Subject Identification from Electroencephalogram (EEG) Signals During Imagined Speech

Katharine Brigham and B. V. K. Vijaya Kumar

Abstract—**We investigate the potential of using electrical brainwave signals during imagined speech to identify which subject the signals originated from. Electroencephalogram (EEG) signals were recorded at the University of California, Irvine (UCI) from 6 volunteer subjects imagining speaking one of two syllables, /ba/ and /ku/, at different rhythms without performing any overt actions. In this work, we assess the degree of subject-to-subject variation and the feasibility of using imagined speech for subject identification. The EEG data are first preprocessed to reduce the effects of artifacts and noise, and autoregressive (AR) coefficients are extracted from each electrode's signal and concatenated for subject identification using a linear SVM classifier. The subjects were identifiable to a 99.76% accuracy, which indicates a clear potential for using imagined speech EEG data for biometric identification due to its strong inter-subject variation. Furthermore, the subject identification appears to be tolerant to differing conditions such as different imagined syllables and rhythms (as it is expected that the subjects will not imagine speaking the syllables at exactly the same rhythms from trial to trial). The proposed approach was also tested on a publicly available database consisting of EEG signals corresponding to Visual Evoked Potentials (VEPs) to test the applicability of the proposed method on a larger number of subjects, and it was able to classify 120 subjects with 98.96% accuracy.**

I. Introduction

Functional brain imaging techniques that are designed to measure an aspect of brain function can be employed to obtain tangible information related to brain activity. Electroencephalogram (EEG) is one such technique, which measures the electric fields that are produced by the activity in the brain. From EEG measurements, it is possible to extract information and determine the intent of the user for a number of different mental activities (e.g., motor imagery, motor planning), using a variety of electrophysiological signals such as slow cortical potentials, P300 potentials, and mu or beta rhythms recorded from the scalp, and cortical neuronal activity recorded by implanted electrodes [1]. The use of EEG for the communication of intent is one of the bases of Brain-Computer Interface (BCI) research, which is geared towards the development of systems to afford people with disabilities or severe neuromuscular disorders the capability of basic communication (by operating word processing programs or through neuroprostheses) [1]. Companies such as Emotiv Systems, Uncle Milton, and Mattel have also been developing video games that use dry electrodes to capture EEG signals that are decoded to determine the user's intended commands. And more recently, studies have been performed on interpreting EEG signals recorded during imagined speech [2]-[6]. The capability to translate imagined speech has the clear potential for a variety of uses such as in BCI applications designed to aid those who are physically unable to talk, or in situations where visual or audible communication is undesirable necessitating the use of silent communication.

It is thus natural to investigate if EEG signals acquired during mental activities can also be used for subject identification to create a more secure environment for applications such as BCIs, game play, or silent communication. For example, in one particular scenario, on the battlefield, military operations may be conducted such that dispersed team members need to covertly exchange information through silent communication. In this type of setting, it may be desirable to enhance the security of the covert operations by verifying the identity of each individual during communication. Or for game play, user profiles may be easily retrieved through automatic subject identification with brain signals without the use of a username and password. It makes sense to utilize the EEG signals already being collected for other purposes (e.g., for imagined speech) to authenticate subjects. Using EEG as a biometric has several advantages, as it cannot easily be forged, it is confidential (as it corresponds to a mental task), and it is also almost impossible to steal [7].

EEG signals have previously been proposed as a biometric for authenticating or identifying individuals. To this end, a variety of EEG data has been studied, such as that collected during motor imagery, during mental tasks (e.g., mental letter composing), while subjects are at rest (with eyes open or shut), and from brain responses to visual stimuli (i.e., Visual Evoked Potentials (VEPs)). To the best of our knowledge, subject identification using imagined speech EEG data has not yet been demonstrated. Paranjape et al. [8] conducted experiments using 40 subjects who simply had their eyes open during the EEG recording. The authors used autoregressive (AR) models of varying orders from 3 to 21, and then applied discriminant function analysis to classify these 40 subjects, attaining correct recognition rates between 49% and 85% depending on the model order.

This work was supported the Army Research Office (under ARO 54228-LS-MUR).

The authors are with the Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA 15213 USA (e-mail: {kbrigham, kumar}@ece.cmu.edu).

978-1-4244-7580-3/10/$26.00 ©2010 IEEE

Poulos et al. [9] also applied AR models of orders ranging from 8 to 12 on EEG data for 79 subjects where the subjects were at rest with their eyes closed during the experiment, thus recording their alpha rhythms. Poulos et al. used a Linear Vector Quantizer neural network to classify the data. To test the subject classification, four experiments were performed where one subject (referred to as subject A) was tested against 75 "non-A" subjects, and repeated for subjects B, C, and D. The classification scores obtained ranged from 72% to 84% in this one vs. rest classification test for 4 subjects.

