Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 133–137
Marseille, 11–16 May 2020

Cortical Speech Databases for Deciphering the Articulatory Code

Harald Höge
Universität der Bundeswehr München
harald.hoege@t-online.de

Abstract

The paper relates to following 'AC-hypotheses': The articulatory code (AC) is a neural code exchanging multi-item messages between the short-term memory and cortical areas as the vSMC and STG. In these areas already neurons active in the presence of articulatory features have been measured. The AC codes the content of speech segmented in chunks and is the same for both modalities - speech perception and speech production. Each AC-message is related to a syllable. The items of each message relate to coordinated articulatory gestures composing the syllable. The mechanism to transport the AC and to segment the auditory signal is based on ϴ/γ-oscillations, where a ϴ-cycle has the duration of a ϴ-syllable. The paper describes the findings from neuroscience, phonetics and the science of evolution leading to the AC-hypotheses. The paper proposes to verify the AC-hypotheses by measuring the activity of all ensembles of neurons coding and decoding the AC. Due to state of the art, the cortical measurements to be prepared, done and further processed need a high effort from scientists active in different areas. We propose to launch a project to produce cortical speech databases with cortical recordings synchronized with the speech signal allowing to decipher the articulatory code.

Keywords: articulatory code, speech perception, speech production, auditory cortex, ventral sensory cortex, short-term memory

1. Introduction

Deciphering the neural code of sensors and actors and deciphering the mechanism of transporting the neural code between the short-term memory[1] and sensor/actor processing neural areas are challenging fields in neuroscience. We regard a special neural code which transports multi-item messages, where each message corresponds to an object and each item corresponds to the parts composing the object. Large progress has been done in understanding the mechanism to transport multi-item messages between distant neural areas (Lisman and Jensen, 2013). The transport is performed by ϴ/ɣ-oscillations. Each ϴ-cycle ports the message of an object. Embedded in each ϴ-cycle are ɣ-cycles porting the parts of the object. The porting mechanism is performed by ϴ/ɣ-oscillations modulated with the neural spike-patterns. Thus, to decipher a neural code porting multi-item messages, following knowledge must be gained:

- the complete set of the parts (items) of all objects
- for each message the ϴ/ɣ-structure (position of the items within a ϴ-cycle)
- the spike-patterns of the items

In rare cases a multi-item neural code has been deciphered. An example is the code generated by 'cell-place' neurons (O'Keefe and Dostrovsky, 1971). This neural code is used to determine the spatial position of a rat.

The paper regards a special multi-item neural code - the articulatory code (AC). It is hypothesized that the AC is the same for both modalities: speech perception and speech production. The objects of its messages are syllables, the parts of the objects (the items) are articulatory gestures composing syllables. To decipher the AC, the set of items composing the neural code, must be explored. This is a great challenge, because the activity of many ensembles of neural cells with the needed spatial resolutions must be measured (see chapter 2).

There exist two areas of application, where a deciphered AC is needed. The first area concerns the development of Brain-Computer-Interfaces (BCI) also called Brain-Machine-Interfaces (BMI) for restoring the communication abilities of handicapped persons (Chaudhary et. al., 2016; Ramsey et. al., 2017). In the second area the knowledge about the AC can be applied to develop cortical models of perception, speech production and speech learning leading to cortical inspired automatic speech recognition systems (Mitra et. al., 2010), speech synthesis systems and learning systems.

To decipher the nature of the AC, several models have been proposed. But a final prove of the correctness of the models is not given. As described in chapter 3, this paper is based on a model of the AC defined by following hypotheses called **AC-hypotheses**:

- Each message relates to a ϴ-syllable[2] as object
- Each item relates to an articulatory gesture composing a ϴ-syllable.
- The size of the set of different articulatory gestures composing ϴ-syllables is in the range of 1000 articulatory gestures
- The AC is the same for perception and production

In order to decipher the AC, these hypotheses must be verified or adjusted by cortical measurements, where the activities of neurons generating or decoding the AC are observed. Due to the high effort needed, the paper proposes to set up a project to produce cortical speech databases containing the neural patterns of the articulatory gestures, synchronized with the ϴ/γ-oscillations and with the speech signal. The resulting databases are the basis to decipher the AC.

