USING ELEMENTARY ARTICULATORY GESTURES AS PHONETIC UNITS FOR SPEECH RECOGNITION

Harald Höge

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

Abstract: Perception and production of speech is linked to a set of basic articulatory gestures related to an articulatory code. Due to the immature methods in measuring cortical activities, the detailed functionalities of these gestures have not been deciphered yet. I hypothesize a set of gestures – *elementary articulatory gestures* (EAGs) – mimicking human articulatory gestures. The concept of EAGs is based on the hypothesis that the gestures are generated by a two level control hierarchy. The upper level is related to '*broad gestures*', the lower to '*narrow gestures*'. The control of broad gestures is used for both, for speech perception and speech production, whereas the concept of narrow gestures is only relevant for speech production. The temporal control of broad gestures is triggered by the quasi rhythmic opening and closing of the mandibular [6] steered by entrained Θ-oscillations [7]. I call one Θ-cycle a '*cortical syllable*', which is defined by 3 states: an opening, middle and a closing state. Each state is related to a set of EAGs, where ɤ-oscillations embedded in the Θ-oscillations steer their temporal dynamics. In speech perception, the EAGs are perceived using the temporal constraints given by the upper level of control hierarchy. A feasibility study is presented, where from a phonetic labeled German speech database a set of 249 opening, 40 middle and 233 closing EAGs are extracted. Using a model mimicking human perception an average classification EAG error rate of 48.6% is achieved.

1 Introduction

The use of articulatory gestures as phonetic units for speech recognition is not new [1]. Due to the principles of evolution [8], theories in phonetics [2] and neuroscience [4] support the hypothesis that perception and production of speech is closely related. The core of the paper is based on the hypothesis, that a specific part of the control mechanism steering the gestures have double use: in speech production this parts determines temporal behavior of gestures; in perception this part partitions the auditory signal into chunks of gestures. The paper concentrates on the neuronal mechanisms active in production of gestures, because in the cortex these mechanisms are explored more than the related mechanisms in perceiving gestures. Given the nature of the gestures observed in production, the mechanisms for perception are derived. The theory [2] – the origin of many gesture theories - claims, that speech is produced by *gestures,* which are specific for each articulator and which are defined by manner & place features. In the ventral sensorimotor cortex (vSMC), somatotopically ordered populations of neurons have been found [4], which control the gestures of the articulators individually as predicted by [2]. In the following I call these articulator specific gestures '*narrow gestures*'. By observing the dynamics of narrow gestures it became clear, that they are embedded in a more complex control of combined actions as described by the articulatory score [3]. The existence of a complex neuronal network in the vSMC coordinating the control of narrow gestures is hypothesized also in [4]. Recent findings in neuroscience [7] claim, that the dynamics of combined narrow gestures are steered by *ɤ*-oscillations*,* which are embedded in entrained *Θ*-oscillations following the quasi rhythmic production of syllables. Further in speech perception the *ɤ*-oscillations are basis to segment the auditory signal into phonetic units. In the following I call gestures, which are controlled by the articulator rhythm (action of Θ- and *ɤ*- oscillations), which are a combination of narrow gestures, and which are related to syllables '*broad gestures*'.

In this paper I propose a specific set of broad gestures called ***E**lementary **A**rticulatory **G**estures* (***EAGs***) as defined in chapter 2. Further this chapter focuses on the plausibility that the EAGs fit to the cortex's activities observed. Chapter 3 presents a feasibility study to use EAGs for ASR. Finally in chapter 4 a cortical control architecture generating EAGs is proposed, which highlights the potential, how the EAGs may improve the performance of ASR-technology with respect to recognition accuracy and reduction of energy consumption.

