Modelling Sensorimotor Knowledge in a Neural Network Model of Speech

Victoria Ivanova
Student ID: 11654198

Supervisor:
Paul Boersma

Thesis Project
RMA Linguistics and Communication

UNIVERSITY OF AMSTERDAM

July 2022

ABSTRACT

Keywords: sensorimotor knowledge, neural networks

This paper explores how sensorimotor knowledge, the learned connection between the motor system and speech production and perception, can be included in a bidirectional neural network model of speech. The training of the network consists of exposure to a large amount of pairs of sound and muscle activity. Our network uses a simplified configuration of four muscles that move the tongue up, down, back and front. The network can perceive and produce vowels, consisting of two formants. By inputting only sound or only muscle activity to the trained network, we mimic perception and comprehension. As anticipated, the acquired sensorimotor knowledge serves successfully as a connection between the muscle nodes and the formant nodes. In production, we observe that the produced vowels fall within clusters (or categories). This could be seen as sticking to preferred articulatory gestures, reminiscent of human strategies and phenomena such as the perceptual magnet effect – but in production.

1 INTRODUCTION

Previously, neural networks have been used to model multiple phenomena, such as phonological feature emergence and auditory dispersion in a manner compatible with the Bidirectional Phonology framework (BiPhon) (Boersma *et al.* 2021, 2020). This compatibility means that the networks, besides being able to showcase the phenomena, would be able to learn and use the “knowledge” both for production and comprehension. Further, the processing of the input would be parallel and allow for each stage of the processing to influence other stages non-chronologically.

These concepts lie in the basis of BiPhon, a model of speech that covers all necessary stages of speech production and comprehension, from the mapping of sound to phonological features, to morphological and lexical processes (Boersma 2011). Figure 1 showcases the structure of the model. Production occurs when a speaker works their way through the stages from top to bottom, whereas comprehension happens from the bottom to the top. We see that what connects the *auditory form* to the *articulatory form* is sensorimotor knowledge – the focus of the current paper. The auditory form is the representation of produced or perceived sound and the articulatory form refers to the gestures the articulators and the activity of the necessary muscles that we use to produce the sound. What connects the two forms is the knowledge we acquire as infants about the effects of particular articulatory gestures on the sounds coming out of our mouths.

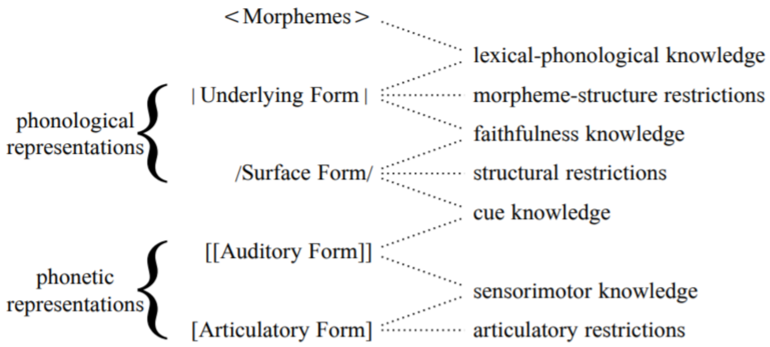


Figure 1: Model of Bidirectional Phonology and Phonetics

In the context of the neural network this knowledge is represented by the connections between the layers of neurons (the nodes). Some of these connections become stronger and some weaker during the learning phase. Strong connections between neurons means that if one of them activates, the other one is likely to activate as well. This concept is comparable with the physiological processes behind the building of synaptic connections between neurons in the human body and their activation. How our network learns from the input it receives is presented in detail in Section 4.

An example of implementation of sensorimotor knowledge in a neural network is the DIVA model by Guenther and Vladusich (2012). This model differs from BiPhon as it is not bidirectional – it models only acquisition and production but no perception. Thanks to the bidirectionality and parallelism our model exhibits, "later" events can influence "earlier" events, essentially providing feedback based on which the eventual output can be adjusted. Examples of this are how the mapping from morphemes to underlying forms can be influenced by phonotactic biases in productions (Boersma and Van Leussen 2017) and the perceptual magnet effect observed in Boersma *et al.* (2021).

Guenther and Vladusich (2012) satisfies the need of feedback by employing a feedback control system, which includes somatosensory and auditory feedback. The need for auditory feedback is straightforward. However, somatosensory feedback is a more complex subject. Guenther and Vladusich (2012); Gallese and Lakoff (2005); Rizzolatti and Craighero (2004) discuss *mirror neurons,* which are said to fire both during perception and production. Their existence supports theories that we are able to comprehend speech (or other actions) *because* its motor representation activates in our brains (Rizzolatti *et al.* 2001, 661). In short, when we perceive speech, the motor commands that would result in us producing the same sound are activated. This activation provides feedback and the sound we hear might be influenced by the activated motor commands. This principle is the basis of the Motor theory of speech perception (Liberman and Mattingly 1985). However, there is empirical evidence against mirror neurons operating according to the claims (Hickok 2009).

Nonetheless, somatosensory feedback should be given consideration as part of a comprehensive model of speech because often we can observe situations where physical articulatory restrictions have an almost immediate effect on the motor commands sent to the articulators. An example of such a situation would be during a *bite-block experiment.* Fowler and Turvey (1980), and subsequently, Gay *et al.* (1981), showed that when speakers are biting down on a block, they succeed in approximating the acoustic characteristics of target vowels as spoken normally by reorganizing their articulation. A more common example (albeit anecdotal) is having hot food in your mouth, yet still managing to produce recognizable vowels by exaggerating certain articulatory gestures.

Thanks to the bidirectional and parallel qualities of our network, the mapping of *intended articulation* to *resultant articulation* can influence an earlier event during production – the mapping between *intended sound* and *intended articulation,* as it is illustrated in Fig. 2. Further, auditory feedback is also part of the model, as *resultant sound* can influence the *intended sound* also thanks to the parallelism of the network.

The physiological mechanism behind the somatosensory feedback in our network can be explained without the necessity for mirror neurons. *Muscle spindles* are receptors, found in muscles, that are sensitive to change in muscle length and velocity (Macefield and Knellwolf 2018). They contribute to the proprioceptive sense of the tongue, as

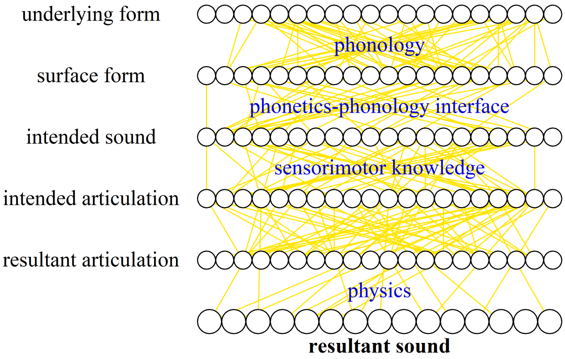


Figure 2: Sensorimotor knowledge integration within BiPhon

found by Cooper (1953). Borden (1980) and Sandyk (1981) find that the feedback that is received through the muscle spindles contributes at the very least to filtration and correction of disturbances to the tongue, possibly playing a role mostly during the speech acquisition process. This proprioceptive sense is what makes it possible for the *resultant articulation* to have an effect on the *intended articulation* during production as depending on how well the actual movement of the muscles goes (as it might be influenced by outside factors as discussed), the speaker is able to change their target articulation.

