Preliminary study for intonation classification of imagined speech for brain-computer interface applications

Isabel Casso*, José Rouillard[†], Hakim Si-Mohammed[†], Nacim Betrouni[‡], François Cabestaing[†] and Anahita Basirat*
* Univ. Lille, CNRS, UMR 9193 SCALab, F-59000 Lille, France
[†] Univ. Lille, CNRS, Centrale Lille, UMR 9189 CRIStAL, F-59000 Lille, France
[‡] Univ. Lille, INSERM, CHU Lille, U1172, Lille Neuroscience & Cognition, F-59000 Lille, France
Emails: isabel.casso@univ-lille.fr; jose.rouillard@univ-lille.fr; hakim.simohammed@univ-lille.fr; nacim.betrouni@inserm.fr
francois.cabestaing@univ-lille.fr; anahita.basirat@univ-lille.fr

***Abstract*—In the current study, we focused on decoding speech prosody from EEG. Prosody (i.e., melody and rhythm of speech) is important during communication as it allows to convey emotion and meaning. However, it has received little attention in the field of brain-computer interfaces. To address this issue, we contrasted the production of two syllables, “ba” and “da”, produced mentally as an affirmation (e.g., “ba.”) or a question (e.g., “ba?”) using two different intonations. We focused on spectral features. After classification in the time-frequency domain, we found above chance-level accuracies in specific frequency ranges of the alpha band (7-12 Hz) early on during the production phase. We also obtained above chance-level results on a range of the low-beta band (16-20 Hz) during a late time window. Based on the visual inspection of topographies and the literature, we suggest that the results during the early time window, but not that during the late time window, reflect a genuine difference between imagined affirmation and question production. Future studies should provide more information about neural markers and underlying neuro-cognitive processes to improve the understanding of the imagined intonation production. This would pave the way for the development of speech-based BCI capable of differentiating intonation and prosody in general.**

***Index Terms*—Intonation, imagined speech, EEG, brain-computer interfaces, prosody**

I. Introduction

Spoken word production involves a set of complex processes such as conceptual preparation, lexical selection, morphological, phonological, and phonetics encoding and articulation [1]. Despite the complexity of the brain’s computations underlying these processes, recent studies have shown that it is possible to decode speech using electrocorticography (ECoG) recordings (e.g., [2]) even during dialogue [3]. These results, currently obtained for the overt speech, pave the way for the development of brain-computer interfaces (BCI) systems using imagined speech for the case of individuals who are unable to speak overtly.

Contrary to overt speech, imagined speech does not involve articulation. According to the dual-stream prediction model [4], the articulatory trajectory is planned in the inferior frontal gyrus and other premotor areas. During imagined speech, the planned trajectory bypasses the primary motor cortex and is simulated internally. The somatosensory consequence of the simulated articulation is then estimated in the inferior parietal cortex. Then, an abstract auditory representation, formed around regions of posterior superior temporal gyrus and superior temporal sulcus, is derived from the estimation.

If the neural activities related to these processes can be detected and decoded, BCI systems could thus use the imagined speech setting for communication purposes, ultimately for developing prosthesis for individuals who cannot articulate (e.g., see [5], for locked-in syndrome due to stroke).

Even though most studies on overt and imagined speech decoding to date have used ECoG recordings, a few studies thus far have shown that some aspects of imagined speech can be decoded using EEG (for a review, see [6]).

Although critical during communication, as it allows to convey emotion and meaning, prosody has not been a common subject of research in the BCI field; generally, paradigms focus on discriminating vowels, syllables, and short words without examining prosodic contrast.

To our knowledge, only one EEG study to date has reported the results on prosody using imagined speech. In this study, Li and Chen [7] examined the decoding of different Mandarin tones. The classification across all four tones reached an 80.1% accuracy using audio-visual stimuli and 67.7% using only visual stimuli.

