Extraction of the Θ- and ɤ-Cycles Active in Human Speech Processing from an Articulatory Speech Database

Harald Höge

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

Abstract: For mimicking human articulatory speech production and perception, entrained Θ- oscillations, related to the rhythm of syllables, and nested ɤ- oscillations, related to the starting points of elementary articulatory gestures (EAGs), must be modeled. To train and test articulatory speech synthesis and speech recognition systems based on cortical principles, speech databases are needed, which are annotated by EAGs aligned to Θ- and ɤ-oscillations. Due to the limited abilities of current technology to measure neural activities in the cortex, such databases are not available. The paper describes a noninvasive methodology working with a phonetic annotated articulatory speech database, to detect the EAGs and align them to cycles of Θ- and ɤ-oscillations. The methodology is based on the combination of two methods, a perceptive and an articulatory method. The perceptive method mimics the extraction of *edge features* as done in the auditory cortex during perception, which determine the duration and phase of each Θ-, and ɤ-cycle. The articulatory method uses the kinematics of the movements of articulators, recorded by electromagnetic articulography, to determine the starting points of EAGs. The starting points retrieved are used to fine tune the Θ-, and ɤ-cycles estimated by the perceptive method. The success of finding ‘correct’ ɤ-and Θ-cycles is measured on the consistency of the starting points of the instances of each EAG found in the database.

1 Introduction

Θ- and ɤ-oscillations play an important role in human speech processing [1]. For mimicking human speech production and perception, the functionality of such oscillations must be integrated. The importance of the Θ- and ɤ-oscillations is given by two facts (see also fig.2):

- In speech production, the Θ- and ɤ-oscillations control the instances of starting **e**lementary **a**rticulatory **g**estures (EAGs, [7]). Mimicking articulatory control requires determination of the relationship between simultaneously recorded activities of neurons responsible for articulatory motor control for EAGs located in the ventral sensorimotor cortex (vSMC, [3,4]) and between the kinematics of the lips, jaw, tongue, and larynx.
- In speech perception, the Θ- and ɤ-oscillations of the speaker are reconstructed by the listener [1]. The main role of the Θ- and ɤ-oscillations is the segmentation of the continuous stream of the auditory signal into chunks produced by the EAGs of the speaker. Combined with the process of segmentation, is the process of transforming the chunks into an articulatory code composed by the articulatory features of the EAGs [5], [6]. In this process the Θ- and ɤ-oscillations organize the transfer of information as done by the clock of a computer. These cortical mechanisms must be implemented when mimicking neural activities transforming the auditory signal to the articulatory code.

Current research is far away from a complete neurobiological understanding of the mechanisms of motor control and perception. The main obstacle for neurobiological understanding is the current state of the art for measuring neural activities in the cortex. To decipher human speech processing, the activity of many neurons (in the order of 100 000 neurons) in different areas of the cortex must be measured simultaneously. Current state of the art is based on invasive measurement methods, where only a few neurons (in the order of 100 neurons) can be observed simultaneously in a single area [1], [3], as seen in fig.1.

To overcome this problem, the paper describes a noninvasive method, which allows to create a speech database labeled by EAGs aligned to Θ- and ɤ-cycles. This is an ambitious task, as neither the nature of the EAGs [3] nor the mechanism to generate Θ- and ɤ- oscillations [1] is deciphered sufficiently. As described in [5], [6], [7], and summarized in chapter 2, a theory based on a set of hypotheses has been developed to define the EAGs and their relation to Θ- and ɤ-oscillations. Based on this theory, in chapter 2 a methodology working with a phonemic labeled articulatory speech database is developed, which extract the EAGs together with the Θ- and ɤ-oscillations. The set of EAGs retrieved and the properties of their alignments to the Θ- and ɤ-oscillations are described in chapter 3.

First steps in this direction have been undertaken in [6], [7]. This paper presents an improved methodology, where neuronal motor control mechanisms are integrated to retrieve more reliable the Θ- and ɤ-oscillations.

