# An EEG Study On The Brain Representations in Language Learning

**Akshara Soman, C R Madhavan, Kinsuk Sarkar, Sriram Ganapathy**

Learning and Extraction of Acoustic Patterns (LEAP) Lab,
Electrical Engineering, Indian Institute of Science, Bangalore-560012

E-mail: sriramg@iisc.ac.in

July 2018

**Abstract.** This paper presents an experimental study to understand the key differences in the neural representations when the subject is presented with speech signals of a known and an unknown language and to capture the evolution of neural responses in the brain for a language learning task. In this study, electroencephalography (EEG) signals were recorded while the human subjects listened to a given set of words from English (familiar language) and Japanese (unfamiliar language). The subjects also provided behavioral signals in the form of spoken audio for each input audio stimuli. In order to quantify the representation level differences for the auditory stimuli of two languages, we use a classification approach to discriminate the two languages from the EEG signal recorded during listening phase by designing an off-line classifier. These experiments reveal that the time-frequency features along with phase contain significant language discriminative information. The language discrimination is further confirmed with a second subsequent experiment involving Hindi (native language of the subjects) and Japanese (unknown language). A detailed analysis is performed on the recorded EEG signals and the audio signals to further understand the language learning processes. A pronunciation rating technique on the spoken audio data confirms the improvement of pronunciation over the course of trials for the Japanese language. Using single trial analysis, we find that the EEG representations also attain a level of consistency indicating a pattern formation. The brain regions responsible for language discrimination and learning are identified based on EEG channel locations and are found to be predominantly in the frontal region.

*Keywords*: Electroencephalography (EEG), Language identification, Language learning, Single-trial analysis.

## 1. Introduction

Speech is the easiest and the most effective way of communication used by humans. Humans are inherently capable of distinguishing between sounds from familiar and unfamiliar languages when they listen to them. Previous work has shown that humans can instantaneously differentiate while listening to songs from known and unknown languages [1]. Also, the studies on brain activations showed interesting differences in the areas of the brain that are activated when exposed to native and non-native languages [2].

With the use of function magnetic resonance imaging (fMRI), it was seen that cerebral activations in the brain are more pronounced when presented with foreign language compared to a known language [3]. Similarly, in speech production it was seen that the right frontal areas are more involved when the subject is attending to speak a new language. The activity in the right pre-frontal cortex was also found to be indicative of the language proficiency of the subject [4].

The difference in response of the human brain to known and unknown stimuli has been of significant interest to facilitate the full understanding of the auditory encoding processes. For example, a stronger P300 peak was observed in electroencephalogram (EEG) signals of the subject when presented with their own names compared to the peak values observed when other stimuli were presented [5]. For infants, the representation in EEG for familiar language and foreign language as well as for familiar and unfamiliar talker was analyzed in [6], where the delta and theta bands showed important differences.

In the case of learning, several studies have shown that with experience, we gain proficiency in an unknown language and the function and structure of the brain changes during this learning process [7, 8]. Similar to the task of musical training, the experience of learning a new language also includes changes in the brain states. The complexity of speech and language causes challenges in understanding the questions related to how, when and where these changes occur in the brain.

The attempts to answer these questions may also throw light on the fundamental questions of primary language acquisition in an infant [9]. With the exemption of a few studies that have attempted to quantify the anatomical changes in the brain during language learning [10, 11, 12], very little is known regarding the changes in the brain during a new language learning in terms of when these changes occur, and how they reflect in the learning. In this work, we attempt to quantify some of these questions at the representation level using electroencephalogram (EEG) recordings.

While the primary language learning for most adults happen at a very young age, the acquisition of a new language can happen at any point in the life time. For the language learning task, the age of acquisition showed little impact in terms of brain representations when normalized for the proficiency levels [13]. The fundamental question whether there is knowledge transfer from a known language to a new language is still open ended. Several studies have shown that the known languages play a key role in acquiring new languages. The first language was found to provide an understanding of the grammar [14, 15]. The popular hypothesis for a secondary language learning is the establishment of a link of the representations of the new language to the features of the already known language. Also, the continued exposure to the foreign language can help in learning the language faster [16, 17]. In the past, studies using MEG signals have shown that there are two major effects seen in the brain when the same words are presented repeatedly. In Repetitive Enhancement (RE), the frontal regions in the brain get activated when the same word from an unknown language is presented to the subject multiple times [18] after which the activations drop leading to Repetitive Suppression (RS). The RS is also observed when a word familiar to the subject is presented. These studies indicate that activations are seen till new brain connections are formed after which the intensity of the activations drop.

The task of learning a new language can be quite complicated in analysis. This can be done at multiple levels like phonemic, syllabic, word-level or sentence level. The evaluation of language learning can also be analyzed for multiple tasks like reading task, spontaneous speaking etc. In this study, we aim to understand the major differences in the brain representations at a word level from a familiar and an unfamiliar language. Additionally, we propose a method to perform trial level analysis to understand the changes in representation of words when the subject listens to words from an unfamiliar language.

We record EEG signals from the subjects when the subjects are presented with word segments from a familiar and an unfamiliar language. Along with EEG signals, we also record behavioral data from the subject where the subject reproduces the stimuli presented to him. The key findings from the work can be summarized as follows,

- With various feature level experiments, we identify that the time frequency representations (spectrogram) of EEG signals carry language discriminative information. These features are also verified for two separate tasks, English versus Japanese and Hindi (native language) versus Japanese.
- The brain regions that contain the most language discriminative information are in the frontal cortex and the temporal lobe (aligned with some of the previous fMRI studies [2]).
- It is seen that the inter-trial variations are more pronounced for the words from unfamiliar language than those from the familiar language in both EEG signals and spoken audio signals. Furthermore, the inter-trial variations in the spoken audio are correlated with those from the listening state EEG representations.
- The EEG signals for the Japanese stimuli are more correlated with the audio signal than those for the English stimuli indicating a higher level of attention to Japanese stimuli.

To the best of our knowledge, this study is one of the first of its kind to probe the linguistic differences in EEG level and in uncovering the language learning process from single trial EEG analysis.

The rest of the paper is organized as follows. In Sec. 2, we describe the data collection procedure along with the pre-processing steps used for EEG data preparation. The feature extraction of EEG signals and the classification between the two languages is described in Sec. 3. A similar analysis is done to extract features and to classify the spoken audio signals and this is also described in Sec. 3. The evidence of language learning is established in Sec. 4. The inter-trial analysis performed on the EEG and the audio signals is described in Sec. 4.2. The relationship between the EEG and the audio signals is analyzed and described in Sec. 4.2.2. In Sec. 5, we provide a discussion of the findings from this work and contrast it with previous studies. Finally, a summary of the paper is also provided in Sec. 6.

## 2. Materials and Methods

### *2.1. Subjects*

All the participants were Indian nationals with self-reported normal hearing and no history of neurological disorders. In the first experiment, English and Japanese language words were used while in the second subsequent experiment Hindi (native language) and Japanese language words were used. The first experiment had 12 subjects while the second experiment had 5 subjects. All the subjects in the first experiment setup had intermediate or higher level of English proficiency. In the English/Japanese experiments, six subjects were males (median age of 23.5) and six were females (median age of 24). The subjects were natives of south Indian languages or Hindi language. In the second experiment of Hindi versus Japanese, all the subjects were natives of Hindi language.

