Key considerations in designing a speech brain-computer interface

Florent Bocquelet, Thomas Hueber, Laurent Girin, Stephan Chabardès, Blaise Yvert

Journal of Physiology - Paris xxx (2017) xxx–xxx

Contents lists available at ScienceDirect

Journal of Physiology - Paris

journal homepage: www.elsevier.com/locate/jphysparis

Key considerations in designing a speech brain-computer interface

Florent Bocquelet [a,b], Thomas Hueber [c], Laurent Girin [c], Stéphan Chabardès [d], Blaise Yvert [a,b,*]

[a] INSERM, BrainTech Laboratory U1205, F-38000 Grenoble, France
[b] Univ. Grenoble Alpes, BrainTech Laboratory U1205, F-38000 Grenoble, France
[c] Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, 38000 Grenoble, France
[d] Grenoble University Hospital, F-38000 Grenoble, France

ARTICLE INFO

Article history:
Received 3 February 2017
Received in revised form 21 June 2017
Accepted 19 July 2017
Available online xxxx

Keywords:
Neural prosthesis
Neural decoding
Rehabilitation
Speech synthesis
Silent speech
Aphasia

ABSTRACT

Restoring communication in case of aphasia is a key challenge for neurotechnologies. To this end, brain-computer strategies can be envisioned to allow artificial speech synthesis from the continuous decoding of neural signals underlying speech imagination. Such speech brain-computer interfaces do not exist yet and their design should consider three key choices that need to be made: the choice of appropriate brain regions to record neural activity from, the choice of an appropriate recording technique, and the choice of a neural decoding scheme in association with an appropriate speech synthesis method. These key considerations are discussed here in light of (1) the current understanding of the functional neuroanatomy of cortical areas underlying overt and covert speech production, (2) the available literature making use of a variety of brain recording techniques to better characterize and address the challenge of decoding cortical speech signals, and (3) the different speech synthesis approaches that can be considered depending on the level of speech representation (phonetic, acoustic or articulatory) envisioned to be decoded at the core of a speech BCI paradigm.

1. Introduction

It is estimated that the prevalence of aphasia is about 0.3% of the population, which corresponds to more than 20 millions people worldwide. Such impairment occurs most often after a brain stroke, but this disability also affects people with severe tetraplegia consequently to an upper spinal cord trauma, locked-in individuals, people suffering from neuro or muscular degenerative diseases (such as amyotrophic lateral sclerosis (ALS), Parkinson's disease, or myopathies), and even comatose patients. For these people, speech loss is an additional affliction that worsens their condition: It makes the communication with caregivers very difficult, and more generally, it can lead to profound social isolation and even depression. Restoring communication abilities is thus crucial for these patients.

Different solutions for communication have been developed, most often consisting of word spelling devices making use of residual physiological signals, for example based on eye-tracking strategies possibly accompanied by a clicking capability. However, these solutions become inappropriate when people have lost too much of their motor functions. Communication systems controlled directly by brain signals have thus started to be developed to overcome this problem. This concept has been pioneered by Farwell and Donchin who proposed a spelling device based on the evoked potential P300 (Farwell and Donchin, 1988), a method that has since been used successfully by an ALS patient to communicate (Sellers et al., 2014). Other EEG-based approaches use steady-state potentials tuned at different frequencies (Middendorf et al., 2000). A great advantage of these approaches is their non-invasiveness. However, they have been limited by a low spelling speed of a few characters per minute, although recent improvements suggest that higher speeds could be achieved (Townsend and Platsko, 2016). Another major limitations of EEG-based BCI systems for communication is that they still require a high level of concentration of the subjects (Käthner et al., 2014; Baykara et al., 2016), imposing a high cognitive workload limiting their easy use over extensive periods of time. Interestingly, with the drawback to require invasive recordings, BCI systems based on intracortical signals seem to alleviate the subject fatigue, the external device becoming progressively embodied after a period of training (Hochberg et al., 2006, 2012; Collinger et al., 2013; Wodlinger et al., 2015). Recently, Jarosiewicz and colleagues showed that incorporating self-recalibrating algorithms into an intracortical brain-computer spelling interface allows spelling performances of about 20–30 characters per minute by people with severe paralysis over long periods of use (Jarosiewicz et al., 2015).

* Corresponding author at: INSERM, BrainTech Laboratory U1205, F-38000 Grenoble, France.
E-mail address: blaise.yvert@inserm.fr (B. Yvert).

http://dx.doi.org/10.1016/j.jphysparis.2017.07.002
0928-4257/

Please cite this article in press as: Bocquelet, F., et al. Key considerations in designing a speech brain-computer interface. J. Physiol. (2017), http://dx.doi.org/10.1016/j.jphysparis.2017.07.002

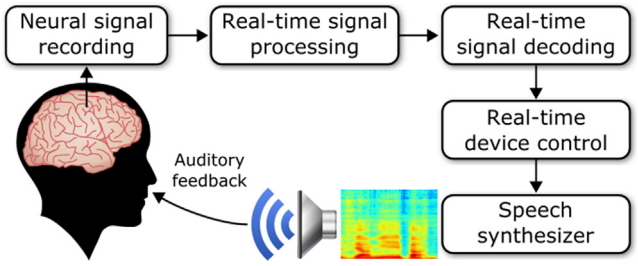

Fig. 1. Principle of a speech brain-computer interface.

The strategy of letter-selection BCI systems remains an indirect way of communicating based on movement direction decoded from the hand and/or arm area of the motor cortex. This is thus conceptually different than using speech, which is the natural and most efficient way of communication of the human species. Moreover, communication is often needed while other motor actions are performed requiring the resources of the hand/arm regions of the motor cortex (e.g. giving a phone call while moving in an environment or reaching for something). Thus, building a "speech BCI" to restore continuous speech directly from neural activity of speech-selective brain areas, as pioneered by Guenther and colleagues (Guenther et al., 2009), is an emerging field in which increasing efforts need to be invested in. As illustrated in Fig. 1, this strategy consists in extracting relevant neural signal features and converting them into input parameters for a speech synthesizer that runs in real time.

In this paper, we discuss several key requirements to restore speech with a BCI, including the choice of the speech cortical areas to record from, the recording techniques and decoding strategies that can be used, and finally the choice of speech synthesis approaches.

2. Choice of a brain region

Speech processing by the human brain involves a wide cortical network, which has been modeled by two main information streams linking auditory areas of the superior temporal plane to articulatory areas of frontal regions, one ventral and the other dorsal (Hickok and Poeppel, 2004, 2007). The ventral stream involves regions of the middle and inferior temporal lobe and maps speech sounds to meaning, while the dorsal stream runs through the dorsal part of the posterior temporal lobe at the temporo-parietal junction and is responsible for the sensori-motor integration of speech by mapping speech sounds to articulatory representations (Friederici, 2011; Hickok et al., 2011). Lesions of ventral stream regions of the temporal lobe result in Wernicke aphasia characterized by impairments of speech comprehension, while lesions of frontal areas result in Broca aphasia characterized by impairments of speech production. Classically, the dorsal stream has been described to be largely left-hemisphere dominant, but several studies indicate that many aspects of speech production activate cortical areas of the dorsal stream bilaterally (Pulvermüller et al., 2006; Peeva et al., 2010; Cogan et al., 2014; Geranmayeh et al., 2014; Keller and Kell, 2016).

Given this broad distribution of the speech network, to build a speech BCI, a choice needs to be made on the cortical areas to record and decode activity from. One possibility is to use signals from auditory areas of the ventral stream, which are known to encode the spectro-temporal representation of the acoustic content of speech, as assessed in both humans (Giraud et al., 2000; Formisano et al., 2008; Leonard and Chang, 2014; Leonard et al., 2015) and animals (Engineer et al., 2008; Mesgarani et al., 2008; Steinschneider et al., 2013). However, these areas are non-selectively involved in the sensory perception and integration of all speech sounds a person is exposed to. This includes self-produced speech but also other people speech, and even of non-speech environmental sounds as in the case for primary auditory areas. Thus, it can be expected that it would be difficult to identify activities specific to self speech intention in these areas. For this reason, probing neural activity in brain locations more specifically dedicated to speech production seems more relevant for conversational applications using a speech BCI (Guenther et al., 2009).

Several speech production conditions can be distinguished, including overt speech production, silent articulation (articulatory movements without vocalization, i.e. with no laryngeal activity), and inner (covert) speech production. The later condition (Perrone-Bertolotti et al., 2014) is likely the one most relevant when envisioning the use of a speech BCI by patients that intend to speak while not being able to produce articulatory movements. Articulatory speech production pathways originate from the speech motor cortex and project to the brainstem trigeminal, facial and ambiguus nuclei. Brainstem nuclei are difficult to access for recordings and there has yet been no evidence for their activation during covert intended speech. Thus, a speech BCI is likely to be easier to achieve by probing cortical areas underlying the production of inner speech.

