Emotion Recognition Robust to Indoor Environmental Distortions and Non-targeted Emotions Using Out-Of-Distribution Detection

YE GAO, University of Virginia
ASIF SALEKIN, Syracuse University
KRISTINA GORDON, University of Tennessee
KAREN ROSE, Ohio State University
HONGNING WANG, University of Virginia
JOHN STANKOVIC, University of Virginia

The rapid development of machine learning on acoustic signal processing has resulted in many solutions for detecting emotions from speech. Early works were developed for clean and acted speech and for a fixed set of emotions. Importantly, the datasets and solutions assumed that a person only exhibited one of these emotions. More recent work has continually been adding realism to emotion detection by considering issues such as reverberation, de-amplification, and background noise, but often considering one dataset at a time, and also assuming all emotions are accounted for in the model. We significantly improve realistic considerations for emotion detection by (i) more comprehensively assessing different situations by combining the 5 common publicly available datasets as one and enhancing the new dataset with data augmentation that considers reverberation and de-amplification, (ii) incorporating 11 typical home noises into the acoustics, and (iii) considering that in real situations a person may be exhibiting many emotions that are not currently of interest and they should not have to fit into a pre-fixed category nor be improperly labeled. Our novel solution combines CNN with out-of-data distribution detection. Our solution increases the situations where emotions can be effectively detected and outperforms a state-of-the-art baseline.

CCS Concepts: • **Applied computing → Health informatics**; • **Computing methodologies → Machine learning**.

Additional Key Words and Phrases: synthetic datasets, convolutional neural networks, out-of-distribution detection, emotion detection, noise, distance, reverberation

ACM Reference Format:
Ye Gao, Asif Salekin, Kristina Gordon, Karen Rose, Hongning Wang, and John Stankovic. 2021. Emotion Recognition Robust to Indoor Environmental Distortions and Non-targeted Emotions Using Out-Of-Distribution Detection. 1, 1 (July 2021), 22 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Many research papers from the field of psychology [31], [15] [32], [9] have shown that emotional health is a crucial part of one's well-being. The rapid development of machine learning in the field of acoustic signal

Authors' addresses: Ye Gao, Department of Computer Science, University of Virginia, yg9ca@virginia.edu; Asif Salekin, Department of Electrical Engineering & Computer Science, Syracuse University, asalekin@syr.edu; Kristina Gordon, Department of Psychology, University of Tennessee, kgordon1@utk.edu; Karen Rose, College of Nursing, Ohio State University, rose.1482@osu.edu; Hongning Wang, Department of Computer Science, University of Virginia, hw5x@virginia.edu; John Stankovic, Department of Computer Science, University of Virginia, jas9f@virginia.edu.

XXXX-XXXX/2021/7-ART $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

processing has resulted in a surge of interest in detecting emotions from speech. Recent publications [36], [24], [16], [55] have classified emotions based only on acoustics with no visual images. These acoustics-based emotion detection algorithms have the potential to monitor people's emotions 24 hours a day and consequently play a vital role in maintaining people's emotional well-being. For example, in one important application, family or informal caregivers for persons with dementia can benefit from these algorithms. These caregivers are twice as likely to experience emotional difficulties compared to the caregivers of patients living with diseases other than dementia [27]. Because long-term untreated emotional difficulties are linked to mental health disorders, helping the caregivers of persons with dementia become aware of their emotions is of paramount importance. However, it is impractical to appoint a human professionally trained to detect the onset of various emotions in the household of each pair of dementia caregiver and care recipient dyad. Therefore, in this application area, the need for emotion detection algorithms is pressing. It is also well known that many other in-home and in-office applications can also benefit from detecting emotions.

The development of machine learning in the field of speech-based emotion recognition has produced solutions such as [2], [50], [53]. The early works in speech-based emotion detection were developed for clean and acted speech and for a fixed set of emotions such as happiness, anger, sadness, and neutrality. One caveat with using such datasets to develop an emotion detection algorithm is that the datasets and algorithms assume that people only exhibit one of those emotions. Recent works have considered the reality of speech-based emotion recognition by taking into account the effect of environmental distortions such as de-amplification, background noise, and reverberation caused by sound signals bouncing on objects. However, these solutions often consider one dataset at a time and consider that all emotions are accounted for in the model developed on that one dataset.

In spite of the progress made on emotion detection from speech, challenges remain to more accurately handle in the wild situations and to identify emotions of interest among the vast array of human emotions exhibited by individuals in every day life. Out of the state-of-the-art algorithms published on emotion recognition [35], [6], [28], [47], [54], [48], in Table 2, only three of them deals with environmental distortions such as deamplification, background noise, and reverberation. Out of the three works that deals with environmental distortions, there is no evaluation on how their algorithm fares when different combination of reverberation factors, such as the decay factor and diffusion, are combined.

Based on the state-of-the-art in this field, **Key Challenges** for speech-based emotion detection are:

- Many speech-based emotion detection algorithms are developed on datasets of either clean speech or speech that are environmentally distorted in various degrees, but they assume that people will only exhibit one of the emotions accounted for in the dataset.
- The effect that environmental distortions has on the classification of different emotions is not well studied, because, despite that there exist works [48], [54] that take environmental distortions into account, they don't study to what extent does each environmental distortion affect the classification of different emotions.

The main contributions of this paper are:

- We created a combined CNN and out of data distribution (OOD) solution that performs in the range of 90% accuracy even in the presence of non-targeted emotions, 11 realistic home noises, deamplification due to distance from the microphones, and reverberations due to different room types. This is a significant improvement over a state-of-the-art model, the hierarchical classifier described in Section 5.2 whose accuracy over the same testing set (that includes environmentally distorted samples and samples of non-targeted class) is 56.2%. The hierarchical classifier outperforms five state-of-the-art models on samples with and without environmental distortions (See table 2).
- To address the first challenge, we show that explicitly training on non-interested confounding emotions where you have data plus employing an OOD technique for non-interested emotions where you don't have data outperforms using just an OOD for all non-interested emotions by more than 10%.

- To address the second challenge, we demonstrate how different environmental distortions affect the classification results: background noise and deamplification have the most impact on the decrease of classification accuracy by 7.3%, followed by room reverberation that results in the decrease of classification accuracy by 4.5%.
- We create a novel synthetic dataset usable by the community that incorporates a padding strategy to better mimic real world speech monitoring than current datasets, and includes deamplification, reverberation, and added noise from 11 types of real home background noise. The information of the synthetic dataset as well as the 5-class CNN and OOD solution is publicly available at: Website.
- Accurate mood detection by our solution allows the emotions of users in a smart home to be automatically detected in a passive manner. Upon detection, while not part of this paper, relaxation and mindfulness techniques that have been shown to alleviate unhealthy mood can be recommended to the users.

2 RELATED WORKS

There have been studies on detecting emotions in smart-home environments. Many of these do not use speech. For example, Fernandez et al. propose a proof-of-concept architecture that detects a user's emotion by analyzing their facial expression and physiological signals [25]. Mano et al. propose to use the facial expression of elderly people to detect their emotions; the emotion detection classifier can identify when an elderly person needs urgent help [43]. Zhao et al. develop EQ-Radio that transmits an RF signal that will bounce off a person's body and back to the analyzing device. The RF signal carries the information of heartbeat and respiration signals that can be teased out and used for emotion recognition [56]. However, it is not always feasible to deploy such systems.

Other in-home systems such as Alexa and Siri are providing voice assistance for access to health-care services and support for reminders, detecting colds, etc. To date, they do not assess mood. In addtion, there have been works that use a smart home assistant to detect mood, such as Chatterjee et al. [13], but this work is evaluated only on clean datasets that are not environmentally distorted. Callisto et al. [11] also propose to use a smart home assistant robot (and available smart home speakers can fill this role) to detect emotions in smart homes, but to date this work has been only a proposal.

The works that have focused on using the acoustic modality to detect emotions on samples that are under various degrees of environmental distortions are by Triantafyllopoulos et al. [50], Dickerson et al. [20], and Vrebcevic et al. [52]. The Deep Residual Network [50] is a scalable deep learning network that can be used to detect emotions in previously unseen environments due to its noise-removal procedure. RESONATE [20] is a reverberation compensation approach that add reverberation (but not noise or deamplification effects) to a training corpus in order for a model trained on this corpus to be able to account for reverberation. After adapting Alex-Net so it is suitable to process sound, Vrebcevic et al. [52] train the classifier on samples that are contaminated with ambient noise (but there are no reverberation or deamplification effects).

The works that have focused on detecting emotions in real-time include Lech et al. [37] that adapts AlexNet for speech processing but does not deal with environmental distortions, Cen et al. [12] that develops a system consisting of voice activity detection and emotion recognition. However, Cen et al. [12] is evaluated on a simulated online learning environment so the testing samples are absent of environmental distortions. Not taking environmental distortions explicitly into account is a very prevalent problem with works on emotion detection in real-time or time-sensitive tasks; works that suffer from this problem also include Stolar et al. [49], Fayek et al. [23], and Bahreini [4].

