THE USE OF TEMPORAL FEATURES IN CORTICAL SEGMENTATION OF SYLLABLES

Harald Höge

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

Abstract: There is high confidence for the hypothesis that in speech perception the cycles of a θ-oscillation segment the auditory signal into syllables [8]. Yet the functionality of the oscillator generating the θ-oscillation is unknown. We follow the finding that, within an auditory scene, speech is perceived as a stream given by temporal coherence [12]. We work with the hypotheses that the θ-oscillator is driven by temporal features providing this coherence. We propose a new temporal feature called *O-distance*, which detects the onset of a syllable - the starting point t_o of a θ-cycle – triggered by the temporal distance from t_o to the instance of the maximal rise of the loudness curve of the vowel. To extract t_o from the auditory signal, we use the statistical properties of this distance based on the C-center hypothesis [25], which predicts a close temporal relation of the onset consonants to the onset of a vowel. The statistics are derived from reference O-distance extracted from an articulatory database, where the minima and maxima of the loudness are related to maxima and minima of the lower incisor and tongue tip. To judge the quality of the O-distance extracted from the auditory signal, we regard the temporal deviation of the O-distance to the reference O-distance. Currently we achieve a mean deviation of 34ms.

1 Introduction

Cortical research relies on measurements of the activity of neurons. In the last 20 years, the knowledge in neuroscience for cortical processing of speech has increased substantially by measuring the activity of neurons. The measurements are performed by cortical electrocorticography (ECoG) and deep brain stimulation (DBS) in clinical settings [1]. ECoG is applied to cure the epilepsy decease, and DBS is applied to cure the Parkinson decease [2]. ECoG allows to measure simultaneously the Local Field Potential (LFP) with high temporal resolution of about 1000 neurons located at the surface of the cortex [3]. Yet, compared to the size of neurons the distance between two neighbored electrodes (ca. 1mm) measuring the LFP is large. Thus, the spatial resolution is low, and the input-output relations of close neurons cannot be measured. DBS allows to measure the LFP generated by a few neurons in deep layers of the brain [3].

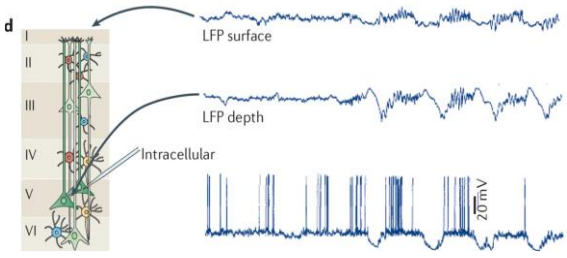

Figure 1 -The left side shows the cortical 6-layer structure (I-VI) of the human cortex (original picture in [3]). Using ECoG, only the LFP, generated from neurons in layer I can be measured with high precision, whereas the intercellular LFP of neurons of the deepest layer VI can be resolved at layer I only poorly. Thus, periodic spike trains generated in layer VI look at layer I like oscillations.

Most publications performing measurements in speech production and speech perception with ECoG for exploring the activities from neurons generated in deep layers, interpret periodic bursts of spike trains generated in layer VI as oscillations (see fig.1).

To overcome the problems of missing information given by current technology in measuring the functionality of the neurons, hypotheses are postulated, whose evidence are checked by

 DOI: 10.35096/othr/pub-7082, URN: urn:nbn:de:bvb:898-opus4-70823

correlations between events of the speech signal and the output of measured neurons. The hypotheses are supported further by psycho-acoustic measurements, and by measurements of the activity of mammal neurons located in deep layers in areas, which have the same functionality as the neurons in human layers. This approach to find evidence for hypotheses is followed also in this paper.

In evolution, speech production and speech perception developed together to perform multimodal communication using facial, manual, and articulatory gestures [4]. Whereas the cortical process of steering and perception of gestures performed by humans is common to mammals, the use of an articulatory stream of gestures for communication is unique for humans. This stream represents a sequence of cognitive units identified as syllables constituting a language. We follow the OVC approach where a syllable is produced by three kinds of articulatory gestures called the Onset (O)-gestures, the vowel cluster (V)-gestures, and Coda (C)-gestures. This approach has been proposed by the *frame/content theory of evolution of speech production* [5]. We extend this theory, that the OVC gestures have the same articulatory code (AC) in speech production and speech perception [6,7]. In the brain the AC is transmitted via θ – oscillation into various areas, where each θ –cycle encodes the AC describing a syllable [7]. In [8] a new principle of speech processing is proposed, where the θ – oscillations perform the segmentation of the auditory signal into syllables. The θ – oscillations have the property that the phase and the duration of a θ – cycle is in synchrony with the temporal position and the duration of a syllable. Thus, there must exist a θ – oscillator, which adapts to the varying speaking rate varies. I.e. each θ – cycle must be adapted (entrained) to each syllable. In the following we call a θ – cycle steering the articulators during speech production by the *articulatory θ –cycles* and we call a θ-cycle segmenting speech in speech perception the *perceptive* θ-cycle. In communication these cycles are in synchrony as discussed in [23].

