RESONATE: Reverberation Environment Simulation for Improved Classification of Speech Models

Robert F. Dickerson
College of William and Mary
Williamsburg, VA, USA
Email: rfdickerson@wm.edu

Enamul Hoque, Philip Asare,
Shahriar Nirjon, and John A. Stankovic
University of Virginia
Charlottesville, VA, USA
Email: {eh6p, asare, nirjon, stankovic}@cs.virginia.edu

***Abstract*—Home monitoring systems currently gather information about peoples activities of daily living and information regarding emergencies, however they currently lack the ability to track speech. Practical speech analysis solutions are needed to help monitor ongoing conditions such as depression, as the amount of social interaction and vocal affect is important for assessing mood and well-being. Although there are existing solutions that classify the identity and the mood of a speaker, when the acoustic signals are captured in reverberant environments they perform poorly. In this paper, we present a practical reverberation compensation method called RESONATE, which uses simulated room impulse responses to adapt a training corpus for use in multiple real reverberant rooms. We demonstrate that the system creates robust classifiers that perform within 5 – 10% of baseline accuracy of non-reverberant environments. We demonstrate and evaluate the performance of this matched condition strategy using a public dataset, and also in controlled experiments with six rooms, and two long-term and uncontrolled real deployments. We offer a practical implementation that performs collection, feature extraction, and classification on-node, and training and simulation of training sets on a base station or cloud service.**

***Keywords*—Speaker Identification, Reverberation Compensation**

I. Introduction

Numerous studies show the quality of speech can indicate certain mood disorders [1]–[4]. In fact, during a mental status examination, a clinician makes a psychological assessment by observing and describing his patient's speech. The report usually includes some comments on its features such as loudness, rhythm, prosody, intonation, pitch, phonation, articulation, quantity, rate, spontaneity, and latency. Some features may indicate a neurological problem: for example, stroke or dementia can slow speech or produce aphonia or dysarthria. People with autism spectrum disorders or Asperger's syndrome show abnormalities in their speech. People with mania or anxiety may have rapid, loud, and pressured speech, while people with depression show prolonged speech latency and speak in a slow, quiet, and hesitant manner and also use only small changes in intonation.

In addition to the speech's features, other information such as how often the patient has conversations with others, and how often the patient actually speaks during these interactions provides a picture of the level of engagement for the speaker. Because conversations involve multiple people, a system must be able to identify who is speaking at any given time in order for a personalized report of speech features to be generated for each user. Although speech information is generally seen as sensitive private information, we employ a strategy in this work where the content of the speech is not needed, only the high-level features. The ability to continuously monitor speech features could benefit an emplaced home monitoring system [5].

There are many technical challenges in designing a system that capture these features accurately and without distortion in real environments. First, there may be significant ambient noise in the home, including that from music, television, appliances and air systems. Second, as with any propagating signal, increasing the distance between the emission source and the microphone attenuates the signal, resulting in a low signal to noise ratio. Third, and the focus of this paper, is that when sound travels through rooms, it becomes distorted by an effect called reverberation. The amount of reverberation is related to the amount of time the original sound spends bouncing off of surfaces before being captured by the microphone. The amount of distortion depends largely on the acoustic characteristics of the room, which are related to the presence of acoustically insulating or reflective materials such hardwood, carpet, furniture, and drapes. The final challenge is that the system is dynamic: users will change their position as they move about the house.

There is a large existing body of work for creating classifiers and completing necessary feature extraction for obtaining the identity of the speaker [6], the number of speakers [7], the speaker's mood [8] as well as general sounds [9], however they all make very limiting assumptions such as that the microphone and speaker are in fairly anechoic (non-reverberant) and non-noisy conditions. Previous studies show how mood detection is very challenging when audio is captured in realistic environments and standard classifiers (SVM and GMMs) are employed [8], [10]. We also show later in our evaluation, Section IV, many examples of how reverberation degrades the performance of SVM classifiers for both speaker identification and mood classification from 80-90% accuracy in non-reverberant conditions to only 20-50% with reverberant conditions.

Speech processing in open, realistic environments is an active and open research problem, but the majority of work to date has concentrated on automatic speech recognition – the task of producing text from speech content. One notable example is how to achieve accurate automatic speech recognition to allow hands-free mobile device interaction while driving a car. A recent survey paper describes the state of the art for controlling reverberation for automatic speech recognition (ASR) [10]. Whereas ASR uses only MFCC features and HMMs for classifiers, mood and speaker recognition approaches use hundreds of features, taken over several frames of audio, with different types of classifiers such as the SVM and GMM.

The main contributions presented in this paper are the following:

- We present a design for a practical platform for monitoring speech: such as speaker identification and mood, for use in home and office environments that can be deployed, trained, and configured quickly.
- We present and thoroughly evaluate a novel system called RESONATE, which combines a matched condition training approach with a unique reverberation impulse response simulator. This system allows a single training corpus to be adapted for various environments, minimizing necessary training and configuration time. We demonstrate that RESONATE performs close to the ideal baseline for accuracy, both in controlled experiments (six different rooms in houses and offices) and in uncontrolled long term deployments in both a home and an office.
- We demonstrate and evaluate how additional knowledge about the environment further improves accuracy, including data about room dimensions and position of the speaker in the room.
- We benchmark various stages of the classification task on different platforms, and offer an analysis of its performance. We show best performance when capture, feature extraction, and classification occurs on-node, while training and simulation is done off-node on a base station or cloud service.

II. Reverberant Environments

Addressing environmental reverberation can be tackled in two main ways [10]. The first strategy is to modify the front-end which tries to reverse or mitigate the effects of reverberation in the the preprocessing and feature extraction stages. Either the audio is preprocessed to explicitly reverse the reverberation, or only features that are robust to reverberation are selected. In this case, the classification model is left untouched. The second strategy takes the opposite approach by changing the classification model in some way to adapt it to handle reverberation.

A. Our Approach: RIR Simulator

Our approach called RESONATE, for **R**everberant **E**nvironment **S**imulation, does not change the frontend nor the classifier, instead it works by transforming the training set to match the testing conditions. The advantage of this approach is that it can work alongside existing approaches for improving the frontend or backend of a classification system, but augmenting the pipeline to first match the testing condition correctly. Our classification system consists of a pipeline of components shown in Figure 1. One basic way to obtain matched training samples would be to record the subjects speaking in a number of different environments and locations. However, in this paper we show that recording each speaker in the requisite number of locations and orientations can be a tedious process which involves over 30-60 minutes for each room in the house, thus makes this approach quite impractical. Of course, this time scales linearly with the number of possible speakers. Ideally, one small set of well conditioned recordings should be captured for a person and it can work in all environments.

