Published in final edited form as:
Phonetica. 2000 ; 57(2-4): 152–169.

Modeling and Perception of ’Gesture Reduction’

René Carré[a] and **Pierre L. Divenyi**[b]

a*ENST, Unité Associée au CNRS, Paris, France*

b*Experimental Audiology Research, VA Medical Center, Martinez, Calif., USA*

Abstract

The phenomenon of vowel reduction is investigated by modeling ‘gesture reduction’ with the use of the Distinctive Region Model (DRM). First, a definition is proposed for the term gesture, i.e. an acoustically efficient command aimed at deforming, in the time domain, the area function of the vocal tract. Second, tests are reported on the perception of vowel-to-vowel transitions obtained with reduced gestures. These tests show that a dual representation of formant transitions is required to explain the reduction phenomenon: the trajectory in the F_1-F_2 plane and the time course of the formant changes. The results also suggest that time-domain integration of the trajectories constitutes an integral part of the auditory processing of transitions. Perceptual results are also discussed in terms of the acoustic traces of DRM gestures.

Introduction

Numerous observations on vowel formant characteristics have been reported [Chiba and Kajiyama, 1941; Peterson and Barney, 1952; House and Fairbanks, 1953; Fant, 1973]. These characteristics have been seen to vary as a function of speaker, style, prosody, and phonetic context. The observations have led to the finding that vowels produced with very different formant characteristics could be perceptually equivalent. Because this multiple-valued nature of vowels was recognized early on, it has been generally assumed [first implicitly and, much later, explicitly, see e.g. Kuhl, 1992] that the average formant values of vowels in isolation represent *targets* that the talker, during continuous speech, always strives, and often fails, to reach. Cases when vocalic targets are not reached have been termed, from the standpoint of production, articulatory *undershoot* [Brownlee, 1996; Lindgren and Lindblom, 1996] and, from the standpoint of acoustic trajectories, ‘vowel reduction’ [Lindblom, 1963]. Vowel reduction is observed, for instance, in fast speech (where the vocal tract is not given enough time to complete the required task and, therefore, it is denied the opportunity to have its shape conform to that of an intended target) as well as in ‘hypospeech’ [Lindblom, 1990]. The intended targets, then, are recovered by the listener either by way of a special compensatory mechanism that produces a ‘perceptual overshoot’ [Lindblom and Studdert-Kennedy, 1967] or by using previously learned and stored templates of vowel formant patterns with the undershoot effects included [Johnson, 1997]. While such a mechanism would predict a one-to-one correspondence between production and perception of unreached target vowels, some results appear to cast the shadow of doubt over this prediction. For example, in production, vowel targets can often be reached even in fast speech [Kuehn and Moll, 1976; Gay, 1978; van Son and Pols, 1992], while normal-rate speech does not necessarily ensure that the targets will be reached at all times [Nord, 1986]. The situation is not less ambiguous in the perceptual domain: under certain conditions, listeners seem to base their judgment on the *average* of the formant values covered by the trajectory. Since averaging the frequencies of a formant trajectory inevitably makes the percept of the final frequency to be displaced toward the initial

Dr. René Carré ENST, Unité Associée au CNRS 46, rue Barrault, F-75634 Paris Cedex 13 (France) Tel. (33) 1 45 81 71 90, Fax (33) 1 45 88 79 35 E-Mail carre@tsi.enst.fr

frequency, such cases yied a perceptual *undershoot* [van Son and Pols, 1993], rather than produce an overshoot stemming from a perceptual compensation for vowel reduction. Other authors argue that first formant transitions are averaged and second formant transitions lead to a perceptual overshoot [Huang, 1987; Di Benedetto, 1989]. The differences between the formant frequencies of carefully produced vowels (that is, those either uttered in isolation or found in laboratory speech) and those observed in spontaneous fluent speech suggest that while vowel reduction is a potentially precious tool for the study of speech phenomena having to do with *speech kinematics*, it is ill-suited for the study of steady-state vowels. Vowel-to-vowel transitions are part of these phenomena. Studying these transitions with vowel reduction views them from the perspective of an approach that, in agreement with Strange [1989], we favor [Carré et al., 1994].