Bruderer *et al.* (2015) found that the impediment of the articulators also affects perception, specifically in pre-verbal babies. When the infant participants were given a teether chew toy that blocked the movements of the tongue, they were less good at recognition of non-native alveolar sounds. This effect is compatible with our network – as the phoneme is non-native, in order for the child to acquire it, it would need to match muscle (articulator) activation with resultant sound, just like the network. If that is impeded, recognition would not be successful.

The following sections detail how these theoretical considerations are implemented in our network. Further, we showcase the process of our network acquiring knowledge during its "babbling" or learning stage and utilizing it to attempt producing and perceiving vowels, similarly to what a human infant learner goes through on their journey to becoming a speaker.

2 STRUCTURE OF THE NETWORK

The network that can be used to simulate the acquisition and use of sensorimotor knowledge according to the bidirectional model of speech has at least four layers. These layers represent only a part of the larger comprehensive model of speech that is BiPhon. However, in order to showcase how the brain creates connections between the intended sound and the movement of the muscles, we need minimally the intended sound and intended articulation layers and the resultant sound and articulation layers.

The following subsections detail the roles and specifics of the layers.

2.1 *Resultant sound layer*

As previously discussed this layer stands for the actual sound that results from the articulation or is heard by the speaker – the auditory form of the utterance. Given the bidirectional character of the network, this layer does not only represent the auditory output of the network (the produced speech) but also the auditory input (the perceived speech). As in Boersma *et al.* (2021), the layer is meant to depict the basilar membrane of the human ear as specific areas of the membrane react to specific frequencies, similarly to activation appearing in a specific area of the layer when there is auditory input to the network. In order to represent the relation between the frequencies and the membrane activation linearly, the frequencies are in an Equivalent Rectangular Bandwidth scale (ERBs).

The scale of 4 to 28 ERBs is denoted by 49 nodes, each 0.5 ERBs apart from the previous. As in Boersma *et al.* (2021), activation on this layer is shown by two Gaussian bumps corresponding to the first two formants, F1 and F2, of the vowel produced or heard.

2.2 *Resultant articulation layer*

The resultant articulation relates to the actual movements a speaker would perform with their articulators. As discussed, this might differ from the intended articulation due to factors which may inhibit the proper use of the articulators, for example physical obstruction of the tongue. This layer consists of four nodes – each representing one of the muscles of our toy anatomical configuration. The nodes can be fully activated, which would be represented by a red disc filling out the node, or less activated, seen as a smaller red disc.

2.3 *Deep layers*

As observable in the figures, the two input/output layers described above are connected to other layers positioned above them. They have an arbitrary number of nodes, equalling 50. This quality of the network classifies it as "deep" and allows it to learn statistically from the data it is fed as activity spreads across all the layers. The goal of our network is to learn that a particular muscle movement (represented by a certain degree of muscle activation) results in a specific sound being produced or, in short, to acquire sensorimotor knowledge. As it is with human babies, it begins with babbling – testing out the relation between articulation and sound. For our network this means being fed pairs of muscle activation and the sound that results from it. Therefore, we activate and *clamp* one or two muscle nodes to a random level and we activate and *clamp* activation on the resultant sound level. *Clamped* means that no matter how the activity spreads through the network, this nodes will remain activated as we inputted it. Their activation will spread to the middle level through the connections. Further, the activations on the middle layer will spread to the top layer. Then, the activation will spread back down to our input levels. This will repeat enough times for the network to settle.

This process results in a network that is bidirectionally knowledgeable. This means that it can utilize the connections between the layers that have been established during the learning to process production *and* perception. In our case, it does not matter whether we input muscle activity for *production* or audio input (formants) for *perception*, the network will be able to give us correct output. This is reflective of the fundamental principle of the theoretical model behind the network. Bidirectionality within the model is also tied with parallelism – the ability of events during the processing to influence each other non-chronologically (Boersma 2009). Letting the network settle allows for the connections between the layers to be re-adjusted multiple times based on what connections occur between other layers – just how the mapping between auditory form and phonological form can be influenced by the lexicon during human perception.

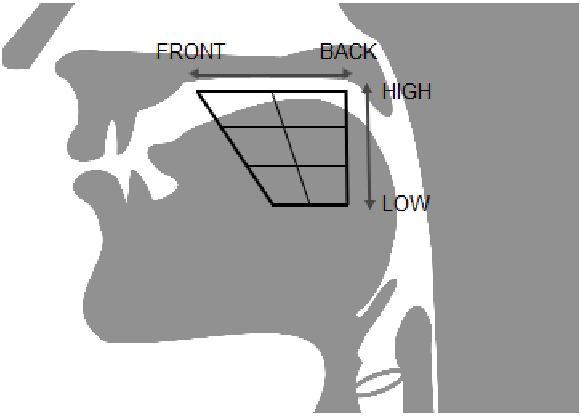


Figure 3: The connection between the toy muscles, the tongue position and the vowel position

3 SIMPLIFIED MUSCLE REPRESENTATION

3.1 *Muscles*

For our first steps towards a network that showcases how we can implement sensorimotor knowledge, we employ a toy muscle configuration and we only focus on vowel production and comprehension. The muscles we include in this model have very simplified functions compared to their anatomically correct counterparts. Instead of the styloglossus, for example, that elevates and draws the tongue back, we opt for two different muscles – one that only pulls the tongue up and another that only pulls the tongue back. Similarly, the two functions of the genioglossus are also given to two different muscles – one that pulls the tongue down, and another that pulls the tongue to the front.

This leaves us with a configuration of four muscles. Their activity is enough to provide us with information about the position of the vowels (front, back, high, low). Consequently, this information allows us to determine the first two formants of the vowels (*F1* and *F2*). The connection between the job of each muscle, the position of the tongue and the vowel position within the trapezium chart is illustrated in Fig. 3.

The relation between the activity of the muscles that control the height of the tongue and the value of the first formant is rather straight-forward. The values for the muscles activation can vary between 0 and 1, where 0 is not activated and 1 is fully activated. In the figures this is represented by the size of the red discs within the activated nodes. A full activation of the muscle that pulls the tongue up would result in a high vowel with a low F1. Respectively, full activation of the muscle that pulls the tongue down results in a low vowel with a high F1. If both muscle nodes are not activated the result is a mid vowel. This relation is expressed through (1):

$$F1 = 10.18 + 3.82 * (muscleLow - muscleHigh) \tag{1}$$

The principle is that if there is no muscle activation, the value will be 10.18 ERBs – precisely in the middle of the range between the highest and lowest possible F1 values. The corner vowel /u/ has the F1 value of 14 ERBs and the lowest vowel /ɑ/ has an F1 of 6.36 ERBs. Would it be that muscleLow is fully activated and the tongue is pulled down, the value would be 14 ERBs, the highest possible of the range and resulting in a low vowel.

How the activation of the muscles that control the frontness or backness of the tongue affect the second formant is more complicated. Due to the physical restrictions posed on our articulators, the F2 range within low vowels is smaller than within high vowels. This is reflected in the shape of the vowel trapezium chart. As a consequence, in order to calculate the correct F2 value based only based on the muscle activity input, we need to take into account the activation of all muscles. The formula we use, (2), reflects this relation.

$$\begin{aligned} F2 = (17.9 * (0.85 + 0.15 * (0.5 + 0.5 * (muscleHigh - muscleLow)))) \\ +(6.39 * (muscleFront - muscleBack) \\ *(0.59 + 0.41 * (0.5 + 0.5 * (muscleHigh - muscleLow)))) \end{aligned} \tag{2}$$

The first line of the equation establishes the middle point of the range of F2 values, as this changes with the change in height of the vowel. This is observable in the incline of the middle line of the vowel trapezium (see Fig. 3). The second line calculates how much would be added or taken from the middle point value depending on the muscle activity, similarly to the way we calculated F1. The last line modulates this addition to the middle point once again depending on the height of the vowel.

