Perception of articulatory dynamics from acoustic signatures

Khalil Iskarous,[a] Hosung Nam, and D. H. Whalen
Haskins Laboratories, 300 George Street, Suite 900, New Haven, Connecticut 06511

(Received 5 October 2009; revised 2 April 2010; accepted 2 April 2010)

This study investigated the degree to which the articulatory trajectory of the tongue dorsum in the production of a vowel-vowel sequence is perceptually relevant. Previous research has shown that the tongue dorsum takes a path that leads to a pattern of area function change, termed the pivot pattern. In this study, articulatory synthesis was used to generate paths of tongue motion for the production of the vowel sequence /ai/. These paths differed in their curvature, leading to stimuli that conform to the pivot pattern and stimuli that violate it. Participants gave naturalness ratings and discriminated the stimuli. The acoustic properties were also compared to acoustic measurements made on productions of /ai/ by 34 speakers. The curvature of the tongue path and the curvature of the F1-F2 trajectory correlate highly with the naturalness-rating task results, but not the discrimination results. However, the particular way in which constriction location changes, particularly whether the change is discrete or continuous, and the maximal velocity of F2 through the transition, explain the perceptual patterns evident in both perception tasks, as well as the patterns in the observed acoustic data. Consequences of these results for the links between production and perception and the segmentation problem are discussed.
 [DOI: 10.1121/1.3409485]

PACS number(s): 43.70.Mn, 43.71.Es [DAB] Pages: 3717–3728

I. INTRODUCTION

When two lingual segments are produced in sequence, e.g., /ai/, the motion of the tongue dorsum through the transition is simultaneously influenced by both segments. Based on an investigation of a large number of VV, CV, CC, and VC transitions, Iskarous (2005) proposed that the tongue trajectory and consequent area function change between two lingual segments is highly systematic, regardless of the phonemic identity of the particular C or V involved. Specifically, the tongue moves maximally at the two constrictions and minimally in between, forming a virtual pivot for the tongue. Even though there is no single flesh point of the tongue that maintains a constant distance from the hard structures, a single location along the hard structures nonetheless does maintain a constant value for the distance from the tongue. The purpose of the current work is to investigate if the specific way in which the tongue moves in a transition between two lingual segments is perceptually relevant. This was accomplished by using an articulatory synthesizer to generate several physiologically possible tongue dorsum trajectories from /a/ to /i/. These trajectories model the possible articulatory dynamic within the limits of the tongue shapes generated by the synthesizer. The trajectories also lead to different ways in which the area function changes as a function of time during /ai/. Several acoustic properties of the resulting F1 and F2 trajectories were then investigated to determine if there is a particular acoustic signature of the articulatory dynamic. The goal of this paper is to investigate which of several articulatory and acoustic descriptions of the /ai/ dynamic explain the perceptual patterns exhibited by participants who listened to the stimuli in naturalness rating and discrimination tasks. Each of the articulatory and acoustic measures used to parameterize the trajectories imposes a similarity metric on the stimuli used. These various metrics were compared to the similarity metrics evident in the perception experiment results to investigate which articulatory or acoustic measure best predicts the perceptual patterns.

In a sequence of two lingual segments, each of the segments is linguistically specified for a place of articulation and a degree of constriction at that place (Wood, 1979). At each point in time during the transition from the first segment to the second, the area function at that point will have a location where the area function is a minimum. The location of that minimum is termed the constriction location (CL), which varies in time during the transition. Also, the degree to which the area function is constricted at that location and that point in time is termed the constriction degree (CD). Since speech is a dynamic phenomenon, two questions arise: how do CL and CD change and how does the tongue bring about these changes? Evidence was presented by Iskarous (2005) that there are two main patterns of tongue motion, the pivot and the arch. In both patterns, the area function changes maximally at the locations of constriction of the two segments in a transition and minimally elsewhere. If the two places of articulation in a transition are sufficiently separated spatially, then there is a region between the two primary constriction locations that experiences little to no change in the distance between the tongue and the fixed structures of the vocal tract—a transition type termed the pivot pattern, based on earlier work by Stone (1991). Figure 1 shows a transition from /a/ to /i/ from a male speaker of Canadian French. The pivot point is highlighted in the middle of the vocal tract for this transition, created as a function of the constriction locations and degrees of the two seg-

[a]Author to whom correspondence should be addressed. Electronic mail: iskarous@haskins.yale.edu

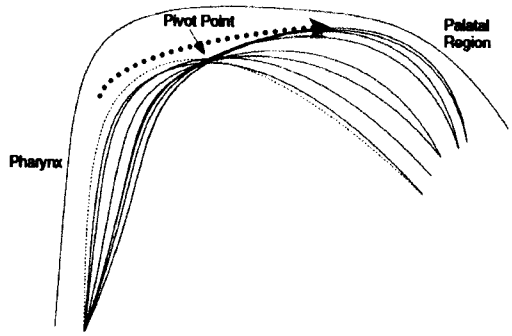

FIG. 1. The transition from /a/ to /i/. The superimposed dotted arrow shows how the constriction would move if its constriction location moved continuously.

ments in the transition. Also superimposed on the transition is a dotted arrow showing the path the tongue would have taken if the constriction were to have moved continuously, which Iskarous (2005) argued does not occur. Constrictions, it was shown, do not move from one location to another, rather formation of the second constriction is synchronous with the release of the first.

The term "pivot" does not refer to mechanical pivoting, where a point or region of the tongue does not move, as might have been due, for instance, to jaw rotation. It also does not refer to a situation where certain fleshpoints are stationary throughout a transition. In the pivot pattern, all points of the dorsum are moving relative to the fixed structures of the vocal tract, and there is no fixed fleshpoint. Rather, the pivoting is exhibited in the lack of change of area function at a point in the vocal tract, even though fleshpoints are moving through that point. Therefore the pivoting is functional, showing a stationarity in the area function, not in the motion of any particular fleshpoint. To accomplish such area function change, the tongue moves orthogonally to the fixed structures for points on the dorsum in the constriction locations (e.g., pharyngeal and palatal locations for the /ai/ transition), and moves parallel to the fixed structures in the area between the primary constriction locations of the two segments in the transition. Motion orthogonal to the fixed structures maximizes area function change, while motion parallel to the fixed structures minimizes it. The area function change resulting from the pivot pattern effectively discretizes the vocal tract into two locations, where CD changes maximally, separated by a region of little to no change. These empirical results are consistent with the Distinctive Regions Model (Mrayati *et al.*, 1988; Carré and Chennoukh, 1995), which predicts that area function change between two lingual segments occurs as two simultaneous acts, release of the first constriction and formation of the second constriction, at two discrete locations. This pattern is termed the transversal pattern. It was also shown by Iskarous (2005) that factor analytic approaches to the simulation of the area function change (Story, 2005) predict pivoting as the linear transition between targets in factor space. There is therefore robust empirical and theoretical evidence for pivoting in speech production.

In this paper, the main question is whether the pivot pattern of area function change is perceptually relevant. That is, if the tongue dorsum trajectory did not move orthogonally to the fixed structures at the constriction, or if area function were to change in a way that does not maintain the discreteness of CL, would listeners notice? Conversely, do they prefer the pivot pattern, which is what they typically hear? The answers to these questions are important for speech production and perception research, since an affirmative answer to the first would imply that this pattern is potentially important in describing speech dynamics. Also an answer to this question is also potentially important for understanding the extent of the dependence of the speech perception system on the speech production system. To test the hypothesis of whether listeners are aware of the coarticulatory dynamics of CL and CD, it is necessary to investigate the possible motions of the tongue dorsum in a sequence like that from /a/ to /i/, and determine if the pivoted coarticulatory dynamics are acoustically and perceptually different from those of the nonpivoted ones.