Functional imaging studies have shown that overt word repetition activates motor and premotor cortices bilaterally (Petersen et al., 1988, 1989; Palmer et al., 2001; Peeva et al., 2010; Cogan et al., 2014). Continuous production of narrative speech was also shown to activate frontal motor speech regions and temporal and parietal areas bilaterally (Silbert et al., 2014). Intraoperative functional mapping data collected in a high number of patients undergoing awake surgery also report bilateral critical motor and premotor regions for overt speech production (Tate et al., 2014). The right hemisphere is also clearly activated during synchronized speaking in several regions including the temporal pole, inferior frontal gyrus, and supramarginal gyrus (Jasmin et al., 2016). When more complex tasks are considered that require additional semantic, lexical, or phonological processing, then specific activations are observed in the left inferior frontal cortex (Petersen et al., 1988, 1989; Price et al., 1994; Sörös et al., 2006; Basho et al., 2007). These findings suggest that speech production becomes left lateralized when inner high-level processing is required. In general, inner speech has been found to activate similar brain areas but with a lesser amplitude than overt speech across most ventral and dorsal stream areas (Price et al., 1994; Ryding et al., 1996; Palmer et al., 2001; Shuster and Lemieux, 2005). In particular, as for high-level overt speech production, cortical activity underlying covert speech production is left lateralized with strong activation of the left motor, premotor and inferior frontal cortex (Ryding et al., 1996; Palmer et al., 2001; Keller and Kell, 2016). The left inferior frontal cortex has further been shown to be specifically activated during covert word retrieval (Hirshorn and Thompson-Schill, 2006) and to be important for inner speech production as assessed using repetitive transcranial magnetic stimulation (Aziz-Zadeh et al., 2005). A careful anatomical voxel-based lesion study further confirmed the importance of this region as well as the white matter adjacent to the left supramarginal gyrus to achieve rhyme and homophone tasks requiring inner speech production (Geva et al., 2011).

Overall, the left inferior frontal region encompassing Broadman areas 4, 6, 44, 45, and 47, thus appears as a pertinent candidate from which to probe and decode neural activity for the control of a speech BCI. It should be noted that this strategy can only apply to aphasic patients whose speech networks remain intact, at least in this region. This is generally the case for instance for locked-in individuals or patients with ALS. To envision a speech BCI in people

Please cite this article in press as: Bocquelet, F., et al. Key considerations in designing a speech brain-computer interface. J. Physiol. (2017), http://dx.doi.org/10.1016/j.jphysparis.2017.07.002

that became aphasic following brain damage, for instance after a stroke, a speech BCI would need to be adapted. In particular, this would require training new brain regions not previously involved in speech production to become active in this task. Thus, adapting a speech BCI to brain-damaged patients will constitute a further challenge beyond the achievement of a speech BCI in brain-intact patients. Here, we thus focus on this latter case, for which no proof of concept has been achieved yet.

3. Choice of a recording technique to monitor speech brain signals

As mentioned above, imaging studies based on PET and fMRI have been extensively used to highlight the brain areas involved in speech production. An important prerequisite toward building a speech BCI is to be able to decode brain signals to predict intended speech. This strategy relies on the temporal dynamics of both speech and brain activity and the correlation that exists between these two dynamics. Several studies have shown that single trial fMRI can be used to successfully predict with an accuracy above chance level which of different speech items or types are perceived by subjects from their BOLD activity recorded in auditory areas (Formisano et al., 2008; Evans et al., 2013; Bonte et al., 2014; Correia et al., 2014). Similarly, speech articulatory features such as place of articulation can also be decoded with this approach (Correia et al., 2015). Although not shown yet, it is possible that fMRI could also be used to predict features of overtly or covertly produced speech. However, in these studies, the number of speech categories that can be discriminated remains limited (typically 2–3), and it is likely that fMRI signals lack the sufficient temporal resolution to allow decoding ongoing sequences of phonemes forming continuous speech. This constitutes a major limitation to envision a real-time speech synthesis from ongoing brain activity recorded with this technique. In addition fMRI equipment makes it not compatible with an everyday life use of a BCI system at home. By contrast, electrophysiological recording techniques can fit into compact portable devices and offer a temporal resolution appropriate to track the time course of brain activity on the scale of the dynamics of speech production.

Non-invasive electro- and magneto-encephalography (EEG/MEG) recording techniques have been used to study the cortical dynamics of speech perception. In particular, several studies have shown that the envelope or rhythm of perceived speech is correlated with oscillatory rhythms composing the activity of the auditory cortex (Luo and Poeppel, 2007; Gross et al., 2013; Di Liberto et al., 2015). It was also recently shown that scalp potentials evoked by different phonemes (phoneme-related potentials or PRPs) show different spatiotemporal distributions over the scalp between 50 and 400 ms after phoneme onset, and that the similarity of PRPs follows the acoustic similarity of phonemes (Khalighinejad et al., 2016). Less data is available for speech production, likely due to experimental limitations and artifacts generated by muscle activity during speech production. Nevertheless, similar observations have been made in this case with low frequency cortical rhythms of the mouth sensorimotor areas also strongly correlating with EMG activity of the mouth during articulation (Ruspantini et al., 2012). Moreover, attention was also found to modulate MEG activity over the left frontal and temporal areas during an overt speech production task (Carota et al., 2010). Non-invasive EEG/MEG techniques have also been used in the quest to decode continuous speech from ongoing brain activity. The fact that brain rhythms get coupled to the envelope of speech during perception could be exploited to classify fragments of speech envelopes from ongoing MEG signals, with longer segments leading to more robust classification (Koskinen et al., 2013). Single trial analysis of EEG responses to speech could also achieve above chance level classification of four speech items differing from their voice onset time (Brandmeyer et al., 2013).

Despite these very informative results, non-invasive electrophysiology techniques likely lack the spatial resolution required to track ongoing neural activity with sufficient details to enable the prediction of continuous intelligible speech. To this end, invasive recordings appear as a promising alternative (Llorens et al., 2011). Intracerebral stereotaxic EEG (SEEG) performed in epileptic patients undergoing presurgical evaluation of their epilepsy has been of great help to detail the functional organization of the human brain auditory system (Liégeois-Chauvel et al., 1991; Yvert et al., 2002, Yvert et al., 2005). This approach has further been used to decipher in more details the cortical dynamics underlying speech and language perception (Liegeois-Chauvel et al., 1999; Basirat et al., 2008; Sahin et al., 2009; Fontolan et al., 2014). In particular it has helped to highlight how brain oscillations encode the rhythmic properties of speech, with a strong coupling of the theta rhythm to the tempo of syllables occurrence in speech and associated nested modulation of gamma-band signals possibly encoding transient acoustic speech features (Giraud and Poeppel, 2012; Morillon et al., 2012). Intracerebral SEEG recordings have also highlighted several aspects of cortical activity underlying silent reading. In particular, it was shown that reading sentences generates broad gamma activity detectable on a single trial basis in the left temporal lobe, supramarginal gyrus and inferior frontal cortex (Mainy et al., 2008; Perrone-Bertolotti et al., 2012; Vidal et al., 2012), the latter region showing an anterior subregion activated by semantic sentences and a posterior subregion more specifically activated by phonologic sentences (Vidal et al., 2012).

A nice feature of SEEG is not only that it offers direct and thus more detailed cortical recordings but also that several regions distant from each other are usually recorded at the same time, thus allowing the analysis of interactions between areas. The drawback of this advantage is that only few electrode contacts can usually be inserted in a given region of interest, for instance the inferotemporal region. This limitation in spatial coverage precludes the access to the detailed dynamics of frontal motor speech areas and may limits the possibility to decode with sufficient details a continuous speech flow produced either overtly or covertly.

Electrocorticographic (ECoG) recordings are also routinely performed in epileptic patients undergoing a pre-surgical evaluation of their pharmaco-resistant epilepsy. A grid housing multiple contacts is positioned over the surface of the cortex, usually subdurally, and allows monitoring the activity of the brain during speech production or imagination. One or several grids may cover a large region encompassing frontal motor areas and temporal auditory areas to advantageously record activity from the cortical speech network during overt and cover speech production. Several ECoG studies have shown that cortical oscillations are relevant correlates of speech processing (Leuthardt et al., 2011; Pei et al., 2011a; Pasley et al., 2012; Bouchard et al., 2013; Pasley and Knight, 2013; Martin et al., 2014; Mugler et al., 2014) (see also Fig. 2). In particular, speech production is classically associated with a decrease of signal power in the beta frequency range (15–25 Hz) and usually an increase in the high gamma frequency range (70–200 Hz) over temporal and motor frontal areas (Canolty et al., 2007; Pei et al., 2011b; Toyoda et al., 2014) while gamma attenuation was observed in more anterior frontal speech cortex including Broca area (Lachaux et al., 2008; Wu et al., 2011; Toyoda et al., 2014). These oscillatory features can thus be used to map functional cortical speech areas, for instance to help delineate functional areas during resection surgeries (Kamada et al., 2014; Tamura et al., 2016). In this respect high-gamma activity has been shown to be informative to map cortical areas activated for different place and manner of articulation (Lotte et al., 2015) and to

Please cite this article in press as: Bocquelet, F., et al. Key considerations in designing a speech brain-computer interface. J. Physiol. (2017), http://dx.doi.org/10.1016/j.jphysparis.2017.07.002