With RESONATE, we use acoustic physical models to characterize reverberation for a particular room if the dimensions and some basic parameters are known. The result of this physical model is a room impulse response (RIR) which is essentially an FIR filter. Once this RIR is obtained, each *clean* recording can then be convolved with this filter to obtain a simulated reverberated sample for training. The difficulty now becomes obtaining these impulse responses. In [10], this process can be done empirically by emitting a very short duration signal into the environment and capturing the signal after it has propogated through the environment for several milliseconds to infer the impulse response. This is a complex and long process that requires expensive audio equipment to do properly because synchronization is important.

Our solution is to use acoustic physical models to synthetically generate RIRs using a unique impulse response simulator that produces acceptable accuracy despite using rooms that are not quite perfectly cuboid or homogeneous in their wall reflectivity. We generate the RIRs by extending Habet's implementation [11] of the Image Method [12]. The necessary parameters for this model are the sound velocity (usually 340m/s, but varies by temperature and humidity), the position of the microphone and the speaker, the room dimensions, and an estimate of the reverberation time $\overline{RT}_{60}$. The technique models the wave function as shown in Equation 1, where X and X' is the position of the source and receiver, respectively, and R represents the 6 wall geometry. R_p represents the distance from the source to the receiver and R_τ is related to the room dimensions.

$$p(t, X, X') = \sum_{p=i}^{8} \sum_{P=-\infty}^{\infty} \frac{\delta t - (|R_p + R_\tau|/c}{4\pi |R_p + R_\tau|} \qquad (1)$$

Obtaining $\overline{RT}_{60}$ "blindly" by analyzing only the received signal is an active research problem. Reverberation is characterized by two components: the early reflections, which depend on the relative positions of the speaker

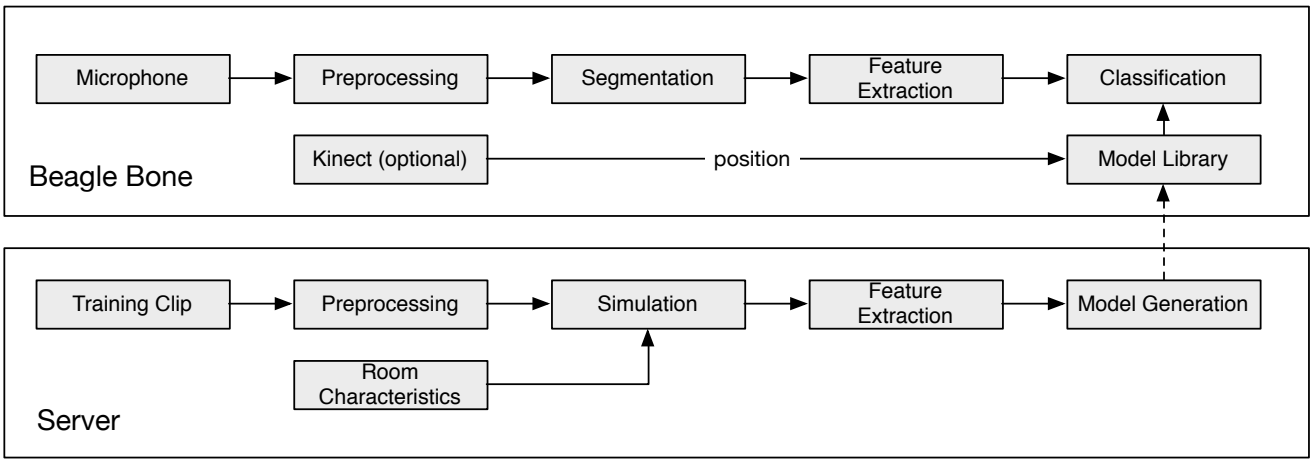


Fig. 1.
The RESONATE classification pipeline augments the currently used approach for speaker identification and mood detection with a simulation step that introduces information such as the room geometry to simulate how the training corpus would sound in a specific room in all locations (RESONATE$_1$) or a specific location (RESONATE$_2$). This model can then be sent to a device for realtime classification.

and the microphone and which can be handled by the model, and $\overline{RT}_{60}$ which is independent of these parameters, but depends on the nature of the materials of the room surfaces, which are not specified to the model. Since $\overline{RT}_{60}$ only depends on the room's properties, it can be inferred from a sample that has been reverberated by a real room with material properties similar to the room being simulated. We used Löllmann's [13] algorithm of blind reverberation time estimation. The approach uses a simple statistical model for the sound decay and $\overline{RT}_{60}$ is estimated by a maximum-likelihood (ML) estimator. We tested this algorithm by creating reverberated samples by convolving the RIRs from the Aachen Impulse Response (AIR) database [14] with 60 speech recordings. We then compare our blindly determined $\overline{RT}_{60}$ value with the true $\overline{RT}_{60}$ value (from the annotation in the AIR database), and found it to be within 60 ms of the true RIR in most cases.

For handling the case where the speaker is positioned in various parts of the room, we used our model to synthesize several RIRs one for each of the various locations the speaker can occupy. This was practically achieved by subdividing the room into a grid pattern on the X-Y plane with a 1 meter offset (the height was set to the average height of a person). In certain setups, sensors, such as trackers, chair sensors, or Kinect, may be available to estimate the location of the speaker in the room. In such a situation, only one of the RIRs above are selected based on position. Several training models are stored in a classifier bank, and during runtime the system adaptively selects the best classifier to use based on the current position. It is important to note, however, that the speaker position is not necessary, but provides additional information that can be used for improved classification accuracy, and will be the focus in the comparison of RESONATE$_1$ and RESONATE$_2$.

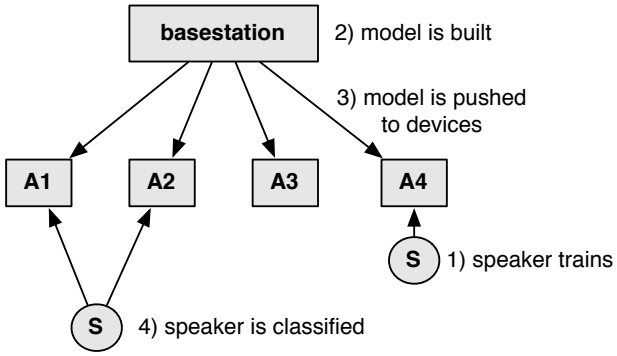


Fig. 2. Four microphones and Beaglebones are installed in different rooms. The speaker trains in front of the microphone in room A4 and the base station builds a model that is pushed down to the other devices for classification later.

III. System Implementation

We show a working example of the system operation in Figure 2. First, a node (with a microphone, beaglebone, and WiFi) is placed in each room where conversation typically occurs. Every person who lives in the home will train on one of the microphone devices close to the microphone to minimize reverberation and distortion. The recordings are sent to the basestation. Frequent visitors to the home can also do training, perhaps on their personal computer or phones, and have their training samples uploaded to the model generator which can either be on a webserver or home basestation. Each of the training samples are transformed to sound as if they came from a particular kind of room, and a tailored classification model is generated and that model is sent to the corresponding node in the system so that classification can be done in realtime without transmitting raw signals from the node. Finally, the classification result is sent back to the basestation.