Having established what our network will be learning, namely mastering the relation between the activity of its four muscles and how this translates to sound, we can proceed with the learning process.

4 LEARNING PROCESS

Before the learning process begins, all network parameters are set to zero – all connections between the layers are set to zero (they are nonexistent) and all the activities of the nodes are also set to zero. This is the beginning of language acquisition and our virtual infant has no knowledge of language at all yet. From here on, just like real infants, it will learn based on the statistical distribution of the input it receives. This happens through the learning steps described below. We provide all of the formulas required to execute the learning phases, making replication of our simulations possible for the reader.

4.1 *Learning algorithm*

First, we apply input, which can be sound, muscle movement, or both. Sound is inputted to the first 49 nodes of the input layer, which represent the basilar membrane of the ear (see Section 2.1). Muscle movement is inputted to the 4 last nodes on the input layer, which represent the muscles (see Section 2.2). If inputting both, the input is activation values for all 53 input nodes. The number of activation values (K) during training in our case is always 53 as we input sound paired with the correct muscle movement as discussed below. The input level activities is x_k.

As mentioned in Subsection 2.3, our network learns in a very similar manner to a human child. By inputting random muscle activations and the correspondent formant values as calculated by (1) and (2), we mimic the babbling stage of an infant. Where a human child would try out vocalising and moving their tongue around or opening and closing their lips to see how that would change the sound that comes out of their mouth, our network learns how different tongue positions (as defined by the activation values of the four muscles) affect the two formants of the resultant vowel.

The weights of the connections between the nodes in the different layers gradually change in accordance to the learning algorithm proposed for Deep Boltzmann machines (Salakhutdinov and Hinton

2009; Goodfellow *et al.* 2016). This algorithm fulfils our requirements – first, it makes use of the Hebbian learning principle and second, it is symmetrical. Hebbian learning (Hebb 2005; Kohonen 1984) means that the weights of the connections between nodes that fire together are strengthened. The symmetry is needed in regards to the bidirectionallity of our network – we want nodes to affect each other equally in both directions so that the same connection weights can be used from production to perception and the other way around.

Even though the *inoutstar* algorithm used in Boersma *et al.* (2020) also covers the requirements, it was found to yield less robust results when used in the neural network in Boersma *et al.* (2021). Since the structure of this network strongly resembles our network structure, we also made use of the Deep Boltzmann machine algorithm.

The learning process has four phases, through which we go below. All of the equations are taken from Boersma *et al.* (2021).

4.1.1 Initial settling

During this first stage, the input layer activities x_k are prevented from being changed by the spreading of activation – the input nodes are clamped. The activation are then spread to the middle level, which has activities y_l, as L is the number of nodes in that level. The top level activities are marked as z_m (where M is the number of nodes of the level) and still equal zero as the activation has not spread there yet.

$$y_l \leftarrow \sigma(b_l + \sum_{k=1}^{K} x_k u_{kl} + \sigma(b_l + \sum_{m=1}^{M} v_{lm} z_m \tag{3}$$

where $\sigma()$ is the logistic function

$$\sigma(x) := 1/(1 + exp(-x)) \tag{4}$$

In (3) u_{kl} is the weight of the connection from the bottom layer to the middle layer and v_{lm} is the weight of the connection between the middle layer and the top layer. The function of equation (3) is to calculate the activation of the middle layer node by adding the sum of all contributions of activity of the nodes connected to it from the input level and the top level. The effect of the activity of a connected node is relative to the weight of its connection – the nodes connected through connections with more weight have a bigger influence and vice versa. The logistic function keeps the outcome between 0 and 1. With the following equation, we spread the activity to the top layer (c_m is the bias of node *m*):

$$z_m \leftarrow \sigma(c_m + \sum_{l=1}^{L} y_1 v_{lm} \tag{5}$$

The calculations with equations (3) and (5) are performed ten times, which leads to a near equilibrium state of the network.

4.1.2 Hebbian learning

The weights of the connections between the nodes change according to the following equations, which represent Hebbian learning (Hebb 2005).

$$a_k \leftarrow a_k + \eta x_k \tag{6}$$

$$b_l \leftarrow b_l + \eta y_l \tag{7}$$

$$c_m \leftarrow c_m + \eta z_m \tag{8}$$

$$u_{kl} \leftarrow u_{kl} + \eta x_k y_l \tag{9}$$

$$v_{lm} \leftarrow v_{lm} + \eta y_l z_m \tag{10}$$

The symbol η stands for the learning step, which, in our case, equals 0.001. The principle of this learning is that whenever two nodes are activated at the same time, the strength or *weight* of their connection increases, leading to them being active together again in the future.

4.1.3 Dreaming

During the *dreaming* phase, the still *clamped* input layer is unclamped. This means that now the spreading of activation and the change in the weights of the connections can influence the input layer activations and change them. This happens by first allowing the activation to spread to the input layer from the middle layer, which happens through this equation (where a_k is the bias of the input level node):

$$s_k \leftarrow a_k + \sum_{l=1}^{L} u_{kl} y_l \tag{11}$$

The new activities for the middle and top layer are calculated to have a Bernoulli random distribution. This results in random connections that the brain of our virtual infant ”imagines”. This happens through the following equations:

$$z_m \sim \mathscr{B}(\sigma(c_m + \sum_{l=1}^{L} y_l v_{lm})) \tag{12}$$

$$y_l \sim \mathscr{B}(\sigma(b_l + \sum_{k=1}^{K} x_k u_{kl} + \sum_{m=1}^{M} v_{lm} z_m)) \tag{13}$$

The dreaming-like spreading is repeated 10 times. The randomness of the dreamed-up activations and the specificity of the real input ensure that the sample the network is exposed to represents a faithful distribution of possible patterns (Boersma 2019).

4.1.4 Anti-Hebbian learning

Finally, through a process that is more or less the opposite of the Hebbian learning step, the network will forget part of the imagined knowledge it acquired during the previous phases. Equations (6) to (10) make sure our virtual brain acquires new knowledge and the following equations ((14) - (18)) ensure that some knowledge is forgotten. When the network has acquired what it could from its environment (the input), equilibrium is reached through the cancelling out of the Hebbian learning phase and the anti-Hebbian learning phase.

$$a_k \leftarrow a_k - \eta x_k \tag{14}$$

$$b_l \leftarrow b_l - \eta y_l \tag{15}$$

$$c_m \leftarrow c_m - \eta z_m \tag{16}$$

$$u_{kl} \leftarrow u_{kl} - \eta x_k y_l \tag{17}$$

$$v_{lm} \leftarrow v_{lm} - \eta y_l z_m \tag{18}$$

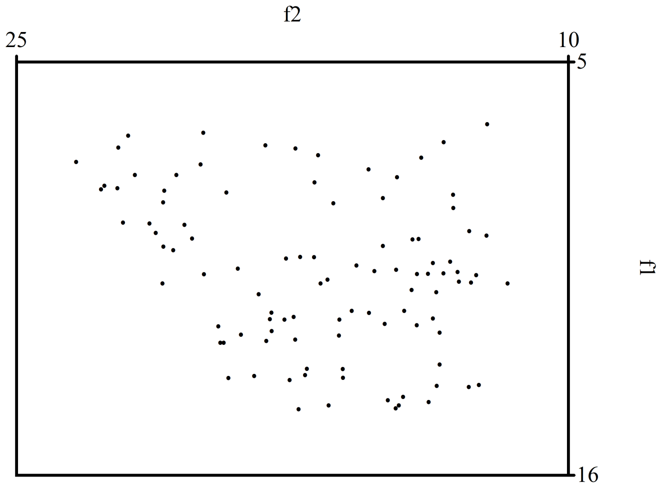


Figure 4: A hundred of the random vowels inputted during training, values in ERBs

