Analysing patterns of right brain-hemisphere activity prior to speech articulation for identification of system-directed speech

Analysing patterns of right brain-hemisphere activity prior to speech articulation for identification of system-directed speech

Fasih Haider

Usher Institute of Population Health Sciences & Informatics, Edinburgh Medical School, the University of Edinburgh, UK

Hayakawa Akira

School of Computer Science and Statistics, Trinity College Dublin, Ireland

Carl Vogel

School of Computer Science and Statistics, Trinity College Dublin, Ireland

Nick Campbell

School of Computer Science and Statistics, Trinity College Dublin, Ireland

Saturnino Luz

Usher Institute of Population Health Sciences & Informatics, Edinburgh Medical School, the University of Edinburgh, UK

Abstract

Autonomous speech-enabled applications such as speech-to-speech machine translation, conversational agents, and spoken dialogue systems need to be able to distinguish system-directed user input from "off-talk" to function appropriately. "Off-talk" occurs when users speak to themselves or to others, often causing the system to mistakenly respond to speech that was not directed to it. Automatic detection of off-talk could help prevent such errors, and make the user's interaction with the system more natural. It has been observed that speech in human-human dialogue and in soliloquy is prosodically different from speech directed at machines, and that the right hemisphere of the human brain is the locus of control of speech prosody. In this study, we explore human brain activity prior to speech articulation alone and in combination with prosodic features

to create models for off-talk prediction. The proposed EEG based models are a step towards improving response time in detecting system-directed speech in comparison with audio-based methods of detection, opening new possibilities for the integration of brain-computer interface techniques into interactive speech systems.

Keywords: multimodal interaction, spoken dialogue systems, speech-to-speech machine translation, brain computer interface (BCI), electroencephalogram (EEG)

1. Introduction

Enabling natural human-computer interaction through speech involves endowing the machine with the ability to distinguish between speech that is addressed to it from speech that is addressed to others [1]. Designers of spoken dialogue systems [2, 3] refer to speech not addressed to the system as "off-talk" [4], and to speech addressed to the system as "on-talk". The latter exhibits distinctive linguistic behaviours. Previous studies have shown that the user's talking behaviour varies depending on whether the interlocutor is a machine or a human being [5, 6], and that talking to a computer is similar to talking to a person who has a hearing impairment [3]. Human-computer communication tends to be more "exaggerated" than human-human communication [5].

In applications such as speech-to-speech machine translation, users often experience communication difficulties due to the errors caused by system misinterpretation of off-talk, especially when facial and gestural cues that normally aid communication are reduced or absent from the system [7]. In such settings, the speaker adapts different strategies such as overarticulating and slowing their speech rate when automatic speech recognition (ASR) fails. However, as modern ASR systems are trained on normal paced speech, this strategy often results in worse ASR performance [8, 9]. Self-talk, or soliloquy which is an example of off-talk, may occur in a dialogue system or speech translation system due to a number of factors. Failure of speech recognition or machine translation

components, for instance, often causes users to talk to themselves in amusement or frustration. In a speech-to-speech machine translation system, users might read out loud the system's textual output, such as feedback or back-translation displayed to them during interaction with the system. Talking to others ("other-talk") is also quite common in several contexts of use of dialogue systems, where addressee detection is an active research area [1]. Therefore, equipping the system with the ability to detect off-talk could enhance system performance by avoiding the processing of off-talk utterances, and using this information (i.e. which utterances are off-talk) as a feedback to the ASR module and other system components. In an audio conference where participants use a machine translation system, for instance, off-talk detection could prevent irrelevant and potentially confusing utterances from being translated and transmitted to the remote participants.

This study extends our previous work [10, 11] where the EEG signal is analysed in overt speech (during articulation) rather than in covert (prior to articulation) as in the present study. We also analyse overt speech (during articulation) in combination with a very high dimensional set of acoustic features in previous studies [10, 11]. This study proposes a new approach for automatic detection of on- and off-talk which could decrease the response time of an interactive speech driven system in accepting or rejecting a speech utterance. This model employs electroencephalography (EEG) features collected prior to articulation (covert speech setting). While EEG signals have been employed before in speech-related brain-computer interaction (BCI), these applications tend to focus on the interpretation of "silent" or covert speech [12], where the user "imagines" the words or phonemes to be produced but does not physically articulate them. This focus on covert speech is due to the fact that muscle activity during speech articulation produces noise that contaminates the EEG signal. By focusing on the preparatory phase of speech production [13, 14] and processing the EEG signal before articulation starts, our approach avoids this difficulty. We implemented and assessed models that employ pre-articulation EEG features in isolation and in combination with prosodic features gathered during articulation for on- and

off-talk detection. The system architecture underlying our method is depicted in Figure 1, where a Voice Activity Detection (VAD) component detects the start and end time of a speech utterance from the incoming audio stream and then extracts acoustic features from the speech utterance for off-talk detection. The EEG based off-talk detection system is triggered as soon as it detects 10-ms of speech and processes the EEG features corresponding to a time window immediately prior to articulation from the memory buffer for off-talk detection. In summary, the main research contributions of this article are:

1. the introduction of a novel method for automatic detection of on- and off-talk utterances;

2. a demonstration of the usefulness of EEG signals recorded up to 2 seconds prior to articulation for on- and off-talk detection;

3. an analysis of the predictive power of the fusion of audio and EEG features with regard to this detection task, and

4. a demonstration of the discriminating power of EEG potential generated by the right mid-front and right mid-back positions of brain for off-talk detection.

At a practical level, improvements in response time yielded by the proposed method could contribute towards the design of interactive speech systems that show attentive behaviour to users and, in the specific case of machine-translated audio conferencing, improve the flow of conversation. In the scope of this study, we mainly focused on prosodic information. Therefore we have focused on electrical signals from the right brain-hemisphere and on basic prosodic features, as it is generally assumed that the right brain-hemisphere preferentially processes prosody [15, 16, 17, 18, 19, 20, 21]. For instance, Heilman et al. assessed the brain of a subject with a right medial frontal cerebral infarction and observed an impairment in expressing emotions using prosody, and in comprehension and repetition of prosody [16]. The left brain-hemisphere is largely involved in speech production aspects other than prosody control [22, 23]. For instance,

Flinker et al. showed that the Broca area activates prior to articulation and stays inactive during articulation. The motor cortex, on the other hand, activates during speech production but remains inactive prior to articulation [22]. Focusing on the right brain-hemisphere also minimises possible confounding from brain patterns related to hand control as the user uses the mouse, since the left brain-hemisphere controls the right hand [17, 24], and the subjects of this study are right-handed. As regards the analysis of prosody on the speech signal for distinguishing on-talk from the two types of off-talk (other-talk and self-talk). It is noted that although other talk and self talk seem to differ prosodically, talking to a system has a distinctively less natural character, marked by features such as over-articulation, louder and slower speaking etc. While further investigation is still needed, these features of off-talk seem to persist across the different languages recorded in our experiment.