2 Methodology for Extracting Θ- and ɤ- Oscillations and Related EAGs

The first two sections of this chapter give a short overview of the theory of the Θ- and ɤ- oscillations and their relations to EAGs. The final section of this chapter describes the methodology to extract the EAGs together with the Θ- and ɤ-oscillations.

2.1 The EAGs and their Relation to Θ- and ɤ- Oscillations

The concept of elementary articulatory gestures (EAGs) used in human speech processing has been introduced in [7]. It is hypothesized that EAGs are defined phonetically and consist of a set of gestures describing the opening and closing of the vocal tract. The phonetic nature of the EAGs is derived from invasive measurements on cortical neurons located in the belt of auditory cortex (superior temporal gyrus (STG), [12]) and located in the ventral sensory motor cortex (vSMC, [3]). In these areas, neurons have been found, whose sensitivity is related to the production and perception of articulatory gestures as shown in fig. 1. These gestures are related to the manner and place features defining phonemes [11]. Equivalent features observed by the sensitivity of the neurons, are called *articulatory features* (AFs) [5]. Thus, each EAGs is composed by a sequence of AF-defined movements of the articulators. Further, due to the equivalence of manner and place features and AFs, EAGs can be related to phonetic structures as syllables as done the C/F-theory [8] followed in this paper.

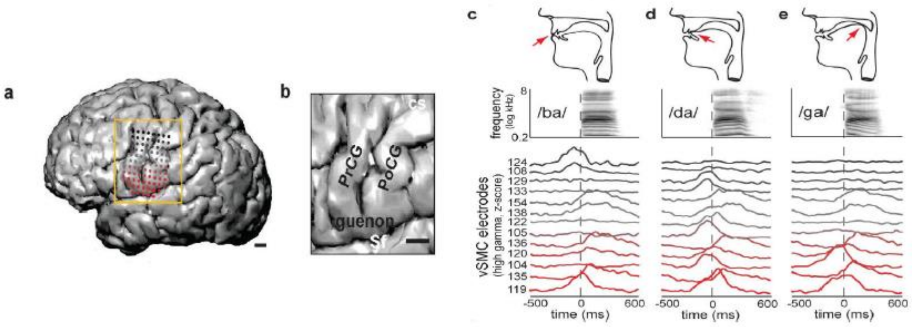

Figure 1 – ECoG measurements in the vSMC [3] **a**: MRI reconstruction of single subject brain with electrodes (dots); about 30 electrodes were connected to neurons delivering useful information **b**: Expanded view of vSMC: pre- and post-central gyri (PrCG and PoCG), central sulcus (CS), sylvian fissure (Sf) **c-d**: activity of selected electrodes during production of CV-syllables with different place of articulation

This theory hypothesizes that the syllable has been developed from gestures performing quasi-rhythmic gestures for opening and closing the vocal tract. In [6] these gestures are defined as **OVC**-*gestures* compost of three sets of gestures: the **O**-gestures describe the **O**pening of the vocal tract, the **V**-gestures describe the transition of opening to closing (**V**owel part) and **C**- gestures describe the **C**losing of the vocal tract. Each open-close cycle of OVC-gestures is related to a C*V*C* syllable. (C*/V* denotes a sequence of n= 0,1, … consonants/vowels; V* should contain at least one vowel). The definition of a syllable is not unique. In [6,7] the concept of a syllable is linked to cortical motor control leading to the concept of a *cortical syllable* (CS): A CS is defined by a C*V*C* structure, where the cortical control of the starting points of the phones realizing this structure is performed within a **single** cycle of an Θ-oscillation. This definition leads to the definition of the OVC-gestures: each OVC-gesture is defined as a specific sequence of AF-features related to a C*V*C* structure building a CS. As argued in [6], each CS defines an *articulatory code*, which is composed by three slots, where each slot is filled with the sequence of AFs given the OVC-gestures of an open-close cycle.

These findings lead to an architecture of human speech processing as shown in Fig. 2 [6].