4.2 *Learning outcome: a trained network*

These four processes of learning occur for each of the 10,000 inputs of formant values and muscle activations that we feed the network. An example of a piece of training data would be full activations of nodes 51 and 52 of the bottom layer, which represent the “up” and “front” muscles, paired with the corresponding activations of bottom layer nodes located around 6.3 ERBs, or around node 6 and around 24.3 ERBs or around node 42. Figure 4 showcases the placement of a 100 out of the 10,000 inputted vowels. It is observable that the distribution follows the shape of the vowel trapezium chart, as to be expected.

The comparison between the network in its initial stage and after the training shows the newly-developed connections between the nodes and layers. The first picture is of a completely untrained network, the second is of a network trained with only a 100 pieces of data, followed by 1000 pieces of input and finally the full 10,000 pairs (Fig. 5). The first picture is comparable to a new-born infant and the last picture shows a fully matured brain, which has established sensorimotor connections between the production of sound and the movement of muscles.

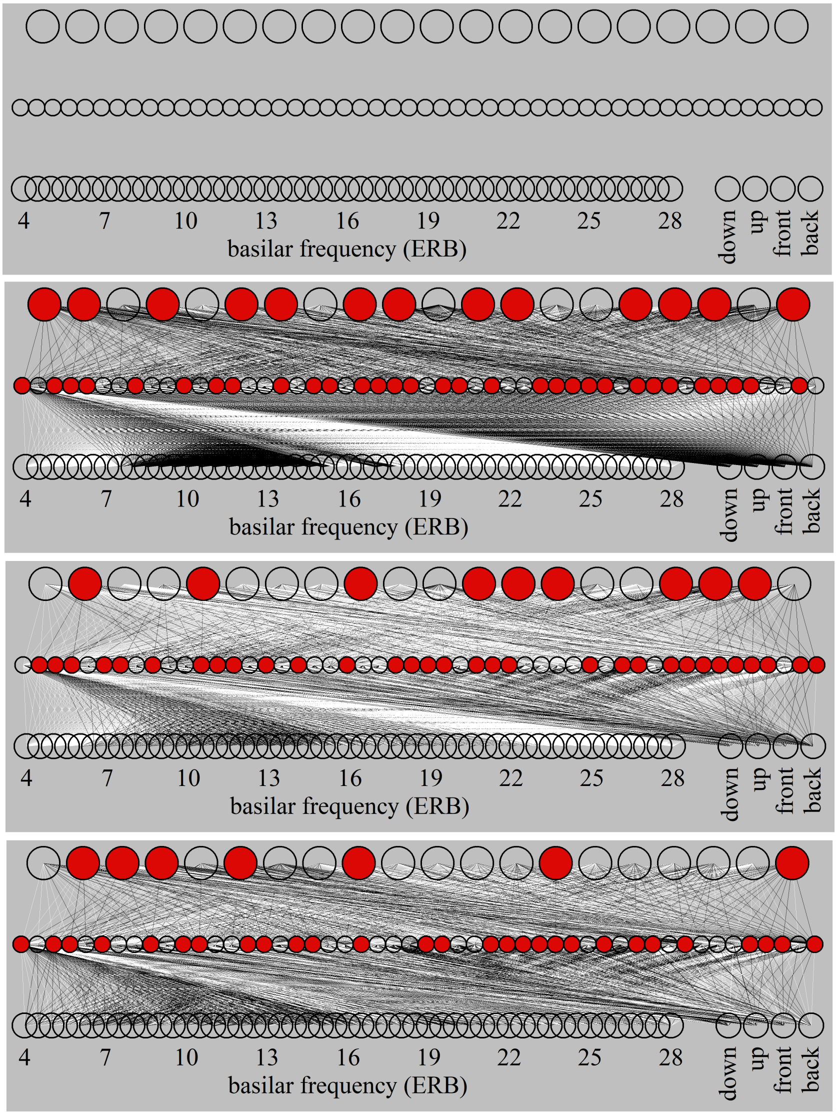


Figure 5:
The network before training and after a 100, a 1000 and 10,000 pieces of training input data

5 RESULTS

After the network has been exposed to the pairs of muscle activity and formants, has learned from them and weighted connections between the layers have been established, we can proceed to test its newly-developed sensorimotor knowledge. In order to establish that the network has learned successfully, we would need to prove that it can produce correct output if it is fed only audio *or* only muscle input. This would show whether the network has learned bidirectionally and is equally good at perceiving as well as producing.

5.1 *Testing production*

We begin by setting the muscle input to the values that correspond to certain vowels, as this allows us to predict what the formants outcome is supposed to be and compare. First, we input high muscle activation of the front and high muscles – aiming to illicit the production of a vowel /i/ with F1 of 6.35 ERBs and F2 of 24.29 ERBs, according to (1) and (2). The upper part of Fig. 6 showcases the result. We observe a higher F1 and a lower F2 than expected. The lower part of Fig. 6 represents the result of the same input with the difference that the muscle layer remained *clamped* during the spreading of the activation among all the layers. Normally we allow for the input to be altered by the activation resonating through the network, sometimes resulting in the network thinking that it "heard" (in the case of perception) different input from what it was fed, which we described as the *dreaming* learning phase. This is in accordance with the parallelism concept and it allows us to account for phenomena such as the Ganong effect or the perceptual magnet effect (Kuhl 1991). By keeping the bottom layer clamped and not allowing for the dreaming to happen, we force the network to show us the exact input activation spreading.

In the case of Fig. 6, the clamping did not change the outcome, but in Fig. 7, we observe a noticeable effect.

The original values for a low back vowel are F1 = 14 ERBs and F2 = 11.5 ERBs, as according to (1) and (2). What we observe in the upper part of Fig. 7 where we leave the muscle input nodes unclamped, is that now they are no longer fully activated, as our input dictates. After the dreaming, the network now thinks that the activation is about mid-level, as seen by the smaller red discs within the activated nodes.