A published summary [22] shows acoustic features or parameters that have been used for emotion detection, encompassing frequency, time, and amplitude. Since a huge variety of features are used in different previous works and different works uses different set of these features, direct comparison and cross-examination on the effectiveness of the features prove challenging. The summary [22] proposes that emotions from the datasets

Work	Approach	Performance	Environmental Distortion
[18]	GentleBoost	86.3%	×
[41]	Linear Discriminant Analyses	81.8%	×
[3]	Support Vector Machine	85.5%	×
[17]	Ensemble of SVM	86.3%	×
[53]	Fourier Transform + SVM	79.51%	×
[19]	Autoencoder-Based Domain Adaptation	62.0%	×
[2]	Convolutional Neural Network	69%	×
[50]	Deep Residual Network	86.3% (unweighted aver. recall)	✓
[20]	RESONATE	80% (approximate)	✓
[52]	Alex-Net + Data Augmentation	34.03%	✓
[8]	Evaluation by Humans	86.0 %	×

Table 1. Evaluation of previous works on EMO-DB. The performance, if not otherwise noted, is measured by accuracy. Three of the works attempt to deal with environmental distortions. The definition of environmental distortions is defined differently in different works.

analyzed can be represented by two features: valence, which represents the degree of positivity or negativity of an expressed emotion, and arousal, which represents the energy or intensity of the expressed emotion. Trigeorgis et al. [51] use valence and arousal for emotion detection from speech. The usage of acoustics-based emotion recognition using valence and arousal extends from emotion classification from speech to classifying emotions based on music [30], which provides a set of features that describe valence and arousal in a musical piece. However, when it comes to realistic deployments and detecting emotions from speech in the wild, the features extracted from the raw audio signal must be resistant to both background noise and various combinations of reverberation factors. The demonstration that the features descriptive of valence and arousal will not be sensitive to environmental distortions remains lacking.

Other popular datasets of acted emotional speech include the Surrey Audio-Visual Expressed Emotion Database (SAVEE) [33], the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS), [40] and the Crowd-sourced Emotional Multimodal Actors Dataset (CREMA-D) [10]. These three datasets consist of both audio-video and audio-only samples. Other works [6], [28], [5], [46] seek to exploit the multi-modality of the datasets, while works such as [35] only use the audio clips from SAVEE and RAVDESS. Despite the high accuracy achieved by some of the works on one or more of the three multi-modal datasets, they still haven't explicitly addressed the issue of environmental distortions. Salekin et al. [47] publish the Distant Emotion Recognition in which they select features that are robust to distance and only extract those features when processing a speech signal. However this work does not evaluate its model on speech that are contaminated with background noise, so it is hard to tell if it is robust to this particular kind of environmental distortion. Another work [54] improves the robustness of speech-based emotion recognition by considering the magnitude spectrogram and the modified group delay spectrogram. This work considers both noise and reverberation, but they simply show that their work is robust noise contaminated speech samples and there is no evaluation on how their model fares when different combination of reverberation factors, such as the decay factor and diffusion, are combined.

Since confounding emotions are largely overlooked in research papers by the acoustic-based emotion detection community, we are unable to find any research that indicates the effort of filtering out confounding emotions so that such samples will not be sent to the classifier and result in a wrong classification. A possible approach to discern an emotional utterance sample that is of emotions that the classifier is not trained to classify is to detect

Work	Dataset(s)	Approach	Accuracy	Distortions
[35]	RAVDESS, SAVEE	Shake-Shake Regularization	60.8 %	×
[6]	SAVEE, CREMA-D	LSTM encoder	41.2 %	×
[28]	RAVDESS	Temporal DNN	67.7 %	×
[28]	CREMA-D	Temporal DNN	74.0 %	×
[54]	RAVDESS	Noise Mitigation + CNN	81.0 %	✓
[48]	RAVDESS	Noise Mitigation	80.0 %	✓
Our baseline	RAVDESS, EMA, CREMA-D, TESS, SAVEE	Hierarhical Classifier	94.7 %	✓
Our baseline	RAVDESS, EMA, CREMA-D, TESS, SAVEE	Hierarhical Classifier	88.4 %	×

Table 2. Evaluation of works on SAVEE, CREMA-D, RAVDESS, EMA, TESS. The performance, if not otherwise noted, is measured by accuracy. The definition of environmental distortions is defined differently in different works. Except for the hierarhical classifier, the other items are published state-of-the-art works on emotion recognition on (a subsets of) the datasets that our solution is trained on. The hierarchical classifier is developed by us to serve as a baseline that we compare against our solution because it is trained on the same datasets that our solution is trained on. On clean samples, our baseline achieves an accuracy of 94.7% which outperforms the first three baselines (two of which use the same algorithm) evaluated on clean speech. On environmentally distorted samples, our baseline achieves an accuracy of 88.47%, which out-performs the last three baselines that are tested with environmental distortions.

if this sample is within the distribution of the training samples. Hendrycks et al. discovers that, when wrongly classified, an out-of-distribution sample results in a generally smaller softmax probability than an in-distribution sample that is correctly classified [34]. Based on this observation, they propose to detect an out-of-distribution sample by comparing it within the context of the statistics of the softmax probability scores by all samples (in the testing set). This approach's performance is improved by Liang et al. [39] who use temperature scaling and input perturbing for out of distribution detection. Both Liang et al. and Hendrycks et al.'s approaches are out-peromed by a Mahalanobis distance-based approach: Lee et al. [38] proposes a way to detect out-of-distribution samples to prevent them from being sent to the softmax layer of a deep neural network and result in a wrong classification. For each training sample, Lee et al. calculate the activation of the penultimate layer (the layer that will forward its activation to the output layer) to get the distribution of the training samples. After training, when a previously unseen sample is sent to the classifier for classification, the activation of the penultimate layer of the neural network is calculated before being sent to the output layer. The Mahalanobis distance [42] from the activation of the penultimate layer by this previously unseen sample to the distribution of the activations of the penultimate layers by training samples thisen be calculated. The Mahalanobis distance measures how many standard deviations a data point is away from the distribution of a group of data points.

3 SYNTHETIC DATASETS

There are 5 well known datasets used for emotion detection: CREMA-D, SAVEE, RAVDESS, TESS, and EMA. Table 3 shows what emotional utterances are in each of the datasets. Even combining these datasets results in too limited data and that data would not account well for noise, reverberation and deamplification. Consequently, we not only combine these 5 datasets to provide more samples and to be more general, but we also develop multiple synthetic datasets. These datasets include clean speech (to serve as a baseline), noisy speech, reverberated speech, and combining all the factors and adding deamplification. This enables the evaluation to show the effects of each factor as well as the overall situation that closely represents realistic environments. The following subsections describe how we built the synthetic datasets from the five publically available sets of emotional speech. We also

Dataset	HAP	ANG	NEU	SAD	SUR	FEA	DIS	CAL
CREMA-D	✓	✓	✓	✓	×	✓	✓	×
RAVDESS	✓	✓	✓	✓	✓	✓	✓	✓
SAVEE	✓	✓	✓	✓	✓	✓	✓	×
TESS	✓	✓	✓	✓	✓	✓	✓	×
EMA	✓	✓	✓	✓	×	×	×	×

Table 3. The components of the original datasets: CREMA-D, SAVEE, RAVDESS, TESS, and EMA. The emotions in this Table are: Happiness, Anger, Neutrality, Sadness, Surprise, Fear, Disgust, and Calm.

describe how we create out of distribution samples to test the case when humans exhibit emotions that are not accounted for in the training.

The fact that all five of the publically available datasets have happy, angry, neutral, and sad classes suggests that these four emotions are acknowledged to be commonplace, but distinctive enough to be separated from other emotions. As a result, these four classes of emotions are also included in our synthetic datasets as emotions of interest. In addition to these four classes, we also have a class of confounding emotions, which are emotions that are distinct enough to not be confused with any of the four commonplace emotions. Because surprise is similar to happiness (for example, TESS lists the surprised emotional utterances are speech samples of "pleasant surprise") [21], and boredom is similar to neutrality, surprise and boredom are not considered confounding emotions, rather variations of happiness and neutrality. This leaves us with two confounding emotions - fear and disgust, and one class of out-of-distribution emotion, calmness. The class of confounding emotions in our synthetic datasets consists of fearful and disgusted speech samples.

3.1 Generate padded samples that are otherwise not environmentally distorted, denoted as $\mathcal{D}_1$

After getting rid of corrupted and otherwise unusable samples from the five sets of emotional speech, we end up with the set of clean audio samples of various lengths. Since most of the audio clips are less than or equal to 4 seconds, we pad each of them into a 5-second window. In order to do the padding, we first generate a 5-second segment of pure silence audio clip and randomly decide the index of a frame in the 5-second segment and overlay a sample of emotional speech with the silence with that frame index as the starting point. This strategy of padding results in a more diverse set of samples of emotional speech in comparison to the strategy in which each sample of emotional speech is overlaid with the silence segment starting universally at a fixed index. Our padding strategy also results in a more realistic dataset that resembles the set of emotional speech segments collected in the wild, as it is unreasonable to expect that each speech segment is perfectly captured by the microphone right when the first syllable of the segment is spoken. By providing a more diverse and more realistic way of padding, we increase the diversity of our dataset and this contributes to the robustness of the classifier.

The complete set of generated clean padded samples is referred as $\mathcal{D}_1$. In $\mathcal{D}_1$, there are 1792 happy samples, 1793 angry samples, 1573 neutral samples, 1793 sad samples, and 1837 samples of confounding speech. In total, there are 8788 samples.

3.2 Generate de-amplified, noise-contaminated samples, denoted as $\mathcal{D}_2$

For each audio clip in $\mathcal{D}_1$, we create two copies of them. For each of the two copies, we randomly de-amplify it with m decibels such that $m \leq 12$. Then, we randomly choose among the dataset of background noise collected in real homes which are around 5-minutes long; within a certain chosen clip of background noise in real home environments, we randomly take a 5-second audio segment from it, and overlay it with the de-amplified sample.

