be easier for a weakly supervised classifier.
- **Continuity:** Representing states in continuous vector space allows powerful gradient-based learning techniques such as backpropagation to be applied effectively. Previous studies [66, 67] have shown that distributed representation of input features improves classification performance compared to discrete representation.

3.3.1 NN2Vec Vector Generation.—This subsection describes our NN2Vec feature generation approach, where NN2Vec vector representations of audio states are learned by a fully connected neural network. Later sections discuss how our NN2Vec neural network model learns similar vector representations for audio words with similar probability of occurring in positive audio clips. Our training set contains (B_i, Y_i) audio clip-label pairs, where the ith audio clip is denoted as B_i and its corresponding label $Y_i \in \{1, 0\}$. We segment these clips into overlapping windows, s_{ij} (jth window in audio clip B_i) and assign an audio word-label pair (w_{ij}, y_{ij} to each of them. Here, w_{ij} is the audio word extracted through the audio-codebook approach (section 3.2) from window s_{ij} and $y_{ij} = Y_i$, label of the respective audio clip. Considering that a codebook size (section 3.2) is V, these audio words w_{ij} are converted to a V dimensional one-hot encoded vector X_{ij}. Suppose audio word w_{ij} is the lth audio word in the codebook. Then its one-hot vector representation would be: $X_{ij} = [x_{ijk}]$, where $k = 1 \ldots V$ and $x_{ijk} = 1$, only if $k = l$ and $x_{ijk} = 0$ otherwise. These one-hot vector-segment level label pairs (X_{ij}, y_{ij}) are our training input set for the NN2Vec vector generation model.

Figure 2 shows our NN2Vec fully connected neural network. Here, the input layer is V dimensional, corresponding to one-hot vectors, and the output layer is a 2-dimensional softmax layer. If the hidden layer has N neurons, generated NN2Vec vectors would be N dimensional. We train the network with (X_{ij}, y_{ij}) pairs from the training set. The weights between the input layer and the output layer of the NN2Vec network can be represented by a $V \times N$ matrix W. Each row of W is the N-dimension vector. After training each row r of W is our NN2Vec vector representation of the rth audio word in the codebook (section 3.2).

Through this approach, if two audio words occur with similar frequency (hence, similar probability) in positive class examples, their corresponding rows in W, hence generated NN2Vec vectors would be similar.

3.3.2 Learning Vector Representations Through NN2Vec Model.—This subsection discusses the approach through which the NN2Vec model learns similar vector representations for the audio words that occur with similar probability in a targeted audio event (positive class examples). Figure 3 shows the simplified form of the NN2Vec network model. Suppose the codebook size (section 3.2) is V and the hidden layer size is N, which means the generated NN2Vec vector size would be N. All the layers are fully connected layers. The input is a one-hot encoded vector, which means that for a given input audio word, only one out of V units, $\{x_1, x_2, \ldots, x_V\}$, will be 1, and the rest will be 0.

The weights between the input layer and the hidden layer can be represented by a $V \times N$ matrix W. Each row W is the N-dimensional vector representation v_w of the associated audio word of the input layer. Given an audio word w_k, $x_k = 1$ and $x_{k'}$ for $k' \neq k$, and:

$$h = W^T x = W^T_{(k,.)} := v^T_{w_k} \quad (1)$$

, which is the k row of W to h. $v_w k$ is the vector representation of the input audio word w_k.

There is a different weight matrix $W' = \left\{w'_{ij}\right\}$ from the hidden layer to the output layer, which is a $N \times 2$ matrix. Using these weights, we calculate the score u_j for each class.

$$u_j = v'^T_{c_j} h \quad (2)$$

Here, v'_{c_j} is the jth column of matrix W'. The NN2Vec architecture uses softmax, a log-linear classification model to calculate the posterior probability, which is a multinomial distribution.

$$p\left(c_j | w_k\right) = y_j = \frac{exp\left(u_j\right)}{\sum_{j'=1}^{2} exp\left(u_{j'}\right)} = \frac{exp\left(v'^T_{c_j} v_{w_k}\right)}{\sum_{j'=1}^{2} exp\left(v'^T_{c_{j'}} v_{w_k}\right)} \quad (3)$$

Here, y_j is the output of the j unit in the output layer (in total 2 classes), $v_w k$ is the vector representation of the input audio word w_k and v'_{c_j} is the representation of class c_j.

Weight updates of this network are performed by backpropagation [33] where the training objective is to maximize the conditional probability of observing the actual output class c_O,

given the input audio word w_I (as shown in equation 3) with regard to the weights. Here, O denotes output class and I denotes the input audio word index. The loss function is $E = -\log p(c_O|w_I)$, which we want to minimize, and the network prediction error of jth-output unit $e_j = \frac{\partial E}{\partial u_j} = y_j - t_j$ is the derivative of E with regard to the jth output layer unit's network input u_j. Here t_j will only be 1 when the j-th unit is the actual output class, otherwise $t_j = 0$.

Using stochastic gradient descent, the weight-updating equation for hidden layer to output weights (W') is:

$$w'^{new}_{ij} = w'^{old}_{ij} - \eta . e_j . h_i \quad (4)$$

or

$$v'^{new}_{c_j} = v'^{old}_{c_j} - \eta . e_j . h \text{ for } j = 1, 2 \quad (5)$$

,where $\eta > 0$ is the learning rate, $h = v^T_{w_I}$, and v'_{c_j} is the vector representation of class j. Hence, if $y_j > t_j$, then a portion of the hidden vector h (i.e.,v_{w_I}) is subtracted fromv'_{c_j}, making v'_{c_j} further away from v_{w_I}; if $y_j < t_j$(only when $t_j = 1$; i.e., $c_j = c_O$), then a portion of the hidden vector h (i.e.,v_{w_I}) is added to v'_{c_O} (here, $j = O$), making v'_{c_O} closer tov_{w_I}.

Moreover, the weight-updating equation for input layer to the hidden layer weights (W) is:

$$v^{new}_{w_I} = v^{\text{old}}_{w_I} - \eta . EH^T \quad (6)$$

Here, v_{w_I} is a row of W, the input vector representation of the audio word I, and is the only row of W whose derivative on the loss function $\left(\frac{\partial E}{\partial W}\right)$ is non-zero, given that inputs of the NN2Vec model are one hot encoded vectors. Hence, all the other rows of W will remain unchanged after this iteration, because their derivatives are zero. Also, $EH_i = \sum^2_{j=1} e_j . w'_{ij}$, where w'_{ij} is the i-th hidden layer unit to j-th output layer unit weight. Hence, vector EH is the sum of output vectors of all classes (two in our case) weighted by their prediction error e_j. Therefore, equation 6 essentially adds a portion of two output vectors to the input vector of the input audio word.

The movement of the input vector of w_I is determined by the prediction error; the larger the prediction error, the more significant effects an output vector of a class will exert on the movement on the input vector of audio word. If, in the output layer, the probability of a class

c_j being the output class c_O is overestimated ($y_j > t_j$), then the input vector of the audio word w_I will tend to move farther away from the output vector representation of class c_j; conversely, if the probability of a class c_j being the output class c_O is underestimated ($y_j < t_j$), then the input vector of the audio word w_I will tend to move closer to the output vector representation of class c_j. If the probability of class c_j is fairly accurately predicted, there will be very small movement on the input vector w_I.

As the model parameters update iteratively in each epoch by going through audio word to target class pairs generated from training data, the effects on the vectors accumulate. The output vector representation of a class c is moved back and forth by the input vectors of audio words w which occur in that class (c) in the training data (equation 5), as if there were physical strings between the vector of c and the vectors of audio words. Similarly, an input vector of an audio word w can also be considered as being moved by two output vectors (equation 6). The equilibrium length of each imaginary string is related to the strength of co-occurrence between the associated audio word and class pair.

Given our proposed NN2Vec is a binary softmax classification model, for an audio word w_k, the $p(c_1 | w_k) = 1 - p(c_0 | w_k)$. Here, c_1 is the positive and c_0 is the negative class. Hence, we can consider that during training an input vector of an audio word w will be moved by an output vector of positive class c_1. After many iterations, the relative positions of the input (for audio words) and output (for class) vectors will eventually stabilize. As stated before, the relative position or similarity of these input-output vector pairs depends on the frequency of these pairs in training data, which means the probability of an audio word w_k occurs in class c_1 (positive event), $p(c_1\ w_k)$.

Now, consider two audio words w_p and w_q that occur with similar probability in class c_1. Their relative vector representation will be similar compared to an output vector of class c_1. That means the vector representations of w_p and w_q will be similar. Hence, all the audio words that occur with similar probability to occur in positive class audio clips (in training set), would have similar NN2Vec vector representation through our proposed NN2Vec model approach.

4 MULTIPLE INSTANCE LEARNING SOLUTION

Our tasks are binary classification tasks where labels are either −1 or 1. MIL is a kind of weakly-supervised learning. Each sample is in the form of labeled bags, composed of a wide diversity of instances associated with input features. Labels are attached to the bags, rather than to the individual instances within them. A positive bag is one that has at least one positive instance (an instance from the target class to be classified). A negative bag contains negative instances only. A negative bag is thus pure, whereas a positive bag is impure. This assumption generates an asymmetry from a learning perspective as all instances in a negative bag can be uniquely assigned a negative label, which cannot be done for a positive bag (which may contain both positive and negative instances).

We represent the bag-label pairs as (B_i, Y_i). Here, the ith bag is denoted as B_i, of size l_i, and the jth instance in the bag as x_{ij} where $j \in 1 \ldots l_i$. The label for bag i is $Y_i \in \{-1, 1\}$, and the label for instance x_{ij} is y_{ij}. The label y_{ij} for instances in bag B_i can be stated as:

$$Y_i = -1 \Rightarrow y_{ij} = -1 \forall x_{ij} \in B_i \quad (7)$$

$$Y_i = 1 \Rightarrow y_{ij} = 1 \text{ for at least one } x_{ij} \in B_i \quad (8)$$

This relation between Y_i and y_{ij} is: $Y_i = \max_j\{y_{ij}\}$

Hence, the MIL problem is to learn a classification model so that given a new bag B_t it can predict the label Y_i. To classify (binary) our weakly labeled data, we break the audio clips into several small contiguous segments. Considering reasonably-sized segments, it is safe to assume that if an audio is labeled as a positive class (anxious mental state above 0 based on self-reported peak anxiety during the speech or high symptom/disorder classification based on screening measure), then at least one of the segments is a positive example, containing a region or pattern indicative of the positive class. On the contrary, if an audio is labeled as a negative class, none of the segments will contain a region or pattern indicative of positive class (i.e., hence all segments are negative examples). Hence, according to the MIL definition, the audio clips can be treated as bags B_i and the segments as instances x_{ij} of the corresponding bag. From the arguments just stated, if the weak information identifies the presence of a positive class in an audio segment, then the label for the corresponding bag is +1. Otherwise, it is −1.

