Distant Emotion Recognition

ASIF SALEKIN, University of Virginia, USA
ZEYA CHEN, University of Virginia, USA
MOHSIN Y AHMED, University of Virginia, USA
JOHN LACH, University of Virginia, USA
DONNA METZ, University of Southern California, USA
KAYLA DE LA HAYE, University of Southern California, USA
BROOKE BELL, University of Southern California, USA
JOHN A. STANKOVIC, University of Virginia, USA

Distant emotion recognition (DER) extends the application of speech emotion recognition to the very challenging situation that is determined by variable speaker to microphone distances. The performance of conventional emotion recognition systems degrades dramatically as soon as the microphone is moved away from the mouth of the speaker. This is due to a broad variety of effects such as background noise, feature distortion with distance, overlapping speech from other speakers, and reverberation. This paper presents a novel solution for DER, addressing the key challenges by identification and deletion of features from consideration which are significantly distorted by distance, creating a novel, called Emo2vec, feature modeling and overlapping speech filtering technique, and the use of an LSTM classifier to capture the temporal dynamics of speech states found in emotions. A comprehensive evaluation is conducted on two acted datasets (with artificially generated distance effect) as well as on a new emotional dataset of spontaneous family discussions with audio recorded from multiple microphones placed in different distances. Our solution achieves an average 91.6%, 90.1% and 89.5% accuracy for emotion happy, angry and sad, respectively, across various distances which is more than a 16% increase on average in accuracy compared to the best baseline method.

CCS Concepts: • **Computer systems organization → Embedded systems**; *Redundancy*; Robotics; • **Networks** → Network reliability;

Additional Key Words and Phrases: Distant emotion detection, word2vec

ACM Reference Format:
Asif Salekin, Zeya Chen, Mohsin Y Ahmed, John Lach, Donna Metz, Kayla de la Haye, Brooke Bell, and John A. Stankovic. 2017. **Distant Emotion Recognition**. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 3, Article 96 (September 2017), 25 pages.
DOI: http://doi.org/10.1145/3130961

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-1319302, and NSF grant IIS-1521722.

2474-9567/2017/9-ART96 $15.00
DOI: http://doi.org/10.1145/3130961

1 INTRODUCTION

Emotion is a fundamental component of health. It is used when investigating people with depression, dementia, cancer, diabetes, obesity, alcoholism, and a myriad of other medical conditions. Many efforts have been undertaken to automatically detect emotion via speech. In almost all solutions, it is assumed that the individual is next to a microphone in an environment with limited ambient noise. However, this severely restricts the monitoring time. As smart space technologies are developed there is a potential to monitor the individual at all times that they are within rooms that have microphones. This increased knowledge of an individuals' emotional state will significantly improve healthcare for these individuals. For this to work, the speech processing solutions must be able to handle the differences that occur in speech due to various distances to a microphone, with ambient noise, with overlapped conversations, with different reverberations of sound caused by room construction and furnishings, and with other realisms found in the wild.

In a realistic indoor speech emotion recognition system, the acoustic sensors, i.e., microphones, capture speech signals originating from sources (humans) situated at various distances. Increasing source-to-microphone distance reduces signal-to-noise ratio and induces noise and reverberation effects in the captured speech signal, thus degrading the quality of captured speech, and hence the performance of the emotion recognizer. While speech to text and speaker ID have addressed the effect of distance with mixed results, distant-emotion-recognition (DER) is an area not previously explored to the best of our knowledge.

This paper presents a novel, robust approach to detect speech emotion (happy, angry, and sad) by addressing the main challenges of DER. The contributions of this paper are:

- The majority of the speech features typically used in classifiers significantly distort with increase of speaker to microphone distances. Hence, using these features complicate and deteriorate the ability to detect the emotional state across variable distances. In our solution we have identified 48 low-level descriptor features which do not significantly distort across variable speaker to microphones distances. We use these features as core elements of the overall solution.
- As shown in figure 1, we segment an audio clip into overlapping small frames and extract these 48 robust low-level descriptor features (LLD) from them. Each small segment of speech represents a state and an emotion is represented by the progression of speech through various states. We develop a novel *Emo2vec* feature modeling approach (section 5) that assigns a similar vector to the small frames (speech states), which appear in a similar context for a specific emotion. The vectors from the small frames represent the states of speech from the small segments. In addition, we exploit temporal dynamics of these states which provides rich information for speech emotion. The temporal information in our emotion detection approach is used through a long short term memory classifier (LSTM), where the sequence of vectors (from small frames using Emo2vec feature modeling) are used as input.
- Most of the existing approaches to automatic human emotional state analysis are aimed at recognition of emotions on acted speech. A common characteristic of all the existing acted emotional speech datasets is that all of them are made of clean speech recorded by closely situated microphones, often in a noise-proof anechoic sound studios. All the existing speech emotion recognition results are based on these clean speech recordings and, hence, these results are inaccurate in a real world environment where acoustic sensors are likely to be situated far from the speakers. To evaluate our approach on speech with various levels of reverberation and de-amplification (generally due to distance and the environment), we used two acted emotion datasets (section 7.1.1). We trained our model on clean training data and evaluated on disjoint modified test data (by introducing various reverberation and de-amplification effects). Through this evaluation we achieved 90.64%. 89.41% and 90.88% accuracy and 92.7%, 90.66% and 90.86% recall for emotions happy, angry and sad, respectively, which is 10.5%, 9.7% and 10.3% improvement in accuracy (and 10.2%, 12.2% and 15.85% improvement in recall) compared to the best baseline solution. (section 7.1.4)

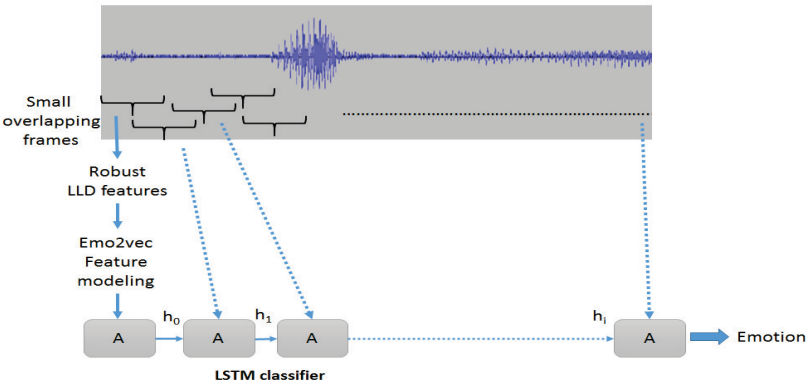

Fig. 1. Overview of our approach

- The fact that acted behavior differs in audio profile and timing from spontaneous behavior has led research to shift towards the analysis of spontaneous human behavior in natural settings [56]. There is no existing spontaneous human emotion speech dataset with audio recorded from multiple microphones placed at different distances. Hence, we have built a new emotional dataset of spontaneous family discussions (Collected from 12 families, a total 38 people) where the audio was collected from 3 different distances: 2.25, 3 and 5.3 meters (section 7.2). Evaluation of our solution on this dataset shows an average 91.42%, 89.1% and 88.04% accuracy and 91.9%, 89.26% and 86.35% recall for emotions happy, angry and sad, respectively, across various distances (Table 6 in section 7.2.2).
- One of the major challenges in the realistic conversations is overlapping of speech, which none of the previous works on speech emotion detection has addressed. This study introduces a novel overlapping speech filtering approach for the DER problem without needing expensive microphone arrays (section 7.2.2) which increases accuracy for happy and angry emotions up to 92.71% and 90.86% (92.43% and 91.21% recall), respectively, across various distances (table 7).
- We have implemented 4 baseline solutions from the literature and compared those solutions with our solution on both acted datasets (section 7.1.4) and our newly created spontaneous family discussion dataset (section 7.2.2). According to our evaluation our novel overall DER solution achieves approximately 16% increase in accuracy compared to the best baseline method.

2 RELATED WORK

Emotion recognition systems can usually be split into three parts, namely feature extraction, modeling, and classification of the emotion. First, the features used in emotion recognition can be divided into two groups according to their time span: low-level descriptors (LLD) are extracted for each time frame, such as Mel-frequency cepstral coefficients, loudness, zero crossing rate, jitter or shimmer. On the other hand, global descriptors are computed using the LLD extracted for the whole audio signal or for an audio segment covering several audio frames, such as the mean, standard deviation, quartile, flatness or skewness, among others.

The modeling stage of an emotion recognition system must obtain a representation of the speech that reflects the emotional information. Depending on the features used, different modeling approaches can be found in the literature. When dealing with LLD, 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, achieved improvement on accuracy using the i-vector representation of speech [13, 22]. The i-vector extraction, which was originally developed for speaker recognition, consists of two separate stages: UBM state alignment and i-vector computation. The role of UBM state alignment is to identify and cluster the similar acoustic content, e.g., frames belonging to a phoneme. The purpose of such alignment is to allow the following i-vector computation to be less affected by the phonetic variations between features. However, the existence of noise and channel variation could substantially affect the alignment quality and, therefore, the purity of extracted i-vectors. The i-vector technique estimates the difference between the real data and the average data and with variance of noise, speaker to microphone distance and reverberation, this difference becomes inconsistent.

Various types of classifiers have been used for speech emotion detection, including hidden Markov models [24], Gaussian mixture models [58], support vector machines (SVM) [4], k-nearest neighbor [41] and many others [37].

Recently, deep learning has revolutionized the field of speech recognition. A deep learning approach called, 'Deep Speech' [17] has significantly outperformed the state-of-the-art commercial speech recognition systems, such as Google Speech API and Apple Dictation. With the Deep Learning breakthrough in speech recognition a number of studies have emerged where Deep Learning is used for speech emotion recognition. Linlin Chao [8] et al. use Autoencoders, which is the simplest form of DNN. Another form of Deep Neural Networks is the Deep Belief Networks, which use stacked Restricted Boltzmann Machines (RBMs) to form a deep architecture. [8] tested DBNs for feature learning and classification and also in combination with other classifiers (while using DBNs for learning features only) like k-Nearest Neighbour (kNN), Support Vector Machine (SVM) and others, which are widely used for classification.