Articulatory synthesis is an ideal tool to use to investigate possible articulatory trajectories, and their constriction, acoustic, and perceptual consequences (Rubin *et al.*, 1981). For the present study, the configurable articulatory synthesizer (CASY) (Rubin *et al.*, 1996) was used. The synthesized /ai/ trajectories began and ended at the constriction location and degree appropriate for /a/ and /i/, respectively, but they varied from each other in the trajectory of the center of the tongue. Using a numerical index that quantifies pivoting, discussed in Sec. II A, it was determined that a linear trajectory of the tongue body corresponds to the "pivot" pattern discussed earlier. This pattern is the one most attested for a transition like /ai/ in the empirical study by Iskarous (2005), in which 600 lingual transitions from Canadian English and French were analyzed for the pattern of change. The convex trajectories (where convex refers to a motion in /ai/ that is first upward, then forward in the vocal tract), and concave trajectories (first forward, then upwards motion), on the other hand, diverge from pivoting. One of the hypotheses pursued in the current work is that the sound corresponding to the linear trajectory would be perceived as the most natural, because it is the most attested in production. The synthesized sequences made it possible to investigate the relation among the tongue trajectory, consequent patterns of change in CL and CD, properties of formant patterns, and perception of the coarticulated /ai/ through controlled variation of the tongue trajectory.

There has been previous work on the perception of the pivoting/transversal pattern within the theory of distinctive regions model (Mrayati *et al.*, 1988). In this theory, the vocal tract is assumed to be discretized into several regions, based on acoustic arguments, and time-varying change in lingual transitions is argued to take place as simultaneous changes in the degree of constriction at discrete locations. To examine whether listeners are aware of how the area function changes, Carré *et al.* (2001) generated two different realizations of /ai/ in the area function domain. In one, the constriction moved continuously from its /a/ location to its /i/ location, whereas in the other, the change at the two constriction

locations is simultaneous. That is in the first transition (longitudinal), constriction location changes continuously, whereas in the second (transversal) constriction location is discrete. The longitudinal transition would correspond in the experiment presented in this paper to a highly convex trajectory, whereas the transverse transition would correspond to either a linear or concave trajectory. The two resulting acoustic patterns were played to native French listeners, who preferred the transversal /ai/ to the longitudinal one. However, listeners in that experiment reported hearing an extra vowel between /a/ and /i/ in the longitudinal transition, which could be why they preferred the transversal transition.

There are several differences between the experiment reported on here and the Carré *et al.* (2001) experiment intended to further investigate the relation between articulatory, acoustic, and perceptual parameters. First, in the current experiment, the synthesis is performed in the articulatory domain in terms of the motion of the dorsum of the tongue, rather than the area function domain. Synthesis in the articulatory domain and investigation of consequent changes in the area function domain allow for the investigation of whether the perception system is tuned to the tongue trajectory or the area function change. This distinction may seem unnecessary, since it may be expected that the tongue trajectory and area-function change are roughly the same variables. However, as will be shown in Sec. II, these variables impose different similarity metrics upon the stimuli in the experiment. Second, instead of using two synthesis steps, either longitudinal or transversal, the curvature of the tongue trajectory varied incrementally, in 13 steps (which will be described in Sec. II), to determine at what stage the preference changes. Third, a discrimination task was included to reveal within-preference distinctions. Fourth, stimuli that sounded to pilot listeners to have an extra vowel between /a/ and /i/, as in the experiment by Carré *et al.* (2001), were excluded (Steps 12 and 13), so that the results are not dependent on the hearing of the extra vowel. Fifth, more articulatory and acoustic parameters of the transition from /a/ to /i/ are investigated in this work than were examined by Carré *et al.* (2001) and Ainsworth and Carré (1997) to determine which of the parameters best explains the perceptual results.

II. ARTICULATORY-ACOUSTIC RELATIONS IN /ai/

A. Relation between tongue dorsum trajectory and constriction dynamics

The Haskins program CASY was used to construct the articulatory trajectories (Rubin *et al.*, 1996; Iskarous *et al.*, 2003). CASY provides a geometric representation of articulators as geometric approximations to the major speech organs: lips, jaw, tongue body, tongue tip, velum, and hyoid bone—in the midsagittal plane, and is based on the earlier articulatory synthesizer (ASY) (Mermelstein, 1973; Rubin *et al.*, 1981). The tongue dorsum is modeled as the arc of a circle, based on the work of Coker and Fujimura (1966), and the tongue tip is represented by another circular arc attached to the dorsum circle. In this study, only the location of the center of the tongue body circle was manipulated, while the radius of tongue body circle was fixed at 20.5 mm. The path

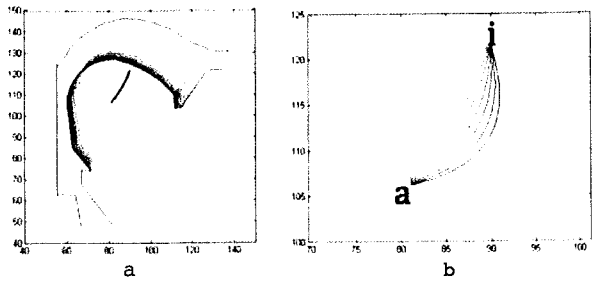

FIG. 2. (a) An example trajectory of the tongue body circle from /a/ to /i/. Superimposed on the figure is a b-spline of the path of the center of the tongue body circle. The gray scale denotes time, with the first tongue trajectory denoting /a/ in black and the last one denoting /i/ in light gray. (b) 13 trajectories of the center of the tongue body. Each V-V transition lasted 350 ms. The most concave is in black and the most convex is in lightest gray. The particular trajectory in (a) is for Step 6.

of the tongue circle center traces a trajectory generated as a cubic b-spline (smooth curve) of a particular curvature. For each transition, the center passes through a particular b-spline path. The paths differed from each other in the amount of curvature. A polar-rectangular grid was then overlaid on the vocal tract, and the area function was calculated for each frame in the sequence. Based on this area function, formants were calculated based on a simple acoustic model, where losses are calculated by a simple formula (Rubin *et al.*, 1981). F3 was held constant at 2500 Hz and F4 was held constant at 3500. Waveforms were then generated from formant values computed by CASY using HLSYN™, a parametric quasiarticulatory synthesizer (Sensimetrics Inc., Malden, MA) (Hanson and Stevens, 2002) which has high quality speech output. The acoustic structure of the stimuli will be discussed in detail in Sec. II B.

The exclusion of F3 is phonetically justified by examination of the amount that formants changes in vowel sequences. In an examination of articulatory-acoustic relations in vowel sequences, Simpson (2002) provides an example of the vocalic portion of the phrase American English "they all" (Fig. 5 in that paper). In that transition, F2 changes by about 4–5 barks, whereas F3 changes by about half a bark or less. Also, Simpson (2001) provides an example of the diphthong /ai/ in American English "light" (Fig. 7 in that paper), where F2 and F3 change by about the same magnitudes in barks as in the example from Simpson (2002). Such small changes in F3 are small from an auditory perspective (Rosner and Pickering, 1994) and would affect the measures used in this work to a minor degree. Therefore, to minimize the less accurate synthesis of F3, a single value for this formant was used.