A variety of MIL algorithms have been proposed in the literature. This paper considers two MIL algorithms as baselines and presents one novel BLSTM-MIL algorithm for MIL. The first baseline algorithm (miSVM) [3] is based on Support Vector Machine (SVM). The standard SVM algorithm is modified to work in the MIL domain. Although a few other formulations of SVM for MIL domain have been proposed [19], the miSVM is the first SVM formulations for MIL and performs well on a variety of MIL tasks. The second baseline algorithm (DNN-MIL) [101] is a deep neural network modified for MIL domain. These MIL classifiers [3, 19, 101] extract a feature vector from each of the segments that are considered to be a representation of an instance. Hence, these classifiers [3, 19, 101] fail to capture the temporal dynamics of speech states, which is indicative of vocal events [52, 56, 76]. In the following section, we discuss our novel MIL method (section 4.1) based on a long short-term memory classifier.

4.1 BLSTM-MIL

Comprehension of temporal dynamics of states throughout a speech segment (which contains the region or pattern indicative of high symptom/disorder classification or state anxiety) requires long-term dependency. The BLSTM classifier takes sequential inputs where the hidden state of one time step is computed by combining the current input with the hidden state of the previous time steps. They can learn from current time step data as well as

use relevant knowledge from the past to predict outcomes. Hence, we present a BLSTM-MIL classifier that uses the temporal information of speech states within an audio segment (which represents an instance) to learn the instance label.

Our BLSTM-MIL classifier is shown in figure 4. An audio clip is segmented into overlapping windows, and feature sets (i.e., NN2Vec vectors) are extracted from each of these windows. In this approach, the feature sets extracted from the windows represent the state of audio from the respective windows. Feature sets from m consecutive windows comprise an instance x_{ij} of bag B_i. For example, the feature set of the kth window of jth instance x_{ij}, from bag B_i is denoted by f_{ijk}, where $k = 1 \ldots m$. Hence, each instance (MIL representation of a segment) contains representations of audio states (feature sets) as well as their changes (sequence of feature sets) throughout time. In figure 4, $m = 4$ with overlapping size 2, which means 4 consecutive feature sets (representation of audio states of 4 consecutive windows) comprise an instance x_{ij}.

A sequence of feature sets f_{ijk}, with $k = 1 \ldots m$, for each instance x_{ij} in a bag B_i is first fed into the 2-layer BLSTM network with an activation function (in this paper we use sigmoid activation [84]). Through this architecture, we can efficiently make use of past features (via forward states) and future features (via backward states) for a specific time window. Hence, we can capture the forward and backward temporal progression of audio states within a time window (which represent the instance x_{ij}). The forward and backward passes over the unfolded network over time are carried out in a similar way to regular network forward and backward passes, except that we need to unfold the hidden states for all time steps. We do forward and backward propagation for entire audio clips, and we only need to reset the hidden states to 0 at the beginning of each audio clip.

The last layer of the network is a MIL Max Pooling Layer. The MIL Max Pooling layer takes the instance level probabilities o_{ij} for instances x_{ij} of a bag B_i as input and predicts bag label denoted as Y_i^o, according to the following equation:

$$Y_i^o = 1 \text{ if } \max_j o_{ij} \geq \tau, \text{ or } Y_i^o = -1 \text{ otherwise} \quad (9)$$

According to this equation, if at least one of the instance level probabilities is greater than the threshold τ, the predicted bag level would be 1, and −1 otherwise.

The MIL adaptation of BLSTM is trained using backpropagation using the gradients of divergence shown in equation 10 & 11.

$$E_i = \frac{1}{2}(\max_{1 \leq j \leq n_i} (o_{ij}) - d_i)^2 \quad (10)$$

$$E = \sum_{i=1}^{N} E_i \quad (11)$$

Here, d_i is the desired output in response to the set of instances from bag B_i. d_i is set to Y_i, the label assigned to B_i.

In BLSTM-MIL training, all the instances x_{ij} of one bag B_i are considered as a batch (input), and single gradient update (updating network parameters) is performed over one batch of samples. During training, once all instances in a bag have been fed-forward through the network, the weight update for the bag is done with respect to the instance in the bag for which the output was maximum. The process is continued until the overall divergence E falls below a desired tolerance. Because all the instances of one bag are inputted as a batch (during training and testing) and the number of instances in a batch size can vary, BLSTM-MIL is adaptable to variable size audio clips. Conventional neural networks (e.g., DNN, CNN) are constrained on fixed size input.

In weakly labeled data, all the instances of a negative bag (training sample) are negative. But in positive training samples, only a small portion of the instances are positive and the rest are negative (noisy instances). Training supervised learning neural network classifiers (e.g., DNN, CNN, BLSTM) considers labels of all the instances of a positive training bag as positive. Due to the significant amount of label noise in positive training samples, supervised learning neural network approaches fail to achieve an optimal solution.

By contrast, in BLSTM-MIL training (equation 10), if at least one instance of a positive training bag is perfectly predicted as positive, the error E_i on the concerned bag is zero and the weights of the network will not be updated. Therefore, the BLSTM-MIL network training avoids weight updates due to noisy instances in positive training samples. Additionally, if all the instances of a negative bag are perfectly predicted as negative, then only the error E_i (equation 10) on the concerned bag is zero and the weights of the network are not updated. Otherwise, the weights are updated according to the error on the instance whose corresponding actual output is maximal among all the instances in the bag.

The ideal output of the network in response to any negative instance is 0, whereas for a positive instance it is 1. For negative bags, equation 10 characterizes the worst-case divergence of all instances in the bag from this ideal output. Minimizing this ensures that the response of the network to all instances from the bag is forced towards 0. In the ideal case, the system will output 0 in response to all inputs in the bag, and the divergence E_i will become 0.

For positive bags, equation 10 computes the best-case divergence of the instances of the bag from the ideal output of 1. Minimizing this ensures that the response of the network to at least one of the instances from the bag is forced towards 1. In the ideal case, one or more of the inputs in the bag will produce an output of 1, and the divergence E_i becomes 0.

Hence, using equations 10 and 9 during training and testing, the MIL adaptation of BLSTM treats negative training samples as supervised learning approaches do, given that negative samples do not contain noisy labels, but effectively avoids weight updates due to noisy labels in positive samples.

5 DATASETS

This section describes the weakly labeled audio datasets for our evaluation of high social anxiety and depression. We discuss the evaluations themselves in sections 6 and 7, respectively.

5.1 Social Anxiety

Because no previous dataset contained sponteneous speech labeled with speakers high in social anxiety, we built our own dataset from a laboratory-based study of a university student sample. The study was approved by the University of Virginia Institutional Review Board (IRB protocol 2013–0262-00) and conducted under the supervision of a licensed clinical psychologist and researcher with expertise in anxiety disorders. Because the collected audio data contains personal content and potentially identifying characteristics, this dataset cannot be shared with outside researchers given the need to protect confidentiality.

5.1.1 Participants.—A total of 105 participants ranging from 17 to 18 years of age (M = 19.24, SD = 1.84) completed the study in exchange for course credit or payment. Participants reported their races as 73.8% White, 13.4% Asian, 6.4% Black, 3.7% multiple, and 2.1% other (0.5% declined to answer) and their ethnicities as 90.9% Non-Hispanic/Latino and 7.0% Hispanic/Latino (2.1% declined to answer).

Participants were selected based on a screening survey at the start of the semester that included the Social Interaction Anxiety Scale (SIAS) and an item from the Social Phobia Scale (SPS) assessing anxiety about public speaking (“I get tense when I speak in front of other people” [57]). We included both measures because the speech task required participants to have anxiety about public speaking (it is possible to have social anxiety symptoms from fearing other social situations, such as dating, without fearing public speaking). Participants with SIAS scores less than or equal to one-quarter of a standard deviation (17 or less) below the mean of a previous undergraduate sample (M = 19.0, SD = 10.1; [57]) and who rated the public-speaking item as 0 (*not at all*), 1 (*slightly*), or 2 (*moderately*) were invited to join the Low Social Anxiety (Low SA) group; 60 enrolled. Those scoring greater than or equal to one standard deviation (30 or greater) above the SIAS mean and who rated the public-speaking item as 3 (*very*) or 4 (*extremely*) were invited to join the High Social Anxiety (High SA) group; 45 enrolled. This screening method or directly analogous ones were used in previous studies [11, 12]. The mean SIAS score for the High SA group (45.9, SD = 10.6) was close to the mean reported for a socially phobic sample (49.0, SD = 15.6; [34]), suggesting a strong analog sample.

5.1.2 Speech Task.—Participants were told researchers are interested in learning how people perceive and predict their own speaking abilities, utilizing a set of guidelines for effective speaking that the researchers were developing. Social anxiety was not mentioned

and participants did not know why they were invited to participate (until full debriefing after the study). As part of a larger study, participants were instructed to give an approximately 3-minute speech to the best of their ability on things they liked and disliked about college or their hometown in front of a large cassette video camera (to make the recording salient) with a web camera used for actual recording mounted on top. They were told the speech was being videotaped so that another researcher in an adjacent room would be able to watch and evaluate their performance. Videotaping was done to make the cover story as believable as possible and to heighten the participants' anxiety (following [18]). The present paper evaluates classification approaches only on these speeches (M length = 3 minutes). Participants were offered 1 minute to prepare their speech. If a participant paused for a significant period of time during the speech, the experimenter encouraged (but did not force) the person to continue talking for the full time.