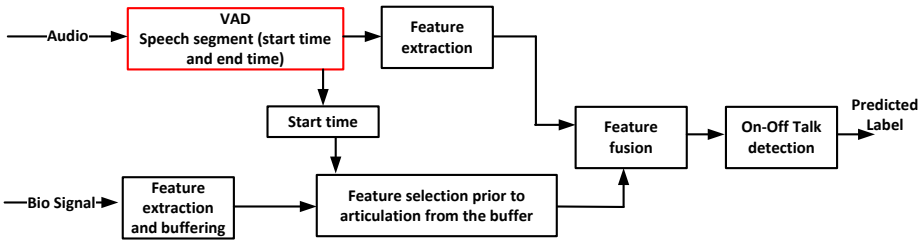

Figure 1: The system architecture where the system processes the EEG features prior to articulation as soon as it received 10 ms of audio which is detected through voice activity detection (VAD) using audio features.

2. Electroencephalography and speech production

With the introduction of less intrusive wireless EEG headsets e.g. EPOC[1], the use of EEG information is now more convenient in human-machine interactions than before. However as mentioned before, the EEG signal is quite susceptible to artefacts caused by talk-related muscle activity, including head

[1] https://www.emotiv.com/epoc/ – Last verified August 2018

movement and eye blinks. This problem is commonly approached by recording signals on several different positions on the scalp, instructing the subjects to avoid moving and to keep calm during recordings, and subsequently employing independent component analysis on the EEG data in order to remove the artefacts [25]. However, in an interactive setting, we cannot restrict user's movements. When engaged in natural interaction people move their heads, speak, display emotions, gesture and laugh. Therefore, preventing EEG artefacts becomes even harder if any amount of naturalness in human-computer interaction is to be preserved.

Muscle activity can introduce noise in EEG signals (e.g. peak frequencies of *masseter* muscle movements are in 50–60 Hz range, and *frontalis* muscles movements are between 30–40 Hz) with the noise band limit ranging between 15 to 100 Hz [26]. Goncharova et al. [27] report the noise range for *frontalis* muscles as 20–30 Hz and temporal muscles as 40–80 Hz. The discrepancies in noise range for frontal muscles can be due to the fact that O'Donnell et al [26] use fewer subjects and electrodes than Goncharova et al. [27]. Posterior head muscle movements have a higher peak frequency close to 100 Hz. but this depends on many factors, such as sex, force and direction of contraction, etc [28]. Muscle activity may also introduce artefacts in a frequency range (20–300 Hz) where most artefacts are at the lower end [29].

Muscle artefact noise is the main reason why the EEG signal has been rarely used for speech related applications. However, a few studies have employed EEG signals in both overt and covert speech settings. In covert speech production settings [14, 30, 31, 32, 12] the subjects are asked to think about a word or phoneme instead of articulating them. While this minimises noise artefacts on the EEG signal, a limitation of this methodology is that it is difficult to verify with certainty that the subjects actually followed the task instructions. In overt speech studies [33, 34], the EEG signal is analysed after a stimulus is presented to the subject up until the start of articulation. Our study adopts a different approach to overcome the issue of noisy EEG signal due to overlap with speech production namely it only uses EEG signal from time periods where there is no

speech activity. In doing so it differs from the above-mentioned studies as it is performed in a setting that is closer to natural communication. Our goal is to detect different speech registers (on- and off-talk) rather than to decode the EEG counterpart to a specific phonetic production.

The Broca area of the human brain plays a role in speech production [35, 36]. The seminal research by Broca, Wernicke and others on the relationship between neural activity and speech production, which highlighted parts of the brain responsible for speech production has been supported by a number of studies [23]. It has been observed that the speech signal is preceded by low variation in the EEG signal up to one second before articulation [37]. The cognitive processes that lead to speech articulation (activate the speech production areas in the brain) are thought to be of three main types [38, 39, 40, 41, 42, 43, 13]: 1. *conceptualization* – the content and pre-linguistic representation of the intended speech, 2. *formulation* – retrieval of the best match between linguistic representation and conceptual structure, and 3. *grammatical and phonological encoding* – selection of lexical items and intonation pattern [14].

Electro-physiological evidence of phonological encoding that leads to articulation has been observed. M. Van Turennout et al. [14] found such evidence in the EEG signal from the mid-line frontal (Fz), central (Cz), and parietal (Pz) sites of the 10-20 system [44] in a picture naming task. Other studies implicated the right brain-hemisphere in the control of speech prosody [19, 20, 21]. The evidence reviewed in the literature therefore suggests that EEG information can be used for modelling the characteristics of speech prior to articulation, and may help distinguish on- and off-talk by anticipating prosodic differences in intonation level, speech rate and lexical words [10, 45].

3. Data Set

A data set was collected which consists of recorded human dialogues mediated through a speech-to-speech machine translation system [46]. The participants communicated remotely through the system to solve a map task problem,

where one participant (the instruction giver) has a complete map and the other (the instruction follower) has a map with missing information [47]. Three different types of talk were observed in this setting: 1) on-talk, where the speaker directed speech to the ASR for transmission to the other participant 2) self-speaking, where participants spoke to themselves (e.g. venting frustration at system component failure) producing utterances not intended for ASR or transmission, and 3) other-talk, where participants spoke directly to other people than their remote task partner (e.g. a colleague that happened to be in the same room). Both self-speaking and other-talk are regarded as off-talk in this study. The data used for the research described in this paper includes precisely synchronised audio and EEG signals, from the Interlingual Map Task (ILMT-s2s) corpus [48].

