HUMAN FEATURE EXTRACTION
THE ROLE OF THE ARTICULATORY RHYTHM

Harald Höge

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

Abstract: Neuro-physical investigations [1] hint to a new paradigm for feature extraction not used in ASR. This paradigm is based on synchronized brain to brain oscillations, active during speech production and speech perception. This mechanism leads to an evolving theory, the author calls the ***Unified Theory of Human Speech Processing*** (***UTHSP***). The core elements of this theory are ***the articulatory rhythm*** and ***the articulatory code***. Speech is produced by activating a sequence of ***articulatory codes***. Each code is transformed to an articulatory gesture steered by entrained gamma and theta oscillations called ***the articulatory rhythm***. During each cycle of the rhythm, an articulator gestures is generated. During perception of speech, the articulatory rhythm of the speaker is reconstructed in the brain of the listener. In the cortex, the stream of spectro-temporal features delivered by the midbrain is aligned to phrases, syllables and phones steered by the articulatory rhythm. During each cycle of the rhythm, the aligned spectro-temporal features are integrated and transformed to a bundle of articulatory features. Each bundle generated in a cycle describes a *cycle-gesture*. In phonetics, each phoneme is described by a *phone-gesture*. The cycle-gestures seem to have another structure than the phone-gestures. Thus, the relation between the cycle-gestures and the related phonetic units is unknown. Human feature extraction is finalized by transforming each bundle of articulatory features to an articulatory code as used in speech production. Based on the UTHSP, an architecture for mimicking the extracting of human features is presented.

1 Introduction

Within the last 20 years, tremendous progress has been achieved in automatic speech recognition (ASR) leading to word error rates (WER) coming closer to the performance of human speech perception (HSP). For measuring progress, benchmark corpora are used representing realistic scenarios in verbal communication. The switch board corpus containing spontaneous telephone conversations is such a corpus, where an ASR-WER of 6.7% has been achieved recently [3]. For this corpus, a HSP-WER of 4% has been measured [2]. Thus, the gap between HSR and ASR becomes closer. The performance of a LVSCR systems depends on its acoustic model, on its language model and on the coupling mechanism between both. In most benchmarks the acoustic models and the language models are tuned to the benchmark corpora to achieve optimal performance. Thus, the comparison between ASR and HSP on these LVSCR tasks is problematic, because human perception is not tuned to specific corpora. A better benchmark approach is the concept of speech intelligibility [4], where phoneme error rates (*PER*) measured from nonsense syllables are compared. This method focuses on the performance of the acoustic model and on features, because no language model is involved[1]. The method to determine intelligibility is based on Fletchers work [5], who measured a *PER* of less than 2% for American phones. This result was achieved for 'clean' speech (no noise, no reverberations, and no band limits). Currently the TIMIT database, which is labeled manually into phonemes,

[1] The author is convinced, that the performance of language models is comparable to the performance achieved by humans (Performance is measured in terms of entropy). Thus, the weakest part in ASR is the acoustic model.

is used as a benchmark corpus for measuring *PER*. On this corpus, a *PER* of 15.7% has been achieved recently [6, (2014)]. Yet the TIMIT corpus is much too small to train acoustic models as described in [3], where very large corpora are needed (e.g. Switchboard: 2000hrs, 543 speakers). Thus, the gap in on intelligibility between HSP and ASR is an open issue. Nevertheless, to close further the gap in performance, it is worthwhile to study human speech processing. Another point of view is the aspect of energy consumption, which is an emerging topic in ASR. As minimizing energy is a principle of evolution, models of HSP can lead to lower energy consumption for ASR.

In ASR, several methods of feature extraction mimicking human speech processing have been implemented already. Examples are

- the Mel Frequency Cepstral Coefficients (MFCC) [7] based on the property of the cochlear
- spectro-temporal features [8] as the Perceptual Linear Prediction (PLP) [9] based on the features processed in the inferior colliculus [10]
- processing in streams [11] performed on the auditory pathway [10]
- articulatory features [12] processed in the belt of the auditory cortex [13]

These methods lead to improvements in ASR especially in noisy environments. The breakthrough to human performance has not been achieved yet.