The paper is organized as follows. Chapter 2 gives a short overview of the state-of-the-art methods of cortical measurement. Chapter 3 reports about investigations done so far leading to the AC-hypotheses. Chapter 4 proposes a project to generate the cortical speech databases needed to decipher the AC.

2. Recording

To measure the activity of neurons non-invasive and invasive methods have been developed. For deciphering the AC, the activity of many neurons must be measured synchronously. For this issue methods with high spatial resolution are needed. The spatial resolution of non-invasive methods as electro-encephalogram (EEG) or

[1] The term 'short-term memory' is equivalent to the term 'working memory'.

[2] The term ϴ-syllable has been introduced by (Ghitza, 2013)

magnetoencephalography (MEG) is too low. Invasive methods as described by (Buzsáki et. al., 2016) have the needed spatial resolution but the number of neurons to be measured synchronously is still quite restricted: *The invasive method 'Electrocorticography' (ECoG) is becoming an increasingly popular tool for studying various cortical phenomena in clinical settings. It uses subdural platinum–iridium or stainless-steel electrodes to record electric activity - the LFP - directly from the surface of the cerebral cortex, thereby bypassing the signal-distorting skull and intermediate tissue. The spatial resolution of the recorded electric field can be substantially improved (<5 mm^2) by using flexible, closely spaced subdural grid or strip electrodes* (see fig. 1).

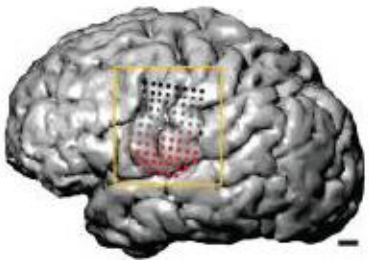

Figure 1: a typical measurement scenario in a clinical setting. The measurements are done on persons during their clinical treatment of epilepsy (Bouchard and et.al.,2013): *ECoG measurements in the ventral sensorimotor cortex (vSMC), MRI reconstruction of single subject brain with electrodes (dots); about 30 electrodes were connected to neurons delivering useful information*

Electrical events at deeper locations can be explored by inserting metal or glass electrodes, or silicon probes into the brain to record the LFP. Recording the wide-band signal (direct current to 40 kHz) — which contains both action potentials and other membrane potential-derived fluctuations in a small neuronal volume — using a microelectrode yields the most informative signal for studying cortical electrogenesis (see fig.2).

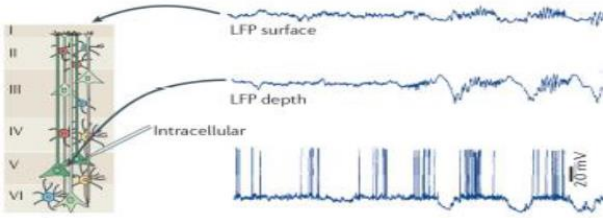

Figure 2: Simultaneously recorded LFP traces from the superficial ('surface') and deep ('depth') layers of the motor cortex in an anaesthetized cat and an intracellular trace from a layer 5 pyramidal neuron (Buzsáki et. al., 2016).

3. The Nature of the Articulatory Code

This chapter gives an overview about the findings from different disciplines as neuroscience, phonetics leading to the AC-hypotheses formulated in the introduction. For speech production and speech perception the definition of the AC is closely related to the coordinated action of the articulators. This action leads to the concept of articulatory gestures. The term 'articulatory gesture' has been defined already by phoneticians describing the action of a single articulator (Browman and Goldstein,1989). In this paper a broader meaning[3] is used for this term: each '**articulatory gesture**' is defined as the actions of **all** articulators producing sequence of segments[4]. As an example, the syllable 'stop' can be split into 3 articulatory gestures /st/, /o/, /p/. If necessary, the 'phonetic defined' articulatory gestures are called 'narrow articulatory gestures' and the 'neural defined' articulatory gestures are called 'broad articulatory gestures'.

This chapter starts in section 3.1 with cortical measurements already done. Section 3.2 treats the findings from phonetics and evolutionary research. Combined with the concept of Θ/ɤ- oscillations in section 3.3 these findings are combined leading to the AC-hypotheses.

3.1 Articulatory Features

ECoG measurements, made in the ventral sensorimotor cortex (vSMC) and in superior temporal gyrus (STG) show, that both modalities – speech perception and speech production - work with manner & place like features as defined by phoneticians (Ladefoged, Johnson, 2015). In the following, these features as implemented in the cortex are called **articulatory features**. As discussed in section 3.3 the use of articulatory features for both modalities is the basis for the hypothesis, that articulatory gestures constitute the items of the AC.

