Perception of complete and incomplete formant transitions in vowels

Pierre Divenyi^{a)}

Veterans Affairs Northern California Health Care Systems and East Bay Institute for Research and Education Martinez, California 04553

(Received 31 October 2007; revised 10 June 2009; accepted 11 June 2009)

In everyday speech, formant transitions rarely reach the canonical frequencies of a target vowel. The perceptual system often compensates for such production undershoots, called vowel reduction (VR), by a perceptual overshoot of the final transition frequencies. The present investigation explored the perceptual parameters and existence region of VR. In a series of experiments a 100-ms steady-state vowel V_1 was followed by a formant transition toward a target vowel V_2 . By manipulating both its duration and velocity, in most stimuli the transition was truncated and only seldom reached the target. After being presented with the vowel V_2 before each block of trials, listeners were asked to rate their confidence that the transition actually reached the V_2 target. Transitions along six trajectories connecting the three cardinal vowels /a/, /i/, and /u/ in both directions as well as the transition /ie/ (halfway along the trajectory /ia/) were examined in experiments in which either the duration of the transition was fixed and its velocity was varied or vice-versa. Results confirmed the existence of perceptual overshoot and showed that, at the point a transition short of reaching the vowel V_2 was just perceived as if it had reached the target, transition duration and transition velocity were inversely related. The amount of overshoot was found to be larger for larger V_1 - V_2 distances and shorter trajectory durations. The overshoot could be reliably predicted by a linear model based on three parameters—the extent of V_1 - V_2 distance, transition velocity, and transition acceleration. These findings suggest that the perceptual dynamics of speech relies on mechanisms that estimate the rate of change in the resonant characteristics of the vocal tract. [DOI: 10.1121/1.3167482]

PACS number(s): 43.71.Es, 43.66.Lj, 43.71.An [RSN]

Pages: 1427–1439

I. INTRODUCTION

To say that speech communication is based on gestures, that is, speech movements, amounts to a tautology. When one individual engages in conveying information to another by way of speech, he/she has to perform a series of articulatory movements that generate a sequence of changes in his/her vocal tract. Although there are brief islands of relatively quiescent periods between consecutive gestures leading to such changes, an overwhelming portion of the information transmitted is encoded in the transitions that connect motion-free quasi-steady-state periods, i.e., in the changes themselves.

Paradoxical as it may sound, because vowels appear to constitute the bulk of such still periods, perhaps nowhere is the importance of transitions better illustrated than in the production and perception of vowels. Studies on vowel production have revealed a considerable variability in the formant values—even within a given language—across different speakers, speaking rates, speaking styles, and contexts (Peterson and Barney, 1952). One of the most intriguing aspects of vowel perception is that, despite large overlaps of the vowel categories when mapped onto the formant domain, vowels retain their perceived identity across physical changes due to variations of different origin (Nearey, 1989). Early investigations on running speech presented evidence that a non-negligible portion of formant variability may be

attributed to dynamic context (Furui, 1986; Lindblom and Studdert-Kennedy, 1967; Strange *et al.*, 1979) and suggest that data obtained on the production and perception of static vowels are inadequate to describe, or model, spoken language. This conclusion is supported by a number of studies focused on the identification and discrimination of V_1 - V_2 , CV, VC, and CVC transitions (Andruski and Nearey, 1992; Carré *et al.*, 2007; Carré and Divenyi, 2000; Hillenbrand and Nearey, 1999; Nearey, 1989; Strange, 1989; van Son and Pols, 1993; van Wieringen, 1995; van Wieringen and Pols, 1994). Also, experiments on the perception of vowels in “silent center” CVC syllables have demonstrated that a vowel replaced by silence remains largely identifiable (Fox, 1989; Jenkins *et al.*, 1999). On the other hand, replacement of consonantal segments in sentences by noise preserves about twice the information as does replacement of vowel segments, suggesting that even vestigial CV and VC transitions in the vowels can lead to identifying the missing consonant or even consonant cluster (Kewley-Port *et al.*, 2007). Interestingly, in contrast to the well-known differences between formants of isolated vowels produced by male, female, and child talkers, formant trajectories from vowel onset to vowel center in CV diphones are remarkably stable across talkers (Hillenbrand *et al.*, 2001). Such stability is also reinforced by an essential lack of overlap between vowel categories in multitalker French V_1 - V_2 utterances, observed when the vowel pair is represented by the rates of F1-F2 change instead of the actual formant frequencies (Carré *et al.*, 2007). It appears, therefore, that vowel transitions contain information

^{a)}Electronic mail: pdivenyi@ebire.org

which not only complements that conveyed by static formants but which, under contextual diversity, is necessary, and may be sufficient, for signaling the vowel's identity. Patterns of inherent diphthongization of vowels provide a dynamic cue that contributes to their identification in American English (Assmann and Katz, 2005; Morrison and Nearey, 2007) while formant contours carry a great amount of phonological information in Australian English (Watson and Harrington, 1999). Important for understanding the perception of diphthongization is to establish the conditions under which altering one or several formants of a steady-state vowel will be perceived as a change to a diphthong or a different vowel. While this change, both from the productive and perceptual points of view, is influenced by the fundamental frequency f_0 (Di Benedetto, 1994), altering the transition appears to affect the percept also when f_0 is held constant, as demonstrated in a series of studies by Nabelek and co-workers. Manipulating the rate and the duration of formant transitions shifted the perceptual boundary between an initial steady-state vowel and a target vowel at the end of a transition (Nabelek *et al.*, 1993) or changed the point at which the initial vowel was heard either as steady-state or diphthongized (Nabelek *et al.*, 1996). Added noise or reverberation also influenced these boundaries (Nabelek *et al.*, 1994, 1996).

Transitions *toward* a vowel, however, whether from a preceding consonant or another vowel, often terminate at formant frequencies that fall short of those of a canonical vowel, i.e., the target vowel in isolation. Such production undershoot, termed vowel reduction (VR, Lindblom and Studdert-Kennedy, 1967), has been associated with casual speaking style ("hypo-speech," Lindblom, 1983), increased speaking rate (Flege, 1988; Gottfried *et al.*, 1990), and reduced vocal effort (Lienard and Di Benedetto, 1999), as well as with consonantal context (Stevens and House, 1963). Interestingly, listeners not only successfully identify these reduced vowels (Macchi, 1980) but also perceive them as *identical* to the canonical vowel when they are preceded by a transition, and as *different* from the canonical vowel when they are presented in isolation (Carré and Divenyi, 2000). This perceptual overshoot has been considered the perceptual system's compensation for the articulatory undershoot of VR (Lindblom and Studdert-Kennedy, 1967; Pols and van Son, 1993). Contrary to citation-form "hyper-speech" in which boundaries of the vowel triangle are expanded (Whalen *et al.*, 2004), in VR the vowel preceded by the transition may have formant values displaced toward the center of the vowel triangle or simply lying on an incomplete trajectory leading to but not reaching the target vowel. VR-type formant changes also underlie diphthongization (Andruski and Nearey, 1992) and CV and VC diphones (Hillenbrand *et al.*, 2001). Three main acoustic factors appear to contribute to VR: formant displacement, transition duration, and transition velocity (Lindblom *et al.*, 1996). Not surprisingly, the same three factors also emerge as those contributing most heavily to identification of silent center vowels in different CVC contexts (Jenkins *et al.*, 1994; Strange and Bohn, 1998), boundary shifts between /a/ and /aI/ percepts (Nabelek *et al.*, 1994), and discrimination of vocalic transitions (van Wierin-

gen and Pols, 1995). VR perception has been observed in /wVw/ and /jVj/ contexts and has been also shown to be affected by nonlinearity, i.e., the acceleration component, of the transition (Nabelek and Ovchinnikov, 1997).

