Real Time Distant Speech Emotion Recognition in Indoor Environments

Mohsin Y Ahmed, Zeya Chen, Emma Fass and John Stankovic
Department of Computer Science, University of Virginia
Charlottesville, VA, USA
{mohsin.ahmed,zeyachen,enf5cb,stankovic}@virginia.edu

ABSTRACT

We develop solutions to various challenges in different stages of the processing pipeline of a real time indoor distant speech emotion recognition system to reduce the discrepancy between training and test conditions for distant emotion recognition. We use a novel combination of distorted feature elimination, classifier optimization, several signal cleaning techniques and train classifiers with synthetic reverberation obtained from a room impulse response generator to improve performance in a variety of rooms with various source-to-microphone distances. Our comprehensive evaluation is based on a popular emotional corpus from the literature, two new customized datasets and a dataset made of YouTube videos. The two new datasets are the first ever distance aware emotional corpuses and we created them by 1) injecting room impulse responses collected in a variety of rooms with various source-to-microphone distances into a public emotional corpus; and by 2) re-recording the emotional corpus with microphones placed at different distances. The overall performance results show as much as 15.51% improvement in distant emotion detection over baselines, with a final emotion recognition accuracy ranging between 79.44%-95.89% for different rooms, acoustic configurations and source-to-microphone distances. We experimentally evaluate the CPU time of various system components and demonstrate the real time capability of our system.

CCS CONCEPTS

• **Computer systems organization** → **Sensors and actuators**; **Real-time system specification**; • **Computing methodologies** → *Machine learning*;

KEYWORDS

Emotion, speech, noise and reverberation

1 INTRODUCTION

Extracting emotional components from human speech (speech emotion recognition) in real time has been a challenging problem for several decades. Speech is the most common and natural communication medium in humans. Therefore, accurate real time speech emotion recognizers have far more potential for real world deployment centric applications because 1) speech has pervasive reachability to nearby sensors (microphone) as opposed to video/facial expression based emotion recognizers, and, 2) speech is less intrusive as opposed to galvanic skin resistance based emotion recognizers. A real time speech emotion recognizer will have profound impact on a wide range of applications if it can be accurate in a wide range of environments with different acoustic configurations and different source-to-microphone distances. If these challenges can be met, then the solution can be used in many applications requiring real time emotion recognition. For example, it can be used in an advanced driver assistance system to detect real time mood of a vehicle driver, as aggressive driving behavior may lead to accidents. Similarly, real time cockpit behavior for airline pilots can be monitored for possible depressive syndromes leading to suicidal tendencies. In general, people with suicidal tendencies can be monitored for mood to support just in time interventions. Certain medical conditions like heart diseases are likely to worsen due to anger and excitement, therefore such patients can be monitored in real time for emotional outbursts and subsequent interventions can possibly avoid heart attacks.

In a real time indoor speech emotion recognition system, the microphones are deployed in certain places of the room. These microphones capture speech signals originating from sources (human) 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. A related area of research is distant-speech-recognition (DSR) i.e. converting speech to text by distant microphones, which is an extension of the automatic-speech-recognition (ASR) problem, where a lot of progress has been made [7, 10, 14] in recent years. However, real time distant emotion recognition (RTDER) is an area not explored before to the best of our knowledge. It is important to note that, the solutions of the DSR problem are not generalizable to solve DER problem because of the difference in the nature of the core problems. DSR targets translating captured speech in distant microphones to text, while RTDER targets classifying captured speech into certain emotional classes in real time. Speech emotion recognition requires a large number of local and global acoustic features, a static or dynamic emotion training set and uses classifiers like support vector machines (SVM), Gaussian mixture models (GMM) and random forest whereas automatic speech recognition needs a limited number of features (MFCC) with hidden Markov models (HMM) and uses a different technique involving phonemes and language models.

Mobiquitous 2017, November 7–10, 2017, Melbourne, VIC, Australia

ACM ISBN 978-1-4503-5368-7/17/11…$15.00
https://doi.org/10.1145/3144457.3144503

A real time speech emotion recognition system is generally evaluated using one or more emotional speech database/corpuses. An emotional speech corpus is generally made from real-world incident recordings or from acted/elicited artificial emotional utterances in sound laboratories by professional/semi-professional/non-professional actors. A majority of the existing emotional speech databases are made artificially because of the legal and moral issues of using real life recordings for research purposes. An extensive list of state-of-the-art emotional corpuses can be found in [4], where the corpuses have been made by professional/nonprofessional actors and from extracted movie clips. A common characteristic of all the existing emotional corpuses is that all of them are made of clean speech recorded by closely situated microphones, often in a noise-proof anechoic sound studio. All the existing real time speech emotion recognition results are based on these clean speech recordings and, hence, these results are not applicable to a real world environment where acoustic sensors are likely to be situated far from the speakers. Therefore, no solution to the RTDER problem exist to date.

In this paper, for the first time, we address different stages of the processing pipeline of a real time speech emotion recognition system to solve the previously unexplored RTDER problem and increase real time emotion recognition accuracy in distant microphones in different room types. The main contributions of our work are:

- We identify several challenges in different stages of a real time distant speech emotion recognition pipeline and provide solutions to them with empirical results obtained over extensive evaluations.
- We create the very first distance aware emotional corpuses to use in our experiments by 1) re-recording a popular emotional corpus with a microphone array with microphones placed at various distances, and 2) by injecting room impulse responses collected in a variety of rooms with various source-to-microphone distances into the same emotional corpus. We plan to make these distance aware emotional corpuses freely available for research purpose.
- Our novel combination of distorted acoustic feature elimination, best feature selection and classifier optimization techniques improves the real time distant emotion detection accuracy between 1.31% (for the worst case scenario of a large church hall) to 6.12% from the baseline in a variety of rooms with various source-to-microphone distances. We perform the most comprehensive feature analaysis for RTDER using the largest known emotional feature set consisting of 6552 acoustic features. Note that, we considered loud background noisy environments and extremely large rooms with very high reverberation effects for possible worst-case situation analysis and achieved improvement from the baseline even in the worst scenarios.
- At the signal acquisition stage, we use 2 state-of-the-art dereverberation and denoising techniques to clean the distant emotional speech signal. In addition, we combine these approaches with our novel combination of distorted feature elimination, best feature selection and classifier optimization techniques to achieve up to 10.84% improvement from the baseline in a variety of rooms with various source-to-microphone distances, with the final classification accuracy ranging between 79.44%-94.95%.
- At the classifier training stage, we train our classifiers with synthetic reverberation obtained from a room impulse response generator to reduce the discrepancy between training and testing conditions in a RTDER environment. Our training approach only requires emotion samples with clean speech at a close microphone. In addition, we combine this with our novel feature and classifier enhancement techniques to obtain up to 15.51% improvement from baselines across all the rooms in a variety of distances, with the final accuracy ranging between 87.85%-95.89%.
- We evaluate the above mentioned techniques on a YouTube video dataset consisting of 37 clips spanning over 3 hours from lectures, public speech, talk-shows, and personal statements from both actors and real people and obtain a maximum of 7.30% of improvement in recognizing real world emotions at various source-to-microphone distances in different rooms, with a maximum accuracy of 93.68%.
- We experimentally evaluate the CPU runtime of each component of our system and demonstrate the real time capability of our system.

