FORMANT DEFLECTION DIRECTIONS OF THE VOICED ALVEOLAR STOP CONSONANT IN DIFFERENT VOWEL CONTEXTS

by

Kang Li

A Dissertation Submitted to the Faculty of the

DEPARTMENT OF SPEECH, LANGUAGE, AND HEARING SCIENCES

In Partial Fulfillment of the Requirements

For the degree of

DOCTOR OF PHILOSOPHY

In the Graduate College

THE UNIVERSITY OF ARIZONA

2006

The University of Arizona

Graduate College

As members of the Dissertation Committee, we certify that we have read the dissertation prepared by Kang Li, titled “Formant deflection directions of the voiced alveolar stop consonant in different vowel contexts”, and recommend that it be accepted as fulfilling the dissertation requirement for the Degree of Doctor of Philosophy.

________________________________ Date: September 28, 2006
Brad Story

________________________________ Date: September 28, 2006
Julie Barkmeier-Kraemer

________________________________ Date: September 28, 2006
Kate Bunton

________________________________ Date: September 28, 2006
Jungmee Lee

Final approval and acceptance of this dissertation is contingent upon the candidate’s submission of the final copies of the dissertation to the Graduate College.

I hereby certify that I have read this dissertation prepared under my direction and recommend that it be accepted as fulfilling the dissertation requirement.

________________________________ Date: September 28, 2006
Dissertation Director: Brad Story

STATEMENT BY AUTHOR

SIGNED Kang Li

ACKNOWLEDGEMENTS

Many thanks are extended to people who have inspired, guided, helped, and laughed with me during the many years I spent at the Department of Speech, Language, and Hearing Sciences, University of Arizona for a great graduate school experience. I would especially like to thank my long-time advisor and committee chair, Brad Story, for his generous time and considerable mentoring. Throughout my doctoral work he continually encouraged me to develop independent thinking, research skills, and scientific writing skills.

I was also fortunate enough to have Kate Bunton, Julie Barkmeier-Kraemer, and Jung-mee Lee on my dissertation committee. They helped me through the grueling process of writing, cutting, editing, and rewriting drafts in the last year.

In large part, my dissertation research has been sponsored by the National Institute on Deafness and Other Communication Disorders through research grant R01-DC04789.

I owe a huge debt of gratitude to my parents for their unconditional love, support, and enthusiasm throughout not only the course of my Ph.D. studies, but my entire life. I am especially grateful to my husband and best friend, Frank, for his patience and for helping me keep my life in proper perspective and balance. This work, and my life, is dedicated to him.

TABLE OF CONTENTS

TABLE OF CONTENTS - Continued

LIST OF TABLES

LIST OF FIGURES

LIST OF FIGURES - Continued

ABSTRACT

A parametric model of the vocal tract area function was used to synthesize a series of alveolar stop consonant constrictions imposed on nine $/V_1V_2/$ transitions (e.g., /idæ/). The constrictions differed in spatial characteristics, places of occurrence along the vocal tract length, and the onset and release times to create different onglide and offglide formant deflection patterns. It was hypothesized that the formant deflection directions caused by the onset and release of the consonant constriction relative to the underlying vowel-to-vowel formant transitions (i.e., without consonant perturbations) provide information about the perceptual identity of the alveolar stop. Perceptual tests were conducted to assess the phonemic identities of the formant patterns produced by the model. The model parameter settings used to create the consonant constrictions that were perceived as /d/'s were analyzed to study the coarticulation between /d/ and different vowel contexts.

SPECIFIC AIMS

The long term goal of this research was to investigate the time-varying relation among the place of stop consonantal constriction, the formant transition patterns, and the perception of the consonant in $/V_1CV_2/$ utterances. The consonant under investigation was the voiced alveolar stop /d/. The vowels were /i/, /æ/, and /ɑ/. The hypothesis was that the onglide and offglide formant deflection directions from the underlying $/V_1V_2/$ formant transitions provide information about the place of articulation for this consonant. The voiced alveolar stop consonant was chosen because its place of articulation is located within a region of the vocal tract where slight changes in the location and geometry of the consonantal constriction may lead to large changes in acoustic features and, correspondingly large changes in perceptual outcomes. The three vowels were chosen to include both front-back and high-low vowel environments to create different formant transition patterns. The study was based on a kinematic area function model that produced time-varying speech sounds by superimposing vowel and consonant perturbations upon the area function of a neutral vowel. The following specific aims were proposed and carried out as three separate experiments:

1. To assess the consonant identification of synthesized /V_1CV_2/ sequences with the hypothesized formant deflection pattern in nine /V_1V_2/ transitions.

2. To characterize synthesis of the alveolar stop /d/ in terms of model parameters.

3. To compare formant transition patterns of synthesized and naturally spoken /V_1dV_2/ sequences.

BACKGROUND AND SIGNIFICANCE

PRODUCTION AND ACOUSTICS OF THE VOICED ALVEOLAR STOP

To produce the stop consonants in an intervocalic position, the vocal tract must become occluded momentarily at some point along the vocal tract. In English, the place of occlusion can be at or in the vicinity of the lips, alveolar ridge, or velum. Posterior to the occlusion, static pressure from the lungs is built up and then released by an abrupt downward motion of the articulators, such as the tongue and the lips. The explosion and aspiration of air serve as the sound source, which is an acoustic disturbance superimposed upon the air flow (Fant, 1960). The disturbance gives rise to frication and a transient release of the air. The vocal tract spatial characteristics anterior to this source create a filter effect that yields the spectral characteristics of this group of sounds.

The articulatory coordinations during the production of the above sequence of movements and acoustic events have been studied and described by many researchers. Stevens (1998) described the coordination required for consonant production as the constriction formed by a major articulator facilitated by a secondary structure. For example, the alveolar occlusion is produced by the upward movement of the tongue tip to contact the alveolar ridge. The production of such a major movement is facilitated by a secondary front movement of the tongue body or a rise of the mandible. Depending on the articulators involved in the closure and the location of the closure, the rate of the movement is different (Stevens, 1998). For example, in a labial stop consonant /b/, the lower lip movement velocity right after the release of the constriction is about 25 cm/s (Fujimura, 1961; Sussman, MacNeilage, & Hanson, 1973); The rate of the mandible movement in the production of /p/, /t/, and /k/ (measured as the rate of the lower incisors movement) is about 10-20 cm/s (Kent & Moll, 1972; Nelson, Perkell, & Westbury, 1984). If the occlusion is formed by the contact between the tongue body and the hard or soft palate, the rate of the articulator movement is slower than that of lips or tongue tip where the occlusion is located (Stevens, 1998). The significance of the different rates of articulatory movements is the change of the spatial characteristics of the occlusions and the consequent aerodynamic changes during the formation of the occlusions, which in turn may result in acoustic and perceptual changes.

In addition to the coordinations among the articulators during the production of an occlusion, the vowels preceding and following a stop consonant also affect the rate of the movement since an articulator has to move from a preceding vowel position to form a constriction at a specific location and move to the position for the following vowel. Depending on the locations of the vowel constrictions and the degrees of the vowel openings, the geometry and location of the stop occlusion are perturbed so that the movements from the preceding and following vowels overlap in time with the consonant occlusion to form smooth articulatory transitions.

The complex fashion of the movement overlap between consonants and vowels has been studied and modeled by many researchers (Öhman, 1966, 1967; Fant, 1969; MacNeilage, 1970; Kent & Minifie, 1977; Löfqvist & Gracco, 1999, 2002). Löfqvist and Gracco (2002) have studied the the kinematics of the tongue tip and tongue body in the production of the alveolar and velar stops in different V_1CV_2 sequences. The tongue movements were measured using a three-transmitter magnetometer system. The receivers were placed at four positions on the tongue surface at the tip, blade, body, and rear. Control positions at the nose, upper incisors, and palate were also placed with receivers to correct the head movements and record the upper occlusion plane. The velocities of the tongue movements were calculated as the square root of the sum of the squares of the first derivatives of the position signal along the vertical and horizontal directions ($\sqrt{\dot{x}^2 + \dot{y}^2}$). The results showed that the tongue movements followed a curved path between the two vowel positions, instead of a direct straight line. Generally, the position of the body at the release of the stop was higher and more forward than that before the release. In addition, the velocity of the tongue movement before, during, and after the occlusion release was different and dependent on the vowel context. For example, the velocity at closure was more influenced by V_1, especially the movement of the tongue tip at the closure for the alveolar stop. Particularly, the tongue movement was slower when the vowel /i/ preceded the consonant than when /ɑ/ or /u/ did. The explanation was that tongue movement velocity was proportional to the distance the tongue had to travel. When /ɑ/ or /u/ preceded the velar or alveolar stop, the tongue had a longer distance to travel in order to complete the occlusion than when /i/ preceded the stop since the tongue was already raised in the oral cavity in the latter case. This study also showed that the tongue body was moving during the consonant occlusion and that the amount of the movement varied with vowel context. These findings agreed with those of Houde (1968, as cited in Löfqvist $et\ al.$, 2002) and Perkell (1969).

Mooshammer, Hoole, and Kühnert (1995) analyzed tongue movements during velar consonant production in various vowel contexts by means of electromagnetic articulography. Three transmitter coils mounted on a helmet generated an alternating magnetic field at three different frequencies. Three sensor coils mounted on the tongue surface detected the strength of the magnetic field, which was inversely proportional to the distance between the sensor coils and the transmitter coils. The raw distances between coils were then converted and calibrated to horizontal-vertical coordinates in the mid-sagittal plane. Two measurements were performed. First, tongue movement trajectories during VCV utterances were traced. Second, the rear tongue coil displacements from the beginning to the end of the consonant closure were measured. The results suggest that the speed of tongue movement at the beginning of the consonant closure depended on the preceding vowel (/ɑ/$>$ /u/$>$/i/). They also confirmed the findings of Houde (1967, in Mooshammer, Hoole, & K$\ddot{u}$hnert, 1995) and Perkell (1969) that the movement of the tongue in the production of VCV for velar stops showed an elliptical path indicating a forward movement of the back of the tongue during the occlusion time. The amount of the forward movements depended on the identity of the initial vowel. When it was /i/, the displacement was greater than when it was /u/ or /ɑ/.

The findings in the above studies are significant because they demonstrate that the articulation of stop consonants is highly vowel dependent. Consequently, the acoustic results of the vocal tract dynamics in the production of consonant-vowel sequences have been studied in association with different vowel contexts. Fant (1969, 1973) divided a $/CV/$ transition into two parts. The first rapidly progressing part was related to the release of the consonantal occlusion (burst). The second part was the formant transition pattern (henceforth, F-pattern) that was related to the tongue movements that formed the coarticulation of the consonant and the following vowel. F-pattern data after the bursts of both voiced and unvoiced stops followed by nine different long vowels in synthetic syllables were collected. The results showed that fluctuations in tongue configuration in the production of CV syllables could predict the formant transitions, especially for F2 and F3. For example, the tongue body configuration for the vowel /i/ in the syllable /bi/ at the release of the consonant was perturbed in a way that the configuration was somewhat like that of a neutral vowel. Acoustically, F2 at the release of the burst for /bi/ (end of the first 5-10 ms) jumped from about 1200 Hz to about 1700 Hz, which was close to the F2 value of a neutral vowel 1500 Hz, but lower than a typical F2 of a steady state vowel /i/ of about 2300 Hz. However, the F2 transition of the utterance /bɑ/ did not show such a perturbation pattern and the transition was small. The difference between /bi/ and /bɑ/ was reproduced in Figure 1. The white squares are the traces of F2 contours in /bi/ (left panel) and /bɑ/ (right panel). The far left square on the /bi/ spectrogram indicates the F2 value of 1600 Hz of the first 10 ms or so right after the occlusion was released.

The coarticulatory influence of vowels and consonants on acoustic characteristics was also studied by Öhman (1966), who measured formant transition patterns of naturally spoken V_1CV_2 sequences in Swedish and American English. He showed that formant transitions from the consonant to V_2 depended on the identity of V_1. For example, when /b/ was preceded by /u/, the starting F2 value of /bV_2/ transition was 500 Hz. When /b/ was preceded by the Swedish /y/, the start F2 value was 1300 Hz. This observation contradicted another popular scenario, the “fixed F2 locus” proposed by Delattre, Liberman,

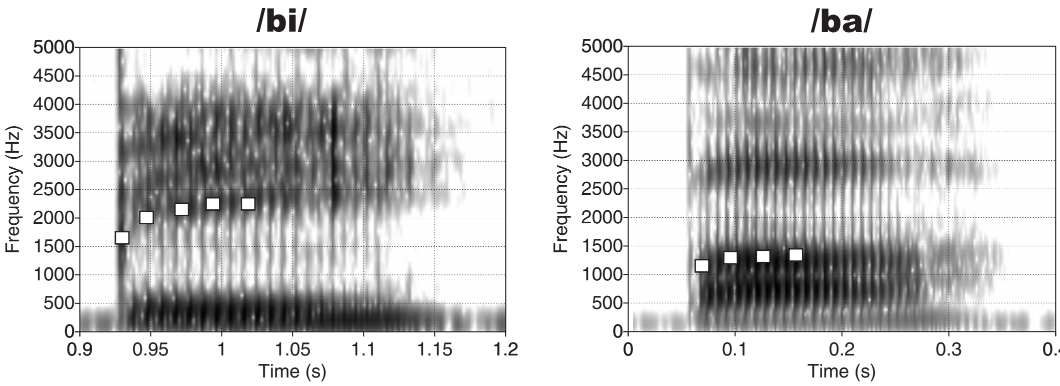


Figure 1: *Spectrograms of the utterances /bi/ (left panel) and /bɑ/ (right panel) by a male speaker. F2 transitions are traced by the white squares. For the /bi/ utterance, the F2 value of 1700 Hz during the first few 10 milliseconds or so (indicated by the far left white square) is similar to the F2 of a neutral vowel (1500 Hz) as a result of the neutral vowel-like tongue configuration. For the /bɑ/ utterance, the F2 transition does not show such perturbation pattern.*

and Cooper (1955). They studied the perception of synthetic stop consonant and vowel syllables and found that each stop had a characteristic and fixed F2 locus corresponding to the place of articulation. They concluded that the CV formant transition could be regarded as the formant movement from the locus to the formant of the following steady state vowel. The "fixed F2 locus" theory has not been widely accepted since there is little evidence that supports the fixed vowel-independent F2 starting locus for a given consonant (Öhman, 1966; Kewley-Port, 1982; Fant, 1969).

Kewley-Port (1982) measured the formant transition patterns in naturally spoken consonant-vowel syllables consisting of eight vowels /i, ɪ, e, ɛ, æ, ɑ, o, u/, and three voiced stop consonants /b, d, g/. Formant transitions were traced from the consonant onset frequencies to the steady-state vowel frequencies. Variance analysis was used to determine if the acoustic measurements were significantly different for the three places of articulation. No fixed loci were found for the voiced stop consonants /b, d, g/ when they were followed by the investigated vowels, which agreed with Öhman 's findings (1966). The results of the study showed that F2 and F3 onset frequencies were not distinctive acoustic correlates of place of articulation. Rather, discriminate analysis indicated that if F1 and F2 onset frequencies were plotted separately for each vowel context, stops were classified correctly at least 97% of the time. In other words, when the context was known, the formant transition patterns presented in F2-F3 space were reliable acoustic correlates of the place of articulation. Figure 2, adapted from Figure 2 in Kewley-Port, 1982, shows the F2-F3 spaces for three vowels /i/, /æ/, and /ɑ/.

It can be seen from Figure 2 that F2 and F3 transition onsets are distinctive for /b, d, g/ in the three panels (separate circles in each panel). The onset F2 and F3 frequencies for the same consonants are different in different vowel contexts. For example, for labial consonant /b/, average F2 and F3 onsets are 1816 Hz and 2462 Hz in context /i/, 1526 Hz and 2156 Hz in context /æ/, and 1069 Hz and 2425 Hz in context /ɑ/. For the alveolar consonant /d/, average F2 and F3 onsets are 1920 Hz and 2587 Hz in context /i/, 1786 Hz and 2548 Hz in context /æ/, and 1660 Hz and 2628 Hz in context /ɑ/. For the velar consonant /g/, average F2 and F3 onsets are 2322 Hz and 2835 Hz in context /i/, 2112 Hz and 2416 Hz in context /æ/, and 1733 Hz and 2367 Hz in context /ɑ/.

Although "fixed" loci were not supported by the measurements in Kewley-Port, 1982, she suggested that the relative positions between the formant loci and the vowel formants might show some invariance. For example, F2 loci of /g/ were higher than those of /d/ in

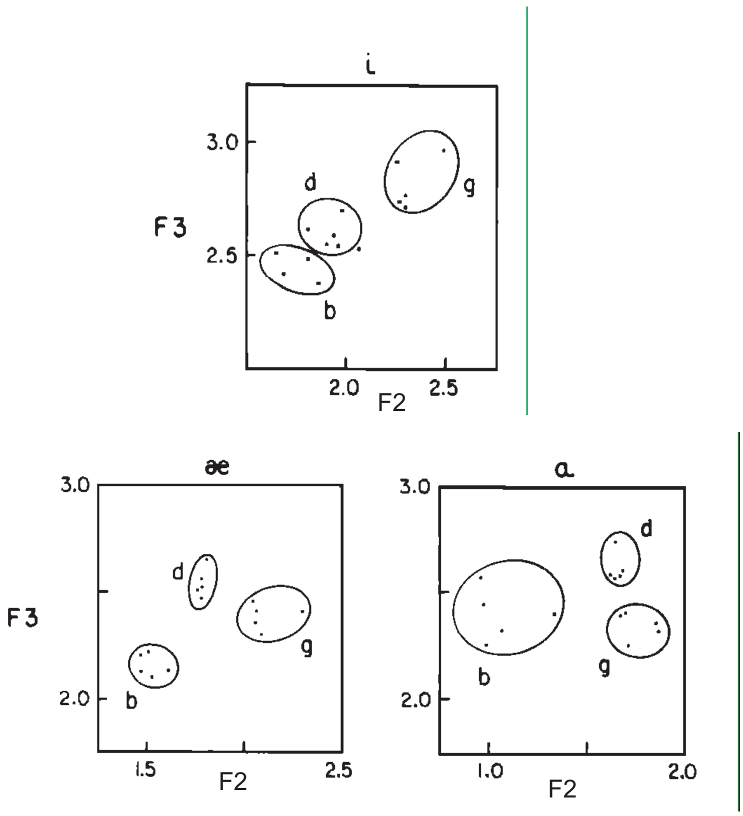


Figure 2: *Place of articulation classified based on F2 and F3 onset frequencies when the vowel context in a CV syllable is known. Frequency axes are in kHz (adapted from Figure 2 in Kewley-Port, 1982, permission acquired from the publisher and the author).*

front vowels (/i, ɪ, e, ɛ, æ/), but lower than those of /d/ in back vowels /ɑ, o, u/. Similarly, F2 loci of /b/ were lower than those of /d/ in front vowel contexts; F2 loci of /b/ were lower than those of both /d/ and /g/ in back vowels. These observations agreed with the relative positions suggested by Delattre, $et\ al.$ (1955).

PERCEPTION OF THE VOICED ALVEOLAR STOP CONSONANT

Perception of stop consonants has been shown to be associated with varied types of initial and transitional changes in spectral patterns. Dorman, Studdert-Kennedy, and Raphael (1977) assessed the roles of burst and formant transitions in the perception of location of articulation for the three voiced stop consonants spoken in nine different vowel contexts. They removed the bursts of naturally spoken $/CVC/$ syllables. The removed bursts were then attached to the front of the corresponding naturally spoken $/VC/$ (same vowel and consonant) syllables to create initial consonants cued with bursts only. $/CVC/$ syllables with the initial noise induced resonances removed (i.e., only voiced transitions and vowels remained) were also created. These syllables were presented to listeners in an identification task. The results suggested that the burst and the transitions were complementary cues for the identification of the place of articulation. In addition, the effects of these cues were context-dependent, influenced primarily by the following vowel. Take the alveolar stop for example, the burst was a moderately effective cue before front vowels, such as /i/ and /ɪ/. This cue became weaker as the lip rounding effect increased, such as the gesture for vowel /u/ because the increasing length of the front cavity strongly attenuated the burst energy. The formant transitions were moderately effective cues (above 60%) when back and central vowels followed the consonant and weak cues when front vowels followed. Their result of vowel dependency agreed with that of Cooper, Liberman, Borst, and Gerstman (1952); and Schatz (1954).

In order to further evaluate the effects of the initial burst and formant transitions on the perception of place of articulation, Stevens and Blumstein (1978), and Blumstein and Stevens (1979, 1980) used an alternative approach of short-term spectral characteristic analysis. They extracted the spectral shapes of the initial bursts sampled at the first 10-20 ms of spoken $/CV/$ utterances. Then a series of $/CV/$ syllables was synthesized where formant transitions and presence of the initial bursts were systematically controlled based on the extracted parameters. Listening tests were conducted to evaluate the perceptual effect of different spectral shapes and formant transitions. They found that the acoustic invariance for each place of articulation for the stops could be specified in terms of its unique context-independent and static spectral shapes of the consonant burst. They summarized the shapes for the three places of articulation as “diffuse falling” for the alveolars; “diffuse rising” for the labials; and “compact” for the velars. Figure 3 depicts the three shapes sampled from the syllables /Cɑ/. Stevens and Blumstein argued that the vowel-independent spectral shape for each place of articulation reflected the acoustic characteristics of the vocal tract cavity in front of the consonant constriction at the moment it was released. Listeners made use of such invariant acoustic characteristics for this group of sounds.

Kewley-Port (1983) extended the time windows of initial burst release to 20-40 ms and confirmed the existence of vowel-independent invariant cues for the voiced stop con-

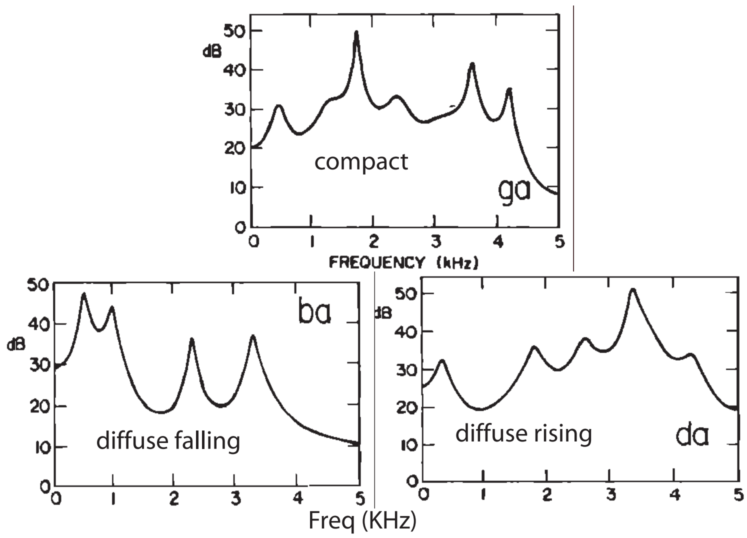


Figure 3: *Spectral shapes sampled at the release of the three voiced stop consonants /b, d, g/ in the utterances /Ca/. They are named as compact, diffuse rising, and diffuse falling as indicated (adapted from Figure 1, Blumstein and Stevens, 1979, permission acquired from the publisher and the first author).*

sonants in the initial portion of natural stop-vowel syllables. In addition, she incorporated temporal information in the spectral analysis and proposed that both temporal and spectral information at the release of the consonant constriction and onset of the following vowel were robust acoustic cues for the detection of the place of articulation. Three time-varying energy change features were proposed and judged to be the criteria for the detection of place of articulation for the voiced stops. They included tilt of the spectral shape at the burst onset (rising for /d/, falling for /b/, and flat for /g/), onset of low frequency energy (separates velars with long VOTs from labials and alveolars), and mid-frequency peaks extending over time into the vowel onset (separates velars with salient mid-frequency peaks from labials and alveolars).

Kewley-Port, Pisoni, and Studdert-Kennedy (1983) conducted three perceptual experiments on natural and synthesized $/CV/$ syllables to test the three dynamic features derived in Kewley-Port (1983). The first experiment confirmed that perceptual cues for the place of articulation could be found in the initial 20-40 ms of a stop-vowel syllable at the release of the burst. The second and third experiments were to compare the perceptual effect of the three time-varying acoustic features proposed in Kewley-Port (1983) and the static spectral shapes (Figure 2) proposed by Stevens and Blumstein (1978), and Blumstein and Stevens (1979, 1980). In experiment two, two sets of $/CV/$ syllables were synthesized using a digital, parallel-resonance synthesizer. One set of syllables presented the static spectral shapes proposed by Stevens and Blumstein. The other set presented the dynamic features proposed by Kewley-Port. Two similar sets of syllables were synthesized for experiment three using the LPC algorithm. The results of these experiments showed that the identifications of the alveolar and velar stops based on the static cues were below the chance level. On the other hand, the identifications of the three stops based on the dynamic spectral cues were much higher and reliable. The results supported the finding that spectral characteristics of stop consonants are dynamic and context-free.

Lahiri, Gewirth, and Blumstein (1984) studied the acoustic invariance for place of articulation for /d/ and /b/ in Malayalam, French, and English, using the static shapes summarized by Stevens and Blumstein (1978), and Blumstein and Stevens (1979, 1980). However, their methods could not differentiate the two consonants in any of the languages studied. Instead, they developed a new metric based on the energy change over time from the burst onset to the onset of voicing. The metric first defined the slopes of the spectra for the burst and the onset of voicing. Next, energy distributions in terms of low and high frequencies were quantified based on the two slopes. Values less than 0.5 indicated the energy change pattern of alveolars, i.e., no more than half of the energy changes from the burst to the voice onset occurred at high frequencies compared to low frequencies. Values greater than 0.5 indicated the energy change pattern for labials, i.e., small energy changes from the burst to the voice onset occurred at low frequencies or at both low and high frequencies. Perceptual evaluations of their metric were conducted on three types of synthesized $/CV/$ syllables between /b, d/ and /i, e, ɑ, o, u/ using the Klatt formant synthesizer with parameters extracted from spoken French $/CV/$ syllables. For the first type, spectral shapes at both the burst onset and voice onset were adjusted so that the energy distributions over time agreed with the corresponding proposed patterns. For the second and third types, the spectral shapes at the burst onset (type 2) and voice onset (type 3) were altered separately, with the spectral shapes characterizing one place of articulation and the formant transitions characterizing the other. Their listeners were all native English speakers with good command of French and Russian. The results confirmed their metrics, i.e., the first type was identified with high percentages in all vowel contexts (above 91%) and the second and third types showed higher identifications indicated by the spectral characteristics, compared to the formant transition indications. The results also showed some vowel-dependent effects. For example, $/CV/$ stimuli that showed the /b/ energy distribution were sometimes identified as /dV/ in vowel contexts /i, e, ɑ, o/ when the spectral shape at the burst onset was altered, and in vowel context /i/ when the spectral shape at the voice onset was altered. However, they attributed the context effects to the speech synthesis artifact.

