IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received May 4, 2021, accepted June 6, 2021, date of publication June 28, 2021, date of current version July 6, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3092763

Improving Daily Routine Recognition in Hearing Aids Using Sequence Learning

THOMAS KUEBERT[1], HENNING PUDER [1,2], AND HEINZ KOEPPL[1]

[1]Department of Electrical Engineering and Information Technology, Technische Universität Darmstadt, 64283 Darmstadt, Germany
[2]Sivantos GmbH, 91058 Erlangen, Germany

Corresponding author: Thomas Kuebert (kuebert@gsc.tu-darmstadt.de)

The work of Thomas Kuebert is funded by the Sivantos GmbH and is supported by the "Excellence Initiative" of the German Federal and State Governments and the Graduate School of Computational Engineering at Technische Universität Darmstadt.

This work involved human subjects or animals in its research. The authors confirms that all human/animal subject research procedures and protocols are exempt from review board approval.

ABSTRACT This work focuses on sequence learning to improve the daily routine recognition in hearing aids (HA), where the goal is to personalize the device configuration for each user. We apply the sequence methods on two large real-world data sets. One publicly available set contains the acceleration (ACC) data of one person, Huynh, over seven working days, whereas our set includes the real life of seven subjects over 104 days with ACC and audio data of a HA. For both sets, we design statistical features to represent the recurring routine behavior well. In our comprehensive simulations, we analyze several sequence classifiers learning the temporal relationships of high-level activities. The multi-layer perceptron (MLP) and random forest (RF) as an observation model for the hidden Markov model (HMM) show the best F-measure performance of 85.3% and 91.6% on our set and the Huynh set, respectively. In particular, the MLP-HMM combination strongly improves on both sets compared to the non-sequence classifier MLP by 6.7% and 10.2%. Within the segment error analysis, we show that the sequence classifiers improve the temporal prediction stability by a reduction of insertion errors. Thus, the improved sequence classification helps the user to better address his condition due to preferred HA settings.

INDEX TERMS Sequence learning, hearing aids, human activity recognition, sensors, sensor fusion.

I. INTRODUCTION

The daily routine is a sequence of high-level activities and contains many recurrent situations and environments. It is periodically performed and is stable for a longer time. In contrast, the acoustic scene is highly non-stationary, changes in short time intervals, and can be ambiguous [1]. Due to this behavior, the acoustic classification of hearing aids can lead to frequent unwanted setting changes that are linked to the sound classes, e.g., speech in noise [2]. These modifications of HA parameters, e.g., frequency gains, can be uncomfortable for the wearer. A stable and reliable situation identification is necessary for a natural and subtle HA control.

Our goal is to personalize the configuration to the user's wishes and requirements. Therefore, the ideal HA device setting is specified by the user's intention in a certain situation, which translates to different hearing needs. For example, the user plays football and someone close to him shouts some commands. Thus, the classification system could decide based on the short-term acoustic cues, that the wearer is in a conversation. Therefore, the system activates a directional processing to emphasize this voice. However, the user wants to monitor his total surroundings. Hence, the short-term acoustic cues can be ambiguous, and, for example, the motion behavior needs to be considered over a longer period to gain more reliable scene information. Therefore, we use an acceleration sensor within a HA for a better scene analysis.

To enhance the user satisfaction, we propose to connect the repetitive daily routine situations and environments to a preferred device setting [3]. Ideal device settings can be found with the described approach of [4], where in a situation the HA parameters are optimized based on subjective A and B comparisons. Since the daily routine consists of recurring activity sequences, the sequence learning problem takes into account the consecutive sample relationships to exploit these routine characteristics. Therefore, we apply sequence

The associate editor coordinating the review of this manuscript and approving it for publication was Wenbing Zhao.

learning methods to improve the classification results compared to a classifier without modeling the temporal relationships. This research is not only valuable for hearing aids, but for other hearables or, in general, wearables as well. These devices can profit from the subject- and environment-specific setting adjustments. The following contributions have been achieved in this article:

- We performed the comprehensive sequence experiments on two real-world data sets and analyzed the routine behavior in detail.
- To model the sequences, we designed a strong hidden Markov classifier with two discriminative observation models, random forest and multi-layer perceptron.
- Thereby, we outperformed the prior work of Huynh using a topic model or Gaussian mixture model with the hidden Markov model.
- A thorough classifier performance assessment of the time and segment-based evaluation criteria shows the strong improvement of temporal prediction stability.

The article has the following structure. In section II, methods for daily routine recognition (DRR) and sequence learning are discussed. In section III, the used data sets are presented. In section IV, the DRR processing scheme is introduced and applied to the routine data. Finally, the results are analyzed, and conclusions are made in sections V and VI, respectively.

II. RELATED WORK

For DRR with body-worn sensors, unsupervised topic models have been used to recognize these high-level activities based on clustered acceleration (ACC) data [5], [6]. Further investigations have been carried out on semi-supervised and supervised approaches to reduce annotation effort and test the recognition performances [7]. All this research has been applied to one public data set, which contains the acceleration data of the author Huynh over seven working days [8]. We also processed this data set with our supervised scheme and outperformed the topic model (TM) approach [9]. The major obstacle for further research on model generalization across multiple subjects is the time-consuming recording of data sets. We bridged this gap and built a large real-life data set of multiple subjects in our earlier work on offline and online classification approaches. Thereby, the random forest (RF) and multi-layer perceptron (MLP) network showed the best performance in both scenarios on our data set of seven people featuring ACC and additional audio data [3]. In particular, the acoustic features are very rich for detecting sound events or characteristic acoustic scenes like certain environments [10], activities of daily living [11], conversations [12], or transportation modalities [13]. This effectively complements the analysis of ACC patterns to differentiate, for example, seated activities like office work vs. having a conversation [14]. We confirmed this fact that our audio features are very informative and improve the routine classification in comparison to only applying ACC features [9].

To further improve DRR, this article focuses on approaches to model the sequence behavior [15]. The classical method is the hidden Markov model (HMM), where a model of each activity state generates the observed data and the transitions between them only depend on the previous state (Markov assumption) [16]. In the domain of routine activities, this assumption does not hold, since the neighboring samples have a correlation that can exhibit longer periods, i.e., from minutes to hours. In contrast to gesture recognition or low-level activities, where these primitive movements have a correlation that can last for a few seconds. For example, an HMM can decode from posture sequences the interest of a child performing tasks [17]. A static posture classifier returns probability scores for the classes, such as lean forward or backward, and the HMM deciphers the posture sequence to one of four interest levels. Furthermore, in assembly tasks, a Gaussian mixture model (GMM) fitted on ACC data was the input for an HMM to model the transitions between different working steps [18]. Additionally, a second classifier trained on audio features was fused to the GMM-HMM for an optimized decision-making. The audio properties showed to be beneficial. For DRR on the Huynh data set, the GMM-HMM was applied to recognize the daily routine and was inferior to the TM on a long observation window of a half-hour shifted by 5 minutes [8]. Thus, the GMM is often used as a generative observation model [19], but discriminative models can also be applied and demonstrated a better performance [16]. We cross-compare the GMM with our well-performing MLP and RF of our earlier work. Further methods are recurrent neural networks, where we evaluate the performance of a long short-term memory (LSTM) network, since it demonstrated in lots of activity studies a good outcome [20], [21]. For example, the activities of daily living or gestures, such as household tasks, physical exercises, opening a door or gait parameters, are accurately detected from sensor readings like acceleration or angular velocity data [22]. In particular, the LSTM net favors learning of long-term relationship in data with a natural ordering, which is limited for an HMM due to the Markov assumption [16].