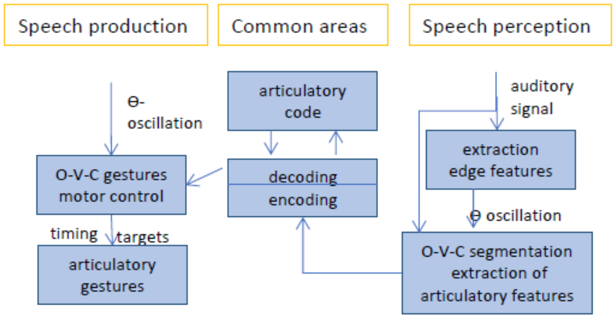

Figure 2 - Architecture of the processing blocks in speech production and speech perception [6].

2.2 The Θ- and ɤ-Oscillation

In [1] the role of Θ-oscillations has been investigated in the context of speech perception. To the author's knowledge, for speech production a similar investigation has not been done yet. It is known, that rhythmic motoric movements as walking or chewing are steered by α-, β- oscillations (α-oscillation in the frequency band of 8–12 Hz; β oscillation in the frequency band of 15–25 Hz) [10]. Speaking is one of the fastest rhythmic motoric actions of humans. Articulatory gestures are steered by Θ-oscillations (frequency band 15–25 Hz) and low ɤ-oscillations (frequency band 25–35 Hz). Mimicking speech perception, fig. 3 shows a cortical model simulating cortical PIN-Θ, and PIN-ɤ complexes reconstructing the Θ-, ɤ- oscillations entrained by the auditory signal of the critical band-channels [2].

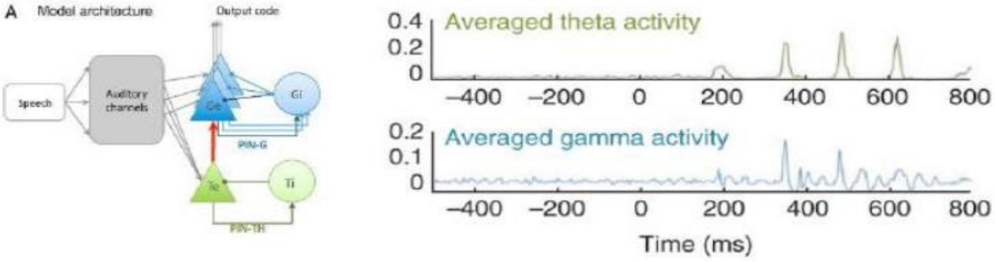

Figure 3 – model of generating Θ- and ɤ-oscillations entrained by auditory channels.

The ɤ-oscillations are generated by a PIN-ɤ complex connected to the output of a PIN-Θ complex realizing a cortical model for nested oscillations. As discussed in [1], the entrainment of the PIN-Θ complex is performed by the envelope of the auditory signal. Compared to the model shown in fig. 3, where all critical channels are fed into a neural network steering the PIN-Θ complex, the use of the envelope is a simplified model used in this paper also. The peaks of the envelope correlates to the instances, where the articulators indicate a switch from opening to closing the vocal tract. This instance is in the V*-segment and defines the center of each CS. Given the consonants cluster C*- C* located between two neighbored CSs with the structure C*V*C*- C*V*C*, a boundary must be defined, which determines, where the split between consonants related to the O-gesture of the first CS and the consonants relating to the C-gesture of the following CS must be done. This is a general problem for defining syllables. As defined in section 2.1 the split of the consonants within neighbored CSs is derived from the middle of a Θ-cycle (see fig.4, M-anchor point).

Nested within a Θ-cycle are a fixed number of ɤ-cycles, which are in phase to the Θ-cycle. The experiments described in chapter 3 indicate, that the number of nested ɤ cycles into a single Θ-cycle depends on the C*VC* structure of an CS. It seems, that for 'full-fletched' C*VC* four ɤ-cycles are nested as shown in fig. 3 and 4, and that for CSs with reduced C*V and VC* structures 2 or 3 ɤ-cycles are nested. But this issue is not explored yet.