Parametric power spectral density (PSD) models are also used in the present study and have been used in other EEG biometric authentication and/or identification methods as well. There are several other related works where AR models are applied using EEG recordings from mental tasks or motor imagery. Palaniappan [10] proposed a two-stage biometric authentication method using the mental task EEG data collected by Keirn and Aunon [11] for 5 subjects. Ten trials were collected in different sessions for each mental task (5 total) for each subject, lasting 10 seconds long. The mental tasks consisted of a baseline where the subject was at rest, visual counting, geometric figure rotation, mental multiplication, and mental letter composing. In [10], each mental task trial was segmented into 0.5 second blocks to obtain a larger dataset. Subject identification was subsequently performed for each mental task separately. A number of features including AR coefficients were extracted from the data. The False Reject Rate (FRR) ranged from 1.5% to 0%, and the False Accept Rate (FAR) was 0% for all 5 subjects. This approach obtained excellent results; however, since each trial was broken up into many smaller segments (creating more data points), it is possible that there were data from within the same trials used in both the training and testing datasets, so the generalization power of this method is largely unknown.

Bao et al. [12] investigated using motor imagery data for subject identification with the motor imagery EEG dataset IIIa from the BCI Competition 2003 (provided by the Graz University of Technology) [13] for 3 subjects where the user was provided with feedback. The authors used a combination of features including AR coefficients, phase synchronization, and energy spectral density, and a neural network classifier to achieve average identification rates ranging from 81.2% to 90.6%. Hu [14], with the same motor imagery data, proposed to use an autoregressive moving average (ARMA) model to authenticate or identify subjects also using a neural network. This approach yielded an average classification rate of 83.9% for authentication, and 81.9% for identification. Another approach to person authentication using motor imagery (and also the generation of words beginning with the same random letter) for 9 subjects was developed by Marcel and Millán, who use a statistical framework based on Gaussian Mixture Models and Maximum A Posteriori model adaptation [7]. These authors report a Half Total Error Rate (HTER) as their performance metric, which is defined as the average of the FAR and FRR. The authors chose to use the Equal Error Rate (EER) where the FAR and FRR are equal to obtain the HTER (which was essentially the EER). Several experiments were performed yielding HTERs from 7.1% to 35.7%. Motor imagery for moving the left arm seemed to provide the best results.

Alternatively, a number of studies (e.g., [15]-[17]) on EEG biometrics have been performed using Visual Evoked Potentials (VEPs), which are brain activity responses to visual stimuli. Zúquete et al. [15] proposed a biometric authentication method using the energy of differential EEG signals as features, and one-class classifiers which consisted of either a k-Nearest Neighbors (k-NN) classifier or a Support Vector Data Description (SVDD) classifier, or a combination of the two. Several experiments were run, resulting in classification accuracies ranging from 50.1% to 93.7% for 70 subjects. Das et al. [16] used VEP data from 20 subjects and extracted discriminative spatio-temporal filters. Classification based on these features was accomplished using Linear Discriminant Analysis (LDA) and Support Vector Machines (SVM) separately, obtaining classification rates ranging from 75% to 94%. Palaniappan and Mandic [17] performed subject identification using VEPs by computing the energy of filtered VEPs as features for 40 subjects, and then performing feature selection using the Davies-Bouldin Index (DBI), which is a metric designed to help find natural partitions of the data [18]. The selected features were then classified using an Elman neural network trained for each subject, and the authors achieved accuracies ranging from 13.63% for 1 channel, approximately 70% for 10 channels, and up to 98.56% for 50 channels. The same authors in a separate work proposed a different approach with the same data where the Multiple Signal Classification (MUSIC) algorithm [19] was used to estimate the dominant frequency and power content for filtered VEP signals (in the gamma band of 25-56Hz) that appeared to have only one dominant sinusoid [20]. For 102 subjects, an average classification rate of 98.12% was achieved again using the Elman neural network as the classifier.

Using brain responses to visual stimuli for identifying or authenticating individuals has yielded good results for a large number of subjects, but lacks universality as a small percentage of the population potentially cannot be authenticated (namely those who are severely visually impaired or blind). Imagined speech presents one of the most intuitive, simple, and universal ways in which to control your brain signaling without actualizing any sort of physical action. A relatively common interpretation of thought to oneself without vocalization is words that are "heard" within one's head. It is plausible that a person may hear his or her own thoughts (i.e., imagine speaking) on a regular basis; thus, it becomes quite an easy task to perform.

In contrast, motor imagery and other mental activities that have been investigated for subject identification using EEG data may be more difficult to perform since these tasks are not frequently executed. Therefore, imagining moving your left or right arm without any overt actions may require a bit of training to hone in on the proper way to imagine movement in order to generate discriminative EEG signals. Successful BCI operation for systems interpreting imagined movements tends to rely on feedback to the user and typically requires that the user develops and maintains the learned skill, which is a skill that consists not of proper muscle control but rather the proper control of specific electrophysiological signals [1]. Studies that have been performed on using motor imagery EEG data for subject identification ([12], [14]) utilized datasets where feedback was provided. In work done by Marcel and Millán [7], however, motor imagery EEG data collected during non-feedback sessions were used for person authentication (though the authors note that the database used is small; there were only 36 trials for each of the 9 subjects), and the authors also state that no conclusions can be drawn for the task of person authentication from their results [7]. It seems to be more practical to use methods for subject identification that do not require training or feedback, and since it is currently unclear how subject identification would perform under non-feedback conditions, the use of imagined movements for subject identification with EEG data may not be a very viable option.