III. DATA SETS

In this article, we propose to improve the daily routine recognition using sequence learning. For this objective, we use the large real-life data set of our earlier work in [3] and the public Huynh set of [8], where the key characteristics are summarized in Table 1. Please refer to the mentioned articles for further details. We give a short overview of our data set containing the real-world routine behavior of seven non-representative subjects. They are younger than typical HA wearers and have a more active lifestyle [23]. The subjects followed their normal routine behavior in an unconstrained way. That is why, the set contains a broad scope of performed activities and reflects the real life. This creates a strong data variability and lots of challenges for our learning task. The total length is $N = 63449$ minutes, which corresponds to a mean duration of about 10 hours a

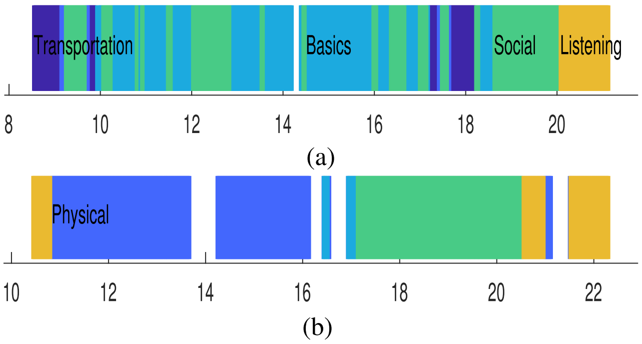


FIGURE 1. Two example days of our data set [h]: (a) a workday and (b) a day on the weekend.

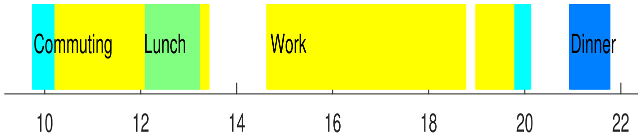


FIGURE 2. Example day of Huynh set [h].

day and over two weeks per participant. We recorded the raw acceleration and precomputed audio features of one hearing aid at a rate of 16 Hz and 2 Hz, respectively. It contains five routine classes: transportation, physical, basics, social, and listening. The users were instructed to annotate the situations based on their intention, i.e., a conversation during a car ride would belong to the social class.

In Fig. 1, two example days, a workday and a day on the weekend, are shown, where the working day has a typical structure. Before and after the office work, there is a transportation scene for commuting and during the day multiple short conversations happen plus two longer ones during lunch and dinner time. These working days follow a natural ordering, which is beneficial to learn these relationships. At the weekend, the day structure is less ordered between different day examples and more free time activities are performed like in Fig. 1 (b), where a longer period of the physical class happens. Besides, people are more socially active during weekends, which is, in particular, the case for young people. Additionally, not all classes are active every day, e.g., the physical class is absent on the working day in Fig. 1 (a). The day structure is variable across subjects and days, i.e., weekends have a different routine order. Analyzing the duration of routine events, most situations have a short duration of a couple minutes and fewer events have a long duration of hours. Thus, the duration probability density function follows an exponentially decaying relationship. In general, we note a variable duration of class events and all 25 possible transitions between the classes occur. Therefore, we have a complex learning task with a high variability and realistic daily routine situations.

The publicly available Huynh data set, mentioned in the related work section, is utilized for reproducibility, evaluation, and comparison [8]. The set contains the real life of the first author, Huynh, in an open setting during working days. The two triaxial ACC sensors were sampling at 100 Hz and were placed at the dominant wrist and in the right pocket. The most frequent routine is (office) work and the remaining three activities - commuting, lunch, and dinner - happen only with a single-digit percentage. Unlabeled segments are not considered for the analysis. The defined classes have a natural order, i.e., only certain transitions are possible, e.g., commuting to work or vice versa, which reduces the possible complexity of the learning task. The class duration has limited variability and a uniform probability density function. The temporal structure is visualized in Fig. 2 and is representative for all seven working days with marginal changes in duration and start times of single class occurrences. In contrast to our set, the Huynh data have less variability of activities and the temporal day structure is comparable between each working day, because it does not contain any days of a weekend or holiday. In both data sets, short recording breaks occur, since the data are uploaded, device batteries are changed, or transmission links need to be restarted. Thus, the learning task is simpler for the Huynh set than for ours and we expect a better detection performance for the Huynh data.

IV. APPROACH

In this section, we introduce our approach to recognize the daily routine. First, we compute a feature representation in two stages and choose the best-performing measures by a feature selection algorithm. Subsequently, the sequence behavior is learned based on a LSTM network and an HMM using various observation models. Finally, we evaluate the performance of these sequence learners.

A. FEATURES

The feature representation of our data set is identical to our earlier work and a detailed description can be found in [3]. It has been designed to separate the routine classes well. Therefore, we extract the ACC features on an activity primitive level, and then build a statistical representation on a routine level for the ACC and audio data. The raw 3D ACC signals have a rate of 16 Hz and the 10D audio features have a rate of 2 Hz. Thus, we fuse the low-level data on the same time grid at a rate of 2 Hz by applying a sliding window of 1 second with 50 percent overlap to the raw ACC data. Thereby, we extract four features: mean, axes correlation, variance, and mean crossing rate. This encodes information about the head and body orientation [24], motion strength [25], conversational gestures [26], and transportation modalities [27]. The window length of 1 second demonstrated in other studies to be beneficial detecting activity primitives, e.g., walking or running [28].

In addition, the precomputed audio features at a rate of 2 Hz are helpful to detect the routine classes using environmental, music, and speech characteristics [12]. These features include 10 measures, such as loudness [29], own voice activation [30], tonality [31], and low-frequency noise [2]. Afterwards, we segment in non-overlapping one-minute frames the

TABLE 1. Overview of the applied the data sets.

Data Set	Daily Routine Classes	Sensors	Duration [Days]	# Users
Our Set [3]	Transportation, Physical, Basics, Social, Listening	Microphone and Accelerometer at the Ear	104	7
Huynh [8]	Work, Commuting, Lunch, Dinner	2 Accelerometers at the Wrist and Pocket	7	1

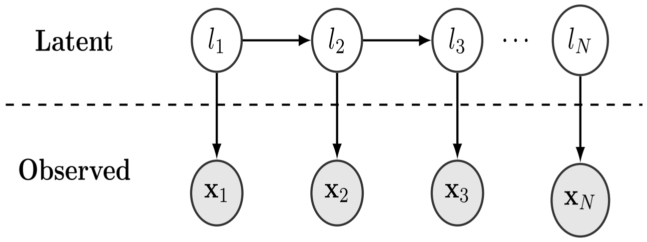


FIGURE 3. The HMM classifier.

22 audio and ACC low-level features to build the high-level routine representation with the statistical quantities: mean, variance, and mean crossing rate. This results in 66 high-level features once per minute, which are normalized to zero mean and unit variance. After a wrapper-based feature selection [32], 16 measures are used for the DRR. With these features, we can distinguish a high number of daily routine situations, for example, a conversation scene from an office work situation. In our prior work [9], we used the feature visualization approach, t-distributed stochastic neighbor embedding, to show these capabilities and analyze the situation clusters in the feature space.

For the Huynh data set, the low-level feature extraction is already done, since the mean and standard deviation of each feature is calculated at a rate of 2.5 Hz due to storage reasons. Afterwards, we apply the same three statistical quantities in one-minute frames, which gives a 36D space. The time-of-day attribute completes the 37D feature space and strongly improved the classification rates due to the very repetitive structure of Huynh's working days [9].

B. SEQUENCE CLASSIFICATION

With the found feature vector $\mathbf{x}$, we classify the routine behavior and environments by exploiting the sequence characteristics. Consequently, we take advantage of the order of feature vectors $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N$ and labels $l_1, l_2, \dots, l_N$ during the learning process, where a label l is chosen out of the K classes $\mathbb{C} = \{c_1, \dots, c_K\}$ for all of the N samples.