The above reports therefore suggest that the perceptual overshoot compensating for the VR undershoot represents not only a noteworthy illusion but also a possibly indispensable aspect of speech perception: it stands to guarantee perceptual invariance by keeping the perceived identity of the vowel constant despite any physical variability imposed by prosodic and contextual differences. Although work over the last 30-or-so years has uncovered characteristics of this phenomenon, it was seldom the focus of investigations but, rather, a tool to test the validity of a large-scale hypothesis—such as the overarching role of context effect (e.g., Holt *et al.*, 2000)—or an ancillary observation while studying another phenomenon—such as inherent spectral changes in vowels (Nearey and Assman, 1986). Yet, knowledge of the parametric regions of existence of VR perception could bring us closer to understanding the dynamic processes operating in speech. For one, identifying the mechanisms behind the perceptual VR overshoot would not be without interest: the controversy concerning the origin—auditory or linguistic—of the overshoot for the identification of vowels in a CVC setting has been around for well over a decade but is still not settled. The currently prevailing auditory explanation states that spectral contrast, whether intrasegmental, i.e., within a CVC syllable, or extrasegmental, i.e., a CVC test item preceded and/or followed by an anchoring sound, can fully account for the perceptual overestimation of the extent of the transition (Holt *et al.*, 2000; Lotto and Holt, 2006). The contrasting linguistic theory states that the perceptual overshoot of the transition is a direct reflection of the articulatory undershoot because speech is perceived by processes that translate the acoustics into the articulatory gestures that were used by the talker (Fowler, 1994; Vishwanathan *et al.*, 2009). Both views have classic predecessors and are backed by modern-day confirmatory evidence. Early auditory explanation based on frequency-specific adaptation along formant sweeps (Lindblom and Studdert-Kennedy, 1967) is consistent with spectrotemporal modulation indices of auditory cortical responses to complex sounds (David *et al.*, 2009). The concept of gestural perception of speech is at least 300 years old¹ and it formed the basis for the analysis-by-synthesis theory proposed in the 1960s (Halle and Stevens, 1962); recently, it has received support from neuroimaging studies showing activation of precentral cortical motor areas mapped to the lips and the larynx, when listening to speech (e.g., Pulvermüller *et al.*, 2006). While testing a hypothesis in favor of either of the two contracting theories is not the objective of the present study, addressing the perception of VR may give support to either theory—or both.

The experiments reported in this article assessed listeners' perception of vowel-to-vowel transitions by asking if they perceived whether a specified target was reached, rather than by looking at conditions in which the presence of the transition resulted in the percept of a diphthong—as was the case in many studies on this topic (e.g., Nabelek *et al.*,

1994). The choice of studying V_1 - V_2 rather than CVC or CV or VC transitions was dictated by a desire to also examine trajectory durations that, albeit brief, could exceed those of consonantal transitions. Aside from determining parameters of complete and incomplete transitions that lead to the percept of a vowel target reached, these experiments also represent a test of whether the Gestalt principle of continuity—the auditory mechanism shown to follow a monotonic rise or fall of the frequency of a sine wave signal (Crum and Hafter, 2008; Nakajima *et al.*, 2000)—can be also demonstrated to operate in the processing of vowels.

Specifically, the experiments reported here used VR to gauge the perception of vowels in a dynamic context by examining in detail the role of three transition parameters: formant trajectory, transition duration, and transition velocity. The specific aim of the experiments was to find for each condition a transition boundary, i.e., the extent to which the transition leading from a steady-state vowel V_1 to formant frequencies of a target vowel V_2 should be truncated so that listeners are just able to perceive the target at the end of the transition. For the first two experiments, the separation of the starting and target vowels was kept as large as possible by using the six trajectories between pairs of the three vowels /a/, /i/, and /u/. Experiment 1 was directed at determining the boundary for a variable transition velocity with its duration held constant within each condition, while experiment 2 measured the boundary for a variable transition duration with its velocity held constant within each condition². In experiment 3, transition velocity and transition duration boundaries were measured for the /ia/ and the /ie/ trajectories, i.e., two trajectories that were collinear and differed only in their lengths. Finally, a control experiment tested the notion that VR-type overshoot for the perception of the extent of transitions can be observed for nonspeech complex sounds having the frequency of a single formant modulated to form a sweep.

II. METHODS

A. Stimuli

In all experiments, the stimulus consisted of an unbroken succession of a steady-state vowel V_1 and a transition toward a second vowel V_2 . The transitions proceeded along a trajectory from a vowel at one corner of the vowel triangle /a/-/u/-/i/ toward a vowel at one of the other two corners. These vowels do occur in American English, although the /u/ vowel is rare in everyday spoken language in America. They were chosen because they are as far from each other as the vowel space allows, thereby facilitating the study of transitions between them. In experiments 1 and 2, the trajectories investigated consisted of the exhaustive set of the six V_1 - V_2 combinations of the three corner vowels—/ai/, /ia/, /au/, /ua/, /iu/, and /ui/.³ All stimuli were digitally synthesized at a 20 kHz sampling rate using the distinctive region model (DRM) (Mrayati *et al.*, 1988)—an articulatory model based on deformations of a vocal tract-analog tube at specific locations that generates a parameter table, which serves as input to a cascade-mode formant synthesizer. The resonators of the first four formants were excited by a train of 100- μ s pulses.

FIG. 1. Schematic time-frequency diagram of the stimuli illustrating the trajectories of the first two formants in four of the eight tokens in each experimental condition. Panel A shows tokens of different transition velocities (i.e., transition slopes) in the constant transition duration experiments (experiments 1 and 3a), while Panel B shows different transition durations for the constant transition velocity experiments (experiments 2 and 3b). The specific example in both panels illustrates the formant trajectories for the /ai/ transition. Note that in both diagrams only one of the transitions reaches the formants of the /i/ target vowel.

The fundamental frequency (f_0) contour of the pulses corresponded to that of a natural V_1 - V_2 utterance by a male talker, with an initial f_0 of 120 Hz linearly rising to 140 Hz for the duration of V_1 and linearly declining back to 120 Hz for the duration of the transition. The duration of the first vowel V_1 was always 100-ms; the duration and the velocity of the transition were the two parameters of interest.

All transitions were synthesized by following, in a piecewise linear fashion, a formant frequency table generated by the DRM, updated every 2 ms. Because formant changes generated by the DRM synthesizer were a monotonic but not linear function of time, the velocity of the frequency change was also compressive-nonlinear (with an exponent between 0.6 and 0.9), with the bulk of the variation concentrated at the onset (at which the given regions of the vocal tract specified by the model are least impeded when the area functions expand or contract). The curvature of the formant change, therefore, was similar to that of the quadratic function used by Nabelek and Ovchinnikov (1997), although the rate in the present experiments was variable.

Experiment 1. In each block of experiment 1, the transition duration was held constant (30, 60, or 90 ms) while eight tokens with different transition velocities were presented 16 times each in random order. The V_2 target was reached, and thus the entire trajectory covered, by only the token having the highest transition velocity; transitions in the other seven tokens had increasingly lower velocities with trajectories covering, on the average, lengths of 7/8, 6/8, 5/8, etc. of the maximum. Figure 1(a) illustrates schematic first and second formant frequency-vs-time trajectories along the /ai/ continuum for four of the eight variable transition velocity tokens used in experiment 1.

Experiment 2. In each block of experiment 2, the transition velocity was held constant (corresponding to the entire V_1 - V_2 trajectory covered in 70, 100, or 130 ms) while eight tokens with different transition durations were presented 16 times each in random order. Again, the V_2 target was reached and the entire trajectory covered by only one token, the one having the longest transition duration; the durations of the transition in the other seven tokens were 7/8, 6/8, 5/8, etc.

of that of the longest-duration token. Figure 1(b) shows frequency-time trajectory plots along the /ai/ continuum for four of the eight variable transition duration tokens used.

Experiment 3. In experiment 3, the perception of transitions along the /ie/ trajectory was investigated. Since the natural American English vowel /e/ has formant frequencies very close to those lying halfway on the trajectory between /i/ and /a/, the formant frequencies chosen for the target vowel V_2 in this experiment were chosen to be exactly at the midpoint of the /ia/ trajectory used in experiments 1 and 2. In experiment 3a, the stimulus was similar to that of experiment 1: a 100-ms initial vowel V_1 , /i/, was followed by a constant-duration (30, 60, or 90 ms) transition toward the target vowel V_2 , that is, /e/. Again, in each run with a given transition duration, eight tokens with different transition velocities were presented in random order. In experiment 3b, as in experiment 2, three transition velocities were used: those corresponding to a fully completed /ie/ transition having a duration of 70, 100, and 130 ms. Just as in experiment 2, in each run in the second condition there were of eight tokens having the transition portion of the stimulus truncated to different durations.

Perceptually, all stimuli in all experiments (even, to some extent, the highest velocity and longest duration tokens) sounded like truncated diphthongs with the truncation percept somewhat attenuated by windowing the envelope of all stimuli with 10-ms cosine-square on- and off-ramps.

B. Subjects

The same eight subjects between the ages of 19 and 29 participated in experiments 1, 2, and 3. A crew of six subjects, four of whom took part in the main experiments, were listeners in the control experiment. They were recruited through advertisement and were paid hourly wages. All participants had pure-tone thresholds of 15 dB HL or better at all frequencies in the range from 250 to 8000 Hz, in both ears. The subjects gave informed consent to participate in the study and the procedures conformed to those mandated by the Institutional Review Board of the VA Northern California Health Care Systems. All subjects had prior experience as listeners in other psychoacoustic tests but were unsophisticated with respect to phonetic awareness.

