IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.Doi Number

Speaker Diarization and Identification from Single Channel Classroom Audio Recordings Using Virtual Microphones

Antonio Gomez[1], Senior Member, IEEE, Marios Pattichis[1], Senior Member, IEEE, and Sylvia Celedón-Pattichis[2]

[1]Department of Electrical and Computer Engineering, University of New Mexico, Albuquerque, NM 87131-0001, USA.
[2]Department of Curriculum and Instruction, The University of Texas, Austin, TX 78712, USA.

Corresponding author: Marios Pattichis (e-mail: pattichi@unm.edu).

ABSTRACT Speaker diarization refers to methods for identifying speakers from audio recordings. An important application comes from the need to assess student interactions in collaborative learning environments. Diarization is very difficult in these environments where a single microphone is used to record multiple voices. Although there have been advancements in this field, little progress has been made in the case of noisy and disorganized multi-speaker environments. Current state-of-the-art methods based on Deep Learning require large training databases and can suffer from significant noise interference and bias due to the speaker's accent, age, and gender. In this paper, we are proposing a new method to identify speakers that does not require the use of large training sets. To this end, we use a virtual array of microphones. The signal at the virtual microphones is simulated by extracting the spatial information of the speakers from a single channel audio recording using approximate speaker geometry observed from a video recording. The Room Impulse Responses (RIRs) at the virtual microphones are then estimated using acoustic scene simulations. The RIRs are then used to compute a cross-correlation matrix of possible audio sources. Speaker diarization is performed using the cross-correlation matrices as input to a classifier. For the task of identifying active student speakers in classroom audio, the proposed method significantly outperformed diarization methods performed by Google Cloud and Amazon AWS services.

INDEX TERMS Speaker Identification, Speaker Diarization, Audio Room Simulation, Virtual Microphone Arrays.

I. INTRODUCTION

Speaker identification in crowded rooms remains very challenging. Crosstalk and large amounts of background noise make speaker separation particularly challenging. The significant variations associated with picking up speakers in crowded rooms makes it very difficult to develop ground truths on large datasets. As a result, the use of Deep Learning methods is fundamentally limited on pre-training datasets that may not be representative of the complexities associated with training for crowded rooms.

For a single speaker in a non-crowded room, a typical speaker identification system involves the extraction of speech features such as formant frequencies, pitch contours, and coarticulation from the test samples and classification against a database of training samples [1]. The datasets still need to contain as many training examples as possible and should be updated periodically to maintain a proper performance level [2]. The accuracy of the identification depends on the size of the dataset; the bigger the dataset, the better the accuracy, but the longer the training times [3].

In addition to long training times, datasets are also prone to bias with respect to spoken language and accent. [4] This biasing is usually unintentional and unconscious, and it is the product of the environment where the speech recognition system is developed [5].

The limitations of speech processing systems are more evident in challenging situations such as classroom environments. In this paper, we restrict our attention to speaker diarization in collaborative learning environments where a small group of 2 to 5 students sits around a table (see Fig. 1).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2022.3177584, IEEE Access

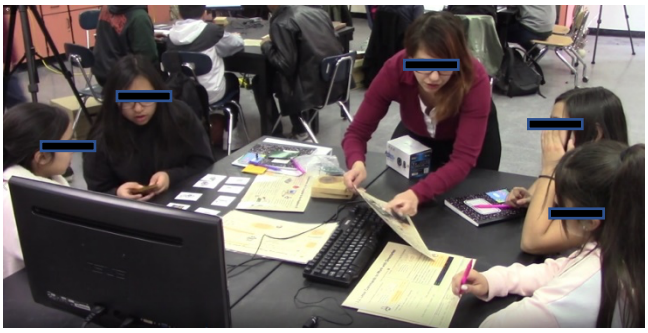

Figure 1: Example of an educational collaborative environment with five speakers in a noisy environment.

In this case, there is strong background interference coming from having up to 5 collaborative groups with over 20 students total, 5 facilitators, 2 teachers, and 5 researchers in the same room. The speakers can take turns to speak, but it is not unusual to have crosstalk, where two or more speakers talk at the same time.

A fundamental problem in educational research is to understand how the classroom material engages the students. To understand how students interact, classroom sessions are recorded and transcribed. An important problem here is to determine which participant is speaking at a particular moment, what she or he has said, and for how long the participant spoke. Manual diarization of meetings is a tedious and time-consuming task, subject in many cases to the interpretation of the transcriber. Automated methods usually require multi-channel audio recordings and are prone to errors due to noise and crosstalk. Also, these systems have limitations in the number of speakers they can process, as well as the length of the audio segments.

While diarization systems do not require enrollment of the speakers, they can only generate abstract labels of a speaker that is active in an audio segment. On the other hand, speaker identification systems can provide non-abstract labels by enrolling the participating speakers. The enrollment process consists of each speaker providing several seconds of noise-free speech without crosstalk. This requirement cannot be met when the data consists of audio recordings of busy meetings with noisy backgrounds. It is thus important to develop speech identification and diarization methods that do not impose any requirement to pre-enroll the speakers.

We present a method for speaker identification and diarization using virtual microphones that does not require prior speaker enrollment. The proposed approach only requires a rough estimate of the speaker geometry that can be derived from video recordings. The approach does not require pre-training, is independent of the spoken language or accent of the participants and works well in noisy environments.

The proposed approach relies on the fact that discriminant information about the 3D geometry of each speaker is embedded in the recorded audio from a single microphone. The basic idea is to recognize speakers using acoustical simulation. As part of the simulation process, the proposed method computes the Room Impulse Response (RIR) for each of the microphones and the speakers and simulates the reception on each of the virtual microphones. The accuracy of the process of computing RIRs is verified through real-life measurements of the correlation patterns. Based on the simulated reception over the virtual microphones, the method computes correlation patterns among the virtual microphones. The recorded audio is then also used to generate different correlation patterns based on hypothesized speaker locations. A classifier is applied to the generated correlation patterns to select the most likely speaker location.

For our approach, we do not consider diarization for multiple speakers within the same group. Our approach however, accounts for significant crosstalk that is the result of strong background interference across groups. Thus, it is possible to address this issue by simply adding an extra microphone for each subgroup of students talking, and then considering the two (or more) subgroups as separate groups. Without an extra microphone, our approach can be adapted for having multiple people speaking simultaneously to the same microphone, as described in our methods section.

This paper is structured as follows: Section II provides background information. Section III describes the proposed method. Section IV describes the implementation of the method, physical validation, and provides experimental results of the proposed method against current state-of-the-art methods. Section V provides concluding remarks.

II. BACKGROUND

Speaker diarization can be summarized as "who said what, and when", and for "how long" [6]. The task of determining for how long one speaker has been active in a multi-participant conversation requires speaker diarization and subsequent identification with non-abstract labels. Most speaker diarization systems work by segmenting the audio using a voice activity detector (VAD), then the segments considered to be only noise are discarded, while those containing speech are analyzed for distinctive features. The different segments are classified with abstract labeling (e.g., speaker 10, speaker 1, etc.), usually by using cluster classification. Speaker identification systems work by enrolling speakers in a database, then extracting speech features to determine if the audio segment contains one of the enrolled speakers. A system that accepts or rejects the identity claim by a speaker is called a *speaker verification system.* In what follows, we present a summary of current state-of-the-art speaker diarization methods. We begin by describing classical speaker diarization of single-channel recordings and continue with speaker diarization using virtual microphone augmentation. We conclude the section with a discussion

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2022.3177584, IEEE Access

of commercial state-of-the-art methods based on Deep Learning.