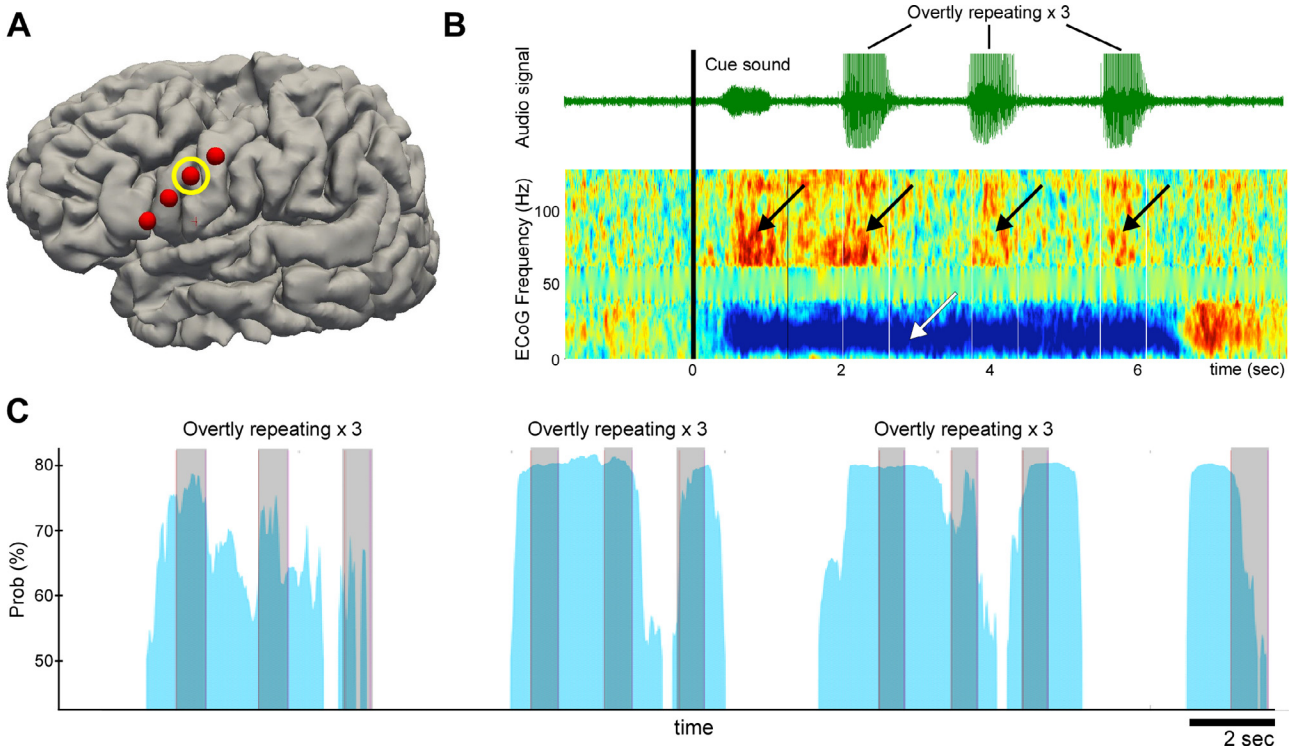

Fig. 2. Example of functional cortical activity underlying overt speech production as recorded by ECoG on peri-tumoral speech motor cortex during awake surgery. A series of isolated vowels and vowel-consonant-vowel speech sounds was presented to the patient with a loudspeaker positioned next to him. The patient was asked to repeat aloud three times each item after its presentation. (A) Position of 4 ECoG electrodes on the reconstructed cortical surface. (B) Time-frequency decomposition of ECoG data underlying speech production recorded on the electrode circled in A, located on the articulatory motor cortex. Top: sample sound recorded by a microphone positioned next to the awake patient. Bottom: time-frequency representation of the ECoG signal showing clear beta suppression (blue, white arrow) and gamma-band responses to the cue and for each sound occurrence (red, black arrows). ECoG data was recorded at 2 kHz and the time-frequency representation was computed using short-time Fourier transform using a Hamming function on 512 samples sliding windows with 95% overlap. The time-frequency representation was then normalized by the 1-s pre-stimulus period and averaged over 83 trials aligned on the beginning of the cue signal. (C) Example of decoding of voice activity using the neural features extracted from the single electrode shown in A and B. The blue area shows the probability that the patient is speaking as continuously predicted by the decoding model. The decoding model consisted of an artificial neural network (ANN) trained on the normalized time-frequency representation by keepings only the frequency bins in the beta (from 10 to 30 Hz) and gamma (from 60 to 90 Hz) frequency bands and averaged over a 500 ms sliding window. The ANN was made of 2 hidden layers of 10 logistic units each, and non-speech segments of the training set were randomly chosen in order to obtain the same number of speech and non-speech segments. This DNN was then continuously applied to the test data (which was not part of the training set) on a frame-by-frame basis by concatenating previous frames over a 500-ms time window. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

determine causal interactions between the motor speech network and the auditory areas (Korzeniewska et al., 2011). Dense ECoG grids have further been used to detail with a higher spatial resolution the functional organization of the ventral sensory-motor cortex with respect to the main speech articulators. Noticeably, it was shown that this region is tuned to the articulatory content of speech during production according to the somatotopic organization of this area (Bouchard et al., 2013), while the auditory content of speech is encoded in a subpart of this region during speech perception (Cheung et al., 2016).

Several studies have further explored the extent to which ECoG signal features could be decoded to predict the content of produced speech. A first level of decoding is the detection of voice activity irrespective of the phonetic content of speech, that is discriminating the time intervals during which the subject is speaking or not. As reported previously, these intervals can be estimated with high reliability from EcoG signals recorded over the frontal motor speech areas or the posterior supratemporal gyrus (Kanas et al., 2014). Fig. 2 also shows an example where voice activity detection can be achieved with 75% reliability from a single electrode site located over the lips-tongue area of the motor cortex. A second level of decoding is the prediction of the actual speech content at the level of individual words or syllables or phonemes. If successful, such decoding could be used in a speech BCI paradigm to reconstruct continuous speech from brain signals. Discriminating between 2 and 10 words could be achieved above chance level using discrete classification algorithms applied to ECoG neural features (Kellis et al., 2010), indicating that ECoG signals contain information that differs from word to word. Continuous spectrograms of speech have further been reconstructed from ECoG signals recorded during overt production over motor frontal and auditory temporal areas (Pasley et al., 2012; Martin et al., 2014). Although the resulting speech intelligibility remained limited, the overall time-frequency structure of the speech spectrograms could be well estimated. Such reconstruction was all the more accurate that the number of electrode sites was high, and the most informative sites were found to be in temporal auditory areas (Pasley et al., 2012). In another study, ECoG signals recorded from the speech motor cortex were also used to decode all phonemes of American English using discrete classification with a success rate of about 20% across 4 different subjects, this rate reaching 36% in one subject using 6 electrodes located over the ventral somatosensory region (Mugler et al., 2014). To a lesser extent, ECoG data could also be used to predict silently articulated or covertly imagined speech not actually overtly pronounced by the subject (Pei et al., 2011a; Ikeda et al., 2014; Martin et al., 2014) (see also Fig. 3 showing an example of inner speech episode decoding from a single electrode located over the articulator motor cortex). The reconstruction of covert speech was in general more limited than for overt speech but above chance level, and more reliable for vowels

Please cite this article in press as: Bocquelet, F., et al. Key considerations in designing a speech brain-computer interface. J. Physiol. (2017), http://dx.doi.org/10.1016/j.jphysparis.2017.07.002

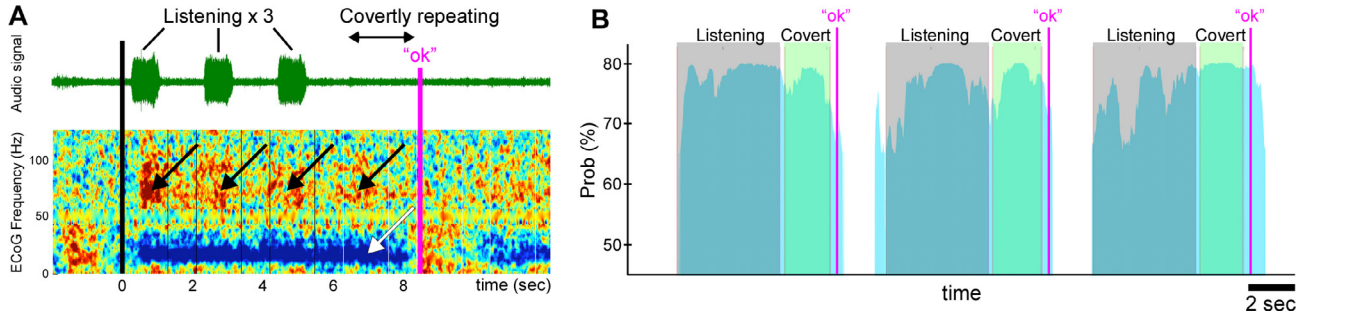

Fig. 3. Example of functional cortical activity underlying covert speech production. A series of isolated vowels and vowel-consonant-vowel speech sounds was presented 3 times to the patient at a regular pace and the patient was asked to continue to imagine pronouncing these items at the same pace after their presentation, and to say "ok" aloud when done. (A) The modulation of beta and high-gamma band activity over the speech motor cortex during speech listening is prolonged during the period the subject is asked to imagine repeating what he has heard. Top: sound recorded by the microphone positioned next to the awake patient. Bottom: time-frequency representation of the ECoG signal averaged over 24 trials on the same electrode and using the same methods as in Fig. 2B. The vertical pink line shows the mean position of the end of the imagination period as notified by the patient by saying "ok" aloud. (B) The decoder previously built on overt speech data (Fig. 2C) still reliably predicts the instants of speech imagination. They correspond closely to actual ones shown at the bottom of the graph. The decoding is performed identically to Fig. 2C. The pink lines indicate the position of the "ok" pronounced by the patient to notify the end of each imagination period. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

than for consonants. Moreover, informative electrodes were localized over both frontal and temporal regions for vowels, but only over temporal sites for consonants. Thus, an open question remains on whether the frontal motor area contains sufficient information to allow an accurate prediction of covert speech, especially for consonants.

