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/TASLP.2021.3050265, IEEE/ACM Transactions on Audio, Speech, and Language Processing

Temporal Dynamics of Workplace Acoustic Scenes: Egocentric Analysis and Prediction

Arindam Jati, *Student Member, IEEE,* Amrutha Nadarajan, Raghuveer Peri, Karel Mundnich, Tiantian Feng, Benjamin Girault, *Member, IEEE,* and Shrikanth Narayanan, *Fellow, IEEE*

***Abstract*—Identification of the acoustic environment from an audio recording, also known as acoustic scene classification, is an active area of research. In this paper, we study dynamically-changing background acoustic scenes from the egocentric perspective of an individual in a workplace. In a novel data collection setup, wearable sensors were deployed on individuals to collect audio signals within a built environment, while Bluetooth-based hubs continuously tracked the individual's location which represents the acoustic scene at a certain time. The data of this paper come from 170 hospital workers gathered continuously during work shifts for a 10 week period. In the first part of our study, we investigate temporal patterns in the egocentric sequence of acoustic scenes encountered by an employee, and the association of those patterns with factors such as job-role and daily routine of the individual. Motivated by evidence of multifaceted effects of ambient sounds on human psychology, we also analyze the association of the temporal dynamics of the perceived acoustic scenes with particular behavioral traits of the individual. Experiments reveal rich temporal patterns in the acoustic scenes experienced by the individuals during their work shifts, and a strong association of those patterns with various constructs related to job-roles and behavior of the employees. In the second part of our study, we employ deep learning models to predict the temporal sequence of acoustic scenes from the egocentric audio signal. We propose a two-stage framework where a recurrent neural network is trained on top of the latent acoustic representations learned by a segment-level neural network. The experimental results show the efficacy of the proposed system in predicting sequence of acoustic scenes, highlighting the existence of underlying temporal patterns in the acoustic scenes experienced in workplace.**

***Index Terms*—Egocentric audio recordings, workplace acoustic scenes, time-delay neural network, gated recurrent unit.**

I. Introduction

THE human auditory system experiences a multitude of sounds, often dynamically changing over space and time, from its ambient environment. These experiences are influenced by the nature of an individual's daily routine, life style and, notably, occupation. For example, a highway maintenance worker might experience traffic noise throughout the day, while a nurse in a hospital might deal with mostly human speech and equipment noises. Different sounds tend to show diverse effects on human health. For example, nature sounds are found to be beneficial for supporting recovery from a psychological stressor [1]. On the other hand, certain ambient noises tend to have detrimental effects on both physiological and psychological well-being; from immediate change in heart rate variability [2] to disturbed sleep patterns [3]. Researchers have also explored the connection between environmental sounds and elicitation of positive and negative emotional reactions [4]. Increased levels of anxiety and depression have been observed in people from diverse age groups due to annoyance caused by undesirable sounds [5]. Office or workplace sounds and noises are found to cause annoyance [6], [7] and decreased concentration [8] depending on subjective noise sensitivity [6], which might eventually result in lower performance and productivity [9], [10], [11].

Technological advances in wearable devices [12], [13] with the capability of capturing multimodal [14] body signals offer a unique opportunity to study the relationship between the ambient acoustic environment and our everyday life and behavioral patterns. An *egocentric* analysis (that is centered on and evolves around an individual) is particularly interesting since it could illuminate auditory experiences directly from the perspective of the user wearing the mobile device. Automated characterization of the sounds and categorization of the dynamic *acoustic scenes* experienced by the individuals using the wearable devices is an essential step into studying the aforementioned relationship. Following the terminology of [15], [16], we define an "acoustic scene" (*e.g.,* home, office, park *etc.*) as a unique acoustic environment that is generally composed of some specific constituent "sound events" (*e.g.,* door closing, human speech, phone ringing, music system, keyboard, birds chirping *etc.*). Although recent progress in deep neural network (DNN)-based approaches facilitate accurate detection and classification of sound events [17], [18], [19], [20], [21] and acoustic scenes [15], [16], [22], they do not address applications based on real-world egocentric audio recordings collected through off-the-shelf wearable devices, especially in the scenario when environmental sounds are possibly overlaid with user's own speech. We term this as *in-talk* acoustic scene identification, since it tries to infer the background acoustic scene when the user's speech is captured by the worn mobile device. This can be useful in providing context-aware user notifications and experiences, and facilitate environment-aware decision making. Furthermore, to the best of our knowledge, past research has mostly focused on categorizing the underlying acoustic scene given an audio recording, and little work has attempted to model the temporally evolving acoustic scenes that a particular individual experiences through the course of their daily life.

In this paper, we focus our study on in-talk acoustic environments experienced by employees in a workplace; specifically

The authors are with the Signal Analysis and Interpretation Laboratory (SAIL), Ming Hsieh Department of Electrical and Computer Engineering, University of Southern California (USC), Los Angeles, CA 90007, USA. e-mail: jati@usc.edu, shri@ee.usc.edu.

of nurses and other clinical providers in a critical care hospital. Every employee in the workplace experiences a variety of, and potential patterns within, acoustic scenes during the work hours day to day. The acoustic scenes under investigation lie inside a larger set of a workplace (hospital) acoustics, and thus represent audio locales with unique acoustic characteristics. In contrast to the commonly used manual acoustic scene annotation schemes [17], [23], we deploy Bluetooth-based acoustic scene tracking devices [24]. We hypothesize the existence of possible temporal patterns in the sequence of acoustic scenes that an individual experiences in a workplace (in this case hospital), from amongst a finite number of acoustic scenes (will be discussed in Section II). The temporal patterns could be associated with, or driven by, the daily routine, demographics, job-roles, and work habits of the person. Motivated by the psychology literature presented above, we also hypothesize that the pattern and duration of exposure to a specific set of acoustic scenes might be correlated with certain behavioral states and traits of the employees. Finally, in the case of existence of temporal patterns, we investigate whether we can capture them by employing a machine learning model.

To pursue the above hypotheses, we organize the work into two parts.

- **Egocentric analysis**: We define some scalar measures of the temporal dynamics given a sequence of acoustic scenes experienced by a certain employee. Then, we perform statistical analyses to verify whether the measures of dynamics are indeed related to some of their underlying factors like job-roles, daily routines, habits, and demographic information of the employee. Furthermore, we perform a correlation study to explore the relationship of the measures with factors such as job performance. The results of this egocentric analysis will reveal the presence of rich temporal patterns in the acoustic scenes experienced by an employee, and a strong, statistically significant relationship between those patterns and the job-roles, habits, and job performance of the employee.
- **Prediction:** We investigate whether machine learning models can capture the temporal dynamics of acoustic scenes, and thus, allow us to predict the sequence of acoustic scenes from the egocentric audio features collected through the wearable device.

The egocentric analysis presented in this work provides an initial evidence of the presence of temporal patterns in the sequence of acoustic scenes, and thus, builds a foundation for developing machine learning models that can learn those temporal patterns for the prediction task. Moreover, in Section VI-D, we perform the same egocentric analysis with the model's prediction (instead of the true scene labels), which underscores the benefits of employing a machine learning model with the capability of learning the underlying temporal dynamics.

For the modeling and prediction task, we propose a two-stage DNN-based framework. A segment-level model is trained to map a segment of low-level audio features to the corresponding acoustic scene label. A recurrent neural network (RNN) subsequently utilizes the *embeddings* (or latent representations)[1] from a pre-trained segment-level model, and learns to map them to a sequence of acoustic scene labels. This two-stage learning framework helps us to separately analyze the capability of the in-talk audio features to infer the background acoustic scenes, and existence of rich temporal patterns in the sequence of acoustic scenes experienced by a specific user. It is worth mentioning that if there is no specific temporal pattern in the sequence of acoustic scenes experienced by the employees, then the incorporation of the RNN model would not help improve the performance beyond the performance of the segment-level model. We will find that in our setting the RNN model, working on top of the segment-level embeddings, helps learn the temporal patterns better than solely employing the segment-level model.

The rest of the paper is organized as follows. Section II describes the dataset collection procedure. Section III presents an egocentric analysis by exploring the association between workplace acoustic scene dynamics experienced by the individuals, and their relevant behavioral patterns and daily routines. Section IV proposes the two-stage framework for predicting sequence of acoustic scenes. Section V provides the details about the experimental setting, and Section VI describes the corresponding results. Finally, we provide concluding remarks and possible future directions in Section VII.

II. Dataset

The data are from the TILES (Tracking IndividuaL performancE with Sensor) project, a part of the IARPA MOSAIC program[2], that aimed assessing the effect of workplace stressors on employees' affective traits, behaviors, and job performance and productivity. As a part of the project, we deployed wearable sensors to capture multimodal (audio, physiological, location *etc.*) [14] data from nurses and other clinical providers in a large critical care hospital[3]. The data collection from each clinical provider participant lasted over a duration of 10 weeks; each provider could be in one of multiple work shifts (e.g., day, night), and each shift spanned 8 to 12 hours. The current study focuses on audio and location data from a set of 170 participants (47 male and 123 female)[4]. More details about the TILES dataset can be found in [26].

Fig. 1 depicts an illustrative schematic of the five acoustically-relevant locales in the hospital environment considered in this study: *nursing station*, *patient room*, *lab*, *lounge*, and *medication room*. Multiple instances of these locales exist in the actual experimental setting distributed across the hospital. In a hypothetical situation, a nurse might experience more than one acoustic locale because of higher mobility in their job, and hence a richer set of acoustic scenes, while a lab technician might encounter fewer of them in a certain work shift due to the more static nature of the job. Moreover, the temporal pattern in which they encounter the

[1]Throughout the paper, "embedding" indicates the output of a specific hidden layer of a DNN, similar to the nomenclature used in [18], [19], [25].