Hu et al. [7] proposed a method to utilize the reverberant information, known as the Direct-to-Reverberant Ratio (DRR), from a single channel recording for speaker diarization. Hu et al. estimate the DRR using the algorithm from Peso Parada et al. [8] and combine it with a Mel-Frequency Cepstral Coefficient (MFCC) Diarization method proposed by Vijayasenan et al. [9]. The method uses both MFCC and DRR features in combination so a trained system can perform a clustering type of classification. The estimates for the DRRs are computed using features such as Signal-to-Noise ratios, MFCCs, power spectrum, and zero-crossing rates, among others. It is important to note that this work was tested only using simulated meeting recordings with clean audio and assumes that the speakers are stationary (they do not change positions).

Yoshioka et al. [10] described a way of linking several recording devices, such as laptops or mobile phones, to emulate a microphone array. After linking the different devices, the multi-channel audio can be used for speaker diarization. Yoshioka et al. claim a 13.6% diarization rate when 10% of the speech duration contains more than one speaker. This approach is innovative but requires the presence of several recording devices in the meeting room, and therefore it is not achievable with a single microphone recording as in our proposed method.

Another approach to virtual microphone emulation was presented by Katahira et al. [11], Del Galdo et al. [12], and Izquierdo et al. [13]. The authors proposed to simulate arrays of microphones by synthesizing virtual microphone signals using two physical microphones. These methods of microphone emulation are not viable when there is only one physical microphone available.

The most recent single-channel methods for speaker identification and diarization are based on Machine Learning. Deep Belief Networks (DBN) are widely used in speech recognition [14, 15]. In [16], the authors claimed the use of X-vectors can achieve state-of-the-art results in speaker recognition. In [17], the authors showed that Deep Neural Networks using X-vectors often outperformed classic i-vector methods in terms of Equal Error Rate (EER) on standard datasets (e.g., VoxCeleb, NIST SRE 2016, and SWBD). To achieve this increase in performance, X-vector DNNs require the data to be augmented by adding noise and reverberation to the training data. This extra step is not necessary for our proposed method, where the only training needed consists of just a few seconds of audio from each of the speakers.

Pawel Cyrta et al. [18] presented a speaker diarization method using a deep learning architecture that builds the speakers embeddings by training a recurrent convolutional neural network applied directly on magnitude spectrograms. The authors evaluated their method using several available datasets consisting of meetings and broadcast materials from news stations, claiming a reduction of the diarization rate error of 30% when compared with the baseline, the LIUMJ Speaker Diarization system. Compared to our proposed approach, this method was tested using clean datasets with very low levels of noise as compared to noisy recordings of classroom environments. The method also demands large datasets for training the deep learning system.

IBM, Google, Amazon, and Microsoft offer speech processing services based on algorithms that use Deep Learning methods. These tech giants offer powerful computer systems and large databases for these services. Amazon's, Google's, and Microsoft's are all closed-source cloud services that provide an API for speech-to-text processing and speaker diarization. In this paper, we reviewed Amazon's Transcribe (AWS) [19], Google's Cloud [20], and Microsoft Azure Speech Services [21], and experimentally compared Amazon's and Google's against our proposed system.

Amazon's Transcribe accepts either audio files or streaming data, single-channel, and outputs text files with speaker diarization based on a specified number of speakers. Amazons Transcribe works better with 2-5 speakers, and it is language dependent. The length of the audio files is limited to a maximum of 120 minutes. Amazon's Transcribe stores the voice data to train the models [22], unless the users select the option to delete this data. Amazon's functionality can be accessed via REST and SOAP protocol over HTTP [23]. Amazon offers a highly trained set of models called Amazon Transcribe Medical which is aimed at medical transcriptions. Users can also customize the vocabulary to better fit their needs, which is a very desirable feature not offered by Google.

Google's Cloud works similarly, with an interface for long speech and single-channel input for transcription purposes [22]. The optimum number of speakers is set at a maximum of 5. As with Amazon Transcribe, Google offers the option of privacy that prevents data logging that could be used to improve the models. Google's models are optimized for phone conversations or videos, accepting 16kHz or 8kHz audio, respectively, depending on the application [23]. It also offers vocabulary customization. Google offers good scalability, infrastructure, and payment schemes that are considered the best among the technology giants [24].

Microsoft offers speaker diarization via its Cognitive Services. Microsoft's Diarization system ranked first at the VoxSRC challenge 2020 by achieving a diarization error (DER) of 3.71% in development and 6.23% in evaluation testing [25]. The datasets consisted of audio collected from YouTube recordings. For the challenge, the network was trained with 1500 hours of simulated mixed training audio. Microsoft Speaker Recognition [21] offers text-independent speaker recognition/verification. The speakers need to be enrolled to create a signature, which is later compared with the audio to be analyzed. The minimum requirements are 20 seconds of speech for training, and 4 seconds of speech for identification, with

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2022.3177584, IEEE Access

unlimited speaker enrollment, with only one speaker present. In the case of diarization, Microsoft can only recognize up to two speakers. Microsoft Transcription requires multi-channel audio for diarization and the signature of the participating speakers for identification, labeling each speech segment with its correspondent speaker. Microsoft does not collect users' voice samples to train its models. Users can customize their vocabulary and the environment they are expecting to operate, meaning that customization must include noise, indoor or outdoor environments, multi-gender speech, etc. [21].

Although the systems we described above perform well under the environments they were tested and designed to operate, they still have some limitations with respect to training requirements, number of identified speakers, and interfacing. First, these systems are paid services that require connectivity to the API and subsequent batch processing. Our proposed system is completely stand-alone, not requiring any connection, thus allowing for implementation in applications where connectivity may be impaired. The system can run on stand-alone computers without the need to access remote computer clusters or databases. Second, we do not require speech databases; our system is based on physical models that are adapted to the scene we are analyzing. Instead of large datasets, our system requires capturing only about 1 to 2 seconds of audio from each speaker for both training and recognition. In contrast, at a minimum, state-of-the-art systems require tens of seconds of clean audio for training and several seconds of identification. In addition, the lack of databases also eliminates privacy issues, as voice logging is not needed to improve the models. Also, the physical model nature allows, at least in theory, to process an unlimited number of speakers, regardless of the language spoken. Finally, our system has been conceived to operate in noisy environments where microphone arrays and cross-correlation analysis have been proven to be efficient methods for speaker discrimination [26,27].