### *2.2. Experimental Paradigm*

Each block of the recording procedure consisted of five phases as illustrated in figure 1. The first phase was the rest period of 1.5s duration followed by a baseline period of 0.5s. The subjects were instructed to attentively listen to the audio signal played after the baseline period. Then, they were given a rest of 1.5s where they were encouraged to prepare for overt articulation of the stimuli. The last phase is the speaking phase where subject was asked to speak the word overtly. The spoken audio was recorded using a microphone placed about one foot from the subject. The subjects were alerted about the change in phase by the display of a visual cue in the center of the computer screen placed in front of them. The participants were asked to refrain from movement and to maintain visual fixation on the center of the computer screen in front of them. All subjects provided written informed consent to take part in the experiment. The Institute Human Ethical Committee of Indian Institute of Science, Bangalore approved all procedures of the experiment.

### *2.3. Stimuli*

In each experiment, the stimuli set contained words from 2 languages. The words were chosen such that they have uniform duration and speech unit variability. In the first experimental setup, the stimuli-set includes 12 English words and 12 Japanese words (Table 1). The duration of all audio stimuli ranges from 0.5s to 0.82s. In the second experimental setup, the stimuli-set includes 12 Hindi words (native language of the subject) and the same 12 Japanese words (Table 1)).

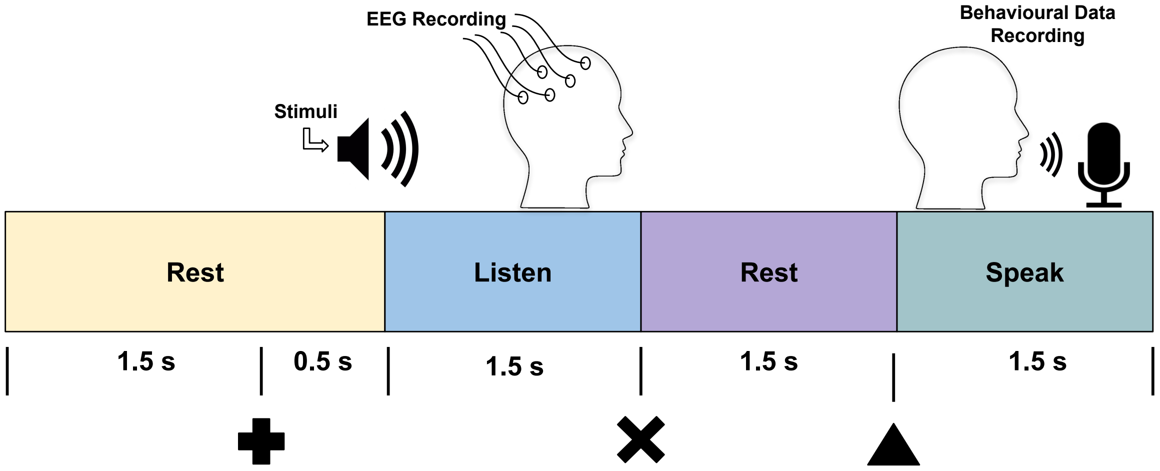


Figure 1: Experimental setup used for EEG and behavioral data collection

| English | | Japanese | | Hindi | |
|---|---|---|---|---|---|
| **Word** | **Duration (s) (# units)** | **Word** | **Duration (s) (# units)** | **Word** | **Duration (s) (# units)** |
| beg | 0.50 (3) | 南極 | 0.82 (4) | दरवासा | 0.73 (4) |
| cheek | 0.67 (3) | 抜き打ち | 0.83 (4) | चावल | 0.62 (3) |
| ditch | 0.70 (3) | 仏教 | 0.77 (3) | कहानी | 0.63 (3) |
| good | 0.50 (3) | 弁当 | 0.72 (3) | धन्यवाद | 0.88 (4) |
| late | 0.77 (3) | 偶数 | 0.76 (2) | आसमान | 0.80 (4) |
| luck | 0.64 (3) | 随筆 | 0.83 (3) | आदमी | 0.61 (4) |
| mess | 0.60 (3) | 先生 | 0.74 (4) | बचपन | 0.74 (4) |
| mop | 0.54 (3) | ポケット | 0.82 (3) | पुजारी | 0.74 (3) |
| road | 0.59 (3) | 計画 | 0.84 (4) | अलमारी | 0.72 (4) |
| search | 0.76 (3) | ミュージカル | 0.83 (4) | सुप्रभात | 0.82 (4) |
| shall | 0.70 (3) | ウィークデイ | 0.76 (4) | परिवार | 0.69 (4) |
| walk | 0.66 (3) | 行政 | 0.80 (3) | किसान | 0.68 (3) |

Table 1: The list of stimuli used for the experiments, the duration of the words in seconds and the number of speech units. English and Hindi are phonetic languages while Japanese is a syllabic language. The first experiment uses the 12 English and 12 Japanese words while the second experiment uses the 12 Hindi and 12 Japanese words.

The Japanese was the unfamiliar language for all the subjects who participated in this experiment. In the first experimental setup, all the trials of English and the first 10 trials of Japanese were presented in random order, while the last 10 trials of Japanese were presented in a sequential manner. In the second experimental setup using Hindi and Japanese language words, all the trials were presented in a random order.

### *2.4. Data Acquisition*

The EEG signals were recorded using a BESS F-32 amplifier with 32 passive electrodes (gel-based) mounted on an elastic cap (10/20 enhanced montage). The EEG data was recorded at a sampling rate of 1024 Hz. A separate frontal electrode (Fz) was used as ground and the average of two earlobe electrodes (linked mastoid electrodes) was used as reference. The channel impedances were kept below 10 kOhm throughout the recording. The EEG data was collected in a sound-proof, electrically shielded booth. A pilot recording confirmed that there was minimal line noise distortion or other equipment related artifact. The pre-processing steps are described in detail in the supplementary material (section A).In this paper, all the analyses are performed with EEG signals recorded at listening state and with the audio signals used in stimuli as well as the spoken audio (behavioral data) collected from the subjects.

