IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received May 11, 2021, accepted June 27, 2021, date of publication July 2, 2021, date of current version July 14, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3094175

Machine-Learning-Based Closed-Set Text-Independent Speaker Identification Using Speech Recorded During 25 Hours of Prolonged Wakefulness

YOUNGSUN KONG[1], (Member, IEEE), HUGO F. POSADA-QUINTERO[1], (Member, IEEE), MATTHEW S. DALEY[2], JEFFREY BOLKHOVSKY[2], AND KI H. CHON[1], (Senior Member, IEEE)

[1]Department of Biomedical Engineering, University of Connecticut, Storrs, CT 06269, USA
[2]Naval Submarine Medical Research Laboratory, Groton, CT 06349, USA

Corresponding author: Ki H. Chon (ki.chon@uconn.edu)

This work was supported in part by the Grant F 1103, and in part by the Defense Health Program.

ABSTRACT We performed machine learning for text-independent speaker identification using speech recorded during the day, evening, and night, from subjects undergoing 25 hours of prolonged wakefulness. Subjects answered casual questions lasting approximately 3 minutes and described pictures presented to them for 0.5 minutes. We extracted 12,515 vocal features using OpenSmile software. For generalization of the training scheme, we segmented the 20 subjects into training and testing sets (10 subjects for each) and repeated testing four times with different subsets. Specifically, we used one set of 10 subjects to find the best feature-sets and the optimal machine-learning method, and the other set of 10 subjects was used to test the trained model. With trained machine-learning models using three speech sessions recorded throughout the day for speaker identification, we obtained 95% and 98.8% for balanced accuracies for daytime and evening speech, respectively, but 84.2% for nighttime-testing speech. With training data from all times of day—daytime, evening, and nighttime—we obtained 97.5%, 98.8%, and 98.1% for balanced accuracies for test data from daytime, evening, and nighttime speech, respectively; the overall accuracy was 98.1%. Prolonged wakefulness deteriorates the performance of machine-learning based speaker identification. This work suggests that machine-learning based speaker identification should be trained using speech data from both daytime and nighttime speech sessions for better overall accuracy. Machine learning can potentially be used for identifying a speaker's voice even when it is affected by tiredness and fatigue which are frequently encountered in scenarios such as the emergency rooms and long-duration repetitive task operations.

INDEX TERMS Prolonged wakefulness, sleep deprivation, speaker identification, speaker recognition, machine learning.

I. INTRODUCTION

Speaker identification is relevant for applications such as military operations, forensic speaker recognition, and phone customer service, among others [1], [2]. For these applications, speaker identification must be independent of the text being spoken, and there can be no reliance on emotional or situational context. This makes speech identification challenging, because external factors like stress, emotions, and fatigue can affect human speech [3]–[5]. [6], [7]. Some studies have already reported accuracies of 90% or more [8]–[12], however, no speaker identification models have been developed and tested using speech recorded while subjects underwent sleep deprivation.

Speaker identification is a category of speech recognition, defined as a method that identifies a speaker amongst a set of speakers [13]. The set of speakers can be either open- or closed-set depending on the purpose of the system.

The associate editor coordinating the review of this manuscript and approving it for publication was Kin Fong Lei.

Closed-set systems have a fixed number of speakers, while open-set systems can have a greater number of speakers than registered (or trained) speakers [14]. Speaker identification can also be either text-dependent or text-independent. Unlike text-dependent systems, text-independent speaker identification does not require speakers to state specific phrases or words, therefore, it can be used for a wider set of applications such as military operations, criminal investigations (e.g., forensic speaker identification), and phone customer service [1], [2]. Unfortunately, given its vast practical applications, text-independent recognition is known to be more challenging than text-dependent [15]. In this work, we developed a closed-set text-independent speaker identification method.

Machine learning is a set of statistical approaches which perform regression, clustering, or classification of unseen data based on a training dataset [16]. As speaker identification is a classification problem, various machine learning methods have been applied to vocal features [8]–[12]. A study involving 30 subjects used both support vector machine (SVM) and multi-layer perceptron (MLP) with a 15th-order linear predictive analysis, and obtained accuracy values of 91.4 and 90.8%, respectively [9]. Mamyrbayev *et al.* performed five different machine-learning approaches on data collected from 20 subjects, including extra-tree, K-nearest-neighbors, and MLP with Mel-frequency cepstral coefficient features, and obtained up to 90% accuracy [10]. Chauhan *et al.* tested various machine-learning methods including MLP, SVM with dynamic time warping, linear predictive coding, and Mel-frequency cepstral coefficients features, and their highest accuracy was 93.1% [8]. Akhsanta and Suyanto developed a text-independent speaker identification system using principal component analysis and SVM with Mel-frequency cepstral coefficients (MFCC) [11]. They obtained 88.97% and 70.93% accuracies for the low- and high-noise levels from 19 subjects, respectively. Chauhan *et al.* trained a feedforward artificial neural network (ANN) and SVM models for text-independent speaker recognition with MFCC and linear predictive coding, and perceptual linear prediction features, and obtained 100% of accuracies for both SVM and ANN classifiers from 23 subjects [12]. In general, the machine-learning approaches have shown robust results for text-independent speaker identification.

