Mel Frequency Cepstral Coefficients Enhance Imagined Speech Decoding Accuracy from EEG

Ciaran Cooney
Intelligent Systems Research Centre
Ulster University
Derry/Londonderry, UK
cooney-c@ulster.ac.uk

Rafaella Folli
Institute for Research in Social Sciences
Ulster University
Jordanstown, UK
r.folli@ulster.ac.uk

Damien Coyle
Intelligent Systems Research Centre
Ulster University
Derry/Londonderry, UK
dh.coyle@ulster.ac.uk

***Abstract*— Imagined speech has recently become an important neuro-paradigm in the field of brain-computer interface (BCI) research. Electroencephalogram (EEG) recordings during imagined speech production are difficult to decode accurately, due to factors such as weak neural correlates and spatial specificity, and signal noise during the recording process. In this study, a dataset of imagined speech recordings obtained during production of eleven different units of imagined speech is used to investigate the relative effects of different features on classification accuracy. Three distinct feature-sets are computed from the data: a linear feature-set, a non-linear feature-set, and a feature-set comprised only of mel frequency cepstral coefficients (MFCC). Each feature-set is used to train a decision tree classifier and a Support Vector Machine classifier. The results indicate that the use of MFCC features provides greater discrimination of imagined speech EEG recordings in comparison with the other features evaluated, and that phonological differences between imagined words can serve as an aid to classification.**

Keywords*— EEG, imagined speech, brain-computer interface, decision tree, support vector machine, mel frequency cepstral coefficients

I. Introduction

Many definitions for imagined speech are present in the literature [1], [2], one of which describes it as internalised, inaudible verbal thought that may or may not reach conscious awareness and may or may not be accompanied by subliminal vocal activity [3]. Related terminology for imagined speech includes self-talk, sub-vocal/covert speech, internal dialogue/monologue, sub-vocalisation, utterance, self-verbalisation, auditory imagery and self-statement [4]. Imagined speech as a neuro-paradigm for communicative BCI has been gaining momentum, with recent developments in the field [5], [6] showing that it may have the potential to improve upon the utility of existing approaches such as the BCI-speller [7]. This kind of Direct-Speech BCI (DS-BCI) [8] offers users the possibility of a naturalistic mode of communication, as neural correlates of imagined speech are becoming a targeted BCI challenge to be addressed. This is in contrast with typical communicative BCI approaches in which some form of modulated brain activity unrelated to speech is harnessed as the modality to relay a BCI users' intended action [9].

However, neural recordings corresponding to imagined speech are extremely challenging to decode and require sophisticated signal processing approaches to obtain sufficient information for effective classification. This problem is amplified when non-invasive recording techniques such as electroencephalography (EEG), and their associated low signal-to-noise ratio, are utilised to determine the users' intent. The difficulty of effectively decoding units of imagined speech from EEG recordings is a constraining factor in progress towards development of a DS-BCI. Therefore, evaluation of feature extraction and selection is of paramount importance for such a system.

In this work, different features and combinations of features, extracted from data recorded whilst fourteen subjects perform imagined speech, are evaluated using two different classifiers. The first of these feature-sets contains several linear time-domain features, including the mean, variance and standard deviation of the signal. The second set contains six non-linear features, including fractal dimension and spectral entropy. The third set of features is derived solely from mel frequency cepstral coefficients (MFCC). MFCCs are based on human hearing perceptions and were primarily developed for use in automatic speech recognition systems [10]. However, they have also been found to be useful in decoding EEG signals for BCI applications [11]. Each feature-set was used to train two different classifiers: a decision tree and a Support Vector Machine (SVM).

The following sections describe the process followed and results obtained in this study. Section II describes the methodology used to acquire and process the data, as well as the feature extraction methods pursued. In Section III the approach to classification is described and the initial parameters of each classifier documented. Sections IV and V present the results and concluding remarks on the findings.

II. Methodology

A. Data Acquisition

The data used in this study were obtained from the KARA ONE database containing EEG data relating to phonological categories in imagined and articulated speech [10]. The data were acquired at the Toronto Rehabilitation Institute and has been made freely available by the University of Toronto here: (http://www.cs.toronto.edu/~complingweb/data/karaOne/karaOne.html). The complete dataset is comprised of three distinct modalities, EEG, face tracking and audio, but for the purposes of this study only the EEG data were required. All EEG data were acquired with a 64-channel Neuroscan Quick-cap