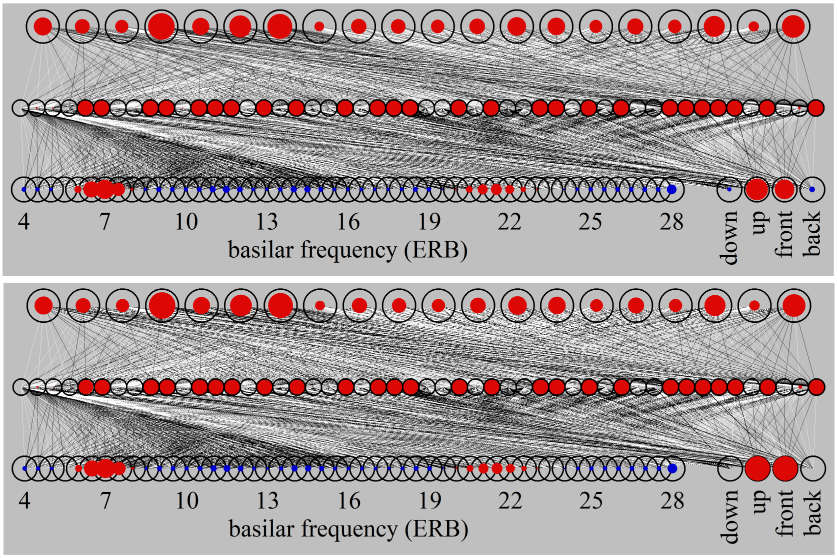


Figure 6: Output, when the input is high activations of the high and front muscles, unclamped (above) and clamped (below)

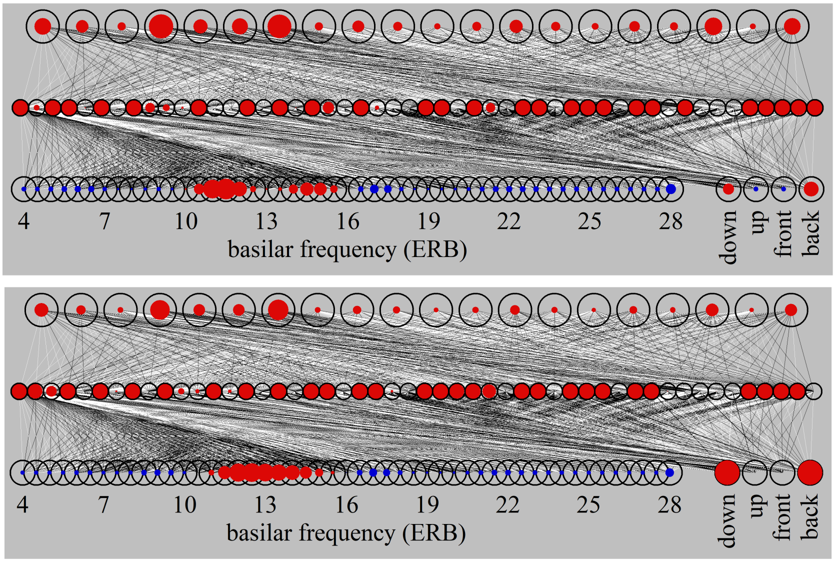


Figure 7: Outcome after inputting high activations of the low and back muscles, unclamped and clamped

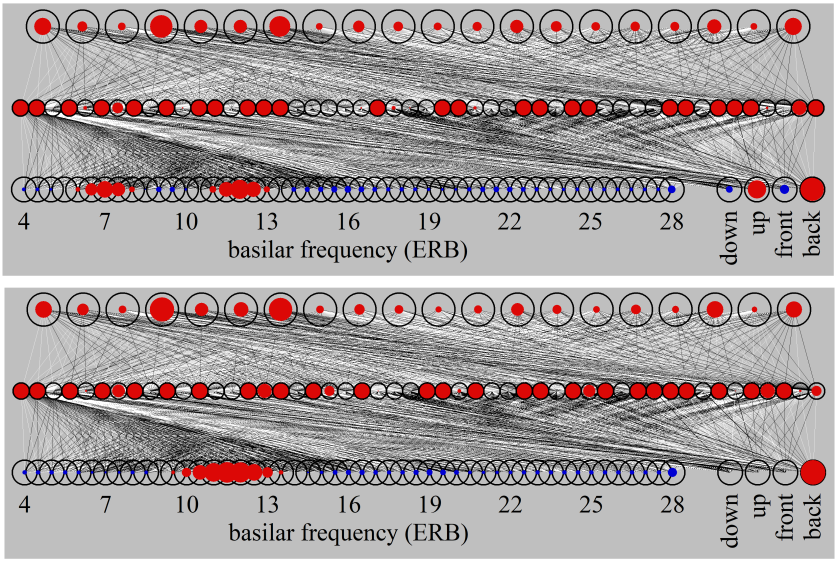


Figure 8: Outcome after inputting high activation of the back muscle and leaving the high and low relaxed, unclamped and clamped

This results in peaks on our "basilar membrane" located around 12 ERBs and 15 ERBs. This is the result of the lower level of muscle activation – the network is producing a less back and less low vowel than expected due to the changed muscle input. The outcome is that the F1 of the produced vowel almost equals the F2 of the goal vowel and the F2 of the produced vowel almost equals the F1 of the goal vowel. Even though we only see the peaks of the activity on sound level without overt sign which peak corresponds to which formant, we know that the muscle activation that changed to mid-level results in a F1 lower than F2.

The change in the inputted muscle activity level could be seen as limitation of very effortful gestures which occurs in real speakers as well. The clamping of the nodes results in the "forced" production of the (more) correct formant values, as seen in the bottom part of Fig. 7. We observe the result of clamping the muscle nodes and having them remain fully activated. The produced formants merge around 13 ERBs – it appears the network still does not allow for F1 to be higher than F2 and they merge around 13 ERBs. Nonetheless, this is much closer to the goal for both formants than when left unclamped.

We also observe surprising output if we input muscle values corresponding to a mid-back vowel as in Fig. 8.

We see that without the clamping, the network ends up producing a high vowel, rather than a truly mid vowel as inputted. The clamping of the bottom layer solves this issue and now we see activation only in the back muscle node. Without the activation of the high muscle induced by the dreaming of the network in the unclamped version, we see that the F1 is now higher, as it should be for a mid vowel. It has almost merged with the F2 around 11.5 - 12 ERBs.

Finally, in order to showcase the effect of only partial activation of a muscle node, we input low activation to the high muscle node, making the vowel slightly higher than mid, and full activation to the front muscle node, producing Fig. 9. This should result is F1 and F2 being affected by the specific height of the vowel.

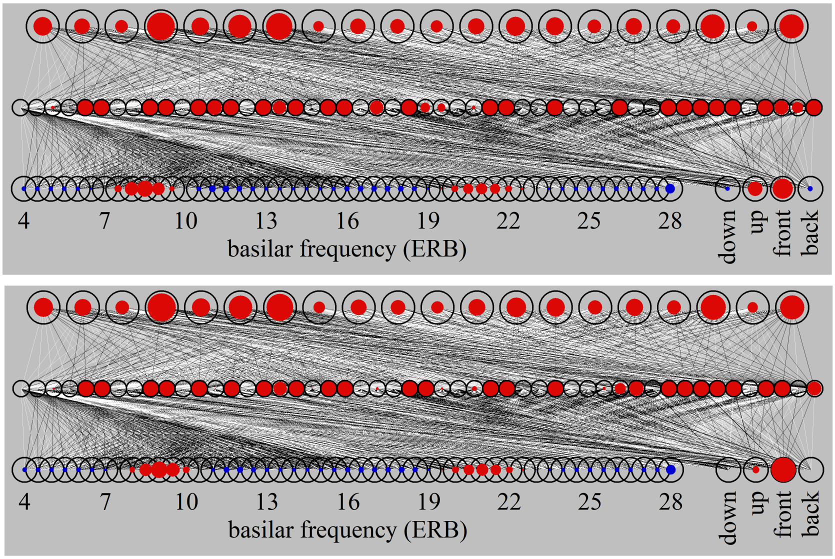


Figure 9: Outcome after inputting high activation of the front muscle and low activation to the high muscle, unclamped and clamped

As we can see, there is a direct effect of the level of activation of the muscle nodes on the formant levels. Without the clamping, the network did produce a slightly higher vowel than when the bottom layer was clamped.