However, speaker identification can fail when confronted with confounding external factors such as stress and emotions [3]–[5]. Wu *et al.* improved the performance of their speaker-identification model by 22% with the addition of 14 types of emotional information, including anger and sadness [4]. Raja and Dandapat showed that stressed conditions (e.g., anger, the Lombard effect with noisy backgrounds, and answering difficult questions) caused poorer performance of text-independent speaker identification methods than resulted from non-stressed conditions [5]. Their method showed more than 80% performance accuracy with neutral conditions but less than 61% accuracy for stressed conditions involving 32 speakers. Speech is also known to be affected by sleep deprivation, which impairs fluency, intonation, and pitch [6], [7]. Moreover, changes in vocal parameters (severity, roughness, breathiness, strain, pitch, and loudness) have been observed throughout a day [17], [18]. Testing of speaker identification methods with consideration of voice changes across a day and with sleep deprivation has not yet been performed.

Thus, in this work, we recorded text-independent speech from 20 subjects during 25-hour sleep deprivation and performed various machine-learning methods to compare the identification results for daytime, evening, and nighttime speech. We hypothesized that diversifying the training speech data by including data from different times of the day would lead to better performance in speaker identification during 25 hours of sleep deprivation.

II. METHODS

A. STUDY PROTOCOL

A total of 20 healthy participants were recruited, between 19 and 32 years old (13 males and 7 females). Consent forms were collected on the days of experiments, and screening questionnaires were used for examining the medical background of each volunteer. Volunteers were asked to keep consistent sleep schedules for a week before their experimental day, and were told to avoid food or drink containing stimulants such as caffeine for two days prior to their experimental day. The experimenters were present throughout the experiment to ensure volunteers were awake for the entire 25 hours and avoided stimulating food and drink. Volunteers were asked to wake up at 6:00 am and arrive at the location within 4 hours. The speech protocol consisted of two text-independent speech samples, which included 1) answering four random questions per session from a library of 54 questions, and 2) describing a provided random picture per session from a pool of 13 photos. Speech samples were recorded using a built-in microphone from Tobii Pro Glasses 2 wearable eye tracker [19]. The protocol took approximately 2-3 minutes for answering four questions and half a minute for describing a picture. The protocol was conducted every two hours for a total of 13 sessions during the 25 hours of sleep deprivation. The Institutional Review Board of the University of Connecticut approved the study protocol.

B. DATA COLLECTION AND VOCAL FEATURE EXTRACTION

We extracted voice recordings from video files (including audio and eye-tracking video) with MP4 format obtained using Tobii Pro Glasses 2. We then segmented the voice recordings of each subject answering the four random questions into 30-second windows with 20 seconds of overlap. The sampling frequency of the audio files was 24,000 Hz. We extracted a total of 12,515 non-duplicated vocal features per speech using OpenSmile software with its built-in feature sets consisting of: the Interspeech 2009 emotion challenge (IS09), the Interspeech 2010 Paralinguistic challenge (IS2010), the Interspeech 2011 Speaker State (IS2011), the Interspeech 2012 Speaker Trait Challenge (IS2012), and the Interspeech 2013 ComParE (IS2013) [20].

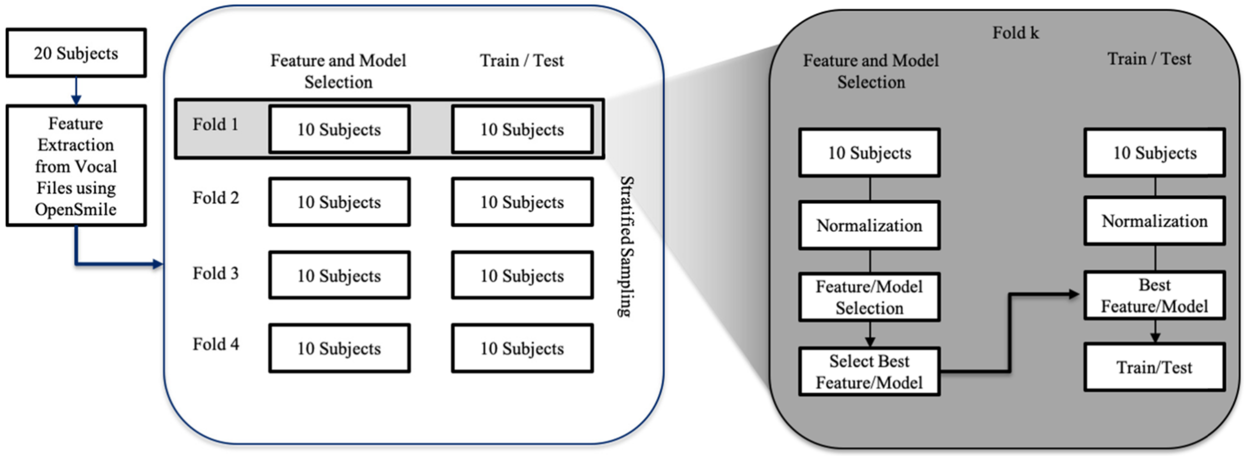

FIGURE 1. Flow chart of our training and testing scheme for machine learning.

C. MACHINE LEARNING

From the vocal features extracted using OpenSmile, we next performed a two-step approach for the closed-set text-independent speaker identification. First, we performed machine learning on 10 subjects with three different feature-selection methods and six classifiers. However, we noticed that these feature selection methods could be biased due to the large amount of features (more than 10,000). Thus, the second step was to test the best feature sets and the classifiers derived from the training data sets from the first ten subjects and then use them to test on the remaining 10 subjects, as shown in Figure 1. This strategy was repeated four times (a total of 4 folds), using the stratified sampling method to avoid training bias of subjects. The four questions and the picture-describing speech were used for both testing and validation.