978-1-5386-6046-1/18/$31.00 ©2018 IEEE

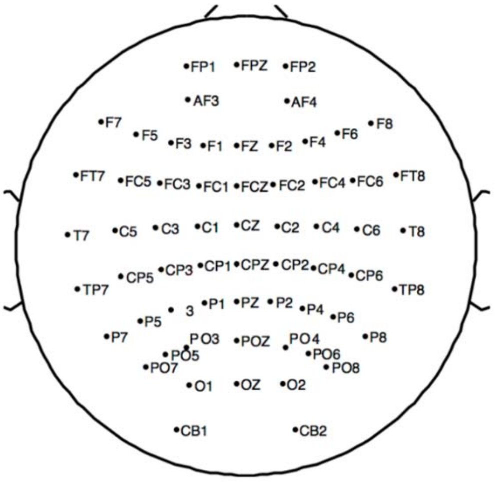


Fig. 1. 62-channel EEG montage using the 10-20 systems.

using the 10-20 system for electrode placement (Fig. 1). A SynAmps RT amplifier was used to record the signals with a sampling rate of 1 KHz.

The experiment conducted by Zhao & Rudzicz [10] required participants to respond to prompts, presented both textually and acoustically, by first reproducing the prompt in an imagined speech state and then in an overt speech state. Eleven prompts were used, subcategorised as seven phonemic/syllabic prompts (*/iy/, /uw/, /piy/, /tuy/, /diy/, /m/, /n/*) and two phonetically-similar pairs of words (*pat, pot, knew* and *gnaw*). Prompts were presented in categorical order, with individual prompts randomly permuted within each category. For each trial, data were recorded from fourteen participants during four distinct states of activity:

1. A 5-second rest state.
2. A stimulus state, where a text prompt would appear on-screen and a corresponding auditory prompt would be played.
3. A 5-second imagined speech state.
4. A speaking state, in which the participant spoke the prompt aloud.

EEG data were recorded during all stages of each trial, with a Kinect sensor being used to capture audio and facial features during the speaking state only. As the subject of the present study is to decode imagined speech content directly from EEG recordings, only the 5-second epochs corresponding to each imagined speech state (state 3) have been extracted from the complete dataset. Of the 14 participants included in the dataset, 3 completed different numbers of trials to the other 11. Therefore, the data from those participants have been excluded from this analysis. Each of the 11 prompts were presented 12 times, resulting in a total of 132 trials per participant.

B. *Preprocessing*

Preprocessing of the EEG data was conducted using the EEGLAB [11] toolbox in MATLAB®. The extracted EEG signals were filtered between 1 and 50 Hz and a small Laplacian filter was also applied to each channel. Each trial epoch, consisting of approximately 5000 samples, was windowed to 500ms frames with a 250ms overlap between frames. Due to small sampling errors resulting in several epochs not consisting of precisely 5000 samples the final window from each epoch was removed to ensure uniform dimensionality across subjects and trials. The first window has also been removed from each trial as this data corresponds to stimulus response rather than imagined speech production.

Independent Component Analysis (ICA) was performed on the dataset to compute signal components that are mutually independent. ICA facilitates extraction of independent components from mixed signals by transforming a multivariate random signal [12]. This approach is effective for the removal of noise from EEG data and is therefore an aid to classification.

C. *Feature Extraction*

In order to maximize the performance of a classifier employed to discriminate between classes of EEG recordings, it is necessary to extract features which accurately describe the information in the data. Typically, there are three types of features associated with decoding approaches to EEG: time-domain features, frequency-domain features and spatial features. In this study, focus has been directed towards the efficacy of time-domain and frequency-domain features in the form of linear, non-linear and mfcc featuresets. Features were calculated for each window in the EEG dataset and for each of the 62 channels within those windows.

1) Linear features: Twelve different time-domain features were computed on the windowed EEG data. Time-domain features calculated for the linear dataset are: *mean, absolute mean, standard deviation, sum, median, variance, maximum, absolute maximum, minimum, absolute minimum, maximum+minimum* and *maximum-minimum*. All twelve linear features are computed for each of the 62 channels in each of the 17 data windows and combined into a feature-set containing 12,648 features.