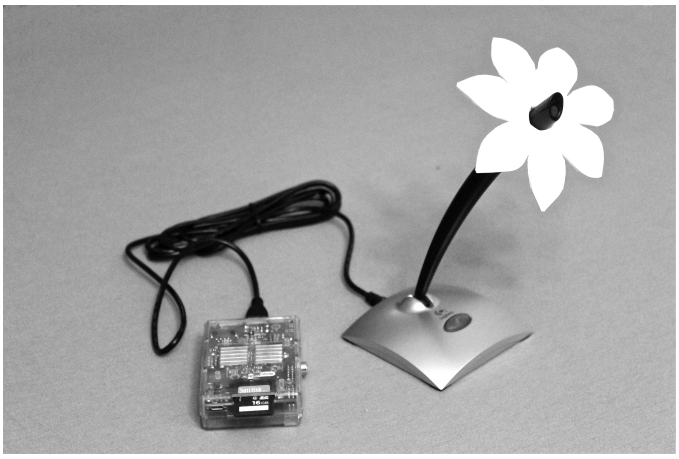


Fig. 3. Our unobtrusive audio capture and classification device uses a USB desktop microphone with a Beaglebone, and one is typically needed per room.

A. Audio Capture

There are many varieties of microphones and they are often referred to by their transducer type such as a condenser, being dynamic, or using MEMS. Most microphones also have a specific directionality (polar pattern) it was designed for (i.e. omnidirectional, unidirectional, cardioid, and shotgun), indicates how sensitive it responds to sounds arriving at various angles about its central axis. Microphones also have a unique dynamic range. Although microphone instrinsics are important to consider, evaluation of microphone selection is out of the scope of this paper. For our testing, we used two types of microphones: high-end dynamic microphones and a desktop USB dynamic microphone, and both have cardioid response patterns.

We built a distributed system based on the device shown in Figure 3. For each node, we used a Beaglebone Black platform eqipped with an ARM7 Cortex-8 microcontroller and 512 MB of memory. We loaded the Linux kernel (v.3.81) compiled for the ARM. We also created a similar device with the Raspberry Pi which we did not use for our experiments. Our experiment example assumes one of these Beaglebone microphones in each room. Additional microphones could be used to increase sensing coverage.

B. Preprocessing

Noise plays a large part in the success of the classification and also for reverberation estimation. After capturing the audio, we perform normalization to remove the DC offset and to keep the maximum amplitude capped at -1.0dB. Preemphasis is applied to reduce the adverse effects of noise and attenuation. All environments have some level of baseline noise typically from the HVAC. Since most signals will contain non-speech (over 95% in our experience), we could build a noise model of the uniform noise in the room and use spectral subtraction on the speech segments. Our system used a Wiener noise suppressor with two-step noise reduction (TSNR) technique [15]. Their approach uses harmonic regeneration noise reduction (HRNR) to refine the SNR a priori to compute a spectral gain to preserve speech harmonics. More sophisticated machine learning-based noise subtractors can extend this approach.

C. Segmentation

The next step, segmentation, obtains discrete chunks of speech for processing. There are many voice activity detectors (VADs), silence detectors, and turn-taking options in the literature [16]–[19]. We used a combination of volume, spectral energy, fundamental frequency (F_0), and spectral flatness for creating a predictor for speech segments. The spectral flatness can be used for characterizing an audio spectrum for how tone-like a sound is, and hence can eliminate signals with a large mixture of sources (such as multiple people talking at once, or music or TV in background).

D. Feature Extraction

We ported the OpenSMILE library (already written in C++) to be compiled on the ARM7 platform (with NEON optimizations). Doing classification on-node decreases network traffic, but also improves privacy concerns about transferring conversation data through the air in which eavesdroppers can intercept, or to the Cloud where other unwanted parties could access the information. We configured the feature extractor to extract a total of 384 functional features, the min, max, mean, stdev of each of the 16 low-level features. The device sends the classification result encrypted over WiFi to the base station. Because frame level features are not sent, reconstructing the speech content using automatic speech recognition would be very hard if not impossible to achieve.

For each segment, we extract the acoustic features described in the Interspeech 2009 Emotion Challenge [20]. By aggregating a series of low level descriptors (such as pitch) recorded at each instance, we compute general statistics over the duration of the utterance, resulting in a total set of 384 features. The OpenSMILE audio feature extractor [21] was used for extracting the features. First, the signal is framed into 20ms chunks using a sliding window of 10ms. A Hamming window is applied to each frame before the fast fourier transform (FFT) is taken. The mel-frequency cepstral coefficients (MFCC) are derived by mapping the powers of the spectrum to the mel scale, taking the logs of the powers at each mel frequency, then finally taking the discrete cosine transform of those log powers. The result of the FFT is also passed to a autocorrelation processor in order to estimate the fundamental frequency (F_0) from the relationship of its harmonic frequencies. The root-mean-square (RMS) energy and the zero crossing rate (ZCR) are also extracted.

We smooth the values of the features into speech contours by using a moving average filter of three window lengths, which minimizes any pops or any abrupt fluctuations in the signal. For each of these contours, various statistical functionals are computed including the maximum, minimum, range, arithmetic mean, standard deviation, skewness, and kurtosis. Additionally, the contour is approximated by the slope and error of a regression.

TABLE I. AIR database Room Characteristics

Room	Dimensions (m)	$\overline{RT}_{60}$ (s)
Office room	5 x 6.4 x 2.9	0.43
Meeting room	8 x 5 x 3.1	0.23
Lecture room	10.8m x 10.9m x 3.15	0.78

E. Classification

We use a support vector machine (SVM) for classification because of the large feature size. The LIBSVM library is used for both the training and testing. Before fitting the model, all features are scaled to the range $[-1, +1]$ so that attributes in greater numeric ranges do not dominate those in smaller numeric ranges. The radial basis function (RBF) kernel then maps the samples onto a higher dimensional space. We configure the parameters of the RBF kernel, C and γ, using the grid-search method using cross-validation to find the best combination with the highest accuracy.

IV. Evaluation

We evaluated RESONATE with three separate sets of experiments: First, we used an impulse response database on a dataset of emotional speech to investigate the effect of reverberation on both speaker identification and mood detection classification, and demonstrated how our system improves accuracy in the presence of reverberation. In the second category of experiments, we collected speech samples from four volunteers in homes and offices in a controlled manner, with a script and a predefined configuration of speaker positions. Finally, we conducted two case studies with the system running continuously for multiple weeks in real environments (one in a home, another in an office). For the controlled study and case study, we only evaluated the speaker identification because of the difficulty of assessing the mood of our speakers empirically.