III. PROPOSED METHOD

We present a top-down diagram of the proposed method in Fig. 2. Our approach relies on estimating the acoustic scene to determine the most likely speaker in each speech segment. In Fig. 2, the acoustic scene is simulated by the room model generator, the source estimator, and the room model estimator. Room model estimation is approximated from a video of the scene (e.g., see Fig. 1). During training, we compute cross-correlation patterns for each possible speaker. Then, during testing, we compute cross-correlation patterns over each audio segment and compare them against the training patterns to determine the speaker that produced the closest correlation pattern. The rest of the current section provides detailed descriptions of each component used in our proposed system. Informed consent was obtained for all study participants.

A. ROOM ACOUSTICS AND SIMULATION

We begin describing our approach using a single source and a single microphone. We then extend our model for several sources and microphones and, finally, we present how we adapt our models to different speaker geometries.

We begin with a simple model based on a single source signal $s(t)$ located in the far-field and recorded by a microphone as a signal $x(t)$ that is the convolution of the Room Impulse Response (RIR), $h(t)$, and additive noise $n(t)$ given by:

$$x(t) = s(t) * h(t) + n(t). \qquad (1)$$

The RIR depends on the locations of the sources, the receiving microphone, the geometry of the room, the absorption of the materials in the room, and the audio frequencies of the sources [28]. The RIR captures audio propagation through a direct path, early reflections, and late reverberations. The direct path component is the Euclidian distance of the source to the microphone, and it is a function of the Time of Arrival (TOA) or the time it takes for the signal to travel from the source to the microphone. The other two components of the RIR are related to the reflections of the sound waves at the walls and objects in the room. The early reflections usually arrive 5 ms after the direct path. The late reverberations arrive 20 or 30 ms after the early reflections begin. The RIR can thus be expressed as the summation of each of the impulse responses corresponding to the direct path and the reflections as given by:

$$h(t) = \sum_{k=1}^{K} h_k(t) + w(t), \qquad (2)$$

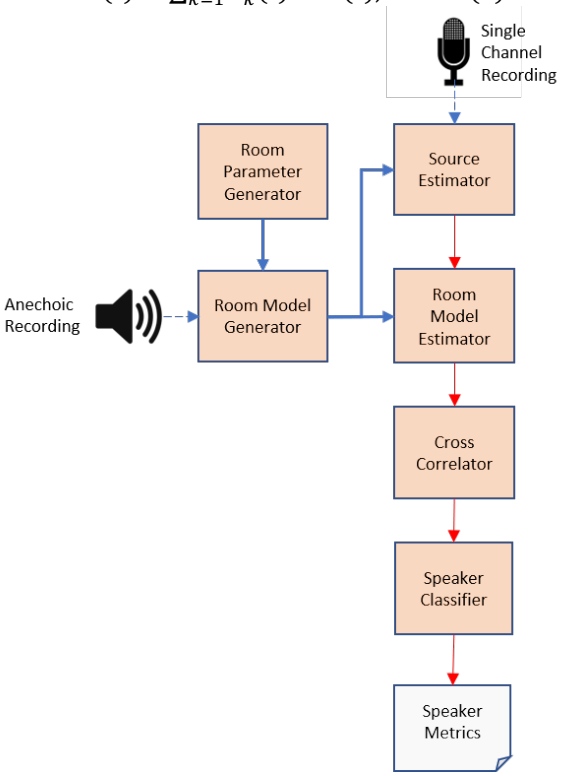

Figure 2: Block Diagram of Proposed Method.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2022.3177584, IEEE Access

where K is the number of reflections, k is used to index specific reflections, and w is measurement noise. The acoustic reflections depend on the absorption of the materials of the room and the frequency components of the acoustic signal [28]. The reverberation signals result from acoustic wave reflections. The late reverberations depend heavily on the frequency components of the sources but, in the case of the early reflections, this influence is minimum [29].

We next extend our model for the case of multiple sources and microphones. Suppose that we have J possible sources: $s_1(t), \dots, s_J(t)$ and N possible microphone signals: $x_1(t), \dots, x_N(t)$. Next, let $h_{j,k}(t)$ denote the RIR that describes the propagation from the j-th source to the k-th microphone. At the k-th microphone, we receive signals from all sources as expressed by:

$$x_k(t) = \sum_{j=1}^{J} s_j(t) * h_{j,k}(t) + n(t), \qquad (3)$$

where $n(t)$ represents additive white noise.

In our collaborative learning environment, we only record $x_1(t)$. We thus need to use (3) to estimate the virtual microphone recordings: $x_2(t), \dots, x_N(t)$ from $x_1(t)$. To use (3), we need estimates for $h_{j,1}(t)$ and their approximate inverses $h_{j,1}^{-1}(t)$. Note that the actual inverses may not exist [28].

We perform the source estimation in two steps. First, we estimate the sources using:

$$s_j(t) \approx x_1(t) * h_{j,1}^{-1}(t). \qquad (4)$$

Second, we plug in the estimated sources from (4) into (3) to compute $x_2(t), \dots, x_N(t)$.

The estimation for $h_{j,k}(t)$ and $h_{j,1}^{-1}(t)$ requires acoustic scene simulation that depends on the geometry of the speakers (sources) and the room where the students are meeting. In what follows, we provide more information on how to set the parameters.

As shown in Fig. 1, we can estimate the relative locations of the speakers and the recording microphone from a single video shot. For example, we can approximate that the table is about 1.5 meters long by 1 meter wide, that the speakers are separated about 0.7 meters from each other, and the speaker's mouths are about 0.24 to 0.25 m from the table. We can also locate the reference microphone in coordinates that are relative to each of the speakers. These are just approximations to create a generic model from where to calculate the RIRs. For the simulation, we consider a simplified model with a small room, large wall absorptions, with a limited number of images due to sound reflections. The acoustic simulation is thus meant to capture early reflections and avoid complex, long-delayed reflections.

B. VIRTUAL MICROPHONES

The spatial locations of the virtual microphones can be directly related to the source audio frequencies. To understand the issues, instead of the classic time-sampling, consider reconstructing an acoustic signal from its 3D spatial samples at a fixed time. In this case, the 3D sampling array separation d between the microphones must be less than half the wavelength λ of the audio source signal. Therefore, d should be

$$d \leq \frac{\lambda_{min}}{2}, \qquad (5)$$

which translates to a maximum frequency of

$$f_{max} \leq \frac{c}{2d}. \qquad (6)$$

For separating the speakers, there is a need to keep the distance between the microphones as large as possible. At larger distances, the correlation patterns will be very different for each speaker. Unfortunately, larger distances imply larger wavelengths and hence smaller spatial frequencies in (6).

For the maximum allowable separation, we select the fundamental frequency of human speech as the smallest spatial frequency that we are interested in. The fundamental frequency of human speech varies from 100 Hz to 120 Hz approximately, with some extreme cases going up to 255-300 Hz (children). Based on a max frequency average of 180 Hz and the speed of sound c = 343 m/s, we set the maximum separation for each microphone to:

$$d \leq \frac{343\ m/s}{2(180Hz)} = 0.95m. \qquad (7)$$

For separating the voices of children, we clearly need to consider much smaller separations that correspond to higher frequencies. After some experimentation, we set $d = 0.05$m for the final collaborative learning environment used for the final classification experiments presented in section IV.C. Here, we note that $d = 0.05$m corresponds to a maximum frequency of 3.43 KHz.