In contrast, Dorman and Loizou (1996) disagreed with the observation of spectral energy distribution being an invariant cue for the identification of place of articulation. They conducted a study on naturally spoken /bV/ and /dV/ English syllables. When stimuli with conflicting cues were presented to the listeners with relative energy distribution specifying one place of articulation and formant transitions specifying the other, their perceptual identifications corresponded to the formant transitions. The results were inconsistent with those of the Lahiri $et\ al.$ (1984) study. Dorman and Loizou (1996) suggested the inconsistency was due to the “non-optimal” parameter settings in the Lahiri $et\ al.$ (1984) study.

Parikh and Loizou (2005) assessed the perceptual significance of the spectral tilt and burst onset frequency in the consonant identification of clean and noise corrupted $/VCV/$

utterances with different signal to noise ratios (-5dB, 0dB, and 5dB S/N). The vowels were /i, ɑ, u/, and consonants were /b, d, g, p, t, k/. Two types of noises were used to corrupt the speech signal. One was multi-talker babble. The other one was a speech shaped noise constructed by filtering white noise through an FIR filter with a frequency response of long-term spectrum of vowels produced my male and female talkers. Measured spectral tilt (slope of the line that connected 1000 Hz and 5000 Hz peaks) and the burst onset frequencies showed heavily corrupted cues at all levels of S/N ratios. However, the perceptual results showed high identification percentages on the consonants at most of the noise levels (S/N of 0 and 5). Context effect was not discussed. They concluded that listeners must be relying on cues other than the spectral tilt and burst onset frequency to identify the consonants.

These experiments on both synthesized and naturally spoken stop-vowel syllables have produced different conclusions regarding the usefulness of the formant transitions and spectral characteristics of the burst as perceptual cues for the place of articulation. The methodology employed in these studies is based on two primary steps. The first step is to make measurements on naturally spoken $/CV/$ or $/VCV/$ syllables and extract certain acoustic patterns as the hypothesized perceptual cues. The second step is to independently control the acoustic parameters to create the hypothesized pattern in synthetic $/CV/$ or $/VCV/$ syllables for perceptual tests. By doing so, the acoustic parameter settings are adjusted without reference to articulation and possibly unknown spectral or temporal correlates that the perceptual system may respond to are altered or removed. Many of the differences in the results of these studies may have been due to the nature of the stimuli used in the perceptual tests. One way to eliminate the possibility of unrealistic acoustic parameter settings is by directly controlling the vocal tract shape with a parametric model of the vocal tract area function. The advantage of such a synthesizer is that it can produce speech signals in terms of the interactions of acoustic pressures within the vocal tract configuration during the normal speech production. The acoustic results can then presented to listeners in perceptual tests.

NUMERICAL MODELS OF THE VOCAL TRACT

One of the purposes of the present study was to demonstrate the relation between the production and acoustics of the alveolar stop consonant in different vowel contexts. If it is believed that natural speech is characterized by overlapping articulation between segments, then observed acoustic correlates must have ways to reflect such contextual effects. Various numerical models have been developed to re-create the articulatory-acoustic relations.

In an attempt to interpret the results of Öhman (1966), Öhman (1967) developed a model for the observed context-dependent shape variations over time in a given /V_1dV_2/ sequence, as shown in Equation 1.

$$s(x,t) = v(x,t) + k(t)[c(x) - v(x,t)]w_c(x) \tag{1}$$

In the equation, $v(x,t)$ represents a unique vowel vocal tract shape (x indicates the position along the vocal tract and t indicates time). At any given time, the shape of the vocal tract $s(x,t)$ is a linear combination of a single context-independent consonant shape $c(x)$

and its associated vowel-dependent "co-articulatory weighting function" $w_c(x)$. The consonant shape is turned on and off by a time interpolation term $k(t)$ that increases from 0 to 1 continuously as the vocal tract begins with a pure vowel shape and moves to a full consonant configuration. At its full consonant configuration, the vocal tract shape still retains the vowel contour to an extent controlled by the weighting function $w_c(x)$. In other words, the temporal and spatial coordination between the vowel and consonant gestures is controlled by the time interpolation term $k(t)$ and weighting function $w_c(x)$. Comparison of the computed shape $s(x,t)$ with X-ray data showed that they matched reasonably well for coarticulation functions for /d/ in symmetric vowel contexts of /i/, /u/, /ɑ/. The model demonstrated that the rules of coarticulation were dynamic and context-dependent in nature (Öhman, 1966, Ericsdotter, Lindblom, and Stark, 1999).

Öhman's (1967) articulatory implementation of coarticulation inspired numerous models that followed this approach. Browman and Goldstein (1987, 1990) applied their task-dynamic model to represent the orchestration of invariant articulatory gestures. Each gesture was a target-specific movement of a tract variable (8 of them) expressed by a second-order linear spring-mass system. The stiffness of the system controlled the duration of the gesture and the damping ratio determined if the movement overshot its target value or if the movement was damped beforehand. The coordination between gestures was specified by phase transitions between the systems. For example, the word "add" had two gestures. One was for /æ/, the other one was for /d/. Changing the stiffness of the gesture for /æ/ or /d/ changed the time the articulator used to reach the vocal tract configuration for that target. Changing the relative phasing between the gesture for /æ/

and the closure gesture for /d/ changed the acoustic transition duration between them. The phase change was continuous to account for varying degrees of overlap between gestures.

Similarly, Fujimura (1992) outlined a theory of phonemic implementation that was an attempt to treat vowel-to-vowel movement as a base function, onto which consonantal gestures could be superimposed as “perturbations”. Båvegård (1995) adopted a triangular function to represent the consonant constriction superimposed on the vowel substrate. The function was specified by the position of constriction, the cross sectional area at the constriction, the width of the triangle, and the symmetry of the triangular constriction. The results on limited number of phonemes verified the possibility of combining the vocal tract area function with a consonantal constriction. Similarly, Carré and Chennoukh (1995) successfully showed the superposition principle that governed the vowel-to-vowel transitions and the consonant closure for $/V_1CV_2/$ syllables. At a specified point in time, the consonantal closure was synchronized with the movement between V_1 and V_2. The duration of the closure was equal to the $V_1 - V_2$ transition.

THEORETICAL OVERVIEW OF AREA FUNCTION MODELS

Area functions

Models that consist of numerous concatenated tubular sections whose cross-sectional areas along its length vary as a function of time have become a standard representation of speech production. The acoustic behavior of a single tube is simple to describe, in terms of the volume velocity, and pressure deviation from the mean tube pressure. Such simplification allows modifications of the effective movement and positions of major articulators (such as the lips, jaw, tongue body, tongue tip, and velum) by providing a set of time-varying areas that represents the vocal tract dimensions along its length.

Fant (1960) developed a 3-section model and demonstrated the resulting F-patterns of different tongue positions as shown in Figure 4. The first shape on the top illustrates the uniform tube that represents the neutral vowel shape. If the length of the vocal tract is 17.6 cm, resonances of such a uniform tube are at every 1 KHz, starting with 500 Hz.[1] The next three configurations are three-tube abstractions of the vocal tract shapes of velar and palatal consonants and their corresponding F-patterns. The narrow middle tube models the vocal tract constriction. Depending on the location of constriction and the lengths of the front and back tubes, the F-patterns are different, especially the values of F2 and F3.

Models with more physiologically-relevant parameters have been developed to account for various vocal tract configurations and articulator movements in speech production. Mrayati, Carré, and Guérin (1988) developed the Distinctive Region Model (DRM) where the vocal tract length was divided into eight regions. The number and size of the regions were deduced based on the sensitivity functions of the vocal tract (Fant & Pauli, 1974). They closely correspond to the major places of constrictions that are commonly engaged in vowel and consonant productions, such as three front regions corresponding to the lips and alveolar; two central regions corresponding to the tongue; and three back regions corresponding to the larynx. Figure 5 illustrates the eight regions along the length

[1] Speed of sound was 352 m/s.

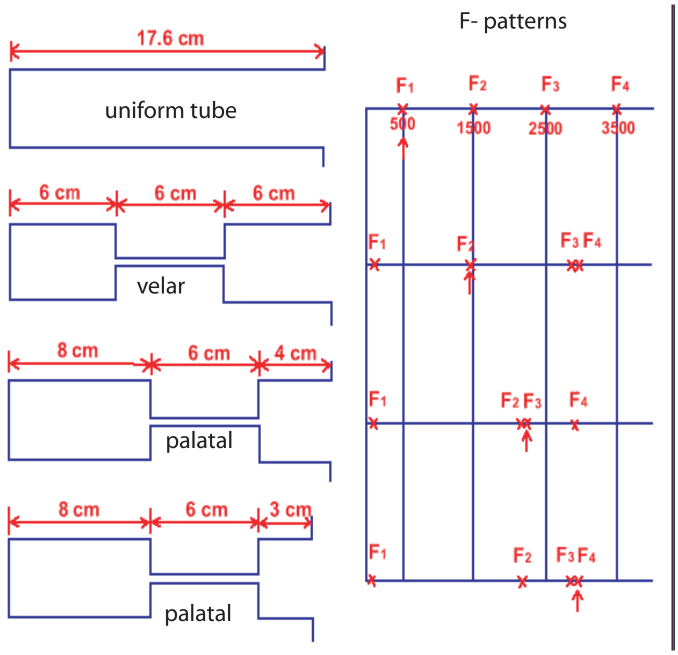


Figure 4: *F-patterns of four different mid-saggital vocal tract profiles represented by 3-tube resonator models (adapted from Figure 1.4-7 in Fant, 1960, p73).*

of the vocal tract and the distinctive formant direction changes resulted from the constriction in a particular region while all other regions remain in their neutral configuration. Although the simple parameters can not capture all the phonetic details of major vowels and consonants, the DRM provides the evidence of simulating the complicated acoustic behavior of the human vocal tract by concatenated tubes.

The vocal tract length can also be divided into more sections of tubes (typically around 40) that are of equal length. As the length becomes small, the shape of the approximation of the series of tubes will converge to that of the continuous vocal tract. The acoustic interactions of each section can be represented by wave equations that describe the relation between the volume velocity and pressure. Such a model provides flexibility in

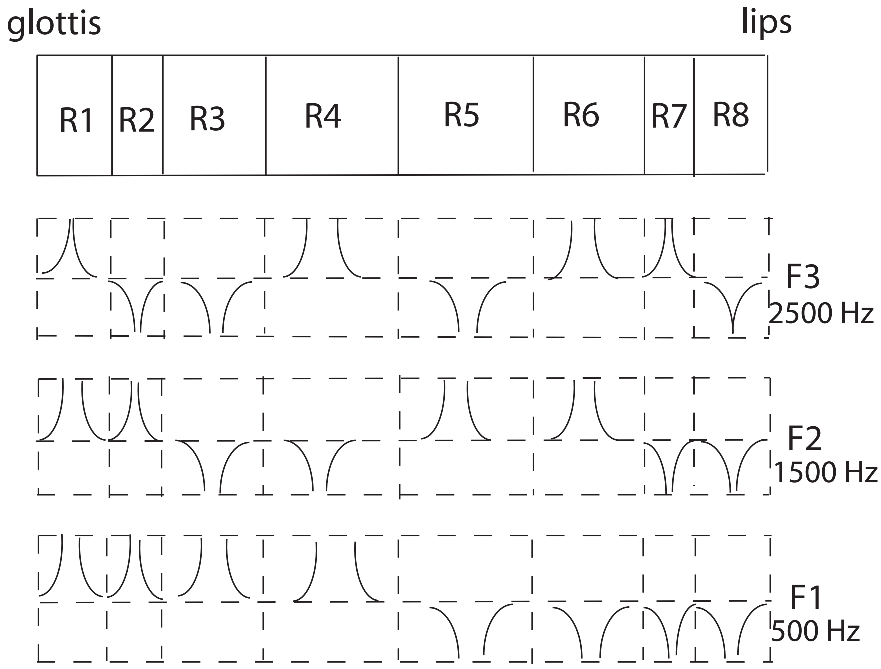
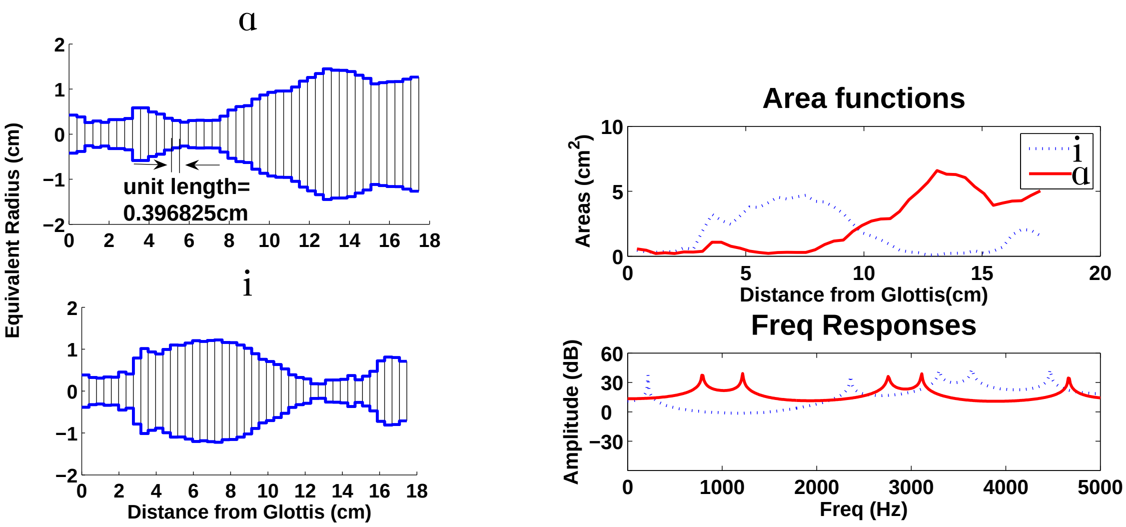


Figure 5: *The distinctive regions along the length of a uniform vocal tract and formant direction changes resulted from a constriction in a particular region while all other regions remain in their neutral configuration. Adapted from Figure 1, Carré and Chennoukh, 1995.*

terms of parameter control (such as the length variations of the vocal tract, Goldstein, 1980, in Story, 2005; Fitch & Giedd, 1999), synthesis of different types of speaker (Story, Titze, & Hoffman, 2001; Story & Titze, 2002; Story, 2002), and choice of simplifying assumptions (how closely the model simulates the speech production mechanism).

The parametric model of the vocal tract area function developed by Story (2005) contains 44 tubular sections. Section one is positioned right above the glottis and section 44 is at the lips. Figure 6 shows the tube sections (panel a) and area functions of vowels /ɑ/ and /i/ and their frequency responses (panel b). The unit section length is $\Delta = 0.396825$ cm. This number originates from a particular type of wave propagation algorithm that can be used to synthesize speech (e.g., Liljencrants, 1985; Story 1995), where $\Delta = \frac{\text{speed of sound}(c)}{2\times\text{sampling frequency}(F_s)}$. Choosing $c = 35000$ cm/s and $F_s = 44100$ Hz, $\Delta = 0.396825$ cm. In addition, representation of an area function by 44 sections is not a requirement of the general structure of the model. The number derives from the approximate spatial resolution obtained in Magnetic Resonance Imaging (MRI)-based reconstructions of vocal tract shape (Story, Titze, & Hoffman, 1996). It is also convenient to use 44 elements for simulating speech with acoustic waveguide models because it allows for a sampling frequency of 44.1 kHz when the tract length is approximately 17.5 cm (typical adult male). The formulation of the model, however, is general enough that area functions with any number of sections can be produced.



(a) (b)

Figure 6: ***(a).*** *Concatenating tube sections approximate the vocal tract configurations of vowels /ɑ/ (upper) and /i/ (lower). For both vowels, section 1 is positioned right above the glottis and section 44 is at the lips. The unit length of each tube is about 0.4cm.* ***(b).*** *Area functions (upper) of vowels /ɑ/ and /i/ and their frequency responses (lower).*

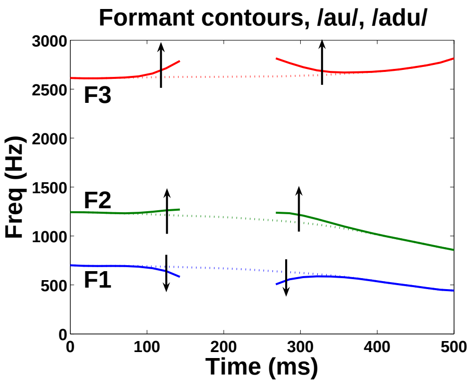
A layered system of speech production

The structure of the model is based on Öhman's (1966) concept that consonant gestures are superimposed on a context-dependent vowel substrate during the production of $/V_1CV_2/$ utterances. To produce a $/V_1CV_2/$ utterance, the vocal tract shape is changed smoothly from the V_1 gesture to the V_2 gesture, then a consonant gesture is activated and superimposed on the $/V_1V_2/$ sequence at a certain time. To synthesize the sequence, the spatial characteristics (location and geometry) of the vowel to vowel transitions are produced by smooth and continuous variations of the cross-sectional areas (area function) along the vocal tract length over time. These time-varying areas are expressed as perturbation functions superimposed on the neutral vocal tract shape (mean shape of 10 English vowel shapes). The consonant constriction is created by superimposing a shaping function to perturb the vowel to vowel transitions. The perturbations are activated and deactivated at different times to control the durations of different gestures and the temporal overlap between gestures.

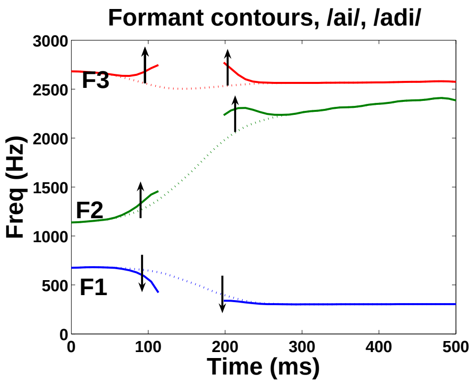
The layered system of speech production and the vowel-dependent formant transition patterns suggest that the acoustic effects should be analyzed in a similar fashion, i. e., the formant transitions caused by the consonant constriction should be analyzed with respect to the underlying vowel-to-vowel formant variations. An example is shown in Figure 7. Time-varying formant transitions of the synthesized /ɑdu/ and /ɑdi/ sequences (solid lines) are plotted against those of the synthesized vowel-to-vowel sequences /ɑu/ and /ɑi/ (dotted lines). The horizontal axis is the time in ms and the vertical axis is frequency in Hz. For both pairs, the onset of the voiced alveolar causes formant deflections from the underlying VV contours. The directions of the deflections relative to the underlying vowel formants (indicated by the arrows) show some similarity between the two vowel contexts (Story, 2005). The onset of the consonant perturbation causes F1 to decrease, and F2 and F3 to increase for both sequences. The offset of the consonant perturbation causes similar deflections. If only the formant variations of the $/V_1dV_2/$ are considered, however, without reference to the underlying vowel formant transitions, F2 trajectories of the two $/V_1dV_2/$ sequences show different directions. For the/ɑdi/ sequence, F2 rises into the steady state F2 of vowel /i/, but for the /ɑdu/ sequence, F2 falls into the steady state F2 of vowel /u/. The different transition directions are one of the examples of varying acoustic features resulting in stable phonemic perception (Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967). This observation suggests that the dynamic and multi-layered mechanism of speech production may be represented at the acoustic level too. The acoustic representation of the constriction should be analyzed within the vowel-to-vowel context onto which the constriction is superimposed (Story, 2005).

The deflection direction pattern shown in Figure 7 can also be observed in the neutral vowel context /əə/ using the DRM (Carré and Chennoukh, 1995), as shown previously in Figure 5. The neutral vowel context is represented by a uniform vocal tract configuration. A constriction at the region corresponding to the place of constriction for the alveolar stop, R6, causes F1 to be deflected downward, F2 and F3 upward relative to the underlying neutral vowel formants at 500 Hz, 1500 Hz, and 2500 Hz.

The formant transition patterns of the voiced stop consonants in naturally spoken CV sequences in Kewley-Port (1982) (see Figure 2) also agreed with the observed offglide



(a)



(b)

Figure 7: *Formant contours of synthesized syllables /ɑdu/ (a) and /ɑdi/ (b). The horizontal axis is time time in ms, the vertical axis is frequency in Hz. The dotted lines indicate the underlying vowel formant transitions of /ɑʊ/ and /ɑi/ extracted from spoken VV syllables. The solid lines show the time-varying formants when the consonant perturbation $/d/$ is generated and superimposed on the vowel to vowel transitions.*

deflection pattern shown in Figure 7. In the sequence /dɑ/, the average first three onset formants were 392 Hz, 1660 Hz, and 2628 Hz. Comparing to the first three steady-state vowel formants of /ɑ/ at 669 Hz, 1166 Hz, and 2523 Hz, the onset of the alveolar constriction caused F1 to be deflected downward, F2 and F3 upward. Similarly, in the sequence /dæ/, the average first three onset formants were 370 Hz, 1786 Hz, and 2548 Hz. Comparing to the first three steady-state vowel formants of /æ/ at 621 Hz, 1696 Hz, and 2481 Hz, the onset of the alveolar constriction caused F1 to be deflected downward, F2 and F3 upward. In utterances /di/, the average onset formants were 313 Hz, 1881 Hz, and 2587 Hz. The average steady-state vowel formants were 340 Hz, 1916 Hz, and 2582 Hz. As can be seen, F1 was deflected downward, but F2 and F3 were almost flat without much deflection. This might be caused by the close proximity of F2 and F3 of the vowel.

The hypothesized deflection direction pattern for $/VdV/$ sequences (Figure 7) is also observed in some of Öhman's data on $/VdV/$ utterances (1966), as shown in Figure 8. The formant transitions show different directions. For example, F2 onglide transitions of /ødu/ and /ødɑ/ show a falling direction; F2 onglide transitions of /udy/ and /udø/ show a rising direction; F2 offglide transitions are falling in /ødu/ and /ødɑ/, rising and flat in /udy/ and /udø/. Similarly, F3 transitions of these sequences also show some variations. But a tendency of F2 and F3 deflections being above and F1 below the vowel-to-vowel transitions can be observed. Figure 9 depicts the underlying transition contours (dotted curves) of the four sequences calculated as the fifth order polynomial function that produces the "minimum jerk" (Hogan, 1984). The calculations were based on the steady-state and formant onset frequencies provided by Öhman (1966). Not all first three formant contours

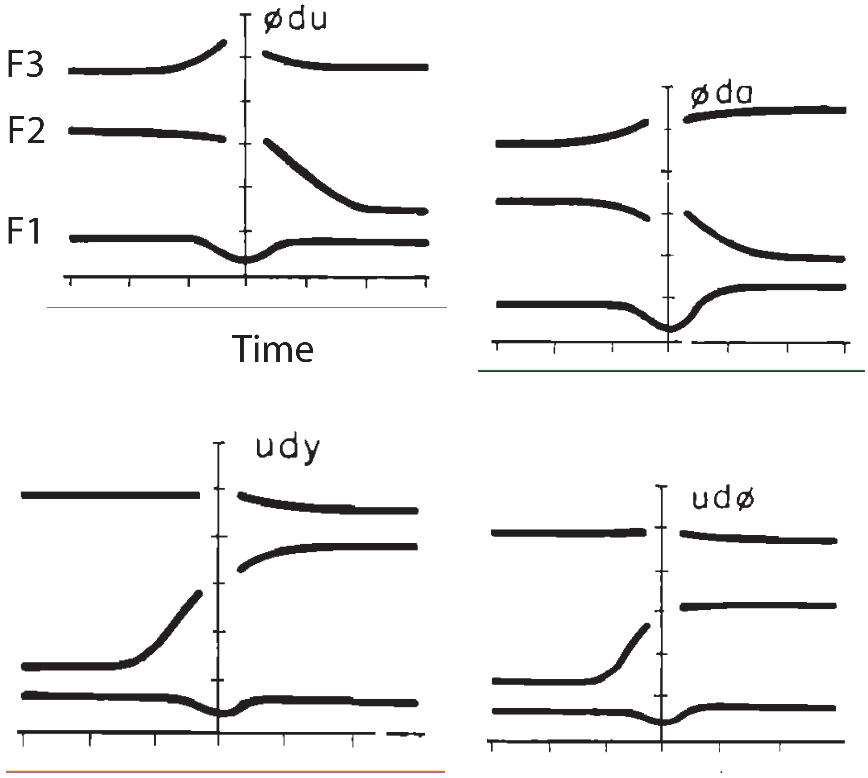


Figure 8: *Formant transition patterns of $/VdV/$ utterances in Öhman, 1966. (Adapted from Figure 8 in Öhman, 1966, permission acquired from the publisher.)*

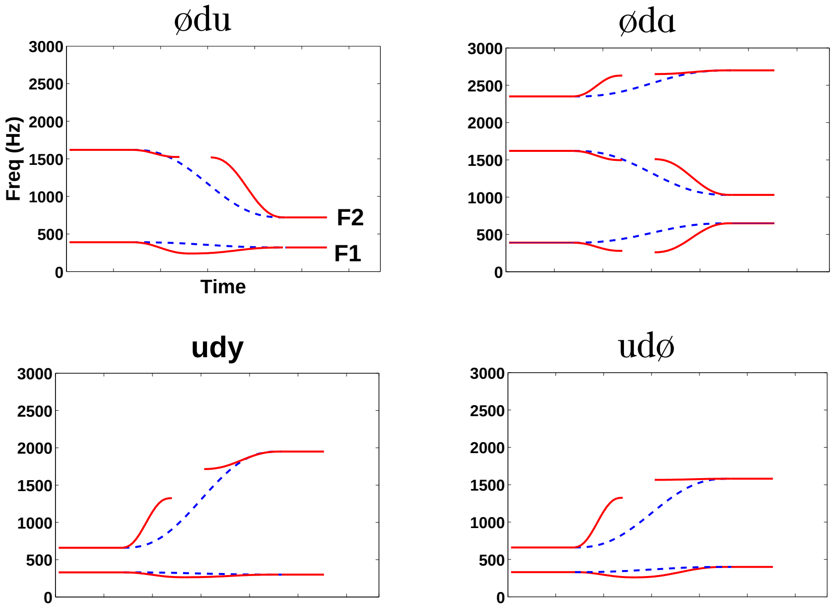


Figure 9: *Formant deflection patterns derived from Öhman, 1966. The underlying vowel-to-vowel transitions were calculated as the fifth order polynomial function that produces a "minimum jerk" contour (Hogan, 1984). Not all first three formant transitions are reproduced due to missing data.*

are plotted due to some missing data. Nonetheless, the visible formant deflection patterns agree with the hypothesized pattern in large part. The onglide F2 deflection in /ødɑ/ is slightly below the underlying vowel-to-vowel transition. Such fluctuation could be due to the signal processing technique.

