

THE TWO TARGETS OF SPEECH PRODUCTION: TWO LEVELS OF SPECIFICATION

By

Mingkang He

Senior Thesis in Electrical Engineering

University of Illinois at Urbana-Champaign

Advisor: Professor Stephen E. Levinson

May 2016

Abstract

A thread of this work is the difference in how articulatory and perceptual features of phonology are integrated into speech production. This idea emerges during the research and simulation of appropriate speech synthesizers for one fellow graduate student to adopt in a proposal for development of an automatic speech acquisition system. While using available synthesizers to produce speech utterances, it occurs that these two features actually determine the two levels of input to speech synthesizers, namely, tasks and muscle activities.

This thesis adheres to two existing models, each accepting one level of input. The TADA approach [1] maintains that input to a speech synthesizer should be tasks, which consist of specifications of tract variables, such as locations and degrees of constrictions, as functions of time. To be more precise, these tasks are given the name of gestural scores, which is explained in the paper later. On the other hand, Praat [2] takes as input muscle activities: the articulatory input specifications initially control the lengths and tensions of the muscles, instead of the positions of articulators.

After making a brief introduction to the above two speech synthesizers, assessment and comparison of their time efficiencies as well as perceptual accuracies are provided (in Part I and Part II), by confronting simulated results from each of them with sounds in the real world. In the end of both parts, suggestion is offered on which category of synthesizers should be adopted with respect to various aspects of research concentrations in articulatory phonology.

Subject Keywords: speech synthesizer; articulatory phonology

Contents

1. The Task Dynamics Approach.....	1
1.1 Introduction.....	1
1.2 Timing.....	1
1.2.1 Timing of Tract Variables Response.....	3
1.2.2 Timing of Speech Production.....	5
1.2.3 Conclusion	7
1.3 Test of Articulation Model.....	7
1.3.1 Amplitude Analysis	7
1.3.2 Spectrogram Analysis	9
1.3.3 Conclusion	11
1.4 Others.....	11
2. The Praat Approach.....	12
2.1 Introduction.....	12
2.2 Timing.....	13
2.2.1 Timing of Muscle Activities.....	13
2.2.2 Timing of Speech Production.....	13
2.2.3 Conclusion	14
2.3 Test of Articulation Model.....	14
2.3.1 Amplitude Analysis	15
2.3.2 Spectrogram Analysis	15
2.3.3 Conclusion	16
2.4 Others.....	16
References.....	17

1. The Task Dynamics Approach

1.1 Introduction

Articulatory Phonology [3] begins with the identification of phonological units with dynamically specified units of articulatory action, called gestures. Thus, an utterance is described as an act that can be decomposed into a small number of primitive units defined as articulatory gestures [4]. Further, since acts are distributed across the various articulator sets of the vocal tract (the lip, tongue, glottis, velum, etc.), each articulator should have its own tier. Further, the gestural score is a representation of the values on all relevant tiers.

For instance, the following gestural score is given for the English word “pawⁿ”, /pɔ:n/.

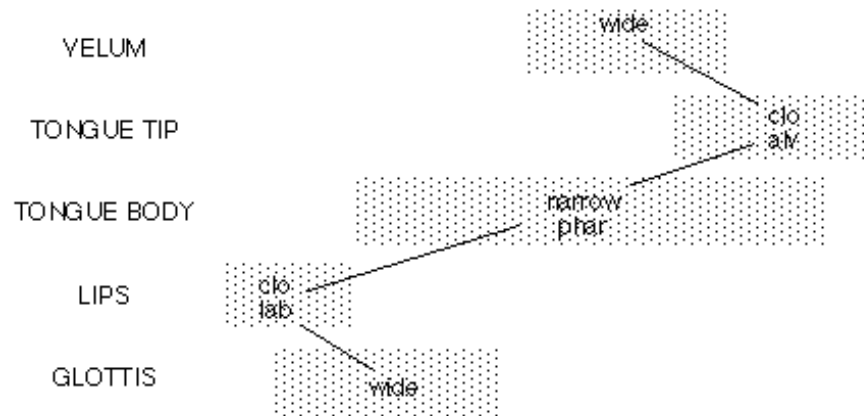


Figure 1.1. An example of gestural score

Figure 1.1 gives an idea of the kind of information contained in the input in type of gestural score. “Each row, or tier, shows the gestures that control the distinct articulator sets: velum, tongue tip, tongue body, lips, and glottis... descriptors... stands for a numerical equilibrium position value assigned to a tract variable... for the tongue tip gesture labeled {clo alv}, {clo} stands for -3.5 mm, and {alv} stands for 56 degrees [5]”.

This articulatory phonology approach is then incorporated into a computational system named Task Dynamics Approach (TADA) [1] being developed at Haskins Laboratories [6].