We present the proposed virtual microphone geometry in Fig. 3. Here, the speakers represent the sources J (J = 3, but this number varies). The dark microphone (labeled M1 in the center) is the only real microphone and represents the recording microphone in the actual physical environment. The rest of the microphones are virtual ($N = 5$).

The distance between the microphones determines the Time Difference of Arrival (TDOA) between the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2022.3177584, IEEE Access

microphones. The TDOA is simply defined as the difference in time a signal takes to reach two points separated by a certain distance in space. Initially, let us assume that fig. 2 is an ideal representation where there are no reflections or room absorptions. Then, the TDOA of an active speaker will be unique to at least a pair of microphones, either virtual or physical. For example, if speaker 3 is active, then the TDOA between M5 and M3 will be the same, and different from the TDOA between M2 and M3. These TDOAs are unique for speaker 3. Without loss of generality, we expect the unique property to hold for more complex models that we consider here.

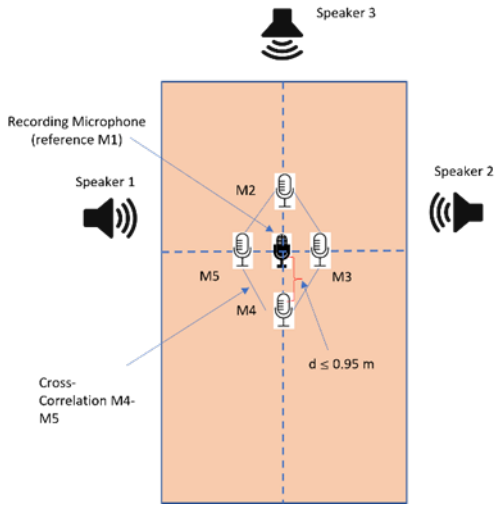

Figure 3: Example placements of the virtual microphones and student speakers for the proposed method.

Let $r_{i,j}(t) = x_i(t) \circledast x_j(t)$ denote the cross-correlation between microphone signals $x_i(t), x_j(t)$. We then define the normalized cross-correlation using:

$$R_{i,j}(t) = \frac{1}{a \cdot b}\, r_{i,j}(t), \qquad (8)$$

where a, b are defined using:

$$a = \sqrt{\sum_t x_i^2(t)} \text{ and } b = \sqrt{\sum_t x_j^2(t)}.$$

We are interested in the location of the peak of the normalized cross-correlation function defined by:

$$T_{i,j} = \text{argmax}\, R_{i,j}(t). \qquad (9)$$

If a source signal propagates to microphones i, j, then $T_{i,j}$ represents the time lag that it takes for the signal to reach j after reaching i. Thus, $T_{i,j} > 0$ implies that the signal arrived at microphone i before j. On the other hand, $T_{i,j} < 0$ implies that the signal arrived at microphone j before i. The cross-correlation matrix of all possible values $T_{i,j}$ will be used for determining the locations of the speakers.

Before using $T_{i,j}$ for speaker recognition, we provide a summary of its properties. First, it is clear that the diagonal is zero. Second, based on the definition, it is clear that $T_{j,i} = -T_{i,j}$. Therefore, the matrix of $T_{i,j}$ values are anti-symmetric. Hence, for differentiating among speakers, we only need to use the entries above or below the diagonal.

To develop a model for the approach, we consider the problem of recognizing one of several possible speakers from a given audio segment. First, we need to construct virtual microphone approximations to $h_{j,k}(t)$. Second, we estimate the correlation matrix features T^m under the assumption that speaker m is talking while all other speakers remain quiet: $s_k(t) = 0, k \neq m$. These T^m models are only computed once here. They do not need to be computed for each audio segment. Third, for each audio segment, we compute T, the cross-correlation matrix of the actual signal. Fourth, we estimate the active speaker by solving:

$$\max_m \text{ match}(T, T^m), \qquad (10)$$

where match (.,.) measures the agreement between T *and* T^m. We thus allocate the speaker $n = m$ that gives the best match among all considered speakers. A simple match function is given by the number of template entries that match as given by:

$$\text{match}(T, T^m) = \sum_i \sum_j \delta\left(T_{i,j} - T_{i,j}^m\right),$$

where $\delta\left(T_{i,j} - T_{i,j}^m\right)$ is the discrete delta function that is 1 when the correlation pattern match with $T_{i,j} = T_{i,j}^m$, and it is 0 when they are different: $T_{i,j} \neq T_{i,j}^m$.

Our approach rejects background noise using hypothesized directions and correlation pattern matching. Firstly, the RIRs model the position of the audio sources. Hence, acoustic sources that do not match the model will generate a different correlation pattern that will not affect our results. We use this approach to model background noise source (e.g., S6 in Fig. 6). Secondly, we note that our use of correlation patterns remains robust with respect to additive white noise. To see this, note that while additive acoustic noise can reduce the cross-correlation coefficient $R_{i,j}$ (see equation (8)), the correlation patterns defined in terms of $T_{i,j}$ only depend on the location of the correlation-pattern maximum (see equation (9)). Thus, a uniform reduction of $R_{i,j}(t)$ throughout time will not be expected to change the location of its maximum.

The proposed method can be extended to address the case when we need to differentiate among more than one speaker talking at the same time within the same group. For this case, we would need to consider a much larger number of correlation patterns. For example, for detecting

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2022.3177584, IEEE Access

up to n active speakers talking at the same time, we have 2^n possibilities. However, the approach can be further complicated by the need to account for having students speaking at very different levels (e.g., loudly versus quietly).

Clearly though, within the group, we are not interested in having multiple speakers talking at the same time. Within the proposed framework, a simple solution would be to place additional microphones within each student subgroup.

IV. IMPLEMENTATION, VALIDATION, AND RESULTS

In this section, we present the experiments conducted to evaluate the capability of the proposed method to identify speakers in audio segments. We begin by applying the principles of section III to an acoustic model based on an approximated room geometry. We validate the physical model using audio experiments. We then provide speaker diarization results and compare our method against Amazon AWS and Google Cloud.

A. ACOUSTIC MODEL PARAMETERS

In Fig. 4, we present the basic setup for our acoustic simulation. We considered a maximum of 5 participants and hence 5 possible source directions. For the cases of 2, 3, or 4 speakers, we simply selected the closest directions from the 5 basic directions of Fig. 4. Hence, we did not re-calibrate our models for every possible variation on the acoustic scene. Furthermore, we also considered all speakers to be at the same height from the table (0.25m). For realistic simulation, we also modeled room noise as a sixth speaker placed at the lower-left part of Fig. 4.

To estimate the RIRs, we used Pyroomacoustics [30,31,32]. Pyroomacoustics is an open-source software system that supports the reproducibility of our results. Pyroomacoustics calculates the RIRs using the Image Source Model Method (ISM) [33]. Image sources are computed based on the distance of each speaker to the absorbing boundaries. For the simulation, the software assumes vertical incidence on the walls and the corners. To control the number of generated sources, we do not consider greatly attenuated sources that are associated with long delays.