5.1.3 State Anxiety Rating.—State affect ratings were collected using Qualtrics [82], a web-based survey data collection tool. Right after the speech, participants were asked to report their peak level of anxiety during the speech (''How anxious/calm did you feel at peak intensity during the speech?") on a visual analog scale (VAS). Participants used the computer mouse to indicate their anxiety from *extremely calm* (coded as −100) to *extremely anxious* (coded as +100), with *neither calm nor anxious* in the middle; only these three labels were visible. A VAS was used to minimize the influence that knowledge of one's prior responses could have on subsequent ratings (e.g., recalling a previous rating of 70). This was needed because multiple state affect ratings were obtained during the session as part of the larger study. Only 2 of the 45 High SA participants reported feeling more calm than anxious (i.e., a negative number); by contrast, 19 of the 60 Low SA participants reported feeling this way.

5.2 Depression

Distress Analysis Interview Corpus - Wizard of Oz (DAIC-WOZ) is a public dataset made available as part of the 2016 Audio-Visual Emotion Challenge and Workshop (AVEC 2016) [92]. It contains audio (with transcription) and video recordings of interviews from 189 participants in English with an animated virtual interviewer operated via a Wizard-of-Oz paradigm [28]. Each participant was assigned a score on the Patient Health Questionnaire-8, a self-report of depressive symptoms [44]. As part of the AVEC challenge [92], 142 participants were assigned to one of two classes: depressed (42) or non-depressed (100); the mean scores were 15.9 and 2.75, respectively. The present paper uses only the audio data (M length = 12 minutes, range= 10–25 minutes).

6 EVALUATION: SOCIAL ANXIETY

This section describes our evaluations of NN2Vec and BLSTM-MIL on social anxiety data (section 5.1) for detecting (a) speakers high in social anxiety symptoms (section 6.1) and (b) state anxiety (section 6.2).

All our evaluations were performed using leave-one-speaker-out cross-validation. To avoid overfitting, we randomly selected 30% of the audio clips for training the audio word

codebook (section 3.2) and 30% of the audio clips to train the NN2Vec model to generate vector representations (section 3.3) in each evaluation.

6.1 Social Anxiety Group

In this section we discuss the efficiency and applicability of our solution by investigating some key questions.

6.1.1 What are beneficial parameter configurations?—There are a number of parameters in our solution. Segment (which represents an instance in MIL) size is one important parameter. If the segment size is too small, it may contain only a fraction of the region indicating the positive class. If segment size is too large, the region indicating the positive class can be only a small fraction of the audio segment, and feature representation and the MIL classifier may fail to comprehend the indicative patterns.

A grid search over window size (section 3.1) from 500 ms to 10 seconds revealed that 1-second window size performs better on average.

Figure 5 shows our evaluation with various MIL instance segment sizes. As shown in figure 5, the F-1 score increases from 83.1% to 90.1% as instance segment size increases from 1 second to 13 seconds under the best parameter configuration. From 13 to 20 seconds the F-1 score is similar and then decreases as segment size increase further. Thus, the optimal segment instance size is 13 seconds.

Our BLSTM-MIL classifier has two layers with [100, 100] nodes, a 20% dropout rate, and two fully connected dense layers [20, 1] with sigmoid activation function. In the BLSTM-MIL approach, an instance segment is comprised of a sequence of NN2Vec vectors, which means that instance segments are sequences of audio states. BLSTM can learn from the current audio state as well as use relevant knowledge from the past to predict outcomes. If a region indicating the positive class (high symptom/disorder classification) is smaller than the segment window, BLSTM can still pass the knowledge through hidden states from one time step to another. Hence, for instance segment sizes 10 seconds to 20 seconds, our F-1 score is similar (88.88% to 90.1%). But as the segment size increases, the regions indicating the positive class decrease more and more relative to the instance segment; hence, BLSTM-MIL performance starts to decrease.

Next, we evaluated our solution with various codebook sizes (table 2), setting segment instance size to 13 seconds and window size to 1 second. When we evaluated NN2Vec vector (section 3.3) dimension from 10 to 100, we found that a vector dimension between 30 to 50 performs better on average for all codebook sizes. Thus, we extracted 30 dimensional NN2Vec vectors in our evaluation of codebook size. According to this evaluation, with an audio codebook size of 3500 we achieve the highest performance of an F-1 score of 90.1% and 91% accuracy.

6.1.2 Is NN2Vec Better?—We evaluate our BLSTM-MIL implementation (section 4.1) using NN2Vec feature representation against four baselines: audio words, I-vector, Emo2vec, and raw audio features (section 3.1). Table 3 shows the results.

The I-vector system is a technique [42] to map the high-dimensional GMM supervector space (generated from concatenating all the mean values of GMM) to low-dimensional space called total variability space. The basic idea of using an I-vector in human event detection [53, 102] is to represent each instance (window) using concatenated I-vector feature vectors extracted based on event-specific (e.g., emotion) GMM super vectors, and then to use these in the classifiers. Hence, the first step is event-specific GMM training. Since our audio clips are weakly labeled, in the positive class audio clips a major proportion of the data does not indicate positive class. Hence, we cannot generate accurate class or event-specific GMM models using weakly labeled data. As shown in table 3, the BLSTM-MIL classifier achieves an F-1 score of 74.7% and 79% accuracy using I-vector features.

Our BLSTM-MIL implementation achieves F-1 scores of 56.82% and 55.55% using raw audio features (272 total) and audio words, respectively. We represent the generated audio words by one hot encoded vector of the size of the codebook. Hence, the feature dimension from each window for our BLSTM classifier is the size of the codebook. BLSTM-MIL performance is low with these high-dimensional discrete feature representations, which do not convey the audio-state-to-class (syntactic) relationship.

Emo2vec is a feature modeling technique that uses audio words (section 3.2) to generate vectors with the characteristics that, if two windows appear in a similar context (i.e., similar surrounding windows) for a specific vocal event (class), then the vectors will be similar. Since our audio clips are weakly labeled, the majority of the co-occurred windows are common for both positive and negative classes. Hence, generated Emo2vec vectors cannot convey the audio-state-to-class relationship. Emo2vec feature modeling reduces the feature space significantly. Hence, Emo2vec performs better than raw audio features and audio words, achieving an F-1 score of 72.3%.

Our NN2Vec approach generates low-dimensional continuous feature representation. NN2Vec vectors generated from the windows represent the state of audio from the respective windows and convey the audio-state-to-class (syntactic) relationship in its representation. This representation makes the classification task from weakly labeled audio clips easier (section 3.3). According to table 3, NN2Vec achieves an F-1 score 17% higher and 12% higher accuracy than those of the I-vector, the best baseline.

6.1.3 Comparison With MIL Baselines.—This section discusses our evaluation of three MIL approaches using NN2Vec feature representation. Table 4 shows the results.

BLSTM-MIL implementation is similar to the approach described in section 6.1.1. The DNN-MIL classifier has three layers with [200, 200, 100] nodes and ReLU activation function, a 30% dropout rate, and one fully connected output dense layer [1] with sigmoid activation function. In this evaluation the mi-SVM implementation of MISVM toolkit [19] is used as the mi-SVM classifier. Previous studies [96] have shown that DNN-based MIL approaches perform better than SVM-based implementations. In our evaluation the DNN-MIL approach achieves an F-1 score of 85%, which is only 2% higher than that of the mi-SVM approach. Our BLSTM-MIL approach achieves an F-1 score 5.6% higher than that of the DNN-MIL approach, the best MIL baseline.

6.1.4 Comparison With Supervised Learning Algorithms.—

This section compares BLSTM-MIL with supervised learning approaches using NN2Vec features. We consider as baselines the four most-evaluated supervised learning algorithms from the recent literature for human vocal event detection: BLSTM, CNN, CNN-BLSTM, and DNN. Table 5 shows the results. Given that input audio clips have variable lengths and the baselines require fixed-length input, input sequences were transformed to fixed length by zero padding. The following network parameter configurations were optimized by performing a grid search of the parameter values.

The CNN implementation has three convolution layers, each with 200 convolution kernels (temporal extension of each filter is 4), and ReLU activation function. The CNN uses a 20% dropout rate and max pooling windows of size 4 and down-scaling factor 2. Two fully connected dense layers [20,1] with sigmoid activation function are attached, which makes a binary classification decision. The network is trained with the mean squared error loss function and RMSprop optimization. The CNN implementation achieves an F-1 score of 83.5% and 85.1% accuracy.

The BLSTM classifier has three layers with [100, 100, 100] nodes, a 20% dropout rate, and two fully connected dense layers [20, 1] with sigmoid activation function. The network is trained with the mean squared error loss function and RMSprop optimization. Because BLSTM can learn from current audio state as well as use knowledge from relevant previous states, it performs better than other approaches for weakly labeled data. In our evaluation, BLSTM is the best baseline approach, achieving an F-1 score of 86.66% and 88.1% accuracy.

The CNN-BLSTM is a serial combination of CNN and BLSTM. Frequency variance in the input signal is reduced by passing the input through two convolution layers, each with 100 convolution kernels (temporal extension of each filter is 4), a 20% dropout rate, and ReLU activation function. The network uses max pooling windows of size 4 and down-scaling factor 2. After frequency modeling is performed, the CNN output (higher-order representation of input features) is passed to the BLSTM layers. Two BLSTM layers [100, 100] and two fully connected layers [20, 1] are stacked at the end of the network architecture for the purpose of encoding long-range variability along the time axis and making the prediction.

The DNN implementation has three fully connected layers with [300, 300, 100] nodes and ReLU activation function, a 20% dropout rate, and one fully connected dense layer [1] with sigmoid activation function to make binary decisions. The DNN implementation achieves an F-1 score of 68.3%.

As shown in table 5, our BLSTM-MIL implementation (similar to section 6.1.1) achieves an F-1 score 3.9% higher than that of the best baseline (BLSTM) when both algorithms use NN2Vec vectors as features.

Our evaluations in section 6.1.2 and 6.1.4 show that the best baseline feature representation and supervised learning algorithm used in the literature are I-vector and BLSTM. Combining I-vector with BLSTM achieves an F-1 score of 71.4% and 76.2% accuracy.

Hence, combining NN2Vec vector features with our BLSTM-MIL approach achieves an F-1 score 20.7% higher than that of the best baseline approach.

6.2 State Anxiety

This section describes our evaluation for detecting state anxiety (anxious vs. calm). We performed a grid search on the model parameters window size, instance segment size, NN2Vec vector dimension, and audio codebook size from 500 ms to 10 seconds, 1 second to 30 seconds, 10 to 100, and 500 to 5000, respectively. The best parameter configuration was window size 1 second, instance segment size 10 seconds, a 30-dimensional NN2Vec vector and audio codebook size 3500. Our BLSTM-MIL approach using NN2Vec feature representation achieves an F-1 score of 93.49% and 90.2% accuracy.