With success in speaker ID and speech transcription using i-vector and deep learning, a study [57] used a combination of prosodic acoustic features and the i-vector features and used Recurrent Neural Network to detect speech emotion. We use this solution as one of the baselines. A recent study [51] proposed a solution to the problem of *context-aware* emotional relevant feature extraction, by combining Convolutional Neural Networks (CNNs) with LSTM networks, in order to automatically learn the best representation of the speech signal directly from the raw time representation. CNNs are mostly used in image recognition. This is because, when dealing with high-dimensional inputs such as images, it is impractical to connect neurons to all neurons in the previous volume because such network architecture does not take the spatial structure of the data into account. Convolutional networks exploit 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. Hence, CNNs are mostly applicable to high-dimensional data. Furthermore, these papers do not consider speaker distance, reverberation or noisy speech. We used this solution as one of our baselines.

Though there is no existing work which addresses the DER problem, however there are a few works on emotion detection from noisy speech [38, 50, 55]. [38] claims that a work is in progress for emotion detection from noisy speech without any details of their approach. [50] used *CFSSubsetEval* with *best first* search strategy feature selection technique to identify features with high correlation with the class, but low correlation among themselves. We use the solution of [50] as one of our baselines.

OpenSMILE is the most extensively used open source feature extraction toolkit. We extract 6552 features, as 39 functionals of 56 acoustic low-level descriptors (LLD) related to energy, pitch, spectral, cepstral, mel-frequency and voice quality, and corresponding first and second order delta regression coefficients according to the most recent INTERSPEECH Computational Paralinguistic Challenge baseline set [47] with the openSMILE toolkit [12]. We train SVM classifiers taking these features as input for emotion recognition (according to [47]). These SVM classifiers with the INTERSPEECH 13 feature set is one of our baselines.

3 PROBLEM FORMULATION

It is difficult to define emotion objectively, as it is an individual mental state that arises spontaneously rather than through conscious effort. Therefore, there is no common objective definition and agreement on the term emotion. This is a fundamental hurdle to overcome in this research area [46]. Additionally, diversity in the way different people speak and express emotions, accents, and age [43] make the task more difficult. However, when speaker to microphone distance increases (as opposed to when the microphone is right next to the speaker), it adds further complexity to the emotion detection problem due to room reverberation, noise, and de-amplification of speech. A realistic emotion detection system which is deployed in open environments such as homes, captures sound waves from distant sources (human subjects). We formally call this a Distant Emotion Recognition (DER) problem. This paper addresses the DER problem by developing new solutions for LLD feature selection and feature modeling, as well as using a LSTM classification technique to exploit the temporal information across low-level speech states represented by minimally distorted features. The following sections describe the challenges of DER and our solutions for each of these stages.

4 CHALLENGES AND SOLUTION: FEATURE DISTORTION

With the increase of speaker to microphone distance, recorded signals start to distort compared to the original signal due to de-amplification of the signal, reverberation and ambient noise of the room. The amount of distortion depends on the speaker to microphone distance, the acoustic properties of the room and the amount of noise. To address this challenge our solution is to identify features not significantly affected by distance. Since these features are robust across distance, their distortion with distance is minimal, hence the reduction of accuracy due to distance is reduced. To measure the distortion of a feature across distance d we use the equation 1 where f_0 and f_d are the feature values for a clean signal and signal from distance d, respectively.

$$distortion_d = |\frac{f_0 - f_d}{f_0}| \times 100\% \tag{1}$$

The following subsections discuss the data collection strategy for our experiment, the extracted features and the identification of robust features from low-level descriptor frames.

4.1 Controlled Lab Experiment

We recruited 12 people to read scripts in a controlled lab experiment. We used a VocoPro UHF-8800 Wireless Microphone System and a transmitter, M-Audio Fast Track Ultra 8R USB 2.0, to record and transmit sound. The microphone setting is shown in Figure 2. It was a rectangular room and 7 microphones were placed facing the speaker. One was 0.5 meters away, three were about 1.5 meters away, two were about 3 meters away and the last one was about 6 meters away. Multiple microphones were placed at the same distances to record sound from different angles. Each microphone recorded speaker's voice separately, but simultaneously, into 44.1kHz, 32-bit wav format files. Each speaker read 64 scripts (duration ranging from 5 to 20 seconds). All the microphones could record the speech. The purpose of this dataset is to identify robust features across distance not emotion detection.

4.2 Select Robust Features

Based on the previous studies on acoustic features associated with emotion (section 2) we considered 77 low-level descriptor (LLD) features shown in Table 1, as well as their delta and delta-delta coefficients (total 231 features). Collected emotional speech audio clips from our controlled experiment are segmented into 25ms small frames (with 10ms overlapping). 231 LLD features typically used in emotion detection are extracted for each of these 25ms small frames. We calculated distortion considering the 0.5 meter distance microphone audio as clean speech

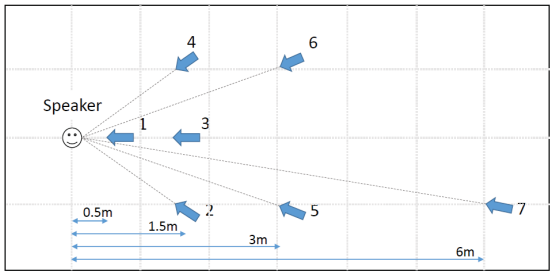

Fig. 2. Microphone setting in Lab

Table 1. Considered acoustic features associated with emotion

Feature	Count
Mel-Frequency cepstral coefficients (MFCC) 1-25	25
Root-mean-square signal frame energy	1
The voicing probability computed from the ACF	1
The fundamental frequency computed from the Cepstrum	1
Pitch	1
Harmonics to noise ratio (HNR)	1
Zero-crossing rate of time signal	1
PLP cepstral coefficients compute from 26 Mel-frequency bands	6
The 8 line spectral pair frequencies computed from 8 LPC coefficients	8
Logarithmic power of Mel-frequency bands 0 - 7	32

in equation 1. Figure 3 shows the number of LLD features for various average distortion ranges. According to our evaluation all delta and delta-delta features distort more than 100%. Also, a majority of the rest of the features distort between 40% to 50% when speaker to microphone distance is 6 meters. Hence, through our evaluation we considered 48 LLD features, with less than 50% distortion through various distances to use as attributes in our DER approach. These features are 5 Mel-Frequency cepstral coefficients, voice probability, fundamental frequency, zero crossing rate, 8 line spectral pair frequencies and 32 Logarithmic power of Mel-frequency bands.

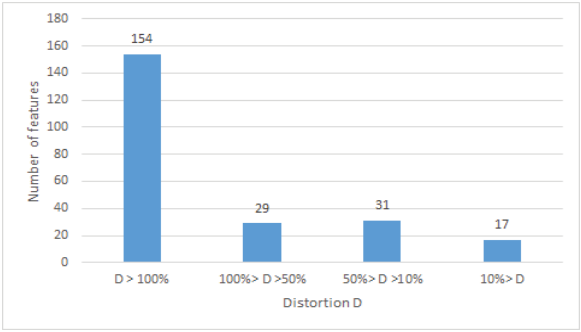

Fig. 3. Number of LLD features on various average distortion range

5 FEATURE MODELING

The modeling stage of an speech analysis system develops a representation of the speech that reflects the speech information for that specific task. Each small segment of speech represents a state, and emotion is represented by the progression of speech through various states. In the DER problem we consider a speech signal from small frames (25ms) to represent a state of speech. The following sections introduce a representation of the speech state from a small frame that handles the small distortion of LLD features due to reverberation or speaker to microphone distance variance, as well as takes into account the relationship of a particular state with its neighbor states for each particular emotion.

5.1 Audio to word

We use the Audio-Codebook model [32, 39] to represent the audio signal from small frames with 'words'. These 'words' represent the state of speech in the small frames and are not words in the normal meaning attributed to words. In our context the Audio-Codebook words are fragments of speech represented by features. We use the k-means clustering method to generate the audio codebook from the LLD feature representations mentioned in section 4.2. K-means is an unsupervised clustering algorithm that tries to minimize the variance between the k clusters and the training data. In the codebook generation step, we first randomly sample points from the audio in the training set and then run k-means clustering. The centroids of the resulting clusters form our codebook words. Once the codebook has been generated, acoustic LLD features within a certain small range of the speech signal are assigned to the closest (Euclidean distance) word in the codebook. As it might be the case that one input frame has a low distance to multiple audio words and, hence, the assignment is ambiguous, we take multiple assignments into account. As shown in the figure 4, the N_c nearest words from codebook are assigned to a small frame.

The LLD features selected from section 4.2 distort up to a certain threshold with variance of speaker to microphone distance and reverberation. Our trained audio codebook places similar points in the feature space into the same words, which reduces the effect of feature distortion to a certain level.

The discriminating power of an audio codebook model is governed by the codebook size. The codebook size is determined by the number of clusters K generated by the k-means clustering. In general, larger codebooks are thought to be more discriminative, whereas smaller codebooks should generalize better, especially when LLD features extracted from small frames can distort with distance, noise and reverberation, as smaller codebooks are more robust against incorrect assignments. We have evaluated with different size of codebooks as described in our evaluation section.

