IMPROVED FEATURES DRIVING AN Θ-OSCILLATOR FOR CORTICAL SEGMENTATION OF SPEECH INTO SYLLABLES

Harald Höge

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

Abstract: The paper describes a model for cortical segmentation of the auditory signal into syllables. Segmentation is based on a θ-oscillator realized by an inter-neuronal network gamma (PING) structure, where the position and duration of each syllable is given by a related θ-cycle. The paper is focused on improving features, which drive the θ-oscillator. We hypothesize that the θ-oscillator is driven by 'V-edge-neurons'. These neurons have been observed in the superior temporal gyrus (STG) which spike at the maximal rise of the envelope of the auditory signal at the onset of the nucleus (vowel) of a syllable. The paper is focused to model the V-edge-neurons. We hypothesize, that the V-edge-neurons have as input two kinds of CB-features [13] processed in critical bands (CB). The first kind are edge features derived from the instances of maximal increase of the partial loudness curve from each CB. The second kind are sustained features derived from CB-modulation features indicating the presents of vowel-onsets. The developed θ-oscillator is evaluated using a labeled speech database. The evaluation is based on the correctness of the match between the position of the syllables and the θ-cycles given by the sequences of θ-spikes emitted by the θ-oscillator. Compared to [2,3] considerable progress in correctness has been achieved from 80% to 90%.

1 Introduction

There is high evidence that speech perception is the transformation of the continuous streams of signals delivered from audio-visual sensors (ears and eyes) to a multi-item neural code of syllables constructed by subcodes describing articulatory gestures of the onset, the syllabic kernel and the coda of a syllable [5,25]. In the following we assume that the syllabic kernel is a vowel or a complex of vowels, as in many languages. Thus, the parts of the syllables are **o**nset, **v**owel(s) and **c**oda lead to the concept OVCs [24,26]. In this framework, the task of perception is to segment and to classify the OVCs. This process is far to be understood.

The paper is focused on mimicking the cortical segmentation of the continuous stream of the audio-signal into syllables using θ-oscillations [5,4]. This cortical process is quite different to the segmentation performed nowadays in automatic speech recognition (ASR), where segmentation is solved by a search algorithm including a language model [15]. In the starting time of ASR, in the 1970th, it was aimed to mimic the human approach in segmenting speech into phonetic units This approach was implemented by 'knowledge-based rules' derived from inspection of the short-term spectra of speech [1]. Yet this approach failed, and still nowadays, no competitive algorithm mimicking the human approach has been found. Nevertheless, it is useful to investigate the human approach as it implements a bottom-up driven interface delivering syllables. This interface separates acoustic from symbolic processing, allows to integrated easier visual input (lip reading) and consumes less computing power.

In the last 20 years, neuroscience increased substantially the knowledge in cortical processing of speech, achieved mainly by measuring the activity of neurons with cortical electrocorticography (ECoG) in clinical settings [16]. With this technology, currently the local field potentials (LFP) of about 200-500 neurons located at the surface of the brain can be measured simultaneously. With this technology the LFPs of neurons located in deeper layers (lamina) of the brain are measured imperfectly (see fig.1 in [16]). To overcome the problem of missing cortical knowledge, hypotheses are postulated, whose evidence are checked by correlations between events of the speech signal and the output of measured neurons, by psycho-acoustic measurements, and by measurements of the activity of mammal neurons located in deep layers of areas, which have the same functionality as human layers.

Publication of papers mimicking the human approach in segmenting speech are scarce. We follow the approach presented in [4] applied in [2,3], where the segmentation is performed by an θ-oscillator driven by features derived from the auditory signal. This paper is focused in improving the models of these features, which include recent cortical measurements and which increase the correctness in segmentation. The paper is organized as follows: In section 2 the architectures and the functionality of the modules constituting the θ-oscillator is given. Section 3 – the core of the paper - describes the features driving the θ-oscillator. Section 4 is devoted to the implementation and evaluation of the proposed θ-oscillator.