For the acoustic simulation, the generation of a large number of simulated sources tends to provide for a better approximation. We simulated the learning environments by assuming acoustic walls with high reflection coefficients located at a short distance behind each speaker. As a result, each speaker generated 2 to 3 reflected sources that were propagated to the virtual microphones. The distance between the table and the ceiling was set to 2m.

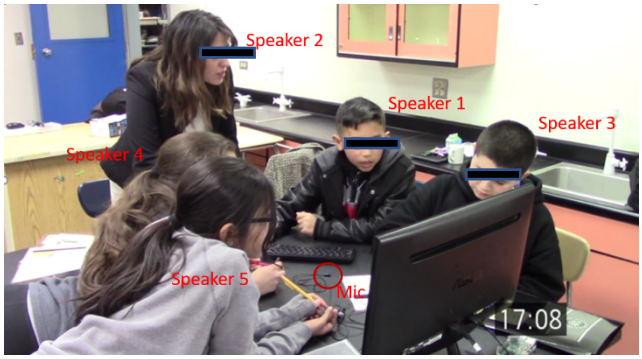

Figure 4: Collaborative environment used for determining speakers from a single central microphone.

B. PHYSICAL MODEL SIMULATION AND VALIDATION

To validate our model simulation approach, we compared correlation patterns generated by our simulation environment and physical measurements using actual microphones and speakers. We consider two setups for validating our approach. First, we compare the performance of the virtual microphone array simulation against a physical microphone array. Second, we perform a controlled audio experiment to understand some of the limitations of the virtual microphone array in collaborative learning environments.

Firstly, for validation using an array of physical microphones, we used the same microphones as the central microphone in our video recordings. The microphones were calibrated using a sinusoidal source of 450 Hz, and we compensated for any physical delay during the audio recordings. The model absorption was empirically set at 0.95. The 2D model included 4 loudspeakers and 5 microphones as depicted in Fig. 5. In the Pyroomacoustics model, sound reflections were simulated using 8 images of the actual audio sources.

As shown in Fig. 5, the physical microphones were placed out closer to the speakers. The larger separations still satisfied the constraint given in equation (5). We note

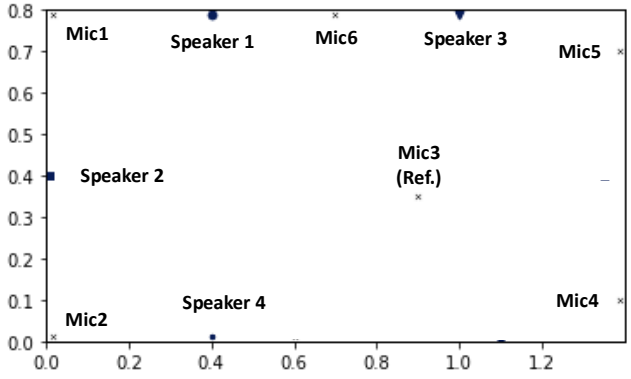

Figure 5: Physical Microphone array setup for validating the virtual microphone simulation environment.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2022.3177584, IEEE Access

TABLE I
CROSS-CORRELATION PATTERN VALIDATION USING PHYSICAL MICROPHONE ARRAY. NUMBERS REFER TO ARRAY INDICES.

	S1		S2		S3		S4	
	SIM.	G.T.	SIM.	G.T.	SIM.	G.T.	SIM.	G.T.
1-3	-29	-17	-62	-81	78	94	36	19
1-6	10	15	-53	-65	91	98	4	3
3-6	40	34	9	15	13	3	-30	-16
2-4	-54	-44	-95	-78	24	25	-99	-78
2-5	-24	-25	-95	-81	81	73	-129	-104
4-5	29	20	0	-1	56	48	-28	-25

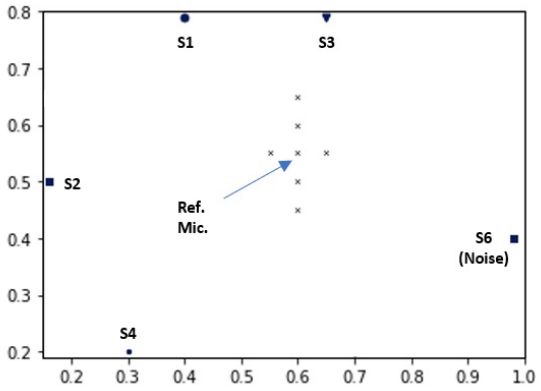

Figure 6: 2-D Model for Controlled Experiments

that larger separations were needed to keep apart the large physical microphones, as opposed to the virtual microphones that do not have such constraints.

To generate the physical measurements, we used an anechoic male voice of 2 seconds duration. The voice was played through the four speakers and was simultaneously recorded through the six microphones. The same signals were simulated using Pyroomacoustics. For each recording, we compute the resulting correlation patterns.

A comparison of the measured correlation patterns is given in Table I. Here, we note that the signals were sampled at 48 kHz, at the same sampling frequency as our video recordings. The results summarized in Table I indicate general agreement between the simulation and the actual physical measurements. In most cases, the error is less than 20%. Most importantly, there are significant differences between the correlation patterns from different speakers. Hence, the simulation model appears to be sufficiently accurate for differentiating speakers based on their positions.

Secondly, we validate our approach in a controlled audio environment. Here, we study the performance of the system in identifying different speakers. For this experiment, we played each source from different loudspeakers in our audio lab and used only the central microphone M3 to capture the audio. To demonstrate the method is not biased to any speech or speaker, speaker 2 (S2) repeats the same speech as speaker 1 (S1) on two occasions. Noise was injected into the environment by playing a compact disk (CD) containing a recording of conversational room noise. The CD player was located at about 2 m from the reference microphone. The audio was segmented using a Voice Activity Detector preserving the noisy segments. The physical dimensions of the model and the location of the microphones were adjusted to better follow the geometry of the acoustic scene depicted in Fig. 4. The final 2-D model is shown in Fig. 6.

We employed the Diarization Error Rate (DER) [34,35] as a metric for Diarization performance. The DER is defined as the fraction of the time that is not attributed correctly to a speaker or non-speech [36]. It is estimated using:

$$\text{DER} = \frac{\text{FA} + \text{Miss} + \text{Overlap} + \text{Confusion}}{\text{Reference Lenght}}, \qquad (11)$$

where FA is the length of False Alarms; Miss is the length of missed speech segments; Overlap is the total length of overlapped speech; Confusion is the total length of misclassified segments, and the Reference Length is the total length of the audio reference. We did not use Overlap for our tests.

The test consisted of playing three separate audio tracks containing only two speakers at a time, and one audio track containing four different speakers. Audio samples A and B were played as speakers 1 and 3, while audio sample C was played as speakers 2 and 4. Audio sample D was played as speakers 1, 2, 3 and 4. The audio was divided into segments with a maximum length of 1.5 s, and a minimum of 0.5 s. We used 1 s long samples from each of the speakers to train the model. As described earlier, for classification, we used a simple match-and-vote classifier where the speaker position with the highest number of cross-correlation matches with respect to the training template is selected as the current speaker.