Figure 2(a) shows an example trajectory of the tongue from /a/ to /i/. The path of the center of the tongue body is shown along with the tongue trajectory from /a/ to /i/. According to the pivoting hypothesis, the tongue moves orthogonally to the fixed structures at the targeted constriction locations and parallel to the fixed structures away from these constriction locations. Therefore the main variable controlled for generating different tongue trajectories from /a/ to /i/ was the curvature of the tongue trajectory. Figure 2(b) shows the 13 trajectories of the tongue body circle center used. At Step 1 (black), the trajectory is highly concave. At the beginning

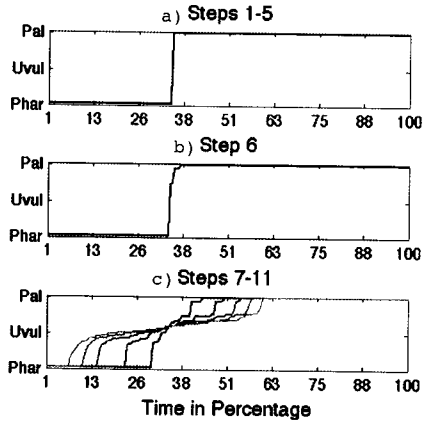

FIG. 3. Location of point of minimum constriction in the vocal tract as a function of time (in percent). Upper panel: Steps 1–5, all of which have identical placement of CL as a function of time. Middle panel: Step 6. Lower panel: Step 7 (darkest gray) to Step 11 (lightest gray).

of the transition, close to /a/, the tongue moves orthogonally to the posterior pharyngeal wall, while at the end of the transition, close to /i/, the tongue moves orthogonally to the palate. At the other end of the scale, Step 13 (lightest gray), the tongue moves parallel to the fixed structures from the beginning to the end of the transition. The intermediate trajectories (in intermediate grays) vary in the amount of curvature of the trajectory. The first panel of Fig. 4 shows how the curvature of the tongue trajectory varies as a function of step. Curvature was measured by fitting a three-point circle to the trajectory and calculating the radius r of the circle, which was measured as positive for concave curvature and negative for convex. Curvature was then calculated as $k = 1/r$.

To investigate the relation between this level of description of coarticulation to that of area function change, specifically CL and CD change, area functions were measured based on a polar-rectangular grid superimposed on the vocal tract for each step of each transition. Since the /ai/ in Steps 12 and 13 sounded to pilot participants as if there were an extra vowel or consonant between the /a/ and the /i/, these two steps were eliminated. Only the first 11 of the 13 steps were subsequently analyzed. To investigate how CL changes as a function of time for each transition, the minimum of the area function from 2 cm above the glottis to 2 cm from the lips was automatically extracted from each area function at each frame of each transition. The location of this minimum was then taken as CL. Figure 3 shows for each transition, how CL varied as a function of time through the transition. As can be seen in Fig. 3, for transitions 1–5, CL switches discretely within the transition from pharyngeal to palatal. For step 6 there is also switching, but the first palatal location is slightly posterior to the main palatal area, however, there is no CL in the uvular area. From Step 7 to Step 11, there are more and more intermediate CLs traversed in the path from pharyngeal to palatal.

To quantify the extent to which CL change is discretely switched, as opposed to continuously changed, the duration for which CL is *not* at the pharyngeal or palatal place was measured for each transition. This duration would be 0 if CL

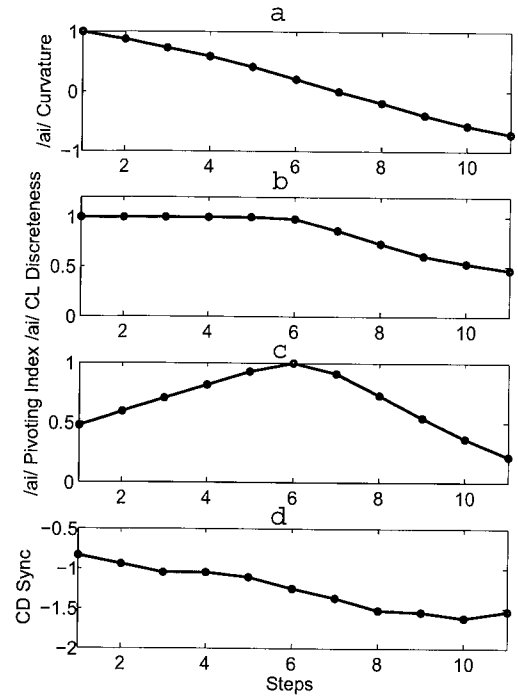

FIG. 4. First panel: Curvature of tongue trajectory as a function of step. Second panel: CL discreteness as a function of step. Third panel: Pivoting index as a function of step. Fourth panel: Degree of synchronization of palatal constriction formation and pharyngeal constriction release.

is switched from pharyngeal to palatal at one point in time, indicating a high value for discreteness. On the other hand, if the duration is long, it is an indication that CL's other than pharyngeal and palatal are being traversed. That duration is then normalized through division by the total duration of the transition (350 ms). This measure is 0 for steps 1 through 5 and becomes greater for subsequent steps. To further normalize the discreteness measure from 0 to 1, with 0 referring to nondiscrete and 1 to fully discrete, the normalized duration of the portion of the transition where CL is not palatal or pharyngeal was subtracted from 1:

$$\text{CL Disc.} = 1 - \frac{\text{Dur(CL not in Palatal or Pharyngeal)}}{\text{Dur(Transition)}}. \quad (1)$$

The second panel of Fig. 4 shows CL discreteness for each step. It can be seen that even though the curvature of the tongue dorsum trajectory changes nearly linearly from concave (positive curvature) to convex (negative curvature), the CL change is highly nonlinear. The first 6 steps show nearly discrete patterns, while the remaining steps show more and more continuous change in CL. Another way of stating the difference between the two variables is that each of them imposes a different similarity metric on the steps in the scale. For instance, the curvature difference between Steps 1 and 6 is the same as the curvature difference between 6 and 11, but the CL Discreteness of Step 6 is far more similar to Step 1 than it is to Step 11, as can be seen in Fig. 4. Several more articulatory and acoustic parameters of the trajectories will

also be investigated to elucidate how these parameters differentiate between the possible /ai/ trajectories.

Iskarous (2005) proposed an index to quantify pivoting. The standard deviations of the change in area function at the two constriction locations through the transition are measured and averaged and divided by the standard deviation of the change at the point in the vocal tract between the two locations for the /ai/ case:

$$\text{Pivoting} = 1 - \frac{\frac{\sigma(\text{Palatal AF}) + \sigma(\text{Pharyngeal AF})}{2}}{\sigma(\text{Uvular AF})}, \quad (2)$$

where σ stands for the standard deviation. This index quantifies how much change in the area function there is at different locations in the vocal tract. It is small when there is a great deal of change in the pivot region (uvular in this case) and is large when there is hardly any change. The third panel of Fig. 4 shows the Pivot Index for each step. Using this definition, Step 6 is the most pivoted step. According to Iskarous (2005), this is the most commonly occurring type of transition. For this step, as can be seen in Fig. 2, the tongue trajectory is relatively flat and as can be seen in Fig. 3, there is little to no change in the area function in the uvular region. In Step 1 and Step 11, on the other hand, there is indeed change in the area function in the uvular region, however, this change is in the opposite directions for the two steps. In Step 11 the change in area function is in the direction toward the fixed structures, whereas for Step 1, the change in the area function in the uvular region is in a direction away from the fixed structures. The CL Discreteness and Pivoting Indices quantify different aspects of area function change, therefore they impose different similarity metrics on the steps. Under the CL discreteness measure, Steps 1–6 are quite similar to each other, while under the pivot index measure, Step 6 is equally different from the steps above and below it. The reason that Steps 9–11 show lower values on the Pivot Index than the lowest steps is due to the differences in the geometry of the fixed structures in the palatal and pharyngeal regions.

