Is Someone There Or Is That The TV? Detecting Social Presence Using Sound

NICHOLAS C GEORGIOU, Social Robotics Lab, Yale University, USA
REBECCA RAMNAUTH, Social Robotics Lab, Yale University, USA
EMMANUEL ADENIRAN, Social Robotics Lab, Yale University, USA
MICHAEL LEE, Social Robotics Lab, Yale University, USA
LILA SELIN, Social Robotics Lab, Yale University, USA
BRIAN SCASSELLATI, Social Robotics Lab, Yale University, USA

Social robots in the home will need to solve audio identification problems to better interact with their users. This paper focuses on the classification between a) *natural* conversation that includes at least one co-located user and b) *media* that is playing from electronic sources and does not require a social response, such as television shows. This classification can help social robots detect a user's social presence using sound. Social robots that are able to solve this problem can apply this information to assist them in making decisions, such as determining when and how to appropriately engage human users. We compiled a dataset from a variety of acoustic environments which contained either *natural* or *media* audio, including audio that we recorded in our own homes. Using this dataset, we performed an experimental evaluation on a range of traditional machine learning classifiers, and assessed the classifiers' abilities to generalize to new recordings, acoustic conditions, and environments. We conclude that a C-Support Vector Classification (SVC) algorithm outperformed other classifiers. Finally, we present a classification pipeline that in-home robots can utilize, and discuss the timing and size of the trained classifiers, as well as privacy and ethics considerations.

CCS Concepts: • **Human-centered computing → Sound-based input / output**.

Additional Key Words and Phrases: human-robot interaction, audio analysis, in-home systems

1 INTRODUCTION

Imagine you are walking around the house when you stumble upon a door that is slightly ajar—opened just enough so that you can hear, but not see, what is going on inside. Opening the door to see if it is appropriate or not to enter is self-defeating. If you do not hear anything, it is very difficult to make any judgments. Suppose, however, that you hear human speech from behind the door. This piece of information can give you insight and can help you in your decision-making.

However, knowing that there is human speech is not enough. Many lower-level characteristics, as well as higher-level conceptual components of this speech, might be important factors in your decision. Do you recognize the voices? Does the speech sound serious or is it more lighthearted? Is there shouting or is the tone normal?

Authors' addresses: Nicholas C Georgiou, nicholas.georgiou@yale.edu, Social Robotics Lab, Yale University, New Haven, Connecticut, USA; Rebecca Ramnauth, Social Robotics Lab, Yale University, New Haven, Connecticut, USA; Emmanuel Adeniran, Social Robotics Lab, Yale University, New Haven, Connecticut, USA; Michael Lee, Social Robotics Lab, Yale University, New Haven, Connecticut, USA; Lila Selin, Social Robotics Lab, Yale University, New Haven, Connecticut, USA; Brian Scassellati, Social Robotics Lab, Yale University, New Haven, Connecticut, USA.

2573-9522/2023/8-ART $15.00
https://doi.org/10.1145/3611658

What emotions can you detect from the speech? How many people can you hear? If you hear two friendly sounding people having a chat, you might be more inclined to knock. If you stop by to relay a message, and you hear yelling coming from the room, it is probably best to steer clear for now. But, imagine that the yelling is from an enthusiastic sportscaster describing a sporting event or that the serious tone that you hear is from a dramatic soap opera. You might make a different decision if you know that the speech is coming from a television show rather than from physically present people in the room conversing. This is an important component of the speech that will influence your understanding of the situation and can affect how you interact, if you do.

Similarly, a social robot that is designed to interact with users in realistic and appropriate ways should have the ability to make this disambiguation. The robot can benefit from knowing whether the speech coming from behind the door is from a physically present human socializing. More generally, knowing when speech is a product of at least one co-located person conversing, or not, can assist social robots in making inferences about users' activities and can help them accommodate their users through a better understanding of their environments. This paper focuses on whether there is (1) *natural* conversation occurring that includes at least one co-located user or (2) *media* playing from electronic sources that does not require a social response. These are common speech scenarios in the home which can assist the robot in detecting the social presence of a user through what the robot hears.

In practice, we imagine countless settings where the ability to make such a classification could be utilized by robots to assist them in accomplishing their goals. For example, a social companion robot in the home may decide to engage a co-located user with a supportive, social interaction if it infers that the user is upset, as opposed to if it knows the speech is *media*. A robot assisting people with Autism Spectrum Disorder may not interrupt when a user is engaged in *natural* conversation (to encourage social interaction), but may attempt to engage if it suspects the user is watching too much *media*. A customer service robot may decide whether or not to head in the direction of customers chatting in a store or may choose to disregard the speech if it is coming from a TV. An in-home robot may reach out for external assistance if a user is distressed, but may not if it realizes the speech is from an action movie on TV. Depending on the end goals of the system, the robot can use such a classification, along with other prudent factors, to help it in making decisions.

To precisely characterize the differences between audio from *natural* and *media* scenarios is a challenge. Both of these audio categories contain human voices. Both categories contain diverse audio with similarities that make it difficult to quantify how we, as humans, usually know which of the two we are listening to. One potential discriminatory criterion, for example, is the speech patterns in the scripted conversation of television shows as opposed to the more spontaneous nature of impromptu conversation. This could be sufficient for categorizing a sitcom as *media*, but this does not help us in correctly classifying a radio podcast where the host is casually interviewing a guest. One could also try to make this classification based on if they hear cleanly engineered audio, like that produced in a studio, versus the noisy, distorted *natural* audio environments of everyday life. This can help with correctly classifying a TV show or movie played on a good sound system as *media*, but will not help when listening to sports, which involve crowd and audience noise. Solely detecting the presence of electronically sourced audio (i.e., coming from the speakers of a computer or television) is also not enough. Video calls with friends are *natural* situations in which there is electronic-sourced audio, along with at least one organically-sourced (i.e., coming directly from human vocal cords) speaker playing an active role in the conversation. If we know that some part of the audio is organically sourced, we can be sure that there is a co-located, physically present person talking. But, it can sometimes be tough to know if this is the case, especially if electronic audio sounds *natural* (e.g., conversational) and is played on a high-quality sound system. Making the classification between audio that is *natural* or *media* is hard.

For this paper, we focus on being able to classify between *natural* and *media* audio from the dynamic environment of the home. We focus on differentiating between speech from popular genres of *media* that is originating from loudspeakers and speech from *natural* conversations including at least one co-located person in the home.

Ideally, robots in real-world environments would have the ability to make this classification, regardless of the acoustic environments they are in (e.g., different rooms, different loudspeakers, distances from the audio source) and the different audio content that they hear (e.g., different voices, different TV/radio shows, background noise). Social roboticists that deploy robots in the home and intend to use audio to make decisions on how their robots interact with users can benefit from this work.

Our main contributions are:

- Describing a salient audio problem that social robots in the home face: the classification between a) *natural* conversation including at least one co-located user and b) *media* playing from electronic sources that does not require a social response
- Training classifiers[1] that use in-home audio to differentiate between *natural* and *media*, and evaluating how well the classifiers generalize to new recordings, acoustic conditions, and environments
- Proposing a classification pipeline that can provide additional, situational context to a social robot by assisting it in detecting social presence using sound

The organization of the paper is as follows: Section 2 offers background and related work. Section 3 describes the methodology in collecting the dataset, in selecting and extracting features of the audio, and in selecting the classification algorithms. Section 4 describes the experiments used to test the generalizability of the classifiers, and discusses the results. Section 5 discusses how these classifiers can be applied in practice, with details on timing and size of each, a proposed classification pipeline, and a discussion on ethics and privacy considerations. Section 6 discusses some limitations of the work and Section 7 concludes the work.

2 BACKGROUND

According to a recent U.S. Bureau of Labor Statistics survey [44], watching television was the most popular and time-consuming leisure activity in an American's average day, with people spending close to three hours watching TV. In comparison, activities such as eating, drinking, socializing, and communicating amount to approximately two hours total a day. These everyday domestic situations involve humans engaging with *media* (e.g., watching television) or *natural* situations (e.g., participating in a conversation at the dinner table).

2.1 In-Home Virtual Assistants

Popular virtual assistants, such as Amazon's Alexa, have already been integrated into many homes around the U.S. They use audio-based techniques that make them effective in the household. Source localization approximates the origin of audio input and wake-word detection [26] prompts sending the speech command to the cloud for natural language processing [25]. These features inform the assistant's decision-making policy to effectively and appropriately respond [29, 35]. These in-home systems do not incorporate much, if any, contextual awareness of their surroundings [40]. In fact, these systems typically require specific and explicit user prompts to engage them (e.g., "Alexa"). Because these systems are user-initiated, the detection of social context is much less necessary. Yet, for systems designed to interact with users autonomously, the ability to garner context about the environment is crucial [28].

We believe that virtual assistants can also benefit from the ideas presented in this paper, especially if developers believe there is value in additional functionality that includes behaving more socially and independently. Although we will focus on social robots in this paper, we note that social presence through sound can be of use to any device in the home that could utilize such context to help it make decisions.

[1]A link to our trained models and the code used to create the input feature vectors for our models: https://github.com/ScazLab/social-presence-sound

2.2 Using Audio for Activity and Event Detection in the Home

Automatic recognition of user activity in dynamic, unstructured environments, like the home, is important for systems whose primary purpose is to support their users through social means. Having some understanding of a user's activity and social context can help the system in its decision making.

Audio scene classification (ASC), or the identification of the environment or activity based on acoustic signals, is important for robotics and can help better facilitate human-robot interaction [3]. ASC has become a trending topic with growing interest because of the advent of smart homes and robots [14, 45, 47]. In recent years, audio analysis capabilities have been added to assistive robotic systems, such as the TIAGo service robot [19] and RiSH, a robot-integrated smart home for elderly care [13], with the goal that audio will provide more contextual awareness. Work for audio analysis in the home includes activity detection specific to helping the elderly by detecting falls [38] or by identifying common activities, to help medical staff monitor people who utilize ambient assisted living services [2, 11, 36]. Audio scene classification has also been used in the context of differentiating between specific kitchen sounds like the mixer, dishwasher, and utensils clanking [45], bathroom sounds like showering, washing hands, and flushing [9], breathing or snoring [17], or common sounds including keyboard typing, applause, and phone ringing [42]. Traditional machine learning classifiers have been used for these classifications with success.

Work has also been done that involves classifying in-home audio with the help of humans-in-the-loop. Some of this work includes human-assisted sound event recognition for home service robots for the elderly, where a human caregiver helps provide a robot with in-the-loop labels to non-voice sounds, in order to help a robot actively learn auditory events [12]. Additional work has used audio to classify different rooms in the home, like the kitchen and office, and also discriminated between nonverbal sounds like clapping and one-word speech scenarios [30].

The research area of voice activity detection (VAD) looks to classify between audio that contains speech and non-speech [20]. Research has been done to use noise cancellation to better implement VAD on smart home devices [22]. Other VAD work includes enhanced speech detection for humanoid robots in sparse dialogue [24] and robust classification between speech and non-speech [39] in noisy environments. Work has been done to recognize emotional states from speech using a support vector machine [41], to separate speech from music [1], and to detect and classify noises in speech signals [33].

There has also been research looking into how to accurately discriminate between speech commands produced from an electronic speaker from organic human speech [6]. This approach was discussed in the context of cybersecurity to better identify replay attacks of certain commands on Internet of Things devices, by focusing on determining the origin of pre-written speech commands, but does not focus on in-home, noisy experimentation.

Our work presents a new tool that can be used by robots in the home to gather more social context about a user's social presence through sound, when presented with human speech. The classification between *natural* and *media* that we focus on in this work encapsulates common speech scenarios in the home, that can give insight into people's activities. Our experimentation focuses on real-world audio recorded in noisy, in-home environments, and this work adds to the research area of activity detection in a dynamic environment.

2.3 Audio Classification of Media

Work has also been done in the classification of different forms of *media*. Audio information has also been utilized when researching genre classification in different forms of media. Music information retrieval methods have explored classifying songs into genres such as pop, rock, or blues [5, 43] and television *media* classification has classified videos into genres such as cartoons, news, or weather forecasts [15]. A key aspect of many of these media approaches, along with the in-home activity detection of Section 2.2, involves extracting time and frequency domain features (e.g., spectral contrasts, spectral roll-offs, Mel-Frequency Cepstral Coefficients, or

chroma features) from the overall audio signal and using these features to inform and train machine-learning classification algorithms. We build on this work by using similar features in our analysis, and discuss more background and motivation of the feature selection in Section 3.2.

3 METHODOLOGY

In this section, we describe how we (a) compiled an audio dataset containing the *natural* and *media* classes, (b) extracted features from each audio sample, and (c) selected the machine learning classifiers that we experimented with. We define two terms that we will be using throughout this paper. First, when discussing a **sample**, we are referring to a 5-second segment of audio that has been recorded and is used in feature extraction. A **recording** is a collection of contiguously captured *samples* during a given time window.

3.1 Audio Sample Collection

We collected audio content from various television genres and radio shows (sound from electronic speakers) and human speakers (sound from human voices). The final dataset contained approximately 30 hours of audio recordings, and was well-balanced between the *media* and *natural* classes.

Both categories were recorded on Kinect One microphones. This was important because any decisions made by a machine learning classifier would be able to focus on the difference of the audio content, rather than discrepancies caused by different recording hardware.

3.1.1 ***Media Recording Set.*** Our *media* (M) recording set consisted of a variety of TV shows or radio recordings, that we recorded on the Kinect One[2]. We focused on collecting audio recordings from popular television genres, which include drama, comedy, participatory/reality, news, and sports [46], as well as audio from radio shows. This category was recorded in different rooms, using a variety of electronic speakers[3], with the microphone capturing audio at varying distances from the speakers, during different contiguous time windows. Recording during different time windows allowed for different background and ambient noise to be captured as a part of the various recordings. All audio recordings were recorded at a rate of 16 kilohertz (kHz) in the waveform audio file format (.wav).

Each room, speaker, and microphone position configuration is referred to as its own unique *label*. These different recording configurations emulate a variety of recording conditions that an in-home agent might face. The distribution of the audio in each label can be seen in Table 1. There are 60 *media* recordings in our dataset, with a total of 10,138 samples, for around 14 hours of audio. Depending on the experiment that we performed, a different split of the recordings in the *media* set was used as training and testing data (explained in more detail in Section 4).

3.1.2 ***Natural Recording Set.*** The *natural* recording set can be broken down into three categories: CHiME5 (C), Video Calls (V), and Family Conversations (F).

Natural Audio from CHiME5. Category C recordings were comprised of content from the CHiME-5 dataset [4], available online. CHiME-5 contains audio captured from dinner parties in different houses. Each dinner party involved a different group of four people, who were told to engage in natural conversation in the house's kitchen, dining room, and living room for at least 2 hours.