2 The θ-Oscillator

There is increasing evidence that the segmentation of the auditory signal into syllables is performed by θ-oscillations observed in the STG [5]. The frequency and phases of the θ-oscillation are event driven depending on the rhythm of the syllables. The position and duration of each syllable is related to the phase and the duration of a single θ-cycle. As discussed later the 'cortical' syllable is different to the phonetical syllable, as the cortical syllable starts with the onset of the vowel(s). We follow the approach [4] that the θ-oscillator is generated by an inter-neuronal network gamma (PING), as found in the thalamus, driven by specific features derived from the auditory signal.

Figure 1: Architecture of the θ-Oscillator: The speech signal is transformed to the auditory signal processed in critical bands (CB). For each CB, edge and sustained features are the input of a V-edge neuron, which outputs spikes driving the PING.

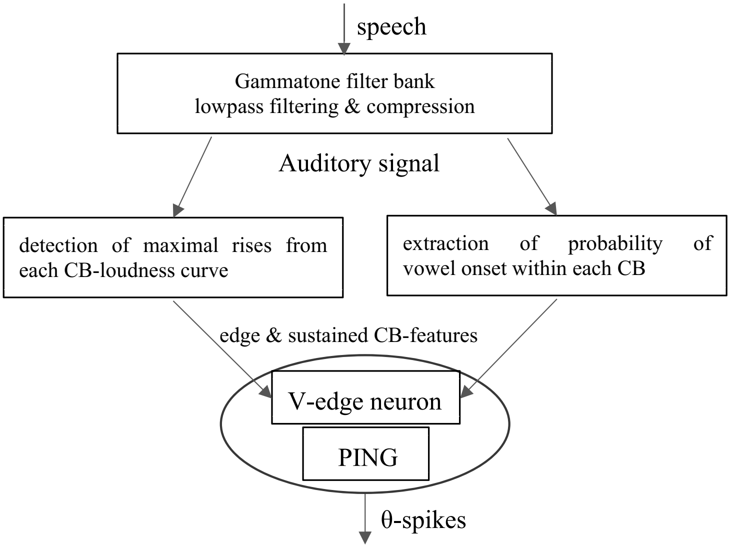

speech
Gammatone filter bank
lowpass filtering & compression
Auditory signal
detection of maximal rises from each CB-loudness curve
extraction of probability of vowel onset within each CB
edge & sustained CB-features
V-edge neuron
PING
θ-spikes

Fig.1. shows the architecture of the proposed θ-oscillator. The output of the θ-oscillator are spikes called θ-spikes spiking at instances t_{PING}. Thus, the θ-oscillator delivers no oscillations, but spikes determining the position and duration of a syllable. For a perfect θ-oscillator, two adjacent instances of t_{PING} define a θ-cycle related to the position and the duration of a syllable. The core of the architecture is the 'V-edge neuron' which models the functionality of neurons measured in [6,7] spiking at the maximal increase at the rise of the envelope of the auditory signal at the onset of the nucleus of a syllable, where in most languages the nucleus is a vowel or a complex of vowels. In the following we call the measured neurons ‘V-edge neurons’. We assume, that the V-edge neurons are not perfect, otherwise no PING would be needed. As shown in section 4, the PING has the property to remove erroneous spikes emitted by the V-edge neuron not relating to the instances t_{PING}, and adds missing spikes not detected by the V-edge neuron.

The properties of the θ-oscillator above are based on some measurements, but still following hypothesis are needed to derive the architecture (fig.1):

- a PING generates θ-spikes at instances t_{PING}, where two adjacent spikes determine the phase and duration of a θ cycle.
- The PING is driven by a V-edge neuron - a model for the V-edge neurons driven by edge and sustained features generated within critical bands (CB-features).
- the θ-spikes are synchronously to the spikes of the V-edge neurons. Thus, a θ-cycle defines the position of a syllable by the position of adjacent instances of maximal rises of the modified loudness curve within syllabic kernels.

