Key considerations in designing a speech brain-computer interface
Florent Bocquelet, Thomas Hueber, Laurent Girin, Stephan Chabardès, Blaise Yvert

To cite this version:
Florent Bocquelet, Thomas Hueber, Laurent Girin, Stephan Chabardès, Blaise Yvert. Key considerations in designing a speech brain-computer interface. Journal of Physiology - Paris, 2016, 110 (4, Part A), pp.392-401. 10.1016/j.jphysparis.2017.07.002 . hal-01978301

HAL Id: hal-01978301
https://hal.science/hal-01978301v1
Submitted on 11 Jan 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Key considerations in designing a speech brain-computer interface

Florent Bocquelet\textsuperscript{a,b}, Thomas Hueber\textsuperscript{c}, Laurent Girin\textsuperscript{c}, Stéphan Chabardès\textsuperscript{d}, Blaise Yvert\textsuperscript{a,b,*}

\textsuperscript{a} INSERM, BrainTech Laboratory U1205, F-38000 Grenoble, France
\textsuperscript{b} Univ. Grenoble Alpes, BrainTech Laboratory U1205, F-38000 Grenoble, France
\textsuperscript{c} Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, 38000 Grenoble, France
\textsuperscript{d} Grenoble University Hospital, F-38000 Grenoble, France

\begin{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.
\end{abstract}

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 consequentially 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 limitation 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).

\begin{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.
\end{abstract}

© 2017 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

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 consequentially 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 limitation 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).
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 performing 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 that 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-hemispheric dominant, but several studies indicate that many aspects of speech production activate cortical areas of the dorsal stream bilaterally (Pulvermüller et al., 2006; Pessa et al., 2010; Cogan et al., 2014; Gerammary 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 ambiguous nuclei. Brainstem nuclei are difficult to access for recordings and there has yet been no evidence for their activation during covert intented 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; Pessa 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 Lemioux, 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 IMRI 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 IMRI 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 IMRI 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 IMRI 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 IMRI 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 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 (Koskinnen 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 stereotactic 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 been further used to decipher in more details the cortical dynamics underlying speech and language perception (Liégeois-Chauvel et al., 1999; Basrati 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 limit the possibility to decode with sufficient details a continuous speech flow produced either overtly or covertly. Electroocorticographic (EOG) recordings are also routinely performed in epileptic patients undergoing a pre-surgical evaluation of their pharmaco-resistant epilepsy. A grid housing multiple contacts 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 EOG 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
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., 2011a; Ikeda et al., 2014; Martin 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.

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.
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, i, e, 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 (Keliris 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 in 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...
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 “erratic”. 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 Haut, 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 a 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 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 multunit 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., Mointet, A., Yamagishi, J., Richmond, K., Ling, Z., King, S., Dutoit, T., 2013. Mage - reactive articulatory feature control of HMM-based parametric speech synthesis. Sow. 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.

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
Boucquet, F., Mesgarani, N., Johnson, K., Chang, E.F., 2013. Functional organization of the human brain. Front. Neurosci. 7, 1–11. Available at: <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=23806414>.

Cheung, C., Hamilton, L.S., Johnson, K., Chang, E.F., 2016. The auditory representation of speech sounds in human motor cortex. Cereb. Cortex 26, 2679–2693. Available at: <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=23680414>.

Fontolan, I., Morillon, B., Liegeois-Chauvel, C., Giraud, A.-L., 2014. The contribution of frequency-specific activity to memory in a declarative task. Nat. Commun. 5, 4694. Available at: <http://www.nature.com/doifinder/10.1038/ncomms6945>.

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

Hickok, G., Poeppel, D., 2004. Dorsal and ventral streams: a framework for understanding aspects of the functional anatomy of language. Cognition 92, 3–17. 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.

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

Hutcheson, E.A., Thompson-Schill, S.L., 2006. Role of the left inferior frontal gyrus in vowel identification. Neuropsychologia 44, 2547–2557. Available at: <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=16402585>.

Ikeda, S., Shibata, T., Nakano, N., Okada, R., Tsuyuguchi, N., Ikeda, K., Kato, A., 2014. Neural decoding of single vowels during covert articulation using fMRI. Nat. Neurosci. 17, 1961–1967.

Jakkamsetti, V., Chang, K.Q., Kilgard, M.P., 2008. Cortical activity patterns of neurons in human superior temporal gyrus. Cereb. Cortex 10, 2679–2693. Available at: <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=18191897>.

Johansen-Berg, H., Jones, D.K., Pajevic, S., Miller, L., Matthews, P.M.D., 2002. Intrinsic connectivity of the human thalamus: a high-resolution diffusion tensor imaging study. J. Neurosci. 22, 3945–3950.

Kaiser, L., Bartenstein, P., Wiesmann, S., Siebner, H.R., 2009. Cortical modulation of brain activity during performance of a virtual object manipulation task. NeuroImage 49, 1197–1208.

Kato, A., Kato, K., 2009. Categorical perception of rhythm: cross-language perspective. J. Acoust. Soc. Am. 125, 1171–1183.

Kato, A., Iwasa, Y., Shimizu, A., 2009. Cortical representation of rhythm: cross-language perspective. J. Acoust. Soc. Am. 125, 1184–1192.

Kato, A., Iwasa, Y., Shimizu, A., 2010. Cortical representation of rhythm: cross-language perspective. J. Acoust. Soc. Am. 127, 3012–3022.

Kato, A., Iwasa, Y., Shimizu, A., 2011. Cortical representation of rhythm: cross-language perspective. J. Acoust. Soc. Am. 129, 263–274.

Kato, A., Iwasa, Y., Shimizu, A., 2012. Cortical representation of rhythm: cross-language perspective. J. Acoust. Soc. Am. 132, 165–174.

Kato, A., Iwasa, Y., Shimizu, A., 2013. Cortical representation of rhythm: cross-language perspective. J. Acoust. Soc. Am. 133, 883–892.

Kato, A., Iwasa, Y., Shimizu, A., 2014. Cortical representation of rhythm: cross-language perspective. J. Acoust. Soc. Am. 135, 3508–3519.