3.1. The ILMT-s2s System

The user of ILMT-s2s system presses a button when they wish to speak to the remote participant. However, they cannot hear each other directly. A speech synthesiser (Apple TTS system with voices Kate, for English, and Joana for Portuguese) provides the output of the ASR and machine translation (MT) components to them. Only one of the dialogue participants uses the physiological recording equipment [48] in any particular session. In total, there are 30 participants (15 English and 15 Portuguese speakers), of which there are 15 subjects who are equipped with *'Mind Media B.V., NeXus-4'* for bio-signal (i.e. skin conductance, heart rate and EEG) recording and the duration of dialogues is between 20 and 74 minutes. In this study, we use the datasets of 13 out of 15 subjects who were fitted with bio-signal recording equipment (i.e. *'Mind Media B.V., NeXus-4'*) because the EEG data of two subject was not recorded properly due to improper fitting of the EEG electrodes. The number of on-off talks produced by all subjects along with the mean and standard deviation of duration values (in seconds) are shown in Table 1. The on-talk utterances were labelled automatically as the speech utterances were sent to the ASR system, and the remaining labels were added and checked manually by a single expert

annotator.

Table 1: Dataset description showing the number of on-talk and off-talk (self-talk and other-talk) instances for each participant, along with the mean and standard deviation of duration (in seconds) for these instances.

Subject	Self Talk			On Talk			Other Talk		
	Instances	mean	Std.	Instances	mean	Std.	Instances	mean	Std.
S1	25 (43.10%)	0.65	0.26	33 (56.90%)	0.87	0.88	0 -	-	-
S2	55 (34.59%)	0.77	0.70	100 (62.89%)	2.1	7.8	4 (2.52%)	0.81	0.54
S3	10 (6.25%)	0.65	0.31	120 (75.00%)	1.87	7.13	30 (18.75%)	0.83	0.90
S4	5 (4.85%)	0.74	0.24	98 (95.15%)	1.65	7.83	0 -	-	-
S5	46 (27.06%)	0.73	0.53	92 (54.12%)	2.2	8.12	32 (18.82%)	0.84	0.74
S6	60 (27.27%)	2.35	10.01	103 (46.82%)	1.21	1.32	57 (25.91%)	0.89	0.80
S7	40 (33.61%)	0.77	0.81	73 (61.34%)	2.27	9.08	6 (5.05%)	0.79	0.25
S8	2 (0.99%)	0.89	0.26	201 (99.01%)	1.52	5.56	0 -	-	-
S9	106 (68.39%)	1.61	7.53	49 (31.61%)	1.66	1.8	0 -	-	-
S10	10 (14.93%)	0.83	0.28	57 (85.07%)	0.76	0.70	0 -	-	-
S11	44 (36.07%)	0.75	0.72	78 (63.93%)	2.2	8.79	0 -	-	-
S12	13 (26.53%)	0.72	0.29	34 (69.39%)	0.79	0.86	2 (4.08%)	0.75	0.38
S13	6 (6.25%)	0.84	0.31	89 (92.71%)	1.63	8.20	1 (1.04%)	0.93	0
Total	422			1127			132		

3.2. Audio Recordings

Two audio and five video streams form part of the ILMT-s2s corpus. For the participants fitted with the EEG sensors audio was recorded from three different sources: a) a Sony HDR-XR500 handy-cam per subject, recording at 1080i, 29.97 fps, b) SMI Eye Tracking Glasses 1.1 recording at 960p, 30 fps, and a push-to-talk microphone, sampled at 96 KHz, 24 bit PCM format. In this study, we used the audio recorded by the two Sony HDR-XR500 handy-cams (sampled at 48 KHz, 16 bit PCM format) rather than the audio captured by the push-to-talk (using the computer's mouse) microphone because the latter records only on-talk.

3.3. EEG Recording

The EEG is recorded using the Mind Media B.V., NeXus-4 with a head fixture EEG cap[2]. The EEG sensors were placed on the F4, C4, P4 sites (located on the right hemisphere of the brain, which is responsible for the control of speech prosody [19, 20]) with a ground channel placed at position A1 (left ear lobe) of the 10-20 location system. The sampling frequency was 1,024 Hz. Henceforth, we refer to the input gathered from the F4-C4 channel as *sensor A: right mid-front* and to the C4-P4 channel input as *sensor B: right mid-back*.

4. Feature Processing and Classification method

4.1. EEG Power Spectrum Features

Feature extraction was performed on the EEG signal two seconds before articulation. A frame length of 250 ms (no overlap with neighbouring frames using a rectangular window) was used for feature extraction which resulted in a total of eight frames (2 seconds) collected prior to articulation. First we took the Fourier transform of the EEG frame and calculate its power spectrum. Then we set a frequency bin resolution of 5 Hz (from 0-40 Hz) that resulted in 8 frequency bins for each EEG frame. We ignored frequencies above $40Hz$ in this study, in line with accepted clinical EEG standards where it is normally assumed that the higher frequencies ($> 40Hz$) do not contain clinically relevant neural activity. We also note in passing that, while contrary to a common misconception the human skull does not filter out higher frequencies [49], neural activity at such frequencies are harder to detect due to attenuation caused by the skull's resistivity [50]. For this reason in the current study we only process frequency bins bellow 40Hz, computing the ratio and range of power between these eight frequency bins, which results in 64 features per frame for each EEG sensor. The processing of EEG features in this study is summarised in Figure 2.

[2]https://www.mindmedia.com/en/products/accessories/minicap/ – Last verified November 2018