This section ends with the description of the first module of the θ-oscillator generating the auditory signal and its envelope. The other modules are described in section 3 and 4.

The generation of the auditory signal is implemented by an Gammatone filter bank organized in critical bands CB (tool-box [19]). For the needs of the θ-oscillator the filter bank is reduced by a set of 10 CBs (CB_1, … CB_{10}) with center frequencies 248, 328, 420, 529, 655, 803, 974, 1175, 1409, 1682 Hz adapted to the frequency ranges of vowels. The filter bank delivers as output the signals $z_k(t, F_k), k = 1, \dots, 10$ in an analytic form. The signal $|z_k(t, F_k)|$ is the amplitude modulated part of $z_k(t, F_k)$. Using a lowpass with an IR impulse response h_{TP},

$$\text{CB-AM}(t, h_{TP}, k) = 20 log_{10} \int_{-\infty}^{+\infty} |z_k(\tau, F_k)| \, h_{TP}(\mathrm{t} - \tau) d\tau; k = 1, \dots., 10 \qquad (1)$$

is a smoothed and log compressed version of $|z_k(t, F_k)|$ defining the auditory signal.

3 The V-edge Neuron

The *V-edge* neuron is implemented as a single bio-physical neuron, which models the complex of neurons measured in [6,7]. This complex is called *V-edge neurons*. It is an open question, whether the input of the V-edge neurons are CB-features related to critical bands. CB-features are generated in the inferior colliculus (IN) [17] or in the ‘speech-region’ of the STG [13,14]. The functionality of the CB-features produced in the IN is common to all mammals [17] and well explored, whereas the CB-features produced in the speech region of the STG are specific for humans and less explored. We assume that the input of the V-edge neurons is given by specific CB-features produced in the IN and further processed in the STG. These features belong to the class of specific edge and sustained features:

- edge features called env-edge features’, which describe ‘edges’ of the envelope of the partial loudness curves [8,9] given by the instants of their maximal rises.
- sustained features, called V-onset features, which describe the presence of the onset of vowels in the auditory signal. In the STG the presence of vowels has been measured in [10].

We assume, that the env-edge features are IN-features indicating edges of the partial loudness [9]. The V-features have the task to prevent the V-edge neuron to spike at rises in non-vowel regions (consonants). In the following subsections the modules generating the env-edge features and the V-onset features are described.

3.1 The env-edge Features

In [2,3] the CB-loudness is defined to be identical to the CB-AM(t, h_{TP}, k) curve (eq. 1). It turned out that this signal is quite noisy concerning instances of maximal rise. Recent measurement [9] come to the conclusion, that for each CB the partial short-term loudness (ST-CB-loudness) is generated as CB-features in the IN. The transformation of the CB-AM(t, h_{TP}, k) curve (1) to the ST-CB-loudness is given by following algorithm [9]:

Samples S_n representing the CB-AM(t, h_{TP}, k) (see eq. (1)) are transformed to the signal S'_n representing the samples of ST-CB-loudness (t_n, k) according to

$$S'_n = \begin{cases} \alpha_a S_n + (1-\alpha_a)S'_{n-1} \text{ for } S_n > S'_{n-1} \\ \alpha_r S_n + (1-\alpha_r)S'_{n-1} \text{ for } S_n \leq S'_{n-1} \end{cases} \tag{2}$$

$$\alpha_a = 1 - e^{-\frac{\Delta t}{Ta}};\ \alpha_r = 1 - e^{-\frac{\Delta t}{Tr}};\ T_a = 0.045\ [s];\ \ T_r = 0.02\ [s];\ \Delta t = \frac{1}{fs}.$$

f_s denotes the sampling frequency of the CB-AM-signal (1).