In particular, more accurate prediction of speech sounds could be expected from even more detailed recordings performed at the cellular or multicellular level using microelectrodes implanted intracortically. The five English vowels (o, a, e, i, u) could be decoded with high accuracy (93%) from spiking data recorded in medial frontal and temporal regions using depth electrodes (Tankus et al., 2012). Ten words could also be classified with 40% accuracy from population unit activity recorded using the intracortical Utah array in the superior temporal gyrus (Chan et al., 2013), a level comparable to the performance of ECoG decoding using 5 optimal electrodes over the frontal motor cortex (Kellis et al., 2010). However, in these intracortical studies, the neural probes were not optimally located in areas specific to speech production. Hence, a higher accuracy is likely to be obtained with microelectrode arrays positioned into specific speech motor areas. To date, only one group has recorded unit activity in the articulatory speech areas using an intracortical Neurotrophic electrode. The recorded signals were used to control a simple vowel synthesizer (Guenther et al., 2009). A further study by the same group reported that it was possible to discriminate 20 out of 38 imagined American English phonemes well above chance level (around 20%) from signals recorded with this single 2-channel microelectrode in the speech motor cortex of a locked-in syndrome individual (Brumberg et al., 2011). These encouraging pioneer studies suggest that high decoding and BCI performance is likely to be expected from denser recordings in this region.

4. Choice of a decoding strategy and associated speech synthesis method

Artificial production of speech can be achieved in several ways, which can be classified based on the type of parameters that are decoded from brain signals to serve as inputs for the speech synthesis. As illustrated in Fig. 4, three different types of parameters can be envisioned, each corresponding to a different representation of speech: phonetic, acoustic, or articulatory. Each of these representations is likely to be more specifically encoded in certain cortical areas than others. For instance the acoustic content of speech is more extensively encoded in temporal auditory areas, while articulatory features are more specifically encoded is the speech motor cortex. Hence, a decoding approach will likely be more efficient if the parametric representation of speech it intends to decode corresponds to the one encoded in the cortical areas from which neural signals are recorded. As a result, different speech synthesis methods can be considered depending on the choice of the decoding strategy.

A first category of speech synthesis consists in concatenating individual discrete phonemes or words. A BCI system based on such synthesis would thus consist in first predicting discrete speech items from brain activity, for instance using discrete classification of neural features as in (Mugler et al., 2014), and then to convert the sequence of predicted phonemes or words into continuous audio speech. This latter step can be done using algorithms used in *text-to-speech synthesis* (TTS) (Taylor, 2009). TTS input is typically a sequence of written words. In most implementations, a natural language processing module converts this sequence into a sequence of phonemes and other features related to the prosody (e.g. whether a syllable is stressed or not). A second module generates the speech waveform from both phonetic labels and prosodic features. The two main strategies currently used in most systems are unit selection and statistical parametric synthesis. In unit selection, as in (Hunt and Black, 1996), the speech signal is obtained by concatenating recorded speech segments stored in a very large database. In statistical parametric synthesis, machine learning techniques (such as hidden semi Markov models (Tokuda et al., 1995) or recurrent deep neural networks (Zen et al., 2013)) are used to directly estimate a sequence of acoustic parameters given a target sequence of phonemes and prosodic features. The speech waveform is finally synthesized by a vocoder. Since a TTS system is driven by a sequence of words, its use in a BCI system requires a front-end module able to decode brain activity at word (or at least phoneme) level. Such front-end module can be seen as an automatic speech recognition (ASR) system driven not by the sound, but directly by the brain activity. Although the design of such decoder and its use in a closed-loop BCI paradigm is still an unsolved issue, a recent study (Herff et al., 2015) reported encouraging results on the offline decoding of EcoG data, with a word-error-rate of 25%. As shown in this work, the major advantage of combining a word-based decoder and a TTS system is probably the possibility to regularize the brain-to-speech mapping by introducing prior linguistic knowledge. Similarly to a conventional audio-based recognition system, such knowledge can be given by

Please cite this article in press as: Bocquelet, F., et al. Key considerations in designing a speech brain-computer interface. J. Physiol. (2017), http://dx.doi.org/10.1016/j.jphysparis.2017.07.002

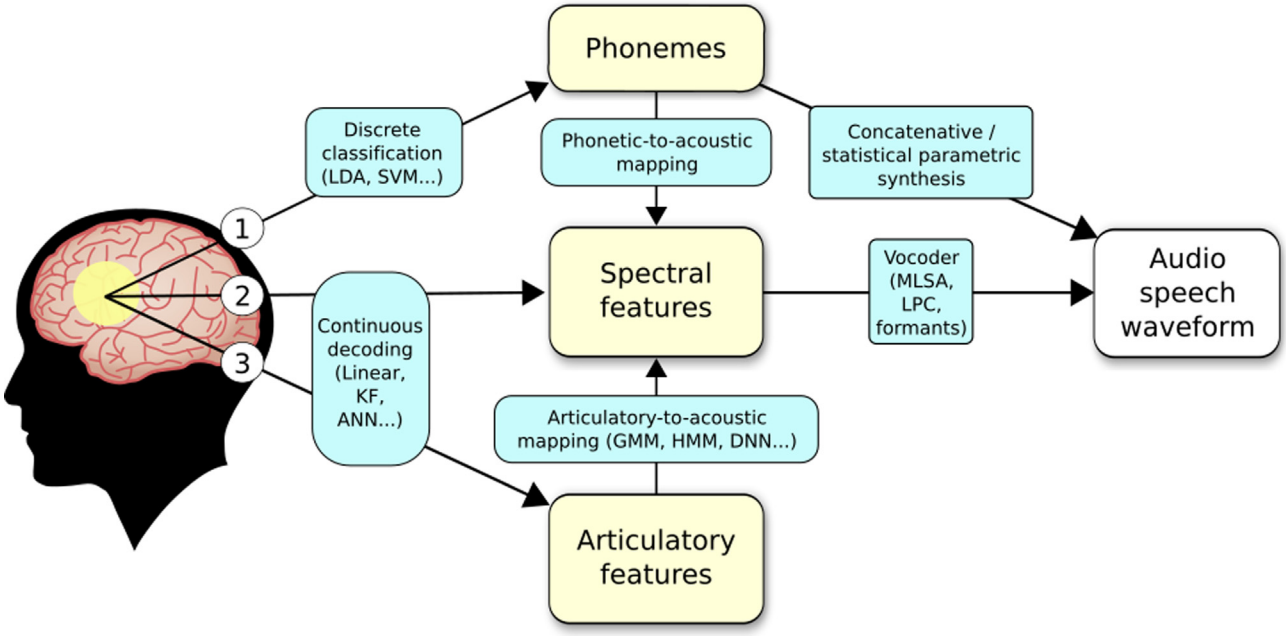

Fig. 4. Three representations of speech (phonetic, acoustic or articulatory) can be decoded from brain signals, each implying the use of specific speech synthesis techniques to build a speech BCI.

a pronunciation dictionary (and thus a limitation on the authorized vocabulary) and a statistical language model (giving the prior probability of observing a given sequence of words in a given language). One limitation of such mapping remains however the difficulty of a real-time implementation. Indeed, even a short-term decoding algorithm will necessarily introduce a delay of one or two words which may be problematic for controlling the BCI in closed-loop.

In a second category of speech synthesis, the input parameters describe the spectral content of the target speech signal. Hence, a BCI system based on this approach would typically convert brain signals into a spectral representation of speech (Guenther et al., 2009; Pasley et al., 2012; Martin et al., 2014), which in turn would be converted into a speech waveform using a vocoder. As for the decoded spectral representation, a privileged choice is to use formants (Flanagan et al., 1962), which are the local maxima of energy in the speech spectrum, since formants are both compact and perceptually relevant descriptors of the speech content. Note that formants are also related to the spatial positions of the speech articulators. Such formantic representation could be used to directly pilot a formant synthesizer such as the Klatt synthesizer (Klatt, 1980). Since this type of synthesizer typically uses several tens of parameters (there exist versions with more than 50 parameters to describe the position and bandwidth of the 6 first formants and the glottal activity), a simplified version should be used. This was the strategy used in the speech BCI described in (Guenther et al., 2009). Since this study focused on vowels, only 2 parameters were estimated from the brain activity: the position of the two first formants (which are sufficient to discriminate vowels), while the other parameters were set to constant values. As mentioned in (Guenther et al., 2009), the formant synthesis is well adapted to vowel synthesis but less to consonants, such as plosives, which require a rapid and accurate control of several parameters to achieve a realistic-sounding closure and burst. In the same category, *vocoders* found in telecommunication systems use other representations of the spectral content of sounds, from which speech can be synthesized. The speech waveform is here obtained by modulating an excitation signal (representing the glottal activity) through a time-varying filter representing the transfer function of the vocal tract (i.e. the spectral envelope). One of the most common techniques is the Linear Predictive Coding (LPC, see (O'Shaughnessy, 1988) for its use in speech processing), where the spectral envelope is modeled by the transfer function of an all-pole filter. In the context of low-bitrate speech coding, good intelligibility can be obtained with a 10th order LPC filter, excited either by a pulse train for voiced sound or by white noise for unvoiced sound (Boite et al., 2000) (nevertheless such simple excitation signal lead to an unnatural voice). An LPC vocoder models the speech spectrum in a compact and accurate way. However, directly mapping LPC prediction coefficients from brain signals in a speech BCI does not appear as a proper choice, since the variation of these coefficients with the speech spectrum content is quite "e rratic". Rather, transcoding predicted formants into an LPC model is an easy signal processing routine. Other models of the spectral envelope can also be envisioned in the same line, among which the mel-cepstrum model with the corresponding digital filter MLSA (Imai et al., 1983).