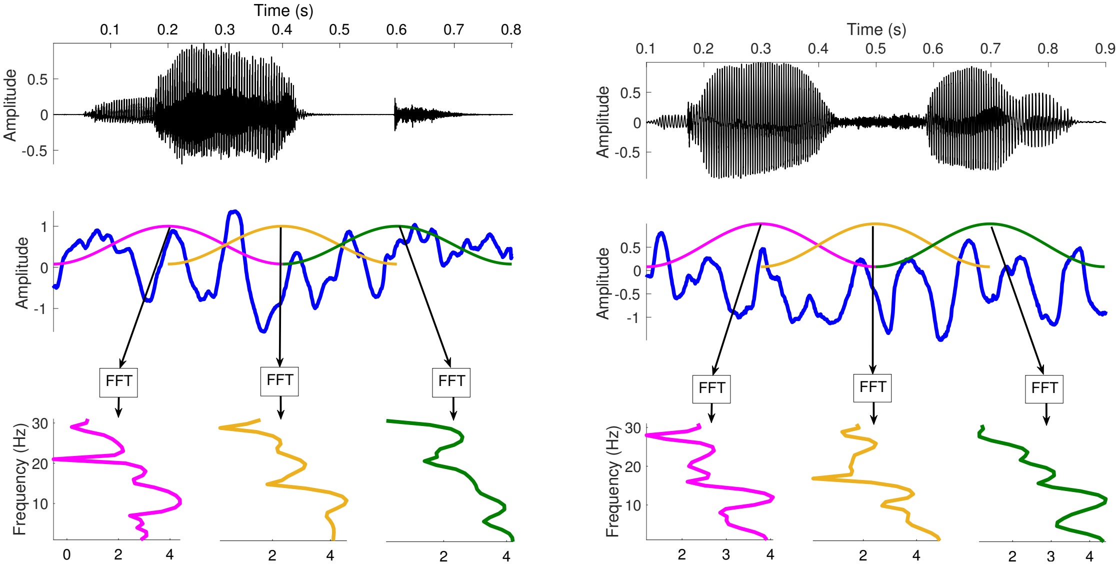
## 3. Language Classification in EEG signals

The language classification approach is used to identify the key features that discriminate the EEG representations of familiar and unfamiliar language. In particular, we try to uncover the best feature and classifier settings for discriminating English and Japanese from EEG signals (and Hindi versus Japanese from the second experimental setup). In these experiments, the chance accuracy is 50%. The training data set consists of 70% of the trials of each stimulus and the rest of the trials form the evaluation set for each subject. A support vector machine (SVM) with a linear kernel has been used as the classifier to validate the performance of different feature extraction methods. The input data to the SVM classifier is normalized to the range of 0 to 1 along each feature dimension. The SVM classifier is implemented using the LIBSVM package [19]. In the later experiments, the SVM classifier is also shown to be the best classification technique. The classification performance on channels with the best accuracies are reported in this paper.

### *3.1. Feature Extraction*

A spectrogram is computed using the Short-Time Fourier Transform (STFT) of a signal. The spectrograms are used extensively in the field of speech processing [20], music, sonar [21], seismology [22], and other areas. The spectrogram has been used to analyze the time frequency characteristics of EEG rhythms [23] and to detect seizures in epileptic patients [24].

In our spectrogram computation, we use a hamming window of duration 400ms and step size of 200ms on the input EEG signal. We compute spectrogram up-to a maximum frequency of 30Hz. In figure 2a, we show the spectrogram with 400ms window of EEG signal recorded in channel 7 when the subject was listening to an English word and figure 2b shows the spectrogram for a Japanese stimulus.



(a) English stimulus

(b) Japanese stimulus

Figure 2: Top row: Audio Signal; Middle row: EEG Signal (channel 7) of subject 1 during listening task (Average of 3 trials); Bottom row: (i) Spectrum of windowed EEG signal centered at 0.2 s; (ii) Spectrum of windowed EEG signal centered at 0.4 s; and (iii) Spectrum of windowed EEG signal centered at 0.6 s (window duration is 0.4s).

### *3.2. Trial Averaging*

In order to reduce the effect of noise and background neural activity, the EEG data from each trial is averaged with two other random trials of the same stimulus, either in temporal domain or in spectral domain. The number of trials averaged is restricted to 3 as it helps to remove noise and at the same time provides enough number of samples to train the classifier. The EEG data recorded for fixed 0.8s duration after the onset of audio stimulus is used for analysis (the duration of all the audio stimuli range from 0.5s to 0.82s). The logarithm of magnitude of spectrogram computed on temporal domain average of EEG trials is termed as *Spec(Avg)* feature. In spectral domain averaging (*Avg(Spec)*), the spectrogram is computed for each trial of the stimulus and then averaged. The logarithm of the average of magnitude spectrum along with the average of cosine of phase of the spectrograms is used as the feature vector (termed as *Avg(Spec+Phase)*).

### *3.3. Results for Language Classification*

*3.3.1.* ***Effect of Temporal Context*** - The English and Japanese languages have phonological dissimilarities like the difference in the production of /r/ and /l/ sounds as well as the presence of unique phoneme sounds in English and Japanese [25]. However, it can be hypothesized that the language specific information may not be evident in shorter segments of speech (phoneme or syllable). The poor performance of language

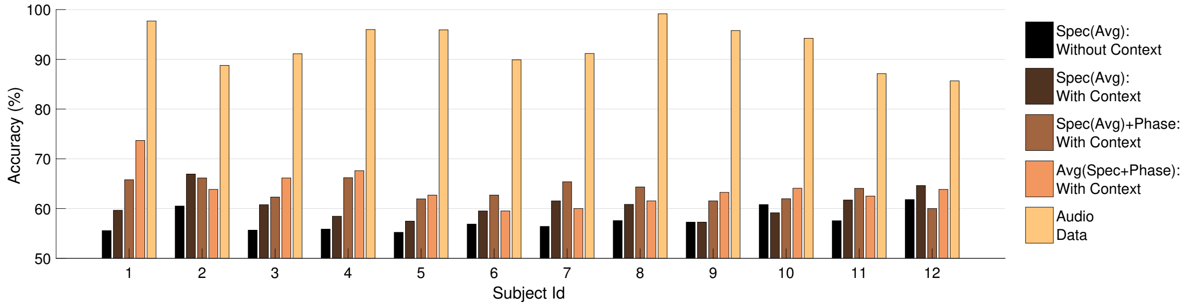


Figure 3: Language classification accuracy obtained for the 12 subjects with different feature extraction techniques on EEG data recorded during the listening state. Different feature types are Spec(Avg): Spectrogram of temporal average of trials (Sec. 3.2)- with and without context, Spec(Avg)+Phase: Phase information appended to the previous feature (Sec. 3.3.2), Avg(Spec+Phase): Average of magnitude and phase of spectrograms of trials. We also compare the performance of language identification from EEG signals to those from the spoken audio data provided by the subjects (Sec. 3.4).