In a second step the ST-CB-loudness is transformed by an automatic gain control (AGC) to a signal $AGC(ST - CB - loudness)$. Compared to [2,3] it turned out that the AGC increases the correctness of detect maximal rises (see section 4). The AGC is implemented by the following algorithm transforming samples x_n to samples y_n.

$$y_n = gain_n * x_n;\ gain_n = a * (level - y_{n-1}) + gain_{n-1} \tag{3}$$

The gain factor a determines the speed of adaptation of y to a reference level (*level*). a is tuned to achieve an adaptation of y to the value '*level*' within the duration of an average syllable (180ms). To avoid overshoots, the adaptation is stopped in regions of pauses (non-speech) detected as low-level signals. Given the curves of $AGC(ST - CB - loudness)$, areas of increase of the curves are determined. Areas of ripples in the curves are concatenated to larger areas. According to some heuristic rules some areas are deleted leading to final areas of increase. For the resulting areas the instant of maximal rise is extracted. The sequences of the instances of the rises constitutes the env-edge features modelled by pulses with a value given by the stiffness of the area of increase as done in [2,3].

3.2 The V-onset Features

The V-onset features should indicate the presence of the onset of a vowels or complexes of vowels. Due to the lack of cortical measurements of neurons generating such features, we implemented a GMM classifier for each CB. For each feature vector $V_k(t_n)$ (eq. 4) the classifier determines the probability of 4 classes*: pause, consonant, vowel*, and *onset of vowels*. The features used to train and test the GMMs are modulation features as found in the STG [14]. In our implementation these features are given by a STFT transformation [18,20, 21] of the CB-AM signal (1) normalized by the AGC (3) leading to the signals

$$y_k^{AM}(t) = \text{AGC}\big(\text{CB} - \text{AM}(t, h_{TP}, k)\big), \text{k} = 1, \ldots.,10.$$

The STFT transforms $y_k^{AM}(t)$ to feature vectors $V_k(t_n)$ with dimension dim_Ω sampled at instances t_n with $\Delta(t_n) = 10ms$. dim_Ω denotes the number of modulation frequencies Ω_i used:

$$Y_k^{AM}(t, w, \Omega_i) = \int_{-\infty}^{+\infty} y_k^{AM}(\tau) w_{\Omega_i}(t-\tau) e^{j\Omega_i \tau} d\tau\ ; k = 1 \ldots 10; \text{i} = 1, \ldots, dim_\Omega$$

$$MFK_k^i(t_n) = \left|Y_k^{AM}(t_n, w, \Omega_i)\right|;\ \Delta(t_n) = 10ms$$

$V_k(t_n) = [MFK_k^1, \ldots, MFK_k^{dim_\Omega}]$; k=1,…,10 (4)

We use as STFT-windows w_Ω Gaussians $N(0, \sigma_\Omega^2)$, which has the best spectro-temporal resolution [20]. The width of the windows is determined by the value of σ_Ω adapted to Ω: for low frequencies $\Omega,$ the window is large to model the modulation of stationary sounds as vowels; for high frequencies Ω, the windows are smaller to model non-stationary sounds as onsets. This implementation seems to be consistent with the measurements in [14].

For training the GMMs, the vectors $V_k(t_n)$ must be aligned to the classes to be classified using the labels of the speech data base. The alignment of the class *onset of a vowel* is defined by an 'onset- region' of 40ms with distances +-20ms from t_{ref} defined in subsection 4.2.2. t_{ref} denotes the instance of maximal increase of a modified loudness at the onset of a vowel.