In the present work, we studied the possibility of decoding a particular component of speech prosody from EEG signals, which is intonation. Our task involved the production of two syllables, ”ba” and ”da”, produced mentally as an affirmation (e.g., ”ba.”) or a question (e.g., ”ba?”). We focused on spectral features. In fact, studies on overt speech detected alpha (7-12.5 Hz) and beta (12.5-30 Hz) power decreases over speech motor areas before and during articulation in tasks such as picture naming or verb generation tasks. In addition to motor aspects of speech production, this phenomenon seems to be related to the process of retrieval of conceptual and lexical information from memory. Increases in theta band (3.5-7 Hz) have also been observed. This has been related to executive control processes and may reflect the need for more control during speech production (e.g., when bilingual participants select the wrong language during word production) [8]. We only focused on motor aspects, as our task does not involve memory retrieval or executive control processes.

In a recent study on short sentence production (e.g., ”Do you understand me?”), Dash and colleagues [9] observed a significant contribution of delta oscillations (0.5-3.5 Hz) in decoding overt and imagined speech. Although the contribution of low-frequency oscillations during speech production remains unclear, such oscillations are thought to be related to the processing of prosodic aspects of utterances during speech perception [10].

Based on the above mentioned literature, we hypothesized that delta, alpha and beta bands could be the frequency bands of interest for observing and decoding the contrast between affirmations and questions in our study.

II. Methods

A. Participants

Six right-handed native french speakers (4 females) were included in this study. All of them were naive to the experiment. Participants had an average age of 22.7 years old (±1.37 years). All had self-reported normal or corrected-to-normal vision without any hearing, language, memory or learning problems. One additional participant was tested but excluded from analyses due to noisy data.

B. Experimental task design

We used four visual prompts: ”ba.”, ”ba?”, ”da” and ”da?”. The syllables, and not words or sentences, were used to minimize semantic processing. The experiment was divided into eight overt and eight imagined speech blocks. An overt speech block is followed by an imagined speech block and vice-versa. The order of blocks was counterbalanced across participants. Each block comprised 24 trials (6 repetitions of each prompt). The order of trials in each block was pseudo-randomized. Overall, 48 trials per prompt have been presented to participants for each block type. In this paper, we focused on imagined speech blocks.

Figure 1 shows the structure of each trial. Participants were first presented with a fixation cross for 1 second. Afterward, a written syllable was presented for 1.5 seconds. The participants were asked to imagine producing the prompt with the appropriate intonation (i.e., affirmation or question) immediately after the disappearance of the visual cue. The next trial started automatically 2.5 s after the disappearance of the prompt.

C. Data acquisition, pre-processing, and analyses

EEG were recorded with the BioSemi 64-channels gel-electrode system at a sampling frequency of 2048 Hz in the open-source software OpenViBE. The recordings were carried out in an isolated room. Participants were seated at approximately 1m from a 15 inch LCD monitor.

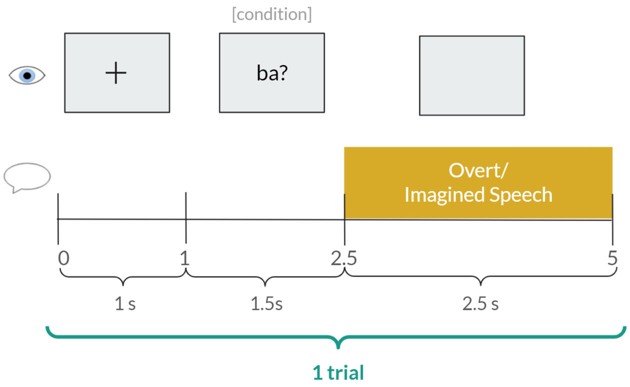

Fig. 1. The structure of experimental trials. Prompts were ”ba.”, ”ba?”, ”da.” and ”da?”.

An average re-referencing was applied to all signals; then, data passed through a one-pass, zero-phase, notch filter from 49.12 to 50.88 Hz to avoid electrical line noise.

All epochs or samples with peaks over 750 μV were rejected and not included in the analyses. Considering previous research results on the decoding of imagined tones [7], we carried our analysis defining a region of interest of six electrodes F5, FC3, P5, CP3 C3 and C4 as in [7]. These electrodes are commonly used in studies on imagined speech as they are thought to capture the activities of regions involved in speech production [6].