A. Public Data Set Evaluation

We investigated the effects of reverberation on both speaker identification and mood detection by artificially introducing reverberation by convolving empirically collected impulse responses (AIR dataset) with recorded speech segments from a popular emotional speech data set (EmoDb). There are limited emotion datasets that are freely available [22]. We selected EmoDb [23] because it contains large number of speakers, is freely available, and is widely accepted in the affective computing community. It contains a collection of utterances spoken by 10 different actors (5 male, 5 female) using a variety of emotions. The recordings include various short phrases taken in an non-reverberant (anechoic) chamber.

The empirical set of RIRs came from the Aachen Impulse Response (AIR) database [14]. A summary room types of the collected RIRs are shown in Table I. For each room type, there were 5 different RIRs corresponding to 5 different speaker positions.

We considered four different scenarios for this evaluation: for the baseline, we assumed that in the real scenario we would have access to clean recordings of the person's speech to properly train our classifiers on, and

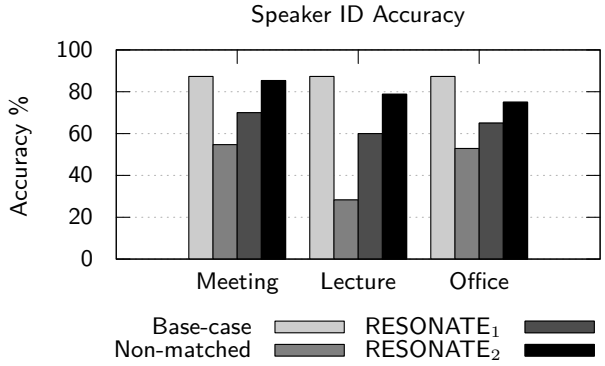


Fig. 4. The performance for the speaker identification task where the classifier was trained and tested on the EmoDB corpus. The effects of reverberation are most notable in larger rooms. Using the RESONATE strategy, the classification results were improved.

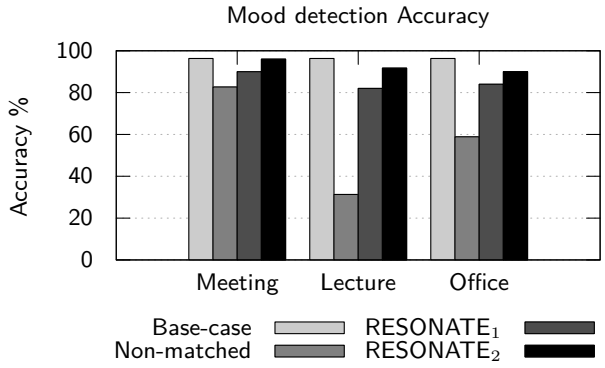


Fig. 5. The performance of the mood detection task trained and tested on the EmoDB corpus. Larger rooms are most impacted from reverberation, but RESONATE can improve the performance of in these cases close to baseline.

that we would have a clean sample to test against such as gathered from a worn microphone. This is a "best case scenario" for the ability to properly classify the user and her mood. The second case introduced reverberation, but no correction was applied i.e., we trained with the clean samples, and tested with the reverberated samples created by simulation. In the third case, RESONATE_1, we used simulation to form a better training set and assumed that we know the dimensions of the room. The final case, RESONATE_2, assumes that along with the dimensions we could determine also the position of the speaker in the real room when speaking.

We evaluated the accuracy of the speaker identification and mood detection classifier under these four different scenarios. We used 10-fold cross-validation on the training and testing set for each scenario. The results are shown in Figures 4 and 5. What we found was that the speaker identification accuracy varied considerably depending upon the room, however the RESONATE approach consistently gave better results, often near the baseline.

In addition to speaker identification and mood detection accuracy, we also evaluated the effects of two impor-

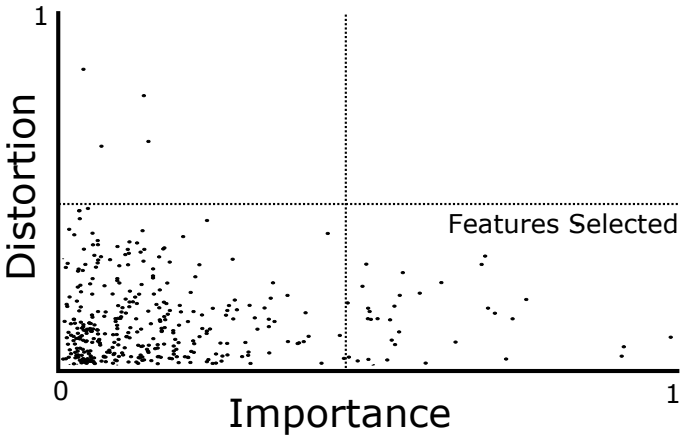


Fig. 6. The 384 features extracted are sorted by their importance for identifying the speaker and the amount of distortion occurring from the effects of reverberation.

tant parameters in the system. The first is the estimation error of $\overline{RT}_{60}$, since our training system must estimate this parameter from the training samples. The second is the effect of the utterance length on classification accuracy since this can vary. We present these evaluations next.

1) Feature Selection: Not all of the 384 features are distorted to the same degree by the effects of reverberation and noise. Also, not all of them are useful for the classification task. To evaluate which features are most important for classification, we used an algorithm by Chen et al [24] that uses F scores to compute the importance of the feature for correct classification. Our goal is to choose features that maximize the number of important features and minimize the number of features prone to distortion.

We plot the importance of feature (from its f-statistic) and the normalized level of distortion in Figure 6. The features in the bottom-right portion of the graph should be selected because they exhibit low distortion, but high importance. In general, among the highest importance features are those related to the MFCC, particularly the higher band frequency (in the 11th band). However when reverberation is introduced, those MFCC suffer from the largest distortions. We discovered that the set of features related to PCM and F_0 to offer a balance between high importance and low distortion.

Next, we evaluated whether choosing a smaller subset of the original 384 features could offer better performance than the list as a whole. In Figure 7 we show how selecting a smaller set of features that have low distortion, but high classification importance improves accuracy. For the speaker identification task, we saw a maximum accuracy of 68.86% when 95 features were chosen versus 63% accuracy when all 384 were used. It is important to note that these accuracy numbers are from uncontrolled long-term collection in real environments, which would explain the low accuracy numbers.

B. Benchmarking

Model building is a computationally intensive task and the Beaglebones take several minutes to complete feature extraction, reverberation simulation, and SVM model fitting on the node, however the base station (multi-core machine with several gigabytes of memory) completes this task in a few seconds as shown in Table II. For these benchmarks, we recorded the time the processing thread spends inside of the user-level of the OS. The Beaglebone has a Cortex A8 ARM processor, and all of our C++ code was compiled for the architecture using aggressive optimizations and the NEON extensions. The results show that realtime classification and feature extraction can be done on-node (for a 5 second clip, classification can be done in less than 3 seconds). We also show how important that the more computationally rigorous tasks such as training from the corpus training and reverberation simulation be done on a more powerful platform such as a base station (or server).