Table II provides a summary of the results. Overall, the results indicate a good DER of not more than 0.27 in the worst case. The results validate the approach on this limited validation experiment. We present a careful comparison against state-of-the-art methods in the following section.

TABLE II
PROPOSED SYSTEM VALIDATION IN ACOUSTIC LAB ENVIRONMENT

Sample & No. of Speakers	No. of Segs	Correct	FA	Miss	Conf.	DER
A: 2 Speak.	116	98	10	0	8	0.12
B: 2 Speak.	9	7	0	2	0	0.19
C: 2 Speak.	15	12	0	2	1	0.19
D: 4 Speak.	37	27	2	0	8	0.27

C. RESULTS FOR COLLABORATIVE LEARNING ENVIRONMENTS

We next present comprehensive validation of our approach based on actual collaborative classroom videos. We provide detailed analysis for complex audio samples collected during the afterschool program [37]. The corresponding videos contain acoustic scenes like the one shown in Fig. 4, with 2, 3, 4, and 5 primary participants in a single collaborative group. The classroom environment

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2022.3177584, IEEE Access

was very noisy with 5 collaborative small groups each consisting of 3 to 4 students, 5 facilitators, 2 teachers, and 5 researchers in the same room (over 32 speakers).

To process the videos, we assume the baseline model presented in Fig. 6. The parameters of the model are set as described in subsection IV.A. We basically made minor adjustments to the baseline model to reflect the number of speakers and their locations, while maintaining the same geometry for the virtual microphone array.

We constructed 8 carefully chosen examples with 2, 3, 4, and 5 speakers. For the ground truth, we reviewed the videos to provide 0.5 second accuracy within a total duration of three minutes. The ground truth involved a manual review of the video clips to associate lip movements to specific speakers. Here, we note that the proposed method allowed us to identify each speaker based on their location. This was not possible for Amazon AWS and Google Cloud. Instead, for comparison purposes, we mapped the results from Amazon AWS and Google cloud to the most likely speaker that would give the best results.

TABLE III.
ACTIVE SPEAKER TIME ESTIMATION IN THE CLASSROOM

Audio Sample	No. of Speakers	Speaker	Ground Truth Time (s)	Proposed Method		Amazon AWS		Google Cloud	
				Time (s)	Error %	Time (s)	Error %	Time (s)	Error %
1	2	S1	117.00	99.99	14.54	94.52	19.21	127.10	8.63
		S2	27.52	34.62	25.80	74.47	170.60	0.00	100.00
2	2	S1	107.00	113.00	5.61	120.90	12.99	73.40	31.40
		S2	18.03	23.44	30.01	45.46	152.14	66.59	269.33
3	3	S1	6.00	20.69	244.83	9.88	64.67	66.59	1009.83
		S2	102.52	100.52	1.95	143.74	40.21	50.80	50.45
		S3	9.26	13.45	45.25	0.00	100.00	10.29	11.12
4	3	S1	65.74	68.93	4.85	106.36	61.79	80.20	22.00
		S2	27.66	25.38	8.24	37.67	36.19	31.39	13.49
		S3	10.86	15.30	40.88	0.00	100.00	0.00	100.00
5	4	S1	28.29	41.61	47.08	52.19	84.48	0.00	100.00
		S2	11.17	14.69	31.51	8.93	20.05	8.30	25.69
		S3	42.27	68.23	61.41	0.00	100.00	35.00	17.20
		S4	73.84	91.57	24.01	0.00	100.00	94.30	27.71
6	4	S1	24.48	25.39	3.72	78.70	221.49	53.59	118.91
		S2	22.28	13.28	40.39	36.95	65.84	29.19	31.01
		S3	25.75	27.69	7.53	0.00	100.00	15.29	40.62
		S4	1.20	4.20	250.00	38.05	3070.83	3.30	175.00
7	5	S1	20.25	7.99	60.54	0.00	100.00	5.09	74.86
		S2	69.19	64.53	6.74	88.77	28.30	24.90	64.01
		S3	9.41	10.71	13.82	0.00	100.00	0.00	100.00
		S4	43.12	48.86	13.31	60.04	39.24	54.70	26.86
		S5	12.27	10.93	10.92	0.00	100.00	46.60	279.79
8	5	S1	14.28	18.80	31.65	0.00	100.00	6.29	55.95
		S2	34.56	42.05	21.67	53.13	53.73	29.59	14.38
		S3	2.50	3.60	44.00	0.00	100.00	7.49	199.60
		S4	15.23	22.27	46.22	17.61	15.63	11.20	26.46
		S5	47.67	27.54	42.23	56.02	17.52	29.59	37.93

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2022.3177584, IEEE Access

To train our system, we used a noisy sample of 1.8 seconds from each speaker. Here, we note that our method does not depend on the specific speakers. We use training to estimate the RIRs that depend on the relative location of the speakers with respect to the physical microphone. Hence, as long as the speakers return to their seats, we can handle any unknown speaker that takes their seat at the table. Furthermore, as discussed earlier, we only require a rough estimate of the sitting arrangement. There is no need to retrain the model unless there are very significant changes in their seating arrangements.

We used simple voice activity detection to segment the audio. We used a maximum audio segment length of 1.2 seconds and discarded audio segments that were shorter than 0.5 seconds.

We present detailed comparative results in Table III and summary results in Table IV. We begin with a summary of the results and then provide a much more detailed analysis.

From the summary results, it is clear that the proposed method significantly outperformed Amazon AWS and Google Cloud. For the results, the percentage error is given in terms of the actual speaker time as given by:

$$\text{PercentError} = \frac{\text{estimated time} - \text{true time}}{\text{true time}} * 100.$$

In all cases, the proposed method gave substantially lower error rates. With two speakers, the error was acceptable at less than 20%. In comparison, the error rates for all alternative methods were much higher in every possible sample. As we shall describe next, alternative methods failed in many instances.

We provide a detailed analysis of the results in Table IV. We use red highlighting to denote cases of dramatic failures. In such cases, we have that a speaker was completely missed, or the estimated talking time of the speaker had more than a 100% error (e.g., an excessive over-estimation of speaker talking time).

Out of 28 possible speakers across all examples, Amazon AWS gave failing results for 14 cases (50%), Google cloud gave failing results for 10 cases (36%), while the proposed method gave failing results for 2 cases (7%). Here, it is interesting to note that the proposed method never failed to detect a speaker (0% error), while Amazon AWS could not detect any talking time for 10 cases (36%). Google cloud failed to detect any talking time for 4 cases (14%). It is also interesting to note that we have dramatic failure cases for all 8 samples for Amazon AWS and Google Cloud. In contrast, for the proposed method, we have 2 samples with examples of over-estimation, with 6 samples being free of dramatic failures. We use green highlighting to denote cases where the total estimated speaking time gave 20% or less error. Based on this criterion, both AWS and Google Cloud gave satisfactory results in 5 cases (18%) versus 11 cases (39%) for the proposed method.

Overall, it is clear that the problem remains challenging. However, the results from the proposed method demonstrate promise in the proposed approach that cannot be matched by the current state-of-the-art methods.