We compared three different decision-tree-based feature-selection methods: gradient boosting, decision tree, and random forests, in terms of their numerical feature importance values [21]. Due to the imbalanced dataset caused by data loss, class weights (determined based on the inverse of the number of samples per class) were applied to the feature-selection methods. In addition, three feature-selection methods needed to be set with the maximum depth of the trees for regularization. To obtain the best maximum depth, we performed the grid-search technique with a stratified 5-fold cross-validation method to find the best parameters between 4 and 10. Based on the numerical importance values provided from these methods, a feature was selected if its value was greater than the average of all features' importance values.

Three feature sets (derived from gradient boosting, decision tree, and random forest) were tested with six different classifiers. We tested logistic regression (LR, with L2 penalty) [22], K-nearest-neighbors (K=5) [23], random forest [24], gradient boosting [25], decision tree [26], MLP [27], linear SVM(L-SVM), the radial basis function (R-SVM), and 3rd order polynomial kernels (P-SVM) [28]. SVM parameters were set to C = 1.0 and gamma = 1 divided by the number of features. These parameters were optimally chosen by a trial and error approach. MLP parameters were set to 3 layers with 100 hidden units, 0.001 learning rate, 200 maximum iterations, the rectifier-linear-unit activation function for regularization [29], and the Adam optimizer which is purported to be computationally efficient, effective, and practical [30]. Class weights based on the inverse of the number of samples per class were applied to account for the imbalance of the training datasets.

Except for the decision-tree-based classification methods, data were standardized with zero mean and unit variance. Finally, a bagging classifier was performed with following parameters: 50 classifiers and 80% of the maximum samples. The random forest was not performed with a bagging classifier, as it is already based on the bagging approach. We evaluated the performance of the machine learning methods by calculating the balanced accuracy as follows:

$$Balanced\ accuracy = \frac{1}{N_{class}} \sum\nolimits_{i}^{N_{class}} \frac{TP_i}{TP_i + FP_i},$$

where N_class, TP, and FP represent the number of classes, true positive, and false positive, respectively.

Although the best option to train models is to use speech from all sessions (i.e., all 25 hours), it is more practically feasible to have a smaller number of training datasets. Therefore, to examine this, we trained machine-learning methods using a different number of speech data samples (protocols 1-3) from various times of the day (day, evening, and night). Note that the protocols in toto involved data from all 13 picture-describing speech samples (1 per session), as shown in Table 1. In this paper, we defined daytime, evening, and nighttime as 10 am-6 pm, 8-10 pm, and 2-6 am, respectively. Protocol 1 consisted of only one session, which occurred in only one of the 3 times: daytime, evening, or nighttime. Protocol 2 consisted of two sessions and up to two times, and protocol 3 used 3 sessions and up to 3 session times. The combinations of times used for training, per protocol, are provided in Table 1. These different protocols were

TABLE 1. Number of daytime, evening, nighttime, and test speech samples for training.

Protocol Index	Number of Speech Samples for Training (Four Questions)				Number of Speech Samples for Testing (Picture Describing)
	Daytime	Evening	Nighttime	Total	
1-1	1	0	0	1	13
1-2	0	1	0		
1-3	0	0	1		
2-1	2	0	0	2	13
2-2	0	2	0		
2-3	0	0	2		
2-4	1	1	0		
2-5	1	0	1		
2-6	0	1	1		
3-1	3	0	0	3	13
3-2	0	0	3		
3-3	2	1	0		
3-4	2	0	1		
3-5	1	2	0		
3-6	1	0	2		
3-7	0	2	1		
3-8	0	1	2		
3-9	1	1	1		

TABLE 2. An example of the protocol index 3-4.

			Find Best Feature/Model		Evaluation	
			Training	Validation	Training	Testing
			Four Questions	*Picture Describing*	*Four Questions*	*Picture Describing*
Session	Category	Time				
1	N/A	08:00		✔		✔
2	Daytime	10:00	✔	✔		✔
3		12:00		✔	✔	✔
4		14:00		✔		✔
5		16:00	✔	✔		✔
6		18:00		✔	✔	✔
7	Evening	20:00		✔		✔
8		22:00		✔		✔
9	N/A	24:00		✔	✔	✔
10	Nighttime	02:00		✔		✔
11		04:00		✔		✔
12		06:00	✔	✔		✔
13	N/A	08:00		✔		✔

designed to examine the effects of the number of training data sessions and which time combinations were most effective for training. The first, last, and 12 am sessions were excluded for training but they were used for testing.

Table 2 shows an example of an experimental protocol configuration for the index 3-4 (two daytime and one nighttime speech samples). Speech samples from the three randomly chosen sessions (two from daytime and one from nighttime) were used for training, while speech samples from all 13 sessions were used to validate and test the machine-learning methods.

