SYNCHRONY OF Θ - OSCILLATIONS IN SPEECH PERCEPTION AND SPEECH PRODUCTION

Harald Höge

Universität der Bundeswehr München

Abstract: In speech perception auditory θ - oscillations play an important role in segmenting speech into syllables [1]. Yet, due to imperfections in cortical measurements, the mechanism to create these θ – oscillations from the auditory signal is far away to be understood. The paper works with the hypothesis, that in speech production the timing to activate motor programs steering articulatory gestures for generating syllables can be modeled by articulatory θ – oscillations. As the movements of the articulators can be measured directly with electromagnetic articulography, we assume that the mechanism to generate articulatory θ – oscillations can be understood easier. This knowledge of this articulatory mechanism can be used to understand the mechanism for generating auditory θ – oscillations. The precondition of this approach demands, that the auditory and articulatory θ – oscillations are in synchrony in phase and duration neglecting a constant delay between the two oscillations. The goal of the paper is to proof this synchrony. First experimental investigations show the feasibility of this new approach.

1 Introduction

In speech perception θ - oscillation play an important role in segmenting speech into syllables [1]. In the following we call these oscillation auditory θ – oscillation. Further, embedded in each cycle of the θ – oscillations are gamma cycles to segment parts of the syllable. In [1] it is claimed, that these parts are phonemes. Based on the S/F theory [7] it is claimed that the parts are related to elementary articulatory gestures (EAGs). The EAGs are articulatory gestures producing the **O**nset, the kernel (**V**owel) and the **C**oda of a syllable leading to the OVC theory [5].

Applying this bottom-up approach – i. e., using auditory θ – oscillations to segment syllables - to automatic speech recognition (ASR), the interface between acoustic and symbolic processing is decoupled. This means the acoustic processing is independent from symbolic processing[1]. Yet, due to imperfections in cortical measurements, the mechanism to create the auditory θ – oscillations from the auditory signal is far away to be understood.

In speech production, the steering of the articulators is performed by motor programs dedicated to produce syllables [9]. The motor programs handle different parts of the syllable in correct sequence and timing. In [10] it is claimed that the timing of the onset of the syllable is performed including the vowel, whereas the timing of the coda depends only on the coda itself. We work with another hypothesis, assuming that the motor programs are activated according to the OVC approach, were the OVC gestures are activated independently. The activation is done at specific temporal instances of the phase of each θ – cycle of the articulatory θ – oscillations.

The core of the paper is the hypothesis that during communication the auditory θ – oscillations extracted by the listener for segmenting speech are synchronous to articulatory θ – oscillations applied from the speaker to activate the motor programs. The synchrony is defined by common instantaneous duration and phase between the two oscillations. For proofing this hypothesis,

[1] In current technology of ASR the acoustic and the symbolic processing is interlinked by a language model. Interlinking is done in a computational costly search process.

the auditory and articulatory θ –oscillations must be compared. To the author's knowledge there exist no cortical measurements for both kinds of θ – oscillations. In the context of hearing the own voice, recently θ – oscillations have been observed in the motor cortex [2].

The approach we are following is the evaluation of models generating auditory θ –oscillations and observation of movements of articulators. For perception, a first model has been proposed [1], which has been implemented by neuronal structures [11]. Due to new cortical findings [6], a new model for generating auditory θ –oscillations has been developed in [4]. In [5] a first model for generating articulatory θ – oscillations is proposed.

In section 2, the models for auditory and articulatory θ –oscillations described above are used to proof synchrony between both oscillations. In section 3 experiments are done using an articulatory database, to proof synchrony.

2 The Method to Proof Synchrony

The model [5] for articulatory θ – oscillations is derived from the measurement of the movement of the mandibular provided by electromagnetic articulography [3]. This model is called the jaw-model. The jaw-model assumes that a syllable is produced by an opening and closing cycle of the mandibular with the timing of an articulatory θ – cycle. In the corpus used for the experiments only 30% of the syllables follow this model. Those syllables are called ideal syllables. A model for the other syllables, where the open-close cycle is performed by other articulators, has still to be developed.

In speech perception 'V-edge neurons' have been measured in the auditory cortex [6] spiking at the maximal increase at the rise of the envelope of the auditory signal at the onset of the nucleus of a syllable. We call these temporal instances **auditory edges**. We follow the hypothesis that the auditory edges drive the auditory θ – oscillations [4].

