DOI: 10.17469/O2102AISV000019

ANNA DORA MANCA, GIORGIO DE NUNZIO, MIRKO GRIMALDI

EEG-Based Recognition of Silent and Imagined Vowels

This work proposes a framework for future Silent Speech Interfaces (SSI) based on non-invasive EEG recordings. Specifically, the information embedded in the brain signals related to the production – overt, covert and imagined production – of the Italian vowels /a/ and /i/ allowed to distinguish the vowels relying on discriminative features calculated by the Ambiguity Function in the context of time-frequency analysis, and ranked by the Fisher contrast. The vowels were classified by using a multilayer feed-forward ANN. Overall, intra-subject classification accuracies, as measured by the area under the ROC curve, ranged from 0.84 to 0.96 for overt production, from 0.83 to 0.96 for covert production, and from 0.89 to 0.98 for imagined vowels. Results indicate significant potential for the use of speech prosthesis controllers for clinical and military applications.

Key words: EEG, speech, neural network, vowels, ambiguity function.

Introduction

Several electrocorticographic, electric and magnetic investigations showed that from the brain signals important information for the discrimination of spoken and perceived speech sounds can be extracted (Bouchard, Mesgarani, Johnson & Chan, 2013; Pei, Barbour, Leuthardt & Schalk, 2011; Wang, Perreau-Guimaraes, Carvalhaes & Suppes, 2012; Obleser, Lahiri & Eulitz, 2004; Obleser, Scott & Eulitz, 2006; Scharinger, Idsardi & Poe, 2011; Luo, Poeppel, 2012). It seems also feasible to recognize neuronal traces of non-audible speech sounds evoked during imagined and mouthed (covert) speech processes; the idea is that the mechanisms underlying such operations rely on the same neuronal substrates involved in the processes of overt speech production, thus, tracing and detecting the related cortical signals seems actually plausible (Tian, Poeppel, 2010).

In the last decades, different attempts have been made to decoding the EEG signals associated to non-audible speech mainly with the interest of testing new methodologies for speech recognition systems such as Silent Speech Interfaces (SSIs). These systems acquire data from brain activity associated with overt and covert speech performance and synthesize information by reproducing a digital representation of the signals necessary for their functioning (for a detailed description of SSIs see Denby, Shultz, Honda, Hueber & Gilbert, 2010). The potential usability of these applications is enormous – from medical to military environments – and leads to explore methodological approaches in support of new portable and user friendly EEG-based SSIs. For researchers, this means resolving some critical steps such as the extraction of the most discriminative features of the brain signals

associated with speech sounds and the choice of accurate classification procedures (Šťastný, Sovka & Stančák, 2003). To date, several approaches have been proposed.

First works go back to the end of 90s when Suppes and colleagues (Suppes, Lu & Han, 1997) succeed in decoding electric and magnetic brain signals recorded during imagined words; some years later however, Porbadnigk and colleagues showed that the classification rates were biased for the effects of the temporal artifacts caused by the experimental protocol (Porbadnigk, Wester & Calliess, 2009). Subsequent studies focused mostly on decoding imagined phonemes. For example, D'Zmura and colleagues (D'Zmura, Deng, Lappas, Thorpe & Srinivasan, 2009) showed with spectral analysis techniques that the brain frequency bands were informative for non-audible sounds classification. They recorded the EEG activity of four subjects performing two imagined syllables /ba/ and /ku/ with three different rhythms and achieved a classification accuracy of 87% only for one of the four subjects included in the experiment. Working on the same data set, Brigham and Kumar extended the result demonstrating that classification rates remarkably improved after an intensive technique of artifacts rejection. Here, the features were extracted by autoregressive coefficients and the classification was done with a k-Nearest Neighbor classifier (Brigham, Kumar, 2010). Meanwhile, DaSalla et al. classified the neuronal activity of three healthy subjects associated to the imagined vocalization of the English vowels /a/ and /u/ as compared to a no-state control condition where subjects were at rest (DaSalla, Kambara, Sato & Koike, 2009). The authors applied spatial filters to the EEG time series and tested a support vector machine (SVM) for the classification of the tasks achieving overall good accuracies (/a/ vs. rest: 68%; /u/ vs. rest: 78%). The same EEG dataset was tested with other kinds of classification algorithms (see Santana, 2015; Iqbal, Shanir, Khan & Farooq, 2016) reporting similar accuracy percentages. Yet, the EEG activity evoked during the imagined production of couples of phonemes differing in patterns of vocal articulation was successfully classified (classification rate above 70%) by exploiting information embedded in spectrogram samples at specific brain frequencies (Chia, Hagedorna, Schoonovera & D'Zmura, 2011). Here, classification was done with a Naive Bayes and Linear Discriminant Analysis (LDA) classifiers. Similar results were found in pairwise classification using SVM of the Japanese vowels /a/ and /u/ (Matsumoto, Hori, 2014). To conclude, to the best of our knowledge, only Riaz and colleagues have discriminated EEG data of three subjects performing mouthing tasks of five different vowels, (a, e, i, o, and u). Results of the pairwise comparisons showed an average accuracy of around 75% with the best separation between vowels /a/ vs. /i/ and /e/ vs. /u/ (Riaz, Akhtar, Iftikhar, Khan & Salman, 2014).

In the present work, we explored a procedure for decoding brain signals associated to different experimental speech tasks: overt production (OP), covert production (CP or mouthing) and imagined production (IP) of the Italian vowels [a] and [i]. These two vowels are suitable for our explorative purposes as they are realized by maximally contrastive tongue gestures: /a/ is pronounced by lowering the tongue body and /i/ by raising the tongue body and advancing the tongue root. Our aims

were (i) to determine whether the patterns elicited by OP were elicited even in absence of audible speech signals (i.e., in CP and IP) and (ii) whether the EEG waves contained discriminant information for vowel classification. To do this, spectral analysis and the (symmetric) Ambiguity Function (AF) were used to represent the EEG signals, and a feed-forward Artificial Neural Network (ANN) was tested for vowel classification in each task. If covert and imagined speech conditions reveal as useful levels of investigation, then the framework may be implemented in new methodological approaches for the development of non-invasive SSIs.