Unfortunately, investigating the live production of vowel transitions and its perceptual consequences poses numerous and serious methodological problems. To bypass these, the method of simulating production by means of a realistic model has been used as an alternative to recorded natural tokens, as long as the model is regarded as capable of generating acceptable approximations. In the present paper, we will resort to such a simulation by way of a simple acoustic production model controlled by command parameters. The model in question, deduced from acoustic theory, is the Distinctive Region Model (DRM), which appears particularly well adapted for the study of the kinematics of speech production [Mrayati et al., 1988]. In this model, simple and efficient commands — similar to those actually realized during speech production — accomplish specific acoustic changes by some specific deformation of the vocal tract. We define efficiency as meaning that a small area deformation should lead to a large acoustic variation. As a consequence of the acoustic properties of a tube closed at one end, deformations applied to all points of the tube will not be equally efficient. Consequently, the model constrains the constrictions of the vocal tract to occur at points that are acoustically the most efficient. DRM commands corresponding to *dynamic* phonetic tasks are assumed to be speech gestures for vowel-to-vowel production. As the articulators are engaged in performing a certain constriction pattern, the area function of the vocal tract and its resonant characteristics are changed in some specific way. Thus, a speech gesture represents the transition in the time domain between two quasistationary segments of an utterance, i.e. the transition between an initial and a target constriction pattern and the corresponding initial and target area functions and their resonance characteristics. For example, the vowel-to-vowel transition /ai/ is described by only one gesture — the one for lingual constriction — while the French /ay/ is described by two gestures — the lingual constriction gesture and the lip gesture. It should be noted that gestures derived using the DRM are in close agreement with the speech gestures derived by Browman and Goldstein [1986] from empirical (i.e. articulatory) data. Because they are distinctive in their nature, DRM gestures also represent units of information analogous to the information units of Browman and Goldstein[1986] that arise from the coordination of articulators when they engage in constricting the vocal tract at specific regions.

The objective of the present article is to investigate the perception of vowel-to-vowel transitions generated by DRM gestures. From the results, we will attempt to determine whether the listener is able to infer the occurrence of a specific vocal tract deformation gesture from the presence of its acoustic consequence, namely, from the presence of a specific temporal-spectral pattern in the formant transition.[1] In the articulatory domain, deformation gestures acting on the area function of the vocal tract are intended to reach articulatory targets corresponding to acoustic targets. In the case of vowel reduction, we use the term ‘gesture reduction’ referring to the case when an articulatory target corresponding to an acoustic target is not reached. If the vocal tract deformation is the joint consequence of several individual gestures (referred to as coproduction

[1] Because the transitions evolve in both the time and the frequency domains, throughout the article we will refer to transitions in two coordinate systems. The first of these

[Kozhevnikov and Chistovich, 1965; Fowler, 1992]), then a reduction may take place in any or all of the component gestures. Therefore, it is essential to scan gestures individually both for the occurrence of reduction in each of the components and for the degree of their relative independence [Mattingly, 1990]. To be precise, vowel production may be said to rely on two largely autonomous supralaryngeal gestures: tongue constriction displacement and lip rounding. If the time course and/or pattern of synchronization of these two gestures are different, the resulting trajectory between a given initial and final vowel will also be different.

The series of experiments described below were designed to investigate the perception of formant trajectories in vowel-to-vowel transitions in situations where vowel reduction could be observed.

Methods

One major objective of the present investigations was to determine the extent to which the transition had to proceed along a given V_1-V_2 trajectory, in order for the listener to perceive the vowel V_2 at the endpoint of the trajectory. To preclude any interference from the recency effect, we resorted to using a $V_1V_2V_1$ stimulus similar to the one reported by Kuwabara [1985] and Beautemps [1993]. The major advantage of such tokens is that the V_1 vowel provides a firm reference anchor[2]. Our $V_1V_2V_1$ tokens were synthesized using the DRM controlled by two gestures: the tongue gesture and the lip gesture. The time course of the transition between two vowels was obtained through a cosine interpolation. The model's output consisted of a sequence of formant frequencies calculated using the algorithm proposed by Badin and Fant [1984]. This algorithm takes into account wall vibrations as well as losses of different origin affecting the vocal tract (e.g. heat, viscosity and the resistive part of the radiation impedance)[3]. The resulting formant frequencies were used to control the first three formants of a cascade formant synthesizer, whereas F_4 and F_5 were fixed at 4,000 and 5,000 Hz, respectively. Bandwidths of the first five formants were, in ascending order, 70, 100, 110, 150 and 200 Hz. Frame duration was held constant at 10 ms. The voice source was represented by a series of 0.1-ms pulses shaped by a second-order glottal filter (F = 100 Hz, BW = 300 Hz). F_0 was varied linearly from 120 Hz at the beginning of V_1, to 130 Hz at the beginning of V_2, and to 100 Hz at the end of V_3. For each experiment, a set of tokens was synthesized, each having different transition characteristics. Each block of trials consisted of ten random-order presentations of each token. Subjects were 5 native French listeners.Their tasks consisted of assigning, by button press, one of three or four labels to the second vowel V_2 in each token. The list of labels, corresponding to French vowels, was assembled by the first author following pilot listening tests. Each trial was initiated by the subjects's response to the previous one. Stimuli were presented through earphones at a comfortable listening level. A PC computer controlled all experiments. Results for each token were expressed as mean percent identification and intersubject standard deviation.