This cortical method to segment speech into syllables via θ – cycles is different to the method applied nowadays in automatic speech recognition (ASR), where segmentation is solved by a search algorithm [9] including a language model. In the starting time of ASR, in the 1970s, 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 [10]. 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 integrate easier visual input (lip reading) and consumes less computing power.

The cortical generation of the *perceptive* θ – cycles is far away from being understood. Due to the high human performance in speech communication, there must exist a neural processing scheme different to the ASR approach [9] and former spectral approach [10]. A hint to the human processing scheme is given in the stream theory [11, 12], which explains how speech from a speaker can be interpreted as a coherent auditory stream, which can be tracked by a listener within a multi-speaker scenario (cocktail party effect). The stream theory postulates that a stream can be tracked if it contains repeating temporal cues. Further it has been demonstrated that the manipulation of the temporal structure of syllables has a great influence in the intelligibility of speech [13].

We assume, that the movements of the articulators produce auditory cues, which provide the coherence and allow to reconstruct the θ – cycles. It is not known, which temporal cues are extracted from the auditory signal to drive an θ – oscillator generating the *perceptive* θ – cycles. A candidate for such cues is the functionality of the V-edge-neurons, which have been detected in the superior temporal gyros (STG) [14]. A V-edge-neuron spike on the maximal rise of the loudness-signal at the onset of a vowel. These neurons belong to the class of *onset neurons* and *sustained neurons* active in the STG [15]. The V-edge neurons must detect the presence of a

vowel and the instance of maximal rise of the loudness. We assume that the V-edge neurons get input from *sustained* neurons detecting the presence of vowel clusters [16] and get input from *onset* neurons detecting the event of an edge given by the maximal rise of the loudness [14]. We assume that in future in the STG many auditory and phonetic features used to drive the θ – oscillator speech perception will be detected.

Publications on cortical models segmenting the auditory signal into syllables are scarce. The first cortical model using a θ – oscillator is described in [17]. The θ – oscillator is implemented by a PING structure as found in the hippocampus [18], where oscillators generate oscillations for various frequency. The θ – oscillator [17] is driven by the envelope of the auditory signal. Thus, the envelope of the auditory signal provides the cue for perceiving speech as a stream. In [19] we propose a model of the θ – oscillator based on the framework of a PING as in [17], but with other features driving the θ – oscillator. The driving features are V-edge-neurons and the curve of the loudness. This approach substantially improved the performance in segmentation.

We aim to improve further the performance of an θ – oscillator segmenting syllables by using a new temporal cue. We hypothesize the existence of neurons providing the feature '*O-distance*'. These neurons spike at the instance of the onset of the syllable. These neurons are driven by V-edge-neurons together with other onset and sustained neurons. The functional core of the O-distance is given by the property to determine the temporal distance between the instance of the onset of a syllable and the instance of the V-edge.

The paper is organized as follows. Section 2 handles the nature of the O-distance, derived from the temporal structure of a syllable. Section 3 describe the methods to extract and evaluate the O-distance and reports on experimental results on the precision of the O-distance.

2 The O-Distance

2.1 Definition of the Syllable and the O-distance

We propose a new temporal feature called *O-distance*, which detects the onset of a syllable t_o triggered by a temporal distance Δ_O, the distance between t_o and the onset of the vowel (V-edge). The V-edge is defined by the instance t_V of maximal rise of the curve of the loudness within the area of a vowel [14]. The instant of the onset of the syllable is defined by the syllable-starting-time t_o, which is the instance, when the movements of the articulators start to produce a syllable. t_o can be either the starting of the movement of the articulators building the O-gesture, or in the case of missing O-gesture the starting time of V-gesture of the articulators producing the vowel. We hypothesize that the start of a θ –cycle is given by the instance t_o.

t_o determines the split of the consonants between two neighbored syllables into coda consonants of the first syllable and into onset consonants of the following syllable. For judging the correctness of the t_o, a definition of a syllable is needed, which provides the 'correct' split of the consonants between neighbored vowels. Yet the definition of the onset-consonants and the coda-consonants is diverse. This problem is discussed in [20] where methods segmenting speech into syllables are evaluated. Whereas it is agreed that the center of a syllable is given by a complex of vowels (V*), the problem of splitting the consonantal clusters C* in V* C* V* constructs is not solved. To solve this problem, in the next subsection we propose a specific syllable called the cortical syllable.

