Study of Neural Mechanism of Mandarin Vowel Perception and Diphthong Production with Neural Network Model

Chao-Min Wu*, Tao-Wei Wang* and Ming-Hung Li*

SUMMARY: The DIVA (Directions Into Velocities Articulator) model provides a computational and neuroanatomical account of speech acquisition and production; however, its prediction of speech perception and production for Mandarin is limited. The aim of this study is to modify the original DIVA model to simulate both normal and speech disordered productions in Mandarin. The proposed version of the model provides additional functions of speech perception, tonal acquisition and diphthong production. Computer simulation of our modified DIVA model verifies its ability to simulate Mandarin tonal production in diphthong and speech perception across vowels.

Key words: speech production, speech perception, neural network model, DIVA model, diphthongs, Mandarin speech

1. Introduction

With the increase in the aging population, along with a need for early intervention, the number of people with various degrees of speech and hearing disorders has also increased. Communicative disorders, including the inability to express needs and feelings, greatly affect the quality of daily life. An understanding of the mechanisms of normal and disordered speech will improve the diagnosis and treatment of speech problems. Much investigation has been done on both normal and disordered speech production, based on analysis of speech sound recordings, articulatory movement tracings, and neuroimaging (e.g., Peterson and Barney 1952, Guenther and Gaja 1996, Heiss et al. 2003, Bartle-Meyer et al. 2009, Kim and Max 2014). Speech production is a complex process where the shapes of articulatory organs are changed continuously to create acoustic waveforms, perceived by listeners as speech sounds. Numerous physiological models of human speech organs (e.g., Wilhelms-Tricarico 1995, Payan and Perrier 1997, Dang and Honda 2004) have been used to examine speech articulation. These speech models account for the degree and location of constriction along the vocal tract while simulating speech production, but they are limited in their ability to simulate neural control. The aim of this study is to develop a speech production model with neural control for analyzing Mandarin speech production.

The neural-network-based directions into velocities of articulators (DIVA) model (Guenther 1994, Guenther et al. 2006) simulates neural correlates of speech production and describes the sensorimotor interactions involved in articulator control during speech production. This model provides interpretation of the speech production (i.e. articulatory movement and speech sound) and perception (i.e. speech recognition) and prediction of the corresponding related-activities of the brain. The DIVA model can be used to investigate motor equivalence (Guenther 1994), coarticulation (Guenther 1995), speaking rate effect (Guenther 1995), brain activities (Guenther et al. 2006), and the role of auditory feedback in normal speech production (Perkell et al. 2000, Nieto-Castanon et al. 2005); also, stuttering (Max et al. 2010) and apraxia of speech (Terband et al. 2009, Terband and Maassen 2010).

However, the DIVA model is limited in its ability to simulate speech production. For one thing, the DIVA model is a learning model which progressively acquires a relationship between speech motor actions and speech sounds but provides no human-like function of speech perception. Secondly, it focuses on simulating and interpreting English speech production with a fixed preset pitch (either male or female pitch). However, Mandarin is a tonal language, where each syllable is assigned a tonal (pitch) contrast. English is not a tonal language in this sense, so the assignment of pitch information between these two languages is different.

*Department of Electrical Engineering, National Central University (國立中央大學電機工程系)
A final point is that the original DIVA model can only produce monophthongs, and is limited in its production of diphthongs. The long range aim of this study is to develop a model to simulate normal and speech disorders production in Mandarin. This study thus adds speech perception, fundamental frequency control, and diphthong production to the DIVA model in order to simulate Mandarin tonal production.

2. Overview of DIVA Model

As shown in Figure 1, the DIVA model is an adaptive neural network model that was designed to simulate speech production based on the human learning process (Guenther et al. 2006). The model consists of a neural network controller whose cells correspond to boxes and synaptic weights correspond to arrows. Each block in the model comprises a set of neurons and corresponds to a cortical region that is involved in the speech process. The “mapping” is used to present a transformation of neural representation from one block to another. A word or a syllable is selected as input and transformed as a time sequence of 8 articulator locations expressed as a vector. The speech synthesizer of the DIVA model, the Maeda model (Maeda 1990), simulates the shape of the vocal tract from the resulting movement commands of speech articulators, and produces speech sounds. The DIVA model produces a sequence of numbers that represent the brain activity regions and levels. These data are analyzed with the SPM2 (statistical parametric mapping) toolbox (http://www.fil.ion.ucl.ac.uk/spm/) and the brain activity regions are displayed based on the output.