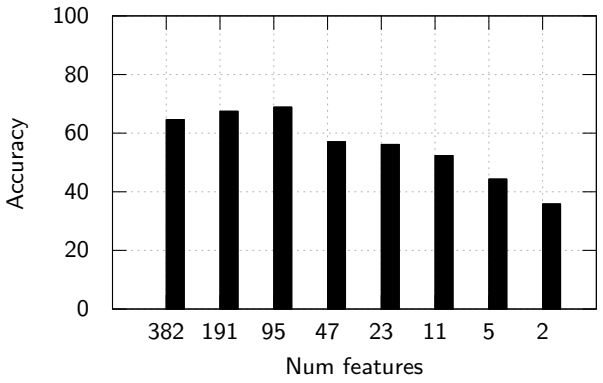


Fig. 7. Using the f-statistic criteria we decreased the number of features used for classification. The peak accuracy was 68.86% with 95 features.

TABLE II. Classification Benchmarks

Task	Beaglebone	Base station
Feature extract (5s clip)	2.51 sec	0.10 sec
Feature extract (corpus)	18.25 min	20.84 sec
SVM Training	4.88 sec	0.25 sec
Classification	5 ms	0.3 ms
Fast Conv. (corpus)	17 min	5 sec
Sim. building (room)	4.92 s	0.40s

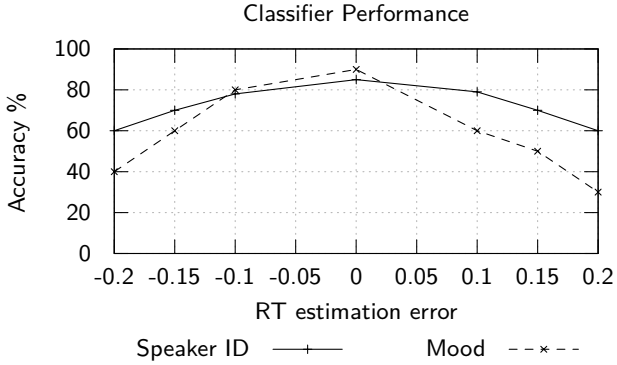


Fig. 8. $\overline{RT}_{60}$ estimation error effect on speaker ID performance.

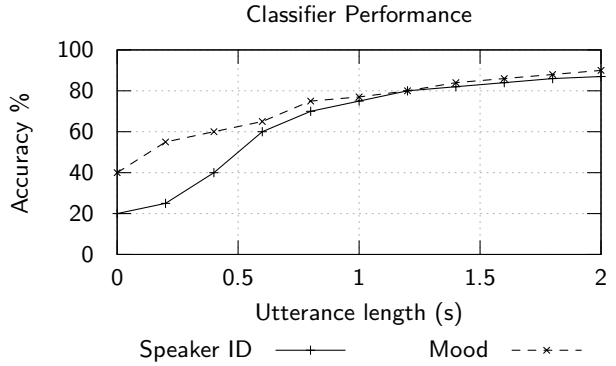


Fig. 9. Relationship of the length and the speaker identification accuracy. As the length of the utterance increases, the accuracy improves. We filtered out short uttererances for this reason.

1) Reverb Estimation and Accuracy: In practice, the $\overline{RT}_{60}$ parameter must be estimated by having no additional information other than the signal itself. The algorithm for blind estimation maximum likelihood estimates the $\overline{RT}_{60}$, but will result in some errors – especially when there is noise in the signal. Therefore, in this section, we evaluate how a poor estimation of $\overline{RT}_{60}$ effects overall accuracy. Because the Aachen Impulse Response database reports the ground truth $\overline{RT}_{60}$ value, we evaluate the accuracy of the speaker identification task as a function of error. We vary the error in milliseconds in steps of 0.05ms. The evaluation was done using the EmoDB speech samples and the results are shown in Figure 8. The results show that if there is zero or very small error in $\overline{RT}_{60}$, then accuracy is above the 80% level. If the $\overline{RT}_{60}$ error is large, e.g. 0.2, then speaker ID accuracy drops to about 60% and mood accuracy drops to about 30%. In the EmoDb data set evaluation, because of the quality of the original recording, rarely did the error exceed 0.05ms. The estimation error our system noticed were within ±0.1s in the living rooms, but in the lecture hall the error was over 0.2s. Our tests have shown that larger rooms make it harder to reliably determine amount of reverberation reliably.

2) Length of segment and Accuracy: While collecting in-home and in-office audio data for many weeks, we observed that the speaking segments vary greatly in duration. Because the classification works by extracting the statistics of the features across frames for the entire utterance length, a large utterance size will increase the accuracy of the classification. We again used the EmoDb corpus, but varied the segment size and observed the classifier accuracy. In Figure 9 we see that if the utterance length is above 2 secs we obtain well over 80% accuracy for the classifier, while utterances under 0.5 secs are in an unacceptable 20-40% range.

C. Controlled Testing in Real Environments

The controlled experiments above show the potential for the RESONATE method for producing favorable accuracy, and in this section we demonstrate how well it performs when collecting audio from our system's microphones in actual environments. We selected a variety of

TABLE III. Experiment Room Characteristics

Type	Dimensions (m)	Floor
Living room	3.9x4.3x3.05	Wood & Rug
Kitchen	2.3x4.3x3.05	Linoleum
Living room	3.7x4.7x2.43	Wood
Kitchen	3.7x4.4x2.43	Linoleum
Meeting	10.0x6.9x2.74	Carpet
Office	15.0x10x2.94	Linoleum

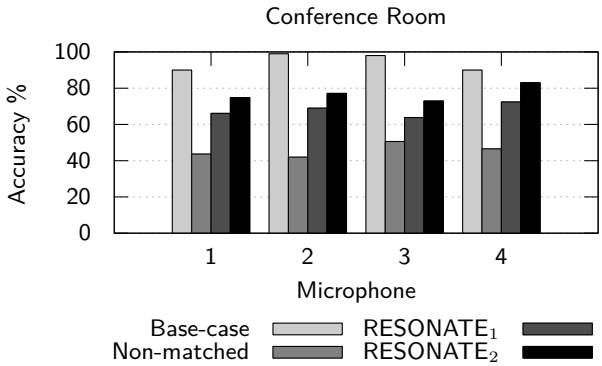


Fig. 11. Conference Room

rooms based on where conversation typically occurs, 2 living rooms, 1 office, 1 conference room, and 2 kitchens. The basic geometry and features of the rooms used in the experiments are shown in Table III. It is important to note that these rooms were furnished with sofas, desks, curtains, and other objects that affect the acoustics of the space. Our test rooms also had typical noise sources such as the hum of the refrigerator and air conditioning system, which could be consistent or intermittent throughout the recordings. Additionally, none of these rooms were precisely cuboid in geometry, and often had open doorways and openings to other rooms, as most real environments do. One of the significant results of this section is that modeling these rooms as simple cuboids in our simulation, despite their small geometric aberrations, was quite successful for accurate speech analysis.