Our model of articulatory θ – oscillations is based on articulatory edges. In [5] an articulatory edge is defined by the temporal instances, where the mandibular changes from opening to closing and from closing to opening. We assume that the articulatory edges T are the instances where motor programs are started. We claim, whenever the instances of the articulatory and auditory edges are in synchrony, the articulatory and auditory θ – oscillations are in synchrony too. This approach can be followed using the OVC theory, where the cycles of the oscillations are related to the OVC gestures of ideal syllables. Both edges are related to specific instances within the θ – cycles located at the onset of vowels as shown in the following section.

3 Experiments

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) [3]. 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 [12] using the Combilex lexicon with its notation of the phone labels [13].

3.2 Extraction of the Articulatory and Auditory Edges

14 756 ideal syllables, representing 30% of the syllables, were found in the articulatory mngu0 corpus. From these syllables the auditory and articulatory edges were extracted. As shown in

fig.1, the auditory edges are extracted at the maximal increase of the envelope of the speech signal within a vowel area. The envelope is given by the loudness as described in [6], where the loudness is calculated by the sum of partial loudness processed in all critical bands. In [4] a modified loudness is calculated, where only by the critical bands covering the frequency band of vowels are regarded.

Fig. 2 highlights the method and results for extracting the articulatory edges. For ideal syllables, maxima of the mandibular-curve denote the instances of articulatory edges. These instances are given, when the mandibular changes his direction either to opening or closing within a syllable.

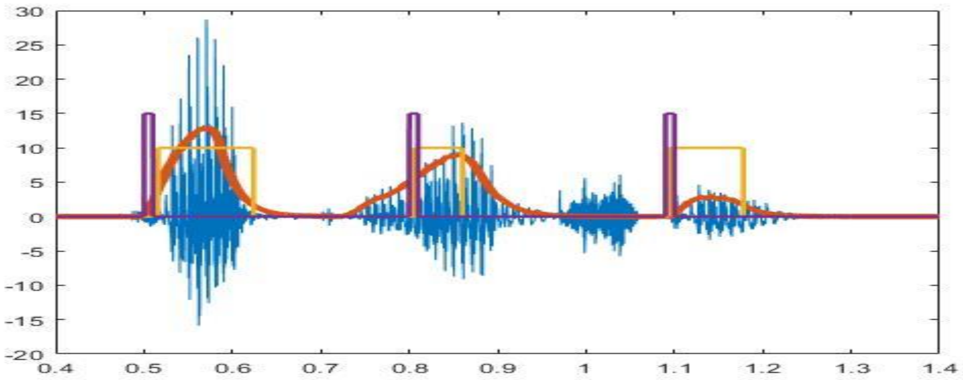

Figure 1 – the picture shows the speech signal together with the modified loudness curve of the utterance /Jack Webster/. Due to the restriction of frequency bandwidth, the modified loudness curve does not show a hill for /s/ in /Webster/. The rectangle curve denotes the start and end of the vowels given by the start-end labels of the vowels. The spikes denote the instances of the auditory edges. For this utterance, all auditory edges are located at the onset of vowels.

At the bottom of the figure the position and duration of the vowels and the consonant clusters at the onset and the coda of the syllables are shown.