2.2 The Cortical Syllable

The definition of a *cortical syllable* (CS) is based on the articulatory OVC approach where OVC gestures are steered by θ – oscillations. This definition needs the temporal position of the syllables and the phases of the θ –cycles. To the author's knowledge there exist no cortical

measurements allowing to study of the temporal relation between the phase of the θ –oscillation and the position of the syllables. Consequently, definition of the CS is based on hypotheses. A definition of a cortical syllable (CS) related to θ –cycle is given in [21], where the CS is called a *theta-syllable* defined by: *The theta-syllable is a theta-cycle long speech segment located between two successive vocalic nuclei.* This definition is centered on a sequence of vowels. This definition has the drawback that it does not describe a syllable spoken in isolation.

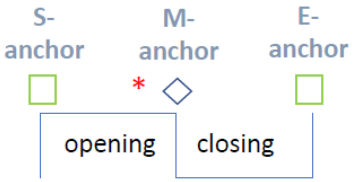

Figure 2 - Anchor-points describing the starting, middle and end of an articulatory Θ-cycle. The * marks the instance of maximum energy of the V complex. The S-anchor point is the instance of the onset of the syllable i.e. the beginning of the O-gesture. The M-anchor determines the instance where the closing of the vocal tract starts. The E-anchor determines the end of the articulatory θ-cycle.

Our definition of the CS is based on articulatory movements in speech production, where the CS is defined by an *articulatory* θ- cycle of opening and closing the vocal tract (see fig.2). This definition allows to handle a syllable spoken in isolation. The approach to move from phonetically defined items as syllables and phones to constructs of articulatory movements is also followed in [22], where the activity of neurons performing the steering of articulators are measured in the ventral senso-motoric cortex. Yet the definition of the CS relies on the existence of *articulatory* θ- cycles.

2.3 The Ideal Cortical Syllable

As discussed above, we relate the O-distance t_o to the start of the movements of the articulators generating the O-gestures and to the start of a *θ –cycle*. Fig.3 shows an example, were the maximum of the curve of the movement of lower incisor is coincident with t_o marked in fig.3 by O-ref-dist.

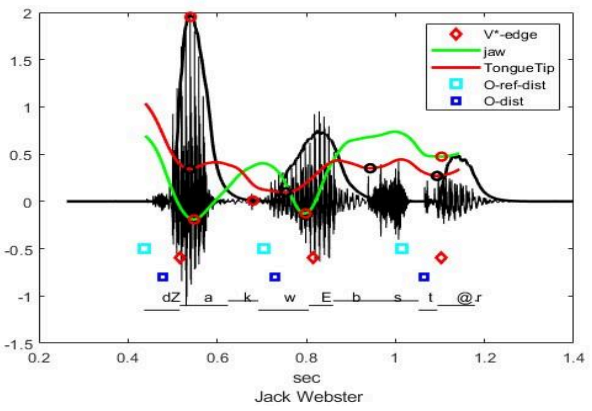

Figure 3 – The figure shows the speech signal of the utterance 'Jack Webster' consisting of 3 ideal syllables together with the loudness. Further the movements of the lower incisor and the tongue tip are shown. The minima and maxima of the curves are candidates for the instances of the articulatory turning between open-close cycles. edges where the opening or closing of the vocal tract starts Those edges determine the timing of the θ –cycles. Fig. 3 show also the incidence of the V-edge marked by a diamond together with the incidences of the O-distance (lower square) and its reference (upper square).

Further we regard the minima and maxima of the loudness curve. We assume, that these minima and maxima are coherence features important for perception. These coherence features must be produced by the articulators. We define a CS to be an *ideal* CS if the minima and maxima of the loudness fit to the minima and maxima of the curves of movement from the lower incisor and/or the tongue tip (see subsection 3.1). As described in section 3 the extraction of the O-distance is performed only with ideal CS, because in these cases t_o can be extracted.