In recent years, large progress in understanding feature processing along the auditory pathway ranging from the cochlea via the midbrain to the cortex has been achieved [10]. Feature processing is a process, where the auditory signal delivered by the hair cells is transformed in several steps along the auditory pathway. The main steps in transformation are the extraction of spectro-temporal features extracted in the inferior colliculus located in the midbrain [10], and articulatory features extracted in the belt of the auditory cortex [13]. Whereas the nature of the spectro-temporal features is sufficiently modelled [8], the nature of articulatory features is still an open issue. Areas of interest for exploring articulatory features are the surrounding neuronal complexes (*nuclei*) of the auditory cortex (A1), and the surrounding nuclei of the sensory motor cortex (SMC). Progress in exploring the functionality of the cortex is fast, as more and more tools are available improving the precision to measure neuronal activity. Yet, as described in chapter 2, the tools available are not sufficient to derive algorithms for feature extraction, but allow to formulate the following two hypotheses for feature extraction:

H1: Oscillations play a dominant role in steering neuronal processes in speech production and speech perception. The author calls the oscillations involved ***the articulatory rhythm***. In speech perception, the articulatory rhythm of the speaker is reconstructed in the listener's brain. In speech production, the articulatory rhythm steers the movements of articulators forming articulatory gestures. In speech perception, the articulatory rhythm controls the process of feature extraction, where spectro-temporal features are transformed to ***articulatory feature*s**.

H2: Basis of human communication are articulatory gestures, which are defined by a code the author calls ***the articulatory code***. In speech production, the sequence of articulatory gestures is generated from sequences of that code. In speech perception, the articulatory features are transformed to a code, which is identical to the articulatory code. Thus, in HSP the articulatory codes of a speaker are mapped to the articulatory codes of the listener.

The two hypotheses are extracted from recent findings in neuroscience. The core elements of the hypotheses are the articulatory rhythm and the articulatory code described in chapter 3. In chapter 4, an architecture of a system simulating human feature extraction is presented.

2 Neuronal Functionalities and their Measurement

The functionality of a single neuron, i.e. the relation between its input and output, is well modelled by the physical relations based on the flow of ions [30]. Due to the electrical potential within a neuron, generated by the ion flow, a neuron can be set to a 'state', where it emits a train of electrical pulses. These pulses are transported via dendrites to other neurons, where they steer as input the ion flow. The algorithm modeling the ion flow are computational too expensive to simulate the functionality of complexes of the neurons in the cortex. Alternatively, the functionality of neurons is modelled by simpler models describing the relation between ingoing and outgoing information contained in the electrical pulse trains [30]. The information is transported by two kinds of codes: the ***rate code*** (number of spikes per sec), and the ***spike code*** (temporal position of spikes). In the following, examples of the codes related to feature extraction are given.

The **rate code** is used for representing the value of features. E.g. for spectro-temporal feature the value is given by the value of the modulation spectrum extracted for a specific frequency and window. Similarly, the value of the rate code of articulatory features can be interpreted as the probability, a specific feature is observed[2]. The rate code can be measured by invasive or non-invasive methods. Invasive methods measure the potential (in the range of ± 100mV) of spikes use micro-electrodes (10 µm needles). Yet these methods are problematic due to ethics problems and limited number of neurons, which can be measured simultaneously[3]. Non-invasive methods (functional MRT, PET, EEG) are mostly used in neuroscience. The resolution of these methods is far away to measure the activity of single neurons[4]. Neuronal activities are observed, if many neurons spike simultaneously. 'Sparse activities' are not detected.

The **spike code** describes the timing of processes implemented neuronally for specific tasks. The timing of the spikes can be observed directly by invasive methods. Using non-invasive methods, the spike code can only be measured, when many neurons spike synchronous. The functions of the spike code – relevant for feature extractions are: measurements of delays, setting clocks, and producing delays. An example of measurements of delays is given by acoustic feature processing in the olive complex [10]. The binaural difference of the timing of spikes is used for detecting the direction of sounds. An example for setting clocks are the oscillations observed in the brain. The oscillations can be described by a spike code characterizing the phases and amplitude of the oscillation. Such spike codes are produced by special neurons [29]. The delay caused by the transmission of electrical spikes via the dendrites is also a functional element. Different delays of the same output of a neuron lead to *receptive fields* [10], on which operations as convolution can be performed neuronally[5].

From the discussion above it can be concluded, that the invasive and non-invasive measurements available are not accurate enough, to decipher the functionality of the neurons involved in feature extraction. As shown in chapter 4, 'intuitive engineering' can be applied, to simulate human feature extraction. The 'correctness' of the simulation can be checked by consistencies to neuronal activities, to psycho-acoustic findings as described in [5,21], and to principles of evolution (e.g. minimal energy consumption).