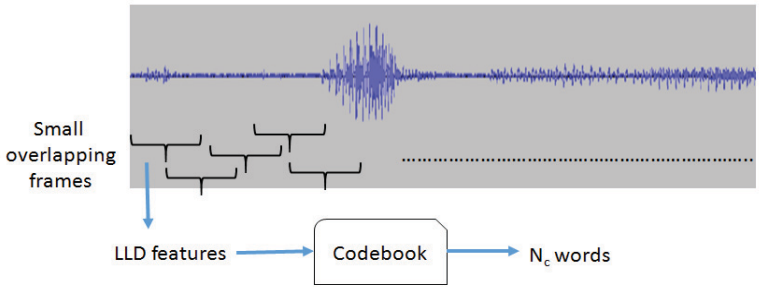

Fig. 4. Extraction of codebook words

5.2 Basic Word2vec Model

In this section, we present a brief description of the skip-gram model [15, 42] and in the following section discuss the novel enhancement of this model for our DER solution. The objective of the skip-gram model is to infer word embeddings (vectors) that are relevant for predicting the surrounding words in a sentence or a document, which means if two different words have very similar 'contexts' (i.e., words which appear around them frequently in training), their vectors are similar.

More formally, given a sequence of training words $w_1, w_2, \ldots, w_T$, the objective of the skip-gram model is to maximize the average log probability, shown in equation 2

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c, j\neq 0} logp(w_{t+j}|w_t) \tag{2}$$

where c is the size of the training context (which can be a function of the center word w_t). The basic Skip-gram formulation defines $p(w_{t+j}|w_t)$ using the softmax function:

$$p(w_O|w_I) = \frac{exp(v_{wO}^{'T} v_{wI})}{\sum_{w=1}^{W} exp(v_w^{'T} v_{wI})} \tag{3}$$

where v_w and $v_w^{'}$ are the 'input' and 'output' vector representations of w, and W is the number of words in the vocabulary.

For example, if we take two N_v dimensional vectors for two words, that are randomly initialized with some values (shown in figure 5a), and if we add a tiny bit of one vector to the other, the vectors get closer to each other, simply by virtue of vector addition. Figure 5b shows this for 2 dimensional vectors OA and OB. If we subtract a tiny bit of one vector from the other the vectors move apart by a tiny bit (shown in figure 5c). During word2vec training, in each step, every word (vector) is either pulled closer to words (vectors) that it co-occurs with, within a specified window or pushed away from all the other words (vectors) that it does not appear with. Word2vec training only brings together words (vectors) that are within the specified window context.

5.3 Novel Emo2vec model

This section introduces a new Speech Emo2vec solution, an acoustic emotion specific word2vec model, which performs more effectively than word2vec (shown in section 7.1.4) for emotion detection. The objective is to generate vectors from the LLD features extracted from small frames. These vectors are the inputs to the classifier. These new vectors for each frame are generated in a manner which indicates that if two frames appear in a similar context (i.e., similar surrounding frames) and for a specific emotion that the vectors will be similar. This

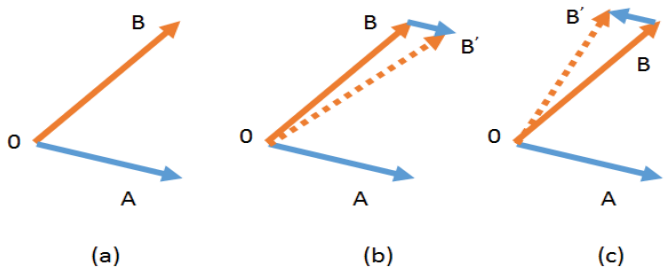

Fig. 5. Explain word2vec

representation of states makes differentiation between the states indicative to a specific emotion from the other emotions easier.

According to this model for every N_c words (from codebook) extracted from a small frame f_i, in the vector identification (word2vec) training, the number of neighbor words is $2 \times (N_c \times c)$, extracted from the left and right c windows of the frame f_i. As shown in figure 6, if $N_c = 3$ and $c = 4$, for each word $w_{i,1}$, $w_{i,2}$ or $w_{i,3}$, the neighbor word set in the word2vec training consists of all the words extracted from $i - b$ and $i + b$ small frames, where $b = 1, 2, 3, 4$.

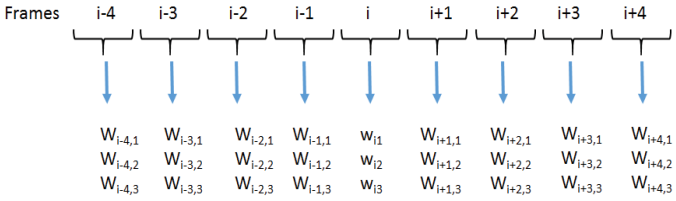

Fig. 6. Emo2vec: vector calculation using neighbour small frames

To achieve similar vectors for the small frames f_i, which occurs in similar context for a specific emotion E, we train a Emo2vec model for each of the emotions with input corpus D_e. D_e is a collection of $(w, Neighbour_w)$ pairs, that occurs in the training speech signal samples for that specific emotion E. Here, w is a word from a small frame f_i, and $Neighbour_w$ is a set consisting of $2 \times (N_c \times c)$ words extracted from the left and right c windows of the frame f_i. Training Emo2vec, generates similar vectors for the $words \in w$ from corpus D_e, if the $words$ have a similar neighbour set $Neighbour_w$ multiple times.

To illustrate further, figure 7 shows an example input training corpus for detection of emotion: happy. Here, $N_c = 2$ and $c = 2$, hence each word has 8 neighbours. In this figure the word-neighbor pairs in the left are from happy speech samples and the right are the ones from other speech samples. According to figure 7 words A, B and, C have similar neighbours. Word2vec training [26, 53] considers samples from the whole corpus (both D_H and D_N). Hence, generated vectors for words: A, B, and, C would be similar. Since, these words occurs with similar neighbour words and generated Word2vec vectors for them are also similar, differentiation task for happy emotion detection classifier is difficult. But, Emo2vec training only considers happy speech samples (corpus D_H), hence generated Emo2vec vectors for only words A and B are similar. Since, according to Emo2vec vector representation, A and B words are closer in feature space (similar vectors) and further from C, the differentiation task for the classifier is easier. Emo2vec put words which occur in similar context (similar neighbour set) for a

specific emotion (in this example happy), closer in feature space. Words which occur with similar neighbours, but for different emotions are pushed further apart. Hence, the classification task for that specific emotion gets easier.

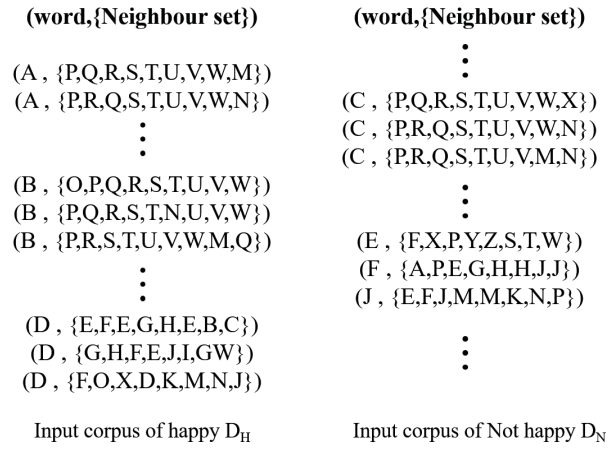

Fig. 7. Example training corpus for emotion: happy

Additionally, we perform the following steps before training our emotion specific Emo2vec model for each of the emotions:

5.3.1 ***Sub-sampling Frequent Words:*** Sub-sampling is a method of diluting very frequent words. Very frequent words are equivalent to stop words in text documents (is, are, the, etc.). Analogously, there are some codebook *words* (section 5.1) which appear in nearly all speech samples of happy, angry and sad emotions. Hence they are not discriminative to any emotion. Our solution deletes these frequent *words* from consideration in our emotion classification.

Before training Emo2Vec models for each of the emotions E_i using training corpus D_i, where i =happy, angry or sad, we sub-sample the frequent *words* which appear more than a threshold t_f times across all training corpuses D_i. The sub-sampling method randomly removes words that are more frequent than some threshold t_f with a probability of p, where f marks the word's corpus frequency: $p = 1 - \sqrt{\frac{t_f}{f}}$

A modified version of the word2vec [31] performs sub sampling frequent words from the corpus it trains on. Applying that approach would eliminate frequent words appearing for our targeted emotion (for example: happy). But, that frequent word can be rare for other emotions, hence highly indicative to our targeted emotion. Elimination of such highly discriminative words would make classification difficult. Hence, we perform sub sampling frequent words removal across all training corpuses (for all emotions), before performing training of Emo2vec.

5.3.2 ***Deletion of rare words across all emotions:*** If a word w_i appears less than a threshold t_r times across all training corpuses (happy, angry and sad), we delete w_i before creating the context windows. The intuition is, rare words are too specific and too small in number, hence they are indicative to some specific audio clips, rather than a generic class of audio clips, in our case audio from specific emotions.

After sub-sampling frequent words and deletion of rare words, through our emotion specific Emo2vec approach we generate vectors for each of those words occurring between t_r to t_f times in the training corpuses (for all emotions). We name this emotion specific word to vector mappings as the Emo2vec dictionary.

5.4 Speech Emo2vec extraction

As shown in figure 8 we extract LLD features from each of the small frames $frame_k$. According to section 5.1, N_c (solution uses $N_c = 3$) words are extracted for that LLD feature set using the generated codebook. Using the Emo2vec dictionary generated in training phase we convert the words to their corresponding vectors of size s. Unseen words w_j, where Emo2vec dictionary does not contain word to vector mapping, are represented with zero vector of size s. Extracted word vectors are added to create final output vector $V = [v_1, v_2, \ldots, v_s]$ of size s. This output vector V represents the state of speech signal from small frame $frame_k$.