Other mental activities that have been used in EEG biometric experiments, such as those performed by Keirn and Aunon [11], involved relatively complex tasks such as mental multiplication of large numbers or geometric figure rotation, which can also be considered more difficult than simply imagining speech. Experiments run by Paranjape et al. [8] did not require any mental task at all; subjects only had to be at rest with their eyes open during the EEG recording, which was the simplest task of all these mental activities. However, in both studies, there were a fairly small number of samples in each EEG dataset. Paranjape et al. only collected (8 trials) for each of the 40 subjects, and Keirn and Aunon only collected 10 trials per subject, per task. Because these datasets are limited, it is difficult to assess how well the methods used to identify the subjects would generalize over larger datasets.

Imagined speech is a rather convenient method for subject identification with EEG data since it is easy to do, as it is typically a regularly executed activity (as opposed to tasks such as complex mental multiplication). It also does not require any external stimuli since the imagined speech brainwave activity is generated at the subject's discretion. The results shown in this work demonstrate that with imagined speech EEG data, subjects can be identified to a high degree of accuracy over a very large number of trials and several different conditions. This outcome indicates that varied speech (e.g., different syllables, different rhythms) have little to no bearing on the ability to accurately identify subjects, which suggests that training is likely unnecessary to generate subject-specific EEG signals. Also, in some situations (e.g., those requiring silent communication), EEG signals are intended to be used for communicating imagined speech, and it is no extra burden to use those same signals for subject authentication or identification. Overall, since imagined speech is a natural, simple task to perform, it seems to be a very practical candidate for subject identification using EEG data.

It should also be noted that the same imagined speech EEG dataset used here for the subject classification experiments was also used in another work by Deng et al. [21], where instead a rhythm classification was performed. More details on this dataset are provided in Section II. Deng et al. aimed to classify the different rhythms at which the syllables were covertly spoken. However, while rhythm classification was done for each subject separately, this differs from the work done here since subject classification was not performed in that only the rhythms at which the syllables were silently spoken were identified as opposed to the subjects themselves.

The remaining sections are organized as follows. Section II provides a description of the EEG datasets used in this work, and Section III details the preprocessing approach used to reduce the effects of unwanted activity (e.g., noise) in the EEG data. Section IV subsequently discusses the proposed approach, which includes feature extraction and subject classification. Section V summarizes the final results and provides a brief analysis, followed by the conclusions in Section VI.

II. Data Collection

Two different datasets were used in this study; one was composed of imagined speech EEG data that was collected at the University of California, Irvine (UCI), and the other dataset was obtained from a publicly available database [22] of VEP EEG data for alcoholic and non-alcoholic subjects, which was used to test the general applicability of the proposed approach. For the VEP EEG data, experiments were conducted in which subjects were presented with visual stimuli consisting of black and white pictures of objects from the Snodgrass and Vanderwart standardized picture set [23]. (More details on the experiment can be found in [24]). EEG signals were measured using 64 electrodes and sampled at 256Hz for 1 second following the stimulus presentation. These signals were also hardware-filtered to a frequency range of 0.1Hz to 50Hz [15]. There were a total of 122 subjects with a varied number of trials per subject ranging from 30 to 120 trials for a total of 11,074 trials.

The imagined speech EEG dataset was collected by a research group in the Department of Cognitive Sciences at UCI. They conducted experiments in which 6 volunteer subjects imagined speaking two syllables, /ba/ or /ku/ while their electrical brainwave activity was being recorded by EEG. These syllables were selected since they contain no semantic meaning so that classification would be performed on the imagined speech instead of the semantic contribution to imagined speech production [3]. The subjects were instructed to covertly speak a given syllable at a certain rhythm, both of which were provided via audio cues. So in each trial, a syllable (either /ba/ or /ku/) was heard through a set of Stax electrostatic earphones followed by a series of clicks at the desired rhythm for the imagined speech. Approximately 1.5 seconds after the last click, the subject was to begin to imagine speaking the spoken syllable at the given rhythm (see Fig. 1 for timelines [3]). During the time segment corresponding to EEG signals of interest, no audio or video stimuli were present.

As described in [3], the EEG data were recorded using a 128 Channel Sensor Net by Electrical Geodesics [25] and sampled at 1024Hz. A single experimental session was typically comprised of 20 trials for each condition, and data were recorded over separate sessions, which varied for each subject. During the recording, the subjects were seated in a dimly lit room and instructed to keep their eyes open and to fixate on a certain point while avoiding any eye blinks and muscle movement.