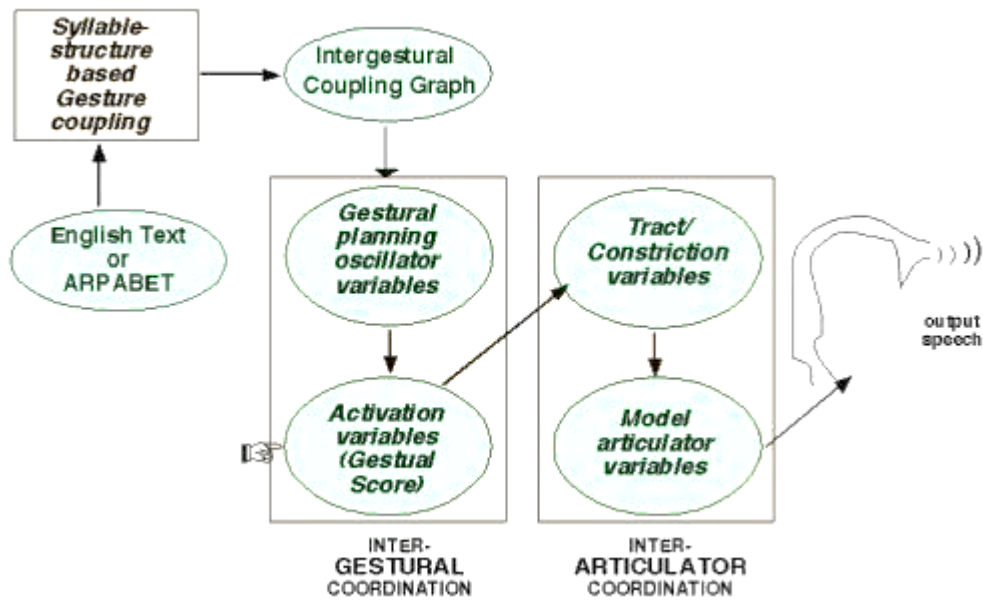



Figure 1.2. Information flow through TADA models

TADA consists of models that can run independently of one another as illustrated by the boxes in Figure 1.2. With the research purpose of investigating different levels of input to speech synthesizes, this thesis explicitly focus on the task dynamic model of inter-articulator coordination, as shown by the  in Figure 1.2. By design it takes as input a gestural score that characterizes the formation and release of local constrictions within the vocal tract (the gesture’s functional task). After having received the input, the task dynamic model calculates the time-varying response of the tract variables and component articulators to the imposition of the dynamical regimes defined by the “task”. The sample response to input gestural score for utterance “paw~~n~~” is shown in Figure 1.3. These time-varying responses are then used to calculate the resulting speech waveform as shown in Figure 1.2.

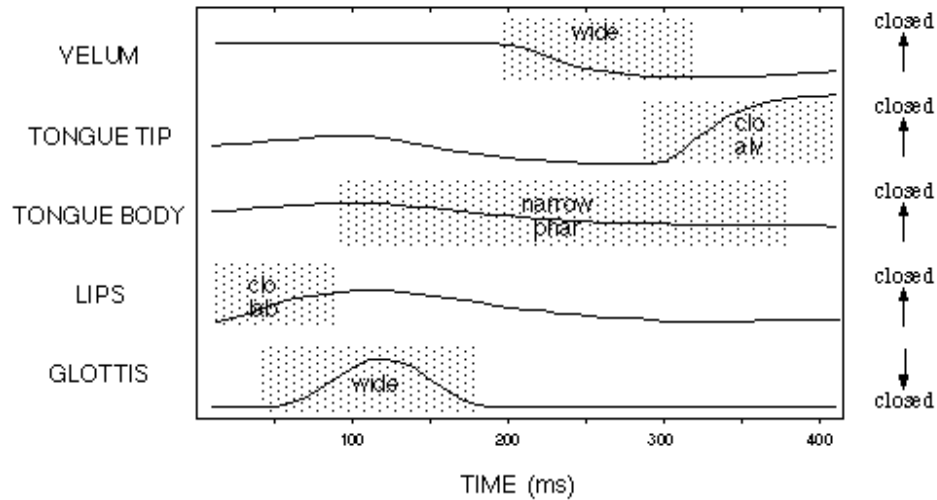


Figure 1.3. An example of calculated response to gestural score

1.2 Timing

” ... the other models represented by boxes in Figure 1.2 (Coupled Oscillator model of inter-gestural coordination and Task Dynamic model of inter-articulator coordination) are meant to be part of a model of the human speech production process [7] ...”. Therefore, the specifications of tract variables are implemented to mimic part of human speech production and it would be reasonable to measure their computed response times in interest of time efficiency of the complete speech production process.

1.2.1 Timing of Tract Variables Response

Fortunately, it becomes possible to isolate exclusively the tract variable computation sub-function from complete TADA source code. Here this thesis takes advantage of the Matlab built-in function of stopwatch timer to measure the computed response times of these tract variables individually.

Note that the first experiment includes gestural scores (tasks) representing four input utterances “chap”, “chip”, “chop” and “chump” that originate from the test data of affricates. Each of the tasks is run consecutively three times and corresponding response times are recorded. The second includes gestural scores (tasks) representing test data of approximants “luck”, “lot”, “lap”, “lip” and the provided pronunciations of the English names of TADA developers. Each of the tasks is run once. While this data could be further adopted in articulatory phonology research, it is used simply and exclusively for the purpose of studying tract variables response times. The experimental results are shown next.

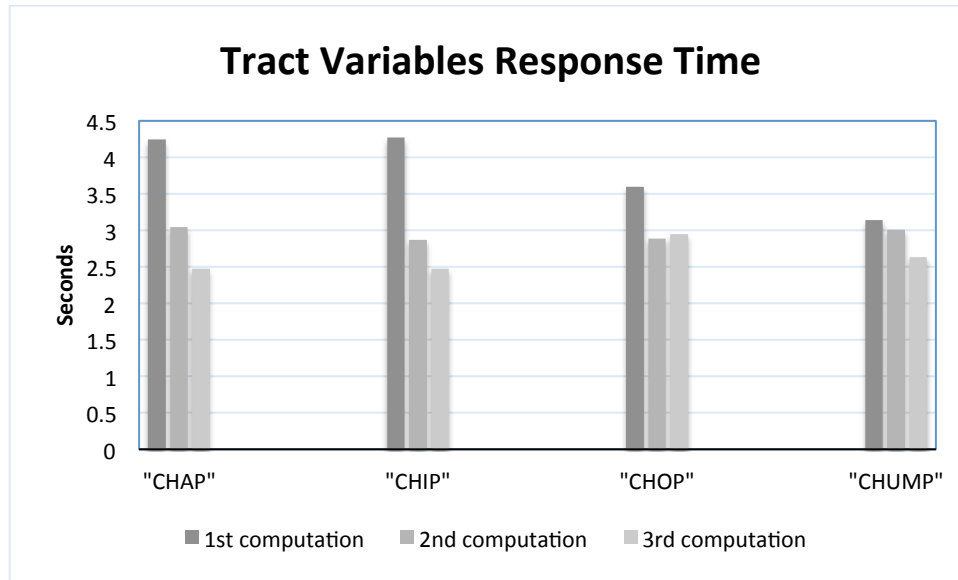


Figure 1.4. Response times for tract variables

Figure 1.4 unveils a general trend of decrease in the computed response times for all tract variables to implement the same task, and therefore to make the same utterance. One conclusion is that the designed nature of tract variables appropriately explains this trend. Firstly, the computation provides different results as the initial conditions of the tract variables are randomly chosen. Relatively different initial conditions result in the varying required times for tract variables to implement seemingly identical tasks. More importantly, running the same task consecutively means that tract variables are able to resume from locations and degrees that are already approximate enough to the tasks, therefore reduces required response times.