To evaluate the effectiveness of NN2Vec representation for detection of state anxiety, we evaluated our BLSTM-MIL (section 4.1) implementation using four baseline feature representations. As shown in table 6, I-vector and Emo2vec achieve F-1 scores of 81.1% and 81.6%, which are higher than the corresponding scores for these feature representations in our evaluation for detecting speakers high versus low in social anxiety (table 3). Because labels in the present evaluation are related to state of speech, a major portion of the audio data indicates anxious vocal state. Hence, these feature modeling approaches perform better than they did in the previous evaluation (section 6.1.2). NN2Vec feature representation achieves an F-1 score about 12% higher than those of I-vector or Emo2vec.

Table 7 shows the results of our evaluation with MIL baselines and supervised learning baselines using NN2Vec features. BLSTM-MIL achieves an F-1 score 5.9% higher than that of DNN-MIL, the best MIL baseline, and an F-1 score 3.5% higher than that of BLSTM, the best supervised learning baseline. Many of the participants who reported feeling more anxious than calm at their peak level of anxiety may have expected to feel anxious for most of their speech. Because BLSTM stores and transfers prediction outcome-related knowledge from state to state, BLSTM better detects state anxiety than it does social anxiety group (discussed in section 6.1.4).

As shown in tables 6 and 7, combining Emo2vec, the best feature detection baseline, with BLSTM, the best of the MIL and supervised learning baselines, achieves an F-1 score of 80.5% and 72.6% accuracy. Hence, combining NN2Vec features with our BLSTM-MIL approach achieves an F-1 score 14.3% higher than that of the best baseline approach.

7 EVALUATION: DEPRESSION

This section describes our evaluation of the NN2Vec and BLSTM-MIL approach for detecting depressed speakers on the DAIC-WOZ dataset (section 5.2). We performed all evaluations using leave-one-speaker-out cross-validation. To avoid overfitting, we randomly selected 30% of the audio clips for training the audio word codebook (section 3.2) and 30% of the audio clips to train the NN2Vec model (section 3.3) in each evaluation.

We performed a grid search on the model parameters window size, instance segment size, NN2Vec vector dimension, and audio codebook size from 500 ms to 10 seconds, 1 second to

60 seconds, 10 to 100, and 500 to 10000, respectively. The best parameter configuration was window size 2 seconds, instance segment size 25 seconds, a 20-dimensional NN2Vec vector, and audio codebook size 5000. The BLSTM-MIL classifier has two layers with [100, 100] nodes, a 20% dropout rate, and two fully connected dense layers [30, 1] with sigmoid activation function.

We evaluated the performance of our BLSTM-MIL implementation (section 4.1) of NN2Vec against that of four baseline feature representations. As shown in table 8, our BLSTM-MIL approach using NN2Vec features achieves an F-1 score of 85.44% and 96.2% accuracy, whereas I-vector and Emo2vec achieve F-1 scores of 70.1% and 74.3%, respectively. Hence, NN2Vec achieves an F-1 score about 13% higher and 8% higher accuracy than those of Emo2vec, the best baseline feature representation.

Table 9 shows the results of our evaluation with MIL baselines and supervised learning baselines using NN2Vec features. Given that our supervised learning baselines require fixed-length input, audio input sequences were transformed to fixed-length by zero padding. The following network parameter configurations were optimized by performing a grid search of the parameter values. Our BLSTM-MIL achieves an F-1 score 10.5% higher than that of DNN-MIL, the best MIL baseline, and an F-1 score 9.7% higher than that of BLSTM, the best supervised learning baseline. The DNN-MIL classifier has three layers with [300, 300, 200] nodes and ReLU activation function, a 20% dropout rate, and one fully connected output dense layer [1] with sigmoid activation function. The BLSTM classifier has two layers with [200, 200] nodes, a 20% dropout rate, and two fully connected dense layers [50, 1] with sigmoid activation function.

We considered two of the most recent depression detection approaches [56, 62] evaluated on the DAIC-WOZ dataset as baselines. First, using I-vector features and Gaussian Probabilistic Linear Discriminant Analysis (G-PLDA) as the classifier [62] achieved an F-1 score of 57%. Second, DepAudioNet [56] encodes the temporal clues in the vocal modality using convolutional layers and predicts the presence of depression using LSTM layers. This serial combination of the CNN and the LSTM achieved an F-1 score of 52%. Hence, our BLSTM-MIL classifier using NN2Vec features achieves an F-1 score 33% higher than that of these other approaches.

8 DISCUSSION

Identifying individuals high in social anxiety and depression symptoms using our NN2Vec features achieves F-1 scores 17% and 13% higher, respectively, than those of the best baselines (I-vector, section 6.1.2; Emo2vec, table 8). Moreover, combining NN2Vec features with our BLSTM-MIL classifier achieves F-1 scores 20% and 33% higher, respectively, than those of the baselines (section 6.1.4 & 7). Detecting state anxiety using NN2Vec features achieves an F-1 score 11% higher than that of the best baseline (Emo2vec), and combining these with our BLSTM-MIL classifier achieves an F-1 score 14% higher than that of the best baseline (section 6.2).

In supervised learning, audio recordings are segmented into small fixed-length windows to train the CNN or CNN-BLSTM model. The labels of these windows are taken to be the same as the long audio clip-level labels. Hence, it is assumed that all the small windows in a positive long audio clip indicate high mental disorder symptoms. This, however, is not an efficient approach as it can result in a significant amount of label noise. Mental disorder symptoms in a long audio clip (segmented into a long sequence of windows) may be indicated by only a few seconds (a small subsequence of windows) of the clip, a fact ignored in assuming the label is strong. Due to the high label noise and limited training samples, convolution layers fail to generate effective higher-order representation of input features. To further support our statement, the DepAudioNet [56] approach applying the CNN-LSTM network using LLDs on the DAIC-WOZ dataset [92] achieved an F-1 score 8% lower than that obtained from using I-vector features and G-PLDA as the classifier [62] on the same dataset (section 7). By contrast, NN2Vec vectors map the audio from segmented windows into low-dimensional continuous feature space encoding the syntactic relationship between audio states (of windows), which facilitates a sequential classifier like BLSTM to effectively model the temporal properties in the speech signal (section 3.3). Hence, BLSTM and BLSTM-MIL classifiers perform better.

BLSTM networks are capable of learning and remembering over long sequences of inputs. This means that if a region (a small subsequence of windows) indicative of the positive class occurs in a long audio clip (long sequence of windows), BLSTM can pass that knowledge through hidden states. Studies [38] have shown that, as the sequence of windows becomes much longer for a limited training set, classifier performance starts to decline. Moreover, in a long positive weakly labeled audio clip, the portion of noise may significantly increase, making network optimization difficult. Hence, an MIL adaptation of BLSTM (BLSTM-MIL) performs better. In detecting speakers high in social anxiety symptoms (section 6.1), the window size was 1 second with 500 ms overlapping. Hence, input sequence size for the baseline BLSTM classifier (section 6.1.4) was 359 for 3-minute audio clips. The BLSTM-MIL classifier with input sequence size 25 (segment instance size 13) achieved an F-1 score about 4% higher than that of the best baseline BLSTM.

By contrast, detecting a depressed speaker (section 7) used a 2-second window size with 1 second overlapping. Hence, the input sequence size for BLSTM was 720 for 12-minute audio clips and achieved an F-1 score of 77.1%. In this evaluation, the BLSTM-MIL classifier with input sequence size 24 (segment instance size 25) achieved a 10% higher F-1 score. Hence, these evaluations (section 6.1.4 and table 9) show that as the input sequence size (audio clip length) increases, BLSTM-MIL performs increasingly better than BLSTM.

The ability to identify symptomatic individuals from their audio data represents an objective indicator of symptom severity that can complement health-care providers' other assessment modalities and inform treatment. Moreover, because vocal analysis does not require extensive equipment and is readily accessible (speech is ubiquitous in natural settings), nonintrusive (it does not require a special wearable monitor), and not burdensome (it does not require additional assessment time or client responses), it is scalable, which is important given the vast number of people with social anxiety and depression who receive no help [13]. Additionally, the unique source of the data (animated virtual clinical interview) for the

depression detection study further supports the possibility of using vocal analysis as a remote assessment tool and one that may eventually be possible to administer via artificial intelligence. This is especially exciting in light of the difficulty in identifying and disseminating care to people facing considerable barriers to seeking treatment for social anxiety [63] and depression [20].

Further, the ability to detect state anxiety from audio data for both individuals high and low in social anxiety symptoms opens new possibilities for assessment, treatment, and prevention. Implementing vocal analysis with mobile technologies (e.g., smartphones) would give health-care providers an objective marker of clients' anxiety as it unfolds outside of the treatment setting, and combining this with other data (e.g., location, actigraphy) could help clarify the antecedents and consequences of clients' anxious states. Providers could even collect such idiographic time series data from clients before treatment to understand each client's dynamic processes and personalize treatment from the start [22]. Moreover, pairing passive outcome monitoring with mobile interventions (e.g., skills training apps) would enable the timely delivery of just-in-time interventions that may offer relief and efficiently promote skills acquisition and generalization. Along these lines, detecting state anxiety from audio data may one day be used to identify changes in speech that suggest a person may be transitioning to a higher-risk state and could benefit from preventive services to avoid the worsening of symptoms.

The present study has several limitations related to sampling. First, we used an analog sample of people high versus low in social anxiety symptoms for whom no formal diagnoses of social anxiety disorder had been established. Second, we analyzed speech audio data from only one situation (a speech stressor task), so future work would benefit from sampling speech from a wider range of both social and nonsocial situations to determine the boundaries of the models' predictive validity. Third, our state anxiety analyses are based on participants' responses to only one question about that situation (their self-reported peak levels of anxiety during the speech).