1. *Methods*

1.1 Subjects

Twelve students of the University of Salento (Lecce, Italy) (7 males and 5 females, 25±3 years) participated in the experiment after providing a written informed consent. They were right-handed according to Handedness Edinburgh Questionnaire and none of them had any known neurological disorder or other significant health problem. The ethical committee of the local health authority of Lecce approved the study.

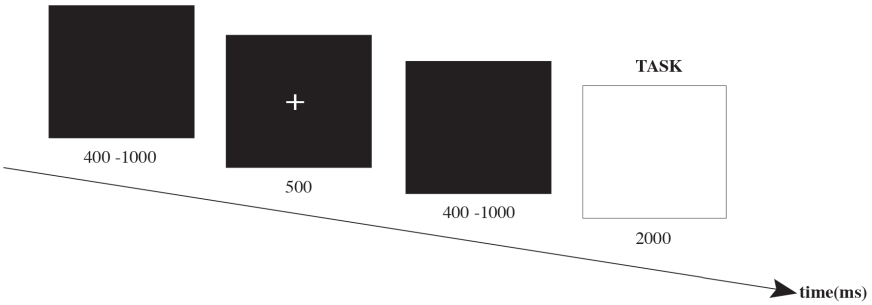
1.2 Experimental procedure

In separate runs, the participants performed three tasks: OP, CP and IP of the vowel /a/ and then of the vowel /i/. In the OP task, the subjects pronounced aloud the vowel, in the CP task, they mouthed the vowel without any emission of sound, and during IP, they had to imagine to produce the vowel without using articulatory muscles (i.e., inertial tongue and mandibular movements) and without uttering any audible sound. The order of the tasks was counterbalanced across participants; the order of the vowels was established before each task.

Each trial began with a black screen displayed for a random time (400-1000 ms) followed by a small white cross (500 ms) in the center of the computer monitor, used to suggest subjects to concentrate and prepare for the task. Another randomized time interval (400-1000 ms) preceded a white screen (2000 ms) which triggered the onset of the task. Each session consisted of 80 visual cues (white screen) and each trial had an average duration of about 3850 ms (Figure 1).

The subjects were instructed to perform as best as possible the experimental task while remaining completely still during imagined phoneme production. All participants took part in a training phase, which was identical to the experimental procedure to ensure an accurate task performance.

Figure 1 - *Schematic illustration of the paradigm employed in the present experiment*



1.3 Data acquisition

Continuous EEG was recorded with a 64-channel actiCAP (10-20 system), a sampling rate of 250 Hz and a band pass filter of 0.1-70 Hz (BrainProducts GmbH, Germany). The vertical Electro-oculogram (EOG) was recorded by means of two electrodes (same type as EEG) just above and below the right eye, and the horizontal EOG was recorded with the FT9 and FT10 electrodes. The online reference was at FCz and impedance was kept under 5 kΩ by electrogel conductant.

2. *Data pre-processing*

2.1 EEG analysis

Off-line signal processing was carried out with MATLAB and the software package EEGLab. Data were digitally filtered at 2–30 Hz, they were re-referenced to the right and left mastoids TP9-T910 and down-sampled to 100 Hz according to EEG studies on the classification of speech stimuli (Wang et al., 2012; Suppes, Han, Epelboim & Lu, 1999). Independent component analysis (ICA) was computed as a pre-processing step to remove muscular and ocular artifacts. A script was written to identify artefactual independent components (ICs) by exploiting their power spectral density (PSD) properties (Vos, Riès, Vanderperren, Vanrumste, Alario, Huffel & Burle, 2010). The basic assumption is that brain EEG signals have lower power at high frequencies whereas muscular EEG signals have higher power at high frequencies. Accordingly, we have considered as potential muscular artifacts the ICs whose average power between 15–30 Hz was at least twice as great as the one between 2-15 Hz. Similarly, keeping in mind that the ocular EEG signal power has a very narrow peak between 0–4 Hz, we considered possible ocular artifacts those ICs whose average power between 2–4 Hz was at least half of that between 4–30 Hz. Finally, we visually inspected the highlighted ICs. Components actually identified as artifacts were rejected, and the original EEG time-courses were reconstructed, using only the preserved ICA components. EEG epochs were extracted with reference to the white screen onset. Each epoch had a duration of 400 ms, 100 ms of pre-stimulus and 300 ms of stimulus.

2.2 Feature extraction

For an accurate characterization of non-stationary signals such as EEG data, Time-Frequency Representations (TFRs) are required. TFRs (Kozek, Hlawatsch, Kirchauer & Trautwein, 1994) are data processing methods in which signals are analyzed simultaneously in the time and frequency domains, in a 2D representation. The rationale for TFRs is that conventional methods as Fourier Transform assume the signals to be periodic or infinite in time, while many real-life signals, such as EEG time series, vary considerably. Known TFRs are the Short-Time Fourier Transform (STFT) and wavelet analysis. In this work, the (symmetric) Ambiguity Function (AF), i.e., the inverse Fourier transform of the Wigner-Ville distribution was proposed to represent EEG signals (Kozek et al., 1994) and an ANN (Haykin, 2008) was used for the vowel classification in the different tasks. EEG epochs for each subject were initially processed by time-frequency analysis in the doppler-delay ambiguity plane. Values of the ambiguity function in the plane were chosen as features for vowel recognition. The most discriminant points in the plane were identified by maximizing the Fisher contrast of the two classes, and the ambiguity values in those points formed the feature vector. A 2-layer, 5-hidden-neuron feedforward ANN was trained and validated for the recognition of the vowels, independently on each subject. ROC (Receiver Operating Characteristic) curves were calculated and the AUC (Area Under the Curve) values were derived as a classification accuracy measure.