Figure 1.5 demonstrates a general trend of increase in tract variables response times with respect to various input utterances, where “Hosung” and “Michael” have two syllables and the rest have one single syllable.

It is important to remember how gestural scores (tasks) are generated and specified in TADA model. “Syllable structure-based gesture coupling model takes as input a text string and generates an intergestural coupling graph ... Coupled oscillator model of inter-gestural coordination takes as input a coupling graph, and generates a gestural score, that specifies activation intervals in time for each gesture [8]”. Therefore, a larger number of syllables eventually results in an increase in the length of time-varying responses of tract variables, as confirmed by choice of input utterances.

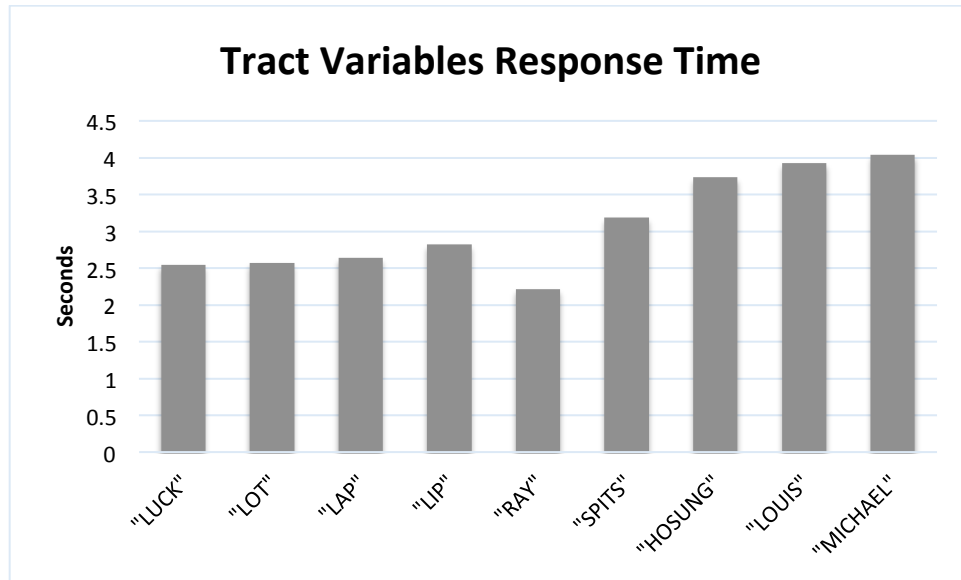


Figure 1.5. Response times for tract variables

1.2.2 Timing of Speech Production

Up to this point the conclusion is that speaker can directly control muscle tensions, muscle lengths, and the locations and degrees of the constrictions in the vocal tract. This thesis has also explained how timing varies in different sets of initial and final conditions in speech production. Next one hypothesis will be made about what role does the group of articulatory specifications (gestural scores, or tasks) play in speech production and perception.

While the existence of feedback loop in perceptual assessment of the simulated results is not specifically stated in TADA, this thesis takes advantage of the bite-block experiments [9] and assumes its existence in simulated speech production process. This experiment specifies the conditions where certain muscle lengths are not the target positions: speakers immediately compensate for the constraints on the jaw, even before phonating, in such a way that the tongue muscles bring about approximately the same area function in the vocal tract as in normally articulated vowels, while having very different shapes.

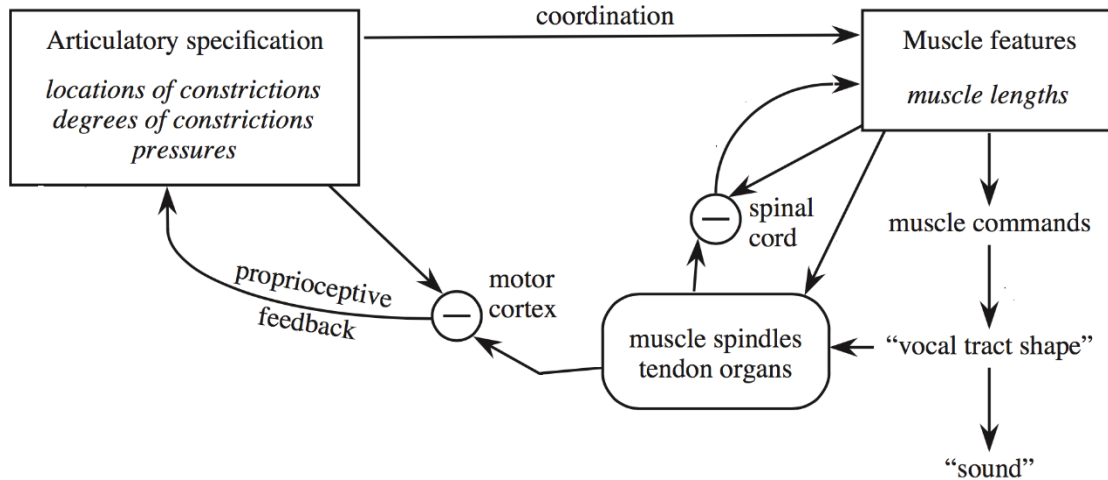


Figure 1.6. Integration of “tasks” into speech production

Rectangles = representations. Rounded rectangles = sensors. Encircled minus signs = comparison centers

Figure 1.6 shows a simplified view of how the articulatory aspects of phonology (gestural scores, or “tasks”) are integrated into speech production. The values on the tiers (see Figure 1.1) represent the articulatory specification of vocal tracts and implement the forward path as we can see in the top side of the Figure 1.6.