TABLE IV
SUMMARY OF COMPARATIVE RESULTS FOR ESTIMATING AUDIO DURATION FOR EACH ACTIVE SPEAKER

No. of Speakers	Proposed Method	Amazon AWS	Google Cloud
2	18.99	88.74	102.34
3	57.67	67.14	201.15
4	58.21	470.34	67.02
5	29.11	65.44	87.98
Total	**42.10**	**184.82**	**108.29**

V. CONCLUSIONS

In this paper, we have demonstrated the advantages of using virtual microphones and cross-correlation patterns to identify speakers in very challenging classroom environments from a single-channel recording. Our method presented an error rate that was significantly better than state-of-the-art systems from Amazon AWS and Google Cloud. Furthermore, in contrast with other methods, our proposed approach does not require extensive training, and it is directly applicable in challenging classroom audio environments where clean audio datasets are not available.

ACKNOWLEDGMENT

This material is based upon work supported by the National Science Foundation under Grant No. 1949230 and Grant No. 1842220.

REFERENCES

[1] S. S. Tirumala, S. R. Shahamiri, A. S. Garhwal, R. Wang, “Speaker identification features extraction methods: A systematic review”. *Expert Systems with Applications*, vol. 90, pp. 250-271, 2017. doi: 0957-4174, https://doi.org/10.1016/j.eswa.2017.08.015.

[2] J. Brownlee, “Impact of Dataset Size on Deep Learning Model Skill and Performance Estimates,” Deep Learning Performance, machinelearningmastery.com, para.4, Jan. 2, 2019. [Online]. Available: https://machinelearningmastery.com/impact-of-dataset-size-on-deep-learning-model-skill-and-performance-estimates/. [Accessed March 12, 2021].

[3] J. Yoon and S. O. Arik “Estimating the Impact of Training Data with Reinforcement Learning,” Cloud AI Team Google Research, googleblog.com, para. 2, Oct. 28, 2020. [Online]. Available: https://ai.googleblog.com/2020/10/estimating-impact-of-training-data-with.html. [Accessed Jul. 6, 2021].

[4] A. Koenecke, A. Nam, E. Lake, J. Nudell, M. Quartey, Z. Mengesha, C. Toups, J.R. Rickford, D. Jurafsky S. Goel, “Racial disparities in automated speech recognition,” *Proceedings of the National Academy of Sciences of the United States of America*, April 7, 2020, 117(14):7684-7689, [Online serial]. Available: https://www.pnas.org/content/117/14/7684. [Accessed Aug. 2, 2021].

[5] J. Martin, K.Tang, “Understanding Racial Disparities in Automatic Speech Recognition: The Case of Habitual “be”. Presented at 21st International Conference on Speech Processing and Applications, Shanghai, China, 2020.

[6] X. Anguera, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, O. Vinyals, "Speaker diarization: A review of recent research," IEEE

Transactions on Audio, Speech, and Language Processing, vol. 20, no. 2, pp. 356-370, 2012.
[7] M. Hu, P.P. Parada, D. Sharma, S. Doclo, T.V Waterschoot, M. Brookes, P.A. Naylor, "Single-channel speaker diarization based on spatial features," In Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), 2015, pp. 1-5.
[8] P. P. Parada, D. Sharma, P. A. Naylor, "Non-intrusive estimation of the level of reverberation in speech," in Proc. IEEE International Conf. on Acoustics, Speech and Signal Processing (ICASSP), Florence, Italy, May 2014, pp. 4718–4722.
[9] D. Vijayasenan, F. Valente, and H. Bourlard, "Multistream speaker diarization beyond two acoustic feature streams," in Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP), Dallas, TX, USA, Mar. 2010, pp. 4950–4953.
[10] T. Yoshioka, Z. Chen, D. Dimitriadis, W. Hinthorn, X. Huang, A. Stolcke, M. Zeng, "Meeting transcription using virtual microphone arrays," Microsoft Technical Report MSR-TR-2019-11, July 2019.
[11] H. Katahira, N. Ono, S. Miyabe, T. Yamada, S. Makino, "Nonlinear speech enhancement by virtual increase of channels and maximum SNR beamformer," *EURASIP Journal on Advances in Signal Processing*, 2016, issue 1, article 11, pp. 1-8, 2016.
[12] G. Del Galdo, O. Thiergart, T. Weller and E. A. P. Habets, "Generating virtual microphone signals using geometrical information gathered by distributed arrays," 2011 Joint Workshop on Hands-free Speech Communication and Communication and Microphone Arrays, Edinburgh, pp. 185-190, 2011.
[13] A. Izquierdo, J. Villacorta, L. del Val, L. Suárez, and D. Suárez, "Implementation of a Virtual Microphone Array to Obtain High Resolution Acoustic Images," *Sensors*, vol. 18, no. 2, p. 25, Dec. 2017.
[14] M. Alam, M.D. Samad, L. Vidyaratne, A. Glandon, K.M. Iftekharuddin," Survey on Deep Neural Networks in Speech and Vision Systems," *Neurocomputing*, Volume 417, 2020, pp. 302-321.
[15] D. Sztahó, G. Szaszák, A. Beke, ''Deep learning methods in speaker recognition: A review,'' *Periodica Polytechnica Electrical Engineering and Computer Science,* vol. 65, no. 4, pp. 310–328, Jan. 2021. [Online]. Available: http://arxiv.org/abs/1911.06615. [Accessed Jun. 3, 2021].
[16] J. Villalba, N. Chen, D. Snyder, D. Garcia-Romero, A. McCree, G. Sell, J. Borgstrom, L. Paola García-Perera, F. Richardson, R. Dehak, P. A. Torres-Carrasquillo, N. Dehak, "State-of-the-art speaker recognition with neural network embeddings in NIST SRE18 and Speakers in the Wild evaluations," *Computer Speech & Language*, vol. 60, March 2020.
[17] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey and S. Khudanpur, "X-Vectors: Robust DNN Embeddings for Speaker Recognition," 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018, pp. 5329-5333.
[18] P. Cyrta, T. Trzciński, W. Stokowiec, "Speaker Diarization using Deep Recurrent Convolutional Neural Networks for Speaker Embeddings," Advances in Intelligent Systems and Computing, pp 107-117, 2017.
[19] Amazon AWS, *Amazon Transcribe*, 2021. [Online]. Available: https://aws.amazon.com/transcribe/?nc=sn&loc=1. [Accessed Jul. 10, 2021].
[20] Google's Cloud, *Separating different speakers in an audio recording,* 2021. [Online]. Available: https://cloud.google.com/speech-to-text/docs/multiple-voices. [Accessed Oct. 24, 2021].
[21] Microsoft Azure Product Documentation, "What is Speaker Recognition?" *Microsoft*, Nov. 3, 2021. [Online]. Available: https://docs.microsoft.com/en-us/azure/cognitive-services/speech-service/speaker-recognition-overview. [Accessed: Nov. 6, 2021]
[22] D. Misal, "Google Speech Vs Amazon Transcribe: The War of Speech Technology," *Analytics India Magazine,* Oct. 22, 2018. [Online], Available: https://analyticsindiamag.com/google-speech-vs-amazon-transcribe-the-war-of-speech-technology/. [Accessed: Nov. 6, 2021]
[23] A. Woollacott, "Benchmarking speech technologies," *Academia.edu*, Feb. 2021. [Online]. Available: Academia, https://www.academia.edu/45165394/Benchmarking_Speech_Technologies. [Accessed: Nov. 6, 2021]
[24] M. Saraswat and R. C. Tripathi, "Cloud computing: comparison and analysis of cloud service providers-AWs, Microsoft and Google," In Proc. 2020 9th International Conference System Modeling and Advancement in Research Trends (SMART), 2020, pp. 281-285.
[25] X. Xiao, N. Kanda, Z. Chen, T. Zhou, T. Yoshioka, S. Chen, Y. Zhao, G. Liu, Y. Wu, J. Wu, S. Liu, J. Li, and Y. Gong, "Microsoft speaker diarization system for the VoxCeleb speaker recognition challenge 2020," In Proc. ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021, pp. 5824-5828.
[26] R. Gupte, S. Hawa, and R. Sonkusare, "Speech recognition using cross correlation and feature analysis using mel-frequency cepstral coefficients and pitch," In Proc. 2020 IEEE International Conference for Innovation in Technology (INOCON), 2020, pp. 1-5.
[27] G. Ekim, N. Ikizler, A. Atasoy, and I. H. Cavdar, "A speaker recognition system using by cross correlation," In Proc. 2008 IEEE 16th Signal Processing, Communication and Applications Conference, 2008, pp. 1-4.
[28] I. Tashev, *Sound Capture and Processing: Practical Approaches*. Chichester, West Sussex: John Wiley & Sons Ltd., 2009, pp 351.
[29] S. Tervo, J. Pätynen, and T. Lokki, "Acoustic reflection localization from room impulse responses," *Acta Acustica united with Acustica*, vol. 98, no. 3, pp. 418-440, 2021.
[30] R. Scheibler, E. Bezzam, and I. Dokmanić, "Pyroomacoustics: a Python package for audio room simulation and array processing algorithms," In Proc. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018, pp. 351-355.
[31] R. Scheibler, I. Dokmanić, S. Barthe, E. Bezzam, and H. Pan, "Room Simulation – Pyroomacoustics 0.5.0 documentation", *pyroomacoustics.readthedocs.io/en/pypi-release*, 2016. [Online]. Available:https://pyroomacoustics.readthedocs.io/en/pypi-release/pyroomacoustics.room.html. [Accessed Nov. 6, 2021].
[32] J.B. Allen, D.A. Berkley, "Image Method for Efficiently Simulating Small-Room Acoustics", *The Journal of the Acoustical Society of America*, 1979. DOI 10.1121/1.382599O.
[33] D. Diaz-Guerra, A. Miguel, and A. J. Beltran, "gpuRIR: A python library for room impulse response simulation with GPU acceleration". Multimed Tools Appl 80, 5653–5671. 2021. https://doi.org/10.1007/s11042-020-09905-3.
[34] O. Galibert, "Methodologies for the evaluation of speaker diarization and automatic speech recognition in the presence of overlapping speech," In Proc. INTERSPEECH 2013, 2013, pp. 1131-1134.
[35] Q. Wang, *SimpleDER: a lightweight library to compute Diarization Error Rate (DER).* [Online]. Available: https://pypi.org/project/simpleder/.
[36] X. A. Miró, "Robust speaker diarization for meetings," Ph.D. thesis, Speech Processing Group Department of Signal Theory and Communications Universitat Politècnica de Catalunya, Barcelona, 2006.
[37] The University of New Mexico, "AOLME: Advancing Out-of-school Learning in Mathematics and Engineering". [Online]. Available: https://aolme.unm.edu/. [Accessed Nov. 6, 2021].