III. RESULTS

A. MACHINE-LEARNING PERFORMANCE

Table 3 shows the testing results of protocol 1 (only 1 speech session) based on the selected feature-set. A set of ten subjects were used for training and the remaining ten subjects were used for testing. Protocol 1-1 (daytime only) exhibited more than 94% balanced accuracy in folds 2 to 4; however, fold 1 showed only 23.8% balanced accuracy. On the other hand, protocol 1-2 (evening) and 1-3 (nighttime) showed more consistent balanced accuracies among folds 1-4 with average balanced accuracies of 86.2 and 85.3%, respectively. Training using either evening or nighttime speech resulted in higher balanced accuracies of 91.3% and 90.8% for the evening and nighttime, respectively. For each fold, the best feature selection method and the classifier (determined based on the balanced accuracies) are provided in the last two columns of Table 3, respectively.

TABLE 3. Balanced accuracies of Protocol 1 (10 subjects).

Protocol (Day /Evening /Night)	Fold	All	Day-time	Evening	Night-time	Feature Selection	Classifier
1-1 (1/0/0)	*1*	0.238	0.240	0.250	0.233	GB	SVML
	2	0.946	0.920	1.000	0.900	GB	SVML
	3	0.946	0.980	1.000	0.867	GB	LR
	4	0.968	0.960	1.000	0.933	GB	LR
	Average	0.775	0.775	0.813	0.733		
1-2 (0/1/0)	*1*	0.815	0.800	0.900	0.767	RF	DT
	2	0.815	0.840	0.850	0.767	GB	RF
	3	0.938	0.940	1.000	0.900	GB	DT
	4	0.877	0.880	0.900	0.867	GB	SVML
	Average	0.862	0.865	0.913	0.825		
1-3 (0/0/1)	*1*	0.776	0.740	0.750	0.867	GB	DT
	2	0.800	0.740	0.800	0.900	RF	LR
	3	0.937	0.940	0.950	0.967	GB	RF
	4	0.899	0.860	0.950	0.900	RF	LR
	Average	0.853	0.820	0.863	0.908		

Table 4 shows the testing results of protocol 2, which used speech data from two sessions within the daytime, evening, or nighttime or all possible combinations of these three time zones to make up two sessions. By using two sessions for training the machine learning methods, all test sets showed more than 90% balanced accuracies. Training using two speech samples from the evening and nighttime showed the highest balanced accuracy of 95.6% (protocol index 2-6), followed by 94.8% for protocol index 2-2 (two speech samples from the evening). As shown in protocol 1, training using samples from either evening or nighttime resulted in the highest balanced accuracies of 98.8% and 95.0%, respectively. Also, training with only daytime or evening sessions' data showed poorer performance on the nighttime speech speaker identification, as the accuracies were only 81.7% and 80.8% for protocol 2-1 and 2-4, respectively.

Table 5 shows the results of protocol 3, which used three speech sessions. Training with three speech sessions from daytime, evening, and nighttime outperformed protocols 1 and 2. Training using speech sessions consisting entirely of either daytime or nighttime showed lower

TABLE 4. Balanced accuracies of Protocol 2 (10 subjects).

Protocol (Day /Evening /Night)	Fold	All	Day-time	Evening	Night-time	Feature Selection	Classifier
2-1 (2/0/0)	*1*	0.869	0.860	0.950	0.767	GB	RF
	2	0.954	0.980	1.000	0.867	GB	SVML
	3	0.946	1.000	1.000	0.867	RF	DT
	4	0.846	0.860	0.900	0.767	GB	SVML
	Average	0.904	0.925	0.963	0.817		
2-2 (0/2/0)	*1*	0.946	0.960	1.000	0.900	GB	LR
	2	0.985	0.960	1.000	1.000	GB	SVML
	3	0.922	0.920	0.950	0.900	GB	SVML
	4	0.938	0.940	1.000	0.900	RF	SVML
	Average	0.948	0.945	**0.988**	0.925		
2-3 (0/0/2)	*1*	0.861	0.860	0.850	0.900	GB	LR
	2	0.938 0.962	0.900	1.000	0.967	GB	SVML
	3	0.937	0.940	1.000	0.967	RF	LR
	4	0.938	0.920	0.850	0.967	GB	SVML
	Average	0.925	0.905	0.925	**0.950**		
2-4 (1/1/0)	*1*	0.885	0.940	0.950	0.800	DT	MLP
	2	0.876	0.880	0.900	0.733	GB	SVML
	3	0.923	0.940	1.000	0.833	RF	LR
	4	0.938	0.980	0.950	0.867	RF	RF
	Average	0.905	0.935	0.950	0.808		
2-5 (1/0/1)	*1*	0.846	0.860	0.850	0.800	RF	RF
	2	0.931	0.900	0.950	0.933	GB	SVML
	3	0.985	1.000	1.000	0.967	RF	RF
	4	0.915	0.900	0.950	0.900	RF	LR
	Average	0.919	0.915	0.938	0.900		
2-6 (0/1/1)	*1*	0.908	0.900	0.900	0.900	GB	DT
	2	0.969	0.960	1.000	0.967	GB	LR
	3	0.969	1.000	1.000	0.867	RF	DT
	4	0.977	0.980	1.000	0.967	RF	RF
	Average	**0.956**	0.960	0.975	0.925		

TABLE 5. Balanced accuracies of Protocol 3 (10 subjects).