The observation of the formant deflection with reference to the underlying vowel transitions can also be explained, to some degree, by the quantal theory of speech developed by Stevens (1972, 1989). His work on the relation between the vocal tract area function perturbation and formant frequency shift showed that the acoustic patterns change from one state to another as the vocal tract structure is manipulated at evenly spaced steps along a certain continuum. Within each acoustic state, the general shape of the acoustic representation remains the same and is associated with a phoneme. Across different acoustic states, the acoustic change is quantal in a sense that the associated phoneme changes. For example, Stevens (1989) demonstrated that when the place of constriction is varied along the vocal tract length, there are regions of acoustic stability (quantal regions) where particular formant frequencies approach their minimum or maximum values and are insensitive to the changes of place of constriction. These quantal regions are associated with the place of articulation for vowels.

Some evidence of the quantal relation between the area perturbations and the resulting formant shifts and perceptions has been obtained using this model (Li & Story, 2003). A series of stop-vowel syllables that differed only in the place of constriction was constructed using three vowel contexts. The place of constriction was varied at evenly spaced locations along the vocal tract length from the lips to about 8 cm above the glottis (slightly behind the soft palate). It was predicted that the continuum of place of constriction would result in the discontinuity of acoustic representation and categorical perception of the constriction continuum. The predicted perception of each synthesized syllable was based on its formant transition patterns. In the perceptual tests, 10 native English speakers served as listeners. Each synthesized syllable was presented 10 times and all syllables were presented in a random order. The listeners had to identify the consonant by selecting /b/, /d/, or /g/. The results showed that the perception of the constriction continuum was categorical, in other words, the quantal acoustic changes resulted in perceptual phonemic boundaries and identification was greater than 70% accurate. Stimuli at boundaries were perceived as one or the other phoneme with roughly equal frequency.

Figure 10 illustrates the nomograms (three upper panels) of the synthesized syllables and the results of the identification tests (three lower panels) for the three vowel contexts. The horizontal axis is the distance from the glottis in centimeters. In the upper panels, the horizontal broken lines are the steady state vowel formants of the stop-vowel syllables. The solid lines are the onset frequencies at the release of constrictions at the locations indicated by the horizontal axis. For example, on the nomogram for the vowel /æ/ (upper left panel), a constriction located at about 15 cm above the glottis results in an F1 onset frequency of 339 Hz, F2 onset of 1537 Hz, and F3 onset of 2646 Hz. The first three steady state formants for the vowel /æ/ are indicated by the horizontal dotted lines as 687 Hz, 1564 Hz, and 2420 Hz respectively. Therefore, the F1 transition shows an increase direction from 339 Hz to 687 Hz, F2 transition also shows an increase direction from 1537 Hz to 1564 Hz, and F3 transition shows a decrease direction from from 2646 Hz to 2420 Hz. Based on the direction patterns of the transitions from the onset to the steady-state vowel formants, each continuum is divided by the vertical broken lines to indicate the predicted phoneme boundaries. The three lower panels show the identification functions of the three continua. The vertical lines indicate the perceptual boundaries of the continua.

There are several observations that can be made about the identification functions. The categorical identification of stop consonants on the /æ/ continuum is more prominent than on the /ɑ/ continuum. For the /æ/ vowel, the function is sharper and the percentage response within a category is higher. The /d/ category is never fully developed in the /i/ continuum (even though there is a small peak at about 15.87 cm from the glottis), but it reaches relatively high peaks in the other two vowel environments. We speculate that

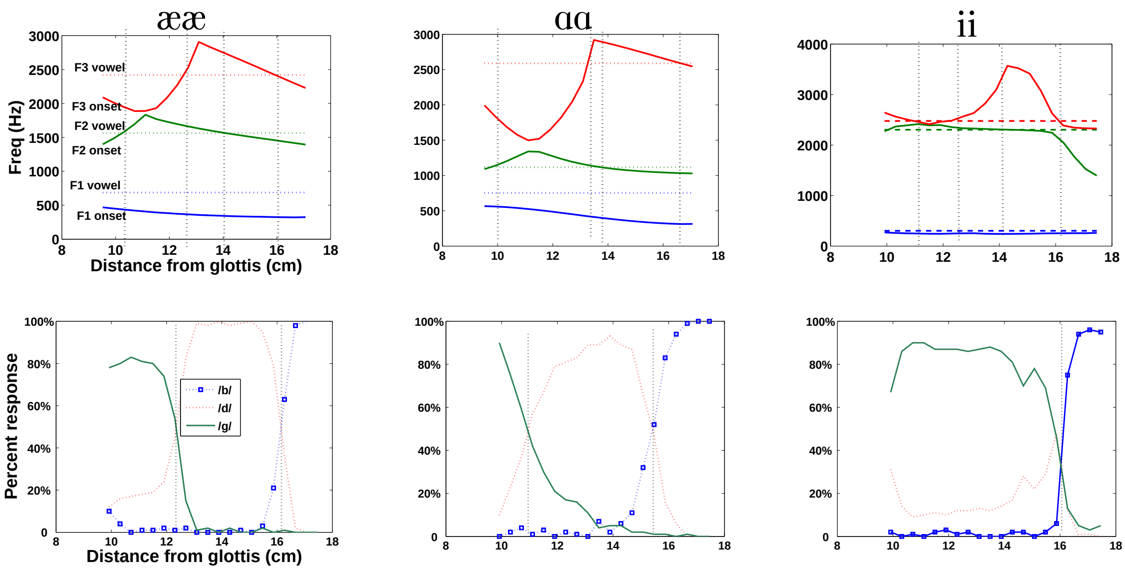


Figure 10: *Nomograms (upper panels) lined up with the identification functions (lower panels). The three upper panels show the first three onset frequencies (solid lines) as a function of place of constriction in the three vowel contexts respectively. Based on the patterns of the transitions, each continuum is divided by the broken vertical lines to indicate the predicted phoneme boundaries. The three lower panels show the identification functions of the three continua. The vertical lines are perceptual boundaries of the continua. Modified from Li and Story, 2003.*

the difference in the prominence of the categories might be caused by the arbitrary set of parameters used to create the constriction geometry. In other words, the constrictions in all the syllables were created with the same set of geometric and temporal features, which might not be the optimal set for all the vowel environments. A different set of parameters might create different acoustic and perceptual effects. Informal listening of syllables created at about 15 to 16 cm sounded somewhat like the voiced alveolar fricative /ð/, a consonant that shares the same place of articulation with the alveolar stop /d/.

The alveolar ridge is a relatively sensitive region of the vocal tract where many sounds are produced, such as /s/, /z/, /l/, /n/, /t/, and $/\theta/$, in addition to /d/ and /ð/. Often times, the place and manner of articulation for the alveolar stop /d/ is similar to those of the voiced alveolar fricatives /z/ and /ð/. Both groups of sounds require a narrow constriction in the vicinity of the alveolar ridge. The constriction is adjusted in size from a small opening for the fricatives to a complete closure for the stop. Despite the similarity in the articulatory manners of the two groups of sounds, the perceptual outcomes are distinctively different. It might be the slight changes in the production dynamics (location and geometry of the consonantal constriction) that lead to the quantal changes in acoustic features and perceptual outcomes between the alveolar stop /d/ and alveolar fricative /z/ and /ð/.

OVERVIEW OF THE RESEARCH PLAN

The project was completed in three phases. The first phase was to record spoken $/V_1CV_2/$ and $/V_1V_2/$ utterances and to extract the time-varying acoustic features. These features were used for synthesis (phase two) and comparison purposes. The consonants in the recorded $/V_1CV_2/$ utterances were the voiced alveolar stop /d/ and voiced alveolar fricatives /z/ and /ð/. The voiced alveolar fricatives were included because of the similar production dynamics and acoustic features they shared with the alveolar stop (Stevens, 1998). The second phase was to complete the model-based computation to create $/V_1CV_2/$ audio samples with different spatial and temporal consonantal constrictions. The temporal parameters were set based on the time-varying transition features extracted in phase one. The spatial parameters (location and geometry of the consonant constrictions) were set to vary at even steps within appropriate ranges. The synthesized speech samples were then grouped based on their formant deflection directions with reference to the underlying vowel formant transitions. A database was created to establish a mapping between the parameter settings and their resulting formant deflection patterns. The final phase was to test the identification of the synthesized consonant by conducting perceptual tests. Due to the large number of audio samples created in phase two, a pool was generated that included 90 samples. These samples represented the hypothesized formant deflection pattern (Figure 7) in all nine vowel contexts generated by different parameter settings in phase two. The listeners responded to the identity of the consonant by choosing from six choices that included the alveolar stop /d/, the fricatives /z/ and /ð/, /v/, and the other two voiced stop consonants /b/ and /g/. The percentage response to each of the presented stimuli was calculated. The acoustic features of the recorded and synthesized syllables with same consonant identities were compared.

SIGNIFICANCE

The long term goal of the proposed research was to understand the time-varying relation among the place of stop consonantal constriction, the formant transition patterns, and their perceptual outcomes for $/V_1CV_2/$ utterances. Knowledge about the relation advances fundamental understanding of human speech production and research and clinical measurements of articulation. Because the model used in this research is based on the spatial characteristics of vocal tract resonance, it characterizes speech signals in terms of their acoustic interactions between the pressure release and the mechanical compliance of the vocal tract configuration. Therefore, the knowledge helps both researchers and clinicians understand normal speech production.

In addition, the data collected in this study contribute to the design of a speech synthesis system based on articulatory parameters since the data describe the speech production mechanism itself in terms of the time-varying vocal tract dimensions as well as the acoustic result of such variations. Specifically, the data provide useful information in the parameter control, synthesis of different types of speaker, and choice of simplifying assumptions when an articulatory synthesizer is designed.

Although speculative at this point, the results of this study may eventually contribute to the design of a speech recognition system based on articulatory space. There has been growing interest in the automatic speech recognition system using articulatory parameters, either as a supplement to or substitute for acoustically based input parameters. Many popular techniques of speech processing for the purpose of recognition, such as hidden Markov modeling (HMM), rely heavily on the use of statistics to model the variability of speech, such as the coarticulation effect. These techniques have little relation to the actual mechanism of speech production or perception. It is desirable to incorporate information provided by an articulatory feature representation such as the place of articulation, length variability of the vocal tract (rounding vs non-rounding), and timing of articulatory gestures into the speech recognition system. Data resulted from this study provide valuable information about the invariant factors in speech perception and recognition.

EXPERIMENT 1

METHOD

This experiment was completed in four steps. The first step was to record spoken $/V_1\mathrm{d}V_2/$ and V_1V_2 sequences and to extract the time-varying formant contours. The second step was to set the model parameters based on the acoustic features extracted in step one. The third step was to synthesize the V_1CV_2 syllables. The fourth step was to generate a sample pool and conduct perceptual tests.

Recording the utterances

All possible combinations of $/V_1CV_2/$ syllables between the three vowels /i, æ, ɑ/ and the alveolar consonant /d/ have been recorded (9 utterances). Their corresponding $/V_1V_2/$ syllables have also been recorded (9 utterances). The 40-year-old male speaker was the same speaker whose neutral area function and the spatial eigenmodes for each vowel had already been obtained in previous studies for different purposes (Story, Titze, & Hoffman, 1996; Story & Titze, 1998). A hearing screening at 25 dBHL for the octave frequencies 500 Hz, 1 kHz, 2 kHz, and 4 kHz was conducted by the investigator to confirm normal hearing at the time of recording (ASHA, 1997).

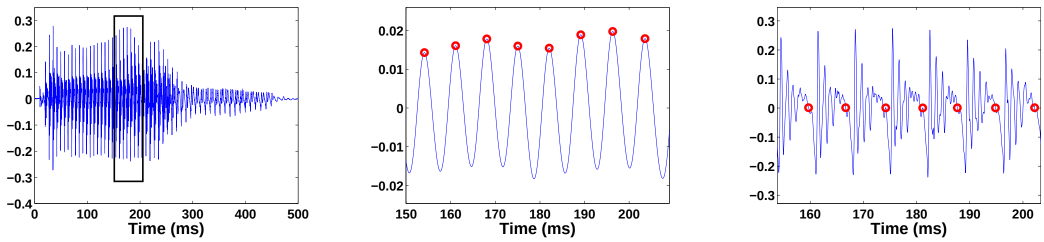
Each $/V_1\mathrm{d}V_2/$ or $/V_1V_2/$ sequence was embedded in a carrier sentence: “Say the word — baby” to minimize the speaking rate effect on the durations of the target sequences. Their non-final position was to prevent the falling fundamental frequency often present at the end of a sentence. The stressed position of the target sequences was to reduce the coarticulation effect between V_1 and the preceding final consonant /d/ in “word”. Each sentence was repeated consecutively three times at the speaker’s comfortable conversational rate and loudness level. The recording was performed in a sound-attenuated booth using an AKG SE300B microphone. The microphone was placed approximately 20 cm from the lips at a 45 degree angle. The speech signals were recorded directly as digital files on a computer using the Computerized Speech Lab (CSL) analysis system (model 4500) by KayPENTAX, with a sampling rate of 44,100 Hz.

In addition to the $/V_1 \mathrm{d} V_2/$ sequences, $/V_1CV_2/$ sequences between consonants /b, g, v, z, ð/ and the vowels of interest were also recorded so that the formant deflection patterns of these sequences could be extracted and compared to those of the synthesized $/V_1CV_2/$ sequences that might yield high perceptual rates of the same consonants.

Extracting the acoustic features

Pitch synchronous analysis (where the analysis segment duration was equal to one pitch period) was used to extract the first three formant contours of the recorded V_1V_2 and V_1CV_2 utterances. The utterances before and after the target sequences in the carrier sentence were discarded. Pitch synchronous analysis was chosen over the pitch asynchronous technique, where the segment duration would be fixed because of the following reasons. First, the synchronous technique closely tracks the transient dynamics of the speech signal so that the last glottal cycle before the vocal tract is fully occluded can be found and analyzed. Second, the segmentation locations in the synchronous technique are flexible and can be positioned at zero-crossing points to avoid sharp truncations that corrupt the parameter estimation. Such corruptions can be significant when the frame size is small (two or three periods in one frame) and the total percentage of the overall error is increased (Rabiner, Atal, & Sambur, 1977; Krishnamurthy, 1984). In addition, the open phase of the glottal cycle can be positioned at the center of any preprocessing window so that the formants are more distinct. The only shortcoming of the pitch synchronous analysis is that it results in a variable-rate coder where the time resolution of the extracted formant contours is uneven. However, this was easily overcome by a re-sampling process at a desired sampling frequency based on a cubic spline technique.

The first step in the process was to extract the time-varying period from the recorded waveform. This was performed by low-pass filtering the speech signal to eliminate the high frequency components so that the fundamental period was distinct in the waveform. The cutoff frequency was set slightly above the estimated fundamental frequency of the speaker to make sure that the fundamental periods were intact. The pitch periods were then detected as the piecewise local maximum values of the low pass filtered waveform. Within an estimated pitch period, every five consecutive values were compared. If the middle value was greater than both the two preceding and the two following ones, a peak was detected. If more than one peak was detected in a single pitch period, the highest peak was selected. The index numbers and the values of the peaks were stored in an array. The index numbers were then transfered to the original utterance waveform and the closest zero-crossing points were found to mark the boundaries between periods. Setting the zero-crossing points to be the boundaries avoided sharp truncations that would otherwise corrupt the formant estimation later. Matlab (release 14 by the Mathworks) rou-

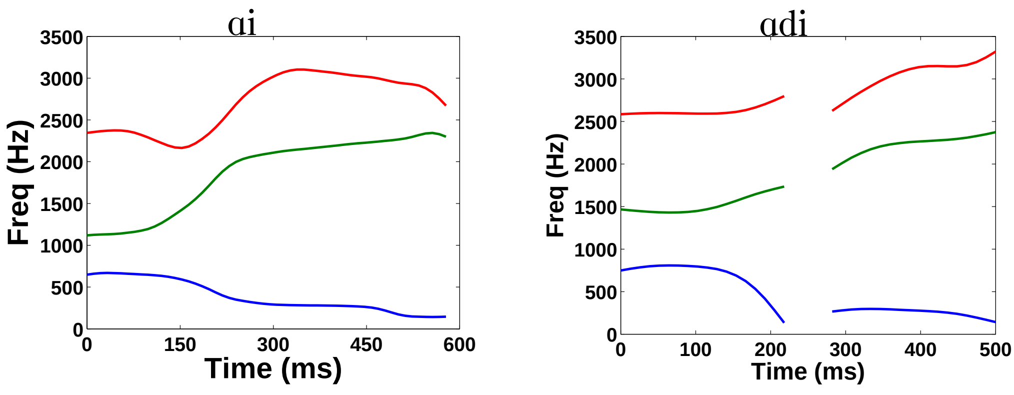


(a) (b) (c)

Figure 11: *Steps of extracting the formants.* **a.** *Waveform of the naturally spoken sequence /ɑi/. The portion within the rectangular window is magnified in panel c.* **b.** *Low pass filtered waveform and the detected peaks indicated by small circles.* **c.** *The period boundaries are marked by the closest zero-crossing points to the peaks shown in panel b.*

tines were written to automate this process. Figure 11 shows an example of the process. Panel (a) is the waveform of the utterance /ɑi/ spoken by the male speaker. The portion of waveform included in the rectangular window is magnified in panels (b) and (c). Panel (b) shows the low-pass filtered waveform with a cutoff frequency of 200 Hz (estimated fundamental frequency was 150 Hz) and detected peaks are indicated by the circles. Panel (c) shows the zero-crossings (indicated by the circles) that mark the period boundaries on the original utterance waveform.

Once the waveform of the signal was segmented based on the pitch periods, the first three formants of each period segment were extracted using LPC. The order of the algorithm was chosen to be about 50, given a sampling frequency of 44,100 Hz. The formants extracted from each segment were then connected together to generate the time-varying formant contours over the entire $/VV/$ or $/VCV/$ utterance. Next, each formant contour was smoothed by a median filter so that the spurious points on the contours due to signal



(a) (b)

Figure 12: *Formant contours of the spoken utterances /ɑi/ (a) and /ɑdi/ (b). The discontinuity of the contour in the right panel indicates the moment when the vocal tract is fully occluded for the production of the stop consonant /d/.*

processing techniques were eliminated. Figure 12 shows the formant contours extracted from the recorded spoken utterances /ɑi/ and /ɑdi/. The y-axis is the frequency in Hertz. The x-axis is time in milliseconds. The gaps on the /ɑdi/ formant contours were due to the deletion of data points when LPC algorithm could not extract the formants of a closed vocal tract configuration. The extracted $/V_1V_2/$ and $/V_1CV_2/$ formant contours were used as a guide to determine the model parameters.

The model

The structure of the vocal tract area function model used in this study (Story, 2005) is based on the perspective that continuous speech is a string of vowel-to-vowel transitions perturbed by superimposed consonant constrictions (Öhman, 1966, 1967). The model consists of four tiers to realize the vowel-to-vowel transitions and possible consonant constrictions at various places along the vocal tract. Each tier is a shaping function that controls different dimensions of the vocal tract to produce perturbations required for a particular vowel or consonant. Such a perturbation might be the result of a constriction at a certain location along the vocal tract length, rounding of the lips, or lengthening of the tract. Each perturbation realized within a tier can be activated or deactivated in a sequential or overlapping fashion with other perturbations in other tiers since the parameters are specified as functions of time.

The four tiers of the model are shown in Figure 13 (Story, 2005). The left column shows the base structural components of each tier. The middle two columns indicate the input parameters at each tier. The input parameters in Tiers I and II are functions of time (t) and distance from the glottis along the length of the vocal tract. The distance is expressed as the section index (i), which extends from just above the glottis at section 1 to section 44 at the lips. The area function is represented as the cross sectional areas of 44 equal length piecewise cylindrical tube sections. In this study, the total vocal tract length was fixed at 17.46 cm, with a length of 0.396825 cm for each section. Tiers III and IV control the vocal tract length variation and nasalization when needed. They were not used in this study and were set to a deactivated status.

In Tier I, the vowel substrate is defined as the sum of the neutral area vector and several characteristic basis area vectors of a desired vowel. The neutral area vector is the mean area vector of ten static English vowels produced by a particular speaker (/i, ɪ, ɛ, æ, ʌ, ɑ, ɔ, u, o, u/). It is similar to the area vector of schwa (/ə/) in configuration. It also shows similar resonances to those of a uniform tube, i.e., formants at about 500 Hz, 1000 Hz,

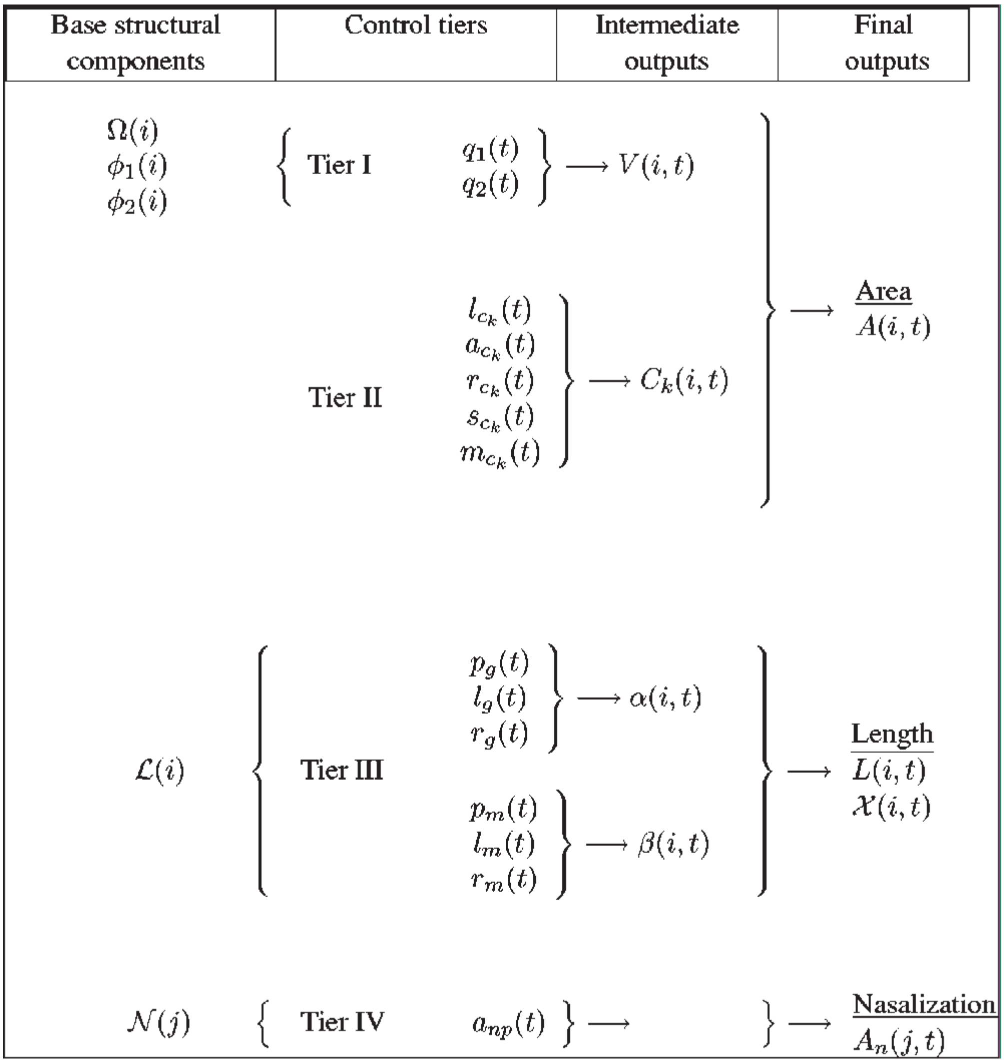


Figure 13: *Structural diagram of the four-tier area function model (taken from Story, 2005, p.3233). Tier I defines and controls the vowel substrate, onto which a consonant constriction generated at Tier II can be superimposed. Tiers III and IV control the vocal tract length variation and nasalization.*

and 1500 Hz. Such similarities suggest that any vowel area vector can be decomposed as the sum of a set of normalized orthogonal basis area vectors and the mean area vector. The basis area vectors are derived from the covariance matrix of the 10 vowel vectors by singular value decomposition (SVD). They are called the empirical orthogonal eigenmodes of a particular vowel and can be considered as the “building blocks” (p228, Story & Titze, 1998) to retrieve the original area vectors. The basis area vectors are a few main factors that account for most of the variance of the original area vector around the mean area vector. The significance of such a decomposition is that each orthogonal mode, or each basis area vector can be associated with a vocal tract perturbation gesture superimposed on the neutral configuration to produce an area vector of a desired vowel, such as the effect of the dropping of the jaw superimposed on the neutral area vector for the area vector of vowel /ɑ/. The acoustic result of such a gesture can be computed and considered as an acoustic perturbation to the resonances of the neutral vowel shape. In other words, the association provides a mapping between the acoustic domain (such as a set of formants) of a desired vowel and a series of empirical orthogonal eigenmodes, which in turn, can be superimposed on a pre-acquired neutral area function to retrieve the desired vowel area vector. Tier I of the model is constructed based on this mapping relationship, as shown in Equation (2).

$$V(i,t) = \frac{\pi}{4}[\Omega(i) + q_1(t)\phi_1(i) + q_2(t)\phi_2(i)]^2;\ i = [1, N_{area}] \tag{2}$$

In the equation, $\Omega(i)$ is the section diameter vector of the mean area function at each tubular section i; $\phi_1(i)$ and $\phi_2(i)$ are the two most significant normalized eigenmodes for a desired vowel; $q_1(t)$ and $q_2(t)$ are the time-varying coefficients of the modes respectively. They represent the time evolution of the vocal tract shape from the mean shape to the desired vowel shape. The areas are computed by two steps. First, $\Omega(i)$ and the multiplication between the coefficients and modes are summed. Second, the sum is squared and scaled by $\frac{\pi}{4}$. $\phi_1(i)$, $\phi_2(i)$, and their corresponding coefficients $q_1(t)$ and $q_2(t)$ are acquired by the pre-established speaker-specific mapping between the measured formants from the acoustic signal and the empirical orthogonal eigenmodes of a desired vowel. The mapping is illustrated in Figure 14 (adapted from Figure 11 in Story, 2005). For each pair of F1 and F2 values in the lower panel, a pair of q_1 and q_2 values can be found in the upper panel, and vice versa.