Experiment 1: Gesture Reduction for [iai]

In the first experiment (experiment 1a), the V_1 V_2 V_1 sequence has the [i] vowel for V_1 and vowels along the [ia] trajectory for V_2. Only one gesture was involved in the transition. The objective of the experiment was to measure the minimum displacement toward the [a] at which the listener will perceive with certainty the distant target vowel, just as if, in fact, it had been reached. Figure 1 illustrates the two successive gestures performed to obtain the constriction patterns that generate the desired [iV_2i] sequence. It is obvious that different degrees of constriction (indicated as the endpoint of the left solid arrow) will lead to different vowels

[2] In contrast, CVC tokens used by a number of investigators fail to provide a well-defined spectral reference on either side of the vowel ands thus raise the possibility of th

[3] In the computation, the total effective length L_e of the tube is identical to the physical length L_p plus the length corresponding to the radiation inductance L_r. L_p is a func

V_2. Using the DRM synthesizer, eight [iV_2i] sequences with different degrees of constriction were generated. The global command amplitude for the eight stimuli is illustrated as a function of time in figure 2a, where as figure 2b shows the corresponding F_1 and F_2 transitions. The value of F_3 obtained by the model is 2,944 Hz for the vowel [i] and 2,795 Hz when the target [a] is reached (for illustrative purposes, the F_3 temporal trajectory is shown in figure 2a for the token in which the target [a] is reached). The duration of the initial V_1 vowel was always 100 ms, whereas that of the final V_1 vowel was 150 ms; the duration of both the V_1V_2 and the V_2V_1 transitions was 100 ms. Two of the eight items actually reached the target[a] before returning to [i] (i.e. the duration of the target in these two items was 10 ms and 0 ms, respectively). In contrast, the six other tokens did not reach the [a] (and thus represent different degrees of reduction), with the first transition truncated at 10, 20, 30, 40, 50, and 60 ms *before* it would have reached the [a] target, had this transition been complete. Saying it differently, the onset of the gesture to return to [i] was advanced, i.e. it occurred at -10, - 20,..., - 60 ms with respect to the instant at which the target [a] would have been reached by a completed transition. (Throughout the article, the negative values in milliseconds refer to the time before the target would be reached, were the transition to continue its course.) These characteristics were chosen by assuming that the command for the first task was to reach the target by means of a quasi-ballistic motion, which was cut short by an opposite command before the target was reached. [In fact, we will see later that the precise choice of the movement between two targets is irrelevant (experiment 1c).] The formant trajectory in the F_1-F_2 plane is plotted in figure 2c. Note that the trajectory in this coordinate system is the same for all eight tokens except for the V_2 endpoint. The figure also attempts to show that the trajectory traverses regions corresponding to several French vowels. For example, formant values on the trajectory 40 ms before reaching [a], i.e. halfway between [i] and [a], correspond to those of the French vowel [ε].

Results of the perception tests are shown in figure 3a. The task of the listeners was to label the extreme vowel V_2 as one of the three French vowels /e/, /ε/, or /a/[4]. The specific aim of this experiment was to determine the extent to which listeners identify the extreme vowel V_2 as /ε/ or /e/ when the transition does not go all the way to the endpoint [a]. Figure 3a illustrates average results of 5 listeners showing the percentage of /iai/ judgments for the eight different degrees of [V_2i] transition cutback. As the figure indicates, the subjects made more than 80% /iai/ judgments for reductions corresponding to [V_2i] transitions ranging from 0 to up to - 30 ms *before* a complete [ia] transition would have been reached. The - 30 ms cutback corresponds to formant values of F_1= 607 Hz and F_2=1,803 Hz which are quite different from those characteristic to a typical [a]. For [V_2i] transitions returning to [i] more than 30 ms before completing the [ia] transition, first an /iεi/ and then an /iεi/ percept were reported. The standard deviations, shown as the error bars in figure 3a and most subsequent data graphs, carry more intersubject than intrasubject variability, suggesting the presence of individual differences as regards the internal reference for the vowels /a, ε, e, i/, even for subjects who belong to the same language group. A proof of this is illustrated in figure 3b that shows data of 1 individual listener averaged over five blocks of trials; the spread of these data is much smaller than the one seen in figure 3a for the data of all subjects combined.