2) Non-linear features: Six non-linear features were computed, not including MFCC features. The frequency-based features in this set are utilised to obtain a transformed representation of the EEG signals in the frequency domain. The six non-linear features computed are: *Hurst Exponent*, *Higuchi's Algorithm of Fractal Dimension*, *Spectral Power*, *Spectral Entropy*, *Magnitude* and *Phase*. Each of the six non-linear features are briefly decribed below:

a) The Hurst Exponent: Otherwise know as Rescaled Range Statistics (R/S), it is a measure of the correlation of the points in a time-series. A Hurst Exponent value greater than 0.5 indicates long range correlations in a given window, whereas a value less than 0.5 indicates long range anticorrelations [13]. It is computed by calculating the accumulated deviation from the mean of the time-series *x(t)* over time *T* such that:

$$X(t,T) = \sum_{i=1}^{t} x(i) - \bar{x} \; where \; \bar{x} = \frac{1}{T}\sum_{i=1}^{t} x(t) \qquad (1)$$

R(T) is then calculated as the difference between the maximum and minimum value of *X(t,T)* and *S(T)* calculated as the standard deviation of time series over time *T*:

$$\frac{R(T)}{S(T)} = \frac{(\max(X(t,T)) - \min(X(t,T)))}{\sqrt{\frac{1}{T}\sum_{t=1}^{T}(x(t) - \bar{x})}^{2}} \tag{2}$$

The Hurst Exponent is obtained by plotting the log[R(n)/S(n)] as a function of log n [14].

b) Fractal Dimension: Calculated using Higuchi's algorithm [15], fractal dimension provides a measure of the complexity of the EEG signal. A Fractal is a shape that retains its structural dimension when scaling and it is included as a non-linear feature due to its relationship to spectral shape. The method generates an estimate of curve length at different scale values q, as follows [16]:

$$C_m(q) = \frac{(N-1)}{\lfloor\frac{N-m}{q}\rfloor q^2} \sum_{i=1}^{\lfloor(\mathrm{N}-\frac{m}{q})\rfloor} |x[m+iq] - x[m+(i-1)q]| \tag{3}$$

The term, $N - 1/\lfloor N - m/q \rfloor q\hat{}2$ represents the normalization factor for the curve length of subset time series. The length of the curve is defined for the time interval q, $C(q)$, as the average value over q sets of C_m(q). If $C(q) \propto q{-}D$, then the curve is fractal with the dimension D.

c) Spectal Entropy: Shannon entropy is ustilised to compute the spectral entropy of the time-domain EEG. The *fast fourier transform* is first computed for a data window and the power spectral density obtained using the periodogram method*:*

$$\bar{P}[k] = \frac{1}{Nf_s}\left|\sum_{N=0}^{N-1} x[n]e^{-j2\pi kn/N}\right|^2 \tag{4}$$

for EEG signal $x[n]$, of length N, where fs is sampling frequency. Spectral entropy is then computed as follows:

$$\bar{E} = -\frac{1}{logl_i}\sum_{k=a_i}^{b_i} \bar{P}_i\,[k]log\bar{P}_i[k] \tag{5}$$

where l is the length of the sequence representing the range of the frequency band.

d) Spectral Power: The spectral power was computed by taking the periodogram (4) and applying the formula:

$$Pow = \frac{s[k]f_s}{N}\sum_{k=0}^{b_i}\bar{P}[k] \tag{6}$$

where $s[k]$ is a scaling factor applied to conserve total power in the spectrum when using only positive frequencies.

e) Magnitude and Phase: The magnitude and phase of the EEG signal was computed by implementing the Hilbert transform to convert the time-domain sequence into a complex time sequence [17] and calculating the mean phase and magnitude for each window. The Hilbert transform is based on the following formula:

$$Y(t) = H(x(t)) = \int_{-\infty}^{+\infty} x(\tau) * \frac{1}{t-\tau}d\tau \tag{7}$$

As with the time-domain features, each non-linear feature is calculated for each of the 62 channels within each of the 17 data windows, and combined into a single feature-set, with a total of 6,324 features.

f) Mel Frequency Cepstral Coefficients: MFCCs have been heavily utilised as features in automatic speech recognition [18] and there has been an increase in their use in EEG-based BCI applications [19], including attempts to classify imagined speech [20]. Thirteen cepstral coefficients were obtained for each window in the EEG dataset from a filterbank consisting of nine filters. The input EEG signal is first transformed into the frequency domain using the *fast fourier transform* and then applied to a bank of triangular filters to compute a weighted sum of filter spectral components approximating a Mel scale [18]. The MFCCs are obtained by converting the log Mel spectrum into the time domain using the Discrete Cosine Transform:

$$C_i = \sqrt{\frac{2}{N}}\sum_{j=1}^{n} m_j cos\left(\frac{\pi i}{N}(j-0.5)\right) \tag{8}$$

where N is the number of filterbank channels. Each of the 13 MFCCs calculated for all 62 channels and all 17 data windows results in a total of 13,702 features which are used as input training features to the classification models.