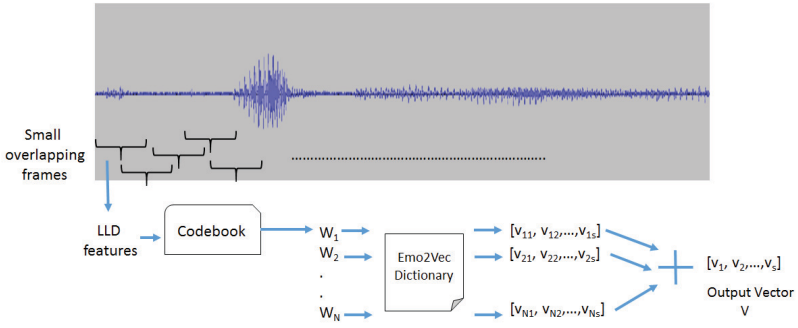

Fig. 8. Emo2vec vector extraction from speech

6 CLASSIFICATION: RECURRENT NEURAL NETWORK AND LSTM

Recurrent Neural Networks [30] are a type of Neural Network 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 knowledge from the past that are relevant to predict outcomes. RNNs are networks with loops in them, allowing information to persist. In the figure 9, a chunk of a neural network, A, looks at some input x_t and outputs a value h_t. A loop allows information to be passed from one step of the network to the next. If we unroll the loop, a RNN can be thought of as multiple copies of the same network, each passing a message to a successor. This chain-like nature reveals that recurrent neural networks are intimately related to sequences and lists. They are the natural architecture to use for such data.

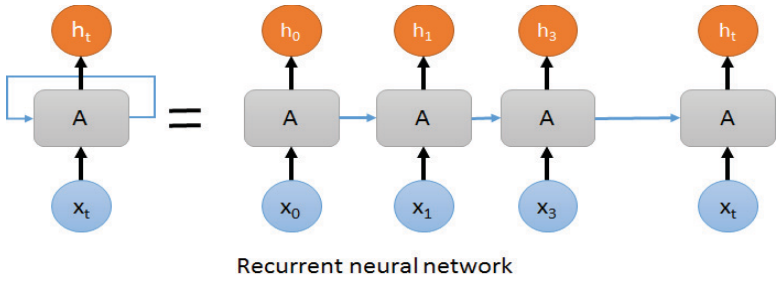

Fig. 9. Recurrent Neural Network

Formally, the output of an RNN is computed as:

$$\hat{y_t} = \sigma(W_o h_t) \tag{4}$$

Here, W_o is a parameter matrix of the model, h_t is the hidden state vector of the Neural Network, σ is some differentiable function that is applied element-wise and $\hat{y_t}$ is a vector containing the predicted output. If the input is a sequence, $x = x_1, x_2, \ldots, x_t$, the hidden state of a step t is computed by:

$$h_t = f(h_{t-1}, x_t) \tag{5}$$

Since, the memory of RNNs is essentially unlimited, RNNs can in principle learn to remember any length of states. RNNs capture long-term temporal dynamics using time-delayed self-connections and are trained sequentially. In our approach, an audio signal is segmented into overlapping small frames, and an Emo2Vec V vector is extracted from each of these small frames. Emo2vec V represents the state of speech from a small frame (25ms in our solution). Intuitively, it is not possible to perceive speech emotion from a 25ms small frame, but emotions can be recognized by the the temporal dynamics across these states (represented by Emo2vec vector V). Consequently, we use the temporal information in our emotion detection through a recurrent neural network (RNN). The sequence of vectors (Emo2Vec vectors), each of which represents the state of speech from a small overlapping frame, are the input sequence x_i, $i = 0, 1, .., t$ to a RNN.

RNNs can detect the final output (in this case an emotion) using Emo2Vec V vectors as input from any length of small frame sequences. This means that our emotion detection using Emo2vec features is not limited to any fixed window size; it can detect emotion capturing the progression of speech states (represented by Emo2Vec) from any variable size windows.

6.1 Long Short-Term Memory Units

A study [33] showed that it is not possible for standard RNNs to capture long-term dependencies from very far in the past due to the vanishing gradient problem. The vanishing gradient problem means that, as we propagate the error through the network to earlier time steps, the gradient of such error with respect to the weights of the network will exponentially decay with the depth of the network. In order to alleviate the gradient vanishing problem, [19] developed a gating mechanism that dictates when and how the hidden state of an RNN has to be updated, which is named as long short-term memory units (LSTM).

The LSTM contains special units called memory blocks in the recurrent hidden layer. The memory blocks contain memory cells with self-connections storing the temporal state of the network in addition to special multiplicative units called gates to control the flow of information. Each memory block in the original architecture contains an input gate and an output gate. The input gate controls the flow of input activations into the memory cell. The output gate controls the output flow of cell activations into the rest of the network.

Emotion is represented by the progression of speech through various states. Our solution represents a state of speech by a Emo2vec vector. Comprehension of temporal dynamics of states throughout the entire speech segment (that we want to classify) requires long-term dependency. Hence, to avoid vanishing gradient problem, our solution uses the LSTM defined by [16] as our Recurrent Neural Network classifier. The implementation (equations 6,7,8,9, and 10) adds the additional forget gate, which addressed a weakness of LSTM models preventing them from processing continuous input streams that are not segmented into sub-sequences.

$$i_t = \sigma(W_{xi} x_t + W_{hi} h_{t-1} + W_{ci} c_{t-1} + b_i) \tag{6}$$

$$r_t = \sigma(W_{xr} x_t + W_{hr} h_{t-1} + W_{cr} c_{t-1} + b_r) \tag{7}$$

$$c_t = r_t \odot c_{t-1} + i_t \odot tanh(W_{xc} x_t + W_{hc} h_{t-1} + b_c) \tag{8}$$

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_{t-1} + b_o) \tag{9}$$

$$h_t = o_t \odot tanh(c_t) \tag{10}$$

Here, $\odot$ is the element-wise product and i, r and o are the input, forget and output gates, respectively. As it can be seen, the gating mechanism regulates how the current input and the previous hidden state must be combined to generate the new hidden state.

6.2 Dropout Regularization

During the training phase of a neural network, neuron weights settle into their context within the network. Weights of neurons are tuned for specific features providing some specialization. This makes neighboring neurons dependant on this specialization, and with further training it can result in a fragile model too specialized to the training data. This dependency on context for a neuron during training is referred to as complex co-adaptations. If neurons are randomly dropped out of the network during training, neighboring neurons would have to step in and handle the representation required to make predictions or classification for the missing neurons. This is believed to result in multiple independent internal representations being learned by the network. The effect is that the network becomes less sensitive to the specific weights of neurons. This, in turn, results in a network that is capable of better generalization and is less likely to overfit the training data. This regularization is named as dropout regularization. [49].

Our solution uses a two layer LSTM model. The first layer is the LSTM layer with N_{neuron} neurons. 20% dropout rate is set for this layer, which means two in 10 neurons were randomly excluded from each update cycle. Since this is a classification problem, we use a dense output layer with a single neuron and a sigmoid activation function to make 0 or 1 predictions for the two classes (Emotion E or not emotion E). Log loss is used as the loss function and the efficient gradient descent optimization algorithm is used.

7 EVALUATION

Our evaluation consists of two parts. First, in section 7.1, two existing acted emotion datasets from the literature are used and distance issues are emulated by incorporating artificial reverberation and de-amplification. We use these generated segments to evaluate our approach. This section addresses different questions as well as compares our solution with the state-of-the-art solutions. Second, since, there is no existing spontaneous human emotion speech dataset with audio recorded from multiple microphones placed at different distances, in section 7.2 we describe the creation of a new dataset. We recruited 12 families and let them exhibit different emotions in spontaneous unscripted discussions. Multiple microphones were used to record the discussion sessions. Several challenges and our solutions to address them, and comparison with state of the art baselines on this family discussion data are discussed in section 7.2.

7.1 Experiment on Acted Data

7.1.1 ***Emotional Datasets.*** We use two emotion speech datasets: EMA and SAVEE (with annotations for 3 emotions: happy, angry, sad and others) where the spoken language was English and speakers were close to the microphone. The Electromagnetic Articulography (EMA) dataset [25] includes articulatory motions recorded by an EMA system where talkers produced simulated (acted) emotional speech. There were 3 participants: a male and two females who are native speakers of American English. In total, the male participant produced 280 utterances (14 sentences with 5 repetitions for each of the 4 emotions), and each of the female participants produced 200 utterances (10 sentences with 5 repetitions for each of the 4 emotions). The SAVEE dataset [18] contains 15

sentences for each of the 4 emotion categories, from 4 male participants. In this dataset some sentences are natural while others are acted or elicited. In total there are 230 audio clips for each of the 4 categories: Happy, angry, sad, and others. These two datasets were merged for the evaluation.

7.1.2 ***Acoustic Pre-processing.*** This section describes the noise filtering, acoustic de-amplification, and reverberation of speech we used in the evaluation of these two datasets.

Filtering and Removing Noise: The first step of pre-processing is to remove unvoiced audio segments using zero crossing rate (ZCR) [6]. To capture the pause in spoken sentences, the detected voiced segments are lengthened by 1 second on both sides. If another consecutive voiced segment starts within the lengthened 1 second segment portion, both the voice segments are merged into one.

Noise which is out of human voice frequency range were removed using a bandpass filter with a low frequency of 80Hz and a high frequency of 3000Hz. Hiss, hum or other steady noises were reduced using a spectral noise gating algorithm [3].

Reverberation and de-amplification. Reverberation refers to the way sound waves reflect off various surfaces before reaching the listener's ear. In recent years a few dereverberation methods has been developed, though blind one-microphone dereverberation approaches are not so accurate yet [54]. However, there are different artificial reverberation effect parameters to model how sound waves reflect from various types of room size and characteristics. In this study, we use different combinations of reverberation parameters: wet / dry ratio, diffusion, and decay factor. Wet and dry ratio is the ratio of the reverberated signal to the original signal. The more reverberation the room has, the larger this ratio is. Diffusion is the density rate of the reverberation tail. A higher diffusion rate means the reflection is closer and the sound is thick. A lower diffusion rate has more discrete echoes. Decay factor is used to measure the time duration that reflection runs out of energy. A larger room has longer reverberation tails and lower decay factors. A smaller room has shorter tails and higher decay factors.

De-amplification decreases the loudness or volume levels of the audio clip and produces the effect of increase in speaker to microphone distance.