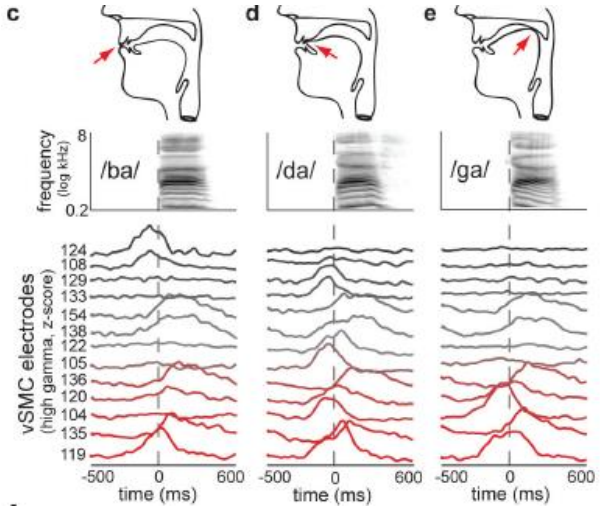

Figure 3: ECoG measurements in the vSMC during speech production (Bouchard and et.al., 2013). Rows c, d, e: Above: the articulatory position and the spectra of the utterances /ba/, /da/ and /ga/, with different articulatory features for 'place'. Below: The z-score of electrodes active for different articulatory places.

Experiments for measuring articulatory features have been performed by (Bouchard and et.al., 2013) for speech

[3] This definition is inspired by the neural implementation of complex steering processing of combined actions of articulators (see section 3.1)

[4] Segments are phoneme-like phonetic units (Browman and Goldstein,1989)

production and by (Mesgarani et.al., 2014) for speech perception. The experiments were performed with utterances of syllables with different articulatory features as shown in fig. 3.

In the vSMC (Bouchard and et.al., 2013) found somatotopically[5] ordered populations of neurons, which correlate to the concept articulatory features in the control of single articulators[6]. But additionally, ensembles of populations of neuron were found whose activity hint to more complex mechanisms for steering the articulators*: It is not any single articulator representation, but rather the coordination of multiple articulator representations across the vSMC network that generates speech. Analysis of spatial patterns of activity revealed an emergent hierarchy of network states, which organized phonemes by articulatory features.* From these findings it can be concluded, that the mechanism of the steering the articulators in speech production is based on a transformation of a complex structure of bundles of articulatory features describing articulatory gestures to the coordinated steering process of single articulators. This concept is in line with the phonetic framework of scores (Nam et. al., 2012) describing the relation between articulatory features and segments.

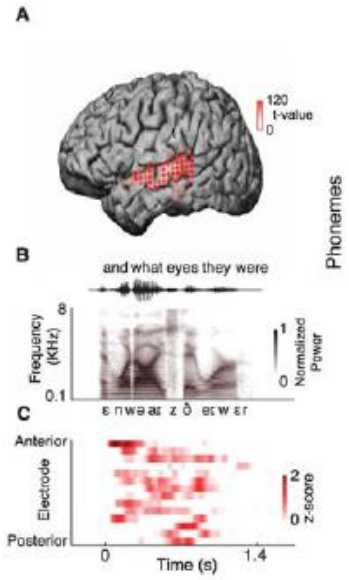

Figure 4: A: Position of the electrodes in the STG; B: recorded sentence together with short term spectrum; C: response (z-score) of the electrodes

In speech perception, (Mesgarani et.al., 2014) measured the activity of neurons in the STG (see fig.4) and related them to articulatory features (see fig. 5) and phonemes. They found*: Furthermore, selectivity of phonemes is organized primarily by manner of articulation distinctions and secondarily by place of articulation, corresponding to the degree and the location of constriction in the vocal tract, respectively.* Given these findings, the neurons in the STG perform a complex transformation from the speaker dependent auditory signal to speaker independent articulatory features. In (Höge, 2017) it is hypothesized, that this transformation is done in the STG with the use of by Θ/ɤ-oscillations generated in the vSMC.