An important aspect of the coarticulatory dynamics of /ai/ is the time at which the constriction formation for /i/ starts. If CD formation for /i/ starts at the same time as CD release for /a/, then CD change is synchronous, whereas if formation of the constriction for /i/ starts after the release of the /a/, then the transition is asynchronous. The first step for measuring CD synchronization (CD sync) is to extract the area of most *closed* part of the vocal tract (in the pharyngeal region for /a/ at the start of the transition), and the area of the most *open* part of the vocal tract (in the palatal region at the start of the transition) at each frame in the first half of the transition (where /a/ is releasing and /i/ is forming). Then the areas of the most closed part and the most open part are regressed against each other. If the slope of the regression line is near -1, it means that as the /a/ is released (opening the most closed part of the vocal tract), the /i/ is synchronously being formed (closing the most open part of the vocal tract). If the slope of the regression line is very different from -1, then the changes in the CD for /a/ and /i/ are asynchronous. The fourth panel of Fig. 4 shows the regression slope as a function of step, as a measure of CD synchronization for /a/ release and /i/ formation. For the lower steps, the slope is indeed close to -1 providing evidence for synchronous change in CD for /a/ and /i/ at those steps. In contrast, the higher steps show change in area of point of maximal constriction almost twice as much as change in the area at the point of minimal constriction—evidence of asynchronous change in CD for /a/ and /i/ at those steps. A few consecutive steps are similar to each other under this metric (e.g., 3 and 4), however the slope decreases relatively consistently across the scale. In that respect, the similarity metric induced by CD sync is more similar to the trajectory curvature metric, but different from the CL metrics. All of these metrics on the steps will later be compared to the metrics evident in the perceptual patterns.

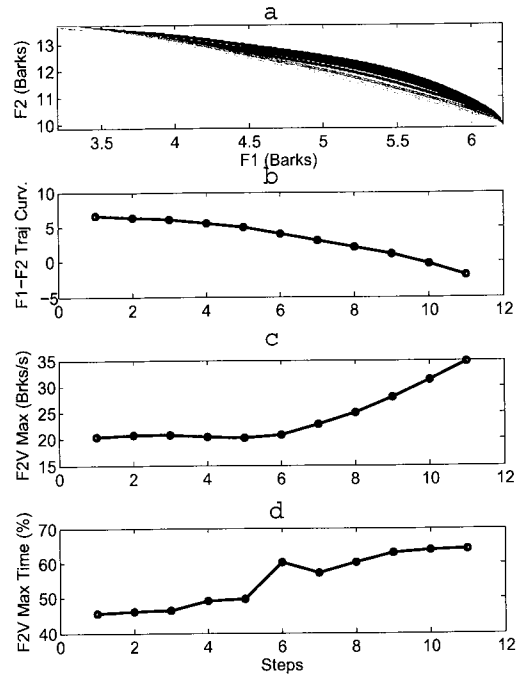

FIG. 5. (a) Trajectories in F1-F2 space (frequency in barks). Step 1 is in black and Step 11 is in lightest gray. (b) Curvature of F1-F2 trajectories as a function of step. (c) F2 maximal velocity as a function of step. (d) Time (in percent) at which F2 maximal velocity occurs.

B. Acoustic interpretation of /ai/

There are many ways to quantify the acoustic dynamic of each trajectory. In this section three methods for quantifying the acoustic trajectories are presented that seem to correlate to a relatively high degree with the articulatory and perceptual parameters considered in this work. F1 and F2 were extracted from the acoustic waveforms using LPC analysis after pre-emphasis and Hamming windowing, using 22 pole coefficients with a 25 ms window, repeated every 5 ms. The formants were then found by peak picking. The first measure used is the curvature of the F1-F2 trajectory in F1-F2 space (Ainsworth and Carré, 1997). The upper panel of Fig. 5 shows the F1-F2 trajectories on a Bark scale, with the most concave (black) trajectory corresponding to the

most concave tongue trajectory and the most convex (lightest gray) corresponding to the most convex tongue trajectory. The second panel of Fig. 5 shows the curvature of each F1-F2 trajectory, as measured using the three-point-circle method discussed earlier for tongue motion trajectories. The F1-F2 curvature function is monotonically decreasing, but it falls more slowly at the beginning of the scale than later in the scale.

The two other dynamic parameters are the maximal velocity of F2 (F2V Max) and the location in time where that peak occurs (F2V Max Time).[1] These parameters have been argued to be important for diphthong production and perception (Gay, 1968, 1970), which can be assumed to be related to the transition between two full vowels. For each of the 11 F2 transitions, the derivative of F2 with respect to time was computed, and the maximal F2 velocity and the time at which this maximum occurs were automatically picked. It can be seen from Panel 3 of Fig. 5 that for Steps 1–6, F2V Max is relatively stable at 20 barks/s, but the higher the transition on the scale, the higher it becomes. Moreover, the first six steps on the scale show virtually identical values. Therefore, despite the fact that the tongue curvature trajectories and F1-F2 curvature trajectories are quite different for Steps 1–6, F2 V Max does not distinguish between these steps—i.e., these steps form a class with respect to this measurement, as they do with respect to CL discreteness. The fourth panel of Fig. 5 demonstrates that Steps 1–5 show an early achievement of maximal velocity, whereas Steps 6–11 show a late achievement.

III. EXPERIMENTS

Each of the articulatory and acoustic measures examined imposes a similarity metric on the stimuli. Several hypotheses will therefore be tested through the perception experiments. Each hypothesis predicts that the speech perception system distinguishes between the stimuli in the same way as one of the measures. Three perception experiments will be presented in this section and their results will be interpreted in light of the similarity metrics imposed by the articulatory and acoustic measures. In addition, data on the acoustic properties of observed /ai/ tokens will be presented to compare to the perception results. The purpose is to test the hypothesis that the perceptual patterns are a reflection of the statistical properties of the various acoustic dimensions of naturally occurring data.

Both discrimination and rating tasks were used in the three experiments. The discrimination task, which took place first, was performed with an AXB task, in which participants listened to a triad of stimuli. A and B were /ai/ stimuli from different steps and X was either A or B, therefore the possible triads are: AAB, BBA, BAA, ABB. The components of the triad were separated by 900 ms. This value is unusually long, but it was difficult to keep the three sequences in the triad separate, when the interstimulus interval was made shorter. Participants were instructed to press the left-arrow button on the computer keyboard if they judged the first two sounds more similar, and the right-arrow button if they judged the second two sounds more similar. There was no time pressure, but they could hear each stimulus only once. After making a decision on one triad by pressing a key, another triad would be played. AXB was chosen as the discrimination task, because it is a low bias task and does not make as much of a demand on short-term memory as ABX. Stimuli differing by one and three steps were used in Experiments 1 and 3 and stimuli differing by five steps were used in Experiment 2.

In the naturalness-rating task, a single stimulus was played at a time and the participant was asked to rate the naturalness of that stimulus as being natural (left-button press) or not natural (right button press). The instructions were given to the participants in writing and the participants were not told what a natural or not natural transition would sound like. Their only way to decide was based on the fact that they had already listened to all the stimuli in the AXB part of the experiment, which served as a demonstration of the universe of /ai/ sounds they would hear in the experiment.