Finally, the third category of approaches for synthesizing speech is the so-called *articulatory synthesis*. The control parameters are here the time-varying positions of the main speech organs, such as the tongue, the lips, the jaw, the velum and the larynx. A BCI based on such synthesis would thus consist in predicting the movements of the articulators from brain activity and then to convert these movements into acoustic speech. Two main approaches have been proposed for articulatory speech synthesis. The first one is a "physical" approach, in which the geometry of a generic vocal tract (including the articulators) is described in two or three dimensions (Birkholz et al., 2011). This geometry is converted into an area function describing how the cross sectional area of the vocal tract varies between the glottis and the mouth opening. Then, an acoustic model of sound propagation is used to calculate the speech wave from the sequence of area functions and corresponding sound sources. In the second approach, supervised machine learning is used to model the relationship between articulatory and acoustic observations. Articulatory and acoustic data are typically recorded simultaneously on a reference speaker, using a motion-capture technique such as electromagnetic-articulography (EMA). Then these data are used to train a mapping

Please cite this article in press as: Bocquelet, F., et al. Key considerations in designing a speech brain-computer interface. J. Physiol. (2017), http://dx.doi.org/10.1016/j.jphysparis.2017.07.002

model. Several models have been proposed in the literature to model the relationship between articulatory positions captured by EMA and corresponding speech spectral features: Artificial neural networks (ANN) (Kello and Plaut, 2004; Richmond, 2006), Gaussian Mixture Models (GMM) (Toda et al., 2008), and Hidden Markov Models (HMM) (Hiroya and Honda, 2004; Hueber and Bailly, 2016). Once calibrated, these models are used to estimate acoustic trajectories from time-varying articulatory trajectories, and a standard vocoder is finally used to generate the speech waveform.

It should be noted that, after the initial neural signal decoding into one of the three representations described above, speech synthesis may further cascade or even combine other representations to optimize the quality of synthesized speech. For instance, articulatory features have to be mapped into acoustic features, which correspond to a different representation, before using a vocoder. Another example could be the simultaneous decoding of both an articulatory and a phonetic representations that could then be combined before speech synthesis (Astrinaki et al., 2013). Moreover, beyond these three categories of speech representations, one could also consider a higher representation of language at higher linguistic level to shape the prosody of synthesized speech.

5. The special case of articulatory-based speech synthesis

The use of an articulatory speech synthesizer can be of particular interest for a BCI application for several reasons. First, as discussed in the previous section, an area of choice to probe the neural activity in a BCI paradigm is the frontal speech motor region. This area is activated during both speech production and perception (Pulvermüller et al., 2006) but it has been shown that it is tuned to the articulatory content of speech during speech production and to the acoustic content of speech during speech listening (Pasley and Knight, 2013; Cheung et al., 2016). In particular, the activity of the sensorimotor speech cortex was found to be similar for similar places of articulation but not for similar acoustic content during speech production, and similar for similar acoustic content and not similar place of articulation during speech perception. This result has not been extended to the case of imagined speech but according to preliminary data showing similar activations during covert and overt speech, it might be expected that this region would also be tuned to articulatory features during speech imagination. This hypothesis remains to be tested but if true, then building a BCI paradigm relying on the activity of this region would benefit from relying on an articulatory-based speech synthesis. A second advantage of articulatory synthesis is that articulatory features vary more slowly and more smoothly than spectral features. It is thus possible to speculate that their time evolution might be easier to estimate from brain activity. Moreover, as mentioned in (Guenther et al., 2009), a third advantage of an articulatory synthesizer is its ability to produce consonants with a limited amount of control parameters. This is notably the case for some plosives that can be estimated from relatively slowly time-varying control parameters corresponding to the movement of an articulator (e.g. the tongue) producing a vocal tract closure. Such pattern is more difficult to produce with a formant synthesizer. Several articulatory speech synthesis systems have been described in the literature including the model proposed by Maeda (Maeda, 1990) that was further implemented in a compact analog electronic circuit board compatible with BCI applications with 7 control parameters (Wee et al., 2008).

In line with these considerations, we recently developed an articulatory speech synthesizer adapted to a BCI application (Bocquelet et al., 2016). This system is based on a deep neural network (DNN) for the articulatory-to-acoustic mapping, which we previously evaluated as being more robust to noisy inputs than state-of-the-art GMM models (Bocquelet et al., 2014). The DNN was trained on a dataset of EMA and audio recordings simultaneously acquired from a reference speaker. Once trained this DNN was then able to convert the movement trajectories of the tongue, lips, jaw and velum into continuously varying spectral parameters, which, in turn, could be transformed by a vocoder to generate a continuous speech waveform (with a proper excitation signal). With a future BCI application in mind, we showed that this system could (i) produce intelligible speech with no restriction on the vocabulary and with as few as 7 control parameter (as the Maeda articulatory synthesizer) (ii) run in real time, (iii) be easily adapted to any arbitrary new speaker after a short calibration phase, and (iv) be controlled in a closed-loop paradigm by several subjects to produce intelligible speech from their articulatory movements monitored using EMA while they articulated silently. Further studies should further expand this study to situations where such synthesizer is controlled in real time from brain signals.

6. Conclusion

Designing a speech BCI requires targeting appropriate brain regions with appropriate recording techniques and to choose a strategy to decode neural signals into continuous speech audio signals. In this respect, the inferior frontal region appears as a key region from which to decode activity specific to the covert production of speech. This region being tuned to the articulatory content of speech, we propose that a speech BCI controlled from this region could use an articulatory-based speech synthesizer as developed recently (Bocquelet et al., 2016). Because such synthesizer is typically controlled by a ten of parameters, neural activity should be sufficiently detailed to allow the simultaneous control of such a number of degrees of freedom (DoF). Recent advances in motor BCI have shown that, provided careful training, a ten of DoF could indeed be controlled from unit or multiunit activity recorded using microelectrode arrays (Collinger et al., 2013; Wodlinger et al., 2015). High-dimensional BCI control of a speech synthesizer from microelectrode array recordings in the frontal speech network could thus be a key challenge for future translational studies. Such proof of principle would directly benefit to aphasic people with preserved cortical speech networks as in Locked-In Syndrome or ALS disease.

Acknowledgments

This work was supported by the Fondation pour la Recherche Médicale (www.frm.org) under grant No DBS20140930785, by the French National Research Agency (www.ANR.org) through projects Neuromeddle ANR-15-CE19-0006 and Brainspeak ANR-16-CE19-0005, and by the European Community's Horizon 2020 Programme (H2020/2014-2020) under grant agreement n° 732032 through the BrainCom project. The authors also wish to thank Marie-Pierre Gilotin and Manuela Oddoux for clinical help during awake surgery and the patient who participated in the study.

References

Astrinaki, M., Moinet, A., Yamagishi, J., Richmond, K., Ling, Z., King, S., Dutoit, T., 2013. Mage - reactive articulatory feature control of HMM-based parametric speech synthesis. Ssw, pp. 207–211.

Aziz-Zadeh, L., Cattaneo, L., Rochat, M., Rizzolatti, G., 2005. Covert speech arrest induced by rTMS over both motor and nonmotor left hemisphere frontal sites. J. Cogn. Neurosci. 17, 928–938.

Basho, S., Palmer, E.D., Rubio, M.A., Wulfeck, B., Müller, R.A., 2007. Effects of generation mode in fMRI adaptations of semantic fluency: paced production and overt speech. Neuropsychologia 45, 1697–1706.

Basirat, A., Sato, M., Schwartz, J.L., Kahane, P., Lachaux, J.P., 2008. Parieto-frontal gamma band activity during the perceptual emergence of speech forms. Neuroimage 42, 404–413.

Baykara, E., Ruf, C.A., Fioravanti, C., Käthner, I., Simon, N., Kleih, S.C., K??bler, A., Halder, S., 2016. Effects of training and motivation on auditory P300 brain-computer interface performance. Clin. Neurophysiol. 127, 379–387. http://dx.doi.org/10.1016/j.clinph.2015.04.054.

Birkholz, P., Kroger, B.J., Neuschaefer-Rube, C., 2011. Model-based reproduction of articulatory trajectories for consonant-vowel sequences. IEEE Trans. Audio, Speech Lang. Process. 19, 1422–1433.

Bocquelet, F., Hueber, T., Girin, L., Badin, P., Yvert, B., 2014. Robust Articulatory Speech Synthesis using Deep Neural Networks for BCI Applications. In: Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech), pp. 2288–2292.

Bocquelet, F., Hueber, T., Girin, L., Savariaux, C., Yvert, B., 2016. Real-time control of an articulatory-based speech synthesizer for brain computer interfaces. PLOS Comput. Biol. 12, e1005119. Available at: <http://dx.doi.org/10.1371%2Fjournal.pcbi.1005119>.

Boite, R., Bourlard, H., Dutoit, T., Hancq, J., Leich, H., 2000. Traitement de la parole. Presses Polytechniques et Universitaires Romandes, Lausanne.

Bonte, M., Hausfeld, L., Scharke, W., Valente, G., Formisano, E., 2014. Task-dependent decoding of speaker and vowel identity from auditory cortical response patterns. J. Neurosci. 34, 4548–4557. Available at: (Accessed April 29, 2014) <http://www.ncbi.nlm.nih.gov/pubmed/24672000>.

Bouchard, K.E., Mesgarani, N., Johnson, K., Chang, E.F., 2013. Functional organization of human sensorimotor cortex for speech articulation. Nature 495, 327–332.

Brandmeyer, A., Farquhar, J.D.R., McQueen, J.M., Desain, P.W.M., 2013. Decoding speech perception by native and non-native speakers using single-trial electrophysiological data. PLoS One 8, e68261. Available at: (Accessed July 17, 2014) <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3708957&tool=pmcentrez&rendertype=abstract>.

Brumberg, J.S., Wright, E.J., Andreasen, D.S., Guenther, F.H., Kennedy, P.R., 2011. Classification of intended phoneme production from chronic intracortical microelectrode recordings in speech motor cortex. Front. Neurosci. 5. Available at: <http://www.frontiersin.org/Journal/Abstract.aspx?s=763&name=neuroprosthetics&ART_DOI=10.3389/fnins.2011.00065>.