The AF of signal x(u), denoted by Ax, is defined as:

$$A_x(\tau,\nu)=\int_{\Re} x\left(u+\frac{\tau}{2}\right)x^*\left(u-\frac{\tau}{2}\right)e^{-2\pi i\nu u}\,du \tag{1}$$

Where t is the time delay, n is the doppler frequency shift, and x* is the complex conjugate of x. The AF can be considered as an autocorrelation function in joint time-frequency domain, which transforms a signal to time delay and frequency shift plane (Ambiguity Plane). Its most useful properties are:

i. The AF modulus is independent of time and frequency shift, that is, if y is a time- and frequency-shifted copy of x: $y(t)=x(t-t_1)e^{2\pi i f_1 t}$, then $A_y(\tau,\nu)=A_x(\tau,\nu)\,e^{2\pi i(f_1\tau-t_1\nu)}$, so that $|A_y(\tau,\nu)|=|A_x(\tau,\nu)|$;
ii. The AF modulus is symmetric with respect to the origin: $|A_x(t,\nu)|=|A_x(-\tau,-\nu)|$. If x is real, then: $|A_x(\tau,\nu)|=|A_x(\tau,-\nu)|$ and $|A_x(-\tau,\nu)|=|A_x(-\tau,-\nu)|$, $|A_x(\mathrm{t},\nu)|=|A_x(-\tau,\nu)|$, $|A_x(\tau,-\nu)|=|A_x(-\tau,-\nu)|$;
iii. The largest AF value is in the axes origin, and equals signal energy: $\forall\tau,\nu$: $|A_x(\tau,\nu)|\le|A_x(0,0)|=\int|x(t)|^2\,\mathrm{dt}$;

Time shift and frequency shift invariance (property (i)) indicates that even if the arriving times and center frequencies of the signal vary from each other, the moduli of their AFs are the same. Therefore, extracting features from the ambiguity plane does not require time alignment and frequency transform. The symmetry properties of AF with real signals (property (ii)) allowed considering only a quarter of the

ambiguity plane without information loss. Some literature exists on the subject of AF applications to pattern recognition and signal classification, where discriminant features are taken from the Ambiguity plane. If the length of a signal (number of samples) is Ns, the AF of the signal is a Ns×Ns matrix, which is generally large. Therefore, it is convenient to project the ambiguity function to a lower-dimensional space. In some studies, (McLaughlin, Droppo & Atlas, 1997; Atlas, Droppo & McLaughlin, 1997; Gillespie, Atlas, 2001), kernel function methods were proposed that extracted features from the ambiguity plane by designing time-frequency kernel functions, which preserved the location of the ambiguity plane that maximized class separability. In Garcia et al. (Garcia, Ebrahimi & Vesin, 2003) and in Ebrahimi et al. (Ebrahimi, Vesin & Garcia, 2003), this method was applied to Brain-Computer Interfacing. As a means of reducing feature-space dimensionality (Garcia, Ebrahimi & Vesin, 2002) used the Fisher Contrast (or Fisher's discriminant ratio, FDR) to locate the N most discriminant locations on the ambiguity plane. Thus, N locations from the ambiguity plane are chosen, in such a way that the values in these locations are very similar for signals from the same class, but they vary significantly for signals from different classes. For our two-class classification of vowels, we followed this methodology, which proved simple but effective. The procedure consisted in determining the coordinates of a number of highest contrast points between two given TFRs in the ambiguity planes (representing the two classes), then using the values of the AF in those points as features for classification. Steps are as follows. Firstly, calculate the FDR for the training sets of the two classes (/a/ and /i/ vowels) in the ambiguity plane (doppler ν, delay τ), for each rebuilt EEG channel c:

$$K_{Fisher}(c,\tau,\nu)=\frac{\left|\overline{A}_{1,c}(\tau,\nu)-\overline{A}_{2,c}(\tau,\nu)\right|^2}{\overline{\overline{A}}^2_{1,c}(\tau,\nu)+\overline{\overline{A}}^2_{2,c}(\tau,\nu)} \tag{2}$$

In the above expression:

$$\overline{A}_{i,c}(\tau,\nu)=\frac{1}{n_i}\sum_{j=1}^{n_i}A_{x_j^{i,c}}(\tau,\nu) \tag{3}$$

$$\overline{\overline{A}}_{i,c}(\tau,\nu)=\frac{1}{n_i}\sum_{j=1}^{n_i}\left(A_{x_j^{i,c}}(\tau,\nu)-\overline{A}_{i,c}(\tau,\nu)\right)^2 \tag{4}$$

are respectively the mean and the variance of the AFs of all epochs (belonging to the training set), calculated for each class and for each channel: i indexes the two classes, n_i is the total number of signals for class i, $A_{x_j^{i,c}}$ is the AF of the j-th epoch of class i, in channel c. The rationale of using the FDR is to optimize the representation space by maximizing the value of $K_{Fisher}(c, \tau, \nu)$, which means increasing the distance between the mean of the two classes, while reducing intra-class dispersion. Secondly, chose a number of points Np in the AF planes, according to a criterion of maximum

discriminant power as measured by the KFisher contrast between the average AF planes for the two classes.

After some tests with a variable number of features (from 10 to some hundreds) we concluded that no important gain could be obtained by using too large Np values, so we set Np = 100 which appeared as a good compromise between accuracy and feature space dimensionality. Each point had coordinates $F_m(c_m, \tau_m, \nu_m)$, m = 1, ..., Np. Discriminant features were chosen as the AF calculated in $\{F_m\}$. Then, by considering the channels in which most frequently the features were chosen by the FDR, information about the most discriminant EEG electrodes was collected. The analysis of the EEG channels more frequently chosen by the FDR value put in evidence a noticeable variability between the subjects, but it allowed, anyway, to derive some interesting common information. In particular, the most discriminative sites for vowel discrimination were Cz, CP2, CP4, CP6 for the OP task, FT7-FT8-T7-T8-C6 for CP, and CP1-CP2-CP3 for the IP task.