The DIVA model includes feedforward and feedback subsystem. The feedforward subsystem is described as the speech-production-related process and hypothesizes that the motor commands project from the ventral premotor cortex to the primary motor cortex, both directly and via the cerebellum. This command could be stored from previous attempts to produce the sound or randomly generated movements in the babbling stage. The sound is represented as a vector of the fundamental frequency (F0) and the first three formant frequencies in the Speech Sound Map (SSM) block. Through auditory-to-articulator mapping, this vector was transformed into the motor commands of eight parameters of articulators. The motor commands in the Articulator Velocity and Position Vector Map block contain both articulatory directions and positions. These commands are used by the Maeda model to simulate the shape of the vocal tract and produce the corresponding speech sound. Physiologically, the feedforward control subsystem simulates human-like speech production.

The feedback control system consists of two traces, namely somatosensory and auditory feedbacks. This feedback system focuses on returning the error between the target and current states and modifies the motor commands of articulation. The former returns the cur-
rent states of the vocal tract to the somatosensory cortex (somatosensory state map in Figure 1) and the latter collects the auditory signals into the higher-order auditory cortex (auditory state map in Figure 1). The somatosensory and auditory feedbacks are hypothesized as the inferior parietal cortex and superior temporal gyrus, respectively. Feedforward and feedback-based control signals are combined in the model’s motor cortex. The feedforward system provides a faster production process, and the feedback system monitors the production and corrects the errors of the motor commands. Combing the feedforward and feedback systems, the DIVA model is capable of simulating speech production and its learning process.

3. Modification of DIVA Model

The DIVA model provides interpretation of the speech production and perception and prediction of the corresponding related-activities of brain. The DIVA model has limitations in simulation of speech production, (a) human-like speech perception, (b) tonal control and acquisition, and (c) diphthong production. In this section, the theories and methods how to resolve these three limitations are described and incorporated in the DIVA model.

3.1 Speech Perception

Information in the sensory cortices (like visual, auditory, and somatosensory regions) is hypothesized as topographic maps. The maps are inherently generated in the self-organization process when receiving and learning external stimulation. For example, in the normal speech perception process, when the basilar membrane receives a vowel, the auditory cortex produces a topographic map that is represented as the formant frequencies of this vowel. Kuhl (Kuhl 1991, Iverson and Kuhl 1995) proposed the “perceptual magnet effects” on the phonetic categories and explained the effects as a warping map that was generated from exposure to the phonemes of an infant’s native language. To describe the abstract relationship between the sensory mapping and neural property, Kohonen (1982, 1990, 1993) developed the SOM network model with competitive learning to construct the multidimensional sensory space. The SOM model provides parallel signal processing to implement the self-organizing mechanism of the brain and describe the sensory mechanism. Physiologically, the self-organizing mechanism is defined as the brain categorizes the unknown external stimuli based on its captured features. The process of the SOM model uses the updated weights that are based on the characteristics of the input data (i.e. the captured features) to simulate the self-organizing mechanism. During the training stage, the weights of the neurons are compared with an input \( \bar{X}_i, \bar{X}_j \in X \) to determine a winner neuron and the weight of this winner neuron and its neighborhood weights \( \bar{W}_j \) are updated with the learning rule, as described in (1). The update rule includes the neighborhood function, \( \Pi_i(t) \), and learning rate, \( \eta(t) \). The trained weights \( \bar{W}=[\bar{W}_1, \bar{W}_2...\bar{W}_3] \) are more similar to the inputs and represented as the categorical features of the input data. This output layer was used to describe the distribution of neuron cells.