[2]Multimodal Objective Sensing to Assess Individuals with Context (MOSAIC): https://www.iarpa.gov/index.php/research-programs/mosaic

[3]USC Keck Hospital, Los Angeles, CA, USA.

[4]All the data were collected in accordance with USC's Health Sciences Campus Institutional Review Board (IRB) approval (study ID HS-17-00876).

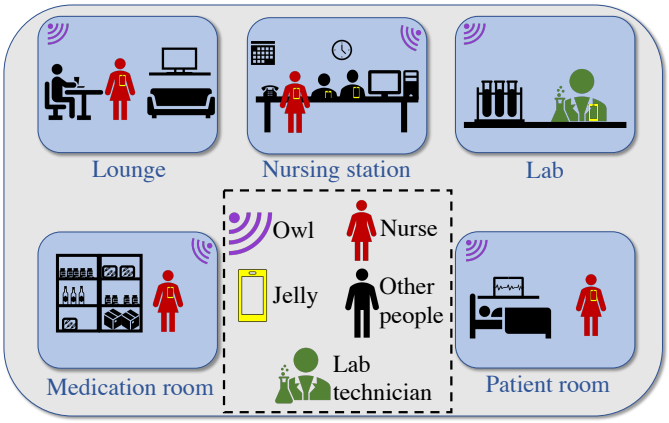

Fig. 1: An illustrative schematic of the hospital acoustic scenes: *nursing station*, *patient room*, *lab*, *lounge*, and *medication room*. Every scene has its own sources of sound events, and thus a unique acoustic characteristics. Note that all the acoustic scenes might have more than one instances (*e.g.,* multiple patient rooms). A nurse (red) might experience several acoustic scenes in a certain work shift, while a lab technician (green) might experience fewer acoustic scenes due to relatively lesser mobility. *Owl* is the Bluetooth hub recording the location context of the user, and they are installed in different places having different acoustic scenes. *Jelly* is the wearable device that captures audio features, and assists Owl in registering the location of the user. The sequence of acoustic scenes a user encounters is derived from the location information captured by multiple Owls. The figure is best viewed in color.

acoustic scenes might vary from one job-type to another. The analyses of the acoustic scene characteristics with respect to job-type, daily routines, and individual behavioral patterns (Section III) lay the foundation for subsequent automated prediction (Section IV) of the temporal sequence of acoustic scenes from audio features.

A. Acoustic features

Acoustic features were collected through a wearable audio recorder (an audio badge called the "TILES Audio Recorder") programmed in-house using a *Jelly*[5] phone device as described in [12]. Audio recordings were triggered with a custom online energy-based Voice Activity Detector (VAD). The HIPAA regulations [27] and the sensitive scenario of the study dictated us from not storing the raw audio signals. Instead, the audio badge sampled the audio signal at 16 KHz, and then extracted (online) several low-level descriptive features from the audio using the OpenSMILE toolkit [28]. The online feature extraction was performed at 60 ms window length with 10 ms shift. The feature set [12] includes energy, prosodic features like pitch, vocal jitter and shimmer, and spectral features like MFCCs. We use the energy and 13 MFCC features along with their derivatives (delta and delta-delta) for the current study, thus creating 42 dimensional feature vectors. The raw audio signal is discarded after the online feature extraction.

[5]Jelly phone device from Unihertz [26], [12].

B. Acoustic scene tracking

In similar previous studies acoustic scenes are annotated by humans after the recording is done [15], [23]. The expensive and time consuming nature of the manual labeling process prohibited us from performing human annotations on our data, especially since our interest was on closely sampled scene labeling for tracking for the entire duration of a work shift. Toward this end, we had installed Bluetooth-based transceivers in all instances of every acoustic scene, the *Owl-in-ones* (*Owl* in short) sensor [24] (see Fig. 1). The Jelly sends Bluetooth pings that are received by the Owls in terms of Received Signal Strength Indicator (RSSI) values. The maximum strength of the RSSI values is determined to register the location, and hence, the background acoustic scene of the participant at a certain time instant.

C. Contextual and user demographic measures

As introduced in Section I, we hypothesize that the temporal patterns of the sequence of acoustic scenes experienced by an employee might be associated with factors such as job-roles, habits, daily routines, and demographics of the individual. Therefore, at the beginning of the TILES study, we collected self-reported information from volunteer participants about their demographics and daily routines [26]. This included information about work shift, hours of work, current position in the hospital, extra jobs held *etc.* A detailed list is presented in Appendix A.

We also hypothesized in Section I that the dynamics of exposure to various acoustic scenes may also be related to some of the behavioral traits and characteristics of the employee. As a part of the study, the participants also completed a comprehensive self-reported assessment (called here, Initial Ground Truth Battery (IGTB) assessment) using a variety of standard psychological instruments to measure multiple traits and behavioral aspects such as personality, organizational behavior, executive function, and smoking and drinking habits [13]. A complete list of the measured IGTB constructs are provided in Appendix B. More detailed description can be found in [26].

III. Egocentric Analysis of Acoustic Scene Dynamics

As introduced in Section I, we investigate the presence of temporal patterns in the acoustic scenes experienced by an employee in the workplace. In this section, we propose various scalar measures that capture the temporal variability in the scene dynamics, and investigate how the underlying temporal patterns are associated with factors such as job-roles, daily routines, and demographics of the employees (as introduced in Section II-C). Moreover, we explore whether some of those measures are related to employee job-performance and cognitive ability.

We represent the temporally varying acoustic scenes experienced by a certain participant by an ordered sequence of

TABLE I: Abbreviation and description of different measures of acoustic scene dynamics incorporated in the egocentric analysis. All of the following measures are computed on a sequence of acoustic scenes.

Abbreviation	Description
std	Standard deviation
range	Range
iqr	Inter-quartile range
change	Normalized number of changes
1-gram-x	1-gram count for acoustic scene class 'x'
2-gram-xy	2-gram count for acoustic scene class pair 'x' and 'y'
entropy	Shannon's entropy measure
tfidf_x	Term frequency–inverse document frequency for class 'x'

enumerated scene labels as defined below. [*lounge*: 1, *patient room*: 2, *nursing station*: 3, *medication room*: 4, *lab*: 5]. For example, in a hypothetical situation, if a participant experiences [*nursing station*, *nursing station*, *patient room*, *patient room*, *patient room*, *lab*] in order, then we represent the temporally varying sequence of acoustic scenes by $[3, 3, 2, 2, 2, 5]$. The enumeration is fixed throughout the experiment.

Definition III.1. *A sequence of acoustic scenes* experienced by the i^{th} participant is given by

$$\mathcal{Y}^i = \left[y_0^i, y_1^i, y_2^i, \ldots, y_{(T_i-1)}^i\right] = \left[y_t^i\right]_{t=0}^{T_i-1}, \tag{1}$$

where T_i is the length of sequence, and $y_t^i \in \{1, 2, \ldots, C\}$ for $C = 5$ acoustic scene *classes*.

Note that $\mathcal{Y}^i$ might not contain uniformly spaced acoustic scenes because of our in-talk analysis (in Section IV-A, we will discuss how in-talk audio segments are extracted), but they are temporally ordered.

A. Acoustic scene dynamics

We propose a set of measures to capture the dynamics of a sequence of acoustic scenes, $\mathcal{Y}^i$. In general, each measure quantifies a certain pattern of a sequence of acoustic scenes for a particular user in terms of a scalar score. TABLE I summarizes the abbreviations and descriptions of the measures we employ for the analysis. The abbreviations will be used in Fig. 2 and Fig. 3. Some of the measures are defined below in detail.

To quantify the mobility of an employee between different acoustic scenes we look at the number of changes in $\mathcal{Y}^i$.

Definition III.2. The *Normalized number of changes* is defined as the total number of changes in acoustic scenes normalized by the length:

$$\Delta\mathcal{Y}^i = \frac{1}{T_i - 1} \sum_{t=1}^{T_i-1} \mathbb{I}\left(\delta_i[t] \neq 0\right), \tag{2}$$

where $\delta_i[t] = \mathcal{Y}^i[t] - \mathcal{Y}^i[t-1]$ is the 1^{st} order difference sequence, and $\mathbb{I}(\cdot)$ is an indicator function *i.e.,* $\mathbb{I}(b) = 1$ if b is True, otherwise $\mathbb{I}(b) = 0$. Higher values of $\Delta\mathcal{Y}^i$ indicate higher mobility of the participant between different acoustic scenes.

The normalized number of changes gives an aggregated measure of variation in the sequence of acoustic scenes, $\mathcal{Y}^i$. A more fine-grained information about the amount of time spent in a particular acoustic scene, or frequency of movement between two different acoustic scenes might reveal important characteristics of $\mathcal{Y}^i$. These can be formally quantified by n-gram counts which are frequently employed for language modeling in the natural language processing domain [29]. The *1-gram count* quantifies the number occurrences of a particular acoustic scene class normalized by the sequence length. Higher values of '1-gram-x' (as abbreviated in TABLE I) indicate the user spends more time in acoustic scene 'x'. The *2-gram count* measures the frequency of scene changes from one class to another, normalized by sequence length. Higher values of '2-gram-xy' (as abbreviated in TABLE I) indicate that the user frequently moves from scene 'x' to scene 'y'.

Furthermore, we try to quantify the uncertainty in the perceived acoustic scenes present in $\mathcal{Y}^i$ through the "entropy" measure. Intuitively, $\mathcal{Y}^i$ for a participant who mostly stays in the same acoustic scene should have different characteristics than the $\mathcal{Y}^i$ experienced by another participant who frequently moves between different scenes in the workplace.