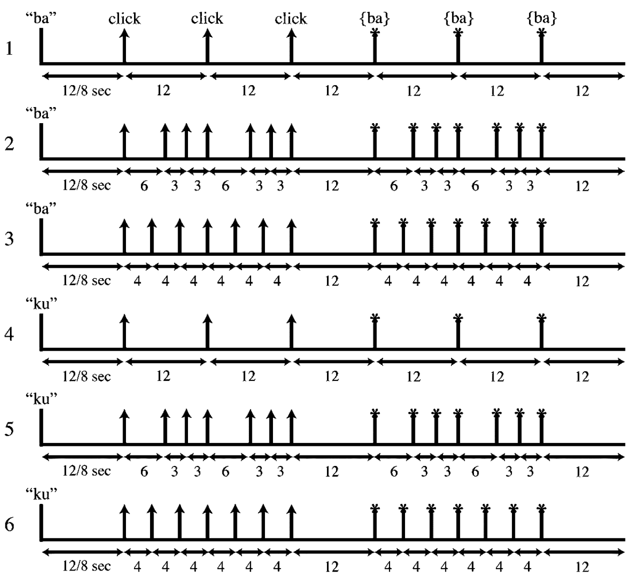

Fig. 1. Timelines for six different conditions in the covert speech experiment [3] (either /ba/ or /ku/ was covertly spoken at three different rhythms).

III. Data Preprocessing

Although the subjects attempt to keep movement to a minimum during these recordings, the EEG data inevitably contains some presence of artifacts (i.e., changes in EEG amplitudes that do not correspond to brainwave activity but to eye movements or muscle movements instead). These artifacts tend to dominate and obscure the actual cortical signal. Additionally, in some cases these artifacts can be fairly predictive. This may result in deceptively high recognition rates since a classifier would succeed by identifying these artifacts as opposed to the portions of the signal that reflect the true brain activity. Therefore, the EEG data is first preprocessed to remove artifacts and also to reduce noise (e.g., 60Hz line noise).

For the imagined speech EEG data, electromyographic (EMG) artifacts (i.e., muscle artifacts) are first considered for removal using the same preprocessing steps suggested by D'Zmura et al. in [3]. EEG signals from 18 of the 128 electrodes that are closest to the neck, eyes, and temple are discarded since they are the most prone to EMG artifacts. Furthermore, since EMG artifacts are typically present in frequencies greater than 25Hz, the remaining EEG signals are filtered to a frequency range of 4 to 25Hz, which additionally removes the 60Hz line noise from these signals. The data is then detrended to remove baseline drift and downsampled to a more manageable sampling rate of 256Hz. In addition, 4 electrodes were found to be faulty in a number of trials (as no data were collected by these electrodes during these trials), so these electrodes were completely discarded as well.

For the VEP EEG dataset, only eye blink artifacts (which are characterized by large signal spikes) are considered for removal. Since evoked brain activity from visual stimuli is known to produce a significant change in gamma band oscillations (which is considered to be between 25Hz and 100Hz) [20], and 60Hz line noise appeared to be adequately suppressed in the raw VEP EEG data, there was no additional band-pass filtering done for this dataset in order to retain potentially important information that is thought to lie in the gamma band. However, any trials containing EEG signals deemed to be contaminated by eye blink artifacts (if the signal amplitudes exceeded the thresholds of +/- 100μV) were discarded, as was done in [20]. This also resulted in the complete removal of data from two of the subjects.

For the imagined speech EEG dataset, eye blink artifacts were similarly detected and discarded. The same thresholds of +/-30μV were used here as in [3] (though these thresholds can vary as some subjects naturally emanate high amplitude signals). For each subject, a percentage of trials (ranging from 10% to 35%) with the largest quantity of offending signals are discarded. After preprocessing the imagined speech EEG data, the data were reduced to the number of trials listed and broken down in Table I. For the VEP EEG dataset, after eye blink artifact removal, the number of trials remaining for each subject ranged from 16 to 118 with a total of 9,596 trials for 120 subjects.

TABLE I
Number of Trials Remaining After Artifact Removal

	Percent discarded	# of remaining trials
S1	10%	743
S2	15%	594
S3	35%	783
S4	15%	957
S5	20%	381
S6	30%	329
Total:		3,787

IV. Feature Extraction and Subject Identification

A. Univariate Autoregressive (AR) Model

After preprocessing, we used the signal model shown in Fig. 2 along with some assumptions to estimate the power spectral density (PSD) of each EEG signal (here represented by $x[n]$, the observed signal).

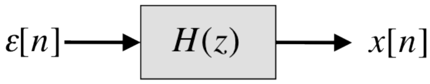

Fig. 2. Block diagram of the signal model

With this model, we assume that the EEG signals can be modeled by a wide-sense stationary random process and that each signal is generated by inputting a zero-mean white noise random process $\varepsilon[n]$ with variance σ^2 into a linear shift-invariant all-pole filter. The corresponding time series model [26] is given in equation (1).

$$x[n] = -\sum_{k=1}^{p} \alpha_k x[n-k] + \varepsilon[n] \qquad (1)$$

where $x[n]$ is the observed signal at time n, α_k are the model coefficients, and $\varepsilon[n]$ is the white noise random process. The integer p is the order of this model. As can be seen in this equation, this autoregressive (AR) model attempts to predict the current time sample given previous time samples, and its transfer function is as given in equation (2).

$$H(z) = \frac{1}{1 + \sum_{k=1}^{p} \alpha_k z^{-k}} \qquad (2)$$

Consequently, the AR coefficients α_k completely determine the spectrum of the model output, since the spectrum of the model output is the product of the squared magnitude of the transfer function, $|H(e^{j\omega})|^2$ and the variance of the white noise input, σ^2. Therefore, applying an AR model to a signal and computing its autoregressive coefficients is one method for estimating a signal's PSD.