Considering only the data from these region, we performed a time-frequency analysis with a frequency-dependent time window length, which consisted in taking ten intervals ranging from 0.5 Hz to 45 Hz, comprising bands delta (0.5-3.5 Hz), theta (3.5 - 7 Hz), low-alpha (7-9 Hz), high-alpha (9-12.5Hz), low-beta (12.5-20 Hz), high-beta (20-30 Hz), and gamma (30-45 Hz). The pre-processed data were filtered within each frequency range. We extracted the epochs of the two conditions (affirmation and question) considering the entire 2.5 production phase adding 0.2 s prior to production phase onset. Baseline correction was applied considering this time window (0.2 s). We obtained the time-frequency representations (TFRs) of every epoch with window size dependent on the upper and lower values of the ten frequency intervals with a resolution of 10 Hz and concatenated in a sliding window manner with a 0.11 s overlap.

We considered a BCI benchmark algorithm [11] [12] also used in the analysis by [7] composed of a Common Spatial Pattern (CSP) as a filtering step and a Linear Discriminant Analysis (LDA) algorithm for class discrimination. We selected six filters per class (i.e., affirmation and question).

The TFR data, consisting of spatial features in the time-frequency domain from different time windows of the production phase, was shuffled and split into train and test sets using 67% and 33% of data, respectively, ensuring the equal representation of classes. We calculated the covariance matrices using the train data.

The CSP + LDA pipeline was trained and tested over a 5-fold cross-validation. The chance level was calculated considering the number of test samples fed into the classifier for a 95% classification confidence as proposed in [13].

III. Results

Figure 2 shows the time-frequency classification accuracies obtained with test TFR data from all subjects. We obtained a maximum classification accuracy of 57% above chance-level early on during the production phase around 0.1 s in 7-12 Hz frequency band and during a late time window at about 1.8 s in 16-20 Hz frequency band. These bands correspond to alpha and low-beta, respectively.

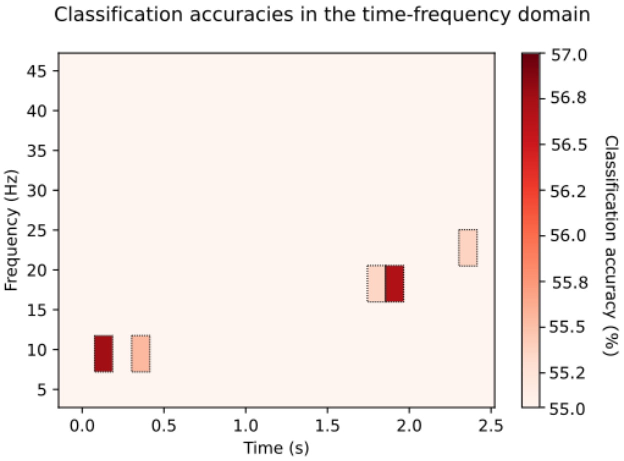

Fig. 2. Classification results of imagined speech intonation in the time-frequency domain. The above chance level results (over 55%) are indicated by dotted rectangles.

To better understand the classification results, we checked the topography maps on these bands and time windows for imagined affirmation and question production. Figures 3 and 4 show the topography averaged across subjects. After visual inspection of these figures, we can observe distinctions between the two intonation conditions on the electrodes we have selected for analysis based on previous research [6], [7]. Regarding the alpha band, the contrast between two conditions can be observed during the time window when the classification accuracy was above chance level (i.e., around 0.1 s and 0.3 s). Regarding the low-beta band, the contrast can be seen around 1.8 s corresponding to the time window during which the classification accuracy was also above chance level. However, other electrodes, not included in our classification analysis and thus not tested, seem to show larger contrasts (e.g., occipital electrodes).

We also examined the topography maps of each subject on all the frequency bands defined in section II for imagined affirmation and question production. For brevity, these topographies are not shown, but it is noteworthy to mention that we observed inter-individual variability regarding both timing and localization of activities.