C. Procedure

Control of the experiments as well as collection and analysis of the data were accomplished by a PC using an integrated psychoacoustic data collection and analysis software developed at our laboratory. Listeners were tested one at a time. Prior to the beginning of each block of trials, they were presented the target vowel V_2 ten times in succession and were told that the vowel stimuli in the following block will glide toward this target. For example, if a given block of trials investigated the perception of the /au/ trajectory, the target was the /u/ vowel; this vowel was the one presented in isolation ten times before the block. The listeners' task was to rate on a four-point scale their confidence that this target was actually reached, with "1" being very sure and "4" not at all. They were encouraged to use all four categories within

each block of trials. They were seated in a sound-treated testing booth and gave their rating responses by pressing one of four horizontally aligned labeled buttons on a special response box. In each trial the stimulus token was presented only once, i.e., repeated listening was not allowed. The listeners were allowed to respond any time (minimum 600 ms) after the termination of the stimulus; their key press response initiated the next trial. Presentation of the stimuli was done through earphones at a maximum instantaneous level of 80 dB sound pressure level. In all experiments, in each block of trials each of the eight tokens occurred 16 times in random order. For each condition and each subject, the results reported below are based on the average of six to eight blocks of 128 trials (8 targets times 16 repetitions per target).

D. Control experiment

Stimuli and targets in experiments 1–3 were synthesized vowels. In order to determine whether overestimates observed for the perception of VR would be restricted to stimuli recognized as speech or whether similar results could be also obtained for nonspeech stimuli, a control experiment was conducted. Stimuli for the control experiment were designed to maintain certain features of the stimuli used in the main experiments: a 100-ms steady-state first portion with a given frequency profile was followed by a glide toward a target—a sound with a different frequency profile. The target sound was explicitly identified to the listeners as such and presented ten times before the block of trials. The listeners' task was to indicate their confidence that the glide reached the target. Inspired by a classic study (Brady *et al.*, 1961), stimuli were sinusoidal complexes generated by passing a pulse train through a single resonator having its resonant frequency kept constant for 100 ms at 2452 Hz and then linearly swept toward 1078 Hz at a rate and duration that varied from condition to condition. Thus, stimuli of the control experiment were buzz-like sounds with some vowel-like characteristics; they were variants of the /ia/ stimuli used in experiments 1, 2, and 3, with only the second formant resonator excited. Two constant transition duration conditions (30 and 90 ms) with eight different glide velocities and two constant transition velocity conditions (velocities needed to reach the target in 70 and 130 ms) with eight different durations were investigated. Pilot listening tests with /ai/-like upward going single-formant transitions yielded results similar to the /ia/-like control experiments.

E. Data analysis

The objective of the data analysis was to assess the transition duration and transition velocity boundaries at which the subjects heard the presence of the target vowel, based on their confidence rating responses. Analysis of the results revealed that the subjects' rating criteria were not identical: some were tilted toward a stricter and others toward a more lax judgment regarding the perceived presence of the target. In order to determine the duration and velocity boundaries while neutralizing the effect of the subjects' idiosyncratic response criteria, the rating scale receiver-operating characteristic (ROC) method was used (Egan *et al.*, 1959). Rating

scale ROC analysis represents performance across a number of different strict-to-lax response criteria which are gauged by asking the subject to rate his/her confidence that the signal was present at a given trial. The rating ROC curve plots performance as the cumulative proportion of a certain confidence rating or stricter, given that the stimulus was the signal, as a function of the cumulative proportion of the same rating or stricter given that the stimulus was not the signal. The assumption underlying ROC analysis is that some portion (most often 50%) of the stimuli presented in a block contains a stimulus recognized by the subject as the signal, whereas the other stimuli are regarded as “noise”—in the narrow or broad sense—thus dividing the stimulus ensemble into two categories. Since in the present experiments the task was to judge the similarity of the final formant frequencies of the transition and those of the target, meaning that the “real signal” (i.e., the steady-state target vowel V_2 as presented before each block) was never actually present, a different definition of signal and noise was needed. Because in each condition the final frequencies of the eight stimulus tokens, from Token 1 to Token 8, got closer and closer to V_2 , we could divide the tokens into signal and noise subsets on the basis of similarity to V_2 in seven ways, each of which yielding an ROC curve. Thus, the Token 1 plus Token 2 subset is the least similar to the target but Token 2 is more so than Token 1—i.e., in this subset Token 2 is the signal and Token 1 the noise. Similarly, in the Token 1 through Token 3 subsets Token 3 is the most similar to the target and becomes the signal while the ensemble of Tokens 1 and 2 will be the noise. As we go on this way to the last, i.e., the entire set of the eight tokens in which Token 8 will be the signal and the ensemble of all others the noise, we will have created seven partially overlapping sets in each of which the signal and the noise are defined. While it is true that the seven noise categories are increasingly overlapping, such overlap is not uncommon in signal detection analysis of categorical data where common false alarm pools are used for different hit (i.e., signal) categories (see e.g., [Anderson and Neill, 2002](#)). Thus, from the seven signal-noise sets seven ROC curves are generated. Each curve plots the cumulative proportion of a given rating being higher than or equal to “highly confident that V_2 was not reached” for a given signal-noise set; when represented on a z-scale, the relationship of these proportions can be approximated by a straight line. Since there is no basis to assume that the variance of the signal and noise distributions is identical (and therefore the ROC curve is parallel to the diagonal), we followed the method suggested by [MacMillan and Creelman \(1991, pp. 65ff\)](#) and calculated the ROC assuming that the slopes were other than unity. From the ROC curves we calculated, for each condition and each subject, Egan’s d_e ([Egan, 1975](#)) and the response bias β . An example of the seven z-transformed ROC curves is illustrated in Fig. 2(a) (the data represent the average results of the eight subjects in one of the constant transition velocity conditions). According to signal detection theory, detection performance expressed as d_e is the distance from the origin to the point at which the ROC curve crosses the negative diagonal. Therefore, the points at which the seven ROC lines in Fig. 2(a) intersect the negative diagonal indicate the de-

FIG. 2. Panel A shows one (constant transition velocity) condition as an example to display the seven ROC curves representing all subjects’ average rating responses of the eight tokens k , truncated to different durations along the trajectory, heard as having their transition reach the target vowel. Each of the seven straight lines was fitted to three cumulative “positive” rating response probabilities as a function of the three corresponding cumulative “negative” response probabilities, where the positive responses signify the subject giving a certain rating when the token number was k and the negative responses mean that the subject gave the same rating to token numbers less than k . Panel B illustrates the points at which the seven positive response tokens intersected the negative diagonal in Panel A. The $1.41d_e$ intersection point is considered as the VR boundary, i.e., the token value corresponding to the subject saying that he/she heard the transition reaching the target vowel at the $d_e=1.41$ level. Thus, tokens with serial numbers higher than indicated by this boundary always led to the subject perceiving that the target vowel was, in fact, reached.

gree to which the “signal” token and the “noise” tokens could be discriminated with respect to the presence of the target vowel at the end of the transition. Distances of these points from the origin, i.e., the d_e values, are re-plotted as a function of the token number in Fig. 2(b), with the abscissa indicating the stimulus value—transition duration in this example—corresponding to the token numbers. (A nonlinear regression line fitted to these points illustrated as the curve in Fig. 2(b) accounts for 89% of the variance.) Except for error, this d_e scale, therefore, is monotonically related to the likelihood that the listener perceives the final formant frequencies of the transition trajectory as being those of the V_2 vowel. The abscissa value at which the curve in Fig. 2(b) crosses the d_e value of 1.41 was adopted as the performance level at which the listener judged, with fair certainty, the final formant frequencies of the stimulus as being those of the target.⁴ As shown in Sec. III, for most experimental conditions the boundary corresponded to a token with a trajectory shorter than that of the eighth token (i.e., the only one of the eight tokens to reach the target frequencies), thus indicating an overestimation of the trajectory length—or VR, as it was described earlier. Consequently, all other trajectories with lengths between the one corresponding to the boundary and that of the eighth token are bound to lead to perceptual overestimation, i.e., the area under the curve between the bound-

TABLE I. Steady-state formant frequencies of the four vowels used in the experiments and their corresponding BM distance from the helicotrema.

Vowel	F1 (Hz)	F2 (Hz)	F1 BM	F2 BM
			pos. (mm)	pos. (mm)
/a/	729	1079	12.24	14.63
/i/	239	2452	6.48	20.02
/u/	271	640	4.03	11.48
/e/	533	1959	10.47	18.51

TABLE II. Trajectory lengths and F1-F2 vector sums of the four V_1 - V_2 vowel pairs used in the experiments, in Hertz and in corresponding BM millimeter distance.