For each dataset, AR coefficients were computed for each electrode's signal using the Burg method as described in [27] and concatenated to form a feature vector. Orders 2 through 6 were tested to see which order gave the best classification accuracies. An AR model order of 4 appeared to be optimal for the VEP EEG dataset, and an AR model of order 2 was optimal for the imagined speech EEG dataset.

B. *Subject Classification*

Classification of the subjects was performed using a linear Support Vector Machine (SVM) classifier. 10 iterations of 4-fold cross-validation were run to obtain the final average identification rates. The software package LibSVM [28] was used for training and testing the SVMs, and each testing set was kept distinct from the training set. The linear SVM classifiers were compared to the k-Nearest Neighbors classifier based on the Euclidean distance between AR model coefficients in the training and testing set. The results for varying values of k are also summarized in the next section.

V. Results

Subjects from the imagined speech EEG dataset were classified using a 2nd order AR model. The classification performance using a linear SVM classifier is shown in Table II with the corresponding Cumulative Match Characteristic (CMC) curve shown in Fig. 3.

TABLE II

Average rank-1 identification accuracies and corresponding 95% confidence intervals for the Imagined Speech EEG dataset using a linear SVM classifier

EEG Dataset	Average Classification Rate ± 95% confidence interval
Imagined Speech (6 subjects)	**99.76% ± 0.04%**

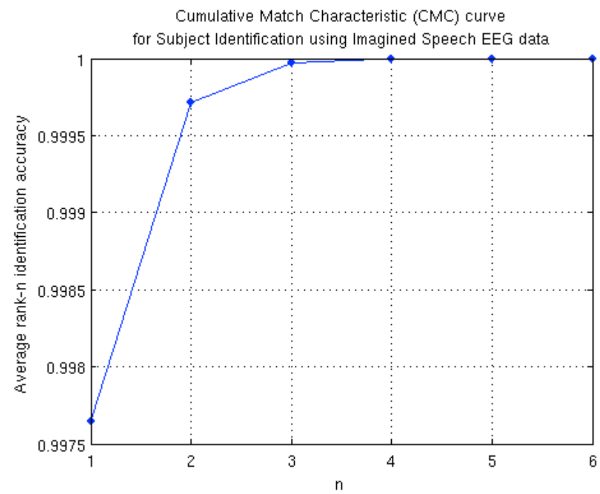

Fig. 3. Plot of the average rank-n identification accuracies using a linear SVM classifier for the imagined speech EEG dataset.

A k-NN classifier was also tested for different values of k for comparison purposes, and the results are summarized in Table III. The corresponding Cumulative Match Characteristic (CMC) curve for the 1-Nearest Neighbor classifier is shown in Fig. 4.

TABLE III

Average rank-1 identification accuracies and corresponding 95% confidence intervals for the Imagined Speech EEG dataset using a k-NN classifier

EEG Dataset	*k*-NN	Average Classification Rate ± 95% confidence interval
Imagined Speech (6 subjects)	1	**99.41% ± 0.05%**
	5	**99.11% ± 0.04%**
	10	**98.86% ± 0.03%**
	15	**98.64% ± 0.04%**
	20	**98.52% ± 0.06%**

As can be seen, the best identification accuracy for the imagined speech EEG dataset was obtained using the linear SVM classifier for an average classification rate of 99.76%.

Classification performance was also tested with a linear SVM classifier for each rhythm separately (with both /ba/ and /ku/ contained within each rhythm), and for each syllable separately (with all 3 rhythms). The results are shown in Table IV and Table V.

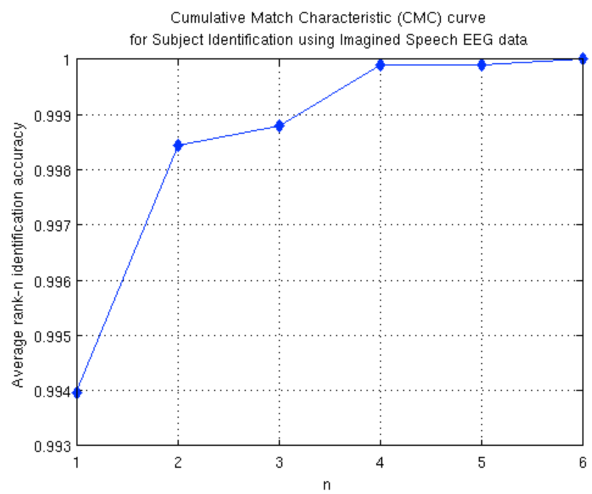

Fig. 4. Plot of the average rank-n identification accuracies using the imagined speech EEG dataset.