Therefore, the sequence learners, HMM with different observation models and LSTM, are selected for the evaluation, which are computationally feasible to use in a HA. We perform sequence learning on the entire training data and apply the fixed model on the unknown test data for the evaluation.

HMM describes the observed temporal sequences of feature vectors as outcomes of hidden states generating these observations as shown in Fig. 3. Since in our supervised case the hidden states correspond to the classes, the HMM is represented by the joint probability distribution [16]:

$$p(l_1, \dots, l_N, \mathbf{x}_1, \dots, \mathbf{x}_N) = p(l_1) \prod_{n=2}^{N} p(l_n|l_{n-1}) \left[\prod_{n=1}^{N} p(\mathbf{x}_n|l_n) \right]. \quad (1)$$

This simplifies the learning procedure of the three quantities on the training data:

- The **initial probability distribution** $p(l_1)$ describes the probability to start in a certain class. This is set to uniform distribution with probability $\frac{1}{K}$ that the observation model solely determines the class decision for the first sample, i.e., $p(l_1, \mathbf{x}_1) = \frac{p(\mathbf{x}_1|l_1)}{K}$. Alternatively, it can be determined by the class prior $p(l)$, which is estimated by the frequency of class labels.
- The **transition probability** $a_{ij} = p(l_n = c_j | l_{n-1} = c_i)$ defines the probability to switch from class c_i to c_j and is estimated by the maximum likelihood (ML) approach. That means the expected number of transitions from c_i to c_j is divided by the expected number of times c_i occurs. Two example transition graphs are shown in Fig. 9 and 13.
- The **observation probability** $p(\mathbf{x}_n|l_n)$ expresses the class likelihood that a feature vector is generated by a class.

We use three different models to generate the observation probabilities, which are the following classifiers trained in a supervised learning scheme:

- random forest trains an ensemble of 20 decision trees using randomization by bootstrapping samples for each tree and a random feature selection per binary split,
- multi-layer perceptron iteratively trains a non-linear decision boundary with 100 hidden neurons, and
- Gaussian mixture model fits a mixture model of 8 components per class and a diagonal covariance matrix.

These classifiers decide for the routine class that has the maximal posteriori probability $p(l_n|\mathbf{x}_n)$. Hence, we compare the recognition performance of the sole classifier model against the combination with an HMM. To apply these classifiers as observation models in Equation (1), we need to convert their output to the class likelihood

$$p(\mathbf{x}_n|l_n) = \frac{p(l_n|\mathbf{x}_n)p(\mathbf{x}_n)}{p(l_n)} \quad (2)$$

via the Bayes rule [33]. The evidence term $p(\mathbf{x}_n)$ is a constant and can be ignored in the decoding of the most likely class

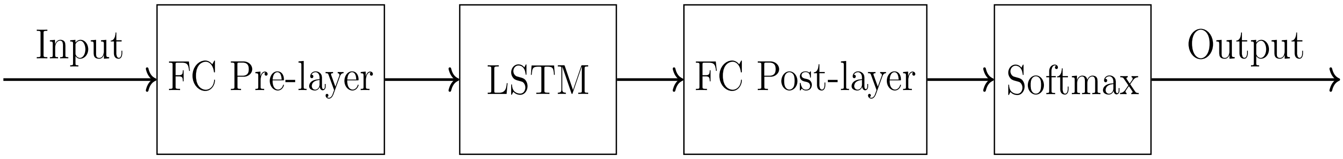


FIGURE 4. The LSTM network consists of a fully-connected (FC) pre-layer and post-layer around the LSTM unit plus a softmax output stage.

sequence, which is given by

$$\arg\max_{l_1,\dots,l_N} p(l_1,\dots,l_N,\mathbf{x}_1,\dots,\mathbf{x}_N). \tag{3}$$

According to [16], the Viterbi algorithm decodes the most likely sequence by setting (2) in (1), taking the logarithm, and ignoring constant factors:

$$\log p(l_1,\dots,l_N,\mathbf{x}_1,\dots,\mathbf{x}_N) \tag{4}$$

$$\propto \sum_{n=2}^{N} \log p(l_n|l_{n-1}) + \left[\sum_{n=1}^{N} \log \frac{p(l_n|\mathbf{x}_n)}{p(l_n)}\right], \tag{5}$$

which can be rewritten in a recursive way:

$$\omega(l_{n+1}) = \log \frac{p(l_{n+1}|\mathbf{x}_{n+1})}{p(l_{n+1})} \tag{6}$$

$$+ \max_{l_n}\left[\log p(l_{n+1}|l_n) + \omega(l_n)\right] \tag{7}$$

$$\text{with initialization } \omega(l_1) = \log \frac{p(l_1|\mathbf{x}_1)}{p(l_1)}. \tag{8}$$

This allows to find the most likely sequence for each time step and the optimal sequence is found by backtracking the gone steps.

LSTM is capable of learning long-term relationships in data sequences [21], [22]. Therefore, we test these capabilities in the routine domain, where activities last for longer periods of minutes to hours. To evaluate these recurrent networks, we need to specify the network architecture and further hyper-parameters such as the number of neurons per layer. Thus, we analyzed the effects of a fully-connected pre-layer and post-layer around the LSTM layer with a softmax output unit, which is illustrated in Fig. 4. Hereby, we varied the number of neurons in a range from 32 to 128 units. After an empirical architecture evaluation with a L2-regularization to decay the weights, an Adam optimizer, and an early-stopping criterion, the LSTM layer started very early to overfit and learned the training data by heart if it contained too many neurons. Thus, we optimized the final architecture and the number of neurons to learn the sequence relationship while keeping it as small as possible to avoid overfitting. The results show that the net with 64, 64, and 32 neurons for pre-layer, LSTM-layer, and post-layer is the best and we use this network for our evaluation.

We performed the experiments in MATLAB R2019b and used the LSTM (from the deep learning toolbox), self-implemented HMM and GMM classifiers with the fitting functionality of Python library scikit-learn 0.22.2 for the MLP network, RF trees, and GMM probability distribution. The hyper-parameters of the classifiers, such as the number of RF trees, are empirically optimized to provide the best classification performance while keeping the computational complexity as low as possible.

C. EVALUATION

For the supervised sequence evaluation, we assess the model performance based on a **cross-validation** (CV) scheme. We are interested in the model generalization abilities of sequence models across multiple subjects. Since the Huynh data set only contains one person, we can only perform a leave-one-day-out scheme, i.e., the personalized model capabilities are evaluated on the seven weekdays. In contrast, in our data set we can assess the model generalization abilities across the seven subjects with a leave-one-person-out scheme. In general, a personalized model has a better performance than a person-independent one [25], which we also showed for the high-level daily routine recognition [3]. To deal with the strong class-imbalance in both data sets, we require recognition metrics that make it obvious if the classifier ignores the minority classes [34]. Thus, we compute

- the confusion matrix containing the four events per class: true positive (TP), true negative (TN), false positive (FP), and false negative (FN),
- the accuracy (A) $\frac{TP+TN}{TP+TN+FP+FN}$, and
- the class-averaged F-measure F_1 as harmonic mean of recall $\frac{TP}{TP+FN}$ and precision $\frac{TP}{TP+FP}$ per class.

The accuracy criterion is affected by the class-imbalance, but the class averaged F_1-measure is independent of the class distribution [25]. Thus, we can judge from the ratio between the two metrics how well a classifier is doing in the overall performance as well as for the minority classes. This means if a classifier has a stronger gap between the A and F_1-measure, the model does not recognize the minority classes well. However, if both metrics are on a par, all classes are equally well detected. We use the confusion matrix for a detailed picture of misclassified routine activities with the best-performing algorithms.