7.1.3 ***Training and Data Preparation.*** We perform filtering and noise removal (section 7.1.2) on the datasets (section 7.1.1) to remove noise and unvoiced audio segments. Our evaluation performs 5-fold cross validation using 90% of the dataset for training and other 10% for testing. In each of the 5 iterations, we train our model on clean, close-to-microphone training data (which is the original recordings in the dataset from section 7.1.1). We apply different combinations of de-amplification along with reverberation parameters: wet / dry ratio, diffusion, and decay factor on the audio clips of the test dataset (remaining 10% of dataset) to artificially introduce the effect of different speaker to microphone distances with different room characteristics. Finally, we evaluate our classifier on this artificially reverberated and de-amplified test data. This experiment measures the robustness of our approach in terms of reverberation and de-amplification, i.e, if a model is trained using clean (no reverberation), close of microphone audio data, how it performs on data with reverberation with different distant speakers (artificially generated).

Following previous audio-codebook generation approaches [32, 39] we randomly select 30% of the training data from all emotions to generate the audio-codebook. We have trained an individual binary classifier for each of the emotions: happy, angry and sad. To generate emotion specific Emo2vec dictionary our approach randomly selects 50% clips from the training set of that respective emotion.

7.1.4 ***Results.*** In this section we describe the efficiency and the applicability of our solution by investigating some of the pertinent questions related to the DER problem.

What are beneficial parameter configurations? There are a number of parameters in our emotion detection solution. They are sub-sampling parameter t_f (section 5.3.1) which defines the threshold of frequent words, t_r

Table 2. Accuracy of Emotion detection with various codebook and Emo2vec vector sizes

Codebook size	Happy Emo2Vec size			Angry Emo2Vec size			Sad Emo2Vec size		
	30	50	100	30	50	100	30	50	100
500	89.4	90.64	90.176	88.8	89.41	89.19	79.61	80.1	81.84
1000	85.9	86.29	89.55	85.2	86.02	88.35	79.1	80.1	83.94
1500	85.9	86.29	88.12	82.9	85.92	87.2	85.33	87.72	90.88
2000	83.2	85.17	87.29	82.9	84.6	86.9	86.72	89.5	85.33

Table 3. Comparison between Emo2vec and generic word2vec approach

Emotion	Angry		Happy		Sad	
	Accuracy	Recall	Accuracy	Recall	Accuracy	Recall
Emo2vec	89.41	90.66	90.64	92.7	90.88	90.86
Generic word2vec	81.19	81.8	81.09	82.17	82.8	82.88

from section 5.3.2 which defines the threshold of rare words and number of neighbour window parameter c (section 5.3). Through our evaluation we identify that the beneficial value of t_f is 880, t_r is 4 and c is 4. Also, our two layer LSTM classifier has N_{neuron} neurons in its' first (hidden) layer. Through our evaluation we identify the beneficial value of N_{neuron} is 100 neurons. Identification of the beneficial values were performed through iterating multiple values within a range. For example, we iterate the value of the sub-sampling parameter t_f over a range from 400 to 1000, and identify that using t_f=880 as sub-sampling parameter facilitates higher accuracy for all of our targeted emotion recognitions.

As shown in section 5.1, the codebook size influences the discriminative characteristics of our feature modeling approach. If the codebook size is too large, it will be more discriminative, but lose the reduction capability of feature distortion. Also, a small codebook will be too generic, hence, will not be able to capture the change of emotional speech states in various small frames. Suppose we select codebook size CB (e.g. 1000), if we represent the generated words by one-hot encoded vector of size of vocabulary CB, the number of features from each small frames for our LSTM classifier will be CB. Emo2vec condenses that dimensional vector. Smaller Emo2vec vectors are better for computation and performance on a dataset of limited size, but too small a vector size loses relational and discriminative information between the words.

As shown in table 2, we evaluated (according to section 7.1.3) our solution with various codebook and Emo2vec vector sizes. Table 2 shows our average 5-fold cross validation results where classifiers were trained on clean, close to microphone data and tested on artificially reverberated and de-amplified data. According to this table, accuracy for emotion happy and angry achieved up to 90.64% and 89.41% (92.7% and 90.66% recall) with codebook size 500 and Emo2vec vector size 50. For the emotion sad, the highest 90.88% accuracy and 90.86% recall is achieved with codebook size 1500 and Emo2vec vector size 100.

Is Emo2vec better than generic word2vec model? Emotion specific Emo2vec training generates similar vectors for the small frames f_i, which occur in similar contexts for a specific emotion E. To evaluate the importance of such emotion specific vector generation, we compare the performance of Emo2vec against the baseline word2vec model described in section 5.2 under the best parameter configuration for each emotion. Our evaluation is shown in table 3. According to this evaluation, Emo2vec approach achieves 10.13%, 11.77% and 9.75% higher accuracy and 10.83%, 12.8% and 9.6% higher recall for emotions angry, happy and sad, respectively, compared to the baseline word2vec model.

Table 4. Evaluation with or without distorted feature removal in Emo2vec approach

Emotion	Angry		Happy		Sad	
	Accuracy	Recall	Accuracy	Recall	Accuracy	Recall
With distorted feature removal	89.41	90.66	90.64	92.7	90.88	90.86
With all features	80.32	81.1	79.19	78.21	80.98	81.32

Elimination of distorted features helpful? This section compares two cases: with and without (distorted) feature selection -i.e., considering all 231 LLD features and without the distorted ones. In section 4.2 we identified 48 LLD features, with less than 50% distortion over various distances to use as attributes in our robust emotion detection approach. This evaluation showed that the majority of the other LLD features we considered distort more than 100%. Including those features in our codebook generation training phase would result in assignment of various states (small frame LLD feature vectors) to wrong codebook words during testing phase, due to feature distortion in realistic settings (with reverberation, variable speaker distance, and noise).

In evaluating this issue, Table 4 shows that if we consider all the 231 features from section 4.2 in our emotion detection approach under the best parameter configuration for each emotion, the accuracy achieved for angry, happy and sad is 80.32%, 79.19%, and 80.98%, respectively. The majority of these 231 features distort significantly with noise, de-amplification and reverberation. This means a state that represents a small frame LLD feature vector can deviate significantly in 231 dimensional feature space with variable speaker to microphone distances. Significant deviation of a state in feature space may result in wrong 'word' assignment from audio codebook, which may lead to wrong classification. According to this evaluation, the elimination of distorted features improves 11.3%, 14.4% and 12.2% accuracy and 11.78%, 18.5% and 11.73% recall for emotions angry, happy and sad, respectively, compared to using all 231 features. Elimination of distorted features reduces the deviation of a state in feature space, which reduces the chance of wrong 'word' assignment, hence increases accuracy.

Comparison with baselines. We implemented and compared our solution to four baseline works [47, 50, 51, 57] and on our acted dataset (from section 7.1.1). A generic correlation based feature selection approach [50] performs the feature selection method to identify features with high correlation with the specific emotion class and at the same time with low correlation among themselves. This technique is not robust under realistic settings (with reverberation, variable speaker distance and noise), since many of the highly correlated features distort extensively. Hence, as shown in table 5, the accuracy using this approach for DER emotion detection is very low. Another baseline [51] performs 'context-aware' emotional relevant feature extraction by combining Convolutional Neural Networks (CNNs) with LSTM networks. As shown in the table, this technique of feature generation using CNN over-fits on the training data very easily and cannot adapt with feature distortion in realistic settings, hence do not perform well. [57] uses a combination of prosodic acoustic and i-vector features and uses a Recurrent Neural Network to detect speech emotion. Using this *ivector + RNN* approach we can achieve up to 81.43%, 82.01%, and 82.32% accuracy (and 80.8%, 84.1%, and 78.43% recall) for angry, happy and sad emotion detection. According to our evaluation of section 4.2, the majority of the LLD features extracted in [47] distort significantly with increase of speaker to microphone distance. 39 functionals applied on those distorted features are also not robust. Since, the majority of the 6552 features distort with variable speakers distance, using this INTERSPEECH Computational Paralinguistic Challenge 13 baseline approach, we achieve low accuracy of 70.2%, 70.4% and 72.83% (68.32%, 70.88% and 70.1% recall) for angry, happy and sad emotion detection. According to table 5, our DER solution achieves 9.7%, 10.5%, and 10.3% higher accuracy and 12.2%, 10.2% and 15.85% higher recall compared to the best baseline solution for emotions: angry, happy and sad.

Table 5. Comparison with baseline

Approaches	Angry		Happy		Sad	
	Accuracy	Recall	Accuracy	Recall	Accuracy	Recall
Our solution	89.41	90.66	90.64	92.7	90.88	90.86
ivector+RNN	81.43	80.8	82.01	84.1	82.32	78.43
CNN+LSTM	74.1	76.68	73.9	73.78	76.7	73.55
Correlation based feature selection	66.1	59.8	68.9	64.6	72.99	67.54
INTERSPEECH 13	70.2	68.32	70.4	70.88	72.83	70.1

7.2 Evaluation: Spontaneous Family Discussions

There is no available existing mood dataset containing spontaneous speech from a realistic setting with audio recorded from multiple microphones placed at different distances. Hence, in order to evaluate our DER approach, we built our own dataset (IRB number UP-16-00227). Our in-lab protocol was developed by using examples from other in-lab family discussion protocols. Twelve families were invited to our lab and were instructed to discuss set topics related to family meals and family eating habits. Experimenters helped them initiate the discussion by providing some issues (with flexibility to discuss other related topics) that they might have wanted to change in their family and encouraged them to select a few of the topics to discuss with other members. Intuition was that, discussion about change in their existing condition would raise conflicting opinions, which in turn would encourage them to express various emotions. After initiating the discussion, experimenters left the room and let the family members have spontaneous discussions. For each of these families, we recorded 15 to 20 minutes of spontaneous discussion.