Category C contained audio from 10 different CHiME-5 sessions. Each session contained audio from six Kinect microphone arrays, placed in different locations (bedroom, kitchen, living room) in each home, with audio input from each channel of each microphone. We used audio from the different Kinect microphones within the same

[2]We recorded the media recordings being emitted through electronic speakers, instead of inputting the media audio file directly into the classifier, because this is how a robot in the home would be capturing the media audio.

[3]The specific speaker models are as follows: Bose SoundLink 359037-1300 Mini Bluetooth Speaker (Bose), MacBook Pro 13" (Mac), iPhone 11 Pro (iPhone), Bose Wave Music System II (BigBose), 40" Eco Bravia VE5 Series LCD HDTV (SonyTV)

dinner party in our dataset because we wanted a diverse set of audio captured from different locations with varying acoustic properties. For the C category, we considered a *recording* to be all of the audio collected from a unique CHiME-5 session. The CHiME-5 audio files were in the waveform audio file format (.wav), with a recording rate of 16kHz. We chose CHiME-5 because it captured *natural*, social scenarios that one can expect to find in a home environment. We input the CHiME-5 files directly into the classifier because this is how *natural* audio would be captured by the robot. In total, category C contained 10,130 samples (1013 samples per recording). This sample number is equivalent to approximately 1.4 hours per CHiME-5 session, for a total of almost 14 hours of audio. Samples from the C category were used as our *natural* training data.

Natural Audio from Our Home Environments. We also captured *natural* audio from our own homes. We had Institutional Review Board approval to record audio in homes and to extract and analyze acoustic features. There were two categories that we experimented with, involving *natural* scenarios from 6 rooms in 3 different homes. We left a recording microphone in locations that we deemed appropriate for an in-home robot or device to be placed, recorded audio, and later inspected the audio. Audio from these two categories was used as our *natural* testing data.

Category V captured audio from video calls taking place in a home's office, dining room, and living room. These recordings involved conversations between members of a family consisting of two children and three adults. Members of the family congregated in their dining room and spoke over a video call on a laptop and phone using Zoom or Facebook Messenger. The calls were all on speaker. As a result, voices were variably distant from the microphone and the recordings captured by the Kinect included a mixture of voices coming from an organic source (the person in the same room as the Kinect microphone) and from electronic sources (the people on the video call). The same person was physically in the room with the Kinect for each of these recordings. Category V included six separate recordings, with a total of 917 samples.

Category F consisted of audio collected from family conversations in kitchens and living rooms, in three different homes. The microphone was placed close to where people were dining and conversing. An example location for the microphone was on a counter in an open, spacious kitchen. The kitchen recordings included some background noises such as the running sink, clanking utensils, and plates and glasses moving, while the living room recordings happened with little to no noise in the background. Category F included 965 samples and seven separate recordings, including voices from 11 different people.

There are multiple reasons that we decided to also collect *natural* audio that we recorded ourselves, despite having an extensive corpus of in-home, *natural* audio from CHiME5. Even though we tried to collect our *media* sample set with similar recording characteristics (i.e., microphone and sampling frequency) to CHiME5, we wanted to see whether or not classifiers trained solely on CHiME-5 could generalize to classifying other *natural* audio from outside of that corpus. This could show that these classifiers are able to correctly disambiguate between *natural* and *media* recorded by us, and that that the classification is not just a result of some discrepancies in how CHiME-5 was collected and how we recorded our audio. Lastly, we wanted to be able to experiment with the case of social presence that includes a mixture of electronic audio and organic-sourced *natural* audio, captured in the V dataset. This circumstance indicates social presence because at least one user that is co-located with the robot is engaged in a natural conversation, while chatting on a call with others. Samples from the V and F categories were used as our *natural* testing data.

3.2 Feature Extraction

We split our entire audio dataset into 5-second samples. From each sample, we extracted features to create an input vector that was used to train machine learning classifiers. We used the LibRosa Python package [31] to extract audio features. These are commonly used features in audio analysis (as mentioned in Section 2.3), which was the motivation for using them.

Table 1. Media Data Set Composition

Label	Room	Speaker	Kinect Distance	Total Samples	Recordings
A	Bedroom	Bose	1 ft	617	6
B	Bedroom	Bose	9 ft	852	7
C	Bedroom	Mac	6 ft	2075	11
D	Playroom	Bose	1 ft	627	5
E	Playroom	Bose	5 ft	517	9
F	Playroom	Bose	10 ft	332	3
G	Playroom	iPhone	1 ft	423	3
H	Playroom	iPhone	4 ft	372	2
I	Kitchen	Bose	9 ft	543	2
J	Kitchen	BigBose	9 ft	1185	5
K	Kitchen	SonyTV	4 ft	1432	2
L	Kitchen	iPhone	6 ft	201	1
M	Kitchen	Mac	6 ft	551	3
N	Kitchen	Mac	1 ft	411	1

In total, 83 features were extracted from each audio sample. We performed a standard transformation of each feature to normalize the feature set. The input vector contained the features below for each audio sample:

- *Mel-frequency cepstral coefficients (MFCCs)*: These are dominant features that have been historically used in speech recognition and they have been explored in separating music and speech [27]. It is typical that 13 coefficients are used for speech representation [43], so we use the means and standard deviations for each of the first 13 coefficients over the sample, for a total of 26 features.
- *Chroma Energy Normalized Statistics (CENS)*: These are features that have been used in audio analysis research to match similar audio [34]. There are 12 chroma classes and we use the mean and standard deviation for each chroma class over the sample, for a total of 24 features.
- *Root-mean-square (RMS) energy values*: Energy features are commonly used in audio analysis, with some prior work finding that the combination of energy with MFCC is better than using MFCCs alone [23]. We use the range, standard deviation, and skewness of this feature, for a total of 3 features.
- *Zero-crossing rates*: These are features that are commonly used in audio analysis [15] and can help provide a measure of noisiness of the audio sample [43]. We use the mean, standard deviation, and skewness, for a total of 3 features.
- *Tempo*: This feature estimates the beats per minute in the audio sample. The motivation behind adding this is that music from TV or radio commercials typically have more tempo than conversational audio in the home. This is 1 feature.
- *Spectral centroid, flatness, rolloff, and bandwidth*: These are also commonly used low-level components of the audio signal [10, 43]. We use the mean, standard deviation, and skewness for each, for a total of 12 features.
- *Spectral contrast*: These are features that have been shown to discriminate among different music genres [23], so we use the means and standard deviations for seven sub-bands, for a total of 14 features.

Note that none of these features involve transcription or semantic representation of dialogue/words in the audio environment. This way, the audio is translated into a machine readable format that has little to no meaning to a human, as opposed to words, which are used in lexical analysis in Natural Language Processing. This is an

arguably less invasive and more privacy-sensitive approach than using words, especially if the robot is intending on sending the input vector to the cloud to be analyzed.

3.3 Classification Algorithms

In our experiments to determine if our classification problem can be solved, we trained and tested different models with six traditional machine learning classification algorithms, using the sci-kit learn Python library [37]. These are commonly used algorithms for audio classification tasks (see Section 2 for more details). We performed an experimental evaluation of various approaches, to see which classifiers would be best suited to tackle the problem. We experimented with the following algorithms:

- KNeighborsClassifier [18]
- DecisionTreeClassifier [7]
- QDA (Quadratic Discriminant Analysis) [21]
- LogisticRegression [49]
- GaussianNB (Gaussian Naive Bayes) [48]
- SVC (C-Support Vector Classification) [8, 16]

We use these traditional classifiers instead of deep learning techniques, which have gained popularity in recent years in the audio analysis space, for multiple reasons. First, our dataset is modestly sized, and traditional ML algorithms have a much better chance at performing successfully than deep learning when the dataset is not very large. Second, we know the feature space that we want to use for this classification task. Lastly, we are hoping to be able to use these trained classifiers on real-time systems, so the response time needs to be quick and the complexity and space taken by the classifier needs to be reasonable (many social robots have limited compute power).

A gridsearch on each classification algorithm measured what hyperparameter combination was the best for each algorithm on our first experiment (described in Section 4.1). The different hyperparameter combinations for each classifier that were experimented with can be found in Appendix A. The hyperparameters that led to the highest performance, and were subsequently selected for the classifier in all of the following tests can be seen in Appendix B.

4 EXPERIMENTS AND RESULTS

In this section, we describe how the various classifiers performed on experiments that tested the classifiers' abilities to generalize to novel recordings, environments, and conditions. We test how well classifiers perform on a leave-one-recording-out cross validation, where we test on recordings that were left out of the training set. We also test how well the classifiers generalize to classifying *natural* recordings from outside of the training corpus and to *media* recordings from (1) rooms, (2) speakers, (3) microphone positions, and (4) combinations of all three, that they were not trained on.

4.1 Leave-One-Recording-Out Cross Validation

We performed an evaluation similar to a leave-one-out cross-validation (LOOCV), but in our case, leave-one-recording-out cross-validation **(LOROCV)**[4]. To perform LOROCV, we trained models using *natural* recordings from our CHiME-5 category (C) and *media* recordings from our media (M) recording set. For each fold of LOROCV,

[4]A conventional splitting of all of the samples into a train, test, and validation set would not be very insightful because many of our data samples were part of the same contiguously recorded audio clips (recordings). For any given recording in our dataset, there were at least 15 samples that were a part of the same original audio recording. When randomly shuffling the dataset for the train/test/validation splits, it is likely that some of a recording's 5-second samples land in each of the folds and the test and validation sets. Since audio from the same recording is inherently similar, we performed a cross validation per recording.

we trained on all recordings except for one from C and one from M. We did this for all possible pairs of recordings from C and M, which resulted in 600 folds (the Cartesian product of the 10 recordings in C and the 60 recordings in M). For each fold, we tested our classifier on the 1) left-out {C,M} recording pair, 2) left-out M recording and *natural* audio sampled from V, 3) left-out M recording and *natural* audio sampled from F, and 4) left-out M recording and *natural* audio sampled from both F and V. Because recordings can be of different lengths, we randomly sampled from the larger recording to match the size of the smaller recording. This ensured that we had balanced test sets each time.

The metrics that we recorded for all of our experiments are below. TP is a true positive, TN is a true negative, FP is a false positive, and FN is a false negative.

- Accuracy=(TP+TN)/(TP+TN+FN+FP)
- Precision = TP/(TP+FP)
- Recall = TP/(TP+FN)
- F1 Score = (2*Precision*Recall)/(Precision+Recall)

We recorded the precision, recall, and F1 scores for both the *media* and the *natural* classes (i.e., we treated both as the positive class). Both the macro averages (arithmetic mean) and micro averages (weighted average) were recorded across all folds. The full results for LOROCV can be found in Table 13 in Appendix D, with a summary in Table 4 in Appendix C.

With LOROCV, we test on *natural* audio from left-out CHiME-5 sessions (new voices and rooms from new homes within the CHiME-5 corpus), or better yet, on *natural* audio from the V or F categories that we recorded ourselves. We also test on unseen *media* recordings that the classifiers have not trained on and that we have recorded ourselves. This provides insight into how the trained algorithms can generalize to classifying novel recordings of *media* and *natural* audio.

4.2 Leave Out Rooms, Speakers, and Microphone Positions in the *Media* Set

We can gain further insight into how robustly the classifiers can differentiate between *natural* and *media* audio, if *media* in the training set contains recordings from different acoustic conditions (e.g., rooms, loudspeakers, microphone distances) than *media* in the testing set. In the experiments in this section, we evaluate how our classifiers perform when toggling which condition(s) of the *media* recording set to leave out of the training set. We also use the *natural* audio from the C category to train our models. We test on the *natural* V and F categories that we recorded ourselves and on the left-out *media*.

We left all of the *media* samples of a specific (1) room, (2) speaker, (3) microphone position, or (4) combinations of the three, out of the training set, and tested on the left out *media* samples and on *natural* samples from the V and F test categories. We matched the number of *media* samples in the training set with an equally distributed, random sample of 5-second samples from each *natural* recording in category C. We randomly sampled from all of the recordings in the larger test subset to match the size of the smaller subset. This ensured that we had balanced test sets each time. We recorded the micro and macro averages of precision, recall, and F1 scores for both the *media* and *natural* classes, as in LOROCV. The following paragraphs describe each experiment that we performed.

In Leave One Label Out (**LOLO**), we wanted to see how well classifiers would perform when they trained on *media* from specific *labels*, or specific room, speaker, and Kinect distance configurations (see Table 1), along with *natural* from category C, and then were tested against configurations that they were not trained on. We performed a Leave One Label Out (LOLO) experiment on all labels of our *media* data, where we trained different models using all the recordings from all combinations of labels, and tested against the held out labels. The left out media data at each fold was tested along with *natural* audio from the category V, F, and V+F datasets. The full results for each classifier can be found in Table 14 of Appendix D, with a summary in Table 5 of the Appendix C.

In Leave One Room Out (**LORO**), we wanted to see how well classifiers would perform when they trained on *media* from specific rooms, along with *natural* from category C, and then were tested against *media* from a room they had not trained on. This is important because each room has a different acoustic environment and layout. The classifiers should be able to make accurate predictions regardless of if they have trained on audio from the room in which they are deployed. In LORO, classifiers test on *media* recordings from a room that they have not trained on, but the test set includes loudspeakers and microphone distances that they have trained on. The left out media data at each fold was tested along with *natural* audio from the category V, F, and V+F datasets. The full results for each classifier can be found in Table 15 of Appendix D, with a summary in Table 6 of Appendix C.

In Leave One Speaker Out (**LOSO**), we wanted to see how well classifiers would perform when they trained on *media* from specific loudspeakers, along with *natural* from category C, and then were tested against *media* from loudspeakers they had not trained on. This is important because each loudspeaker has different hardware properties and the classifiers should be able to make accurate predictions regardless of if they have trained on audio from the loudspeaker from which they hear audio. In LOSO, classifiers test on *media* recordings from a loudspeaker that they have not trained on, but the test set includes rooms and microphone distances that they have trained on. The left out media data at each fold was tested along with *natural* audio from the category V, F, and V+F datasets. The full results for each classifier can be found in Table 16 of Appendix D, with a summary in Table 7 of Appendix C.

In Leave One Distance Out (**LODO**), we wanted to see how well classifiers would perform when they trained on *media* from certain microphone distances from a loudspeaker, along with *natural* from category C, and then were tested against *media* from microphone distances they had not trained on. This is important because the robot might be at variable distances from the sound source. In LODO, classifiers test on *media* recordings from a microphone distance that they have not trained on, but the test set includes loudspeakers and rooms that they have trained on. The left out media data at each fold was tested along with *natural* audio from the category V, F, and V+F datasets. The full results for each classifier can be found in Table 17 of Appendix D, with a summary in Table 8 of Appendix C.