Trajectory	F1 (Hz)	F2 (Hz)	F1 BM distance (mm)	F2 BM distance (mm)	F1-F2 vector sum	F1-F2 BM distance vector sum
/ai/	490	1373	5.76	5.39	1458	7.89
/au/	458	439	5.20	3.16	634	6.09
/iu/	21	1812	0.55	8.55	1812	8.57
/ie/	294	493	3.96	1.51	574	4.24

ary line and the eighth is the one of VR. Since essentially for all experimental conditions this abscissa crossing occurred between two tokens, transition durations or velocities corresponding to the VR boundary were obtained by linear interpolation between those corresponding to the tokens on either side.

Similar ROC analyses were performed for the data obtained in each experimental condition to obtain estimates for boundaries of transition velocity (in experiments 1 and 3a) and transition duration (in experiments 2 and 3b). These estimates, by definition, minimize the effects of individual tendencies toward too strict or too lax certainty responses. Across all experimental conditions, intersubject variability of the d_e threshold estimates was 0.150 (0.077) while that of intersubject variability of the response bias β was 2.243 (1.012). In other words, the difference among listeners was almost 15 times larger as regards their strict-lax response criteria than as regards their ability (and proclivity) to associate a given transition duration/velocity combination with the target.

To tie the data on the perception of transitions to physical dimensions of the stimuli, the frequency change in the first two formants (F1 and F2) was selected. Although a transition between any two vowels represents simultaneous changes that involve all formants, because proportionally the largest changes occur in the first two formants (F1 and F2), and because F1 and F2 frequencies are sufficient to unam-

biguously specify a vowel, the results were gauged on the F1-F2 change in the transition. Since auditory sensitivity to frequency change as well as to formant frequency discrimination in vowels under ordinary listening conditions (Kewley-Port and Zheng, 1999)⁵ is a constant proportion of critical bands (Moore, 1973), we represented the extents of the formant changes in the transitions as F12, the vector sum of the F1 and F2 changes on a scale analogous to the Bark scale: the basilar membrane (BM) distance between the start and the end of the transition. The algorithm used was the one by Greenwood (1990)⁶ that estimates positions of maximum excitation on the BM from the helicotrema to the base. Frequencies and corresponding BM positions of the four V_1 and/or V_2 vowels included in the study are shown in Table I, whereas maximum transition lengths (i.e., trajectories between V_1 and V_2 vowels) as well as F1 and F2 vector sums are shown in Table II.

III. RESULTS

Combined results of all subjects in experiment 1, that is, transition velocities for constant-duration transitions estimated to yield VR boundaries, are illustrated in Fig. 3 for the three transition durations investigated. Panel A shows the velocity estimates for the /ai/ and /ia/ trajectories, panel B for the /au/ and /ua/ trajectories, and panel C for the /iu/ and /ui/ trajectories. With the exception of two among the eighteen

FIG. 3. Average results of the eight subjects in experiment 1, the constant transition duration experiment. On the abscissa: transition duration; on the ordinate: transition velocity boundary estimated with the ROC method illustrated in Fig. 2(b). This boundary corresponds to the lowest velocity of the $[F1, F2]$ vector's change (in BM mm/ms) at which the listeners perceived a complete V_1 - V_2 trajectory, i.e., heard the terminal formant frequencies of the transition glide as if they had reached those of the target vowel presented before the block of trials. The error bars stand for intersubject standard errors of the mean. The three panels illustrate thresholds for transitions proceeding in both directions along the three trajectories defined by the /a/-i/-u/ vowel triangle. The dashed line in each panel shows transition velocities across all durations at which the V_1 - V_2 trajectory is complete. Data points lying below this line indicate perceptual overestimate of the length of the transition's trajectory, i.e., perceived VR, and the vertical distance between the line and any data point gives the degree of overestimation.

FIG. 4. Average results of the eight subjects in experiment 2, the constant transition velocity experiment. On the abscissa: transition velocity in BM mm/ms; on the ordinate: transition duration corresponding to the VR boundary. The error bars stand for intersubject standard errors of the mean. The three panels illustrate VR thresholds for transitions of either direction on the three trajectories defined by the /a/-/i/-/u/ vowel triangle. The dashed line in each panel shows transition durations across all velocities at which the V_1 - V_2 trajectory is complete. Data points lying below this line indicate perceptual overestimate of the length of the transition's trajectory, i.e., perceived VR, and the vertical distance between the line and any data point gives the degree of overestimation.

conditions, intersubject variability was low to very low. The most striking feature of the data is that velocity estimates for the two opposing trajectory directions were essentially identical in each of the three panels. As shown by the heavy dashed lines in Fig. 3, at the 30-ms transition duration for all three trajectories and at the 60-ms transition duration for the /ai/-/ia/ and /au/-/ua/ trajectories, velocities estimated to be at VR boundary are below those of the trajectories actually reaching V_2 . One must therefore conclude that in those conditions a perceptual overshoot, indicating VR, occurred. At the longest, 90-ms, duration and at the 60-ms duration for the /iu/-/ui/ transitions the velocity estimates overlapped with the velocity of the complete V_1 - V_2 trajectory's velocity, indicating that in those conditions perception of the transition was almost veridical. In other words, for all three vowel pairs, VR was largest when the duration of the transition was the shortest.

Combined results of all subjects in experiment 2, that is, transition durations for constant-velocity transitions estimated to yield VR boundaries, are illustrated in Fig. 4 for the three transition velocities investigated. In Fig. 4, panel A shows these boundaries for the /ai/ and /ia/ trajectories, panel B for the /au/ and /ua/ trajectories, and panel C for the /iu/ and /ui/ trajectories. Duration estimates were very similar for the two opposing directions along all three trajectories, although not as perfectly overlapping as it was seen in experiment 1. Intersubject variability was even lower than in experiment 1. In all three panels velocities of V_1 - V_2 durations over the complete trajectory at the three constant transition velocities are also shown. Since the duration estimates always fell well below those of the durations of the complete V_1 - V_2 transitions (shown by the heavy dashed lines), it may be concluded that a strong perceptual overshoot leading to VR was manifest, with the strongest VR observed for the lowest of the three constant velocities.

Combined results of the eight subjects in experiment 3 are presented in Fig. 5. Panel B shows results of experiment 3a, with transition velocity at VR boundary as a function of transition duration, whereas results of experiment 3b are shown in panel C, with transition duration at VR boundary as a function of transition velocity. For comparison purposes, data for the /ia/ trajectory from experiments 1 and 2 are also

shown on the same graph. Panel B of the figure suggests that when listeners had to determine which transition of different velocities was most likely to reach the /e/ target, VR was absent; i.e., the perception of the /ie/ transition was veridical for all three fixed transition durations. In contrast, as shown in panel C, VR is clearly present at all three constant transition velocities of the /ie/ trajectory, although the perceptual overshoot is much smaller than the one observed for the /ia/ trajectory. One observation in panel C worth noting is the apparent continuity of the VR boundary durations for the transitions perceived to be /ia/ (on the right) and /ie/ (on the left). This continuity may be due to the fact that, as illustrated in the vowel triangle shown in panel A, the /ia/ and /ie/ trajectories were collinear with the /ie/ trajectory being half

FIG. 5. Average results of the eight subjects in experiment 3, the /ie/ transition experiment. Panel A shows the formant frequencies of the /i/, /e/, and /a/ vowels in the F1-F2 plane (in linear Hz), in order to illustrate the fact that the /ie/ and /ia/ trajectories are collinear and that the length of the /ie/ trajectory is $\frac{1}{2}$ of that of the /ia/ trajectory. Panel B: results of experiment 3a. On the abscissa: transition duration; on the ordinate: transition velocity corresponding to the VR boundary shown as open squares. For comparison purposes data for the /ia/ trajectory in experiment 1 [Fig. 3(a)] are also shown as filled circles. Velocities corresponding to the complete /ie/ trajectory are shown as the line with short dashes, and those corresponding to the complete /ia/ trajectory as the line with long dashes. Panel C: results of experiment 3b. On the abscissa: transition velocity; on the ordinate: transition duration corresponding to the VR boundary shown as open squares. Data for the /ia/ trajectory in experiment 2 [Fig. 4(a)] are shown as filled circles. Data points lying below the short-dashed lines indicate perceptual overestimate of the /ie/ trajectory, whereas those lying below the long-dashed lines overestimate of the /ia/ trajectory. The vertical distance between the short-dashed line and the open squares indicates the degree of overestimation for the /ie/ trajectory, whereas the distance between the long-dashed line and the closed circles the degree of overestimation for the /ia/ trajectory. Error bars in all panels represent intersubject standard error of the mean.