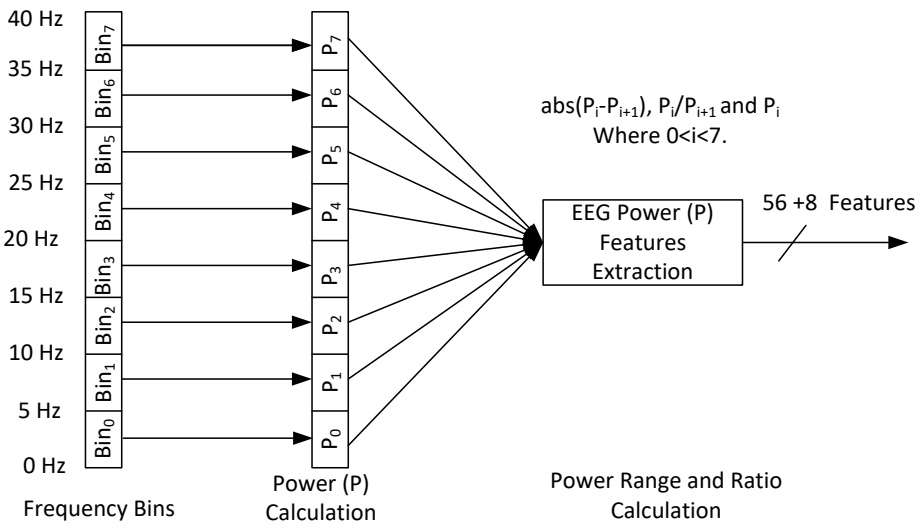

Figure 2: *Frame (250 ms) level feature extraction from the EEG Signal*

4.2. Prosodic Features

For the analysis of possible differences in phonetic characteristics during articulation of on- and off-talk utterances we extracted functionals of prosodic features of the participant's speech. The mean and standard deviation values of sound intensity, loudness (normalised intensity raised to a power of 0.3) and fundamental frequency were extracted from each speech utterance of on- and off-talk using the openSMILE toolkit [51, 52]. Although prosodic differences in the broad sense includes speaking pace and pauses, are distinguishing features of off-talk from a listener perspective, many off-talk utterances are too short (see Table 1) for the feature to be useful in the machine learning approach adopted in this study, which is applied at the utterance level. The duration of speech utterance is not constant and the duration statistic of speech utterances are shown in Table 1. While in previous studies [10, 11] we evaluated the acoustic feature sets of the Compare challenge (6373 features) [52] and Emobase (988 features) [51], in this study we limited acoustic features to fewer basic prosodic features (4 features). The motivation for this is to compare prosodic features of speech, which can be more easily interpreted than those large feature sets,

to EEG features extracted from right brain-hemisphere, which controls speech prosody.

4.3. Classification Methods

In this study, we used the Scikit-learn [53] implementation of the Random Forest (RF) classifier [54, 55] for model training and testing in a 10-fold cross validation setting using 50 trees in the forest. The RF classifier was chosen because this classification method has been shown to be robust in tasks where the number of features approaches the number of training instances, as compared to other methods such as discriminant analysis, support vector machines and neural networks [55]. For comparison we also employed a K-nearest neighbour classifier (KNN, with K=3). The results were evaluated using the A-weighted F_1-score statistic which is the average of F_1-score of both classes (on- and off-talk). The baseline of A-weighted F_1-score is 50%.

5. Results and Discussion

We have evaluated the discrimination power of EEG signals two seconds prior to articulation with a frame length of 0.25 seconds on 0-40Hz frequency bands by conducting the following three different experiments:

Experiment 1: In this setting, we evaluated the discriminative power of eight frames (250 ms) of EEG (2 seconds before articulation) for on-off talk classification using the KNN and the RF classifiers. The EEG signals from both sensors are used in this experiment. The predictive power of each sensor is evaluated individually and in combination. The fusion of Sensor A and Sensor B features was evaluated for EEG signals one second before articulation (1s) and one second before 1s (2s), separately. We then fused the features of all frames for classification. The results are presented in Table 2. The best result (80.25%) was obtained using the frame level features of both EEG sensors extracted two seconds prior to articulation.

Experiment 2: The prosodic features (during articulation) were used for this classification task. The results for these features are shown in Table 3. The

Table 2: Results of the 10-fold cross validation experiment 1 (A-Weighted F-Score %) for each frame before articulation, and feature fusion of one second (4 frames) and two seconds (8 frames) before articulation. Bold face indicates the best results.

EEG Features	Window (sec)	Sensor A (F4-C4)		Sensor B (C4-P4)		Fusion	
		KNN	RF	KNN	RF	KNN	RF
EEG Frame 1	0.00-0.25	53.90	72.09	52.26	63.27	53.12	79.40
EEG Frame 2	0.25-0.50	54.33	73.00	55.68	64.92	53.94	79.08
EEG Frame 3	0.50-0.75	51.24	70.99	52.61	63.86	56.17	79.21
EEG Frame 4	0.75-1.00	54.88	72.20	51.99	62.08	56.45	78.70
EEG Frame 5	1.00-1.25	53.65	71.24	54.23	64.67	54.20	79.05
EEG Frame 6	1.25-1.50	55.38	71.86	52.99	64.57	55.49	79.38
EEG Frame 7	1.50-1.75	55.66	72.75	55.52	66.36	57.57	79.21
EEG Frame 8	1.75-2.00	55.98	71.66	54.65	64.75	56.22	78.08
EEG Frame 1S	0.00-1.00	57.52	73.75	52.96	64.01	55.75	79.50
EEG Frame 2S	1.00-2.00	51.80	72.65	53.84	66.82	56.21	79.46
EEG Frame (1S+2S)	0.00-2.00	54.12	**74.04**	52.79	**64.93**	54.97	***80.25***

best result (81.83%) for audio modality was obtained using the fusion of all prosodic functionals (mean and standard deviation of loudness, intensity and fundamental frequency).

Table 3: Result of 10-fold cross validation experiment 2 (A-Weighted F-Score %). Bold face indicates the best results.

Features	KNN	RF
Intensity	74.37	NaN
Loudness	75.89	75.24
Fundamental Frequency (fo)	67.49	68.63
Audio Fusion	68.29	**81.83**
Audio + EEG Frame 1	53.44	86.38
Audio + EEG Frame 2	54.87	86.72
Audio + EEG Frame 3	56.56	***86.76***
Audio + EEG Frame 4	56.48	86.39
Audio + EEG Frame (1S+2S)	54.93	85.95

Experiment 3: The results of experiments 1 and 2 suggest that the EEG signal of all eight frames and their combinations is predictive of on- and off-talk utterances, scoring well above the 50% baseline. It was also observed that the fusion EEG signals from both sensors improves accuracy. Therefore, we fused the acoustic information with the EEG features of both sensors (the latter 2s before articulation, as in Experiment 1), and trained a new model. The fusion of EEG and prosodic features improved the results as expected, by about 5% with respect to the best performing model of experiments 1 and 2. A summary of results is shown in Table 3.