Finally, we wish to emphasize that implementation of our approach, even if designed to support health-care providers, must include the informed consent of clients, who should be allowed to discontinue the monitoring at any time, and robust privacy protections. It is important to note that our approach does not use the semantics (transcribed text) of the client's speech and that the proposed feature extraction is irreversible (section 3.2), thereby ensuring clients' privacy. Any feedback provided to the client about increases in symptoms or state anxiety would ultimately be paired with treatment resources or other services (e.g., interventions) that the client can use to seek relief. Further, future research is needed to evaluate the feasibility, acceptability, and safety of our approach before providers implement the approach on a large scale in the community.

9 CONCLUSION

Although depression and social anxiety disorder are highly prevalent, many depressed and socially anxious people do not receive treatment. Current assessments for these disorders are typically based on client self-report and clinical judgment and therefore are subject to

subjective biases, burdensome to administer, and inaccessible to clients who face barriers to seeking treatment. Objective indicators of depression and social anxiety would help advance approaches to identification, assessment, prevention, and treatment. We propose a weakly supervised learning framework for detecting symptomatic individuals and state affect from long speech audio data. Specifically, we present a novel feature modeling technique named NN2Vec that identifies and exploits the inherent relationship between vocal states and symptoms/affective states. In addition, we present a new MIL adaptation of the BLSTM classifier, named BLSTM-MIL, to comprehend the temporal dynamics of vocal states in weakly labeled data. We evaluated our framework on 105 participants' spontaneous audio speech data weakly labeled with speakers high in social anxiety. Our NN2Vec and BLSTM-MIL approach achieved an F-1 score of 90.1% and 90% accuracy in detecting speakers high versus low in social anxiety symptoms, and an F-1 score of 93.49% and 90.2% accuracy in detecting state anxiety (anxious vs. calm). These F-1 scores are 20.7% and 14.3% higher, respectively, than those of the best baselines. To our knowledge, this study is the first to attempt such detection using weakly labeled audio data. Using audio clips from virtual clinical interviews, our approach also achieved an F-1 score of 85.44% and 96.7% accuracy in detecting speakers high versus low in depressive symptoms. This F-1 score is 33% higher than those of the two most recent approaches from the literature. Readily accessible, not intrusive or burdensome, and free of extensive equipment, the NN2Vec and BLSTM-MIL framework is a scalable complement to health-care providers' self-report, interview, and other assessment modalities.

ACKNOWLEDGMENTS

This work was supported, in part, by DGIST Research and Development Program (CPS Global center) funded by the Ministry of Science, ICT and Future Planning, NSF CNS-1646470, and NSF grant IIS-1521722. It was also supported in part by NIMH grants R34MH106770 and R01MH113752 to B. Teachman.

REFERENCES

[1]. Alden Lynn E and Wallace Scott T. 1995. Social phobia and social appraisal in successful and unsuccessful social interactions. Behaviour Research and Therapy 33, 5 (1995), 497–505. [PubMed: 7598670]

[2]. Anagnostopoulos Christos-Nikolaos, Iliou Theodoros, and Giannoukos Ioannis. 2015. Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011. Artificial Intelligence Review 43, 2 (2015), 155–177.

[3]. Andrews Stuart, Tsochantaridis Ioannis, and Hofmann Thomas. 2003 Support vector machines for multiple-instance learning. In Advances in neural information processing systems 577–584.

[4]. American Psychiatric Association et al. 2013 Diagnostic and statistical manual of mental disorders (DSM-5®) American Psychiatric Pub.

[5]. Bengio Yoshua, Ducharme Réjean, Vincent Pascal, and Jauvin Christian. 2003. A neural probabilistic language model. Journal of machine learning research 3, Feb (2003), 1137–1155.

[6]. Boukhechba Mehdi, Huang Yu, Chow Philip, Fua Karl, Teachman Bethany A, and Barnes Laura E. 2017 Monitoring social anxiety from mobility and communication patterns. In Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers. ACM, 749–753.

[7]. Chao Linlin, Tao Jianhua, Yang Minghao, and Li Ya. 2014 Improving generation performance of speech emotion recognition by denoising autoencoders. In Chinese Spoken Language Processing (ISCSLP), 2014 9th International Symposium on. IEEE, 341–344.

[8]. Cheplygina Veronika, Tax David MJ, and Loog Marco. 2015. Multiple instance learning with bag dissimilarities. Pattern Recognition 48, 1 (2015), 264–275.

[9]. Chow Philip, Xiong Haoyi, Fua Karl, Bonelli Wes, Teachman Bethany A, and Barnes Laura E. 2016. SAD: Social anxiety and depression monitoring system for college students (2016).

[10]. Chung Junyoung, Gulcehre Caglar, Cho KyungHyun, and Bengio Yoshua. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv: 1412.3555 (2014).

[11]. Clerkin Elise M and Teachman Bethany A. 2010. Training implicit social anxiety associations: An experimental intervention. Journal of anxiety disorders 24, 3 (2010), 300–308. [PubMed: 20102788]

[12]. Clerkin Elise M and Teachman Bethany A. 2011. Training interpretation biases among individuals with symptoms of obsessive compulsive disorder. Journal of Behavior Therapy and Experimental Psychiatry 42, 3 (2011), 337–343. [PubMed: 21371415]

[13]. Collins Kerry A, Westra Henny A, Dozois David JA, and Burns David D. 2004. Gaps in accessing treatment for anxiety and depression: challenges for the delivery of care. Clinical psychology review 24, 5 (2004), 583–616. [PubMed: 15325746]

[14]. Cummins Nicholas, Epps Julien, Breakspear Michael, and Goecke Roland. 2011 An investigation of depressed speech detection: Features and normalization. In Twelfth Annual Conference of the International Speech Communication Association.

[15]. Cummins Nicholas, Scherer Stefan, Krajewski Jarek, Schnieder Sebastian, Epps Julien, and Quatieri Thomas F. 2015. A review of depression and suicide risk assessment using speech analysis. Speech Communication 71 (2015), 10–49.

[16]. Cummins Nicholas, Vlasenko Bogdan, Sagha Hesam, and Schuller Björn. 2017 Enhancing speech-based depression detection through gender dependent vowel-level formant. In Proc. of Conference on Artificial Intelligence in Medicine. Springer 5.

[17]. Depression 2017 World Health Organizsation http://www.who.int/mediacentre/factsheets/fs369/en/.

[18]. Dickerson Sally S and Kemeny Margaret E. 2004. Acute stressors and cortisol responses: a theoretical integration and synthesis of laboratory research. Psychological bulletin 130, 3 (2004), 355. [PubMed: 15122924]

[19]. Doran Gary and Ray Soumya. 2014. A theoretical and empirical analysis of support vector machine methods for multiple-instance classification. Machine Learning 97, 1–2 (2014), 79–102.

[20]. Endicott Jean et al. 1996. Barriers to seeking treatment for major depression. Depression and anxiety 4, 6 (1996), 273–278. [PubMed: 9166655]

[21]. Eyben Florian, Scherer Klaus R, Schuller Björn W, Sundberg Johan, André Elisabeth, Busso Carlos, Devillers Laurence Y, Epps Julien, Laukka Petri, Narayanan Shrikanth S, et al. 2016. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. IEEE Transactions on Affective Computing 7, 2 (2016), 190–202.

[22]. Fisher Aaron J. 2015. Toward a dynamic model of psychological assessment: Implications for personalized care. Journal of Consulting and Clinical Psychology 83, 4 (2015), 825. [PubMed: 26009781]

[23]. Flint Alistair J, Black Sandra E, Campbell-Taylor Irene, Gailey Gillian F, and Levinton Carey. 1993. Abnormal speech articulation, psychomotor retardation, and subcortical dysfunction in major depression. Journal of psychiatric research 27, 3 (1993), 309–319. [PubMed: 8295162]

[24]. France Daniel Joseph, Shiavi Richard G, Silverman Stephen, Silverman Marilyn, and Wilkes M. 2000. Acoustical properties of speech as indicators of depression and suicidal risk. IEEE transactions on Biomedical Engineering 47, 7 (2000), 829–837. [PubMed: 10916253]

[25]. Garcia-Romero Daniel and Espy-Wilson Carol Y. 2011. Analysis of i-vector Length Normalization in Speaker Recognition Systems. In Interspeech, Vol. 2011 249–252.

[26]. Gillespie Stephanie, Moore Elliot, Laures-Gore Jacqueline, Farina Matthew, Russell Scott, and Logan Yash-Yee. 2017 Detecting stress and depression in adults with aphasia through speech analysis. In Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on. IEEE, 5140–5144.

[27]. Goldberg Yoav and Levy Omer. 2014. word2vec explained: Deriving mikolov et al.'s negative-sampling word-embedding method. arXiv preprint arXiv:1402.3722 (2014).

[28]. Green Paul and Wei-Haas Lisa. 1985 The rapid development of user interfaces: Experience with the Wizard of Oz method. In Proceedings of the Human Factors Society Annual Meeting, Vol. 29. SAGE Publications Sage CA: Los Angeles, CA, 470–474.

[29]. Hall Judith A, Harrigan Jinni A, and Rosenthal Robert. 1995. Nonverbal behavior in clinician-patient interaction. Applied and Preventive Psychology 4, 1 (1995), 21–37.

[30]. Hamilton Max. 1960. A rating scale for depression. Journal of neurology, neurosurgery, and psychiatry 23, 1 (1960), 56.

[31]. Hannun Awni, Case Carl, Casper Jared, Catanzaro Bryan, Diamos Greg, Elsen Erich, Prenger Ryan, Satheesh Sanjeev, Sengupta Shubho, Coates Adam, et al. 2014. Deep speech: Scaling up end-to-end speech recognition. arXiv preprint arXiv:1412.5567 (2014).

[32]. Hawton Keith, Casañas i Comabella Carolina, Haw Camilla, and Saunders Kate. 2013. Risk factors for suicide in individuals with depression: a systematic review. Journal of affective disorders 147, 1 (2013), 17–28. [PubMed: 23411024]

[33]. Hecht-Nielsen Robert et al. 1988. Theory of the backpropagation neural network. Neural Networks 1, Supplement-1 (1988), 445–448.

[34]. Heimberg Richard G, Holt Craig S, Schneier Franklin R, Spitzer Robert L, and Liebowitz Michael R. 1993. The issue of subtypes in the diagnosis of social phobia. Journal of Anxiety Disorders 7, 3 (1993), 249–269.