Protocol (Day /Evening /Night)	Fold	All	Day-time	Evening	Night-time	Feature Selection	Classifier
3-1 (3/0/0)	*1*	0.914	0.900	1.000	0.867	GB	SVML
	2	0.915	0.960	0.950	0.767	RF	RF
	3	0.954	0.960	1.000	0.900	RF	LR
	4	0.938	0.980	1.000	0.833	RF	DT
	Average	0.930	0.950	**0.988**	0.842		
3-2 (0/0/3)	*1*	0.853	0.820	0.900	0.900	GB	SVML
	2	0.844	0.740	0.850	0.933	GB	DT
	3	0.930	0.900	1.000	0.933	RF	LR
	4	0.969	0.960	1.000	1.000	RF	RF
	Average	0.899	0.855	0.938	0.942		
3-3 (2/1/0)	*1*	0.931	0.920	1.000	0.900	RF	DT
	2	0.946	0.960	1.000	0.867	RF	DT
	3	0.961	1.000	1.000	0.900	RF	RF
	4	0.954	0.980	0.950	0.933	RF	RF
	Average	0.948	0.965	**0.988**	0.900		
3-4 (2/0/1)	*1*	0.931	0.920	0.950	0.900	RF	RF
	2	0.931	0.920	0.950	0.867	GB	RF
	3	0.992	1.000	1.000	0.967	RF	RF
	4	0.946	0.960	0.900	0.933	DT	SVMR
	Average	0.950	0.950	0.950	0.917		
3-5 (1/2/0)	*1*	0.954	0.980	1.000	0.867	RF	DT
	2	0.946	0.940	1.000	0.867	GB	LR
	3	0.969	1.000	1.000	0.933	RF	RF
	4	0.938	0.960	0.900	0.933	DT	MLP
	Average	0.952	0.970	0.975	0.900		
3-6 (1/0/2)	*1*	0.938	0.920	0.950	0.967	RF	RF
	2	0.954	0.920	0.950	0.967	RF	RF
	3	0.953	0.960	0.950	0.967	RF	DT
	4	0.954	0.920	1.000	1.000	GB	SVML
	Average	0.950	0.930	0.963	0.975		
3-7 (0/2/1)	*1*	0.890	0.860	0.950	0.833	DT	SVMR
	2	0.962	0.960	1.000	0.900	GB	RF
	3	0.992	1.000	1.000	1.000	RF	RF
	4	0.946	0.980	0.950	0.933	RF	RF
	Average	0.947	0.950	0.975	0.917		
3-8 (0/1/2)	*1*	0.953	0.960	1.000	0.900	RF	DT
	2	0.985	0.980	1.000	1.000	GB	SVML
	3	1.000	1.000	1.000	1.000	RF	RF
	4	0.929	0.920	0.900	0.967	DT	SVMR
	Average	0.967	0.965	0.975	0.967		
3-9 (1/1/1)	*1*	0.977	0.960	1.000	0.967	RF	RF
	2	0.985	0.960	1.000	1.000	RF	RF
	3	0.992	1.000	1.000	1.000	RF	RF
	4	0.969	0.980	0.950	0.967	RF	RF
	Average	**0.981**	**0.975**	**0.988**	**0.983**		

performance than did training using speech sessions from different times, as the balanced accuracies showed 93.0% and 89.9% in protocol 3-1 and 3-2, respectively, while other balanced accuracies were at least 94.7%. Table 6 shows results from 20 subjects with three feature selection methods and nine classifiers using speech data sessions from the daytime, evening, and nighttime (Protocol 3-9 with 20 subjects). Overall, the random-forest-based feature selection with either MLP or decision tree showed the highest balanced accuracy as both methods showed 92.2%. With the gradient boosting feature selection method, the MLP provided the best accuracy value of 90.3%; for the decision tree feature selection method, both R-SVM and GB provided the highest accuracy value of 88%.

B. INFORMATIVE FEATURES

By training with one session each from daytime, evening, and nighttime speech data (protocol 3-9), we obtained the highest balanced accuracies (see Table 5). For protocol 3-9, the random forest was consistently chosen as the best method for both feature selection and classifier, and the number of features ranged between 539 and 816. Since we performed

TABLE 6. Balanced accuracies from 20 subjects using a daytime, an evening, and a nighttime speech sessions for training (Protocol 3-9 with 20 subjects).

Feature Selection Method	Machine Learning	All	Daytime	Evening	Nighttime
Gradient Boosting	GB	0.717	0.720	0.800	0.733
	RF	0.899	**0.920**	**0.950**	0.900
	DT	0.748	0.730	0.800	0.767
	L-SVM	0.911	**0.920**	0.900	**0.917**
	P-SVM	0.605	0.580	0.700	0.533
	R-SVM	0.857	0.870	0.850	0.800
	KNN	0.899	**0.920**	**0.950**	0.900
	MLP	**0.903**	**0.920**	0.850	**0.917**
	LR	0.884	0.880	0.900	0.883
Random Forest	GB	0.919	0.910	**0.925**	**0.933**
	RF	0.787	0.800	0.900	0.733
	DT	**0.922**	**0.960**	**0.925**	0.867
	L-SVM	0.833	0.860	0.850	0.817
	P-SVM	0.888	0.890	0.875	0.883
	R-SVM	0.632	0.630	0.650	0.600
	KNN	0.818	0.810	0.800	0.833
	MLP	**0.922**	**0.960**	**0.925**	0.867
	LR	0.857	0.880	0.850	0.800
Decision Tree	GB	**0.880**	0.860	**0.950**	0.850
	RF	0.876	0.870	0.875	0.850
	DT	0.659	0.650	0.850	0.633
	L-SVM	0.783	0.830	0.825	0.717
	P-SVM	0.740	0.780	0.725	0.700
	R-SVM	**0.880**	**0.890**	0.900	**0.900**
	KNN	0.682	0.700	0.750	0.650
	MLP	0.864	0.880	0.925	0.833
	LR	0.783	0.830	0.825	0.717

TABLE 7. Feature list frequently selected in random forest feature-selection (protocol 3-9).