The proprioceptive sensory system, consisting of muscle spindle and tendon organs sends the information about the realized shapes back to the motor cortex, where it is compared to the intended articulatory specifications, so that appropriate action can be taken if there are any differences.

This forms a typical representation of skilled motor behavior, where the speaker/listener will monitor and assess the faithfulness of the computed acoustic result based on the original perceptual specification of certain utterance. And in fact, this is in accordance with Fowler [10], who argues that the timing of articulatory gestures is not extrinsically controlled by things like syllable boundaries or incompatible articulatory specifications. Instead timing is an integrated part of the mental specification of the motor plan for each segment.

1.2.3 Conclusion

Section 1.2.2 has demonstrated that intrinsic timing resides at the articulatory task level and is implemented in TADA model. The experiment results as described in section 1.2.1 have shown how syllable lengths affect timing of speech production extrinsically.

And therefore from the timing aspect, this thesis concludes that TADA would not be a perfect model in simulating the speech production of a skilled speaker since its process is much too slow for that. However, it could be an appropriate model in simulating language learners at later ages, as a child finishes learning to speak and starts to compare his/her own utterance with what he/she hears through proprioceptive feedback.

1.3 Test of Articulation Model

Next, how the articulation model of TADA performs in simulation of speech production is investigated. Again the test data of affricates is adopted, “chap”, “chip”, “chop” and “chump”. In each of the tests, the speech sound of a certain utterance produced by TADA is compared to the actual sound of that utterance in real world. First, evaluation of the faithfulness of the computed perceptual result is provided by looking at the amplitude plots as well as the input gestural scores. After that, a simple spectrogram analysis will be performed in order to visualize possible similarities and differences between the simulated and real sound signals.