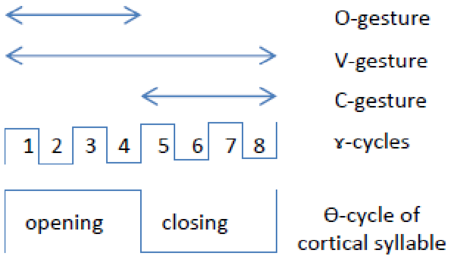

Figure 4 – 8 ɤ states are nested in a single Θ-cycle defining the range of starting points of the gestures

Whereas the relation of the Θ-oscillations to the rhythm of syllables is widely accepted, the relation of the ɤ-oscillations to phonetic units is not settled. In [1], [2] it is assumed, that the start of each ɤ-cycle is the start of an AF-defined gesture representing a phoneme. Yet this interpretation assumes that the number of ɤ-cycles depends on the number of phones realizing a C*V*C* structure. This approach contradicts the assumption, that the ɤ-cycles are nested with constant number of ɤ-cycles within a single Θ-cycle. A solution of this problem is proposed in [6], where the starting points of the ɤ-cycles are interpreted as the starting points of the OVC-gestures. In this concept the number of nested ɤ-cycles depends only on the number of OVC-gestures activated by a C*V*C* structure (e.g. missing O-gesture for VC-syllables).

2.3 The Articulatory and the Perceptive Method

This section describes two methods – the articulatory and the perceptive method - to extract the Θ-cycles from an articulatory speech database. Both methods work with the concept, that each CS is described by an open-closing cycle of the vocal tract. The starting point and the duration of an open-close cycle is retrieved and specify start and end of an Θ-cycle.

For detecting the open-close cycle, **the articulatory method** analyses the kinematics of the articulators recorded by electromagnetic articulography. Given the movements of the articulators, the starting points of articulatory gestures indicate the instances for opening and closing the vocal tract. The current implementation assumes, that the observation of the yaw is enough for detecting the starting point and the duration of the open-close cycle defining ***ideal*** CS. Especially the closing of the vocal tract is performed often by other articulators e.g. tongue tip and the lips. The analysis as performed in chapter 3, shows, that only 30% of the CSs are ideal. The current implementation discards non ideal CSs.

For detecting the open-close cycle **the perceptive method** uses the envelope of the auditory signal in a similar way as shown in fig. 3 mimicking cortical perception. Instead of using a model of the auditive signal, easy to implement algorithms as detectors of maxima/minima of the energy of the envelope of the speech samples itself are used. Due to the inertia of the jaw, the minima and maxima are delayed instances of the starting point of the cortical control of the jaw. Further 'noisy' minima and/or maxima could be detected within a CS. Thus, the minima and maxima determined with the perceptive method are rough estimates for the starting point and the duration of an open-close cycle. The following section describes, how the both methods are combined to estimate optimally the starting/end point of Θ-cycles.

2.4 The Procedures of the Analysis

The analysis starts with searching the speech segments containing V* complexes as annotated is the articulatory database. First the energy and duration of the V* is analyzed to check, if the V* is a candidate for the kernel of an CS. If duration is very short and energy is low, the candidate is discarded, as the notated V* could be a reduced syllable not related to an open-close cycle. The next analysis step determines the slopes of the movement of the jaw. If they are consistent with an opening-closing jaw-cycle, the CS is regarded as an ***ideal*** CS and analysis is continued. Next, for ideal CSs, the non-speech segments around the CS are analyzed determining the instances, where their Θ-cycle start and/or end[1]. These segments are either detected by the annotations given in the articulatory database, or by a non-speech-detector.

After this preparatory analysis, the perceptive method is started leading to a rudimentary position of the Θ-cycles. The final position of the Θ-cycle is found by the articulatory method. As shown in fig. 5, this task is equivalent to the extraction of the **M**iddle, **S**tart and **E**nd of a Θ-cycles, leading to the notation of M-, S-, and E-anchor-points. If no non-speech segment is located between two neighbored CSs, the E-anchor point of an CS is the same as to the S-anchor point of the following CS.