D. Prinicpal Component Analysis

The features detailed in the previous section were calculated independently for each participant, thus allowing comparison of classifier performance across subjects. Variability across participants is often quite high in BCI applications so it can be informative to evaluate results independently. Before training, Principal Component Analysis (PCA) was applied to the data set to reduce dimensionality and identify the components with greatest variance. Dimensionality reduction is particularly important when used with computationally-expensive training algorithms, such as multi-class SVM. The PCA algorithm extracts the components from a dataset which are most responsible for the variance in that data. The first principal component contains the greatest variance, with the second containing less, and so on. Typical approaches include retention of the first k principal components (where k = 1,2,3 etc.), or retention of k number of components such that a specified fraction of the total variance is explained. In this case, due to the variability of explained variance across participants, the number of components corresponding to 95% of the total variance have been retained for classifier training.

III. Classification

Two classifiers have been trained for each participant to obtain classification accuracy for imagined speech trials from EEG. Three primary tests have been performed for each participant and for each classifier. First of these was to train and test each of the classifiers on the linear features, e.g. mean, standard deviation, etc. The second test was performed on the dataset with non-linear features only. The third test was to train the classifiers on the MFCC features computed. The three separate approaches to training facilitate comparison of the impact of those features on imagined speech classification.

A. 5-fold Cross-Validation

A *k*-fold cross-validation approach to splitting data into training and test sets provides a much more robust estimate of a classifier's ability to generalize than more basic validation techniques. A 5-fold cross-validation scheme was selected to ensure that robust estimates of classification accuracy have been obtained. EEG signals were randomly divided into 5 sets, 4 of which were used for training the classifier. The other set would then be used as a test-set. This process is repeated through a total of 5 iterations, with each classifier's accuracy retained to compute an average value for the final accuracy.

B. Decision Tree

The first classifier trained on the imagined speech EEG data was a decision tree. A decision tree is a non-parametric supervised-learning technique which enables transparency in the model obtained and reduces decision-making ambiguity in comparison with some other methods. Decision trees have also been associated with overfitting, meaning that overly-complex trees may not generalize well to new data. Decision tree classifiers are typically initialized with one of two parameters for measuring the quality of a split: *Gini Impurity* and *Information Gain Entropy*. In this study, Gini Impurity has been selected as the splitting criteria due to it being less computationally intensive than Information Gain Entropy. A second important parameter required when initializing a decision tree is the maximum number of splits/leaf nodes. There is an implicit trade-off in performance associated with this parameter, as a relatively small value for maximum number of splits will require less computational resource but result in poor model performance. Empirical study of the classifier's performance when initialized with several different values for maximum number of splits led to the selection of 600 for this parameter. The classifier demonstrated good performance increase up to this number, with plateauing of performance beyond.

C. Support-Vector Machine

The second classifier to be trained on the EEG dataset was the SVM. The SVM classifier has often been employed in research relating to DS-BCI, with some promising results [10], [5]. There are many possible configurations of an SVM classifier, particularly when faced with a multiclass problem, as in this case. There are several possible kernels which may be utilized in the algorithm, including radial basis function and sigmoid. The SVM classifier employed in this study was initialized with a linear kernel. The multiclass SVM must also be selected to apply a one-vs-one or one-vs-all training algorithm. These methods determine the number of classifiers that must be trained and effect the decision boundary they compute. Here, we have selected a one-vs-all SVM, which is the less computationally expensive of the two approaches.

IV. Results & Discussion

The results obtained from the 5-fold cross-validated training models indicate that the MFCC features provide stronger discrimination between imagined speech EEG signals than do either combination of linear or non-linear features. Figures 2A and 2B present the results obtained from the two trained models for all participants and all features-sets. Improved performance when using MFCC features is seen for both the decision tree and SVM classifiers, with the effect on the classification accuracy of the SVM particularly strong. In Fig. 2A, classification accuracies resulting from training with a decision tree on linear, non-linear, and MFCC features are presented. Average accuracy for each of the three feature-sets is greater than chance level accuracy (9.09% for 11 classes) with MFCC exhibiting the best performance with average accuracy of 19.69%. Interestingly, and perhaps unexpectedly, the classification accuracy resulting from use of the linear features is greater than that obtained from use of the non-linear features, 15.91% versus 14.67%. A t-test performed on these results indicates that the difference is statistically significant ($p < 0.05$). There are several possible reasons for this. One possibility is that the non-linear features forming the feature-set are not well-suited to the problem of classifying imagined speech from EEG recordings. Another is dimensionality. Prior to application of PCA, the time-domain feature-set is constructed of 12,648 (62 channels x 17 windows x 12 features) features, with the non-linear feature-set only made up of 6,324 (62x17x6) features. A t-test performed on the results from the SVM classifier indicated that the difference between the average accuracies of the linear and non-linear features was not statistically significant. Future work to ascertain which of these features are truly discriminative in this field is required.