1.3.1 Amplitude Analysis

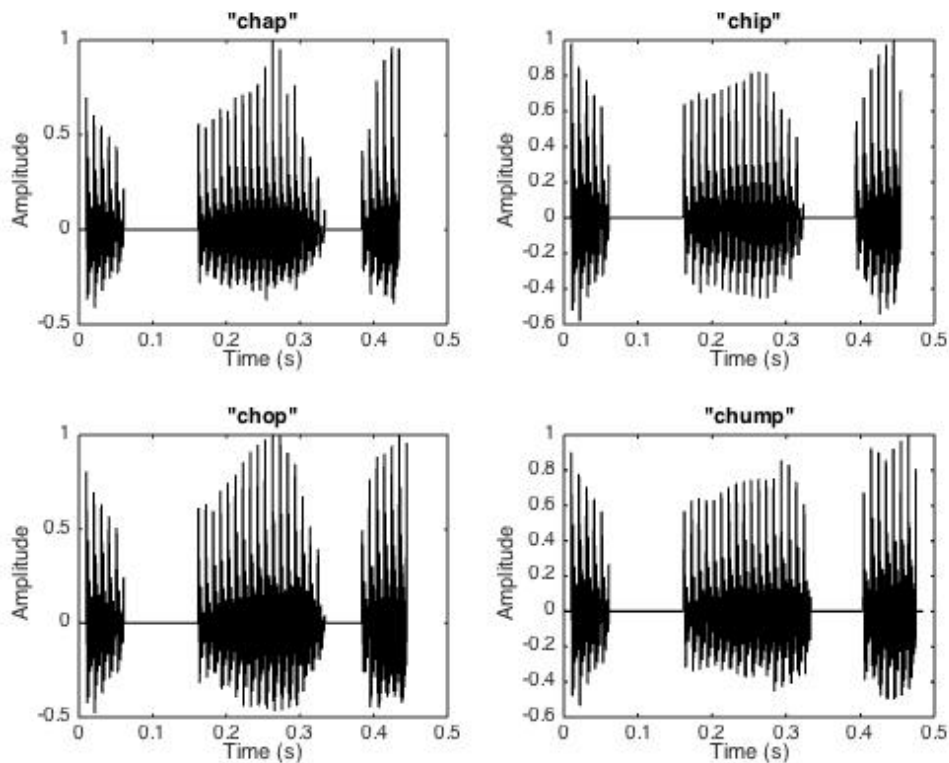


Figure 1.7. Amplitude plots of utterances simulated by TADA

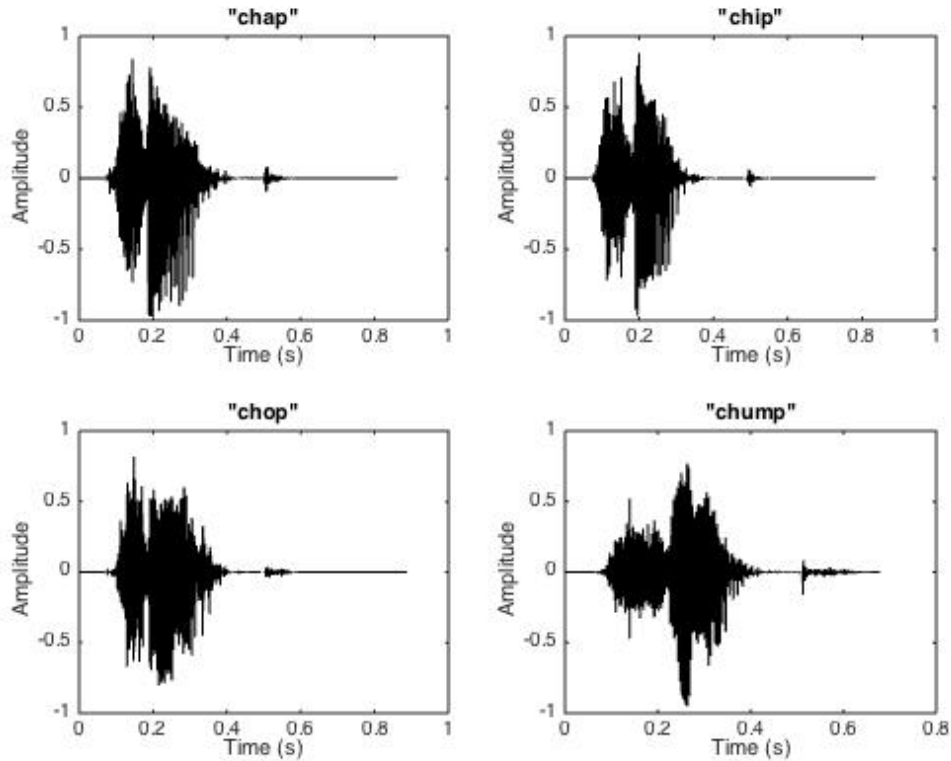


Figure 1.8. Amplitude plots of identical utterances in real-world

The time-domain plots of TADA performance demonstrates the presence of three bursts of energy, corresponding to three articulations. However, trying to read out these words prompts that they are less accurate. The “*ch*” and “*p*” should each correspond to one burst of energy, implying that one articulation might be redundant. In fact, the same plots of real-world utterance sounds confirm the presence of only two bursts of energy.

It is possible to tell that the first articulation simulated by TADA is the redundant one, by simply listening to the two sets of sounds. This thesis will use this conclusion as a premise of the following comparisons and try to validate this conclusion later with input gestural scores.

Also, the shapes and lengths of the first burst are consistent within all four amplitude plots of simulated sounds. This could be another indicator that it is merely a redundant noise due to inaccurate presetting. The middle burst, corresponding to the articulation of “*ch*”, has similar shapes (low-high-low) and lengths (around 0.1 seconds) in amplitude plots of both simulated and real sounds. The last burst has slightly different shapes in simulated sound amplitude plots (ending with a sudden halt) as compared to the real ones (ending with a fading effect).

Figure 1.9 shows part of the input gestural scores for utterance “chap”. The articulatory specifications for glottis and velum are critical in the explanation for the first redundant articulation.



Figure 1.9. Part of the input gestural scores for utterance “chap”

It is clear that trajectory evaluated from the opening glottis and velum at the beginning gives rise to the simulated noise. Therefore the noise is due to inaccurate input specifications instead of miscalculation during simulation process. Since the input gestural scores consistently specify opening glottis and velum initially within all four data sets, this thesis makes the hypothesis that it could be because of limitations of the model in forming certain shapes.