A. Experiment 1

1. Methods

11 native speakers of American English, 7 males and 4 females, participated. Consent to participate in the experiment was obtained from all the participants, and they were reimbursed for their participation. None had a reported history of speech or hearing disorders. All instructions were given in writing to the participants. Since there are 11 steps in the scale, there were 10 one-step pairs and 8 three-step pairs of triads in the AXB task. Each pair was represented eight times in the experiment (2 AAB+2 BBA+2 BAA +2 ABB) for a total of 144 triads, which were divided into two blocks. The one-step and three-step stimuli were randomized together into each of the blocks. In the naturalness-rating task, 6 tokens of each step were presented for a total of 66 stimuli, which were all randomized into one block.

2. Results and discussion

Figure 6 presents the results for all three tasks. The data is presented in means and standard errors for each step or pair. The 1-Step discrimination rises above 60% only for the 10–11 pair; otherwise it is near chance. The 3-Step discrimination rises above 60% for the 7–10 and 8–11 pairs; otherwise it is near chance. The naturalness rating is between 65% and 80% for Steps 1–7, then drops to 42% at Step 11 and then progressively lower to 25% at Step 11. To test whether the rating for the first 4 steps is significantly higher than for the last 4 steps, a one way repeated measures analysis of variance (ANOVA) was performed. The result is that Steps 1–4 are significantly higher than Steps 8–11 in naturalness rating with $F(1,10)=18.72$, $p<0.005$.

The naturalness rating was consistently high for the concave steps and low for the convex steps. This is not unusual in perception experiments where a phoneme boundary is crossed, but in this experiment, all the stimuli began and ended at the same targets, so no such boundary is crossed.

All participants reported that the AXB discrimination task was exceedingly difficult. This difficulty could have led

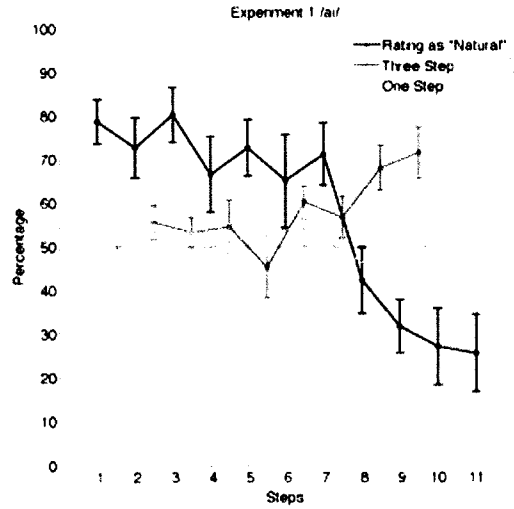

FIG. 6. Results of perception Experiment 1 as a function of step: Naturalness rating (black), three-step discrimination (gray), and 1-Step discrimination (lightest gray).

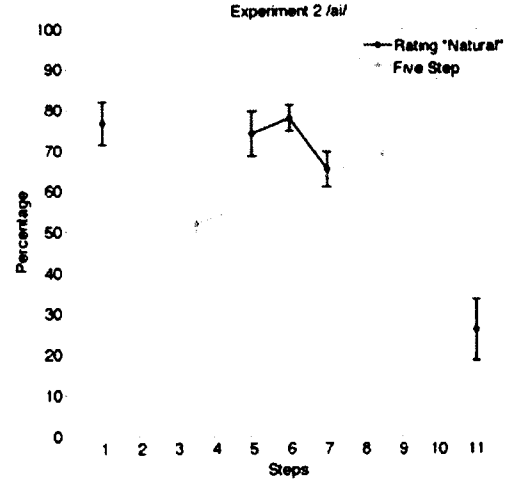

FIG. 7. Results of perception Experiment 2 as a function of step: Naturalness rating (black) and 5-Step discrimination (gray).

to the poor discrimination performance on the 1-Step and 3-Step tasks at the low end of the scale. There are two possible reasons for this difficulty: (1) The step size, even the three-step may have been two small. (2) There were only eight tokens of each discrimination pair, which may have been too low, leading to nonrobust results. Therefore, two additional experiments were performed. In Experiment 2, discrimination between Step 1 and 6 and between Step 6 and 11 were tested. In Experiment 3, the same pairs were tested as in Experiment 1, but more tokens were included of each pair.

B. Experiment 2

1. Methods

17 native speakers of American English, 7 males and 10 females, participated. None had participated in Experiment 1. Consent to participate in the experiment was obtained from all the participants, and they were reimbursed for their participation. None had a reported history of speech or hearing disorders. All instructions were given in writing. In this experiment, AXB task was the 5-Step: 1–6 and 6–11. Limiting the AXB to two pairs allowed us to use 40 tokens of each pair. For the naturalness-rating task, the steps that were used were 1, 5, 6, 7, and 11. Steps 1, 6, and 11 already occurred in the discrimination task, but Steps 5 and 7 had not. Each token was included 20 times for a total of 100 stimuli that were randomized into one block.

2. Results and discussion

Figure 7 presents the results of Experiment 2. Discrimination of pair 1–6 is at chance level, at 51%, whereas the mean discrimination for pair 6–11 is 68%. The two pairs are discriminated significantly differently according to a one way repeated measures ANOVA with $F(1,16)=14.69$, $p<0.005$. Furthermore, Steps 1, 5, 6, and 7 are rated as natural on average higher than 60% of the time, whereas Step 11 is rated natural only 26% of the time. Therefore, discrimination at the low end of the scale is still chance, even when the distance between steps is increased from 3 to 5. The naturalness-rating results are similar to those in Experiment 1.

C. Experiment 3

1. Methods

Seven native speakers of American English, three males and four females, participated. Consent to participate in the experiment was obtained from all the participants, and they were reimbursed for their participation. None had a reported history of speech or hearing disorders. All instructions were given in writing to the participants. All of the participants in this experiment had already participated in Experiment 2. The reason for limiting participation in this experiment to participants that had already been in Experiment 2 was to attempt to overcome the poor performance in Experiments 1. There, participants had poor discrimination performance at the lower end of the scale and had given higher naturalness ratings to the lower end of the scale than the higher end of the scale. It is therefore possible that previous exposure to the scale would allow participants to improve discrimination at the lower end and to equalize naturalness rating across the scale. That is, the patterns in the perceptual data that arise in Experiment 1 and 2 could disappear with greater familiarization with the data. In Experiment 3, the same tasks were run as in Experiment 1, except that each AXB pair was represented 36 times in the data (vs. 8) for a total of 576 stimuli randomized into eight blocks that were completed in two sessions. For the rating task, each step was included 18 times (vs. 6) for a total of 198 tokens randomized into two blocks.

2. Results and discussion

Figure 8 shows the results for all three tasks. The 1-Step function is all still near chance. However, while the 3-Step function is near chance at the lower end of the scale, it gradually improves until it reaches about 65% for the 4–7 and 5–8 pairs and continues to improve until it levels near

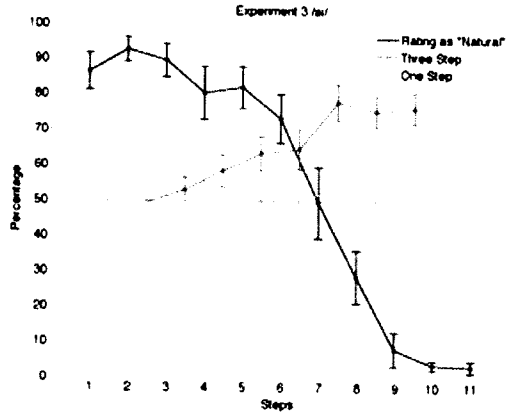

FIG. 8. Results of perception Experiment 3 as a function of step: Naturalness rating (black), three-step discrimination (gray), and 1-Step discrimination (lightest gray).