2.4 The Nature of the O-distance

The O-distance is derived from the activation-spin model of Tilson [24]. Tilson states in the summary*: Studies of the control of complex sequential movements have dissociated two aspects of movement planning: control over the sequential selection of movement plans, and control*

over the precise timing of movement execution. This distinction is particularly relevant in the production of speech: utterances contain sequentially ordered words and syllables, but articulatory movements are often executed in a non-sequential, overlapping manner with precisely coordinated relative timing. Thus, due to the definition of the CS and the investigations in [22], the activation-spin model moves from phonetic constructs to articulatory movements constructs. In our approach the movements plans perform the selection of the OVC-gestures. Due to [26], the selection of the gestures is given by the composition of articulatory codes of each of the selected gestures resulting in a composed code for a syllable. This code is input to motor-programs steering the movements of the articulators. Thus, the motor programs can produce articulatory movements of the OVC – gestures providing optimal vocal tract configurations to allow to encode the articulatory code for all gestures of the syllable in perception. The activation-spin model is a model for the motor programs. This model explains the difference between the articulatory movements of the O-gestures versus the C-gestures. We conclude that the O-gestures and the V-gestures are generated by motor programs, which have as input the articulatory code of the O- gesture and the V-gesture generating combined movements of the articulators for both gestures. Thus, the OV- movements have a rigid timing described by the C-center effect [25]. In contrast, the motor programs of the C-gestures have as input solely their own code.

3 Methods and Experiments

3.1 Method to Determine the O-distance and Reference O-Distance

The O-distance must be derived from the auditory signal bottom up. Due to the definition of the O-distance, the instance of the V-edge t_V and the instance of the start of the syllable i.e. the start of the θ-cycle t_o must be estimated. t_V is determined as described in [19] using the loudness [14]. To avoids handling insertion or deletion errors of vowels t_V is determined knowing the position of the vowel.

The instance t_o is determined in two steps. In the first step the range Δ_O of the of potential instances of t_o is determined. Δ_O is given by the statistics of the reference O-distances generated as described below. The statistics deliver the mean value and the standard deviation of t_o. These values determine the range of the position of potential instances of t_o.

In the second step we use several heuristic methods to determine a value of t_o in Δ_O. These methods are coarse phonetic features describing the probability of areas of non-speech or consonants in Δ_O. The methods use the properties described by the C-center hypothesis and the minima of the loudness curve.

To evaluate the accuracy of the O-distance, we generate reference O-distances using an articulatory database. This database allows to extract ideal CS together with the articulatory θ-cycles using a method already described in [27] where extrema of the envelope of the auditory signal are compared to the extrema of the lower incisor to test fitting. This method is improved by using in addition to the lower incisor the tongue tip and by replacing the envelope of the auditory signal by the curve of the loudness as done in [19]. By integrating the tongue tip additionally, we use the maxima and minima of the movement of both articulators to judge the fit to the maxima and minima of the loudness. Further the threshold deciding if the deviation of the extrema of the loudness and the articulators is small enough to be an ideal CS, has be diminished. This procedure leads to more precise position of the reference t_o. The edge of the vowel t_V is determined as described for the extraction of the O-distance above.

3.2 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) [28]. 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 labelled automatically Labelling is performed by forced alignment [29] using the Combilex lexicon with its notation of the phone labels [30].

3.3 Results

Using the method described in 3.1. we found 3 542 ideal CS in the database mngu0.

Comparing the reference O-distances with the O-distances a mean deviation of 34ms was achieved.

To judge the statistical differences of the reference O-distances and the O-distances, the table below shows the values of means and standard deviations for both distances. The high value of the standard deviation for the O-distances (180ms) shows, that there is a large range of the values of the extracted O-distances hinting to estimation errors.

	Reference O distance	**O-distance**
Mean	102ms	79ms
Standard deviation	52ms	180ms

4 Conclusion

To improve the functionality of a cortical θ-oscillator in segmenting the auditory signal into syllables, we propose a new feature called *O-distance* driving the θ-oscillator. This feature is extracted bottom up from the auditory signal. We hypothesize the existence of *O-distance* neurons spiking at the instance t_o of the start of articulatory gestures producing the onset of a syllable. The temporal range Δ_O of t_o is given by the mean and standard deviation of t_o, which is determined from reference-O-distances extracted from an articulatory database. Within Δ_O the final position of t_o is determined using feature related to the C-center hypothesis. Comparing t_o the reference O-distance to the O-distance a mean deviation of 34ms was achieved. This result encourages us to work further in the direction of including statistical knowledge about articulatory timing into the method to extract the O-distance feature to achieve less deviation.