In Tier II, a consonant perturbation is generated and superimposed on the vowel substrate by the multiplication between the two tiers element-by-element. The geometry of the perturbation is defined as the mid-sagittal profile of the constriction controlled by input parameters that describe the location, area, range, and skewness of the constriction. The profile is expressed by two Gaussian functions, one for the upstream (glottal side) of the constriction and the other one for the downstream (lip side), as shown in Equations (3) and (4).

$$C(i,t) = 1 - m_c(t)d_c(t)e^{ln(16)[(i-l_c)/r_{cb}(t)]^2}; \qquad for \quad i < l_c(t); \tag{3}$$

$$C(i,t) = 1 - m_c(t)d_c(t)e^{ln(16)[(i-l_c)/r_{cf}(t)]^2}; \qquad for \quad i > l_c(t); \tag{4}$$

In the equations, the location of the constriction l_c is defined as the distance between the glottis and the center of the constriction, indicated by the number of sections; d_c is the degree of the constriction, and is defined as the ratio between the desired cross-

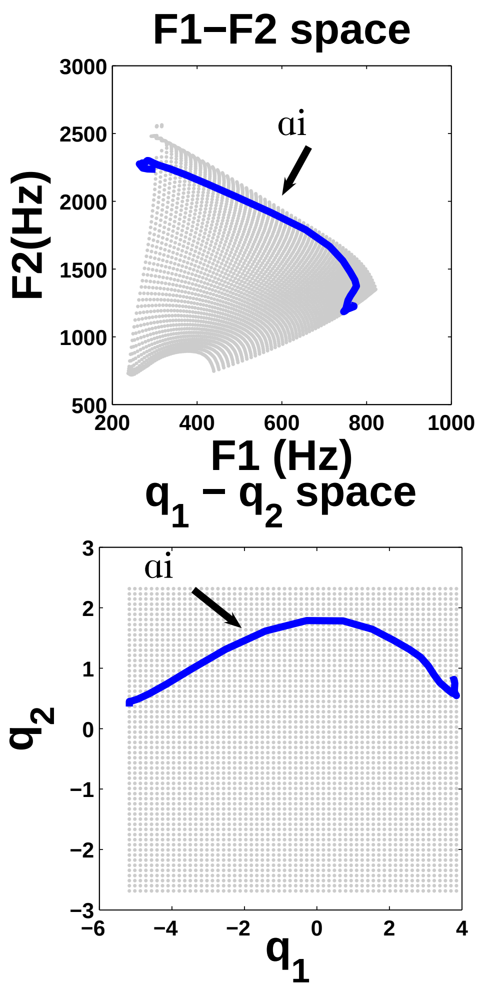


Figure 14: *Mapping between formant space (F1-F2, upper panel) and mode coefficient space (q_1-q_2, lower panel). Each dot in the upper panel indicating a pair of F1-F2 values has a corresponding dot in the lower panel indicating a pair of q_1-q_2 values. The thick curve in the upper panel shows measured F1 and F2 contours in a recorded utterance /ɑi/. The curve in the lower panel is the corresponding coefficient trajectory. Adapted from Figure 10 in Story, 2005*

sectional area $a_c(t)$ at the point of maximum constriction, and the area of the vowel substrate $V(l_c(t), t))$ at $l_c(t)$. The computation of $d_c(t)$ is shown as Equation (5).

$$d_c(t) = 1 - \frac{a_c(t)}{V(l_c(t), t))} \tag{5}$$

For stop consonants /b/, /d/, and /g/, d_c is set at 1, indicating 100% full occlusions ($a_c(t) = 0$). For fricatives, the cross-sectional area at maximum constriction point may be as small as 0.05 cm^2. Therefore, $d_c(t)$ may have a value smaller than but very close to 1.

The parameters $r_{cb}(t)$ and $r_{cf}(t)$ define the upstream constriction range and downstream constriction range seen on the mid-sagittal profile of the vocal tract. The sum of the two is the range of the constriction, defined as the distance between points along the vocal tract where the consonant function is 50% constricted (half way to the maximal constriction point). When $r_{cb}(t)$ equals $r_{cf}(t)$, the upstream has the same range as that of the downstream, indicating a symmetrical constriction shape. When they are not equal, the constriction is skewed to the side with a shorter range. The skewness of the constriction is defined as the ratio between the upstream range $r_{cb}(t)$ to the downstream range $r_{cf}(t)$. A skewed constriction shape is illustrated in Figure 15. The horizontal axis indicates the distance from the glottis by section numbers. The vertical axis indicates the scaling values for the area at a location.

The parameter $m_c(t)$ in Equations (3) and (4) serves as a time function to activate or deactivate the consonant constriction and control the magnitude of the constriction when it is activated. A value of 0 gives the constriction function a value of 1 over the entire vocal tract (see Equations (3) and (4)), indicating a non-constricted vowel tract for a vowel

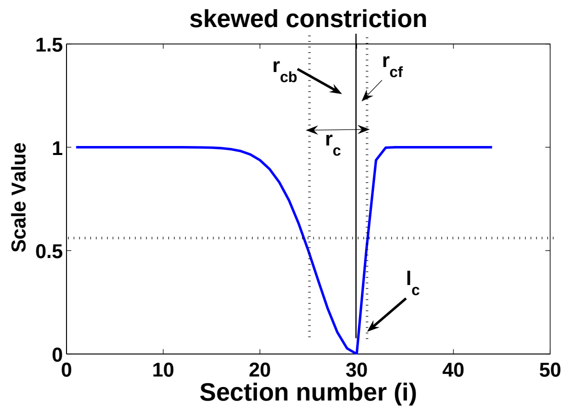


Figure 15: *A skewed full constriction generated at section 30:* $l_c = 30$, $r_c = 8$, $m_c = 1$. *Modified after Figure 4 in Story, 2005.*

articulation. When $m_c(t)$ takes a non-zero value, a constriction is superimposed over the vocal tract range specified by the sum of $r_{cb}(t)$ and $r_{cf}(t)$. The larger the value of $m_c(t)$, the tighter the constriction is. When $m_c(t)$ is 1, the maximum constriction area set by $a_c(t)$ in Equation (5) is fully realized, such as the example illustrated Figure 15. The value of $m_c(t)$ is also allowed to be slightly smaller or greater than 1 (e.g., 0.9, 0.8, or 1.1, 1.2). When it is smaller than 1, the cross-sectional area assigned by $a_c(t)$ is not fully realized. When $m_c(t)$ is greater than 1, the contact area of a full occlusion forms a flat surface that occupies more than one section, such as the one illustrated in Figure 16. To simulate the gradual articulator movements in connected speech, $m_c(t)$ varies smoothly between 0 and its maximum value and then back to 0 to activate or deactivate the consonant constriction at appropriate time points. The smoothness of $m_c(t)$ value is governed by a fifth order polynomial function that produces a “minimum jerk” movement (Story, 2005; Hogan, 1984).

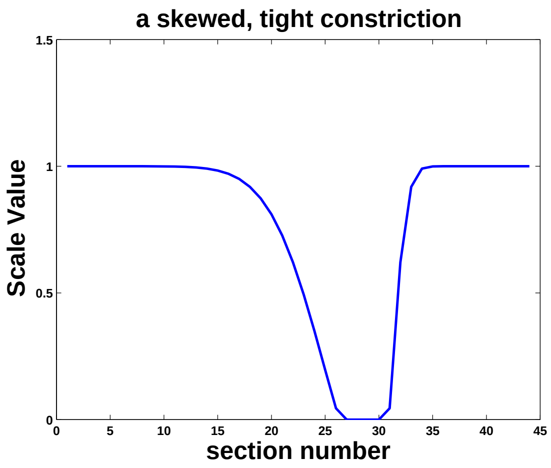


Figure 16: *A skewed full constriction with a flat surface at the contact area, generated at section 30,* $m_c = 1.2$*.*

The use of Gaussian-based functions and corresponding input parameters to approximate the consonant constriction geometry obtained from MRI experiments has been verified by Story (2005). The correlation coefficient r and the rms values between the measured and model generated consonant shapes were calculated for 10 consonants (/p, t, k, m, n, ŋ, f, ɵ, s, ʃ/), indicating reasonably good matches for most consonants, especially the stops ($r > 0.88$ and $rms < 0.5cm^2$ for the stops). The model parameters that govern the shape of the constriction can be related to the classical description of the articulation of consonants in terms of place and manner of articulation.

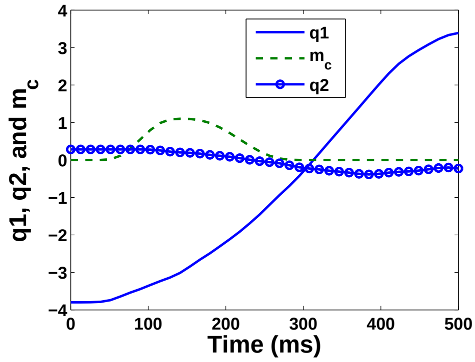
The vowel substrate $V(i,t)$ generated in Tier I and the consonantal constriction $C(i,t)$ generated in Tier II are multiplied element-by-element to synthesize the time-dependent area perturbations superimposed on the vowel substrate, as indicated in Equation (6):

$$A(i,t) = V(i,t)C(i,t), \quad i = [1, N_{area}]; \tag{6}$$

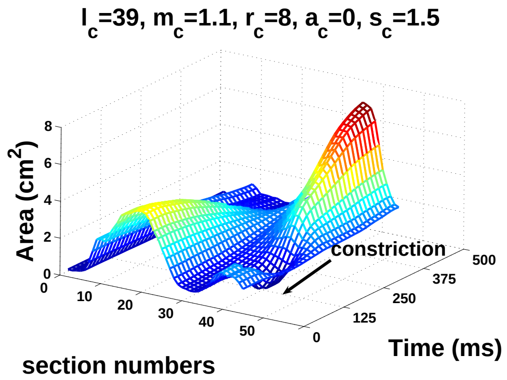
In Tier III, lengthening or shortening of various regions of the vocal tract can be realized if necessary. It is well known that the tract length is typically longer for rounded vowels such as /u/ and shorter for spread vowels such as /i/. As a simplification for this study, however, the vocal tract length was held at a constant value of 17.46 cm. Whereas this might degrade the realism of the model's output somewhat, the absence of tract lengthening was not expected to change the formant transition pattern in a significant way. Furthermore, the vowels used in the V_1 and V_2 positions were /i/, /æ/, and /ɑ/. The length differences among these vowels was much less than they would be if rounded vowels had been included (Story, Titze, & Hoffman, 1996; 1998). Tier IV controls the coupling of the nasal cavity, which was not involved in this study.

To synthesize the vocal tract area changes for a $/V_1CV_2/$ sequence, a consonant constriction generated in Tier II is superimposed on a series of time-varying area functions that approximate the $/V_1V_2/$ transitions generated in Tier I. The time course of the $/V_1V_2/$ transition is governed by the time-varying coefficients $q_1(t)$ and $q_2(t)$ that initially represent the coefficients of the first two modes for vowel V_1 and change over time to those of V_2. The initiation and cease of the consonant constriction are controlled by the parameter $m_c(t)$. In other words, $m_c(t)$ governs the amounts of temporal overlap between the consonant and vowel gestures. Other parameters that indicate the location (l_c), range (r_c), skewness (s_c), maximum magnitude (m_c), and the minimum area (a_c) of the constriction are kept constant over the duration of the $/V_1CV_2/$ synthesis.

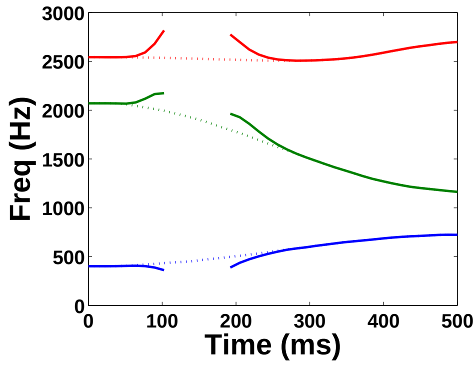
Figure 17 illustrates the simulation of the sequence /idɑ/ with a duration of 500 ms. The time-varying parameters $q_1(t)$ and $q_2(t)$ in panel (a) are the eigenmode coefficients that



(a)



(b)



(c)

Figure 17: *The simulation of the sequence /idɑ/.* **(a):** *The time-varying mode coefficients* q_1, q_2, *and* m_c *for the synthesis of /idɑ/.* **(b):** *Time-varying area functions. The model parameters for the constriction are indicated on the top.* **(c):** *Time varying formants of synthesized /iɑ/ (dotted lines) and /idɑ/ (solid lines).*

approximate the VV transitions from vowel /i/ to vowel /ɑ/. Also plotted in this panel is the time interpolation term $m_c(t)$ for the consonant constriction. It is set to 0 initially when the vocal tract is engaged in the vowel production of /i/. It then starts to rise at about 60 ms and gradually reaches its full value of 1.1 at 140 ms to indicate a full and tight occlusion. It falls back to 0 again at 280 ms when the constriction is released. The rise and fall of the constriction at different times will create significant acoustic changes, and potentially, have perceptual effects. The area function series over the duration of the sequence is shown in panel (b) of Figure 17. All the constant parameters to set the location and geometry of an alveolar constriction required for /d/ are indicated at the top. The constriction that occurs at section 39 (indicated by the arrow) is about 1.98 cm from the lips (17.46 cm - 0.3968 cm $\times$ 39 sections = 1.98 cm). The formant variations of the sequence corresponding to the area function variations are shown in panel (c) as the solid lines. The gap in the middle of the contours indicates the moment when the vocal tract is fully occluded. The underlying dotted lines are the /iɑ/ formant transitions. The formant contours of the synthesized VV and VCV sequences are determined from the frequency response of each area function during the simulation. It can be seen that the consonant constriction perturbs all three formants in different directions, with falling F1, and rising F2 and F3, comparing to the underlying vowel-to-vowel formant transitions.

The set of parameters used to create the constriction in the example sequence /idɑ/ shown in Figure 17 were chosen based on the experience with the model. A different set of parameters will create a constriction with different spatial (location and geometry) and temporal characteristics. Consequently, formant transitions between the consonant and the preceding/following vowels may show some differences in the directions and durations. These changes in the transition characteristics may result in perceptual changes in terms of the phonemic identity.

Determining the model parameter values

To synthesize $/V_1CV_2/$ sequences, the following model parameters were controlled:

1). The time-varying mode coefficients $q_1(t)$ *and* $q_2(t)$

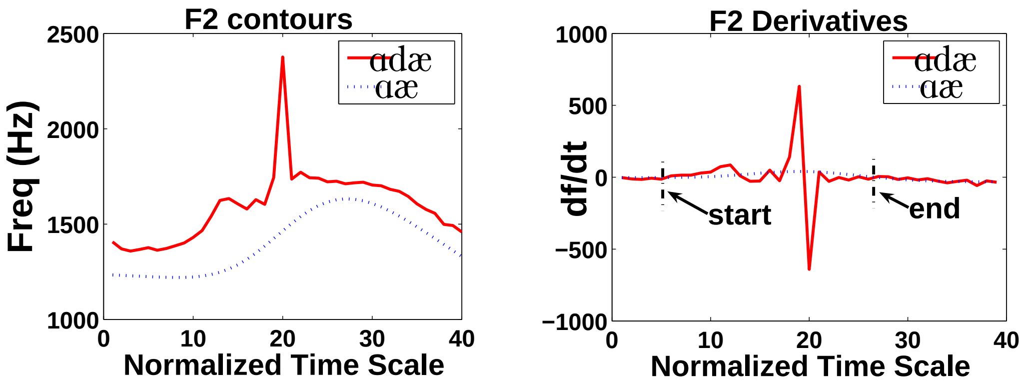
The mode coefficients that governed the vowel-to-vowel transitions were acquired from the pre-established one-to-one mapping between F1-F2 values and mode coefficients (see Figure 14).

2). The time course of $m_c(t)$

The time course of $m_c(t)$ was defined by 3 time instants, i.e., the onset and release times of the constriction and the moment the constriction reaches its full magnitude. The three time points were then interpolated by a fifth order polynomial function to produce smooth transitions between movements (Story, 2005; Hogan, 1984). The three points were decided based on the extracted F2 contours of the spoken $/V_1\mathrm{d}V_2/$ and $/V_1V_2/$ utterances. F2 was chosen over other formants because the rising and falling of F2 is the basic F-pattern that reflects the tongue constriction location (Fant, 1980). In addition, the F2 transition from vowel midpoint to vowel offset has been shown to be a reliable predictor of temporal overlap between consonant and vowel gestures (Weismer, Kent, Hodge, & Martin, 1988; Kent, Kent, Weismer, Martin, Sufit, Brooks, & Rosenbek 1989; Tjaden & Weismer, 1998).

The following steps were used to determine the three time points. First, F2 contours of a recorded $/V_1\mathrm{d}V_2/$ utterance and its corresponding V_1V_2 utterance were re-sampled to the same length of 40 (resulting in a sampling period of 12.5 ms) so that they could be compared on the same time scale. The lengths of the utterances were approximately 500 ms in most cases. The last few points of the re-sampled formant contours due to sharp truncations were manually corrected. Second, the derivatives of the two re-sampled F2 contours were calculated and plotted on the same panel to show the rates of changes. The instant when the derivative of F2 contour of $/V_1dV_2/$ started to deviate from or converge back to that of the corresponding $/V_1V_2/$ utterance marked the initiation and termination of the consonant gesture. The index numbers of the points of deviation and convergence on the derivative contours were picked manually. The moment when the constriction reached its full value was chosen to be the relative position on the $/V_1dV_2/$ F2 contour when the LPC algorithm began to produce anomalous values (the last glottal cycle before the vocal tract occlusion).

Figure 18 illustrates the steps discussed above for the utterance /ɑdæ/ and /ɑæ/. Panel (a) shows F2 contours extracted from the recorded /ɑdæ/ (solid line) and /ɑæ/ (dotted line) utterances, re-sampled to the same time length (40 sample points). The sharp peak at the middle of the /ɑdæ/ F2 contour indicates the time duration when the vocal tract was fully occluded and the LPC algorithm failed to estimate the model coefficients. The index number before the peak is 18, indicating the last glottal cycle before the full occlusion. Its proportional position relative to the whole utterance length of 40 is 18/40=0.45. Panel (b) shows the derivatives of the two F2 contours. The dashed vertical lines indicate the



(a) (b)

Figure 18: *Determining the time course of* $m_c(t)$. **(a):** *F2 contours of the spoken utterances /ɑdæ/ (solid line) and /ɑæ/ (dotted line). The peak of the contour in the /ɑdæ/ contour indicates the moment when the vocal tract is fully occluded for the production of the stop consonant /d/.* **(b):** *Derivatives of the F2 contours shown in the left panel. The short vertical lines mark the times when the consonant constriction is started and ended.*

points where the derivative of /ɑdæ/ contour started to deviate and converge back to that of the /ɑæ/ contour. The index numbers of the two points are 6 and 27, their proportional positions relative to the whole utterance length of 40 are 6/40=0.15 and 26/40=0.675. Therefore, for a synthesized /ɑCæ/ sequence of 500 ms long, the consonant constriction was activated at the instant of 75 ms (0.15 $\times$ 500 = 75 ms) and deactivated at the instant of 337.5 ms (0.675 $\times$ 500 = 377.5 ms). It reached its full occlusion at the instant of 225 ms (0.45 $\times$ 500 =225 ms).

3). The location (l_c), range (r_c), skewness (s_c), magnitude (maximum value of m_c), and the minimum area (a_c) of the constriction

These parameters were varied within estimated ranges, as shown in Table I. The values of the 5 parameters were kept constant during the simulation of one $/VCV/$ sequence. Across sequences, they were different. The maximum value of m_c that indicated the magnitude of the constriction were varied from 1.0 to 1.3 at an interval of 0.1. This gave four values for this parameter. The desired minimum area $a_c(t)$ at the center of the constriction had two values: 0 and $0.05\ cm^2$. The value $0.05\ cm^2$ was included to test if an absolute full occlusion of the vocal tract was necessary to yield a stop consonant perception with the model. The location of the constriction l_c varied from section 37 to 41 at one section intervals, which gave a distance range of 2.77 cm to 1.18 cm from the lips with 5 places of articulation. This range covered the range of place of articulation for the alveolar stop between 1.5 and 2 cm from the lips observed from real speech by Stevens (1998). The large range of locations was to guarantee that all possible constriction locations for the production of /d/ were included. The range of constriction $r_c(t)$ which is the sum of the upstream range and downstream range in section numbers had two values: 8 and 6, corresponding to 3.17 cm and 2.38 cm. These numbers were decided based on the comparison between measured MRI data on consonants and model generated area functions (Story, 2005). For each constriction range, three symmetry situations were generated: symmetrical, left skewed, and right skewed. The upstream constriction range $r_{cb}(t)$ and the downstream constriction range $r_{cf}(t)$ were adjusted (equal or one greater than the other) to create different symmetries. All possible combinations of the values of the five parameters were executed to create 240 consonant constrictions (4 magnitudes x 2 minimum areas x 5 locations x 2 ranges x 3 symmetries = 240) that differed in their location and geometry. Each of the 240 constrictions was superimposed on each of the 9 vowel-to-vowel transitions to generate a total of 2160 V_1CV_2 sequences (240 constrictions $\times$ 9 contexts = 2160 sequences).

Symbol	**Parameter**	**Varying values**
$m_c(t)$	Activation magnitude of the constriction superposition function	1.0, 1.1, 1.2 or 1.3. (4 values)
$a_c(t)$	Desired minimum cross-sectional area of the constriction	0 or 0.05 cm^2. (2 values)
$l_c(t)$	Location of the constriction specified as the section number	sections 37-41 at 1 section intervals. (5 values)
$r_c(t)$	Range of the constriction along the tract length	8 or 6 sections. (2 values)
$s_c(t)$	Skewing quotient of the constriction	anteriorly skewed (1.5), symmetrical (1), or posteriorly skewed (0.6). (3 values)

TABLE I: *Parameters controlled in the synthesis of the audio samples.*

Synthesizing the utterances

Synthesis of all the 2160 stimuli was automated by Matlab routines. To generate the actual acoustic signals, a glottal source model was coupled to a digital waveguide representation (Liljencrants, 1985; Story, 1995) of the vocal tract and trachea. The glottal source consisted of a parametric model of the time-varying glottal area based on Rosenberg (1970) (this was originally intended as a glottal flow model but has been modified here to serve as a glottal area generator). This model allows for control of the skewing quotient of each glottal area cycle as well the open quotient and fundamental frequency.

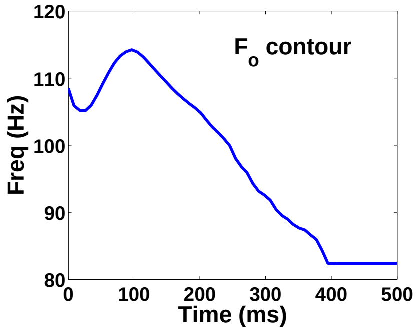


Figure 19: *The Fo contour used for all synthesized stimuli.*

The glottal flow generated was time synchrony with the wave propagation in the trachea and vocal tract as described in Story (1995) and more recently in Titze (2002). All synthesized stimuli had the same source parameters: the skewing quotient of the glottal area signal was 1.5 and the open quotient was 0.6. The fundamental frequency contour for all $/VCV/$'s was identical but approximated that of the recorded VCV sequences, as shown in Figure 19.

All the generated stimuli and their model parameters were saved as digital files on a computer.

The perceptual tests

The purpose of the perceptual experiments was to test if the hypothesized deflection pattern synthesized with different constriction parameter settings would yield /d/ identifications in different vowel contexts. The hypothesized pattern was shown in Figure 7, with F1 being deflected downward, F2 and F3 upward, relative to the underlying vowel-to-vowel transitions. In order to fulfill the task, the formant deflection patterns of the 2160 synthesized sequences were examined. The direction of an onglide formant deflection was determined in three steps. First, the vector difference between each time-dependent formant track (one each for F1, F2, and F3) of a $/VCV/$ sequence and that of the underlying $/VV/$ sequence was calculated for the period extending from the time instant of constriction onset (the first non-zero value of $m_c(t)$) to the instant when the constriction was 50% accomplished (the moment when $m_c(t)$) was increased to 50% of its maximum value). The next step was to sum the elements of each individual formant difference vector. The final step was to make the decision concerning the deflection direction. If the sum of the difference vector for a given formant showed a positive sign, the deflection direction was defined as upward. Otherwise, the deflection direction was defined as downward. The offglide direction was determined in a similar fashion, except that the time duration was from the moment when the consonant constriction was 50% released (the moment when $m_c(t)$) was decreased to 50% of its maximum value) to the moment when it was fully released (last non-zero value of $m_c(t)$). The formant deflection direction was determined based on this mathematical metric only. No specific amount of deflection was involved in the decision due to the uncertainties on the perceptual significance of the amount of formant transitions.

Four patterns out of 64 possibilities (if both onglide and offglide of the first three formants were considered) emerged. They are illustrated in Figure 20 and numbered as pattern 1 to pattern 4, with pattern 3 matching the hypothesized pattern. In the figure, downward arrows indicate that the formant deflection direction of the V_1CV_2 sequence is going down relative to the underlying V_1V_2 sequence. Upward arrows indicate the opposite direction. Of the four patterns, patterns 1 to 3 appeared in all nine vowel contexts and pattern 4 appeared only in contexts /æi/ and /ii/. The distribution in terms of the number of stimuli of the four patterns in the nine vowel contexts are summarized in Table II. The hypothesized pattern for a /d/ (pattern 3) was represented by 743 VCV sequences, out of the total of 2160 stimuli.

VV	Ptn1	Ptn2	Ptn3	Ptn4	Total
/æi/	16	79	75	70	240
/ii/	12	119	10	99	240
/æɑ/	8	145	87		240
/ææ/	8	152	80		240
/ɑi/	6	138	96		240
/iɑ/	6	92	142		240
/iæ/	20	156	64		240
/ɑɑ/		142	98		240
/ɑæ/		149	91		240
Total	76	1172	743	169	2160

TABLE II: *A summary of the distributions of the four deflection patterns in terms of the number of stimuli in the nine vowel contexts.*

Among the 743 stimuli showing the hypothesized formant deflection pattern, 90 stimuli were chosen to generate the audio samples for the perceptual tests, with 10 samples from each vowel context. The number 90 was chosen because the distribution of stimuli

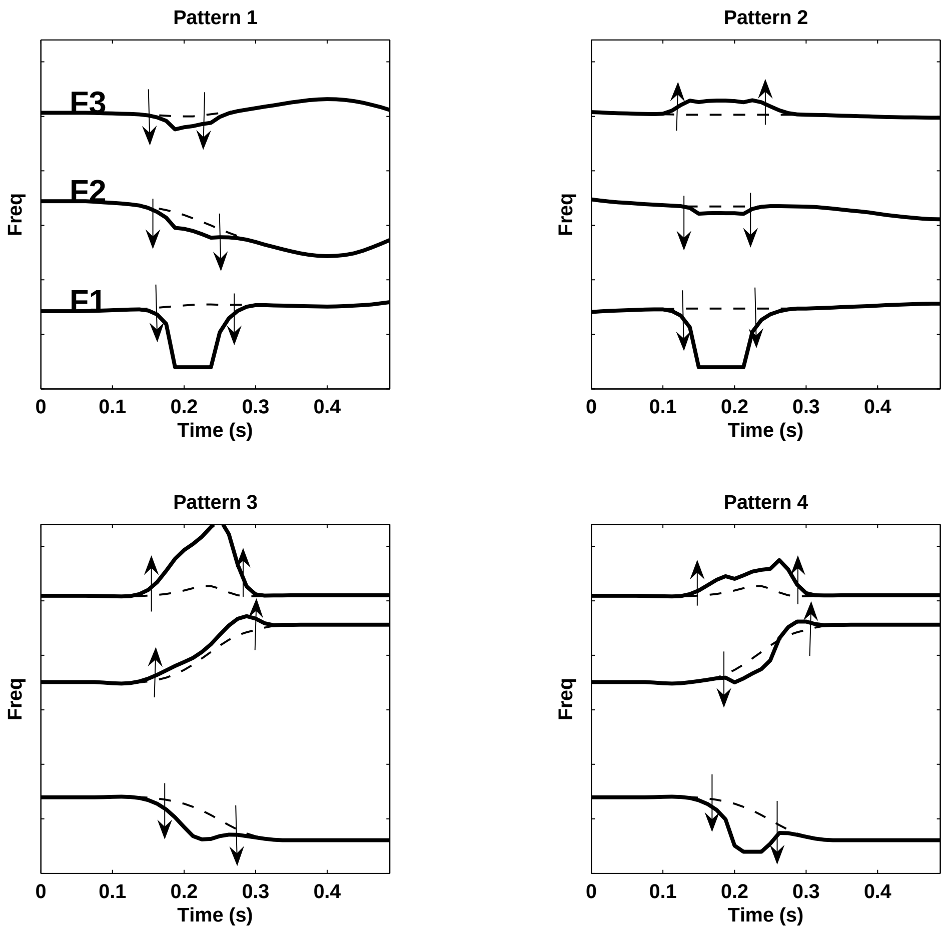


Figure 20: *Four formant deflection patterns shown in all 2160 synthesized* V_1CV_2 *sequences. The broken curves in each panel show the vowel-to-vowel formant transitions. The solid thick curves show the formant transitions when a consonant constriction is superimposed. The patterns differentiate from each other based on the onglide and offglide deflection directions. Downward arrows indicate that the formant deflection direction of the* $/V_1CV_2/$ *sequence is going down relative to the underlying* $/V_1V_2/$ *sequence. Upward arrows indicate the opposite direction.*

showing Pattern 3 in the nine contexts was uneven and the smallest number of samples was 10 in the context /ii/ (refer to Table II). If each vowel context had to be represented with equal chance to study the context effect of the formant deflection pattern using variance analysis, the number of samples in each context should be no more than 10. The samples in each vowel context were picked in a semi-random fashion, meaning that several sets of samples in each vowel context were generated randomly and the one set that had the maximum representation of all parameter values was picked as the sample set for this vowel context. This was to maximize the chance that each parameter value was represented in the vowel context.