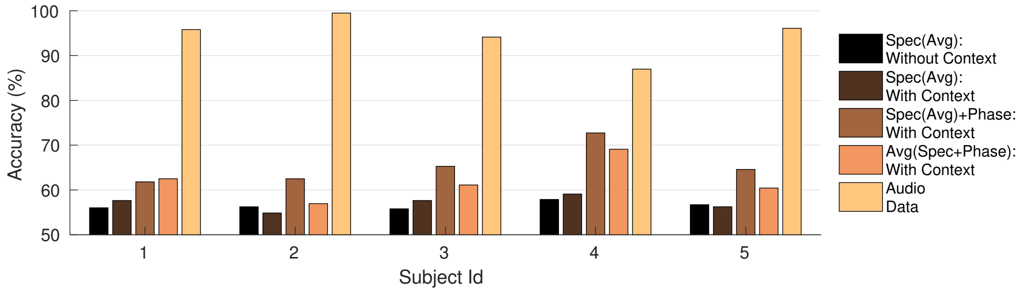


Figure 4: Hindi vs Japanese Language classification accuracy obtained for the 5 subjects with different feature extraction techniques on EEG data recorded during the listening state. Different feature types are Spec(Avg): Spectrogram of temporal average of trials (Sec. 3.2)- with and without context, Spec(Avg)+Phase: Phase information appended to the previous feature (Sec. 3.3.2), Avg(Spec+Phase): Average of magnitude and phase of spectrograms of trials. We also compare the performance of language identification from EEG signals to those from the spoken audio data provided by the subjects (Sec. 3.4).

identification at syllabic level (using a single window of 400ms without context) from neural signals confirms this hypothesis. The language variabilities are more pronounced at the interaction between different sounds which is referred to as co-articulation. Hence, incorporating context aids in language identification. Figure 3 shows the performance of *Spec(Avg)* features with SVM classifier with and without context padding. The feature extraction with context of size 3 provides better accuracy than using the features extracted from single window of EEG signal (of duration 0.4s). The features with context that provided the best accuracy in Figure 3 are also shown to perform the best in the classification of Hindi versus Japanese shown in Figure 4.

*3.3.2.* ***Effect of Phase Information in Language Recognition*** - Given the long duration of spectrogram window, we hypothesize that the phase of spectrum in the 400ms windows is also a useful feature for classification. We concatenate the cosine of the phase to the magnitude of spectrogram feature for each frame of the input signal and use it as feature vector using temporal domain averaging (*Spec(Avg)+Phase*) or using spectral domain averaging (*Avg(Spec+Phase)*). Our experiments indicate that the phase adds meaningful information to the feature regarding the familiarity of the language as shown in figure 3. We can observe that adding the phase information provides better language classification accuracy than using the magnitude of spectrogram alone, for most of the subjects. This observation is also confirmed with the experiments reported on Hindi versus Japanese (second experiment) reported in Figure 4.

As seen in figure 3, all subjects achieve language classification accuracy above 59.5% for *Avg(Spec+Phase)* features. Subject 1 attains the highest classification accuracy (73.68%). The average language classification (across subjects) obtained by *Avg(Spec+Phase)* is approximately 64% which is significantly better than chance level. The t-test conducted at a significance level of 0.05 obtained a p-value less than $10^{-5}$. This suggests that significant cues exist in the listening state EEG regarding the language identity of the stimuli. In the Hindi vs Japanese language classification, subject 4 attains the highest classification accuracy (72.73%). The classification performance for the rest of the subjects are also above 60% with phase information added to the feature vector.

*3.3.3.* ***Performance of Different Classifiers*** - As shown previously, the spectrogram magnitude information is more meaningful along with the phase information.

The performance of different classifiers for the *Avg(Spec+Phase)* features in terms of average accuracy is shown in table 2 (a). It is seen that the SVM provides best performance amongst them ($p < 10^{-4}$). The Gaussian mixture model (GMM) with two mixtures performs better than a single Gaussian model or a GMM model with 4 mixtures. The input to all four classifiers other than SVM is standardized to zero mean and unit variance along each dimension. For the LDA based classifier, we use the mean of the two classes of training data as the threshold. The statistical significance of the difference in performance of classifier models has been evaluated using paired sample t-test with significance level of 0.05 (with p $< 10^{-4}$). In this statistical test, the SVM classifier is found to be significantly better than the rest. In the second subsequent experiment on classifying Hindi and Japanese words, the Gaussian classifier provides the best performance ($p < 10^{-4}$). The Gaussian classifier, being a simpler classifier, shows better performance in classifying Hindi (L1) versus Japanese while the SVM classifier performs better for the relatively harder task of classifying English (L2) versus Japanese. Also, the data for Hindi-Japanese experiments came from only 5 subjects compared to the data from 12 subjects used in English-Japanese experiments.

| **a. Performance of Different Classifiers** | | | | | |
|---|---|---|---|---|---|
| **Model Type** | **Discriminative** | | **Generative** | | |
| **Classifier** | SVM | LDA | Gaussian | GMM<br>2 mix. | GMM<br>4 mix. |
| `I. English vs Japanese Classification` | | | | | |
| **Average Accuracy (%)** | **64.06** | 62.79 | 58.64 | 60.46 | 59.99 |
| `II. Hindi vs Japanese Classification` | | | | | |
| **Average Accuracy (%)** | 62.57 | 52.18 | **65.09** | 62.19 | 59.86 |

| **b. Language Classification in EEG Spectral Bands** | | | | | | |
|---|---|---|---|---|---|---|
| **Spectral Band** | $\delta$<br>(0.1-4Hz) | $\theta$<br>(4-8Hz) | $\alpha$<br>(8-13Hz) | $\beta$<br>(13-30Hz) | $\gamma$<br>(30-50Hz) | ALL<br>(0.1-50Hz) |
| `I. English vs Japanese Classification` | | | | | | |
| **Average Accuracy (%)** | 62.52 | 61.54 | **64.21** | 63.19 | 63.06 | 62.83 |
| `II. Hindi vs Japanese Classification` | | | | | | |
| **Average Accuracy (%)** | 61.55 | **63.71** | 61.33 | 62.44 | 62.44 | 62.73 |

Table 2: (a) Performance of different classifiers with Avg(Spec+Phase) features (Spectral Band: 0.1-30Hz). (b) Classification accuracy of SVM classifier with Avg(Spec+Phase) features in different spectral bands.

*3.3.4.* ***Language Classification in Different Spectral Bands of EEG -*** The accuracy of the language identification task varies depending on the different spectral bands of EEG signal. The analysis indicates that $\alpha$ and $\beta$ bands capture more language discriminative information as compared to $\theta$ and $\gamma$ band (Table 2 (b)). We obtain the highest classification accuracy of 64.21% in the $\alpha$ band. In the classification experiment involving Hindi versus Japanese, the $\theta$ band provides the best performance. This indicates that the language discriminative information is spectrally selective and the dominant language information is present in $\alpha$ and $\theta$ bands. The best performing sub-band rhythms has statistically significant difference in performance compared to the next best one (with $p < .005$).

### *3.4. Comparison of Language Classification in Spoken Audio and EEG*

We also perform the language classification experiment on the behavioral signals (spoken audio) from the subjects. We use the Mel Frequency Cepstral Coefficients (MFCC) [26] as the features for this experiment. The MFCC features with a context size of 53 (800ms) is concatenated and a linear discriminant analysis (LDA) is performed at word-level to reduce the dimension of these features to 23. With these features and SVM classifier, we obtain an average accuracy of 93% (for both the experiments). The comparison of resul)ts between audio and EEG shows that, while the spoken audio contains significant information for language classification, the EEG signals at the listening phase can also provide language discriminative cues which are statistically significant.

## 4. Language Learning and EEG

In the rest of analysis provided, we only use the data collected from the first experiment involving English and Japanese words.

### *4.1. Evidence of Language Learning*

In this section, we attempt to establish the evidences for Japanese language learning using the behavioral data (spoken audio signals). The aspect of language learning may cover many facets like memory, recall, semantics and pronunciation etc. In this paper, we limit the scope of language learning to improvement in pronunciation of the spoken audio. We use an automatic pronunciation scoring setup as well as human expert evaluation for this purpose.