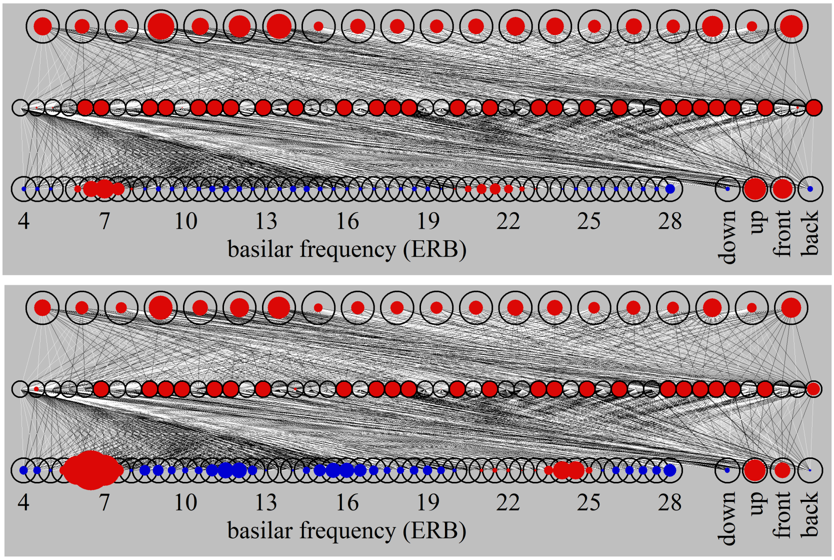
Generally, the network is able to produce sounds based on the muscle activation we input, albeit often altered according to what it “thinks” it should produce as we leave the input layer unclamped.

5.2 *Testing perception*

To test perception, we input sound to the trained network and observe which muscles activate and how much. The aim is, of course, that inputting formant frequencies that we know correspond to specific muscle activations (according to (1) and (2)) would activate the required muscle nodes.

First, if we input the ERB values characteristic to a maximally high front vowel, we expect full activation of the muscles in question. Figure 10 showcases the outcome.

Figure 10: Outcome after inputting formant values 6.3 ERBs and 24.3 ERBs, unclamped and clamped



As we can see, the results are as expected. When the bottom layer is held unclamped, the network ends up “thinking” that it has heard a slightly less high and front vowel, similarly to how the network produces a slightly less high and front vowel when inputted full activation of the relevant muscles. The clamping slightly changes the results, by actually yielding less activation of the front muscle.

Next, we apply the values of 10 ERBs for F1 and 18 ERBs for F2, which according to our formulas, should correspond to a central vowel, produced without the engagement of any of the muscles. Figure 11 shows the outcome first with the bottom layer unclamped and then – clamped. The result is no activation of the muscle nodes, which signifies relaxed muscles as wanted.

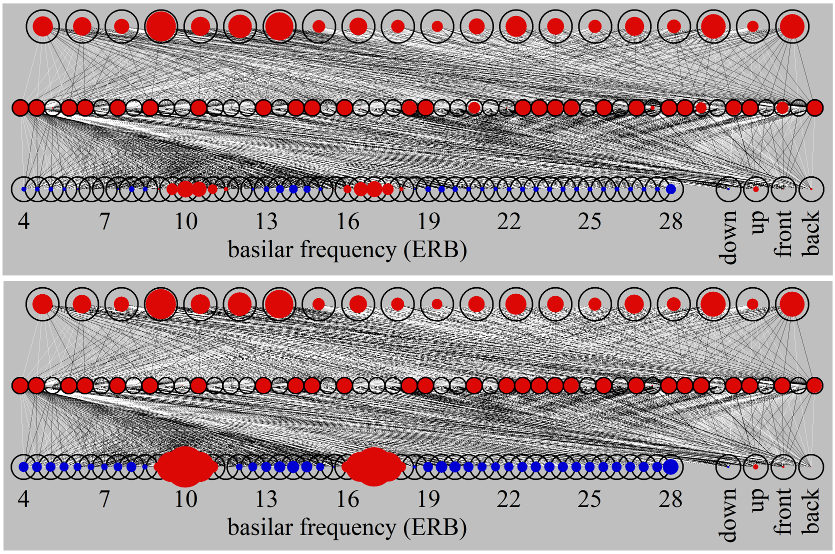


Figure 11: Outcome after inputting formant values 10 ERBs and 17 ERBs, unclamped (above) and clamped (below)

We also test whether the network has picked up well on the fact that low vowels are restricted in how front they can be. We input the lowest possible F1 of 14 ERBs and F2 of 20.5 ERBs, which is the maximum at this height, according to (2). If the network has learned well, we expect to see complete activation of the front muscle, rather than partial.

Figure 12 showcases the desired result, especially when unclamped – the front muscle node is fully activated.

Overall, the network shows that it has effectively acquired knowledge that allows it to perceive and spread information about the perceived throughout the layers correctly.

5.2.1 Clustering in production

The behaviour we overall observed, specifically in production, can be seen as muscle movement categorization, similarly to phonetic categorization. The network appears to group some movements together, as it can be observed in Fig. 13. The plot shows vowels, produced by the trained network when it was given a 100 different muscle input combinations, varying in level of activation and in which muscle is ac-

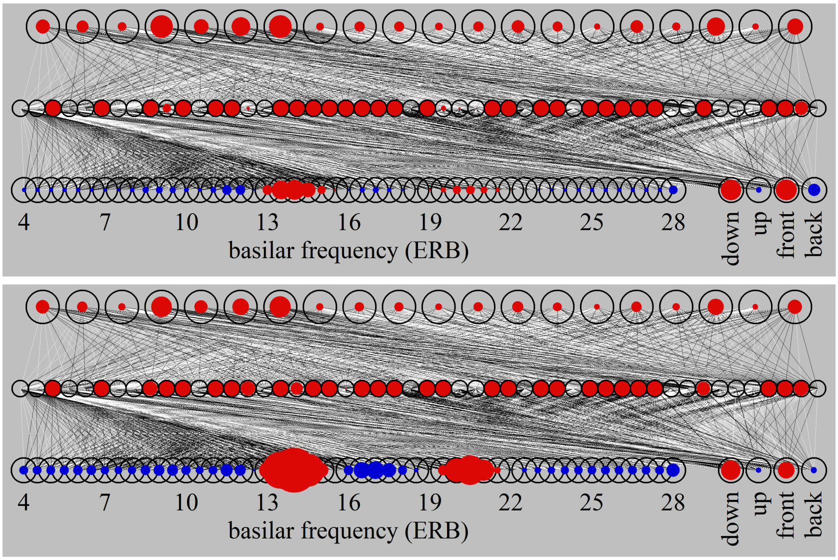


Figure 12: Outcome after inputting formant values 14 ERBs and 20.5 ERBs, unclamped (above) and clamped (below)

tivated. The red circles represent a 100 vowels the network produced when the random input was clamped and the black crosses represent a 100 vowels that the network produced when the same random input was unclamped. We observe that clamping results in a more spread out distribution, which is expected as the network is "forced" to produce a vowel according to the input, even if when left unclamped, it would alter the muscle movement input to a preferred one (perhaps an "easier" one). Nonetheless, in both cases we see very apparent clustering of the produced vowels into 5 categories.