\[
\bar{W}_j(t+1) = \begin{cases} \bar{W}_j(t) + \eta(t)[\bar{X}_i - \bar{W}_j(t)] & \text{if } j \in \Pi_i(t) \\ \bar{W}_j(t) & \text{if } j \notin \Pi_i(t) \end{cases}
\]  

(1)

The sensorial response of neuron cells includes the distribution of neuron cells and the activity level of neuron cells, but the transitional SOM model only simulates the distribution of neuron cells. Wu (Wu et al. 2013) added a similarity index \( S_j \) in the original SOM model to provide the neural activity representations and used this modified model to develop the auditory perception model. The index is determined by the Euler distances between an input \( \bar{X}_i \) and weights \( \bar{W}_j \) of the trained model, as shown in (2).

\[
S_j = |\bar{X}_i - \bar{W}_j|
\]  

(2)

The trained weights of neuron cells were seen as categorical targets of training data and used to determine the similarity (or difference) between the trained weights and test data, since the weights of the SOM model are trained to represent the input and to display the distribution. Our auditory perception model maintained the ability to represent cell distribution and additionally showed the activity level of the neuron cells. The current work will combine our auditory perception model with the DIVA model to simulate the language acquisition and speech production to provide a model to interpret the speech mechanisms.

3.2 Tone Acquisition

Previous studies (Howie 1976, Ladefoged and Johnson 2011) have confirmed that the control factor of Mandarin tones is primarily the fundamental frequency (F0). The function of time varying fundamental frequency is necessary in tonal speech production. The
pitch synchronous overlap and add (PSOLA) algorithm is a speech process to modify the pitch and duration of speech signals in speech synthesis (Charpentier, and Stella 1986). The segments of speech signals are moved further (see in Figure 2(a)) or closer together (see in Figure 2(b)) to decrease or increase the pitch. Our pervious study (Wu and Wang 2012) added the PSOLA algorithm in the DIVA model to produce the varied tonal speech.

During the babbling stage, infants only produce pitch contours to imitate their parents. Many neuro-imaging studies have investigated the mirror neurons of the human brain and have suggested human beings can inherently imitate speech production and movement (Rizzolatti et al. 1996, Kohler et al. 2002). Tone variation is based on the F0 contour and considered as a curve learning of F0 variation. The back-propagation neural network (BPNN) is established to model the nervous system (i.e. the brain) and neural signal processing in the brain for the simulation of curve learning. Based on the BPNN model, the tone model was developed and used to simulate tone acquisition of infants during pre-babbling stage (Wang and Wu 2014). To simulate tone acquisition, this tone model is combined with the DIVA model to provide the abilities of both tone control and tone acquisition.

3.3 Diphthong Production

Diphthongs are types of vowels in which two or more vowels connect in one syllable. In a diphthong, “vowel inherent spectral change (VISC)” is an important role in speech perception (Morrison 2013). Vowel inherent spectral change (VISC) has been defined as slow variations in formant frequencies associated with two vowels compared to monophthongs. Some studies have examined the characteristics of VISC and indicated that the onset+offset hypothesis is superior in terms of leading to higher correct-classification rates and higher correlation with listeners’ vowel identification responses. In addition, Morrison (Morrison 2013) used discrete cosine transform (DCT) to describe the formant transition between two vowels. Therefore, the current work used Morrison’s approach to improve the DIVA model in the production of diphthongs. We hypothesized that the diphthong is generated by the selected vowels /ai/ and /i/ rather than considered as the diphthong /ai/. As shown in Figure 3, the diphthong is composed of three parts: (1) the formant target of the first vowel, (2) the formant transition of the two vowels, and (3) the formant target of the second vowel.

4. Simulation

The purpose of the current work is to add speech perception, fundamental frequency control, and diphthong production to the original DIVA model for simulation of Mandarin tonal production. This study applied the
Study of Neural Mechanism of Mandarin Vowel Perception and Diphthong Production with Neural Network Model

![Graph showing vowel frequencies](image)

Figure 4  (a) Implied six Mandarin vowels for categorization. (b) One thousand speech sounds for learning process with a 50 Hz interval.