Table 4: Confusion Matrix of the best results obtained in the three experiments

	Fusion		Audio		EEG	
	Off-Talk	On- Talk	Off-Talk	On- Talk	Off-Talk	On- Talk
Off-Talk	407	146	378	175	391	162
On- Talk	40	1086	82	1044	126	1000

5.1. Discussion

The classification results show that EEG features prior to articulation can effectively classify on- and off-talk. In our previous study [11], the EEG features during articulation (overt seech) provides a result of 74.80% using discrete wavelet transform and this study (covert speech) improves the results up to 80.25%. While our current method would still require 10 ms of speech detection for the analysis of (buffered) pre-articulation EEG to be triggered (see Figure 1), it still represents a substantial time improvement in comparison to audio-only off-talk detection and our previous study [11], as shown in Figure 3. In terms of brain location sources, the EEG signal from the F4-C4 channel (sensor A) provides better results (74.04%) than the signal from the C4-P4 channel (sensor B, 66.82%). We tested these differences in predictive accuracy using the mid-p-value McNemar test with a null hypothesis that sensor A and sensor B have equal accuracy for predicting the target (on- and off-talk). The statistical test rejects the null hypothesis ($p = 0.01$).

The RF classifier provides better results than KNN using EEG features. We also compared these methods using the mid-p-value McNemar test with a null hypothesis that the KNN classifier and the RF classifier have equal predictive accuracy. The statistical test rejects the null hypothesis ($p = 3.27 \times 10^{-39}$).

To gain further insight into timings, we have investigated different intervals of the EEG signal within the 2-second window prior to articulation. The most discriminative time period in the EEG signal for classification is frame 1 (0.00–0.25 seconds before utterance), and the fusion of 8 frames (2 seconds prior to articulation) yields an increase in performance. Fusing both sensor features improves the performance even further. The confusion matrices of the best results obtained in these three experiments are shown in the Table 4. We compared the best results of the three experiments under a null hypothesis that audio, EEG and fusion features have equal accuracy for predicting on- and off-talk. The mid-p-value McNemar test rejected this null hypothesis for 'EEG and fusion' ($p_{Exp.1-Exp3} = 2.73 \times 10^{-12}$), and 'audio and fusion' ($p_{Exp2-Exp3} = 4.48 \times 10^{-7}$) but was unable to reject the null hypothesis for 'EEG and audio' ($p_{Exp2-Exp1} = 0.11$).

The prosodic features produce only slightly better results than the EEG features, while the fusion of EEG and acoustic features improves the accuracy with respect to both. However, it should be noted that the EEG system has a much quicker response time (RTEEG) compared to the prosodic system (RTAudio), as depicted in Figure 3. If we assume that the processing time (as depicted in Figure 3) is 0 ms then the $RTEEG = 10$ ms (first 10 ms of speech articulation) and *RTAudio = duration of speech utterance*, and duration of speech utterances is not constant as shown in Table 1. This advantage of EEG is particularly relevant in ASR applications to natural speech, where fast response is essential to the correct processing of speech input. There may be other factors than off-talk detection that affect response time of a speech driven system. The improvement potential of the proposed method lies in the difference between off-talk detection in EEG verses in speech (where the system needs to record an entire utterance in order to make a decision).

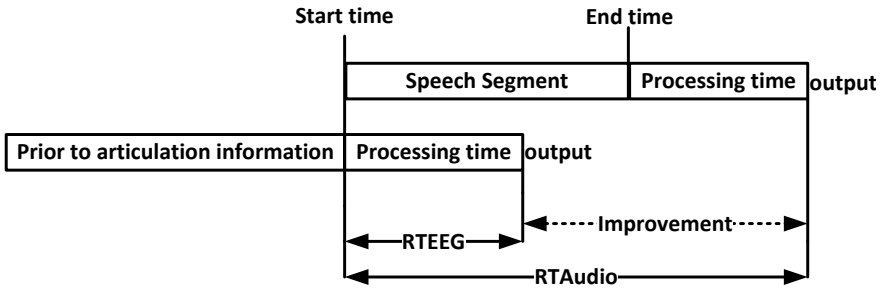

Figure 3: The baseline of the response time (RTAudio) and the proposed system response time (RTEEG). The *output* is the predicted label and *processing time* is the time taken by a machine's processor for classification purpose.

To further explore the relations among the best results of each experiment, we drew the Venn diagram shown in Figure 4. The blue ('Target') circle represents the labels (target), yellow ('Audio') circle represents the predicted labels using audio, green ('EEG') circle represents the predicted labels using EEG and the red ('Fusion') circle represents the predicted labels using fusion of audio and EEG. From this Venn diagram, it can be seen that there are 1207 instances which are correctly recognised by all three experiments (EEG, Audio and fusion of 'EEG and Audio'). However there are 75 instances (70 off-talk instances and 5 on-talk instances) which have not been recognised. These instances belong to S1 (7 off-talk), S2 (2 on-talk), S3 (20 off-talk), S4 (1 on talk), S5 (25 off-talk), S6 (3 off-talk and 2 on-talk), S8 (2 off-talk), S9 (1 off-talk), S10 (6 off-talk) and S13 (6 off-talk). The fusion is able to recognise 11 instances (7 on-talk and 4 off-talk) correctly which have not been recognised by EEG or audio alone. These belong to S4 (1 off-talk), S5 (2 off-talk and 1 on-talk), S8 (1 off-talk and 2 on-talk), S10 (1 on-talk) and S11 (3 on-talk). It is observed that off-talk (70 instances) is misclassified more frequently than on-talk (5 instances), but this is not true for all subjects as S2, S4, S7, S11 and S12 off-talk instances are correctly captured by the model. There are two subjects whose off-talk behaviour is difficult to capture using any modality (EEG, audio and fusion). They are S3 (20 out of 40 instances misclassified) and S5 (25 out of 78 instances are misclas-

sified). We note that contrary to the majority of participants, who produced no instances of other-talk at all, S3 and S5 produced a substantial amount of other-talk (75% and 42%, respectively). Further research and data collection are needed to investigate the question of whether other-talk is fundamentally distinct from self-talk, as these differences of model performance on S3 and S5 seem to indicate.

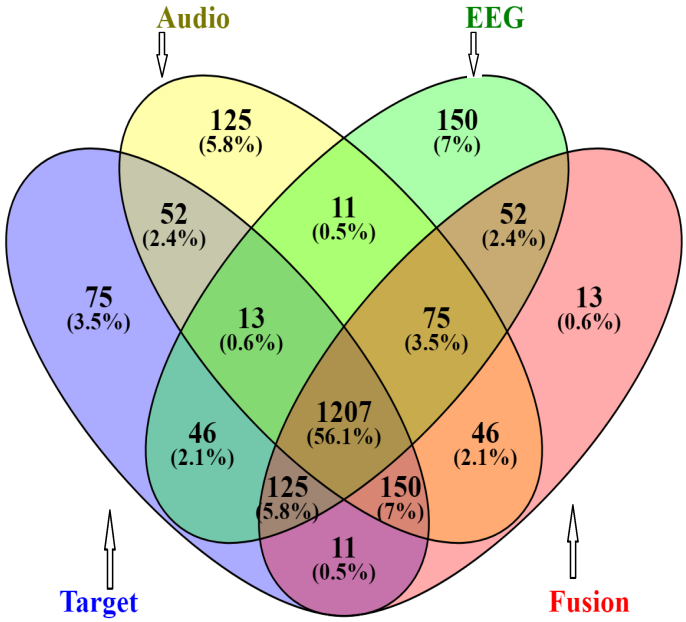

Figure 4: Venn Diagram displaying the shared relationship between the best results of the three experiments.