2 Evaluating the Concept of EAGs

The concept of EAGs is described by following four hypotheses (H1-H4):
H1-articulatory code: *Human communication is based on the articulatory code defined by manner and place features structured syllabically. In speech production sequences of articulatory codes are transformed to gestures. In speech perception the auditory signal is partitioned into chunks of articulatory gestures, from which the articulatory code is reconstructed.*
H2-articulatory rhythm: *Perception and production of speech is controlled by entrained ϴ- and ɤ-oscillations called the articulatory rhythm. In communication this rhythm is the carrier of the information –the articulatory code - and is synchronized between speaker and listener. In speech production, the articulatory rhythm steers the dynamics of the movements of articulators. In speech perception, the articulatory rhythm is reconstructed from the auditory signal and segments the stream of auditory features into chunks of articulatory gestures.*
H3-control of EAGs: *The EAGs are produced by a 2 level hierarchy, where the upper level transforms the articulatory code into EAGs and the lower level transforms the EAGs to **narrow gestures**. The EAGs are linked to the concept of cortical syllables defined by the quasi rhythmic opening and closing gesture of the mandibular. Each cortical syllable is defined by three states: the **o**pening, **m**iddle and **c**losing state. Each state is related to a set of EAGs denoted as **o**EAGs, **m**EAGs and **c**EAGs. Each cortical syllable is represented by a sequence {**o**EAG, **m**EAG, **c**EAG}, where positions within a sequence may be empty (e.g. CV syllables).*
H4-Context of EAGs: *The control of an EAG relating to a given state is independent from the control of EAGs from the following states. In speech perception the features, which are extracted from the auditory signal from an EAG-segment, are statistic independent from those extracted from neighbored EAG-segments.*

The scope of this paper does not allow evaluating the concept of the EAGs against the tremendous amount of papers characterizing state of the art. Instead, four key papers [2, 6, 4, 7] are selected. The papers [2, 6, 4] concentrate on the speech production; the paper [7] on the perception. The knowledge gained from measurements in the cortex is very limited due the immature status of 'brain-machine interfaces (BMIs)' currently used[1]. For compensating this lack of knowledge, other resources of knowledge as findings from evolution and psycho-acoustic measurements are used additionally.
I start with the concept of narrow gestures, as defined in H3. The origin of the concept of narrow gestures is based on theory of '*gesture*' [2, 1989]. As described in the introduction this theory is supported by cortical measurements by Bouchard et. al. [4, 2013]. Fig. 1 shows the

[1] In order to decipher the functionality of the cortex, the activity of about 100 000 neurons performing a specific task in a nucleus has to be observed simultaneously together with the neuronal activity of the connections between the nuclei. The BMIs available are based on invasive and non-invasive measurement methods. The BMIs based on non-invasive methods as fMRT, PET, EEG have a resolution of about 2 mm^2 per pixel measuring the averaged activity of neuronal population in the range of about 200000 neurons. The invasive methods using high density electrocorticography grids (ECoG) allow simultaneous observation of the activity of a small amount of neurons in the order of 100 neurons.

activity of single neurons controlling place features. The organization of the control is discussed in chapter 4 further.

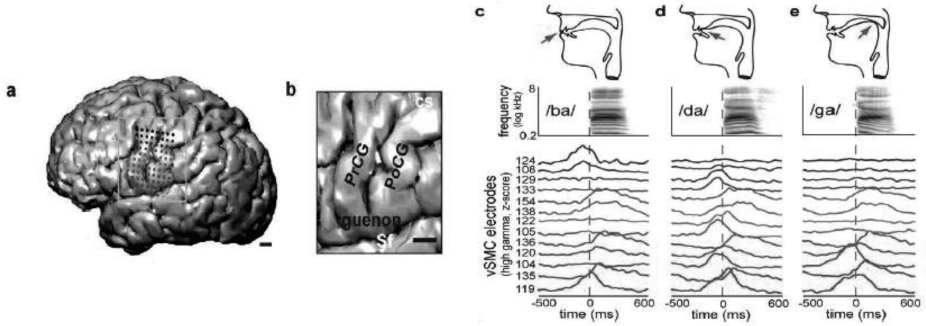

Figure 1 – ECoG measurements in the ventral sensorimotor cortex (vSMC) **a:** MRI reconstruction of single subject brain with electrodes (dots); about 30 electrodes were connected to neurons delivering useful information **b:** Expanded view of vSMC: pre- and post-central gyri (PrCG and PoCG), central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