2.3 Classification method

Each validation EEG epoch *x* was classified by a supervised classifier as class i, according to the set of features $AFx(c_m, \tau_m, \nu_m)$, m = 1, ..., Np. Supposing that the overall number of training trials for a subject is NT = NT1 + NT2 (the summation of trials for class 1 and 2 respectively, 160 per subject in our experiments), the classifier input is a NT×Np matrix (i.e. 160×100). Finally, a feed-forward ANN was chosen as the classifier (1 hidden layer with 5 hidden neurons, HNs). In intrasubject experiments, the training set was randomly divided into two subsets of equal cardinality, and the training-validation process was repeated 50 times, each time calculating the ROC curve and its AUC as a measure of accuracy. The result was a mean AUC with an associated error (the standard deviation). This step was repeated for each subject for the OP, CP, and IP tasks. Vowel classification was also performed intersubject in LOSO (Leave One Subject Out) cross validation.

3. *Results*

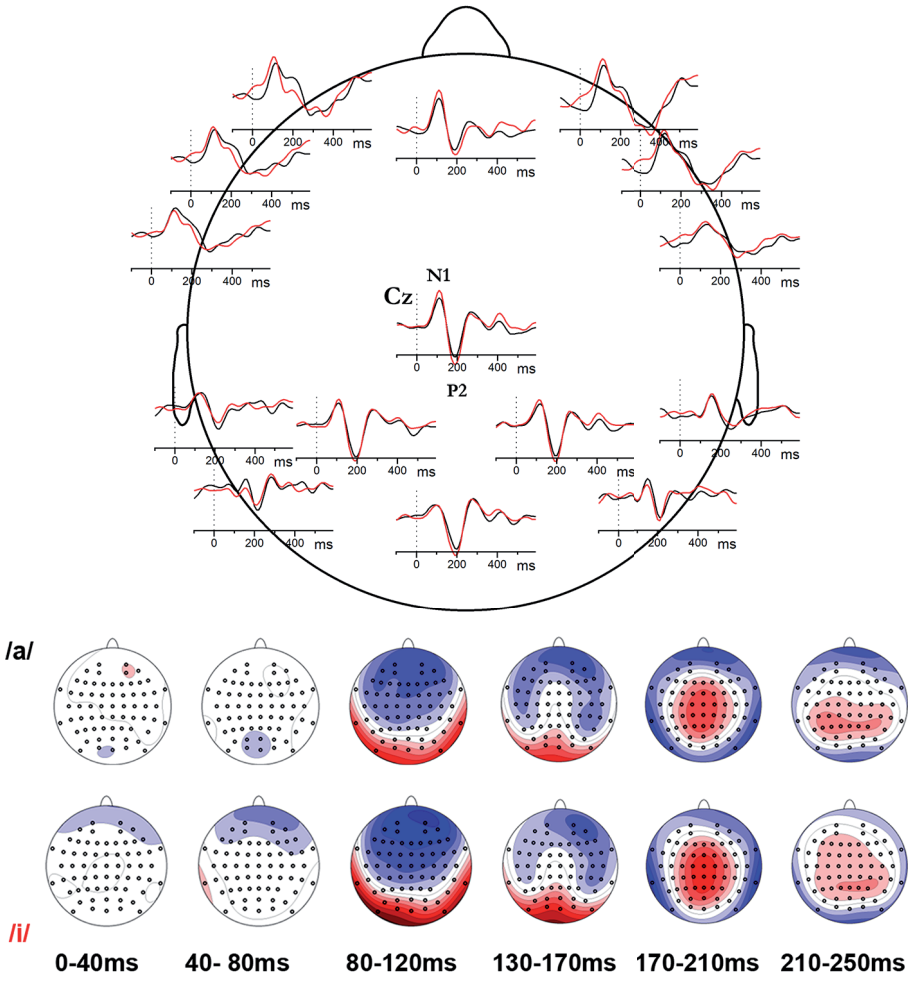
3.1 EEG patterns

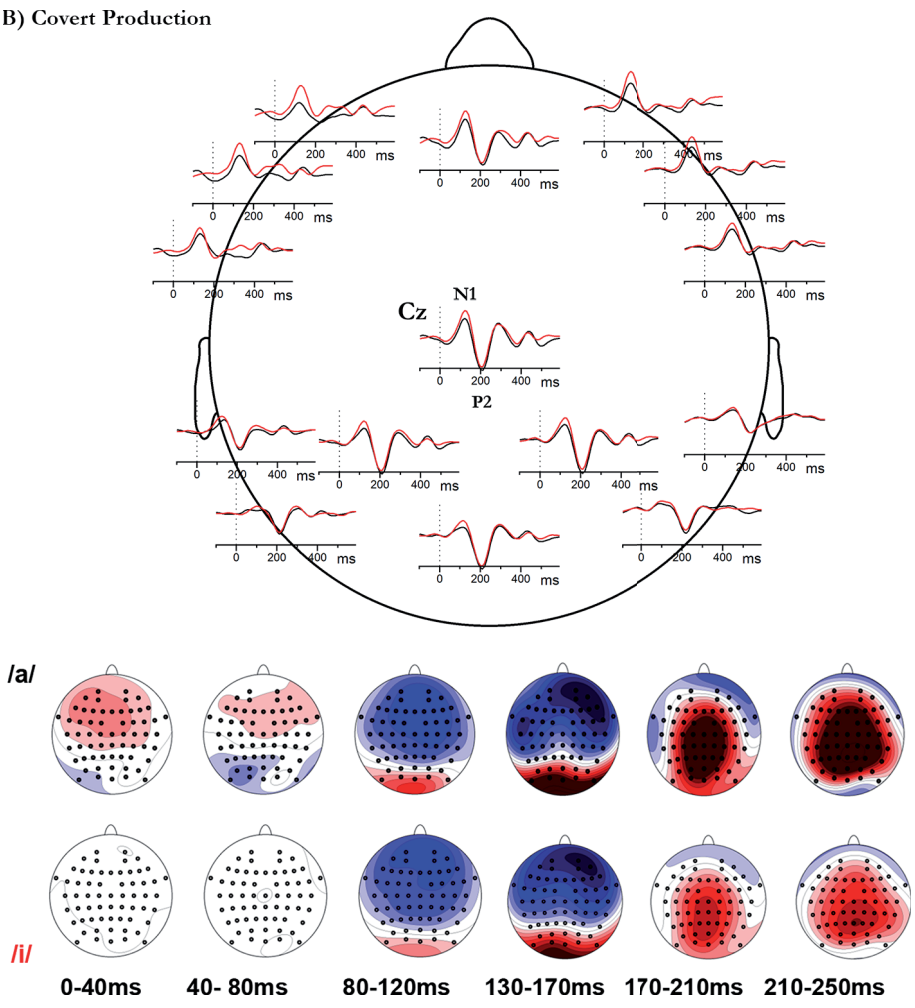
To visualize the responses to each vowel we calculated the average of all epochs in the three tasks separately. Figure 2 (A, B, C) shows the grand averages at the most important electrodes for vowel classification (Section 2.2); time 0 ms coincides with the appearance of the visual cue triggering the onset of the task execution. In each experimental condition, we recognized a negative trend reaching the most negative peak between 80-120 ms over the fronto-central electrodes, and a late positive shift peaking between 170-200 ms at central and parietal sites. These peaks resemble the typical neuronal audito-