Definition III.3. The *entropy* [30] is defined as the average amount of uncertainty present in the signal. Denoting Y as a random variable for the observed acoustic scene with possible outcomes $y_k \in \{1, 2 \ldots, C\}$, each with probability $P_Y(y_k)$, Shannon's entropy is defined as[6]:

$$H(Y) = -\sum_k P_Y(y_k) \log_2 P_Y(y_k). \tag{3}$$

Finally, we borrow the "tf-idf" measure from information retrieval literature [31]. In the current context, intuitively it denotes how important a particular acoustic scene, c is to a certain sequence of acoustic scene, $\mathcal{Y}^i$ in a collection of several sequences, $\mathcal{S}$.

Definition III.4. The *tf-idf* for a particular acoustic class, c can be defined as:

$$\text{tf-idf}\left(c, \mathcal{Y}^i, \mathcal{S}\right) = \text{tf}\left(c, \mathcal{Y}^i\right) \times \text{idf}\left(c, \mathcal{S}\right). \tag{4}$$

Here, $\mathcal{S} = \left\{\mathcal{Y}^i\right\}_{i=1}^N$ denotes the collection of all sequences of acoustic scenes in the dataset containing N participants. The term frequency $\text{tf}\left(c, \mathcal{Y}^i\right)$ denotes the frequency of occurrence of acoustic class c in the sequence $\mathcal{Y}^i$. The inverse document frequency $\text{idf}\left(c, \mathcal{S}\right)$ measures how much information the scene c provides, and penalizes if it occurs frequently in most of the sequences.

$$\text{idf}\left(c, \mathcal{S}\right) = \frac{1 + N}{1 + \text{df}\left(c\right)} + 1, \tag{5}$$

where $\text{df}\left(c\right)$ is the number of sequences in which the class c is present. The final score is l_2 normalized.

[6]Ideally this is valid for *i.i.d.* observations, which might not be always satisfied in our dataset.

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/TASLP.2021.3050265, IEEE/ACM Transactions on Audio, Speech, and Language Processing

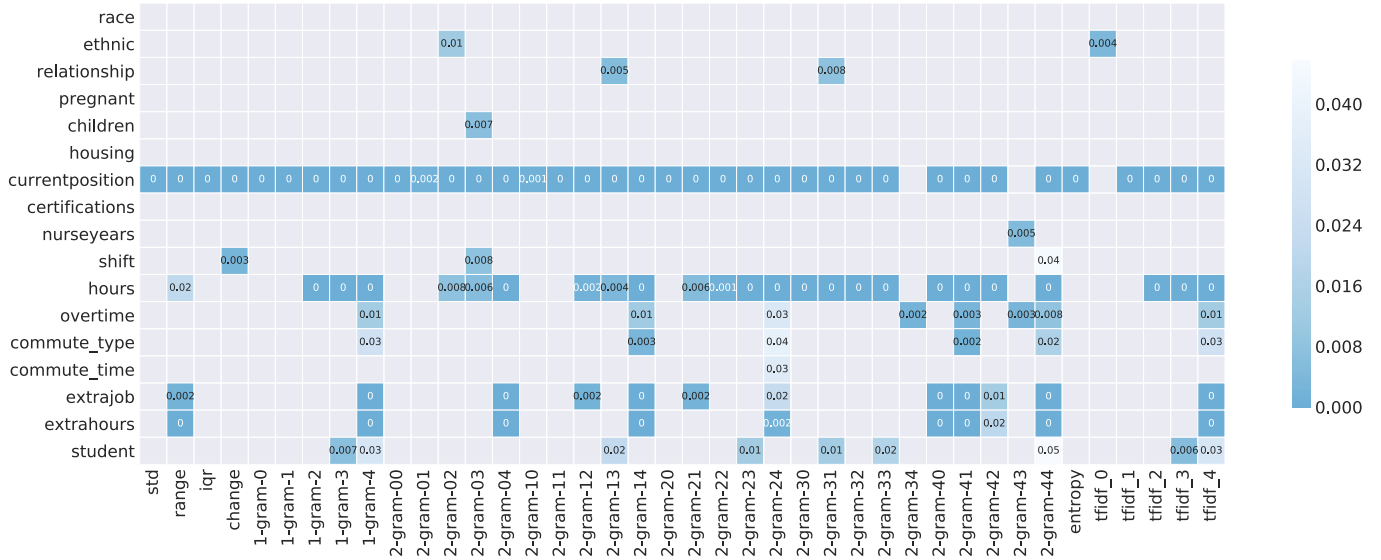

Fig. 2: p-values obtained from the Kruskal–Wallis hypothesis tests between **scene dynamics** (horizontal axis), and individual **daily routines and demographics** (vertical axis). The multiple comparison problem is corrected by the Benjamini–Hochberg procedure. All the indicated p-values are statistically significant, and for the cases when the null hypothesis is rejected. Cases with $p < 0.001$ are shown as 0 for clearer visualization. Empty cells denote observations that fail to significantly reject the null hypothesis as determined by the Benjamini–Hochberg procedure. Please see Section III-B1 and III-B2.

B. Relationship of acoustic scene dynamics with individual demographic and behavioral constructs

In this part, we report correlation analysis and hypothesis tests to explore relationships between different measures of acoustic scene dynamics, and individual demographics and behavioral constructs.

1) Multiple comparisons: All the tests involve *multiple comparisons* [32], [33] *i.e.,* we have multiple variables (*e.g.,* several demographic and daily routine variables as shown in the vertical axis of Fig. 2), and hence, an outcome of a statistically significant observation (in general low p-value, *i.e.,* $p < \alpha = 0.05$) might be purely by chance. As a correction technique, we incorporate the Benjamini–Hochberg procedure [34] to control the False Discovery Rate (FDR). The FDR is defined as the proportion of significant results or "discoveries" that are actually false positives [32]. In brief, Benjamini–Hochberg procedure ranks all the test outcomes by sorting the p-values in increasing order. The maximum p-value which satisfies $p < (i/m)Q$ is a significant observation, and all the smaller p-values are also significant. $(i/m)Q$ is known as the critical value, where i is the assigned rank, and m is total number of tests, and Q is the chosen FDR.

In the following two experiments, we apply Benjamini–Hochberg procedure with 10% FDR to choose the statistically significant observations[7]. For each dynamics measure (*e.g.,* normalized number of changes or "change" column in Fig. 2), the Benjamini-Hochberg procedure is performed for all demographic and daily routine variables (*i.e.,* for all rows in Fig. 2).

[7]Please note that, 10% FDR does *not* mean $\alpha = 0.1$. Significant p-values are chosen based on ranking, and the Benjamini-Hochberg critical value, $(i/m)Q$ [32].

2) Relationship with individual demographics: We attempt to find the association between the proposed measures of scene dynamics and underlying factors such as job-roles, daily routines, habits, and other demographics information. We hypothesize that different groups of employees (*e.g.,* with different job-roles) might experience different patterns in their acoustic scene dynamics. Toward this end, we perform *hypothesis test* to reveal this relationship. Most of the constructs for demographics, job-roles, and daily routines introduced in Section II-C (full list in Appendix A) are categorical in nature. Therefore, we perform Kruskal–Wallis hypothesis test [35] between these individual demographics and the measures of experienced scene dynamics for all the participants. This test is a non-parametric version of one way ANOVA test. The null hypothesis assumes that *data* (here, a particular scene dynamics, *e.g.,* normalized number of changes or "change") from different *categories* (here, different groups of a particular demographic, *e.g.,* different job-roles or `currentposition` in the hospital) come from the same distribution. A resultant low p-value casts doubt on the validity of null hypothesis, and those observations are of particular interest since they denote all the data samples do *not* come from the same distribution. In other words, different groups of participants, with respect to a certain demographic construct, experience different acoustic scene dynamics.

Fig. 2 shows the observations which reject the null hypothesis in Kruskal–Wallis test (observations with $p < 0.001$ are denoted as 0 for clearer visualization). Some notable observations are summarized below:

- Current occupation (`currentposition`) in the hospital has low p-value for most of the measures of acoustic scene dynamics, which intuitively makes sense since 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/TASLP.2021.3050265, IEEE/ACM Transactions on Audio, Speech, and Language Processing