In Leave One Room and Speaker Out (**LORSO**), we wanted to see how well classifiers would perform when they were tested on *media* rooms and speakers that they had not trained on. This is a more robust test than the previous ones. In LORSO, classifiers test on *media* recordings from a room and speaker that they have not trained on, but the test set includes microphone distances that they have trained on. The left out media data at each fold was tested along with *natural* audio from the category V, F, and V+F datasets. The full results for each classifier can be found in Table 18 of Appendix D, with a summary in Table 9 of Appendix C.

In Leave One Room and Distance Out (**LORDO**), we wanted to see how well classifiers would perform when they were tested on *media* rooms and microphone distances that they had not trained on. In LORDO, classifiers test on *media* recordings from a room and microphone distances that they have not trained on, but the test set includes microphone distances that they have trained on. The left out media data at each fold was tested along with *natural* audio from the category V, F, and V+F datasets. The full results for each classifier can be found in Table 19 of Appendix D, with a summary in Table 10 of Appendix C.

In Leave One Speaker and Distance Out (**LOSDO**), we wanted to see how well classifiers would perform when they were tested on *media* speakers and microphone distances that they had not trained on. In LOSDO, classifiers test on *media* recordings from a loudspeaker and microphone distances that they have not trained on, but the test set includes rooms that they have trained on. The left out media data at each fold was tested along with *natural* audio from the category V, F, and V+F datasets. The full results for each classifier can be found in Table 20 of Appendix D, with a summary in Table 11 of Appendix C.

In Leave One Room, Speaker, and Distance Out (**LORSDO**), we wanted to see how well classifiers would perform when they were tested on *media* speakers, rooms, and microphone distances that they had not trained on. This is the most challenging test that we perform for the classifier. In LORSDO, classifiers test on *media*

recordings from a room, loudspeaker, and microphone distance that they have not trained on. The left out media data at each fold was tested along with *natural* audio from the category V, F, and V+F datasets. The full results for each classifier can be found in Table 21 of Appendix D, with a summary in Table 12 of Appendix C.

Table 2. Experiment Summary. The table shows the average of the macro average F1 scores ($(F_{natural} + F_{media})/2$) for each classifier across all folds of each experiment. The table shows the average results of the trained classifiers being tested on the left out *media* sets along with *natural* recordings from the V and F categories. The classifier with the best average performance on each test set and experiment is in bold. More comprehensive results can be found in the Appendix.

Experiment	Test Set	KNN	QDA	DT	GNB	LR	SVC
LOROCV	V+M	**94.5**	89.0	87.9	86.0	87.9	91.3
	F+M	87.1	**99.5**	98.9	96.2	98.9	96.6
	F+V+M	91.3	93.3	92.8	90.5	92.8	**93.7**
LOLO	V+M	78.5	**99.3**	96.0	90.4	91.7	93.1
	F+M	85.6	88.8	77.9	82.4	82.9	**90.4**
	F+V+M	82.5	**93.4**	85.8	85.5	86.4	91.7
LORO	V+M	77.4	**99.2**	94.3	81.0	85.0	94.9
	F+M	83.1	86.1	75.1	83.5	82.7	**86.5**
	F+V+M	80.4	**92.6**	84.8	82.9	84.5	90.6
LOSO	V+M	76.9	98.5	**99.0**	93.8	97.1	93.4
	F+M	83.6	84.4	75.2	80.7	84.3	**87.5**
	F+V+M	80.9	**91.0**	86.6	87.0	90.4	90.3
LODO	V+M	78.9	**98.8**	97.3	91.6	97.8	94.8
	F+M	74.3	83.6	63.9	71.6	81.6	**84.1**
	F+V+M	77.7	**91.0**	81.4	81.8	89.6	89.4
LORSO	V+M	67.9	86.6	**87.9**	70.1	85.6	86.3
	F+M	82.1	82.9	78.1	76.7	87.2	**88.1**
	F+V+M	76.1	85.1	83.1	74.9	86.9	**87.6**
LORDO	V+M	70.0	92.5	**95.3**	78.5	88.8	87.4
	F+M	78.3	86.7	77.9	77.1	89.4	**90.5**
	F+V+M	75.0	89.4	85.8	78.0	**89.6**	89.5
LOSDO	V+M	76.3	90.9	90.1	85.2	**95.3**	94.5
	F+M	79.1	80.8	76.6	78.3	83.3	**84.8**
	F+V+M	77.9	85.5	82.8	81.4	89.0	**89.5**
LORSDO	V+M	65.8	82.4	**87.2**	70.0	86.5	85.3
	F+M	73.0	77.1	72.7	66.7	83.7	**85.2**
	F+V+M	69.9	79.5	79.6	68.2	85.0	**85.3**

4.3 Selecting a Classifier

In general, we see that most of the trained classification algorithms perform well on our experiments. We see that most of the classifiers have average F1 scores in the 90s or 80s for a majority of the experiments. Table 2 summarizes the results for all our experiments, for each classifier.

4.3.1 Results. We see that SVC has the best performance on the most tests throughout our experiments. SVC has the highest average F1 score on 12 out of the 27 tests, with the highest average F1 score on 7 out of the 12 more difficult tests (where two or three of the *media* parameters are left out of the test set in LORSO, LORDO, LOSDO, and LORSDO). SVC has the highest performance on the F+V+M test sets on all but one of the more difficult experiments, and SVC has the highest F1 score on the F+M test sets for almost all of the experiments. On LORSDO, the most difficult experiment, SVC has the best performance on two out of three of the tests (V+M and F+M). Despite not having the highest scores on V+M, it does consistently well on the test set, throughout all of the experiments. Generally, SVC is the most consistent classifier across the different test sets and experiments, and is always performing with high F1 scores.

The next best classifier in terms of leading F1 scores is QDA, which has 7 of the best F1 scores. For QDA, all of these top results come in the first five experiments, where the training data includes more of the acoustic environment and conditions than in the last four experiments. QDA performs very strongly on the V+M test sets and on the F+V+M test sets for these experiments. This shows that if the training set has certain qualities similar to the test set, QDA could be a legitimate option for classifying between *natural* and *media*. However, the classifier that performs the best when the test data is most dissimilar to the training data is SVC. QDA does reasonably well, but performs overall worse than SVC in the last four experiments, especially on the F+M and the F+V+M datasets. QDA could be a good option alongside SVC if we know that the testing environment and conditions will have similarities to the training set.

DT has the top average F1 scores on 4 of the tests. DT does very well when classifying the V+M test set, with high scores on three out of four of the V+M tests in the more difficult experiments. Except for LOROCV, DT performs very well on the V+M test sets on all of the experiments. However, there is a significant tradeoff seen in how well DT performs on the F+M test sets. DT might be very good at classifying between *natural* and *media* with *natural* video calls and *media* in the test set, but does very poorly at classifying *natural* family conversations. In this regard, SVC is better overall for its consistency across both the V+M and F+M test sets.

LR has the top average F1 score on only 2 of the tests, however we see that LR is able to generalize well to new *media* and *natural* audio. LR performs very well in many of the experiments, with F1 scores that are close to, albeit slightly worse than, SVC in most of the experiments. Especially in LORSO, LORDO, LOSDO, and LORSDO, we see that LR is able to perform consistently well on V and F data, with scores similar to that of SVC on the F+V+M datasets. LR does a good job at generalizing to new environments that it has not trained on for left out *media* data, and video calls and family conversations. However, QDA is better than LR when the training set is more similar to the test set, and SVC is better than LR when the test set is more dissimilar.

KNN and GNB have the worst performances on our experiments. KNN performs the best on V+M in LOROCV, but besides that, KNN and GNB show substantially worse performance than the other classifiers. They perform particularly poorly on LORSDO, which tests how well they can generalize when training on very dissimilar *media* data to the test set. We would not recommend KNN or GNB, especially when compared to our other trained classifier.

4.3.2 Discussion. Overall, SVC is best able to generalize to new recordings. We see this both in SVC's ability to perform well on *natural* data that we recorded in our own homes, which was outside of the *natural* audio from the CHiME-5 corpus that the model was trained on, as well as good performance of the classifier to *media* from loudspeakers, microphone distances, and rooms that it was not trained on (Table 21). SVC performs consistently

well when tested on in-the-home, *natural* audio of both video calls (V) and family conversations (F). SVC performs with accuracies of over 85% on LORSDO, with recall scores of over 90% for natural V or F audio, and recall of over 81% for *media* data from a different room, loudspeaker, and microphone distance than it was trained on. We believe that SVC is the best classification algorithm that we experimented with at disambiguating between *natural* and *media*. It does the most consistently well across our tests sets in our experiments, and does the best at generalizing to new environments and conditions that it has not trained on.

LR also performs well on both of the *natural* test sets and on many of the experiments, but performs worse than SVC overall. QDA performs very well when tested against data with some similar characteristics to what it is trained with, but does more poorly on stricter generalizability tests. DT performs very well on video calls, but very poorly when tested against family conversations. KNN and GNB do not perform well.

Since QDA, LR, and SVC all perform well across all of our test sets and experiments, with QDA showing particularly strong performance when the *media* testing conditions have some similarities to their training coonditions, it could be an option to use an ensemble of classifiers in making the *natural* vs. *media* prediction. We need to verify that the classifiers do not take too long to make predictions and that they do not take too much space in memory. If these two statements hold true, it could be reasonable to use all three in predicting *natural* vs. *media*. We perform these timing and size experiments in Section 5.1.

5 PROPOSED APPLICATION

A critical criterion when selecting a classification algorithm is that it can perform in close to real-time to be suitable for a robot in the home or in the real world. A robot should provide a naturalistic and intuitive interaction for human users, so real-time classifications and responses are essential. Taking too much time to analyze the audio environment, extract features, make predictions, and act on those predictions may negatively affect the overall interaction. Keeping these factors in mind, we (a) perform several timing and size tests on various steps of the audio collection and decision-making process, (b) suggest an overall classification pipeline for a robot to implement this approach, and (c) present ethics and privacy considerations that were taken into account for this pipeline. For these timing and size experiments, we train the classifier on the entire *natural* C category that we compiled, and all of our *media* recordings.

5.1 Timing and Size Experiments

We measured the speed of feature extraction and prediction using around 45 minutes of audio data (540 5-second samples). Extracting features from each of the 540 audio samples took an average of 0.557 seconds (STD=0.0442 seconds) on a Dell Laptop with an Intel i5-5200U CPU @ 2.2GHz and 8GB RAM. To measure the average prediction time for each audio sample, we measured the time that it took to standardize and predict the entire (540x83) input vector and divided it by 540. The trained standardization scaler had a size of 4 kB. The average prediction times and the sizes on disk for each trained classifier can be seen in Table 3 below.

Table 3. Classifier Size and Prediction Times

Model	Avg. Prediction Time (ms)	Size (kB)
KNN	2.524	13,545
QDA	0.01064	115
DT	0.00117	12
GNB	0.00312	4
LR	0.00366	4
SVC	0.17480	668

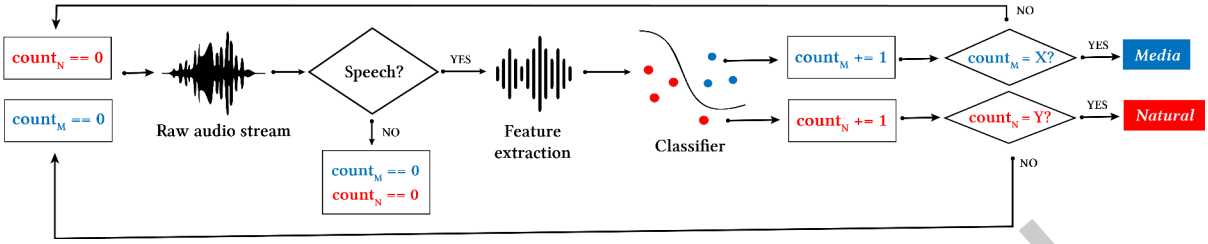

Fig. 1. Proposed classification pipeline. See Section 5.2 for description.

We see that all of the classifiers that we trained have fast prediction times. DT and GNB are the fastest, with LR and QDA next, then SVC, and KNN last. However, all the classifiers, except for KNN, are considerably faster than a millisecond, so we believe that any of the classifiers would be sufficient in that respect.

With respect to size on disk, LR and GNB are the smallest, with DT as next smallest. SVC is the second largest, but still not prohibitively big.

These sizes (and predictions) are also promising in that if the dataset were to get substantially larger, that most of these classification algorithms seem like they would be able to scale and still be reasonable to use on-board and real-time. This might not be true for KNN, but that was eliminated due to its poor performance on generalization.

This also means that after recording a 5-second sample, the whole classification process could be used on-board a robot, even on one with little memory. The whole classification process, after recording a 5-second sample, can take less than a second for feature extraction, standardization, and the prediction, making it possible to use this in real-time.

Furthermore, a robot could reasonably include multiple trained classifiers on disk and require less than one megabyte (MB) of space. If using an ensemble of classifiers, the prediction time still remains substantially lower than one millisecond. Both the timing and size of the classifiers together allow for an ensemble to be used.

5.2 Classification Pipeline

In a real-world setting, we suggest our classifier be used as a part of a greater classification pipeline, shown in Figure 1. A Kinect One microphone would be required[5], along with minimal onboard computing power. All audio collection, analysis, and computation can take place locally, without needing to offload any data to online services.

The system begins by recording a 5-second raw audio stream of the environment and initializing the count variables to 0. The system stores the recording and checks it for speech.[6] If speech is not detected, the system should loop back to the start by resetting the counts, and deletes the recording. If speech is detected, the feature extraction is performed, the audio is deleted, and a corresponding *natural* or *media* prediction is made. After a prediction, the corresponding *count* is incremented, and the other count is reset to 0. Only after X, or Y, consecutive predictions in a certain category will the decision be "final". Otherwise, the corresponding *count* is reset to 0. Once a final decision is output by the pipeline, the process starts again, with both counts initialized to 0.

[5]We did not test multiple microphones so we cannot say whether or not our classifier would have any success recording with a different microphone.

[6]Speech could be checked for by using a voice activity detection (VAD) algorithm, trained on the Kinect, that can detect when human speech is a part of the acoustic environment.

Depending on how sensitive we want the system to be to the classifier's predictions, we can alter the values of X and Y. For example, with $X = 3$, the classifier will have to predict close to 15 consecutive seconds (three decisions in a row) as *media*. This approach does not allow for one false positive to ruin the final classification, but rather the classifier would have to get the audio scene wrong three times in a row in order to make a mistake.

An alternative approach is to set both X=1 and Y=1, in which case the pipeline will be returning a final prediction on every 5-second audio sample, unless it does not detect speech. This will give a robot using this pipeline more frequent data points to use in its final decision making.

After the system determines whether or not the speech that it hears in its environment is *media* or *natural*, it can use this classification, along with other contextual information to make decisions on how to act. For example, the robot could also have other tools available to it that can detect characteristics from human speech such as tone, emotion, and intensity. The robot could also utilize context like the time of day, the day of the week, its location in the home, the current weather, and more.