1.3.2 Spectrogram Analysis

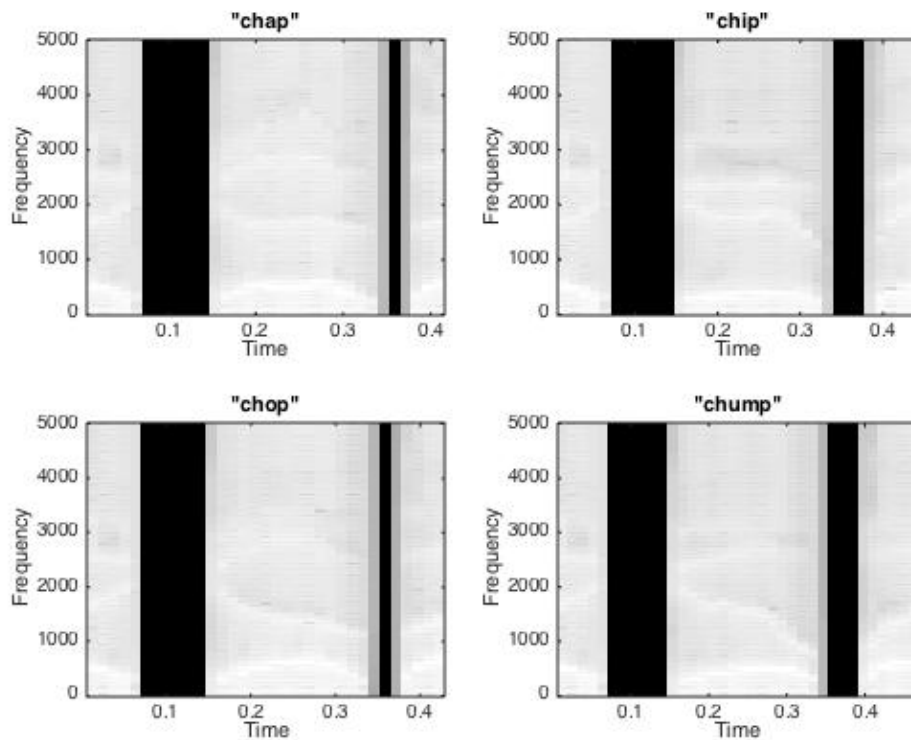


Figure 1.10. Spectrograms of utterances simulated by TADA

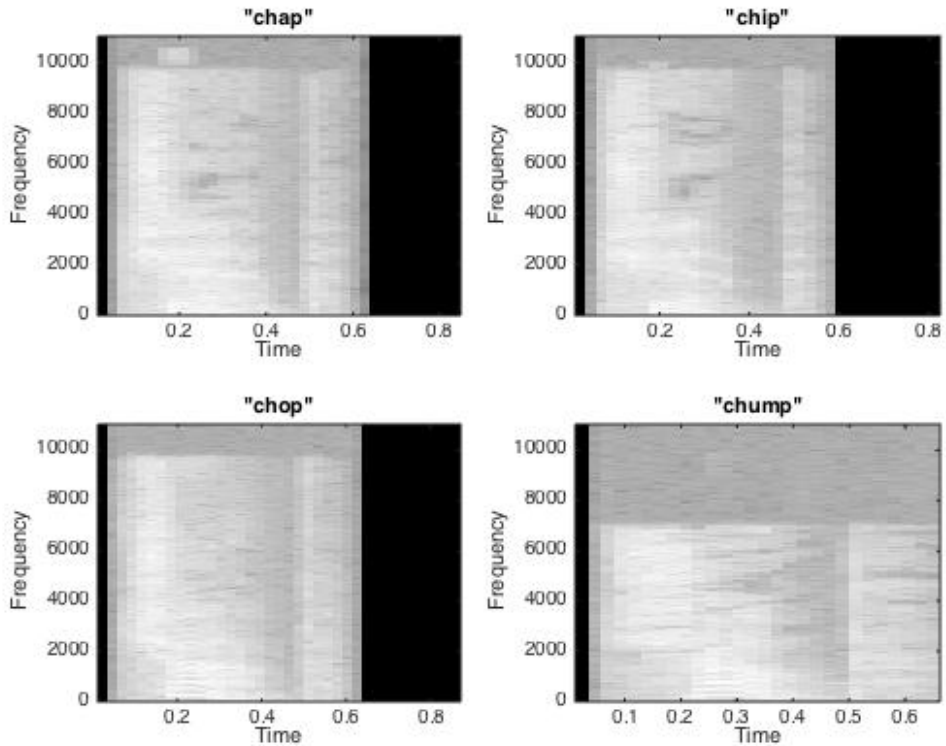


Figure 1.11. Spectrograms of identical utterances in real-world

Figure 1.10 and Figure 1.11 displays spectrograms using Short-Time Fourier Transform (STFT), original sampling frequency and Gaussian window length of 128. Moreover, the colors of the spectrogram encode frequency power levels. Brighter colors indicate frequency content with higher power, while darker colors indicate frequency content with very low power.

Due to the difficulty in finding real-world sounds having the same nature frequencies of TADA simulation, the following comparisons are only described qualitatively and relatively.

First of all, both plots confirm the separation of higher-frequency contents into two intervals. As can be seen in the spectrograms, higher frequencies are well represented in the articulations of “ch” and “p”.

Moreover, in the spectrogram plots of real sounds, the darker contents that appear during transition could be caused by oral wall vibration, and is not audible due to its low frequency yet still recognizable in the spectrogram. While these frequency contents in amplitude plots of real sounds are rather smooth in transition, some of these contents are seen completely voiceless in those of simulated sounds.

1.3.3 Conclusion

TADA consistently outputs acoustic results with an initial noise due to the limitation of input gestural scores. Also sudden halts of the motion of the vocal tracts make the results less natural and less smooth as compared to real-world sounds, especially during transitions.

1.4 Others

TADA specifies positions and degrees of articulators (tasks) as immediate targets of speech production. With the implementation of proprioceptive feedback used to evaluate differences between these specifications and perceptual results, it follows the listener-oriented principle of minimization of perceptual confusion. Therefore, TADA functions as a result-oriented articulatory synthesizer and the tasks would be the appropriate level of inputs. And researchers working in development of text-to-speech systems might want to take advantage of TADA.

2. The Praat Approach

2.1 Introduction

Praat is a comprehensive model of speech-production apparatus with focus on physical phenomena that are used in speech, including lungs, glottis, and vocal and nasal tracts [11], [12], [13]. The foremost difference between the two models described in this thesis is that inputs to Praat model are muscle activities: the articulatory input specifications initially control the lengths and tensions of the muscles, not the positions of the articulators. Starting from the activities of the main muscles involved, the Praat model then computes the target positions and tensions of the articulators. These parameters finally synthesize acoustic output.

The vocal tract is viewed a structure of ducts (channels that contain air). “We model the muscles (and, therefore, the walls of the ducts) as mass-spring systems. As the input to the model is formed by the activities of these muscles... the stiffness of a muscle [14] ...”. Figure 2.1 gives a simplified view of this design: the input into Praat is stiffness (equivalent to spring constant k).

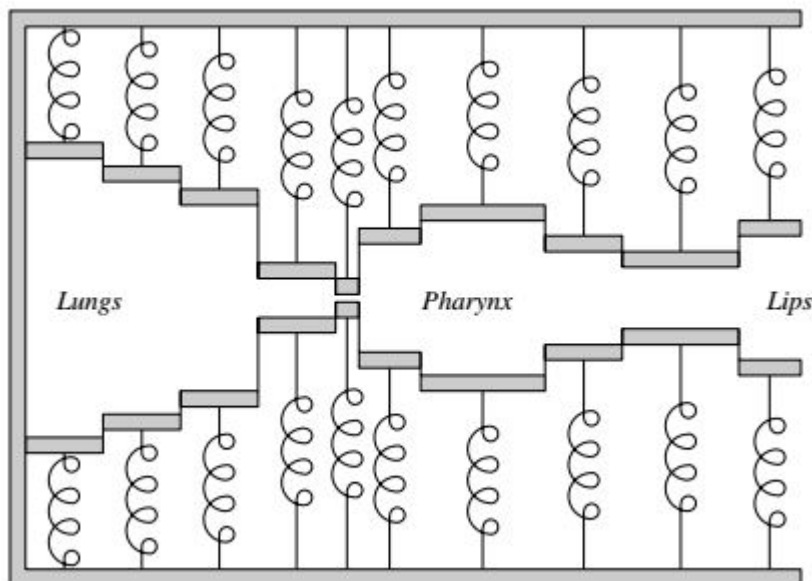


Figure 2.1 Simplified mid-sagittal view of Praat model of speech apparatus

2.2 Timing

Section 1.2.2 describes how intrinsic timing resides at the higher level of articulatory tasks and since it is not modeled into Praat, this thesis will look into some other timing strategies here.