80%. The naturalness rating is above 70% for Steps 1–6 and then drops to 50% for Step 7 and then lowers and is less than 10% for Steps 9–11. As in Experiments 1 and 2, discrimination within the first six steps is still at chance; however the discrimination between steps in the 1–6 part of the scale and steps higher than 6 are above chance (starting at the 4–7 pair). The naturalness rating is again quite similar to the first two experiments, except that Step 7 is now identified as natural only at around 50% of the time. Apparently, when the noise in the data is reduced by increasing the number of each triad in each block, Step 7 becomes more like the higher steps than the lower steps.

D. Acoustic analysis of x-ray microbeam data

An assumption of several theories is that speech perception involves auditory/acoustic pattern matching. Therefore, it would be useful to compare the various acoustic measures used to observed measures from typical productions of /ai/ to determine if the perceptual measures were based on a simple comparison between the acoustic properties of the synthetic stimuli and acoustic properties of real productions.

1. Methods

The vowel sequence /ai/ is a relatively rare sequence in American English. But a speech database, the Wisconsin X-ray Microbeam Database (Westbury, 1994), contains productions by American English Speakers of VV sequences, including /ai/. The acoustic waveform for /ai/ was segmented out automatically for 49 participants. F1 and F2 were extracted using Linear Predictive Coding (LPC) analysis after pre-emphasis and hamming windowing using 22 coefficients with a 25 ms window, repeated every 5 ms. After visual analysis of the data, only data from 34 participants were kept, since the formant extraction procedure did not work well for the other speakers and resulted in highly discontinuous estimates. There are, of course, many methods for smoothing formant trajectories, but smoothing trajectories prior to measuring analysis-sensitive parameters like curvature and maximal velocity would potentially bias the data; therefore data with discontinuous trajectories were excluded.

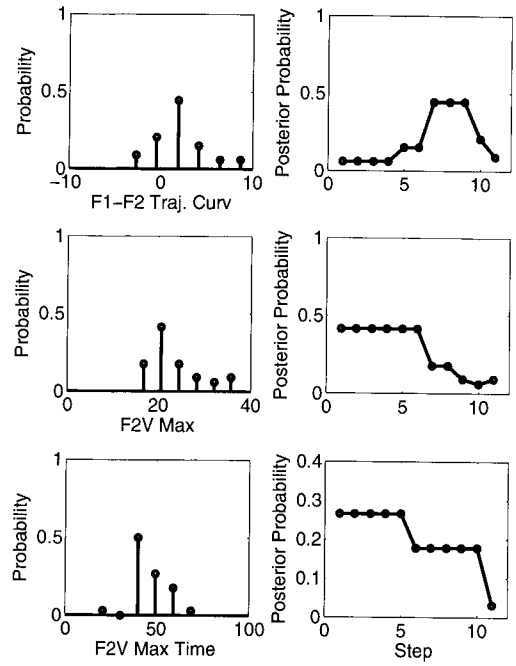

FIG. 9. Left: Probability histograms of the chance of occurrence of each of the measured acoustic quantities in the observed 34 tokens of /ai/. Right: Posterior probability of each of the steps on the scale of synthetic stimuli, given the likelihood of observing that measured quantity in the observed data.

Curvature of F1-F2, F2V max, and F2V Max Time were measured in the same way as for the synthetic data.

2. Results and discussion

The left side of Fig. 9 shows probability histograms of each particular F1-F2 curvature, F2V Max, and F2V Max Time among the 34 tokens. The histograms are unimodal, indicating that each of the acoustic quantities has a most probable value in naturally occurring /ai/ utterances. Of course this data is from 34 particular speakers; however, it will be assumed that the probabilities are representative of natural occurrences of /ai/. To test the hypothesis that perceptual patterns are derived from the probabilities of occurrence of each of the acoustic parameters of possible /ai/'s, the probability for preferring a stimulus of a particular step was derived from the probability of hearing each of the acoustic quantities in naturally occurring /ai/'s. For instance, approximately 50% of the tokens had a maximal F2 Velocity of 20 barks/s. Comparison of this value to the F2V max values for the synthetic data shows that this particular value of F2V Max is approximately the value for Steps 1–6. Under the assumption that the 34 tokens chosen here are representative of /ai/ productions in American English, and if the hypothesis is true that perceptual judgments depend on the average values of various acoustic quantities of naturally produced tokens in a listener's environment, then the probability of a listener perceptual judgments should be a function of the probability of hearing that stimulus (Lisker and Abramson, 1970; Nearey and Hogan, 1986). For each of the acoustic quantities measured, and each of the steps, the probability of that acoustic quantity being measured was computed, based

TABLE I. Relation between articulatory, acoustic, and perceptual variables. First ten rows show r^2 for each two pairs of variables. Pairs with explained variance higher than 85% are in bold large type. Last two rows show the difference (in z-scores) for each of the articulatory and acoustic variables between Steps 6 and 1 (row 11) and 11 and 6 (row 12), for comparison with discrimination data. Variables that show insignificant difference for the 1–6 pair and a highly significant difference for the 6–11 pair are in bold large type. Abbreviations: TngCurv =curvature of tongue trajectory, CLDisc=CL discreteness, Piv=pivoting index, CDS=CD sync, AcCurv =F1-F2 curvature, F2VMx=F2V Max, F2VMxT=F2V Max Time, Exp1NR=Experiment 1 naturalness rating, and Exp3NR=Experiment 3 naturalness rating.

	TngCurv	CLDisc	Piv	CDS	AcCurv	F2VMx	F2VMxT	Exp1NR	Exp3NR
				Articulatory factors					
TngCurv									
CLDisc	**87**								
Piv	15	48							
CDS	**95**	79	9						
				Acoustic factors					
AcCurv	**97**	**94**	29	**88**					
F2VMx	80	**96**	57	67	**91**				
F2VMxT	**92**	74	8	**91**	**86**	65			
				Naturalness rating					
Exp1NR	**85**	**93**	38	79	**88**	**86**	77		
Exp3NR	**92**	**96**	32	**89**	**94**	**86**	83	**92**	
				Discrimination					
Dsc 6–1	−1.33	**−0.13**	2.11	−1.47	−0.89	**−0.07**	−1.90		
Dsc11–6	−1.55	**−2.40**	−3.16	−1.04	−2.05	**−2.74**	−0.52		

on a simple identity function. That is, the probability of preferring a stimulus was calculated as the probability of hearing that stimulus in naturally occurring tokens. The posterior probabilities for each of the steps (having heard the acoustic stimuli) are shown on the right hand side of Fig. 9. For instance, since 20 barks/s is the most probable F2 MaxV in the observed data, and this value is shared by the Steps 1–6, then a listener would prefer these steps, assigning them a high posterior probability of occurrence, as can be seen in higher probabilities for the lower that for the higher steps. The predictions of the hypothesis are therefore: (1) based on F1-F2 curvature, the most preferred stimuli should be 7–9; (2) based on F2V Max, the most preferred stimuli should be 1–6; (3) based on F2V Max Time, the most preferred stimuli should be 1–5.