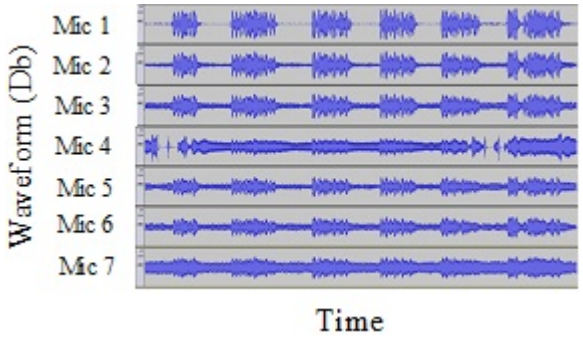

Figure 1: Raw waveforms of an angry utterance containing 6 separate sentences in 7 microphones situated at different distances.

2 PROBLEM FORMULATION: REAL TIME DISTANT EMOTION RECOGNITION

Speech based real time emotion detection is a complex problem due to the diversity in the way different people speak and express emotions, linguistics of different languages and accents, and the expression of a wide range of emotions by human. However, when speech is captured by distant microphones (as opposed to right next to the speaker), it adds further complexity to the real time emotion detection problem due to room reverberation, noise, and reduced signal-to-noise ratio.

In the past 2 decades, various emotional corpuses have been made by the affective computing community from clean emotional speech recorded in anechoic (non-reverberant) and noise-free sound studios simulated by professional or non-professional actors, and hence, all the existing emotion detection results are based on clean emotional recordings. However, for a realistic real time emotion

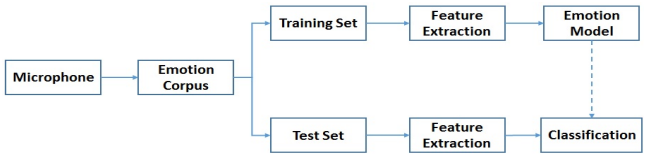

Figure 2: Stages of a standard acoustic emotion detection pipeline.

detection system deployed in open environments, one or more microphones will be situated at certain places of a room, and hence capture sound waves coming from distant sources (human subjects). We formally call this a *Real Time Distant Emotion Recognition* (RTDER) problem. As an example, we record 6 sample angry utterances from an emotional corpus in front of a microphone array consisting of 7 microphones situated at various distances, and Figure 1 shows the recorded waveforms. The waveform of microphone 1 demonstrates clear recording with highest signal-to-noise-ratio (SNR), where microphone 1 is situated nearest to the speaker. However, as the speaker to microphone distance increases, background noise and room reverberation are injected into the recordings, and hence the SNR decreases, as observed in recordings from microphones 2-7. This induced noise and reverberation drastically affect the emotion detection performance, as we demonstrate in later sections.

Figure 2 demonstrates a standard acoustic emotion detection pipeline. A clean speech emotional corpus is created by induced or acted emotional utterances from professional or non-professional actors. A training and test set is generated from the emotional corpus, and emotional models are generated from the training set after extracting emotion correlated acoustic features. Finally, the same features are extracted from the test set and the emotional models are applied on the test set to classify emotion. In the context of RTDER, the training set is obtained from a clean speech signal while the test set is obtained from distant speech; hence the challenge arises because of discrepancy in training and test data. In this paper, we address different stages of the acoustic emotion detection pipeline in context of RTDER with an objective of making training and testing conditions similar, and thus increase emotion classification accuracy. The challenges we address and solve are:

- **Challenge 1:** Can we find a set of emotion correlated acoustic features which are robust against microphone-to-speaker distance?
- **Challenge 2:** Can we clean the test speech recorded over distance from noise and room reverberation?
- **Challenge 3:** Can we add artificial room acoustic configuration into the clean speech training to reduce the discrepancy between training and test scenario?
- **Challenge 4:** Can we execute various system components of our solution in real time on standard available hardware of our target safety centric applications?

3 FINDING DISTANCE ROBUST FEATURES

When emotional speech signal is recorded by a distant microphone, the recorded signal becomes distorted compared to the original signal because of room ambient noise and reverberation. The amount of distortion depends on the acoustic properties of the room and amount of noise. For our solution, we empirically find a set of acoustic features which are robust across distance as well as correlated to target emotions. We then use these distance robust features for both training with clean emotional speech and testing on distant speech. Since these features are robust across distance, their distortion with distance is minimal, hence the discrepancy between training and testing in RTDER is also minimal, and accuracy is improved.

In our solution, we calculate the distortion of a particular feature f at distance d using the following formula:

$$distortion_d = \left| \frac{f_0 - f_d}{f_0} \right| \times 100\% \quad (1)$$

Where, f_0 = feature value for clean signal (distance 0),
f_d = feature value for signal at distance d.

For the rest of this section, we discuss the data preparation strategy for our experiments, feature extraction, distorted feature filtering, best feature selection and classifier optimization methods.

3.1 Data Preparation

We used the Berlin Emotional Speech Database, also known as Emo-DB [2], in our experiments to find distance robust emotional speech features. Emo-DB is a well-known and widely used freely available emotional corpus in the affective computing domain. It contains short sentences in German each spanning between 2-5 seconds in 7 different emotion categories: anger, anxiety, boredom, disgust, happiness, neutrality and sadness. There are 535 utterances in total in Emo-DB spanning these 7 emotions spoken by 10 professional actors (5 males and 5 females).

Just like most other emotional corpuses, Emo-DB contains only clear speech recordings. We apply the following 2 techniques to impose distance effect in Emo-DB recordings.

Re-record Emo-DB with a microphone array: We played the Emo-DB recordings with a loudspeaker and recorded them with a VocoPro UHF-8800 microphone array consisting of 4 microphones placed at distances of 1m, 3m, 5m and 7m from the loudspeaker. The recording was done in a 10m x 5m lab at 16 KHz sampling rate and 24-bit precision. There was loud background HVAC noise present while recording. Most indoor applications would fall within 7m of speaker-to-microphone distance hence we did not record for any further distance. To the best of our knowledge, this is the first emotional corpus recorded with different speaker-to-microphone distances, and we plan to make this dataset freely available to the research community. We refer this dataset as "Emo-DB-Array" for the remaining of the paper.