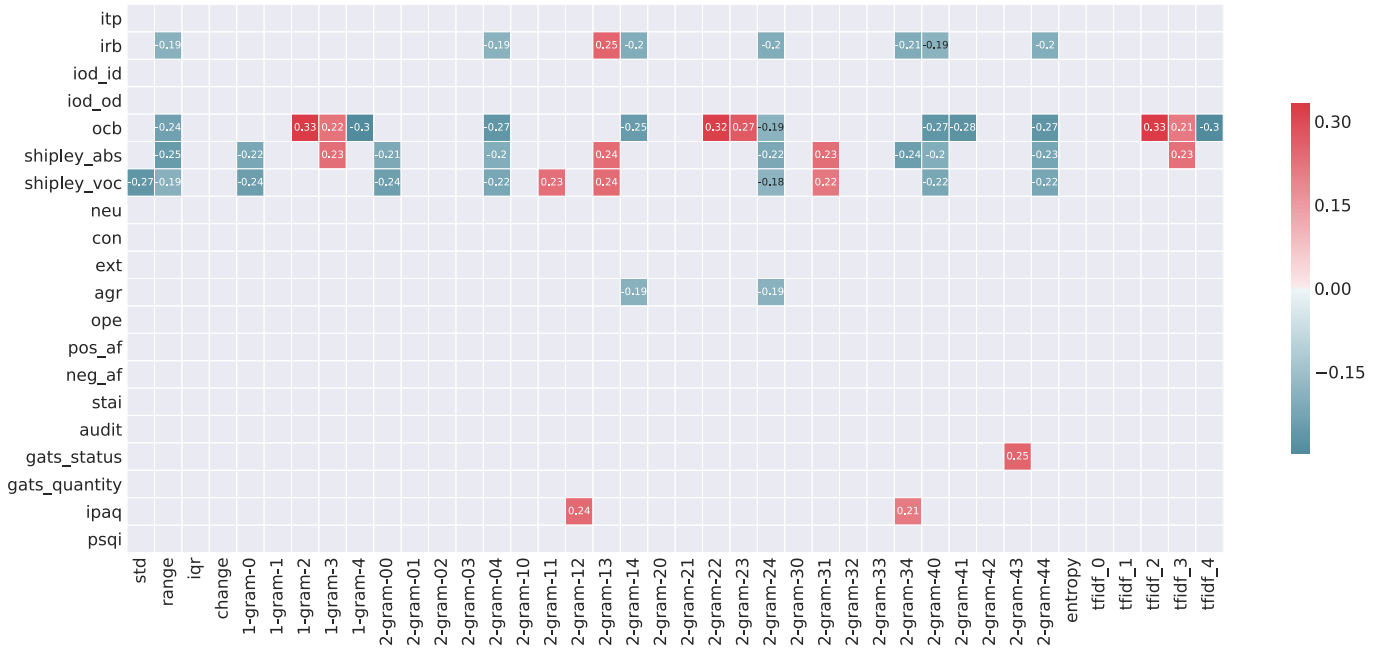

Fig. 3: Spearman's correlation between **scene dynamics** (horizontal axis) and individual **behavioral constructs** (vertical axis). The multiple comparison problem is corrected by the Benjamini–Hochberg procedure. Empty cells denote zero correlations or statistically insignificant correlations. All indicated nonzero correlations are statistically significant as determined by the Benjamini–Hochberg procedure. Please see Section III-B1 and III-B3. Best viewed in color.

job-roles presumably determine the mobility patterns of the hospital employees.

- Rejections of null hypothesis are also observed for work shift (`shift`), hours of work (`hours`), `overtime`, type of commute (`commute_type`), extra jobs held (`extrajobs`), extra hours (`extrahours`), and student status (`student`).

3) Relationship with individual behavioral constructs: We compute Spearman's rank correlation between the scene dynamics measures, and the individual behavioral constructs (Section II-C, and Appendix B) over all the participants. Fig. 3 shows the statistically significant correlations between IGTB constructs and the measures of acoustic scene dynamics. Some notable findings therein are:

- Some job performance related constructs (In-Role Behavior: `irb`, and Organizational Citizenship Behavior: `ocb`) are significantly correlated with a number of the acoustic dynamics measures.
- Cognitive ability measures (Shipley Abstraction: `shipley_abs` and Shipley Vocabulary: `shipley_voc`) tend to manifest a similar trend.
- Physical activity related measures (`ipaq`) also show significant correlations with a couple of acoustics dynamics measures.
- Maximum absolute correlation of $+0.33$ is observed between `ocb`, and 1-gram-2/tfidf_2.
- The only personality construct "aggreableness" (`agr`) shows significant correlations with two dynamics measures.
- Affect-related constructs (`pos_af` and `neg_af`) do *not* show significant correlations with any of the scene dynamics measures.

To summarize, we defined a set of scalar measures that capture specific temporal patterns in a dynamically varying sequence of acoustic scenes experienced by a certain employee. The rejection of the null hypothesis in the Kruskal–Wallis test in Section III-B2 shows that those measures are associated with some of the underlying factors like job-roles and daily routines. Moreover, the presence of significant correlations in Section III-B3 indicates the presence of an inherent relationship between the measures and certain behavioral constructs of the employees, specifically those related to job performance and cognitive ability. These observations confirm the hypotheses introduced in Section I, and also motivate us to investigate further the potential of machine learning methods for modeling the temporal patterns, and inferring the acoustic scene classes directly from the audio features collected via the wearable device.

IV. Automated Prediction of Temporally Varying Acoustic Scenes

As explained in Section I, in contrast to existing acoustic scene classification works, we deal with long egocentric in-talk audio recordings to predict the sequence of observed acoustic scene classes. For this, we propose a two-stage modeling framework. A segment-level model first processes the raw acoustic features. The intermediate representations (or embeddings) learned by the segment-level model are passed on to a recurrent model to learn the temporal dynamics of the sequence of acoustic scenes. Fig. 4 shows an overview of the employed framework.

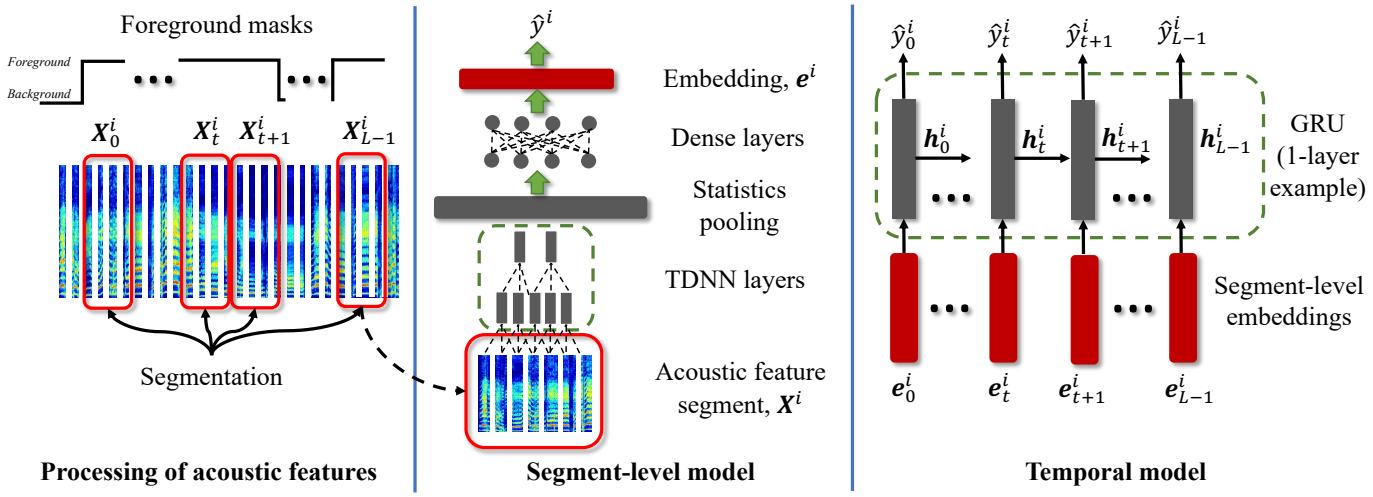

Fig. 4: A two-stage modeling framework for identifying sequence of *in-talk* acoustic scenes. **Left:** Acoustic feature stream is masked with the output of the foreground detection model. This keeps the portions of the stream when there is a possible foreground activity, which are then segmented in windows of fixed length, T_s. **Middle:** The segment-level TDNN model takes a segment of length T_s, and learns to predict the corresponding acoustic scene. **Right:** A GRU model is then trained on top of the segment-level embeddings for learning the sequence of acoustic scenes.

A. Processing of acoustic features

1) In-talk acoustic scene identification: The acoustic features obtained from the TILES audio recorder correspond to any audio activity that happened near the participant during the work shift (due to the presence of VAD module, see Section II-A). As introduced in Section I, in this work we focus on the problem of in-talk acoustic scene identification, *i.e.,* classification of the background acoustic scene while the user (wearing the audio recording device) is presumably talking. The main difference with the traditional acoustic scene identification is that the in-talk acoustic signals originating from ambient sound sources are supposed to be overlaid with speech coming from the user wearing the microphone (in Section I, we have discussed several possible applications). Another distinction comes from the egocentric collection of audio over a long duration, which opens up opportunities for encountering dynamically evolving acoustic scenes.

2) Foreground activity detector: The collection of in-talk acoustic signal requires selecting the portions of the entire audio recording which correspond to possible speech activity by the participant wearing the mobile device. We apply a foreground speaker (*i.e.,* the person wearing the mobile device) detection model developed in [36] to extract the portions in the audio recordings when there is a possible speech activity from the participant. The foreground detection model is trained on a labeled dataset of meeting speech (see [36] for details), and the pre-trained model is used in the current work. The model provides smoothed binary masks that are employed to extract the foreground speech activity (see the left part of Fig. 4). The audio features obtained after the masking contain user's own speech overlaid on background audio coming from different sound sources like machine beeps, door slamming, clock, telephone, and speech from other people. A certain combination of some of these sound events makes an acoustic scene to possess unique characteristic. The masked audio features are segmented in chunks of length T_s, which are subsequently employed for segment-level modeling.

B. Modeling segment-level acoustic scene

We represent a sequence of all temporally ordered segments (might not be uniformly spaced) for the i^{th} participant by

$$\mathcal{X}^i = \left[\mathbf{X}_0^i, \mathbf{X}_1^i, \ldots, \mathbf{X}_{T_i-1}^i\right] = \left[\mathbf{X}_t^i\right]_{t=0}^{T_i-1} \tag{6}$$

where, $\mathbf{X}_t^i$ represents a segment of acoustic feature vectors of length T_s. Note that the corresponding acoustic scene labels are given by $\mathcal{Y}^i$, as defined in (1). The segment-level model ignores the time information, and treats all the segments in $\mathcal{X}^i$ to be independent and identically distributed (*i.i.d.*). Dropping the time index, the segment-level model takes $\mathbf{X}^i$ at input and learns to predict the corresponding scene label y^i.