We recruited four volunteers (2 male, 2 female) and recorded them in each of the rooms. We placed four microphones in each of the corners of the room. Additionally, the speaker carried a hand-held microphone in order to simultaneously capture the signal with minimal reverberant effects. This signal was used as the 'clean' sample for base case training and testing, as well as later for the signal on which simulated RIRs would be applied. We divided the room into a grid (similar to the method described in producing simulated sampling), and at each point, the speaker spoke facing the four ordinal orientations (approximately north, south, east, and west). Speakers read the same three-sentence passage from a book to ensure consistency in our experiment. The speakers were instructed to remain in a neutral speaking tone. We did not evaluate the case of a moving speaker in this experiment.

Here, in the 'Baseline' case, we trained and tested with the samples recorded by a particular microphone at any corner. In the 'Non-matched' case we trained with

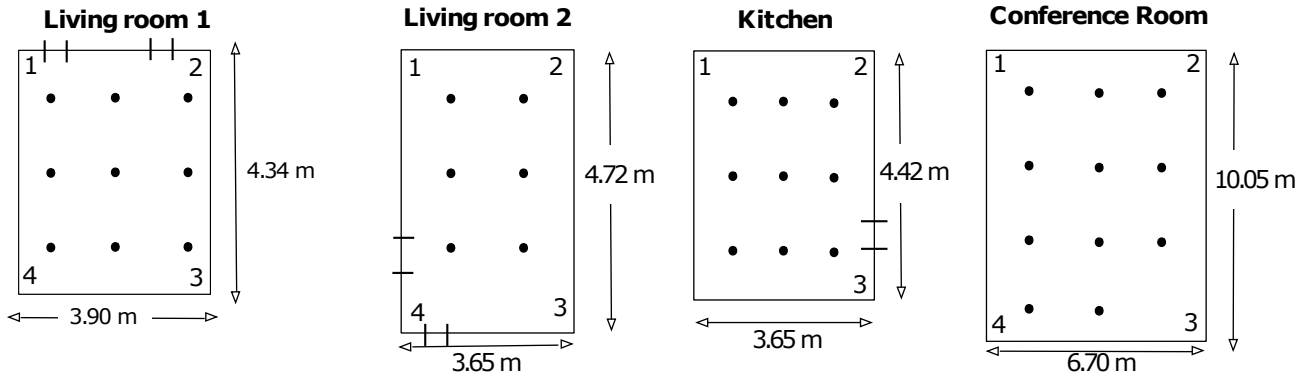


Fig. 10. Our approach was evaluated in 2 living rooms, a kitchen, and an office for the controlled experiments where the each participant would occupy the positions indicated by the black dots.

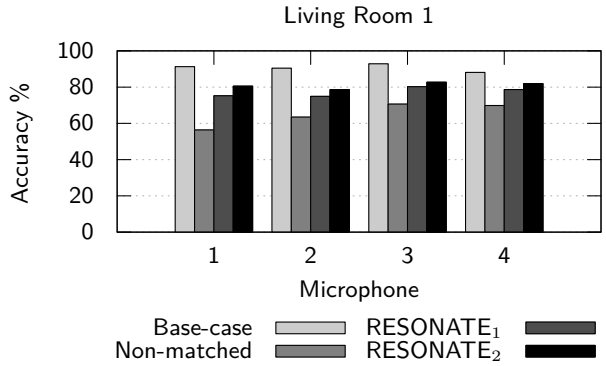


Fig. 12. Living Room 1

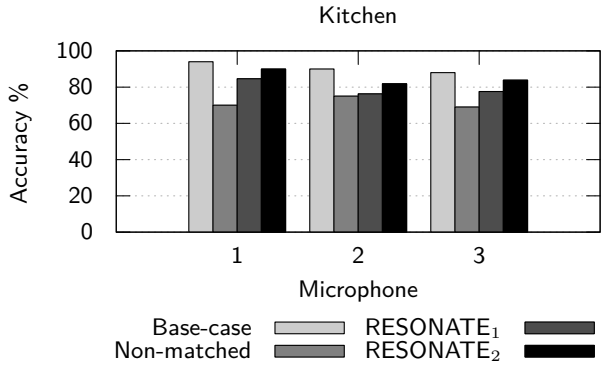


Fig. 14. Kitchen 1

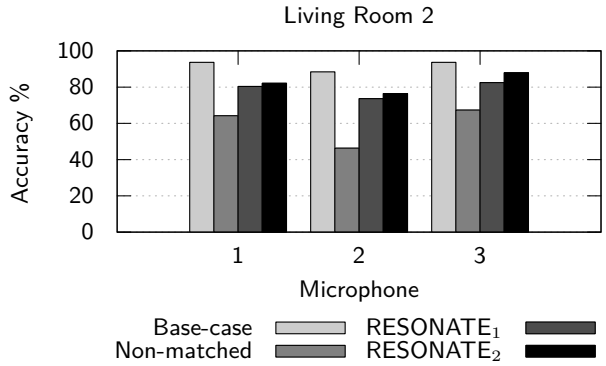


Fig. 13. Living Room 2

the 'clean' samples recorded by the handheld microphone, and tested with the samples recorded by a particular microphone at any corner. In the RESONATE_1 case (simulation without known location), we converted the 'clean' recordings of the speaker's microphone to a reverberated version, assuming the receiver is a corner microphone and the speaker can be in any of grid positions of the room; then we tested on the recordings of the corner microphones. Finally, in the RESONATE_2 case (simulation with known location), we used the location of the speaker in the room and generated a reverberated signal accordingly from the 'clean' signal from the speaker's microphone. In all four cases, we first removed the ambient noise from the recordings by all microphones using a noise removal algorithm [15] before processing. We applied 10-fold cross-validation as before, and again only evaluated speaker identification accuracy. The results are shown in Figures 11, 12, 13 and 14.

The accuracy for the 'Baseline' scenario was typically around 90% in each room. However, this scenario requires training for each user in each room where the system will be deployed. The 'Non-matched' scenario (where nothing was done to compensate for reverberation) shows results below 70% and as low as 40%. This shows the dramatic impact of reverberation on speaker identification accuracy. In the RESONATE_2 scenario (simulation with known location), the classification accuracy rises to within 5% of the base case for most of the tests, except for the conference room. In addition to having an unusual variety of sound-reflective surfaces, the conference room had considerable HVAC noise in the background. This noise was difficult to subtract using our Wiener filter and when reverberation was applied, the noise was amplified. This is another example why noise subtraction is a fundamental step to this strategy. As we do not know the location of the speaker in the RESONATE_1 scenario, the speaker identification accuracy drops. However, it is still significantly better than the non-matched scenario. Our results here show

TABLE IV. Training Time

Case	Training Time	Accuracy
Base Case	45 min	90%
Do Nothing	1.5 min	60%
$RESONATE_1$	1.5 min	80%
$RESONATE_2$	1.5 min	85%

that our techniques to address reverberation significantly improve speaker identification accuracy compared to the non-matched scenario by roughly 30%, and in most cases closely approach the baseline accuracy of training all of the speech in the environment that it will be tested in.