To further consider the temporal order of the predictions and ground truth, we apply the segment error assignments of [35]. It defines a segment as a change in the ground truth or prediction output. Thus, the three following error types are computed and illustrated in Fig 5:

- **Insertion** is a wrong class transition at the start or end of segment or fragments a segment of the same class in parts. For example, before or after a social class

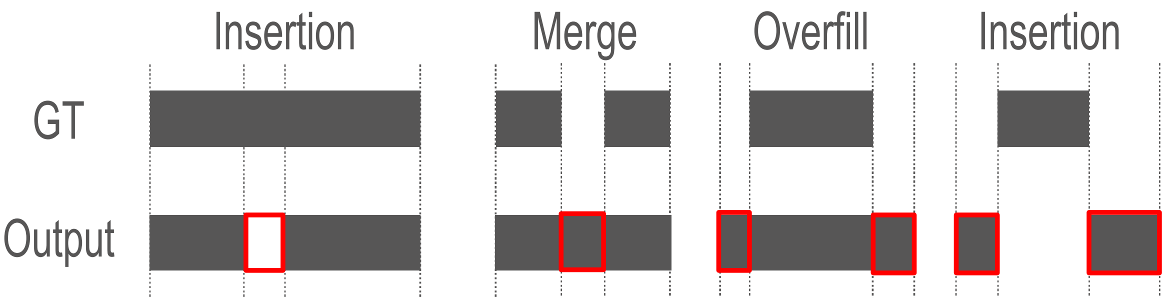


FIGURE 5. The segment error types are shown for a binary classification example with the ground truth (GT) and an output of a classifier. The vertical dashed lines denote the segment boundaries and the red rectangles mark the erroneous segments.

segment, the classifier wrongly predicts a physical segment instead of transportation or within a social segment the predictions change to the listening class. Thus, a high number of insertions means a classifier is not stable in time and often changes its predictions between the classes.

- **Overfill** is a segment that extends a ground truth segment over its boundaries, i.e., a class segment starts too early or ends too late. That is why, a high number of overfills stand for a classifier that changes its predictions too less.
- **Merge** is a special case of an overfill, where between two occurrences of the same class no change to another class happens, e.g., the ground truth has a sequence of social, listening, and social, but the classifier just outputs social.

Since the segments in both data sets have a variable length, we normalize the three error types per sample duration and not per number of segments, which would be misleading, since the variable segment lengths have a different weight. Since we face a multi-class problem, the output is a special kind of confusion matrix, the so-called segment error table. That is why, we simplify the analysis by summing up the error patterns over all classes. Therefore, we do not distinguish if, e.g., an insertion error happened for one class not the other. Thus, the sum of these three segment errors is 1 minus the accuracy value. Based on the three segment errors, we can judge a classifier's tendency to change predictions a lot or be stable over time.

V. RESULTS

The DRR results for the LSTM and HMM sequence learners with three observation models are compared against the performance of classifiers without exploiting the temporal relationship on two data sets, Huynh and our set.

A. HUYNH SET

Starting the leave-one-day-out CV analysis on the Huynh set with acceleration data and the time-of-day feature, we show the personalized classifier evaluation with recognition rates and segment error analysis plus the confusion matrix of the best performing algorithm in Fig. 6, 7, and 8. We first analyze the **performance of the non-sequence classifiers used as the HMM observation models** in Fig. 6 without modeling the sequence relationships, and then check for a possible improvement marked as the red bar by sequence learning approaches. Therefore, the RF classifier is the best non-sequence learning model with an F_1 and accuracy performance of 86.6% and 93.0% compared to GMM and MLP with an accuracy of 87.8% and 87.4%. However, they both have detection problems with the minority classes resulting in a lower F_1 rate of 71.8% and 75.4%, since the minority classes strongly overlap within the ACC space. The MLP has a higher capability to model a complex decision boundary than the GMM, which explains the better minority class detection. In this case, further features such as audio could ease the detection problem. The reason for the big performance gap between the classifiers is, that the decision trees of the RF can effectively profit from the time-of-day feature. The decision trees can derive rules like from 12 a.m. to 1 p.m. it is lunch, because the Huynh set has a very structured daily routine.

After **adding the HMM sequence learner** to the three classifiers, all metrics improve shown by the red bar in Fig. 6. The GMM-HMM classifier has the smallest F_1 and A increase of 1.2% and 1.1%, because it has the most problems with the class overlap in the feature space. Then, the RF-HMM combination follows and strongly improves by 5% and 2.8%. The MLP-HMM classifier nearly doubles both metrics by 10.2% and 5%. However, the MLP model starts from a lower level of correct decisions than the RF. Thus, adding the HMM gives more possibilities to smooth out erroneous class transitions. In an overall comparison, the RF-HMM combination is the best classifier even compared to the LSTM, which has an F_1 and A margin of 13.9% and 9.4% compared to the best. To enhance the LSTM, the amount of training data needs to be increased for a better model generalization. The ranking stays the same as without the sequence modeling since the class overlap too much within acceleration feature space and RF is the only classifier that can effectively deal with the time-of-day feature.

In comparison to the **prior work of Huynh** [8], we outperform both of his methods, GMM-HMM and TM, with an F_1 rate of 64.6% and 74.3%, i.e., our GMM-HMM and

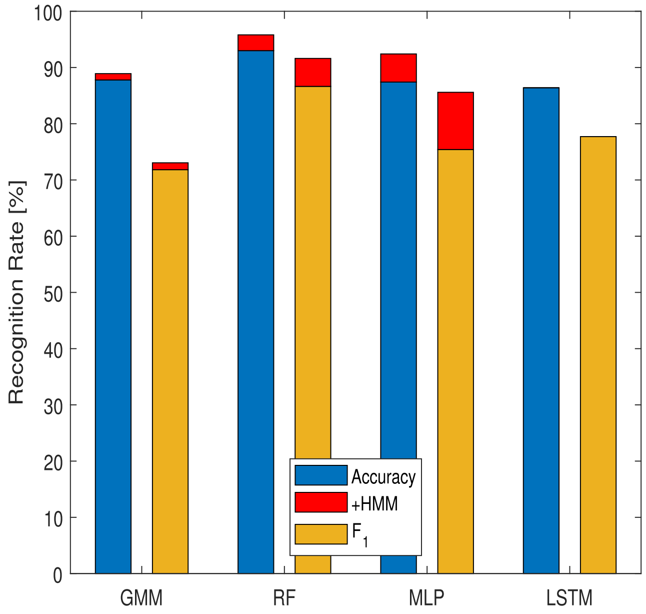


FIGURE 6. Classifier evaluation on Huynh set. The red bar marks the sequence learning improvement of the added HMM to the non-sequence classifiers, such as RF.