TABLE IV
AVERAGE RANK-1 IDENTIFICATION ACCURACIES AND CORRESPONDING 95% CONFIDENCE INTERVALS FOR THE IMAGINED SPEECH EEG DATASET FOR DIFFERENT RHYTHMS (USING BOTH /BA/ AND /KU/)

EEG Dataset	**Rhythm**	**Average Classification Rate ± 95% confidence interval**
Imagined Speech	1	**99.67% ± 0.09%**
	2	**99.41% ± 0.11%**
	3	**99.47% ± 0.10%**

TABLE V
AVERAGE RANK-1 IDENTIFICATION ACCURACIES AND CORRESPONDING 95% CONFIDENCE INTERVALS FOR THE IMAGINED SPEECH EEG DATASET FOR DIFFERENT SYLLABLES (USING ALL 3 RHYTHMS)

EEG Dataset	**Syllable**	**Average Classification Rate ± 95% confidence interval**
Imagined Speech	/ba/	**99.49% ± 0.07%**
	/ku/	**99.73% ± 0.08%**

The classification performance does not change much when using only the same rhythms or the same syllables, although the identification accuracies are slightly lower. These results suggest the proposed biometric method works for different rhythms and different syllables.

Experiments were also run to investigate the session-to-session variability by using one session as the training dataset, and another session as the testing dataset. Since the number of sessions varied for each subject (ranging from 4 to 10), shown in Table VI are the results from comparing only the first 4 sessions for each subject.

In Table VI, session 1 corresponds to each subject's respective first session, session 2 to each subject's respective second session, and so on. The results shown in Table VI imply that the imagined speech EEG data is most likely not fully stationary with respect to time. For example, when using the first session from each subject to predict a subject's identity in the subsequent sessions 2, 3, and 4, the classification rates decrease as the sessions grow farther apart in time. It also appears that using future sessions to predict previous sessions results in even lower classification rates, with the exception of sessions 3 and 4, where almost all subjects had both their third and fourth sessions recorded on the same day (hence the higher classification rates). These results indicate that when performing subject identification with EEG data, it may be desirable to either update the training database with current data after identification, or to use some adaptive classification techniques to account for the non-stationarity of the EEG data.

TABLE VI
RANK-1 IDENTIFICATION ACCURACIES FOR THE IMAGINED SPEECH EEG DATASET, WHEN TRAINING AND TESTING ON SEPARATE SESSIONS

		Testing Data			
	Session	**1**	**2**	**3**	**4**
Training Data	**1**	**--**	**98.4%**	**86.0 %**	**84.9%**
	2	**97.0%**	**--**	**95.7 %**	**94.0%**
	3	**86.5%**	**90.7%**	**--**	**99.8%**
	4	**78.6%**	**89.3%**	**99.8%**	**--**

It is also noted that the subject identification could have performed well due to differing head shapes (and therefore, potentially different electrode placements on each subject resulting in unequal electrode comparisons). However, EEG is known to have poor spatial resolution so adjacent electrodes tend to be highly correlated for the 128 Channel Sensor Net, and furthermore, subject classification was performed using a large number of trials that were recorded in sessions that took place at different times (and from session to session, even for one subject the electrodes would not be placed in the same exact locations). Therefore, it is unlikely that the classification benefited from differences in electrode placement.

Subject classification was also performed using the VEP EEG dataset with an AR model order of 4, and the results from using the linear SVM classifier are shown in Table VII with the corresponding CMC curve plotted in Fig. 5. Again, for comparison purposes subject identification was also performed using a *k*-NN classifier for the VEP dataset. These results are given in Table VIII, and the corresponding Cumulative Match Characteristic (CMC) curve for the 1-Nearest Neighbor classifier is shown in Fig. 6.

TABLE VII
AVERAGE RANK-1 IDENTIFICATION ACCURACIES AND CORRESPONDING 95% CONFIDENCE INTERVALS FOR THE VEP EEG DATASET USING A LINEAR SVM CLASSIFIER

EEG Dataset	**Average Classification Rate ± 95% confidence interval**
VEP (120 subjects)	**98.96% ± 0.05%**

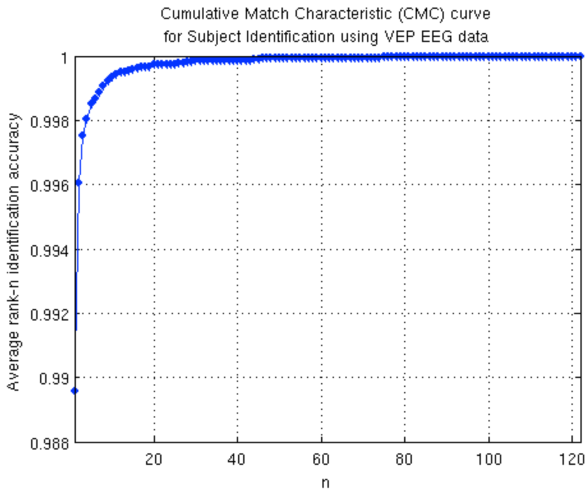

Fig. 5. Plot of the average rank-n identification accuracies using the VEP EEG dataset with a linear SVM classifier.