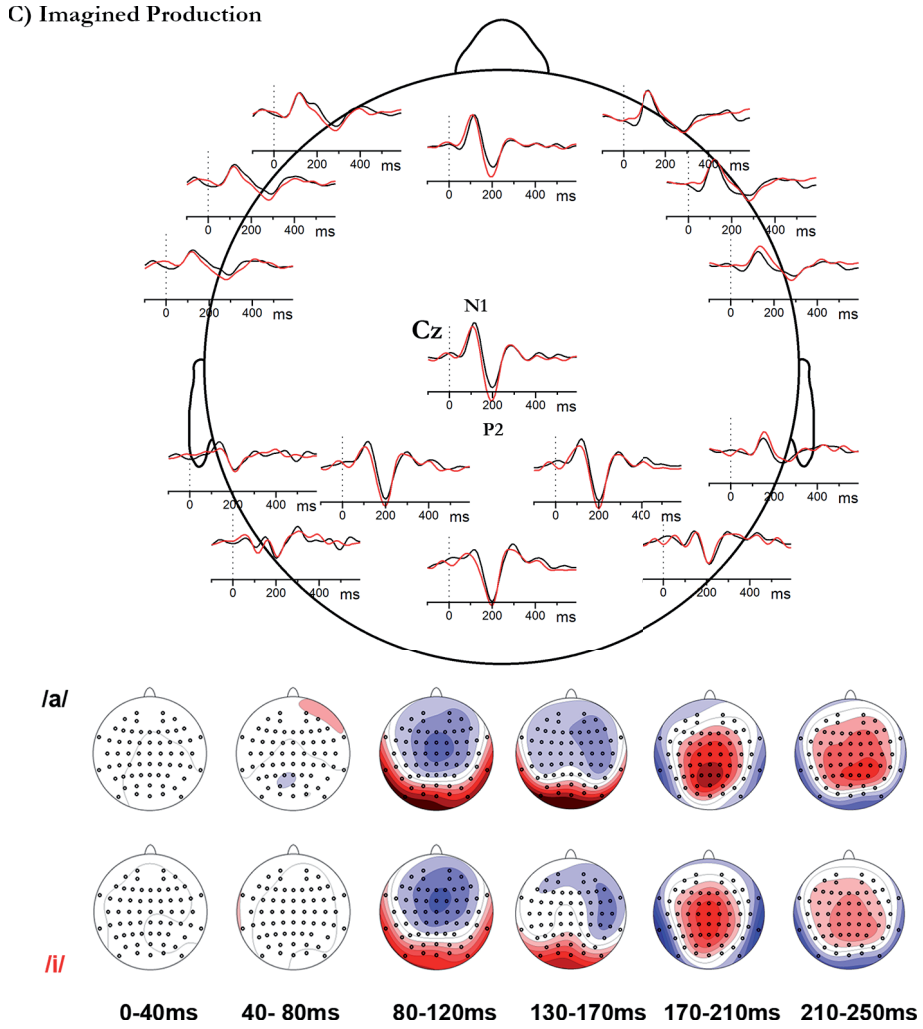
ry N1/P2 pattern (e.g., in Manca, Grimaldi, 2013). The mean amplitude of these peaks was calculated considering an interval of 60 ms centered at the maximum peak.

Figure 2 - *Grand average waveforms (N=12 subjects) at fronto-central and centro-parietal electrodes and topographic maps for the vowel /a/ (in black) and /i/ (in red)*

A) Overt Production




B) Covert Production
Cz
N1
P2
0 200 400 ms
/a/
/i/
0-40ms
40- 80ms
80-120ms
130-170ms
170-210ms
210-250ms



The presence of the N1 and P2 pattern of response was verified by a series of t-tests against zero at the midline electrodes FCz, Fz, and Cz ($p < 0.05$). Data were normality distributed ($p > 0.5$) as evaluated by a series of Shapiro-Wilk tests on the N1 and P2 amplitude and latency values at Cz electrode (n=12) where the components had the maximum distribution (Table 1).

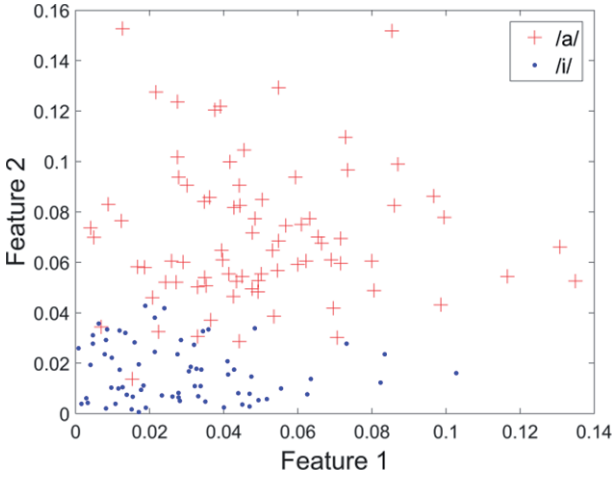
Table 1 - *Normality tests on N1 and P2 values for each experimental tasks*