Inject distance effect into Emo-DB from AIR impulse response database: The Aachen Impulse Response (AIR) [6] database is a collection of room impulse responses (IR) measured in a variety of rooms with various acoustic configurations and with different source-to-microphone distances. A room impulse response can be used to describe the acoustic properties of a room in terms of sound propagation and reflections for a specific source-microphone configuration. The distant reverberated signal *s(n)* is represented as a convolution of the source (clean) signal *x(n)* with the room IR *r(n)*

$$s(n) = x(n) * r(n) \quad (2)$$

We convolved the Emo-DB recordings with room impulse responses obtained from the AIR database. Convolving the Emo-DB

Table 1: Room Configurations of IRs from AIR Database

Room	Dimensions	Speaker-to-Microphone Distances
Office room	5m x 6.4m x 2.9m	1m, 2m, 3m
Meeting room	8m x 5m x 3.1m	1.45m, 1.7m, 1.9m, 2.25m, 2.8m
Lecture room	10.8m x 1.09m x 3.15m	2.25m, 4m, 5.56m, 7.1m, 8.68m, 10.2m
Aula Carolina church	19m x 30m	1m, 2m, 3m, 5m, 15m, 20m

recordings with these IRs injects the acoustic and various distance effects of AIR database rooms into the Emo-DB database recordings. We refer this dataset as "Emo-DB-AIR" for the remaining of the paper. Table 1 summarizes the dimensions and speaker-to-microphone distances of various rooms from the AIR database whose IRs were convolved with Emo-DB recordings to construct Emo-DB-AIR.

3.2 Feature Extraction

We used the widely used OpenSMILE feature extraction toolkit [5] to extract a large number of 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. The 39 statistical functionals are applied to the LLDs computed from each of the emotional utterances to map a time series of variable length onto a static fixed size (6552) feature vector. Features were extracted on the utterance level, i.e., 1 feature vector per sentence. These 6552 features constitute the Emo-Large feature set of OpenSMILE toolkit, which is the largest emotion specific feature set known to date in terms of number of features. We chose the largest feature set because it would allow us to know more emotion correlated features which get distorted over distance, and hence would help to build a feature set robust to distance by keeping the distance agnostic emotion correlated features only. Such an approach to find distance robust emotional features has not been attempted before, to the best of our knowledge.

3.3 Iterative Distorted Feature Cut (IDFC) Procedure

For each of the 6552 features, distortion of each feature with respect to its corresponding clean signal feature value is calculated using equation 1, for both the Emo-DB-Array and Emo-DB-AIR datasets. The features are then sorted by their distortion value from highest to lowest, and iteratively discarded (cut) from the train and test sets one by one. In each step of the iteration, a new emotion model is built with the updated reduced-by-1 feature set, and tested upon the test set, and corresponding classification accuracy is logged. A support vector machine (SVM) classifier from the Weka data mining toolkit is used for training and testing, as SVM is reported to have best performance in emotion detection in prior works. The features are normalized to the range [-1, 1] before training and testing. This procedure iterates across all 6552 features and returns the best accuracy achieved across all iterations and the corresponding best feature cut.

3.4 Feature Selection and SVM Parameter Optimization

We chose a large feature set consisting of 6552 features so that we can identify the highest number of distance sensitive distorted

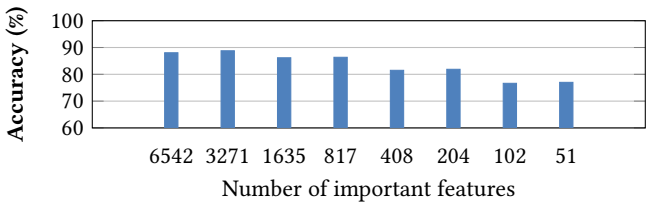

Figure 3: We picked the subset of most important features. Choosing the most important 3271 features yielded highest cross-validation accuracy of 88.78%.

features. Although the Emo-Large feature set is an emotion detection feature set, not all the 6552 features are equally important for the emotion detection task. Allowing a lot of less-correlated features overfits the classification model resulting in greater errors, in addition to increased latency for real time operation. Therefore, we used an algorithm presented in [3] which ranks the features by their importance to the classification problem by calculating their F-scores. The larger the F-score is, the more likely this feature is more discriminative. We also calculated optimized hyperparameters for the SVM using grid search, with the cost $c = 4$ and gamma $\gamma = 0.00195$.

For a 2-class (binary) classification problem, given training vectors $x_k, k = 1, ..., m$, if the number of positive and negative instances are n^+ and n^-, respectively, then the F-score of the i^{th} feature is defined as:

$$F(i) = \frac{(\bar{x_i}^{(+)} - \bar{x_i})^2 + (\bar{x_i}^{(-)} - \bar{x_i})^2}{\frac{1}{n_+ - 1}\sum_{k=1}^{n^+}(x_{k,i}^{(+)} - \bar{x_i}^{(+)})^2 + \frac{1}{n_- - 1}\sum_{k=1}^{n^-}(x_{k,i}^{(-)} - \bar{x_i}^{(-)})^2} \quad (3)$$

where $\bar{x_i}, \bar{x_i}^{(+)}, \bar{x_i}^{(-)}$ are the mean of the i^{th} feature of the whole, positive and negative data sets respectively, and $\bar{x_{k,i}}^{(+)}, \bar{x_{k,i}}^{(-)}$ are the i^{th} feature of the k^{th} positive and negative instance, respectively. The Emo-DB dataset has 7 different emotion labels; hence this is a multi-class classification problem, as opposed to binary classification. For multi-class classification, the algorithm constructs $C(k, 2) = \frac{k(k-1)}{2}$ binary classifiers between each possible pair of the original k classes. F-score based feature ranking is calculated for each of these binary classifiers, and finally it selects the same feature subset for every binary classifier to maximize the average accuracy over all classes.

Next, we sort all the features by their F-score importance, and evaluate if choosing a smaller subset of more important features improves the classification performance. We iteratively chose larger subsets of important features, and do a 10-fold cross validation on the training set. Figure 3 shows that choosing a subset of 3271 most important features yields highest cross-validation accuracy (88.78%) with the optimized SVM model ($c = 4, \gamma = 0.00195$) with a Radial-Basis kernel.

Finally, we again apply the iterative distorted feature cut procedure on the best 3271 features to eliminate the most distorted features among these best 3271 features to further increase the emotion detection accuracy, as seen from results in the next section.