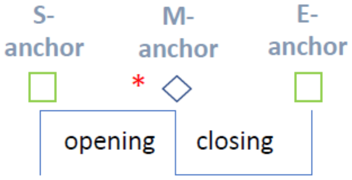

Figure 5 - Anchor-points describing the starting, middle and end of a Θ-cycle. The * marks the instance of maximum energy of the V* complex.

As shown in fig. 5 the instance of maximal energy of V* should be delayed to the instance of the M-anchor point. This delay is caused by the inertia of the jaw, where the cortical control, initiated by the M-anchor point of the Θ-cycle, is earlier than the maximal opening of the vocal tract delivering maximal energy. For ideal CSs, the analysis is continued by followings procedural steps (results are annotated by the example shown in fig. 6):

1. For each ideal CS the instance of maximal energy of the V* is evaluated (marked by * in fig. 6). Left to this instant (i.e. at earlier time instances), instances of minima of the speed of the jaw are searched and marked as M-anchor point (marked by a ◊ in fig. 6).
2. Between two vowels, minima of signal energy are searched and correlated with left-neighbored maxima of the speed of the jaw leading to the S-E-anchor points (□-mark).
3. Between two neighbored M-anchor points 4 ɤ- cycles (8 ɤ-states) are integrated and adjusted (entrained) at the M-anchor points.
4. The vowels and consonants between two neighbored vowels are aligned to the ɤ-states using the starting points of the phones as annotated in the speech database. Depending on the alignment, sequences of phones are categorized as O-, V-, or C-gestures.

[1] In the STG [12] neurons detecting *edge features* are found, which detect instances, where speech activity or change in AFs start. Input to that neurons are the auditory signals provided by the critical bands [16]. These neurons are event-driven, whereas the neurons transforming the auditory signal to the articulatory code are driven by the Θ- and ɤ- oscillations.

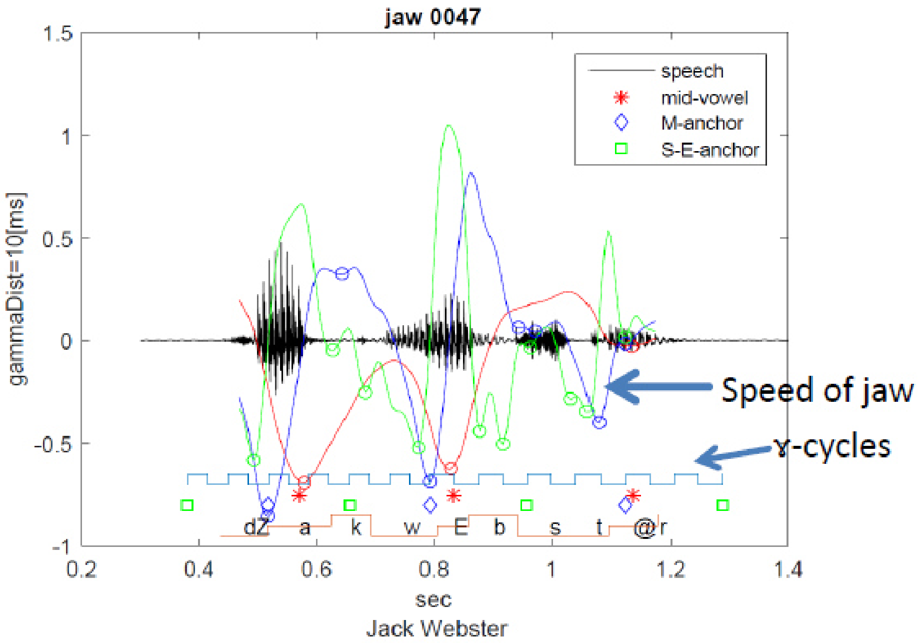

Figure 6 - Speech signal and jaw trajectories (minima indicated by 'o') of the utterance 'Jack Webster' aligned to the ɤ-cycles, the anchor-points and the position of maximal energy of the vowels /a/, /E/, /@/; at the bottom the duration of the OVC-gestures together with the composing phonemes as annotated. The length of lowest (highest) line denotes the duration and the position of the O- (C-) gesture.