	Overt Production		*Covert Production*		*Imagined Production*	
	Statistic	*Sig.*	*Statistic*	*Sig.*	*Statistic*	*Sig.*
N1 Amplitude /a/	,930	,376	,662	,089	,662	,089
N1 Latency /a/	,982	,990	,893	,129	,893	,129
P2 Amplitude /a/	,972	,926	,700	,098	,700	,098
P2 Latency /a/	,920	,282	,928	,355	,928	,355
N1 Amplitude /i/	,911	,221	,927	,347	,927	,347
N1 Latency /i/	,986	,998	,960	,785	,960	,785
P2 Amplitude /i/	,942	,520	,920	,286	,920	,286
P2 Latency /i/	,884	,098	,930	,375	,930	,375

3.2 Vowel classification

Vowel classification was obtained by a 1-hidden-layer feed-forward ANN with back-propagation (5 HNs). Tests with less or no HNs (in the hypothesis that the discrimination problem might be liner) gave poor results. Increasing the number of HNs gave no relevant accuracy improvement. As an example of the discriminant power of the chosen features, Figure 3 shows the scatter plot of the two classes (vowel /a/ and vowel /i/) in the plane of the two best features (named Feature 1 and Feature 2 in the graph), for the IP task of one of the subjects. The good class separation is evident.

Figure 3 - *Scatter plot of the two vowel classes in the plane of the two best features, for the IP task of one of the subjects*



AUC values for vowel classification, for each task and for each subject are reported in Table 2. AUC were averaged on 50 iterations of the ANN training/validation process, and standard deviations are reported as the uncertainties. The CP and IP

tasks showed an average classification accuracy (as measured by the ROC AUC) of 0.91 and 0.93 respectively, which suggests slightly better performance compared to the OP task (i.e., 0.89). This was statistically significant for OP vs IP comparison (according to a paired t-test applied to the mean values in Table 2, giving $p = 10^{-3}$). This finding needs anyway deeper investigation and confirmation.

Intersubject classification in LOSO cross validation was finally tested with very poor results (AUC about 0.50-0.60).

Table 2 - *Average AUC values and standard deviations calculated with 50 runs of the training/validation process, for each subject (S) and each task*

S	AUC – OP	AUC - CP	AUC – IP
1	0.85±0.03	0.95±0.02	0.93±0.03
2	0.85±0.04	0.96±0.02	0.89±0.03
3	0.88±0.03	0.85±0.04	0.92±0.03
4	0.88±0.03	0.95±0.02	0.94±0.02
5	0.87±0.05	0.93±0.03	0.91±0.03
6	0.95±0.02	0.84±0.05	0.94±0.02
7	0.83±0.05	0.90±0.03	0.93±0.02
8	0.94±0.02	0.95±0.02	0.96±0.02
9	0.96±0.02	0.91±0.03	0.96±0.02
10	0.93±0.03	0.95±0.02	0.97±0.02
11	0.93±0.03	0.86±0.05	0.98±0.02
12	0.87±0.03	0.90±0.03	0.89±0.03
Mean	0.89	0.91	0.93

4. *Discussion*

The current study shows that is feasible to recognize the features distinguishing the vowels /a/ and /i/ from information embedded in the EEG signals generated during covert and imagined production. Two main conclusions derived from the results.

First, the CP and IP tasks elicit similar neural responses to the OP showing the typical auditory N1/P2 responses to speech sounds (Näätänen, Picton, 1987) as in previous studies on Italian vowels. In particular, an event–related study on the perception and production processes of the same vowel pairwise (Manca, Grimaldi, 2013) suggested the N1/P2 pattern as index of the activation of auditory neurons to the linguistically relevant properties of sounds; yet, the generators of the auditory activity to the perceived vowels was localized in the supratemporal auditory cortex of both hemispheres (Manca, Di Russo & Grimaldi, 2015). Furthermore, models of speech production (Guenther, Hampson & Johnson, 1998; Tian, Poeppel, 2010) have established that during speech production two efferent copies – auditory and motor – are created from stored models of previous speech motor acts; when the speech command is executed, auditory feedback of the spoken sound is heard at the level of

the peripheral auditory system and processed through the central auditory pathway to the bilateral auditory temporal lobe (Tourville, Reilly & Guenther, 2008). Yes, further MEG studies revealed that the auditory cortical potentials at a latency of approximately 100 ms are modulated during speech execution (Gunji, Hoshiyama & Kakigi, 2000; Heinks-Maldonado, Nagarajan & Houde, 2006). In the present work, the topography of our waves resembles the same neuronal pattern revealing some hints of auditory activation also in the tasks where there exists no auditory feedback (Figure 2). That is, since the subjects in the present experiment were instructed to generate different forms of speech production, we can speculate that the early activity elicited during the CP and IP tasks (as described by N1/P2 pattern) represents the output of sensory-motor circuits along which the auditory system is activated even when motor activity is inhibited or absent. Anyway, the role played by the motor areas in speaking cannot be ruled out: studies on movement-related cortical potentials (Deecke, Engel, Lang & Kornhuber, 1986) reported negatives potentials with symmetric activities at 100 ms post vocalization that overlap the auditory activity (Gunji et al., 2000). In our study, the early frontal activity (from 80 to 170 ms) may be actually related to the motor act of speaking (as in OP and CP) that requires interconnection among the frontal, temporal, and parietal lobes of the brain (Guenther, 2007). Further investigations are needed to improve the understanding of the activities that are recruited in the speech production tasks and, in this perspective, it will be necessary to take into account recent intracranial investigations finding no activation of the Broca's area during actual articulation (Flinker, Korzeniewska, Shestyuk, Franaszczuk, Dronkers, Knight & Crone, 2015).

The second finding is that information extracted by the early dynamics contains sufficient discriminative features for the vowel cortical classification. As to features, D'Zmura (2009) used matched filters, DaSalla et al. (2009) and Matsumoto and Hori (2014) applied the Common Spatial Patterns (CSP) methods that exalts more discriminative EEG channels, taking variance as the discriminating feature. We used the Ambiguity Function that gave us a large number of features (the values of the Ambiguity Function at each point of the ambiguity planes) then reducing the feature space dimensionality by FDR.