1) Training Time: Our results show that the baseline case still provided the greatest speaker identification accuracy. However, this came at the cost of a lengthy training period. For example, in one room, each speaker must occupy 34 total positions at 4 orientations each, totalling 136 recordings. At 20 seconds per recording, the minimum amount of time it would take to complete the training for 4 rooms would be approximately 45 minutes per person. Although this method provides good accuracy, this time investment is not always convenient especially considering multiple rooms and many speakers. In addition, there are also some situations where training in the real environment is not even possible. One such example is where the classifier has been trained from a preexisting corpus (such as EmoDB) that cannot be trained in an environment. RESONATE is able to solve this problem by giving close to baseline accuracy with minimal training time (1.5 minutes), without requiring access to the real environment for training.

D. Long-term Real Deployment Evaluation

We now test the system in a completely uncontrolled manner in the long term, by testing it in two real deployments: one in a living room in a home, and the other in our office space. The Beaglebone system shown in Figure 3 was used, and was ideal for this purpose since it is compact, unobtrusive, and powered by a wall outlet. Although the system can do classification on-node, for post-experiment analysis of the data, we captured the signal and compressed it using the libVorbis codec at 44.1 kHz sampling rate and stored onto the 16 GB microSD card.

For the home, data was collected for 4 weeks and for the office for 6 weeks. The floor plan of the living room and office room is shown in Figure 15. In the living room was a large sofa, a TV, and an electric keyboard. The microphone was placed on the table next to the TV. Adjacent to the living room was a long hallway and an entrance to the kitchen. The office space was a large room (almost 10m x 10m) with cubicles down its center line, and the microphone was placed in the last cubicle next to the far wall.

Since our system does not have a robust signal selector, we selected only speech segments that were over 2s long and manually removed laughter or TV noises in the background since these were complicated to remove automatically. The signals often had pops, knocks, and clicks in the audio, there were also examples of typing and some appliances that were filtered out as well. We show a table of the types of sounds we came across other than voice in our listening stage in Table V. Voice vs. noise detection is outside of the scope of this paper, however the literature uses many machine learning approaches similar to speaker identification and mood detection, and with similar feature sets and classifier types. The RESONATE approach might help augment those machine learning techniques as well.

1) Challenges and Solutions: Real deployments offered a number of challenges that did not occur in the controlled recordings. For speech itself, many observations came to light: First, that real utterances are most often brief statements averaging 1 second long. This duration is insufficient for reasonable accuracy from our classifier. However, since mood detection only requires a small number of longer speech samples over the entire day, eliminating short samples might not affect overall mood detection. Second, the speech of multiple people is often mixed and overlapping. If these instances of speech are not separated by silence, the system cannot detect that they are separate utterances by different people. Third, in a real environment, people do not speak in the same consistent manner as they do when creating their training set. The occupants of the home in particular often took on different affects, and raised their voice into a higher register when talking to their cat. People are also prone to making many vocal noises that could be confused for speech, such as laughter and coughing. These issues have a negative effect on the system's overall accuracy. Another problem was that the microphone in the living room was able to pick up sounds and speech signals from adjacent rooms like the kitchen and the hallway. However, when using the networked configuration of Beaglebone devices, one in each room, only the cleanest signal is used for analysis. Advancements made in the area of blind source separation (BSS) can be used to separate mixed sources in a signal because the received signal is a linear mixture of statistically independent sources. However, to date, BSS tends to not produce good results in reverberant environments.

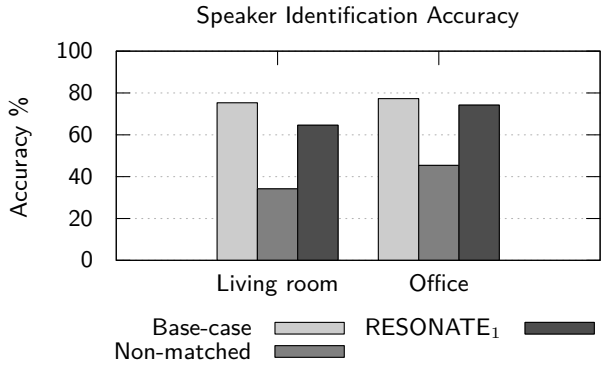
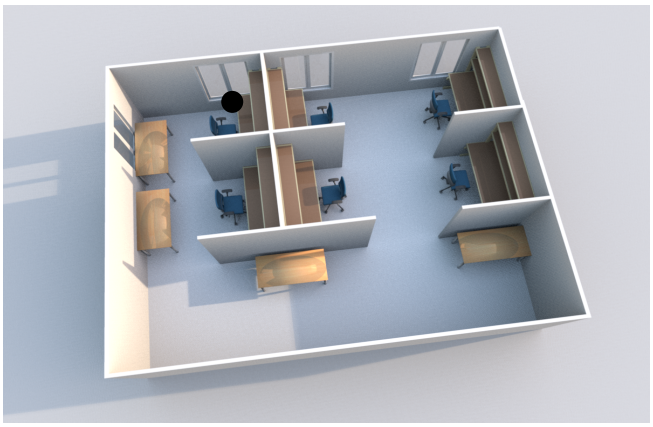


Fig. 16. Speaker ID classification for the long-term 2-month studies in the home and the office.



Office Room

Living room and Kitchen

Fig. 15. We tested RESONATE for several weeks with a single microphone in an office and also a living room in an uncontrolled setting.

TABLE V. Sounds Encountered

Type	Examples
Physiological	Sneezing, nose blowing, sniffling, clearing throat, hiccup, eating, burp, humming, laughter, drinking
Objects	Phone vibrating or ringing, typing, mouse wheel, unwrapping food, papers rustling, clothes rustling, television, piano, moving furniture, doors opening and closing, objects dropping or moving, footsteps, pouring liquid, coffee percolation, dishwasher
Ambient	Truck backing up, siren, birds chirping, passing airplane, traffic, motorized tools (lawnmower, etc)

2) Speaker ID Performance: For establishing ground truth for the speaker identification, an occupant living in the home and also working in the office listened to each audio clip and assigned a speaker label. The classifier was trained with three people who work in the office, and the two people who live in the home. In Figure 16, we show the performance of the classifier in the home and the office. Since the speakers' location was not tracked in this experiment, we did not use distance information as in our previous evaluations for RESONATE$_2$. The accuracy was roughly 75% for classifying among four speakers in the best case. It is important to consider that the accuracy even for state-of-the-art approaches for speaker identification is poor in the case for an unconstrained freeform speech. However, we demonstrate that RESONATE is able to resolve almost all the problems resulting from reverberation to match within 5-10% of the baseline. The accuracy of speaker ID in the controlled setting was significantly higher than in the long-term deployment because there was more consistency in the input to the classifier. As discussed, there are two main sources of variation: the first, which our system helps to overcome, is the reverberation and sounds from the environment. The second, however, is the variety of different ways that speakers talk in a real environment, in comparison to the consistent tone and content used during a training session. In our controlled experiments, the scripted content and tone during the training and testing cases were identical; however, if testing had been performed instead on spontaneous speech, it can be projected that the accuracy would have been much lower.