[35]. Huang Yu, Gong Jiaqi, Rucker Mark, Chow Philip, Fua Karl, Gerber Matthew S, Teachman Bethany, and Barnes Laura E. 2017 Discovery of Behavioral Markers of Social Anxiety from Smartphone Sensor Data. In Proceedings of the 1st Workshop on Digital Biomarkers. ACM, 9–14.

[36]. Huang Zhengwei, Dong Ming, Mao Qirong, and Zhan Yongzhao. 2014 Speech emotion recognition using CNN. In Proceedings of the 22nd ACM international conference on Multimedia. ACM, 801–804.

[37]. Jager Mark, Knoll Christian, and Hamprecht Fred A. 2008 Weakly supervised learning of a classifier for unusual event detection. IEEE Transactions on Image Processing 17, 9 (2008), 1700–1708. [PubMed: 18701402]

[38]. Jozefowicz Rafal, Vinyals Oriol, Schuster Mike, Shazeer Noam, and Wu Yonghui. 2016 Exploring the limits of language modeling. arXiv preprint arXiv:1602.02410 (2016).

[39]. Karafiát Martin, Burget Lukáš, Matějka Pavel, Glembek Ondřej, and Černocky Jan`. 2011 iVector-based discriminative adaptation for automatic speech recognition. In Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on. IEEE, 152–157.

[40]. Kaushik Lakshmish, Sangwan Abhijeet, and Hansen John HL. 2015 Laughter and filler detection in naturalistic audio. In INTERSPEECH 2509–2513.

[41]. Kazdin Alan E. 2017. Addressing the treatment gap: A key challenge for extending evidence-based psychosocial interventions. Behaviour research and therapy 88 (2017), 7–18. [PubMed: 28110678]

[42]. Kenny Patrick, Boulianne Gilles, and Dumouchel Pierre. 2005. Eigenvoice modeling with sparse training data. IEEE transactions on speech and audio processing 13, 3 (2005), 345–354.

[43]. Kessler Ronald C, Petukhova Maria, Sampson Nancy A, Zaslavsky Alan M, and Wittchen Hans-Ullrich. 2012. Twelve-month and lifetime prevalence and lifetime morbid risk of anxiety and mood disorders in the United States. International journal of methods in psychiatric research 21, 3 (2012), 169–184. [PubMed: 22865617]

[44]. Kroenke Kurt, Strine Tara W, Spitzer Robert L, Williams Janet BW, Berry Joyce T, and Mokdad Ali H. 2009. The PHQ-8 as a measure of current depression in the general population. Journal of affective disorders 114, 1 (2009), 163–173. [PubMed: 18752852]

[45]. Kumar Anurag, Hegde Rajesh M, Singh Rita, and Raj Bhiksha. 2013 Event detection in short duration audio using gaussian mixture model and random forest classifier. In Signal Processing Conference (EUSIPCO), 2013 Proceedings of the 21st European. IEEE, 1–5.

[46]. Kumar Anurag and Raj Bhiksha. 2016 Audio event detection using weakly labeled data. In Proceedings of the 2016 ACM on Multimedia Conference. ACM, 1038–1047.

[47]. Kumar Anurag and Raj Bhiksha. 2016 Weakly supervised scalable audio content analysis. In Multimedia and Expo (ICME), 2016 IEEE International Conference on. IEEE, 1–6.

[48]. Laukka Petri, Linnman Clas, Åhs Fredrik, Pissiota Anna, Frans Örjan, Faria Vanda, Michelgård Åsa, Appel Lieuwe, Fredrikson Mats, and Furmark Tomas. 2008. In a nervous voice: Acoustic analysis and perception of anxiety in social phobics' speech. Journal of Nonverbal Behavior 32, 4 (2008), 195.

[49]. Lavner Yizhar, Cohen Rami, Ruinskiy Dima, and IJzerman Hans. 2016 Baby cry detection in domestic environment using deep learning. In Science of Electrical Engineering (ICSEE), IEEE International Conference on the. IEEE, 1–5.

[50]. Le Duc, Aldeneh Zakaria, and Provost Emily Mower. 2017 Discretized continuous speech emotion recognition with multi-task deep recurrent neural network. Interspeech, 2017 (to apear) (2017).

[51]. Le Duc and Provost Emily Mower. 2013 Emotion recognition from spontaneous speech using hidden markov models with deep belief networks. In Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on. IEEE, 216–221.

[52]. Lim Wootaek, Jang Daeyoung, and Lee Taejin. 2016 Speech emotion recognition using convolutional and recurrent neural networks. In Signal and information processing association annual summit and conference (APSIPA), 2016 Asia-Pacific. IEEE, 1–4.

[53]. Lopez-Otero Paula, Docio-Fernandez Laura, and Garcia-Mateo Carmen. 2014. iVectors for continuous emotion recognition. Training 45 (2014), 50.

[54]. Losiak Wladyslaw, Blaut Agata, Klosowska Joanna, and Slowik Natalia. 2016. Social Anxiety, Affect, Cortisol Response and Performance on a Speech Task. Psychopathology 49, 1 (2016), 24–30. [PubMed: 26650543]

[55]. Low Lu-Shih Alex, Maddage Namunu C, Lech Margaret, Sheeber Lisa B, and Allen Nicholas B. 2011 Detection of clinical depression in adolescents' speech during family interactions. IEEE Transactions on Biomedical Engineering 58, 3 (2011), 574–586. [PubMed: 21075715]

[56]. Ma Xingchen, Yang Hongyu, Chen Qiang, Huang Di, and Wang Yunhong. 2016 DepAudioNet: An Efficient Deep Model for Audio based Depression Classification. In Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge. ACM, 35–42.

[57]. Mattick Richard P and Clarke J Christopher. 1998. Development and validation of measures of social phobia scrutiny fear and social interaction anxiety. Behaviour research and therapy 36, 4 (1998), 455–470. [PubMed: 9670605]

[58]. McGirr Alexander, Paris Joel, Lesage Alain, Renaud Johanne, and Turecki Gustavo. 2007. Risk factors for suicide completion in borderline personality disorder: a case-control study of cluster B comorbidity and impulsive aggression. The Journal of clinical psychiatry (2007).

[59]. Mikolov Tomáš. 2012. Statistical language models based on neural networks. Presentation at Google, Mountain View, 2nd April (2012).

[60]. Mitchell Alex J, Vaze Amol, and Rao Sanjay. 2009. Clinical diagnosis of depression in primary care: a meta-analysis. The Lancet 374, 9690 (2009), 609–619.

[61]. Mundt James C, Snyder Peter J, Cannizzaro Michael S, Chappie Kara, and Geralts Dayna S. 2007 Voice acoustic measures of depression severity and treatment response collected via interactive voice response (IVR) technology. Journal of neurolinguistics 20, 1 (2007), 50–64. [PubMed: 21253440]

[62]. Nasir Md, Jati Arindam, Shivakumar Prashanth Gurunath, Chakravarthula Sandeep Nallan, and Georgiou Panayiotis. 2016 Multimodal and multiresolution depression detection from speech and facial landmark features. In Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge. ACM, 43–50.

[63]. Olfson Mark, Guardino Mary, Struening Elmer, Schneier Franklin R, Hellman Fred, and Klein Donald F. 2000. Barriers to the treatment of social anxiety. American Journal of Psychiatry 157, 4 (2000), 521–527. [PubMed: 10739410]

[64]. World Health Organization et al. 2016 International Statistical Classification of Diseases and Related Health Problems 10th Revision (IDC-10). 2010

[65]. Pancoast Stephanie and Akbacak Murat. 2012 Bag-of-Audio-Words Approach for Multimedia Event Classification. In Interspeech 2105–2108.

[66]. Pennington Jeffrey, Socher Richard, and Manning Christopher. 2014 Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). 1532–1543.

[67]. Plate Tony. [n. d.]. Distributed representations. Encyclopedia of Cognitive Science ([n. d.]).

[68]. Ramakrishnan S and El Emary Ibrahiem MM. 2013. Speech emotion recognition approaches in human computer interaction. Telecommunication Systems (2013), 1–12.

[69]. Rani Barkha. 2016 I-Vector based depression level estimation technique. In Recent Trends in Electronics, Information & Communication Technology (RTEICT), IEEE International Conference on. IEEE, 2067–2071.

[70]. Rao K Sreenivasa, Kumar Tummala Pavan, Anusha Kusam, Leela Bathina, Bhavana Ingilela, and Gowtham SVSK. 2012. Emotion recognition from speech. International Journal of Computer Science and Information Technologies 3, 2 (2012), 3603–3607.

[71]. Rapee Ronald M and Lim Lina 1992. Discrepancy between self-and observer ratings of performance in social phobics. Journal of abnormal psychology 101, 4 (1992), 728. [PubMed: 1430614]

[72]. Rawat Shourabh, Schulam Peter F, Burger Susanne, Ding Duo, Wang Yipei, and Metze Florian. 2013. Robust audio-codebooks for large-scale event detection in consumer videos (2013).

[73]. Rieger Steven A, Muraleedharan Rajani, and Ramachandran Ravi P. 2014 Speech based emotion recognition using spectral feature extraction and an ensemble of kNN classifiers. In Chinese Spoken Language Processing (ISCSLP), 2014 9th International Symposium on. IEEE, 589–593.

[74]. Rong Xin. 2014. word2vec parameter learning explained. arXiv preprint arXiv:1411.2738 (2014).

[75]. Sainath Tara N, Kingsbury Brian, Saon George, Soltau Hagen, Mohamed Abdel-rahman, Dahl George, and Ramabhadran Bhuvana. 2015. Deep convolutional neural networks for large-scale speech tasks. Neural Networks 64 (2015), 39–48. [PubMed: 25439765]

[76]. Salekin Asif, Chen Zeya, Ahmed Mohsin Y, Lach John, Metz Donna, De La Haye Kayla, Bell Brooke, and Stankovic John A. 2017. Distant Emotion Recognition. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 1, 3 (2017), 96.

[77]. Salekin Asif, Wang Hongning, Williams Kristine, and Stankovic John. 2017 DAVE: Detecting Agitated Vocal Events. In Connected Health: Applications, Systems and Engineering Technologies (CHASE), 2017 IEEE/ACM International Conference on. IEEE, 157–166.

[78]. Sapra Ankur, Panwar Nikhil, and Panwar Sohan. 2013. Emotion recognition from speech. International Journal of Emerging Technology and Advanced Engineering 3 (2013), 341–345.