IV. GENERAL DISCUSSION

A. Comparison of perception experiment results

The results for the three experiments differ in many details, but they are consistent in the basic patterns. 1-Step discrimination goes above 60% only between Steps 10 and 11 (Experiment 1). 3-Step discrimination is chance until the 5–8 pair in Experiment 1 and the 4–7 pair in Experiment 3. Discrimination stays at the same level or improves for higher 3-Step pairs. Therefore, for the 3-Step, discrimination is poor for the lowest steps and starts to improve near the middle of the scale. This is despite the fact that Experiment 3 participants had a great deal of exposure to the data, since they had already participated in Experiment 2. Also, despite the increase in the number of steps in Experiment 2, 1–6 discrimination is still at chance, whereas 6–11 discrimination is significantly above chance. Therefore the discrimination results are relatively consistent in all three experiments, showing poor discrimination at the low end and higher discrimination at the high end. The naturalness rating for all three experiments shows high rating from Step 1 until the middle of the scale (Step 7 for Experiments 1 and 2, and Step 6 for Experiment 3), and then sharply drops for higher steps. Therefore the three experiments are also relatively consistent regarding naturalness rating.

B. Articulatory-acoustic-perceptual relations

In order to examine the relation between the articulatory, acoustic, and perceptual rating patterns, the amount of variability that each factor explains of the others is calculated as r^2 for each factor pair. The results are shown in the first 10 rows of Table I, which lists r^2 as percentages. Rows 2–8 of the table show the relation among and between the articulatory and acoustic factors and Rows 9 and 10 show how each of those variables explains the variability seen in the rating tasks of Experiments 1 and 3. Cells for which r^2 is greater than 85% are in bold face. The last two rows of the table show the difference in z-scores between the value for each investigated function at Step 1 and Step 6 (Row 11) and Step 6 and Step 11 (Row 12). These differences are provided in comparison to the discrimination pattern in Experiment 2. In that experiment, 1–6 were discriminated at chance, whereas 6–11 were discriminated significantly above chance. Articulatory and acoustic variables that explain that discrimination pattern would show near 0 difference (in z-scores) between Steps 1 and 6 and a significant difference between Steps 6 and 11. There are only two factors (CL Discreteness and F2V Max) that show this pattern and their z-score differences are emphasized by being in bold face type.

Most of the articulatory and acoustic factors, except for the pivot index, predict the asymmetry in the naturalness task, but the results for the more difficult discrimination task

distinguish among the articulatory and acoustic variables more than the rating task. Tongue trajectory curvature, CD sync, and F1-F2 curvature all show z-score differences of about 1 standard deviation or higher for both the 1–6 pair and 6–11 pair, which would predict that 1–6 and 6–11 would both be discriminable; however, the discrimination results showed that only the former pair could be distinguished. Only CL discreteness and F2V Max show a near 0 standard deviation difference for 1–6 and higher than 2 standard deviation difference for 6–11, predicting the exact discrimination pattern. Indeed CL discreteness and F2V Max have a 96% shared explained variation allowing us to conclude that F2V Max is the main acoustic measure that tracks the discreteness of CL change and that CL Discreteness, signaled by F2 V Max, is the main quality of the /ai/ articulatory dynamic that predicts both the naturalness rating and discrimination perceptual patterns.

The preference for discrete switching from pharyngeal to palatal could be interpreted as a preference for static realizations of /ai/, since CL switching leads to two static values of CL. However, even though the CD sync measure is not a good predictor of the discrimination patterns, the parameterization for the lower end of the scale shows that /i/ begins at the same time as the release of the constriction for /a/ as was seen in the fourth panel of Fig. 4. That is, even though CL switches at some point, CD change occurs continuously through the transition. Therefore the most highly rated steps are not at all static sequences of /a/ and /i/. This can be seen clearly also in the continuity of F1 and F2 change.

F1-F2 curvature is very highly correlated with tongue dorsum trajectory curvature (r^2=97%). The extremeness of these values suggests that F1-F2 curvature is tracking the tongue trajectory curvature in the same way that F2V Max is tracking CL discreteness. But only the latter two factors predict both the rating and discrimination patterns. This complex relation between factors allows for distinguishing between high level (more abstract) and low level (less abstract) measures. Low-level descriptors of /ai/, specifically tongue and F1-F2 trajectory curvature, specify the kinematic changes in /ai/ very well, but do not explain the entire set of perceptual patterns. High level variables on the other hand, namely CL discreteness and F2V Max, describe all the perceptually relevant aspects of the transition. To answer the questions posed at the beginning of this paper, it seems that the total of the perceptual judgements are not based on the exact trajectory of the tongue or the exact trajectory of F1 and F2 in F1-F2 space, the low level variables, but on the higher level descriptors of CL change and its acoustic signature, F2V Max.

The pivoting index does make the largest differentiation between Steps 6 and 11, −3.16 z-scores, which agrees with the results of the discrimination task. The index also makes the prediction that Step 6 would be highly rated as natural, which is indeed the case in all three experiments. However, the pivoting index based prediction differentiates greatly between Steps 1 and 6 (2.1 z-score) and predicts that the lowest steps would not be highly rated, both of which are untrue. Instead, what seems to be the case is that change in area function in the pivot region leads to high naturalness rating as long as it is *anticonstricting*, that is, as long as it is away from the fixed structures, which is what happens in the lower steps. In the earlier study of pivoting (Iskarous, 2005), the principle of CL discreteness was thought to drive pivoting. However, the use of synthetic trajectories shows that CL discreteness can also occur in nonpivoted trajectories whose path moves maximally away from the hard structures, as opposed to nonpivoted trajectories that move following the curvature of the hard structures. It appears from the current work that the most salient aspect of produced /ai/ trajectories is their discreteness of CL change, which can be high, even when the pivoting index is low (as happens in Steps 1–5). CL discreteness is therefore a better measure of what is most articulatorily relevant.

Of course acoustic speech patterns are structured by dynamic vocal tracts, so it may seem that the acoustic and articulatory patterns are entirely interchangeable. However there are nonlinear relations between the two domains, so the map between the two domains is not one-to-one. Specifically, since F2V Max explains the perceptual patterns, it might be said that no reference to articulation needs to be made. Exposure to that acoustic quantity by listeners is all that is required. Indeed as seen in Fig. 9, the posterior probability of Steps 1–6 is higher than that for the other steps, since their value for F2V Max is approximately the same as that in the most probable in the observed /ai/'s. However, the other acoustic quantities do not show the same pattern and reliance on these posterior probabilities in a multidimensional pattern recognition task would lead to different patterns than the observed perception patterns. For instance, reliance on the most probable F1-F2 Curvature would lead to preference for the high steps, not the low steps, exactly in opposition to what is observed. It is therefore concluded that, in this case at least, perceptual judgements are not based on the total acoustic pattern, but are biased toward certain acoustic parameters and against certain others. The extremely high correlation between F2V Max and CL discreteness suggests that the bias toward F2V Max is based on its being the signature of a crucial aspect of the production.