ANTONIO GOMEZ holds a BS in Electrical Engineering (1986) and a MS in Engineering Management from Florida International University (2007). He also holds a Graduate Certification in Systems Engineering from New Mexico State University (2009), and he will be receiving his PhD in Engineering in May 2022 from The University of New Mexico. He works as a Principal Systems Engineer for Sandia National Laboratories in both nuclear weapons development, and satellite data processing for nuclear test treaty surveillance. Previously, he worked at Honeywell Federal Manufacturing and Technologies as a Principal Engineer

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2022.3177584, IEEE Access

developing technologies for sub-surface structure detection and characterization using electromagnetic and seismic methods. He has also worked in developing testing platforms for tampering sensors performance evaluation.

MARIOS S. PATTICHIS received the B.Sc. degree with high honors and special honors in Computer Sciences, the Bachelor of Arts with high honors in Mathematics, and a minor in Electrical Engineering from the University of Texas at Austin in 1991. He received a M.S. in Electrical Engineering and a Ph.D. in Computer Engineering from the University of Texas at Austin in 1993 and 1998 respectively. He is currently a Professor in the Department of Electrical and Computer Engineering at the University of New Mexico. His current research interests include biomedical image analysis, image and video processing, video communications, audio processing, and dynamically reconfigurable computer architectures. He holds the 2019-2022 ECE Gardner Zemke Professorship for teaching.

Dr. Pattichis was a Fellow of the Center for Collaborative Research and Community Engagement with the UNM College of Education during 2019 and 2020. He was a recipient of the 2016 Lawton-Ellis and the 2004 Distinguished Teaching Awards from the Department of Electrical and Computer Engineering at UNM. For his development of the digital logic design labs at UNM, he was recognized by the Xilinx Corporation, in 2003 and by the UNM School of Engineering's Harrison Faculty Excellence Award, in 2006. At UNM, he also serves as the Director of the Image and Video Processing and Communications Lab (ivPCL). He was the General Chair of the *2008 IEEE Southwest Symposium on Image Analysis and Interpretation* (SSIAI) and served as a General Co-Chair for the same conference in 2020. He is currently serving as a Guest Associate Editor for the special issue on "Large scale video analytics for clinical decision support," to be published by the *IEEE Journal of Biomedical and Health Informatics* and for the special issue on "Teaching and learning mathematics and computing in multilingual contexts," to be published by *Teachers College Record*.

He has served as a Senior Associate Editor for the *IEEE Transactions On Image Processing* and a Senior Associate Editor for *IEEE Signal Processing Letters*, Associate Editor for *IEEE Transactions on Image Processing*, *Pattern Recognition*, *IEEE Transactions on Industrial Informatics*, and a Guest Associate Editor for two additional special issues published in the *IEEE Transactions on Information Technology in Biomedicine*, and another special issue published in *Biomedical Signal Processing and Control*. In 2022 he was elected Fellow of the European Alliance of Medical and Biological Engineering and Science (EAMBES).

SYLVIA CELEDON-PATTICHIS is Professor of Bilingual/Bicultural Education in the Department of Curriculum and Instruction at the University of Texas at Austin. Her research interests focus on studying linguistic and cultural influences on the teaching and learning of mathematics, particularly with multilingual students. She has led several funded projects that focus on turning language and culture into educational assets in teaching and learning mathematics. Currently, she collaborates in interdisciplinary projects to integrate mathematics and computing in bilingual middle schools.