*4.1.1.* ***Automatic Pronunciation Scoring -*** The automatic rating of speech pronunciation has been a problem of interest for many analysis tasks as well as for applications like computer assisted language learning (CALL) [27]. Several methods have been proposed for automatic pronunciation rating based on stress placement in a word [28, 29], learning-to-rank approaches [30] etc. In this paper, we use a modified version of log-likelihood based pronunciation scoring with the force-alignment of hidden Markov models(HMM) [31].

A HMM based speech recognition system is trained using the Corpus of Spontaneous Japanese(CSJ) [32]. A Hybrid HMM-Deep Neural Network (DNN) model is developed using the Kaldi toolkit [33]. For the given Japanese word used in our EEG experiments, the word level HMM is formed by concatenation of the phoneme HMMs that correspond to the phonetic spelling of the word (obtained from the dictionary of the CSJ corpus). Using the word level HMM (denoted as $\lambda$), the likelihood of the speech data $\mathbf{O} = \{\mathbf{o}_1, \mathbf{o}_2, ..., \mathbf{o}_T\}$ is approximated as [34],

$$P(\mathbf{O}|\lambda) = \sum_{\mathbf{Q}} P(\mathbf{O}, \mathbf{Q}|\lambda) \approx \max_{Q} P(\mathbf{O}, \mathbf{Q}|\lambda). \tag{1}$$

where $\mathbf{Q} = \{\mathbf{q}_1, \mathbf{q}_2, ..., \mathbf{q}_T\}$ denotes the state-sequence of the HMM and $T$ denotes the time duration. The above likelihood can be efficiently solved using the Viterbi algorithm [34]. In this work, the log-likelihood of the behavioral data (spoken audio from the subjects) and the stimuli audio are computed with force alignment and are used as confidence estimates of pronunciation. The main modification of our approach compared to the previous work in [31] is the use of state-of-art speech acoustic modeling

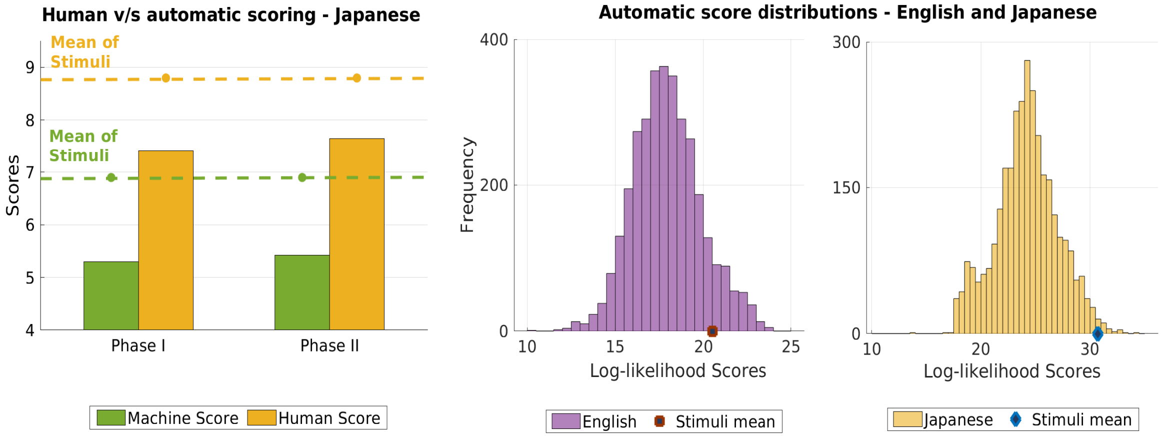


Figure 5: The left panel depicts the comparison of human and machine pronunciation scoring for Japanese language audio data. The right panel depicts the histogram of log-likelihood scores (raw machine scores) for English and Japanese spoken audio data. In both cases, mean of the stimuli is also highlighted for reference.

using deep neural networks. The details of the HMM-ASR based pronunciation rating system are provided in supplementary material (Section E).

*4.1.2.* ***Pronunciation Scoring by Human Expert -*** We also evaluate the pronunciations using a human expert[‡] based pronunciation rating (for Japanese audio). Given the large number of spoken audio recordings (20 recordings per subject per word), we use a smaller subset of this audio (4 recordings per subject per word from the 1st, 6th, 11th and 16th trial) for evaluation from the human expert in Japanese language. This was done in a scale of 1-10 (where 1 indicates a poor pronunciation and 10 indicates a native speaker pronunciation). In this evaluation, the human expert was also provided with the stimuli (in a hidden randomized manner similar to hidden reference in audio quality testing [35]) in order to ensure the effectiveness of the rating. Out of the 12 Japanese words, the 3 words for which the stimuli recording had a pronunciation rating of less than 8 were excluded from further analysis.

*4.1.3.* ***Improvement of Pronunciation over Trials -*** In figure 5, we compare the evaluations from the human expert for Japanese language recordings along with the automatic pronunciation scores. For this plot, the logarithm of likelihood scores are normalized and are linearly mapped to the range of $1-10$ in order to make the comparison with the human scores. The average rating of all the spoken audio data from the subjects (12 subjects) is plotted for two phases separately - Phase-I ($1-10$

[‡] The human expert used in our study was a professional Japanese language tutor. The text used in the stimuli was provided before the pronunciation evaluation.

trials) and Phase-II ($11-20$ trials). The stimuli ratings are also recorded for both human expert and automatic rating.

As seen in figure 5 (left panel), both the human scoring and the automatic scoring indicate an improvement in the pronunciation of the Japanese words for the Phase-II over the Phase-I. At the subject level, we also find that 10 out of 12 subjects showed an increase in scores (both human expert and automatic method) for the Phase-II over the Phase-I. Also, using the approach of log-likelihood with forced alignment shows a good match with the human expert based scoring. We also find the score improvement to be statistically significant for the human expert scoring and the machine scoring (with $p = 0.027, 0.017$ respectively).

In both cases, the mean of the log-likelihood scores for the stimuli are different from the mean of the spoken audio recordings from the subjects. This is expected as the stimuli are clean speech utterances which were recorded in a close talking microphone setting while the spoken audio recordings from the participants were collected in a far-field microphone setting. However, in the case of English spoken audio recordings, the mean of the log-likelihood scores for the stimuli is more similar to the rest of distribution compared to the Japanese language (the percentage of data below the mean value of the stimuli is 70 % in the case of English while is 95 % in the case of Japanese). This difference between the two languages is also statistically significant.

### *4.2. Understanding Language Learning via the EEG*

In this section, we use two types of analyses, (i) based on inter-trial distances and (ii) based on distance between audio and EEG envelopes.

*4.2.1.* ***Inter-trial Distance Analysis -*** We use the inter-trial distance between EEG signals to quantify the change in representation while listening to the same word over time. The hypothesis here is that in the case of a known language like English the inter-trial distance is some what random (due to the measurement noise in EEG) but small in value throughout. However, in the case of Japanese, the inter-trial distances may show a pattern of reduction over trials as a consistent representation is formed in the brain.