3.5 Evaluations

We experimented with the IDFC, feature selection with F-score and SVM parameter optimization procedures on Emo-DB-AIR and Emo-DB-Array datasets. We set the baseline as when no feature

and classifier enhancements are done, and training is done on clean speech from Emo-DB and testing is done on noisy and reverberated speech from Emo-DB-AIR and Emo-DB-Array. On average, we get 2.15%, 2.93%, 2.09% and 1.31% classification improvement for Aula Carolina church, lecture, meeting and office rooms with a final average classification accuracy of 85.14%, 81.53%, 93.19%, and 90.97%, respectively, for the Emo-DB-AIR dataset, and 6.12% average improvement for the Emo-DB-Array dataset with a final average accuracy of 87%. Small distances (like 1m) in small rooms (meeting, office, lab) yields least improvement, as the signal gets little distorted with such small distance, hence the IDFC procedure is less effective. But in larger distances, the IDFC procedure accompanied with best feature subset and optimized SVM is effective in most cases.

While the average accuracy increase may seem low, we show in sections 4 and 5 that these distorted feature elimination, best feature selection and classifier optimization techniques, when accompanied by signal cleaning and training transformation techniques, yield accuracy improvement as much as 15.51% from the baseline.

4 CLEANING SIGNAL FROM REVERBERATION AND NOISE

As stated earlier, speech signals captured in distant microphones are infected with reverberation and noise. In this section, we address the signal acquisition stage of the pipeline presented in Figure 2. We use 2 state-of-the-art dereverberation and denoising techniques to clean the distant signal, as described below.

4.1 Dereverberation and Denoising Algorithms

4.1.1 Weighted-Prediction Error (WPE). WPE performs inverse filtering of room acoustics based on linear prediction. For each sample time t, the WPE method [13] linearly predicts the reverberation component contained in an observed speech sample, $x(t)$ from its preceding samples $x(u); u < t$. Let $y(t)$ be the distant speech signal at time t containing reverberation and background noise. Let $y_n[k]$ denote a short-time-Fourier-transform (STFT) coefficient calculated from $y(t)$, where n and k are the time frame and frequency bin indices, respectively. $y_n[k]$ is dereverberated at each frequency bin k using a linear filter as follows:

$$x_n[k] = y_n[k] - \sum_{\tau=T_\perp}^{T_T} g_\tau^*[k] y_{n-\tau}[k] \tag{4}$$

where * is the complex conjugate operator, and $T_\perp$ and T_T is the effective time period of the filter. We set $T_\perp = 3$ and $T_T = 50$ to deal with long-term reverberation. $G = (g_{T_\perp}, ..., g_{T_T})$ is a set of filter coefficients optimized to minimize the following objective function:

$$F_{WPE} = \sum_{n=1}^{N} \left(\frac{\left| y_n[k] - \sum_{\tau=T_\perp}^{T_T} g_\tau^*[k] y_{n-\tau}[k] \right|^2}{\theta_n} + \log \theta_n \right) \tag{5}$$

where N is the total number of time frames and $\theta = (\theta_1, ..., \theta_N)$ is a set of auxiliary variables optimized jointly with G. The optimized filter F_{WPE} is applied to $y_n[k]$ to generate dereverberated and denoised STFT coefficient $x_n[k]$.

4.1.2 Coherent-to-Diffuse Power Ratio Estimation (CDR). This method has been proposed by Schwarz and Kellermann [12] to clean the speech signal from reverberation and noise. This technique estimates the ratio between direct and diffuse (reverberation and noise) signal components, also called as coherent-to-diffuse power ratio (CDR), from the measured coherence between speech captured in two omnidirectional microphones. The CDR estimators are applied in a spectral subtraction postfilter for reverberation suppression.

We experimented with 3 different CDR estimators to suppress reverberation:

- Known direction of arrival (DOA) and noise coherence
- Unknown DOA
- Unknown noise coherence

DOA is the angle between the received sound wave axis and microphone axis. Sounds which propagate directly to microphone have a DOA of 0, but reverberated sound being reflected from room walls and objects have a non-zero DOA. The details of the CDR estimators are beyond the scope of this paper, and interested readers should refer to [12] for the details.

4.2 Results

We applied the CDR dereverberation technique on Emo-DB-AIR dataset and WPE dereverberation and denoising technique on both the Emo-DB-AIR and Emo-DB-Array datasets. The CDR algorithm requires having 2 omnidirectional microphones for coherence and CDR estimation. The AIR database is a binaural impulse response database, which means IRs were collected with 2 microphones, which justifies our use of Emo-DB-AIR dataset for CDR based dereverberation.

4.2.1 Emo-DB-AIR Evaluation. Figure 4 shows the comparative results of different dereverberation and denoising techniques on the Emo-DB-AIR dataset. For these evaluations, we set the baseline as when no signal cleaning is done, i.e. training on clean speech from Emo-DB and testing on noisy and reverberated speech from Emo-DB-AIR. The metric for these analyses is the percentage accuracy of correctly classified emotional utterances (total 535). For baseline, we use the SVM classifier with original 6552 features for classification.

The room dimensions of different rooms in the AIR dataset are provided in Table 1. From Figure 4, we can see that, for the Aula Carolina church which has the largest dimension and very high reverberation effect, none of the dereverberation and denoising technique can improve the baseline. For the other 3 rooms, the CDR with unknown noise coherence technique consistently improves from the baseline in most cases, the improvement ranging between 1 to 10 utterances. Best improvement was obtained in the office room, which has the smallest dimensions. As expected, the number of correctly classified utterances decreases in all rooms with increasing speaker-to-microphone distance.

The performance of the WPE algorithm in most cases was below the baseline and in some instances it severely degraded the performance. To investigate the issue, we listened to some of the emotional clips from EMO-DB-AIR dereverberated by the WPE algorithm. Our perception was that this algorithm distorts the original signal significantly as a side effect of dereverberation, which causes performance degradation. The performance of other 2 CDR estimation based techniques also ended up being sub baseline.

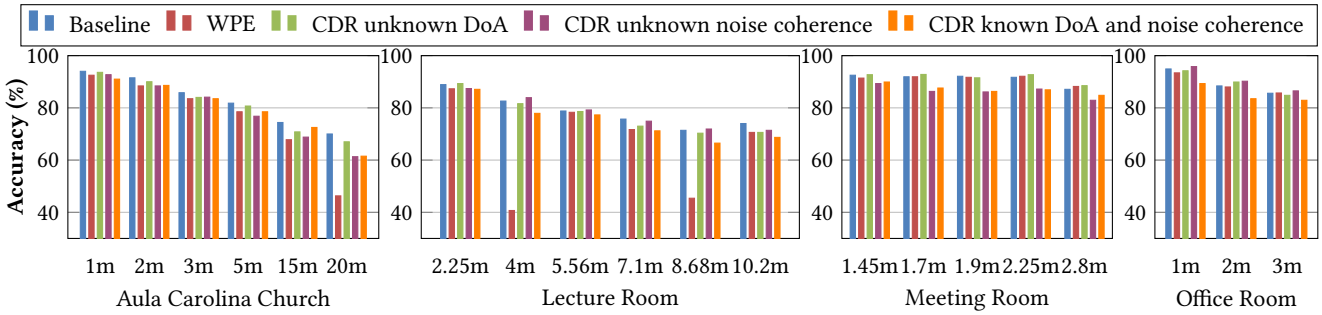

Figure 4: Performance of WPE and CDR dereverberation and denoising techniques on Emo-DB-AIR dataset. CDR with unknown noise coherence consistently outperformed the baseline in most cases (except Aula Carolina church).