Fig. 2B presents classification accuracies obtained from training a linear SVM classifier on each of the three feature-sets. Again, the MFCC feature-set produces the best performance from the classifier, with an average accuracy of 20.80%. However, the linear and non-linear features fail to produce average classification accuracies significantly above chance level. The results from both the decision tree and SVM classifiers clearly exhibit the relative performance improvement when MFCCs are used, with the SVM exhibiting an increase of almost 10% in classification accuracy. It is clear from the results presented in the bar-charts summarizing results in Figure 2 that there is substantial variation in classifier performance among participants, as well as feature-sets. This variation is clearly visible in Fig. 2A, where it can be seen that participant number 2 reaches a high classification accuracy above 30% while participants 3, 9 and 10 fail to produce particularly strong classification accuracies with any of the three feature-sets. Inter-participant variation in performance is a common issue in BCI applications, and imagined speech, as the mode of communication, is no different. In fact, given current understanding of the phenomenology of imagined speech,

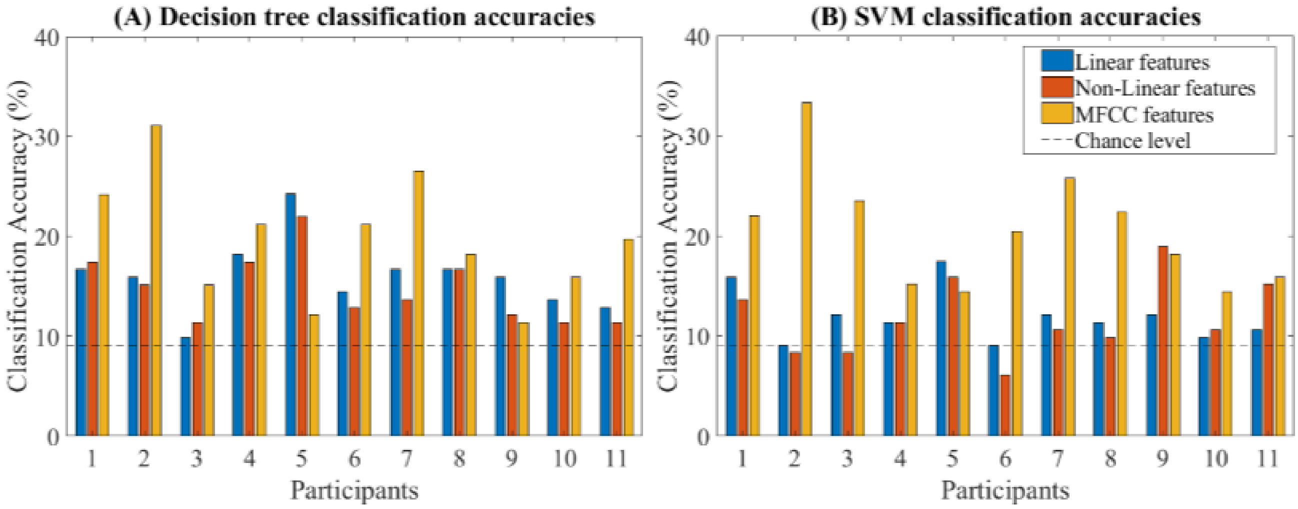


Fig. 2. Classification accuracy of imagined speech EEG signals using different feature-sets with decision tree and SVM classifiers.

this variation may even be exacerbated. Thus, future BCI experiments involving imagined speech must be rigorously designed and all participants effectively prepared to minimize this issue. The peak classification accuracies of participant 2 in particular (decision tree: 31.06%; SVM: 33.33%), indicate that there is some potential to accurately classify units of imagined speech directly from EEG activity.

Pairwise t-tests were performed on each pair of feature-sets and across classifiers. As stated above, the difference in classification accuracy between the linear and non-linear feature-sets when applied to the decision tree classifier was statistically significant ($p < 0.05$). The pairwise t-tests conducted between the linear features and MFCCs, and the non-linear features and MFCCs, when used with the decision tree classifier, both showed statistical significance with p-values of $p < 0.01$ and $p < 0.05$ respectively. Statistical significance was not apparent when the t-test was applied to the linear and non-linear features classified by the SVM. However, t-tests conducted between the linear features and MFCCs, and the non-linear features and MFCCs, when used with the SVM, indicated statistical significance with p-values of $p < 0.005$ and $p < 0.01$. The t-tests conducted across the two classifiers indicate that the difference in classification accuracy between the two when using the MFCC feature-set is not statistically significant ($p > 0.5$). This was also the case when the t-test was performed on the non-linear features across the two classifiers ($p > 0.05$). However, the result of the t-test conducted between the linear features across both classifiers indicated statistical significance with $p < 0.001$.