For testing this hypothesis, the EEG signals recorded during each trial are converted into a log magnitude spectrogram (window length of 100 ms and shifted by 50 ms). The magnitude spectrogram of each channel is converted into a single long vector and a pairwise distance between trials is computed using Euclidean distance between spectrogram vectors. An inter-trial distance matrix of size $20 \times 20$ is computed for each channel separately. This is a symmetrical matrix whose elements contain the inter-trial distances between any pair of trials. An example of inter-trial distance in EEG is shown in figure S3 of supplementary material.

In order to further analyze the inter-trial distances, the trials are broken down into two phases as before - Phase-I (trials $1-10$) and Phase-II (trials $11-20$). The mean

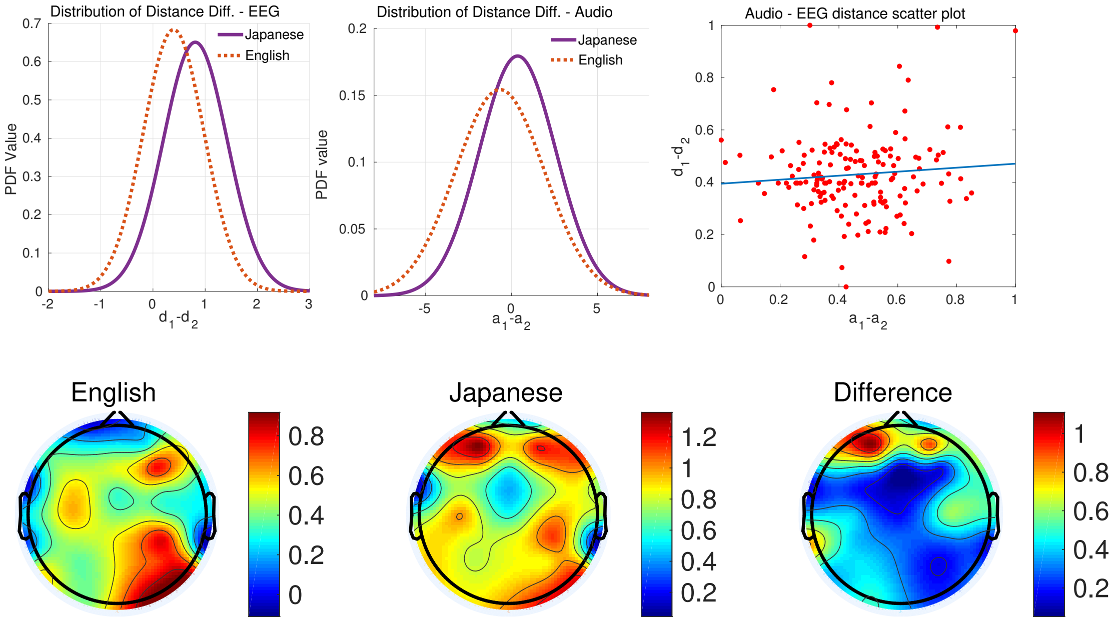


Figure 6: (above) Histograms plotted using a Gaussian fit depicting difference between the mean inter-trial distance in the Phase-I and Phase-II ($d_1 - d_2$) for EEG signals.A two-sample t-test is performed between the distribution of English and Japanese in the case of both EEG and audio. It is observed that in both the cases the distributions are statistically significant($\alpha = 0.05$).On the right, correlation between audio and EEG inter-trial distance differences for Japanese trials is shown (EEG data from electrode site FC4 is plotted here)
. (below) Scalp plots indicating the channels with higher $d_1 - d_2$ difference for English, Japanese and the difference of the two languages.

of the inter-trial distances in Phase-I (denoted as $d_1$) and the Phase-II (denoted as $d_2$) are calculated. The difference $d_1 - d_2$ is indicative of change in inter-trial distances over the course of 20 trials. We compare $d_1 - d_2$ averaged over all words and all subjects.

As hypothesized, the inter-trial distances reduce over time in the case of Japanese but remains more or less uniform in the case of English as seen in (figure 6). The histograms depicting the difference values ($d_1 - d_2$) for all the channels and the stimuli are plotted separately using a Gaussian fit for Japanese language and English language . In order to confirm the statistical significance, we performed a two-tailed test with null hypothesis being that the values of $d_1 - d_2$ for both English and Japanese come from the same distribution and the alternative hypothesis being that Japanese measurements of $d_1 - d_2$ come from a different distribution compared to English. The tolerance level alpha was set to 0.05. It is seen the distributions of the difference values for English and Japanese are statistically different.

The brain regions that show the language differences the most are shown in the scalp-plot of the difference in terms of ($d_1 - d_2$) (for English and Japanese separately for each channel averaged over all the subjects) in figure 6. A plot which differentiates the two language level scalp plots is also shown here. The regions that show more changes in English stimuli are in the temporal region while the frontal regions also show this effect in the case of Japanese stimuli. The regions that have higher difference between the two languages are predominately in the frontal brain regions.

An extension of this analysis performed for the spoken audio data is done using the audio recorded during the speaking phase of each trial. The silence portion of each recorded audio is removed. Each audio signal is converted into a sequence of MFCC feature vectors. Similar to analysis done in the previous section, a symmetrical distance matrix of size 20 × 20 is computed for each word. Since the duration of the spoken audio for the same word differs each time, an Euclidean metric based Dynamic Time Warping (DTW) distance is calculated for the pair-wise trial distance. Similar to the EEG analysis, the trials are divided into Phase-I and Phase-II. The mean inter-trial distance in Phase-I (denoted as $a_1$) and Phase-II (denoted as $a_2$) are calculated. The difference $a_1$- $a_2$ is computed similar to EEG and the histogram of the difference in the case of audio for Japanese and English (using the spoken audio data from all subjects) is plotted using a Gaussian fit (shown in the middle of the top panel of figure 6). As seen in the case of EEG, the difference ($a_1$- $a_2$) in the mean distance between the two phases is greater in the case of Japanese than English. The distribution obtained in the case of Japanese has a mean that is significantly larger than zero but not for English. Similar to the case of EEG signals, a two-tailed t-test was performed on the Gaussian fit of English and Japanese (alpha=0.05) and the two distributions were found to be statistically different.

We also analyze the correlation between EEG recorded during the listening state and the spoken audio in terms of the mean inter-trial difference in the Phase-I and Phase-II (i.e correlation between ($d_1$- $d_2$) and ($a_1$- $a_2$)). A scatter plot is shown with the difference values for EEG signals along the y-axis and the corresponding difference for audio signals along the x-axis (i.e. ($d_1$- $d_2$) versus ($a_1$- $a_2$)). An example of the scatter plot difference of ($d_1$- $d_2$) versus ($a_1$- $a_2$) for the frontal EEG channel (FC4) is shown in figure 6. Each point on the plot indicates a (subject,word) pair. The values along both the axes are normalized between 0 and 1. A line of best fit is plotted through the points. The slope of the line (denoted by m) of best fit is positive for most of the channels. Since the scales for the EEG spectrogram and the audio MFCC features are different, the amount of correlation between the listening state EEG and the audio spoken may be unnormalized. Additionally, the mean slope of best fit lines for Japanese words is found to be higher than English. These observations indicate that the pattern formation seen in the behavioral data is also correlated with the patterns seen in EEG recordings.