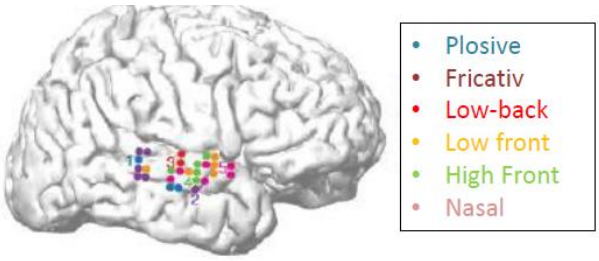

Figure 5: STG-position of electrodes active for specific manner and place features

In speech perception to the author knowledge no cortical measurements have been done hinting to ensembles of neurons with a complex structure of composed articulatory features. A hint is given by the dual stream hypothesis (Hickock, Poeppel, 2007) shown in fig. 6.

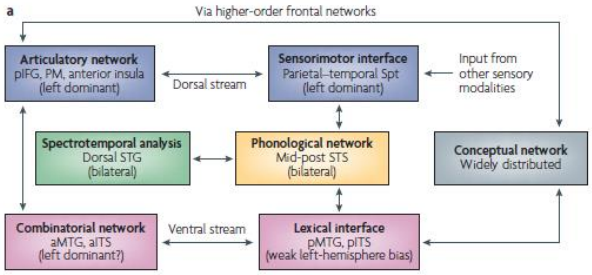

Figure 6: the dual-stream model of the functional anatomy of language (Hickock, Poeppel, 2007)

The auditory signal (termed 'spectro-temporal analysis') is transformed to a 'phonological network', which performs the processing of articulatory features. Following the 'Dorsal stream', the articulatory features enter the sensorimotor interface (vSMC)[7] and enter finally in the 'articulatory network'. Within the articulatory network the generation of articulatory gestures is performed, as given by (Hickock and Poeppel, 2007) hypotheses: *We suggest that there are at least two levels of auditory–motor interaction — one involving speech segments[4] and the other involving sequences of segments. Segmental-level processes would be involved in the acquisition and maintenance of basic articulatory phonetic skills.* The sequences of segments constitute articulatory gestures.

3.2 Phonetic and Evolutionary Findings

*P*honeticians *defined a*rticulatory gestures, *which* describe the content *of* speech *by sequences of segments*. The first

[5] Using electrical stimulation, (Penfield et. al,1937; Foerster, O., 1931) already described the somatotopic organization of face and mouth representations in human vSMC

[6] In (Ramsay et.al., 2017) this somatotopically ordered neural area is used for a BCI.

[7] It is hypothesized that in this area input from other sensory modalities as lip-reading (Park et. al., 2016) is integrated.

step in this development was the definition of phonemes defined by articulatory features. This concept has been extended to describe additionally the dynamics of each articulator leading to the concept of *narrow* articulatory gestures (Browman and Goldstein, 1989) and has been extended further to the concept of 'scores' defining the combined action of the articulators to produce segments (Nam et.al., 2012). Measuring the dynamics of gestures (Byrd, 1994; Mooshammer et.al., 2012) it became evident, that *articulatory* gestures are related to the positions (onset, vowel, coda) within a syllable. Thus, it is concluded that the *activity of the neuronal network* steering the articulators is related to the structure of a syllable. This concept was used to defining three kinds of articulatory gestures (Höge, 2018): The onset-gestures (O-gesture), defined by the consonants of the beginning of a syllable, the vowel gesture (V-gesture) defined by the vowels of the center of the syllable and the coda-gestures (C-gesture) defined by the consonants building the coda of a syllable. Using EMA-data from a *German and an English* articulatory speech database, *for both languages* about 1000 OVC-gestures have been detected. This approach is limited by the precision needed to determine the temporal boundaries (end, beginning) of the open-close cycle of a syllables by observing the articulators. The boundaries of the syllables are needed to determine the set of O-V-C-gestures. Especially when regarding the consonants between two neighbored syllables, the problem arises, which consonants belong to the C-consonants and which belong to the O-consonants. As discussed in section 3.3, the temporal boundaries of a syllable as processed by the cortex, are given by the phase of the Θ-oscillations leading to a neural definition of a syllable: the Θ-syllable[2].