IV. Discussion

Several BCI studies to date have focused on decoding phonemes, syllables, and words from brain signals. These studies are interesting for developing speech prostheses and speech synthesizers for various medical applications, from reeducation to communication devices (e.g., [5]). However, the focus on decoding prosody is scarce.

Prosody is an essential aspect of speech communication allowing to convey meaning and emotion. The goal of the current exploratory study was to examine EEG markers of imagined intonation production and test whether the intonation of affirmations and questions could be decoded using an EEG signal.

Given the scarcity of EEG research on speech intonation production, we must first ensure that distinguishing features on cortical activity for various intonations will allow a BCI to perform its decoding task. We explored data from a set of electrodes previously used in the decoding of imagined mandarin tones production [7], namely F5, FC3, P5, CP3, C3, C4; above cortical regions involved in prosody production [14] [15] and imagined speech production [4].

We performed a time-frequency analysis with a time window dependent on frequency range and classification with a CSP+LDA pipeline. As mentioned in the Introduction, we expected to observe results in alpha and beta bands as they have been reported in speech production tasks [8]. The delta band was another frequency band of interest as it has been reported in decoding sentences including imagined affirmations and questions [9] (although the focus of the study was not on prosody decoding).

The results showed above chance-level accuracies during an early time window in the alpha band (between 7 Hz and 12 Hz) and during a late time window in the low-beta band (between 16 Hz and 20Hz). The contrast between affirmations and questions during the early time window could be related to the planning phase of speech production and is consistent with speech production literature. However, the results during the late time window would occur too late. We believe that this latter result may be related to processes other than imagined speech production *per se*. Contrary to our hypothesis, we did not observe above chance-level results on the delta band. This result could be because the stimuli we used were too short for a slow frequency band. Dash and colleagues [9] who reported the importance of the delta band for imagined speech decoding had used more prolonged stimuli (e.g., ”Do you understand me?”). The involvement of the delta band in prosody production should be investigated in future studies.

In sum, although our results should be interpreted cautiously due to a limited number of participants, they show for the first time that (1) intonation could be decoded using EEG in an imagined speech setting and (2) spectral features would be appropriate features for decoding speech prosody. It is noteworthy that the study of Li and Chen [7] reported the results on tone production without using spectral features and

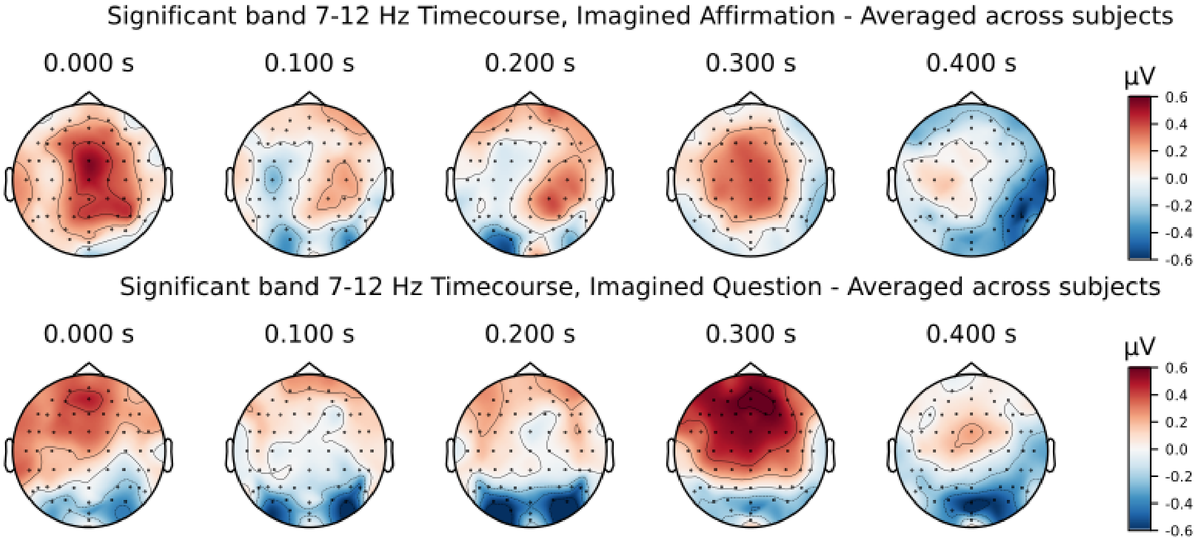

Fig. 3. Topography maps, averaged across subjects, on 7-12 Hz frequency band between 0 to 0.4 s of production phase for imagined affirmations (top) and imagined questions (bottom).