We employ a Time-Delay Neural Network (TDNN), currently popular in automatic speech recognition [37] and speaker verification [25], for the segment-level modeling. The middle part of Fig. 4 shows a schematic of the TDNN model. The TDNN learns a nonlinear mapping or *embedding* from the input features: $\mathbf{e}^i = f\left(\mathbf{X}^i\right)$. The embedding $\mathbf{e}^i$ is projected to the final output layer with $C = 5$ softmax units emitting posterior probabilities for every acoustic scene class, $P\left(\hat{y}^i = c|\mathbf{X}^i\right)$. The model is trained with an acoustic scene classification objective, and the training is done with all segments from all the participants in the training set (see Section V-C for details). Generally, the embedding learned in this fashion captures the semantic class-level information (for example, in speaker verification the speaker embeddings carry speaker characteristics [25], [38], in our case they are the acoustic scene classes), and thus they could be subsequently used for temporal modeling.

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/TASLP.2021.3050265, IEEE/ACM Transactions on Audio, Speech, and Language Processing

C. Modeling temporal sequence of acoustic scenes

The embeddings learned from the segment-level model help compress a chunk of audio features of length T_s into a fixed dimensional vector (typically 128 dimensional, see Section V). But the segment-level model does not exploit the temporal dependencies available in the data. We hypothesize that the way a particular participant encounters acoustic scenes during his/her work shift might possess a certain temporal pattern. Therefore, a recurrent modeling framework might be more effective in predicting the sequence of acoustic scenes as experienced by that participant.

We adopt Gated Recurrent Units (GRU) [39] neural network to map the sequence of segment-level embeddings $[\mathbf{e}^i_0, \mathbf{e}^i_1, \ldots, \mathbf{e}^i_{L-1}]$ into the sequence of acoustic scene labels $[y^i_0, y^i_1, \ldots, y^i_{L-1}]$ (see the right part of Fig. 4). Dropping participant's index i for simplicity, and considering $\mathbf{e}_t$ to be the input at t^{th} time step, the recurrent transformations for a single layer GRU and the output transformation in this work can be summarized as (see [39] for details):

$$
\begin{aligned}
\text{Reset:} \quad & \mathbf{r}_t = \sigma(\mathbf{W}_{er}\mathbf{e}_t + \mathbf{b}_{er} + \mathbf{W}_{hr}\mathbf{h}_{(t-1)} + \mathbf{b}_{hr}) \\
\text{Update:} \quad & \mathbf{z}_t = \sigma(\mathbf{W}_{ez}\mathbf{e}_t + \mathbf{b}_{ez} + \mathbf{W}_{hz}\mathbf{h}_{(t-1)} + \mathbf{b}_{hz}) \\
\text{Candidate:} \quad & \tilde{\mathbf{h}}_t = \tanh(\mathbf{W}_{e\tilde{h}}\mathbf{e}_t + \mathbf{b}_{e\tilde{h}} + \\
& \qquad \mathbf{r}_t * (\mathbf{W}_{h\tilde{h}}\mathbf{h}_{(t-1)} + \mathbf{b}_{h\tilde{h}})) \\
\text{Hidden:} \quad & \mathbf{h}_t = (1 - \mathbf{z}_t) * \tilde{\mathbf{h}}_t + \mathbf{z}_t * \mathbf{h}_{(t-1)} \\
\text{Output:} \quad & \mathbf{o}_t = \mathbf{W}_o \times \ \text{relu}(\mathbf{h}_t) + \mathbf{b}_o \\
\text{Posterior:} \quad & \hat{\mathbf{y}}_t = \text{softmax}(\mathbf{o}_t)
\end{aligned}
$$

where, $\sigma(\cdot)$ is the sigmoid function, and $*$ is the Hadamard product operation. For a multi-layer GRU (not shown in Fig. 4 for clarity, but employed in our experiment), the hidden state $\mathbf{h}_t$ of a layer becomes the input for the next layer.

V. Experimental Setting

A. Model parameters

Segment-level models: For the segment-level modeling we compare the performance of TDNN model with two other model architectures:

1) *Multi-layer perceptron (MLP)*: It consists of three layers with hidden dimensions $[1024 \rightarrow 1024 \rightarrow 512]$. The embedding dimension is 512. The model has a total of $\sim$ 1.6 M trainable parameters. This model is fed with a concatenated temporal mean and standard deviation of the T_s seconds segments. The remaining models are provided with the temporal features.
2) *Resnet-18*: To investigate the potential of 2D time-frequency convolutions, we experiment with a Resnet-18 model [40]. Two modifications are made: use of 16×2 average pooling to comply with the 42 dimensional features, and having 5 output units for 5 acoustic scene classes. The embedding dimension is 512. The model has $\sim$ 11.1 M parameters.
3) *TDNN small:* It follows the TDNN architecture of [25], except the use of fewer CNN filters and lower statistics dimension. We use 128 filters at every CNN layer, and set the statistics and embedding dimensions to 256 and 128, respectively. The model has $\sim$ 280 k parameters.
4) *TDNN big:* It has 256 filters at every CNN layer, and the statistics and embedding dimensions are 512 and 256, respectively. The model has $\sim$ 954 k parameters.

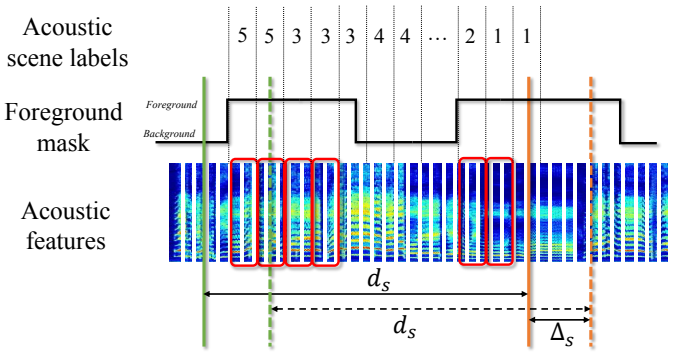

Fig. 5: Mining sequence of acoustic features and scene labels for temporal modeling and evaluation.

Temporal models: We experiment with GRUs of different size and depth. The grid search for model selection is performed over hidden dimensions of $[64, 128, 256]$, and number of hidden layers of $[1, 2, 3]$. The best model is selected from the validation set performance (see Section V-C).

The following parameters are common for all the above models (both segment-level and temporal):

- Relu activation between any two hidden layers.
- Softmax activations in the output layer.
- 30% dropout (for GRU, only applicable if it is a multi-layer GRU).
- Cross entropy loss as the minimization objective. For the temporal model, this is done over all time steps of the sequence.
- Adam optimizer with learning rate 0.001, $\beta_1 = 0.9$, and $\beta_2 = 0.999$.
- Mini-batch size of 64.

B. Data subsets

1) Segment-level experiment: For segment-level training and testing, we mine $T_s = 5$ seconds continuous segments from the foreground masked acoustic features (see Section IV-A). Segments shorter than 5 seconds are ignored, and segments longer than 5 seconds are chunked with no overlap. This creates a total of 269, 170 samples. The acoustic scene class labels are *not* uniformly distributed: $\sim$ 43% samples coming from patient rooms, $\sim$ 37% from nursing station, $\sim$ 11% from lounge, $\sim$ 5% from lab, and $\sim$ 4% from medication rooms.

2) Temporal modeling experiments: For temporal modeling, we mine sequences of audio segments from the foreground masked recordings (Section IV-A) along with the corresponding acoustic scene labels. Fig. 5 shows a schematic for easier interpretation. At a certain time point, we look behind d_s time units (called the *context duration*), and accumulate all the segments (each of length T_s, similar to Section V-B1) which fall under any possible foreground speech activity region. This

TABLE II: Details of the sequences mined from audio recordings. The sequence lengths are in terms of total number of constituent audio segments.

Context duration, d_s	Number of sequences mined	Minimum sequence length	Mean sequence length	Maximum sequence length
15 minutes	$\sim$ 15 k	12	28.00	424
30 minutes	$\sim$ 19 k	12	29.79	678
1 hour	$\sim$ 21 k	12	32.97	1209
2 hours	$\sim$ 21 k	12	39.54	1839

TABLE III: Performance of different models in segment-level acoustic scene classification.

Model	Accuracy (%)	F1 score	McNemar's test
MLP	54.99	0.53	+
Resnet-18	60.14	0.57	+
TDNN small	**63.49**	**0.62**	+
TDNN big	**63.62**	**0.62**	−

serves the purpose of obtaining temporal sequence of acoustic features and scene labels required for the temporal modeling (Section IV-C). For the next sequence, we move forward in time by Δ_s time units. This moving window approach with nonzero Δ_s helps skipping repetitive sequences. In this work, we set Δ_s to be equivalent to 4 audio segments *i.e.,* $\Delta_s \geq 4 \times 5 = 20$ seconds.

An ablation study was performed with different values of context duration, d_s: 15 minutes, 30 minutes, 1 hour, and 2 hours. TABLE II provides different statistics of the data subsets we mine from the audio recordings for the temporal model training and evaluation. For training and testing, we incorporate sequences that have at least 12 time steps (*i.e.,* total duration of all the segments in the sequence is $\geq 5 \times 12 = 60$ seconds)[8] in a given context duration. In other words, this removes all sequences having less than 12 foreground masked segments. For example, if a 15 minutes time window has only two 5 seconds segments with foreground speaker activity, it is ignored in our current analysis.

C. Data splits, cross validation, and model selection

For all the experiments (both segment-level and temporal), we perform 5-fold cross validation ensuring *no participant overlap* between any two folds. This helps mitigating modeling bias that could arise from speaker-related characteristics. We randomly create a validation split from the training participants each time we train. The validation and train splits also have no overlap of participants. We run the training for 50 epochs (because the training loss seems to saturate around that point), and choose the model with the best validation performance. We repeat the overall experiment 5 times, and report the mean accuracy for both segment-level and temporal modeling.