FIG. 6. Results of six listeners for the control experiment with /ia/-like glide stimuli but with only the second formant present and all others suppressed. Data from this experiment are shown as the open squares, whereas data for the corresponding full /ia/ stimuli (from experiments 1b and 2b) are represented as the solid circles. The left panel shows data for the constant transition duration, and the right panel for the constant transition velocity conditions. The short-dashed lines display velocities (left) and durations (right) corresponding to complete trajectories in the control experiment stimuli, whereas the long-dashed lines show velocities (left) and durations (right) corresponding to complete /ia/ trajectories from Fig. 3(a) and 4(a), respectively; data points below the lines indicate perceptual overestimation. Error bars represent intersubject standard errors of the mean.

of the length of the /ia/ trajectory. The interesting data points are the nearly overlapping ones in the middle: the duration estimate for the lowest /ia/ velocity (chosen to reach the /a/ target in 130 ms) is almost on top of the one for the highest /ie/ velocity (chosen to reach the /e/ target in 70, i.e., close to 130/2 ms). This overlap means that, for that particular velocity and at apparently similar (61–66 ms) durations, the transition can be perceived to lead either to the /e/ or to the /a/ vowel, depending on which one was designated as the target, played ten times pre-block, and remembered as such.

Results of six listeners in the control experiment are shown in Fig. 6 with the constant-duration conditions on the left and the constant-velocity conditions on the right as the open squares. In addition to the velocity estimates (on the left) and duration estimates (on the right) of the stimuli judged to have their terminal formant frequency reach 1078 Hz, data for the perception of the /ia/ trajectory in experiments 1 [Fig. 3(b)] and 2 [Fig. 4(b)] are also displayed for comparison. Overestimation of the trajectory's length is assessed as the difference between the velocity or duration estimates at the VR boundary and those corresponding to the complete trajectories displayed as the heavy dashed lines. While for constant-duration transitions, no overestimation of the trajectory length was observed (the difference between the velocities of full trajectory and judged trajectory at the 30-ms duration is not significant); the duration estimates for the constant-velocity conditions are significantly below those of the complete physical trajectory (the short-dashed line), thus indicating that the length of the trajectory was overestimated. Since for the /ia/ vowel stimuli the trajectory was overestimated for all conditions (illustrated by the distance between the data points and the long-dashed line that represents the target values), although less for the constant duration than the constant-velocity conditions, we have to assume

that when all formants change simultaneously along their /i/ toward /a/ trajectory, the totality of simultaneous formant transitions generates a percept in which the course of frequency change is more overestimated than when the second formant alone was present. But, despite of the lack of overestimation in the constant-duration control condition, this experiment suggests that perceptual overestimation of a monotonic frequency change can also occur for complex nonspeech sounds, i.e., it is not the exclusive property of vowels.

IV. DISCUSSION

A. Perception of vowel-to-vowel transitions: A model

Listeners in the experiments reported here were given the task of finding either the velocity for constant-duration transitions or the duration for constant-velocity transitions that were perceived to lead to a target vowel specified preceding the trial block. Since knowing both the duration and the velocity of a transition can reveal the extent of the trajectory, the difference between the starting and final frequencies, i.e., the frequency trajectory traveled, represents a perceptual cue—dimension ΔF , the format change⁷—in addition to the duration t and the velocity v of the transition. The question arises: on which dimension or dimensions of the stimulus did the listeners base their decision? Because the duration of the steady-state vowel V_1 was 100-ms, i.e., long enough to form a perceptual unit (Hirsh, 1974), the transition portion of the stimuli was perceptually distinct, allowing us to focus our attention on perceptual cues of the transition alone. Since the subjects were instructed to gauge the extent of the transition trajectory and compare its final formant frequencies to those of the target, the explicitly stated cue was ΔF —a cue reported to be the most salient in the discrimination of formant transitions (van Wieringen and Pols, 1995).

Comparing the results in Figs. 3 and 5(a) with those in Figs. 4 and 5(b) suggests that, for constant-duration variable velocity transitions, the perceptual overshoot was present only when the transitions were brief, whereas for transitions of constant-velocity and variable duration it was consistent and often substantial. This discrepancy may be at least partially attributed to the different salience of these two cues, suggesting that a constant velocity v in trial blocks with randomly mixed durations t provided a more salient cue than the opposite, i.e., constant-duration stimuli with different velocities. Such larger overshoots for rapid and brief transitions were also observed by Nabelek and co-workers for the perceptual boundary of diphthongized transitions (Nabelek et al., 1994). The finding is also consistent with results reported by van Wieringen and Pols (their experiment 2, 1995) as well as with a study showing that, while discrimination of frequency change velocities in the absence of other cues for sinusoidal sweeps is possible, it is also rather inefficient (Divenyi, 2005). In contrast, this latter study as well as others (e.g., van Wieringen and Pols, 1994) demonstrated that transition duration discriminability of pure tone sweeps and formants in speech stimuli are comparable (6–10% of the base duration).

FIG. 7. Incomplete F1-F2 trajectory length necessary to generate VR normalized by the length of the complete trajectory, for the six trajectories in experiments 1 and 2 and the trajectory in experiment 3. Solid and open bars stand for the constant transition duration and constant transition velocity experiments, respectively, whereas the descending and ascending hatched bars display respective constant-duration and constant-velocity predictions of the model in Eq. (2). While the trajectory lengths are reasonably similar along the two directions of a transition between any two vowels, they differ significantly between trajectories. VR is perceived for all trajectories except the constant-duration condition for the /ie/ trajectory. Error bars represent intersubject standard errors of the mean.

With no exception, all panels in Figures 3–5 display a negative slope of the perceptual data: at the VR boundary, increasing transition durations t are associated with decreasing transition velocities v and vice versa. Such trend hints at the existence of a t - v trade-off. However, a strict trade-off means that the product of transition duration and velocity, i.e., the trajectory covered, would have to be constant for all V_1 - V_2 trajectories and across all conditions, which was not the case in the present data. In fact, illustrated as the dark and the empty bars in Fig. 7 for the constant-duration and constant-velocity conditions, respectively, trajectory lengths (the normalized lengths of the $|F1, F2|$ frequency change vectors, i.e., ΔF , perceived as the complete trajectories) were not only consistently shorter for the constant-velocity than the constant-duration experiments but they were also different across the different vowel pair trajectories. Obviously, a strict trade-off $\Delta F = tv$ model does not account for the data. To better understand the auditory processes responsible for the perception of the complete and incomplete formant transitions in the present stimuli, a different model had to be found.

To search for an adequate but simple model, it was first assumed that we could find a linear combination of maximum trajectory (the V_1 - V_2 trajectory leading to the target), velocity at VR boundary, and transition duration at VR boundary that explained a large proportion of the data variance. A stepwise regression analysis that included linear and quadratic terms found that the model

$$\Delta F_{\text{incompl}} = C + \Delta F_{\text{compl}} + t + v \quad (1)$$

(where C is a constant, ΔF_{compl} is the complete V_1 - V_2 trajectory, t and v are the transition duration and velocity, respectively, and $\Delta F_{\text{incompl}}$ is the incomplete trajectory length perceived to be equal to ΔF_{compl}) accounted for 0.789 of the variance [$F(3, 38) = 52.18$, $p < 0.0001$], although the model fit was only moderate (Durbin–Watson D statistic of 2.103). To look for a better fit, some basic assumptions of the model had been modified. One previous finding pointed to the high salience of acceleration, specifically to that of a steady-state

frequency suddenly starting to change into a monotonically increasing or decreasing one (Divenyi, 2005). If such an acceleration monitor were present, it could possibly explain the large perceptual overshoot in the Fig. 3 data at the shortest fixed duration. For this, it was assumed that the stimuli were processed through a running 66-ms asymmetric window (half-power width of 34 ms)⁸ and the running average of velocity and acceleration were computed. A stepwise regression analysis that included linear and quadratic terms found that the model

$$\Delta F_{\text{incompl}} = C + \Delta F_{\text{compl}} + v_{\text{runav}} + a_{\text{runav}} \quad (2)$$

(where v_{runav} and a_{runav} are the velocity and the acceleration of the transitions calculated with the running averaging time window) accounted for 0.964 of the variance [$F(3, 38) = 365.02$, $p < 0.0001$] and displayed a good fit (Durbin–Watson D statistic of 2.486). The model output and the data shown in adjacent columns in Fig. 7 suggest that the target trajectory cue combined with the velocity and acceleration of the formant transition processed by a realistic window provide a reasonably good explanation of the perception of vowel-to-vowel transitions, including the perceptual overshoot. The presence of acceleration among the factors, despite the fact that the stimuli contained only linear transitions, is consistent with results by Nabelek and Ovchinnikov (1997) on overestimation of CVC trajectories using nonlinear transitions.