Figure 10 shows the lab setting, and indicates where members were seated at a round table and how far the 3 microphones were placed from the table's center: 2.25, 3 and 5.3 meters. The radius of the table was 1.5 meters, so speaker to microphone distances for Microphone 1, 2 and 3 varied from 0.7 to 3.75, 1.5 to 4.5, and 3.8 to 6.8 meters. Speakers were free to move around the table. Hence, our collected data contains speeches of moving and steady speakers from distances varied between 0.75 to 6.8 meters. Since, this paper aims to recognize distant speech emotion in indoor realistic setting, variable microphone to speaker distance in range of 0.75 to 6.8 were determined considering the size of average indoor rooms [2].

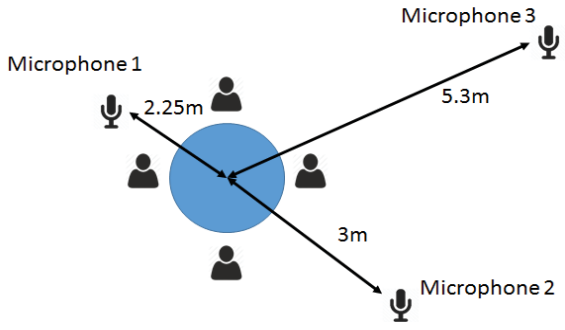

Fig. 10. Data collection lab setting

In total, we collected spontaneous speech from 38 people (9 male, 29 female), with age ranging from 10 to 56 years. All of the conversations were spoken in English.

The twelve videos were coded using the Noldus Observer XT event logging software [60]. The three emotions happy, anger, and sad were coded on a 3-point scale. The emotions were labeled as one of the following three degrees: borderline emotional tone, emotional tone, or very emotional tone (e.g., borderline happy, happy, or very happy). If a coder was unsure about one of the three emotions, then were instructed to classify it as a borderline emotional tone. All the emotions were tagged exclusively with 'start' and 'stop' times, meaning that two or more emotions were never coded at the same time for a single person. We did allow for emotions to overlap among participants (to account for participants speaking at the same time). Codes were not applied to segments where none of these three emotions were exhibited.

The two coders were research assistants with backgrounds in behavioral science. They trained on practice tapes until they achieved an inter-rater reliability rate (specifically, Cohen's kappa statistic) of 73% (0.73). This is considered moderately high [29]. The Observer XT software calculates the inter-rater reliability value automatically. After achieving a kappa statistic greater than 0.70, the two coders 'double-coded' a video every three to four discussions to ensure their inter-rater reliability rate remained above 70%. If the kappa statistic was lower than 0.70, they re-coded and re-trained until they once again achieved at least 0.70.

7.2.1 ***Training and Data Preparation.*** Our collected family discussions from different distance microphones contain humming sound from home appliances such as air conditioners, knocking on the table, slamming doors, squeaky hinges, dragging of chairs, etc. Hence, speech signal to noise ratio was low, specially for microphone 3 (5.3 meters distance). We have performed noise removal and filtering as shown in section 7.1.2 to reduce the steady noises. Also to adopt to home settings, we collected home environmental sounds from UrbanSound dataset [44] and use a subset of these data clips as negative samples during the training phase of our classifiers.

Our evaluation performed N fold cross validation where we trained an emotion specific classifier for each of the emotions E considering audio clips from microphone 1 (2.25 meters distance) of $N-1$ families for training and tested our trained model on audio clips from microphone 1,2, and 3 of N th family. In this evaluation we use softmax activation function in the output layer of the LSTM classifiers, which is basically the normalized exponential probability of class observations represented as neuron activations. Softmax classifiers give the probabilities for each class label. Hence, if two classifiers (Example: classifiers for happy and sad) provide positive output for an audio segment, we consider the classifier with higher positive output probability as the detected emotion of our DER solution.

7.2.2 ***Results.*** In this section we discuss the evaluation of our solution using the spontaneous family discussions dataset.

Change of beneficial parameters due to large variety of data? In our evaluation with the spontaneous family discussions dataset, we identify the beneficial values of t_r (threshold of rare words from section 5.3.2) and c (number of neighbour window parameter from section 5.3) are the same as section 7.1.4 (4, 4). The beneficial value of sub-sampling parameter t_f (from section 5.3.1) is 800. Also, our two layer LSTM classifier has 100 neurons in its first (hidden) layer.

Codebook size (section 5.1) influences the discriminative characteristics of our feature modeling approach. Also, our DER solution wants to identify an Emo2vec vectors size which is smaller (hence, better for computation and performance on a dataset of limited size), but not too small to loose relational and discriminative information between the words. Table 6 shows our evaluation with various codebook sizes and Emo2vec vector sizes on 3 different distances. According to this table beneficial codebook size for all addressed emotions and distances is 2000 and Emo2vec size is 100. Our family discussion dataset from different distance microphones contains speech

Table 6. Evaluation on family discussion data with various codebook and Emo2vec vector sizes

Microphone Distance		2.25					3					5.3				
		Codebook size														
Emotion	Vector size	500	1000	1500	2000	2500	500	1000	1500	2000	2500	500	1000	1500	2000	2500
Angry	50	87.3	85.19	85.11	87.25	87.42	82.3	81.41	81.41	83.5	82.82	80.1	81.47	81.47	83	82.35
	100	92.48	90.1	90.1	92.52	92.52	85.19	87.28	87.28	88.9	88.86	83.98	85.11	85.19	85.9	85.9
	200	90	87.28	87.28	90.1	90.1	80	84.32	84.32	84.4	84.4	78	79.11	79.11	81.44	81.44
Sad	50	81.4	83.32	85.16	86.35	86.35	79.1	81.44	81.44	83.55	82.34	77.2	77.88	80	82.92	82.34
	100	86.36	88.72	90	90.12	90.12	85.11	85.9	85.9	88.9	87.36	82.9	84	84.31	85.11	84.31
	200	80	84.32	87.28	88.72	88.72	80	80.1	80.1	82.34	80.1	75.32	79.1	79.1	81.44	81.44
Happy	50	83.31	85.83	85.21	88.7	88.7	83.3	83.32	83.32	88.71	88.71	80	77.2	77.2	82.35	81.77
	100	88.7	90.1	90.1	94.5	94.5	88.7	88.7	88.7	92.5	92.5	84.33	86.37	86.37	87.28	86.37
	200	80	88.72	88.72	92.5	92.5	77.3	84.17	84.17	90.05	90.05	77.93	81.79	81.79	82.88	82.88

from 31 individuals with variety of ages and accents. Hence, beneficial codebook and Emo2vec size is relatively larger compare to our result in section 7.1.4.

One of the significant observations from our spontaneous family discussions is, about 40% of the audio segments labeled as emotion 'happy' contains laughter. Acted emotional speech datasets (from section 7.1.1) do not contain laughter in 'happy' emotional speeches and majority of the 'false positives' for emotion: happy classifier in section 7.1.4 are 'angry' speech. Hence, presence of laughter makes classification easier for the happy emotion, which leads to higher accuracy of 94.5%, 92.5% and 87.28% (94.6%, 91.87% and 87.1% recall) for distance 2.25, 3 and 5.3 meters, respectively.

Since, our collected family discussion audio contains significant noise and reverberation, signal to noise ratio is relatively low. Low signal to noise ratio makes it harder to distinguish between sad and other categories of speech. Hence, sad emotion detection achieved upto 90.12%, 88.9% and 85.11% accuracy (88.21%, 87.54% and 83.32% recall) for distance 2.25, 3 and 5.3 meters, respectively, which is low compared to the other two emotions (happy and angry).

Challenge: overlapping speech. Approximately 15% of the voiced speech segments in our family discussion speech samples contain overlapping speech. Through our evaluation we identify that a majority of false positives from happy and angry emotion classifiers are overlapping speech. There are few studies that address the issue of identifying overlapping speech from audio, but none of them have achieved significant high accuracy. These studies [23, 48] have used spectral autocorrelation peak-valley ratio, harmonic to noise ratio, fundamental frequency, average spectral aperiodicity, MFCC, perceptual linear predictive coefficients (PLP), and spectral flatness measure as features. Using our distorted feature identification approach (section 4.2) we identify that MFCC, perceptual linear predictive coefficients (PLP), fundamental frequency, and spectral flatness measure distort less than 50% across various distances. Hence, we extract mel-frequency cepstral coefficients (MFCC) 1-6 in full accordance to htk-based computation, fundamental frequency computed from the cepstrum, 6 perceptual linear predictive coefficients (PLP) and spectral flatness measure from 25ms overlapping small frames. Next the 12 functionals: mean, standard deviation, kurtosis, skewness, minimum and maximum value, relative position, and range as well as two linear regression coefficients with their mean square error (MSE) are applied on a chunk of small frames (from 1 second audio segments).

Our solution trains a binary neural network classifier (with 1 hidden layer consisting of 100 neurons) with two output classes: overlapping speech and single speech, which uses the features discussed above from a 1 second audio segment as input. Softmax activation function is used in the output layer of the neural network classifiers, which gives the probabilities for each class label and the sum of probabilities of all the classes is 1. Our solution identifies a 1 second audio segment as overlapping speech and does not consider it for emotion detection only if

Table 7. Evaluation with overlapping speech filtering

Mood	Distance (meters)	2.25		3		5.3	
		Accuracy	Recall	Accuracy	Recall	Accuracy	Recall
Angry	With filtering	94.14	95	90.9	91.1	87.55	87.53
	Without filtering	92.52	92.7	88.9	89.2	85.9	85.88
Sad	With filtering	90.1	88.3	88.9	87.54	85.9	83.87
	Without filtering	90.12	88.21	88.9	87.54	85.11	83.32
Happy	With filtering	95.26	95.77	94.11	94.2	88.76	87.34
	Without filtering	94.5	94.6	92.5	91.87	87.28	87.1

the probability of class 'overlapping speech' for that audio segment is higher than 75%. Though our evaluation we identify that, considering this threshold enables us to detect overlapping speech with 98.1% precision and 74.8% recall. Between precision and recall trade-offs we choose to achieve the highest possible precision with a cost of lower recall. The intuition is we want to identify and exclude as much overlapping audio segments as possible, but do not (if possible) want to exclude any single speech segments from consideration for emotion classification. Table 7 shows the increase of DER accuracy due to our overlapping speech filtering approach. According to this table overlapping speech filtering improves about 2% accuracy for the emotion: angry across all distance. But, there is no significant improvement for emotion sad, since very few false positives for sad classifier were overlapping speech and our overlapping speech filtering approach excludes them from emotion detection consideration.