Opensmile Feature sets	Features	Counts
IS2011	audSpec_Rfilt_sma[17]_percentile1.0	4
IS2009	voiceProb_sma_linregerrQ	4
IS2012	F0final_sma_peakMeanAbs	4
IS2011	F0final_sma_quartile2	4
IS2013	audSpec_Rfilt_sma[0]_flatness	4
IS2011	F0final_sma_qmean	4
IS2010	F0finEnv_sma_quartile1	4
IS2010	lspFreq_sma[5]_linregerrA	4
IS2013	pcm_fftMag_spectralEntropy_sma_peakMeanRel	4
IS2011	pcm_fftMag_mfcc_sma[10]_amean	4
IS2011	audSpec_Rfilt_sma_de[17]_pctlrange0-1	4
IS2011	audSpec_Rfilt_sma[12]_pctlrange0-1	3
IS2011	pcm_fftMag_spectralVariance_sma_pctlrange0-1	3
IS2011	pcm_RMSenergy_sma_quartile3	3

IS: InterSpeech, audSpec: auditory spectrum, Rfilt: Relative Spectral Transform (RASTA)-style filtered, sma: smoothing by moving average, voiceProb: power of harmonics divided by the total power, linregerrQ: quadratic error computed as the difference of the linear approximation and the actual contour, peakMeanAbs: the mean of absolute peak, F0final: The smoothed fundamental frequency contour, quartile2: the 50% percentile, qmean: the mean glottal flow rate, F0finEnv: The envelope of the smoothed fundamental frequency contour, qurtile1: the 25% percentile, lspFreq: The 8 line spectral pair frequencies computed from 8 LPC coefficients, spectralEntropy: a signal is a measure of its spectral power distribution, peakMeanRel: the mean of relative peak, mfcc: Mel-frequency cepstral coefficients, amean: the arithmetic mean of the contour, de: delta, pctlrange0-1: the outlier robust signal range 'max-min' represented by the range of the 1% and the 99% percentile, quartile3: the 75% percentile, RMSenergy: the root-mean-square signal frame energy

feature selection methods for each session, frequently chosen features are most likely informative features for the classification task. Table 7 shows features that have been selected at least three times. Fig. 2 shows a significant difference between sessions for the protocol index 3-9, which resulted in the highest balanced accuracy (98.1%). The outlier robust signal range 'max-min' represented by the range of the 1% and the 99% percentile of the smoothed relative spectral transform style filtering on auditory spectra features (i.e., the audSpec_Rfilt_sma[12]_pctlrange0-1) showed a significant difference between sessions 3 to both 10 and 11, which are daytime and nighttime speech data, respectively. The measure of the outlier of the signal range 'max-min' represented by the 1% and the 99% percentile of the smoothed spectral variance (i.e., pcm_fftMag_spectral-Variance_sma pctlrange0-1) also showed significant difference between session 11 to both 3 and 6, which are daytime and nighttime speech data, respectively. The 75% percentile of the smoothed root-mean-square signal frame energy (i.e., pcm_RMSenergy_sma_quartile3) showed significant difference between 10 (nighttime) and to among 2-8 sessions (daytime and evening), and sessions between 3, 4, and 7 (daytime and evening) to 9, 11, and 12 (nighttime). Moreover, we found that different vocal features, including relative spectral transform, fundamental frequency, and Mel-frequency cepstral coefficients (MFCC), were frequently selected using the feature selection methods. Details of these feature sets provided by OpenSmile software are shown in Table 7.

IV. DISCUSSION

We used machine learning methods to identify speakers from spontaneous speeches recorded during 25 hours of sleep deprivation. Several different machine learning methods were trained and tested using different numbers of speech samples from the daytime, evening, and nighttime over the 25 hours. The machine learning methods trained with speech sessions from different times resulted in higher performance for identifying speakers than when trained using speech samples from only one of the three times. With three speech sessions each from the day, evening, and night used for training, the best balanced accuracy of 98.1% was attained. respectively.

Our method consists of feature selection, training, validation, and testing of classifiers. When feature selection methods were used without appropriate validation, they were likely to over-fit based on the set of features. The closed-set speaker identification systems are supposed to over-fit for given subjects because they are designed to identify unknown speakers from a list of enrolled people. For example, the training has to be carried out using data from all speakers so that they can be identified when spoken. Moreover, previous studies tested a wide set of classification methods, but questions remain about which method is most appropriate in practice [8]–[10]. Therefore, to test the generalization of the classification methods, we segmented the number of subjects into two groups for validating and testing data. We also prevented possible bias that can happen when splitting data into two groups by using four different data time segments (i.e., four folds), as one of our results showed poor performance in only one fold (fold 1 of protocol 1).