As for the choice of the classifier, we used a feed-forward ANN to recognize the features distinguishing the vowels /a/ and /i/ and showed that a 2-layer, 5-HNs feed-forward ANN can be successfully trained for the intrasubject recognition of overt, covert and imagined vowel production, in line with previous studies using different approaches. Other studies preferred the use of SVMs as classifiers (DaSalla et al., 2009; Matsumoto, Hori, 2014; Riaz et al., 2014) that has some advantages: e.g., the guaranty of finding the global minimum during training, but we preferred ANNs because they are naturally fit for problems with nonlinear decision hyperplanes, while SVMs require the selection of appropriate kernels and parameters. Surprisingly, we found that the pairwise comparison performed slightly better in the CP and IP tasks (i.e., 0.91 and 0.93 respectively) than in the OP task (i.e., 0.89), and the difference between OP and IP was judged as statistically significant by the t-test. This was tenta-

tively explained by the reduced presence of motor artifacts in the IP task where motor activity should be absent. It is likely that the CP and IP trials contain a large number of useful information for vowel cortical distinction since they are less affected by muscular activity as compared to OP signals. Furthermore, the location of the most discriminative electrodes of the scalp showed that the most informative sites are placed over both sensorimotor areas in CP, very close to the motor cortex (Riaz et al., 2014), and over posterior regions for IP task suggesting that the classification of these signals may be based on mostly on the imagined speech muscle movements as shown in other speech imagery studies (DaSalla et al., 2009). Further works are needed to provide additional validation to out hypothesis.

5. *Conclusions and Limitations*

The current work proposes a method that may be used for classifying speech sounds from brain signals and suggests the EEG technique as a pursuable and a necessary approach for developing frameworks for EEG-based SSIs systems. Other techniques result limiting in that perspective: functional magnetic resonance imaging (fMRI) has reduced temporal resolution, magnetoencephalography (MEG) is not sensitive to all the currents generated by the brain and ECoG requires the implantation of electrodes in the brain during neurosurgical operations.

However, in this study a series of limitations has to be taken into account, at least because finding highly significant results in such an experiment is new. For example, there is good reason to believe that the fixed order of vowel affected the accuracy performance as suggested by Porbadnigk and colleagues (2009) although, a more recent study has also revealed no significant difference between the imagined vocalization of vowels presented in fixed and in random order (Matsumoto, Hori, 2013). In future works, we are going to extend our probes on the activities involved during the motor preparation and to select small temporal windows (shorter than 300 ms) in order to provide a more fine-grained picture of the phenomenon under investigation.

Yet, data filtering needs to be much intensive: for example, building a composite system in which CSP is used as a preprocessing step before time-frequency analysis, may be a good solution; the reduction of the number of prominent electrodes as well as their physical significance for classification, remain subjects for future studies. To conclude, it will be also important to compare the performance of well-known procedures (e.g. SVM *vs*. ANN) and to test other efficient classification techniques moving beyond the pairwise classification of vowels. Currently, we are working to examine all these points in an EEG study with all Italian vowels.

Knowledgements

This work is supported by the "Programma Operativo Nazionale (PON) 254/Ric – Ricerca e competitività 2007/2013" of the Italian Ministry of Education, University,

and Research (upgrading of the "Centro ricerche per la salute dell'uomo e dell'ambiente" PONa3_00334)."

Bibliography

Atlas, L., Droppo, J. & McLaughlin, J. (1997). Optimizing time-frequency distributions via operator theory. In *Proceeding of SPIE*, 3162, 161-171.

Bouchard, K., Mesgarani, N., Johnson, K. & Chan, E. (2013). Functional organization of human sensorimotor cortex for speech articulation. In *Nature*, 495, 327-332.

Brigham, K., Kumar, B.V. (2010). Imagined speech classification with EEG signals for silent communication: a preliminary investigation into synthetic telepathy. In *Bioinformatics and Biomedical Engineering (iCBBE)*, 4th International Conference on. IEEE, 1-4.

Chia, X., Hagedorna, J.B., Schoonovera, D. & D'zmura, M. (2011). EEG-based discrimination of imagined speech phonemes. In *International Journal of Bioelectromagnetism*, 13, 201-206.

DaSalla, C.S., Kambara, H., Sato, M. & Koike, Y. (2009). Single-trial classification of vowel speech imagery using common spatial patterns. In *Neural Networks*, 22, 1334-1339.

Deecke, L., Engel, M., Lang, W. & Kornhuber, H.H. (1986). Bereitschaftspotential preceding speech after holding breath. In *Experimental Brain Research*, 65(1), 219-223.

Denby, J.B., Shultz, T., Honda, K., Hueber, T. & Gilbert, J.M. (2010). Silent speech interfaces. In *Speech Communication*, 52, 270-287.

D'Zmura, M., Deng, S., Lappas, T., Thorpe, S. & Srinivasan, R. (2009). Toward EEG sensing of imagined speech. In *International Conference on Human-Computer Interaction*. Springer: Berlin Heidelberg, 40-48.

Ebrahimi, T., Vesin, J. & Garcia, G. (2003). Brain-Computer Interface in Multimedia Communication. In *IEEE Signal Processing Magazine*, 20, 14-24.

Flinker, A., Korzeniewska, A., Shestyuk, A.Y., Franaszczuk, P.J., Dronkers, N.F., Knight, R.T. & Crone, N.E. (2015). Redefining the role of Broca's area in speech. In *Proceedings of the National Academy of Sciences*, 112(9), 2871-2875.

Garcia, G., Ebrahimi, T. & Vesin, J.M. (2002). Classification of EEG signals in the ambiguity domain for brain-computer interface applications. In *14th International Conference on Digital Signal Processing (DSP2002)*, 301-305.

Garcia, G., Ebrahimi, T. & Vesin, J. (2003). Joint Time-Frequency-Space Classification of EEG in a Brain Computer Interface Application. In *Eurasip Journal on Applied Signal Processing – Special issue on Neuromorphical Signal Processing*, 7, 713-729.

Gillespie, B., Atlas, L. (2001). Optimizing time-frequency kernels for classification. In *IEEE Transactions on Signal Processing*, 49, 485-496.