Another approach exploring the neural process of steering the articulators is based on the principles of evolution (Darwin, 1871). (MacNeilage ,1998) treats the quasi rhythmic opening and closing gesture of the mandibular as a Frame with Slots to be filled (F/S-theory). Each frame represents a syllable with slots representing clusters of phonemes relating to articulatory gestures. The main argument of the F/S-theory is given by the observation that errors observed in speech production have following properties: *most errors in speech production are exchange errors. The central fact about exchange errors is that in virtually all segmental exchanges, the units move into a position in syllable structure similar to that which they vacated: syllable-initial consonants exchange with other syllable-initial consonants, vowels exchange with vowels, and syllable-final consonants exchange with other syllable-final consonants.*

3.3 The Articulatory Code

As stated by the AC-hypotheses, it is claimed that the AC is a multi-item neural code transmitted by Θ/ɤ-oscillations and that the objects coded are syllable whose items are articulatory gestures. Further it is claimed, that the AC is the same for production and perception.
The close relationship between both modalities is supported by the findings of (Assaneo and Poeppel, 2018), where during perception, the Θ/ɤ-oscillations have been observed synchronously in auditory and motor cortices. The concept of syllables is already stated in section 3.1 and 3.2. Concerning the cortical implementation of syllable-oriented processing during perception (Giraud and Poeppel, 2015) state: *Recent data show that delta, theta and gamma oscillations are specifically engaged by the multi-timescale, quasi-rhythmic properties of speech and can track its dynamics. We argue that they are foundational in speech and language processing, 'packaging' incoming information into units of the appropriate temporal granularity…The faster 'phonemic' gamma oscillations are 'nested' in the slower 'syllabic' oscillations. Through theta-gamma nesting, concurrent* ***syllabic*** *and* ***phonemic*** *analyses can remain hierarchically bound. Nesting is manifest and can be functionally relevant only if there is a minimum ratio across frequencies. In the theta-gamma nesting pattern that emerges in the human primary auditory cortex in response to speech,* ***there is a frequency ratio of about 4****, suggesting that about 4 cycles of the higher frequency occur during one cycle of the lower one*.
Thus, during the process of the segmentation of the auditory signal into syllables, items of the AC are constructed: Within each Θ-cycle, each syllable is segmented into windows given by the boundaries of the ɤ-cycles. Within these windows the articulatory gestures constituting the syllable are classified leading to the items of the AC (Höge, 2018).
To summarize, the main findings leading to the AC-hypotheses are derived from perception and the interaction of the Θ/ɤ-oscillations in the STG and the vSMC.

4. Project Proposal

To verify the AC-hypotheses the feasibility to measure the activity of the neurons involved in coding/decoding the AC must be checked. For this issue a pre-project must be started. Due to the knowledge gained so far in the construction of AC, most information is given by the cortical processing steps performed in perception (see section 3.3). But the precise neural area, where the AC is generated and sent to the short-term memory is unknown. Due to the property of the vSMC to integrate all information from all sensors it seems that the vSMC is the area, where the AC is constructed (Höge, 2017). In that area, all ensembles of neurons active for the set of articulatory gestures must be accessed. Due to the assumption that about 1000 ensembles must be measured, advanced pads must be used.

As soon as the feasibility is proven, a project can be set up with following activities: the Θ- and ɤ- oscillations together with the modulated spike patterns must be recorded. Synchronous to the neural measurements the speech signal must be recorded and notated with phonetic annotations. Given these recordings, the phones uttered during a ɤ-cycle can be collected, which define the articulatory gestures coded within each of the ɤ-cycles nested in a Θ-cycle. Helpful in aligning the annotated speech signal to the Θ-oscillation would be the simultaneous recording of the articulatory features in the STG. These issues lead to the work-packages

- Defining a corpus of sentences covering the structure of the syllables of a language.
- Detecting the areas, where the AC is built.

- Recording the neuronal activities including the Θ/ɤ-oscillations and the speech signal.
- Label the speech signal into phonemes
- Synchronize the Θ/ɤ-oscillations to the phonetic labels
- Determine the set of articulatory gestures

5. Conclusion

Although many hints concerning the structure of the articulatory code as defined in the AC-hypotheses, we are far away to decipher the articulatory code. Due to state of the art for measuring the activities of the human brain, nowadays only invasive measurement as ECoG deliver the needed spatial and temporal resolution to observe the activity of ensembles of neurons synchronously. A main challenge is the large number of neurons, which must be measured to capture the activity of all neurons relating to the items of the code. To overcome this problem a large project must be launched. The resulting cortical speech databases should be made publicly accessible.