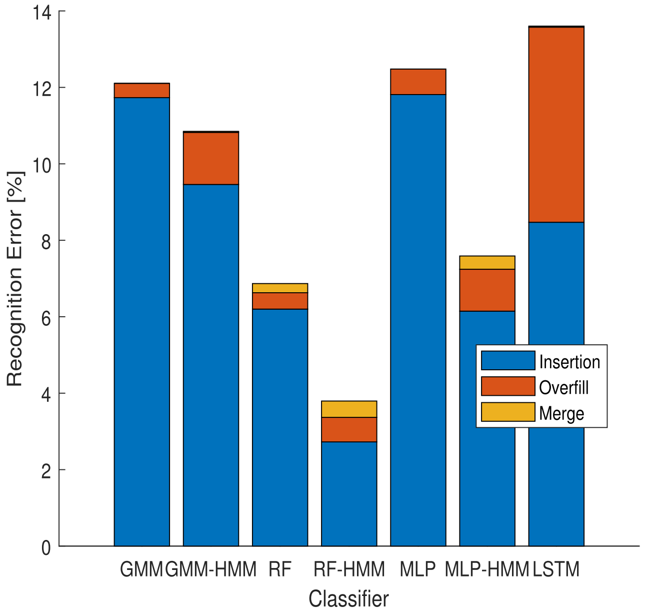
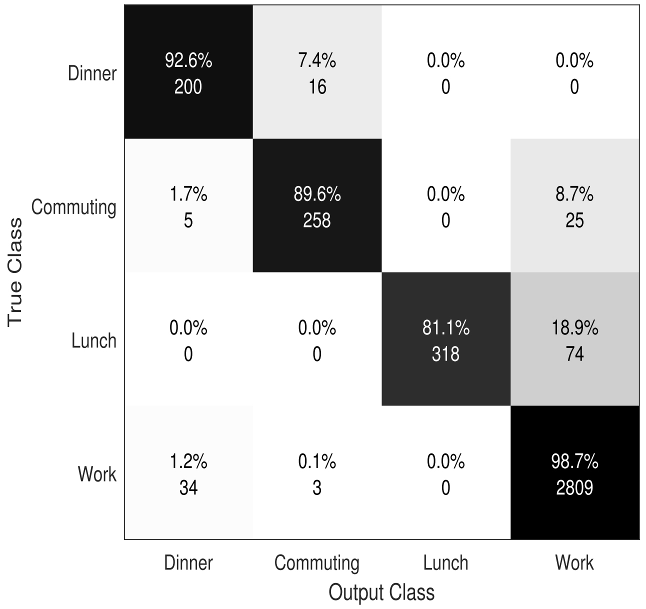


FIGURE 7. Classifier segment error evaluation on Huynh set.

RF-HMM have an F_1 rate of 73.0% and 85.6%. The high margin is a result of our well-performing high-level feature representation and the appropriate window length of 1 minute, where Huynh used a length of 30 minutes. Additionally, the superior observation model RF has a strong contribution to the improved HMM performance.

Analyzing the **time behavior of predictions**, the results of the segment evaluation are depicted in Fig. 7. The main source of errors are insertions with the highest number of cases for the non-sequence classifiers, GMM and MLP, with 11.7% and 11.8%. Adding the HMM highly decreases the insertion percentage by 2.2% and 5.7% while slightly increasing the overfill error by 1% and 0.4%. Thus, the stability of classifier predictions is improved as expected, which increases the number of overfill events. This is also the case for the RF, which has a similarly low quantity of insertions like the MLP-HMM and even lower for the RF-HMM with the best result of 9.5%. The reason is that the decision trees of the RF can efficiently deal with the time-of-day feature, which has a very high predictive power for the Huynh set due to the structured daily work routine. Merge errors only occur for the RF, RF-HMM, and MLP-RF with a small percentage of 2.7% to 6.2%, because of the mostly long class duration, which makes it difficult to merge segments. The LSTM has the worst performance with the biggest overfill error of 5.1% and a medium insertion error of 8.5%, since it has a strong tendency to stay with its class predictions over a longer period and changes them too less. Therefore, the sequence learning approaches improve as expected the temporal stability of the predictions. In some cases, the stability is too strong for the LSTM algorithm.

Furthermore, we analyze in detail the **confusion matrix** of the best-performing RF-HMM, where the class-wise recall is shown in the rows in Fig. 8. The majority class, work, is particularly well detected with a high recall of 98.7%. Only a few errors occur due to other situations, which also consist of seated activities, e.g., lunch or dinner, since the activity patterns of the ACC data are very similar. Here, different sensors could be beneficial to distinguish these kinds of situations. Some confusions, such as between lunch and commuting or dinner and lunch, do not happen, even though they could have similar activity patterns. This is the case, since Huynh's working routine is very structured and the classes contain an implicit time order: commute, work, lunch,



	Dinner	Commuting	Lunch	Work
Dinner	92.6% 200	7.4% 16	0.0% 0	0.0% 0
Commuting	1.7% 5	89.6% 258	0.0% 0	8.7% 25
Lunch	0.0% 0	0.0% 0	81.1% 318	18.9% 74
Work	1.2% 34	0.1% 3	0.0% 0	98.7% 2809

FIGURE 8. Confusion matrix of the best-performing RF-HMM sequence learner on Huynh set.

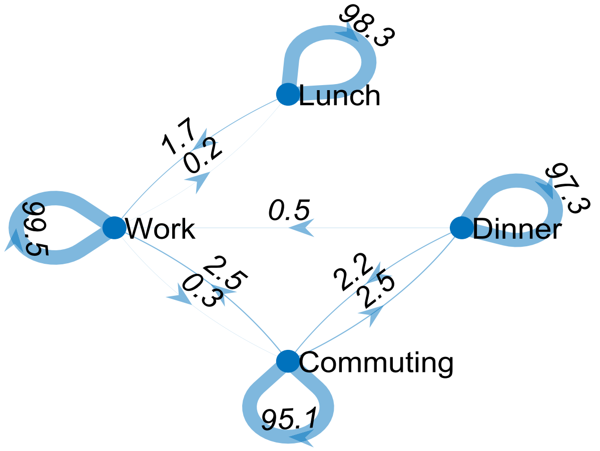


FIGURE 9. Transition graph of the first cross-validation fold on Huynh set [%].

work, commute to dinner. This knowledge is also present in the transition matrix of the HMM, since some transitions, e.g., lunch to dinner, do not happen. To show this switching behavior between the classes, we plot the **transition graph** of the first cross-validation fold in Fig. 9, where the weight of a transition parameter a_{ij} is displayed as the thickness of an arrow. No transition event between two classes corresponds to no arrow in the graph, e.g., between lunch and commuting. Obviously, the strongest transition is staying in the same class, e.g., 97.3% for dinner, i.e., the probability mass is strongly concentrated on the diagonal elements of the transition matrix due to the long duration of routine events. Thus, the duration density for each class is uniformly distributed with a small spread. Therefore, it is not optimal for an HMM, which models the transitions to decay exponentially [16]. The only exception occurs for commuting that has a peak at the typical period.

B. OUR DATA SET

On the contrary to the Huynh set, we perform a leave-one-person-out CV and test the person-independent model generalization across subjects on the acceleration and audio data. The results of the classifier evaluation are depicted in Fig. 10 and 11 as well as the confusion matrix of the best performing algorithm in Fig. 12.

Again, we start the analysis with **the performance of the non-sequence classifiers** in Fig. 10 that are used for the observation models. Afterwards, the possible gain of sequence modeling is assessed. The RF classifier is the best non-sequence learning model with an F_1 and accuracy performance of 79.4% and 83.9%. These rates are slightly superior to MLP with 78.6% and 83.2%. The GMM lies within a margin of 6 to 7% in both metrics. In comparison to the Huynh results, we see a lower overall performance due to the person-independent training and the more complex problem. The minority class recognition works relatively better because of the smaller difference between the accuracy and F_1 metrics. Here, the rich audio features are beneficial to distinguish the routine classes.