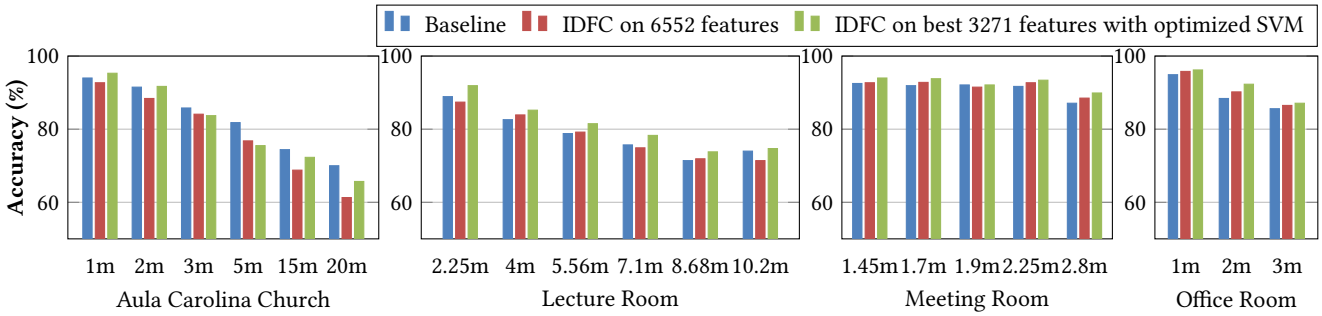

Figure 5: IDFC on original 6552 features improves performance in most cases (except Aula Carolina church due to its very large dimensions) for the Emo-DB-AIR dataset under CDR with unknown noise coherence. Another IDFC on best 3271 features (from Figure 3) with optimized SVM parameters ($c = 4, \gamma = 0.00195$) further boosts performance in all source-to-microphone distances.

The CDR with unknown noise coherence technique was found to be the best performing one from the comparative performance study of all the dereverberation and denoising techniques. We further improve its performance by applying the iterative distorted feature cut procedure and feature selection and classifier parameter optimization techniques introduced in section 3. The results are shown in Figure 5. For the Aula Carolina church, the IDFC procedure with best 3271 features and optimized SVM improves performance for 1m and 2m source-to-microphone distances, although without any feature enhancement the performances were sub-baseline for all distances. For lecture, office and meeting rooms, there is a 2.34%, 1.57% and 2.25% accuracy improvement, with an average final accuracy of 80.90%, 92.70%, and 91.96%, respectively.

4.2.2 Emo-DB-Array Evaluation. The result of WPE dereverberation technique on Emo-DB-Array dataset is shown in Figure 6. Unlike Emo-DB-AIR dataset, WPE technique improves from the baseline in all distances for Emo-DB-Array except 1m as the signal distortion due to reverberation and noise at 1m distance is too small to be dereverberated. The lab environment where we recorded Emo-DB-Array using a microphone array had more surrounding noise component from HVAC than reverberation, while the Emo-DB-AIR dataset has stronger reverberation effect than noise. Hence, the lesson learned is that WPE technique performs better on noisy signals rather than reverberated signals.

We further applied the iterative distorted feature cut, best feature selection and classifier optimization procedures from section 3 to further improve the WPE dereverberation and denoising performance on the Emo-DB-Array dataset. An improvement of 1.12%, 8.79%, 10.84% and 8.41% was obtained with a final accuracy of 94.95%, 92.15%, 86.17%, and 79.44% for 1m, 3m, 5m and 7m distances, respectively, when we applied the IDFC procedure on the best 3271 features with an optimized SVM classifier.

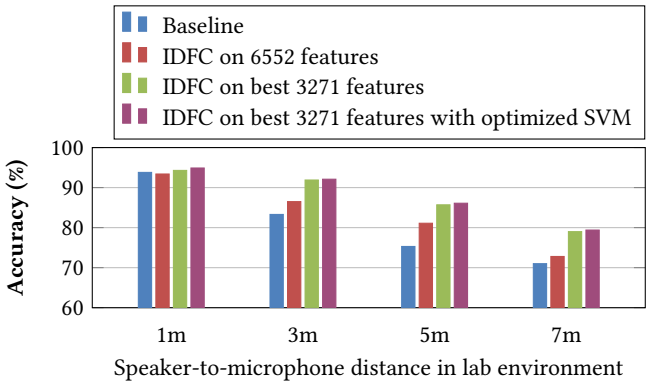

Figure 6: Performance of WPE dereverberation and denoising technique on Emo-DB-Array dataset. Combining WPE with feature selection, IDFC and optimized SVM results in a performance boost.

From these results, we can conclude that the best feature selection, IDFC and classifier optimization techniques combined with the CDR and WPE techniques significantly improve emotion detection performance in various indoor environments across a wide range of source-to-microphone distances even in worst reverberant (Aula Carolina) and noisy (lab) conditions.

5 TRAINING WITH ARTIFICIAL REVERBERATION

The objective of our approach in section 4 was to make the testing condition similar to the training condition by reducing noise and reverberation from the distant speech and making it as clean as possible like the clear speech training. In this section, we take the opposite approach: making the training condition as similar as possible to testing condition. We do this by injecting artificial reverberation into clear emotional speech and use this artificially reverberated speech for training the classifier.

5.1 The Artificial Room Impulse Response Generator

In equation 2, we showed that the distant reverberated sound signal is the convolution of the clean speech signal and the room impulse response. Impulse response represents the acoustic physical property of a room in terms of sound propagation and reflection which is essentially a FIR filter. Room impulse responses can be synthetically generated using the image source model [1] (ISM), which acts as a transfer function between a sound source and an acoustic sensor (microphone) in a given environment if some acoustic parameters of the room are known. Once such a room impulse response is available, a sample of distant audio data at any distant microphone can be obtained by convolving the impulse response with the clean speech signal, as in equation 2.

We used the Lehmann's modified and improved ISM simulation technique [8] to generate artificial room impulse responses. The model requires the following room specific parameters:

- Room dimensions
- Source and microphone positions
- Reverberation time T_{60}
- Absorption coefficients of 6 wall surfaces of the room (optional)
- Sound velocity in the room (optional)

Absorption coefficients represent the acoustic absorption capability of a surface. The value is in the range of 0 to 1. The higher the value, the more sound absorbing the surface is (hence, less reverberant). Anechoic chambers are made of fully absorbing wall, floor and ceiling surfaces. This parameter to the ISM simulation model is optional. If omitted, the model is built with equal relative absorption coefficients for all 6 wall surfaces.