Event	Instances
(object) rustling	60
(object) snapping	57
cupboard	40
cutlery	76
dishes	151
drawers	51
glass jingling	36
object impact	250
people walking	54
washing dishes	84
water tap running	47

Table 4. Events that are present in the background noise collected from real homes from the dataset [44]. All of them are covered in the process of contaminating audio samples with background noise.

Table 4 illustrates the events that are present in these audio clips of home environments. The way we overlay the background noise clips with the duplicate samples ensures that the household events in Table 4 are present in the resulted audio clips of the overlaying.

The set of de-amplified, noise-contaminated samples is referred as $\mathcal{D}_2$. In $\mathcal{D}_2$, there are 3584 happy samples, 3586 angry samples, 3146 neutral samples, 3586 sad samples, and 3674 samples of confounding speech. In total, there are 17576 samples.

3.3 Generate reverberated samples, denoted as $\mathcal{D}_3$

For each audio clip in $\mathcal{D}_1$, we duplicate it once. The duplication is then reverberated. The reverberation effect is generated by the combination of three reverberation factors: the wet/dry ratio r, diffusion d, and the decay factor f. Each time an audio sample is reverberated, a random set of values for the wet/dry ratio, diffusion, and the decay factor is generated. The choice of the reverberation parameters and the set of their values is also used in DER (Distant Emotion Recognition) [14] which seeks to solve the problem of speaker identification by generating different reverberation models to represent difficult rooms in which a speaker might be present. Their approach yields almost zero error rate when tested in real-time when human participants were speaking in a room where a microphone array was present. The performance achieved by DER indicates that the different combinations of the given reverberation parameters and their given ranges are able to describe typical indoor environments. Therefore, we adopt the same reverberation model to change the audio samples as if they were collected in the indoor environments described by the different combinations of the reverberation parameters.

The set of reverberated samples is referred as $\mathcal{D}_3$. In $\mathcal{D}_3$, there are 1792 happy samples, 1793 angry samples, 1573 neutral samples, 1793 sad samples, and 1542 samples of confounding speech. In total, there are 8493 samples.

3.4 Generate samples that are de-amplified, noise-contaminated, and reverberated, denoted as $\mathcal{D}_4$

For each audio clip in $\mathcal{D}_2$ that is not of a confounding emotion, we duplicate it once. For each duplication, we reverberate it in the same way as we obtain the reverberated samples in $\mathcal{D}_3$. The set of samples that are de-amplified, noise-contaminated, and reverberated is denoted as $\mathcal{D}_4$. In $\mathcal{D}_4$, there are 3584 happy samples, 3586 angry samples, 3146 neutral samples, and 3586 sad samples. In total, there are 13902 samples.

$$\mathcal{D} = \cup_{i=1}^{n} \mathcal{D}_i, n = 4 \tag{1}$$

$$\cap_{i=1}^{n} \mathcal{D}_i = \varnothing, n = 4 \tag{2}$$

3.5 Training and testing sets, denoted as $\mathcal{D}_{train}$ and $\mathcal{D}_{test}$

After generating the above synthetic datasets we take the union of them for overall evaluation. We split the union of the synthetic datasets into training and testing sets. To ensure that all the emotion classes are equally represented in both the training and testing sets, we randomly select 80% of confounding samples, 80% of happy samples, 80% of angry samples, 80% of neutral samples, 80% of sad samples and use them for training, while the rest of the samples are used for testing. The 80% of samples selected from an emotion consists of samples from $\mathcal{D}_1$, $\mathcal{D}_2$, $\mathcal{D}_3$, $\mathcal{D}_4$ so the environmental distortions and their combination are accounted for the testing set.

Let $\mathcal{D}_{train}$ denote the training set and $\mathcal{D}_{test}$ denote the testing set. Equation 3 describes the relationship between the entirety of the synthetic datasets, $\mathcal{D}$, and $\mathcal{D}_{train}$ and $\mathcal{D}_{test}$. The mutual exclusion of $\mathcal{D}_{train}$ and $\mathcal{D}_{test}$ is described in Equation 4.

$$\mathcal{D} = \mathcal{D}_{test} \cup \mathcal{D}_{train} \tag{3}$$

$$\mathcal{D}_{test} \cap \mathcal{D}_{train} = \varnothing \tag{4}$$

Equations 5 and 6 describe the subsets of the training and testing sets based on the subsets' relationship to $\mathcal{D}_1$, $\mathcal{D}_2$, $\mathcal{D}_3$, and $\mathcal{D}_4$. Each of $\mathcal{D}_i$, $i \in \{1, 2, 3, 4\}$ shares some members with the training set. Similarly, it also shares some members with the testing set. For example, $\mathcal{D}_{train,1}$ denotes the set of samples in the training set that are not distorted in anyway, and $\mathcal{D}_{test,3}$ denotes the set of samples in the testing set that are reverberated, but are neither deamplified nor contaminated with noise.

$$\mathcal{D}_{train,i} = \mathcal{D}_{train} \cap \mathcal{D}_i, i \in \{1, 2, 3, 4\} \tag{5}$$

$$\mathcal{D}_{test,i} = \mathcal{D}_{test} \cap \mathcal{D}_i, i \in \{1, 2, 3, 4\} \tag{6}$$

3.6 Generate samples that are out of the distribution of the training and testing sets

We have addressed the importance of considering confounding emotions. However, the number of confounding emotion classes to be included in the training and testing sets is finite. If an emotional speech segment is not within the distribution of the samples of interested and confounding emotions in the training set, this speech segment does not belong to any of the interested or confounding classes. Classifying such a sample will be pointless, as the classifier will make mistakes on classifying it. Therefore, out-of-distribution detection is crucial, and it will further assist in filtering out emotional utterances that are not of the interested classes.

We generate out-of-distribution samples to test the performance of the out-of-distribution technique. We use calm speech segments from TESS as out-of-distribution samples, because in-distribution samples will be happy, angry, neutral, sad, or confounding (disgusted or fearful). We generate synthetic calm samples in the same way that we generate synthetic samples of other emotions. First, we pad them to 5-second window, following the exact step described in Section 3.1. This group of padded but not otherwise environmentally distorted calm samples is represented by O_1. Then, we create three copies of them. For the two copies of each padded calm sample, we de-amplify them and contaminate them with noise, as described in Section 3.2. This group of calm samples is represented by O_2. The other copy of the padded calm sample is reverberated, following the same procedure described in Section 3.3. This group of calm samples is represented by O_3. Then, we make a copy of the

Low-Level Descriptors (LLD's)	Amounts
Mel-Frequency cepstral coefficients (MFCC) 1-13	104
Delta coefficients for MFCC 1-13	104
Zero-crossing rates	8
Delta coefficients for zero-crossing rates	8
Root-mean-square signal frame energy	8
Delta coefficients for root-mean-square signal frame energy	8
Spectral centroid features	8
Delta coefficients for the spectral centroid related features	8
Pitch-related features	8
Delta coefficients for the pitch-related features	8
Total amount	272

Table 5. The 272 low-level descriptor features

calm samples that are de-amplified, noise-contaminated, but not reverberated. For all the copies, we reverberate them, so they become de-amplified, noise-contaminated, and reverberated. This group of calm speech segments is represented by O_4. Collectively, the entire set of calm speech segments is represented by O, which is the union of O_1, O_2, O_3, and $O4$. Note that these calm speech segments are in neither training nor testing sets.

4 FEATURE SELECTION

In this work we use 272 low-level descriptor features associated with emotions because these are common to many previous solutions such as [47]. Table 5 provides a summary of these features.

5 SOLUTION

5.1 Overview

To illustrate the importance of considering confounding emotions, we have developed two emotion detection algorithms. One of the algorithms serves as a baseline and is a hierarchical structure that consists of 3 CNN's and this algorithm only has four classes (section 5.4). It is able to achieve high accuracy even in the presence of background noise, de-amplification, and reverberation, under the condition that all audio clips passed to this structure are happy, angry, neutral, or sad utterances. However, as the Evaluation section shows, the accuracy significantly drops once the condition no longer holds - in other words, the structure fails to perform adequately if speech samples of confounding emotions are added which is the typical situation in the real world.

The second algorithm (sections 5.2 and 5.3) is our solution. Due to the importance of considering confounding emotions, our solution is a CNN classifier that has five classes - confounding emotions (emotions that we are not interested in, but for which we have data), happiness, anger, neutral, and sadness and to these 5 classes we add an OOD component to capture confounding emotions where we do not have data.

Our solution works with 5-second audio clips. If the input is shorter than 5-seconds, we pad it with the padding algorithm described in the synthetic datasets section. Note that the authors of [47] show that 5-second audio samples obtained from padding the original samples from datasets yields high accuracy across various speaker-to-microphone distances. After making the input to be exactly 5-seconds long, we slice it into 48 small, overlapping frames, each of which lasts for 25 ms. From each small frame, we obtain 272 LLD features. As a result, the input of our classifier is a tensor of 48 frames × 272 channels per frame. The hyper-parameters we choose are provided in [47], as these values for the hyper-parameters are shown to achieve good classification accuracy.

	Softmax Distribution	ODIN	Mahalanobis
Accuracy	85.0%	91.0%	95.7

Table 6. The detection accuracy of three out-of-distribution detection algorithms. The in-distribution samples are CIFAR-10 samples. Samples of SVHN are out-of-distribution samples [38]. The metric is accuracy.