B. Perception of VR: An auditory explanation

The model thus shows that the extent of incomplete trajectory heard as the complete trajectory is a function of (1) the length of the complete trajectory connecting the initial and target vowels, (2) the velocity of the transition, and (3) the acceleration of the transition. The velocity and acceleration components are consistent with the existence of a putative frequency velocity detector/monitor (Divenyi, 2005; Kay and Matthews, 1972; Pollack, 1968), suggesting a frequency differentiation process. Due to the time course of the formation of excitatory and inhibitory contours in spectrotemporal receptive fields (STRFs) (David et al., 2009), such velocity detector should account for an overshoot in the perception of formant transitions and the spectral asymmetry in labeling the vowel in jVj or wVw syllables, as suggested by Linblom and Studdert-Kennedy (1967) and confirmed by others (Holt et al., 2000; Nabelek and Ovchinnikov, 1997). The fact that no asymmetry was found in the present data (see in Figs. 4 and 5 the practically colinear pairs of curves for either direction of the transition along the trajectories) does not invalidate the peripheral explanation, it only points out that V_1 - V_2 transitions may differ from CVC transitions in ways that should be examined in the future. While the build-up and decay of excitatory and inhibitory contours in STRFs is a continuous process, the sampling of frequencies at times coinciding with glottal pulses and making judgments from predictions based on a mechanism that averages instantaneous frequency samples, i.e., a frequency integration process, has also been proposed (Brady et al., 1961). Thus, if duration cues are not available or are uncertain, the rate of

frequency change can be discriminated by sampling the instantaneous frequency at the beginning and end of the transition and taking the difference between them (Dooley and Moore, 1988; Pollack, 1968). In fact, a frequency sampling/integration hypothesis may be consistent with the present data: since they show larger overestimation for constant velocity and variable duration formant sweeps than for sweeps with variable velocity and constant duration, the percept of a prolonged trajectory is more likely to be induced by stimuli with a fixed rate of formant change than with a fixed duration. Chistovich as well as Feth (Chistovich, 1985; Chistovich *et al.*, 1979; Feth *et al.*, 2006) proposed a variant of the frequency sampling hypothesis, a “center of gravity” effect, consisting of averaging samples of instantaneous frequency, to account for the perception of vowels as well as CV and V_1V_2 formant transitions. Trying to account for the present results by postulating a frequency sampling/averaging runs into the problem that the length of the trajectory perceived to reach the target is at times three-to-four pulse durations shorter than the complete trajectory, i.e., frequencies contributing to the percept are not part of the sample. Thus, a frequency differentiation, i.e., velocity detector, explanation could be given more credence. Nevertheless, in either case the results appear to be interpretable by invoking auditory processing. The fact that the single-formant nonspeech stimuli of the control experiment also induced overestimation of the trajectory (although to a lesser degree than the synthesized four-formant speech stimuli in the main experiments) supports an interpretation that does not need to rely on linguistic processes.

Auditory processing also constitutes the cornerstone of the spectral contrast theory that explains effects of coarticulation and other stimulus context on phonetic labeling (Leek *et al.*, 1987; Liberman *et al.*, 1957; Lotto *et al.*, 1997). This theory could also cover the present data, except for one element: memory. Although memory in the theory is not altogether absent—spectral contrast is based on memory for spectrum within a syllabic-range time scale (as in CVC stimuli) or between different items in a perstimulatory or poststimulatory time range short enough to be part of a single trial (Holt *et al.*, 2000)—remembering the target presented to listeners prior to the block of trials in the present study would require them to have memory lasting several minutes (if the possibility of linguistic encoding is not taken into account) or be altogether part of the life-long language acquisition process (if one assumes that the target triggers the memory of a long-ago encountered phonetic exemplar). The memory issue is perhaps the most troubling when attempting to give experiment 3 a spectral contrast explanation: transitions along the /ia/ trajectory elicited different responses depending on whether the target was /a/ or /e/, which is possible only if the contrast is explicitly defined as one between the stimulus and the memory of a target heard up to 5 min earlier (128 trials times 2.5 s, the average trial duration). Therefore, spectral contrast could provide an auditory explanation for the results but only if it were to be supplemented with a long term memory process, one which is outside the realm of a purely auditory domain.

A more general perceptual explanation is also a possi-

bility. The formant transitions having endpoints the subject must evaluate represents an instance of monotonic motion the perception of which is likely to be governed by the Gestalt principle of continuity (Koffka, 1935). The existence of such “perceptual inertia” has been demonstrated over the years mainly in the visual modality (e.g., Krekelberg and Lappe, 1999), though comparable instances in audition have also been reported. For example, when an upward frequency glide is broken by a noise burst that also marks the onset of a similar glide shifted in frequency, the first glide is perceived as if it were continuous (Nakajima *et al.*, 2000); the threshold of a probe tone forward-masked by a broad-band noise burst following a pure tone glide is lower at the frequency that lies on the extrapolated path traversing, but not physically present in, the masking noise (Crum and Hafter, 2008). Further support to the inertia notion is provided by results showing high sensitivity to acceleration of a sinusoid’s frequency change (Divenyi, 2005), consistent with the significant acceleration component in the model of Eq. (2). It is therefore possible that the perceptual extrapolation of formant transitions toward an expected target is just part of an ensemble of processes through which the nervous system enables a person to function in a dynamically changing environment.

C. Perception of VR: An auditory explanation?

Despite its plausibility, there are two problems with an exclusively auditory explanation of VR perception. First, an exclusively auditory interpretation of the results of the control experiment is tenuous because the overestimation of vowel trajectories was consistently larger than overestimation of single-formant glides which, in one of the conditions, was altogether missing. Although these differences could be due to nonlinguistic auditory processes—the target, presented pre-block, could have been kept in memory regardless of whether it was a vowel or a buzz having a given spectral prominence—an auditory explanation would require the simultaneous existence of a memory that is better for vowels than for buzzes.⁹ This memory requirement may become a bigger problem if one assumes that (i) compensation for VR routinely occurs and (ii) Eq. (2) representing VR magnitude also holds when a listener is presented real-life speech, for how else would the listener know what the putative target is without having access to his/her stored language references? Second, noting that the vowel-plus-transition stimuli are present in the repertoire of speech sounds generated by articulatory gestures, one cannot exclude the possibility that, at a higher cortical level the stimuli transmitted by the auditory system will evoke the very gestures that would have to be produced by the articulatory mechanism. That is, an auditory model would have to explain activity in the speech-motor cortical areas during speech perception, documented by recent fMRI and TMS studies (D’Ausilio *et al.*, 2009; Fadiga *et al.*, 2002; Pulvermuller *et al.*, 2006). Because the mechanisms of speech production and of speech perception have long been considered to be optimally matched complementary systems (de Cordemoy, 1668; Guenther *et al.*, 1998), an articulatory-gestural interpretation of the VR-bound percep-

tual overshoot should represent a complement, rather than an alternative, to an auditory-acoustic interpretation. Viewing the two not as opposite but as complementary has already led to discoveries by computational speech scientists for the development of articulatory-based low rate speech coding (Atal *et al.*, 1978) and for automatic speech recognition (Rose *et al.*, 1994). Another area of auditory-linguistic exploration of VR perception, not yet undertaken, could be the study of the phenomenon in other languages and across languages.¹⁰

V. CONCLUSION

The perception of VR prolongs the perceived extent of the formant transition leading to a vowel. This phenomenon affords the perceptual system to compensate for the articulatory undershoot of a target vowel's formant frequencies that often occurs in everyday conversational speech, in rapid speech, and in diverse phonetic context. The present experiments identified some rules underlying such compensation and suggest that they may be valid not only in the diphthong-type context investigated but could also apply to the perception of CV and VC transitions. The perception of VR is consistent *both* with current understanding of auditory mechanisms involved in the processing of monotonic frequency changes *and* with a mechanism tracking articulatory gestures that produce such changes, which situates the phenomenon in the general framework of speech dynamics.

ACKNOWLEDGMENTS

The author thanks Alvin Liberman, René Carré, Jean-Sylvain Liénard, Björn Lindblom, Brian Gygi, Associate Editor Rochelle Newman, and four anonymous reviewers for their helpful comments and advice. Portions of the paper were presented at the 135th meeting of the Acoustical Society of America and the 16th International Congress on Acoustics in Seattle, WA, in June 1998. The research was supported by Grant No. R01-07998 from National Institute on Aging, a grant from the Air Force Office of Scientific Research, and by the Veterans Administration Biomedical Laboratory Research.

¹For example, de Cordemoy, 1668.

²Although Nabelek and co-workers studied the effect of transition duration and transition velocity on the boundary for /a/ to /aI/ diphthongization (Nabelek *et al.*, 1994), their main focus was the effect of noise and reverberation in normal-hearing and hearing-impaired listeners and included only a limited parametric investigation of the trajectory, duration, and velocity of formant transitions.