The reverberation time, T_{60} is the time required for the sound energy to decay by 60 dB after the sound source has been turned off. A standard method for measuring T_{60} from the room impulse response has been presented by Schroeder [11]. But in our case, we have to estimate T_{60} blindly, as we do not know the impulse response. T_{60} only depends on the room's physical properties, and hence it can be estimated from a signal reverberated in that particular room. We used a blind T_{60} estimation method proposed in [9] which estimates reverberation time from a reverberated sound signal using a statistical model for sound decay. The advantage of this method is that it can be used to estimate T_{60} just from the (reverberated) sound recordings and without any additional measurement. The limitation of this method is that, this algorithm allows estimating the T_{60} within a range of 0.2 s to 1.2 s and assumes that source and receiver are not within the critical distance. Hence, for very large rooms or halls (like Aula Carolina in the AIR dataset) having high T_{60}, this method will not work. But for almost every practical in-house scenario, the method works.

Table 2: Average True T_{60} vs. average estimated T_{60} in different rooms

Room	Speaker-to-Microphone Distances	Average True T_{60} (Schroeder's method)	Average Estimated T_{60}
Office room	1m, 2m, 3m	0.56s	0.63s
Meeting room	1.45m, 1.7m, 1.9m, 2.25m, 2.8m	0.30s	0.25s
Lecture room	2.25m, 4m, 5.56m, 7.1m, 8.68m, 10.2m	0.84s	0.80s

To test the accuracy of the T_{60} estimation, we took a number of clean recordings from the Emo-DB dataset and convolved them with the real impulse responses from the lecture, meeting and office rooms from AIR dataset with different source-to-microphone distances. Then we blind estimated the corresponding T_{60} of the reverberated signal for that particular room and source-to-microphone distance. We also measured the true T_{60} from the impulse responses in the AIR dataset using Schroeder's method [11], and compared the estimated T_{60} with true T_{60}, as shown in Table 2. We found the discrepancy between average true and estimated T_{60} being 7 ms, 5 ms and 4 ms for office room, meeting room and lecture room, respectively. It must be noted that, the T_{60} estimation for Aula Carolina church was not possible because of its very large dimensions (18m x 30m x 15m), and the true T_{60} for Aula Carolina using Schroeder's method was found to be 4.5+ seconds, where the estimation can estimate T_{60} up to 1.2 second.

5.2 Results

Evaluations were done on meeting, lecture and office rooms from the Emo-DB-AIR dataset and on the Emo-DB-Array dataset.

5.2.1 Emo-DB-AIR evaluation. We used the room dimensions, different source and microphone positions and T_{60} reported in [6] for the rooms in Emo-DB-AIR dataset as input to the ISM simulation model. Figure 7 shows results of when training is done with speech from Emo-DB and testing on noisy and reverberated speech from Emo-DB-AIR (baseline). The 2nd series in Figure 7 is IDFC on the best 3271 features with optimized SVM which we showed in earlier sections and kept here for comparison. The 3rd series is training with reverberation (but no feature or classifier enhancement) which significantly improves average classification performance of 7.41%, 3.07% and 2.31% for lecture, meeting and office rooms, respectively, across all the source-to-microphone distances. When we incorporate training with synthetic reverberation with IDFC on the best 3271 features and optimized SVM, there is a performance boost, as seen from Figure 7. On average, we get 10.44%, 3.74% and 4.24% final performance increase compared to baseline for lecture room, meeting room and office room, respectively, when we use a fusion of training with synthetic reverberation and apply IDFC on best 3271 features with optimized SVM parameters. The final emotion classification accuracy achieved has an average accuracy of 88.75%, 94.84% and 93.89% for lecture, meeting and office rooms, respectively.

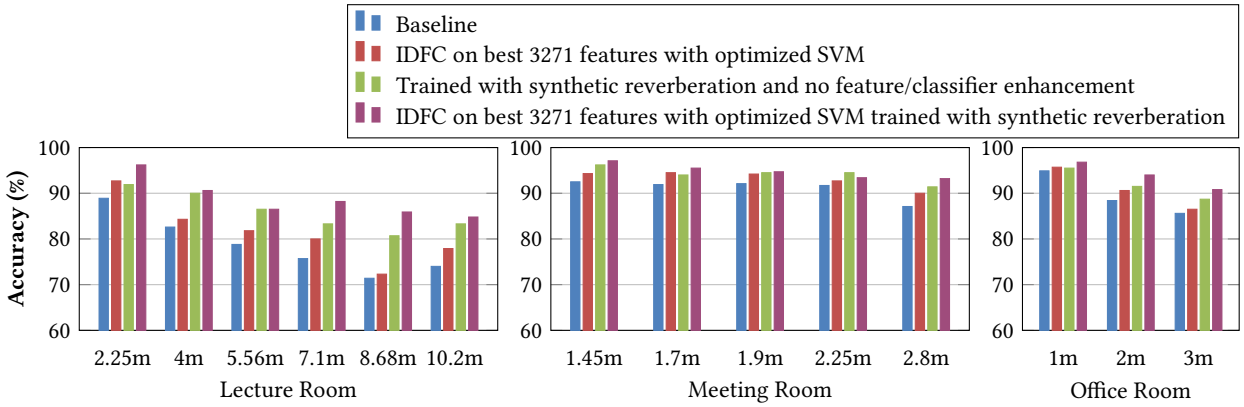

Figure 7: Training with synthetic reverberation accompanied with IDFC on best feature selection and classifier optimization results in 3.74%-10.44% average improvement across different rooms for Emo-DB-AIR dataset.

5.2.2 Emo-DB-Array evaluation. We measured the lab dimensions where we recorded the Emo-DB-Array dataset, with positions of the loudspeaker and microphones for various source-to-microphone distances, and estimated T_{60} from Schroeder's method as input to the ISM simulation model. However, in contrast to Emo-DB-AIR dataset, we used the default setup of absorption coefficients in the ISM simulation model with equal relative absorption coefficients (instead of real absorption coefficients) for the lab walls to observe the performance under this constraint.