Guenther, F.H. (2007). Neuroimaging of normal speech production. In Ingham, R.J. (Ed.), *Neuroimaging in Communication Sciences and Disorders*. San Diego: Plural Publishing Inc., 1-51.

Guenther, F.H., Hampson, M. & Johnson, D.A. (1998). Theoretical investigation of reference frames for the planning of speech movements. In *Psychological review*, 105(4), 611.

Gunji, A., Hoshiyama, M. & Kakigi, R. (2000). Identification of auditory evoked potentials of one's own voice. In *Clinical Neurophysiology*, 111(2), 214-219.

Haykin, S.O. (2008). In Haykin, S.O. (Ed.). *Neural Networks and Learning Machines.* USA: Pearson.

Heinks-Maldonado, T.H., Nagarajan, S.S. & Houde, J.F. (2006). Magnetoencephalographic evidence for a precise forward model in speech production. In *Neuroreport*, 17(13), 1375. https://www.ncbi.nlm.nih.gov/pubmed/16932142.

Iqbal, S., Shanir, P.M., Khan, Y.U. & Farooq, O. (2016). Time Domain Analysis of EEG to Classify Imagined Speech. In *Proceedings of the Second International Conference on Computer and Communication Technologies*. Springer India, 793-800.

Kozek, W., Hlawatsch, F., Kirchauer, H. & Trautwein, U. (1994). Correlative time-frequency analysis and classification of nonstationary random processes. In *Proceedings of the IEEE-SP International Symposium on Time-Frequency and Time-Scale Analysis*, 417-420.

Luo, H., Poeppel, D. (2012). Cortical oscillations in auditory perception and speech: evidence for two temporal windows in human auditory cortex. In *Frontiers in Psychology*, 3(1), 170. https://www.ncbi.nlm.nih.gov/pubmed/22666214.

Manca, A.D., Grimaldi, M. (2013). Perception and production of Italian vowels: an ERP study. In *Proceedings of INTERSPEECH*, 916-920.

Manca, A.D., Di Russo, F. & Grimaldi, M. (2015). Orderly organization of vowels in the auditory brain: the neuronal correlates of the Italian vowels. In Vayra, M., Avesani, C. & Tamborini, F. (Eds.), *Acquisizione, mutamento e destrutturazione della struttura sonora del linguaggio/Language acquisition and language loss. Acquisition, change and disorders of the language sound structure*. Milano: AISV, 357-368.

Matsumoto, M., Hori, J. (2013). Classification of silent speech using adaptive collection. In *Computational Intelligence in Rehabilitation and Assistive Technologies (CIRAT)*, 5-12.

Matsumoto, M., Hori, J. (2014). Classification of silent speech using support vector machine and relevance vector machine. In *Applied Soft Computing*, 20, 95-102.

McLaughlin, L., Droppo, J. & Atlas, L. (1997). Class-dependent time-frequency distributions via operator theory. In *Proceeding of ICASSP*, 3, 2045-2048.

Näätänen, R., Picton, T. (1987). The N1 wave of the human electric and magnetic response to sound: a review and an analysis of the component structure. In *Psychophysiology*, 24(4), 375-425.

Obleser, J., Lahiri, A. & Eulitz, C. (2004). Magnetic Brain response mirrors extraction of phonological features from speakers vowels. In *Journal of Cognitive Neuroscience*, 16, 31-39.

Obleser, J., Scott, S.K. & Eulitz, C. (2006). Now you hear it, now you don't: Transient traces of consonants and their unintelligible analogues in the human brain. In *Cerebral Cortex*, 16, 1069-1076.

Pei, X., Barbour, D.L., Leuthardt, E.C. & Schalk, G. (2011). Decoding vowels and consonants in spoken and imagined words using electrocorticographic signals in humans. In *Journal of neural engineering*, 8(4). https://www.ncbi.nlm.nih.gov/pubmed/21750369.

Porbadnigk, A., Wester, M. & Jan-p Calliess, T.S. (2009). EEG-based speech recognition impact of temporal effects. In *Biosignals 2009*, 376-381.

Riaz, A., Akhtar, S., Iftikhar, S., Khan, A.A. & Salman, A. (2014). Inter comparison of classification techniques for vowel speech imagery using EEG sensors. In *Systems and Informatics (ICSAI), 2014 2nd International Conference*, 712-717.

Santana, R. (2015). Supervised classification of vowel speech imagery. In *Actas de la XVI Conferencia CAEPIA*, Albacete, 951-961.

Scharinger, M., Idsardi, W.J. & Poe, S. (2011). A Comprehensive Three-dimensional Cortical Map of Vowel Space. In *Journal of Cognitive Neuroscience*, 23, 3972-3982.

Šťastný, J., Sovka, P. & Stančák, A. (2003). EEG signal classification: introduction to the problem. In *Radioengineering*, 12(3), 51-55.

Suppes, P., Lu, Z.L. & Han, B. (1997). Brain wave recognition of words. In *Proceedings of the National Academy of Sciences*, 94(26), 14965-14969.

Suppes, P., Han, B., Epelboim, J. & Lu, Z.L. (1999). Invariance between subjects of brain wave representations of language. In *Proceedings of the National Academy of Sciences*, 96(22), 12953-12958.

Tourville, J.A., Reilly, K.J. & Guenther, F.H. (2008). Neural mechanisms underlying auditory feedback control of speech. In *Neuroimage*, 39(3), 1429-1443.

Tian, X., Poeppel, D. (2010). Mental imagery of speech and movement implicates the dynamics of internal forward models. In *Frontiers in Psychology*, 1, 166.

Vos, D.M., Riès, S., Vanderperren, K., Vanrumste, B., Alario, F.X., Huffel, V.S. & Burle, B. (2010). Removal of muscle artifacts from EEG recordings of spoken language production. In *Neuroinformatics*, 8(2), 135-150.

Wang, R., Perreau-Guimaraes, M., Carvalhaes, C. & Suppes, P. (2012). Using phase to recognize English phonemes and their distinctive features in the brain. In *Proceedings of the National Academy of Sciences*, 109(50), 20685-20690.