Another interesting contextual tool that could be incorporated into this pipeline is sound source localization (SSL), which utilizes the microphone array of the Kinect. SSL could help the robot get an approximation of where the speech is coming from. This extra context, combined with the *natural* vs. *media* classification, could further assist the robot in making a more informed decision on social presence and providing it with a better understanding its environment. VAD and SSL could be combined to localize and individually classify multiple speakers in a noisy audio scene, but such VAD for multi-speaker diarization in real-world scenarios remains an open research problem[32].

This classification pipeline can provide the robot with an understanding of if speech is *natural* or *media* in its environment, helping it in inferring social presence. The robot can use this information, along with other context, to make appropriate decisions about how to interact, or not, and to best accommodate its user(s) and to reach its goals.

5.3 Ethics and Privacy Considerations

In home data is inherently sensitive, and the audio pipeline presented in our paper is considerate of that. We believe our solution is minimally invasive. Using one modality (i.e., just audio) to make decisions is undoubtedly less invasive than using more. In fact, our suggested solution is computed locally (it is lightweight and would not require sending any sensitive data to online services), only needs to store a 5-second sample of audio at a time (which can be deleted immediately after features are extracted from it), and does not use any semantic representation or transcription of the audio (which could contain sensitive information) as a part of its decision making. These are important factors that keep users' privacy in mind.

6 LIMITATIONS

There are several limitations to this work that we believe are important to make clear. First, the dataset that we compiled could be more diverse and representative. Our *natural* training data is only comprised of audio from the CHiME-5 dataset, even though it does contain audio from different homes, rooms, and voices. Our *media* dataset contains three different rooms from within one home and five different electronic devices. Obviously, there are countless other possible devices from which audio can be emitted in the home, which were not included in our training set. Despite these limitations, our results showed that classifiers were able to make accurate *media* classifications on audio from recording devices, rooms, microphone distances, and combinations of the three that they were not trained on, and the classifiers were able to classify *natural* audio from outside of the CHiME-5 training corpus, that included new rooms and voices in the V and F test sets. Another limitation is that the recordings in our V and F categories could be more diverse and comprehensive, with the inclusion of

audio from more homes, families, and people. Also, we only focus on audio from the home, when ideally, such a classification tool should be able to make predictions in other dynamic, human environments as well.

Additionally, our dataset does not include examples of scenarios where *media* from television or radio shows is playing at the same time that *natural* conversation (that includes at least one co-located person) is occurring.[7] Further testing would be needed to see how our classifiers would perform when both *media* and *natural* audio are overlaid. We did see that in situations where electronic and organic speakers are conversing with each other in the audio scene (in our video calls test category), the classification algorithms classified the audio as *natural*. It could be beneficial if a robot could garner more detailed context of identifying, indexing, and classifying between each organic and electronic speaker engaged in the conversation, but we leave this as a future research direction. Regardless, through our experimentation in this paper, we see that the classifiers can provide important context to a robot by accurately differentiating between common speech scenarios in the home from which social presence can be implied: popular genres in *media* originating from loudspeakers and *natural* conversation including a co-located user.

7 CONCLUSIONS

Detecting social presence using sound involves being able to classify audio as containing either 1) *natural* conversation including at least one co-located user or 2) *media* playing from electronic sources that does not require a social response, such as television shows. It is important for in-home social robots to have such a capability, as the additional context can help them in their decision making. We perform an experimental evaluation that tests the robustness of several traditional machine learning classifiers on data from our compiled *natural* vs. *media* dataset. We conclude that a C-Support Vector Classification (SVC) algorithm outperforms other classifiers, and we propose a classification pipeline that can be utilized by social robots in the home to help them in detecting social presence using sound.

8 ACKNOWLEDGMENTS

Supported by the National Science Foundation (NSF) under award numbers 1813651, 2106690, 1928448, and 1955653 and the Office of Naval Research (ONR) award #N00014-18-1-2776. Rebecca Ramnauth is supported by the NSF Graduate Research Fellowship and the NASEM Ford Predoctoral Fellowship. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF and NASEM.

We would like to thank Kate Candon, Tesca Fitzgerald, Sydney Thompson, Marynel Vázquez, Debasmita Ghose, Morgan Vanderwall, and Eleonora Serbeti who helped improve the paper through their feedback.

REFERENCES

[1] Abdullah I. Al-Shoshan. 2006. Speech and Music Classification and Separation: A Review. *Journal of King Saud University - Engineering Sciences* 19, 1 (2006), 95–132. https://doi.org/10.1016/S1018-3639(18)30850-X

[2] Rosa Ma Alsina-Pagès, Joan Navarro, Francesc Alías, and Marcos Hervás. 2017. homeSound: Real-Time Audio Event Detection Based on High Performance Computing for Behaviour and Surveillance Remote Monitoring. *Sensors* 17, 4 (2017). https://doi.org/10.3390/s17040854

[3] Sumair Aziz, Muhammad Awais, Tallha Akram, Umar Shahbaz Khan, Musaed A. Alhussein, and Khursheed Aurangzeb. 2019. Automatic Scene Recognition through Acoustic Classification for Behavioral Robotics. *Electronics* (2019).

[4] Jon Philip Barker, Shinji Watanabe, Emmanuel Vincent, and Jan Trmal. 2018. The fifth 'CHiME' Speech Separation and Recognition Challenge: Dataset, task and baselines. *Interspeech* (2018).

[5] Roberto Basili, Alfredo Serafini, and Armando Stellato. 2004. Classification of musical genre: a machine learning approach. *ISMIR* (2004).

[7] Because the end goal of our *natural* vs. *media* classification is to help a robot in detecting a co-located user's social presence using sound, we would consider labeling this situation as *natural* because it includes conversational audio from a co-located person, and it implies that a user is physically present with the robot. However, it could be beneficial if a social robot could detect that there is both *natural* and *media* audio in the environment. Such knowledge could give it more nuanced context than purely a *natural* classification, but we leave this for future work.

[6] Logan Blue, Luis Vargas, and Patrick Traynor. 2018. Hello, Is It Me You're Looking For? Differentiating Between Human and Electronic Speakers for Voice Interface Security. In *Proceedings of the 11th ACM Conference on Security & Privacy in Wireless and Mobile Networks*. 123–133.

[7] L. Breiman, Jerome H. Friedman, Richard A. Olshen, and C. J. Stone. 1984. Classification and Regression Trees.

[8] Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A Library for Support Vector Machines. *ACM Trans. Intell. Syst. Technol.* 2, 3, Article 27 (may 2011), 27 pages. https://doi.org/10.1145/1961189.1961199

[9] Jianfeng Chen, Alvin Harvey Kam, Jianmin Zhang, Ning Liu, and Louis Shue. 2005. Bathroom Activity Monitoring Based on Sound. In *Pervasive Computing*, Hans W. Gellersen, Roy Want, and Albrecht Schmidt (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 47–61.

[10] Selina M. Chu, Shrikanth S. Narayanan, C.-C. Jay Kuo, and Maja J. Matarić. 2006. 'Where am i? scene recognition for mobile robots using audio features. *IEEE International Conference on Multimedia and Expo* (2006), 885–888.

[11] Vladimir Despotovic, Peter Pocta, and Andrej Zgank. 2022. Audio-based Active and Assisted Living: A review of selected applications and future trends. *Computers in Biology and Medicine* 149 (2022), 106027. https://doi.org/10.1016/j.compbiomed.2022.106027

[12] Ha Do, Weihua Sheng, and Meiqin Liu. 2016. Human-assisted sound event recognition for home service robots. *Robotics and Biomimetics* 3 (12 2016). https://doi.org/10.1186/s40638-016-0042-2

[13] Ha Manh Do, Minh Pham, Weihua Sheng, Dan Yang, and Meiqin Liu. 2018. RiSH: A robot-integrated smart home for elderly care. *Robotics and Autonomous Systems* 101 (2018), 74–92. https://doi.org/10.1016/j.robot.2017.12.008

[14] Bo Dong, Cristian Lumezanu, Yuncong Chen, Dongjin Song, Takehiko Mizoguchi, Haifeng Chen, and Latifur Khan. 2020. At the Speed of Sound: Efficient Audio Scene Classification. In *Proceedings of the 2020 International Conference on Multimedia Retrieval* (Dublin, Ireland) *(ICMR '20)*. Association for Computing Machinery, New York, NY, USA, 301–305. https://doi.org/10.1145/3372278.3390730

[15] Hazım Kemal Ekenel and Tomas Semela. 2013. Multimodal genre classification of TV programs and YouTube videos. *Multimedia Tools and Applications* (2013).

[16] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. LIBLINEAR: A Library for Large Linear Classification. *J. Mach. Learn. Res.* 9 (jun 2008), 1871–1874.

[17] Tim Fischer, Johannes Schneider, and Wilhelm Stork. 2016. Classification of breath and snore sounds using audio data recorded with smartphones in the home environment. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 226–230. https://doi.org/10.1109/ICASSP.2016.7471670

[18] Evelyn Fix and Joseph L. Hodges. 1989. Discriminatory Analysis - Nonparametric Discrimination: Consistency Properties. *International Statistical Review* 57 (1989), 238.

[19] Lacrimioara Grama and Corneliu Rusu. 2018. Adding audio capabilities to TIAGo service robot. In *2018 International Symposium on Electronics and Telecommunications (ISETC)*. 1–4. https://doi.org/10.1109/ISETC.2018.8583897

[20] J.A. Haigh and J.S. Mason. 1993. Robust voice activity detection using cepstral features. In *Proceedings of TENCON '93. IEEE Region 10 International Conference on Computers, Communications and Automation*, Vol. 3. 321–324 vol.3. https://doi.org/10.1109/TENCON.1993.327987

[21] Trevor Hastie, Robert Tibshirani, Jerome Friedman, and James Franklin. 2004. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. *Math. Intell.* 27 (11 2004), 83–85. https://doi.org/10.1007/BF02985802

[22] Seok-Hoon Kim Jeong-Sik Park. 2020. Noise Cancellation Based on Voice Activity Detection Using Spectral Variation for Speech Recognition in Smart Home Devices. *Intelligent Automation & Soft Computing* 26, 1 (2020), 149–159. https://doi.org/10.31209/2019.100000136

[23] Dan-Ning Jiang, Lie Lu, Hong-Jiang Zhang, Jian-Hua Tao, and Lian-Hong Cai. 2002. Music type classification by spectral contrast feature. *In Multimedia and Expo, 2002. ICME'02* 1 (2002), 113–116.

[24] Hyun-Don Kim, Jinsung Kim, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G. Okuno. 2008. Target speech detection and separation for humanoid robots in sparse dialogue with noisy home environments. In *2008 IEEE/RSJ International Conference on Intelligent Robots and Systems*. 1705–1711. https://doi.org/10.1109/IROS.2008.4650977

[25] Veton Këpuska and Gamal Bohouta. 2018. Next-generation of virtual personal assistants (Microsoft Cortana, Apple Siri, Amazon Alexa and Google Home). *2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC)* (2018).

[26] Ken'ichi Kumatani, Sankaran Panchapagesan, Minhua Wu, Minjae Kim, Nikko Strom, Gautam Tiwari, and Arindam Mandal. 2017. Direct modeling of raw audio with DNNS for wake word detection. *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)* (2017), 252–257.

[27] Beth Logan. 2000. Mel frequency cepstral coefficients for music modeling. *Proc. 1st Int. Symposium Music Information Retrieval.*

[28] Christof Mahieu, Femke Ongenae, Femke De Backere, Pieter Bonte, Filip De Turck, and Pieter Simoens. 2019. Semantics-based platform for context-aware and personalized robot interaction in the internet of robotic things. *Journal of Systems and Software* 149 (2019), 138–157.

[29] Thanassis Mavropoulos, Georgios Meditskos, Spyridon Symeonidis, Eleni Kamateri, Maria Rousi, Dimitris Tzimikas, Lefteris Papageorgiou, Christos Eleftheriadis, George Adamopoulos, Stefanos Vrochidis, et al. 2019. A context-aware conversational agent in the rehabilitation domain. *Future Internet* 11, 11 (2019), 231.