5.2 The 5-Class CNN Classifier

Although neural networks sometimes yield good performance with minimally tuned hyper-parameters, [29] suggests that performance will significantly increase if a thorough tuning of hyper-parameters is performed. Grid search is recommended by [29] as an approach to perform hyper-parameter optimization when there are relatively few hyper-parameters. When performing grid search over a set of hyper-parameters, the programmer assigns a small, finite set of values for each hyper-parameter. The grid search algorithm loops over the combinations of the specified hyper-parameters and trains the model multiple times to iterate through the combinations. The combination of the hyper-parameters that results in the model that yields the least validation set error is the final choice of the hyper-parameters.

The hyper-parameters that we choose to tune are a standard choice for tuning CNN's. Due to the relatively small number of the hyper-parameters we have, we choose to perform grid search on them, instead of random search, another hyper-parameter optimization algorithm that is quicker but less thorough. We have performed grid search over the following hyper-parameters: epoch, batch size, the number of convolutional layers, the number of kernels in each layer, the size of kernels in each layer, the stride size of max-pooling, the activation function (ReLU and Leaky ReLU), the optimizer, the learning rate, the decay ratio, and the size of the dense layers after the convolutional neural nets are flattened. Our final model consists of four convolutional layers. After the output of the final CNN is flattened, three dense layers of 2048 neurons are attached. We choose the adam optimizer with the learning rate as 1e-4 and decay ratio as 0. The optimal batch size is 128. The optimal epoch is 1000. The optimal filter size is 3 (filters are measured by the number of small frames).

The classifier is implemented with Keras, a neural network API using Tensorflow backend and trained on the samples in the training set, $\mathcal{D}_{train}$. For each sample $\boldsymbol{X}$ in $\mathcal{D}_{train}$, we slice them into 48 overlapping small frames $x_1 ... x_n, n = 48$ with hop length of 5. Hop length is defined as the amount of samples between two small frames. Librosa, a python package on acoustic signal analysis, is used to extract the 272 features for each small frames s_i such that $i \in \{1, ..., 48\}$. As we have described in Table 5, the features can be categorized into (1) MFCC's 1-13, (2) the delta coefficients of MFCC's 1-13, (3) zero-crossing rates, (4) the delta-coefficients of zero-crossing rates, (5) RMSE, (6) delta-coefficients of RMSE, (7) spectral centroids, (8) delta-coefficient of spectral centroids, (9) pitch-related features, and (10) their delta-coefficients. For each of these categories, we calculate the minimum, the maximum, the median, the mean, the standard deviation, the variability, the skewness, and the kurtosis. This indicates that, for the first category, we have 13×8 features, for the second category, we have 13×8, for the rest of the categories, each category has 1×8 features. In total, we have $2 \times 13 \times 8 + 8 \times 1 \times 8 = 272$ features.

5.3 The Detection of Out-Of-Distribution Samples

The fifth class (the confounding emotions class) of the classifier described above is to prevent disgusted and fearful emotional speech segments (as examples of emotions for which we are not interested in but have data) from being classified as happiness, anger, neutrality, or sadness. However, this classifier can handle only six most basic emotions (happiness, anger, neutrality, sadness, fear, and disgust), and human beings are capable of expressing other emotions. Table 3 shows that the TESS dataset considers calmness to be an emotion category that is different from the aforementioned six emotions.

In order to further filter out emotional speech segments that are not included by the 5 emotions, we use the out-of-distribution technique [38] using Mahalanobis distance. We select this technique because it outperforms two other major out-of-distribution detection algorithms - softmax Distribution and ODIN as indicated in Table 6. In the following paragraphs, we describe in detail the Mahalanobis distance-based OOD detection technique.

Let $\boldsymbol{x}_i$ and y_i represent a sample in the training set and its label. Let N_c represent the number of all samples in the training set whose label is of class c. Equation 7 defines the empirical mean of a class c and f represents the activation of the penultimate layer of our 5-class CNN by the input $\boldsymbol{x}_i$. $\hat{\mu}_c$ is the mean of the activations of the penultimate layer of our 5-class CNN by all samples in the training set whose label is of class c.

$$\hat{\mu}_c = \frac{1}{N_c} \sum_{i:y_i=c}^{N} f(\boldsymbol{x}_i) \tag{7}$$

Equation 8 [38] defines the empirical covariance matrix, an important part required to represent the distribution of the training set. After the empirical means for all $c \in C$ are calculated, whereas C is the set of all possible labels, $(f(\boldsymbol{x}_i) - \hat{\mu}_c)$ is used to measure how each samples in the training set deviates from the empirical mean of all samples in the same class, c. $(f(\boldsymbol{x}_i) - \hat{\mu}_c)(f(\boldsymbol{x}_i) - \hat{\mu}_c)^\top$ results in a symmetric matrix with numbers of the rows and columns equal to the number of neurons of in the penultimate layer of our 5-class CNN. By looping through all classes, the empirical covariance matrix of the training set is calculated.

$$\hat{\Sigma} = \frac{1}{N} \sum_{c}^{C} \sum_{i:y_i=c}^{N} (f(\boldsymbol{x}_i) - \hat{\mu}_c)(f(\boldsymbol{x}_i) - \hat{\mu}_c)^\top \tag{8}$$

After the empirical means for all classes and the covariance matrix for the training set are computed, we have obtained the two crucial components that collectively describe the distribution of the training set so that a new incoming sample that is two many standard deviations away from the this distribution is considered abnormal and should be prevented from being sent to the output layer of the classifier. Equation 9 [38] describes how to calculate the Mahalanobis distance of a previously unseen sample $\boldsymbol{x}$ to our 5-class CNN after the training phase of the CNN. Since there are multiple classes, we need to calculate the Mahalanobis distance for each class c based on its μc. After the calculation of the Mahalanobis distances for all the classes, we pick the class such that the Mahalanobis distance from the sample $\boldsymbol{x}$ is the smallest.

$$C_M(\boldsymbol{x}) = \underset{c}{\operatorname{argmin}}\, (f(\boldsymbol{x}) - \hat{\mu}_c)^\top \hat{\Sigma}^{-1} (f(\boldsymbol{x}) - \hat{\mu}_c) \tag{9}$$

After we obtain the Mahalanobis distance from x, the previously unseen sample that we want to classify, we check if its Mahalanobis distance is in a threshold obtained during the validation phase with direct experiment. If yes, the this sample is in the distribution of the training set, indicating that it is of happiness, anger, neutrality, sadness, disgust, or fear. Since it is in distribution, the activation of the penultimate layer when given x will be forwarded to the output layer of the 5-class CNN. If the sample is out of the distribution of the training set, then it is not of happiness, anger, neutrality, sadness, disgust and fear. Therefore, forwarding it to the output layer of our 5-class CNN will result in a wrong classification no matter what, so we do not forward the penultimate layer's activation of this sample to the output layer.

5.4 The Hierarchical Structure of Classifiers

As mentioned above, to further demonstrate the value of our solution, we compare it to a hierarchical set of 3 classifiers as a baseline. Hierarchical structures attempt to leverage the information given by the higher-level classifiers to reduce the complexity of the problem that lower-level classifiers need to solve [7]. Our hierarchical set of three classifiers is trained on $\mathcal{D}_{train}$, the same training set on which our solution is trained, for the purpose of a better comparison on the performance of the hierarchical structure and our solution. This baseline demonstrates how a false sense of accuracy can be achieved by only testing on targeted emotions when other emotions also exist. We intend to use the hierarchical structure to prove the hypothesis that a mood detection algorithm that achieves high accuracy on interested emotions - in our case, happiness, anger, neutrality and sadness, will not perform adequately in a scenario where confounding emotions are present.

The hierarchy baseline consists of 3 binary CNN classifiers: C_1, a classifier that separates happy and angry audio clips from the neutral and sad audio clips, C_2, the classifier that separates happy and angry audio clips, and C_3 the classifier that separates neutral and sad audio clips. The three classifiers are trained with the same batch size, which is 128, the same optimizer, adam, and the same learning rate (1e-4) and decay (0), and the same kernel size (3 small frames). All of them are trained in 1000 epochs with an early stop of 50 epochs.

The classifier that separates happy and angry audio clips from the neutral and sad audio clips, represented as C_1. This C_1 decides if a given audio input is in either of the two classes: (1) the input is of happy or angry speech; (2) the input is of neutral or sad speech. This classifier returns a vector of size 2, whereas the first value represents the score computed by the CNN that this input of emotions is happiness or anger and the second value, the score for neutrality or sadness. Based on these scores, the audio input is passed to C_2 or C_3. C_1 is trained on samples of happiness, anger, neutrality, and sadness in $\mathcal{D}_{train}$.

The classifier that classifies happy and angry audio clips, represented as C_2. If C_1 decides that a given input is either a happy speech or an angry speech, that input is passed to this classifier which further determines if the speaker that produced the audio clip is happy or angry. The output is a vector of size 2, whereas the first value represents the score that the input is a happy speech and the second value represents the score that the input is an angry speech. C_2 is trained on samples of happiness and anger in D_{train}.

The classifier that classifies neutral and sad audio clips, represented as C_3. If C_1 decides that a given input is either a neutral speech or a sad speech, the input is passed to this classifier which further determines if the speaker that produced the audio clip was neutral or sad. The output is a vector of size 2, whereas the first value represents the score that the input is a neutral speech and the second value represents the score that the input is sad speech. C_3 is trained on samples of neutrality and sadness in D_{train}.

6 EVALUATION

After training, the evaluation is conducted on $\mathcal{D}_{test}$ for both our 5-class CNN solution and our baseline, the hierarchical classifier.

6.1 Evaluation to Show the Necessity of Adding Environmental Distortions to Training Samples

We place an emphasis on the importance/necessity of adding reverberation, background noise, and deamplification effect into samples in the training set. Before we start evaluating our solution, the 5-class classifier, and the baseline we built, which is the hierarchical classifier, we present experimental results to demonstrate the importance/necessity of adding environmental distortions; see Table 7.