V. Conclusions

We present and thoroughly evaluate a novel system called RESONATE, which combines a matched condition training approach with a unique reverberation impulse response simulator. This system allows a single training corpus to be adapted for various environments, minimizing necessary training and configuration time. We have practically demonstrated how it mitigates the negative effects of reverberation in real home or office environments for speech classification applications such as speaker identification and mood detection. Our results show that reverberation has a significantly negative effect on the performance of these applications in real environments, and we also show how our approach improves performance considerably in the presence of reverberation using only very basic room information. RESONATE minimizes training effort for users using a shared large corpus of voices and then creates a tailored training set by generating reverbed samples of their voice considering different room acoustics, based solely on simple room acoustic models. We believe that this solution is extensible and in the future can be used in conjunction with other machine learning strategies such as multiple classifier models, improved feature sets, improved noise elimination, and blind source separation (BSS). We have evaluated RESONATE using public data sets, collecting voice samples from volunteers in different rooms in homes and offices in controlled settings, and finally by deploying our system for two long-term studies.

Acknowledgment

This work was supported, in part, by NSF grant CNS-1319302, DGIST, Ministry of Education, Science, and Technology, Korea, and a gift from Parc, Palo Alto, Calif.

References

[1] J. Mundt, P. Snyder, M. Cannizzaro, K. Chappie, and D. Geralts, "Voice acoustic measures of depression severity and treatment response collected via interactive voice response (IVR) technology," *Journal of neurolinguistics*, vol. 20, no. 1, pp. 50–64, 2007.

[2] N. Cummins, J. Epps, M. Breakspear, and R. Goecke, "An investigation of depressed speech detection: features and normalization," in *Interspeech*, 2011, pp. 6–9.

[3] A. Flint, S. Black, I. Campbell-Taylor, G. Gailey, and C. Levinton, "Abnormal speech articulation, psychomotor retardation, and subcortical dysfunction in major depression," *Journal of psychiatric research*, vol. 27, no. 3, pp. 309–319, 1993.

[4] M. Alpert, E. Pouget, and R. Silva, "Reflections of depression in acoustic measures of the patient's speech," *Journal of affective disorders*, vol. 66, no. 1, pp. 59–69, 2001.

[5] R. F. Dickerson, E. Gorlin, and J. A. Stankovic, "Empath: a continuous remote emotional health monitoring system for depressive illness," in *Conference on Wireless Health*, San Diego, CA, 2011.

[6] H. Lu, A. Brush, B. Priyantha, A. Karlson, and J. Liu, "SpeakerSense: energy efficient unobtrusive speaker identification on mobile phones," in *Pervasive*, San Fransisco, CA, 2011.

[7] C. Xu, S. Li, G. Liu, and Y. Zhang, "Crowd ++ : Unsupervised Speaker Count with Smartphones," in *Ubicomp*, 2013, pp. 43–52.

[8] B. Schuller, D. Seppi, A. Batliner, A. Maier, S. Steidl, and S. S. De, "Toward more reality in the recognition of emotional speech," in *IEEE Acoustics, Speech, and Signal Processing*, no. 101, Honolulu, HI, 2007, pp. 941–944.

[9] S. Nirjon, R. F. Dickerson, P. Asare, Q. Li, D. Hong, and J. A. Stankovic, "Auditeur : A Mobile-Cloud Service Platform for Acoustic Event Detection on Smartphones," in *International Conference on Mobile Systems, Applications and Services (MobiSys 2013)*, Taipei, Taiwan, 2013.

[10] A. Sehr, M. Delcroix, and K. Kinoshita, "Making machines understand us in reverberant rooms," *IEEE Signal Processing*, vol. 29, no. 6, pp. 114–126, 2012.

[11] E. Habet, "Room impulse response generator for Matlab," 2012.

[12] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small room acoustics," *Acoustic Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[13] H. W. Lollmann, E. Yilmaz, M. Jeub, and P. Vary, "An improved algorithm for blind reverberation time estimation," in *Acoustic Echo and Noise Control*, no. 2, 2010, pp. 1–4.

[14] P. Jeub, Marco and Schafer, Magnus and Vary, "A binaural room impulse database for the evaluation of dereverberation algorithms," in *Digital Signal Processing*, Santorini, Greece, 2009.

[15] C. Plapous, C. Marro, and P. Scalart, "Improved signal to noise ratio estimation for speech enhancement," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 6, pp. 2098–2108, Nov. 2006.

[16] P. Khoa, "Noise robust voice activity detection," Ph.D. dissertation, Nanyang University, 2012.

[17] J. Sohn, S. Member, N. S. Kim, and W. Sung, "A statistical model based voice activity detection," *Signal Processing*, vol. 6, no. 1, pp. 1998–2000, 1999.

[18] J. Ramírez, J. Segura, J. M. Górriz, and L. Garcia, "Improved voice activity detection using contextual multiple hypothesis testing for robust speech recognition," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 15, no. 8, pp. 2177–2189, 2007.

[19] A. Cho, Yong Duk and Al-Naimi, Khaldoon and Kondoz, "Improved voice activity detection based on a smoothed statistical likelihood ratio," in *Acoustics, Speech, and Signal Processing*, Salt Lake City, UT, 2001, pp. 737–740.

[20] B. Schuller, S. Steidl, and A. Batliner, "The interspeech 2009 emotion challenge," *Interspeech*, pp. 312–315, 2009.

[21] F. Eyben, M. Wöllmer, and B. Schuller, "openSMILE: the Munich sersatile and fast open-source audio feature extractor," in *ACM Proceedings of Multimedia (MM)*, Florence, Italy, 2010, pp. 1459–1462.

[22] G. McKeown, M. F. Valstar, R. Cowie, and M. Pantic, "The SEMAINE corpus of emotionally coloured character interactions," *IEEE Multimedia and Expo*, pp. 1079–1084, Jul. 2010.

[23] F. Burkhardt, A. Paeschke, and M. Rolfes, "A database of german emotional s peech," in *European Conference on Speech Communication and Technology*, 2005, pp. 3–6.

[24] Y. wei Chen, "Combining SVMs with various feature selection strategies," in *Feature Extraction*, no. 1. Springer-Verlag, 2005.