Recognition experiments on the 4 classes showed that the highest recognition rate was achieved for the class *onsets of the vowels* (see subsection 4.2.1). Due to this property, we use as V-onset feature the probability of the presence of the class *onset of a vowel* and not the class *vowel.*

4 Experiments

4.1 Implementation of the Modules of the θ-Oscillator

The V-edge neuron and the PING are modelled by a simplified bio-physical Hodgkin-Huxley model [22,4], whereas the other modules are implemented as 'engineering models' as described below. The modelling of oscillations is done easier in bio-physical neuronal models, because the phase and duration of cycles of the θ-oscillations corresponds directly to the instances of the spikes provided by the V-edge neuron and the PING. To decrease the time needed for simulation, the generation of the auditory signal is done by an engineering approach not using bio-physical implementations. This approach is justified by the precise models available [18]. Due to missing measurements in the speech area of the STG, the generation of the edge and sustained features shown in fig.1 are engineering models implemented by 'invented' algorithms designed for optimal performance in generating correct θ-spikes.

To couple an engineering model to a bio-physical neuron we use two methods. Either the output of the engineering model are pulses, which are interpreted as spikes entering as input to bio-physical modeled synapses, or the output of the engineering model are values, where the values are interpreted as ion-currents fed directly into the cell of the Hodgkin-Huxley model. The θ-oscillator is implemented in matlab running in real time on a 4-core laptop.

4.2 Evaluation

The θ-oscillator is evaluated by a speech database containing 1300 phonetically diverse utterances (read speech from a professional British speaker) [23, 2]. The database is phonetically labeled and designed for articulatory research as done in [24] to detect the θ-oscillation in speech production.

Four kinds of evaluation are performed to determine the quality of the V-onset features, the env-edge features, the V-edge neuron and the PING. The V-onset features are evaluated in subsection 4.2.1. The other evaluations are done by comparing trains of spikes as described in subsection 4.2.2.

4.2.1 Evaluation of the V-onset Features

As intermediate result, the GMM-classification determines for each feature vector $V_k(t_n)$ (4) the probability of presence of all the four classes *pause, consonant, vowel, onset of vowel.* If the class with the highest probability is not consistent with the label of the database an error is counted. The error rates averaged over all $V_k(t_n), k = 1 \dots 10$ are shown in tab.1. The low error rates of 4% for the class *onset of vowels* is quite astonishing. It seems that the feature vectors $V_k(t_n)$ fit well to detect the onset of a vowel or a complex of vowels.

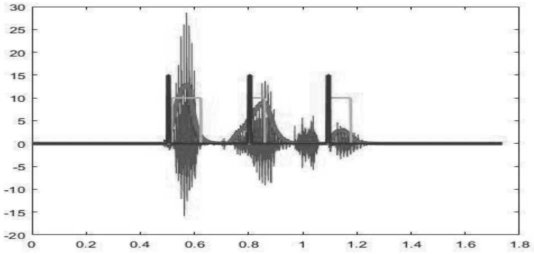
Table 1: Error rates in % for the 4 classes in dependence on the number of modes used for the GMMs. In these experiments $\dim_\Omega$=9 is chosen.

number of modes	pause	consonant	vowel	onset of vowels
4	21	8	12	4
8	0	31	15	4
32	0	29	15	4

4.2.2 Evaluation of Train of Spikes

For evaluation we compare the train of pulses or spikes emitted at instances t_{edge} with reference trains at instances t_{ref}. An instance t_{ref} is defined as the maximal rise of a modified loudness curve at the vowel onset. The modified loudness is defined as the sum of the partial loudness of CB$_1$, …, CB$_{10}$. The position of the region of the vowel onset is derived from the labels of the vowels as described in subsection 3.2. Ideally, the instance t_{ref} should be in accordance to the spiking of the V-edge neurons. Whether this is true is an open question. For illustration the instances t_{ref} are depicted in fig. 2 for the utterance 'Jack Webster'.

Figure 2: curves: speech signal of utterance 'Jack Webster' with the envelope of the modified loudness, the pulses at instances t_{ref} and the position (rectangle curve) of the 3 vowels. As the partial loudness of the selected CBs do not cover the frequencies of the fricative 's' in Webster', the modified loudness shows no maximum at this sound.