*4.2.2.* ***Distance between EEG and Audio Envelopes -*** A direct relationship between the EEG signals recorded during the listening and the audio spoken by the subject during the speaking phase may also present useful insights. Previous studies

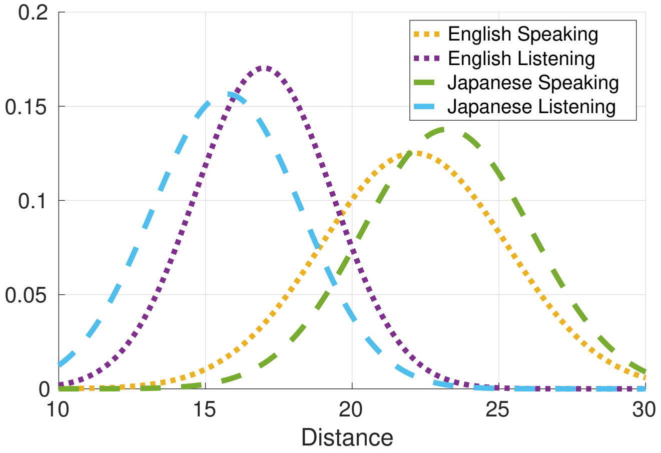


Figure 7: Probability Distribution Function of distances between the envelope of listening state EEG and envelope of stimuli presented and the envelope of spoken audio. A two-sample t-test is performed and it is seen that the distributions of the two languages are statistically different for each state.($\alpha = 0.05$)

have attempted to predict the audio envelope using EEG [36] or to perform a correlation analysis between the EEG and audio envelope [37]. In our study, we try to align the EEG and audio envelopes (after down-sampling to the same rate) and measure a distance between the two. Thus, the distance measure is inversely proportional with the correlation measures used in the past, as smaller distances between audio and EEG envelopes are associated with higher correlations and vice-versa. The choice of distance is to maintain consistency with the previous analysis based on distances. A sample plot of EEG and audio envelopes that are time aligned is shown in figure S4 of supplementary material.

For the distance computation, the silence portion of the audio is removed and the length of the EEG signals are kept to stimuli length plus $100ms$. Both the signals are converted to their corresponding Hilbert envelopes and the envelopes are down-sampled to 64Hz. The DTW distance between the two aligned envelopes is calculated. It is seen that the mean distance between the two envelopes is greater in the case of Japanese than English (figure 7).

As a follow up to the comparison done between the envelope of the EEG signals and the audio spoken, a similar analysis is done between the envelope of the EEG signals and the stimuli presented to the subject. A DTW distance is computed between envelope of listening EEG and envelope of stimuli. A histogram of all the distances (between listening stimuli and EEG as well as those between the spoken audio and EEG) for both the languages are shown in figure 7. The average distance values between envelopes (for listening state) is less in the case of Japanese compared to speaking state. The distance between the envelopes of the EEG signal and the spoken audio is more than the distance between the envelopes of the EEG and stimuli presented as well.

A two-tail t-test was performed on the distributions of distance between the envelopes of EEG and audio for English and Japanese. This was done for both distance measures between EEG and stimuli envelope as well as EEG and spoken audio envelope. In both the cases, the null hypothesis was that the distributions of English and Japanese are not statistically different and the alternative hypothesis being that the two distributions are statistically different. In both the cases, the t-test indicated statistically significant deviation from the null hypothesis. This supports our claim that the distributions obtained for the relationship between the envelopes of EEG and audio are statistically different for the two languages.

## 5. Discussion

As seen in figure 6, the inter-trial distances reduce over time in the case of Japanese but remains more or less uniform in the case of English. The familiarity of the subjects to the words from English language may have resulted in generating invariant EEG responses when presented with these stimuli. In the case of Japanese stimuli, subjects are listening to those words for the first time. Over the trials, subjects form a consistent neural representation for the unfamiliar stimulus. It is evident from the reduction of inter-trial distances of the EEG responses.

The stimuli presentation and listening state EEG recording happens in parallel. Hence, a higher correlation is expected between the two compared to the correlation between the envelopes of listening state EEG and the spoken audio. This is seen in figure 7. Since Japanese is unfamiliar, the spoken audio is not well aligned to the stimuli. Hence, the distance between spoken audio and EEG envelopes may be more for Japanese compared to English.

All the subjects who participated in our recordings were not exposed to Japanese before but had a good proficiency to English. We hypothesize that due to their unfamiliarity of Japanese their attention while listening to Japanese stimuli is much more than English resulting in lesser distance between envelopes of EEG and Japanese stimuli compared to English. The absence of semantic processing in Japanese could also explain the reduced distance between stimuli envelope and EEG envelope for Japanese. In the speaking state, the subjects tend to reproduce audio that is less correlated with stimuli for Japanese language than English language. This may explain the rightward shift of the distribution of distances for Japanese spoken audio in figure 7.

## 6. Conclusions

The key findings from this work are the following,

- A consistent neural representation is formed when exposed repeatedly to words from an unfamiliar language. This is also consistent with language learning established using pronunciation rating.
- In the listening state, the correlation between audio stimuli and EEG envelope is more for Japanese trials than English trials (smaller distance values). The

correlation between EEG envelope of listening state and envelope of the spoken audio is less for Japanese than English.

- The discriminative signatures of the language are encoded in the time-frequency representation of the EEG signals in the range of 0-30Hz both in magnitude and phase.

We have additionally performed analysis to find out the channels that capture the language learning the most. The channels are identified as the ones that show the maximum difference between the Phase-I and Phase-II. The top five channels are found to be $(O2, AF3, F8, AF4, F7)$, located primarily in the frontal region of the brain.

In the current setup, the unfamiliar words are presented to the subjects without the semantic meaning or the context of the word. In future experiments, we plan to see how the neural responses change when the unfamiliar words are provided with semantics. Additionally, longer content is expected to provide future insight between language level differences compared to word level analysis. This can be achieved with stimuli containing longer words, phrases and sentences. The current setup lacks any feedback to the subjects on how well they perform the learning task. In future, we also plan to introduce a scoring model which rates and gives feedback to the subject depending on how well they pronounce the words during the experiment itself.

## 7. Acknowledgements

We would like to thank Axxonet System Technologies, Bangalore for providing the facilities for data collection and helping us for the same. Additionally we would also like to thank Dr. Aravind Kumar, NIMHANS, Bangalore for critical discussions and useful insights regarding EEG recordings and experimental design.

## 8. Funding Details

This work was partly funded by grants from the Department of Science and Technology under the Early Career Award and from the Prathiksha Young Investigator Award. The views expressed in the paper are those of the authors and do not reflect the views of the funding agencies. The authors declare no competing financial interests.