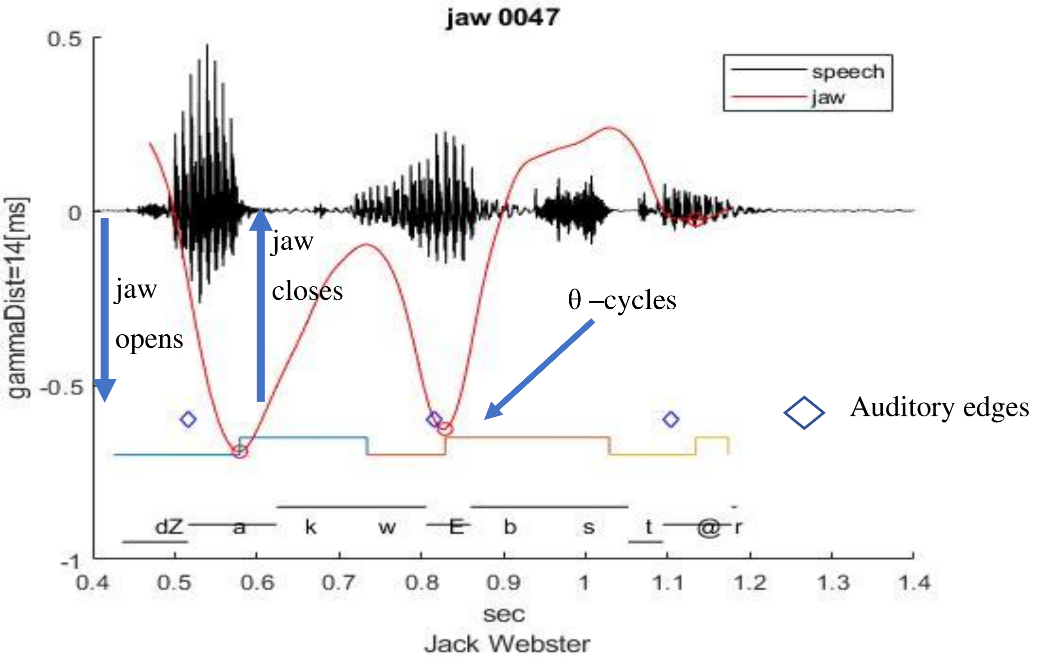

Figure 2 – At the top of the figure the speech signal of the utterance 'Jack Webster' consisting of 3 ideal syllables is shown. Attached to the signal of the speech is the curve of the movement of the mandibular. The minima of the curve of the mandibular denote the instance of the articulatory open-close edge, when the mandibular turns from opening to closing of the vocal tract. Those edges are assumed to determine the start of each θ –cycle located at the onset of a vowel or vowel cluster. The corresponding auditory edges are marked by diamonds. Their position

is located before the articulatory edge at the minima of the jaw-curve. The distance between the 3 auditory edges to the minima of the jaw-curve is not constant leading to the variance of the distances.

By comparing the temporal instances of corresponding articulatory and auditory edges extracted from ideal syllables we get following results: The average difference of corresponding auditory edges are 15ms with a standard deviation of 20.6 ms.

4 Conclusions

The experiments are focused on edges located around the kernel of a syllable. In average the auditory edge appears15ms earlier than the articulatory opening-closing edge. This delay is caused by the different phases of the articulatory and auditory edges. If the delay is constant, synchrony can be easily achieved by shifting the oscillations by a constant phase. Yet for synchrony, the variance of the distance between corresponding shifted edges should be zero. In the measurement a standard deviation of 20.6 ms has been achieved. This rather high value may be caused by imperfections in the measurements and methods implemented. Nevertheless, the experiments give strong evidence for the hypothesis of synchrony. For more precise detection of the timing of a θ –cycle, edges are needed to detect the begin and end of a syllable. For an ideal syllable this timing is given by the articulatory edges performed by the movements of the mandibular at the closure of the vocal tract.

For non-ideal syllables the closure is performed by other articulators. The current implementation assumes that the corresponding auditory edges indicating closures of the vocal tract are minima of the loudness. But these minima are quite fuzzy. The author is still waiting for cortical measurements of neurons detecting auditory edges at the begin and end of syllables. It is quite probable that those neurons use the acoustic properties of consonants in addition to the loudness.

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] CASAS, A.S.H., T. LAJNEF, A. PASCARELLA, H. GUIRAUD-VINATEA, H. LAAKSONEN: *Neural oscillations track natural but not artificial fast speech: Novel in- sights from speech-brain coupling using MEG*. NeuroImage, Elsevier, 2021, 244, pp.118577. 10.1016/j.neuroimage.2021.118577. hal-03373459

[3] 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.

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

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

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

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

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

[9] GUENTHER, F. H. AND G. HICKOK: *Speaking*. In *Neurobiology by Language* edited by Hickok, G., Small, S.L., chapter 58. Elsevier, 2016.

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

[11] 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

[12] CLARK, R. R., RICHMOND, K. and KING, S.: Multisyn: *open domain unit selection for the Festival speech synthesis system*. In *Speech Communication,* Vol.49, no.4, pp. 317-330, 2007.

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