Comparison with baseline. As discussed in section 7.1.4 we implemented the solutions of four baseline works [47, 50, 51, 57]. Table 8 shows the comparison of these baseline solutions with our DER solution on the family discussion dataset (section 7.2). Since, the generic correlation based feature selection approach [50] and 'context-aware' emotional relevant feature extraction, by combining Convolutional Neural Networks (CNNs) with LSTM networks [51] are not robust in realistic settings (discussed in section 7.1.4), their performance is very low in our evaluation on spontaneous family discussion data. Also, majority of the 6552 features extracted in the INTERSPEECH 13 baseline approach [47], distort significantly with variable speaker to microphone distances hence, this approach achieves low accuracy in our evaluation on spontaneous family discussion data. [57] uses a combination of prosodic acoustic and i-vector features and uses Recurrent Neural Network to detect speech emotion. This *ivector + RNN* approach achieves significantly low accuracy for our considered emotions (happy, angry and sad) compared to our evaluation in section 7.1.4, specially on 5.3 meters distance (22.5%, 24.1%, 20.1% lower accuracy and 20.8%, 30%, 22.9% lower recall on 5.3 meters distance compared to our DER approach).

8 DISCUSSION

Several researches on speech analysis have used The Electromagnetic Articulography (EMA) dataset [25] in their evaluation. [28] used i-vector method and GMM classifier for speech emotion recognition and applied on EMA dataset. They did not report accuracy for particular emotions, their highest reported emotion recognition accuracy achieved was 86.765% for male speakers. Also, [5] proposed a shape-based modeling of the fundamental frequency contour for emotion detection and evaluated on EMA dataset. This paper achieved 91.3%, 77.7% and 63.3% accuracy and 96%, 78% and 69.3% recall for emotions happy, angry and sad, respectively. Our Emo2vec emotion recognition approach achieves 97.2%, 96.4% and 93.8% accuracy and 98.8%, 98.2% and 96.16% recall for emotions happy, angry and sad, respectively, when applied on EMA dataset (without introducing any reverberation and de-amplification effect).

Table 8. Comparison with baseline on family discussion data

Mood	Distance (meters)	2.25		3		5.3	
		Accuracy	Recall	Accuracy	Recall	Accuracy	Recall
Angry	Our approach	94.14	95	90.9	91.1	87.55	87.53
	ivector+RNN	79.1	78.21	77.78	77.18	70.5	67.3
	CNN+LSTM	65.76	60.3	62.89	59.19	59	56.32
	INTERSPEECH 13	66.9	67.17	63.9	64.2	56.5	59.2
	Correlation based feature selection	57.21	58.2	54.45	53.43	51.1	48.77
Sad	Our approach	90.1	88.3	88.9	87.5	85.9	83.87
	ivector+RNN	76.21	75.76	75.3	75	71.51	68.2
	CNN+LSTM	67.2	61.27	68.7	62	60	54.98
	INTERSPEECH 13	66.4	66	61.3	60.73	55.3	53.9
	Correlation based feature selection	60	58.4	56.18	55.23	53.19	53.8
Happy	Our approach	95.26	95.77	94.11	94.2	88.76	87.34
	ivector+RNN	80.6	80.61	77.78	78.43	72.4	72.3
	CNN+LSTM	70	64.9	66.8	61.84	62.67	58.61
	INTERSPEECH 13	68.47	68.4	66.91	66	59.5	56.4
	Correlation based feature selection	58.39	56.9	55.9	54.5	54.6	52.74

Variable speaker to microphone distances introduce noise, de-amplification of speech and room reverberation in the captured speech signal. Speech enhancement is the area of study, which aims to improve speech intelligibility and overall perceptual quality of degraded speech signal using audio signal processing techniques. Our paper uses standard speech enhancement techniques such as, unvoiced audio segments removal using zero crossing rate (ZCR), noise removal using a band-pass filter and steady background noise removal using spectral noise gating algorithm to reduce noise. Hence, using these basic speech enhancement techniques we reduce background noise. However, to de-amplify speech due to increased speaker to microphone distance, it is necessary to know the distance. In general distance is not known. To the best of our knowledge there is no work which preforms blind, single channel (single microphone) speech amplification of a moving speaker across varying distances. Hence, the problem of speech de-amplification due to varying distance of a moving speaker cannot be done using speech enhancement. Further, although there are a wide variety of blind single channel (microphone) dereverberation techniques to handle room reverberation, only a few state of the art works addressed blind single channel dereverberation with moving subjects [11, 20]. Performance of these works significantly degrade with the increase of speaker to microphone distance [10]. Since, there is no good solution for either speech de-amplification or dereverberation using speech enhancement, we need a distance emotion detection (DER) solution, which is robust in emotion detection from speech with de-amplification and dereverberation.

Speech recognition is the process of capturing spoken words using a microphone and converting them into a digitally stored set of words. Speech recognition systems convert raw speech signal into a sequence of vectors (which are a representation of the signal from small frames) which are measured throughout the duration of speech. Then, using feature modeling and a classifier these vectors are matched with phonemes. In recent years some studies have address the distant speech recognition problem [40, 59]. While speech recognition is the process of converting speech to digital data, emotion detection is aimed toward identifying the mental states of a

person through speech. In this task, there is no specific definition or digital code for a vector (representation of a signal from small frames through features) to match with. Emotion detection works by analyzing the features of speech that differ between emotions. Hence, the challenges and approaches of speech recognition and emotion recognition are markedly different. Distant-emotion-recognition (DER) is an area not explored before to the best of our knowledge.

Linguistic content of the spoken utterance is also an important part of the conveyed emotion. Hence, in recent years, there has been a focus on hybrid features, i.e., the integration of acoustic and linguistic features [36, 45, 52]. But, all of these works used clean data without addressing the challenges of variable speaker to microphone distance. Although linguistic content of the spoken utterance can help the acoustic emotion features improve the accuracy of emotion recognition, current speech recognition systems still cannot reliably recognize the entire verbal content of emotional speech. Thus, the most popular feature representation for speech recognition is acoustic features such as prosodic features (e.g., pitch-related feature, energy-related features and speech rate) and spectral features (e.g.,Mel frequency cepstral coefficients (MFCC) and cepstral features).

To evaluate automated speech to text transcription accuracy on distance speech data, in this study we asked two volunteers to translate two different spontaneous family discussions (total four family discussions, 15-20minutes each) and label the words (of their translated text) from 1 to 3, where 1 was easy to understand and 3 is extremely difficult to understand from audio clips. We performed automated speech transcription using Google Cloud API [1] (uses the most advanced deep learning neural network algorithms) on these 4 family discussions. The accuracy of automated translation on words with different difficulty levels are shown in figure 11. According to the figure transcription accuracy on medium and difficult words (according to human labeling) is very low, where as transcription accuracy on acted clean emotion dataset (section 7.1.1) is 96.4%.

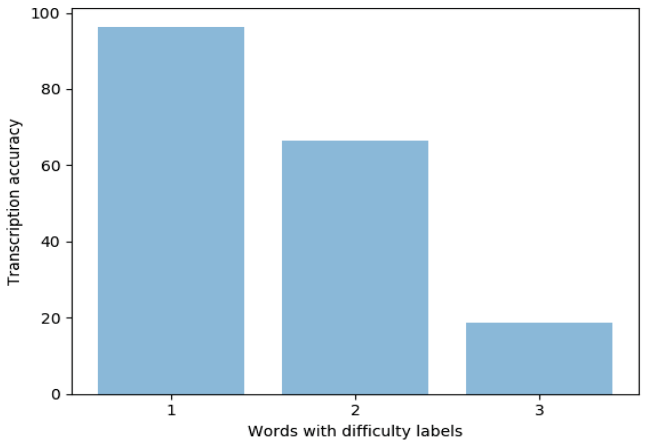

Fig. 11. Transcription accuracy of words with different difficulty levels

Since, accurate transcription of distance speech signal was out of scope of this study, we focused only on distance emotion detection using acoustic signal features.

Human emotion detection from acoustic speech signal has significant potential due to its non-intrusive nature (compare to wearables) and pervasive reachability to sensors (compared to video based emotion recognitions). Hence, in recent years speech emotion detection is receiving attention with progress of advanced human-computer

interaction systems [34, 35]. Also, emotion detection has paramount importance in the entertainment industry, either for the development of emotionally responsive games or toys [21] or for the development of serious games for aiding people with problems to understand social signs [7, 14]. Additionally some potential use of speech emotion detection can be in smart homes, e-learning [9], smart vehicles [27], etc. All of these applications need distant emotion recognition and our solution has applicability for them.

9 CONCLUSION

Automated DER has usability in non-intrusive monitoring of people with a variety of physical and mental health problems, as well as for human computer interaction systems, smart homes, the entertainment industry, and many other domains. Our novel DER solution consists of the identification of 48 features that minimally distort with distance and includes novel feature modeling that generates similar vectors for small frames which appear in a similar context for a specific emotion. These generated vectors represent the state of the small overlapping speech segments and then the temporal dynamics across these states are used in a LSTM classifier. The resulting solution addresses various challenges of DER. We evaluate our solution on two acted datasets (with artificially generated distance effect) as well as on our newly created emotional dataset of spontaneous family discussions with audio recorded from multiple microphones placed at different distances. Our solution achieves about 90% accuracy for our addressed emotions (happy, angry and sad), which is more than 16% on average better than the best baseline solution (among 4 baselines).

REFERENCES

[1] 2017. Google cloud speech API. https://cloud.google.com/speech/. (10 Feb 2017).