A mouse-driven Matlab GUI (Graphic User Interface) was created to present the 90 samples to the listeners. Each sample was repeated five times. Altogether 450 tokens (90 $\times$ 5 = 450) were played to each listener in random order. The listeners were instructed to identify the consonant of each token by mouse clicking on one of the choice buttons. Six choices were given: /d, b, g, v, z, ð/. Voiced stops /b/ and /g/ were included because the wide range of the constriction location for /d/ (from section 37 to 41, corresponding to a distance range between 2.77 cm and 1.18 cm from the lips) might have extended to the locations of constrictions for /b/ or /g/. The rest of the choices were possible voiced consonants that shared a similar location of constriction as /d/. The listeners were encouraged to guess if they were not sure about what they heard. A training demonstration was provided before each test. During the training session, six naturally spoken $/V_1CV_2/$ syllables that contained the six above listed consonants were presented to the listeners five times in random order. The listeners were asked to identify the consonants. There were two purposes of the demonstration session. One was to make sure that the listeners were able to assign a phoneme to the target consonant. The other was to verify the validity of the consonant identities of the recorded $/V_1\mathrm{d}V_2/$ utterances of which the formant contours were used as a guidance to the model parameter setting. Each subject's identification function was saved as a matrix.

Eleven native speakers of American English (7F, 4M) of ages from 21 to 35 years who self-reported no any past or present speech and hearing disorders or neurological impairment were recruited as listeners for the perceptual tests. A hearing screening at 25 dBHL at octave frequencies of 500 Hz, 1 kHz, 2 kHz, and 4 kHz were conducted by the investigator to confirm their normal hearing (ASHA, 1997) at the time of the test.

The perceptual tests were performed in a sound booth in the Speech Acoustic Lab, Speech, Language, and Hearing Sciences, University of Arizona. The listeners heard the audio samples through a loud-speaker approximately 0.5 m in front of them. The audio samples were played through a 16-bit Soundblaster sound card.

RESULTS AND DISCUSSION

Individual listeners' responses to the spoken $/V_1CV_2/$ sequences in the training session were examined. There was one listener who consistently labeled consonant /v/ as /ð/. Therefore, the perceptual data of this listener were excluded from the analysis. All other listeners showed 100% agreement between the perceptual phonemes and the intended ones. The result confirmed the validity of the consonant identities of the recorded $/V_1\mathrm{d}V_2/$ utterances of which the formant contours were used as a guidance to the model parameter setting.

Plotted in Figure 21 are the response percentages of the six phonemes as functions of the stimuli. The stimuli are arranged along the x-axis in the nine vowel contexts divided by vertical lines. Each point on the identification functions represents a percentage based on 50 responses (5 repetitions x 10 subjects = 50 responses). First, the ANOVA F test was performed to test if /d/ responses were overall significantly higher than other responses. The results supported the hypothesis ($F(8, 72)=58.782$, $p<0.05$). Then, paired t-tests between /d/ response rates and each of the other five phoneme response rates were performed to determine whether they differed from each other in a significant way under the assumptions that the paired response rates were independent and identically normally distributed. The tests showed that /d/ response rates were significantly different from the other consonants within each pairwise comparison ($p<0.05$).

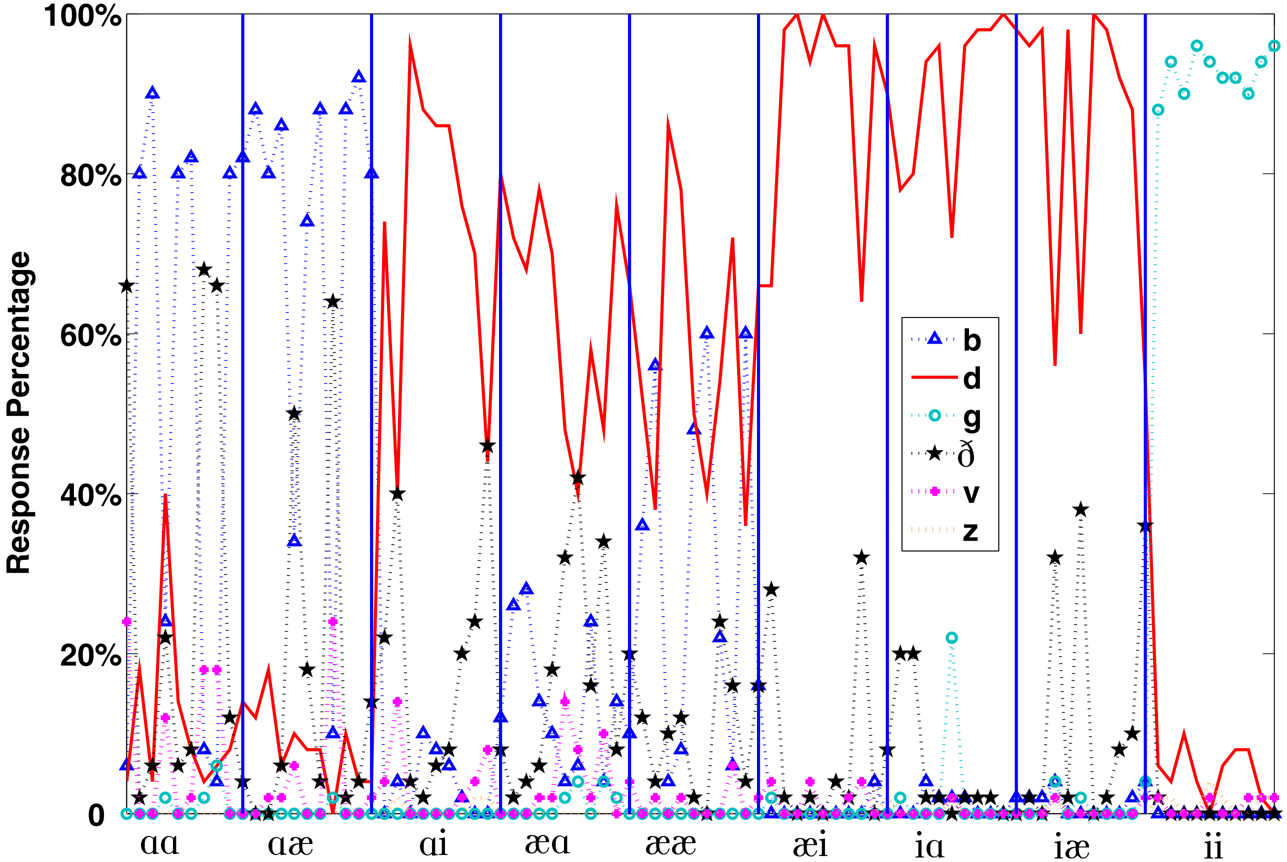


FIG. 21: *Perceptual responses (y-axis) as a function of stimuli (x-axis). The stimuli are arranged in nine vowel contexts divided by the vertical lines. Data are based on the 10 valid individual identification functions.*

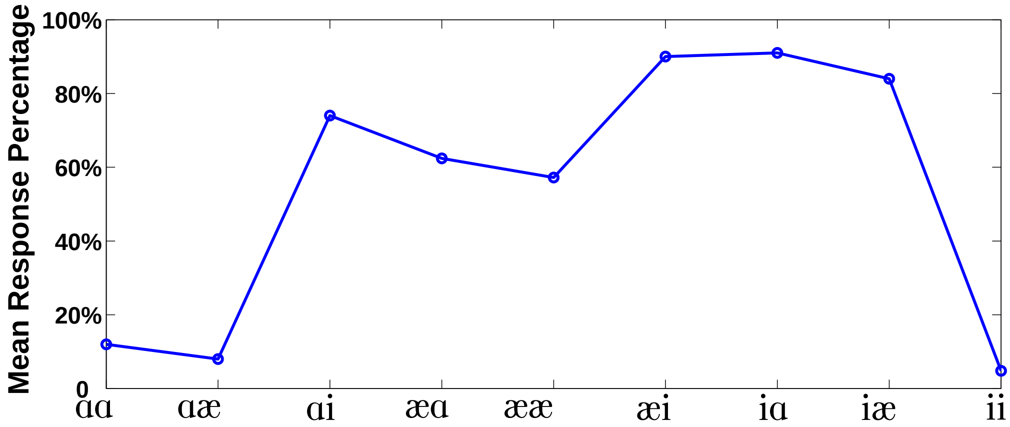


Figure 22: *The averaged /d/ response as a function of vowel contexts.*

In order to understand the vowel context effect on the identification of the alveolar stop consonant, the /d/ responses of all stimuli within each vowel context were averaged and plotted in Figure 22. The analysis of a repeated-measure general linear model (within subject) was performed on the average responses in the nine vowel contexts. The effect of the vowel context was significant ($F(8, 72) = 53.11$, $p = 0.019$). Post-hoc contrast tests showed that /d/ responses in contexts /ɑi, æɑ, ææ, æi, iɑ, iæ/ were significantly higher than those in contexts /ɑɑ, ɑæ, ii/ ($t = -20.553$, $p < 0.05$).

EXPERIMENT 2

METHOD

The model parameter values of those synthesized $/VCV/$ sequences in which the consonant had a /d/ response rate above the chance level were entered in a contingency table. The correspondence analysis was performed to extract all the relevant dimensions that represented the relationship between the model parameter values and vowel contexts. The dimensions were ordered based on the amount of variance accounted. The contributions of the largest three dimensions to the explanation of the vowel context profiles will be discussed.

The correspondence analysis was chosen to analyze the spatial (location and geometry) characteristics of the consonantal constriction based on several advantages it offers. First, it is a non-linear, multi-dimensional technique designed to describe the relationship between or among both qualitative (nominal) and quantitative variables. The quantitative variables can be either discrete (categorical) or continuous (H$\ddot{a}$rdle and Simar, 2003). In this study, the five independent variables of interest were of a mixed nature. They could be treated as either quantitative variables with discrete values (such as the 5 places of articulations) or qualitative variables (such as the symmetry of the constrictions). Second, the correspondence analysis is used to reduce the dimensional representation of the relationship between the predetermined independent variable categories and the dependent variables. In this study, the independent variables were the five model parameters. For each of the five parameters, multiple values were used to synthesize the audio samples.

Therefore, the total dimensions of the independent variables were 16 (sum of all possible values). The correspondence analysis reduces the model parameter categories and extracts the most "relevant" parameter settings for the synthesis of the consonant. Third, the practical strength of the analysis is that the data do not have to conform to a normal distribution (Shavelson, 1988), which was not a presumption of the variable distributions in this study.

The correspondence analysis technique

Correspondence analysis computes the χ^2-test statistics on a contingency table to search for the association between rows and columns. The contingency table records the frequency of occurrence of one variable as a function of another. This computation is accomplished in the following standard steps. First, the contingency table is transformed into a matrix of row and column profiles, where rows and columns are vectors of conditional densities, or relative frequencies (actual minus expected cell values as a proportion of marginal totals). Second, χ^2 distances (weighted Euclidean distances) are calculated for each element in the profile matrix. Third, singular value decomposition (SVD) is performed on the χ^2 distance matrix to find the eigenvectors (coordinates) that account for majority of the χ^2 distance. The correspondence strength between the vowel context and a model parameter depends on the row and column index of an eigenvalue and how much percentage the eigenvalue has on the recomposition of the corresponding χ^2.

According to the results of the perceptual tests, all the stimuli where the consonant was labeled as /d/ were entered in the contingency table and subject to the correspondence analysis. The criterion for a /d/ label was decided as above the chance level, i.e., any stimulus that had a /d/ response rate above 16.7% was included in the contingency table (1/6 choices = 16.7%). Altogether 63 out of 90 audio samples were included in the contingency table illustrated in Table III. The 63 samples were grouped based on their vowel contexts indicated by the rows. Contexts /ɑi, æɑ, ææ, æi, iɑ/, and /iæ/ contained 10 stimuli each. Context /ɑɑ/ contained two stimuli. Context /ɑæ/ contained one stimulus. Context /ii/ was not included in the analysis since none of the samples in this context had a response rate above the chance level. The 16 columns indicate the 16 parameter values of the five model parameters (refer to Table I): range of the constriction (r_c), skewness of the constriction (s_c), location of the constriction (l_c), magnitude of the constriction (m_c), and minimum area of the constriction (a_c). Each cell of the table records the frequency of occurrence of the corresponding parameter value. For example, in vowel context /ɑi/ in the third row, there were two stimuli synthesized with a location of section 37 of which the consonants were responded to as /d/ with higher than the chance level, regardless of the values of other parameters. Similarly, there were three stimuli synthesized with a location of section 38 of which the consonants were responded as /d/ with higher than the chance level, regardless of the values of other parameters, and so on. Indicated in the right column of the table are the total numbers of coded parameters in each vowel context. For each vowel context, this total number is five times the number of samples that were entered in the table since each sample corresponds to a combination of five parameter values. The marginal totals of the parameter values are summarized in the bottom row respectively. The total number of coded parameter values for the 63 stimuli is 315.

	location					magnitude				min area		range		skewness			Totals
VV	37	38	39	40	41	1.0	1.1	1.2	1.3	0	0.05	6	8	1.5	1	0.6	
/ɑɑ/	2	0	0	0	0	1	1	0	0	1	1	2	0	1	1	0	10
/ɑæ/	1	0	0	0	0	0	0	1	0	0	1	0	1	1	0	0	5
/ɑi/	2	3	3	2	0	5	1	1	3	4	6	2	8	5	3	2	50
/æɑ/	3	4	3	0	0	5	1	3	1	5	5	1	9	3	5	2	50
/ææ/	3	3	4	0	0	4	2	2	2	6	4	3	7	6	2	2	50
/æi/	4	3	3	0	0	3	2	3	2	6	4	2	8	7	3	0	50
/iɑ/	2	2	2	3	1	4	2	3	1	6	4	3	7	5	3	2	50
/iæ/	3	4	3	0	0	5	2	1	2	5	5	4	6	4	5	1	50
Totals	20	19	18	5	1	27	11	14	11	33	30	17	46	32	22	9	315

TABLE III: *The contingency table that summarizes the occurrence frequency of each model parameter value for the 63 stimuli of which the consonants were responded as /d/ with rates higher than the chance level. The rows indicate the vowel contexts of the 63 stimuli and the columns indicate the parameter values. Each cell of the table records the frequency of occurrence of the corresponding parameter value in a certain context.*

The row and column profiles

The row and column profiles are the frequency of each model parameter value recorded in each table cell relative to the total frequency of coded values, in this case, the number 315. The row profiles are the relative proportions of each cell within each vowel context. These relative proportions allow within-context comparisons of the variables. Similarly, the column profiles are the relative proportions of each cell within each parameter value. They allow between-context comparisons. Table IV and V illustrate the row and column profiles respectively.

In Table IV, the proportion of each parameter value appearing in a vowel context for the synthesis of /d/ is recorded. For example, in vowel context /ɑi/, it can be seen that of all the model parameter values, 0.04 (or 4%) referred to location 37; 0.06 (or 6%) referred to location 38; 0.06 (or 6%) referred to location 39; 0.04 (or 4%) referred to location 40; and 0% referred to location 41. The marginal totals in the far right column equate to 100% of all coded parameter values in a vowel context. The marginal totals at the bottom row of Table IV indicate the sum of relative frequency of occurence of each parameter value across all vowel contexts. For example, 0.74 in the first column of the bottom row is the sum of relative frequencies of occurence of location 37 across all vowel contexts.

	location					**magnitude**				**min area**		**range**		**skewness**			Marginal
VV	37	38	39	40	41	1.0	1.1	1.2	1.3	0	0.05	6	8	1.5	1	0.6	Total
/ɑɑ/	.2	0	0	0	0	.1	.1	0	0	.1	.1	.2	0	.1	.1	0	1.00
/ɑæ/	.2	0	0	0	0	0	0	.2	0	0	.2	0	.2	.2	0	0	1.00
/ɑi/	.04	.06	.06	.04	0	.1	.02	.02	.06	.08	.12	.04	.16	.1	.06	.04	1.00
/æɑ/	.06	.08	.06	0	0	.1	.02	.06	.02	.1	.1	.02	.18	.06	.1	.04	1.00
/ææ/	.06	.06	.08	0	0	.08	.04	.04	.04	.12	.08	.06	.14	.12	.04	.04	1.00
/æi/	.08	.06	.06	0	0	.06	.04	.06	.04	.12	.08	.04	.16	.14	.06	0	1.00
/iɑ/	.04	.04	.04	.06	.02	.08	.04	.06	.02	.12	.08	.06	.14	.1	.06	.04	1.00
/iæ/	.06	.08	.06	0	0	.1	.04	.02	.04	.1	.1	.08	.12	.08	.1	.02	1.00
Marginal Total	.74	.38	.36	.10	.02	.62	.30	.46	.22	.74	.86	.50	1.10	.90	.52	.18	8.00

TABLE IV: *Row profiles: relative proportion of each variable in each vowel context.*

Table V illustrates the column profiles that describe the proportion of each parameter value in one vowel context relative to the total occurrence of that parameter value across all vowel contexts. For example, of the total occurrence of location 37, 0.1 (or 10%) of occurences contributed to vowel context /ɑɑ/, 0.05 (or 5%) to /ɑæ/, and 0.1 (or 10%) to /ɑi/, etc. Similarly, of all the occurences of magnitude 1.3, none of them came from contexts /ɑɑ/ or /ɑæ/, 0.27 (or 27%) of them were from context /ɑi/, 0.09 (9%) occurences were from context /æɑ/, 0.18 (18%) were from context /ææ/, etc. Table V also indicates the proportion of context occurences relative to the sum of all vowel context occurences (marginal totals in the far right column). For example, of all the stimuli in which the consonant was labeled as /d/ above the chance level, 3% came from vowel context /ɑɑ/, 1% from context /ɑæ/, and 15% from context /ɑi/, and so on. Marginal totals close to 100% in the bottom row of Table V indicate the sums of relative frequencies of the corresponding parameter values across all vowel contexts. Some totals are not exactly 1 because of rounding error.

The numerical information contained in the row and column profiles could be transformed into graphs to search for patterns of occurrence of parameter values for each vowel context. An example is shown in Figure 23 for context /iɑ/ across all parameter values. Such a graph could be generated for each of the eight vowel contexts and 16 parameter values. It is apparent that information presented in this way would become too large to manage. Therefore, singular value decomposition procedure was used to reveal the relative contribution of each variable in a low-dimensional representation.

	location					magnitude				min area		range		skewness			Marginal
VV	37	38	39	40	41	1.0	1.1	1.2	1.3	0	0.05	6	8	1.5	1	0.6	Total
/ɑɑ/	.1	0	0	0	0	.04	.09	0	0	.03	.03	.12	0	.03	.05	0	.03
/ɑæ/	.05	0	0	0	0	0	0	.07	0	0	.03	0	.02	.03	0	0	.01
/ɑi/	.1	.16	.17	.4	0	.19	.09	.07	.27	.12	.2	.12	.17	.15	.14	.22	.15
/æɑ/	.15	.21	.17	0	0	.19	.09	.21	.09	.15	.17	.06	.20	.09	.23	.22	.15
/ææ/	.15	.16	.22	0	0	.15	.18	.14	.18	.18	.13	.18	.15	.19	.09	.22	.15
/æi/	.2	.16	.17	0	0	.11	.18	.21	.18	.18	.13	.12	.17	.22	.14	0	.14
/iɑ/	.1	.11	.11	.6	1	.15	.18	.21	.09	.18	.13	.18	.15	.16	.14	.22	.23
/iæ/	.15	.21	.17	0	0	.19	.18	.07	.18	.15	.17	.24	.13	.13	.23	.11	.14
Marginal Total	1.00	1.01	1.02	1.00	1.00	1.02	.99	.98	.99	.99	.99	1.02	.99	1.00	1.02	.99	

TABLE V: *Column profiles: relative proportion of each variable across vowel contexts.*

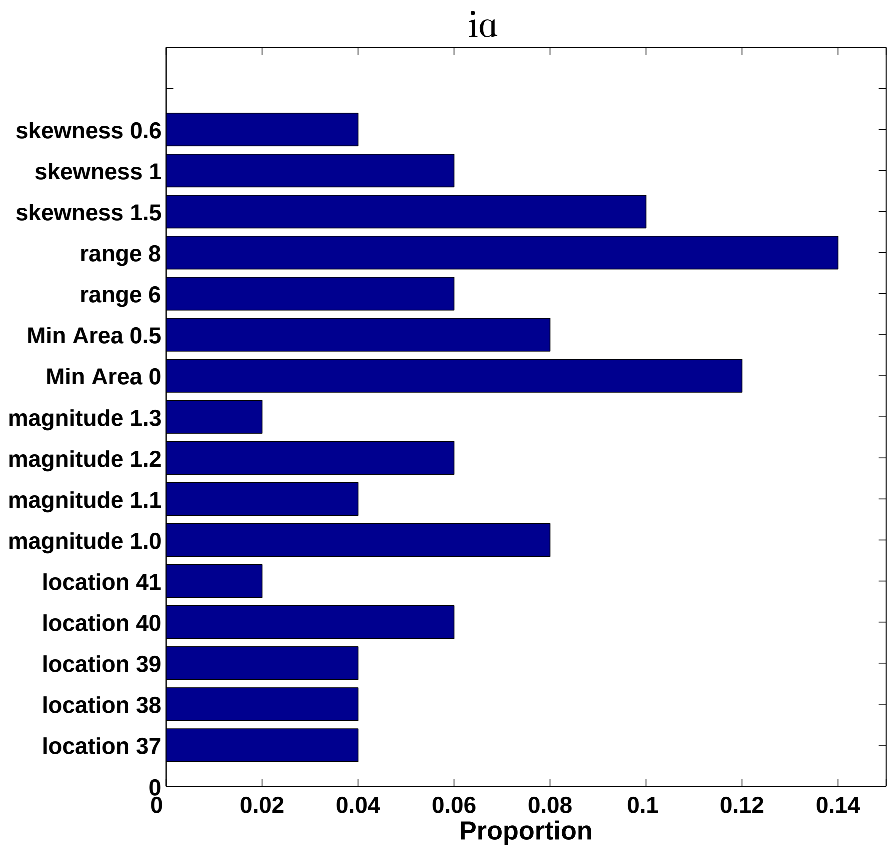


Figure 23: *Proportional marginal total of each parameter value for vowel context /iɑ/.*

Singular value decomposition

The correspondence analysis transforms and normalizes the frequency data in the contingency table by calculating χ^2 distances from the row and column profiles. Then a singular value decomposition (SVD) is performed to factor the basic structure of the χ^2 distance matrix, resulting in a set of row vectors, column vectors, and singular values. The singular value decomposition reduces the complexity contained in the row and column profiles and creates a low-dimensional representation of the dependence between the model parameters, as shown in Table VI. In the table, the singular values indicate the relative contribution of each dimension to the explanation of the variance. The inertia indicates how much of the variance in the original data is accounted for in the dimensional solution. As in principal components analysis, the first dimension explains as much variance as possible, the second dimension is orthogonal to the first dimension and it explains as much of the remaining variance, and so on.

Dimension	**Singular Value**	**Inertia**	**Proportion Explained**	**Cumulative Proportion**
1	.242	.058	33.22	33.22
2	.214	.046	25.85	59.08
3	.191	.037	20.58	79.65
4	.127	.016	9.05	88.71
5	.114	.013	7.33	96.04
6	.075	.006	3.18	99.22
7	.037	.001	.78	100

TABLE VI: *Singular values and proportion of variance explained.*

The cut-off point for how many dimensions should be included in the analysis is a balance between the number of dimensions and their cumulative proportion of variance explained. The purpose of running the correspondence analysis was to reduce the complexity of the data. However, the reduced complexity would not be helpful if important features were overlooked. Hair, Anderson, Tatham, and Black (1995) recommended a

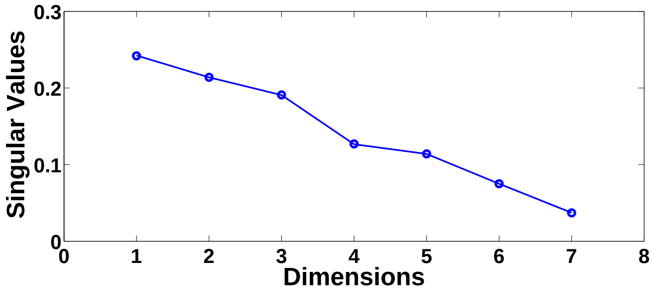


Figure 24: *Scree plot of singular values.*

cut-off singular value of 0.2, indicating that any dimensions with a singular value greater than 0.2 should be included in the analysis. In this case, such a cut-off value would include only the first two dimensions with a cumulative percent of 59.08% (refer to Table VI). In this case, such a cumulative percentage seemed insuficient. In order to assist in the decision about the number of dimensions, a scree plot is presented in Figure 24 to visualize the drop of the singular value. In the figure, the value drops away more rapidly after the third dimension indicating that the amount of variance being accounted for drops rapidly after the third dimension. At this dimension, the cumulative percentage was 79.65% (refer to Table VI). Therefore, three dimensions were included in the analysis.