Canolty, R.T., Soltani, M., Dalal, S.S., Edwards, E., Dronkers, N.F., Nagarajan, S.S., Kirsch, H.E., Barbaro, N.M., Knight, R.T., 2007. Spatiotemporal dynamics of word processing in the human brain. Front. Neurosci. 1, 185–196.

Carota, F., Posada, A., Harquel, S., Delpuech, C., Bertrand, O., Sirigu, A., 2010. Neural dynamics of the intention to speak. Cereb. Cortex 20 (Available at:), 1891–1897 <http://www.ncbi.nlm.nih.gov/pubmed/20008453>.

Chan, A.M., Dykstra, A.R., Jayaram, V., Leonard, M.K., Travis, K.E., Gygi, B., Baker, J.M., Eskandar, E., Hochberg, L.R., Halgren, E., Cash, S.S., 2013. Speech-specific tuning of neurons in human superior temporal gyrus. Cereb. Cortex 10, 2679–2693. Available at: (Accessed July 31, 2014) <http://www.ncbi.nlm.nih.gov/pubmed/23680841>.

Cheung, C., Hamiton, L.S., Johnson, K., Chang, E.F., 2016. The auditory representation of speech sounds in human motor cortex. Elife 5, 1–19. Available at: <http://elifesciences.org/lookup/doi/10.7554/eLife.12577>.

Cogan, G.B., Thesen, T., Carlson, C., Doyle, W., Devinsky, O., Pesaran, B., 2014. Sensory-motor transformations for speech occur bilaterally. Nature 507, 94–98. Available at: (Accessed January 20, 2014) <http://www.ncbi.nlm.nih.gov/pubmed/24429520>.

Collinger, J.L., Wodlinger, B., Downey, J.E., Wang, W., Tyler-Kabara, E.C., Weber, D.J., McMorland, A.J.C., Velliste, M., Boninger, M.L., Schwartz, A.B., 2013. High-performance neuroprosthetic control by an individual with tetraplegia. Lancet 381, 557–564. Available at: (Accessed November 11, 2013) <http://www.ncbi.nlm.nih.gov/pubmed/23253623>.

Correia, J., Formisano, E., Valente, G., Hausfeld, L., Jansma, B., Bonte, M., 2014. Brain-based translation: fMRI decoding of spoken words in bilinguals reveals language-independent semantic representations in anterior temporal lobe. J. Neurosci. 34, 332–338. Available at: (Accessed August 1, 2014) <http://www.ncbi.nlm.nih.gov/pubmed/24381294>.

Correia, J.M., Jansma, B.M.B., Bonte, M., 2015. Decoding articulatory features from fMRI responses in dorsal speech regions. J. Neurosci. 35, 15015–15025. Available at: <http://www.jneurosci.org/cgi/doi/10.1523/JNEUROSCI.0977-15.2015>.

Di Liberto, G.M., O'Sullivan, J.A., Lalor, E.C., 2015. Low-frequency cortical entrainment to speech reflects phoneme-level processing. Curr. Biol. 25, 2457–2465. http://dx.doi.org/10.1016/j.cub.2015.08.030.

Engineer, C.T., Perez, C.A., Chen, Y.H., Carraway, R.S., Reed, A.C., Shetake, J.A., Jakkamsetti, V., Chang, K.Q., Kilgard, M.P., 2008. Cortical activity patterns predict speech discrimination ability. Nat. Neurosci. 11, 603–608. Available at: <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=18425123>.

Evans, S., Kyong, J.S., Rosen, S., Golestani, N., Warren, J.E., McGettigan, C., Mourão-Miranda, J., Wise, R.J.S., Scott, S.K., 2013. The pathways for intelligible speech: multivariate and univariate perspectives. Cereb. Cortex, 1–12. Available at: (Accessed February 15, 2014) <http://www.ncbi.nlm.nih.gov/pubmed/23585519>.

Farwell, L., Donchin, E., 1988. Talking Off the Top of Your Head: toward a mental prosthesis utilizing event-related brain potentials. Electroencephalogr. Clin. Neurophysiol. 70, 510–523.

Flanagan, J.L., Coker, C.H., Bird, C.M., 1962. Computer simulation of a formant-vocoder synthesizer. J. Acoust. Soc. Am. 34. http://dx.doi.org/10.1121/1.1937133. 2003.

Fontolan, L., Morillon, B., Liegeois-Chauvel, C., Giraud, A.-L., 2014. The contribution of frequency-specific activity to hierarchical information processing in the human auditory cortex. Nat. Commun. 5, 4694. Available at: http://www.nature.com/doifinder/10.1038/ncomms5694%5Cn http://www.ncbi.nlm.nih.gov/pubmed/25178489%5Cn http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4164774.

Formisano, E., De Martino, F., Bonte, M., Goebel, R., 2008. "Who" is saying "what"? Brain-based decoding of human voice and speech 80Available at: Science 322, 970–973 <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=18988858>.

Friederici, A.D., 2011. The brain basis of language processing: from structure to function. Physiol. Rev. 91, 1357–1392. Available at: (Accessed July 11, 2014) <http://www.ncbi.nlm.nih.gov/pubmed/22013214>.

Geranmayeh, F., Wise, R.J.S., Mehta, A., Leech, R., 2014. Overlapping networks engaged during spoken language production and its cognitive control. J. Neurosci. 34, 8728–8740. Available at: (Accessed July 14, 2014) <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4069351&tool=pmcentrez&rendertype=abstract>.

Geva, S., Jones, P.S., Crinion, J.T., Price, C.J., Baron, J.C., Warburton, E.A., 2011. The neural correlates of inner speech defined by voxel-based lesion-symptom mapping. Brain 134, 3071–3082.

Giraud, A.-L., Poeppel, D., 2012. Cortical oscillations and speech processing: emerging computational principles and operations. Nat. Neurosci. 15, 511–517. Available at: (Accessed April 28, 2014) <http://www.ncbi.nlm.nih.gov/pubmed/22426255>.

Giraud, A.L., Lorenzi, C., Ashburner, J., Wable, J., Johnsrude, I., Frackowiak, R., Kleinschmidt, A., 2000. Representation of the temporal envelope of sounds in the human brain. J. Neurophysiol. 84, 1588–1598.

Gross, J., Hoogenboom, N., Thut, G., Schyns, P., Panzeri, S., Belin, P., Garrod, S., 2013. Speech rhythms and multiplexed oscillatory sensory coding in the human brain. PLoS Biol. 11, e1001752. Available at: (Accessed July 28, 2014) <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3876971&tool=pmcentrez&rendertype=abstract>.

Guenther, F.H., Brumberg, J.S., Wright, E.J., Nieto-Castanon, A., Tourville, J.a., Panko, M., Law, R., Siebert, S.A., Bartels, J.L., Andreasen, D.S., Ehirim, P., Mao, H., Kennedy, P.R., 2009. A wireless brain-machine interface for real-time speech synthesis. PLoS One 4, e8218.

Herff, C., Heger, D., de Pesters, A., Telaar, D., Brunner, P., Schalk, G., Schultz, T., 2015. Brain-to-text: decoding spoken phrases from phone representations in the brain. Front. Neurosci. 9, 1–11.

Hickok, G., Houde, J., Rong, F., 2011. Sensorimotor integration in speech processing: computational basis and neural organization. Neuron 69, 407–422. Available at: (Accessed November 11, 2013) <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3057382&tool=pmcentrez&rendertype=abstract>.

Hickok, G., Poeppel, D., 2004. Dorsal and ventral streams: a framework for understanding aspects of the functional anatomy of language. Cognition 92, 67–99. Available at: <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=15037127>.

Hickok, G., Poeppel, D., 2007. The cortical organization of speech processing. Nat. Rev. Neurosci. 8, 393–402.

Hiroya, S., Honda, M., 2004. Estimation of articulatory movements from speech acoustics using an HMM-based speech production model. IEEE Trans. Speech Audio Process. 12, 175–185.

Hirshorn, E.A., Thompson-Schill, S.L., 2006. Role of the left inferior frontal gyrus in covert word retrieval: neural correlates of switching during verbal fluency. Neuropsychologia 44, 2547–2557.

Hochberg, L.R., Bacher, D., Jarosiewicz, B., Masse, N.Y., Simeral, J.D., Vogel, J., Haddadin, S., Liu, J., Cash, S.S., van der Smagt, P., Donoghue, J.P., 2012. Reach and grasp by people with tetraplegia using a neurally controlled robotic arm. Nature 485, 372–375. Available at: <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=22596161>.

Hochberg, L.R., Serruya, M.D., Friehs, G.M., Mukand, J.A., Saleh, M., Caplan, A.H., Branner, A., Chen, D., Penn, R.D., Donoghue, J.P., 2006. Neuronal ensemble control of prosthetic devices by a human with tetraplegia. Nature 442, 164–171. Available at: <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=16838014>.

Hueber, T., Bailly, G., 2016. Statistical conversion of silent articulation into audible speech using full-covariance HMM. Comput. Speech Lang. 36, 274–293.

Hunt, A.J., Black, A.W., 1996. Unit selection in a concatenative speech synthesis system using a large speech database. In: 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings, pp. 373–376 Available at: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=541110>.

Ikeda, S., Shibata, T., Nakano, N., Okada, R., Tsuyuguchi, N., Ikeda, K., Kato, A., 2014. Neural decoding of single vowels during covert articulation using electrocorticography. Front. Hum. Neurosci. 8, 125. Available at: (Accessed August 5, 2014) <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3945950&tool=pmcentrez&rendertype=abstract>.