TABLE VIII
AVERAGE RANK-1 IDENTIFICATION ACCURACIES AND CORRESPONDING 95% CONFIDENCE INTERVALS FOR THE VEP EEG DATASET USING A K-NN CLASSIFIER

EEG Dataset	*k*-NN	Average Classification Rate ± 95% confidence interval
VEP (120 subjects)	1	**96.34% ± 0.11%**
	5	**96.17% ± 0.19%**
	10	**95.21% ± 0.15%**
	15	**94.14% ± 0.18%**
	20	**93.24% ± 0.09%**

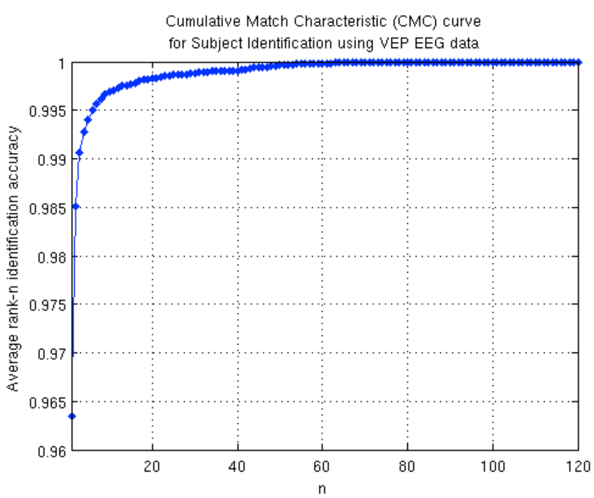

Fig. 6. Plot of the average rank-n identification accuracies using the VEP EEG dataset with a k-NN classifier.

The performance of the linear SVM classifier is still superior to the *k*-NN classifier for the VEP EEG dataset as it was with the imagined speech EEG dataset.

These results show that the algorithm presented here can be used to identify a larger number of subjects as well. The best results seen thus far using the same VEP EEG database (with nearly all of the available subjects) were presented in [20]. Palaniappan and Mandic [20] were able to attain 98.12% classification accuracy (with a standard deviation of 1.26%) with data from 102 subjects and 3,950 trials. The method proposed here was tested with 120 subjects and 9,596 trials, and the maximum classification accuracy achieved was 98.96% ± 0.05%, which demonstrates the general applicability of this approach for subject identification with EEG data.

VI. Conclusions

While electrical brainwave activity as recorded by EEG is not typically expected to be the same for different subjects even under the same conditions (since the anatomical layouts of peoples' brains may differ), subjects can actually be identified with EEG data to a high degree of accuracy. However, when using EEG signals for decoding imagined speech, one of the complications is that it is currently unknown if and how imagined speech and its characteristics (e.g., pitch, volume, articulation) manifest in EEG data, and what kind of impact each subject's particular variation of covert speech may have on these signals. These imagined speech characteristics could vary from trial to trial, which may or may not result in EEG signals that provide sufficient information for any kind of classification. However, despite this, in using the proposed approach with the imagined speech EEG dataset, we were still able to use all of the trials even under varying conditions (e.g., different syllables, different rhythms, different sessions at different times) to accurately identify each subject, which has potential applications such as increased security for covert communication. Using imagined speech EEG data as a biometric has its advantages; it is seemingly unaffected by different variations of imagined speech, and it is relatively easy to do (i.e., it does not seem to require training). Being able to accurately identify subjects using imagined speech is an interesting result, as it seems to suggest that people may have unique brain patterns in simply the way that they think.

Acknowledgments

We thank Dr. Michael D'Zmura, Dr. Ramesh Srinivasan, Siyi Deng, Tom Lappas, and Samuel Thorpe of UCI for collecting and providing us with the imagined speech EEG data, and for their guidance in processing the EEG data. We would also like to thank Professor Henri Begleiter from the Neurodynamics Laboratory at the State University of New York Health Center in Brooklyn, USA for providing the VEP EEG data.

References

[1] J. Wolpaw, N. Birbaumer, D. J. McFarland, G. Pfurtscheller, T. M. Vaughan, "Brain-computer interfaces for communication and control," *Clinical Neurophysiology*, vol. 113, pp. 767-791, 2002.
[2] *MURI: Synthetic Telepathy*. (2009). [Online]. Available: http://cnslab.ss.uci.edu/muri/index.html
[3] M. D'Zmura, S. Deng, T. Lappas, S. Thorpe, and R. Srinivasan. "Toward EEG sensing of imagined speech," *Human Computer Interactions and New Trends*, pp. 40-48, 2009.