C. Why discreteness?

It has still not been determined, however, why listeners prefer stimuli from the lower end of the scale, rather than the more frequently occurring midpoint. The answer could follow from an appreciation of the traditional problem of the discrete vs. the continuous in speech. Each linguistic community seems to make use of a relatively small number of phonological contrasts to distinguish between linguistic units. These discrete linguistic units seem to be learned very early and influence speech perception in children and adults (Jusczyk and Luce, 2002). However, in the production of speech, these linguistic units seem to be intermixed in a way that obscures their linguistic individuality due to coarticulation (Hockett, 1955). In contrast to linguistic discreteness, we therefore seem to have phonetic continuity—tongue trajectories are continuous and formant trajectories are continuous. A major problem in speech perception is how this continuous phonetic flow can serve as evidence for discrete

contrastive linguistic units. In the current investigation of the vowel sequence /ai/, F1 and F2 change continuously as the tongue trajectory changes continuously in all the possible trajectories for /ai/. Why, however, is /ai/ considered a sequence of two contrastive entities, rather than a sequence of a large number of intermediate sounds resolved by the auditory system in the course of the 350 ms stimulus? The answer that this experiment points to is that /ai/ is parceled into two segments, because despite the continuousness of motion of the tongue dorsum, CD, and F1 and F2, CL changes *discretely*, rather than changing continuously. That is despite the presence of some continuously changing parameters in the production of speech, there are also discrete switches in some parameters. It is possible that such discrete switches could be the foundation for how the speech production system serves to communicate discrete contrasts. The relation between contrastive linguistic entities and the articulatory signal that communicates them that emerges here is, therefore quite different from the one proposed by Hockett (1955) and assumed in much of the literature. Discrete contrastive linguistic entities, namely, CL, remain discrete in the articulatory output, despite the continuity of other output parameters, including the acoustic ones.

The results presented here therefore have an implication to the solution of the segmentation problem (Fowler and Smith, 1986; Liberman and Whalen, 2000): linguistic and perceptual discreteness could arise from the discreteness of a critical contrastive linguistic parameter CL, despite continuous change in other articulatory parameters, as well as F1 and F2. Perceptual tuning to this pattern, evident in the results of the discrimination tasks, could also serve to explain the preference for the lower end of the scale: Stimuli 1–6 are segmentable into /a/ and /i/ due to CL discreteness, using the F2V Max cue, despite the continuous change of the acoustic parameters. In future work, the experimental verification of the importance of CL switching for speech segmentation will be pursued by investigating other transitions.

V. CONCLUSION

Ratings of naturalness of synthetic /ai/ sequences show preference for stimuli with concave-to-pivoted movement patterns, all of which had discrete changes in CL. Listeners were unable to discriminate pairs of vowel sequences that did not differ in CL discreteness. The perceptual judgements closely correspond to the acoustic and articulatory properties of transitions from /a/ to /i/. One particular articulatory parameter identified that explains the basic perceptual patterns is the CL discreteness parameter. The acoustic signature of that parameter was identified as the maximal velocity of F2. It is proposed that the discreteness of CL change is a crucial aspect of how contrastive linguistic units are communicated through the speech production system, and may serve to segment the speech signal into overlapping discrete entities.

ACKNOWLEDGMENTS

This work was supported by NIH-NIDCD Grant No. DC-02717. We would like to thank David Berry and two anonymous reviewers for their insightful critiques. Many thanks also go to Carol Fowler, Philip Rubin, Mark Tiede, Arthur Abramson, Christine Mooshammer, Bruno Repp, and Louis Goldstein for many helpful discussions and thoughtful comments

[1]Syrdal and Gopal (1986) show that F2 in American English is auditorily relevant through its distance to F3. Therefore, given the fixing of F3 to a constant, the measures that depend on F2 could also be regarded as measures that depend on the distance F3-F2. The small change in F3 for American English vowel sequences such as /ai/ exhibited in Simpson (2001, 2002) justifies this approximation since most of the change in F3-F2 would be accomplished by F2.

Ainsworth, W. A., and Carré, R. (**1997**). "Perception of synthetic two-formant vowel transitions," Speech Commun. **21**, 273–282.

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

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

Coker, C., and Fujimura, O. (**1966**). "Model for the specification of the vocal-tract area function," J. Acoust. Soc. Am. **40**, 1271.

Fowler, C., and Smith, M. R. (**1986**). "Speech perception as vector analysis: An approach to the problems of invariance and segmentation," in *Invariance and Variability in Speech Processes*, edited by J. S. Perkell and D. H. Klatt (Lawrence Erlbaum Associates, Hillsdale, NJ).

Gay, T. (**1968**). "Effect of speaking rate on diphthong formant movement," J. Acoust. Soc. Am. **44**, 1570–1573.

Gay, T. (**1970**). "A perceptual study of American English diphthongs," Lang Speech **13**, 65–88.

Hanson, H., and Stevens, K. (**2002**). "A quasiarticulatory approach to controlling acoustic source parameters in a Klatt-type formant synthesizer using HLsyn," J. Acoust. Soc. Am. **112**, 1158–1182.

Hockett, C. (**1955**). *A Manual of Phonology* (Waverley, Baltimore).

Iskarous, K. (**2005**). "Patterns of tongue movement," J. Phonetics **33**, 363–381.

Iskarous, K., Goldstein, L., Whalen, D. H., Tiede, M., and Rubin, P. (**2003**). "CASY: The configurable articulatory synthesizer," in Proceedings of the 15th International Congress of Phonetic Sciences, pp. 185–188.

Jusczyk, P., and Luce, P. (**2002**). "Speech perception and spoken word recognition: Past and present," Ear Hear. **23**, 2–40.

Liberman, A., and Whalen, D. H. (**2000**). "On the relation of language to speech," Trends Cogn. Sci. **4**, 187–196.

Lisker, L., and Abramson, A. (**1970**). "The voicing dimension: Some experiments in comparative phonetics," in Proceeding of the 6th International Congress of Phonetic Sciences, pp. 563–567.

Mermelstein, P. (**1973**). "Articulatory model for the study of speech production," J. Acoust. Soc. Am. **53**, 1070–1082.

Mrayati, M., Carré, R., and Guerin, B. (**1988**). "Distinctive regions and modes: A new theory of speech production," Speech Commun. **7**, 257–286.

Nearey, T., and Hogan, J. (**1986**). "Phonological contrast in experimental phonetics: Relating distributions of production data to perceptual curves," in *Experimental Phonology*, edited by J. Ohala and J. Jaeger (Academic, New York).

Rosner, B. S., and Pickering, J. B. (**1994**). *Vowel Perception and Production* (Oxford University Press, New York).

Rubin, P., Baer, T., and Mermelstein, P. (**1981**). "An articulatory synthesizer for perceptual research," J. Acoust. Soc. Am. **70**, 321–328.

Rubin, P., Saltzman, E., Goldstein, L., McGowan, R., Tiede, M., and Browman, C. (**1996**). "Casy and extensions to the task-dynamic model," in Proceedings of the 1st ESCA ETRW on Speech Production Modeling the Speech Production Seminar, pp. 163–178.

Simpson, A. (**2001**). "Dynamic consequences of differences in male and female vocal tract dimensions," J. Acoust. Soc. Am. **109**, 2153–2164.

Simpson, A. (**2002**). "Gender-specific articulatory-acoustic relations in vowel sequences," J. Phonetics **30**, 417–435.

Stone, M. (**1991**). "Toward a model of three-dimensional tongue movement," J. Phonetics **19**, 309–320.

Story, B. (**2005**). "A parametric model of the vocal tract area function for vowel and consonant simulation," J. Acoust. Soc. Am. **117**, 3231–3254.
Syrdal, A. K., and Gopal, H. S. (**1986**). "A perceptual model of vowel recognition based on the auditory representation of American English vowels," J. Acoust. Soc. Am. **79**, 1086–1100.
Westbury, J. (**1994**). *X-ray Microbeam Speech Production Database Users Handbook* (University of Wisconsin, Madison, WI).
Wood, S. (**1979**). "A radiographic analysis of constriction locations for vowels," J. Phonetics **7**, 25–43.