The graphical display

The visual presentation of the first three dimensions are plotted in Figure 25, Figure 26 (dim1-dim2), and Figure 27 (dim3). The coordinates and contribution to the inertia of each variable are summarized in Table VII. The coordinates used to plot the figures resulted from the symmetrical normalization that spreads the inertia across both the row and column profiles. Such normalization permits the placement of parameter values (column variables) and vowel contexts (row variables) in the same graphical display. In such display the distances between parameters and vowel contexts are not Euclidean distances, and therefore, it is not appropriate to measure the distance between row and column variables. However, comparing the general orientation of row points and column points is well served by the symmetrical normalization option.

Parameter& Context	Margin	**Dimension 1**		**Dimension 2**		**Dimension 3**	
		coordinate	contribution	coordinate	contribution	coordinate	contribution
/ɑɑ/	.03	.6033	.1966	.9016	.5640	.0140	.0002
/ɑæ/	.01	.2002	.0108	-.1827	.0116	1.1646	.5911
/ɑi/	.15	-.1926	.1002	-.0732	.0186	-.1335	.0776
/æɑ/	.15	.0469	.0059	-.2352	.1918	-.0298	.0039
/ææ/	.15	.0950	.0244	-.0487	.0082	.0223	.0022
/æi/	.14	.1630	.0718	-.0737	.0188	.1683	.1234
/iɑ/	.23	-.4320	.5039	.2286	.1813	.0923	.0371
/iæ/	.14	.1790	.0865	.0402	.0056	-.1943	.1645
L37	.0925	.3655	.1443	.2116	.0621	.2736	.1305
L38	.0475	.0512	.0027	-.2238	.0660	-.1859	.0572
L39	.0450	.0241	.0006	-.1983	.0491	-.1375	.0297
L40	.0125	-1.3868	.5192	.5043	.0882	-.0104	.0000
L41	.0025	-1.7817	.1714	1.0687	.0792	.4835	.0204
M10	.0775	-.0136	.0003	.0102	.0002	-.1919	.0867
M11	.0375	.1753	.0183	.3765	.1082	-.0292	.0008
M12	.0575	-.0853	.0055	-.1850	.0332	.5557	.3768
M13	.0275	-.0333	.0007	-.1661	.0211	-.2070	.0411
A0	.0925	-.0102	.0002	.0384	.0034	-.0334	.0032
A005	.1088	.0112	.0002	-.0422	.0037	.0368	.0035
R6	.0625	.2182	.0437	.5430	.3477	-.1539	.0351
R8	.1375	-.0806	.0161	-.2007	.1285	.0569	.0130
S15	.1125	.0320	.0018	.0208	.0010	.1884	.0990
S1	.0650	.1009	.0121	.0158	.0004	-.1883	.0644
S06	.0225	-.3604	.0631	-.1127	.0079	-.2217	.0386

TABLE VII: *Correspondence analysis solution.*

In Figure 25, the model parameters are marked by various symbols as explained in the legend inlet: locations are indicated by “$\bigtriangledown$”, magnitudes are indicated by “$\bigcirc$”; minimum areas are indicated by “$*$”, ranges are indicated by “$\times$”, skewness values are indicated by “$\bullet$”, and vowel contexts are indicated by “+”. In the figure, the name of a vowel context is indicated next to the “+” symbol; the combination of a letter and a number next to a parameter symbol indicates the parameter name and value. The letter is the first letter of the parameter name except that “A” means minimum area. For example, L41 means that the “$\bigtriangledown$” symbol next to it represents location 41, A0 means that the “$*$” symbol next to it represents a minimum area of 0 cm^2, A005 means that the “$*$” symbol next to it represents a minimum area of of 0.05 cm^2. The bottom cluster in Figure 26 is magnified and replotted in Figure 26 for a better visualization.

RESULTS AND DISCUSSION

Dimensions 1 and 2

Dimension 1, which accounted for 33.22% of the variance (refer to Table VI), appears to contain parameters that relate to the anterior-posterior geometry of a consonant constriction profile. In Figures 25 and 26, the left-hand pole (negative pole) of Dimension 1 contains location values of 40 and 41, indicating anterior places of articulation, compared to the posterior locations of 37, 38, and 39 on the right-hand pole of this dimension (see Figure 26 for the magnifed portion of Figure 25). The parameter skewness that defines the symmetry of a consonant constriction profile is also located along the first dimension,

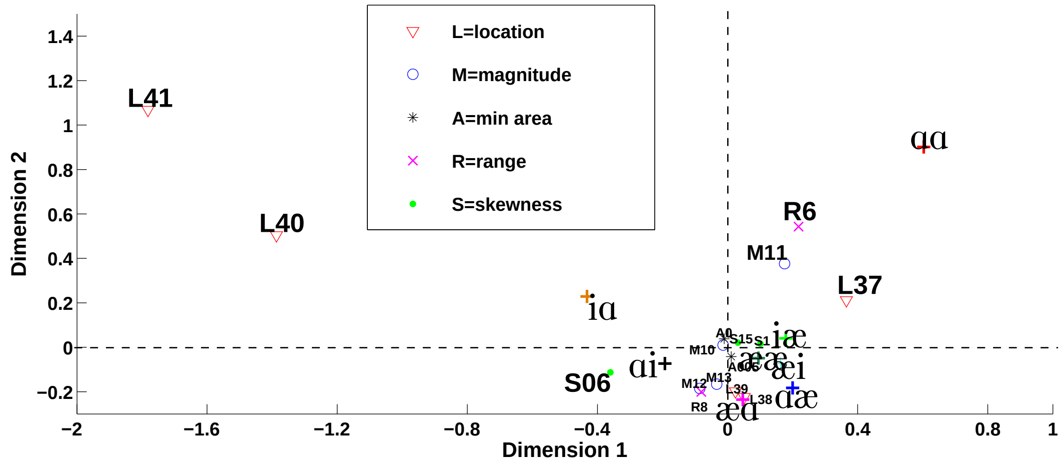


Figure 25: *Correspondence analysis scaling solution coordinates for model parameters and vowel contexts. Vowel contexts are indicated by "+" symbols and labeled beside their symbols. The model parameters are marked by various symbols as explained in the legend inlet. They are also labeled by the combination of the first letter of their parameter name and the values except that "A" represents minimum area. For example, "M10" indicates the magnitude value of 1.0; "A005" indicates the minimum area of 0.05* cm^2. *Dimension 1 is on the horizontal axis and Dimension 2 is on the vertical axis. The dashed vertical and horizontal lines at zero values of the axes divide the two dimensions into positive and negative poles. The bottom cluster of variables is magnified in Figure 26 for a better visualization.*

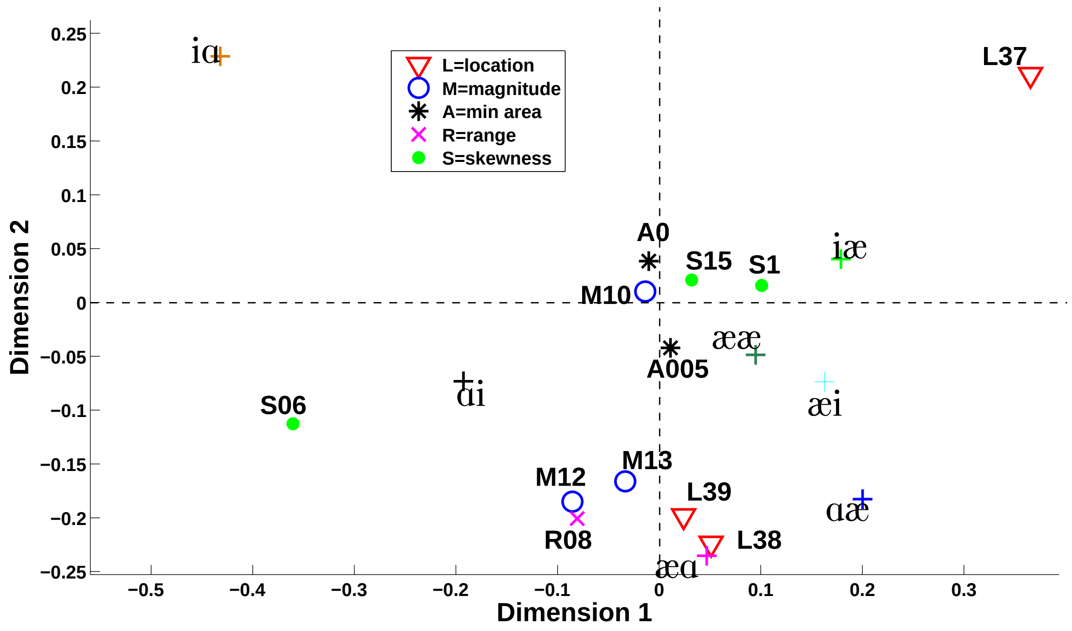


Figure 26: *The magnified bottom cluster of variables in Figure 25.*

with the value 0.6 (posteriorly skewed) being located at the left-hand side and the value 1.5 and 1 (anteriorly skewed and symmetrical) at the right-hand side. Although the two values of constriction ranges are not obviously located along the first dimension, they do reside on each side of the dimension, with the value 8 on the left side and 6 on the opposite side.

Along Dimension 2, which accounted for 25.85% of the variance (refer to Table VI), are parameters that relate more to the vertical axis of the consonant constriction profile, such as the magnitude and minimum area. For example, magnitude values of 1.0 and 1.1 are located on the upper pole (positive pole) of Dimension 2, and the two greater values of 1.2 and 1.3 are located on the lower pole (negative pole) of the dimension. Similarly, the minimum area value of 0 cm^2 is located at the upper pole and the value of 0.05 cm^2 is located at the lower pole.

It can be seen from Figures 25 and 26 that the poles of Dimensions 1 and 2 divide the variables into four quadrants. The general orientation of row points and column points within each quadrant provides information about the association between the row and column variables. In the upper right quadrant of the map are the vowel contexts /ɑɑ, iæ/, location 37, two skewness values of 1.5 and 1, a range value of 6, and a magnitude value of 1.1, indicating that these parameter values are most associated with the synthesis of /d/ in vowel contexts /ɑɑ/and /iæ/. Similarly, the lower right quadrant associates vowel contexts /ææ, æi, æɑ, ɑæ/ with locations 39, 38 and the minimum area of 0.05 cm^2. The upper left quadrant connects the vowel context /iɑ/ to the locations 41 and 40, a magnitude of 1.0, and a minimum area of 0 cm^2. The lower left quadrant reveals the association between the vowel context /ɑi/, and the skewness value of 0.6, the range value of 8, magnitude values of 1.2 and 1.3.

It is important to note that the two-dimensional chart in Figure 25 resulting from the correspondence analysis provides a graphical display of the association between column and row variables, but the analysis does not provide a measure of how close this association is between a particular vowel context and a parameter value. In other words, only non-precise general statements between the variables can be made. This is because of two reasons in addition to the normalization limitation mentioned previously in this documentation. First, the $/VCV/$ sequences of which the model parameters were entered in the contingency table for analysis were any sequences with a /d/ response rate higher than the chance level. In other words, no information about how “good” the consonants were perceived was involved in the analysis. Therefore, the association between vowel contexts and parameter values derived from Figure 25 should not be taken as an interpretation of a causal relationship between a definite /d/ identification and synthesis parameters. Rather, the association provides a guidance of how the parameter values could be chosen to yield a possible /d/ identification. Second, vowel context /ii/ was completely excluded from the analysis because of the low identification rates for all the samples in this vowel context. In other words, none of the model parameter combinations was able to characterize the /d/ synthesis in this context. Given the assumption that the correspondence analysis extracts the most important dimensions that were already included in the variables, variables that were overlooked or not included at the beginning of the analysis would not be "discovered" in the final scaling solution. Therefore, no association could be provided between the synthesis of /d/ and vowel context /ii/.

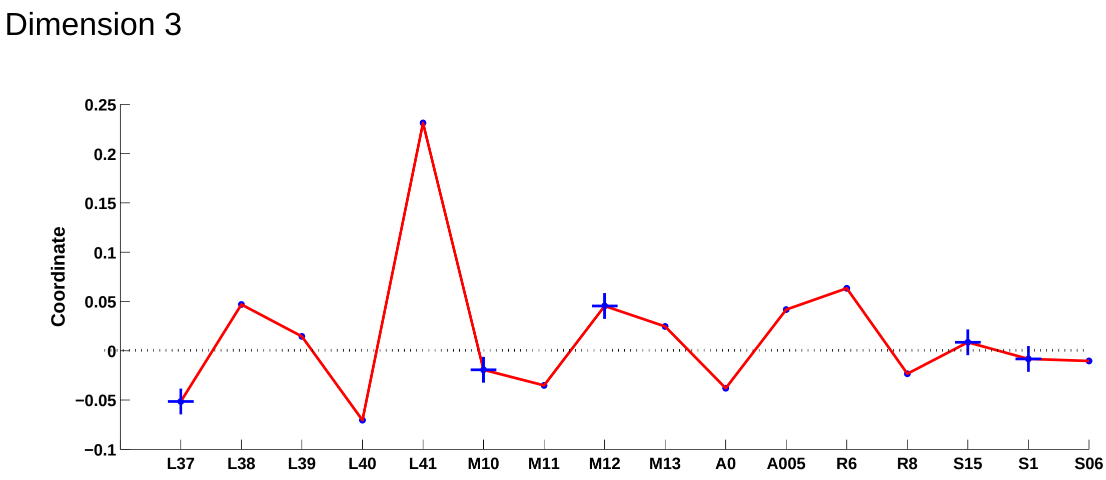


Figure 27: *Correspondence analysis scaling solution, Dimension 3. The "+" symbols indicate the parameter values that contribute more than their expected proportions to this dimension.*

Dimension 3 accounts for 20.58% of the variance (refer to Table VI). It is displayed in Figure 27. The “+” symbols indicate the parameter values that contribute more than their expected proportions to this dimension. Expected proportion of a variable is calculated by dividing 1 (the possible total contribution) by the total number of variables, in this case 16 (total number of parameter values), giving an expected proportion of contribution per variable of 0.0625. Therefore, variables that contribute more than 0.0625 are considered as important for the interpretation of this dimension, in so far as they contribute more than would be expected by chance.

The dashed horizontal line in Figure 27 divides Dimension 3 into the negative pole (lower portion) and positive pole (upper portion) of Dimension 3. The negative pole has location 37, magnitude 1.0, and skewness 1.0 (symmetrical) as its major contributors. At the positive pole of Dimension 3 lie magnitude 1.2 and skewness 1.5 (anteriorly skewed) as its major contributors. It is difficult to find an interpretation of this dimension that is different to what has been indicated by Dimensions 1 and 2. However, Dimension 3 accounts for a substantial proportion of the variance in vowel context /ɑæ/’s profile (59.11%, Table VII). Dimension 3 also contributes substantially to the variances of contexts /iæ/ (16.45%, Table VII) and /æi/ (12.34%, Table VII).

Dimension 3 seems to capture the geometry of a constriction along both the horizontal and vertical directions of a constriction profile. For example, the dimension divides the magnitudes values of 1.0 and 1.1 at the negative pole, from those of 1.2 and 1.3 at the positive pole. It also divides the skewness value of 1.5 at the positive pole from that of 0.6 at the negative pole.

EXPERIMENT 3

METHOD

Formant contours of the spoken $/V_1dV_2/$ and corresponding $/V_1V_2/$ sequences were extracted and plotted. The method of extracting the formant contours of the utterances was documented under section 3.1.2. The time scales of all formant contours were normalized to the length of 40 sample points, the same length as the synthesized sequences. The formant deflection patterns of the naturally spoken $/V_1dV_2/$ sequences relative to the underlying $/V_1V_2/$ sequences were compared with those of the synthesized ones. Time-varying vocal tract modes during the $/V_1CV_2/$ sequences were also compared between the natural utterances and the synthesized ones.

Figure 28 shows the formant contours of the nine pairs of spoken $/V_1dV_2/$ and $/V_1V_2/$ sequences. Similar to previous figures, the x-axis indicates the sample points and the y-axis indicates frequency in Hertz. Formant contours of $/V_1dV_2/$ sequences are plotted as solid lines. The gap in the middle resulted from deletion of data points when the LPC algorithm could not extract the formants of a closed vocal tract configuration. Formant contours of the corresponding $/V_1V_2/$ sequences are plotted as the dotted lines. The $/V_1V_2/$ context of the formant contours is labeled at the top of each panel. The arrows show the formant deflection directions from the underlying $/V_1V_2/$ sequence to the $/V_1dV_2/$ sequence.

Formant contours of nine synthesized $/V_1CV_2/$ sequences (solid lines) and their corresponding $/V_1V_2/$ (dashed lines) sequences are plotted in Figure 29. Each $/V_1CV_2/$

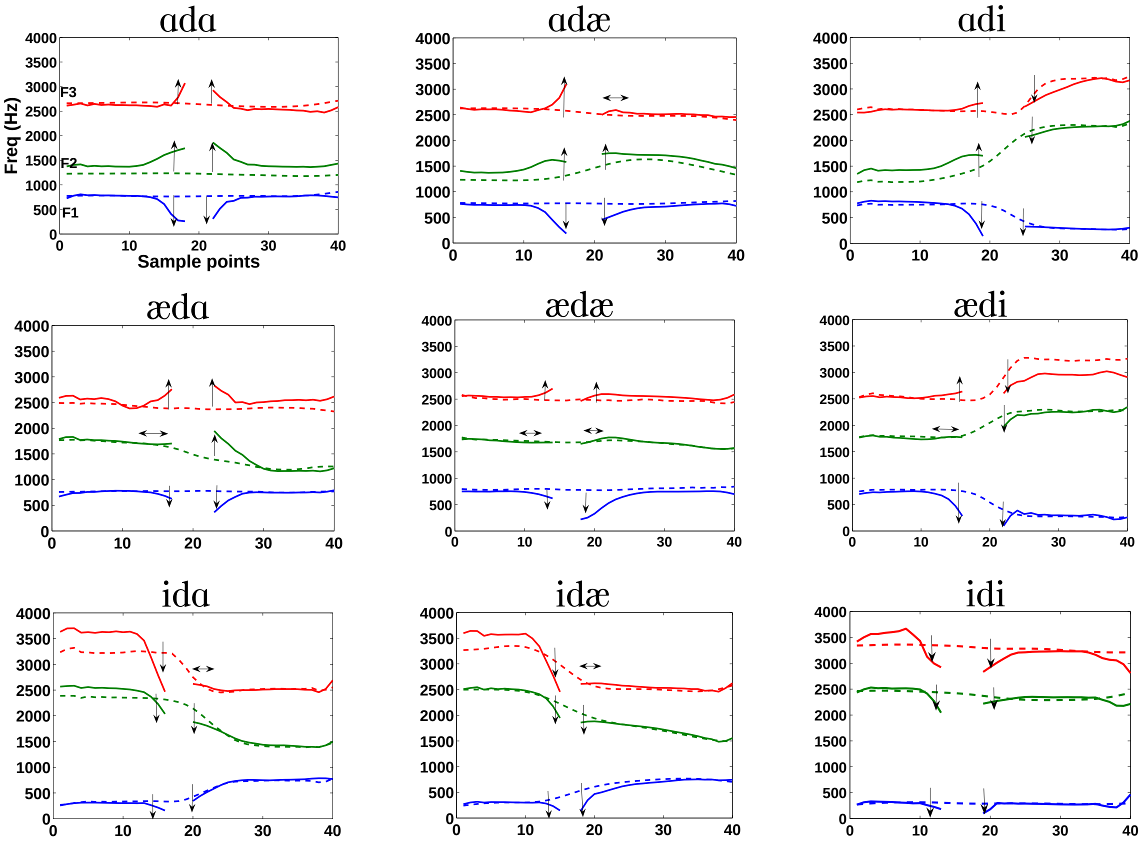


Figure 28: *Formant contours of the nine pairs of spoken* $/V_1dV_2/$ *(solid lines) and* $/V_1V_2/$ *(dotted lines) sequences. The gaps in the middle of the* $/V_1dV_2/$ *contours were resulted from deletion of data points when the vocal tract was fully closed. The x-axis indicates sample points. The y-axis indicates frequency in Hertz.*

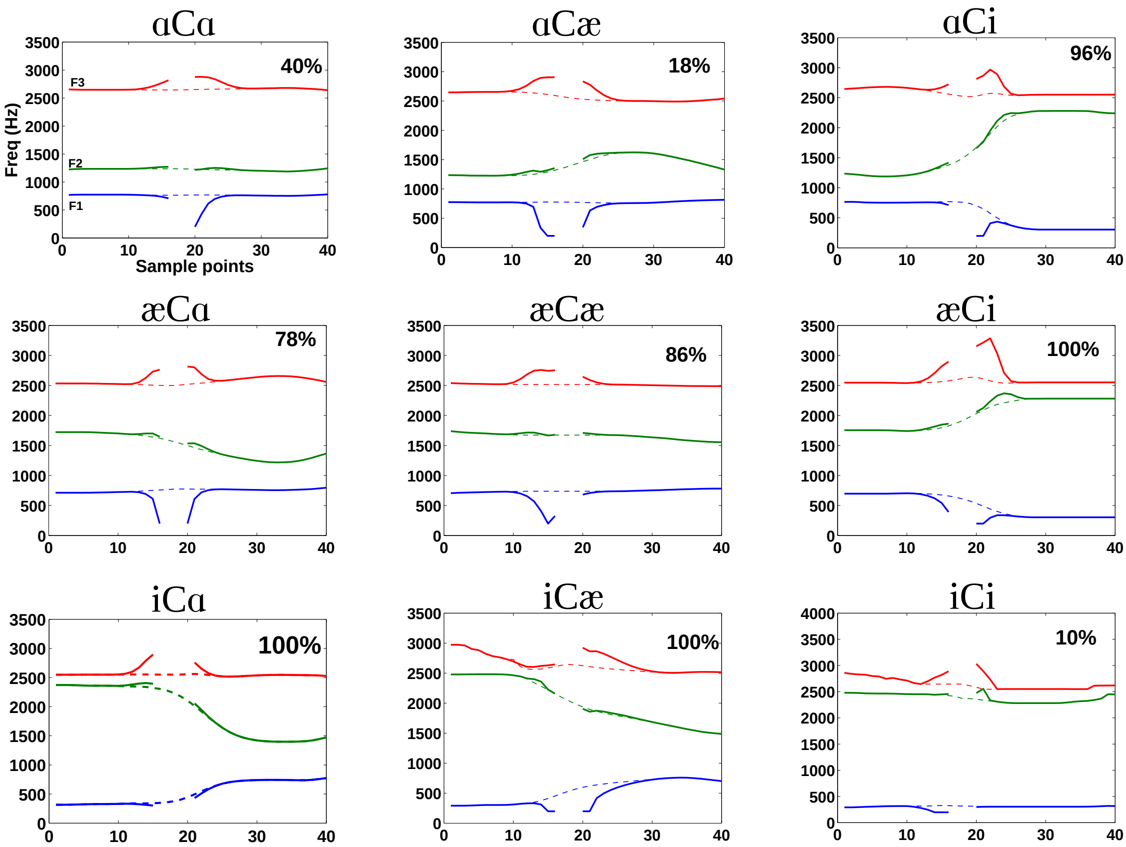


Figure 29: *Formant deflection patterns of nine synthesized* $/V_1CV_2/$ *sequences (one sequence from each context). Each sequence plotted had the highest /d/ response rate among all the samples in its vowel context. The response rate is indicated at the upper right corner of each panel.*

sequence of which the pattern is plotted had the highest /d/ response rate among all the samples in its vowel context. The response rate is indicated at the upper right corner of each panel. For example, for context /ææ/, the plotted /æCæ/ sequence had a /d/ response rate of 86%, which was the highest /d/ rate among all 10 samples in this context. All the sequences in Figure 29 show the same deflection pattern since all the audio samples in the perceptual tests were chosen from sequences that showed the hypothesized pattern.

RESULTS AND DISCUSSION

In the naturally spoken samples, shown in Figure 28, only one sample, /ɑdɑ/, showed the formant deflection pattern that agreed with the hypothesized one, i.e., F1 transition in /ɑdɑ/ was deflected downward from that of /ɑɑ/ (indicated by the downward arrows), and F2 and F3 transitions were deflected upward from those of /ɑɑ/ (indicated by the upward arrows). Deflection patterns of sequences /ɑdæ/ and /ædɑ/ are very similar to the hypothesized pattern. For sequence /ɑdæ/, the F3 offglide closely followed the underlying vowel-to-vowel formant contour as indicated by the double arrow without an obvious direction. The beginning of the transition is slightly above the underlying /ɑæ/ transition. But such fluctuation more likely resulted from the signal processing technique or between-utterance variations than from the true formant deflection. Similarly, the F2 onglide of utterance /ædɑ/ followed the underlying F2 transition of /æɑ/, instead of being deflected upward in the hypothesized pattern. For all other sequences, the deflection directions show some variations from the hypothesized pattern. For example, in sequence /ædæ/, the consonant constriction did not cause much deflection to the F2 onglide and offglide;

in sequences /ɑdi/, /ædi/, and /idi/, F2 and F3 offglides were deflected downward; in /idɑ/ and /idæ/, F3 offglides followed the vowel-to-vowel transition, and all other formant transitions were deflected downward. Based on these observations, it is concluded that the hypothesized formant deflection pattern did not provide consistent cues for the identification of the place of articulation in the naturally spoken $/V_1dV_2/$ sequences.

It is also observed that the influence of a consonant constriction on the vowel formant could start as early as the beginning of a vowel. For example, in contexts /ɑɑ/, /ɑæ/, and /ɑi/ (first row of Figure 28), the entire F2 contours of vowel /ɑ/ in the $/V_1dV_2/$ sequences shifted up about 200 Hz above the F2 contours of the underlying /ɑ/ in $/V_1V_2/$ sequences. Similarly, F3 contours of initial /i/'s in /idɑ/, /idæ/, and /idi/ shifted up about 200-300 Hz compared to the F3 contours of the underlying $/V_1V_2/$ sequences. The formant contours of vowel /æ/ did not shift significantly as shown in the panels of the second row of Figure 28. The small frequency variations between the $/VV/$ and $/VdV/$ contours for vowel /æ/ could be caused by between-utterance variations and artifacts of the signal processing technique.

The synthesized $/V_1CV_2/$ sequences showed the same formant deflection pattern as plotted in Figure 29. The formant contours before the onset and after the offset of the consonant constriction were the exact replica of the underlying V_1 and V_2 contours since the model parameter settings before the onset and after the offset of the consonant constriction were exactly the same for each pair of $/V_1CV_2/$ and $/V_1V_2/$ sequences. The hypothesized pattern in the synthesized $/V_1CV_2/$ sequences yielded high /d/ response rates in most vowel contexts except in /ɑɑ/, /ɑæ/, and /ii/ (see Figure 21). It is noted that the pattern in context /ii/ yielded consistently high /g/ response rates for all samples in the context. The average /g/ response rate was 92.6% in this context. In order to find out if the naturally spoken utterance /igi/ showed the same formant deflection pattern as that of the synthesized /iCi/ sequences, the formant deflection pattern of the spoken /igi/ utterance was extracted and plotted in Figure 30, along with the pattern of a synthesized /iCi/ sequence that had the highest /g/ response rate of 96%. All other synthesized /iCi/ sequences showed very similar deflection patterns.