3 Experimental Results

3.1 The Articulatory mngu0 Corpus

From a professional British speaker, 1300 phonetically diverse utterances (read speech) were recorded together with a Carstens AG500 electromagnetic articulography (EMA) [13]. The EMA data are delivered from six midsagittal coils positioned at the upper lip, lower lip, lower incisor, tongue tip, tongue body, and tongue dorsum, and from two reference coils for correcting head movements. The processed EMA data are sampled at 200Hz. Further, the corpus provides the velocity and acceleration of the coils. The audio samples are down-sampled to 16 kHz and are labeled automatically Labelling is performed by forced alignment [14] using the Combilex lexicon with its notation of the phone labels [15].

3.2 Quality Measure

The quality measure developed concerns the correct determination of the starting points of the ɤ- cycles. The processing step 2, as described in section 2.4, is the most critical. As soon as the instances of the S/E-anchor points are estimated incorrectly, the nesting of the ɤ-cycles fails leading to wrong alignments of the phones, leading to a wrong set of OVC-gestures together with a wrong alignment. To detect these errors reliably, a reference given by cortical activities would be the best procedure. But those measurements are not available.

In this paper a quality measure is developed based on the assumption, that neighbored OVC-gestures are statistic independent [16]. This hypothesis is derived from psychoacoustic measurements made by Fletcher investigating the intelligibility of consonants and vowels embedded in nonsense CV, VC and CVC syllables. Given the properties of the error rates of the phonemes building these syllables, it is concluded [16], that the AFs of the OVC-gestures producing the nonsense syllables are statistic independent. Further it is concluded, that statistic independence is equivalent to a motor control mechanism, where neighbored gestures are steered independently. From this property it is assumed, that each instance of a given OVC-gesture has the same starting point controlled by the same ɤ-state (e.g. as shown in tab. 3, for correct alignments, all instances of O-gestures /t/, should be aligned to ɤ-state number 1). In this paper it is assumed, that the intelligibility results of Fletcher could be extended to more complex C*V*C* structures, leading to the hypothesis that all neighbored OVC-gestures are statistic independent. The percentage of correct state-alignment is used as measure for **A**lignment **Q**uality (**AQ**-quality). This measure assumes that for a given OVC-gesture, the correct aligned ɤ-state is that state, to which most of the instances of an OVC-gesture are aligned. Taking the ideal CSs, the current implementation leads to an AQ of about 60%.

3.3 The OVC Gestures and Their Timing

From the training part of the mng0 corpus, 14 814 syllables have been analyzed yielding 4292 ideal CS syllables (29%). As shown in tab.1 in total 502 OVC-gestures have been found with an average AQ of 57% (the AQ is evaluated by weighting with the number of hits; see tab. 4).

O-gestures-AQ [%]	V-gestures-AQ [%]	C-gestures – AQ [%]
277 - 62	37 - 43	188 - 54

Table 1 - number of O-, V-, C-gestures and the related values of the AQ

The first 3 lines of tab. 2 show OVC-gesture gestures, which occurred most in the ideal CSs. The next two lines are instances of gestures composed of more than one phoneme.

O-gestures		V-gesture		C-gestures	
phone	hits	phone	hits	phone	hits
/t/	215	/@/	726	/n/	317
/r/	147	/I/	638	/s/	272
/D/	136	/i/	535	/t/	259
/p r/	19	/i @/	9	/n s/	25
/t r/	16	/aU @/	4	/s t/	25

Table 2 - selected OVC-gestures with their number of occurrences (hits) analyzed in the database.

For some selected OVC-gestures, tab. 3 shows the number of alignment of instances (*hits*) to one of the 8 ɤ-states of a Θ-cycle. For ideal alignments all hits should be concentrated on a single ɤ-state (AQ=100%).