[30] J. Maxime, X. Alameda-Pineda, L. Girin, and R. Horaud. 2014. ‘Sound representation and classification benchmark for domestic robots. *IEEE International Conference on Robotics and Automation (ICRA)* (2014), 6285–6292.
[31] Brian McFee, Colin Raffel, Dawen Liang, Daniel P. W. Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. 2015. librosa: Audio and music signal analysis in python. *In: Proceedings of the 14th python in science conference*.
[32] Ivan Medennikov, Maxim Korenevsky, Tatiana Prisyach, Yuri Khokhlov, Maria Korenevskaya, Ivan Sorokin, Tatiana Timofeeva, Anton Mitrofanov, Andrei Andrusenko, Ivan Podluzhny, Aleksandr Laptev, and Aleksei Romanenko. 2020. Target-Speaker Voice Activity Detection: A Novel Approach for Multi-Speaker Diarization in a Dinner Party Scenario. 274–278. https://doi.org/10.21437/Interspeech.2020-1602
[33] Nobuyuki Miyake, Tetsuya Takiguchi, and Yasuo Ariki. 2007. Noise Detection and Classification in Speech Signals with Boosting. In *2007 IEEE/SP 14th Workshop on Statistical Signal Processing*. 778–782. https://doi.org/10.1109/SSP.2007.4301365
[34] Meinard Müller and Sebastian Ewert. 2011. Chroma Toolbox: MATLAB implementations for extracting variants of chroma-based audio features. *In Proceedings of the International Conference on Music Information Retrieval (ISMIR)* (2011).
[35] Bauyrzhan Ospan, Nawaz Khan, Juan Augusto Wrede, Mario Quinde, and Kenzhegali Nurgaliyev. 2018. Context Aware Virtual Assistant with Case-Based Conflict Resolution in Multi-User Smart Home Environment. *International Conference on Computing and Network Communications (CoCoNet)* (2018), 36–44.
[36] Sharnil Pandya and Hemant Ghayvat. 2021. Ambient acoustic event assistive framework for identification, detection, and recognition of unknown acoustic events of a residence. *Advanced Engineering Informatics* 47 (2021), 101238. https://doi.org/10.1016/j.aei.2020.101238
[37] Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, Edouard Duchesnay, and Gilles Louppe. 2011. E.Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.
[38] Héctor Peiteado, Inmaculada Hernáez, Artzai Picon, Javier Camarena, and Eva Navas. 2010. Audio Classification Techniques in Home Environments for Elderly/Dependant People. 320–323. https://doi.org/10.1007/978-3-642-14097-6_51
[39] T. Lakshmi Priya, N.R. Raajan, N. Raju, P. Preethi, and S. Mathini. 2012. Speech and Non-Speech Identification and Classification using KNN Algorithm. *Procedia Engineering* 38 (2012), 952–958. https://doi.org/10.1016/j.proeng.2012.06.120 INTERNATIONAL CONFERENCE ON MODELLING OPTIMIZATION AND COMPUTING.
[40] Reza Rawassizadeh, Taylan Sen, Sunny Jung Kim, Christian Meurisch, Hamidreza Keshavarz, Max Mühlhäuser, and Michael Pazzani. 2019. Manifestation of virtual assistants and robots into daily life: Vision and challenges. *CCF Transactions on Pervasive Computing and Interaction* 1 (2019), 163–174.
[41] Kai-Tai Song, Meng-Ju Han, and Shih-Chieh Wang. 2014. Speech signal-based emotion recognition and its application to entertainment robots. *Journal of the Chinese Institute of Engineers* 37, 1 (2014), 14–25. https://doi.org/10.1080/02533839.2012.751330 arXiv:https://doi.org/10.1080/02533839.2012.751330
[42] Andrey Temko, Robert Malkin, Christian Zieger, Dušan Macho, Climent Nadeu, and Maurizio Omologo. 2006. CLEAR Evaluation of Acoustic Event Detection and Classification Systems. In *Proceedings of the 1st International Evaluation Conference on Classification of Events, Activities and Relationships* (Southampton, UK) *(CLEAR'06)*. Springer-Verlag, Berlin, Heidelberg, 311–322.
[43] George Tzanetakis and Perry R. Cook. 2002. Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing* (2002), 293–302.
[44] United States Bureau of Labor Statistics. 2019. American Time Use Survey Summary. https://www.bls.gov/news.release/pdf/atus.pdf.
[45] Anastasios Vafeiadis, Konstantinos Votis, Dimitrios Giakoumis, Dimitrios Tzovaras, Liming Chen, and Raouf Hamzaoui. 2020. Audio content analysis for unobtrusive event detection in smart homes. *Engineering Applications of Artificial Intelligence* 89 (2020), 103226. https://doi.org/10.1016/j.engappai.2019.08.020
[46] A. Watson. 2019. Genre breakdown of the top 250 TV programs in the United States in 2017. https://www.statista.com/statistics/201565/most-popular-genres-in-us-primetime-tv/.
[47] Jie Xie and Mingying Zhu. 2019. Investigation of acoustic and visual features for acoustic scene classification. *Expert Systems with Applications* 126 (2019), 20–29. https://doi.org/10.1016/j.eswa.2019.01.085
[48] Harry Zhang. 2004. The Optimality of Naive Bayes. In *The Florida AI Research Society*.
[49] Ciyou Zhu, Richard H. Byrd, Peihuang Lu, and Jorge Nocedal. 1997. Algorithm 778: L-BFGS-B: Fortran Subroutines for Large-Scale Bound-Constrained Optimization. *ACM Trans. Math. Softw.* 23, 4 (dec 1997), 550–560. https://doi.org/10.1145/279232.279236

A HYPERPARAMETERS USED FOR GRIDSEARCH ON LEAVE-ONE-RECORDING-OUT CROSS VALIDATION

A.1 KNN

- "n_neighbors": [1,3,5,7,9]

- "weights": ['uniform', 'distance']
- "p": [1,2]

A.2 QDA

- "reg_param": [0.00001, 0.0001, 0.001,0.01, 0.1
- "tol": [0.0001, 0.001,0.01, 0.1]

A.3 DT

- "criterion": ['gini','entropy']
- "max_depth": [1, 5,10,None]
- "min_samples_split": [2,5,10]
- "min_samples_leaf": [1,2,5]

A.4 GNB

- "var_smoothing": [1e-2, 1e-3, 1e-4, 1e-5, 1e-6, 1e-7, 1e-8, 1e-9, 1e-10, 1e-11, 1e-12, 1e-13, 1e-14, 1e-15]

A.5 LR

- "solver": ['lbfgs','liblinear', 'newton-cg']
- "penalty": ['l1', 'l2']
- "C": [0.001,0.01,0.1,1,10,100,1000]

A.6 SVC

- "kernel": ['linear','rbf']
- "gamma": ['scale', 'auto']
- "C": [0.1,1,10,1000]

B HYPERPARAMETERS OF MODELS PRESENTED IN SECTION 4 RESULTS

Model	**Hyperparameters**
KNN	'algorithm'='auto', 'leaf_size'=30, 'metric'='minkowski', 'metric_params'=None, 'n_jobs'=None, 'n_neighbors'=8, 'p'=1, 'weights'='distance'
QDA	'priors'=None, 'reg_param'=0.01, 'store_covariance'=False, 'tol'=0.0001.
DT	'ccp_alpha'=0.0, 'class_weight'=None, 'criterion'='entropy', 'max_depth'=None, 'max_features'=None, 'max_leaf_nodes'=None, 'min_impurity_decrease'=0.0, 'min_impurity_split'=None, 'min_samples_leaf'=2, 'min_samples_split'=2, 'min_weight_fraction_leaf'=0.0, 'random_state'=0, 'splitter'='best'
GNB	'priors': None, 'var_smoothing': 0.001
LR	'C': 0.1, 'class_weight': None, 'dual': False, 'fit_intercept': True, 'intercept_scaling': 1, 'l1_ratio': None, 'max_iter': 100, 'multi_class': 'auto', 'n_jobs': None, 'penalty': 'l2', 'random_state': None, 'solver': 'lbfgs', 'tol': 0.0001, 'verbose': 0, 'warm_start': False.
SVC	'C': 10, 'break_ties': False, 'cache_size': 200, 'class_weight': None, 'coef0': 0.0, 'decision_function_shape': 'ovr', 'degree': 3, 'gamma': 'scale', 'kernel': "rbf", 'max_iter': -1, 'probability': False, 'random_state': None, 'shrinking': True, 'tol': 0.001, 'verbose': False

C EXPERIMENT F1 SCORE SUMMARIES

Leave-One-Recording-Out Cross Validation (LOROCV) Summary. We present the average F1 scores between each of the two classes across all LOROCV folds. For each fold, all of a room's *media* recordings were held out of the training set, and used in the testing set along with the *natural* audio from our own homes.

Test Set	Metrics	KNN	QDA	DT	GNB	LR	SVC
V + M	$F1_{natural}$	**95.1**	87.7	86.4	85.5	86.4	90.6
	$F1_{media}$	**93.8**	90.2	89.4	86.5	89.3	91.9
	Avg. F1	**94.5**	89.0	87.9	86.0	87.9	91.3
F + M	$F1_{natural}$	86.9	**99.5**	98.9	96.4	99.0	96.5
	$F1_{media}$	87.2	**99.5**	98.8	95.9	98.8	96.6
	Avg. F1	87.1	**99.5**	98.9	96.2	98.9	96.6
F + V + M	$F1_{natural}$	91.7	93.1	92.3	90.4	92.3	**93.4**
	$F1_{media}$	90.9	93.4	93.2	90.5	93.2	**93.9**
	Avg. F1	91.3	93.3	92.8	90.5	92.8	**93.7**

Leave-One-Label-Out (LOLO) Summary. We present the average F1 scores between each of the two classes across all 14 LOLO folds. For each fold, a *media* recording and *natural* C recording were held out of the training set, and used in the testing set along with the *natural* audio from our own homes.

Test Set	Metrics	KNN	QDA	DT	GNB	LR	SVC
V + M	$F1_{natural}$	81.5	**99.3**	96.9	92.9	94.9	94.1
	$F1_{media}$	75.5	**99.2**	95.1	87.9	88.5	92.1
	Avg. F1	78.5	**99.3**	96.0	90.4	91.7	93.1
F + M	$F1_{natural}$	88.4	87.5	75.6	84.0	85.0	**90.2**
	$F1_{media}$	82.7	90.0	80.2	80.7	80.8	**90.6**
	Avg. F1	85.6	88.8	77.9	82.4	82.9	**90.4**
F + V + M	$F1_{natural}$	85.4	**93.0**	85.8	87.8	89.4	92.0
	$F1_{media}$	79.5	**93.8**	85.8	83.2	83.3	91.3
	Avg. F1	82.5	**93.4**	85.8	85.5	86.4	91.7

Leave-One-Room-Out (LORO) Summary. We present the average F1 scores between each of the two classes across all three LORO folds. For each fold, all of a room's *media* recordings were held out of the training set, and used in the testing set along with the *natural* audio from our own homes.

Test Set	Metrics	KNN	QDA	DT	GNB	LR	SVC
V + M	$F1_{natural}$	77.8	**99.2**	94.8	85.6	88.9	94.7
	$F1_{media}$	76.9	**99.2**	93.7	76.3	81.0	95.0
	Avg. F1	77.4	**99.2**	94.3	81.0	85.0	94.9
F + M	$F1_{natural}$	85.2	84.3	71.8	82.9	82.5	**86.5**
	$F1_{media}$	81.0	**87.8**	78.3	84.0	82.9	86.4
	Avg. F1	83.1	86.1	75.1	83.5	82.7	**86.5**
F + V + M	$F1_{natural}$	81.5	**92.1**	84.1	84.0	85.2	90.4
	$F1_{media}$	79.3	**93.0**	85.4	81.7	83.7	90.7
	Avg. F1	80.4	**92.6**	84.8	82.9	84.5	90.6

Leave-One-Speaker-Out (LOSO) Summary. We present the average F1 scores between each of the two classes across all five LOSO folds. For each fold, all of a loudspeaker's *media* recordings were held out of the training set, and used in the testing set along with the *natural* audio from our own homes.

Test Set	Metrics	KNN	QDA	DT	GNB	LR	SVC
V + M	$F1_{natural}$	78.8	98.5	**99.0**	94.1	97.1	93.4
	$F1_{media}$	75.0	98.4	**98.9**	93.4	97.1	93.4
	Avg. F1	76.9	98.5	**99.0**	93.8	97.1	93.4
F + M	$F1_{natural}$	87.0	83.1	72.7	82.0	83.8	**87.2**
	$F1_{media}$	80.1	85.7	77.6	79.4	84.8	**87.8**
	Avg. F1	83.6	84.4	75.2	80.7	84.3	**87.5**
F + V + M	$F1_{natural}$	83.3	**90.6**	85.9	87.6	90.2	90.1
	$F1_{media}$	78.4	**91.3**	87.2	86.4	90.5	90.4
	Avg. F1	80.9	**91.0**	86.6	87.0	90.4	90.3

Leave-One-Distance-Out Cross Validation (LODO) Summary. We present the average F1 scores between each of the two classes across all three LODO folds. For each fold, all of a microphone distance's *media* recordings were held out of the training set, and used in the testing set along with the *natural* audio from our own homes.

Test Set	Metrics	KNN	QDA	DT	GNB	LR	SVC
V + M	$F1_{natural}$	79.7	**98.8**	97.4	92.1	97.7	94.5
	$F1_{media}$	78.1	**98.8**	97.1	91.1	97.8	95.0
	Avg. F1	78.9	**98.8**	97.3	91.6	97.8	94.8
F + M	$F1_{natural}$	81.5	82.5	70.9	76.9	82.2	**84.6**
	$F1_{media}$	67.0	**84.6**	56.8	66.3	80.9	83.6
	Avg. F1	74.3	83.6	63.9	71.6	81.6	**84.1**
F + V + M	$F1_{natural}$	79.9	**90.7**	83.6	83.9	89.6	89.2
	$F1_{media}$	75.4	**91.3**	79.2	79.6	89.5	89.5
	Avg. F1	77.7	**91.0**	81.4	81.8	89.6	89.4

Leave-One-Room and Speaker-Out (LORSO) Summary. We present the average F1 scores between each of the two classes across all nine LORSO folds. For each fold, all of a room's *media* recordings were held out of the training set, and used in the testing set along with the *natural* audio from our own homes.

Test Set	Metrics	KNN	QDA	DT	GNB	LR	SVC
V + M	$F1_{natural}$	75.5	**92.2**	91.7	81.4	89.0	89.5
	$F1_{media}$	60.3	81.0	**84.0**	58.8	82.2	83.1
	Avg. F1	67.9	86.6	**87.9**	70.1	85.6	86.3
F + M	$F1_{natural}$	85.8	84.5	76.1	80.5	87.6	**89.1**
	$F1_{media}$	78.4	81.2	80.0	72.9	86.7	**87.0**
	Avg. F1	82.1	82.9	78.1	76.7	87.2	**88.1**
F + V + M	$F1_{natural}$	81.1	88.2	83.6	80.6	88.1	**89.0**
	$F1_{media}$	71.0	81.9	82.5	69.1	85.7	**86.2**
	Avg. F1	76.1	85.1	83.1	74.9	86.9	**87.6**

Leave-One-Recording and Distance-Out (LORDO) Summary. We present the average F1 scores between each of the two classes across all nine LORDO folds. For each fold, all of a room and microphone distance's *media* recordings were held out of the training set, and used in the testing set along with the *natural* audio from our own homes.

Test Set	Metrics	KNN	QDA	DT	GNB	LR	SVC
V + M	$F1_{natural}$	76.1	94.0	**96.4**	85.4	91.5	89.6
	$F1_{media}$	63.8	90.9	**94.2**	71.6	86.0	85.1
	Avg. F1	70.0	92.5	**95.3**	78.5	88.8	87.4
F + M	$F1_{natural}$	83.9	86.3	75.6	80.3	89.2	**90.5**
	$F1_{media}$	72.7	87.0	80.2	73.8	89.6	**90.5**
	Avg. F1	78.3	86.7	77.9	77.1	89.4	**90.5**
F + V + M	$F1_{natural}$	80.2	90.0	85.8	82.6	**90.2**	90.0
	$F1_{media}$	69.8	88.8	85.8	73.4	**88.9**	**88.9**
	Avg. F1	75.0	89.4	85.8	78.0	**89.6**	89.5

Leave-One-Speaker and Distance-Out (LOSDO) Summary. We present the average F1 scores between each of the two classes across all nine LOSDO folds. For each fold, all of a speaker and microphone distance combination's *media* recordings were held out of the training set, and used in the testing set along with the *natural* audio from our own homes.

Test Set	Metrics	KNN	QDA	DT	GNB	LR	SVC
V + M	$F1_{natural}$	79.7	94.6	93.7	89.6	**95.2**	94.3
	$F1_{media}$	72.8	87.2	86.4	80.7	**95.4**	94.6
	Avg. F1	76.3	90.9	90.1	85.2	**95.3**	94.5
F + M	$F1_{natural}$	84.0	82.7	77.4	81.6	85.5	**87.0**
	$F1_{media}$	74.2	78.9	75.7	75.0	81.0	**82.5**
	Avg. F1	79.1	80.8	76.6	78.3	83.3	**84.8**
F + V + M	$F1_{natural}$	82.1	88.3	85.4	85.2	89.7	**90.0**
	$F1_{media}$	73.7	82.6	80.2	77.5	88.3	**89.0**
	Avg. F1	77.9	85.5	82.8	81.4	89.0	**89.5**

Leave-One-Room and Speaker and Distance-Out (LORSDO) Summary. We present the average F1 scores between each of the two classes across all 14 LORSDO folds. For each fold, all of a room, speaker, and microphone distance combination's *media* recordings were held out of the training set, and used in the testing set along with the *natural* audio from our own homes.