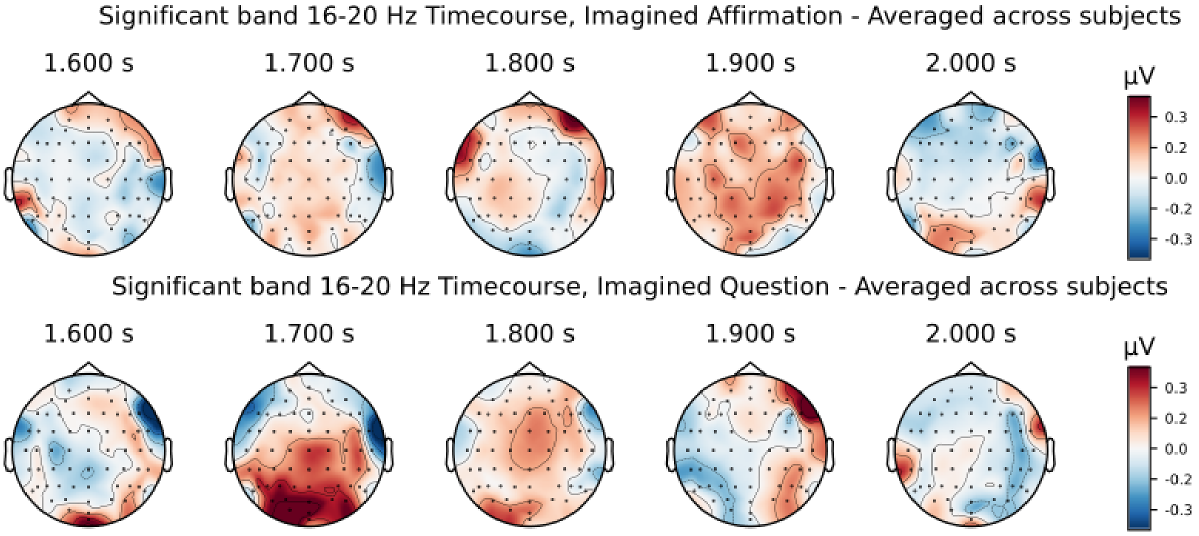

Fig. 4. Topography maps, averaged across subjects, on 16-20 Hz frequency band between 1.6 to 2 s of production phase for imagined affirmations (top) and imagined questions (bottom).

the study of Dash and colleagues [9] on decoding imagined speech using spectral features did not focus on prosody.

One of the challenges when examining imagined speech, and in BCI studies in general, is that the results are not time-locked to the beginning of the target process, i.e., speech production in the current study. As in most studies on imagined speech, our analyses were time-locked on the disappearance of the prompt: subjects were asked to begin to produce speech mentally at this moment. In future studies, to better investigate EEG markers, it would be necessary to find ways to estimate more accurately the time-frame of the production by, for example, asking participants to report the end of their production as in [16]. The inter-subject variability that we observed in the current study may be related to this issue.

In addition, given the limited literature on prosody production in the field of BCI, we decided to use the same electrodes as selected in the study of Li and Chen [7] which was on imagined tone production. This low spatial density may have contributed to the classification's poor performance. Based on the visual inspection of topographies, including other electrodes in the analyses seems required. Another critical issue that should be addressed in future studies on prosody production for BCIs concerns appropriate EEG markers and underlying neurocognitive processing. We believe that focusing on ”blind” classification approaches as in the current study as well as the study of [7] would not be sufficient to develop BCIs as reliable systems need to be based on knowledge on underlying cortical processes.

In conclusion, this preliminary study addressed the high complexity and feasibility of intonation decoding of imagined speech from data obtained through EEG.