In this experiment, we have trained a classifier using *only clean samples* of the five classes (happiness, anger, neutrality, sadness, fear/disgust) from the synthetic training dataset. For this classifier, we evaluate it in two ways. First, we evaluate it on samples that are not environmentally distorted. These clean samples are the samples from the testing set that have not been environmental distorted in any way. Second, we evaluate it on a subset of a real-life dataset, VoxCeleb.

VoxCeleb [45] contains utterances extracted from YouTube videos in which celebrities give talks or attend interviews. We evaluate the clean classifier on a subset of VoxCeleb. Voxceleb is a very imbalanced dataset, for there are only 216 angry samples in the entire testing set. Since we want to have a more balanced dataset, the subset of VoxCeleb we choose consists of: 500 happy samples, 216 angry samples, 500 neutral samples, 500 sad

Dataset	f1 score
Clean samples in testin set	73.84%
VoxCeleb	17.51%

Table 7. The left column consists of the two datasets that the classifier trained on only the clean samples is evaluated on. The difference in the f1 score demonstrates the necessity of adding environmental distortions. The metric is f1 score.

Classifier	Happy	Angry	Neutral	Sad	Overall
C_1	99.7%	99.6%	100%	100%	99.8%
C_2	95.5%	95.0%			95.3%
C_3			92.9%	96.1%	94.6%

Table 8. The hierarchical structure evaluated on $(\mathcal{D}_{test} \setminus C) \cap \mathcal{D}_1$ that contains only samples of interested emotions that are not environmentally distorted. On the entire testing set, the hierarchical structure achieves an accuracy of 94.7%. The metric is accuracy.

samples, and 500 samples that are not of the happy, angry, neutral, and sad emotions As seen in Table 7, the clean classifier yields an f1 score of 73.84% on the not environmentally distorted samples, but it drops to 17.51% on the subset of the real-life dataset. The decrease in the performance is by 56.33%. This demonstrates that the classifier trained on clean speech does not maintain the same level of performance on real-life dataset. **As a result, adding environmental distortions are necessary.**

6.2 Evaluation on the Hierarchical Structure of CNN's, the Baseline

The purpose of the hierarchical structure solution is to serve as the baseline, as stated before. It's performance is superior to the state-of-the-art on clean samples without non-targeted emotions. This baseline achieves an accuracy of 94.7%, and it outperforms 5 state-of-art algorithms on mood detection (see Table 2) on clean samples that do not include non-targeted samples.

We now show that the hierarchy of classifiers, only trained on happy, angry, neutral, and sad examples, are able to achieve very high accuracy when only evaluated on these emotions. However, it will always make mistakes if the input is of a confounding emotion, such as disgust. This is the case of many state-of-the-art classifiers on emotion detection; they are able to perform accurately only on certain emotions, but they are not expected to be as accurate as they are in environments where different kinds of emotions are omnipresent. In addition, by evaluating the hierarchical classifier, we have further confirmed that environmental distortions affect the classifier's performance.

6.2.1 The hierarchical structure evaluated on $(\mathcal{D}_{test} \setminus C) \cap \mathcal{D}_1$. Table 8 shows the performance of the hierarchical structure evaluated on $(\mathcal{D}_{test} \setminus C) \cap \mathcal{D}_1$. Recall that $\mathcal{D}_{test}$ is the testing set, C is the set of every samples of confounding emotions, and $\mathcal{D}_1$ is the set of samples that are not distorted. Thus $(\mathcal{D}_{test} \setminus C) \cap \mathcal{D}_1$ is the subset of the testing set that contains only happy, angry, neutral, and sad samples that are not distorted. The three CNN classifiers in the hierarchy achieve 99.8%, 95.3%, and 94.6% of accuracy respectively, which suggests that, without environmental distortions, the hierarchy can achieve a very high level of accuracy, as many state-of-the-art algorithms on mood detection do.

Classifier	Happy	Angry	Neutral	Sad	Overall
C_1	95.9%	98.4%	98.4%	98.5%	98.2%
C_2	95.4%	94.3%			94.8%
C_3			92.2%	92.8%	92.5%

Table 9. The hierarchical structure evaluated on $(\mathcal{D}_{test} \setminus C) \cap \mathcal{D}_2$ that consists of only samples of interested emotions that are de-amplified and mixed with noise. The metric is accuracy.

Classifier	Happy	Angry	Neutral	Sad	Overall
C_1	99.4%	99.3%	99.6%	99.6%	99.5%
C_2	88.9%	89.2%			89.1%
C_3			86.1%	93.2%	89.6%

Table 10. The hierarchical structure evaluated on $(\mathcal{D}_{test} \setminus C) \cap \mathcal{D}_3$ that contains only samples of interested emotions that are reverberated. The metric is accuracy.

6.2.2 The hierarchical structure evaluated on $(\mathcal{D}_{test} \setminus C) \cap \mathcal{D}_2$. $(\mathcal{D}_{test} \setminus C) \cap \mathcal{D}_2$ is the subset of the testing set that contains only happy, angry, neutral, and sad samples that are de-amplified and then mixed with background noise. Table 9 describes the performance of C_1, C_2, and C_3 when evaluated only on samples of interested emotions that are de-amplified and contaminated with background noise. The three CNN classifiers in the hierarchy achieve 98.2%, 94.8%, and 92.5% of accuracy respectively. The accuracy of the three classifier drops by 1.6%, 0.5%, and 2.1% when compared to their performance on samples of targeted emotions that are not distorted, demonstrating that de-amplification and noise have impact on the performance of the hierarchy but the hierarchical structure is still able to compete with state-of-the-art mood detection algorithms in terms of accuracy.

6.2.3 The hierarchical structure evaluated on $(\mathcal{D}_{test} \setminus C) \cap \mathcal{D}_3$. Table 10 shows the hierarchy's performance in terms of weighted accuracy on the samples of interested emotions that are reverberated. Compared to their performance on samples of interested emotions that are not environmentally distorted, the accuracy of the three classifiers all drop: C_1's accuracy drops to 99.5% by 0.3%, C_2's accuracy drops to 89.1% by 6.2%, C_3's accuracy drops to 89.6% by 5%. Reverberation distorts the samples, and the drop of accuracy is expected. The fact that the drop in accuracy is insignificant illustrates that the hierarchical structure is robust to reverberation.

Compared to their performance when evaluated on samples of interested emotions that are distorted by de-amplification and background noise shown in Table 9, C_1 is more resistant to reverberation (drop by only 0.3% on reverberated sample), while C_2 and C_3 are more resistant to de-amplification and background noise (drop by 0.5% and 2.1% on samples that are de-amplified and mixed with background noise), as indicated in Table9 and Table 10.

6.2.4 The hierarchical structure evaluated on $(\mathcal{D}_{test} \setminus C)$. Table 11 illustrates the performance of the classifiers evaluated on the samples of interested emotions in $\mathcal{D}_{train}$. C_1 achieves an accuracy of 97.6 %. C_2 achieves an accuracy of 91.0%. C_3 achieves an accuracy of 90.3%. Since $\mathcal{D}_{test} \setminus C$ is the set of samples of interested emotions that are either not environmentally distorted or environmentally distorted by the myriad combinations of the de-amplification amount measured in decibels, various segment of background noise collected from real home environments, and the three reverberation factors, this set is descriptive of speech samples of interested emotions that will take place in home environments. The high accuracy of the three classifiers demonstrate that the

Classifier	Happy	Angry	Neutral	Sad	Overall
C_1	95.9%	98.1%	98.4%	98.1%	97.6%
C_2	92.3%	89.8%			91.0%
C_3			88.4%	91.9%	90.3%

Table 11. The hierarchical structure evaluated on $\mathcal{D}_{test} \setminus C$, that contains all samples of interested emotions. The metric is accuracy.

Classifier	Happy	Angry	Neutral	Sad	Confounding	Overall
C_1	95.9%	98.1%	98.4%	98.1%	0%	83.3%
C_2	92.3%	89.8%			0%	68.4%
C_3			91.9%	90.3%	0%	66.6%

Table 12. The hierarchical structure evaluated on $\mathcal{D}_{test}$, the entire testing set that contains both samples of interested and confounding emotions. On the entire testing set, the hierarchical structure achieves an accuracy of 56.2%. The metric is accuracy.

hierarchical structure on emotion detection of the interested emotions (happiness, anger, neutrality, and sadness) is robust to environmental distortions in home environments.

6.2.5 The hierarchical structure evaluated on $\mathcal{D}_{test}$. Table 12 shows the performance of the three classifiers of the hierarchy when tested on four of the emotions of interest and the confounding emotion. Table 12 shows the disadvantage of a common theme of emotion detection classifiers: emotions outside those of interest are assumed to not exist. However, many other emotions may exist in the real world and, thereby, reduce the overall accuracy of classifiers solely trained on the emotions of interest.

In Table 12, the three classifiers still achieve the exact same performance on the interested emotions, happiness, anger, neutrality, and sadness as Table 11. However, since confounding emotions are introduced and the hierarchical structure must classify the samples of confounding emotions (fear and disgust) as one of the interested emotions, it is bound to make mistakes. As a result, C_1, C_2, and C_3 achieves an accuracy of 0% on confounding samples. This result in a significantly decrease of the their overall performance measured in accuracy: C_1's accuracy drops from 97.6% to 83.3%, C_2's accuracy from 91.0% to 68.4%, and C_3's accuracy from 90.3% to 66.6%. The deterioration of the performance of the three classifiers in the hierarchy suggests that classifiers that are trained to achieve very high accuracy on only interested emotions will not perform adequately when given samples of confounding emotions.