For naturally spoken /igi/ shown in the left panel of Figure 30, the F1 transition was deflected downward relative to the F1 contour of /ii/. F2 and F3 followed those of the underlying vowel-to-vowel contours without any obvious transition deflections. The synthesized /iCi/ in the right panel of Figure 30 with a high /g/ response rate (96%) showed some formant deflection. But it was suspected that the formant transitions of the synthesized /iCi/ might not be resolved by the auditory system. At the frequency range of 2500 - 3000 Hz, the critical bandwidth of the human auditory system is roughly 300 to 350 Hz(Zwicker, 1957; Glasberg & Moore, 1990), which means that the time-varying formant transitions between 2500 Hz and 2700 Hz at the onset and offset of the consonant constriction are perceptually equivalent to one strong formant centered at about 2600 Hz, roughly the region of F2 and F3 of vowel /i/. Such transition pattern at the constriction release has been described as the acoustic cue for /g/ identifications repeatedly (Blumstein, et. al., 1979; Fant, 1960; Lehiste, & Peterson, 1961). Therefore, the synthesized /iCi/ yielded high /g/ response rates, even though the acoustic pattern showed some variations from that the naturally spoken one.

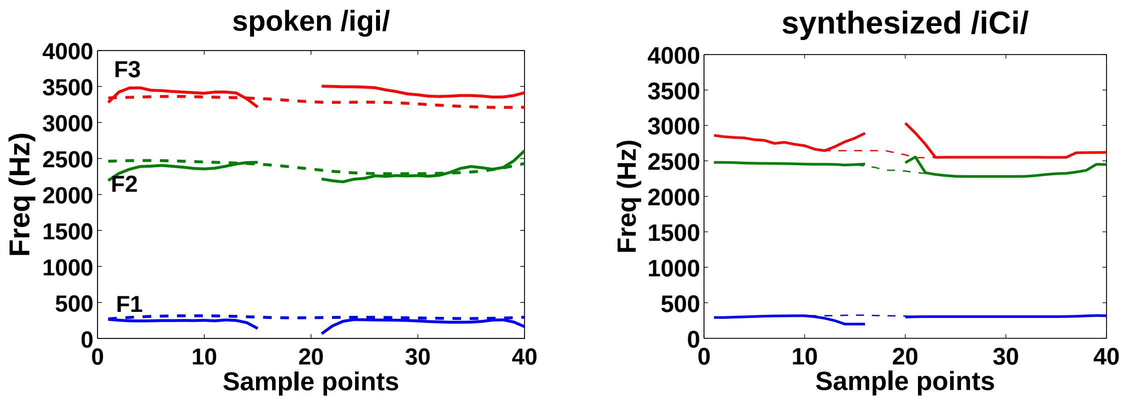


Figure 30: *The formant deflection pattern of the spoken /igi/ utterance (left) and a synthesized /iɕi/ sequence that had a /g/ response rate of 96% (right).*

In order to understand the articulatory interpretation of the vowel context effect, the time-varying coefficients (q_1 and q_2) of the two most significant modes were studied. The modes were derived from the principal component analysis of the covariance matrix of ten vowel area functions (see section 2.5). The mode coefficient trajectories of a vowel-to-vowel transition have been shown to capture some shaping pattern changes of the vocal tract during the transition (Story and Titze, 1998, Story, 2005, 2006). The first mode largely represents the forward and upward movement of front the tongue and /or upward jaw movement and the second mode is associated with the up-down movement in the middle of the vocal tract, such as the arching of the tongue body.

Figure 31 plots the time-varying coefficients (q_1 and q_2) of the two most significant modes (ϕ_1 and ϕ_2), along with their corresponding F1-F2 contours used for the synthesis of $/VCV/$ sequences. The percentages in the formant space panels indicate the average /d/ response rate in that vowel context. The solid horizontal line in the $q_1 - q_2$ spaces (left columns) indicates the range of q_1 values and the dashed vertical line indicates the range of q_2 values. The intersection point of the two lines is the point where both modes have a value of zero, representing the mean area vocal tract shape of the ten vowel area vectors used to derive the mode coefficients. The corresponding time-varying F1-F2 contours of each $/VV/$ transition used to synthesize the steady state vowels in $/VCV/$ sequences are plotted in the right columns. The vowel context is indicated at the top of each pair of panels.

Figure 31 demonstrates the mapping relationship between the mode trajectories and the F1-F2 contours for each VV transition. For example, for the context /ɑi/, the mode trajectory in the $q_1 - q_2$ space (left panel) starts at the upper right quadrant of the space (marked by an empty circle). It then traverses across the upper part of the space and covers essentially the whole range of q_1 and a lesser range of q_2. The trajectory ends at the edge of the upper left quadrant marked by the '+' sign at the end. The corresponding F1-F2 contour in the panel next to it shows a similar shape of trajectory rotated at a certain angle. Take the /ɑɑ/ transition for another example, the mode coefficient trajectory does not show much of a "trajectory" but rather a dot, indicating little vocal tract shape change from the initial /ɑ/ to the final /ɑ/. As a result, F1 and F2 show little variation.

Comparisons of the mode coefficient trajectories between the three vowel contexts (/ɑɑ/, /ɑæ/, and /ii/) that had low /d/ response rates and those that had high /d/ response rates indicate that trajectories in the low response rate contexts tend not to approach the neutral points of q_1 (points on the dashed vertical line). For example, the trajectories of /ɑɑ/ and /ii/ transitions are essentially constant points that are relatively distant from the dashed vertical line; the trajectory of /ɑæ/ is very much positioned in the upper right corner

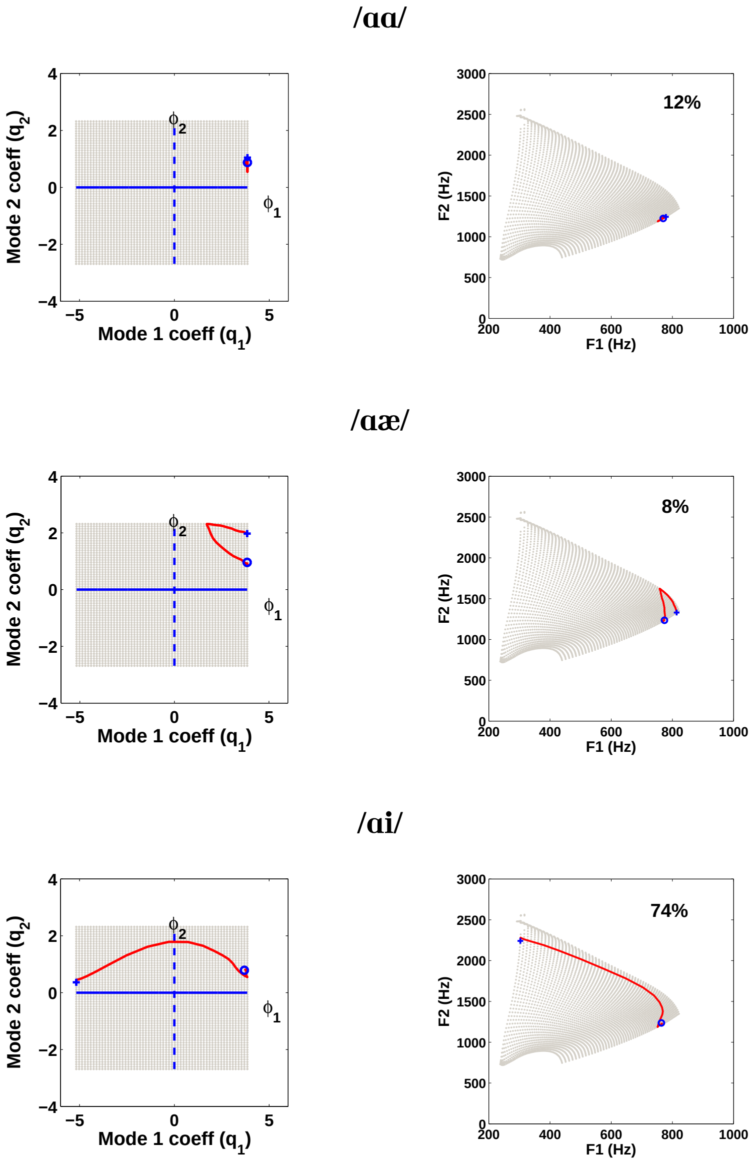


Figure 31: *Coefficient and formant spaces for the synthesized VCV sequences. In each pair of panels, the left panel is q_1-q_2 space and the right one is the F1-F2 space. The percentages in the formant space panels indicate the average /d/ response rate in that vowel context. The starting and ending points of the coefficient trajectories and formant trajectories are marked by "o" and "+" symbols respectively.*

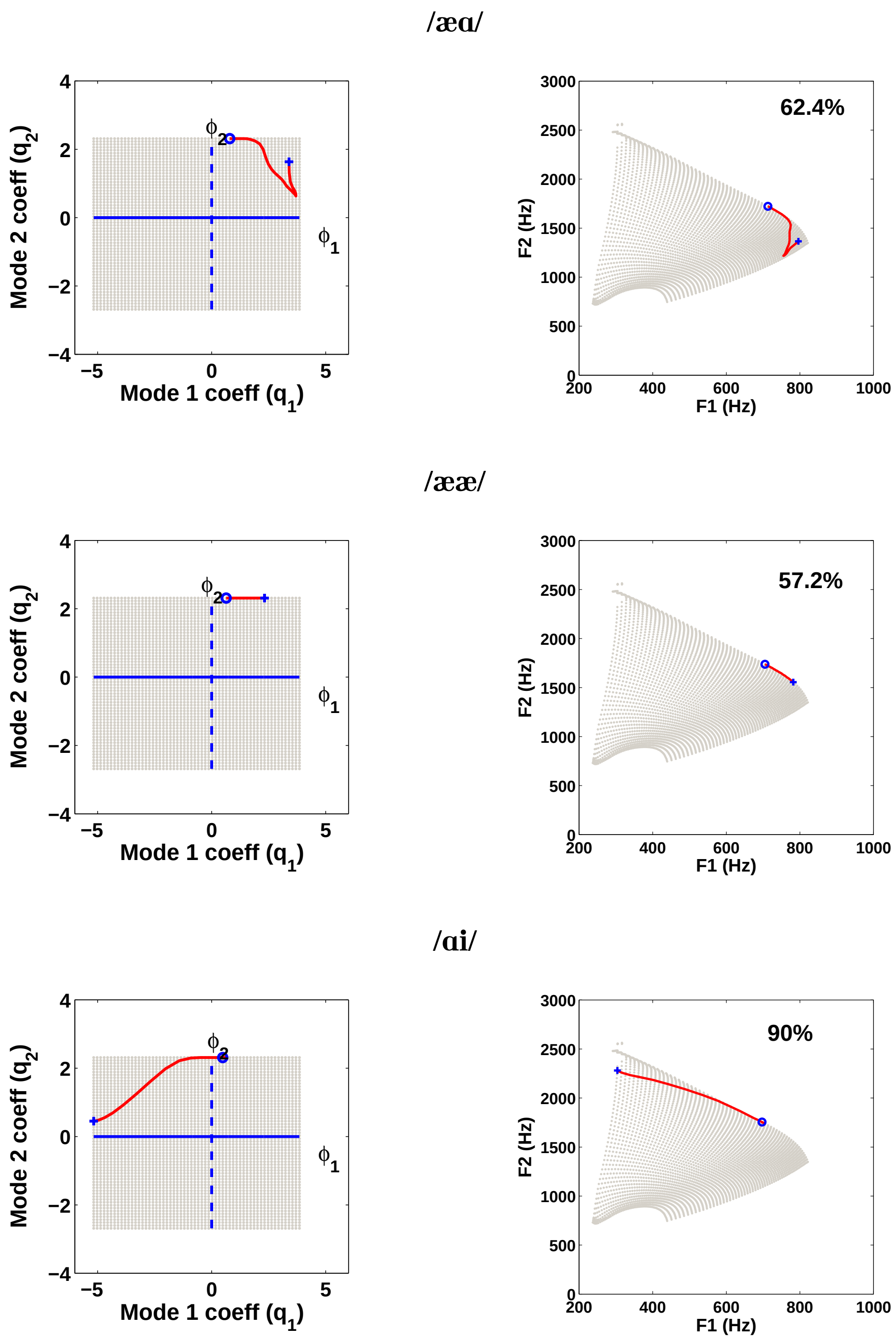
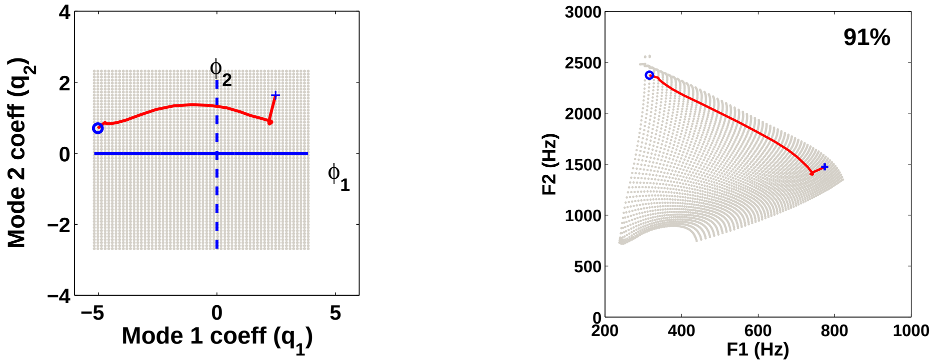
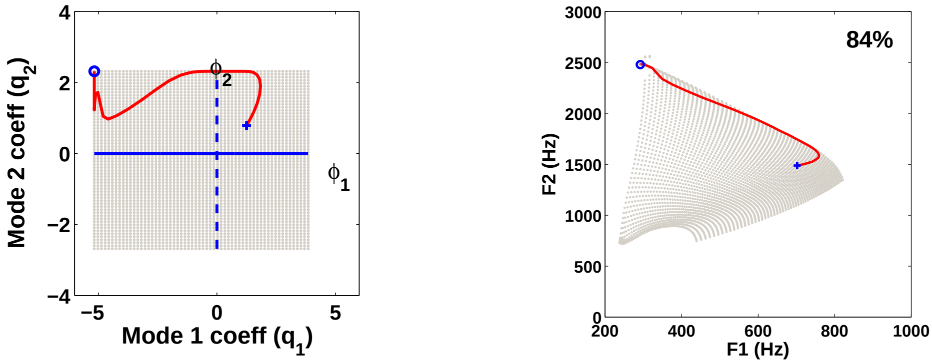


Figure 31: *To be continued*

/iɑ/



/iæ/



/ii/

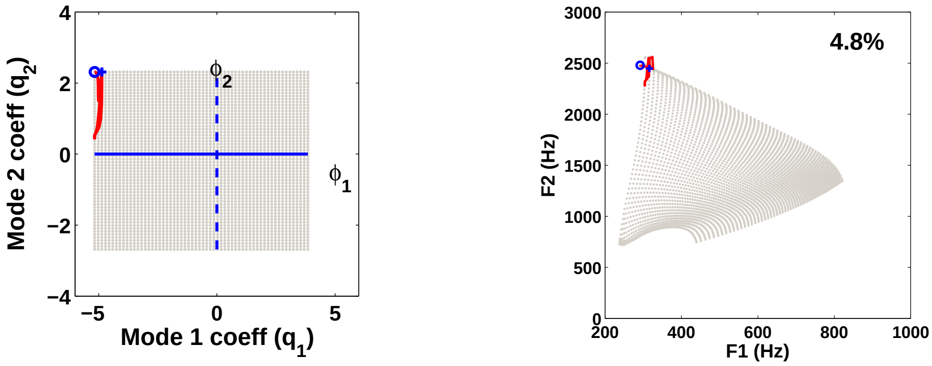


Figure 31: *To be continued*

of the quadrant and relatively far away from the neutral point of q_1. On the other hand, the /ææ/ trajectory that shows a high average /d/ response rate (86%) approaches the zero points of mode one (dashed vertical line) very closely. All other $/VV/$ trajectories either pass a zero point of q_1 (such as /ɑi, æi, iɑ, iæ/) or approach it very closely (such as /æɑ/). From the articulatory point of view, a zero q_1 value results in a zero value of mode one ($q_1 \times \phi_1 = 0$), i.e., no or nearly no area expansion or reduction is superimposed on the neutral vocal tract shape from mode one at the moment the coefficient trajectory passes or approaches the zero point. Although higher modes with non-zero values may still perturb the neutral vocal tract shape somewhat, such perturbations will not significantly deviate the vocal tract configuration from the neutral state since mode one alone accounts for more than 80% of the total variance of the area vector set. Based on this relationship, passing or approaching to a zero value of q_1 indicates that a neutral vocal tract shape or nearly a neutral shape is formed or intended during the VV transition. It was speculated that the intended or actually formed neutral vocal tract shape during the VV transition might have had an effect on the formant transtions, and consequently, on the perception of the consonant.

The speculation was supported indirectly from our previous work on the perceptual categories of constriction regions along the vocal tract for voiced stop consonants /b, d, g/ in neutral vowel vocal tract shape (Li and Story, 2002). The neutral vowel shape was the mean shape of 10 English vowel shapes derived from the Magnetic Resonance Imaging (MRI), the same vowels used to derive the mode coefficients. Using an early version of the model, a series of 25 stop-neutral (/Cə/) syllables that differed only in the place of constriction at even spaces from the lip end to about 8 cm (section 20) above the glottis was constructed. The consonant constriction profiles represented by a Gaussian function were kept the same at all places of constriction. It was predicted that the continuum of place of constriction would result in the discontinuity of acoustic representations and therefore, categorical perceptions of the constriction continuum. The predicted perception of the consonant in each synthesized syllable was based on the formant transition directions. Perceptual tests were conducted with 10 normal native speakers of American English. The subjects had to identify the initial consonant of each /Cə/ syllable by clicking on one of the four choices /b/, /d/, /g/, and “unidentified” using a computer mouse. Figure 32 presents the response percentages of the four choices (y-axis) as a function of the place of constriction (x-axis) indicated by section numbers. The identification percentages were based on 100 responses (10 subjects $\times$ 10 repetitions =100). Consonant identification rates of /b, d, g/ are presented by the dashed curve marked with “+” signs, solid curve without marks, and dashed curve marked with circles respectively. The rates for the un-identified consonants are indicated by the dashed curve without marks.

It can be seen from Figure 32 that the identification functions for the three stop consonants show obvious peaks and valleys in the anterior portion of the vocal tract from the lip end at section 44 to section 29 (about 11 cm above the glottis). The vertical dashed lines mark the crossover points of the three consonant identification functions. From sections 41 to 44, the function of /b/ response shows a prominent peak (76%). The function of /g/ response shows a peak of 83% at section 33. The function of /d/ response shows the most prominent peak (99%) at the place of constriction of section 37 (14.7 cm above

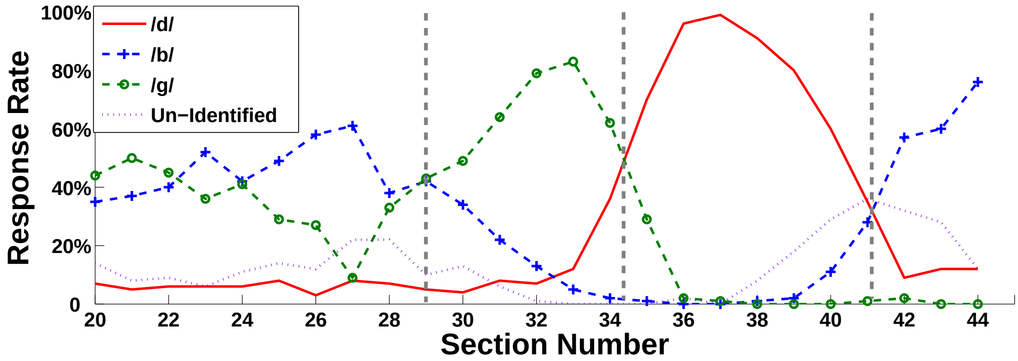


Figure 32: *Consonant identification functions for the constriction continuum in the neutral vowel context. Adapted from Figure 8 in Li and Story, 2002*

the glottis). Unambiguous /d/ identification was achieved when the consonant constriction was embedded in the neutral vowel vocal tract at this place of constriction. The q_1 and q_2 coefficients of modes 1 and 2 (-0.28 and 0.11 respectively) were kept constant at values very close to zero throughout the vowel portion of the syllable.

It has also been documented in some other studies that the synthesis of consonant /d/ in extreme vowel contexts /i/ and /ɑ/ are not as successful as in other vowel contexts. For example, Carr$\acute{e}$, Li$\acute{e}$nard, Marsico, and Serniclaes (2002) studied the consonant identifications of various formant transitions using the Distinctive Region Model (DRM). The model divides the vocal tract into eight regions with unequal lengths, where an expansion or constriction of the vocal tract area in one region results in a unique combination of the first three formant transition directions relative to the neutral vowel formants. For example, a constriction (area reduction) at the eighth region located at the lip end results in monotonic decreases in all of the first three formants from the neutral vowel formants. A constriction at the seventh region located posterior to the eighth region results in a falling F3, but rising F1 and F2. In their study, various formant transition directions in /əə/, /əɑ/, and /ɑɑ/ contexts were generated by superimposing the constrictions at three different regions that were considered as the places of articulation for the three voiced stop consonants. Perceptual tests were conducted to test the consonant identifications of the various formant transitions. The results showed that for context /əɑ/, when the vocal tract transition from the initial neutral vowel to the final vowel started at the same time as the superposition of the consonant gesture, /b/ and /g/ perception could be achieved, but not /d/ perceptions. When formant transitions started after the superposition of the consonant gesture (i.e., when the consonant constriction was overlapped with the initial neutral vowel shape only), instead of the vowel transitional shape, /d/ perceptions were obtained. Similarly, for context /ɑɑ/, a neutral shape had to be created during the consonant constriction superposition to yield the /d/ perception. They concluded that a neutral vowel shape was necessary to create a neutral "reference" environment for the synthesis of /d/. Consequently, they suggested that F2 and F3 transition directions relative to the neutral F2 and F3 around 1500 Hz and 2500 Hz should be used to categorize the perception of place of articulation. Their perceptual data confirmed their suggestions.

In order to search for the temporal relations between the vowel-to-vowel transitions and the stop constrictions, L$\ddot{o}$fqvist and Gracco (1999, 2002) studied the stop consonant production in V_1CV_2 sequences using a three-transmitter magnetometer system (the methods of the studies were reviewed in section 2.1). The middle consonant was one of the six stops /p, b, t, d, k, g/, and the vowel contexts were all combinations of /i, ɑ, u/. They reported that the velocity of the consonant constriction was heavily influenced by the first vowel in the VCV sequence. The velocity decreased in the order of /ɑ/ $>$ /u/ $>$/i/. This result also confirmed the finding by Gay (1977) that the onset of consonant closure was earlier when the initial vowel was /ɑ/ than when it was /i/ in V_1CV_2 sequences. L$\ddot{o}$fqvist and Gracco (1999, 2002) also found that the tongue movement during the production of a VCV sequence followed a loop-like curve, instead of a straight path which would be thought of as a more “efficient” strategy. When V_1 was either /u/ or /ɑ/, the tongue showed a forward movement during the consonant closure. When V_1 was /i/, the movement was backward. It was suggested that the time difference between the tongue movements and the consonant constriction had to be within a very short time window so that an extra vowel could not be perceived.

L$\ddot{o}$fqvist and Gracco (1999, 2002) studies have produced results that are open to interpretation. The vowel-dependent velocity and the looping pattern of tongue movements may suggest that the vocal tract approaches a neutral shape during the consonant constriction. Specifically, the degree of openness of the front cavity decreased for initial vowels /ɑ/, /u/, and /i/. If a neutral shape had to be achieved or approached during the consonant constriction, the degree that the front cavity of the vocal tract would have to be narrowed fell in the order of /ɑ/ $>$ /u/ $>$/i/. Therefore, the velocity of constriction had to fall in the same order, as suggested in L$\ddot{o}$fqvist and Gracco (1999, 2002). In addition, the backward movement of the tongue in $/VCV/$ sequences when the initial vowel was /i/ further expands the oral cavity during the consonant constriction when the cavity was very narrowed during the production of the initial vowel /i/. The forward movement of the tongue observed when the initial vowel was either /u/ or /ɑ/ might also have resulted from the neutralization of the front cavity.

Stevens and House (1963) showed the tendency of neutralization in his study of the effect of consonant perturbation on the vowel articulations in symmetrical $/CVC/$ sequences. The initial and final consonants included both voiced and voiceless fricatives and stops. The middle vowel was one of the 8 vowels (/i, ɪ, ɛ, æ, ɑ, ʌ, i, ʊ, u/). He reported that the consonant context always shifted the steady-state F2 of the middle vowel toward a more central position, compared to the situation when the vowel was produced in isolation. Specifically, when the vowel was a front vowel with high F2, such as /i/, the shift was downward. When the vowel was a back vowel with a low F2, such as /ɑ/, the shift was upward. These directions were also observed in the naturally spoken $/V\mathrm{d}V/$ sequences presented in Figure 28. Of all the consonant contexts, front vowels showed the greatest F2 shifts when the consonant contexts were labials and alveolars; vowel /u/ showed the greatest shifts when the consonant contexts were alveolars.

If the neutralization of the vocal tract is necessary for the synthesis and production of the stop consonants in certain vowel contexts, the temporal coordination between the execution of the consonant constriction and the V_1-to-neutral and neutral-to-V_2 has to be in such a fashion that no extra "neutral" vowel is perceived in the transitional process. Few studies have addressed the relationship between the temporal variations of the transitions in VCV sequences and their perceptual results. Among the few, Carr$\acute{e}$, Ainsworth, Jospa, Maeda, and Pasdeloup (2001) studied the perceptual effects of vowel-to-vowel transitions with different temporal characteristics. The vowel-to-vowel pairs of /iɑ/-/ɑi/, /iu/-/ui/, and /uɑ/-/ɑu/ with the transitional durations of 50 ms, 100 ms, 150 ms, 200 ms, 250 ms, and

300 ms were synthesized using the DRM. The steady-state durations of the initial and final vowels were kept at 100 ms and 150 ms respectively for all the sequences. The results of the listening tests showed that a transitional duration shorter than 150 - 200 ms was necessary to avoid the perception of extra vowels during the transition. A similar study by Carr$\acute{e}$, Chennoukh, Jospa, and Maeda (1996, in Carr$\acute{e}$, $et.$ $al.$ 2001) demonstrated that the synchronization between the vowel-to-vowel gesture and the consonant gesture in VCV sequences had to be within a certain temporal range to maintain the reliable VCV perception. More studies are needed in order to draw conclusions on the temporal characteristics of neutralization of the vocal tract configuration during the production of vowel-stop-vowel sequences.