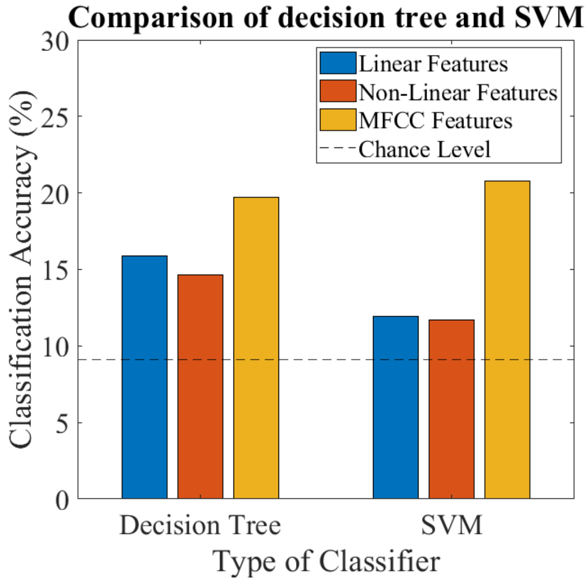
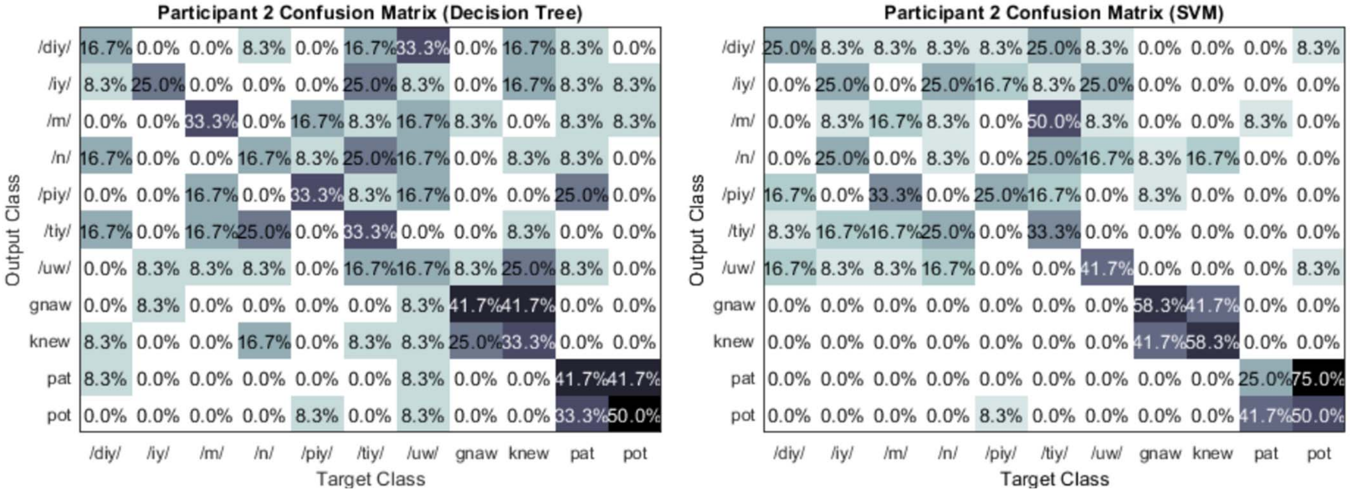


Fig. 3. Classification accuracies for each feature-set using decision tree and SVM.

The relative performance of the decision tree and SVM classifiers is presented in Fig. 3. When trained on the MFCC featureset, the two classifiers exhibit similar performance, with average accuracies of 19.69% and 20.80% respectively. However, this is not the case in relation to the linear or non-linear features. When these features are used to train the two classifiers, it results in a much stronger performance from the decision tree. There is no current consensus in the literature on the best methods for decoding imagined speech from EEG. Determining the best candidates should therefore be the subject of future work in this field, including evaluation of more complex models using deep learning techniques.

The confusion matrices for participant 2, presented in Fig. 4, show that almost all of the eleven phonemes and words are classified above chance level (9.09%) and that the two phonological pairs of words (gnaw/knew, pat/pot) achieve an average accuracy of 44.79%. It is this classification performance in relation to the phonological pairs which is most striking when viewing the results in Fig. 4. Not only are these words classified with considerable accuracy, but the confusion matrices also indicate that, in general, they are being misclassified as their most phonologically similar words. This is particularly apparent from the SVM confusion matrix in Fig. 4, where all but one instance results in either a word being classified correctly or being misclassified as its phonologically- similar pair. These results suggest that with methodological improvements to experimental design, signal processing and machine learning there is potential for DS-BCI to yield enhanced performance.