6.3 Evaluation on the 5-class classifier Without Out-Of-Distribution Samples

Next, we evaluate our solution, the 5-class CNN, to understand its performance on recognizing each emotion with and without environmental distortions.

Tables 13 and 14 show the evaluation of the 5-class classifier in four different scenarios that account for environmental distortions. The first column is the evaluation on this classifier on samples in the testing set that are padded, but no environmental distortion is introduced to the sound. The second column is the evaluation on this classifier on padded samples that are reverberated. The third column is the evaluation on this classifier only on padded samples that are de-amplified and mixed with noise. The last column is the evaluation on this classifier

Class	$\mathcal{D}_{test,1}$	$\mathcal{D}_{test,2}$	$\mathcal{D}_{test,3}$	$\mathcal{D}_{test}$
Happy	92.6%	92.2%	80.2%	89.6%
Angry	92.3%	92.4%	84.8%	88.9%
Neutral	95.9%	91.3%	90.7%	88.7%
Sad	94.9%	88.6%	93.8%	89.1%
Confounding	88.9%	79.0%	78.4%	81.4%
Mean	92.9%	88.4%	85.6%	88.0%

Table 13. Evaluation on the 5-class classifier that categorizes its input into categories: Happiness, Anger, Neutrality, Sadness, and Confounding emotions.The metrics for evaluation is accuracy. The average is weighted.

	$\mathcal{D}_{test,1}$	$\mathcal{D}_{test,2}$	$\mathcal{D}_{test,3}$	$\mathcal{D}_{test}$
f1 score	93.2%	86.1%	87.6%	87.9%

Table 14. Evaluation on the 5-class classifier on different subsets of the testing set. **The metric for evaluation is f1 score**. The average is weighted. Since our synthetic dataset is very well balanced, we can see that the f1 scores are very similar to the scores in Table 13. The metric is f1 score.

on the entirety of the testing set. The samples used for evaluation in Table 13 are denoted as $\mathcal{D}_{test,k}, k = 1, 2, 3$, defined as Equation 10.

$$\mathcal{D}_{test,k} = \mathcal{D}_{test} \cap \mathcal{D}_k, k = 1, 2, 3 \tag{10}$$

The average accuracy of the 5-class CNN drops from 92.9% when evaluated on $\mathcal{D}_{test,1}$, the set that resembles the idealistic scenario in which the speaker is close to the microphone and the room reverberation and background noise have minimal impact on the quality of the sound signal captured by the microphone, to 85.6% when reverberation is introduced, and to 88.4%, when de-amplification and background noise are introduced. This indicates that environmental distortions such as reverberation, background noise, and de-amplification still affect the performance of the 5-class CNN - which is as expected, since the signal is distorted, the classification result of any classifier is expected to deteriorate slightly, moderately, or severely depending on the robustness of the classifier. Evaluated on $\mathcal{D}_{test}$, the weighted average of accuracy of our classifier is 88.0%, which is only a 4.9% drop of accuracy from 92.9% obtained from the evaluation on the ideal $\mathcal{D}_{test,1}$ dataset. In other words, our 5-class CNN deteriorates only by 4.9%, despite the various combinations of environmental distortions that are collectively illustrated by the reverberation factors (the wet/dry ratio, decay factor, and diffusion), the number of value measured in decibels deduced from the amplitude, and the various background noise from home environments.

The previous paragraph discusses the overall performance of the classifier in $\mathcal{D}_{test,i}, i = \{1, 2, 3\}$ and $\mathcal{D}_{test}$, but we can also observe how the classifier performs in the four scenarios on each emotion. Its accuracy on happiness drops from 92.6% on happy samples in $\mathcal{D}_{test,1}$ to 80.2% on happy samples in $\mathcal{D}_{test,2}$ when reverberation is introduced; meanwhile, the accuracy on happy samples in $\mathcal{D}_{test,1}$ drops only 0.4% when compared to its accuracy on happy samples in $\mathcal{D}_{test,3}$. The difference between the two drops in accuracy suggests that reverberation is harder to deal with by our classifier than noise and de-amplification for happiness. Similarly, when classifying angry samples, the drop from the accuracy achieved on angry samples in $\mathcal{D}_{test,1}$ to the accuracy achieved on angry samples in $\mathcal{D}_{test,2}$ is 7.5%, while the accuracy on angry samples that are distorted and de-amplified actually increases 0.1% compared to the accuracy achieved on angry samples that are not distorted at all. The classifier's performance on happy and angry samples suggests that noise and de-amplification have minimal influence on our classifier's performance when compared to reverberation.

On neutral and sad samples, the observation that noise and de-amplification have less influence on the classifier's performance than reverberation no longer holds. Compared to the classifier's performance on neutral samples that are not altered, the accuracy drops from 95.9 % to 90.7% by 5.2% and 91.3% by 4.6%, respectfully on reverberated neutral samples and neutral samples that are de-amplified and contaminated with noise. Reverberation and amplification with noise result in similar deterioration of the classifier's accuracy on neutral samples. On sad samples, reverberation results in a drop in accuracy from 94.8% to 93.8% caused by reverberation, and to 88.6% caused by de-amplification and noise. The observation on the classifier's performance on distorted sad samples is that reverberation has minimal influence on the classifier's performance when compared to de-amplification and noise.

On confounding samples, the classifier achieves an accuracy of 88.9% on clean samples, but it drops to 78.4% by 10.5% and 79.0% by 9.9% on reverberated samples and samples that are de-amplified and contaminated with noise. For happy, angry, neutral, and sad samples, the drop of accuracy obtained from clean samples to distorted samples is always less than 5%, which is also expected, because the set of confounding emotions C is more complex than the sets $\mathcal{H}, \mathcal{A}, \mathcal{N}, \mathcal{S}$. C consists of more samples of more than one emotions, while the others consists of samples of only one emotion.

Note that the entirety of the testing set $\mathcal{D}_{test}$ is descriptive of the actual environment in which the classifier is envisioned to be deployed, because it encompasses a variety of different combinations of the environmental distortions illustrated by different factors that range from having no effect on the acoustic signals to significantly distorting the acoustic signals.

Our solution reaches an accuracy of 92.9% on $\mathcal{D}_{test,1}$, the set of samples that are not environmentally distorted, for all emotions, and it achieves 88.0% high accuracy in $\mathcal{D}_{test}$, for all emotions. The environmental distortions only reduce the overall accuracy of our solution by 4.9%. Therefore, our solution is highly robust to environmental distortions.

The evaluation of our solution described in Table 13 illustrates the pattern that noise and de-amplification have less impact on the performance of our solution than reverberation for happy, angry, neutral, and confounding emotions. However, sad utterances do not following the same pattern, as noise and de-amplification decrease our solution's performance by 6.3% while reverberation decreases our solution's performance by only 1.1%. This is because the original sad utterances have lower volume (measured in decibels) compared to the other emotions. During the process of adding environmental distortions, we subject all utterances to the same standards (for example, the same range measured in decibels of possible de-amplification). As a result, the sad utterances, whose volumes are lower than the other emotions, are more susceptible to de-amplification.

6.4 5-class classifiers with out-of-distribution detection technique versus 4-class classifiers with out-of-distribution detection technique

We have stated that, out of the 5 classes of our solution, there is one class that we are not interested in. The aforementioned performance that the 5-class classifier achieved with out-of-distribution detection technique prompts us to ask the question - what if we train a classifier only on the four interested classes and use the out-of-distribution detection technique to intercept emotional utterances of other classes? In other words, we investigate the value of having a confounding class. In the following paragraphs, we answer the question:

With the out-of-distribution technique and samples of classes that we are not interested in, which approach is better: (1) should we include those samples during training, or (2) should we exclude them from training and let the out-of-distribution algorithm intercept samples from this "confounding"/uninterested class during training?

	without OOD	with OOD
f1 score	65.86%	76.66%

Table 15. Evaluation on the 4-class classifier on the combined set of the testing set and the set of the Calm emotion. Samples of the calmness class and the confounding class are considered out-of-distribution, since the 4-class classifier is trained on and can only assign any given input to the four targeted classes. The metric is f1 score.

To do so, we train a 4-class classifier (the four classes being happiness, anger, neutrality, sadness). To make it comparable to the 5-class classifier, the 4-class classifier is required to achieve the similar level of performance on testing samples of those four emotions as the 5-class classifier on the testing samples of five emotions (happiness, anger, neutrality, sadness, and confounding). Via direct experiment, we have obtained a 4-class classifier that achieves an f1 score of 87.18% on all the samples of the 4 classes in the testing set. It achieves a similar level of performance the 5-class classifier achieves (with an f1 score of 87.9%).

Having obtained a 4-class classifier, we evaluate it the same way we evaluate the 5-class classifier - first, just itself without out-out-distribution detection technique; second, this 4-class classifier paired up with the out-of-distribution detection technique.

Without the out-of-distribution detection technique, the 4-class classifier's performance drops from an f1 score of 87.9% to 65.86%, by 22.04%, as every sample of the calmness class and the confounding class is predicted to be of the happy, anger, neutrality, and sadness class. As a result, the predictions on calm samples and samples of the confounding class are always wrong.

With the out-of-detection technique to intercept samples that are potentially of classes unknown by the 4-class classifier, the f1 score improves to 76.66%, by 10.8% compared to its performance without the out-of-distribution detection technique. However, it is still 11.24% lower than the 4-class classifier's performance on testing samples from its 4 targeted classes.