³Although these vowel combinations sound like diphthongs, they do not necessarily figure in the repertory of diphthongs in American English. They were chosen because they encompass the longest possible trajectories in the vowel space and thus best allow for studying the perception of transitions along those trajectories.

⁴We say "fair certainty" because this boundary is stricter than the 50% subjective equality criterion most often used for the definition of phonetic category shifts, including the one by Nabelek and colleagues (Nabelek *et al.*, 1993) in their study on the shift of vowel identity in CVC settings.

⁵In their paper, Kewley-Port and Watson (1994) measured discrimination of either F1 or F2 in synthetic vowels. Although they proposed a two-tier function of the formant difference limen (DL) as a function of formant frequency, with the DL remaining constant up to about 1100 Hz and the

Weber fraction constant thereafter, the variability of the data permit representing the DL as a linear function of formant frequency, i.e., as a constant Weber fraction.

⁶Greenwood, perhaps more than any other researcher, devoted much time and effort to defining and clarifying the auditory bandwidth's dependence on frequency, debunking on his way several widely accepted theories, such as the Mel scale (e.g., Greenwood, 1997).

⁷The symbol ΔF is to be interpreted as a matrix containing all salient formants, specifically F1 and F2.

⁸Other window durations were also tried but, when inserted in the various models, they accounted for less of the variance and resulted in a lower Durbin-Watson index of fit.

⁹Unfortunately, the subjects were not asked how they remembered the target, so the possibility exists that they could have used a phonetic image.

¹⁰The vowel space used in the present study may differ from those of other languages, such as the relatively restricted space of Arabic or the complex space of Swedish, to cite only two examples. It is, therefore, quite likely that changing the stimulus space to one with Mandarin vowels and repeating the study in Beijing with Mandarin speakers and listeners (L2-L2), or using native Mandarin speaking listeners with the present stimuli (L1-L2), would have yielded different results. However, since the study's aim was to characterize the *mechanism* responsible for the perception of transitions, and since such mechanism is likely to be shared by people regardless of their language background, it is likely that the conclusions to which such L2-L2 or L1-L2 results would point would be similar to those proposed here. The generalization should hold even for languages not having any V1-V2 transitions because the present results point to the duration-velocity field of CV and VC transitions present in practically every language.

Anderson, C. J., and Neill, W. T. (2002). "Two bs or not two Bs? A signal detection theory analysis of repetition blindness in a counting task," *Percept. Psychophys.* **64**, 732–740.

Andruski, J. E., and Nearey, T. M. (1992). "On the sufficiency of compound target specification of isolated vowels and vowels in /bVb/ syllables," *J. Acoust. Soc. Am.* **91**, 390–410.

Assmann, P. F., and Katz, W. F. (2005). "Synthesis fidelity and time-varying spectral change in vowels," *J. Acoust. Soc. Am.* **117**, 886–895.

Atal, B. S., Chang, J. J., Mathews, M. V., and Tukey, J. W. (1978). "Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique," *J. Acoust. Soc. Am.* **63**, 1535–1555.

Brady, P. T., House, A. S., and Stevens, K. N. (1961). "Perception of sounds characterized by a rapidly changing resonant frequency," *J. Acoust. Soc. Am.* **33**, 1357–1362.

Carré, R., and Divenyi, P. L. (2000). "Modeling and perception of "gesture" reductions," *Phonetica* **57**, 152–169.

Carré, R., Pellegrino, F., and Divenyi, P. (2007). "Speech dynamics: Epistemological aspects," paper presented at the 16th International Congress of Phonetic Sciences, Saarbrücken, Germany.

Chistovich, L. A. (1985). "Central auditory processing of peripheral vowel spectra," *J. Acoust. Soc. Am.* **77**, 789–805.

Chistovich, L. A., Sheikin, R. L., and Lublinskaya, L. L. (1979). "Centres of gravity and spectral peaks as the determinants of vowel quality," in *Frontiers of Speech Communication Research*, edited by B. L. a. S. Ohman (Academic, London), pp. 143–157.

Crum, P. A., and Hafter, E. R. (2008). "Predicting the path of a changing sound: Velocity tracking and auditory continuity," *J. Acoust. Soc. Am.* **124**, 1116–1129.

D'Ausilio, A., Pulvermüller, F., Salmas, P., Bufalari, I., Begliomini, C., and Fadiga, L. (2009). "The motor somatotopy of speech perception," *Curr. Biol.* **19**, 381–385.

David, S. V., Mesgarani, N., Fritz, J. B., and Shamma, S. A. (2009). "Rapid synaptic depression explains nonlinear modulation of spectro-temporal tuning in primary auditory cortex by natural stimuli," *J. Neurosci.* **29**, 3374–3386.

de Cordemoy, G. (1668). *Discours physique de la parole (Physical Discourse on Speech)* (Bibliothèque nationale de France, Paris, France).

Di Benedetto, M.-G. (1994). "Acoustic and perceptual evidence of a complex relation between F1 and F0 in determining vowel height," *J. Phonetics* **22**, 205–224.

Divenyi, P. L. (2005). "Frequency change velocity detector: A bird or a red herring?," in *Auditory Signal Processing: Physiology, Psychology and Models*, edited by D. Pressnitzer, A. d. Cheveigné, S. McAdams, and L. Collet (Springer-Verlag, New York), pp. 176–184.