Participant 2 Confusion Matrix (Decision Tree)

	/diy/	/iy/	/m/	/n/	/piy/	/tiy/	/uw/	gnaw	knew	pat	pot
/diy/	16.7%	0.0%	0.0%	8.3%	0.0%	16.7%	33.3%	0.0%	16.7%	8.3%	0.0%
/iy/	8.3%	25.0%	0.0%	0.0%	0.0%	25.0%	8.3%	0.0%	16.7%	8.3%	8.3%
/m/	0.0%	0.0%	33.3%	0.0%	16.7%	8.3%	16.7%	8.3%	0.0%	8.3%	8.3%
/n/	16.7%	0.0%	0.0%	16.7%	8.3%	25.0%	16.7%	0.0%	8.3%	8.3%	0.0%
/piy/	0.0%	0.0%	16.7%	0.0%	33.3%	8.3%	16.7%	0.0%	0.0%	25.0%	0.0%
/tiy/	16.7%	0.0%	16.7%	25.0%	0.0%	33.3%	0.0%	0.0%	8.3%	0.0%	0.0%
/uw/	0.0%	8.3%	8.3%	8.3%	0.0%	16.7%	16.7%	8.3%	25.0%	8.3%	0.0%
gnaw	0.0%	8.3%	0.0%	0.0%	0.0%	0.0%	8.3%	41.7%	41.7%	0.0%	0.0%
knew	8.3%	0.0%	0.0%	16.7%	0.0%	8.3%	8.3%	25.0%	33.3%	0.0%	0.0%
pat	8.3%	0.0%	0.0%	0.0%	0.0%	0.0%	8.3%	0.0%	0.0%	41.7%	41.7%
pot	0.0%	0.0%	0.0%	0.0%	8.3%	0.0%	8.3%	0.0%	0.0%	33.3%	50.0%

Participant 2 Confusion Matrix (SVM)

	/diy/	/iy/	/m/	/n/	/piy/	/tiy/	/uw/	gnaw	knew	pat	pot
/diy/	25.0%	8.3%	8.3%	8.3%	8.3%	25.0%	8.3%	0.0%	0.0%	0.0%	8.3%
/iy/	0.0%	25.0%	0.0%	25.0%	16.7%	8.3%	25.0%	0.0%	0.0%	0.0%	0.0%
/m/	0.0%	8.3%	16.7%	8.3%	0.0%	50.0%	8.3%	0.0%	0.0%	8.3%	0.0%
/n/	0.0%	25.0%	0.0%	8.3%	0.0%	25.0%	16.7%	8.3%	16.7%	0.0%	0.0%
/piy/	16.7%	0.0%	33.3%	0.0%	25.0%	16.7%	0.0%	8.3%	0.0%	0.0%	0.0%
/tiy/	8.3%	16.7%	16.7%	25.0%	0.0%	33.3%	0.0%	0.0%	0.0%	0.0%	0.0%
/uw/	16.7%	8.3%	8.3%	16.7%	0.0%	0.0%	41.7%	0.0%	0.0%	0.0%	8.3%
gnaw	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	58.3%	41.7%	0.0%	0.0%
knew	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	41.7%	58.3%	0.0%	0.0%
pat	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	25.0%	75.0%
pot	0.0%	0.0%	0.0%	0.0%	8.3%	0.0%	0.0%	0.0%	0.0%	41.7%	50.0%

Fig. 4. Confusion matrices for Participant 2 when MFCCs used as features for classification.

V. Conclusions

In this work, we investigated the effect of different types of features on the decoding accuracy of EEG recordings of imagined speech production. The EEG data was extracted from the KARAONE dataset and processed to obtain 17 500ms windows containing 62 channels of EEG data for each trial. Features were extracted from the data to obtain linear, non-linear and MFCC feature-sets. Each of the three feature-sets were used to train decision tree and SVM classifiers. The accuracies obtained from the 5-fold cross-validated models indicate that MFCCs are superior to the other features in discriminating between EEG recordings of imagined speech. This finding was consistent across both classifiers. The linear features performed better that the non-linear features across both classifiers and more work is required to understand why this was the case. Results also suggest that the phonological features of imagined words can aid decoding accuracy. Future work in this area will include filtering of the raw EEG data into distinct frequency bands to determine where the most discriminative information resides and to investigate the effect of spatial filtering on decoding accuracy. Evaluation of different classification methods is also an important area where further work is required.