---

[2] This interpretation is equivalent to the functionality of NNs used in ASR (e.g. the estimation of the emission probability of a state of a phoneme)

[3] Concerning the number of neurons progress in this field is rapid. Current micro-electrode arrays contain about 4000 micro-electrodes (64x64 pats) at a sampling rate of about 8kHz. Larger arrays are in development.

[4] Current resolution of MRT is about 2 mm^2 per pixel relating to the activity of 200 000 neurons.

[5] convolutional NNs as used in ASR simulate those operations, yet the delay is constrained to the fixed durations of frames (e.g. 10ms). The length of the dendrites can be optimized for the tasks.

3 The Unified Theory of Human Speech Processing

The Hypotheses H1 and H2 are part of an evolving theory the author calls the ***United Theory of Human Speech Processing*** (***UTHSP***) [6]. This theory treats speech production and speech perception as processes, which are interweaved neuronally. Both processes use common principles in organization [26] and even common neuronal complexes [25,15]. The articulatory rhythm and the articulatory code, the core elements of the UTHSP, are treated in the following two sections.

3.1 H1 – the Articulatory Rhythm

The movements of many human (motoric) actions as walking, eating (chewing), and speaking are quasi-rhythmic. The movements are steered by neuronal oscillations – *the action-oscillations*. In parallel, sensory systems as feeling and hearing analyze these actions. In these processes, oscillations – *the sensory-oscillations* - are involved. It is hypothesized, that the action-oscillations and the related sensory-oscillations are synchronous. The neurons producing the synchronous oscillations are either located in the same brain as in walking, or located in different brains (*brain-to-brain synchronization*). The latter case is given in a speaker-listener scenario, where oscillations can be interpreted as a motor control to perform 'predictive' speech perception [14,28]. The neuronal origin of the oscillations is unclear. It is assumed, that the oscillations are generated in a hierarchical manner with master clocks located in the hippocampus. These master clocks are projected to the superficial layer of the nuclei, where they are transformed to oscillations, which are adapted (*entrained*) to the tasks of the neurons involved [29]. Modeling the oscillations as sinus waves, the positive part and the negative part of a wave can be mapped to two state, which alternate within each cycle. The states can be interpreted as modes of a cyclic *attention* as treated in the *dynamic attending theory* [16,14]. Attention coordinates the information flow between different processes. Slow and fast oscillations are used to coordinate processes with different timing. The function of the two states can be modelled by an alternating process of inhibition and excitation of a neuron [1]:

- an input state**,** when the neuron is open for integrating incoming spike trains
- an output state**,** during which the neuron emits spike trains and during which input processing is stopped.

The states are equal in duration. Consequently, the incoming spike trains are blocked for half of a cycle. In the following, this kind of processing is called the ***cyclic attention processing***.

The articulatory rhythm is defined as the actions of different oscillation involved in speech production and speech perception. These oscillations can be described by their **nature** and its **role**. The nature describes the relation between the oscillations and phonetic units. The role describes, how the oscillations control the process of feature extraction and articulation.

The **nature** of the articulatory rhythm is the same for speech production and speech perception. The articulatory rhythm is composed of three kinds of oscillations. Each cycle of such an oscillation corresponds to a specific articulatory/phonetic unit as well in perception [31,19] as in production. The oscillations and the related units are: delta oscillations [1- 3 Hz] ↔ phrasal units, theta oscillations [4 - 8 Hz] ↔ syllables, and gamma oscillations [25 - 45 Hz] ↔ phones. Further the oscillations are nested, i.e. the slower oscillations influence the faster oscillations. This hints to a hierarchical concept in speech processing. In speech perception, the listener's articulatory rhythm must be synchronized with the articulatory rhythm of the speaker by the

[6] In [32] a more ASR oriented but in intention similar theory – the Unified Approach for Speech Synthesis and Recognition (UASR) - is presented

process of *entrainment* done in the ventral part of the sensory motor cortex (vSMC). Entrainment is performed using *edge features* [1,29], which are assumed to be extracted in the belt of the A1[7].

The **role** of the articulatory rhythm is different in speech production and speech perception. In speech production, the articulatory rhythm steers the timing of the movements of the articulators [18]. Neuronally the commands for movements are generated in the vSMC. In speech perception, the articulatory rhythm must perform two roles: segmentation of the stream of spectro-temporal features into phonetic/articulatory units [1], and control of features extraction based on cyclic attention processing as described above. Both roles are performed in a single process, where the stream of spectro-temporal features delivered by the midbrain are transformed to articulatory features by neurons located in the STG [13,22]. The nature of the articulatory features is treated in the next section 3.2.

3.2 H2 - the Articulatory Code

Phoneticians have hypothesized, that speech production can be modelled as a sequence of articulatory gestures [17]. The gestures are described by distinctive features characterized by manner and place features [20]. Nowadays this hypothesis is supported by neuronal activities in the vSMC [18], yet the nature of the gestures seems to be different as assumed in [17]. The origin of this difference is caused by the involvement of the articulatory rhythm in generating articulatory gestures. Due to [1] during each cycle of theta oscillation a syllable, and during each gamma cycle a phone is produced. From the articulatory view, a phonetically defined syllable is related to a *syllable-gesture*, where the jaw opens and closed. Embedded in a *syllable-gesture* are faster *gestures*, the author calls ***cycle-gestures***. Within each cycle of a gamma oscillation such a gesture produced. In the phonetic theory [17] each gesture represents a **single** phoneme called a phone-gesture. The author hypotheses that the phone-gestures are not equivalent to the cycle-gestures. This is motivated by assumption that consonant cluster as {/s/, /t/} with the same place of articulation are generated during a single cycle of a gamma oscillation. Thus, it is assumed that clusters of phones are generated by a single cycle-gestures fitting better to a continuous rhythm, which can be entrained more easily.

During speech perception, articulatory features are extracted. The articulatory features are defined by that bundle of features related corresponding to a cycle-gestures. Due to [13] manner and place features have been detected in the STG. Thus, the articulatory features are composed by a bundle of distinctive features characterizing a cycle-gesture. Due to hypothesis H2 the articulatory features are transformed to an articulatory code, which is identical to those to activate cycle-gestures. As the nature of the cycle-gestures is not deciphered, this holds also for the ***articulatory code***. This problem is discussed also in [33], addressed as an old, yet unsolved problem. The transformation of the articulatory code seems to be located in the vSMC.

4 Implementing Human Feature Extraction

This chapter comes up with an architecture of a system mimicking human feature extraction. The architecture of the system is presented in fig. 1. The functionally of the building blocks are based on the hypotheses H1 and H2. Due to discussion of the chapters 2 and 3, the exact functionally of the building blocks as performed by human feature extraction is not known. Thus, the implementation of the building blocks is not straight forward and has to be done still by 'intuitive engineering'. The resulting functionality of the system must be checked against the functionality, as derived from knowledge sources as described at the end of chapter 2.

[7] In lip reading edge features are generated in the visual cortex and are used to support the articulatory rhythm [27]

The extraction of spectro-temporal features is well explored and implementations are proposed [8,12]. Implementation of the building block 'articulatory feature extraction' has been described in [12] based on distinctive features modeled statistically [12, 21]. The statistical model is compatible with psycho-acoustic findings. The biggest issue is the implementation of the articulatory rhythm based on edge features. This issue opens a new field in neuronal implementation of entrained oscillations. A first implementation for speech perception is presented in [29]. This approach simulates the articulatory rhythm by simulating the neuronal flow of ions. This approach is computational very expensive. Below a less computational expensive implementation is proposed. Feature extraction described in [12,13] is done without cyclic attention processing. But this kind of processing can be implemented straightforward, because only the kind of temporal windows used must be modified. As the nature of the articulatory features is not known yet (see chapter 3), the transformation of the spectro-temporal features to articulatory features and the transformation to the articulatory code is still an open issue.

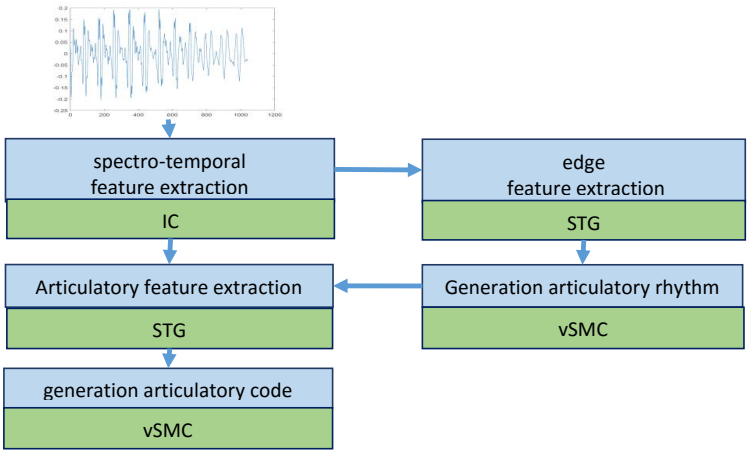

figure 1 – Systemarchitecture of human feature extraction. Five buildings blocks are involved in human feature extraction. Neuronal areas: IC=Inferior colliculus (midbrain), STG= superior temporal gyrus (belt of auditory cortex), vSMC=ventral sensory motor cortex

4.1 The Generation of the Articulatory Rhythm

In ASR segmentation of speech into phonetic units, is an old problem implemented by algorithms called *search*. In the beginning of ASR, it was tried to solve the search problem using decisions, based on acoustic features (bottom-up approach). But this approach was not successful. Around 1980 a breakthrough has been achieved using the method of dynamic programming for recognition of continuous speech [23]. Using additionally sophisticated language models, and applying the framework of HMMs [24], this kind of search is still used nowadays (top-down approach). But this method has two main draw-backs. First it uses a first order Markov model, assuming, that the features used in the search algorithms are statistic independent. Yet temporal adjacent feature vectors used in ASR are strongly statistic dependent. Second, the top-down approach is computational very expensive leading to high energy consumption.

Astonishingly evolution has solved the search problem with a bottom up approach. This approach does not rely only on decisions on acoustic features, but this approach integrates the process of speech production into the process of feature extraction as described in chapter 3.

The implementation of the buildings blocks *generation of edge features* and *generation of articulatory* is inspired of [29], especially the implementation of the edge features. For implementing the articulatory rhythm, the starting point is the implementation of the entrained theta

oscillations modelling syllables. The oscillation can be described by sin-wave with variable amplitude $a_\theta(t)$ and the phase $\varphi_\theta(t)$:

$$\theta(t) = a_\theta(t)\sin\big(\varphi_\theta(t)\big) \quad (1)$$

The edge features drive $a_\theta(t)$ and $\varphi_\theta(t)$. Thus, a transformation from the edge features to the amplitude and the phase must be found. The performance of the match between $\theta(t)$ and the rhythm of syllables can be determined with speech databases, where the position of the syllables is labelled. In [29] the TIMIT database was used, and a good match was achieved. This result hints, that phonetically defined syllables are consistent with cycle syllable gestures.

The γ - oscillations have the same structure as (1), but they must match to cycle gestures and they must be nested into the θ- oscillations. As mentioned above, the phonetic code of the cycle gestures is not known. A hint, how the cycle gestures are organized is given by the nested structure: each cycle of a γ- oscillation corresponds to a cycle-gesture and a single θ- cycle relates to *n* γ- cycles, where *n* is the number of cycle gestures building a syllable. In [1] for *n* a value 4 was observed. If this is a general rule, perhaps the nature of the cycle gestures and the articulatory code can be deciphered straight forward.

5 Conclusion

Human feature extraction is an interweaved process of speech production and speech perception. Both processes are implemented neuronally by the same mechanisms described by the articulatory rhythm and the articulatory code, which are part of a *United Theory of Human Speech Processing*. Based on this theory an architecture implementing human feature extraction system is presented. An open issue is a phonetic description of the set of cyclic gestures.

6 References

[1] 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.

[2] R.P. Lippmann: *Speech Recognition by Machines and Humans*. In *Speech Communication,* 22 (1), pp. 1-15, 1997.

[3] G. Sam, T. Sercu, S. Rennie, and J.K. Hong-Kwan: *The IBM 2016 Conversional Tele phone Speech Recognition System.* In *Proc. Interspeech*, 2016.

[4] European Committee for Electrotechnical Standardization: *Objective rating of speech intelligibility by speech transmission index.* In European standard IEC60268-16: *Sound system equipment – Part 16,* 2011.

[5] 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.

[6] L. Toth: *Combining time- and frequency-domain convolution in convolutional neural network-based phone recognition*. In *ICASSP*, pp. 190-194. 2014.

[7] S. B. Davis, and P. Mermelstein: *Comparison of parametric representation for monosyllabic word recognition in continuously spoken sentences*. In *IEEE Transactions on Acoustic, Speech and Signal Processing*, vol. 28, no. 4, pp. 357–366. 1980.

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

[9] H. Hermansky: *Perceptual linear predictive (plp) analysis of speech*. In *Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752. 1990.

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

[11] N. Mesgarani, S. Thomas, and H. Hermansky: *A Multistream Multiresolution Framework for Phoneme Recognition*. In *Proc. INTERSPEECH*, pp. 318–321. 2010.

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

[13] N. Mesgarani, C. Cheung, K. Johnson, and E.F. Chang: *Phonetic Feature Encoding in Human Superior Temporal Gyrus*. In *Science*, 343(6174), pp.1006–1010. 2014.
[14] B. Morillon, A. Troy, T.A. Hackett, Y. Kajikawa, E. Charles, and C. E. Schroeder: *Predictive motor control of sensory dynamics in Auditory Active Sensing*. In *Curr Opin Neurobiol*., 31, pp230–238. April 2015.
[15] C. Cheung, L.S. Hamilton, K. Johnson, and E.F. Chung: *The auditory representation of speech sounds in human motor cortex*. In *Elife*, 12577, 2016.
[16] E.W. Large, and M.R. Jones: *The dynamics of attending: How people track time-varying events.* In *Psychol Rev.*, 106, pp.119–159. 1999.
[17] C.P. Browman, and L. Goldstein: *Articulatory Gestures as Phonological Units*. In: *Haskins Laboratories Status Report on Speech Research*, 99, pp. 69–101. 1989.
[18] 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.
[19] C. Kayser, R.A. Ince, and S. Panzeri: *Analysis of Slow (Theta) Oscillations as a Potential Temporal Reference Frame for Information Coding in Sensory Cortices*. In *PLoS Comput Biol*,8(10), e1002717, doi:10.1371/journal.pcbi.1002717. 2012.
[20] P. Ladefoged, and K. Johnson: *A Course in Phonetics*. In: *Wadsworth Cengage Learning*, 7th Edition, Boston, 2015.
[21] H. Höge: *On the Nature of the Features Generated in the Human Auditory Pathway for Phone Recognition.* In *Proc. Interspeech*, Dresden 2015.
[22] M. Steinschneider, K.V. Nourski, H. Kawasaki, H. Oya, J.F. Brugge, and M.A. Howard: *Intracranial Study of Speech-Elicited Activity on the Human Posterolateral Superior Temporal Gyrus*. In *Cerebral Cortex*, V. 21, N 10, pp.2332-2347. October 2011.
[23] H. Ney: *The use of a one-stage dynamic programming algorithm for connected word recognition*. In *IEEE Trans. Acoustic, Speech and Signal Processing*, 32, pp.263-271. 1984.
[24] X.D. Huang, Y. Ariki, and M.A. Jack: *Hidden Markov Models for Speech Recognition.* In *Information Technology Series*, Edinburg University Press, 1990.
[25] F. Pulvermüller, M. Huss, F. Kherif, F. Martin, O. Hauk, and Y. Shtyrov: *Motor cortex maps articulatory features of speech sounds*. In *PNAS*, Vol.103, pp. 7865-7870. May 2006.
[26] S. Evans, and M.H. Davis: *Hierarchical Organization of Auditory and Motor Representations in Speech Perception: Evidence from Searchlight Similarity Analysis*. In *Cerebral Cortex*,25, pp. 4772–4788. December 2015.
[27] H. Park, C. Kayser, G. Thut, and J. Gross: *Lip movements entrain the observers' low-frequency brain oscillations to facilitate speech intelligibility*. eLife, 5, e14521. DOI: 10.7554/eLife.14521 pp.1-17, May 2016
[28] A. M. Liberman, and I. G. Mattingly: *The motor theory of speech perception revised*. In *Cognition*, 21, pp. l-36. 1985.
[29] 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
[30] W. Gerstner, and W. Kistler: *Spiking Neuron Models, Single Neurons, Populations, Plasticity*. In: Cambridge University Press, Cambridge UK, 2002
[31] O. Ghitza: *linking speech perception and neurophysiology: speech decoding guided by Cascaded oscillators locked to the speech rhythm.* In *Frontiers in Psychology - Auditory Cognitive Neuroscience*, V 2, Article 130, pp.1-13. June 2011.
[32] M. Eichner, M. Wolff, and R. Hoffmann: *A unified approach for speech synthesis and recognition using stochastic Markov graphs*. In *Proc. ICSLP*, vol.1, pp.701-704. Bejing 2000.
[33] D. Conant, K.E. Bouchard, and E.F. Chang: *speech map in the human ventral sensory-motor cortex*. In *current opinion in neurobiology*,24, pp. 63-67. 2014.