Test Set	Metrics	KNN	QDA	DT	GNB	LR	SVC
V + M	$F1_{natural}$	74.1	89.6	**91.1**	81.2	89.1	88.1
	$F1_{media}$	57.4	75.1	83.3	58.8	**83.8**	82.5
	Avg. F1	65.8	82.4	**87.2**	70.0	86.5	85.3
F + M	$F1_{natural}$	80.9	83.6	72.3	76.9	86.7	**88.2**
	$F1_{media}$	65.0	70.5	73.1	56.4	80.6	**82.1**
	Avg. F1	73.0	77.1	72.7	66.7	83.7	**85.2**
F + V + M	$F1_{natural}$	78.0	86.4	82.0	78.9	87.8	**88.1**
	$F1_{media}$	61.7	72.5	77.2	57.5	82.2	**82.4**
	Avg. F1	69.9	79.5	79.6	68.2	85.0	**85.3**

D EXPERIMENT COMPREHENSIVE RESULTS

Leave-One-Recording-Out CV Results. The table presents the macro and micro averages across all LOROCV folds for each classifier.

				Our In-Home Natural Recordings					
		C+M		F+M		V+M		V+F+M	
Model	Metrics	Macro	Micro	Macro	Micro	Macro	Micro	Macro	Micro
KNN	Accuracy	96.8	96.3	94.7	94.0	87.3	86.9	91.5	90.9
	Precision_N	95.7	94.8	95.7	94.7	94.8	93.7	95.3	94.2
	Recall_N	99.5	99.6	95.0	94.6	81.0	81.1	89.0	88.7
	F1_N	97.3	96.8	95.1	94.4	86.9	86.5	91.7	91.1
	Precision_M	99.4	94.8	94.1	93.7	82.5	82.3	88.7	88.4
	Recall_M	94.1	93.1	94.4	93.4	93.6	92.7	94.1	93.1
	F1_M	95.8	95.3	93.8	93.1	87.2	86.7	90.9	90.2
QDA	Accuracy	99.6	99.6	89.1	88.9	99.5	99.4	93.6	93.5
	Precision_N	99.8	99.7	99.8	99.7	99.7	99.7	99.7	99.7
	Recall_N	99.5	99.5	78.4	78.0	99.3	99.1	87.3	87.2
	F1_N	99.6	99.6	87.7	87.4	99.5	99.4	93.1	93.0
	Precision_M	99.5	99.7	82.4	82.1	99.3	99.1	88.8	88.7
	Recall_M	99.8	99.7	99.8	99.8	99.7	99.7	99.8	99.7
	F1_M	99.6	99.6	90.2	90.0	99.5	99.4	93.4	93.4
DT	Accuracy	99.0	99.0	88.2	88.3	98.9	98.8	92.8	92.9
	Precision_N	99.2	99.2	99.0	99.1	99.3	99.3	99.1	99.2
	Recall_N	99.1	99.1	77.6	77.7	98.7	98.6	86.7	86.7
	F1_N	99.1	99.1	86.4	86.6	98.9	98.9	92.3	92.3
	Precision_M	99.1	99.2	82.0	82.0	98.7	98.6	88.3	88.4
	Recall_M	99.0	99.0	98.8	98.9	99.1	99.1	99.0	99.0
	F1_M	99.0	99.0	89.4	89.4	98.8	98.8	93.2	93.2
GNB	Accuracy	92.2	92.8	86.2	87.4	96.2	96.7	90.5	91.4
	Precision_N	93.7	94.8	93.0	94.1	94.7	95.7	93.6	94.7
	Recall_N	91.2	91.1	79.7	80.8	98.5	98.4	87.8	88.4
	F1_N	91.7	92.2	85.5	86.7	96.4	96.9	90.4	91.3
	Precision_M	92.7	94.8	81.2	82.9	98.5	98.3	88.4	89.0
	Recall_M	93.2	94.4	92.6	94.0	93.9	95.0	93.2	94.4
	F1_M	92.3	93.1	86.5	87.9	95.9	96.4	90.5	91.5
LR	Accuracy	99.0	99.0	88.2	88.3	98.9	98.8	92.8	92.9
	Precision_N	99.2	99.2	99.0	99.1	99.3	99.3	99.1	99.1
	Recall_N	99.1	99.1	77.5	77.7	98.7	98.6	86.7	86.7
	F1_N	99.1	99.1	86.4	86.6	99.0	98.9	92.3	92.3
	Precision_M	99.1	99.2	82.0	82.0	98.7	98.6	88.3	88.4
	Recall_M	98.9	99.0	98.8	98.9	99.1	99.0	98.9	99.0
	F1_M	98.9	98.9	89.3	89.4	98.8	98.8	93.2	93.2
SVC	Accuracy	99.4	99.4	91.4	91.5	96.6	96.5	93.6	93.7
	Precision_N	99.4	99.4	99.4	99.4	99.4	99.4	99.4	99.4
	Recall_N	99.6	99.6	83.5	83.8	94.0	93.9	88.0	88.2
	F1_N	99.4	99.5	90.6	90.9	96.5	96.5	93.3	93.4
	Precision_M	99.6	99.4	85.5	86.0	94.4	94.2	89.3	89.4
	Recall_M	99.2	99.2	99.2	99.2	99.1	99.2	99.2	99.2
	F1_M	99.3	99.3	91.9	92.1	96.6	96.5	93.9	93.9

Leave-One-Label-Out Results. The table presents the macro and micro averages across all LOLO folds for each classifier.

Model	Metrics	LOLO Average					
		V+M		F+M		V+F+M	
		Macro	Micro	Macro	Micro	Macro	Micro
KNN	Accuracy	79.5	78.8	86.4	86.1	83.4	82.9
	$Precision_N$	84.1	83.9	85	85.3	84.5	84.5
	$Recall_N$	81.2	80.5	94	92.8	88.4	87.3
	$F1_N$	81.5	80.9	88.4	88	85.4	84.9
	$Precision_M$	76.2	75.4	91.6	90.5	84.4	83.5
	$Recall_M$	77.8	77.2	78.8	79.4	78.3	78.4
	$F1_M$	75.5	74.7	82.7	82.6	79.5	79.1
QDA	Accuracy	99.2	99.2	88.9	87.7	93.4	92.8
	$Precision_N$	99.2	99.3	99.3	99.3	99.3	99.3
	$Recall_N$	99.3	99.1	78.3	76	87.5	86.3
	$F1_N$	99.3	99.2	87.5	86	93	92.3
	$Precision_M$	99.3	99.1	82.2	80.7	88.9	87.9
	$Recall_M$	99.2	99.3	99.5	99.5	99.3	99.4
	$F1_M$	99.2	99.2	90	89	93.8	93.3
DT	Accuracy	95.6	96.3	78.6	78	86.1	86.1
	$Precision_N$	94	94.8	93.2	93.4	93.6	94.1
	$Recall_N$	99.8	99.8	64.9	63.1	80.2	79.5
	$F1_N$	96.4	96.9	75.6	74.5	85.8	85.7
	$Precision_M$	99.6	99.7	71.9	71.1	81.8	81.4
	$Recall_M$	91.5	92.7	92.4	92.8	92	92.8
	$F1_M$	94.2	95.1	80.2	79.9	85.8	86
GNB	Accuracy	90.7	91.5	83.2	84	86.5	87.3
	$Precision_N$	88.4	89.4	87.1	88.9	87.5	89
	$Recall_N$	98.2	97.9	82.4	81.7	89.3	88.9
	$F1_N$	92.4	92.9	84	84.5	87.8	88.4
	$Precision_M$	91.1	91.9	78	78.4	83.2	83.8
	$Recall_M$	83.1	85	84	86.3	83.6	85.7
	$F1_M$	86.4	87.9	80.7	82	83.2	84.6
LR	Accuracy	92.3	93.2	84.1	84	87.7	88.1
	$Precision_N$	92.6	93.3	92.6	93.4	92.6	93.3
	$Recall_N$	98.3	98.2	80.5	79.2	88.4	87.7
	$F1_N$	94.4	94.9	85	84.5	89.4	89.5
	$Precision_M$	92.5	93.3	76.6	76.4	82.2	82.6
	$Recall_M$	86.4	88.1	87.6	88.8	87	88.5
	$F1_M$	86.7	88.5	80.8	81.2	83.3	84.3
SVC	Accuracy	93.4	93.5	90.5	90.1	91.8	91.6
	$Precision_N$	95.6	95.5	97.1	97.5	96.3	96.5
	$Recall_N$	93.6	93.5	84.9	83.5	88.7	88
	$F1_N$	94.1	94.1	90.2	89.6	92	91.7
	$Precision_M$	92.7	92.8	86.4	85.4	89.1	88.5
	$Recall_M$	93.3	93.5	96.2	96.7	94.9	95.3
	$F1_M$	91.8	92.1	90.6	90.3	91.3	91.2

Leave-One-Room-Out Results. The table presents the results of the three LORO folds (each room column is the left-out room), and the macro averages across all LORO folds for each classifier. Only macro averages are presented because the test sets were the same size (the left out room *media* set was larger than the *natural* testing subset, so *media* was sampled to match the size of the natural sets).

		Kitchen			Bedroom			Playroom			Average		
Model	Metrics	V+M	F+M	V+F+M	V+M	F+M	V+F+M	V+M	F+M	V+F+M	V+M	F+M	V+F+M
KNN	Accuracy	82.4	84.3	83.4	73.1	93.2	83.4	76.7	72.8	74.7	77.4	83.4	80.5
	$Precision_N$	82.1	77.4	79.4	71.4	97.3	83.6	76.3	67.1	71	76.6	80.6	78
	$Recall_N$	82.9	97	90.1	77.2	88.8	83.2	77.4	89.5	83.6	79.2	91.8	85.6
	$F1_N$	82.5	86.1	84.4	74.2	92.8	83.4	76.9	76.7	76.8	77.8	85.2	81.5
	$Precision_M$	82.7	96	88.6	75.2	89.7	83.2	77.1	84.3	80.1	78.3	90	84
	$Recall_M$	81.9	71.6	76.6	69	97.5	83.6	76	56.2	65.8	75.6	75.1	75.4
	$F1_M$	82.3	82	82.2	72	93.4	83.4	76.6	67.4	72.3	76.9	81	79.3
QDA	Accuracy	98.9	89.6	94.2	99.6	82.5	90.8	99	86.9	92.8	99.2	86.3	92.6
	$Precision_N$	99.7	99.1	99.4	100	99.7	99.9	98.3	94.4	96.5	99.3	97.7	98.6
	$Recall_N$	98.1	80	88.8	99.2	65.2	81.8	99.7	78.4	88.8	99	74.5	86.5
	$F1_N$	98.9	88.5	93.8	99.6	78.8	89.9	99	85.7	92.5	99.2	84.3	92.1
	$Precision_M$	98.2	83.2	89.9	99.2	74.1	84.6	99.7	81.6	89.6	99	79.6	88
	$Recall_M$	99.7	99.3	99.5	100	99.8	99.9	98.3	95.3	96.8	99.3	98.1	98.7
	$F1_M$	98.9	90.5	94.5	99.6	85.1	91.6	99	87.9	93.1	99.2	87.8	93
DT	Accuracy	98.4	81.7	89.8	86.3	72.4	79.2	98.4	73	85.3	94.3	75.7	84.8
	$Precision_N$	96.9	99	97.8	79.5	88.3	82.4	97.1	73.7	85.5	91.2	87	88.6
	$Recall_N$	99.9	63.9	81.5	97.7	51.7	74.1	99.7	71.3	85.1	99.1	62.3	80.2
	$F1_N$	98.4	77.7	88.9	87.7	65.2	78.1	98.4	72.5	85.3	94.8	71.8	84.1
	$Precision_M$	99.9	73.4	84.1	97	65.9	76.5	99.7	72.2	85.2	98.9	70.5	81.9
	$Recall_M$	96.8	99.4	98.1	74.8	93.2	84.2	97.1	74.6	85.5	89.6	89.1	89.3
	$F1_M$	98.3	84.4	90.6	84.5	77.2	80.2	98.3	73.4	85.4	93.7	78.3	85.4
GNB	Accuracy	87.9	82.5	85.1	64.1	85	74.8	95	83.1	88.9	82.4	83.5	82.9
	$Precision_N$	82.5	80.2	81.4	58.6	96.2	70.9	91.6	85.1	88.5	77.6	87.2	80.3
	$Recall_N$	96.3	86.2	91.1	96.1	72.8	84.2	99	80.2	89.4	97.1	79.8	88.2
	$F1_N$	88.9	83.1	86	72.8	82.9	77	95.2	82.6	88.9	85.6	82.9	84
	$Precision_M$	95.5	85.1	89.9	89.1	78.1	80.5	98.9	81.3	89.3	94.5	81.5	86.6
	$Recall_M$	79.6	78.8	79.2	32.2	97.1	65.5	90.9	85.9	88.4	67.6	87.3	77.7
	$F1_M$	86.9	81.8	84.2	47.3	86.6	72.2	94.8	83.5	88.8	76.3	84	81.7
LR	Accuracy	98.5	87.9	93	66	87.9	77.2	94.3	72.8	83.3	86.3	82.9	84.5
	$Precision_N$	99.9	95.9	98	59.8	100	73	96.9	70	81.9	85.5	88.6	84.3
	$Recall_N$	97.1	79.2	87.9	97.7	75.8	86.5	91.6	79.9	85.6	95.5	78.3	86.6
	$F1_N$	98.5	86.7	92.7	74.2	86.2	79.1	94.2	74.6	83.7	88.9	82.5	85.2
	$Precision_M$	97.1	82.3	89	93.7	80.5	83.4	92	76.6	84.9	94.3	79.8	85.8
	$Recall_M$	99.9	96.6	98.2	34.2	100	68	97.1	65.8	81	77.1	87.5	82.4
	$F1_M$	98.5	88.8	93.4	50.2	89.2	74.9	94.5	70.8	82.9	81	82.9	83.7
SVC	Accuracy	94.9	93.5	94.2	95.9	89.6	92.7	93.7	76.3	84.8	94.8	86.5	90.6
	$Precision_N$	98.6	95.3	96.8	100	100	100	96.4	73.6	83.7	98.3	89.6	93.5
	$Recall_N$	91.2	91.5	91.3	91.7	79.3	85.3	90.8	82.2	86.4	91.2	84.3	87.7
	$F1_N$	94.7	93.3	94	95.7	88.4	92.1	93.5	77.6	85	94.7	86.5	90.4
	$Precision_M$	91.8	91.8	91.8	92.3	82.8	87.2	91.3	79.8	85.9	91.8	84.8	88.3
	$Recall_M$	98.7	95.4	97	100	100	100	96.6	70.5	83.2	98.4	88.6	93.4
	$F1_M$	95.1	93.6	94.3	96	90.6	93.2	93.9	74.8	84.6	95	86.4	90.7

Leave-One-Speaker-Out Results. The table presents the results of the five LOSO folds (each speaker column is the left-out speaker), and the macro (M) and micro (μ) averages across all LOSO folds for each classifier.