VI. Results and Discussion

First, we present the result for segment-level prediction (modeling is in Section IV-B). Intuitively, this is equivalent to existing approaches that infer an acoustic scene label given an audio recording (in this case a segment of length $T_s = 5$ seconds). Next, we present the results for the proposed two-stage temporal model (modeling is in Section IV-C) for a much longer context duration, d_s (discussed in Section V-B2). The best performing segment-level model becomes a baseline for the proposed two-stage temporal framework.

A. Segment-level prediction

TABLE III shows the unweighted classification accuracy scores and weighted F1 scores in the 5-fold cross validation experiments for all the models. We report the mean scores over 5 random repetitions as explained in Section V-C. We perform McNemar's test [41] to verify statistical significance of the results. It is a non-parametric paired hypothesis test to check whether the difference between the error rates of two classifiers is statistically significant. In TABLE III, a positive $(+)$ outcome indicates that the model outperforms the previous model (1-level higher row in the table), and the difference between their mis-classification rates is statistically significant. The model in the first row has been compared with the chance accuracy which is $\sim 43\%$. We can see the basic MLP model is ahead of chance by 12% in classification accuracy. The Resnet-18 model significantly outperforms the MLP by an absolute 5.15% in accuracy, and 0.04 in F1 score. Both the TDNN models significantly outperform the Resnet-18 model by $\sim 3.4\%$ in accuracy, and 0.05 in F1 score. We hypothesize that the relatively lower performance of Resnet-18 might be because of the large number of trainable parameters compared to the TDNN models, and possible overfitting issues. Both the TDNN models perform similarly as verified by the negative $(-)$ outcome in McNemar's test when we move from TDNN small to TDNN big model. Therefore, we use segment-level embeddings extracted from the TDNN small model for the subsequent temporal analysis due to their lower embedding dimension (helps to train RNN faster).

B. Temporal sequence prediction

The sequence models are trained with the embeddings extracted from the segment-level TDNN small model (embedding dimension is 128). TABLE IV shows the unweighted accuracy scores and weighted F1 scores for the best performing temporal model and the segment-level model for different values of context duration. The segment-level model denotes the already trained TDNN small model (Section VI-A). The performance of the segment-level model is equivalent to splitting the entire audio recording into multiple chunks and inferring the acoustic scene independently for every segment. Therefore, any performance gain achieved by employing a temporal RNN model would highlight the existence of an underlying temporal pattern in the sequence of acoustic scenes experienced by the employees. We report mean scores over 5 random repetitions of the experiment. The chance accuracy is $\sim 45\%$ for all values of context duration. As explained in Section V-B2, for the experiment with sequential acoustic scene labels, we discarded the short sequences. Interestingly, we see an increase in the performance of the segment-level

[8]Note that the time steps might not be equally spaced, as can be seen in Fig. 5

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/TASLP.2021.3050265, IEEE/ACM Transactions on Audio, Speech, and Language Processing

TABLE IV: Performance of different models in predicting temporal sequence of acoustic scenes for different context duration. "Segment-level" denotes the performance of the *TDNN small* model aggregated over all 5 second windows (it does not employ temporal information). "Temporal" denotes the performance of the two-stage model (TDNN embeddings + GRU) which learns and utilizes the temporal pattern.

Context duration, d_s	**Model**	**Accuracy (%)**	**F1 score**	**McNemar's test**
15 min	Segment-level	80.24	0.79	+
	Temporal	**83.52**	**0.83**	+
30 min	Segment-level	78.06	0.77	+
	Temporal	**81.24**	**0.81**	+
1 hour	Segment-level	76.26	0.75	+
	Temporal	**79.33**	**0.79**	+
2 hour	Segment-level	74.97	0.74	+
	Temporal	**77.72**	**0.77**	+

model from the performance reported in Section VI-A. This indicates the inability of the segment-level model in learning the acoustic scenes for the short isolated segments. This might be happening because of errors coming from the foreground detection module (see Section IV-A)[9].

In TABLE IV, a positive $(+)$ outcome of the McNemar's test indicates that the corresponding model significantly outperforms the previous model (1-level higher row in the table). The first row (segment-level model) is compared with chance accuracy. It is evident that the GRU-based temporal models significantly outperform the best segment-level model for all values of context duration in terms of both accuracy and F1 score. We hypothesize that the performance gain arises from the presence of temporal dependencies between the acoustic scenes observed by a certain participant. For example, within a given context duration, a nurse might experience the acoustic scenes in a specific pattern *e.g.,* (s)he might mostly move between nursing station and patient room.

TABLE IV shows us another interesting (somewhat intuitive though) characteristic. The performance of the temporal model decreases as the context duration increases. This might be happening because it is harder to find patterns in the data for longer sequences. The decrease in the performance of segment-level model might also be a factor, since the segment-level embeddings are utilized as features in temporal modeling. An end-to-end temporal modeling might be more helpful in this situation, and this will be discussed in Section VII.

The results of both segment-level and temporal experiments show that the DNN models are able to classify the acoustic scenes with reasonable accuracy from in-talk acoustic features, representing potentially a mixture of ambient sounds and the user's speech. This shows the feasibility of addressing the task of in-talk ambient acoustic scene classification, as well as the ability of the DNN models to learn the mapping from the in-talk acoustic cues.

[9]But, this untraceable because of the absence of human annotated labels, and raw audio signal.

C. Sequence visualization

Fig. 6 visualizes 4 different sequences of acoustic scenes (for 1 hour context duration) along with their predicted versions at different accuracy levels. A closer inspection reveals that there are different types of errors made by the temporal model, including failures at sudden change in the acoustic scenes (*e.g.,* the second subplot from the top), and failures at isolated segments (*e.g.,* third subplot from the top). The errors could arise from either the segment-level embeddings that are utilized to train the temporal model, or the inability of the GRU models to learn the temporal dependencies in those specific situations.

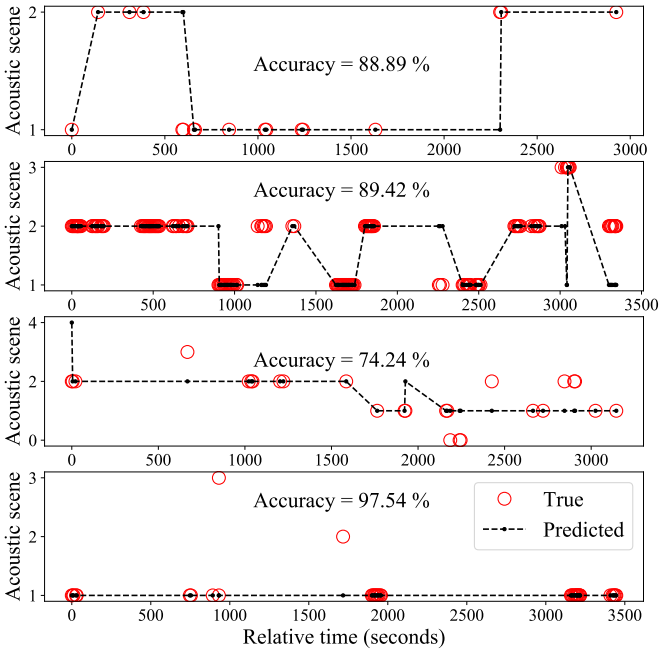

Fig. 6: Visualization of 4 true (red circles, no line) and predicted (black dots and dotted lines) sequences. The accuracy value shown in each subplot indicates the prediction accuracy for that particular sequence.

D. Egocentric analysis with predicted scene sequence

The egocentric analysis presented in Section III was performed with the true sequence of acoustic scenes captured by the Bluetooth trackers. In Section VI-B, we compared performances of both the segment-level and the proposed two-stage temporal model in predicting the sequence of acoustic scene labels from audio features. Here, we analyze how the prediction error affects the egocentric analysis if we perform the analysis with a model's predictions instead of the true scene labels. Any prediction error will have a direct effect on the measures of acoustic scene dynamics that we compute from a sequence of acoustic scenes (in Section III-A). Table V shows the Mean Absolute Error (MAE) between the true measures of scene dynamics, and the predicted ones. The MAE is averaged over all the participants. Thus, a lower MAE indicates better performance by the prediction model since the predicted measures of dynamics are closer to 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/TASLP.2021.3050265, IEEE/ACM Transactions on Audio, Speech, and Language Processing

TABLE V: Mean Absolute Error (MAE) between the measures of temporal dynamics computed with true scene labels and predicted scene labels.

Model	MAE at different context duration d_s			
	15 mins	30 mins	1 hour	2 hours
Segment-level	0.062	0.063	0.066	0.068
Temporal	**0.048**	**0.050**	**0.054**	**0.055**
Relative improvement	22.21%	20.57%	17.68%	18.90%

TABLE VI: Total number of constructs with statistically significant outcome in the egocentric analyses performed with the predicted scene labels. The number inside the parenthesis is the number of constructs that are also observed in the same statistical test with the true labels. The rows with *True Labels* are for comparison purpose.

Construct	Model	#Constructs at different context duration d_s			
		15 mins	30 mins	1 hour	2 hours
Demographics	*True labels*			*13 (13)*	
	Segment-level	5 (5)	6 (6)	7 (7)	6 (6)
	Temporal	8 (**7**)	9 (**8**)	12 (**10**)	10 (**9**)
Behavior	*True labels*			*7(7)*	
	Segment-level	3 (1)	3 (1)	4 (2)	7 (**4**)
	Temporal	5 (**3**)	7 (**4**)	13 (**5**)	8 (2)

true measures. We compare performances of the segment-level model and the proposed two-stage temporal framework. It is evident that MAE is smaller for the temporal model than the segment-level model, which is expected since the former one has achieved better prediction performance (Table IV). The higher value of MAE for longer context duration might possibly be because of higher error in prediction (as can be seen in Table IV).

