Phonetic-Phonological Feature Emerges by Associating Phonetic with Semantic Information — a GSOM-Based Modeling Study

Mengxue Cao[1], *Aijun Li*[1], *Qiang Fang*[1], *Bernd J. Kröger*[2]

[1]Laboratory of Phonetics and Speech Science, Institute of Linguistics,
Chinese Academy of Social Sciences, Beijing, China
[2]Neurophonetics Group, Department of Phoniatrics, Pedaudiology, and Communication Disorders,
Medical School, RWTH Aachen University, Aachen, Germany

mengxuecao@outlook.com

Abstract

Born with an innate neural architecture built specially for language learning, young children have the ability to distinguish sounds in a variety of languages. As they are exposed to native language environment, perceptual reorganization occurs, and native language system gradually establishes. Phonology knowledge, which is language-specific, emerges during this process. In this study, based on the further developed Interconnected Growing Self-Organizing Maps (I-GSOM) model, we present a series of computational modeling experiments which simulate the early language acquisition process for young children. A universal model that has the ability to distinguish phonemes in different languages (two in our case) is built based on English and Standard Chinese data. The native language learning is simulated based on Standard Chinese data with both phonetic and semantic inputs for children between 12 and 18 months old. Experiment results show that the conceptual based top-down process contributes to the reorganization of phonetic knowledge, and phonetic–semantic association helps the emergence of phonetic-phonological knowledge. It can be hypothesized from these findings that the phonetic-phonological interface does not appear as a clean cut within the speech processing system but as a broader zone located between sensorimotor and semantic processing.

Index Terms: child language acquisition, phonetic–semantic association, neural model of speech processing

1. Introduction

The mechanism of language acquisition is still an unrevealed mystery in modern science. Young children, however, are able to learn their mother tongue without making special efforts. The development of children's language abilities follows a general path as described by Kuhl [1], where young children, with an innate ability necessary for language acquisition, generally tweak their language system to the language they are exposed to. The basic theory behind this is the Chomskian *Principles and Parameters (P&P)* model. With universal *principles*, young children are able to distinguish different sounds (*or* phonetic units) in both native and nonnative language on phonetic level [1, 2, 3, 4]. Meanwhile, by categorical perception [1], young children are able to group perceptually distinct sounds into the same category. These abilities change at 10–12 months old as children gradually explore their language-specific *parameters* with exposure to their mother tongue. Taking Japanese children as a typical example, their perceptual ability change of distinguishing /r/ and /l/ [5] indicates a decline in children's ability of distinguishing nonnative phonetic contrasts as their native language system develops. This process is termed as perceptual reorganization [6].

In parallel with phonetic learning, young children are relating sound patterns to meanings. During language acquisition, apart from auditory information, children receive various kinds of stimuli simultaneously through their interactions with communication partners and the surrounding environment, such as visual information, tactile information, and olfactory information. By observing and feeling, children begin to understand the property of an object or the effect of a movement. Furthermore, as demonstrated by the triangular attention model [7], children learn to relate sounds they hear to objects or movements they see or feel during direct face-to-face communication. Therefore, phonetic information (i. e., pronunciation of a word) and semantic information (i. e., meanings of the sound) are linked together during language acquisition.

Behavior studies have limitations in revealing the perceiving path in details. With the help of neurocomputational techniques, we want to explain the emergence of phonetic-phonological features following the idea proposed in [7]. In previous study [8], we have shown the importance of associating auditory information with semantic information during language acquisition by using computational modeling based on I-GSOM model. In this study, we will further demonstrate our idea by modeling the process of perceptual reorganization. The target language (i. e., mother tongue) of our model (the "child") is set to Standard Chinese. First, we build a universal model, which has the "universal perceiving ability", based on English and Standard Chinese data. Then our hypothesis is that with the exposure to native language (i. e., Standard Chinese), language-specific phonetic-phonological features (e. g., consonantal, vocalic and tonal features) can be established based on phonetic–semantic association.

2. Materials

We use both English and Standard Chinese data in this modeling study. English data is used in combine with Standard Chinese data to train the universal model, so that the model has the ability to perceive phonetic distinctions in both English and Standard Chinese. This initial model acts as the innate universal perceiving ability of children. Based on that, Standard Chinese data is further used as native language exposure for the language acquisition process.

2.1. English data

The English data is in audio form only, since the initial distinguishing ability is based on information from phonetic level [1]. The recording script is mainly comprised of minimal pairs of English RP, including vowel pairs of [ɑː]–[ʌ], [iː]–[ɪ], [uː]–[ʊ], and consonant pairs of [b]–[p]/[p^h], [d]–[t]/[t^h], [g]–[k]/[k^h]. Distinctive features in vowel pairs mainly include duration and quality [9]. Although English does not use "aspiration/non-aspiration" as distinctive feature, we still transcript them phonetically in this study. In total, 350 words, consisting 150 minimal pairs (e. g., *beat–bit*) and 50 redundant words (e. g., *but*, *see*), are listed. The recording is done by a 22 years old female British speaker. Each word is read once, and the sampling rate for the recording is 44.1 kHz.

2.2. Standard Chinese data

The Standard Chinese data consists of two parts, which are word corpus and audio recordings. The corpus we use is the *Mandarin Multimedia Child Speech Corpus (CASS_CHILD)* [10]. This corpus is based on a longitudinal recording (with one-month interval) of speech production of 23 Mandarin-speaking children starting from one year old all the way to their four years old. The interactive communications between children and caretakers are recorded in both audio and video mode. Recordings are transcribed and annotated by experts. In this study, we focus on restrict syllable types and word classes (*see section 2.3.1* for details) of the input speech from caretakers, so we extract words which fit into our requirements together with their frequency counts, month by month, and build the word data set for this study.

Although the corpus is excellent for child language development research, the recording quality is not good enough for phonetic analysis, due to highly-overlapped speech and indifferent sound quality of portable microphones clipped below the collar of both caretakers and children. Therefore, a separate audio dataset is recorded by a 22 years old female Standard Chinese speaker, with sampling rate at 22.05 kHz, and mapped to the words we selected from the *CASS_CHILD* corpus.

2.3. Training Sets

2.3.1. Word data set

By analyzing the caretakers' input speech in the *CASS_CHILD* corpus for the period that children grow from 12 to 36 months old, a significant change in the frequency pattern of consonants, vowels and tones is observed at the age of 18 months old. This indicates a change point of caretakers' input speech pattern. A possible explanation for this is that at around 18 months old, children are capable of relating sounds to meanings [11, 12, 13] and perceptual reorganization has formulated their language system native enough [6, 14], so that more communicative interactions between children and caretakers occur. Therefore, in this study, as the first step of our series study, we use caretakers' input speech data of the period that children grow from 12 to 18 months old as the source of our training data.

In the corpus, we only focus on nouns, verbs, and adjectives with *CV* syllable type that consist of consonants including [p], [p^h], [t], [t^h], [k], [k^h] and vowels including [a], [i], [u], [o], [ɤ]. Words fit into this requirement are extracted and the word data set is built. In total, 49 high-frequency words (e. g., *big–[ta4]*, *father–[pa4]*, and *rabbit–[t^hu4]*) ranked by occurring frequency are selected as the vocabulary to be learned for the model.

2.3.2. Audio data set

Audio recordings of both English and Standard Chinese are converted to 16 kHz sampling rate, chopped into consonant and vowel segments, and their pitch contours are extracted. For Standard Chinese data, the syllable–word (i. e., sound–word) relations are built according to the mappings between audio recordings and word corpus as mentioned in *section 2.2*. Voicing in plosive, such as [b], [d] and [g], is common in continuous speech of Standard Chinese, so audio tokens of these voiced plosive are also considered as valid pronunciations of words consisting of voiceless plosive. The syllable construction rules are kept while mapping sounds to words. Therefore, one word is linked to a specific meaningful sound sequence (i. e., a realization of the pronunciation of that word), and the sound sequence (i. e., its syllable structure) is linked to a specific combination of meaningless elements of consonant, vowel and pitch contour. The idea of this dual level representation is inspired by the *Duality of Patterning* theory [15].

2.3.3. Feature representations

As connectionist proposed, knowledge should be represented by distributed representations [16]. Therefore, we represent the meaning of words by a set of semantic features and the sound of words by a set of phonetic (i.e auditory-based) features.

The semantic feature of each word is developed by five research students with linguistic background who speak Standard Chinese. Instructions are given to them that their task is to describe the observable and perceptible features of the word presented to them based on sensations such as visual, auditory, tactile and olfactory, by using simple descriptions that can be accepted by young children. In total, 311 features (e. g., *has two legs*, *is circular*) are developed. To build the training data, each word is represented by 311 feature vectors using binary coding as described in [8].

The auditory feature of each word is represented by its syllable elements: consonant, vowel and tone. Features of consonants and vowels are represented by spectral state neuron feature map as described in [8]. The time step for neurons is set to 2 ms for consonant representations and 10 ms for vowel representations as human brain uses different timing windows to process transient and steady signals [17, 18]. Feature of tones is represented by the frequency values of pitch contour. Pitch durations are normalized, and pitch values are normalized by z-score normalization between the English speaker and Standard Chinese speaker. In total, 2,948 neurons are used to represent consonantal feature, 1,144 neurons are used to represent vocalic feature, and 10 pitch points are used to represent tonal feature.

3. Methods

In this study, we further developed the I-GSOM model in both architecture and algorithm perspectives. Major changes are presented in *section 3.1* and *section 3.2* respectively.

3.1. Description of the Model

The current model consists of a Semantic Map, a Phonetic Map, a Consonant Map, an Vowel Map and a Tone Map (*see* Figure 1). All these maps except the Phonetic Map are GSOM based self-organizing neural network. The Phonetic Map is specially designed as a phonetic knowledge processing unit, which records the combination rules of consonants, vowels and tones (i. e., the syllable constructing rule) and the associations be-

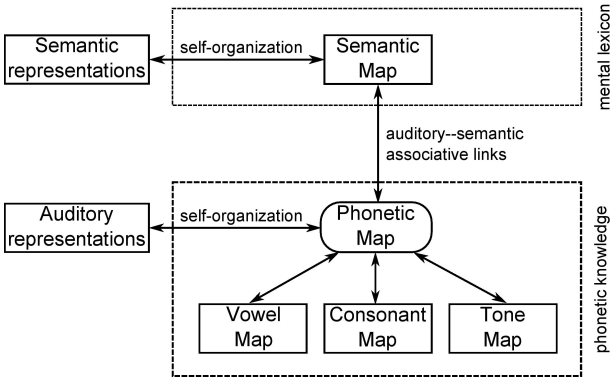

Figure 1: *The structure of the modified I-GSOM model.*

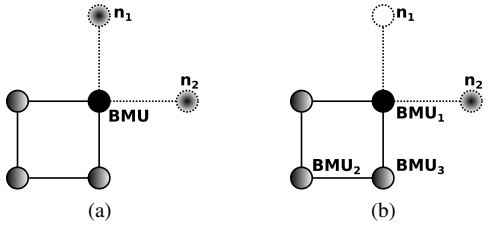

Figure 2: *Different growing strategies. (a) GSOM adds new nodes to all available positions (n_1 and n_2). (b) Optimized growing strategy only adds one new node at the best selected position (n_2) at a time. BMU (equally BMU_1) represents the best matching unit of the training token, and BMU_2 and BMU_3 represent the second-best and third-best matching unit of the training token respectively.*

tween each phonetic combination (i. e., syllable structure) and semantic representation in the Semantic Map (i. e., meaning of a word). In this study, we model not only the bottom-up path but also the top-down path. Therefore, the links between each map are truly bidirectional and maps co-activate with each other.

Language acquisition is not a one-way bottom-up process. Top-down process contributes too. "*The top-down or conceptually driven processing works simultaneously and in conjunction with bottom-up or data driven processing to provide a sort of multiplicity of constrains that jointly determine what we perceive*" [19]. Therefore, the semantic knowledge children learned can interactively affect the learning of phonetic knowledge. Moreover, since phonetic–semantic correlations are language specific, it is reasonable to deduce that the top-down process helps with formulating children's native language system.

Each word (i. e., a neuron in the Semantic Map) should have a stable, although not unique, pronunciation. Therefore, syllables linked by the same node of Semantic Map should have similar pronunciations. Based on this logic, the top-down process acts as follows. First, for a neuron in the Semantic Map, its linked syllable(s) in the Phonetic Map is(are) located. Next, a "target pronunciation" (ω_{target}) is calculated for the word by Equation 1, where ω_i represents the weight vector of each linked neuron, and ω_{link_i} represents the weight of each associative link. Then, weight update is applied to corresponding neurons in the Consonant Map, Vowel Map and Tone Map respectively by the rule described in Equation 2, where $\omega_{link_{mean}}$ represents the mean value of the weight of associative links linked to these neurons. Following this procedure, the learned knowledge from top semantic level reformulates the bottom phonetic level, and language-specific perceiving emerges gradually.

$$\omega_{target}(t) = \sum_{i=1}^{n} \frac{\omega_{link_i}(t)}{\sum_{i=1}^{n} \omega_{link_i}(t)} \omega_i(t) \tag{1}$$

$$\omega_i(t+1) = \omega_i(t) + \omega_{link_{mean}}(t) \times \big(\omega_{target}(t) - \omega_i(t)\big), i \in N \tag{2}$$

3.2. Description of the Algorithm

It is pointed out that when applying GSOM to high-dimensional data, the spiral growth problem can arise and leads the network to an inefficient growing path [20]. As a solution provided by [20], a calibrating phase that goes through the training data is applied before the growing phase. During calibrating phase, no growing is allowed and weight update only applies to the BMU node and its direct neighbors. The main aim of the calibrating phase is to get the four initial nodes in the network positioned strategically over the input data space rather than being dragged as a whole around the input space [20]. According to different requirements, several cycles of calibrating phase can be applied. The growth threshold of the growing phase is set by the highest error between the training data and neuron nodes in the network of the last calibrating phase, and therefore, more reliable.

When GSOM grows, new nodes will be added at all available positions around the BMU node (*see* Figure 2a). This kind of growing strategy lacks of clear direction and can increase the number of redundant nodes in the network. Inspired by [21, 22], we propose an optimized growing strategy, that only add nodes to the best selected positions. It compares the distance between each available position and the BMU core nodes, and the node who is the closet to the BMU core is selected as the position for network growing. Meanwhile, as described in [21], we also modified the error distribution function of the GSOM algorithm.

4. Experiments and Results

The experiment consists of two stages, which are pre-learning stage and learning stage. Following the same procedures, five simulations with identical model parameters are performed. From our analysis, all five simulations behave similarly and lead to comparable results. Therefore, we only report on and discuss the results of the first simulation.

4.1. Pre-learning Stage

As *P&P* theory describes, children have innate neural architecture built specially for language learning tasks, so that they are born with universal language perceiving ability. Therefore, at pre-learning stage, we build a universal model by using audio data from both English and Standard Chinese, expecting the model to have the ability to distinguish phonemes from both English and Standard Chinese on phonetic level. The Semantic Map and Phonetic Map is not introduced into this stage, since the universal perceiving ability is based on phonetic level and no semantic knowledge is involved. Two calibrating phases and ten growing phases are applied in this training.

Trained networks and checking results are shown in the left column of Figure 3. As shown in Figure 3a, aspirated plosive and non-aspirated plosive are clearly separated. Small clusters of voiced plosive are located at the mid-left and bottom areas. As shown in Figure 3c, each vowel cluster is clearly separated.

The spatial relation of vowel distributions is consistent with the acoustic vowel space, where "low-high" and "front-back" relations are clearly observed. In the Vowel Map, specifically, low vowels [a], [ɑː] and [ʌ] are located at the bottom area, high-front vowels [i], [iː] and [ɪ] are at the left and upper-left areas, high-back vowels [u] and [o] are at the right and upper-right areas, and [ʊ] and [ɤ] are in the middle areas. Since tonal category is language specific, no phonology-based tonal perceiving can be learned at this stage. We examine the tonal learning result by analyzing the network's representation of tones' auditory features, shown as tone contours in Figure 3e. We find the network can classify tones by their contour features, such that four main clusters emerge as *level*, *rising*, *falling & rising*, and *falling*. Furthermore, in the *falling* area, detailed clusters can be found as *convex falling*, *falling*, and *concave falling*. At this stage, although some small overlaps exist, each phoneme has its central cluster area and are phonetically distinguishable by the model.

4.2. Learning Stage

By exposing to native language space, gradually, children's perceiving ability becomes more language-specific. Therefore, at learning stage, we model the development of children's language system by using caretakers' input speech (Standard Chinese) for children growing from 12 to 18 months old. Semantic Map and Phonetic Map begin to acquire knowledge starting from this stage. Therefore, auditory–semantic links are built, and the top-down processing path is activated as children begin to associate sounds with meanings [11, 12, 13]. Two calibrating phases and fifteen growing phases are applied in this training.

Trained networks and checking results are shown in the right column of Figure 3. As shown in Figure 3b, by exposing to Standard Chinese stimuli, due to the top-down effect from Semantic Map, the perceiving of voiced plosive becomes very weak. Only a few nodes are identified as representing the feature of voiced plosive and they are mixed with non-aspirated plosive in the Consonant Map. Due to the large number of stimuli tokens with [t], [t] takes over the space of [p], and disturbs the perceiving of [p]. The cluster of [k] is developed and forms into as stable region. As shown in Figure 3d, English vowel clusters are merged into Standard Chinese perceiving categories, and the "low-high" and "front-back" relations are clearly kept. As shown in Figure 3f, four clusters of Standard Chinese tones are established, and the three types of *falling* tones are merged into the Tone 4 (falling tone) of Standard Chinese. Therefore, by exposing to native language, perceptual reorganization occurs as a consequence of the top-down effect.

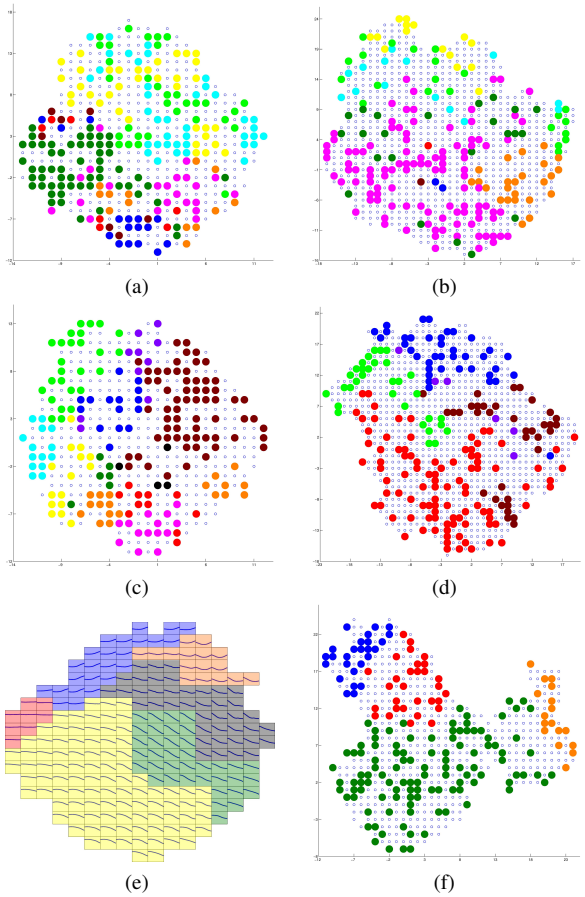

Figure 3: *The trained network structures and checking results of the Consonant, Vowel and Tone Maps at pre-learning stage (left column) and learning stage (right column). In Consonant Maps (top line), blue stands for [b], red for [d], brown for [g], dark-green for [p], pink for [t], orange for [k], light-green for [p^h], light-blue for [t^h], and yellow for [k^h]. In Vowel Maps (middle line), red stands for [a], pink for [ɑː], orange for [ʌ], light-green for [i], light-blue for [iː], yellow for [ɪ], brown for [u], black for [uː], dark-green for [ʊ], purple for [o], and blue for [ɤ]. In Tone Maps (bottom line), in (e), each block represents the tonal feature of that node, shown as the blue curve within it, red shaded blocks stand for level contour, blue for rising, orange for falling & rising, gray for concave falling, green for falling, and yellow for convex falling. In (f), red stands for Tone 1, blue for Tone 2, orange for Tone 3, and dark-green for Tone 4. In (a)–(d) and (f), features represented by small empty nodes are not typical enough as those of the big solid nodes.*

5. Discussion and Conclusions

In this study, we have simulated the early language acquisition process for young children from 12 to 18 months old. A universal model that can distinguish phonetic features in both English and Standard Chinese is built first. Then by exposing to native language (i. e., Standard Chinese), reorganization occurs in consonantal, vocalic and tonal spaces. During this process, we can observe that the semantic-based top-down process helps the perceptual reorganization by dragging phonetically different items into the same category if these phonetic differences do not affect meaning understanding.

This study still has limitations in many aspects. In further studies, we can expand our work to more types of word classes and syllable structures, and more age stages can be modeled. Despite of these limitations, it can be concluded already from the simulation results presented in this study that the phonetic-phonological interface does not appear as a simple cut but as a broader zone comprising self-organizations at the level of phonetic repository as well as at the level of mental lexicon and comprising neural associations between these two main levels.

6. Acknowledgements

We thank Dr. Gao Jun for providing us the *CASS_CHILD* corpus. This study is supported and funded by the NSFC Key Project with No. 61233009, the National 973 Project with No. 2013CB329301, and the CASS Innovation Program.

7. References

[1] P. K. Kuhl, "Early language acquisition: cracking the speech code," *Nature reviews neuroscience*, vol. 5, no. 11, pp. 831–843, 2004.

[2] ——, "Theoretical contributions of tests on animals to the special-mechanisms debate in speech." *Experimental biology*, vol. 45, no. 3, pp. 233–265, 1985.

[3] P. K. Kuhl, E. Stevens, A. Hayashi, T. Deguchi, S. Kiritani, and P. Iverson, "Infants show a facilitation effect for native language phonetic perception between 6 and 12 months," *Developmental science*, vol. 9, no. 2, pp. F13–F21, 2006.

[4] S. E. Brauth, W. S. Hall, and R. J. Dooling, *Plasticity of development*. MIT Press, 1991.

[5] T. Tsushima, O. Takizawa, M. Sasaki, S. Shiraki, K. Nishi, M. Kohno, P. Menyuk, and C. T. Best, "Discrimination of english /rl/ and /wy/ by japanese infants at 6-12 months: language-specific developmental changes in speech perception abilities." in *The 3rd International Conference on Spoken Language Processing, ICSLP*, 1994.

[6] J. F. Werker and R. C. Tees, "Cross-language speech perception: Evidence for perceptual reorganization during the first year of life," *Infant behavior and development*, vol. 7, no. 1, pp. 49–63, 1984.

[7] C. Eckers, B. J. Kröger, and M. Wolff, "Semantic, phonetic, and phonological knowledge in a neurocomputational model of speech acquisition," *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung*, pp. 244–251, 2012.

[8] M. Cao, A. Li, Q. Fang, E. Kaufmann, and B. J. Kröger, "Interconnected growing self-organizing maps for auditory and semantic acquisition modeling," *Frontiers in psychology*, vol. 5, 2014.

[9] P. Ladefoged and K. Johnson, *A course in phonetics*. Cengage learning, 2014.

[10] J. Gao, A. Li, and Z. Xiong, "Mandarin multimedia child speech corpus: Cass_child," in *Proceedings of* 15^{th} *International Conference on Speech Database and Assessments (Oriental COCOSDA)*. IEEE, 2012, pp. 7–12.

[11] D. A. Baldwin, "Infants' contribution to the achievement of joint reference," *Child development*, vol. 62, no. 5, pp. 874–890, 1991.

[12] ——, "Infants' ability to consult the speaker for clues to word reference," *Journal of child language*, vol. 20, no. 02, pp. 395–418, 1993.

[13] D. A. Baldwin, E. M. Markman, B. Bill, R. N. Desjardins, J. M. Irwin, and G. Tidball, "Infants' reliance on a social criterion for establishing word-object relations," *Child development*, vol. 67, no. 6, pp. 3135–3153, 1996.

[14] D. L. Mills, C. Prat, R. Zangl, C. L. Stager, H. J. Neville, and J. F. Werker, "Language experience and the organization of brain activity to phonetically similar words: Erp evidence from 14-and 20-month-olds," *Journal of Cognitive Neuroscience*, vol. 16, no. 8, pp. 1452–1464, 2004.

[15] C. F. Hackett, "The origin of speech," *Scientific American*, vol. 203, pp. 88–96, 1960.

[16] D. E. Rumelhart, J. L. McClelland, P. R. Group *et al.*, "Parallel distributed processing: Explorations in the microstructures of cognition. volume 1: Foundations," *MIT Press, Cambridge, MA*, vol. 2, pp. 560–567, 1986.

[17] A.-L. Giraud and D. Poeppel, "Cortical oscillations and speech processing: emerging computational principles and operations," *Nature neuroscience*, vol. 15, no. 4, pp. 511–517, 2012.

[18] B. Morillon, C. Liégeois-Chauvel, L. H. Arnal, C.-G. Bénar, and A.-L. Giraud, "Asymmetric function of theta and gamma activity in syllable processing: an intra-cortical study," *Frontiers in psychology*, vol. 3, 2012.

[19] J. L. McClelland and D. E. Rumelhart, "An interactive activation model of context effects in letter perception: I. an account of basic findings." *Psychological review*, vol. 88, no. 5, p. 375, 1981.

[20] R. Amarasiri, D. Alahakoon, and K. Smith, "Applications of the growing self organizing map on high dimensional data," *IITC*, 2004.

[21] W.-S. Tai and C.-C. Hsu, "A growing mixed self-organizing map," in *Proceedings of Sixth International Conference on Natural Computation*, vol. 2, 2010, pp. 986–990.

[22] ——, "Improving visualization of mixed-type data with a dynamic som," in *Proceedings of Seventh International Conference on Natural Computation (ICNC)*, vol. 1. IEEE, 2011, pp. 431–435.