Compared to the concept of narrow gestures the concept of broad gestures is very ‘fuzzy’. This concept has been presented in its rudimentary form by MacNeilage speculative paper [6] based on an evolutionary background. He treats the quasi rhythmic opening and closing gesture of the mandibular as a **F**rame with **S**lots to be filled (F/S-theory). The frame is the carrier of information to be transmitted and received. The slots are clusters of phonemes described by broad 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*. Thus this concept comes close to the concept of EAGs. The concept of broad gestures is also confirmed by [4]: *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.* With respect to evolution, [4] notes: *However, we found an **additional** laryngeal representation located at the dorsal-most end of vSMC. This dorsal laryngeal representation appears to be absent in **non-human primates**, suggesting a unique feature of vSMC for the specialized control of speech.* As the larynx is linked to the manner features this finding hint to the structure of the articulatory code (see [4], Figure 4: Phonetic Organization of Spatial Patterns). Hints for mandibular state dependent broad gestures are given by the statement: *dynamics of individual phonemes were superimposed on a slower oscillation characterizing the transition between consonants and vowels, which occupied distinct regions of the cortical state-space. Although trajectories could originate or terminate in different regions, they consistently passed through the same (target) region of the state-space for shared phonetic features. Large state-space distances between consonant and vowel representations may explain why it is more common to substitute consonants with one another, and same for vowels, but very rarely across categories in speech errors.*

Further this last statement hints to the articulatory rhythm (H2) which is focus of Giraud’s and Poeppel’s paper [7]. The paper states: *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* (see Fig. 2).

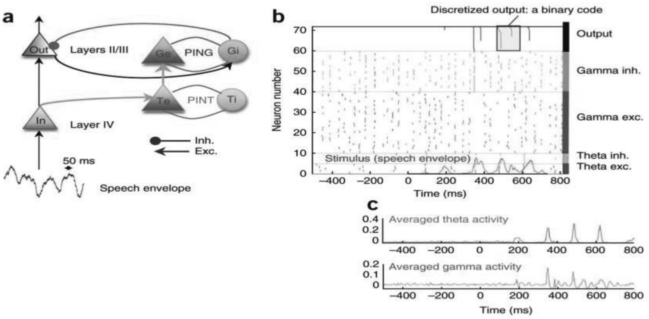

Fig. 2 – a: generation of the oscillations; b: neuronal activities; c: ɤ- oscillations nested in Θ-oscillations

The assumption that each ɤ-cycle is related to a single phoneme and that four ɤ-cycles are embedded in a single Θ-cycle would implicate that each syllable is constructed by 4 phonemes. As will be discussed in chapter 4, the ɤ-cycles fit better to EAGs, as the number of EAGs within a syllable is fixed. A further unexplored field is the independence properties of the EAGs as stated in H4. This issue is discussed in chapter 4.

3 Classification Experiments – a Feasibility Study

This feasibility study is a first step to work with EAGs in ASR. The system used to classify EAGs is a modification of the system described in [11], which classifies phonemes with manner & place features extracted from the auditory signal. For this feasibility study an approximation of the unknown EAGs has to be found. I use a syllabic-phonemic approach, where **o**EAGs and **c**EAGs are approximated by the consonant clusters before and after the central vowel-cluster – the **c**EAGs. The EAGs are extracted from a speech database labeled in syllables and phonemes (Kiel Corpus; the database contains 3868 fluently spoken phonetic balanced German sentences; the set of phonemes used are shown in tab. 2 (appendix). Ideally the syllable should approximate the cortical syllable, where the Θ-oscillations have to be extracted from the auditory signal. Such an approach is presented in [9], where simulated neuronal PIN-loops generate Θ-oscillations with nested ɤ- oscillations according to fig. 2a. I use a simpler approach, where a script [10] is applied on the database, which mimics the mandibular oscillation by the analysis of increasing and decreasing energy of the speech signal. From the training part of the database the resulting *pseudo cortical syllables* delivered 76832 ‘*pseudo-EAGs*’ leading to three sets of pseudo-EAGs tabulated in tab.2 (appendix). Given the auditory signal the segmented pseudo-EAGs are classified with the system [11], where the phonemic approach is adapted to an EAG-approach: the analysis window covering a segmented phoneme is exchanged by an analysis window covering a segmented EAG, and for each set of EAGs a different LDA and different GMMs is generated. For each critical band the probabilities P(EAG|critical band) are determined. Assuming, that these probabilities are statistic independent with respect to the different bands, the product of these probabilities gives the probability P(EAG). Applying Bayes rule as described in [11], following '*all band EAG-error rates (EAG-ER)*' have been achieved on the test part of the database:

	# classes	# of modes	EAG-ER
oEAG	142	568	43,7
mEAG	35	140	64,6
cEAG	117	468	37,6
		average	48,6

Table 1 – all band EAG error rates (EAG-ER) achieved for the three sets (pseudo-**o**EAGs, -**m**EAGs and -**c**EAGs) with equal a priori probabilities of the set specific pseudo-EAGs.

4 Discussion and Conclusion

In this chapter the use of EAGs in ASR is discussed according to three aspects: first, their feasibility to be trained; second, their potential to increase recognition rate and third, their potential to decrease power consumption.

First - feasibility: The pseudo-EAGs regarded in chapter 3 are a crude approximation of the 'real' EAGs as assumed to be implemented in the cortex. The approximation depends on the set of phonemes (tab.2) and depends on the kind of pseudo cortical syllable used. As the resulting amount of different pseudo-EAGs is smaller than 300 (see table 3), it is feasible to train acoustical models for the EAGs with moderate amount of speech. I assume that the sizes of the sets of 'real' EAGs are in the same range. It should be noted, that the set of 'real' EAGs as implemented in the cortex is speaker (articulator) dependent, as the sensitivity of neurons (see fig.1c) realizing the control of EAGs are adapted by learning. For ASR standard EAGs have to be defined in equivalence to the IPA standard defining phonemes. An acoustic model of the EAGs has to model their acoustic variations.

Second – potential of higher accuracy: higher accuracy can be achieved, if inherent constraints are incorporated in the acoustic model. A strong constraint of EAGs is their state dependency within a cortical syllable (chapter 2). Assuming that the segmentation given by the ϴ- and ɤ- oscillations is error free, for each state of the cortical syllable the related set of EAGs is known restricting the number of EAGs to be classified simultaneously. State of the art acoustic models use context dependent phones, which incorporate partly these restrictions. As seen in tab. 3, there exist pseudo-EAGs, which are constructed with the same sequence of phonemes, but belong to different sets. Thus EAGs have the potential to lead to sharper acoustic models.