The fact that the out-of-distribution detection technique improves the performance significantly by 10.8% demonstrates that it is still advantageous to pair the classifier with the out-of-distribution detection technique.

However, in Section 6.6, we demonstrate that, with the out-of-distribution detection technique, the 5-class classifier's performance (an f1 score of 87.71%) on its targeted classes and a previously unseen class is almost identical to its performance (an f1 score of 87.9%) without the out-of-distribution detection technique. The same improvement is not observed when we pair the 4-class classifier with out-of-distribution detection technique. The less significant improvement (by 10.8%) suggests that the samples in the confounding class are very similar to one or more of the 4 targeted classes. As a result, the out-of-distribution detection technique is not effective at picking them out and allows them to be passed to the 4-class classifier.

As a result, we conclude that **With the out-of-distribution technique and samples of classes that we are not interested in, we should include those samples during training**: there is always a possibility that the samples of the uninterested classes resemble one or more of the samples of the interested classes. In this case, the out-of-distribution is not effectively at distinguishing samples of interested classes from samples from uninterested classes, as **the vector representations of samples of interested classes can be within the distribution of the vector representations of samples of interested classes**.

6.5 Evaluation on the 5-class classifier With Out-Of-Distribution Samples

We now evaluate our full solution by including the detection of samples that are out of the distribution of the training set. Tables 12 and 14 show that our classifier is able to achieve an accuracy score of 88.0% and an f1 score of 87.9% on the testing set, which contains not only clean samples but also samples that are environmentally

	without OOD	with OOD
f1 score	77.78%	86.64%

Table 16. Evaluation on the 5-class classifier on the combined set of the testing set and the set of the Calm emotion, which is not one of the 5 classes and therefore considered out-of-distribution (OOD). The metric is f1 score.

distorted. We have achieved these scores using the testing set, which also has five classes. Despite that we have 4 interested emotion classes and 1 uninterested class, we must acknowledge the possibility that the classifier encounters samples that are not of the 5 classes. Since the classifier can only assign a class out of the five, its prediction is always wrong because the true label is not among the five.

We hypothesize that many samples not belonging to our recognized 5 classes are out of distribution of our training set. Therefore, if we intercept the out-of-distribution samples, we can significantly improve the performance of our classifier in a more realistic setting.

Table 16 shows the evaluation result of the 5-class solution on the combined set of the testing set and the set of the Calm emotion, which is not one of the 5 classes and therefore considered out-of-distribution(OOD) with and without the out-of-distribution detection technique. This combined set has 9995 samples. All the samples in the testing set are among those samples. The newly added samples are all of the Calm emotion. The Calm samples have been preprocessed in the same way as the samples in the synthetic dataset: we have clean Calm samples and environmentally reverberated Calm samples.

Without the out-of-distribution detection technique, we have achieved an f1 score of 77.78%. This is a significant drop (10.12%) from the f1 score achieved only on the testing set that has 5 targeted emotions (87.9%). The drop is to be expected, since the classifier can only assign the Calm samples to be happiness, anger, neutrality, sadness, and confounding emotions (fear and disgust), its prediction on a Calm samples is always wrong.

With the out-of-distribution detection technique, the f1 score improves to 86.64%, a significant increase (8.86%) compared to the experiment result in which no out-of-distribution detection technique is used and consequently all the out-of-distribution samples are wrongly assigned by the classifier. Note that the f1 score (86.64%) achieved on five targeted emotions and one previously unseen emotion class with the out of distribution technique is very close to the f1 score achieved on the five targeted emotions (87.9%) only. This indicates that, with the out-of-distribution detection technique, the classifier can maintain its level of good performance on sets that have out-of-distribution samples, compared to its performance on sets that only have samples that are of one of the classifier's targeted classes.

6.6 Real Time Computation

Our solution is currently part of a home Patient-Caregiver monitoring, modeling, and interactive recommendation system for caregivers of dementia patients [26], titled the Patient-Caregiver Relationship (PCR) system. One of the objectives of this system is to detect the onset of anger of the caregiver using only the acoustical modality (which our 5-class solution is a part of) and immediately notify the recommender system which will recommend mindfulness/relaxation techniques intelligently. In the PCR system, we use an external microphone and an off-the-shelf laptop to detect emotions with our 5-class solution with OOD.

Real-time applications such as PCR on requires an emotion detection algorithm to notify the caregiver promptly when their difficult emotions are expressed in their own speech. The notification would not be helpful to the caregiver if the notification is sent too late. Since the PCR problem is time-sensitive, it is imperative that the emotion classifier is capable of real-time computation.

The obvious computation architecture to choose is a standard architecture for smart home speakers such as Google Home speaker. However, smart home speakers do not perform the classification themselves; they only serve as acoustic sensors that stream data. The actual classification is handled using cloud computing: the acoustic signals are sent to a powerful cloud computing server from which the smart home speakers receive classification results. In other words, even in the case of smart home speakers, the acoustic signals are also processed by a powerful computing architecture. By using a standard PC architecture, we demonstrate that an off-the-shelf PC is able to run our CNN in real-time. There are also small form factor processors such as Toshiba's dynaEdge DE-100 [1] that have the same capabilities as a laptop, but without a screen and our solution can execute on one of these when we want the deployment equipment to take up less space than a laptop.

The average running time of the components of our 5-class CNN on 184 samples of 5-second duration on a Intel(R) Core(TM) i5-9300H CPU at 2.40 GHz is 2.62 seconds. The average time consists of the time required to perform feature extraction, the time required to perform out-of-distribution detection, and the time to perform classification. If the caregiver's speech starts to become angry or sad, it will take 5-seconds for their speech to be recorded, and 2.62 second for our 5-class CNN to generate the classification result. Within the amount of time required to compute the emotion of the caregiver based on their speech, the caregiver will likely still be under the influence of the emotion; thus a notification to them is more likely to help than a notification that is sent to them long after they are no longer under the influence of the emotion. In other words, our 5-class CNN is capable of real-time computation, which is a necessary condition that an emotion detection algorithm must meet in order to be used in applications like PCR.

7 CONCLUSION

Emotional health is a crucial part of one's well-being. The rapid development of machine learning in the field of acoustic signal processing has resulted in a surge of interest in detecting emotions from speech. We have created a combined convolutional neural network (CNN) and out of data distribution (OOD) solution that is robust to environmental distortions such as reverberation, noise, distance, and handles emotions that are not the targeted emotions through a class we call confounding emotions. To test our solution we created synthetic datasets that combined five standard datasets and enhanced them with de-amplification, home noises, reverberation, and a real-world padding scheme. Our solution outperforms a state of art baseline and achieves high accuracy in the presence of environmental distortions and confounding emotions.

REFERENCES

[1] [n.d.]. dynaEdge DE-100. https://asia.dynabook.com/laptop/dynaedge-de100/overview.php

[2] Starlet Ben Alex, Ben P Babu, and Leena Mary. 2018. Utterance and Syllable Level Prosodic Features for Automatic Emotion Recognition. In *2018 IEEE Recent Advances in Intelligent Computational Systems (RAICS)*. IEEE, 31–35.

[3] Halis Altun and Gökhan Polat. 2007. New frameworks to boost feature selection algorithms in emotion detection for improved human-computer interaction. In *International Symposium on Brain, Vision, and Artificial Intelligence*. Springer, 533–541.

[4] Kiavash Bahreini, Rob Nadolski, and Wim Westera. 2016. Towards real-time speech emotion recognition for affective e-learning. *Education and information technologies* 21, 5 (2016), 1367–1386.

[5] Pablo Barros and Stefan Wermter. 2016. Developing crossmodal expression recognition based on a deep neural model. *Adaptive behavior* 24, 5 (2016), 373–396.

[6] Rory Beard, Ritwik Das, Raymond WM Ng, PG Keerthana Gopalakrishnan, Luka Eerens, Pawel Swietojanski, and Ondrej Miksik. 2018. Multi-Modal Sequence Fusion via Recursive Attention for Emotion Recognition. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*. 251–259.

[7] Paul N Bennett and Nam Nguyen. 2009. Refined experts: improving classification in large taxonomies. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*. ACM, 11–18.

[8] Felix Burkhardt, Astrid Paeschke, Miriam Rolfes, Walter F Sendlmeier, and Benjamin Weiss. 2005. A database of German emotional speech. In *Ninth European Conference on Speech Communication and Technology*.

[9] Ru Ying Cai, Amanda L Richdale, Cheryl Dissanayake, and Mirko Uljarević. 2018. Brief report: Inter-relationship between emotion regulation, intolerance of uncertainty, anxiety, and depression in youth with autism spectrum disorder. *Journal of autism and developmental disorders* 48, 1 (2018), 316–325.

[10] Houwei Cao, David G Cooper, Michael K Keutmann, Ruben C Gur, Ani Nenkova, and Ragini Verma. 2014. CREMA-D: Crowd-sourced emotional multimodal actors dataset. *IEEE transactions on affective computing* 5, 4 (2014), 377–390.

[11] José Carlos Castillo, Álvaro Castro-González, Fernándo Alonso-Martín, Antonio Fernández-Caballero, and Miguel Ángel Salichs. 2018. Emotion detection and regulation from personal assistant robot in smart environment. In *Personal assistants: Emerging computational technologies*. Springer, 179–195.

[12] Ling Cen, Fei Wu, Zhu Liang Yu, and Fengye Hu. 2016. A real-time speech emotion recognition system and its application in online learning. In *Emotions, technology, design, and learning*. Elsevier, 27–46.