Even though it showcases a 100 results from random muscle input, due to the fact that many data points overlap, the plot appears less dense than Fig. 4. Furthermore, the ERB formant values differ by steps of 0.5 as they relate to the most strongly activated nodes on the audio input/output layer, which are 0.5 ERB apart, which creates a less scattered plot. In order to visualise the data more efficiently, we interpolate the peak values that constitute the data points – they are the peaks in the vector, comprised of all x_m activities of the nodes of the bottom layer (m = 1 to 49, since we only take the sound input part of the layer). Through a parabolic interpolation formula, we obtain the new data points that do not overlap as in Fig. 13. Once again, red circles represent the 100 vowels when the input was clamped and

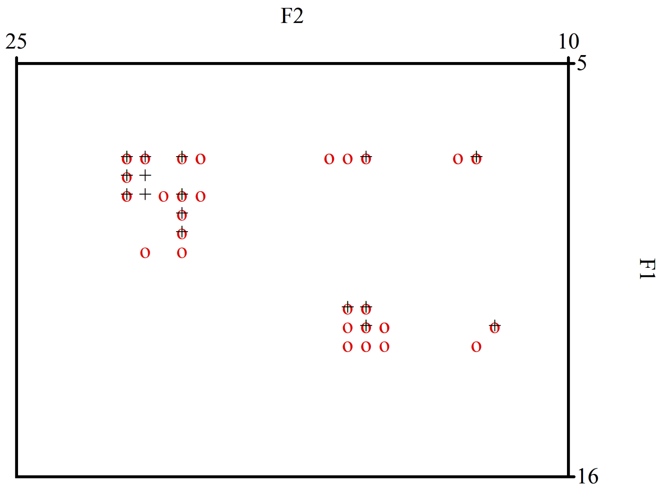


Figure 13: A hundred random vowels as output of the trained network when inputted with a 100 random muscle activations

the black crosses represent the vowels when the input nodes were unclamped.

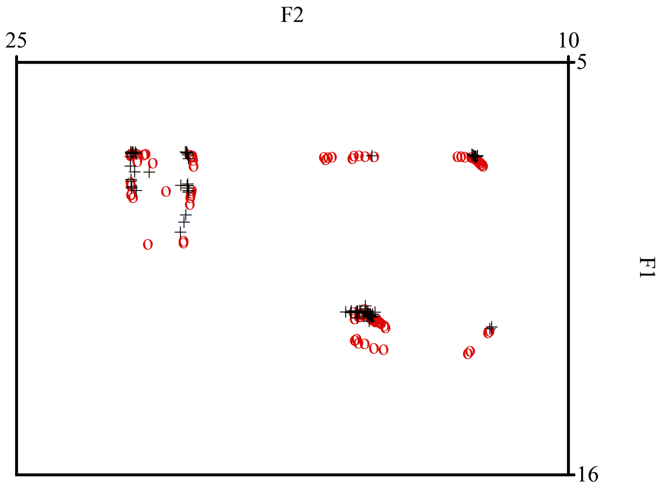


Figure 14: Interpolated output formant values of a hundred random inputted muscle activations

In order to compare the produced vowels to the vowels that the network has been trained on, we plotted together the data from Fig. 4 and the interpolated data from Fig, 14.

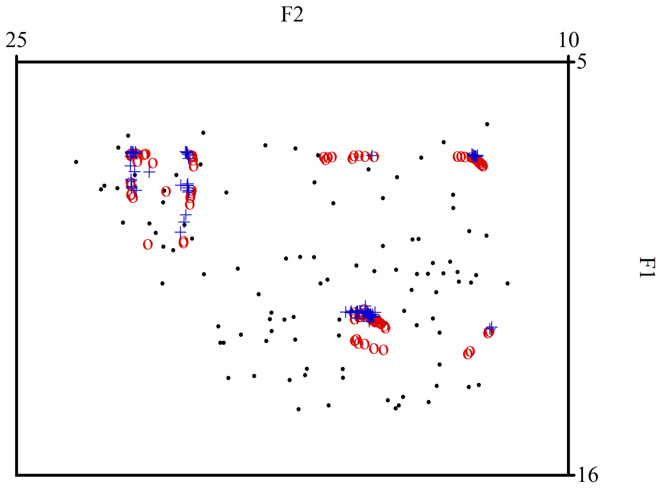


Figure 15: A 100 vowels used for training overlapped with a 100 interpolated values of produces vowels (clamped and unclamped)

This combined figure, Fig. 15, clearly showcases that even though the network was exposed to a randomly distributed vowels, it learned to produce vowels within categories. The black dots represent the vowels, used for training, the red circles represent a 100 random produced vowels with the input clamped, and finally the blue crosses are for a 100 random produced vowels with the input unclamped.

To further analyse the production process, Fig. 16 shows the difference between the expected formant output given each level of muscle activation (0,5 increments). More specifically, we input different levels of activation of the muscle that pulls the tongue down, as at the same time we keep the back muscle clamped at maximum.

It is observable that after the 3.5 point of muscle activation, the formant output stagnates around 11.5 ERBs. This is reminiscent of the *the perceptual magnet effect*, normally a phenomenon seen in perception. Here, in production, the network does not discern between muscle activation above 3.5 and the maximum, in this case. Related to real life, this could manifest at the inability to produce a foreign sound that is drawn towards a familiar prototypical sound and, therefore, is not properly perceived.

By plotting the *articulatory parameter* that is the movement of the toy "low" muscle of the tongue, against *the perceptual parameter* – that is, how front or back the vowel is, in Fig. 16, we approach the pro-

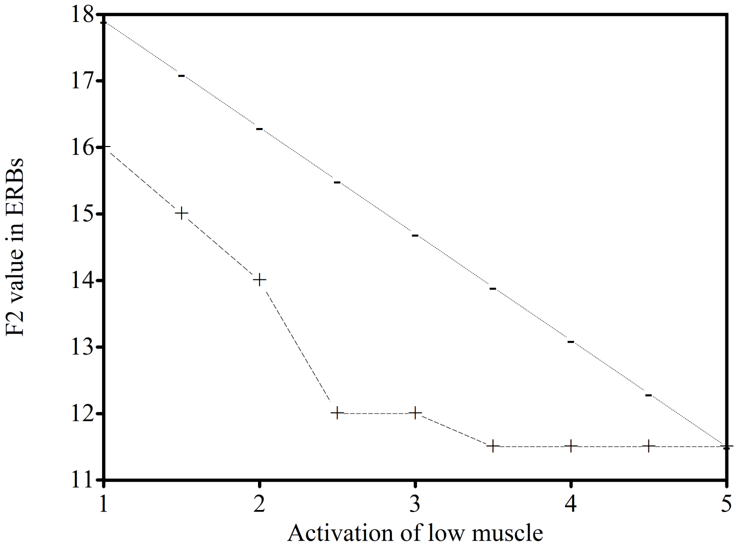


Figure 16: Resultant (dashed line) vs expected (dotted line) formant outputs when the low muscle is manipulated

duction similarly to how the Quantal theory of speech (Stevens 1989) approaches articulation. Interestingly, the slope of the outcome values resembles the visualisations that Stevens provides. If we look at Fig. 17, we see that there are sections with less steep slope, namely I and III.