2.2.1 Timing of Muscle Activities

The change in stiffness of springs controls Praat as a mass-spring system. This is identical to the idea of Perrier [15], who use a linear behavior of stiffness control where at least two targets have to be specified: (1) the starting values of equilibrium dimensions at $t = 0$ and (2) the ending values at $t = T$. For instance, K is the linear spring constant (in N/m) of the spring and is specified as k_1 at time t_1 and as k_2 at time t_2 , then the spring at every time t between t_1 and t_2 can be expressed as

$$k(t) = k_1 + \frac{t - t_1}{t_2 - t_1}(k_2 - k_1)$$

Hence, the activities of the articulating muscles become varying functions of time. Moreover, this is very similar to the timing strategy of tract variables used in TADA, where larger difference between initial and final values leads to longer simulation time.

2.2.2 Timing of Speech Production

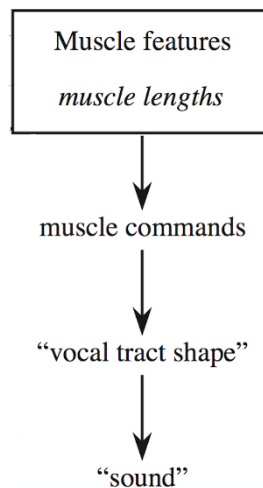


Figure 2.2. Integration of muscle activities into speech production

rectangles = representations

Figure 2.2 show that the speaker can control the tension of muscles and thereby setting his/her muscles to a specified length. For this, a direct muscle command is sent to the muscle fibers, whose contraction then results in a change in the vocal tract shape.

This forward control process involves more collective effort of neuron fibers and brain stem, however here this thesis simplifies it in order to concentrate on the time efficiency of the complete simulation process. Notice that the speech production process shown in Figure 2.2 is completely integrated into the TADA process (see Figure 1.6). One major difference is the fact that Praat does not implement the proprioceptive feedback that is used to compare and compensate the difference between target and actual articulating parameters.

2.2.3 Conclusion

Figure 2.2 shows how ultimate targets of muscle activities directly lead to acoustic sound output. With no proprioceptive feedback used in check and maintenance, the model of Praat functions as a language speaker who has highly organized articulatory specifications in terms of degrees of tract variables and air pressures. This could help explain why the Praat model advances in time efficiency as compared to TADA model. And by minimizing articulation effort, Praat honors the principle of economy as stated by Passy: “languages tend to get rid of anything that is superfluous [16]”.

2.3 Test of Articulation Model

Similar to section 1.3, this thesis will investigate how the articulation model of Praat performs in simulation of speech production. In this contrast test, the utterance being simulated and compared is /əpə/. Firstly, evaluation of the faithfulness of the computed perceptual results will be provided by looking at the amplitude plots. After that, a simple spectrogram analysis will be performed in order to visualize possible similarities and differences between the simulated and real sound signals. Conclusions will follow in the end.

2.3.1 Amplitude Analysis

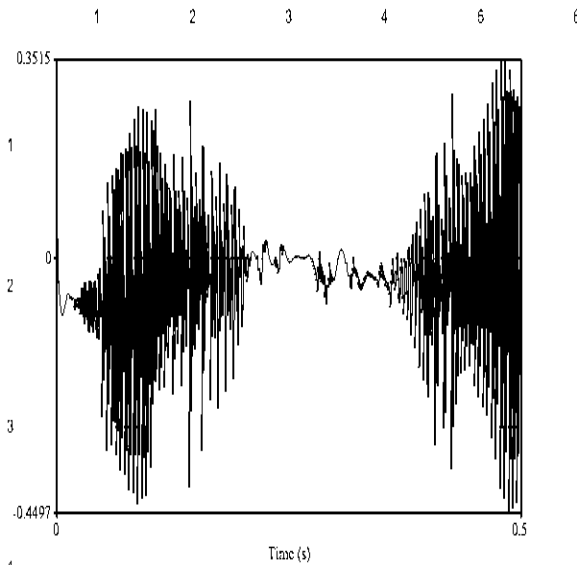


Figure 2.3. Amplitude plots of utterances simulated by Praat

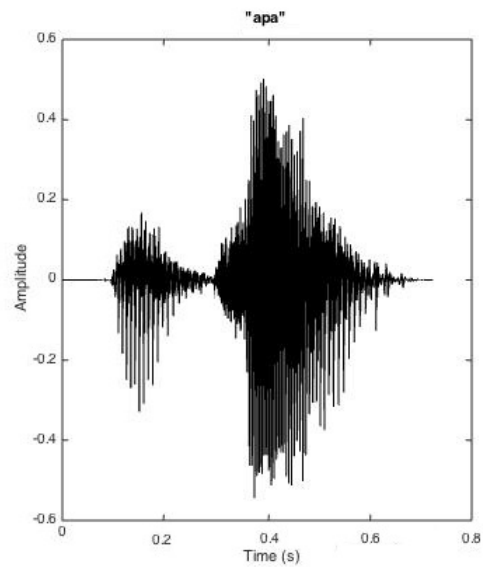


Figure 2.4. Amplitude plots of identical utterances in real-world

The time-domain plots of both Praat performance (Figure 2.3) and real world sound (Figure 2.4) confirm the presence of two bursts of energy, corresponding to two articulations of /ə/. And the two bursts are similar in shape and length.

More importantly, Praat's acoustic outputs have more natural transitions during the starting and ending of energy bursts as compared to those of TADA in section 1.3.1. The bursts in Praat output appear and fade in a gradual process as they are expected to be in real world. This could be one of the advantages for modeling human vocal tracts as mass-spring system, where the springs initially guarantee smoother transitions in simulation process.

2.3.2 Spectrogram Analysis

Figure 2.5 and Figure 2.6 shows spectrograms using Short-Time Fourier Transform (STFT), original sampling frequency and Gaussian window length of 512. Again, brighter colors indicate frequency content with higher power and darker color indicates lower.