		Bose			iPhone			BigBose			Sony			Mac			Average					
																	V+M		F+M		V+F+M	
Model	Metrics	V+M	F+M	V+F+M	V+M	F+M	V+F+M	V+M	F+M	V+F+M	V+M	F+M	V+F+M	V+M	F+M	V+F+M	M	μ	M	μ	M	μ
KNN	Accuracy	68.4	59.8	64	87.1	95.7	92	87.1	95.5	91.8	56.2	73.9	65.9	86.6	95.8	91.3	77.1	76.2	84.1	82.5	81	79.6
	Precision_N	66.2	56.1	60.2	91.1	100	96.1	96.7	98.9	98	54.2	67.9	61.4	89.1	96.7	93.1	79.4	77.8	83.9	81.8	81.7	79.7
	Recall_N	75.5	89.6	82.7	82.3	91.5	87.4	76.9	92.1	85.3	79.6	90.7	85.7	83.5	94.8	89.3	79.5	79.5	91.7	91.8	86.1	86
	F1_N	70.5	69	69.7	86.4	95.5	91.6	85.6	95.4	91.2	64.5	77.6	71.5	86.2	95.8	91.2	78.7	77.9	86.7	85.5	83	81.9
	Precision_M	71.4	74.3	72.4	83.8	92.1	88.5	80.8	92.6	87	61.6	86	76.3	84.5	94.9	89.7	76.4	76.1	88	87.3	82.8	82.1
	Recall_M	61.4	29.9	45.3	91.9	100	96.5	97.3	98.9	98.2	32.8	57.1	46.1	89.7	96.8	93.4	74.6	72.9	76.5	73.3	75.9	73.1
	F1_M	66	42.7	55.7	87.7	95.9	92.3	88.3	95.7	92.3	42.8	68.6	57.5	87	95.8	91.5	74.4	73.4	79.7	77.4	77.9	76
QDA	Accuracy	94.4	65.2	79.4	99.2	85.9	91.7	99.3	85.6	91.7	99.5	86.5	92.4	98.9	84.7	91.6	98.3	97.9	81.6	80.6	89.4	88.7
	Precision_N	90.1	61.2	73.8	99.8	99.8	99.8	99.8	99.2	99.5	100	100	100	99	98.8	98.9	97.7	97	91.8	90	94.4	92.9
	Recall_N	99.8	83.1	91.2	98.6	71.9	83.5	98.9	71.7	83.8	99.1	73	84.8	98.8	70.3	84.2	99	99.1	74	74.4	85.5	85.9
	F1_N	94.7	70.5	81.6	99.2	83.6	90.9	99.3	83.2	91	99.5	84.4	91.8	98.9	82.1	91	98.3	98	80.8	80.1	89.2	88.7
	Precision_M	99.8	73.7	88.5	98.6	78	85.8	98.9	77.9	86	99.1	78.7	86.8	98.8	76.9	86.2	99	99.1	77.1	76.8	86.7	86.8
	Recall_M	89	47.4	67.6	99.8	99.8	99.8	99.8	99.4	99.6	100	100	100	99	99.2	99.1	97.5	96.7	89.1	86.8	93.2	91.4
	F1_M	94.1	57.7	76.7	99.2	87.6	92.3	99.3	87.3	92.3	99.5	88.1	92.9	98.9	86.6	92.2	98.2	97.8	81.5	80.1	89.3	88.4
DT	Accuracy	95.7	55.8	75.2	99.3	80.7	88.8	100	79.4	88.6	99.8	79.1	88.5	99	69.6	83.9	98.8	98.5	72.9	71.3	85	84
	Precision_N	92.2	54.6	71.5	98.6	99.7	99.1	100	100	100	99.7	100	99.8	98.2	72.9	86.3	97.7	97.2	85.4	82.2	91.3	89.1
	Recall_N	100	68.8	84	100	61.6	78.3	100	58.8	77.1	100	58.2	77.1	99.8	62.3	80.6	100	99.9	61.9	62.4	79.4	79.9
	F1_N	95.9	60.9	77.2	99.3	76.1	87.5	100	74.1	87.1	99.8	73.6	87	99	67.2	83.3	98.8	98.5	70.4	69.3	84.4	83.7
	Precision_M	100	57.8	80.6	100	72.2	82.1	100	70.8	81.4	100	70.5	81.3	99.8	67.1	81.8	100	99.9	67.7	66.8	81.4	81.4
	Recall_M	91.5	42.7	66.5	98.6	99.8	99.3	100	100	100	99.7	100	99.9	98.1	76.9	87.2	97.6	97	83.9	80.3	90.6	88.1
	F1_M	95.6	49.1	72.9	99.3	83.8	89.9	100	82.9	89.7	99.8	82.7	89.7	99	71.7	84.4	98.7	98.4	74	72	85.3	84.1
GNB	Accuracy	93.6	55.5	74.1	96.4	86.8	91	96.7	87.8	91.8	95.7	88.3	91.7	80.7	85.1	83	92.6	91.4	80.7	79.2	86.3	84.9
	Precision_N	88.9	53.8	68.7	94.3	92.6	93.4	96.1	95.9	96	94.6	98.5	96.5	73.2	83.1	77.7	89.4	87.6	84.8	82.5	86.5	84.1
	Recall_N	99.7	77.8	88.5	98.8	80.1	88.3	97.3	79	87.2	97.1	77.8	86.5	96.8	88.2	92.4	98	98	80.6	80.9	88.6	88.9
	F1_N	94	63.6	77.3	96.5	85.9	90.8	96.7	86.7	91.4	95.8	87	91.2	83.4	85.6	84.4	93.3	92.2	81.7	80.7	87	85.9
	Precision_M	99.6	59.9	83.8	98.8	82.4	88.9	97.3	82.2	88.3	97	81.7	87.8	95.3	87.4	90.6	97.6	97.5	78.7	77.9	87.9	87.7
	Recall_M	87.6	33.2	59.7	94	93.6	93.8	96	96.7	96.4	94.4	98.9	96.9	64.6	82.1	73.5	87.3	84.8	80.9	77.5	84	80.9
	F1_M	93.2	42.7	69.7	96.3	87.7	91.3	96.7	88.8	92.1	95.7	89.5	92.1	77	84.7	81.2	91.8	90.3	78.7	76.5	85.3	83.5
LR	Accuracy	91.2	73.1	81.9	99.3	88.1	93	99.3	87.7	92.9	96.8	88.3	92.1	99.2	87.3	93.1	97.2	96.6	84.9	84.2	90.6	90
	Precision_N	91.7	69	78.9	100	100	100	99.8	100	99.9	95.1	99.8	97.3	99.5	99.6	99.5	97.2	96.7	93.7	92.3	95.1	94
	Recall_N	90.6	83.6	87	98.6	76.2	85.9	98.9	75.4	85.8	98.6	76.7	86.6	98.9	74.9	86.6	97.1	96.6	77.4	77.7	86.4	86.5
	F1_N	91.2	75.6	82.8	99.3	86.5	92.4	99.3	86	92.3	96.8	86.7	91.6	99.2	85.5	92.6	97.2	96.6	84.1	83.5	90.4	89.9
	Precision_M	90.7	79.2	85.6	98.6	80.7	87.7	98.9	80.2	87.6	98.6	81.1	87.9	98.9	79.9	88.1	97.1	96.6	80.2	80.2	87.4	87.3
	Recall_M	91.8	62.5	76.8	100	100	100	99.8	100	99.9	94.9	99.9	97.6	99.5	99.7	99.6	97.2	96.7	92.4	90.7	94.8	93.5
	F1_M	91.3	69.9	80.9	99.3	89.3	93.4	99.3	89	93.3	96.7	89.5	92.5	99.2	88.7	93.5	97.2	96.6	85.3	84.4	90.7	90.1
SVC	Accuracy	85.4	76	80.6	95.9	91.5	93.4	96.7	90.7	93.4	91.7	90.3	91	97.1	89.8	93.4	93.4	92.8	87.7	86.9	90.3	89.7
	Precision_N	85.6	72	78.1	100	100	100	100	100	100	89.8	100	94.7	100	100	100	95.1	94.3	94.4	93.1	94.6	93.4
	Recall_N	85.1	85.1	85.1	91.7	83.1	86.8	93.4	81.5	86.8	94.1	80.6	86.7	94.2	79.7	86.8	91.7	91.3	82	82	86.4	86.3
	F1_N	85.3	78	81.4	95.7	90.8	93	96.6	89.8	92.9	91.9	89.3	90.6	97	88.7	92.9	93.3	92.7	87.3	86.7	90.2	89.5
	Precision_M	85.2	81.8	83.6	92.3	85.5	88.4	93.8	84.4	88.3	93.8	83.8	87.8	94.5	83.1	88.3	91.9	91.5	83.7	83.5	87.3	87
	Recall_M	85.7	66.9	76.1	100	100	100	100	100	100	89.3	100	95.2	100	100	100	95	94.2	93.4	91.9	94.3	93
	F1_M	85.4	73.6	79.7	96	92.2	93.8	96.8	91.5	93.8	91.5	91.2	91.3	97.2	90.8	93.8	93.4	92.8	87.9	87	90.5	89.7

Leave-One-Distance-Out Results. The table presents the results of the three LODO folds (each microphone distance is the left-out distance), and the macro averages across all LODO folds for each classifier. Only macro averages are presented because the test sets were the same size (the left out microphone distance *media* set was larger than the *natural* testing subset, so *media* was sampled to match the size of the natural sets).

		1 ft			4-6 ft			8-10 ft			LODO Average		
Model	Metrics	V+M	F+M	V+F+M	V+M	F+M	V+F+M	V+M	F+M	V+F+M	V+M	F+M	V+F+M
KNN	Accuracy	83.3	55.3	68.9	87.9	84.4	86.1	65.8	91.6	79	79	77.1	78
	$Precision_N$	87.7	53	64	88.9	80.6	84.3	63.1	93.5	76.9	79.9	75.7	75.1
	$Recall_N$	77.5	94.8	86.4	86.7	90.7	88.7	75.9	89.3	82.8	80	91.6	86
	$F1_N$	82.3	68	73.6	87.8	85.3	86.5	68.9	91.4	79.8	79.7	81.5	79.9
	$Precision_M$	79.9	75.2	79.1	87	89.3	88.1	69.8	89.8	81.4	78.9	84.8	82.9
	$Recall_M$	89.1	15.8	51.5	89.2	78.1	83.5	55.6	93.8	75.2	78	62.6	70.1
	$F1_M$	84.2	26	62.4	88.1	83.4	85.8	61.9	91.7	78.2	78.1	67	75.4
QDA	Accuracy	98.6	75.3	86.7	99.1	87	92.9	98.5	88.7	93.5	98.8	83.7	91
	$Precision_N$	97.5	76.2	86.9	99.6	98.8	99.2	98.8	99.5	99.1	98.6	91.5	95.1
	$Recall_N$	99.8	73.5	86.3	98.7	74.9	86.5	98.3	77.8	87.8	98.9	75.4	86.9
	$F1_N$	98.7	74.8	86.6	99.1	85.2	92.4	98.5	87.3	93.1	98.8	82.5	90.7
	$Precision_M$	99.8	74.4	86.4	98.7	79.8	88	98.3	81.8	89	98.9	78.7	87.8
	$Recall_M$	97.5	77.1	87	99.6	99.1	99.3	98.8	99.6	99.2	98.6	91.9	95.2
	$F1_M$	98.6	75.7	86.7	99.1	88.4	93.3	98.5	89.8	93.8	98.8	84.6	91.3
DT	Accuracy	93.2	41.1	66.5	99.5	79.4	89.2	99	82.6	90.6	97.2	67.7	82.1
	$Precision_N$	88.2	45	61.3	99.1	92.8	96.5	98.1	97.6	97.9	95.1	78.5	85.2
	$Recall_N$	99.8	80.2	89.7	99.9	63.7	81.3	100	66.8	83	99.9	70.3	84.7
	$F1_N$	93.6	57.7	72.8	99.5	75.6	88.3	99	79.3	89.8	97.4	70.9	83.6
	$Precision_M$	99.7	9.48	80.8	99.9	72.4	83.9	100	74.8	85.2	99.9	52.2	83.3
	$Recall_M$	86.6	2.07	43.3	99.1	95	97	98	98.3	98.2	94.6	65.1	79.5
	$F1_M$	92.7	3.4	56.4	99.5	82.2	90	99	85	91.3	97.1	56.8	79.2
GNB	Accuracy	91.7	46	68.2	88.1	84.3	86.2	95	89.1	92	91.6	73.1	82.1
	$Precision_N$	86.1	47.4	63.4	83.6	82	82.8	93.8	96.4	95	87.8	75.3	80.4
	$Recall_N$	99.3	73.7	86.2	94.9	87.9	91.3	96.5	81.1	88.6	96.9	80.9	88.7
	$F1_N$	92.3	57.7	73.1	88.9	84.8	86.8	95.1	88.1	91.7	92.1	76.9	83.9
	$Precision_M$	99.2	40.9	78.4	94.1	86.9	90.3	96.4	83.7	89.3	96.6	70.5	86
	$Recall_M$	84	18.2	50.3	81.4	80.7	81	93.6	97	95.3	86.3	65.3	75.5
	$F1_M$	91	25.2	61.3	87.3	83.7	85.4	95	89.9	92.2	91.1	66.3	79.6
LR	Accuracy	95.6	67.3	81.1	98.8	88.8	93.7	98.8	89.3	93.9	97.7	81.8	89.6
	$Precision_N$	99.9	63.8	78.6	99	97.8	98.5	99	99.9	99.4	99.3	87.2	92.2
	$Recall_N$	91.4	79.9	85.5	98.6	79.4	88.7	98.6	78.8	88.4	96.2	79.3	87.5
	$F1_N$	95.4	71	81.9	98.8	87.6	93.3	98.8	88.1	93.6	97.7	82.2	89.6
	$Precision_M$	92.1	73.1	84.1	98.6	82.7	89.7	98.6	82.5	89.6	96.4	79.4	87.8
	$Recall_M$	99.9	54.7	76.7	99	98.2	98.6	99	99.9	99.5	99.3	84.3	91.6
	$F1_M$	95.8	62.6	80.2	98.8	89.8	94	98.8	90.3	94.3	97.8	80.9	89.5
SVC	Accuracy	94	71.1	82.3	96.9	90.3	93.5	93.3	91.5	92.3	94.7	84.3	89.4
	$Precision_N$	99.8	67.4	80.6	100	99	99.5	96.6	99.4	98	98.8	88.6	92.7
	$Recall_N$	88.2	81.9	85	93.9	81.3	87.5	89.7	83.4	86.5	90.6	82.2	86.3
	$F1_N$	93.6	73.9	82.7	96.9	89.3	93.1	93	90.7	91.9	94.5	84.6	89.2
	$Precision_M$	89.4	76.9	84.1	94.2	84.2	88.8	90.4	85.7	87.9	91.4	82.3	86.9
	$Recall_M$	99.8	60.4	79.6	100	99.2	99.6	96.8	99.5	98.2	98.9	86.4	92.5
	$F1_M$	94.3	67.7	81.8	97	91.1	93.9	93.5	92.1	92.8	95	83.6	89.5

Leave-One-Room+Speaker-Out Results. The table presents the macro and micro averages across all LORSO folds for each classifier.