- Dooley, G. J., and Moore, B. C. (1988). "Duration discrimination of steady and gliding tones: A new method for estimating sensitivity to rate of change," *J. Acoust. Soc. Am.* **84**, 1332–1337.
- Egan, J. P. (1975). *Signal Detection Theory and ROC-Analysis* (Academic, New York).
- Egan, J. P., Schulman, A. I., and Greenberg, G. Z. (1959). "Operating characteristics determined by binary decisions and by ratings," *J. Acoust. Soc. Am.* **31**, 778–773.
- Fadiga, L., Craighero, L., Buccino, G., and Rizzolatti, G. (2002). "Speech listening specifically modulates the excitability of tongue muscles: A TMS study," *Eur. J. Neurosci.* **15**, 399–402.
- Feth, L. L., Fox, R. A., Jacevicz, E., and Iyer, N. (2006). "Dynamic center-of-gravity effects in consonant-vowel transitions," in *Dynamics of Speech Production and Perception*, edited by P. L. Divenyi, S. Greenberg, and G. Meyer (IOS, Amsterdam, The Netherlands), pp. 103–111.
- Flege, J. E. (1988). "Effects of speaking rate on tongue position and velocity of movement in vowel production," *J. Acoust. Soc. Am.* **84**, 901–916.
- Fowler, C. A. (1994). "Speech perception: Direct realism theory," in *Encyclopedia of Language and Linguistics*, edited by R. E. Ashler (Pergamon, New York), pp. 4199–4203.
- Fox, R. A. (1989). "Dynamic information in the identification and discrimination of vowels," *Phonetica* **46**, 97–116.
- Furui, S. (1986). "On the role of spectral transition for speech perception," *J. Acoust. Soc. Am.* **80**, 1016–1025.
- Gottfried, T. L., Miller, J. L., and Payton, P. E. (1990). "Effect of speaking rate on the perception of vowels," *Phonetica* **47**, 155–172.
- Greenwood, D. D. (1990). "A cochlear frequency-position function for several species—29 years later," *J. Acoust. Soc. Am.* **87**, 2592–2605.
- Greenwood, D. D. (1997). "The Mel scale's disqualifying bias and a consistency of pitch-difference equisections in 1956 with equal cochlear distances and equal frequency ratios," *Hear. Res.* **103**, 199–224.
- Guenther, F. H., Hampson, M., and Johnson, D. (1998). "A theoretical investigation of reference frames for the planning of speech movements," *Psychol. Rev.* **105**, 611–633.
- Halle, M., and Stevens, K. N. (1962). "Analysis by synthesis," *Proceedings on the Seminar on Speech Compression and Processing* (USAF Cambridge Research Center, Cambridge, MA), AFCRC-TR-59-198, Vol. 2, p. D7.
- Hillenbrand, J. M., and Nearey, T. M. (1999). "Identification of resynthesized /hVd/ utterances: effects of formant contour," *J. Acoust. Soc. Am.* **105**, 3509–3523.
- Hillenbrand, J. M., Clark, M. J., and Nearey, T. M. (2001). "Effects of consonant environment on vowel formant patterns," *J. Acoust. Soc. Am.* **109**, 748–763.
- Hirsh, I. J. (1974). "Temporal order and auditory perception," in *Sensation and Measurement: Papers in Honor of S. S. Stevens*, edited by H. R. Moskowitz, B. Scharf, and J. C. Stevens (D. Reidel, Dordrecht, The Netherlands), pp. 251–258.
- Holt, L. L., Lotto, A. J., and Kluender, K. R. (2000). "Neighboring spectral content influences vowel identification," *J. Acoust. Soc. Am.* **108**, 710–722.
- Jenkins, J. J., Strange, W., and Miranda, S. (1994). "Vowel identification in mixed-speaker silent-center syllables," *J. Acoust. Soc. Am.* **95**, 1030–1034.
- Jenkins, J. J., Strange, W., and Trent, S. A. (1999). "Context-independent dynamic information for the perception of coarticulated vowels," *J. Acoust. Soc. Am.* **106**, 438–448.
- Kay, R. H., and Matthews, D. R. (1972). "On the existence in human auditory pathways of channels selectively tuned to the modulation present in frequency-modulated tones," *J. Physiol. (London)* **225**, 657–677.
- Kewley-Port, D., and Watson, C. S. (1994). "Formant-frequency discrimination for isolated English vowels," *J. Acoust. Soc. Am.* **95**, 485–496.
- Kewley-Port, D., and Zheng, Y. (1999). "Vowel formant discrimination: Towards more ordinary listening conditions," *J. Acoust. Soc. Am.* **106**, 2945–2958.
- Kewley-Port, D., Burkle, T. Z., and Lee, J. H. (2007). "Contribution of consonant versus vowel information to sentence intelligibility for young normal-hearing and elderly hearing-impaired listeners," *J. Acoust. Soc. Am.* **122**, 2365–2375.
- Koffka, K. (1935). *Principles of Gestalt Psychology* (Harcourt Brace Jovanovich, New York).
- Krekelberg, B., and Lappe, M. (1999). "Temporal recruitment along the trajectory of moving objects and the perception of position," *Vision Res.* **39**, 2669–2679.
- Leek, M. R., Dorman, M. F., and Summerfield, Q. (1987). "Minimum spectral contrast for vowel identification by normal-hearing and hearing-impaired listeners," *J. Acoust. Soc. Am.* **81**, 148–154.
- Liberman, A. L., Harris, K. S., Hoffman, H. S., and Griffith, B. C. (1957). "The discrimination of speech sounds within and across phoneme boundaries," *J. Exp. Psychol.* **54**, 358–368.
- Lienard, J. S., and Di Benedetto, M. G. (1999). "Effect of vocal effort on spectral properties of vowels," *J. Acoust. Soc. Am.* **106**, 411–422.
- Lindblom, B. (1983). "Economy of speech gesture," in *The Production of Speech*, edited by P. F. MacNeilage (Springer-Verlag, New York), pp. 217–246.
- Lindblom, B., and Studdert-Kennedy, M. (1967). "On the role of formant transitions in vowel recognition," *J. Acoust. Soc. Am.* **42**, 830–843.
- Lindblom, B., Brownlee, S. A., and Lindgren, R. (1996). "Formant undershoot and speaking styles: An attempt to resolve some controversial issues," in *Sound Patterns of Connected Speech, Models and Explanation*, edited by A. P. Simpson and M. Pätzold (Institut für Phonetik und Digital Sprachverarbeitung, Universität Kiel, Kiel, Germany), pp. 119–130.
- Lotto, A. J., and Holt, L. L. (2006). "Putting phonetic context effects into context: a commentary on Fowler (2006)," *Percept. Psychophys.* **68**, 178–183.
- Lotto, A. J., Holt, L. L., and Kluender, K. R. (1997). "Effect of voice quality on perceived height of English vowels," *Phonetica* **54**, 76–93.
- Macchi, M. J. (1980). "Identification of vowels spoken in isolation versus vowels spoken in consonantal context," *J. Acoust. Soc. Am.* **68**, 1636–1642.
- Macmillan, N. A., and Creelman, C. D. (1991). *Detection Theory: A User's Guide* (Cambridge University Press, Cambridge, UK).
- Moore, B. C. J. (1973). "Frequency difference limens for short-duration tones," *J. Acoust. Soc. Am.* **54**, 610–619.
- Morrison, G. S., and Nearey, T. M. (2007). "Testing theories of vowel inherent spectral change," *J. Acoust. Soc. Am.* **122**, EL15–EL22.
- Mrayati, M., Carré, R., and Guérin, B. (1988). "Distinctive Region and Modes: a new theory of Speech Production," *Speech Commun.* **7**, 257–286.
- Nabelek, A. K., Czyzewski, Z., and Crowley, H. J. (1993). "Vowel boundaries for steady-state and linear formant trajectories," *J. Acoust. Soc. Am.* **94**, 675–687.
- Nabelek, A. K., Czyzewski, Z., and Crowley, H. J. (1994). "Cues for perception of the diphthong /aI/ in either noise or reverberation. Part I. Duration of the transition," *J. Acoust. Soc. Am.* **95**, 2681–2693.
- Nabelek, A. K., and Ovchinnikov, A. (1997). "Perception of nonlinear and linear formant trajectories," *J. Acoust. Soc. Am.* **101**, 488–497.
- Nabelek, A. K., Ovchinnikov, A., Czyzewski, Z., and Crowley, H. J. (1996). "Cues for perception of synthetic and natural diphthongs in either noise or reverberation," *J. Acoust. Soc. Am.* **99**, 1742–1753.
- Nakajima, Y., Sasaki, T., Kanafuka, K., Miyamoto, A., Remijn, G., and ten Hoopen, G. (2000). "Illusory recouplings of onsets and terminations of glide tone components," *Percept. Psychophys.* **62**, 1413–1425.
- Nearey, T. M. (1989). "Static, dynamic, and relational properties in vowel perception," *J. Acoust. Soc. Am.* **85**, 2088–2113.
- Nearey, T., and Assman, P. (1986). "Modeling the role of inherent spectral change in vowel identification," *J. Acoust. Soc. Am.* **80**, 1297–1308.
- Peterson, G., and Barney, H. (1952). "Control methods used in a study of the vowels," *J. Acoust. Soc. Am.* **24**, 175–184.
- Pollack, I. (1968). "Detection of rate of change of auditory frequency," *J. Exp. Psychol.* **77**, 535–541.
- Pols, L. C. W., and van Son, R. J. J. H. (1993). "Acoustics and perception of dynamic vowel segments," *Speech Commun.* **13**, 135–147.
- Pulvermüller, F., Huss, M., Kherif, F., Moscoso del Prado Martin, F., Hauk, O., and Shtyrov, Y. (2006). "Motor cortex maps articulatory features of speech sounds," *Proc. Natl. Acad. Sci. U.S.A.* **103**, 7865–7870.
- Rose, R. C., Schroeter, J., and Sondhi, M. M. (1994). "An investigation of the potential role of speech production models in automatic speech recognition," paper presented at the ICSLP-1994, Yokohama, Japan.
- Stevens, K. N., and House, A. S. (1963). "Perturbation of vowel articulations by consonantal context: an acoustical study," *J. Speech Hear. Res.* **6**, 111–128.
- Strange, W. (1989). "Dynamic specification of coarticulated vowels spoken in sentence context," *J. Acoust. Soc. Am.* **85**, 2135–2153.
- Strange, W., and Bohn, O. S. (1998). "Dynamic specification of coarticulated German vowels: perceptual and acoustical studies," *J. Acoust. Soc. Am.* **104**, 488–504.
- Strange, W., Edman, T. R., and Jenkins, J. J. (1979). "Acoustic and phono-

- logical factors in vowel identification," *J. Exp. Psychol. Hum. Percept. Perform.* **5**, 643–656.
- van Son, R. J. J. H., and Pols, L. C. W. (1993). "Vowel identification as influenced by vowel duration and formant track shape," paper presented at the Proceedings of Eurospeech-93, Berlin.
- van Wieringen, A. (1995). "Perceiving dynamic speechlike sounds," Doctoral dissertation, University of Amsterdam, the Netherlands.
- van Wieringen, A., and Pols, L. C. W. (1994). "Frequency and duration discrimination of short first-formant speechlike transitions," *J. Acoust. Soc. Am.* **95**, 502–511.
- van Wieringen, A., and Pols, L. C. W. (1995). "Discrimination of single and complex consonant-vowel- and vowel-consonant-like formant transitions," *J. Acoust. Soc. Am.* **98**, 1304–1312.
- Vishwanathan, N., Fowler, C. A., and Magnuson, J. S. (2009). "A critical examination of the spectral contrast account of compensation for coarticulation," *Psychon. Bull. Rev.* **16**, 74–79.
- Watson, C. I., and Harrington, J. (1999). "Acoustic evidence for dynamic formant trajectories in Australian English vowels," *J. Acoust. Soc. Am.* **106**, 458–468.
- Whalen, D. H., Magen, H. S., Pouplier, M., Kang, A. M., and Iskarous, K. (2004). "Vowel production and perception: Hyperarticulation without a hyperspace effect," *Lang Speech* **47**, 155–174.