## References

[1] Shi S J and Lu B L 2009 EEG signal classification during listening to native and foreign languages songs *4th Int. IEEE/EMBS Conf. on Neural Engineering* pp 440–443 ISSN 1948-3546

[2] Pallier C, Dehaene S, Poline J, Lebihan D, Argenti A, Dupoux E and Mehler J 2003 Brain imaging of language plasticity in adopted adults: can a second language replace the first? *Cerebral cortex* **13** 155–61

[3] Vingerhoets G, Van Borsel J, Tesink C, van den Noort M, Deblaere K, Seurinck R, Vandemaele P and Achten E 2003 Multilingualism: an fMRI study *NeuroImage* **20** 2181–2196

[4] Videsott G, Herrnberger B, Hoenig K, Schilly E, Grothe J, Wiater W, Spitzer M and Kiefer M 2010 Speaking in multiple languages: Neural correlates of language proficiency in multilingual word production *Brain and language* **113** 103–112

[5] Berlad I and Pratt H 1995 P300 in response to the subject's own name *Clinical Neurophysiology* **96** 472–474

[6] Radicevic Z, Vujovic M, Jelicic L and Sovilj M 2008 Comparative findings of voice and speech: language processing at an early ontogenetic age in quantitative EEG mapping *Experimental brain research* **184** 529–532

[7] Münte T F, Altenmüller E and Jäncke L 2002 The musician's brain as a model of neuroplasticity *Nature Reviews Neuroscience* **3** 473–478

[8] Van Praag H, Kempermann G and Gage F H 2000 Neural consequences of enviromental enrichment *Nature Reviews Neuroscience* **1** 191–198

[9] Kuhl P K 2000 A new view of language acquisition *Proceedings of the National Academy of Sciences* **97** 11850–11857 ISSN 0027-8424

[10] Osterhout L, Poliakov A, Inoue K, McLaughlin J, Valentine G, Pitkanen I, Frenck-Mestre C and Hirschensohn J 2008 Second-language learning and changes in the brain *Journal of Neurolinguistics* **21** 509–521

[11] Li P, Legault J and Litcofsky K A 2014 Neuroplasticity as a function of second language learning: anatomical changes in the human brain *Cortex* **58** 301–324

[12] Mårtensson J, Eriksson J, Bodammer N C, Lindgren M, Johansson M, Nyberg L and Lövdén M 2012 Growth of language-related brain areas after foreign language learning *NeuroImage* **63** 240–244

[13] Perani D, Paulesu E, Galles N S, Dupoux E, Dehaene S, Bettinardi V, Cappa S F, Fazio F and Mehler J 1998 The bilingual brain. proficiency and age of acquisition of the second language. *Brain: A Journal of Neurology* **121** 1841–1852

[14] Weber K, Christiansen M H, Petersson K M, Indefrey P and Hagoort P 2016 fmri syntactic and lexical repetition effects reveal the initial stages of learning a new language *Journal of Neuroscience* **36** 6872–6880

[15] Butzkamm W 2003 We only learn language once. the role of the mother tongue in FL classrooms: death of a dogma *Language Learning Journal* **28** 29–39

[16] Potter C E and Saffran J R 2015 The role of experience in children' s discrimination of unfamiliar languages *Frontiers in Psychology* **6**

[17] Prat C S, Yamasaki B L, Kluender R A and Stocco A 2016 Resting-state qEEG predicts rate of second language learning in adults *Brain and language* **157** 44–50

[18] Almeida D and Poeppel D 2013 Word-specific repetition effects revealed by MEG and the implications for lexical access *Brain and language* **127** 497–509

[19] Chang C C and Lin C J 2011 LIBSVM: A library for support vector machines *ACM Transactions on Intelligent Systems and Technology* **2** 27:1–27:27

[20] Rabiner L R and Schafer R W 1978 *Digital Processing of Speech Signals* (Prentice Hall)

[21] Lourens J 1990 Passive sonar detection of ships with spectrograms *Proc. of the South African Symp. on Communications and Signal Processing, (COMSIG 90)* (IEEE) pp 147–151

[22] Parrot M, Berthelier J, Lebreton J, Sauvaud J, Santolik O and Blecki J 2006 Examples of unusual ionospheric observations made by the demeter satellite over seismic regions *Physics and Chemistry of the Earth, Parts A/B/C* **31** 486–495

[23] Van Hoey G, Philips W and Lemahieu I 1997 Time-Frequency analysis of EEG signals *Proc. of the ProRISC Workshop on Circuits, Systems and Signal Processing*

[24] Schiff S J, Colella D, Jacyna G M, Hughes E, Creekmore J W, Marshall A, Bozek-Kuzmicki M, Benke G, Gaillard W D, Conry J *et al.* 2000 Brain chirps: spectrographic signatures of epileptic seizures *Clinical Neurophysiology* **111** 953–958

[25] Carruthers S W 2006 Pronunciation difficulties of japanese speakers of english: Predictions based on a contrastive analysis *Hawaii Pacific University TESOL Working Paper Series* **4** 17–24

[26] Davis S B and Mermelstein P 1990 Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences *Readings in Speech Recognition* (Elsevier) pp 65–74

[27] Stouten F and Martens J 2006 On the use of phonological features for pronunciation scoring *Proc. of the Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)* vol 1 (IEEE) pp 329–332

[28] Kawahara T, Dantsuji M and Tsubota Y 2004 Practical use of English pronunciation system for Japanese students in the call classroom *Eighth Int. Conf. on Spoken Language Processing* pp 1689–1692

[29] Tepperman J and Narayanan S 2005 Automatic syllable stress detection using prosodic features for pronunciation evaluation of language learners *Int. Conf. on Acoustics, Speech, and Signal Processing, Proceedings (ICASSP)* vol 1 (IEEE) pp 937–940

[30] Chen L Y and Jang J S R 2015 Automatic pronunciation scoring with score combination by learning to rank and class-normalized dp-based quantization *IEEE/ACM Transactions on Audio, Speech, and Language Processing* **23** 1737–1749

[31] Neumeyer L, Franco H, Weintraub M and Price P 1996 Automatic text-independent pronunciation scoring of foreign language student speech *Fourth Int. Conf. on Spoken Language, ICSLP Proceedings* vol 3 (IEEE) pp 1457–1460

[32] Maekawa K 2003 Corpus of Spontaneous Japanese: Its design and evaluation *ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition (SSPR)* pp 7–12

[33] Povey D *et al.* 2011 The Kaldi speech recognition toolkit *IEEE 2011 workshop on Automatic Sspeech Recognition and Understanding* (IEEE Signal Processing Society)

[34] Rabiner L R and Juang B H 1993 *Fundamentals of Speech Recognition* vol 14 (Prentice Hall Englewood Cliffs)

[35] Rix A W, Beerends J G, Kim D, Kroon P and Ghitza O 2006 Objective assessment of speech and audio quality technology and applications *IEEE Transactions on Audio, Speech, and Language Processing* **14** 1890–1901

[36] Horton C, Srinivasan R and D’Zmura M 2014 Envelope responses in single-trial EEG indicate attended speaker in a ‘cocktail party’ *Journal of Neural Engineering* **11** 046015

[37] de Cheveigné A, Wong D D, Di Liberto G M, Hjortkjær J, Slaney M and Lalor E 2018 Decoding the auditory brain with canonical component analysis *NeuroImage* **172** 206–216