Firstly, both plots confirm the separation of higher-frequency contents into two intervals. As can be seen in the spectrograms, higher frequencies are well represented in the articulations of "ə". Secondly, this time the simulated results of Praat are similar to real sounds from the aspect that transitions caused

by oral wall vibration are not audible however still recognizable in the spectrograms. They are no longer frequency content with almost zero power at all as viewed in spectrogram plots.

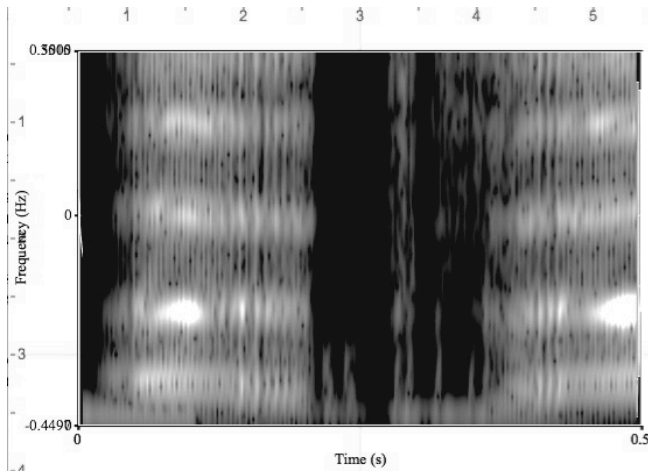


Figure 2.5. Spectrogram plots of utterances simulated by Praat

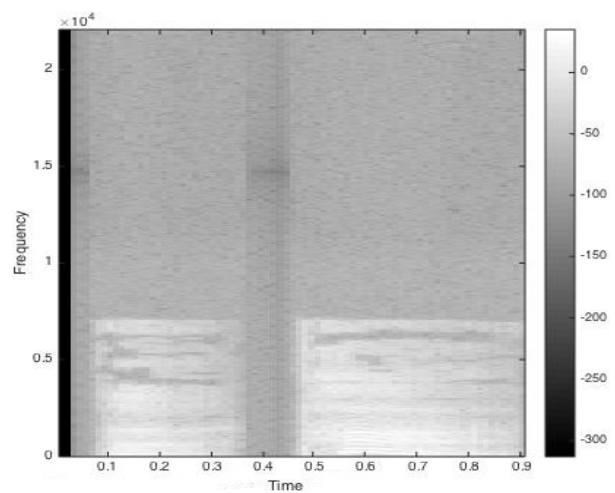


Figure 2.6. Spectrogram plots of identical utterances in real-world

2.3.1 Conclusion

Praat is capable of producing acoustic outputs in a more natural manner as compared to TADA. Therefore the hypothesis is made that the mass-spring model by design contributes to smoother transitions.

2.4 Others

The sound production process implemented by Praat refers not to learned motor behaviors but instead to independent control of all relative muscles. Praat aims at describing the acoustic consequences of articulatory activities from the standpoint of early age language learners who have no built-in coordinative articulatory tasks. It is able to appropriately mimic the babbling stage of young children. Therefore, it is recommended that researchers with interest in language acquisition process of early age children take a look at Praat.

References

- [1] TADA: Task Dynamic Application, web page. Available at: http://www.haskins.yale.edu/tada_download/index.php.
- [2] Praat: doing phonetics by computer, web page. Available at: <http://www.fon.hum.uva.nl/praat>.
- [3] Browman, Catherine P. & Louis Goldstein, "Dynamics and articulatory phonology," in *Status Report on Speech Research*, Haskins Laboratories, New Haven, no. 113, pp. 51–62.
- [4] Kelso, J.A. Scott, Elliot L. Saltzman & Betty Tuller, "The dynamical perspective on speech production: data and theory," in *Journal of Phonetics*, vol. 14, pp. 29–59, 1986.
- [5] Gestural Model, web page, p. 20. Available at: <http://www.haskins.yale.edu/research/gestural.html>.
- [6] Haskins Laboratories, web page. Available at: <http://www.haskins.yale.edu>.
- [7] *TADA (Tash Dynamics Application) manual*, Haskins Laboratories, Inc., 300 George Street, New Haven, CT. Available at: <http://www.haskins.yale.edu>.
- [8] Gestural Model, web page, p. 1. Available at: <http://www.haskins.yale.edu/research/gestural.html>.
- [9] Lindblom, Björn, James Lubker & Thomas Gay, "Formant frequencies of some fixed-mandible vowels and a model of speech motor programming by predictive simulation," in *Journal of Phonetics*, vol. 7, pp. 147–161, 1979.
- [10] Fowler, Carol A., "Coarticulation and theories of extrinsic timing," in *Journal of Phonetics*, vol. 8, p. 113, 1980.
- [11] Boersma, Paul, "Synthesis of speech sounds from a multi-mass model of the lungs, vocal tract, and glottis," in *Proceedings of the Institute of Phonetic Sciences*, University of Amsterdam 15, pp. 79–108, 1991.
- [12] Boersma, Paul, "An articulatory synthesizer for the simulation of consonants," in *Proceedings Eurospeech '93*, pp. 1907–1910.
- [13] Boersma, Paul, "Interaction between glottal and vocal-tract aerodynamics in a comprehensive model of the speech apparatus," in *Proceedings of the International Congress of Phonetic Sciences*, vol. 2, pp. 430–433, 1995.
- [14] Boersma, Paul, *Functional Phonology: Formalizing the Interactions between Articulatory and Perceptual Drives*. The Hague: Holland Academic Graphics, 1998.
- [15] Perrier, Pascal, Hélène Lœvenbruck & Yohan Payan, "Control of tongue movements: The equilibrium-point hypothesis perspective," in *Journal of Phonetics*, vol. 24, pp. 53–75, 1996.

[16] Passy, Paul, *Etude sur les changements phonétiques et leurs caractères généraux*. Paris, Librairie Firmin – Didot, 1891.