As regards speech features alone, the loudness feature provides better results (75.24%) than other prosodic features, highlighting the importance of speech volume variations in distinguishing between on- and off-talk. This is consistent with the observation by Batliner *et al.* that users tend to interact with an ASR system as they would with a person who has a hearing impairment [3]. However, it should be noted that we collected the data in a controlled acoustic environment, and therefore prosodic models may perform less well in a more realistic, noisy environment.

As with most studies involving the use of EEG in interactive situations, our study has limitations. The number of participants is relatively small (though not unusually so for an EEG study) and drawn from a narrow group (university students and researchers). Nevertheless, it should be noted that the number of labelled talk events (1682) is sufficient to enable inductive learning of these classes through the EEG features and learning method used. Another limitation is the use of a push-to-talk button to activate the ASR. While the actual act of pushing the button could not have confounded the analysis by generating mechanical EEG noise (as the classifier input signal was collected prior to the pushing of the button) it remains possible that the EEG signal reflected motor-control neural activity related to preparation for pushing the button, rather than preparation for on-talk. Confounding due to the latter is unlikely since EEG signals relating to limb movements seem to be restricted to channels C3 and CZ [24], none of which were used in the present study. In hindsight, if we had used more EEG channels we might have been able to resolve this issue. However, a different experimental design or further experiments may be necessary to minimise the possibility of confounding in future studies. An alternative design might involve the use of a voice activation prompt rather than push-to-talk. This, however, would introduce preparation for uttering the ASR activation prompt as a potential confounder. An additional experiment could involve participants communicating with a push-to-talk interface under two conditions: with ASR (as in our study) or directly. Capturing EEG in interactive speech communication remains challenging, and further research is needed.

Despite these limitation, the fact that the results of models created from pre-articulator EEG data practically match those based on speech features of the actual utterances (81.83% and 80.25%, respectively) suggest interesting possibilities for further experimental exploration. It is possible that the results obtained from the EEG signal reflect prosodic information processing in the brain, as the right hemisphere is thought to be responsible for the control of speech prosody. If this is the case, our results would be consistent with the hypothesis that the intonation pattern of produced speech is defined before articulation, as

suggested by previous studies [38, 39, 40, 41, 42, 43, 13]. While more research is needed in this area, by providing for the first time (to the best of our knowledge) evidence of consistent and distinguishable EEG activity prior to the articulation of prosodically distinct utterance types, our study may provide useful methods and models for future studies in this area.

In terms of practical applications (for instance, in BCI-enhanced interactive speech systems), the fact that models based on EEG signal features predict off-talk instances more accurately than prosodic features (Table 4) makes the EEG signal a better candidate than audio signal for situations where the misclassification cost for off-talk is higher than on-talk.

The failure of system components (e.g. ASR) results in a kind of behaviour that is not common in human-human communication, as discussed above. Therefore, we may assume that a brain-computer interface (where there is no overt speech) might experience the same situations (the brain signal reading components fails) that results in a neural activity (we might call such activity "off-thoughts") which should not be processed by the system. While the proposed models may also work in a covert speech situation (on- and off-thoughts), the brain has different activity patterns for overt and covert speech [56, 57], which may cause a decrease in accuracy of the proposed models for those kinds of brain-computer interfaces.

6. Conclusion

The EEG signal from the right hemisphere of the brain is able to classify 'on- and off-talk' at an accuracy of 80.25%; with the F4-C4 channel of the 10-20 system providing better results than the C4-P4 channel. For on- and off-talk detection, accuracy based on EEG alone is practically as high as detection based on prosodic features. This result could have interesting practical implication, given that the EEG signal in our data is captured in a fairly unconstrained interaction setting, suggesting the possibility that this technology could be deployed "in the wild" in speech-based interactive systems. Also interesting is the fact

that the fusion of prosodic and EEG information resulted in an improvement of performance. This indicates that whatever information is contributed by the EEG features regarding brain activity during the preparatory phase of speech production, this information is partially complementary to the produced speech itself.

Prosodic features extracted from the whole utterance add latency because the system needs to wait until the utterance is finished, while EEG features extracted prior to articulation do not have this problem. However, the current system still relies on detection of 10 ms of audio to trigger the start of processing of buffered EEG information corresponding to the pre-articulation period. A possible topic for further investigation is the use of EEG in settings that do not rely on this short audio detection interval.

Although the results presented in this paper point to promising directions in the use of BCI in interactive speech-based systems, more data and research are needed to elucidate the possible mechanisms behind these results. As the setting we used for data collection is fairly unconstrained, it is possible that factors other than brain activity related to speech articulation, such as visual or haptic feedback, might have confounded the EEG signal. Further investigation, with data collected in more constrained settings, is necessary for a closer examination of such factors. In future work we also plan to investigate the use of higher EEG frequency bands, the detection of on-talk using mid-line and left hemisphere signals which encode activation information from the phonological and motor area, and the task of distinguishing self-talk from other-talk.

7. Acknowledgement

This research has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 769661, SAAM project at the University of Edinburgh, UK, and "ADAPT 13/RC/2106" project (http://www.adaptcentre.ie/) in the Speech Communication Lab at Trinity College Dublin, the University of Dublin, Ireland.

References

[1] T. J. Tsai, A. Stolcke, M. Slaney, A study of multimodal addressee detection in human-human-computer interaction, IEEE Transactions on Multimedia 17 (9) (2015) 1550–1561. doi:10.1109/TMM.2015.2454332.

[2] A. Batliner, C. Hacker, E. Nöth, To talk or not to talk with a computer: On-Talk vs. Off-Talk, in: How People Talk to Computers, Robots, and Other Artificial Communication Partners, SFB/TR 8 Spatial Cognition, Hansewissenschaftskolleg, Delmenhorst, Germany, 2006, pp. 79–100.