The correctness of an instance t_{edge} is defined by the distance to a related reference instance t_{ref}: we define an instance t_{edge} as correct, whenever the relation

$$\left| t_{ref} - t_{edge} \right| < maxDist_{ref} \; ; \; maxDist_{ref} = 0.070 \; [sec] \qquad (5)$$

holds. The value of $maxDist_{ref}$ is quite conservative due to the lack of cortical measurements. Applying this definition, missing (deleted) and inserted θ-spikes can be defined leading to the percentage of correct detection and the percentage of deletion- and insertion errors. This measure of correctness is applied to pulses/spikes t_{edge} from the env-edge features, the V-edge neuron and the PING.

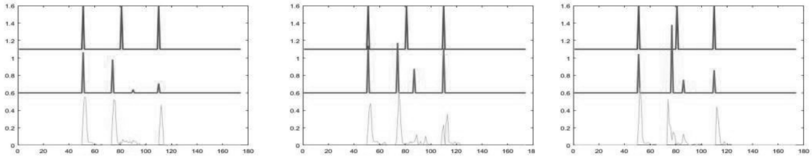
As baseline for judging improvements of the correctness, we use the results achieved in [2] for the θ-spikes of the PING as depicted in tab.2.

Table 2: Correctness of instances t_{edge} of the PING for 181 utterances (2491 syllables) using two settings S1 and S2 of parameters

Setting	Correct %	Insertion %	Deletion %
S1	79	30	19
S2	82	36	15

As shown tab.3, the correctness of the instances t_{PING} depends on the correctness of the instances t_{V-edge} of the V-edge neuron. The output of the V-edge neuron is determined by its input – the env-edge and V-onset features for all CBs. For illustration of the natures of these features, fig. 3 shows their plot for the same utterance 'Jack Webster' as above for 3 CBs.

Figure 3: From left to right pulses/spikes t_{edge} together with t_{ref} from CB$_1$ (248Hz), CB$_2$ (328Hz), CB$_3$ (420Hz); Top: instances of t_{ref}; Middle: env-edge pulses weighted with stiffness; Bottom: probabilities of V-onset features.



Tab. 3 shows the correctness of instances t_{edge} of the env-edge features, the V-edge neuron and the PING for specific settings of parameters. The percentage of error rates and correct detected of $t_{env-edge}$ from the env-edge features are averaged over all 10 CBs. They show high error rates in missing spikes. As the spikes of all CBs are input to the V-edge neuron many spikes are recovered, but many insertions pop up. Finally, the PING removes many of the inserted spikes, but loose some already detected correct spikes. As shown for setting 2, the correctness of the PING can be increased, but to the cost of increase of insertion errors. The mechanism for deleting and adding spikes is described in detail in [2,3]. The mechanism for deleting inserted θ-spikes is performed by deleting all spikes, which are close together. Only the first spike is maintained. The insertion of missing θ-spikes is performed by injecting a constant current into the PING, which increase its sensitivity for input in the range of the instances of the expected θ-spike. Comparing the of instances t_{edge} of the PING in tab.2 with those of tab. 3, the achieved improvements in correctness are evident.

Table 3: correctness for different setting S1, S2.

Setting	t_{edge}	Correct %	Insertion %	Deletion %
S1	env-edge feature	62	6	37
S1	V-edge neuron	93	139	6
S1	PING	89	16	10
S1	env-edge feature	62	6	37
S2	V-edge neuron	94	163	5
S2	PING	91	19	9

5 Conclusion

The paper describes a model to mimic cortical generation of θ-oscillations. Compared to [2,3] large improvements on correctness of the θ-spikes are achieved. The improvements were gained by using the AGC transformed partial loudness to generate the instances of maximal rise and the use of the V-onset features to detect the presence of vowel onsets. Yet the gap to human performance is still large.

The architecture proposed allows many improvements to be implemented. Hopefully, in the near future spectro-temporal receptive fields (STRFs) as described in [13,14], relating to the env-edge and V-onset features, will be measured, which can be used to achieve human performance.