ɤ-state	1	2	3	4	5	6	7	8
O: /t/	107	79	26	3	0	0	0	0
O: /r/	42	38	41	26	0	0	0	0
V: /@/	22	9	26	98	366	171	27	7
V: /I/	51	8	42	162	307	59	6	3
C: /n/	0	0	0	0	5	73	131	108
C: /s/	0	0	0	0	2	81	126	63

Table 3 - hits of OVC-gestures aligned to the ɤ-states 1 – 8; gesture O: /t/ has a of: AQ=107/215=50%

Most errors of wrong alignments may be caused by the complex behavior of the articulators leading to wrong anchor points and from the assumption, that the number of ɤ-states within a Θ-cycle is fixed to 8. Further it must be considered, that the labeling of the phones is done automatically leading to alignment errors in the range of 10ms.

4 Conclusion

The paper proposes a new noninvasive methodology working with an articulatory speech database for extracting articulatory gestures and related Θ- and ɤ-cycles. The methodology proposed has been implemented for ideal cortical syllables (CS) covering 29% of all CS of the articulatory database investigated. The alignment quality AQ, evaluating the consistency of aligning of ɤ-cycles to instances of OVC-gestures, predicts that about 60% of the ɤ-cycles have been determined correctly. Both measures, the syllable coverage and the value of the AQ, indicate much room for improvements. Future topics would be the interaction of the movements of **all** articulators to model also non-ideal CS, the extraction of more precise edge features detecting more reliable the starting and ending of the Θ-cycles, and the adaption of the number of nested ɤ-cycles on the C*V*C* structure of the CSs. A high value of the QA of near 100% for all CS would also support the correctness of the hypotheses used.

5 References

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

[2] Hyafil, A., Fontolan, L., Kabdebon, C., Gutkin, B., and Giraud, A.: *Speech encoding by coupled cortical theta and gamma oscillations.* In *eLife*, DOI: 10.7554/eLife06213, 2015.

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

[4] Bouchard, K. E., and Chang, E. F.: *Control of Spoken Vowel Acoustics and the Influence of Phonetic Context in Human Speech Sensorimotor Cortex.* In *The Journal of Neuroscience*, 34(38): pp. 12662–12677, September 17, 2014.

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

[6] Höge, H.: *The Articulatory Code and Related OVC-Gestures.* In *Proc. ITG*, 2018.

[7] Höge, H.: *Using Elementary Articulatory Gestures as Phonetic Units for Speech Recognition.* In *Proc. Konferenz Elektronische Sprachsignalverarbeitung (ESSV)*, 2018.

[8] MacNeilage, P. F.: *The frame/content theory of evolution of speech production.* In *Behavioral and Brain Sciences* 21, S. 499–511. 1998.

[9] Kayser, C., Ince,R., and Panzeri, S.: *Analysis of Slow (Theta) Oscillations as a Potential Temporal Reference Frame for Information Coding in Sensory Cortices.* In *PLOS Computational Biology*, Vol.8, I.10, e1002717. October. 2012.

[10] Brinkman, L., Stolk, A., Dijkerman, H. C., de Lange, F.P., and Toni, I.: *Distinct Roles for Alpha- and Beta-Band Oscillations during Mental Simulation of Goal-Directed Actions.* In *The Journal of Neuroscience*,34(44), pp.14783–14792, October 29. 2014.

[11] Ladefoged, P., and Johnson, K.: *A Course in Phonetics.* Wadsworth Cengage Learning, 7th Edition, Boston, 2015.

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

[13] Richmond, K., Hoole, P., and King, S.: *Announcing the Electromagnetic Articulography (Day 1) Subset of the mngu0 Articulatory Corpus.* in Interspeech, pp. 1505-1508, 2011.

[14] Clark, R. R., Richmond, K. and King, S.: Multisyn: *open domain unit selection for the Festvial speech synthesis system.* In *Speech Communication,* Vol.49, no.4, pp. 317-330, 2007.

[15] Fitt, S., Richmond, K., and Clark, R.: *The Combilex lexicon.* www.cstr.ed.ac.uk/research/projects/combilex

[16] Höge, H.: *On the Nature of the Features Generated in the Human Auditory Pathway for Phone Recognition.* In *Interspeech Dresden*, 2015.