Table 1  The first three average formant frequencies of six Mandarin vowels from 24 male college students. The Chinese vowels are listed in the first column and pronounced as the IPA symbols shown in the parentheses.

<table>
<thead>
<tr>
<th>Mandarin phonetic symbol (IPA)</th>
<th>F1(Hz)</th>
<th>F2(Hz)</th>
<th>F3(Hz)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( /a/ )</td>
<td>713</td>
<td>1,154</td>
<td>2,475</td>
</tr>
<tr>
<td>( /i/ )</td>
<td>307</td>
<td>2,384</td>
<td>3,094</td>
</tr>
<tr>
<td>( /u/ )</td>
<td>326</td>
<td>739</td>
<td>2,087</td>
</tr>
<tr>
<td>( /e/ )</td>
<td>498</td>
<td>2,057</td>
<td>3,123</td>
</tr>
<tr>
<td>( /o/ )</td>
<td>518</td>
<td>888</td>
<td>2,250</td>
</tr>
<tr>
<td>( /e/ )</td>
<td>295</td>
<td>1,697</td>
<td>2,150</td>
</tr>
</tbody>
</table>

The modified DIVA model was used to perform three simulations. The first two simulations were used to simulate language acquisition of the newborn along with categorical responses to verify the auditory perception model; the third simulation was used to determine whether the modified DIVA model can produce the diphthong from learning Mandarin tones.

The language categories of the newborn were shown as a clustering effect such that the categories formed during the development (Kuhl 1991). In the first simulation, the model without training (MWT) was used to simulate the newborn without vowel categories. One hundred speech sounds for each Mandarin vowel were randomly generated and used as the input data for the learning process which is analogous to infant babbling. As shown in Figure 4(a), six Mandarin vowels were based on their first three formant frequencies of vowels from recording of 24 male college students (as shown in Table 1). With the preset one thousand neurons for this simulation, MWT was exposed to six Mandarin vowels and trained to form six corresponding vowel categories to imitate the process of language acquisition of the newborn. After the learning process, the speech sounds shown in Figure 4(b) were used as the test sounds for our auditory perception model to show its learning capability.

In the second experiment, the modified DIVA model was used to simulate tone acquisition for the infant. For this simulation, first, the four Mandarin tones were given to simulate the feature of the perceived speech
(as shown in Figure 5). Then, as in Figure 6, the tone acquisition of the modified DIVA model was applied in three steps as follow:

I. First, a tone contour of the model is generated randomly. In the simulation, two kinds of initial contours were selected as input of the model; a tone similar to Mandarin Tone 1 and then randomly generated tones (as shown in Figure 7).

II. Determining whether the error between the generated contours of the model's output and the expected contours are below the given threshold.

II. The contours of the modified DIVA model were repeatedly updated until the difference between the error is below a pre-defined threshold.

The modified DIVA model was simulated for tone acquisition of infants and diphthong production with the learned tone.

5. Results of Simulation

In the first experiment, the modified DIVA model was used to simulate the language acquisition of infant babbling. The six Mandarin vowel categories (as shown in Figure 4a) were used to train the MWT and the speech sounds (F1: 200–1,000 Hz, F2: 500–2,500 Hz, F3: 1,500–3,500 Hz, shown in Figure 4b) were used to test this trained model. After the learning process, the speech sounds shown in Figure 8 display six categories on the F1–F2–F3 plane where the speech perception space is easily recognized. As shown in Figure 8, the modified DIVA model indicates its learning capability of Mandarin vowels.

The purpose of the third experiment was to simulate the tone acquisition and produce the tonal diphthongs. In the simulation, two kinds of initial contours were selected as inputs to the model, the tone similar to Mandarin Tone 1 and randomly generated tones (as shown in Figure 7). These initial states were used to represent the pitch of pre-babbling, which could be any tone. The training process is like a learning or practicing stage. After training, the simulated results indicated similar contours with the given contours. Based on the learned tones, the DIVA model produced the diphthongs (/'ia/) and the Mandarin tones and these sounds were analyzed with Praat 4.312. Figure 9 showed the spectrograms of the diphthong with four tones and the solid lines represent the pitch trajectories.