[13] Rajdeep Chatterjee, Saptarshi Mazumdar, R Simon Sherratt, Rohit Halder, Tanmoy Maitra, and Debasis Giri. 2021. Real-Time Speech Emotion Analysis for Smart Home Assistants. *IEEE Trans. Consumer Electron.* 67, 1 (2021), 68–76.

[14] Zeya Chen, Mohsin Y. Ahmed, Asif Salekin, and John A. Stankovic. 2019. ARASID: Artificial Reverberation-Adjusted Indoor Speaker Identification Dealing with Variable Distances. In *Proceedings of the 2019 International Conference on Embedded Wireless Systems and Networks (EWSN '19)*. Junction Publishing, USA, 154–165. http://dl.acm.org/citation.cfm?id=3324320.3324339

[15] Ming Cheng, Andrew Friesen, and Olalekan Adekola. 2019. Using emotion regulation to cope with challenges: a study of Chinese students in the United Kingdom. *Cambridge Journal of Education* 49, 2 (2019), 133–145.

[16] Akash Roy Choudhury, Anik Ghosh, Rahul Pandey, and Subhas Barman. 2018. Emotion Recognition from Speech Signals using Excitation Source and Spectral Features. In *2018 IEEE Applied Signal Processing Conference (ASPCON)*. IEEE, 257–261.

[17] Taner Danisman and Adil Alpkocak. 2008. Emotion classification of audio signals using ensemble of support vector machines. In *International Tutorial and Research Workshop on Perception and Interactive Technologies for Speech-Based Systems*. Springer, 205–216.

[18] Dragos Datcu and Léon JM Rothkrantz. 2005. Facial expression recognition with relevance vector machines. In *2005 IEEE International Conference on Multimedia and Expo*. IEEE, 193–196.

[19] Jun Deng, Xinzhou Xu, Zixing Zhang, Sascha Frühholz, and Björn Schuller. 2017. Universum autoencoder-based domain adaptation for speech emotion recognition. *IEEE Signal Processing Letters* 24, 4 (2017), 500–504.

[20] Robert F Dickerson, Enamul Hoque, Philip Asare, Shahriar Nirjon, and John A Stankovic. 2014. Resonate: reverberation environment simulation for improved classification of speech models. In *IPSN-14 Proceedings of the 13th International Symposium on Information Processing in Sensor Networks*. IEEE, 107–117.

[21] Kate Dupuis and M Kathleen Pichora-Fuller. 2010. *Toronto Emotional Speech Set (TESS)*. University of Toronto, Psychology Department.

[22] Florian Eyben, Klaus R Scherer, Björn W Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y Devillers, Julien Epps, Petri Laukka, Shrikanth S Narayanan, et al. 2015. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE Transactions on Affective Computing* 7, 2 (2015), 190–202.

[23] Haytham M Fayek, Margaret Lech, and Lawrence Cavedon. 2015. Towards real-time speech emotion recognition using deep neural networks. In *2015 9th international conference on signal processing and communication systems (ICSPCS)*. IEEE, 1–5.

[24] V Fernandes, L Mascarehnas, C Mendonca, A Johnson, and R Mishra. 2018. Speech Emotion Recognition using Mel Frequency Cepstral Coefficient and SVM Classifier. In *2018 International Conference on System Modeling & Advancement in Research Trends (SMART)*. IEEE, 200–204.

[25] Antonio Fernández-Caballero, Arturo Martínez-Rodrigo, José Manuel Pastor, José Carlos Castillo, Elena Lozano-Monasor, María T López, Roberto Zangróniz, José Miguel Latorre, and Alicia Fernández-Sotos. 2016. Smart environment architecture for emotion detection and regulation. *Journal of biomedical informatics* 64 (2016), 55–73.

[26] Ye Gao, Meiyi Ma, Kristina Gordon, Karen Rose, Hongning Wang, and John Stankovic. 2020. A monitoring, modeling, and interactive recommendation system for in-home caregivers: demo abstract. In *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*. 587–588.

[27] Joseph Gaugler, Bryan James, Tricia Johnson, Allison Marin, and Jennifer Weuve. 2019. 2019 Alzheimer's disease facts and figures. *Alzheimers & Dementia* 15, 3 (2019), 321–387.

[28] Esam Ghaleb, Mirela Popa, and Stylianos Asteriadis. 2019. Multimodal and Temporal Perception of Audio-visual Cues for Emotion Recognition. In *8th International Conference on Affective Computing & Intelligent Interaction (ACII 2019), Cambridge, United Kingdom.*

[29] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. The MIT Press.

[30] Jacek Grekow. 2018. Music Emotion Maps in the Arousal-Valence Space. In *From Content-based Music Emotion Recognition to Emotion Maps of Musical Pieces*. Springer, 95–106.

[31] James J Gross and Ricardo F Muñoz. 1995. Emotion regulation and mental health. *Clinical psychology: Science and practice* 2, 2 (1995), 151–164.

[32] James J Gross, Helen Uusberg, and Andero Uusberg. 2019. Mental illness and well-being: an affect regulation perspective. *World Psychiatry* 18, 2 (2019), 130–139.

[33] S. Haq and P.J.B. Jackson. 2010. *Machine Audition: Principles, Algorithms and Systems.* IGI Global, Hershey PA, Chapter Multimodal Emotion Recognition, 398–423.

[34] Dan Hendrycks and Kevin Gimpel. 2016. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136* (2016).

[35] Che-Wei Huang and Shrikanth Narayanan. 2018. Stochastic Shake-Shake Regularization for Affective Learning from Speech.. In *Interspeech.* 3658–3662.

[36] Amin Jalili, Sadid Sahami, Chong-Yung Chi, and Rassoul Amirfattahi. 2018. Speech Emotion Recognition Using Cyclostationary Spectral Analysis. In *2018 IEEE 28th International Workshop on Machine Learning for Signal Processing (MLSP).* IEEE, 1–6.

[37] Margaret Lech, Melissa Stolar, Christopher Best, and Robert Bolia. 2020. Real-time speech emotion recognition using a pre-trained image classification network: Effects of bandwidth reduction and companding. *Frontiers in Computer Science* 2 (2020), 14.

[38] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. 2018. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems.* 7167–7177.

[39] Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. 2017. Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv preprint arXiv:1706.02690* (2017).

[40] Steven R Livingstone and Frank A Russo. 2018. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PloS one* 13, 5 (2018), e0196391.

[41] Marko Lugger and Bin Yang. 2007. An incremental analysis of different feature groups in speaker independent emotion recognition. In *16th Int. congress of phonetic sciences.*

[42] Prasanta Chandra Mahalanobis. 1936. On the generalized distance in statistics. National Institute of Science of India.

[43] Leandro Y Mano. 2018. Emotional condition in the Health Smart Homes environment: emotion recognition using ensemble of classifiers. In *2018 Innovations in Intelligent Systems and Applications (INISTA).* IEEE, 1–8.

[44] Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen. 2016. TUT database for acoustic scene classification and sound event detection. In *2016 24th European Signal Processing Conference (EUSIPCO).* IEEE, 1128–1132.

[45] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. 2017. Voxceleb: a large-scale speaker identification dataset. *arXiv preprint arXiv:1706.08612* (2017).

[46] Fatemeh Noroozi, Marina Marjanovic, Angelina Njegus, Sergio Escalera, and Gholamreza Anbarjafari. 2017. Audio-visual emotion recognition in video clips. *IEEE Transactions on Affective Computing* 10, 1 (2017), 60–75.

[47] 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. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (2017), 96.

[48] Eugene Yu Shchetinin, Leonid A Sevastianov, Dmitry S Kulyabov, Edik A Ayrjan, and Anastasia V Demidova. 2020. Deep Neural Networks for Emotion Recognition. In *International Conference on Distributed Computer and Communication Networks.* Springer, 365–379.

[49] Melissa N Stolar, Margaret Lech, Robert S Bolia, and Michael Skinner. 2017. Real time speech emotion recognition using RGB image classification and transfer learning. In *2017 11th International Conference on Signal Processing and Communication Systems (ICSPCS).* IEEE, 1–8.

[50] Andreas Triantafyllopoulos, Gil Keren, Johannes Wagner, Ingmar Steiner, and Björn Schuller. 2019. Towards Robust Speech Emotion Recognition Using Deep Residual Networks for Speech Enhancement. *Proc. Interspeech 2019* (2019), 1691–1695.

[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 *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP).* IEEE, 5200–5204.

[52] N Vrebčević, I Mijić, and D Petrinović. 2019. Emotion Classification Based on Convolutional Neural Network Using Speech Data. In *2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO).* IEEE, 1007–1012.

[53] Kunxia Wang, Ning An, Bing Nan Li, Yanyong Zhang, and Lian Li. 2015. Speech emotion recognition using Fourier parameters. *IEEE Transactions on Affective Computing* 6, 1 (2015), 69–75.

[54] Lahiru Wijayasingha and John A Stankovic. 2021. Robustness to noise for speech emotion classification using CNNs and attention mechanisms. *Smart Health* 19 (2021), 100165.

[55] Adib Ashfaq A Zamil, Sajib Hasan, Showmik MD Jannatul Baki, Jawad MD Adam, and Isra Zaman. 2019. Emotion Detection from Speech Signals using Voting Mechanism on Classified Frames. In *2019 International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST).* IEEE, 281–285.

[56] Mingmin Zhao, Fadel Adib, and Dina Katabi. 2016. Emotion recognition using wireless signals. In *Proceedings of the 22nd Annual International Conference on Mobile Computing and Networking.* 95–108.