Using Elementary Articulatory Gestures as Phonetic Units for Speech Recognition

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

1 Summary

The paper investigates the use of neuronal defined elementary articulatory gestures as phonetic units for ASR. In the sensorimotor cortex, complexes of neurons are found [4], where each complex steers the temporal dynamics of a specific articulatory gestures representing a set of 'elementary articulatory gestures'. For speech production current research focuses on the nature of the function and the control mechanisms of articulatory gestures. I make the hypothesis, that each of the elementary gestures can be described by three classes of gestures relating to a syllable: the onset gestures, the vowel gestures, and the offset gestures steered by oscillations. The offset/onset gestures relate to the consonant clusters, which are before/after the central vowel of a syllable. The paper treats two aspects:

- Theoretical aspect: this part of the paper investigates the plausibility of the elementary articulatory gestures based on recent cortical measurements
- Practical aspect: This part of the paper presents a computer simulation performing classification of the proposed gestures using a German speech database.

Theoretical Aspect

As outlined in [1], in a speaker-listener scenario speech production and speech perception is synchronized by the articulatory rhythm. In speech production the articulatory rhythm organize the timing of the articulatory gestures. In speech perception the articulatory rhythm segments the speech according the timing of the articulatory gestures. The articulatory rhythm is defined by entrained theta oscillations representing the timing of the mouth open-close gesture and by the embedded faster gamma oscillations representing the timing of the embedded phone gestures [2]. These embedded phone gestures are the elementary articulatory gestures postulated above. It is accepted widely, that the mouth open-close gesture relates to a syllable, yet the nature of the phone gestures is not clear. Traditionally it is assumed that each phone gesture represents a single phoneme defined by manner and place features. Due to the embedding mechanism of the gamma oscillation as described in [2], and the dynamics observed in [4] I argue, that an elementary phonetic gesture has to be regarded in the context of the temporal structure of a syllable and not as a stream of gestures representing phonemes.

Practical Aspect

The simulations presented are in the spirit of those described in [3] using the framework of theta and embedded gamma oscillations. Aiming to build a speech recognition system a complete set of elementary phonetic gestures is extracted from a phoneme/syllable segmented database. Using the framework [5] classification experiments of those gestures are presented.

References

[1] H. Höge, "Human Feature Extraction - The Role of the Articulatory Rhythm," proc. ESSV2016
[2] A.L. Giraud, D. Poeppel, "Cortical oscillations and speech processing: emerging computational principles and operations' Nat. Neurosc.: 15(4) pp. 511-517, 2015
[3] A. Hyafil, L. Fontolan, C. Kabdebon, B. Gutkin, and A. Giraud 'Speech encoding by coupled cortical theta and gamma oscillations', DOI: 10.7554/eLife06213, 2015
[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," Modeling of Phone Features for Phoneme Perception," proc. ITG 2016