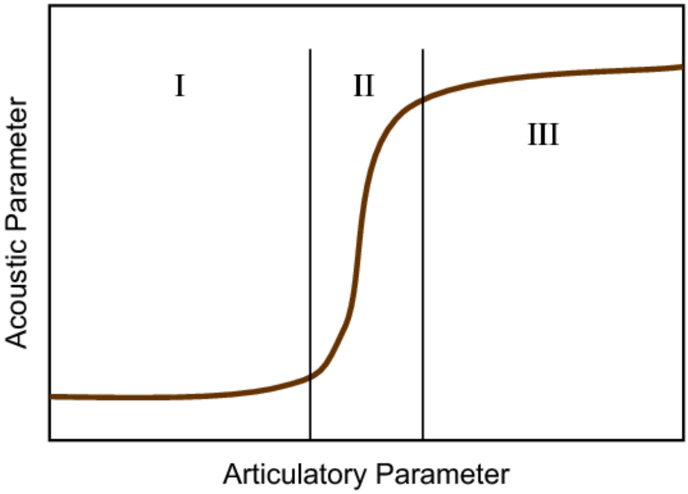


Figure 17: Having the articulatory parameter on the y axis and the perceptual parameter on the x axis, this graph showcases the different slopes of the line, (Stevens and Keyser 2010)

According to the theory, those sections are more stable as changes in the articulatory parameter have a smaller effect on the perceptual parameter. Therefore, they are preferred by the speakers of the world. Our network also seems to prefer them, as we can observe from the strong clustering in Fig. 13.

6 DISCUSSION

The aim of this project was to explore how we can include the sensorimotor knowledge and connections in a bidirectional neural network as a model of speech. This was done by first establishing a structure for the network with an input/output bottom layer and deep layers. Then, we describe the learning algorithm that allows the network to acquire knowledge from the large amount of random input it receives during training.

Following its training, we tested both the perception and production as it is a bidirectional network. The results showed that the network is capable of producing output based on its acquired knowledge. However, without clamping the bottom layer, the network will often “mishear” the input we use to elicit the spreading of activation. A deeper look into the data shows us that the network will produce vowels only from a few specific clusters, thus categorizing its output.

Thus, the outcome of the project is positive and curious. Sensorimotor knowledge can be included in a BiPhon neural network, enriching it with new outcomes.

Many different directions could be taken to take this project further. The toy muscle configuration we worked with can be expanded and made more anatomically correct, as already touched upon. The focus could be shifted from vowels to sibilants, for example. Further, different learning algorithms and network structures could be explored. Different phenomena, such as the *McGurk* effect, could be considered within the context of a neural network model which includes sensorimotor knowledge.

To conclude, this project showcased how the formalisation of speech processes often results in gaining new insights on phenomena. In our case we were to observe a perceptual magnet effect in production and the categorization of the output of the network. Further expansion of this model will no doubt shine light on other behaviours that humans and machines might have in common.

REFERENCES

Paul Boersma (2009), Cue constraints and their interactions in phonological perception and production, in Paul Boersma and Silke Hamann, editors, *Phonology in perception*, volume 15, pp. 55–110, Mouton de Gruyter, Berlin, New York.

Paul Boersma (2011), A programme for bidirectional phonology and phonetics and their acquisition and evolution, in Anton Benz and Jason Mattausch, editors, *Bidirectional optimality theory*, volume 180, pp. 33–71, John Benjamins Publishing Company, Amsterdam.

Paul Boersma (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.

Paul Boersma, Titia Benders, and Klaas Seinhorst (2020), Neural network models for phonology and phonetics, *Journal of Language Modelling*, 8(1):103–177, ISSN 2299-8470, doi:10.15398/jlm.v8i1.224, URL `https://jlm.ipipan.waw.pl/index.php/JLM/article/view/224`, number: 1.

Paul Boersma, Kateřina Chládková, and Titia Benders (2021), Phonological features emerge substance-freely from the phonetics and the morphology1, *Submitted to a journal*.

Paul Boersma and Jan-Willem Van Leussen (2017), Efficient evaluation and learning in multilevel parallel constraint grammars, *Linguistic Inquiry*, 48(3):349–388.

Gloria J. Borden (1980), Use of feedback in established and developing speech, in Norman J. Lass, editor, *Speech and Language*, volume 3, pp. 223–242, Elsevier, Amsterdam.

Alison G. Bruderer, D. Kyle Danielson, Padmapriya Kandhadai, and Janet F. Werker (2015), Sensorimotor influences on speech perception in infancy, *Proceedings of the National Academy of Sciences*, 112(44):13531–13536, ISSN 0027-8424, 1091-6490, doi:10.1073/pnas.1508631112, URL `https://pnas.org/doi/full/10.1073/pnas.1508631112`.

Sybil Cooper (1953), Muscle spindles in the intrinsic muscles of the human tongue, *The Journal of Physiology*, 122(1):193–202.

Carol A. Fowler and Michael T. Turvey (1980), Immediate compensation in bite-block speech, *Phonetica*, 37(5-6):306–326.

Vittorio Gallese and George Lakoff (2005), The brain's concepts: The role of the sensory-motor system in conceptual knowledge, *Cognitive Neuropsychology*, 22(3-4):455–479.

Thomas Gay, Björn Lindblom, and James Lubker (1981), Production of bite-block vowels: Acoustic equivalence by selective compensation, *The Journal of the Acoustical Society of America*, 69(3):802–810.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville (2016), *Deep learning*, MIT press.

Frank H. Guenther and Tony Vladusich (2012), A neural theory of speech acquisition and production, *Journal of Neurolinguistics*, 25(5):408–422.

Donald O. Hebb (2005), *The organization of behavior: A neuropsychological theory*, Psychology Press.

Gregory Hickok (2009), Eight problems for the mirror neuron theory of action understanding in monkeys and humans, *Journal of Cognitive Neuroscience*, 21(7):1229–1243.

Teuvo Kohonen (1984), Phonotopics maps insightful representation of phonological features of speech recognition, in *Proceedings of Seventh International Conference on Pattern Recognition, Montreal, 1984*, pp. 182–185.

Patricia K. Kuhl (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.

Alvin M. Liberman and Ignatius G. Mattingly (1985), The motor theory of speech perception revised, *Cognition*, 21(1):1–36.

Vaughan G. Macefield and Thomas P. Knellwolf (2018), Functional properties of human muscle spindles, *Journal of Neurophysiology*, 120(2):452–467.

Giacomo Rizzolatti and Ljarn Craighero (2004), The mirror-neuron system., in Gary G. Bernston and John T. Cacioppo, editors, *Handbook of neuroscience for the behavioral sciences*, volume 1, John Wiley & Sons, Hoboken, NJ.

Giacomo Rizzolatti, Leonardo Fogassi, and Vittorio Gallese (2001), Neurophysiological mechanisms underlying the understanding and imitation of action, *Nature Reviews Neuroscience*, 2(9):661–670.

Ruslan Salakhutdinov and Geoffrey Hinton (2009), Semantic hashing, *International Journal of Approximate Reasoning*, 50(7):969–978.

R. Sandyk (1981), Somatosensory feedback in tongue and speech muscle movements, *South African Medical Journal*, 60(26):992–993.

Kenneth N. Stevens (1989), On the quantal nature of speech, *Journal of Phonetics*, 17(1):3–45.

Kenneth Noble Stevens and Samuel Jay Keyser (2010), Quantal theory, enhancement and overlap, *Journal of Phonetics*, 38(1):10–19.

Sensorimotor Knowledge in Neural Network