Acknowledgment

The authors would like to acknowledge the KARAONE dataset's authors for providing the data used in this work. Ciaran Cooney is supported by a Department for the Economy studentship.

References

[1] C. Herff *et al.*, "Brain-to-text: Decoding spoken phrases from phone representations in the brain," *Front. Neurosci.*, vol. 9, no. JUN, pp. 1–11, 2015.

[2] B. Alderson-Day and C. Fernyhough, "Inner speech: Development, cognitive functions, phenomenology, and neurobiology.," *Psychol. Bull.*, vol. 141, no. 5, pp. 931–965, 2015.

[3] C. L. Marvel and J. E. Desmond, "From storage to manipulation: How the neural correlates of verbal working memory reflect varying demands on inner speech," *Brain Lang.*, vol. 120, no. 1, pp. 42–51, 2012.

[4] A. Morin and J. Michaud, "Self-awareness and the left inferior frontal gyrus: Inner speech use during self-related processing," *Brain Res. Bull.*, vol. 74, no. 6, pp. 387–396, Nov. 2007.

[5] S. Martin *et al.*, "Word pair classification during imagined speech using direct brain recordings," *Sci. Rep.*, vol. 6, no. 1, 2016.

[6] C. Herff, A. de Pesters, D. Heger, P. Brunner, G. Schalk, and T. Schultz, "Towards continuous speech recognition for {BCI}," *Brain-computer interface Res. a state-of-the-art Summ. 5*, pp. 1–9, 2017.

[7] X. Chen, Y. Wang, M. Nakanishi, X. Gao, T.-P. Jung, and S. Gao, "High-speed spelling with a noninvasive brain–computer interface," *Proc. Natl. Acad. Sci.*, vol. 112, no. 44, pp. E6058–E6067, 2015.

[8] O. Iljina *et al.*, "Neurolinguistic and machine-learning perspectives on direct speech BCIs for restoration of naturalistic communication," *Brain-Computer Interfaces*, vol. 4, no. 3, pp. 186–199, 2017.

[9] D. J. Krusienski *et al.*, "A Comparison of Classification Techniques for the P300 Speller," *J. Neural Eng.*, vol. 3(4), pp. 299–305, 2006.

[10] S. Zhao and F. Rudzicz, "Classifying phonological categories in imagined and articulated speech," *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, vol. 2015–Augus, pp. 992–996, 2015.

[11] A. Delorme and S. Makeig, "EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis," *J. Neurosci. Methods*, vol. 134, no. 1, pp. 9–21, 2004.

[12] A. Subasi and M. I. Gursoy, "EEG signal classification using PCA, ICA, LDA and support vector machines," *Expert Syst. Appl.*, vol. 37, no. 12, pp. 8659–8666, 2010.

[13] M. J. Aguila, H. D. V Basilio, P. V. C. Suarez, J. P. E. Dueñas, and S. V Prado, "Comparative Study of Linear and Nonlinear Features Used in Imagined Vowels Classification Using a Backpropagation Neural Network Classifier," *Proc. 7th Int. Conf. Biosci. Biochem. Bioinforma.*, pp. 7–11, 2017.

[14] T. Balli, R. Palaniappan, and A. N. Measures, "A Combined Linear & Nonlinear Approach for Classification of Epileptic EEG Signals," pp. 714–717, 2009.

[15] T. Higuchi, "Approach to an irregular time series on the basis of the fractal theory," *Phys. D Nonlinear Phenom.*, vol. 31, pp. 277–283, 1988.

[16] J. M. O. Toole and G. B. Boylan, "NEURAL: quantitative features for newborn EEG using Matlab," *arvix.org*, 2017.

[17] Y. Y. Wang, P. Wang, and Y. Yu, "Decoding English alphabet letters using EEG phase information," *Front. Neurosci.*, vol. 12, no. FEB, pp. 1–10, 2018.

[18] L. Muda, M. Begam, and I. Elamvazuthi, "Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques," vol. 2, no. 3, pp. 138–143, 2010.

[19] P. Nguyen, D. Tran, X. Huang, and W. Ma, "Age and Gender Classification Using EEG Paralinguistic Features," *2013 6th Int. IEEE/EMBS Conf. Neural Eng.*, pp. 1295–1298, 2013.

[20] A. Riaz, S. Akhtar, S. Iftikhar, A. A. Khan, and A. Salman, "Inter comparison of classification techniques for vowel speech imagery using EEG sensors," *2014 2nd Int. Conf. Syst. Informatics, ICSAI 2014*, no. Icsai, pp. 712–717, 2015.