Also, we discovered informative features that were frequently chosen by our various feature selection methods, including the fundamental frequency, MFCC, and the relative spectral transform style filtering on auditory spectra features. This is in agreement with previous studies as these features have also been used in many speaker identification and verification approaches [8], [10], [31]. MFCC features, especially, have been applied to detect fatigue caused by sleep deprivation [32], [33]. Greeley *et al.* used 36 MFCC features to detect fatigue from fixed words, comprised of 12 cepstral coefficients along with their first and second time derivatives [7]. Baykaner *et al.* used 19 MFCC features to detect fatigue from speech data from a book [33]. Given that the purpose of the study was to examine how 25-hour sleep deprivation affects speech recognition, it is not surprising that features including MFCC and other frequently chosen features from different times throughout the day resulted in higher performance than just using speech data from a particular time of the day or even a combination of two time sessions. While we have limited the effect of increasing the number of protocol sessions from one to three to examine their effect on the accuracy of speaker identification, including more than three training sessions from any of the combinations of daytime, evening, and nighttime may not necessarily increase the accuracy since we have already reached greater than 98% balanced accuracy, as shown in Table 5.

Recording nighttime voice samples to have them for training data can be challenging, as most people are inactive during these hours. We aim to conduct a future study looking into features that are not affected by sleep deprivation, allowing higher accuracies, in order to develop applications that can work with speech recorded at any time of the day.

V. CONCLUSION

We found that diversifying training speech sessions from different time segments in a given day resulted in better performance for closed-set text-independent speaker identification during 25 hours of prolonged wakefulness. With training based on speech sessions during the daytime, evening, and nighttime, machine-learning methods were able to obtain robust performance during 25-hour sleep deprivation experiments. This work has shown some promise in that machine learning can potentially be used for applications that require identifying a speaker's voice even when it is affected by tiredness and fatigue among other stress factors (e.g., 24-hour phone services) which are frequently encountered in scenarios such as emergency rooms and long-duration repetitive task operations.

ACKNOWLEDGMENT

The views expressed in this article reflect the results of research conducted by the author(s) and do not necessarily reflect the official policy or position of the Department of the Navy, Department of Defense, nor the U.S. Government. The study protocol was approved by the University of Connecticut and Naval Submarine Medical Research Laboratory Institutional Review Board in compliance with all applicable Federal regulations governing the protection of human subjects. Jeffrey Bolkhovsky is an employee of the U.S. Government. This work was prepared as part of his official duties. Title 17 U.S.C. §105 provides that 'Copyright protection under this title is not available for any work of the United States Government.' Title 17 U.S.C. §101 defines a U.S. Government work as a work prepared by a military service member or employee of the U.S. Government as part of that person's official duties.

REFERENCES

[1] E. MacAskill, ''Did'Jihadi John'kill Steven Sotloff,'' *Guardian*, Sep. 2014. Accessed: Jul. 5, 2021. [Online]. Available: https://nam10.safelinks.protection.outlook.com/?url=https%3A%2F%2Fwww.theguardian.com%2Fmedia%2F2014%2F sep%2F02%2Fsteven-sotloff-video-jihadi-john&data=04%7C01%7C%7Ce874cbbe74724774c66808d93fa4616e %7C17f1a87e2a254eaab9df9d439034b080%7C0%7C0%7C637610797609975344%7CUnknown%7CTWFpbGZsb3d8ey JWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C1000&sdata=uQMYlpDyrprsatlNDN6YUC %2BdAhWwogP7myQ%2B7NeTgE4%3D&reserved=0 and https://www.theguardian.com/media/2014/sep/02/steven-sotloff-video-jihadi-john

[2] B. Beranek, ''Voice biometrics: Success stories, success factors and what's next,'' *Biometric Technol. Today*, vol. 2013, no. 7, pp. 9–11, Jul. 2013, doi: 10.1016/S0969-4765(13)70128-0.

[3] J. H. L. Hansen, A. Sangwan, and W. Kim, ''Speech under stress and lombard effect: Impact and solutions for forensic speaker recognition,'' in *Forensic Speaker Recognition: Law Enforcement Counter-Terrorism*, A. Neustein and H. A. Patil, Eds. New York, NY, USA: Springer, 2012, pp. 103–123, doi: 10.1007/978-1-4614-0263-3_5.

[4] T. Wu, Y. Yang, and Z. Wu, ''Improving speaker recognition by training on emotion-added models,'' in *Affective Computing and Intelligent Interaction*. Berlin, Germany: Springer, 2005, pp. 382–389, doi: 10.1007/11573548_49.

[5] G. S. Raja and S. Dandapat, ''Speaker recognition under stressed condition,'' *Int. J. Speech Technol.*, vol. 13, no. 3, pp. 141–161, Sep. 2010.

[6] Y. Harrison and J. A. Horne, ''Sleep deprivation affects speech,'' *Sleep*, vol. 20, no. 10, pp. 871–877, Oct. 1997.

[7] H. P. Greeley, E. Friets, J. P. Wilson, S. Raghavan, J. Picone, and J. Berg, ''Detecting fatigue from voice using speech recognition,'' in *Proc. IEEE Int. Symp. Signal Process. Inf. Technol.*, Aug. 2006, pp. 567–571.

[8] N. Chauhan, T. Isshiki, and D. Li, ''Speaker recognition using LPC, MFCC, ZCR features with ANN and SVM classifier for large input database,'' in *Proc. IEEE 4th Int. Conf. Comput. Commun. Syst. (ICCCS)*, Feb. 2019, pp. 130–133, doi: 10.1109/CCOMS.2019.8821751.