6 References

[1] Lowerre, P.T.: *The Harpy speech recognition system*. In *Ph.D. Thesis* Carnegie-Mellon Univ., Pittsburgh, PA. Dept. of Computer Science,1976.

[2] Höge, H.: *A Cortical Model for a θ-Oscillator Segmenting Syllables*. In *Proc. ITG*, 2021.

[3] Höge, H.: *Cortical Segmentation of Syllables*. In Proc. ESSV, 2021.

[4] Hyafil, A., 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.

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

[6] Oganian, Y. and E. F. Chang: *A speech envelope landmark for syllable encoding in human superior temporal gyrus.* In *Science Advances*, 2019.

[7] Kojima, K., Y. Oganian, C. Cai, A. Findlay, E. Chan and S. Nagarajan: *Low-frequency neural tracking of speech envelope reflects evoked responses to acoustic edges*. In *Preprint,* 2020.

[8] B. Moore, B.R. Glasberg, A. Varathanathan and J. Schlittenlacher: *A Loudness Model for Time- Varying Sounds Incorporating Binaural Inhibition*. *Trends in Hearing*, Vol. 20, 2016.

[9] Thwaites, A., J. Schlittenlacher , I. Nimmo-Smith, W. D. Marslen-Wilson , B. C. J. Moore: *Tonotopic representation of loudness in the human cortex.* In *Hear. Res. 344*: 244–254, 2017.

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

[11] Höge, H.: *Using Elementary Articulatory Gestures as Phonetic Units for Speech Recognition*. In *Proc. ESSV*, 2018.

[12] Opoku-Baah, C., Schoenhaut, A. M., Vassall, S. G., Tovar, D., A., Ramachandrannd, R., Wallace M., T.: *Visual Infuences on Auditory Behavioral, Neural, and Perceptual Processes: A Review.* In *JARO,* 2021.

[13] L.S. Hamilton, E. Edwards, F. Edward, E.F. Chang: *A Spatial Map of Onset and Sustained Responses to Speech in the Human Superior Temporal Gyrus*. In *Current Biology* 28, pp. 1860–1871, 2018

[14] Hullett, P. W., L. S. Hamilton, N. Mesgarani, C. E. Schreiner and E. F. Chang: Human Superior Temporal Gyrus Organization of Spectrotemporal Modulation Tuning Derived from Speech Stimuli. In *Journal of Neuroscience*, 36 (6) 2014 – 2026, 2016.

[15] Ney, H.; *The Use of a One Stage Dynamic Programming Algorithm for Connected Word Recognition*. In *IEEE Trans.* In *Acoustics, Speech and Signal Processing*, Vol. ASSP-32, No.2, pp.263-271, 1984.

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

[17] Winer, J. A. and C. E. Schreiner: *The Inferior Colliculus*. New York: Springer, 2005

[18] Chi, T., Ru, P., Shamma, S. A.: *Multiresolution spectrotemporal analysis of complex sounds*. In *J. Acoust. Soc. Am.* 118, August 2005, pp. 887–906.

[19] Hohmann,V.: *Frequency analysis and synthesis using a Gammatone filterbank*. In *Acta Acoustica United with Acustica*, Vol 88, pp.433-442 2002.

[20] T.F. Quatieri, *Discrete Time Speech Signal Processing*. Upper Saddle River, NJ: Prentice Hall PTR, 2002.

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

[22] W. Gerstner and W. Kistler, “Spiking Neuron Models,” *Cambridge University Bridge*, UK 2002.

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

[24] Höge, H.: *Extraction of the Θ- and ɤ-Cycles Active in Human Speech Processing from an Articulatory Speech Database*. In *ESSV*, 2019.

[25] Höge, H.: *The nature of the articulatory code*. In *Proc. Konferenz Elektronische Sprachsignalverarbeitung* (*ESSV*), 2020

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