Future work must be done in two directions. First, the method to extract the reference O-distances must be improved. This can be done, by regarding the movements of all articulators performing a closure of the vocal tract. Second, the methods to find the final position of t_o in the range of Δ_O must be improved.

5 References

[1] Chrabaszcz, A., W. J. Neumann, O. Stretcu, W. J. Lipski, A. Bush, C.A. Dastolfo-Hromack, D. Wang, D. J. Crammond, S. Shaiman, M. W. Dickey, L. L. Holt, R.S. Turner, J. A. Fiez, and R.M Richardson: *Subthalamic Nucleus and Sensorimotor Cortex Activity During Speech Production*. In *The Journal of Neuroscience,* 39 (14), pp. 2698 – 2708, 2019.

[2] Chen, Y. C., H-T Wu, P-H Tu, C-H Yeh, T-C Liu, M-C Yeap, Y-P Chao, P-L Chen, C-S Lu and C-C Chen: *Theta Oscillations at Subthalamic Region Predicts Hypomania State After Deep Brain Stimulation in Parkinson's Disease*. In *Front. Hum. Neurosci*. 15, 2021.

[3] Buzsáki, G., 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.

[4] Arbib, M. A., Liebal, K., and Pika, S.: *Primate Vocalization, Gesture, and the Evolution of Human Language*. In *Current Anthropology* Vol. 49, 6, 2008.

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

[6] Höge, H.: *The nature of the articulatory code*. In *Proc. ESSV2020*, 2020.

[7] Höge, H.: *Towards a Brain Computer Interface for Speech Perception*, In *ITG 2023*, 2023.

[8] 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.

[9] 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.

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

[11] Zion Golumbic, E. M., D. Poeppel, and C.E. Schroeder: *Temporal Context in Speech Processing and Attentional Stream Selection: A Behavioral and Neural perspective*. In *Brain Lang.,* pp.151–161, 2012.

[12] Elhilali, M., L. Ma, C. Michey, A.J. Oxenham A.J., and S.A. Shamma: *Temporal Coherence in the Perceptual Organization and Cortical Representation of Auditory Scenes,* In *Neuron*, 61(2), pp.317–329, 2009

[13] Ghitza O.: *On the role of theta-driven syllabic parsing in decoding speech: intelligibility of speech with a manipulated modulation spectrum*. In *Frontiers in Psychology*, 3:238, 2012.

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

[15] L.S. Hamilton, E. Edwards, F. Edward, and 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

[16] 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.

[17] 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.

[18] Buzsáki, G.: *Theta Oscillations in the Hippocampus*. In *Neuron*, Vol. 33, pp. 325–340, 2002.

[19] Höge, H.: *Improved feature driving a θ-oscillator for cortical segmentation of speech into syllables*. In *Proc. ESSV2022*, 2022

[20] Villing, R, T. Ward, and J. Timoney: *Performance limits for envelope based automatic syllable segmentation*. In *IET Irish Signals and Systems Conference (ISSC*), pp. 521–526, 2006.

[21] GHITZA, O.: *The theta-syllable: a unit of speech information defined by cortical function*. In *Frontiers in Speech psychology,* Article 138:1-5, 2013

[22] CHARTIER, J., G.K. ANUMANCHIPPALLI, K. JOHNSON, E.F. CHANG: *Encoding of Articulatory Kinematic Trajectories in Human Speech Sensorimotor Cortex*. In *Neuron* 98, pp.1042–1054, 2018

[23] HÖGE, H.: *Synchrony of θ - Oscillations in Speech Perception and Speech Production*. In *Proc. ESSV2023*, 2023.

[24] TILSEN, S.: *Selection and coordination of articulatory gestures in temporally constrained production*. In *Journal of Phonetics 44*: 26–46, 2014.

[25] Marin, S. and M. Pouplier: *Temporal Organization of Complex Onsets and Codas in American English: Testing the Predictions of a Gestural Coupling Model*. In: *Motor Control*, 14, 380-407, 2010.

[26] HÖGE, H.: *Towards a Brain Computer Interface for Speech Perception*. In *ITG 2023*, 2023.

[27] HÖGE, H*.: Extraction of the ϴ- and ɤ-Cycles active in Human Speech Processing from an Articulatory Speech Database*. In *Proc. ESSV2019*, 2019.

[28] 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.

[29] 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.

[30] FITT, S., RICHMOND, K., and CLARK, R.: *The Combilex lexicon*. `www.cstr.ed.ac.uk/research/projects/combilex`