Model	Metrics	LORSO Average					
		V+M		F+M		V+F+M	
		Macro	Micro	Macro	Micro	Macro	Micro
KNN	Accuracy	70.3	67	83	81.9	77.3	75.1
	$Precision_N$	71.6	68.8	79.5	78.2	75.6	73.4
	$Recall_N$	82.7	81.6	95.1	94	89.6	88.4
	$F1_N$	75.5	73.2	85.8	84.7	81.1	79.3
	$Precision_M$	67.8	63	93.2	91.9	83.6	81.5
	$Recall_M$	57.9	52.3	71	69.8	65.1	61.9
	$F1_M$	60.3	54.6	78.4	77.4	71	68.1
QDA	Accuracy	89.1	90.1	84.2	85	86.4	87.3
	$Precision_N$	88.4	89.3	92.5	94.5	90	91.2
	$Recall_N$	99.3	99.2	80.3	77.5	88.9	87.4
	$F1_N$	92.2	92.8	84.5	84.1	88.2	88.3
	$Precision_M$	99.3	99.2	80.8	80.3	87.3	86.8
	$Recall_M$	79	80.9	88.1	92.5	84	87.2
	$F1_M$	81	83	81.2	84.2	81.9	84.7
DT	Accuracy	89.2	86.9	78.6	79.2	83.3	82.7
	$Precision_N$	86.3	83.6	90.2	90.2	87.3	85.8
	$Recall_N$	100	100	67.4	67.5	82	82.2
	$F1_N$	91.7	90	76.1	76.6	83.6	83.2
	$Precision_M$	99.9	99.9	73.5	73.8	82.4	82.3
	$Recall_M$	78.5	73.9	89.8	90.9	84.6	83.2
	$F1_M$	84	80.3	80	81	82.5	81.9
GNB	Accuracy	75.1	75.3	78.5	80.9	76.9	78.4
	$Precision_N$	71.7	72.1	80.5	83.9	74.8	76.3
	$Recall_N$	97.6	97.6	83.2	81.8	89.7	88.9
	$F1_N$	81.4	81.6	80.5	81.8	80.6	81.3
	$Precision_M$	95.7	95.6	78.7	80.1	83.6	84.3
	$Recall_M$	52.5	53	73.7	80.1	64.1	67.8
	$F1_M$	58.8	59	72.9	78	69.1	72.8
LR	Accuracy	86.8	86.3	87.6	88	87.2	87.2
	$Precision_N$	86.3	85.7	93.9	95.4	89.4	89.7
	$Recall_N$	93.5	94	83.6	81.6	88.1	87.2
	$F1_N$	89	88.9	87.6	87.4	88.1	87.9
	$Precision_M$	87.7	87	86	84.3	87.4	86.7
	$Recall_M$	80.1	78.7	91.5	94.4	86.3	87.3
	$F1_M$	82.2	80.8	86.7	88	85.7	86.1
SVC	Accuracy	87.4	86.5	88.4	88.4	88	87.5
	$Precision_N$	88.4	88	91.1	91.9	89.1	89.2
	$Recall_N$	92.5	91.6	88.6	86.4	90.4	88.7
	$F1_N$	89.5	88.8	89.1	88.5	89	88.3
	$Precision_M$	87.4	85.3	89.3	87.3	89.6	88
	$Recall_M$	82.4	81.4	88.3	90.4	85.6	86.3
	$F1_M$	83.1	81.4	87	87.8	86.2	86.1

Leave-One-Room+Distance-Out Results. The table presents the macro and micro averages across all LORDO folds for each classifier.

Model	Metrics	LORDO Average					
		V+M		F+M		V+F+M	
		Macro	Micro	Macro	Micro	Macro	Micro
KNN	Accuracy	71.7	74.2	80.2	81.4	76.4	78.1
	$Precision_N$	71.8	74.1	77.3	79.2	74.2	76
	$Recall_N$	83.3	83.9	94.2	94.3	89.3	89.5
	$F1_N$	76.1	77.8	83.9	84.9	80.2	81.3
	$Precision_M$	72	75.1	90.7	92	83.1	84.8
	$Recall_M$	60.2	64.6	66.2	68.6	63.6	66.7
	$F1_M$	63.8	68.1	72.7	73.9	69.8	72.6
QDA	Accuracy	92.8	94.8	86.8	87.5	89.5	90.9
	$Precision_N$	90.2	93.3	92.8	95.1	91.3	93.9
	$Recall_N$	99.4	99.1	81.5	80.2	89.6	88.9
	$F1_N$	94	95.6	86.3	86.6	90	91
	$Precision_M$	99.4	99.1	83.1	82.7	89.2	89.1
	$Recall_M$	86.2	90.6	92	94.7	89.5	92.8
	$F1_M$	90.9	93.5	87	88	88.8	90.5
DT	Accuracy	95.6	96.3	78.6	78	86.1	86.1
	$Precision_N$	94	94.8	93.2	93.4	93.6	94.1
	$Recall_N$	99.8	99.8	64.9	63.1	80.2	79.5
	$F1_N$	96.4	96.9	75.6	74.5	85.8	85.7
	$Precision_M$	99.6	99.7	71.9	71.1	81.8	81.4
	$Recall_M$	91.5	92.7	92.4	92.8	92	92.8
	$F1_M$	94.2	95.1	80.2	79.9	85.8	86
GNB	Accuracy	81.4	81.3	78.5	80.3	79.7	80.8
	$Precision_N$	77.9	78.2	80	81.6	78.5	79.3
	$Recall_N$	97.2	95.7	82.6	83	89.1	88.9
	$F1_N$	85.4	85.1	80.3	81.5	82.6	83.1
	$Precision_M$	84.7	87.4	74	78	77.9	81.8
	$Recall_M$	65.5	66.8	74.3	77.6	70.3	72.6
	$F1_M$	71.6	73.1	73.8	77.4	73.4	76.2
LR	Accuracy	89.9	91.6	89.4	88.9	89.7	90.1
	$Precision_N$	91.3	92.4	92.8	93.4	91.7	92.5
	$Recall_N$	93.2	95	86.4	84.5	89.6	89.4
	$F1_N$	91.5	93.1	89.2	88.5	90.2	90.5
	$Precision_M$	88.3	90.4	87.3	85.9	89.1	89
	$Recall_M$	86.7	88.1	92.5	93.3	89.9	90.9
	$F1_M$	86	87.9	89.6	89.2	88.9	89.4
SVC	Accuracy	88.2	90	90.6	91	89.6	90.5
	$Precision_N$	90.3	92.5	92.5	93.5	91.1	92.5
	$Recall_N$	90.3	91.2	89.1	88.8	89.7	89.9
	$F1_N$	89.6	91.2	90.5	90.9	90	90.9
	$Precision_M$	86.9	88.2	89.5	89.4	89.3	89.7
	$Recall_M$	86	88.8	92.1	93.3	89.5	91.2
	$F1_M$	85.1	87.2	90.5	91.1	88.9	90

Leave-One-Speaker+Distance-Out Results. The table presents the macro and micro averages across all LOSDO folds for each classifier.

Model	Metrics	LOSDO Average					
		V+M		F+M		V+F+M	
		Macro	Micro	Macro	Micro	Macro	Micro
KNN	Accuracy	77.5	74.7	80.8	78.5	79.3	76.7
	$Precision_N$	78.7	75.4	77.4	75.1	77.7	75
	$Recall_N$	82.7	82	93.6	93	88.7	88
	$F1_N$	79.7	77.6	84	82.3	82.1	80.2
	$Precision_M$	75.9	73.5	85.9	83.5	80.6	78.1
	$Recall_M$	72.2	67.4	68	63.9	69.8	65.5
	$F1_M$	72.8	69	74.2	70.5	73.7	70
QDA	Accuracy	92.6	92.3	81.8	81.1	86.6	86.2
	$Precision_N$	91.7	91.8	88.6	88.1	89.9	89.8
	$Recall_N$	99.5	99.3	80.2	79.4	88.9	88.5
	$F1_N$	94.6	94.4	82.7	82.1	88.3	88
	$Precision_M$	98.4	98	78.6	77	84.9	83.6
	$Recall_M$	85.8	85.3	83.3	82.8	84.4	83.9
	$F1_M$	87.2	86.1	78.9	77.9	82.6	81.7
DT	Accuracy	91.8	91.4	77.8	74.8	84.2	82.4
	$Precision_N$	91.3	90.5	87.4	85.9	89.3	88.2
	$Recall_N$	97.9	98.9	71.9	68.1	83.7	82.1
	$F1_N$	93.7	93.6	77.4	74.1	85.4	83.9
	$Precision_M$	98	98.9	70.1	66	77.3	74.5
	$Recall_M$	85.7	84	83.7	81.6	84.6	82.7
	$F1_M$	86.4	85.1	75.7	72.4	80.2	77.7
GNB	Accuracy	87.1	85.5	79.6	78.8	82.9	81.9
	$Precision_N$	83.8	82.4	81.5	81.1	82.4	81.6
	$Recall_N$	98	97.5	83.8	83.2	90.1	89.7
	$F1_N$	89.6	88.4	81.6	81	85.2	84.5
	$Precision_M$	86.6	84.1	75.7	73.5	80.4	78.2
	$Recall_M$	76.1	73.5	75.4	74.4	75.7	74
	$F1_M$	80.7	78	75	73.4	77.5	75.5
LR	Accuracy	95.3	95.3	84.2	82.9	89.2	88.6
	$Precision_N$	97.2	96.3	89.6	88.7	92.2	91.5
	$Recall_N$	93.5	94.4	84	82.5	88.3	87.9
	$F1_N$	95.2	95.3	85.5	84.2	89.7	89.1
	$Precision_M$	93.8	94.6	80.9	78.7	88	87.5
	$Recall_M$	97.1	96.3	84.5	83.2	90.1	89.2
	$F1_M$	95.4	95.4	81	79.1	88.3	87.5
SVC	Accuracy	94.5	93.6	85.7	83.9	89.6	88.3
	$Precision_N$	97.2	96.1	89.5	88.1	91.8	90.5
	$Recall_N$	91.7	91.2	86.7	85.3	89	88
	$F1_N$	94.3	93.5	87	85.5	90	88.8
	$Precision_M$	92.2	91.7	83.8	81.6	88.9	87.7
	$Recall_M$	97.2	96.1	84.7	82.5	90.2	88.7
	$F1_M$	94.6	93.8	82.5	80	89	87.6

Leave-One-Room+Speaker+Distance-Out Results. The table presents the macro and micro averages across all LORSDO folds for each classifier.

Model	Metrics	LORSDO Average					
		V+M		F+M		V+F+M	
		Macro	Micro	Macro	Micro	Macro	Micro
KNN	Accuracy	68.5	69.3	75.8	77.2	72.6	73.7
	$Precision_N$	67.2	68.4	71.2	72.7	69.4	70.8
	$Recall_N$	84.9	85.1	95.9	96.4	91.1	91.4
	$F1_N$	74.1	74.8	80.9	82.1	78	78.9
	$Precision_M$	69	68.4	88.8	90	79.4	80.2
	$Recall_M$	52	53.5	55.6	58.1	54	56
	$F1_M$	57.4	57.8	65	67	61.7	63
QDA	Accuracy	85.6	88.7	80	81.1	82.5	84.5
	$Precision_N$	83.9	87.5	84.1	87	84	87.2
	$Recall_N$	99.5	99.5	87.5	84.9	92.9	91.4
	$F1_N$	89.6	91.8	83.6	83.8	86.4	87.6
	$Precision_M$	99.5	99.5	79.2	78.6	85	85
	$Recall_M$	71.7	77.9	72.4	77.3	72.1	77.6
	$F1_M$	75.1	80.4	70.5	73.4	72.5	76.4
DT	Accuracy	88.7	89.2	74.4	71.2	80.8	79.3
	$Precision_N$	85.6	85.8	81.1	78.3	83.9	83.2
	$Recall_N$	99.5	99.5	69.8	62.7	83	79.1
	$F1_N$	91.1	91.4	72.3	67	82	79.8
	$Precision_M$	99.4	99.5	71.6	68.1	80.3	77.8
	$Recall_M$	78	79	78.9	79.7	78.5	79.4
	$F1_M$	83.3	84.5	73.1	71.5	77.2	76.6
GNB	Accuracy	74.9	78.7	71	74.8	72.7	76.5
	$Precision_N$	71.2	75	72	75.9	71.5	75.3
	$Recall_N$	97.7	96.6	87.1	86.5	91.8	91
	$F1_N$	81.2	83.4	76.9	79.2	78.9	81.1
	$Precision_M$	85	86.8	68	71.9	73.1	76.6
	$Recall_M$	52.1	60.8	54.9	63.2	53.6	62.1
	$F1_M$	58.8	67	56.4	64	57.5	65.3
LR	Accuracy	87.5	89.1	85	86.1	86.1	87.4
	$Precision_N$	87.9	89.2	87.3	89.9	87.2	89.2
	$Recall_N$	91.6	93.1	88.7	87.3	90	89.9
	$F1_N$	89.1	90.5	86.7	87.4	87.8	88.8
	$Precision_M$	86.4	88.6	85	84.5	85.3	85.9
	$Recall_M$	83.3	85.2	81.3	84.8	82.1	85
	$F1_M$	83.8	86	80.6	82.5	82.2	84.2
SVC	Accuracy	86.2	86.9	86.3	87.6	86.3	87.3
	$Precision_N$	86.3	87.6	87.4	89.8	86.7	88.5
	$Recall_N$	91.4	90.9	91.3	90.4	91.4	90.6
	$F1_N$	88.1	88.5	88.2	89.1	88.1	88.7
	$Precision_M$	85.6	86.1	88	87.7	86.8	86.8
	$Recall_M$	81	82.9	81.2	84.8	81.1	84
	$F1_M$	82.5	83.7	82.1	84.1	82.4	84.1