Imai, S., Sumita, K., Furuichi, C., 1983. Mel Log Spectrum Approximation (MLSA) filter for speech synthesis. Electron. Commun. Japan 66-A, 10–18.

Jarosiewicz, B., Sarma, A.A., Bacher, D., Masse, N.Y., Simeral, J.D., Sorice, B., Oakley, E. M., Blabe, C., Pandarinath, C., Gilja, V., Cash, S.S., Eskandar, E.N., Friehs, G., Henderson, J.M., Shenoy, K.V., Donoghue, J.P., Hochberg, L.R., 2015. Virtual typing by people with tetraplegia using a self-calibrating intracortical brain-computer interface. Sci. Transl. Med. 7, 1–11.

Please cite this article in press as: Bocquelet, F., et al. Key considerations in designing a speech brain-computer interface. J. Physiol. (2017), http://dx.doi.org/10.1016/j.jphysparis.2017.07.002

Jasmin, K.M., McGettigan, C., Agnew, Z.K., Lavan, N., Josephs, O., Cummins, F., Scott, S.K., 2016. Cohesion and joint speech: right hemisphere contributions to synchronized vocal production. J. Neurosci. 36, 4669–4680. Available at: <http://www.jneurosci.org/cgi/doi/10.1523/JNEUROSCI.4075-15.2016>.

Kamada, K., Ogawa, H., Kapeller, C., Prueckl, R., Guger, C., 2014. Rapid and low-invasive functional brain mapping by realtime visualization of high gamma activity for awake craniotomy. Conf ProcAnnu Int Conf IEEE Eng Med Biol Soc IEEE Eng Med Biol Soc Annu Conf 2014, 6802–6805.

Kanas, V.G., Mporas, I., Benz, H.L., Sgarbas, K.N., Bezerianos, A., Crone, N.E., 2014. Joint spatial-spectral feature space clustering for speech activity detection from ECoG signals. IEEE Trans. Biomed. Eng. 61, 1241–1250. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/24658248>.

Käthner, I., Wriessnegger, S.C., Müller-Putz, G.R., Kübler, A., Halder, S., 2014. Effects of mental workload and fatigue on the P300, alpha and theta band power during operation of an ERP (P300) brain-computer interface. Biol. Psychol. 102, 118–129. http://dx.doi.org/10.1016/j.biopsycho.2014.07.014.

Keller, C., Kell, C.A., 2016. Asymmetric intra- and interhemispheric interactions during covert and overt sentence reading. Neuropsychologia, 1–18. http://dx.doi.org/10.1016/j.neuropsychologia.2016.04.002.

Kellis, S., Miller, K., Thomson, K., Brown, R., House, P., Greger, B., 2010. Decoding spoken words using local field potentials recorded from the cortical surface. J. Neural Eng. 7, 56007. Available at: <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=20811093>.

Kello, C.T., Plaut, D.C., 2004. A neural network model of the articulatory-acoustic forward mapping trained on recordings of articulatory parameters. J. Acoust. Soc. Am. 116, 2354.

Khalighinejad, B., da Silva, G.C., Mesgarani, N., 2016. Recurrent Representation of Acoustic Phonetic in Neural Responses to Continuous Speech. Press 37, 2176–2185.

Klatt, D.H., 1980. Software for a cascade/paralell formant synthesizer. J. Acoust. Soc. Am. 67, 971–995.

Korzeniewska, A., Franaszczuk, P.J., Crainiceanu, C.M., Kuś, R., Crone, N.E., 2011. Dynamics of large-scale cortical interactions at high gamma frequencies during word production: event related causality (ERC) analysis of human electrocorticography (ECoG). Neuroimage 56, 2218–2237.

Koskinen, M., Viinikanoja, J., Kurimo, M., Klami, A., Kaski, S., Hari, R., 2013. Identifying fragments of natural speech from the listener's MEG signals. Hum. Brain Mapp. 34, 1477–1489. Available at: (Accessed August 5, 2014) <http://www.ncbi.nlm.nih.gov/pubmed/22344824>.

Lachaux, J.P., Jung, J., Mainy, N., Dreher, J.C., Bertrand, O., Baciu, M., Minotti, L., Hoffmann, D., Kahane, P., 2008. Silence is golden: transient neural deactivation in the prefrontal cortex during attentive reading. Cereb. Cortex 18, 443–450.

Leonard, M.K., Bouchard, K.E., Tang, C., Chang, E.F., 2015. Dynamic encoding of speech sequence probability in human temporal cortex. J. Neurosci. 35, 7203–7214. http://dx.doi.org/10.1523/JNEUROSCI.4100-14.2015.

Leonard, M.K., Chang, E.F., 2014. Dynamic speech representations in the human temporal lobe. Trends Cogn. Sci. 18, 472–479. http://dx.doi.org/10.1016/j.tics.2014.05.001.

Leuthardt, E.C., Gaona, C., Sharma, M., Szrama, N., Roland, J., Freudenberg, Z., Solis, J., Breshears, J., Schalk, G., 2011. Using the electrocorticographic speech network to control a brain-computer interface in humans. J. Neural Eng. 8, 36004. Available at: <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=21471638>.

Liegeois-Chauvel, C., de Graaf, J.B., Laguitton, V., Chauvel, P., 1999. Specialization of left auditory cortex for speech perception in man depends on temporal coding. Cereb. Cortex 9, 484–496.

Liégeois-Chauvel, C., Musolino, A., Chauvel, P., 1991. Localization of the primary auditory area in man. Brain 114, 139–153.

Llorens, A., Trébuchon, A., Liégeois-Chauvel, C., Alario, F.X., 2011. Intra-cranial recordings of brain activity during language production. Front. Psychol. 2, 1–12.

Lotte, F., Brumberg, J.S., Brunner, P., Gunduz, A., Ritaccio, A.L., 2015. Electrocorticographic representations of segmental features in continuous speech. Front. Hum. Neurosci. 9, 1–13.

Luo, H., Poeppel, D., 2007. Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex. Neuron 54, 1001–1010.

Maeda, S., 1990. Compensatory articulation during speech: evidence from the analysis and synthesis of vocal tract shapes using an articulatory model. In: Hardcastle, W.J., Marchal, A. (Eds.), Speech Production and Speech Modeling. Kluwer Academic Publishers, pp. 131–149.

Mainy, N., Jung, J., Baciu, M., Kahane, P., Schoendorff, B., Minotti, L., Hoffmann, D., Bertrand, O., Lachaux, J.P., 2008. Cortical dynamics of word recognition. Hum. Brain Mapp. 29, 1215–1230.

Martin, S., Brunner, P., Holdgraf, C., Heinze, H.-J., Crone, N.E., Rieger, J., Schalk, G., Knight, R.T., Pasley, B.N., 2014. Decoding spectrotemporal features of overt and covert speech from the human cortex. Front. Neuroeng. 7, 14. Available at: (Accessed July 24, 2014) <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4034498&tool=pmcentrez&rendertype=abstract>.

Mesgarani, N., David, S.V., Fritz, J.B., Shamma, S.A., 2008. Phoneme representation and classification in primary auditory cortex Available at: (Accessed December 24, 2013) J. Acoust Soc Am 123, 899–909 <http://www.ncbi.nlm.nih.gov/pubmed/18247893>.

Middendorf, M., McMillan, G., Calhoun, G., Jones, K.S., 2000. Brain-computer interfaces based on the steady-state visual-evoked response. IEEE Trans. Rehabil. Eng. 8, 211–214.

Morillon, B., Liégeois-Chauvel, C., Arnal, L.H., Bénar, C.G., Giraud, A.L., 2012. Asymmetric function of theta and gamma activity in syllable processing: an intra-cortical study. Front. Psychol. 3, 1–9.

Mugler, E.M., Patton, J.L., Flint, R.D., Wright, Z.a., Schuele, S.U., Rosenow, J., Shih, J.J., Krusienski, D.J., Slutzky, M.W., 2014. Direct classification of all American English phonemes using signals from functional speech motor cortex. J. Neural Eng. 11, 35015. Available at: (Accessed May 28, 2014) <http://www.ncbi.nlm.nih.gov/pubmed/24836588>.

O'Shaughnessy, D., 1988. Linear predictive coding. IEEE Potentials 7, 29–32. Available at: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1890>.

Palmer, E.D., Rosen, H.J., Ojemann, J.G., Buckner, R.L., Kelley, W.M., Petersen, S.E., 2001. An event-related fMRI study of overt and covert word stem completion. Neuroimage 14, 182–193.

Pasley, B.N., David, S.V., Mesgarani, N., Flinker, A., Shamma, S.a., Crone, N.E., Knight, R.T., Chang, E.F., 2012. Reconstructing speech from human auditory cortex. PLoS Biol. 10, e1001251. Available at: (Accessed December 14, 2013) <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3269422&tool=pmcentrez&rendertype=abstract>.

Pasley, B.N., Knight, R.T., 2013. Decoding Speech for Understanding and Treating Aphasia. Elsevier BV. Available at: (Accessed August 5, 2014) <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4043958&tool=pmcentrez&rendertype=abstract>.

Peeva, M.G., Guenther, F.H., Tourville, J.A., Nieto-Castanon, A., Anton, J.L., Nazarian, B., Alario, F.X., 2010. Distinct representations of phonemes, syllables, and supra-syllabic sequences in the speech production network. Neuroimage 50, 626–638. Available at: <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=20035884>.

Pei, X., Barbour, D.L., Leuthardt, E.C., Schalk, G., 2011a. Decoding vowels and consonants in spoken and imagined words using electrocorticographic signals in humans. J. Neural Eng. 8, 46028. Available at: <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=21750369>.