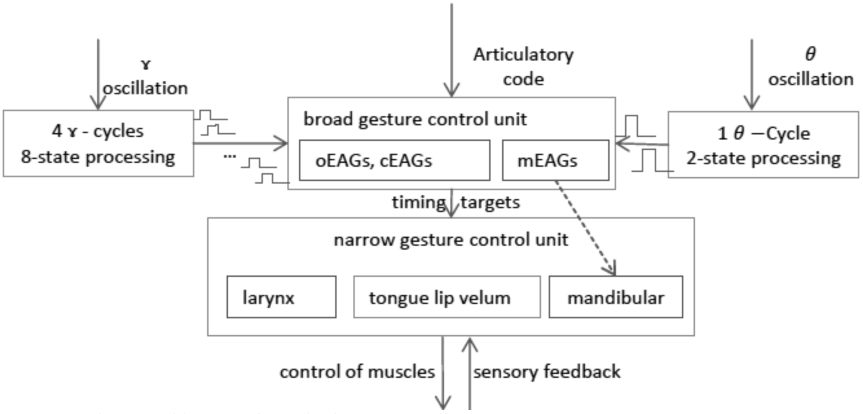

Figure 3 – 2-layer architecture for articulatory control

Additional constraints of an acoustic model are given by the kind of acoustic and temporal features used. In the cortex, manner & place features are extracted from the auditory signal. As the EAGs are composed by narrow gestures defined by manner & place features, the related '*EAG-features'* seem to be clusters of manner & place features. Further specific neurons may exist, which are sensitive a specific clusters of narrow gestures. In the feasibility study no EAG-features are used but a modulation feature vector constructed from the auditory signal (see [11]). Thus the transformation of the modulation feature vector to the EAG-features is still an open issue. Due to the articulatory rhythm, strong temporal constraints are active. Such temporal constraints have not yet been used in state of the art acoustic models. As shown in Fig. 3, I propose a hierarchical architecture with two layers, the upper control the production of broad gestures, and the lower producing narrow gestures. The control is organized by states generated from the Θ-cycles and the ɤ-cycle, where each cycle generates 2 states [5]. As four ɤ-cycles are within a single *Θ*-cycle, 8 states of the ɤ-cycles and 2 states of the Θ-cycles are available to steer the control for producing narrow gestures and EAGs. This control mechanism has still to be explored to model the inherent constraints in the acoustic model. Taking into account evolution, it can be speculated, that the control of mandibular gestures is performed with the 'old' Θ-cycles. Later in evolution the mandibular gestures are enhanced with manner features leading to **m**EAGs. Thus the **m**EAGs may have a separate extended 'old' control.The last step in evolution was the development of faster gestures given by the flexibility by the tongue and the lips and the development of the speech specific larynx [4] leading to the control of **o**EAGs and **c**EAGs using ɤ-cycles. How this control works in detail is still an open issue. I assume that the starting and ending points of narrow gestures is controled by the Θ- and ɤ-cycles, whereas the dynamics of the narrow gestures is realized articulator specific by intrinsic delays. The Θ- and ɤ-cycles can be used to design optimal windows for feature extraction. I assume, that the intrinsic delays are adapted to the speaker dependent properties of the articularors and thus contain less constraints.

Third-power consumption: As hypothesized in H4 the features extracted from an EAG-segment are statistic independent. This hypothesis is presented in [12] derived from error rates [13] measured for consonants and vowels embedded in nonsense syllables with CV and CVC structure. I assume that the findings in [13] can be extended to EAGs, i.e. the error rate of {**o**EAG, **m**EAG, **c**EAG} syllables is given as the product of the error rates of the sets of **o**EAG, **m**EAG, and **c**EAG. As no context has to be modeled across EAGs, the search space needed for ASR systems based on EAGs is extremely small saving computer power. Further a bottom up architecture can be realized with the articulatory code as interface.

Conclusion: A first step for using EAGs in ASR and the potential to achieve improved performance has been presented. But it is a long way to develop a system competitive with state of the art ASR systems: robust algorithms for extracting the articulatory rhythm, suited EAG-features and acoustic models integrating the temporal constraints given by articulatory control have to be found.

5 Appendix

Phoneme	Place (coarse)	Place (detailed)	Phoneme	Place (coarse)	Place (detailed)
d	alveolar		a:6	front	open front to central
h	glottal		U	back	near-close near-back
i:	front	close front	u:	back	close back
z	alveolar		r	alveolar	
O	mid	open-mid back	C	palatal	
n	alveolar		i:6	front	close front to central
@	central		j	palatal	
l	alveolar		Y	front	near-close near-front

a	front	open front	u:6	back	close back to central
x	velar		O6	back	open-mid back to central
t	alveolar		e:6	front	close-mid front to central
Q	glottal		p	bilabial	
m	labial		ASpause		
b	bilabial		2:	front	close-mid front
aU	front	open front to near-close near-back	9	front	open-mid front
I	front	near-close near-front	E6	front	open-mid front to central
s	alveolar		U6	back	close back to central
v	labiodental		o:6	back	close-mid back to central
k	velar		I6	front	near-close near front to central
N	velar		OY	back	open-mid back to near-close near-front
y:	front	close front	a6	front	open front to central
6	central		2:6	front	close-mid front to central
f	labiodental		y:6	front	close front to central
E	front	open-mid front	E:6	front	open-mid front to central
e:	front	close mid front	*		
aI	front	open front to near-close front	AShesit		
g	velar		E:	front	open-mid front
S	postalveolar		96		open-mid front to central
Y6	front	near-close front to central	Z	postalveolar	
o:	back	open-mid back	a~	front	open front nasalized
a:	front	open	=6	central	central syllabic