[3] A. Batliner, C. Hacker, E. Nöth, To talk or not to talk with a computer: Taking into account the user's focus of attention, Journal on multimodal user interfaces 2 (3–4) (2009) 171–186.

[4] D. Oppermann, F. Schiel, S. Steininger, N. Beringer, Off-Talk — a Problem for Human-Machine-Interaction?, in: Proceedings of EUROSPEECH 2001 Scandinavia: the 7th European Conference on Speech Communication and Technologyand the 2nd INTERSPEECH Event, ISCA, Aalborg, Denmark, 2001, pp. 2197–2200.

[5] H. P. Branigan, M. J. Pickering, J. Pearson, J. F. McLean, Linguistic alignment between people and computers, Journal of Pragmatics 42 (9) (2010) 2355–2368.

[6] K. Fischer, How people talk with robots: Designing dialog to reduce user uncertainty, AI Magazine 32 (4) (2011) 31–38.

[7] L. Cerrato, A. Hayakawa, N. Campbell, S. Luz, A Speech-to-Speech, Machine Translation Mediated Map Task: An Exploratory Study, Springer International Publishing, Cham, 2016, pp. 53–64. doi:10.1007/978-3-319-33500-1_5.

[8] A. Hayakawa, L. Cerrato, N. Campbell, S. Luz, A Study of Prosodic Alignment In Interlingual Map-Task Dialogues, in: The Scottish Consortium

for ICPhS 2015 (Ed.), Proceedings of ICPhS XVIII (18th International Congress of Phonetic Sciences), no. 0760.1.5, University of Glasgow, Glasgow, United Kingdom, 2015.

[9] S. Goldwater, D. Jurafsky, C. D. Manning, Which words are hard to recognize? prosodic, lexical, and disfluency factors that increase speech recognition error rates, Speech Communication 52 (3) (2010) 181–200.

[10] A. Hayakawa, F. Haider, S. Luz, L. Cerrato, N. Campbell, Talking to a system and oneself: A study from a Speech-to-Speech, Machine Translation mediated Map Task, in: Proceedings of Speech Prosody 2016 (SP8), ISCA, Boston, Massachusetts, USA, 2016, pp. 776–780.

[11] F. Haider, H. Akira, S. Luz, V. Carl, N. a. Campbell, On-talk and off-talk detection: A discrete wavelet transform analysis of electroencephalogram, in: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, Calgary, Canada, 2018, pp. 960–964.

[12] A. R. Sereshkeh, R. Trott, A. Bricout, T. Chau, Eeg classification of covert speech using regularized neural networks, IEEE/ACM Transactions on Audio, Speech, and Language Processing 25 (12) (2017) 2292–2300. doi:10.1109/TASLP.2017.2758164.

[13] W. J. Levelt, Speaking: From intention to articulation, MIT press, 1989.

[14] M. van Turennout, P. Hagoort, C. M. Brown, Electrophysiological evidence on the time course of semantic and phonological processes in speech production., Journal of Experimental Psychology: Learning, Memory, and Cognition 23 (4) (1997) 787–806. doi:10.1037/0278-7393.23.4.787.

[15] J. Kreitewolf, A. D. Friederici, K. von Kriegstein, Hemispheric lateralization of linguistic prosody recognition in comparison to speech and speaker recognition, Neuroimage 102 (2014) 332–344.

[16] K. M. Heilman, S. A. Leon, J. C. Rosenbek, Affective aprosodia from a medial frontal stroke, Brain and language 89 (3) (2004) 411–416.

[17] P. F. MacNeilage, L. J. Rogers, G. Vallortigara, Origins of the left & right brain, Scientific American 301 (1) (2009) 60–67.

[18] A. D. Friederici, K. Alter, Lateralization of auditory language functions: a dynamic dual pathway model, Brain and language 89 (2) (2004) 267–276.

[19] B. E. Shapiro, M. Danly, The role of the right hemisphere in the control of speech prosody in propositional and affective contexts, Brain and Language 25 (1) (1985) 19–36. doi:10.1016/0093-934X(85)90118-X.

[20] S. Weintraub, M.-M. Mesulam, L. Kramer, Disturbances in prosody: A right-hemisphere contribution to language, Archives of Neurology 38 (12) (1981) 742–744. doi:10.1001/archneur.1981.00510120042004.

[21] E. D. Ross, M.-M. Mesulam, Dominant language functions of the right hemisphere?: Prosody and emotional gesturing, Archives of Neurology 36 (3) (1979) 144–148. doi:10.1001/archneur.1979.00500390062006.

[22] A. Flinker, A. Korzeniewska, A. Y. Shestyuk, P. J. Franaszczuk, N. F. Dronkers, R. T. Knight, N. E. Crone, Redefining the role of brocas area in speech, Proceedings of the National Academy of Sciences 112 (9) (2015) 2871–2875.

[23] S. C. Blank, S. K. Scott, K. Murphy, E. Warburton, R. J. Wise, Speech production: Wernicke, Broca and beyond, Brain 125 (8) (2002) 1829–1838. doi:10.1093/brain/awf191.

[24] C. Neuper, G. Pfurtscheller, Evidence for distinct beta resonance frequencies in human eeg related to specific sensorimotor cortical areas, Clinical Neurophysiology 112 (11) (2001) 2084–2097.

[25] A. Delorme, S. Makeig, EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis, Journal of Neuroscience Methods 134 (1) (2004) 9–21.

[26] R. D O'Donnell, J. Berkhout, W. R. Adey, Contamination of scalp EEG spectrum during contraction of cranio-facial muscles, Electroencephalography and Clinical Neurophysiology 37 (2) (1974) 145–151.

[27] I. I. Goncharova, D. J. McFarland, T. M. Vaughan, J. R. Wolpaw, EMG contamination of EEG: spectral and topographical characteristics, Clinical neurophysiology 114 (9) (2003) 1580–1593.

[28] S. Kumar, Y. Narayan, T. Amell, Power spectra of sternocleidomastoids, splenius capitis, and upper trapezius in oblique exertions, The Spine Journal 3 (5) (2003) 339–350.

[29] E. Criswell, Cram's introduction to surface electromyography, Jones & Bartlett Publishers, Sudbury, Massachusetts, USA, 2010.