Furthermore, we believe that future research on imagined speech decoding should provide more information about neural markers and frequency bands of interest. This would improve the understanding of the underlying neurocognitive process and favor the development of speech-based BCIs capable of differentiating prosody.

V. Acknowledgment

We would like to thank all the participants in this study, Laurent Ott for his help in preparing the experiment, and David Devos for discussions. This work was supported by a grant from the Carnot Cognition Institute.

References

[1] W. J. Levelt, A. Roelofs, and A. S. Meyer, “A theory of lexical access in speech production,” *Behavioral and brain sciences*, vol. 22, no. 1, pp. 1–38, 1999.

[2] E. M. Mugler, J. L. Patton, R. D. Flint, Z. A. Wright, S. U. Schuele, J. Rosenow, J. J. Shih, D. J. Krusienski, and M. W. Slutzky, “Direct classification of all american english phonemes using signals from functional speech motor cortex,” *Journal of neural engineering*, vol. 11, no. 3, p. 035015, 2014.

[3] D. A. Moses, M. K. Leonard, J. G. Makin, and E. F. Chang, “Real-time decoding of question-and-answer speech dialogue using human cortical activity,” *Nature communications*, vol. 10, no. 1, pp. 1–14, 2019.

[4] X. Tian and D. Poeppel, “The effect of imagination on stimulation: the functional specificity of efference copies in speech processing,” *Journal of cognitive neuroscience*, vol. 25, no. 7, pp. 1020–1036, 2013.

[5] F. H. Guenther, J. S. Brumberg, E. J. Wright, A. Nieto-Castanon, J. A. Tourville, M. Panko, R. Law, S. A. Siebert, J. L. Bartels, D. S. Andreasen, *et al.*, “A wireless brain-machine interface for real-time speech synthesis,” *PloS one*, vol. 4, no. 12, p. e8218, 2009.

[6] J. T. Panachakel and A. G. Ramakrishnan, “Decoding covert speech from eeg-a comprehensive review,” *Frontiers in Neuroscience*, vol. 15, p. 392, 2021.

[7] H. Li and F. Chen, “Classify imaginary mandarin tones with cortical eeg signals.,” in *INTERSPEECH*, pp. 4896–4900, 2020.

[8] V. Piai and X. Zheng, “Speaking waves: Neuronal oscillations in language production,” in *Psychology of learning and motivation*, vol. 71, pp. 265–302, Elsevier, 2019.

[9] D. Dash, P. Ferrari, and J. Wang, “Role of brainwaves in neural speech decoding,” in *2020 28th European Signal Processing Conference (EUSIPCO)*, pp. 1357–1361, IEEE, 2021.

[10] O. Ghitza, “Acoustic-driven delta rhythms as prosodic markers,” *Language, Cognition and Neuroscience*, vol. 32, no. 5, pp. 545–561, 2017.

[11] F. Lotte, M. Congedo, A. Lécuyer, F. Lamarche, and B. Arnaldi, “A review of classification algorithms for eeg-based brain–computer interfaces,” *J Neur Eng*, vol. 4, no. 2, p. R1, 2007.

[12] C. S. Nam, A. Nijholt, and F. Lotte, *Brain–computer interfaces handbook: technological and theoretical advances*. CRC Press, 2018.

[13] G. Müller-Putz, R. Scherer, C. Brunner, R. Leeb, and G. Pfurtscheller, “Better than random: a closer look on bci results,” *International journal of bioelectromagnetism*, vol. 10, no. ARTICLE, pp. 52–55, 2008.

[14] F. H. Guenther, *Neural control of speech*. Mit Press, 2016.

[15] L. Aziz-Zadeh, T. Sheng, and A. Gheytanchi, “Common premotor regions for the perception and production of prosody and correlations with empathy and prosodic ability,” *PloS one*, vol. 5, no. 1, p. e8759, 2010.

[16] F. Bocquelet, T. Hueber, L. Girin, S. Chabardès, and B. Yvert, “Key considerations in designing a speech brain-computer interface,” *Journal of Physiology-Paris*, vol. 110, no. 4, pp. 392–401, 2016.