6. Bibliographical References

Assaneo M.F., and Poeppel, D. (2018). The coupling between auditory and motor cortices is rate-restricted: Evidence for an intrinsic speech-motor rhythm. *Science Advanced*

Byrd, D. (1994). Articulatory Timing in English Consonant Sequences. *Thesis UCLA California*

Bouchard, K.E., Mesgarani, N., Johnson., K. and Chang E.F. (2013). Functional organization of human sensorimotor cortex for speech articulation. *Nature*, 21, 495 (7441): 327–332

Browman, C.P. and Goldstein L. (1989). Articulatory Gestures as Phonological Units. *Haskins Laboratories Status Report on Speech Research*, 99: 69–101.

Buzsáki, G., Anastassiou, C. A., and Koch, C. (2016). The origin of extracellular fields and currents — EEG, ECoG, LFP and spikes. *Nat Rev Neurosci*ence 13(6): 407–420.

Chaudhary, U. et. al. (2016). Brain-computer interfaces for communication and rehabilitation. *Nat. Rev. ANeurol*., 12: 513-525.

Darwin, C. (1871). The descent of man. In Great Books, Encyclopedia Britannica, 1871.

Foerster, O. (1931) The cerebral cortex in man. *Lancet*, 221: 309–312

Giraud, A.L., and Poeppel, D. (2015). Cortical oscillations and speech processing: emerging computational principles and operations. *Nat. Neuroscience*, 15(4): 511-517.

Ghitza,O. (2013). The theta-syllable: a unit of speech information defined by cortical function. *Frontiers in Speech psychology*, Article 138:1-5

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

Höge, H. (2017). Human Feature Extraction - The Role of the Articulatory Rhythm. *Proc. Konferenz Elektronische Sprachsignalverarbeitung (ESSV)*

Höge, H. (2018). Using Elementary Articulatory Gestures as Phonetic Units for Speech Recognition. *Proc. ESSV2018*

Ladefoged, P. and Johnson K. (2015). A Course in Phonetics. *Wadsworth Cengage Learning*, 7th Edition, Boston.

Lisman, J. E., and Jensen, O. (2013). The Theta-Gamma Neural Code. *Neuron*, 77(6): 1002–1016

MacNeilage, P. F. (1998). The frame/content theory of evolution of speech production. *Behavioral and Brain Sciences*, 21:499–511

Mesgarani, N., Cheung, C., Johnson K., and Chang. E.F. (2014). Phonetic Feature Encoding in Human Superior Temporal Gyrus. *Science*, 343(6174):1006–1010

Mitra, V., Nam, H., Tiede, M., Epsy-Wilson, C., Saltzmann, E. and Goldstein, L. (2010). Robust word recognition using articulatory trajectories and gestures. *Proc. Interspeech*:2038-2041

Mooshammer, C., Goldstein, L., Nam H., and McLure S. (2012). Bridging planning and execution: Temporal planning of syllables. *Journal of Phonetics*: 374-389

Nam, H., Mitra, V., Tiede, M., Hasegawa-Johnson, M., Epsy-Wilson, C., Saltzmann, E. and Goldstein, L. (2012). A procedure for estimating gestural scores from speech acoustics. *J. Acoust. Soc. Am.*, Vol.132, No.6:3980-3989.

O'Keefe, J., and Dostrovsky, J. (1971). The hippocampus as a spatial map. Preliminary evidence from unit activity in the freely-moving rat. *Brain Res.*; 34:171–175.

H. Park, C. Kayser, G. Thut, and J. Gross: Lip movements entrain the observers' low-frequency brain oscillations to facilitate speech intelligibility. eLife, 5, e14521. DOI: 10.7554/eLife.14521 pp.1-17, May 2016

Penfield, W., Boldrey, E. (1937). Somatic Motor and Sensory Representation in The Cerebral Cortex of Man Studied by Electrical Stimulation. *Brain*, 60: 389–443

Ramsey, N. F., Salari, E., Aarnoutse, E.J., Vansteensel, M. J., Bleichner, M. G. and Freudenburg, Z. V. (2017). Decoding spoken phonemes from sensorimotor cortex with high density ECoG grid. *NeuroImage*, 2017.10.011: 1-11