The results are shown in Figure 8. The first 2 series in Figure 8 show the performance for baseline (no training transformation done) and IDFC on the best 3271 features with optimized classifier, as described in earlier sections. The final 3 series are for training with synthetic reverberation, adding IDFC on the original 6552 features, and adding IDFC on best 3271 features with optimized classifier. At the end, a final classification accuracy of 95.89%, 93.08%, 87.85% and 86.54% are achieved for 1m, 3m, 5m and 7m source-to-microphone distances, respectively, with an improvement of 2.06%, 9.72%, 12.52% and 15.51% from the baseline, respectively. The improvement increases with increasing source-to-microphone distance, as the signal distortions at near distances are too small for the training transformation to be as effective at further distances. Note that, the lab had loud background HVAC noise present, under which these improvements were obtained.

From these results, we conclude that the feature and classifier enhancement techniques combined with training with synthetic reverberation improves distant emotion recognition in a variety of situations, even with loud background noise.

6 CPU TIME BENCHMARKING FOR REAL TIME EXECUTION

We did all our experiments on a workstation having a Core i7-2600 CPU with 3.40 GHz clock frequency and 8 GB memory. We benchmark the computation time for SVM model building and classification, feature extraction, CDR and WPE dereverberation and denoising, T_{60} estimation using Schroeder's method and blind T_{60} estimation and impulse response simulation using ISM method with fast convolution, as shown in Table 3. Computation time is the time spent running the particular task plus running OS code on behalf of the task.

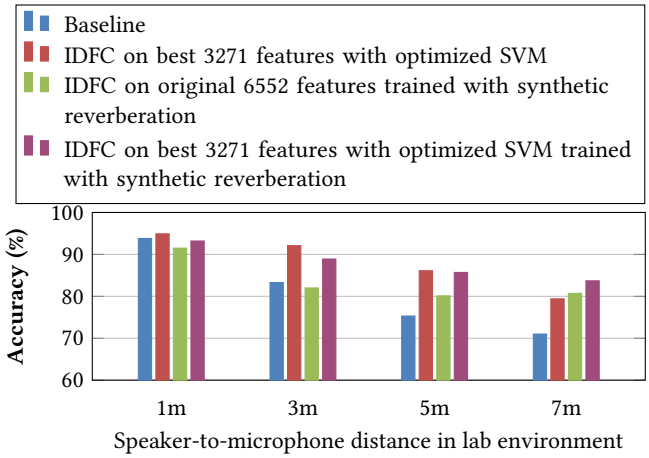

Figure 8: Training with synthetic reverberation accompanied with IDFC on best feature selection and classifier optimization results in 2.06%-15.51% improvement across various source-to-microphone distances for Emo-DB-Array dataset.

Table 3: Computation Time for Various System Tasks

Task		Computation time (s)
SVM training & classification		0.15
6552 feature extraction		0.03
CDR	No noise coherence	0.21
	No DoA	0.21
	Known DoA and noise coherence	0.26
WPE (per utterance)		1.91
T_{60} estimation	Schroeder's method	0.02
	Blind estimation	11.17
ISM simulation & convolution		1.28

As seen from Table 3, some tasks (like WPE dereverberation, blind T_{60} estimation, ISM simulation) have high CPU execution times even for a powerful workstation we used in our experiments, and pose a challenge for RTDER. Other tasks have low CPU execution times and can run in real time. Blind T_{60} estimation, which has extremely high latency, is needed only once for a particular room for input into the ISM model, and needs not to be run in real time. The ISM simulation and convolution computations are also

needed once for a particular room and a speaker-microphone configuration. If the speaker position is static (like sitting by a dining table or in a living room), ISM model needs to be computed once for training and hence needs not run in real time. However, if the speaker is moving, ISM needs to be updated as the speaker moves, which needs real time ISM computation based on speaker position. Schemes like advanced computation of all possible ISM models can be incorporated to minimize real time ISM computation (discussed in section 8). And, low latency CDR can be used instead of comparatively higher latency WPE for dereverberation and denoising for smooth real time operation.

Note that, the CPU times reported in Table 3 will increase in orders of magnitude if run on more resource constrained hardware like the Arduino/Raspberry Pi or a smartphone. In such cases, the corresponding system components may not be able to execute real time without a cloud service. However, we argue that the most likely applications of a RTDER system are safety-critical in nature (vehicle/aircraft safety, patient safety, occupant safety in smart homes) and therefore it is expected that the system components would run on powerful machines to ensure real time execution.

7 YOUTUBE DATASET EVALUATION

All our evaluations in sections 3, 4, and 5 were based on the Berlin Emotional Speech Database Emo-DB and its 2 distance aware variants created by us: Emo-DB-AIR and Emo-DB-Array. In this section, we experimented with a different dataset made from both acted and real emotional incidents taken from a number of YouTube video clips, as opposed to only acted artificial utterances of Emo-DB.

We collected 37 emotional clips from YouTube spanning more than 3 hours of emotional speech. 2 persons labeled them into 4 emotion categories: angry, happy, neutral and sad. The angry recordings included clips from the talk show "The Daily Show" with the host reacting on the South Carolina church shooting in 2015, a talk show from CNN with the participants reacting about gun control, a heated speech from a presidential candidate of the US national election 2016, and a number of clips taken from TV shows and movies from YouTube. Among the happy recordings are some personal statements released by a number of people and some funny clips taken from NBC talk shows. The neutral recordings consisted of a documentary recording about the US constitution, with a few others. The sad recordings included a number of personal statements about abuse, depression, monologues about deceased people and a number of clips from TV shows and movies.

We split these recordings into a total of 1124 10 second utterances and convolved them with the meeting and office room impulse responses from the AIR database with various source-to-microphone distances. We transformed the training by injecting synthetic reverberation from the impulse response simulator, as described in section 5. We also selected the best 3255 features from the F-scores of the original 6552 features and applied the IDFC procedure on the features of both training and test set with the optimized SVM. We achieve an average accuracy increase of 3.11% and 7.30% from the baseline (no training transformation or feature enhancement done) in the meeting room and office room, respectively, with a final emotion detection accuracy of 93.68% and 93.15%, respectively, across all the source-to-microphone distances in corresponding rooms.

From these results, we verify that our approach increases distant emotion recognition accuracy for realistic emotional speech data as well as acted corpus as shown in prior sections.

8 DISCUSSION, LIMITATION, AND FUTURE WORK

8.1 Public Dataset vs. Real Deployment

We limited our experiments to the 2 distance aware variants of the Emo-DB public dataset: Emo-DB-Array and Emo-DB-AIR, which we customized according to our needs. These 2 are the very first distance aware emotional datasets, to the best of our knowledge. Since RTDER is a new area of research, we present our preliminary results based on variants of the public Emo-DB dataset in this paper. In our future subsequent work, we plan to include RTDER results based on real deployments in real families.