[79]. Scherer Stefan, Morency Louis-Philippe, Gratch Jonathan, and Pestian John. 2015 Reduced vowel space is a robust indicator of psychological distress: a cross-corpus analysis. In Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on. IEEE, 4789–4793.

[80]. Scherer Stefan, Stratou Giota, Gratch Jonathan, and Morency Louis-Philippe. 2013 Investigating voice quality as a speaker-independent indicator of depression and PTSD. In Interspeech 847–851.

[81]. Silber-Varod Vered, Kreiner Hamutal, Lovett Ronen, Levi-Belz Yossi, and Amir Noam. 2016 Do social anxiety individuals hesitate more? The prosodic profile of hesitation disfluencies in Social Anxiety Disorder individuals. Proceedings of Speech Prosody 2016 (SP2016) (2016), 1211–1215.

[82]. Snow Jonathan and Mann M. 2013. Qualtrics survey software: handbook for research professionals http://www.qualtrics.com (2013).

[83]. Sobin Christina and Sackeim Harold A. 1997. Psychomotor symptoms of depression. The American journal of psychiatry 154, 1 (1997), 4. [PubMed: 8988952]

[84]. Specht Donald F. 1990 Probabilistic neural networks. Neural networks 3, 1 (1990), 109–118.

[85]. Stasak Brian, Epps Julien, Cummins Nicholas, and Goecke Roland. 2016 An Investigation of Emotional Speech in Depression Classification. In INTERSPEECH 485–489.

[86]. Stolar Melissa N, Lech Margaret, and Allen Nicholas B. 2015 Detection of depression in adolescents based on statistical modeling of emotional influences in parent-adolescent conversations. In Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on. IEEE, 987–991.

[87]. Stopa Lusia and Clark David M. 1993. Cognitive processes in social phobia. Behaviour Research and Therapy 31, 3 (1993), 255–267. [PubMed: 8476400]

[88]. Sturim Douglas, Torres-Carrasquillo Pedro A, Quatieri Thomas F, Malyska Nicolas, and McCree Alan. 2011 Automatic detection of depression in speech using gaussian mixture modeling with factor analysis. In Twelfth Annual Conference of the International Speech Communication Association.

[89]. Su Ting-Wei, Liu Jen-Yu, and Yang Yi-Hsuan. 2017 Weakly-supervised audio event detection using event-specific Gaussian filters and fully convolutional networks. In Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on. IEEE, 791–795.

[90]. Teachman Bethany A. 2014. No appointment necessary: Treating mental illness outside the therapist's office. Perspectives on Psychological Science 9, 1 (2014), 85–87. [PubMed: 26173245]

[91]. Thomée Sara, Härenstam Annika, and Hagberg Mats. 2011. Mobile phone use and stress, sleep disturbances, and symptoms of depression among young adults-a prospective cohort study. BMC public health 11, 1 (2011), 66. [PubMed: 21281471]

[92]. Valstar Michel, Gratch Jonathan, Schuller Björn, Ringeval Fabien, Lalanne Dennis, Torres Mercedes Torres, Scherer Stefan, Stratou Giota, Cowie Roddy, and Pantic Maja. 2016 Avec 2016: Depression, mood, and emotion recognition workshop and challenge. In Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge. ACM, 3–10.

[93]. Vlasenko Bogdan, Sagha Hesam, Cummins Nicholas, and Schuller Björn. 2017. Implementing gender-dependent vowel-level analysis for boosting speech-based depression recognition. Proc. Interspeech 2017 (2017), 3266–3270.

[94]. Voci Sabrina C, Beitchman Joseph H, Brownlie EB, and Wilson Beth. 2006 Social anxiety in late adolescence: The importance of early childhood language impairment. Journal of anxiety disorders 20, 7 (2006), 915–930. [PubMed: 16503112]

[95]. Wang Philip S, Berglund Patricia, Olfson Mark, Pincus Harold A, Wells Kenneth B, and Kessler Ronald C. 2005. Failure and delay in initial treatment contact after first onset of mental disorders in the National Comorbidity Survey Replication. Archives of general psychiatry 62, 6 (2005), 603–613. [PubMed: 15939838]

[96]. Wang Xinggang, Yan Yongluan, Tang Peng, Bai Xiang, and Liu Wenyu. 2016. Revisiting Multiple Instance Neural Networks. arXiv preprint arXiv:1610.02501 (2016).

[97]. Weeks Justin W, Lee Chao-Yang, Reilly Alison R, Howell Ashley N, France Christopher, Kowalsky Jennifer M, and Bush Ashley. 2012. "The Sound of Fear": Assessing vocal fundamental frequency as a physiological indicator of social anxiety disorder. Journal of anxiety disorders 26, 8 (2012), 811–822. [PubMed: 23070030]

[98]. Weeks Justin W, Srivastav Akanksha, Howell Ashley N, and Menatti Andrew R. 2016. "Speaking More than Words": Classifying Men with Social Anxiety Disorder via Vocal Acoustic Analyses of Diagnostic Interviews. Journal of Psychopathology and Behavioral Assessment 38, 1 (2016), 30–41.

[99]. Weiller E, Bisserbe J-C, Boyer P, Lepine J-P, and Lecrubier Y. 1996. Social phobia in general health care: an unrecognised undertreated disabling disorder. The British Journal of Psychiatry 168, 2 (1996), 169–174. [PubMed: 8837906]

[100]. Wells Adrian, Clark David M, Salkovskis Paul, Ludgate John, Hackmann Ann, and Gelder Michael. 1995. Social phobia: The role of in-situation safety behaviors in maintaining anxiety and negative beliefs. Behavior Therapy 26, 1 (1995), 153–161.

[101]. Wu Jiajun, Yu Yinan, Huang Chang, and Yu Kai. 2015 Deep multiple instance learning for image classification and auto-annotation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 3460–3469.

[102]. Xia Rui and Liu Yang. 2012 Using i-vector space model for emotion recognition. In Thirteenth Annual Conference of the International Speech Communication Association.

[103]. Zhang Teng and Wu Ji. 2015 Speech emotion recognition with i-vector feature and RNN model 524–528.

[104]. Zhang Wan Li, Li Guo Xin, and Gao Wei. 2014 The Research of Speech Emotion Recognition Based on Gaussian Mixture Model. In Applied Mechanics and Materials, Vol. 668 Trans Tech Publ, 1126–1129.

CCS Concepts:

- Computing methodologies → Neural networks
- **Applied computing** → *Health informatics*

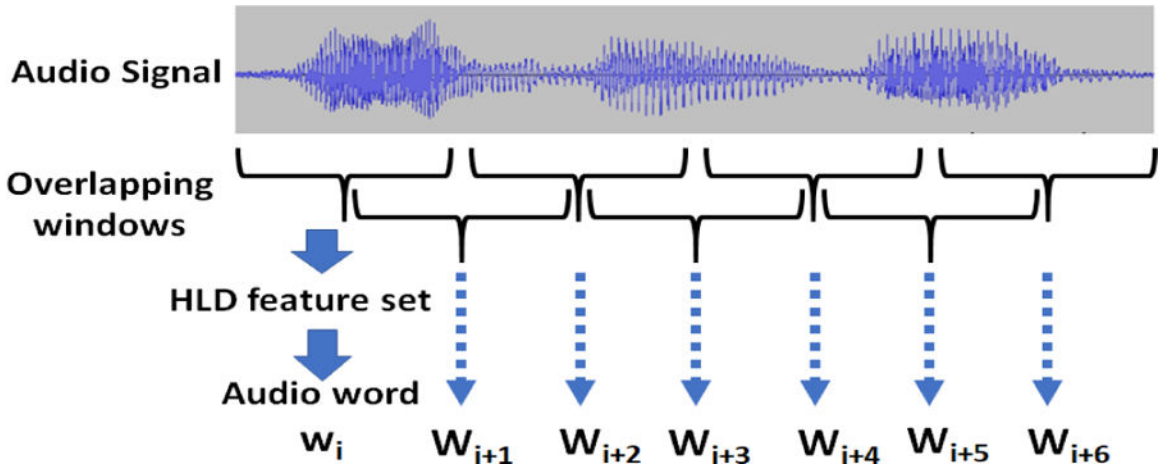

Fig. 1.
Conversion of audio signal to sequence of audio words using audio-codebook method