[2] 2017. Spaces in New Homes. goo.gl/1z3oVs. (10 Feb 2017).

[3] 2017. spectral noise gating algorithm. http://tinyurl.com/yard8oe. (1 Jan 2017).

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

[5] Juan Pablo Arias, Carlos Busso, and Nestor Becerra Yoma. 2014. Shape-based modeling of the fundamental frequency contour for emotion detection in speech. *Computer Speech & Language* 28, 1 (2014), 278–294.

[6] RG Bachu, S Kopparthi, B Adapa, and BD Barkana. 2008. Separation of voiced and unvoiced using zero crossing rate and energy of the speech signal. In *American Society for Engineering Education (ASEE) Zone Conference Proceedings*. 1–7.

[7] Emilia I Barakova and Tino Lourens. 2010. Expressing and interpreting emotional movements in social games with robots. *Personal and ubiquitous computing* 14, 5 (2010), 457–467.

[8] Linlin Chao, Jianhua Tao, Minghao Yang, and Ya Li. 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.

[9] Marti Cleveland-Innes and Prisca Campbell. 2012. Emotional presence, learning, and the online learning environment. *The International Review of Research in Open and Distributed Learning* 13, 4 (2012), 269–292.

[10] Christine Evers. 2010. Blind dereverberation of speech from moving and stationary speakers using sequential Monte Carlo methods. (2010).

[11] C Evers and JR Hopgood. 2008. Parametric modelling for single-channel blind dereverberation of speech from a moving speaker. *IET Signal Processing* 2, 2 (2008), 59–74.

[12] Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*. ACM, 1459–1462.

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

[14] Ofer Golan, Emma Ashwin, Yael Granader, Suzy McClintock, Kate Day, Victoria Leggett, and Simon Baron-Cohen. 2010. Enhancing emotion recognition in children with autism spectrum conditions: An intervention using animated vehicles with real emotional faces. *Journal of autism and developmental disorders* 40, 3 (2010), 269–279.

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

[16] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. 2013. Speech recognition with deep recurrent neural networks. In *Acoustics, speech and signal processing (icassp), 2013 ieee international conference on*. IEEE, 6645–6649.

[17] Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, et al. 2014. Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567* (2014).
[18] S. Haq and P.J.B. Jackson. 2010. *Machine Audition: Principles, Algorithms and Systems*. IGI Global, Hershey PA, Chapter Multimodal Emotion Recognition, 398–423.
[19] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.
[20] James R Hopgood and Christine Evers. 2007. Block-based TVAR models for single-channel blind dereverberation of speech from a moving speaker. In *Statistical Signal Processing, 2007. SSP'07. IEEE/SP 14th Workshop on*. IEEE, 274–278.
[21] Christian Jones and Jamie Sutherland. 2008. Acoustic emotion recognition for affective computer gaming. In *Affect and emotion in human-computer interaction*. Springer, 209–219.
[22] Martin Karafiát, Lukáš Burget, Pavel Matějka, Ondřej Glembek, and Jan Černockỳ. 2011. iVector-based discriminative adaptation for automatic speech recognition. In *Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on*. IEEE, 152–157.
[23] Kasturi Rangan Krishnamachari, Robert E Yantorno, Jereme M Lovekin, Daniel S Benincasa, and Stanley J Wenndt. 2001. Use of local kurtosis measure for spotting usable speech segments in co-channel speech. In *Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on*, Vol. 1. IEEE, 649–652.
[24] Duc Le and Emily Mower Provost. 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.
[25] Sungbok Lee, Serdar Yildirim, Abe Kazemzadeh, and Shrikanth Narayanan. 2005. An articulatory study of emotional speech production.. In *Interspeech*. 497–500.
[26] ZHOU Lian. 2015. Exploration of the Working Principle and Application of Word2vec. *Sci-Tech Information Development & Economy* 2 (2015), 145–148.
[27] Harold Lunenfeld. 1989. Human factor considerations of motorist navigation and information systems. In *Vehicle Navigation and Information Systems Conference, 1989. Conference Record*. IEEE, 35–42.
[28] Lenka Macková, Anton Čižmár, and Jozef Juhár. 2016. Emotion recognition in i-vector space. In *Radioelektronika (RADIOELEKTRONIKA), 2016 26th International Conference*. IEEE, 372–375.
[29] Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica* 22, 3 (2012), 276–282.
[30] Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernockỳ, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model.. In *Interspeech*, Vol. 2. 3.
[31] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. 3111–3119.
[32] Stephanie Pancoast and Murat Akbacak. 2012. Bag-of-Audio-Words Approach for Multimedia Event Classification.. In *Interspeech*. 2105–2108.
[33] Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. 2013. On the difficulty of training recurrent neural networks. *ICML (3)* 28 (2013), 1310–1318.
[34] Rosalind W Picard. 2000. Toward computers that recognize and respond to user emotion. *IBM systems journal* 39, 3.4 (2000), 705–719.
[35] Oudeyer Pierre-Yves. 2003. The production and recognition of emotions in speech: features and algorithms. *International Journal of Human-Computer Studies* 59, 1 (2003), 157–183.
[36] Tim Polzehl, Alexander Schmitt, Florian Metze, and Michael Wagner. 2011. Anger recognition in speech using acoustic and linguistic cues. *Speech Communication* 53, 9 (2011), 1198–1209.
[37] S Ramakrishnan and Ibrahiem MM El Emary. 2013. Speech emotion recognition approaches in human computer interaction. *Telecommunication Systems* (2013), 1–12.
[38] Rajib Rana. 2016. Emotion Classification from Noisy Speech-A Deep Learning Approach. *arXiv preprint arXiv:1603.05901* (2016).
[39] Shourabh Rawat, Peter F Schulam, Susanne Burger, Duo Ding, Yipei Wang, and Florian Metze. 2013. Robust audio-codebooks for large-scale event detection in consumer videos. (2013).
[40] Steve Renals and Pawel Swietojanski. 2014. Neural networks for distant speech recognition. In *Hands-free Speech Communication and Microphone Arrays (HSCMA), 2014 4th Joint Workshop on*. IEEE, 172–176.
[41] Steven A Rieger, Rajani Muraleedharan, and Ravi P Ramachandran. 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.
[42] Xin Rong. 2014. word2vec parameter learning explained. *arXiv preprint arXiv:1411.2738* (2014).
[43] Melissa Ryan, Janice Murray, and Ted Ruffman. 2009. Aging and the perception of emotion: Processing vocal expressions alone and with faces. *Experimental aging research* 36, 1 (2009), 1–22.
[44] Justin Salamon, Christopher Jacoby, and Juan Pablo Bello. 2014. A dataset and taxonomy for urban sound research. In *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 1041–1044.
[45] Asif Salekin, Hongning Wang, Kristine Williams, and John Stankovic. 2017. DAVE: Detecting Agitated Vocal Events. In *The Second IEEE/ACM International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE)*. IEEE/ACM.

[46] M Schroder and R Cowie. 2006. Issues in emotion-oriented computing toward a shared understanding. In *Workshop on emotion and computing*.

[47] Björn Schuller, Stefan Steidl, Anton Batliner, Alessandro Vinciarelli, Klaus Scherer, Fabien Ringeval, Mohamed Chetouani, Felix Weninger, Florian Eyben, Erik Marchi, et al. 2013. The INTERSPEECH 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism. (2013).

[48] Navid Shokouhi, Amardeep Sathyanarayana, Seyed Omid Sadjadi, and John HL Hansen. 2013. Overlapped-speech detection with applications to driver assessment for in-vehicle active safety systems. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2834–2838.

[49] Nitish Srivastava, Geoffrey E Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research* 15, 1 (2014), 1929–1958.

[50] Ashish Tawari and Mohan M Trivedi. 2010. Speech emotion analysis in noisy real-world environment. In *Pattern Recognition (ICPR), 2010 20th International Conference on*. IEEE, 4605–4608.

[51] George Trigeorgis, Fabien Ringeval, Raymond Brueckner, Erik Marchi, Mihalis A Nicolaou, Björn Schuller, and Stefanos Zafeiriou. 2016. Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network. In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 5200–5204.

[52] Chung-Hsien Wu and Wei-Bin Liang. 2011. Emotion recognition of affective speech based on multiple classifiers using acoustic-prosodic information and semantic labels. *IEEE Transactions on Affective Computing* 2, 1 (2011), 10–21.

[53] Bai Xue, Chen Fu, and Zhan Shaobin. 2014. A study on sentiment computing and classification of sina weibo with word2vec. In *Big Data (BigData Congress), 2014 IEEE International Congress on*. IEEE, 358–363.

[54] Takuya Yoshioka, Xie Chen, and Mark JF Gales. 2014. Impact of single-microphone dereverberation on DNN-based meeting transcription systems. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 5527–5531.

[55] Mingyu You, Chun Chen, Jiajun Bu, Jia Liu, and Jianhua Tao. 2006. Emotion recognition from noisy speech. In *Multimedia and Expo, 2006 IEEE International Conference on*. IEEE, 1653–1656.

[56] Zhihong Zeng, Maja Pantic, Glenn I Roisman, and Thomas S Huang. 2009. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE transactions on pattern analysis and machine intelligence* 31, 1 (2009), 39–58.

[57] Teng Zhang and Ji Wu. 2015. Speech emotion recognition with i-vector feature and RNN model. In *Signal and Information Processing (ChinaSIP), 2015 IEEE China Summit and International Conference on*. IEEE, 524–528.

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

[59] Yu Zhang, Guoguo Chen, Dong Yu, Kaisheng Yaco, Sanjeev Khudanpur, and James Glass. 2016. Highway long short-term memory rnns for distant speech recognition. In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 5755–5759.

[60] Patrick H Zimmerman, J Elizabeth Bolhuis, Albert Willemsen, Erik S Meyer, and Lucas PJJ Noldus. 2009. The Observer XT: A tool for the integration and synchronization of multimodal signals. *Behavior research methods* 41, 3 (2009), 731–735.

Received February 2017; revised May 2017; accepted June 2017