[30] B. M. Schmitt, T. F. Münte, M. Kutas, Electrophysiological estimates of the time course of semantic and phonological encoding during implicit picture naming, Psychophysiology 37 (4) (2000) 473–484. doi:10.1111/1469-8986.3740473.

[31] B. M. Schmitt, K. Schiltz, W. Zaake, M. Kutas, T. F. Münte, An electrophysiological analysis of the time course of conceptual and syntactic encoding during tacit picture naming, Journal of Cognitive Neuroscience 13 (4) (2001) 510–522. doi:10.1162/08989290152001925.

[32] R. Abdel Rahman, M. van Turennout, W. J. Levelt, Phonological encoding is not contingent on semantic feature retrieval: an electrophysiological study on object naming., Journal of Experimental Psychology: Learning, Memory, and Cognition 29 (5) (2003) 850–860. doi:10.1037/0278-7393.29.5.850.

[33] C. C. Duncan-Johnson, B. S. Kopell, The stroop effect: brain potentials localize the source of interference, Science 214 (4523) (1981) 938–940. doi:10.1126/science.7302571.

[34] M. Liotti, M. G. Woldorff, R. Perez, H. S. Mayberg, An ERP study of the temporal course of the Stroop color-word interference effect, Neuropsychologia 38 (5) (2000) 701–711. doi:10.1016/S0028-3932(99)00106-2.

[35] J. L. Stone, Paul Broca and the first craniotomy based on cerebral localization, Journal of Neurosurgery 75 (1) (1991) 154–159. doi:10.3171/jns.1991.75.1.0154.

[36] H. A. Whitaker, A Model for Neurolinguistics, University of Rochester, 1970.

[37] D. W. McAdam, H. A. Whitaker, Language production: Electroencephalographic localization in the normal human brain, Science 172 (3982) (1971) 499–502.

[38] J. K. Bock, Toward a cognitive psychology of syntax: Information processing contributions to sentence formulation., Psychological Review 89 (1) (1982) 1–47. doi:10.1037/0033-295x.89.1.1.

[39] G. S. Dell, A spreading-activation theory of retrieval in sentence production., Psychological Review 93 (3) (1986) 283–321. doi:10.1037/0033-295x.93.3.283.

[40] M. F. Garrett, The Analysis of Sentence Production, Psychology of Learning and Motivation 9 (1975) 133–177. doi:10.1016/S0079-7421(08)60270-4.

[41] M. F. Garrett, Processes in language production., Linguistics: The Cambridge survey, Vol. 3, Cambridge University Press, 1988, pp. 69–96.

[42] G. Kempen, Conceptualizing and Formulating in Sentence Production., Erlbaum, 1977, pp. 259–274.

[43] G. Kempen, E. Hoenkamp, An incremental procedural grammar for sentence formulation, Cognitive Science 11 (2) (1987) 201–258. doi:10.1207/s15516709cog1102_5.

[44] H. H. Jasper, The ten twenty electrode system of the international federation, Electroencephalography and clinical Neurophysiology 10 (1958) 371–375.

[45] A. Hayakawa, S. Luz, N. Campbell, Talking to a System and Talking to a Human: A Study from a Speech-to-Speech, Machine Translation Mediated Map Task, in: Proceedings of INTERSPEECH'16: the 17th Annual Conference of the International Speech Communication Association, ISCA, San Francisco, California, USA, 2016, pp. 1422–1426. doi:10.21437/Interspeech.2016-1623.

[46] A. Hayakawa, N. Campbell, S. Luz, Interlingual Map Task Corpus Collection, in: Proceedings of INTERSPEECH'14: the 15th Annual Conference of the International Speech Communication Association, ISCA, Singapore, 2014, pp. 189–191.

[47] A. H. Anderson, M. Bader, E. G. Bard, E. Boyle, G. Doherty, S. Garrod, S. Isard, J. Kowtko, J. McAllister, J. Miller, C. Sotillo, H. S. Thompson, R. Weinert, The hcrc map task corpus, Language and Speech 34 (4) (1991) 351–366. doi:10.1177/002383099103400404.
URL http://las.sagepub.com/content/34/4/351

[48] A. Hayakawa, S. Luz, L. Cerrato, N. Campbell, The ILMT-s2s Corpus — A Multimodal Interlingual Map Task Corpus, in: N. Calzolari, K. Choukri, T. Declerck, M. Grobelnik, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, S. Piperidis (Eds.), Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), European Language Resources Association (ELRA), Paris, France, 2016, pp. 605–612.

[49] J. Gotman, High frequency oscillations: The new EEG frontier?, Epilepsia 51 (0 1) (2010) 63–65. doi:10.1111/j.1528-1167.2009.02449.x.
URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3786932/

[50] T. F. Oostendorp, J. Delbeke, D. F. Stegeman, The conductivity of the human skull: results of in *vivo* and in *vitro* measurements, IEEE

transactions on bio-medical engineering 47 (11) (2000) 1487–1492. doi:10.1109/TBME.2000.880100.

[51] F. Eyben, F. Weninger, F. Gross, B. Schuller, Recent developments in openSMILE, the munich open-source multimedia feature extractor, in: Proceedings of the 21st ACM International Conference on Multimedia, MM '13, ACM, New York, NY, USA, 2013, pp. 835–838. doi:10.1145/2502081.2502224.

[52] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, et al., The INTERSPEECH 2013 Computational Paralinguistics Challenge: Social Signals, Conflict, Emotion, Autism, in: Proceedings of INTERSPEECH'13: the 14th Annual Conference of the International Speech Communication Association, ISCA, Lyon, France, 2013, pp. 148–152.

[53] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830.

[54] A. Liaw, M. Wiener, Classification and Regression by randomForest, R News 2 (3) (2002) 18–22.

[55] L. Breiman, Manual on setting up, using, and understanding random forests v3.1, University of California Berkeley, CA, USA.

[56] I. K. Christoffels, E. Formisano, N. O. Schiller, Neural correlates of verbal feedback processing: an fMRI study employing overt speech, Human Brain Mapping 28 (9) (2007) 868–879. doi:10.1002/hbm.20315.

[57] X. Pei, E. C. Leuthardt, C. M. Gaona, P. Brunner, J. R. Wolpaw, G. Schalk, Spatiotemporal dynamics of electrocorticographic high gamma activity dur-

ing overt and covert word repetition, NeuroImage 54 (4) (2011) 2960–2972. doi:10.1016/j.neuroimage.2010.10.029.