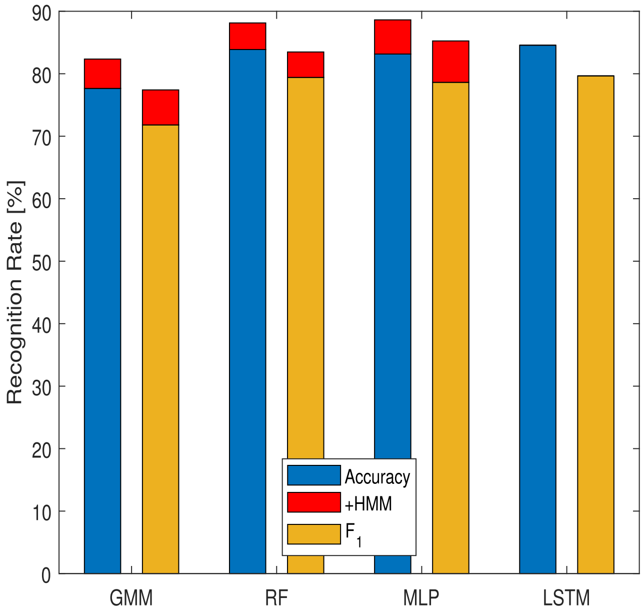


FIGURE 10. Classifier evaluation on our data set. The red bar marks the sequence learning improvement of the added HMM to the non-sequence classifiers, such as RF.

To assess the **gain of sequence modeling**, we use the three classifiers as observation models for the HMM and all metrics strongly improve about 4 to 7%. Thus, the best non-sequence classifier, RF, enhances the rates about 4% by including the HMM, whereas the GMM-HMM gains an upgrade of almost 5%. The MLP-HMM has the strongest F_1 and A improvement of 6.7% and 5.4%. Thus, the MLP-HMM combination is the winner even against the LSTM, which has an F_1 and A margin of 3.8% and 3.5% compared to the best. To enhance the LSTM performance, the amount of training data needs to be increased for a better model generalization.

Analyzing the **time behavior of predictions**, the results of the segment evaluation are depicted in Fig. 11. The main source of errors are insertions with the highest number of cases for the non-sequence GMM classifier with 20.9%. Adding the HMM to the GMM, it highly decreases the insertion percentage by 6.8% while slightly increasing the overfill error by 1.7%. Thus, the GMM-HMM has a similar level of insertion errors like the RF and MLP of about 14-15%. The RF and MLP including the HMM strongly decrease the insertions by 9.1% and 9.5% while enhancing the number of overfills by about 4%. Thus, the stability of classifier predictions is improved as expected, which increases the number of overfill events. The best overall performance is achieved by the MLP-HMM. Merge errors only occur for the RF and the sequence learners with a small percentage of 0.2% to 2.2%. The LSTM has a medium overfill performance of 4.2% and a medium insertion error of 10.5%, since it has a strong tendency to stay with its class predictions over a longer period and rarely changes.

Furthermore, we analyze in detail the **confusion matrix** of the best-performing MLP-HMM in Fig. 12. Obviously, the two majority classes, social and basics, and transportation

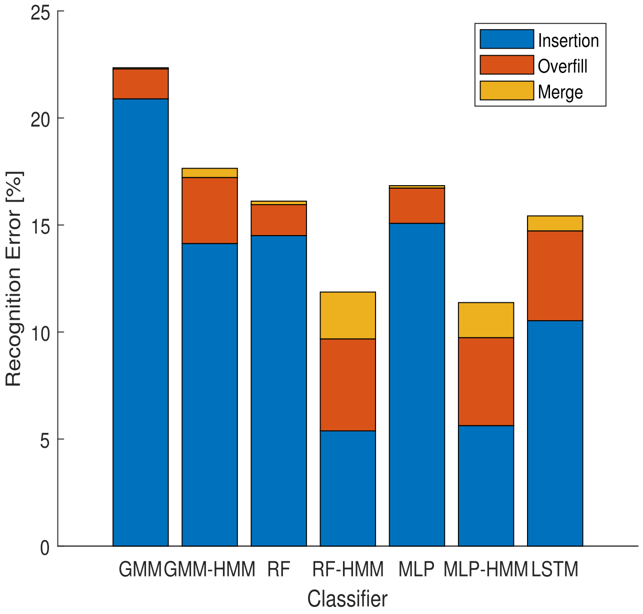
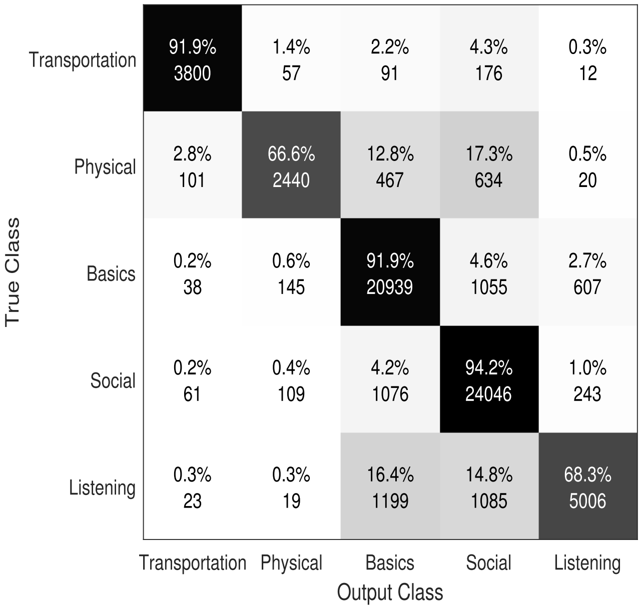


FIGURE 11. Classifier segment evaluation on our data set.



	Transportation	Physical	Basics	Social	Listening
Transportation	91.9% 3800	1.4% 57	2.2% 91	4.3% 176	0.3% 12
Physical	2.8% 101	66.6% 2440	12.8% 467	17.3% 634	0.5% 20
Basics	0.2% 38	0.6% 145	91.9% 20939	4.6% 1055	2.7% 607
Social	0.2% 61	0.4% 109	4.2% 1076	94.2% 24046	1.0% 243
Listening	0.3% 23	0.3% 19	16.4% 1199	14.8% 1085	68.3% 5006

FIGURE 12. Confusion matrix of the best-performing MLP-HMM sequence learner on our data set.

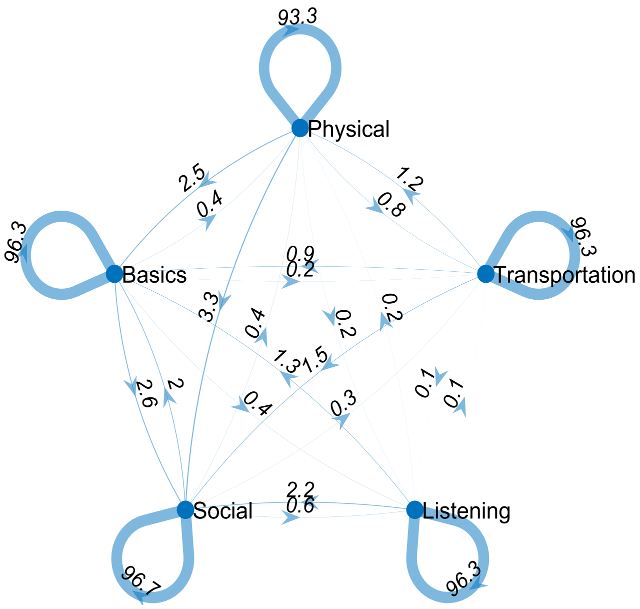


FIGURE 13. Transition graph of the first cross-validation fold on our set [%].

are very well recognized with a recall over 90% and contribute to the high overall accuracy of 88.1%. They are mainly distinguishable through audio characteristics, such as low-frequency car noise or own voice activation. In contrast, the strong confusion between listening and basics (16.4%) or social (14.8%) stems from the high similarity within the audio features and the strong dependency of the reference class on the subjective user intention. This means a background conversation can be either listening or basic depending if the subject wants to follow it. Additionally, it can quickly change to social if the subject decides to participate in the conversation. Thus, we have many transitions between these classes and, in general, all routine transitions are possible, which is different to the Huynh set and makes the problem more complicated. To show this switching behavior between the classes, we plot the **transition graph** of the first cross-validation fold in Fig. 13. Obviously, the strongest transition is staying in the same class, e.g., for social 96.7%, i.e., the main portion of the probability mass lies on the diagonal elements of the transition matrix. However, on some days, not all transitions happen since an activity class is not performed every day. The HMM inherently models a duration probability density that is exponentially decaying [16], which is a good fit to our data set. This is, because many events have a short duration of a few minutes and a small number of events have a long duration of hours. Similarly, during sport activities, we have a high intensity in the ACC signal and a voice activation, which leads to the bigger mismatch of 17.3% between physical and social. Thus, both classes can also occur simultaneously and then the user's intention decides. Here, the situational intention needs to be better decoded from suitable motion patterns or further sensors [36], e.g., electromyograms for listening attention [37], that can deliver a more reliable input.