6. Discussion

The DIVA model provides a computationally explicit account of the interactions between the brain regions involved in speech acquisition and production. This model also provides a valuable tool for studying the mechanisms underlying normal and disordered speech. This study focused on adding speech perception, fundamental frequency control, and diphthong production.
Study of Neural Mechanism of Mandarin Vowel Perception and Diphthong Production with Neural Network Model

Figure 8 The trained model clustered with the testing vowels (small dots) and formed six vowel categories (the circles).

Figure 9 The synthesized speech sounds with four tones: (a) The first tone, (b) second tone, (c) third tone, and (d) fourth tone. The solid lines were the pitch trajectories with Praat 4.312.

to the original DIVA model for simulation of Mandarin tonal production.

In the first experiment, MWT was used to simulate the newborn without vowel categories and uses six vowel categories trained with the MWT to simulate language acquisition by infants. The simulated results (in Figure 8) showed that the trained perception model produced six similar vowel categories as the target categories. Saffran et al. (1996) suggested that infants have access to a powerful mechanism for computation of statistical properties of the language input. In this simulation, the mechanism is hypothesized as unsupervised learning. The unsupervised learning rule of the perception model is used to simulate that infants can find the hidden acoustic feature and relationship in unlabeled data. Neurophysiologically, the native perceptual mechanism is verified in Molfese and Molfese (1997).

Since the language process should be involved with various mechanisms and related-learning processes in the production and perception, we used the SOM neural model to form a nonlinear projection in an unsupervised-learning (adaptive) algorithm in the neural-network category. The model carries out a clustering process for visualization and abstraction of the multidimensional input data (i.e. sensory). However, tone variation is based on the F0 contour and considered as a curve learning of F0 variation. The BPNN is
a suitable model for the simulation of curve learning. The modified version of DIVA model maintains the original functions of speech simulation and combines perception (based on the SOM) and tone (based on the BPN model) functions to provide the human-like perception and learning process.

To handle the tonal contrasts of Mandarin, we added tonal control function with the TD-PSOLA algorithm in order to modify the motor commands of the corresponding articulator to generate tonal speech. Furthermore, we used the BPN model to develop a tonal learning function and simulate the tone acquisition of infants during pre-babbling stage. The results showed that the modified DIVA model could use the learning tone curves to produce the diphthongs. The current modified DIVA model can simulate the auditory perception and predict the speech production in English and Mandarin; however, the current modified DIVA model cannot inherently obtain the tone information (or pitch variation) from the perceived sounds. Language are capable of obtaining the pitch information and imitating the perceived sound. The targets of four Mandarin tones were first given to simulate the feature of the perceived speech sound in the second experiment. Further study will focus on the development of pitch perception.

7. Conclusion

This aim of this study is attempt to add speech perception, fundamental frequency control, and diphthong production to the DIVA model to simulate Mandarin tonal production. In addition to the original capability of speech simulation, the current modified DIVA model can be used to produce tonal speech and simulate the auditory perception to represent the corresponding neural correlates. Future work will focus on the investigation of normal speech mechanism and simulation of disordered speech. Although this study shows no direct connection to the Converter–Distributor (C/D) model (Fujimura 2000), the concept and recent progress of the C/D model would be considered in the future development of our neural network model.

Acknowledgements

A special thanks to my advisor, Osamu Fujimura, for showing me the world of phonetic speech science. Thanks also to Donna Erickson and Reiner Wilhelms-Tricarico for their encouragement and proofreading on a previous version of the paper and an anonymous reviewer for useful suggestions. This work was supported by the grants of NSC 102-2221-E-008-046 and NSC101-2221-E-008-005 from the National Science Council, Taiwan.

References


Study of Neural Mechanism of Mandarin Vowel Perception and Diphthong Production with Neural Network Model


(Received Mar. 18, 2015, Accepted Jun. 30, 2015)