[9] H. Fenglei and W. Bingxi, ‘‘Text-independent speaker recognition using support vector machine,’’ in *Proc. Int. Conf. Info-Tech Info-Net*, Nov. 2001, pp. 402–407.

[10] O. Mamyrbayev, N. Mekebayev, M. Turdalyuly, N. Oshanova, T. I. Medeni, and A. Yessentay, ‘‘Voice identification using classification algorithms,’’ in *Intelligent System and Computing*. London, U.K.: IntechOpen, 2019.

[11] M. F. Akhsanta and S. Suyanto, ‘‘Text-independent speaker identification using PCA-SVM model,’’ in *Proc. 3rd Int. Seminar Res. Inf. Technol. Intell. Syst. (ISRITI)*, Dec. 2020, pp. 525–528.

[12] N. Chauhan, T. Isshiki, and D. Li, ‘‘Speaker recognition using fusion of features with feedforward artificial neural network and support vector machine,’’ in *Proc. Int. Conf. Intell. Eng. Manage. (ICIEM)*, Jun. 2020, pp. 170–176.

[13] S. Sujiya and D. E. Chandra, ‘‘A review on speaker recognition,’’ *Int. J. Eng. Technol.*, vol. 9, pp. 1592–1598, 2017.

[14] A. M. Ariyaeeinia, J. Fortuna, P. Sivakumaran, and A. Malagaonkar, ‘‘Verification effectiveness in open-set speaker identification,’’ *IEE Proc.-Vis., Image Signal Process.*, vol. 153, no. 5, pp. 618–624, 2006.

[15] T. Kinnunen and H. Li, ‘‘An overview of text-independent speaker recognition: From features to supervectors,’’ *Speech Commun.*, vol. 52, no. 1, pp. 12–40, Jan. 2010, doi: 10.1016/j.specom.2009.08.009.

[16] E. Alpaydin, *Introduction to Machine Learning*. Cambridge, MA, USA: MIT Press, 2020.

[17] M. Artkoski, J. Tommila, and A.-M. Laukkanen, ‘‘Changes in voice during a day in normal voices without vocal loading,’’ *Logopedics Phoniatrics Vocology*, vol. 27, no. 3, pp. 118–123, Jan. 2002, doi: 10.1080/140154302760834840.

[18] B. M. Ben-David and M. Icht, ‘‘Voice changes in real speaking situations during a day, with and without vocal loading: Assessing call center operators,’’ *J. Voice*, vol. 30, no. 2, pp. 247.e1–247.e11, Mar. 2016, doi: 10.1016/j.jvoice.2015.04.002.

[19] (Apr. 27, 2015). *Eye Tracking Technology for Research—Tobii Pro*. Accessed: Oct. 9, 2019. [Online]. Available: https://www.tobiipro.com/

[20] F. Eyben, M. Wöllmer, and B. Schuller, ‘‘Opensmile: The munich versatile and fast open-source audio feature extractor,’’ in *Proc. Int. Conf. Multimedia (MM)*, 2010, pp. 1459–1462.

[21] H. Deng and G. Runger, ‘‘Feature selection via regularized trees,’’ in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jun. 2012, pp. 1–8.

[22] P. McCullagh, *Generalized Linear Models*. Evanston, IL, USA: Routledge, 2019.

[23] N. S. Altman, ‘‘An introduction to kernel and nearest-neighbor nonparametric regression,’’ *Amer. Statist.*, vol. 46, no. 3, pp. 175–185, Aug. 1992.

[24] T. K. Ho, ‘‘Random decision forests,’’ in *Proc. 3rd Int. Conf. Document Anal. Recognit.*, Aug. 1995, pp. 278–282.

[25] J. H. Friedman, ‘‘Greedy function approximation: A gradient boosting machine,’’ *Ann. Statist.*, vol. 29, no. 5, pp. 1189–1232, Oct. 2001.

[26] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, ‘‘Classification and regression trees. Belmont, CA: Wadsworth,’’ *Int. Group*, vol. 432, pp. 151–166, Jul. 1984.

[27] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York, NY, USA: Springer, 2009.

[28] C. Cortes and V. Vapnik, ‘‘Support-vector networks,’’ *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.

[29] X. Glorot, A. Bordes, and Y. Bengio, ‘‘Deep sparse rectifier neural networks,’’ in *Proc. 14th Int. Conf. Artif. Intell. Statist.*, 2011, pp. 315–323.

[30] D. P. Kingma and J. Ba, ‘‘Adam: A method for stochastic optimization,’’ 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980

[31] A. Zulfiqar, A. Muhammad, A. M. Martinez-Enriquez, and G. Escalada-Imaz, ‘‘Text-independent speaker identification using VQ-HMM model based multiple classifier system,’’ in *Proc. Mexican Int. Conf. Artif. Intell.*, 2010, pp. 116–125.

[32] H. P. Greeley, J. Berg, E. Friets, J. Wilson, G. Greenough, J. Picone, J. R. Whitmore, and T. Nesthus, ‘‘Fatigue estimation using voice analysis,’’ *Behav. Res. Methods*, vol. 39, no. 3, pp. 610–619, Aug. 2007.

[33] K. Baykaner, M. Huckvale, I. Whiteley, O. Ryumin, and S. Andreeva, ‘‘The prediction of fatigue using speech as a biosignal,’’ in *Statistical Language and Speech Processing*. Cham, Switzerland: Springer, 2015, pp. 8–17, doi: 10.1007/978-3-319-25789-1_2.