To summarize the **findings on sequence learning** for daily routine recognition, it strongly improves the classification performance of all tested non-sequence learners, RF, MLP, and GMM, by adding the HMM to them. For the GMM classifier, the enhancement is less strong than for the others. The RF and MLP outperform the LSTM model. Thus, the sequence learning is particularly beneficial for the RF and MLP classifier as the observation model for the HMM. In addition, the thorough sequence classifier evaluation shows the expected improvement in the temporal prediction stability.

VI. CONCLUSION AND OUTLOOK

In this article, we improved the daily routine recognition (DRR) on two real-world data sets using sequence learning methods. Our large and realistic data set consists of seven non-representative hearing aid (HA) wearers, recording their unconstrained real-life for a total length of 63449 minutes. Thereby, the acceleration and audio features are designed to represent the routine characteristics well. In contrast, the Huynh set contains his real-life of seven working days with two acceleration sensors. On this basis, we perform a comprehensive sequence classifier evaluation. We demonstrated that the multi-layer perceptron (MLP) and random forest as an observation model for the hidden Markov model (HMM) achieved the best F-measure performance of 85.3% and 91.6% on our set and the Huynh set. Thereby, the MLP has the strongest F-measure improvement of 6.7% and 10.2% on both sets by adding the HMM sequence learner. The long short-term memory network has an F-measure of 79.7% and 77.7% on both sets. The segment error analysis discovers for sequence learners the strong improvement of the temporal prediction stability. On our set, the remaining confusion for classifiers mainly stems from the intention-based class decision of the HA users. Thus, we showed the improved classification performance of sequence learners.

For future work, we can investigate a tailored motion representation or apply further sensors to distinguish the intended behavior more precisely [13], [38]. Further algorithmic improvements can be achieved to make the HMM backtracking real-time applicable without the need to store the sequences in memory for an optimal decoding of the most likely path. Additionally, a data set of representative HA users can be evaluated to determine the DRR performance of these elderly wearers. We can investigate, if they perform other activities, have a weaker motion patterns, or the routine changes over a longer period due to a concept drift [39]. Nevertheless, it is expected, that these elderly users rely on more recurring set of activities and environments [23]. Thus, the daily routine recognition should be simplified.

ACKNOWLEDGMENT

The authors would like to thank Sivantos GmbH for funding the work of Thomas Kuebert and support, which made this contribution possible. Their special thanks go to the study participants for their dedication during the long recordings.

REFERENCES

[1] B. Tessendorf, A. Bulling, D. Roggen, T. Stiefmeier, M. Feilner, P. Derleth, and G. Tröster, “Recognition of hearing needs from body and eye movements to improve hearing instruments,” in *Proc. Int. Conf. Pervas. Comput.* Berlin, Germany: Springer, 2011, pp. 314–331.

[2] M. Büchler, S. Allegro, S. Launer, and N. Dillier, “Sound classification in hearing aids inspired by auditory scene analysis,” *EURASIP J. Adv. Signal Process.*, vol. 2005, no. 18, pp. 1–12, Dec. 2005.

[3] T. Kuebert, H. Puder, and H. Koeppl, “Daily routine recognition for hearing aid personalization,” in *SN Computer Science*. Berlin, Germany: Springer, Mar. 2021.

[4] N. S. Jensen, L. W. Balling, and J. B. Nielsen, “Effects of personalizing hearing-aid parameter settings using a real-time machine-learning approach,” in *Proc. 23rd Int. Congr. Acoust.*, Aachen, Germany, Sep. 2019, pp. 1–8.

[5] J. Seiter, O. Amft, M. Rossi, and G. Tröster, “Discovery of activity composites using Topic models: An analysis of unsupervised methods,” *Pervas. Mobile Comput.*, vol. 15, pp. 215–227, Dec. 2014.

[6] R. J. White, “Using topic models to detect behaviour patterns for healthcare monitoring,” Ph.D. dissertation, School Syst. Eng., Univ. Reading, Reading, U.K., 2018.

[7] M. Stikic, D. Larlus, S. Ebert, and B. Schiele, “Weakly supervised recognition of daily life activities with wearable sensors,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 12, pp. 2521–2537, Dec. 2011.

[8] T. Huynh, M. Fritz, and B. Schiele, “Discovery of activity patterns using topic models,” in *Proc. 10th Int. Conf. Ubiquitous Comput.*, vol. 8, 2008, pp. 10–19.

[9] T. Kuebert, H. Puder, and H. Koeppl, “Daily routine recognition with visual interactive labeling by fusing acceleration and audio signals,” in *Proc. IEEE Int. Symp. Signal Process. Inf. Technol. (ISSPIT)*, Dec. 2019, pp. 1–6.

[10] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M. D. Plumbley, “Detection and classification of acoustic scenes and events,” *IEEE Trans. Multimedia*, vol. 17, no. 10, pp. 1733–1746, Oct. 2015.

[11] J. Schroeder, S. Wabnik, W. J. P. van Hengel, and S. Goetze, “Detection and classification of acoustic events for in-home care,” in *Ambient Assisted Living*. Berlin, Germany: Springer, 2011, pp. 181–195.

[12] A. Temko, C. Nadeu, D. Macho, R. Malkin, C. Zieger, and M. Omologo, “Acoustic event detection and classification,” in *Computers in the Human Interaction*. Springer, 2009, pp. 61–73.

[13] B. Fu, N. Damer, F. Kirchbuchner, and A. Kuijper, “Sensing technology for human activity recognition: A comprehensive survey,” *IEEE Access*, vol. 8, pp. 83791–83820, 2020.

[14] J. Parkka, M. Ermes, P. Korpipaa, J. Mantyjarvi, J. Peltola, and I. Korhonen, “Activity classification using realistic data from wearable sensors,” *IEEE Trans. Inf. Technol. Biomed.*, vol. 10, no. 1, pp. 119–128, Jan. 2006.

[15] G. T. Dietterich, “Machine learning for sequential data: A review,” in *Proc. Joint IAPR Int. Workshops Stat. Techn. Pattern Recognit. (SPR) Struct. Syntactic Pattern Recognit. (SSPR)*. Berlin, Germany: Springer, 2002, pp. 15–30.

[16] M. C. Bishop, *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer, 2006.

[17] S. Mota and R. W. Picard, “Automated posture analysis for detecting learner-s interest level,” in *Proc. Conf. Comput. Vis. Pattern Recognit. Workshop*, Jun. 2003, p. 49.

[18] J. A. Ward, P. Lukowicz, G. Tröster, and E. T. Starner, “Activity recognition of assembly tasks using body-worn microphones and accelerometers,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 10, pp. 1553–1567, Oct. 2006.