[4] K. Brigham, B. V. K. Vijaya Kumar, "Imagined speech classification with EEG signals for silent communication: a preliminary investigation into synthetic telepathy," *4th International Conference on Bioinformatics and Biomedical Engineering*, 2010, to be published.

[5] F. Guenther, J. S. Brumberg, E. J. Wright, A. N. Castanon, J. A. Tourville, M. Panko, et al., "A wireless brain-machine interface for real-time speech synthesis," *PLoS ONE* 4(12): e8218, vol. 4, no. 12, Dec. 2009.

[6] C. S. DaSalla, H. Kambara, M. Sato, Y. Koiek, "Single-trial classification of vowel speech imagery using common spatial patterns", *Neural Networks*, vol. 22, no. 9, pp. 1334-1339, November 2009.

[7] S. Marcel and J. D. R. Millan, "Person authentication using brainwaves (EEG) and maximum a posteriori model adaptation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 29, no. 4, pp.743-752, April 2007.

[8] R. B. Paranjape, J. Mahovsky, L. Benedicenti, and Z. Koles, "The electroencephalogram as a biometric," *Proc. Canadian Conf. Electrical and Computer Eng.*, vol. 2, pp. 1363-1366, 2001.

[9] M. Poulos, M. Rangoussi, V. Chrissikopoulos, and A. Evangelou, "Person identification based on parametric processing of the EEG," *Electronics, Circuits and Systems, 1999. Proceedings of ICECS '99. The 6th IEEE International Conference on*, vol.1, pp.283-286, 1999.

[10] R. Palaniappan, "Two-stage biometric authentication method using thought activity brain waves," *International Journal of Neural Systems*, vol. 18, no. 1, pp. 59-66, 2008.

[11] Z. A. Keirn, J. I. Aunon. "A new mode of communication between man and his surroundings". *IEEE Transactions on Biomedical Engineering*, 37(12):1209–1214, December 1990.

[12] X. Bao, J. Wang, J. Hu, "Method of individual identification based on electroencephalogram analysis," *New Trends in Information and Service Science, 2009. NISS '09. International Conference on*, pp. 390-393, 2009.

[13] *BCI Competition III.* (2003). [Online]. Available: http://www.bbci.de/competition/iii/

[14] J. F. Hu, "New biometric approach based on motor imagery EEG signals," *BioMedical Information Engineering, 2009. FBIE 2009. International Conference on Future*, pp. 94-97, 13-14 Dec. 2009.

[15] A. Zúquete, B. Quintela, J. P. S. Cunha, "Biometric authentication using brain responses to visual stimuli," *Proceedings of the International Conference on Bio-inspired Systems and Signal Processing*, pp. 103-112, 2010, to be published.

[16] K. Das, S. Zhang, B. Giesbrecht, and M. P. Eckstein, "Using rapid visually evoked EEG activity for person identification," *Engineering in Medicine and Biology Society, 2009. EMBC 2009. Annual International Conference of the IEEE,* pp.2490-2493, 3-6 Sept. 2009.

[17] R. Palaniappan and D. P. Mandic, "EEG based biometric framework for automatic identity verification," *VLSI Signal Processing*, vol. 49, no. 2, pp. 243-250, 2007.

[18] D. L. Davies, D. W. Bouldin, "A cluster separation measure," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. PAMI-1, no. 2, pp 224-227, 1979.

[19] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," *Antennas and Propagation, IEEE Transactions on*, vol. 34, no. 3, pp. 276-280, 1986.

[20] R. Palaniappan and D. P. Mandic, "Biometrics from brain electrical activity: a machine learning approach," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 29, no. 4, pp. 738-742, 2007.

[21] S. Deng, R. Srinivasan, T. Lappas, and M. D'Zmura, "EEG classification of imagined syllable rhythm using Hilbert spectrum methods," *J. Neural Eng.*, vol. 7, no. 4, 2010.

[22] *EEG Database*. (1999). [Online]. Available: http://kdd.ics.uci.edu/databases/eeg/eeg.html

[23] J. G. Snodgrass and M. Vanderwart, "A standardized set of 260 pictures: norms for naming agreement, familiarity, and visual complexity," *J. Exp. Psychol: Human Learning and Memory*, vol. 6, pp. 174-215, 1980.

[24] X. L. Zhang, H. Begleiter, B. Porjesz, W. Wang, and A. Litke, "Event related potentials during object recognition tasks," *Brain Research Bulletin*, vol. 38, no. 6, pp. 531-538, 1995.

[25] *EGI*. (2008). [Online]. Available: http://www.egi.com/index.php

[26] J. S. Lim and A. V. Oppenheim. *Advanced Topics in Signal Processing*. Englewood Cliffs, NJ: Prentice Hall, pp. 87-89, 1988.

[27] S. M. Kay. *Modern Spectral Estimation: Theory and Application.* Englewood Cliffs, NJ: Prentice Hall, pp. 228-230, 1988.

[28] Chih-Chung Chang and Chih-Jen Lin, "LIBSVM: a library for support vector machines", 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm