Jingnan Yang
Prof. Paul Boersma
Research Tutorial
July 30, 2020

Distributional Learning of Mandarin Lexical Tones in Bidirectional Deep Neural Network

Abstract

This paper presents an artificial bidirectional deep neural network simulation, namely deep Boltzmann machines, to model the distributional learning of four Mandarin Chinese lexical tones with a design of extracting and segmenting the auditory-phonetic information of tonal patterns on a three-dimensional continuum. The virtual learner exhibits the emergence of prototypes and perceptual magnet effect after being trained with an adequate number of sound-only tokens. After being trained with an adequate number of sound-meaning pairing tokens, the virtual learner has developed the ability to handle both comprehension and production of the four tones. Evidence for distributional learning is also found in the variant outcomes of learning four categories of sound tokens that are not distributed equally.

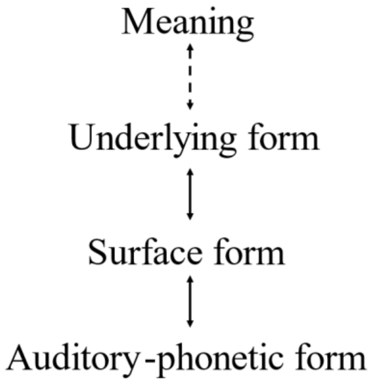
Keyword: neural networks, categories, distributional learning, Mandarin tones, bidirectional, Boltzmann machines, cognitive modeling

1. Introduction

Mandarin Chinse is known as a tone language with four different lexical tones. This means that voice pitch variations are used at the syllable level to convey different lexical meanings (Ching, Williams & Van Hasselt, 1994). In a simplified manner, native speakers' learning of the four tones consists of two simultaneous processes: one is to create categories for the four tones after listening to numerous pieces of distributional sound input, so that they establish the ability of distinguishing tones when listening to new syllables; the other is to link different meanings in the lexicon with the phonological recognition of tonal patterns, so that the meanings of the syllables are ensured to be encoded in the tonal utterances. This present paper presents how these two processes can be modeled in an artificial bidirectional deep neural network.

Bidirectional multi-level models for phonology and phonetics (BiPhon), as adopted in Boersma (2009) and Boersma & Hamann (2009), are models consisting of multiple levels of representations with the capability of processing linguistic information in both comprehension and production directions (Boersma, 2019). A simplified BiPhon framework is presented in Figure 1. Language comprehension starts from the input of auditory-phonetic information. Governed by an interaction between auditory cues and structural constraints in native phonology, the auditory-phonetic form is mapped onto the phonological surface form. The surface form is mapped on to the underlying form under the guidance of faithfulness knowledge as a parallel process. The underlying form pairs up with semantic meaning to be preserved in the lexical storage. Language production works in the reverse direction.

Figure 1: BiPhon Model Structure



A comprehensive model based on the BiPhon framework has to be a neural network due to its accountability for various types of behavioral data and its compatibility with our current state of knowledge about the mechanism of human brains (Boersma, Benders & Seinhorst, 2018). Neural network models presented in previous researches have successfully handled phonological feature emergence (Seinhorst, 2012; Boersma, Benders & Seinhorst, 2018; Seinhorst, Boersma & Hamann, 2019; Benders, 2013; Chladkova, 2014) and auditory dispersion (Seinhorst, 2012; Boersma, Benders & Seinhorst, 2018; Seinhorst, Boersma & Hamann, 2019).

However, all the models mentioned above dealt with representations of auditory-phonetic information on a one- or two- dimensional continuum, such as the VOT of bilabial plosives or the F1 and F2 values of corner vowels. None of them addressed the issue of durational auditory-phonetic information, which characterize Mandarin lexical tones. The originality of this paper is reflected in an attempt to model a type of auditory-phonetic information requiring the processing of temporal sequences.

The paper is organized as follows. Section 2 explains the representations for the auditory-phonetic information of the four tones as input for the current simulation. Section 3 describes the structure of the artificial neural network. Section 4 introduces the training procedure based on restricted Boltzmann machines. Section 5 and 6 presents the performance of the network after learning sound-only and sound-meaning pairing data. Section 7 discusses the findings, issues and concludes the paper.

2. Representation of Mandarin Lexical Tones

This section explains how the auditory-phonetic information and the temporal sequences of the four tones in Mandarin Chinese are represented as input for the current neural network simulation. The virtual learner is expected to establish a phonological interpretation for the four tones as four discrete categories after receiving the representations of the auditory-phonetic information as described in this section.

2.1. Cues: F0 Height and Contour

Tone or pitch is a function of the rate of vocal fold vibration, which is quantified as the fundamental frequency (F0), expressed in Hertz (Hz) (Jongman et al., 2006). F0 is considered to serve as a primary cue to the perception of Mandarin tones (Howie, 1976), and F0 height and F0 contour both serve as important perceptual cues for Mandarin speakers to distinguish the four tones

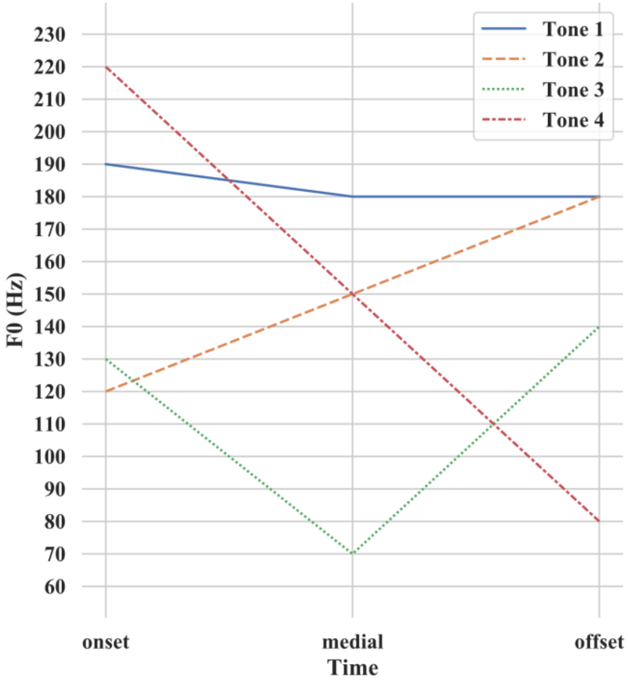
(Gandour, 1984).

For the current simulation, in order to convey the information of both F0 height and F0 contour to the virtual learner, each piece of sound input is divided into three evenly spaced segments, and then represented by three values, which are respectively the F0 value at the onset, medial and offset point of the sound. Table 1 shows the relevant information of the four tones that will be used to train the current simulation. This information is based on the data of average F0 values in Hz produced by a male native speaker, as provided in Wu (1986). It also provides the corresponding lexical meaning for each sound, taking the syllable *ma* for example. Figure 2 visualizes the F0 data provided in Table 1.

Table 1: Representations for four tones

Tone	Onset F0	Medial F0	Offset F0	F0 Contour	Meaning
1	190	180	180	Plain/High	mother
2	120	150	180	Rising	hemp
3	130	70	140	Falling-Rising	horse
4	220	150	80	Falling	scold

Figure 2: Four Mandarin tones



2.2. Distribution of Sounds

The occurrences of the four tones are not perfectly equally frequent in real Mandarin Chinese usage, so the training data for the virtual learner should follow the natural distribution of the four tones in the language. According to the documentation of Zhang & Lai (2010, p.167) based on the corpus of Mandarin syllable frequencies with tones (Da, 2004), the distribution of the four tones is presented in Table 2. The neutral tone is not included in the consideration of the present study.

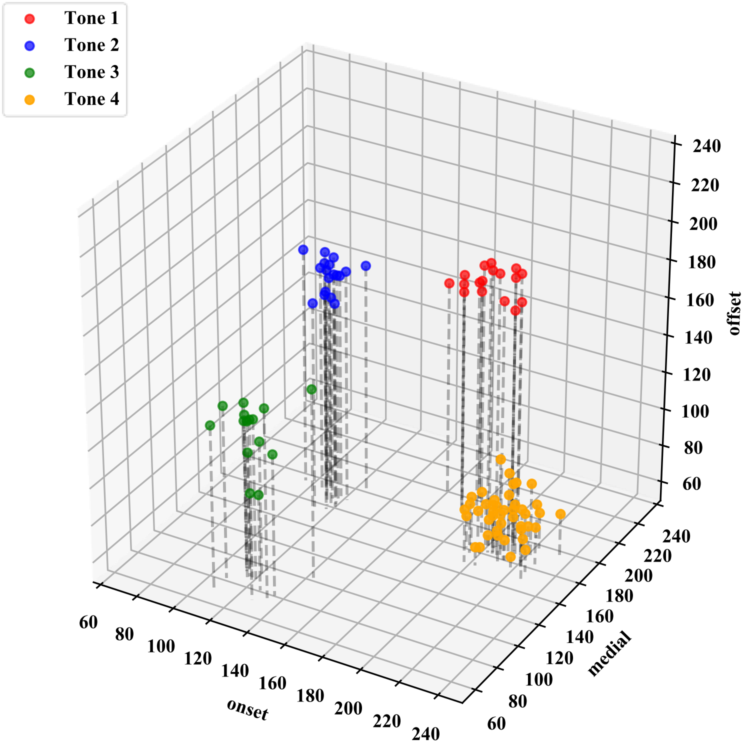
Table 2: Distribution of Mandarin tones

Tone	Percentage
1	16.7
2	18.4
3	14.8
4	42.5
Neutral	7.6
Total	100

The values in Table 1 are only averages of the four sounds. In reality, each sound consists of three F0 values, and each of these F0 values is drawn from a normal distribution. The means of these distributions are the values presented in Table 1, and this paper assumes the standard deviation of each F0 distribution to be 10 Hz.

Since each sound is represented with three F0 values, it can be viewed as a data point in a three-dimensional F0 space. The three axes of this space represent the onset, medial and offset F0 continuum respectively. Figure 3 shows 100 randomly generated sound tokens in a three-dimensional F0 space with the means as shown in Table 1 and a standard deviation of 10 Hz. Each continuum on the axes ranges from 60 to 240 Hz, in line with the normal F0 range of human male voice. The distribution of these sound tokens is in line with the data in Table 2. These tokens serve as a sample of the training input for the virtual learner.

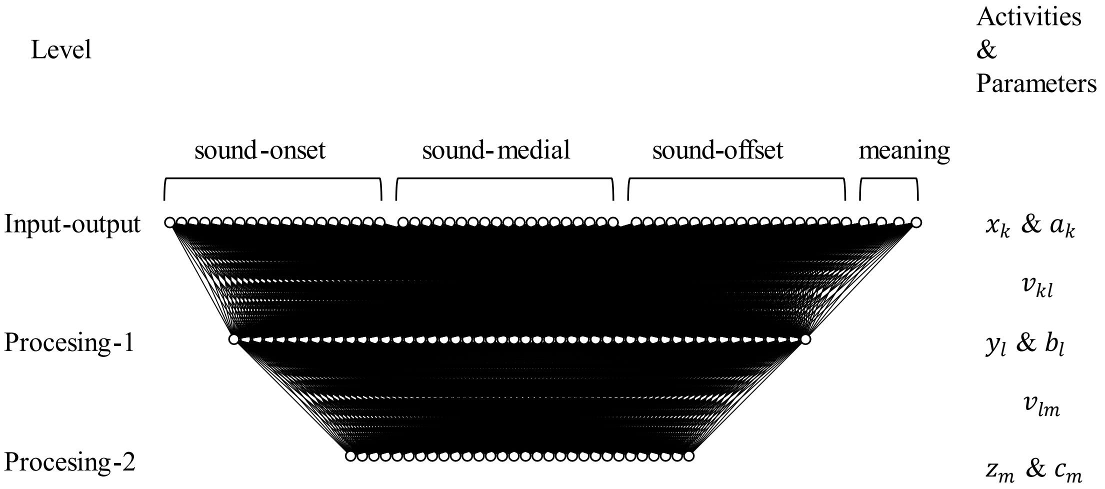
Figure 3: 100 randomly generated sound tokens



3. Network Structure

The simulation adopted in this paper is a deep restricted Boltzmann network with three levels: one level for sound and/or meaning input and output, and two deeper layers for information processing. The design of this network is based on the deep Boltzmann machines presented in Boersma (2019). All nodes on each level connect with all nodes on its neighboring level(s), and there are no within-level connections between nodes. The architecture of this network is shown in Figure 4. This section describes what different groups of nodes manage and what types of representations can be found on each of these levels.

Figure 4: Network architecture



3.1. Input-output Level

The input-output level contains all the sound and meaning nodes. The sound nodes are a representation of the basilar membrane that functions in hearing F0. They are divided into 3 clusters, representing the onset, medial and offset time points of each sound. This design is inspired by Li et al.'s (2006) artificial neural network model for Mandarin tone recognition which divides a sound into 10 evenly spaced segments. Within each cluster, 19 sound nodes represent a F0 continuum running from 60 to 240 Hz, each spaced 10 Hz apart. The four meaning nodes represent the four lexical meanings of the four tones, running from 1 to 4.

The input-output level has activities x_k, where k runs from 1 to K = 61. Although sound and meaning nodes are placed on the same level of the network, they do not receive activation in the same manner, and there is no within-level connection between the two groups of nodes. Table 3 presents the nodes that correspond to the means of the F0 distributions shown in Table 1.

Table 3: Representation for four tones in nodes

Tone	Onset (μ_1)	Medial (μ_2)	Offset (μ_3)
1	13	12 + 19 = 31	12 + 19*2 = 50
2	6	9 + 19 = 28	12 + 19*2 = 50
3	7	2 + 19 = 21	8 + 19*2 = 46
4	16	9 + 19 = 28	2 + 19*2 = 40

Each number represents the node's position in the cluster, counting from left to right. As mentioned in section 2.2, the F0 values in the sound tokens for training have a standard deviation of 10 Hz from the means, which is applied as $\sigma = 1$ node in the simulation. The distribution of sound tokens for training can be represented as:

(1) $s_1 \sim \mathcal{N}(\mu_1, \sigma)$
$s_2 \sim \mathcal{N}(\mu_2, \sigma)$
$s_3 \sim \mathcal{N}(\mu_3, \sigma)$

Whenever a sound comes in, each F0 value of the sound causes a Gaussian bump on the basilar membrane, which is represented by a node cluster, with a spreading half-width of $w = 2$ nodes. For the sampled sound tokens for training, the input activation for sound nodes is then the basilar excitation pattern:

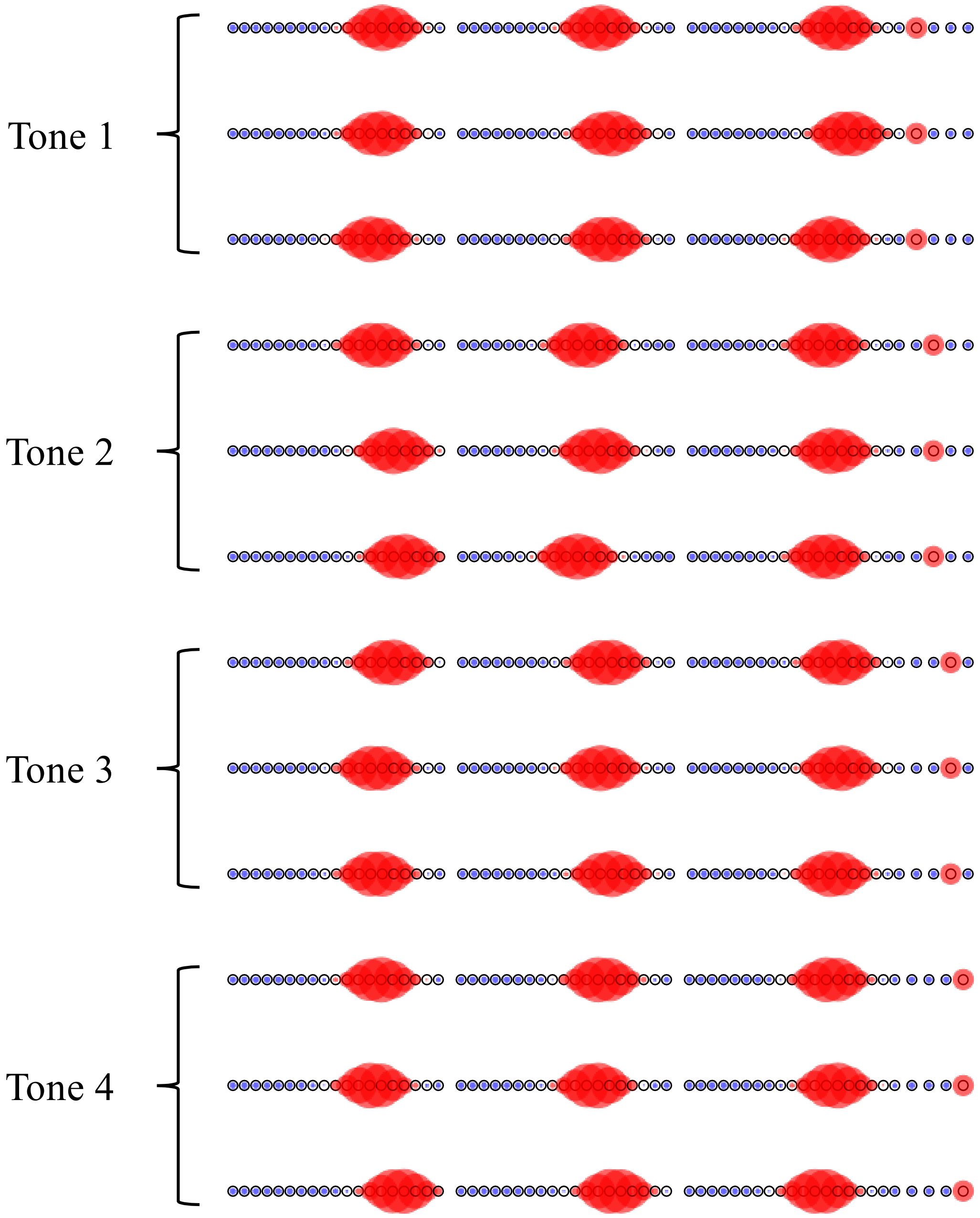
$$(2)\quad x_k = \begin{cases} 5 * \exp\left(-\frac{1}{2} * \left(\frac{k - s_1}{w}\right)^2\right) - 0.5, k = 1 \ldots 19 \\ 5 * \exp\left(-\frac{1}{2} * \left(\frac{k - s_2}{w}\right)^2\right) - 0.5, k = 20 \ldots 38 \\ 5 * \exp\left(-\frac{1}{2} * \left(\frac{k - s_3}{w}\right)^2\right) - 0.5, k = 39 \ldots 57 \end{cases}$$

Whenever a meaning comes in, the activated meaning node receives an activation of 2 while the three other meaning nodes receive a negative activation of -0.5, for *k* running from 58 to 61:

$$(3)\quad x_k = \begin{cases} 2, \; if\ k\ is\ activated \\ -0.5, \; if\ k\ is\ not\ activated \end{cases} \quad (k = 58 \ldots 61)$$

Figure 5 shows the activity patterns on the input-output level after being activated by some sound-meaning pairing tokens in the training data. Each black-lined circle represents a node and they do not differ in size. Each color-filled circle represents an activation and the size of the circle represents the activation strength. Red represents positive and blue represents negative activity.

Figure 5: Examples of activation patterns on the input-output level



3.2. Deeper Levels & Connections

The middle processing level has activities y_l, with l running from 1 to $L = 50$. The deepest level has activities z_m, with m running from 1 to $M = 30$. These deeper levels do not get in direct contact with input or output representations, but they are in charge of processing and preserving information in a relatively long-term memory based on the training input.

Both the sound and meaning nodes are connected to the middle layer. Each node k on the input-output level is connected to each node l on the middle level with a weight of u_{kl}. Each node l is also connected to each node m on the deepest level with a weight of v_{lm}. The three levels from shallow to deep respectively have a bias a_k, b_l and c_m. These weights and biases are updated with a learning rate of 0.001 as the network learns from the training input of sounds or sound-meaning pairs. The nodes in these two deeper levels have binary activities, which means their activity is either 0 or 1.

The network in the current simulation is bidirectional in two senses. One is that it can handle language processing in both comprehension and production directions as a whole. The other is that the same set of weights and biases can handle the flow of information through the three levels no matter it flows from shallow to deep or from deep to shallow levels. None of the level in the network should be interpreted as directly referring to any of the levels shown in Figure 1.

4. Training Procedure

This section describes the training procedure of the current neural network simulation. Training starts with initializing the network to an empty initial state, which means that all the parameters, including the biases on the three levels and the two sets of weights, are set to zero. Then 1000 training steps at the learning rate of 0.001 are executed. In each training step, the input-output level receives a sound token or a sound-meaning pairing token. The network processes each training input through four phases: initial settling and spreading up, Hebbian learning, dreaming, and anti-Hebbian learning. The current paper adopts the terminology for the first phase as "spreading up" in line with the terminology coined in Boersma (2019), despite that it is literally "spreading down" to deeper levels in the current simulation, as shown in Figure 4.

4.1. Initial Settling and Spreading Up

Each training step begins with activating the nodes on the input-output level according to the incoming information. As described in section 3.1, a piece of sound-only training input causes a Gaussian bump on each cluster of sound nodes, and a piece of sound-meaning pair input causes additionally a positive activity on the chosen meaning node.

The activities y_l on the middle level are determined by both the input-output level and the deepest level. While x_k has just been decided by the new piece of input and is temporarily clamped in this phase, z_m still remains the same as from the last training step. The middle level activities y_l, for l from 1 to 50:

$$(4) \quad y_l = \sigma\,(b_l + \textstyle\sum_{k=1}^{K} x_k u_{kl} \;+\; \sum_{m=1}^{M} v_{lm} z_m)$$

The sigmoid function σ is a standard logistic function:

$$(5) \quad \sigma(x) = 1 / (1 + \exp(-x))$$

What comes next is that the activity of y_l then spreads to the deepest level z_m, for m from 1 to 30:

$$(6) \quad z_m = \sigma(c_m + \textstyle\sum_{l=1}^{L} y_l v_{lm})$$

Because both deeper levels only have binary activities, as mentioned in section 3.2, a random Bernoulli function $\mathcal{B}$ is then applied to their activities. This function computes the activities of y_l and z_m stochastically, which means that the binary values 0 and 1 will be assigned to these nodes with probabilities of the values of their activities computed through (4) and (6):

(7) $y_l = \mathcal{B}(y_l)$

(8) $z_m = \mathcal{B}(z_m)$

In order to resonate into a near-final state, the simulation repeats the sequence of (4), (6), (7) and (8) for five times. After that, the network achieves an almost equilibrium state in a deterministic way.

4.2. Hebbian Learning

The second phase of the training procedure allows the network to update the two sets of weights and three biases according to the influence of the training input which has already been spread throughout the network in the first phase. This process is governed by the Hebbian learning rule, which assumes that that the connections between two active nodes should be strengthened, and the strengthened connections will motivate them to be more often simultaneously active later on(Hebb, 1949). At the meantime, active nodes receive stronger biases, so that they will be more often active in the future. As mentioned before, the learning rate *r* is 0.001.

(9) $a_k = a_k + rx_k$

(10) $b_l = b_l + ry_l$

(11) $c_m = c_m + rz_m$

(12) $u_{kl} = u_{kl} + rx_ky_l$

(13) $v_{lm} = v_{lm} + ry_lz_m$

4.3. Dreaming

The network dreams up its own pattern in the third phase. Stochasticity is again applied in the resonation. The activities on the input-output level are not clamped anymore, and they are influenced by the middle level:

(14) $x_k = a_k + \sum_{l=1}^{L} u_{kl}y_l$

The activities on two deeper levels keep resonating in this phase in the same manner as in the first phase of spreading up. The sequence (14), (4), (6), (7), (8) is repeated for five times. The stochasticity and the influence from training input ensure that the network can faithfully process the information brought by the training token.

4.4. Anti-Hebbian Learning

The last phase of the training procedure allows the network to “forget” what it has learned. The parameters are updated with a negative learning rate of the same absolute value:

(15) $a_k = a_k - rx_k$

(16) $b_l = b_l - ry_l$

(17) $c_m = c_m - rz_m$

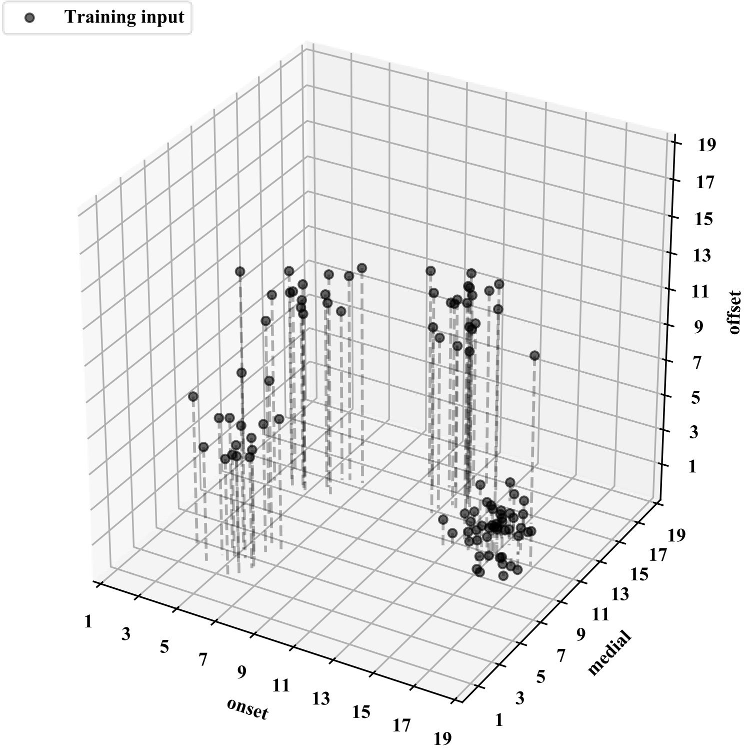
(18) $u_{kl} = u_{kl} - rx_ky_l$

(19) $v_{lm} = v_{lm} - ry_lz_m$

5. Learning Sounds

This section reports the simulation's learning outcomes after being trained with 1000 sound-only tokens at a learning rate of 0.001. Figure 6 illustrates a possible realization of 100 out of 1000 randomly generated training tokens in a three-dimensional F0 space. Although the sound tokens come without any intended phonological or semantic representations, the distribution exhibits clear visual clustering. The virtual learner is expected to generalize the prototypes from the distributional pattern of the sound tokens and establish perceptual magnet effect.

Figure 6: 100 Tokens in a realization of sound-only training input



5.1. Emergence of Prototypes

After being trained with 1000 sound tokens, the virtual learner is tested for its emergence of prototypes.

The testing measure is to feed the virtual learner with a standard sound, i.e. a sound with the F0 values provided in Table 3, and the reflection on the input-output level after the network's processing is regarded as the created prototype. To compare with the standard input, each reflection is represented with the peak nodes among the activities on each sound node cluster. The prototypes are expected to be very close or even identical to the standard sounds. Being a prototype means that the virtual learner considers this sound token as "better" than others on the same continuum (Kuhl, 1991). If the current task finds the prototypes, they will be used as benchmarks for measuring perceptual magnet effect in the next task.

When being tested, the virtual learner processes the information through a spreading up phase and a dreaming phase as described in section 4, namely the sequence of (4) and (6) for five times, followed by the sequence of (14), (4) and (6) for five times, except that the random Bernoulli function is not applied after calculating the activities of y_l and z_m each time. Putting the stochasticity aside allows the two deeper levels to have non-binary activities and provide clearer exhibition of the network's behavior. Hebbian learning and anti-Hebbian learning phases are not performed in any testing, because the virtual learner is not expected to learn from testing input, which means that the parameters are not supposed to be influenced by the testing input.

The learning proves to be effective in that the prototypes reflected in a network after 1000 training steps are consistently extremely close or even identical to the standard sounds. We ran the experiment of training a new virtual learner and examined its created prototype multiple times and observed a consistent pattern of how the prototypes would turn out. Table 4 reports the F0 values of this consistent pattern.

Table 4: Prototypes created in the virtual learner

	Prototypes			cf. Standard Sounds		
	Onset	Medial	Offset	Onset	Medial	Offset
Tone 1	13	11	12	13	12	12
Tone 2	6	10	12	6	9	12
Tone 3	7	2	8	7	2	8
Tone 4	16	9	2	16	9	2

Figure 7 visualizes the four prototypes. The prompts, i.e. testing input of standard sounds, are marked with black colored arrows. Red colored lines connecting nodes represent positive weights, blue colored lines represent negative weights, and black colored lines represent zero weights. Since the training input only has sound tokens, the weights connecting the meaning nodes are naturally not updated at all.

Figure 7: Prototypes Created in the Virtual Learner

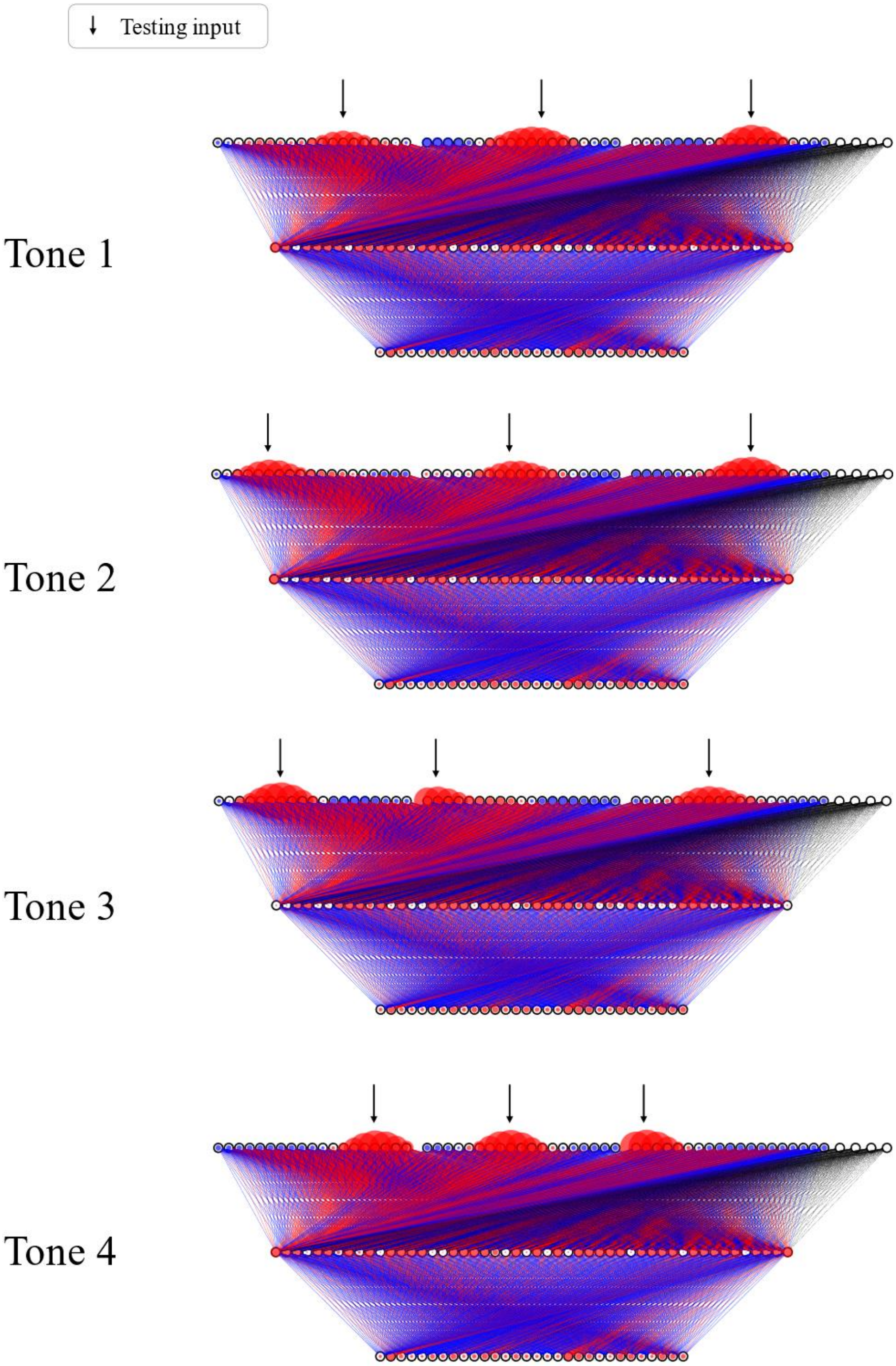
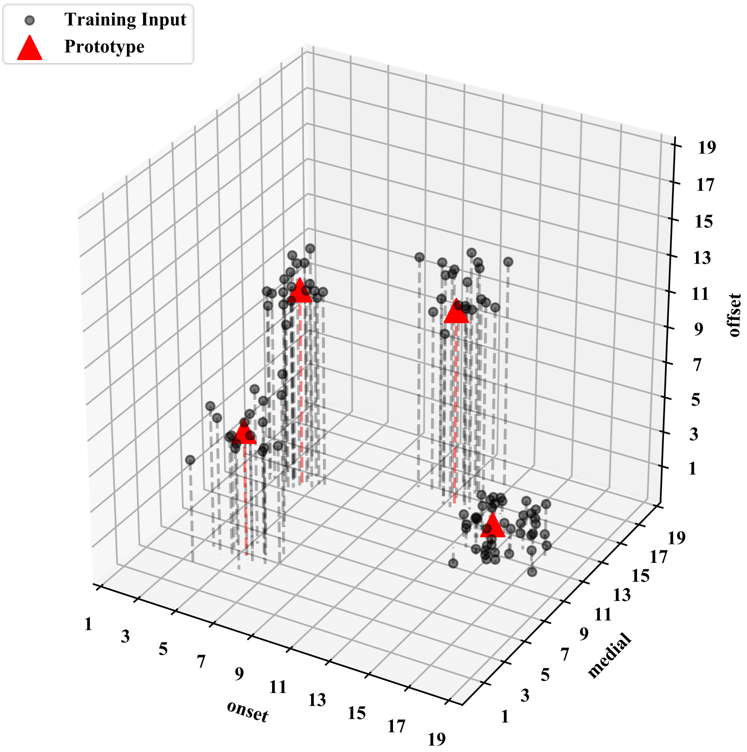


Figure 8 shows the prototypes, which are represented with the three peak nodes, in a three-dimensional F0 space along with 100 training input. This figure shows that the prototypes are roughly located in a central place in each of the visual clustering of training sound tokens.

Figure 8: A Training Input Sample vs. Created Prototypes



5.2. Emergence of Perceptual Magnet Effect

The virtual learner is also tested for its creation of sound categories, exhibited as a perceptual magnet effect which would "drag" the sound they hear much closer to their learned prototypes in the reflection. Although the virtual learner might hear a wide range of sound tokens on the continuum, it is expected to only reflect them as belonging to four discrete categories. perceptual magnet effect has been empirically proved in human adults and infants (Kuhl, 1991).

The prototypes found in the previous task are used in the current task as the benchmarks for the testing of perceptual magnet effect, namely the values shown in Table 4. The testing measure is to give the virtual learner a random sound and see if reflection turns out much closer to one of the prototypes. A more accurate name for the testing prompt is actually a "non-prototypical" sound rather than a "random" sound, because it is not really a sound that can possibly be anywhere in the F0 space, but is drawn from the F0 distributions with the same means as the training input but a larger standard deviation of 3 nodes (cf. the standard deviation for training input is 1 node). Using non-prototypical sounds but not technically random sounds is more of a simulation to real-life situations.

The virtual learner processes the random sound's activation with the sequence of (4) and (6) for five times. Then it dreams up the pattern and reflects the outcome on the input-output level with through repeating the sequence of (14), (4) and (6) for five times.

We judge whether the perceptual magnet effect is shown in the virtual learner's performance by comparing the distance between each testing input and its closest prototype with the distance between each reflection and its closest prototype. The distances d between each training sound token or a reflection token, which are represented with three nodes t_1, t_2, t_3, with the four prototypes, which are represented with three nodes p_1, p_2, p_3, are calculated as Euclidean distance:

(20) $d = \sqrt[2]{(p_1 - t_1)^2 + (p_2 - t_2)^2 + (p_3 - t_3)^2}$

A function *Min* is performed to the four to get the shortest distance out of the four and pinpoint the closest prototype:

(21) $d_{min} = Min(d_1, d_2, d_3, d_4)$

If the distance from the reflection to its closest prototype is significantly shorter than the distance from the testing input to its closest prototype, it is recognized that the virtual learner exhibits perceptual magnet effect after receiving 1000 steps of training.

The learning proves to be effective in that the network provides a reflection which is very close to one of the prototypes almost every time in our testing experiment, which means that when the virtual learner listens to a continuum of sounds, it almost always interprets them as belonging to one of the four discrete categories. Figure 9 shows examples of the virtual learner's categorical behavior when it receives a random sound. In each example, values of the closest prototype with relation to the reflection are marked with green-colored arrows, and the values of the testing prompt are marked with black-colored arrows.

Figure 9: Example reflections of perceptual magnet effect

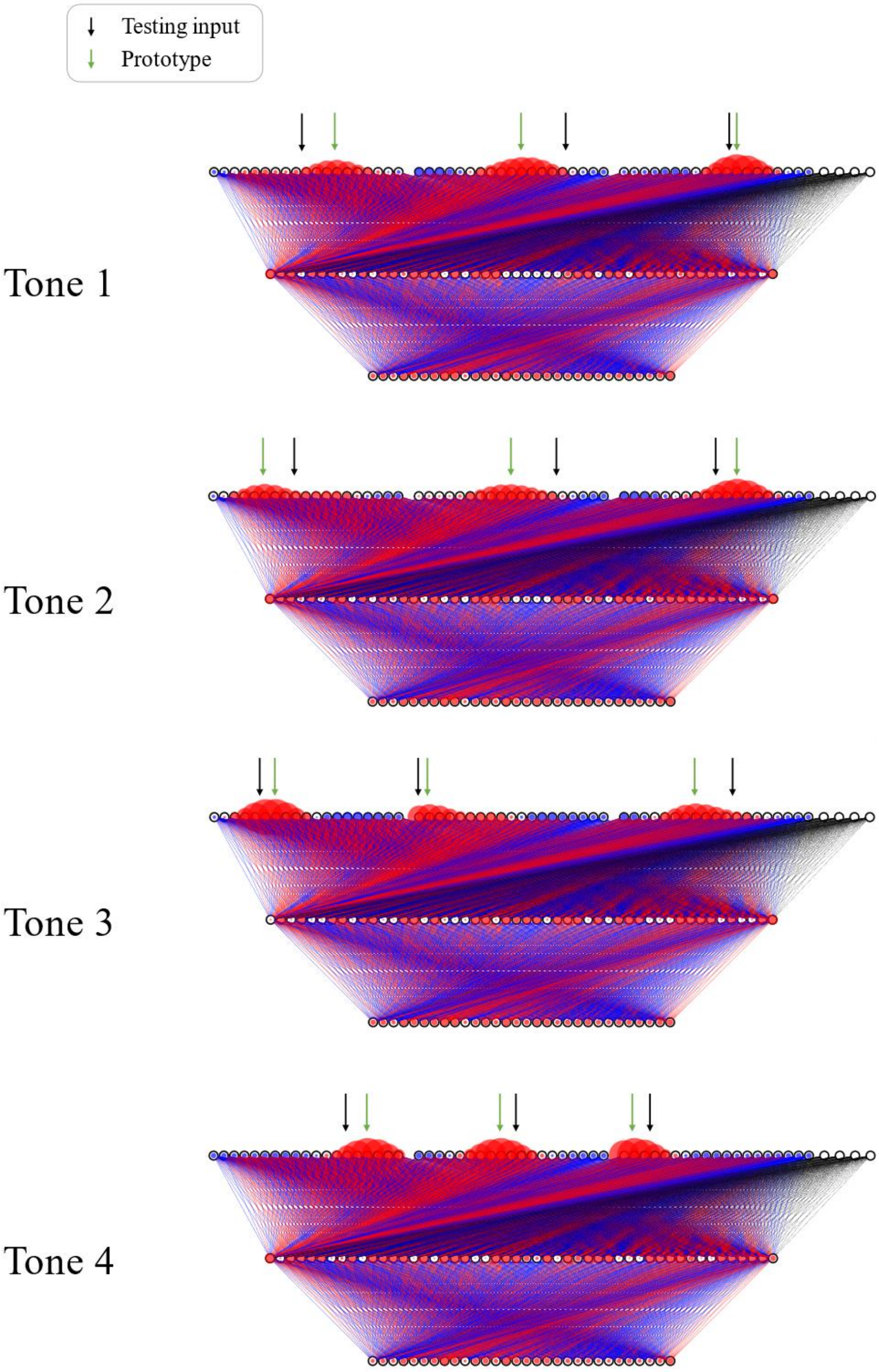
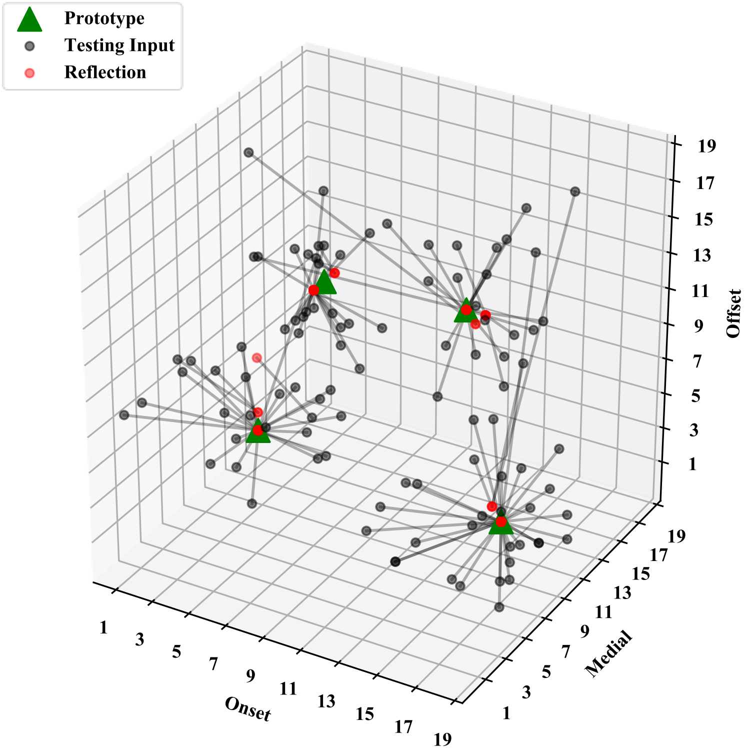


Figure 10 shows an experiment result of testing the virtual learner with 100 pieces of random sounds in a three-dimensional F0 space. The opaque black lines connect each testing sound token and the network's reflection. The distance from each piece of perception to the closest prototype is significantly shorter than from each piece of testing input. Also, many of the perception tokens overlap in the figure and are extremely close to the prototypes, indicating that the perceptual magnet effect in the virtual learner is strong.

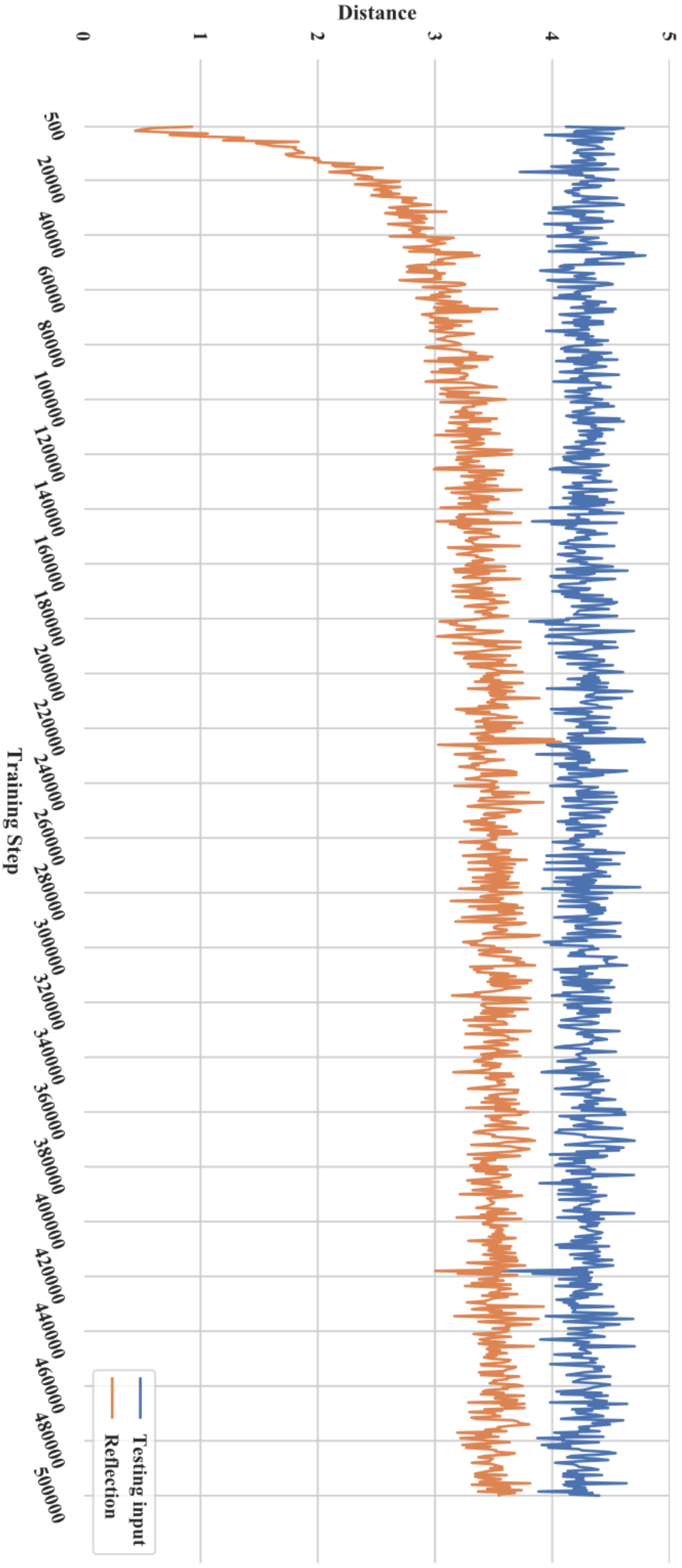
Figure 10: 100 Pieces of testing input vs. reflection



However, this degree of strong categorical behavior is transient. Figure 11 shows the strength of the virtual learner's categorical behavior with regard to the number of training steps it takes. It is trained with 500000 sound tokens in total, and each checkpoint is 500 tokens apart. At each checkpoint, the network receives 100 pieces of random sound as testing input, and the average distance from the 100 testing sound tokens to each of their closest prototype and the average distance from the 100 perceptions to each of their closest prototype are computed.

As Figure 11 shows, the strength of the virtual learner's categorical behavior reaches summit when it is trained with between 1000 and 1500 pieces of data. After that, as the number of training steps increases, the strength of categorical behavior decreases. After being trained with 200000 pieces of data, the strength of the virtual learner's categorical behaviors remains roughly stable, as shown in that the difference between the two sets of distances stays about 0.5 node. This result is different from what was reported in Boersma (2019) that his deep Boltzmann machine learner loses the perceptual magnet effect after being overtrained.

Figure 11: Comparison of distances to prototypes from testing input and reflections



6. Learning Sound-meaning Pairs

As stated in the introduction to the BiPhon framework, this neural network model is supposed to be capable of handling language processing in both comprehension and production directions. This section reports the simulation's learning outcomes after being trained with 1000 pieces of sound-meaning pairing tokens at the learning rate of 0.001. Beyond the creation of sound prototypes and categories as shown in section 5, provided with data with meaning labels, it is also expected to link the meanings with its recognitions of sound patterns.

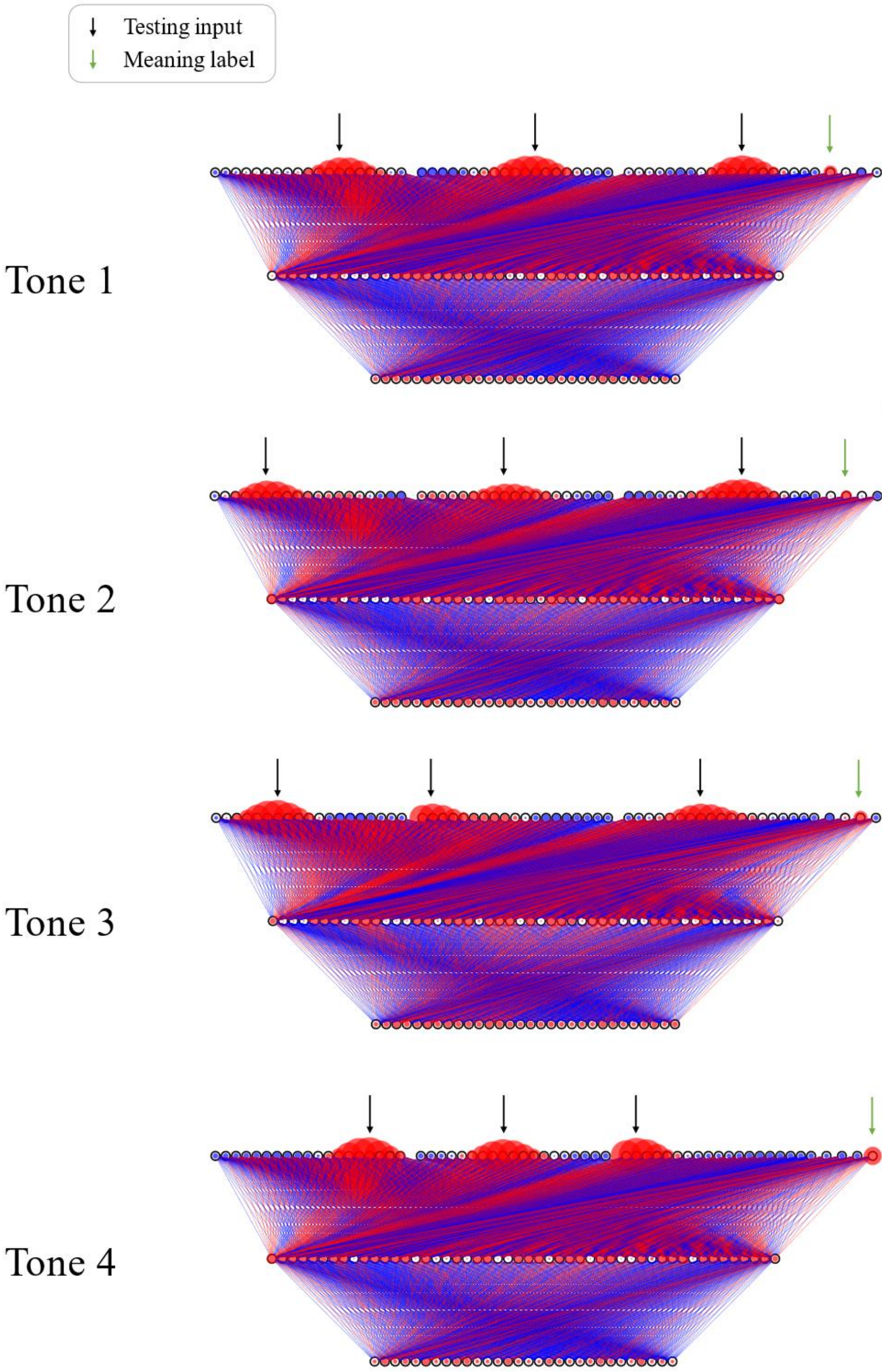
6.1. Comprehension

This section reports the performance of the virtual learner's capacity of comprehending standard sounds and random sounds. The former task aims to test if the virtual learner can correctly reflect the meaning of a labeled sound token. The latter task aims to test if the virtual learner can interpret a random sound as belonging to one of the four discrete categories and simultaneously reflect the meaning of that interpreted category.

For a standard sound, the virtual learner is expected to present an activated meaning node after processing the sound and that meaning node should match the label of the testing prompt. The testing measure is to give the virtual learner a standard sound token. The virtual learner processes the sound information with the sequence of (4) and (6) for five times and (14), (4) and (6) for five times. Finally, the activity pattern reflected on the input-output level will include the activities of both sound and meaning nodes.

We ran this experiment multiple times, and the typical testing results for the four tones are presented in Figure 12. The testing sound prompts are marked with black colored arrows. The results show that the virtual learner is able to distinguish and comprehend the four prototypical sounds, and also a consistent pattern of different strength of activity among the four meaning nodes when they are the chosen one, namely Tone 4 is stronger than the other three.

Figure 12: Results of comprehending standard sounds



For a random sound, which does not come with any meaning label, the virtual learner is expected to interpret the sound as one of the four discrete categories and simultaneously reflect a matching meaning for it. The testing measure of this task is to provide the virtual learner with a random sound, same as the testing prompt in section 5.2. The network's processing is again the same sequence of (4) and (6) for five times followed by (14), (4) and (6) for five times. If the activity pattern of the sound nodes is close enough to a prototype and matches the activated meaning nodes, we consider the virtual learner as capable of comprehending a random sound.

We ran this experiment multiple times, and the testing results show that after being trained with 1000 sound-meaning pairing tokens, the virtual learner can interpret a random sound as belonging to one of the four categories and reflect the meaning of that category. Figure 13 shows a few examples of the testing results. In each example, the testing input is marked with black arrows, while the interpreted meaning and the its corresponding prototypical sound pattern are marked with green arrows.

Figure 13: Results of comprehending random sounds

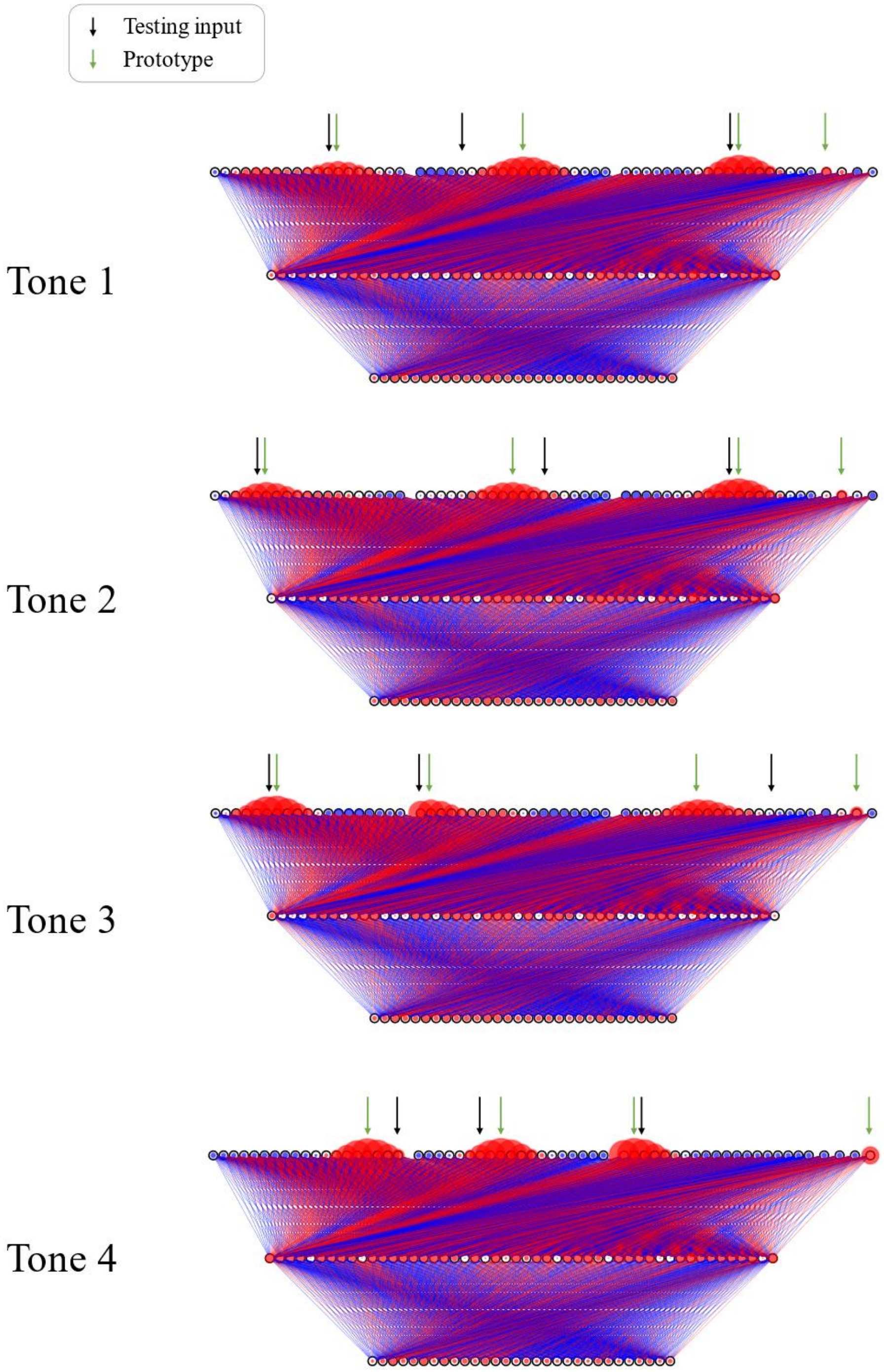
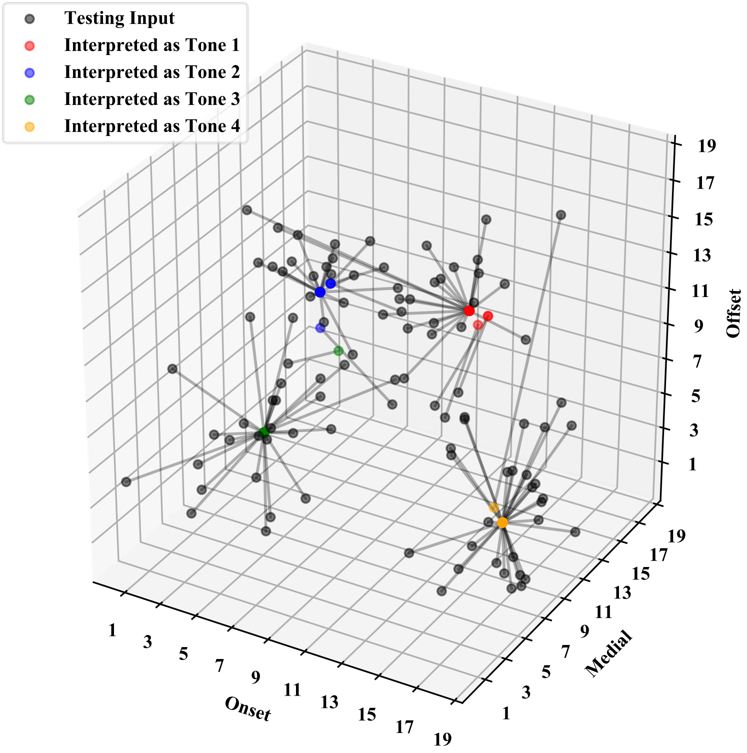


Figure 14 shows 100 pieces of testing input and the network's comprehension of them in a three-dimensional F0 space. Compared with Figure 10, the virtual learner in this task obviously exhibited not only categorical behavior but also the ability to label the category of the sound it hears.

Figure 14: 100 Pieces of testing input vs. interpretations



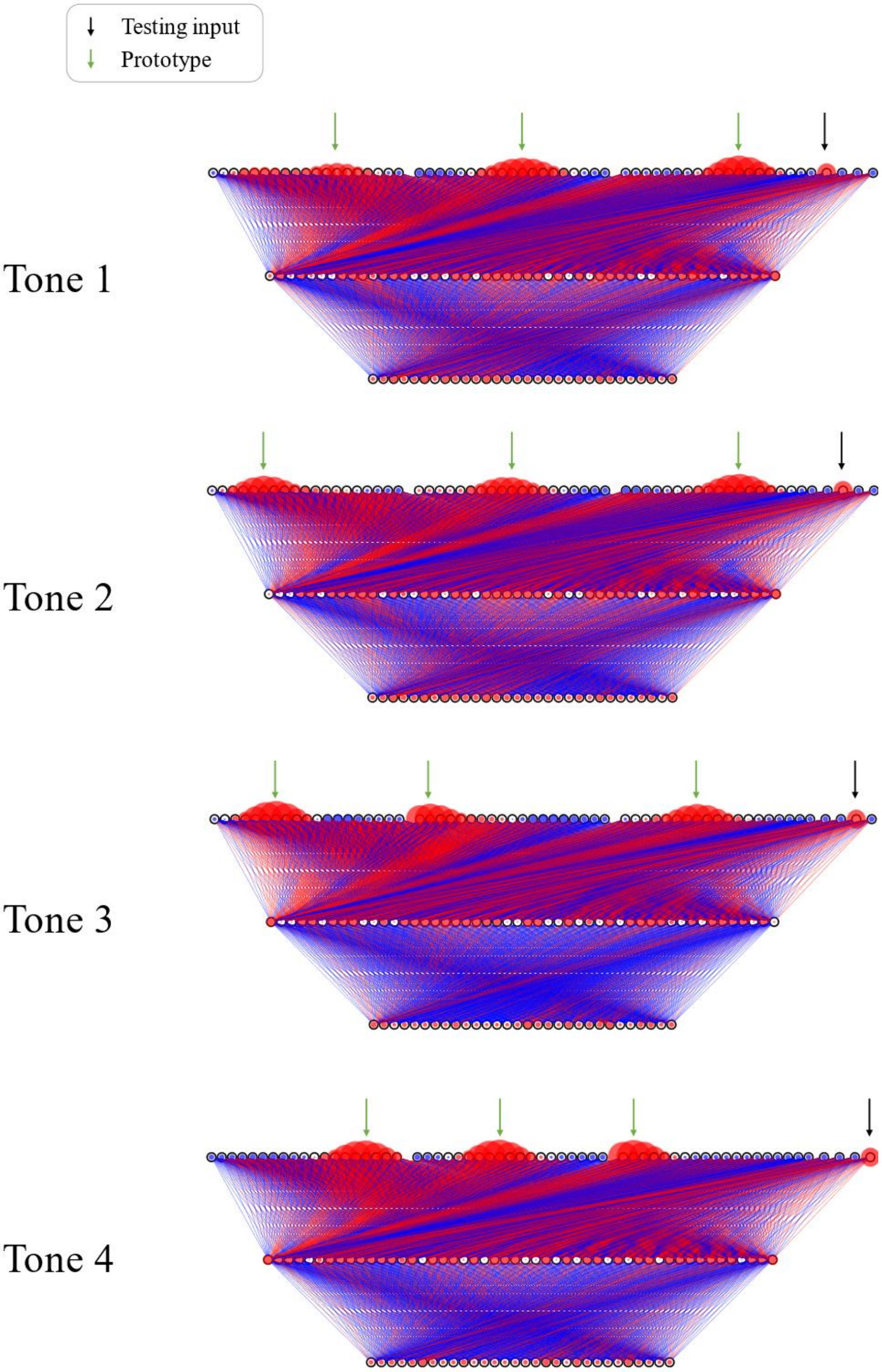
6.2. Production

This section reports the virtual learner's capability of producing sounds with given meanings. As a bidirectional model, when it is provided with a given meaning, it is expected to produce an activity pattern of the corresponding sound on its input-output level.

The testing measure is to turn the activity of the relevant meaning node to 2, those of the other three meaning nodes to -0.5, and those of the sound nodes to 0. Then the overall activity pattern of the input-output level is spread throughout the network as in the dreaming phase. In the few rounds of spreading activities down and up through the network, the activities of the four meaning nodes are clamped, which means that they remain stagnant and unaffected by other activities in the network. The final activity pattern on the input-output level is regarded as the virtual learner's product.

The bidirectionality of this neural network model is attested with its performance of producing the four sounds with given meanings. The results are presented in Figure 15. The activated meaning nodes are marked with black colored arrows and the learned prototypes are marked with green colored arrows for reference. As it is shown, the sounds produced by the virtual learner are extremely close or identical to its created prototypes.

Figure 15: Results of testing production with given meanings



7. Discussion

This section discusses the outcomes reported in the previous two sections and other relevant issues.

7.1. Distributional Learning

The performance of the virtual learner in all the previous tasks in section 5 and 6 supports the hypothesis of distributional learning that discrete sound categories are largely acquired on the basis of stimuli that the learner experiences in the environment, instead of being hardwired innately in their cognitive faculties (Wanrooij, 2015). It is reflected in two aspects in the current simulation.

First, the perceptual magnet behavior of the virtual learner is established on the basis that the distributional patterns of the training data make the four prototypical sounds prominent among an unlimited number of possible incoming sound tokens. After the training process, the virtual learner acquires the prototypes, establishes sound categories and thereafter interprets any incoming sound as belonging to one of the four discrete categories.

An additional finding of the current simulation is the maintaining of categorical behavior. The one-dimensional simulation in Boersma (2019) is reported to have transient categorical behavior: after a certain number of training steps, the virtual learner is trained with too much data to maintain the perceptual magnet effect. This phenomenon is not observed in the current simulation. As shown in Figure 11, after 200000 pieces of training data, the perceptual magnet effect does not vanish but stay stable. In the stable stage, the virtual learner is still able to "drag" an incoming sound closer to one of the prototypes, but the degree of the perceptual magnet behavior is not at its peak strength anymore.

Second, the different distributions of the four tones in the training data in general lead to different degrees of acquisition in the learner. As it is shown in Table 2, Tone 4 has a much higher frequency of occurrence than the other three sounds in the training data. The performance of the virtual learner in the comprehension tasks show that, the learner can acquire all four tones, but the learning of Tone 4 seems to be the most robust. It is shown in Figure 12 and Figure 13 that the activity of the meaning node for Tone 4 is greater than that of any other tones, and the activities of the meaning nodes for the other three tones are similarly big, since they have a similar frequency of occurrence in the training data.

However, a question arises: why does Tone 3 not appear to be the least robust one in both Figure 12 and Figure 13, when it is the least frequent type token in the training data? We look forward to answers that will be provided in future research. A possible answer is that the learning outcome does not only depend on the distributional pattern of the sounds but also how distinctive the sounds are. In Figure 12, Tone 2 seemed to be least robust, and this might due to that Tone 2 is not as distinctive as other sounds. The three values of Tone 2 are all close to the center of each continuum, while the other sounds have more outlier-like values which are closer to the ends of the continua.

7.2. Towards Higher Levels of Representations

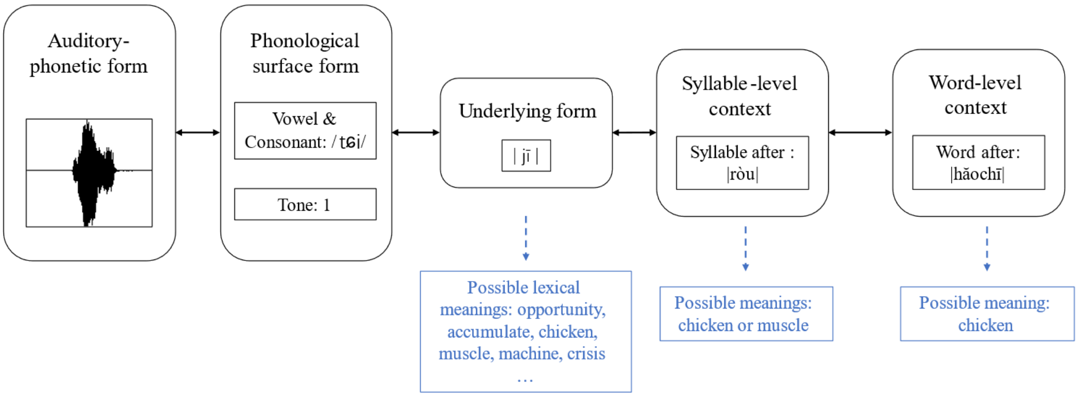
The present simulation models the cognitive process of learning both sounds and meanings only in a simplified manner that each sound, which conveys F0 information, is paired with a lexical meaning, as shown in Table 1.

However, using Mandarin lexical tones as the learning objective is unique in that the sounds and meanings are not necessarily 1-on-1 pairings but rather abstract generalized patterns that can represent an unlimited number of sounds and meanings. When the simulation recognizes the tonal pattern of a sound, it is still far from interpreting its meaning correctly without concerning contextual information, not to mention that the current model does not even involve interpreting consonant or vowels.

The realistic situation is that any Mandarin Chinese syllable's auditory information also includes at least a single vowel. Since many Mandarin words are disyllabic or trisyllabic, the lexical meaning of a word usually cannot be understood with the sound itself, but also requires the facilitation of contextual information. For example, for a Mandarin speaker, the meaning of the syllable *jī* (/tɕi/, Tone 1)" can be understood as opportunity in the disyllabic word *jīhuì*, as machine in the disyllabic word *jīqì*, as chicken or muscle in two disyllabic words that are both pronounced as *jīròu*.

Therefore, the term sound in the current simulation should not be understood as only a mere sound, such as a single vowel /a/ or /u/ in the toy language created in Chladkova (2014), but rather as a generalized tonal pattern that can be applied to any Mandarin syllables. The term meaning in the current simulation should not be understood as a form-meaning pairing entry in the lexicon, but rather an intermediate step that handles tone recognition, which contributes to interpret the true meaning of the utterance with parallel processing of contextual information. The simplified process of a more realistic bidirectional model for the processing of a Mandarin syllable, taking *jī* meaning chicken in *jīròu* for example, is presented in Figure 16.

Figure 16: Model for processing a mandarin syllable


Auditory-phonetic form
Phonological surface form
Vowel & Consonant: / tɕi/
Tone: 1
Underlying form
| jī |
Syllable -level context
Syllable after : |ròu|
Word-level context
Word after: |hǎochī|
Possible lexical meanings: opportunity, accumulate, chicken, muscle, machine, crisis …
Possible meanings: chicken or muscle
Possible meaning: chicken

Each black arrow connects a form in the parallel processing procedures, and the blue words list the possible meanings of this syllable as it is processed. When the listener only has its underlying form, there are many possibilities what this syllable can mean. With the interpretation of syllable-level context, which tells that it is a type of meat, the listener can eliminate the possibilities to two meanings: chicken or muscle. With the interpretation of word-level context, which means delicious, the listener can be certain that it is about chicken but not muscle, because chicken is way more likely to be talked about something edible.

The current virtual learner is naïve in assuming a fixed vowel and consonant combination and a non-confusing context so that it could avoid processing contextual information and neglect the possibility of multiple interpretations for the same syllable. A more advanced modeling requires the integration of vowel (and consonant) information with tonal pattern and higher levels of representations for contextual or even orthographic information.

7.3. Representation of Temporal Sequences

The current simulation represents temporal sequences by dividing a sound into three time points and extracting the three F0 values. This design is inspired by Li et al.'s (2006) artificial neural network model for tone recognition. However, this design is not ideal because a real cognitive perception process does not receive the activation of three F0 values at the same time. This "holistic" activation cannot preserve all the details of the original sound and does not reflect the duration of a sound accurately enough, despite that sound duration can also be used as a cue for tone recognition (Yang et al., 2017).

8. Conclusion

The current bidirectional deep Boltzmann machine neural network successfully modeled the learning of four Mandarin lexical tones, including the emergence of categories and the comprehension and production of them. The current research contributes to the establishment of a comprehensive theory of attempting to understand the cognitive processes of language comprehension, production, acquisition and change through artificial neural network models. Further research is needed to find out better representations of temporal sequences and higher-level forms of information.

References

Ching, T. Y., Williams, R., & Hasselt, A. V. (1994). Communication of lexical tones in Cantonese Alaryngeal speech. *Journal of Speech, Language, and Hearing Research*, 37(3), 557-563.

Benders, T. (2013). *Nature's distributional-learning experiment*. PhD dissertation, University of Amsterdam.

Boersma, P. (2009). Cue constraints and their interactions in phonological perception and production In *Phonology in Perception*, 15, 55-110, P. Boersma, & S. Hamann (Eds.). Berlin: Mouton de Gruyter.

Boersma, P. (2019). Simulated distributional learning in deep Boltzmann machines leads to the emergence of discrete categories. In *Proceedings of the 19th International Congress of Phonetic Sciences*, (pp. 1520-1524). Canberra: Australasian Speech Science and Technology Association Inc.

Boersma, P., Hamann, S. (2009). Loanword adaptation as first-language phonological perception. *Loanword Phonology*, 11-58.

Boersma, P., Benders, T., Seinhorst, K. (2018). Neural networks for phonology and phonetics. Manuscript, University of Amsterdam.

Chladkova, K. (2014). *Finding phonological features in perception*. PhD dissertation, University of Amsterdam.

Da, J. (2004). Chinese text computing: syllable frequencies with tones. Retrieved from: https://lingua.mtsu.edu/chinese-computing/phonology/syllabletone.php

Gandour, J. (1984). Tone dissimilarity judgments by Chinese listeners. *Journal of Chinese Linguistics*, 12, 235-261.

Hebb, D.O. 1949. *The Organization of Behavior*. New York: Wiley.

Howie, J. M., & Howie, J. M. (1976). *Acoustical studies of Mandarin vowels and tones (Vol. 18)*. Cambridge University Press.

Jongman, A., Wang, Y., Moore, C. B., & Sereno, J. A. (2006). Perception and production of Mandarin Chinese tones. Submitted to *Handbook of Chinese Psycholinguistics*. E. Bates, L.H. Tan, & O.J.L. Tzeng (Eds.). Cambridge University Press.

Kuhl, P. K. (1991). Human adults and human infants show a "perceptual magnet effect" for the prototypes of speech categories, monkeys do not. Perception & psychophysics, 50(2), 93-107.

Li, X., Wenle, Z., Ning, Z., Chaoyang, L., Yongxin, L., Xiuwu, C., & Xiaoyan, Z. (2006). Mandarin Chinese tone recognition with an artificial neural network. *Journal of Otology*, 1(1), 30-34.

Seinhorst, K. (2012). The evolution of auditory dispersion in symmetric neural nets. MA thesis, University of Amsterdam.

Seinhorst, K., Boersma, P., & Hamann, S. (2019). Iterated distributional and lexicon-driven learning in a symmetric neural network explains the emergence of features and dispersion. In *Proceedings of the 19th International Congress of Phonetic Sciences*, Melbourne, Australia 2019 (pp. 1134-1138). Canberra: Australasian Speech Science and Technology Association Inc.

Tsai, C.H. (2000). Mandarin Syllable Frequency Counts for Chinese Characters. Retrieved from: http://technology.chtsai.org/syllable/

Wanrooij, K. E. (2015). *Distributional learning of vowel categories in infants and adults*. PhD dissertation, University of Amsterdam.

Wu, Z.J. (1986). The spectrographic album of mono-syllables of Standard Chinese. Social Science Press, Beijing.

Yang, J., Zhang, Y., Li, A., & Xu, L. (2017). On the Duration of Mandarin Tones. In *INTERSPEECH* (pp. 1407-1411).

Zhang, J., & Lai, Y. (2010). Testing the role of phonetic knowledge in Mandarin tone sandhi. *Phonology*, 27(1), 153-201. doi:10.1017/S0952675710000060