If the neutralization of the vocal tract before or at the onset of the consonant constriction for the production of the alveolar consonant is necessary for some vowel contexts, the formant transitions at onglide and offglide positions should show a tendency of either approaching to or deviating from the neutral formants. Figure 33 plots the formant transitions of the naturally spoken $/V\mathrm{d}V/$ sequences along with the first three formants of the neutral vocal tract shape at 600 Hz, 1525 Hz, and 2500 Hz. The neutral formants were obtained from the neutral area function and spatial eigenmodes of the same speaker who produced the $/V\mathrm{d}V/$ sequences in this study. They are plotted as the three dashed horizontal lines in each panel of Figure 33.

It can be seen from Figure 33 that the F2 or F3 contours of $/V_1dV_2/$ sequences before and after the onset of the consonant constrictions either reached or approached the neutral F2 or F3 values. For example, at the onset of the alveolar constriction, the F2

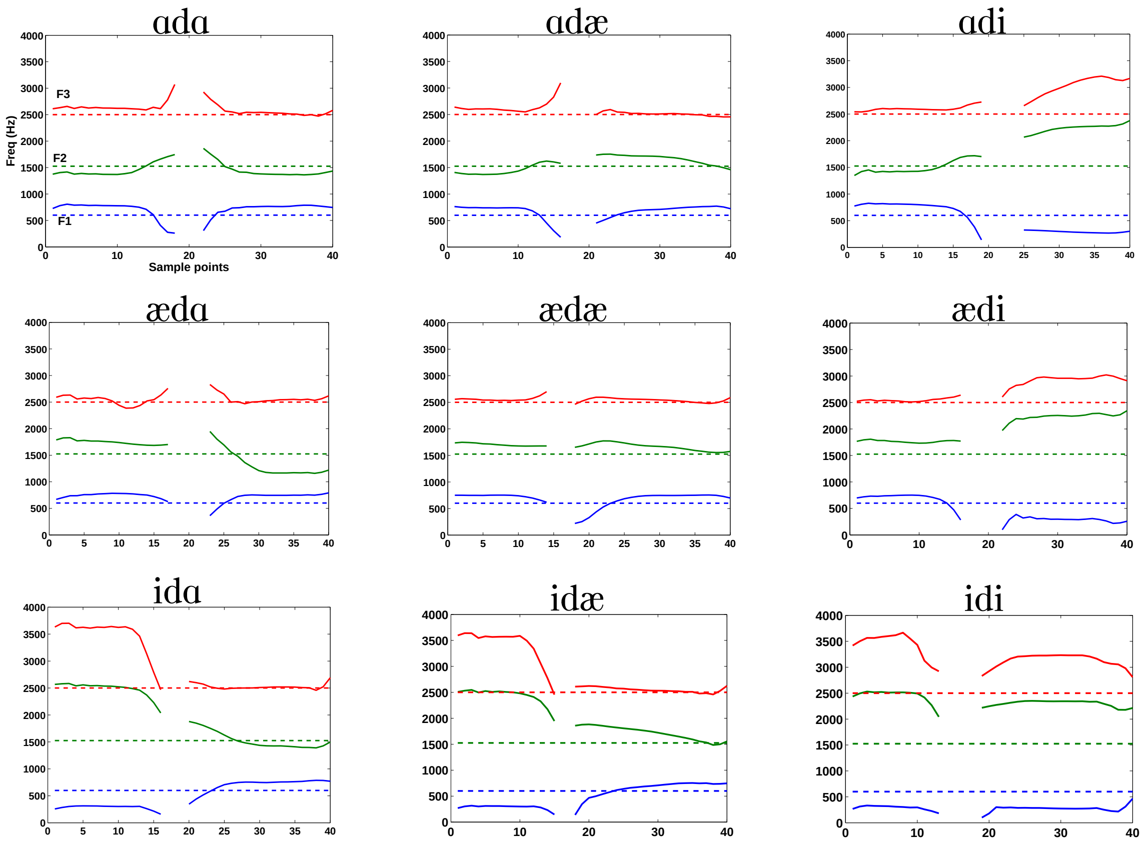


Figure 33: *Formant transitions of the nine spoken* $/V_1dV_2/$ *(solid lines) sequences and the neutral vowel formants (dotted lines).*

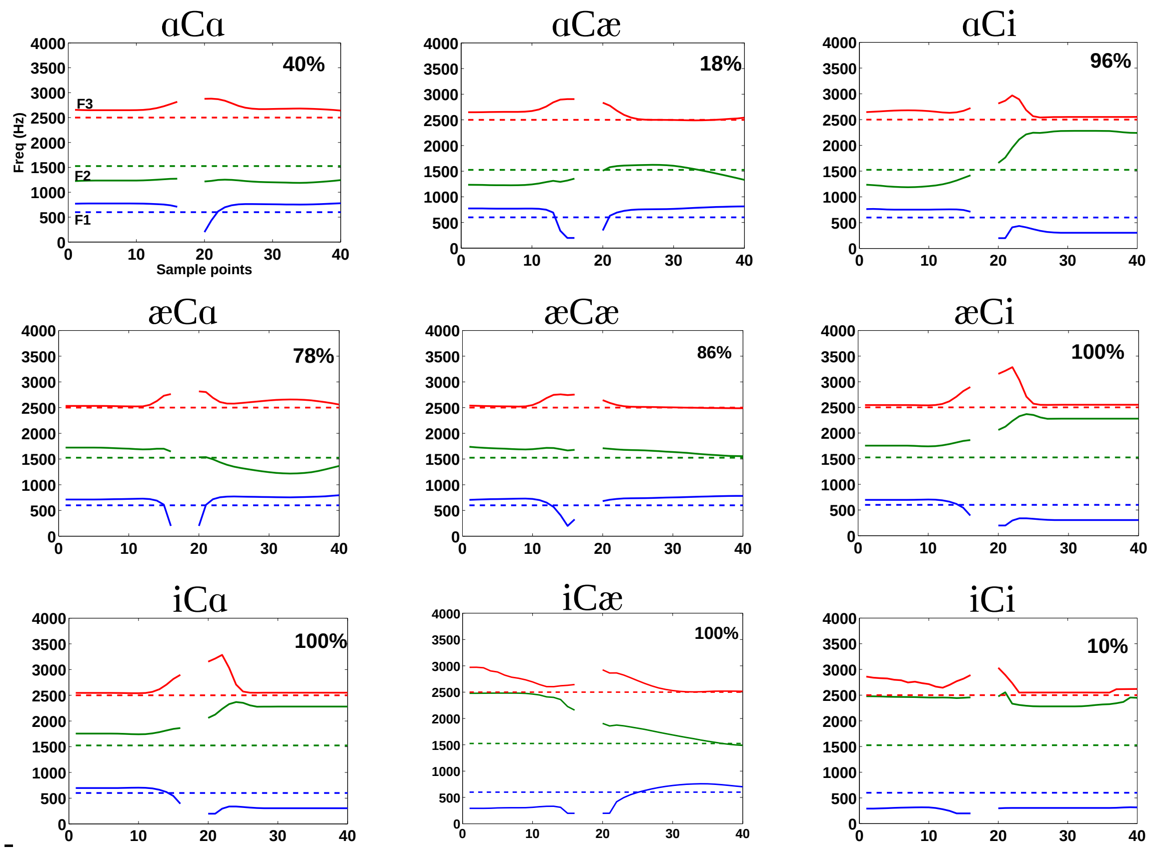


Figure 34: *Formant transitions of nine synthesized* $/V_1CV_2/$ *sequences (one sequence from each context) and the neutral vowel formants (dashed lines). Each sequence plotted had the highest /d/ response rate among all the samples in its vowel context. The response rate is indicated at the upper right corner of each panel.*

transition increased from 1367 Hz to 1742 Hz in /ɑdɑ/, from 1369 Hz to 1605 Hz in /ɑdæ/, and from 1429 Hz to 1712 Hz in /ɑdi/. The starting points of the F2 transitions in these /ɑdV/ sequences were all higher than an ideal F2 value of vowel /ɑ/ and all shifted toward the neutral F2 values. As the vocal tract was further constricted, F2 transitions reached the neutral F2 value of 1500 Hz. These observations suggest that the tongue tip or body configuration at the beginning of the /ɑdV/ sequences and at the onset of the consonant constriction might be perturbed in the direction of a neutral configuration. Another example is formant transitions of utterance /idi/ shown in Figure 32 in the lower right corner. Both F2 and F3 transitions showed sharp decreases approaching the neutral F2 and F3 values at 1500 Hz and 2500 Hz respectively. F2 decreased from 2492 Hz to 2042 Hz and F3 decreases from 3364 Hz to 2927 Hz. It is also observed that the rates of transitions in /idi/ were higher than those of /ɑdV/ sequences. In /idi/, the F2 transition showed a frequency range of 450 Hz from 2492 Hz to 2042 Hz in about 3 sampling points (roughly 30ms). In /ɑdɑ/, the F2 transition of 375 Hz from 1367 Hz to 1742 Hz occurred in about 7 sampling points (roughly 70 ms). The fast transition change in /idi/ might be caused by the high sensitivity of a formant to the dimensional changes of a narrow vocal tract since the vocal tract was much narrower in context /ii/ than in /ɑɑ/.

The formant transitions of the synthesized VCV sequences and the neutral formants were also examined. They are plotted in Figure 34. Each sequence in the figure had the highest /d/ response rate among all the samples in its vowel context (the same ones as in Figure 29). The response rate is indicated at the upper right corner of each panel. As can be seen, the deflection pattern of /iCi/ sequence, which had the lowest /d/ response rate, showed that F2 transition deviated from the neutral F2 nearly 1000 Hz. No obvious trend of F2 transition approaching the neutral F2 can be seen. Similarly, the F2 transition of /ɑɑ/, which also had a low /d/ response rate, did not show any obvious trend approaching the neutral vowel formant either.

It is speculative at this time that formant deflection patterns at the consonant constriction onset and offset relative to the neutral vowel formants may reflect the neutralization process of the vocal tract during the production of the alveolar stop /d/ in certain vowel contexts, especially extreme ones such as /ɑɑ/ and /ii/. From the perceptual point of view, the neutralization of the vocal tract shape at or before the onset of the stop constriction in extreme vowel contexts enlarges the distance between proximate vowel formants so that formant transitions may extend to a larger range. Such extension may enhance the decoding of transitional cues for consonant identification. Future studies on more speech samples and more speakers are needed to understand if the neutralization of the vocal tract consistently affect the production of stop consonants in different vowel contexts, how the temporal and spectral characteristics of the neutralization vary with vowel contexts, how these characteristics can be used to guide the model parameter setting in consonant synthesis, and if and how the periphery auditory system uses the neutral vowel context as a reference point to identify the adjacent consonant.

CONCLUSIONS

Formant transition patterns of $/VCV/$ sequences have been studied with much research effort in order to establish the connection between the “variant” transition patterns and the “invariant” consonant identification. The general conclusion of previous research is that the formant transition patterns for a consonant identification are vowel-dependent. The present study extended the results of previous research to analyze the formant deflections caused by the consonant constriction with respect to the underlying vowel-to-vowel formant variations. The objectives have been to describe the formant deflection directions of the synthesized $/V_1$d$V_2/$ sequences relative to the underlying $/V_1V_2/$ formant contours and to characterize the production of the voiced alveolar stop consonant in nine vowel contexts in terms of the model parameters. It was hypothesized that the formant deflection pattern of a falling F1, rising F2 and F3 relative to the underlying $/V_1V_2/$ formant contours would be perceived as the alveolar /d/ in the nine vowel contexts of interest. An articulatory model instead of a formant synthesizer was used to generate various consonant constriction profiles that were superimposed on the nine vowel-to-vowel vocal tract transitions. The formant contours of the synthesized $/V_1CV_2/$ were calculated and extracted. Perceptual tests were conducted to assess the consonant identifications of the synthesized $/V_1CV_2/$ sequences with the hypothesized deflection pattern. The formant deflection patterns of the spoken $/V_1dV_2/$ and $/V_1V_2/$ sequences were extracted and compared with the hypothesized pattern.

In conclusion, synthesized $/V_1CV_2/$ sequences specified by the formant transition directions relative to the underlying vowel-to-vowel transitions in the nine vowel contexts yielded statistically significant alveolar stop consonant responses. To characterize the spatial characteristics of the consonant constriction for the synthesis of the alveolar stop consonant in terms of the model parameters, correspondence analysis of the parameter settings of those VCV sequences of which the middle consonants yielded an above the change level /d/ responses was conducted. Three principle dimensions were extracted. The first dimension (accounts for 33.22% variance) described parameters that were related to the anterior-posterior geometry (on the mid-saggital profile of the vocal tract) of the consonant constrictions such as the place, range, and skewness of the constriction. The second dimension (accounts for 25.85% variance) was related to the vertical direction on the constriction profile, such as the magnitude and minimum area of the constriction. The third dimension (accounts for 20.58% variance) showed a mixed nature that both the anterior-posterior and the vertical geometries were represented.

The formant deflection patterns of the spoken $/V_1dV_2/$ sequences did not agree with the hypothesized deflection pattern. Comparison between the synthesized and spoken $/V_1dV_2/$ showed that a neutralization tendency in both synthesized and naturally spoken $/VCV/$ sequences with high level of /d/ response rates was observed. It is speculated that the neutral vowel shape at or before the onset of the consonant constriction may provide information for the identification of place of articulation for the alvelolar stop /d/ in some vowel contexts, especially extreme vowel contexts such as /ii/ and /ɑɑ/.

The study cannot be considered as a perceptual test of the hypothesized formant deflection pattern. After all, the underlying vowel-to-vowel transitions in the VV utterance is not available to the auditory system when a VCV sequence is perceived. Rather, the study meant to show if the formant deflection pattern of a phonetic segment could be derived from an articulatory gesture using an articulatory-based synthesizer, and if the formant deflection pattern must reflect the “layered” structure of the production mechanism, upon which the structure of the area function model was based. Such reflection would guide the synthesis of the consonant in different vowel-to-vowel transitions. Future studies should focus on the fine temporal and spectral aspects of the “layered” structure of the production mechanism such as the timing of the vocal tract neutralization at the onset and offset of the alveolar constriction in different vowel contexts.

REFERENCES

American Speech Language Hearing Association Panel on Audiologic Assessment (1997). Guidelines for Audiological Screening. Rockville, MD.

Båvegård, M. (1995). Introducing a parametric consonantal model to the articulatory speech synthesizer, *Proceedings Eurospeech 95, Madrid, Spain.*

Blumstein, S. E., & Stevens, K. N. (1979). Acoustic invariance in speech production: evidence from measurements of the spectral characteristics of stop consonants. *Journal of the Acoustical Society of America, 66,* 1001-1017.

Blumstein, S. E., & Stevens, K. N. (1980). Perceptual invariance and onset spectra for stop consonants in different vowel environments. *Journal of the Acoustical Society of America, 67(2),* 648-662.

Browman, C., & Goldstein, L. (1987). Tiers in articulatory phonology, with some implications for casual speech, *Haskins Laboratories Status Report on Speech Research, SR-92,* 1-30. [Also in *Papers in Laboratory Phonology I: Between the grammar and the physics of speech.* Kingston, J., & Beckman, M. E. editors, pp341-376. Cambridge: Cambridge university Press.]

Browman, C., & Goldstein, L. (1990). Gestural specification using dynamically-defined articulatory structures, *J. Phonetics, 18,* 299-320.

Carré, R., Ainsworth, W. A., Jospa, P., Maeda, S., & Pasdeloup, V. (2001). Perception of vowel-to-vowel transitions with different formant trjectories. *Phonetica, 58,* 163-178.

Carré, R. & Chennoukh, S. (1995). Vowel-consonant-vowel modeling by superposition of consonant closure on vowel-to-vowel gestures, *J. Phonetics, 23,* 231-241.

Carré, R., Liénard, J. S., Marsico, E., & Serniclaes. W. (2002). On the role of the "schwa" in the perception of plosive consonants. *7th International Conference on Spoken Language Processing, September 16-20, Denver, Colorado,*1681-1684.

Cooper, F. S., Delattre, P. C., Liberman, A. M., Borst, J. M., & Gerstman, L. J. (1952). Some experiments on the perception of synthetic speech sounds. *Journal of the Acoustical Society of America, 24,* 597-606.

Delattre, P. C., LibermanA. M., & Cooper, F. S. (1955). Acoustic loci and transitional cues for consonants. *Journal of the Acoustical Society of America, 27(4),* 769-773.

Dorman, M. F., Studdert-Kennedy, M., & Raphael, L. J. (1977). Stop-consonant recognition: release bursts and formant transitions as functionally equivalent, context-dependent cues. *Percept. Psychophys, 22,* 109-122.

Dorman, M. F., & Loizou (1996). Relative spectral change and formant transitions as cues to labial and alveolar place of articulation. *Journal of the Acoustical Society of America, 100 (6),* 3825-3830.

Ericsdotter, C., Lindblom, B., & Stark, J.(1999). Articulatory coordination in coronal stops: Implications for theories of coarticulation. *Proceedings of the XIVth International Congress of Phonetic Sciences (ICPhS 99), San Francisco, August, 1999,* 1885-1888.

Fant, G. (1960). *Acoustic theory of speech production.* The Hague, Netherlands: Mouton.

Fant, G. (1969). Stops in CV-syllables. *STL-Quarterly Progress and Status Report, 4,* 1-25.

Fant, G. (1973). *Speech Sounds and Features.* The MIT Press. Cambridge, Mass.

Fant, G. (1980). The relations between area function and the acoustic signal. *Phonetica,* 55-86.

Fant, G., & Pauli, S. (1974). Spatial characteristics of vocal tract resonance modes. *Proceedings of Speech Communication Seminar, Stockholm,* 121-131.

Fitch, W. T., & Giedd, J. (1999). Morphology and development of the human vocal tract: A study using magnetic resonance imaging. *Journal of the Acoustical Society of America, 106,* 1511-1522.

Fujimura, O. (1961). Bilabial stop and nasal consonants: A motion picture study and its acoustical implications. *Journal of Speech and Hearing Research, 4,* 233-246.

Fujimura, O. (1992). Phonology and phonetics - A syllable based model of articulatory organization. *Journal of Acoust. Soc. Jpn., 13,* 39-48.

Glasberg, B. R., & Moore, B. C. J. (1990). Deviation of auditory filter shapes from notched-noise data. *Hearing Research, 47,* 103-138.

Hair, J. F., Anderson, R. E., Tatham, R. L., & Black, W. C. (1995). *Multivariate Data Analysis.* (4th ed.) Upper Saddle River, NJ: Prentice Hall.

Härdle, W., & Simar, L. (2003). *Applied Multivariate Statistical Analysis*. Springer: Verlag Berlin Heidelberg New York.

Hogan, N. (1984). An organizing principle for a class of voluntary movements. *The Journal of Neuroscience, 4(11),* 2745-2754.

Kent, R. D., & Moll, K. L. (1972). Cineflurographic analyses of selected lingual consonants. *Journal of Speech and Hearing Research, 15,* 453-473.

Kent, R. D. & Minifie, F. D. (1977). Coarticulation in recent speech production models. *Journal of Phonetics, 5,* 115-133.

Kent, R. D., Kent, J. F., Weismer, G., Martin, R. Sufit, R. L., Brooks, B., & Rosenbek, J. C. (1989). Relationships between speech intelligibility and the slope of the second formant trnasitions in dysarthric subjects. *Clinical Linguistics and Phonetics, 3,* 347-358.

Kewley-Port, D. (1982). Measurement of formant transitions in naturally produced stop consonant-vowel syllables. *Journal of the Acoustical Society of America, (72),* 379-389.

Kewley-Port, D. (1983). Time varying features as correlates of place of articulation in stop consonants. *Journal of the Acoustical Society of America, (73),* 322-335.

Kewley-Port, D., Pisoni, D. B., & Studdert-Kennedy, M. (1983). Perception of static and dynamic acoustic cues to place of articulation in initial stop consonants. *Journal of the Acoustical Society of America, 73,* 1779-1793.

Krishnamurthy, A. K. (1984). Two channel (speech and EGG) analysis for formant tracking and glottal inverse filtering. *Proc. IEEE Int. Conf. Acoust. Speech Signal Proc, 36,* 36.6.1-36.6.4

Lahiri, A., Gewirth, L., & Blumstein, S. E. (1984). A reconsideration of acoustic invariance for place of articulation in diffuse stop consonants: Evidence from a cross-language study, *Journal of the Acoustical Society of America, 76 (2),* 391-404.

Lehiste, I., & Peterson, G. (1961). Transitions, glides, and diphthongs. *Journal of the Acoustical Society of America, 33,* 268-277.

Li, K., & Story, B. (2002). Unpublished Master's thesis, An investigation of perceptual tolerance limits of stop constriction regions along the neutral vocal tract.

Li, K., & Story, B. (2003). An investigation of perceptual tolerance limits of stop constriction regions along the vocal tract. Presented at the 146th Acoustical Society Meeting, J. Acoust. Soc. Am., 114 (4). Pt. 2, 2337, (Austin, Texas).

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psych. Rev., 74(6),* 431-461.

Lijencrants, J. (1985). Seech synthesis with a reflection-type line analog. DS Dissertation, Dept. of Speech Comm. and Music Acoust., Royal Inst. of Tech., Stockholm, Sweden.

Löfqvist, A & Gracco, V. (1999). Interarticulator programming in VCV sequences: lip and tongue movements. *Journal of the Acoustical Society of America, 105(3),* 1864-1876.

Löfqvist, A & Gracco, V. (2002). Control of oral closure in lingual stop consonant production. *Journal of the Acoustical Society of America, 111(6),* 2811-2827.

MacNeilage, P. F. (1970). Motor control of the series ordering of speech. *Psychology Review, 77,* 182-196.

Mooshammer, C., Hoole, P., & Kühnert, B. (1995). On loops. *Journal of Phonetics, 23,* 3-21.

Mrayati, M., Carré, R., & Guerin, B. (1988). Distinctive regions and modes: a new theory of speech production. *Speech Communication, 7,* 257-286.

Nelson, W. L., Perkell, J. S., & Westbury, J. R. (1984). Mandible movements during increasingly rapid articulations of single syllables: Preliminary observations. *Journal of the Acoustical Society of America, 75,* 945-951.

Öhman, S.E.G.(1966). Coarticulation in VCV utterances: Spectrographic measurements. *Journal of the Acoustical Society of America, 39,* 151-168.

Öhman, S.E.G.(1967). Numerical model of coarticulation. *Journal of Acoustical Society of America, 41(2),* 310-320.

Parikh, G., & Loizou, P. C. (2005). The influence of noise on vowel and consonant cues. *Journal of Acoustical Society of America, 118(6),* 3874-3888.

Perkell, J. S. (1969). *Physiology of Speech Production: Results and Implications of a Quantitative Cineradiographic Study.* Cambridge, MA: The MIT Press.

Rabiner, L. R., Atal, B. S., & Sambur, M. R. (1977). LPC prediction error— analysis of its variation with the position of the analysis frame. *IEEE Transactions on Acoustics, Speech, and Signal Processing, 25(5),* 434-442.

Rosenberg, A. E. (1971). The effect of glottal pulse shape on the quality of natural vowels. *Journal of the Acoustical Society of America, 49,* 583-590.

Shadle, C. H., & Damper, R. I. (2002) Prospects for articulatory synthesis: A position paper. *Proceedings of 4th ISCA Workshop on Speech Synthesis, Pitlochry, Scotland, August/September 2001,* 121-126.

Shavelson, R. J. (1988). *Statistical Reasoning for the Behavioral Sciences.* (2nd ed.). Boston: Allyn and Bacon.

Schatz, C. D. (1954). The role of context in the perception of stops. *Language, 30,* 47-56.

Stevens, K. N., House, A. S. (1963). Perturbation of vowel articulations by consonantal context: An acoustical study. *Journal of Speech and Hearing Research, 6,* 111-128.

Stevens, K. N. (1972). The quantal nature of speech: Evidence from articulatory-acoustic data. In *Human communication: a unified view,* E.E. David, Jr. & P. B. Denes, Eds., 51-66. McGraw-Hill, New York.

Stevens, K. N. (1989). On the quantal nature of speech. *Journal of Phonetics, 17,* 3-45.

Stevens, K. N. (1998). *Acoustic Phonetics*. Cambridge, MA: The MIT Press.

Stevens, K. N. & Blumstein, S. E. (1978). Invariant cues for place of articulation in stop consonants *Journal of the Acoustical Society of America, 64(5),* 1358-1368.

Story, B. H. (2002). A parametric area function model of three female vocal tracts based on orthogonal modes. *Presented at the Acoustical Society Meeting, Journal of the Acoustical Society of America, 112 (5),* pt 2, 2481.

Story, B. H. (2005). A parametric model of the vocal tract area function for vowel and consonant simulation. *Journal of the Acoustical Society of America, 117 (5),* 3231-3254.

Story, B. H. (2005). Research grant, Quantitative modeling of speech, R01-DC04789.

Story, B. H. (1995). Physiologically-based speech simulation using an enhanced wave-reflection model of the vocal tract, Ph.D. Dissertation, University of Iowa.

Story, B. H., & Titze, I. R. (1998). Parameterization of vocal tract area functions by empirical orthogonal modes. *Journal of Phonetics, 26,* 223-260.

Story, B. H., & Titze, I. R. (2002). A preliminary study of voice quality transformation based on modifications to the neutral vocal tract area function. *Journal of Phonetics, 30,* 485-509.

Story, B. H., Titze, I. R., & Hoffman, E. A. (1998). Vocal tract area functions for an adult female speaker based on volumetric imaging. *Journal of the Acoustical Society of America, 104,* 471-487.

Story, B. H., Titze, I. R., & Hoffman, E. A. (2001). The relationship of vocal tract shape to three voice qualities. *Journal of the Acoustical Society of America, 109,* 1651-1667.

Story, B. H., Titze, I. R., & Hoffman, E. A. (1996). Vocal tract area functions from magnetic resonance imaging. *Journal of the Acoustical Society of America, 100,* 527-554.

Sussman, H. M., MacNeilage, P. F., & Hanson, R. J. (1973). Labial and mandibular dynamics during the production of bilabial consonants: Preliminary observations. *Journal of Speech and Hearing Research, 16,* 385-396.

Titze, I.R, (2002). Regulating glottal airflow in phonation: Application of the maximum power transfer theorem to a low dimensional phonation model. *Journal of Speech and Hearing Research, 111,* 367-376.

Tjaden, K. & Weismer, G. (1998). Speaking-rate-induced variability in F2 trajectories. *Journal of Speech, Language, and Hearing Research, 41,* 976-989.

Weismer, G., Kent, R. D., Hodge, M., & Martin, R. (1988). The acoustic signature for intelligibility test words. *Journal of the Acoustical Society of America, 84,* 1281-1291.

Zwicker, E., Flottorp, G., & Stevens, S. (1957). Critical bandwidth in loudness summation. *Journal of the Acoustical Society of America, 29,* 548-557.