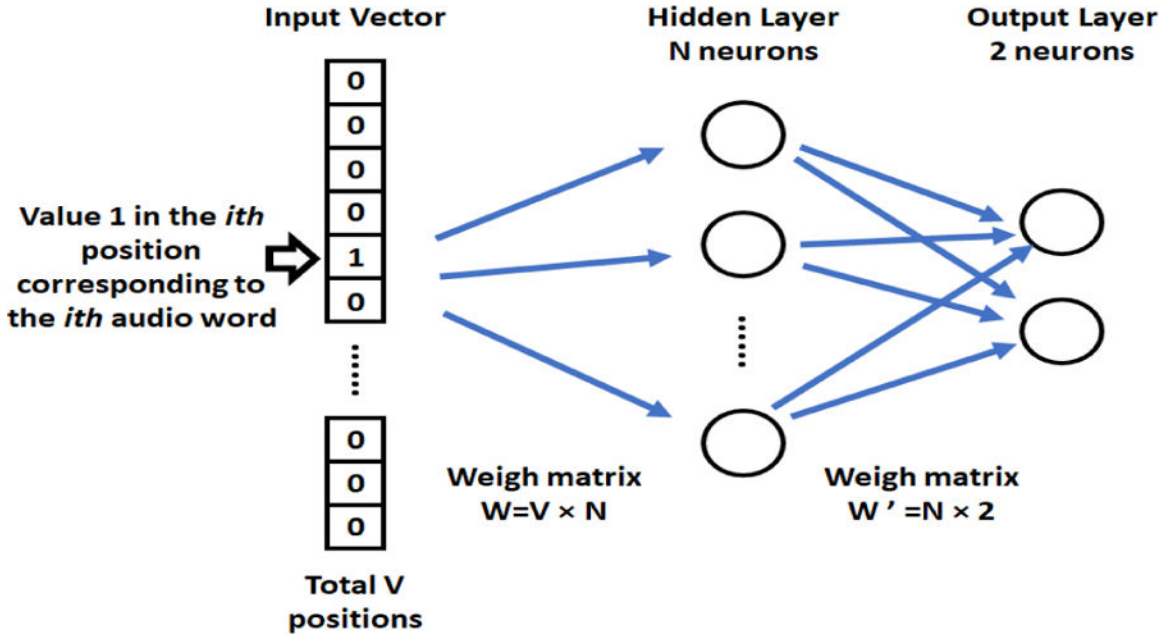

Fig. 2.
NN2Vec fully connected neural network

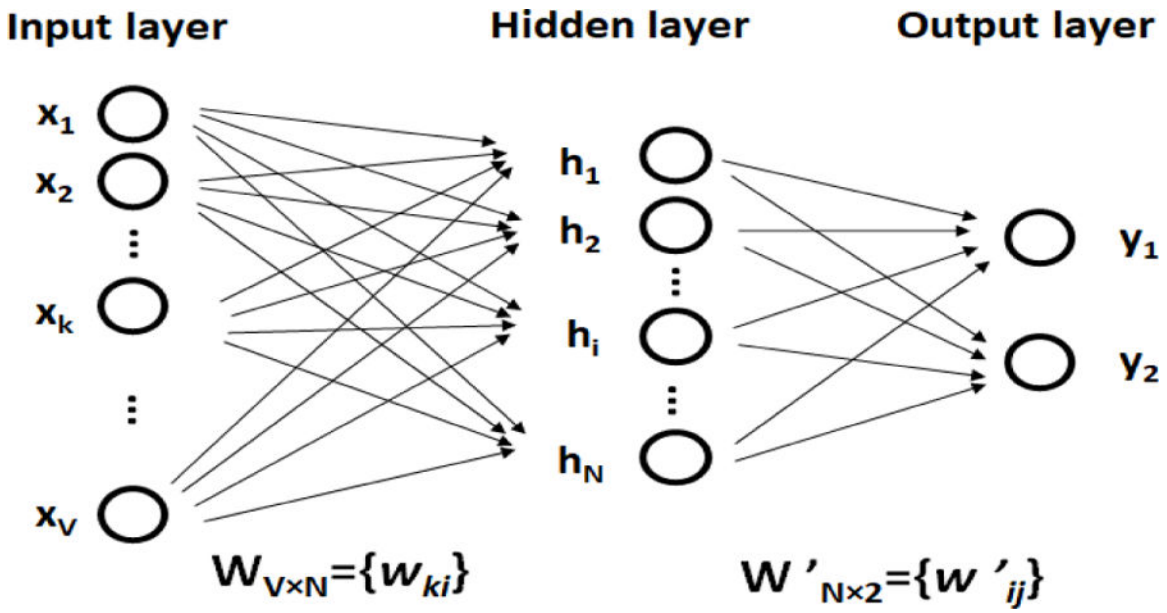

Fig. 3.
NN2Vec model with binary output

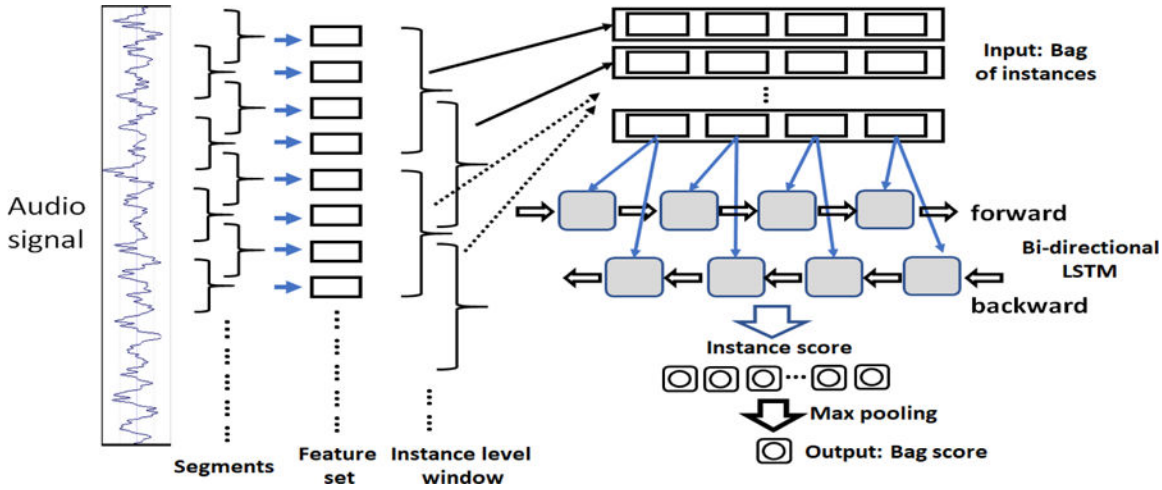

Fig. 4.
Bidirectional LSTM multiple instance learning classifier (BLSTM-MIL)

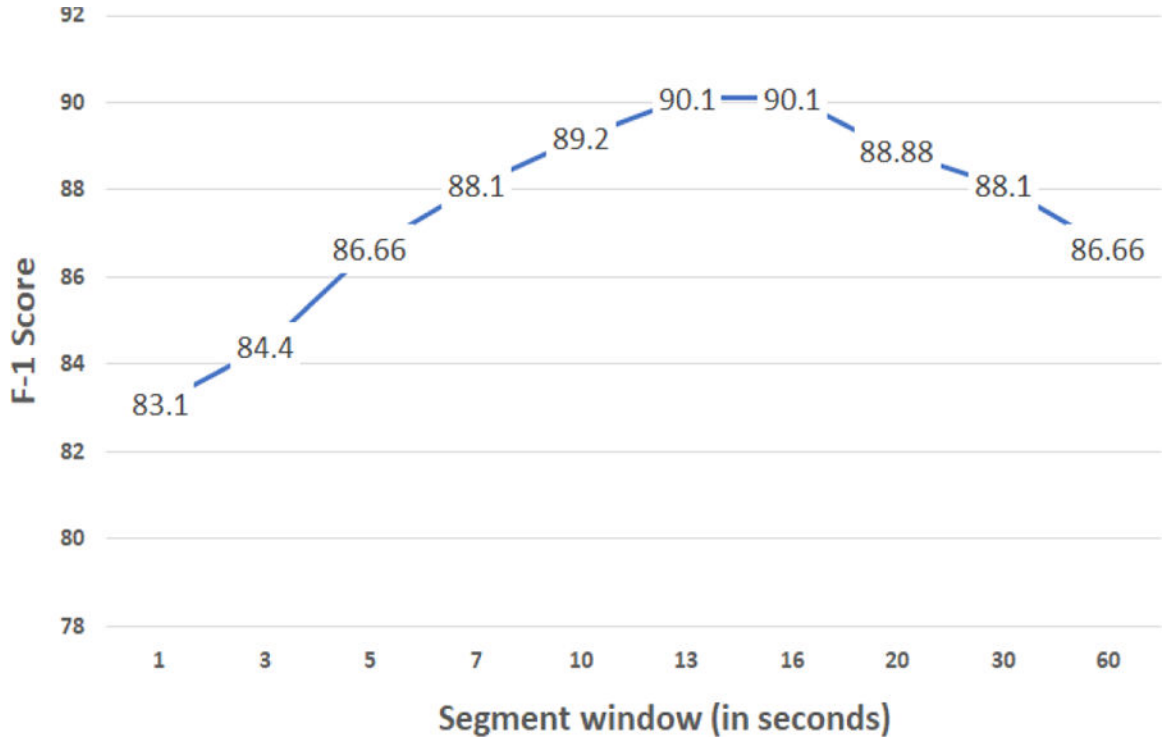

Fig. 5.
Evaluation for social anxiety with variable segment size

Table 1.

Low-level descriptive features and high-level functionals; *std*: standard deviation; *var*: variance; *dim*: dimension

Features	Functionals
Zero crossing rate & Δ (2-dim)	Min, Max, std, var, mean, median, skew, and kurtosis
Energy & Δ (2-dim)	
Spectral centroid & Δ (2-dim)	
Pitch & Δ (2-dim)	
MFCC & Δ (26-dim)	

Table 2.

Evaluation for social anxiety with variable audio codebook size

Audio codebook size	F-1 Score	Accuracy
500	77.2	78.9
1000	82.8	84.1
2000	87.9	89.1
2500	88.17	89.1
3000	89.13	90.1
3500	90.1	91
4000	90.1	91
4500	89.1	90
5000	88.17	89.1

Table 3.

Evaluation for social anxiety with NN2Vec and various feature representations

Feature	F-1 Score	Accuracy
NN2Vec	90.1	90
Emo2vec	72.3	77.22
I-vector	74.7	79
Audio word	55.55	68
Raw features	56.82	62.4

Table 4.

Evaluation for social anxiety with MIL algorithms

Algorithm	F-1 Score	Accuracy
BLSTM-MIL	90.1	90
DNN-MIL	85	88.11
mi-SVM	83.2	85

Table 5.

Evaluation for social anxiety with supervised learning algorithms

Algorithm	F-1 Score	Accuracy
BLSTM-MIL	90.1	90
BLSTM	86.66	88.1
CNN-BLSTM	83.5	85.1
CNN	83.5	85.1
DNN	68.3	73

Table 6.

Evaluation for state anxiety with various feature representations

Feature	F-1 Score	Accuracy
NN2Vec	93.49	90.2
I-vector	81.1	77
Emo2vec	81.6	77.8
Audio word	63.2	58.1
Raw features	62.3	57.5

Table 7.

Evaluation for state anxiety with baseline algorithms

Algorithm	F-1 Score	Accuracy
BLSTM-MIL	93.49	90.2
DNN-MIL	87.91	84.8
mi-SVM	85	83.1
BLSTM	90.24	86.31
CNN-BLSTM	88.86	86.1
CNN	88.86	86.1
DNN	74.2	68

Table 8.

Evaluation for depression with various feature representations

Feature	F-1 Score	Accuracy
NN2Vec	85.44	96.7
I-vector	70.1	86.54
Emo2vec	74.3	88.1
Audio word	51.2	79.1
Raw features	52.76	79.66

Table 9.

Evaluation for depression with baseline algorithms

Algorithm	F-1 Score	Accuracy
BLSTM-MIL	85.44	96.7
DNN-MIL	76.4	90.64
mi-SVM	66	84.1
BLSTM	77.1	91.4
CNN-BLSTM	71	87.6
CNN	68.8	85.1
DNN	56	80.86