To perform the egocentric analysis of Section III with the predicted acoustic scenes, we do the same statistical tests (to recap, Kruskal-Wallis for daily routines and demographics, and Spearman's correlation for behavioral constructs), but this time with a model's predictions. Subsequently, we compare the outcomes of the statistical tests performed with the true and the predicted scene labels. TABLE VI shows the total number of constructs (either demographic or behavioral) that are found to have a statistically significant relationship with *at least* one measure of scene dynamics. The number inside the parenthesis indicates how many of those constructs also have significant outcome in the same statistical test performed with the true labels. We present those values for the segment-level and the proposed two-stage temporal model. We can see that, in most of the cases, the predictions from the proposed temporal model find more constructs (compared to the segment-level model) that are also observed in the experiment performed with the true labels. This highlights the efficacy of the proposed two-stage framework in better predicting the temporal sequence of acoustic scenes. A comparison between the two types of construct (demographics and behavior) shows that the predicted labels are able to get more similar outcome as the true labels for the demographic variables. More details can be found in Appendix C.

TABLE VII: Abbreviation, description, and data type of different demographic and daily routine information incorporated for egocentric analysis.

Abbreviation	Description	Data type
race	Race	Categorical $(1-7)$
ethnic	Ethnicity	Categorical $(1-2)$
relationship	Relationship status	Categorical $(1-4)$
pregnant	Pregnancy status	Categorical $(1-2)$
children	Number of children below 18 years of age	Integer $(0-15)$
housing	Housing status	Categorical $(1-4)$
currentposition	Current position in the hospital	Categorical $(1-8)$
certifications	Certifications regarding occupation	Categorical $(1-7)$
nurseyears	Years in the current profession	Integer $(1-80)$
shift	Work shift	Categorical $(1-2)$
hours	Hours of work per week	Integer $(1-100)$
overtime	Overtime hours per month	Integer $(0-200)$
commute_type	Means of communicating to the workplace	Categorical $(1-6)$
commute_time	Quantized time for communicating to the workplace	Categorical $(1-6)$
extrajob	Having at least one extra job	Categorical $(1-2)$
extrahours	Extra hours spent at extra job(s) per week	Integer $(0-100)$
student	Whether enrolled in a certain student program	Categorical $(1-9)$

VII. Conclusion and Future Directions

We characterized the temporal dynamics of the acoustic scenes observed in a workplace. Specifically, we studied the temporally evolving acoustic scenes experienced by nurses and other clinical providers in a large hospital environment from egocentric audio recordings collected with wearable microphones. The acoustic scene labels are obtained via Bluetooth hubs installed in the hospital.

In the first part of our study, we investigated the presence of underlying temporal patterns in the sequence of acoustic scenes experienced by a certain individual. To this end, we characterized the temporal dynamics of the acoustic scenes by proposing a set of measures that try to capture the variability in the acoustic scenes experienced by an individual. We showed that some of those measures are strongly associated with a number of driving factors including those related to job type, work hours, and extra jobs. Furthermore, we found the patterns of exposure to a set of acoustic scenes are correlated with variables like job performance and cognitive ability.

The second part of the study focused on modeling the temporal dynamics. We proposed a two-stage deep learning framework to predict the sequence of acoustic scenes from egocentric audio features. A TDNN-based segment-level model was trained to learn the acoustic scenes from short segments of audio features. The acoustic scene embeddings extracted from the trained segment-level model were utilized in the next stage of learning *i.e.,* the GRU-based temporal model to directly predict the sequence of acoustic scenes. The extensive experiments and results showed the presence of rich temporal patterns in acoustic scenes encountered by the participants. Specifically, the proposed two-stage temporal model was found to achieve superior performance to the baseline segment-level model in predicting the sequence of

TABLE VIII: Abbreviation, description, domain, and data type of different behavioral constructs incorporated for egocentric analysis.

Abbreviation	Description	Domain	Data type
itp	Individual Task Proficiency	Job Performance	Likert scale $(1-5)$
irb	In-Role Behavior	Job Performance	Likert scale $(1-7)$
iod_id	Interpersonal and Organizational Deviance Scale / Interpersonal Deviance	Job Performance	Frequency scale $(1-7)$
iod_od	Interpersonal and Organizational Deviance Scale / Organizational Deviance	Job Performance	Frequency scale $(1-7)$
ocb	Organizational Citizenship Behavior	Job Performance	Integer $(0-8)$
shipley_abs	Shipley Abstraction	Cognitive ability	Integer $(0-25)$
shipley_voc	Shipley Vocabulary	Cognitive ability	Integer $(0-40)$
neu	Neuroticism	Personality	Likert scale $(1-5)$
con	Conscientiousness	Personality	Likert scale $(1-5)$
ext	Extraversion	Personality	Likert scale $(1-5)$
agr	Agreeableness	Personality	Likert scale $(1-5)$
ope	Open-Mindedness	Personality	Likert scale $(1-5)$
pos_af	Positive Affect	Affect	Likert scale $(1-5)$
neg_af	Negative Affect	Affect	Likert scale $(1-5)$
stai	State-Trait Anxiety Inventory	Anxiety	Likert scale $(1-5)$
audit	Alcohol Use Disorders Identification Test	Health $-$ Alcohol use	Integer $0-40$
gats_status	Global Adult Tobacco Survey $-$ status	Health $-$ Tobacco use	Categorical (current, past, or never)
gats_quantity	Global Adult Tobacco Survey $-$ quantity	Health $-$ Tobacco use	Integer, tobacco units in past week
ipaq	International Physical Activity Questionnaire	Health $-$ Physical activity	Integer, minutes in the past week
psqi	Pittsburgh Sleep Quality Index	Health $-$ Sleep	Float $(0-21)$

acoustic scenes.

In summary, we provided a comprehensive study of dynamically evolving background acoustic scenes from the egocentric perspective of an employee in a workplace. The egocentric analysis revealed rich temporal patterns in the perceived acoustic scenes, which were also found to be strongly associated with a number of underlying job-related factors. This built a foundation for developing machine learning models that can learn those temporal patterns in order to predict the sequence of acoustic scenes directly from audio features. The improvements obtained by employing a temporal model over the segment-level model, in turn, highlighted the existence of rich temporal patterns in the egocentric sequence of acoustic scenes.

There are several future research directions.

- In the online acoustic feature acquisition part, more distinctive features like log mel energies might be considered in the future, since they were found to have superior performance in a variety of sound event detection tasks [42].
- Future research can be performed on disentangling the environmental sounds from user's speech to have better prediction accuracy. Several novel unsupervised disentanglement methods can be found in recent literature [43].
- An extension of the proposed two-stage training framework would be to perform end-to-end training of the temporal model directly from the raw acoustic features. An inspection can be done on seamless acoustic scene detection (no foreground activity detection). The amount of training data to process would be however a challenge for that approach, and thus incorporation of efficient data sub-sampling techniques might be beneficial.
- Real-time implementation of the proposed models, and analyzing the feasible speed of prediction might be another research problem.
- The efficacy of the temporal prediction model, and the association between scene dynamics and job characteristics can provide useful insights in devising novel applications such as frameworks that can automatically generate movement statistics of the employees between different workplace acoustic scenes. Moreover, the findings about the correlation between the scene dynamics and behavioral states of the employees can inspire further work on building behavioral models [44] that can predict the behavioral states and traits directly from acoustic data.

Appendix A
Demographics and daily routines

TABLE VII lists the information about demographics and daily routines that are used for the analysis. The abbreviations are utilized in Fig. 2.

Appendix B
Behavioral constructs

TABLE VIII tabulates the MOSAIC Initial Ground Truth Battery (IGTB) constructs along with their description, domain, and data type. The abbreviations are utilized in Fig. 3.

Appendix C
Detailed results: Egocentric analysis with the sequence of predicted scenes

In Section VI-D we compared the outcomes of the statistical tests performed on true and predicted scene labels by considering the presence of a particular construct if at least one measure of dynamics achieved statistically significant

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/TASLP.2021.3050265, IEEE/ACM Transactions on Audio, Speech, and Language Processing

(a) $d_s = 15$ mins, Segment-level (b) $d_s = 15$ mins, Temporal

(c) $d_s = 30$ mins, Segment-level (d) $d_s = 30$ mins, Temporal

(e) $d_s = 1$ hour, Segment-level (f) $d_s = 1$ hour, Temporal

(g) $d_s = 2$ hours, Segment-level (h) $d_s = 2$ hours, Temporal

Fig. 7: Word clouds showing outcomes of the egocentric analysis of daily routine and demographics performed with model predictions. Predictions of the segment-level model and the proposed two-stage temporal model are compared at different context duration, d_s. *Green*: the construct is also present for true labels, *Red*: not present for true labels, *Word size*: proportional to the total number of measures of dynamics having significant outcome.

(a) $d_s = 15$ mins, Segment-level (b) $d_s = 15$ mins, Temporal

(c) $d_s = 30$ mins, Segment-level (d) $d_s = 30$ mins, Temporal

(e) $d_s = 1$ hour, Segment-level (f) $d_s = 1$ hour, Temporal

(g) $d_s = 2$ hours, Segment-level (h) $d_s = 2$ hours, Temporal

Fig. 8: Word clouds showing outcomes of the egocentric analysis of behavioral constructs performed with model predictions. Predictions of the segment-level model and the proposed two-stage temporal model are compared at different context duration, d_s. *Green*: the construct is also present for true labels, *Red*: not present for true labels, *Word size*: proportional to the total number of measures of dynamics having significant outcome.

observation. Here, we consider all the measures that show statistically significant outcome and reflect that count in the form of word clouds. Figure 7 shows the word clouds for demographics and daily routines. A green word denotes that the particular construct also has a significant outcome in the test performed with true labels, whether a red word indicates that it does not have a significant outcome in the test with true labels. A word with larger size denotes that relatively higher number of measures of scene dynamics give statistically significant outcome (but it might be green or red as described above). From Figure 7, we can see that the proposed two-stage temporal model is able to produce higher number of outcomes that are similar to the test with true labels, although it generates some other outcomes as well. Figure 8 shows similar word clouds for the egocentric analysis with the behavioral constructs. In most of the cases, the temporal model still tends to produce more observations that are similar to those of the true labels (except for $d_s = 2$ hours which is also evident from the counts presented in TABLE VI). A comparison between Figure 7 and Figure 8 shows that more similar (with the true test outcomes) observations are found with demographics than behavioral variables.