8.2 Static vs. Dynamic Speakers

We considered only static speakers in this paper, i.e. though speakers were situated away from the microphones, their position remained static, as opposed to moving/dynamic speakers. Dynamic and moving speakers will impose Doppler effect of changing frequency on the distant microphones, and will cause dynamic noise and reverberation profiles in the indoor environment. Our work is extensible to handle moving speakers, given that we are able to measure speaker-to-microphone distance in real time with good precision. Several indoor positioning and distant measurement techniques in wireless fields (Wi-Fi, Bluetooth, RSSI) exist in literature. We look forward to utilize a suitable real time distance measurement technique to support dynamic speakers in our future work.

8.3 Real Time Computation Scenario for Static vs. Dynamic Speakers

For static fixed position speakers, feature extraction, signal cleaning (WPE or CDR), and SVM classification needs to be done real time with the audio stream. Reverberation adaptation (T_{60} estimation, ISM simulation, convolution) and SVM training need not to be done real time, as they need to be computed just once for static speakers and a certain indoor environment.

However, for dynamically moving speakers, the room IR needs to be updated real time as the speaker moves, therefore ISM simulation, convolution, and SVM training also needs to be computed real time. As discussed in section 6, ISM simulation and convolution are computation heavy tasks, and challenging to be computed real time, especially in resource constrained platforms like Raspberry Pi. One possible solution is advanced computing of all possible room IRs by all possible speaker-to-microphone distances into an IR cache, and use the appropriate IR from the cache depending on the latest speaker position for convolving with the speech signal. Another solution is to use opportunistic room IR computation based on latest speaker position. For example, if the current speaker-to-microphone distance is 3 meters, then compute IRs with 2.9m and 3.1m in advance and use the one whichever happens to be the next distance (assuming 0.1m distance precision) with the opportunistic scheme. T_{60} estimation, even for dynamically moving speakers,

Table 4: Confusion Matrix for Detected Emotions

True	Classified As						
	Anger	**Anxiety**	**Boredom**	**Disgust**	**Happy**	**Neutral**	**Sad**
Anger	117	0	0	1	8	1	0
Anxiety	4	56	0	0	6	3	0
Boredom	0	0	70	0	0	5	5
Disgust	0	2	1	38	1	2	2
Happy	13	3	0	1	53	1	0
Neutral	0	1	3	0	0	75	0
Sad	0	0	2	0	0	2	59

needs to be computed just once for a particular room, hence need not to be computed real time.

8.4 Confusion Matrix

With our proposed techniques, we significantly improved real time emotion recognition performance over distance. However, scope exists for betterment of the core solution i.e. emotion detection with clear speech, upon which our solutions stand. We analyzed the confusion matrix of 7 different emotions from Emo-DB after a 10-fold cross validation on the clear speech data, as shown in Table 4. We noticed that a significant number (around 33%) of misclassifications occur between the angry and happy emotions. Techniques involving hierarchical classifiers with specialized angry vs. happy separators, textual features added with acoustic features after a speech-to-text conversion can improve the misclassification rate between angry and happy classes.

8.5 Important Features for DER

We analyzed 6552 acoustic 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. Among these, various MFCC, FFT, zero crossing rates, and pitch related LLD features were found to be most important for distant emotion recognition. These features have minimal distortion due to distance and can most effectively catch the changes in prosody due to different emotions.

9 CONCLUSION

We present novel solutions to various challenges in the processing pipeline of an acoustic RTDER system. The solutions we present are useful in real time recognition of emotions from distant speech in a variety of rooms with various acoustic configurations and source-to-microphone distances. We provide empirical evidence that our novel combination of feature selection, classifier optimization and distorted feature elimination technique combined with WPE and CDR dereverberation and denoising algorithm is capable of increasing emotion detection accuracy as much as 10.84%, with the final accuracy ranging between 79.44%-94.95%. In addition, our feature and classifier enhancement and distorted feature elimination technique combined with training with synthetic reverberation from a room impulse response generator increase emotion detection accuracy as much as 15.51% across various rooms, acoustic configurations and source-to-microphone distances, with the final accuracy ranging between 87.85%-95.89%. We considered worst-case situations with extremely reverberant church halls and noisy backgrounds with loud HVAC noise and obtained improvement even in worst conditions. A case study on a dataset made from 37 realistic YouTube videos spanning 4 different emotions demonstrates a maximum of 7.30% increase in accuracy for detecting real world distant emotions in different rooms, with a maximum accuracy of 93.68%. We evaluated the CPU runtime of various system components and demonstrate the real time execution capability of our system. We concluded with discussing some limitations of our current solutions, and proposed methods to solve those in future works.

ACKNOWLEDGMENT

This paper was supported, in part, by NSF Grant IIS-1521722 and DGIST Research and Development Program (CPS Global Center) funded by the Ministry of Science, ICT and Future Planning.

REFERENCES

[1] Jont B Allen and David A Berkley. 1979. Image method for efficiently simulating small-room acoustics. *The Journal of the Acoustical Society of America* 65, 4 (1979), 943–950.

[2] Felix Burkhardt, Astrid Paeschke, Miriam Rolfes, Walter F Sendlmeier, and Benjamin Weiss. 2005. A database of german emotional speech.. In *Interspeech*, Vol. 5. 1517–1520.

[3] Yi-Wei Chen and Chih-Jen Lin. 2006. Combining SVMs with various feature selection strategies. *Feature extraction* (2006), 315–324.

[4] Moataz El Ayadi, Mohamed S Kamel, and Fakhri Karray. 2011. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition* 44, 3 (2011), 572–587.

[5] 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.

[6] Marco Jeub, Magnus Schafer, and Peter Vary. 2009. A binaural room impulse response database for the evaluation of dereverberation algorithms. In *Digital Signal Processing, 2009 16th International Conference on*. IEEE, 1–5.

[7] Kenichi Kumatani, John McDonough, and Bhiksha Raj. 2012. Microphone array processing for distant speech recognition: From close-talking microphones to far-field sensors. *IEEE Signal Processing Magazine* 29, 6 (2012), 127–140.

[8] Eric A Lehmann and Anders M Johansson. 2008. Prediction of energy decay in room impulse responses simulated with an image-source model. *The Journal of the Acoustical Society of America* 124, 1 (2008), 269–277.

[9] Heiner Löllmann, Emre Yilmaz, Marco Jeub, and Peter Vary. 2010. An improved algorithm for blind reverberation time estimation. In *Proceedings of International Workshop on Acoustic Echo and Noise Control (IWAENC)*. 1–4.

[10] 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.

[11] M. R. Schroeder. 1965. New Method of Measuring Reverberation Time. *The Journal of the Acoustical Society of America* 37, 6 (1965), 1187–1188. https://doi.org/10.1121/1.1939454 arXiv:http://dx.doi.org/10.1121/1.1939454

[12] Andreas Schwarz and Walter Kellermann. 2015. Coherent-to-diffuse power ratio estimation for dereverberation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 23, 6 (2015), 1006–1018.

[13] 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.

[14] 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.