Table 2 – phonetic symbols used for defining narrow gestures; vowels have a course (place (coarse)) and a detailed (place (detailed)) description

Nr.	opening	#	vowel	#	closing	#
1	/h/	4625	/@/	5385	/n/	6442
2	/d-h/	3704	/I/	5077	/s/	2141
3	/n/	3416	/a/	4303	/t/	2141
4	/v/	2802	/a:/	2758	/C/	1778
5	/m/	2709	/aI/	2594	/m/	1482
6	/Q/	2603	/6/	2374	/x/	1378
7	/f/	2566	/U/	2373	/k/	1124
8	/z/	2102	/i:/	2267	/l/	1059
9	/b-h/	1644	/E/	2067	/t-h/	881
10	/g-h/	1579	/e:/	1900	/N/	755
11	/t-s/	1333	/O/	1535	/k-h/	624
12	/l/	1281	/u:/	1157	/n-t/	602
13	/r/	1196	/aU/	1113	/C-t/	474
14	/k-h/	996	/e:6/	1112	/f/	468
15	/b/	511	/o:/	871	/n-t-h/	301
16	/t-n/	460	/U6/	747	/s-t/	292
17	/j/	459	/OY/	604	/p/	284
18	/f-r/	443	/Y/	550	/S/	279
19	/g-N/	422	/9/	543	/l-n/	252

20	/s/	398	/a:6/	542	/C-t-h/	235
21	/S/	377	/E6/	539	/l-t/	235
22	/S-t-h/	359	/i:6/	530	/t-s/	210
23	/t-h/	348	/O6/	520	/x-t-h/	209
24	/d/	346	/y:/	484	/n-s/	185
25	/g/	338	/2:/	304	/N-k/	160
26	/d-n/	291	/o:6/	302	/s-t-h/	158
27	/b-m/	249	/a6/	293	/l-t-h/	142
28	/C/	227	/y:6/	226	/r/	118
29	/b-r/	202	/Y6/	151	/g-N-s/	115
30	/x/	185	/u:6/	116	/x-t/	107
31	/S-p-h/	166	/E:6/	102	/p-h/	105
32	/b-h-r/	165	/I6/	100	/S-t/	97
33	/S-n/	163	/2:6/	84	/k-s/	97
34	/g-l/	147	/E:/	83	/l-m/	94
35	/s-n/	142	/96/	20	/l-s/	90

Table 3 – the sign '-' is used to separate narrow gestures within an pseudo-EAG. From the training part of the database 76832 EAGs are extracted. The 35 most frequent observed pseudo-EAGs are tabulated. The total number of different pseudo-EAGs per set is: {oEAG}=249; {mEAG}=40; {cEAG}=233

6 References

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

[2] C.P. Browman, and L. Goldstein: *Articulatory Gestures as Phonological Units*. In *Haskins Laboratories Status Report on Speech Research*, 99, pp. 69–101. 1989.

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

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

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

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

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

[8] C. Darwin: *The descent of man*. In *Great Books*, Encyclopedia Britannica, 1871.

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

[10] https://www.phonetik.uni-muenchen.de/~reichelu/publications/ReichelIS2012.pdf

[11] H. Höge: *Modeling of Phone Features for Phoneme Perception*. In *ITG*, Leipzig, 2016.

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

[13] H. Fletcher, and R.H. Galt: *The perception of Speech and Its Relation to Telephony.* In *The Journal of the Acoustic Society of America*, Vol. 22, number 2, pp. 89-151. 1950.