[19] M. R. Schädler and B. Kollmeier, “Separable spectro-temporal Gabor filter bank features: Reducing the complexity of robust features for automatic speech recognition,” *J. Acoust. Soc. Amer.*, vol. 137, no. 4, pp. 2047–2059, 2015.

[20] E. De-La-Hoz-Franco, P. Ariza-Colpas, J. M. Quero, and M. Espinilla, “Sensor-based datasets for human activity recognition—A systematic review of literature,” *IEEE Access*, vol. 6, pp. 59192–59210, 2018.

[21] J. Wang, Y. Chen, S. Hao, X. Peng, and L. Hu, “Deep learning for sensor-based activity recognition: A survey,” *Pattern Recognit. Lett.*, vol. 119, pp. 3–11, Mar. 2019.

[22] Y. N. Hammerla, S. Halloran, and T. Plötz, “Deep, convolutional, and recurrent models for human activity recognition using wearables,” in *Proc. 25th Int. Joint Conf. Artif. Intell.*, 2016, pp. 1–7.

[23] Y.-H. Wu and R. A. Bentler, “Do older adults have social lifestyles that place fewer demands on hearing?” *J. Amer. Acad. Audiol.*, vol. 23, no. 9, pp. 697–711, 2012.

[24] A. Zinnen, U. Blanke, and B. Schiele, “An analysis of sensor-oriented vs. model-based activity recognition,” in *Proc. Int. Symp. Wearable Comput.*, Sep. 2009, pp. 93–100.

[25] A. Bulling, U. Blanke, and B. Schiele, “A tutorial on human activity recognition using body-worn inertial sensors,” *ACM Comput. Surv.*, vol. 46, no. 3, pp. 33:1–33:33, Jan. 2014.

[26] J. Hale, J. A. Ward, F. Buccheri, D. Oliver, and A. F. de C Hamilton, “Are you on my wavelength? Interpersonal coordination in naturalistic conversations,” *J. Nonverbal Behav.*, vol. 44, no. 1, pp. 1–28, 2018.

[27] S. Hemminki, P. Nurmi, and S. Tarkoma, ''Accelerometer-based transportation mode detection on smartphones,'' in *Proc. 11th ACM Conf. Embedded Netw. Sensor Syst.*, 2013, pp. 1–14.

[28] O. Banos, J.-M. Galvez, M. Damas, H. Pomares, and I. Rojas, ''Window size impact in human activity recognition,'' *Sensors*, vol. 14, no. 4, pp. 6474–6499, Apr. 2014.

[29] G. Peeters, ''A large set of audio features for sound description (similarity and classification) in the CUIDADO project, version 1.0,'' Inst. Res. Coordination Acoust./Music, Paris, France, Tech. Rep., 2004.

[30] T. Powers, M. Froehlich, E. Branda, and J. Weber, ''Clinical study shows significant benefit of own voice processing,'' Hearing Rev., Leawood, KS, USA, Tech. Rep., 2018, vol. 25, no. 2.

[31] M. F. McKinney and J. Breebaart, ''Features for audio and music classification,'' in *Proc. 3rd Int. Symp. Music Inf. Retr.* Paris, France: IRCAM, 2003, pp. 151–158.

[32] J. Tang, S. Alelyani, and H. Liu, ''Feature selection for classification: A review,'' *Data Classification: Algorithms and Applications*. Boca Raton, FL, USA: CRC Press, 2014.

[33] C. M. Bishop, M. Svensen, and G. E. Hinton, ''Distinguishing text from graphics in on-line handwritten ink,'' in *Proc. 9th Int. Workshop Frontiers Handwriting Recognit. (IWFHR)*, Oct. 2004, pp. 142–147.

[34] G. Forman and M. Scholz, ''Apples-to-apples in cross-validation studies: Pitfalls in classifier performance measurement,'' *ACM SIGKDD Explorations Newslett.*, vol. 12, no. 1, pp. 49–57, Nov. 2010.

[35] J. A. Ward, P. Lukowicz, and H. W. Gellersen, ''Performance metrics for activity recognition,'' *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 1, pp. 6:1–6:23, Jan. 2011.

[36] R. C. King, E. Villeneuve, R. J. White, R. S. Sherratt, W. Holderbaum, and W. S. Harwin, ''Application of data fusion techniques and technologies for wearable health monitoring,'' *Med. Eng. Phys.*, vol. 42, pp. 1–12, Apr. 2017.

[37] D. J. Strauss, F. I. Corona-Strauss, A. Schroeer, P. Flotho, R. Hannemann, and S. A. Hackley, ''Vestigial auriculomotor activity indicates the direction of auditory attention in humans,'' *eLife*, vol. 9, Jul. 2020, Art. no. e54536.

[38] Y.-L. Zheng, X.-R. Ding, C. C. Y. Poon, B. P. L. Lo, H. Zhang, X.-L. Zhou, G.-Z. Yang, N. Zhao, and Y.-T. Zhang, ''Unobtrusive sensing and wearable devices for health informatics,'' *IEEE Trans. Biomed. Eng.*, vol. 61, no. 5, pp. 1538–1554, May 2014.

[39] G. Ditzler, M. Roveri, C. Alippi, and R. Polikar, ''Learning in nonstationary environments: A survey,'' *IEEE Comput. Intell. Mag.*, vol. 10, no. 4, pp. 12–25, Nov. 2015.

THOMAS KUEBERT received the M.Sc. degree in electrical engineering from Technische Universität Darmstadt, Germany, in 2016, where he is currently pursuing the Ph.D. degree in electrical engineering with a focus on machine learning techniques for hearing aid personalization. His research interests include data visualization, applying unsupervised, semi-supervised, supervised, and sequence methods.

HENNING PUDER received the joint Diploma degree in electrical engineering from the Technische Universität Darmstadt, Germany, and the Grande École Supérieure d'Électricité (Supélec), Paris, France, in 1997, and the Ph.D. degree in electrical engineering from the Technische Universität Darmstadt, Germany, in 2003. Since 2002, he has been with the Sivantos GmbH, former Siemens Audiologische Technik GmbH, Erlangen, Germany. In 2006, he was appointed the Head of the Signal Processing Group within Sivantos Research and Development Department. Since 2011, he has been a part-time Professor, and since 2018, a honorary Professor with Technische Universität Darmstadt, while still holding his position at Sivantos. At Darmstadt, he acts as the Head of the Adaptive Systems for Processing of Speech and Audio Signals Research Group. His research interest includes focuses on audio signal processing applications for hearing aids, such as noise reduction, beamforming, and feedback cancellation. His Ph.D. thesis dealing with the topic of Freisprecheinrichtungen für Kraftfahrzeuge (Hands-free equipment for cars) was honoured with the German ITG Johann Philipp Reis Award, in 2003, shared with Peter Jax from RWTH Aachen. Since 2002, he has been with the Sivantos GmbH, former Siemens Audiologische Technik GmbH, Erlangen, Germany.

HEINZ KOEPPL received the M.Sc. degree in physics from Karl-Franzens University Graz, in 2001, and the Ph.D. degree in electrical engineering from the Graz University of Technology, Austria, in 2004. After that he held postdoctoral positions with UC Berkeley and Ecole Polytechnique Federale de Lausanne (EPFL). From 2010 to 2014, he was an Assistant Professor with the ETH Zurich and a part-time Group Leader at IBM Research Zurich, Switzerland. Since 2014, he has been a Full Professor with the Department of Electrical Engineering and Information Technology, Technische Universität Darmstadt, Germany. He received two awards for the Ph.D. thesis, the Erwin Schrödinger Fellowship, the SNSF Professorship Award, and currently holds an ERC consolidator grant. His research interest includes stochastic models and their inference in applications ranging from communication networks to cell biology.