Pei, X., Leuthardt, E.C., Gaona, C.M., Brunner, P., Wolpaw, J.R., Schalk, G., 2011b. Spatiotemporal dynamics of electrocorticographic high gamma activity during overt and covert word repetition. Neuroimage 54, 2960–2972.

Perrone-Bertolotti, M., Kujala, J., Vidal, J.R., Hamame, C.M., Ossandon, T., Bertrand, O., Minotti, L., Kahane, P., Jerbi, K., Lachaux, J.-P., 2012. How silent is silent reading? Intracerebral evidence for top-down activation of temporal voice areas during reading Available at: (Accessed May 8, 2014) J. Neurosci. 32, 17554–17562 <http://www.ncbi.nlm.nih.gov/pubmed/23223279>.

Perrone-Bertolotti, M., Rapin, L., Lachaux, J.P., Baciu, M., Lœvenbruck, H., 2014. What is that little voice inside my head? Inner speech phenomenology, its role in cognitive performance, and its relation to self-monitoring. Behav. Brain Res. 261, 220–239. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/24412278>.

Petersen, S.E., Fox, P.T., Posner, M.I., Mintun, M., Raichle, M.E., 1988. Positron emission tomographic studies of the cortical anatomy of single- word processing. Nature 331, 585–589. Available at: <http://www.ncbi.nlm.nih.gov/htbin-post/Entrez/query?db=m&form=6&dopt=r&uid=3277066>.

Petersen, S.E., Fox, P.T., Posner, M.I., Mintun, M., Raichle, M.E., 1989. Positron emission tomographic studies of the processing of single words. J. Cogn. Neurosci. 1, 153–170.

Price, C.J., Wise, R.J., Watson, J.D., Patterson, K., Howard, D., Frackowiak, R.S., 1994. Brain activity during reading. The effects of exposure duration and task. Brain 117, 1255–1269.

Pulvermüller, F., Huss, M., Kherif, F., del Prado, Moscoso, Martin, F., Hauk, O., Shtyrov, Y., 2006. Motor cortex maps articulatory features of speech sounds. Proc. Natl. Acad. Sci. USA 103, 7865–7870. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1472536&tool=pmcentrez&rendertype=abstract>.

Richmond, K., 2006. A Trajectory Mixture Density Network for the Acoustic-Articulatory Inversion Mapping. In: INTERSPEECH 2006 and 9th International Conference on Spoken Language Processing, Vols 1–5, pp 577–580.

Ruspantini, I., Saarinen, T., Belardinelli, P., Jalava, A., Parviainen, T., Kujala, J., Salmelin, R., 2012. Corticomuscular coherence is tuned to the spontaneous rhythmicity of speech at 2–3 Hz. J. Neurosci. 32, 3786–3790. Available at: (Accessed May 21, 2014) <http://www.ncbi.nlm.nih.gov/pubmed/22423099>.

Ryding, E., Bradvik, B., Ingvar, D., 1996. Silent speech activates prefrontal cortical regions asymmetrically, as well as speech-related areas in the dominant. Hemisphere 52, 435–451.

Sahin, N.T., Pinker, S., Cash, S.S., Schomer, D., Halgren, E., 2009. Sequential processing of lexical, grammatical, and phonological information within Broca's area. Science 326, 445–449. Available at: (Accessed May 25, 2014) <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4030760&tool=pmcentrez&rendertype=abstract>.

Sellers, E.W., Ryan, D.B., Hauser, C.K., 2014. Noninvasive brain-computer interface enables communication after brainstem stroke. Sci. Transl. Med. 6, 257re7. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/25298323%5Cn25298323%5Cn10.1126/scitranslmed.3007801%5CnSciTransl Med-2014-Sellers-257re7.pdf>.

Shuster, L.I., Lemieux, S.K., 2005. An fMRI investigation of covertly and overtly produced mono- and multisyllabic words 93, 20–31.

Silbert, L.J., Honey, C.J., Simony, E., Poeppel, D., Hasson, U., 2014. Coupled neural systems underlie the production and comprehension of naturalistic narrative

Please cite this article in press as: Bocquelet, F., et al. Key considerations in designing a speech brain-computer interface. J. Physiol. (2017), http://dx.doi.org/10.1016/j.jphysparis.2017.07.002

speech. Proc. Natl. Acad. Sci. 111, E4687–E4696. Available at: <http://www.pnas.org/cgi/doi/10.1073/pnas.1323812111>.

Sörös, P., Sokoloff, L.G., Bose, A., Mcintosh, A.R., Graham, S.J., Stuss, D.T., 2006. Clustered functional MRI of overt speech production. Neuroimage 32, 376–387.

Steinschneider, M., Nourski, K.V., Fishman, Y.I., 2013. Representation of speech in human auditory cortex: is it special? Hear Res. 305, 57–73. Available at: (Accessed March 25, 2015) <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3818517&tool=pmcentrez&rendertype=abstract>.

Tamura, Y., Ogawa, H., Kapeller, C., Prueckl, R., Takeuchi, F., Anei, R., Ritaccio, A., Guger, C., Kamada, K., 2016. Passive language mapping combining real-time oscillation analysis with cortico-cortical evoked potentials for awake craniotomy. J. Neurosurg., 1–9 Available at: <http://www.ncbi.nlm.nih.gov/pubmed/26991386>.

Tankus, A., Fried, I., Shoham, S., 2012. Structured neuronal encoding and decoding of human speech features. Nat. Commun. 3, 1015. Available at: (Accessed June 3, 2014) <http://www.ncbi.nlm.nih.gov/pubmed/22910361>.

Tate, M.C., Herbet, G., Moritz-Gasser, S., Tate, J.E., Duffau, H., 2014. Probabilistic map of critical functional regions of the human cerebral cortex: Broca's area revisited. Brain 137, 2773–2782. Available at: <http://www.brain.oxfordjournals.org/cgi/doi/10.1093/brain/awu168>.

Taylor, P., 2009. Text-to-Speech Synth. Text-to-speech synthesis, 1–597. Available at: <http://www.scopus.com/inward/record.url?eid=2-s2.0-84925160976&partnerID=40&md5=278e4c8e7d44063486654957f388339e>.

Toda, T., Black, A.W., Tokuda, K., 2008. Statistical mapping between articulatory movements and acoustic spectrum using a Gaussian mixture model. Speech Commun. 50, 215–227.

Tokuda, K., Kobayashi, T., Imai, S., 1995. Speech parameter generation from HMM using dynamic features 1995 Int Conf Acoust Speech. Signal Process. 1, 660–663.

Townsend, G., Platsko, V., 2016. Pushing the P300-based brain-computer interface beyond 100 bpm: extending performance guided constraints into the temporal domain. J. Neural Eng. 13, 26024. Available at: <http://iopscience.iop.org/article/10.1088/1741-2560/13/2/026024>.

Toyoda, G., Brown, E.C., Matsuzaki, N., Kojima, K., Nishida, M., Asano, E., 2014. Electrocorticographic correlates of overt articulation of 44 English phonemes: intracranial recording in children with focal epilepsy. Clin. Neurophysiol. 125, 1129–1137. Available at: (Accessed August 5, 2014) <http://www.ncbi.nlm.nih.gov/pubmed/24315545>.

Vidal, J.R., Freyermuth, S., Jerbi, K., Hamame, C.M., Ossandon, T., Bertrand, O., Minotti, L., Kahane, P., Berthoz, A., Lachaux, J.J.-P., Atomique, E., Grenoble, F-., Neuro-cognition, L.D.P., Cnrs, U.M.R., 2012. Long-distance amplitude correlations in the high gamma band reveal segregation and integration within the reading network. J. Neurosci. 32, 6421–6434.

Wee, K.H., Turicchia, L., Sarpeshkar, R., 2008. An analog integrated-circuit vocal tract. Biomed. Circ. Syst. IEEE Trans. 2, 316–327.

Wodlinger, B., Downey, J.E., Tyler-Kabara, E.C., Schwartz, A.B., Boninger, M.L., Collinger, J.L., 2015. Ten-dimensional anthropomorphic arm control in a human brain-machine interface: difficulties, solutions, and limitations. J. Neural Eng. 12, 16011. Available at: (Accessed February 23, 2015) <http://www.ncbi.nlm.nih.gov/pubmed/25514320>.

Wu, H.C., Nagasawa, T., Brown, E.C., Juhasz, C., Rothermel, R., Hoechstetter, K., Shah, A., Mittal, S., Fuerst, D., Sood, S., Asano, E., 2011. Gamma-oscillations modulated by picture naming and word reading: Intracranial recording in epileptic patients. Clin. Neurophysiol. 122, 1929–1942. http://dx.doi.org/10.1016/j.clinph.2011.03.011.

Yvert, B., Fischer, C., Bertrand, O., Pernier, J., 2005. Localization of human supratemporal auditory areas from intracerebral auditory evoked potentials using distributed source models. Neuroimage 28, 140–153.

Yvert, B., Fischer, C., Guénot, M., Krolak-Salmon, P., Isnard, J., Pernier, J., 2002. Simultaneous intracerebral EEG recordings of early auditory thalamic and cortical activity in human. Eur. J. Neurosci. 16, 1146–1150.

Zen, H., Senior, A., Schuster, M., 2013. Statistical parametric speech synthesis using deep neural networks. Int. Conf. Acoust. Speech Signal Process., 7962–7966 Available at: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6639215>.

Please cite this article in press as: Bocquelet, F., et al. Key considerations in designing a speech brain-computer interface. J. Physiol. (2017), http://dx.doi.org/10.1016/j.jphysparis.2017.07.002