ACKNOWLEDGMENT

The research is based upon work supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via IARPA Contract No 2017 - 17042800005. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing official policies or endorsements, either expressed or implied, of ODNI, IARPA, or U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon.

The authors would also like to thank Md Nasir for his valuable comments regarding temporal dynamics measures.

REFERENCES

[1] J. J. Alvarsson, S. Wiens, and M. E. Nilsson, "Stress recovery during exposure to nature sound and environmental noise," *International*

journal of environmental research and public health, vol. 7, no. 3, pp. 1036–1046, 2010.

[2] U. Kraus, A. Schneider, S. Breitner, R. Hampel, R. Rückerl, M. Pitz, U. Geruschkat, P. Belcredi, K. Radon, and A. Peters, "Individual daytime noise exposure during routine activities and heart rate variability in adults: a repeated measures study," *Environmental health perspectives*, vol. 121, no. 5, pp. 607–612, 2013.

[3] A. Muzet, "Environmental noise, sleep and health," *Sleep medicine reviews*, vol. 11, no. 2, pp. 135–142, 2007.

[4] P. Moscoso, A. Eldridge, and M. Peck, "Emotional associations with soundscape reflect human-environment relationships," *Journal of Ecoacoustics*, vol. 1, 2018.

[5] M. E. Beutel, C. Jünger, E. M. Klein, P. Wild, K. Lackner, M. Blettner, H. Binder, M. Michal, J. Wiltink, E. Brähler *et al.*, "Noise annoyance is associated with depression and anxiety in the general population-the contribution of aircraft noise," *Plos one*, vol. 11, no. 5, 2016.

[6] G. Belojević, E. Öhrström, and R. Rylander, "Effects of noise on mental performance with regard to subjective noise sensitivity," *International archives of occupational and environmental health*, vol. 64, no. 4, pp. 293–301, 1992.

[7] C. Clark and S. A. Stansfeld, "The effect of transportation noise on health and cognitive development: A review of recent evidence," *International Journal of Comparative Psychology*, vol. 20, no. 2, 2007.

[8] S. P. Banbury and D. C. Berry, "Office noise and employee concentration: Identifying causes of disruption and potential improvements," *Ergonomics*, vol. 48, no. 1, pp. 25–37, 2005.

[9] V. Siskova and M. Juricka, "The effect of sound on job performance," in *2013 IEEE International Conference on Industrial Engineering and Engineering Management*. IEEE, 2013, pp. 1679–1683.

[10] C. M. Mak and Y. Lui, "The effect of sound on office productivity," *Building Services Engineering Research and Technology*, vol. 33, no. 3, pp. 339–345, 2012.

[11] P. Nassiri, M. Monazam, B. F. Dehaghi, L. I. G. Abadi, S. Zakerian, and K. Azam, "The effect of noise on human performance: A clinical trial," *Int J Occup Environ Med (The IJOEM)*, vol. 4, no. 2 April, pp. 212–87, 2013.

[12] T. Feng, A. Nadarajan, C. Vaz, B. Booth, and S. Narayanan, "TILES audio recorder: An unobtrusive wearable solution to track audio activity," in *4th ACM Workshop on Wearable Sys. and Apps.* ACM, 2018, pp. 33–38.

[13] M. L'Hommedieu, J. L'Hommedieu, C. Begay, A. Schenone, L. Dimitropoulou, G. Margolin, T. Falk, E. Ferrara, K. Lerman, and S. Narayanan, "Lessons learned: Recommendations for implementing a longitudinal study using wearable and environmental sensors in a health care organization," *JMIR mHealth and uHealth*, vol. 7, no. 12, p. e13305, 2019.

[14] B. M. Booth, K. Mundnich, T. Feng, A. Nadarajan, T. H. Falk, J. L. Villatte, E. Ferrara, and S. Narayanan, "Multimodal human and environmental sensing for longitudinal behavioral studies in naturalistic settings: Framework for sensor selection, deployment, and management," *Journal of medical Internet research*, vol. 21, no. 8, p. e12832, 2019.

[15] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M. D. Plumbley, "Detection and classification of acoustic scenes and events," *IEEE Tran. on Multimedia*, vol. 17, no. 10, pp. 1733–1746, 2015.

[16] A. Mesaros, T. Heittola, E. Benetos, P. Foster, M. Lagrange, T. Virtanen, and M. D. Plumbley, "Detection and classification of acoustic scenes and events: Outcome of the dcase 2016 challenge," *IEEE/ACM Tran. on Audio, Speech and Language Processing*, vol. 26, no. 2, pp. 379–393, 2018.

[17] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 776–780.

[18] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold *et al.*, "CNN architectures for large-scale audio classification," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 131–135.

[19] A. Jati, N. Kumar, R. Chen, and P. Georgiou, "Hierarchy-aware loss function on a tree structured label space for audio event detection," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6–10.

[20] G. Richard, S. Sundaram, and S. Narayanan, "An overview on perceptually motivated audio indexing and classification," *Proceedings of the IEEE*, vol. 101, no. 9, pp. 1939–1954, 2013.

[21] S. Chu, S. Narayanan, and C.-C. J. Kuo, "Environmental sound recognition with time–frequency audio features," *IEEE Tran. on Audio, Speech, and Language Processing*, vol. 17, no. 6, pp. 1142–1158, 2009.

[22] S. Chu, S. Narayanan, C.-C. J. Kuo, and M. J. Mataric, "Where am I? scene recognition for mobile robots using audio features," in *Int. conference on multimedia and expo*. IEEE, 2006, pp. 885–888.

[23] A. Mesaros, T. Heittola, and T. Virtanen, "A multi-device dataset for urban acoustic scene classification," in *Proceedings of the Det. and Classif. of Acoustic Scenes and Events 2018 Workshop*, November 2018, pp. 9–13.

[24] K. Mundnich, B. Girault, and S. Narayanan, "Bluetooth based indoor localization using triplet embeddings," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 7570–7574.

[25] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5329–5333.

[26] K. Mundnich, B. M. Booth, M. L'Hommedieu, T. Feng, B. Girault, J. L'Hommedieu, M. Wildman, S. Skaaden, A. Nadarajan, J. L. Villatte, T. H. Falk, K. Lerman, E. Ferrara, and S. Narayanan, "TILES-2018, a longitudinal physiologic and behavioral data set of hospital workers," *Sci Data*, vol. 7, no. 354, 2020.

[27] D. L. Anthony, A. Appari, and M. E. Johnson, "Institutionalizing HIPAA compliance: organizations and competing logics in us health care," *Journal of health and social behavior*, vol. 55, no. 1, pp. 108–124, 2014.

[28] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*. ACM, 2010, pp. 1459–1462.

[29] D. Jurafsky and J. H. Martin, *Speech and Language Processing: An introduction to speech recognition, computational linguistics and natural language processing*. Upper Saddle River, NJ: Prentice Hall, 2008.

[30] T. M. Cover and J. A. Thomas, *Elements of information theory*. John Wiley & Sons, 2012.

[31] Sivic and Zisserman, "Video Google: a text retrieval approach to object matching in videos," in *Proceedings Ninth IEEE International Conference on Computer Vision*, Oct 2003, pp. 1470–1477 vol.2.

[32] J. H. McDonald, *Handbook of biological statistics*. Sparky house publishing Baltimore, MD, 2009, vol. 2.

[33] G. Rupert Jr, *Simultaneous statistical inference*. Springer Science & Business Media, 2012.

[34] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal statistical society: series B (Methodological)*, vol. 57, no. 1, pp. 289–300, 1995.

[35] W. H. Kruskal and W. A. Wallis, "Use of ranks in one-criterion variance analysis," *Journal of the American Statistical Association*, vol. 47, no. 260, pp. 583–621, 1952.

[36] A. Nadarajan, K. Somandepalli, and S. Narayanan, "Speaker agnostic foreground speech detection from audio recordings in workplace settings from wearable recorders," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6765–6769.

[37] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *16th Annual Conference of the International Speech Communication Association*, 2015, pp. 3214–3218.

[38] A. Jati, R. Peri, M. Pal, T. J. Park, N. Kumar, R. Travadi, P. Georgiou, and S. Narayanan, "Multi-task discriminative training of hybrid DNN-TVM model for speaker verification with noisy and far-field speech," in *Proceedings of Interspeech*, 2019, pp. 2463–2467.

[39] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," in *NeurIPS Workshop on Deep Learning*, December 2014.

[40] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2016, pp. 770–778.

[41] Q. McNemar, "Note on the sampling error of the difference between correlated proportions or percentages," *Psychometrika*, vol. 12, no. 2, pp. 153–157, 1947.

[42] E. Cakir, T. Heittola, H. Huttunen, and T. Virtanen, "Polyphonic sound event detection using multi label deep neural networks," in *2015 international joint conference on neural networks (IJCNN)*. IEEE, 2015, pp. 1–7.

[43] R. Peri, M. Pal, A. Jati, K. Somandepalli, and S. Narayanan, "Robust speaker recognition using unsupervised adversarial invariance," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6614–6618.

[44] S. Narayanan and P. G. Georgiou, "Behavioral signal processing: Deriving human behavioral informatics from speech and language," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1203–1233, 2013.