But would a V_2, which was perceived as /a/ in an [iV_2i] context, also be perceived as /a/ when presented in isolation? The objective of experiment 1b was to answer this question. Stable vowels (200 ms in duration) having the same formant values as those of the V_2 endpoints of the transitions illustrated in figure 2c were synthesized and presented to the same subjects for labeling, using the three response categories identical to those in experiment 1a. Results shown in figure 3c indicate that, up to formant values corresponding to cutbacks not exceeding - 20

[4] In order to obtain a more complete picture of the phenomenon, the same experiments should be repeated with native speakers of other languages used as subjects. It is ex

ms, the vowel was predominantly perceived as /a/. Thus, a comparison of the results of experiments 1a and 1b suggests the presence of a perceptual overshoot of about 15%[5]. Note that the identification curves for /ε/ and /e/ in figure 3a are also shifted rightwards with respect to figure 3c and thus they, too, suggest a perceptual overshoot of approximately 15%.

In experiment 1a, the slope of the first and the duration of the second transition were kept constant. Experiment 1c was designed to examine what role the slope plays in the reduction process. We selected one particular V_2 among the most reduced targets in experiment 1a which was still perceived as /a/. This vowel is close to the transition cutback of - 30 ms (with F_1= 596 Hz and F_2=1,803 Hz) and is predominantly perceived as /ε/ in isolation. In the five tokens synthesized, the slopes of the two transitions were changed gradually (fig. 4a) while keeping the other parameters constant — especially the formant trajectory in the F_1-F_2 plane (fig. 4b) and the duration of V_2. The total duration of the transitions of the reduced vowel (i.e. the transition region that included [iV_2] and [V_2i] was fixed at 150 ms. In other words, when the slope of the first transition increased, that of the second decreased. Listening results, shown in figure 4c, indicate that changing the slopes in this fashion left the percept unaffected: all the tokens were perceived as /iai/.

In experiment 1c, the duration of the complete [iV_2i] transition region in which vowel reduction was observed was about 150 ms. In experiment 1d, we addressed the question of whether the perceived V_2 vowel was influenced by the equivalent of speaking rate. To accomplish this, we constructed a stimulus series starting with the token described in experiment 1c (that is, with F_1= 596 Hz and F_2=1,830 Hz) and simultaneously lengthened all segments of the token from the original 400 ms, to 440, 480, 520, 560, 600, and 800 ms, while keeping the proportion of the steady-state portions and transitions constant. (The duration was changed by simply increasing the frame duration from the original 10 ms, to 11, 12, 13, 14, 15, and 20 ms.) Figure 5 illustrates the labeling results obtained in this experiment. As it appears, the reduction persists for token durations up to 600 ms. The duration of the entire transition region for this token was 225 ms. When transition duration increases past 225 ms, i.e. when the speaking rate further decreases, vowel reduction vanishes and the percept becomes /iεi/.

Experiment 2: Gesture Reduction for [aya]

The first experiment examined the reduction for the case of a vowel sequence generated by a single gesture, i.e. tongue constriction. There exist, however, vowel-to-vowel transitions in which more than one gesture participates, e.g. [aya] in French. We chose this sequence to become the stimulus of experiment 2a for two reasons: the presence of two simultaneous gestures in the [ay] transition, i.e. tongue constriction and lip rounding, and the vocal tract shape of [y]. In comparison to the area function shape of the vowel [i] used in experiment 1, the vowel [y] differs by virtue of the presence of a second gesture, i.e. lip rounding. For the sake of simplicity, in the synthesis sheme used for generating the [y] vowel in the present experiment (2a), the two gestures were made strictly synchronous. Stimulus generation was accomplished using the DRM synthesizer described in the ‘Methods’ section.

Figure 6a shows the time domain plot of a set of eight [aya] formant transitions with increasing [V_2a] transition cutbacks, i.e. with an increasingly earlier starting point of the return to [a]. Among the eight tokens there is one in which the transition returns to [a] immediately after reaching the [y] and another in which a 10-ms steady-state [y] is present before returning to [a]. The value of F_3 obtained by the model is 2,795 Hz for the vowel [a], and 2,540 Hz when the target [y] is reached. For illustrative purposes, the time trajectory of F_3 (important for the distinction /i/-/y/) is shown in figure 6a for the case where the target [y] is reached. The F_1-

[5] Actually, the reduction is probably quite a bit larger than 15% because, as a consequence of our smooth transition onsets and offsets, the nominal duration of the steady-s

F_2 trajectory corresponding to the time domain plot is illustrated in figure 6b. When, in the model, the two gestures are produced in perfect synchrony, the acoustic effect caused by lip rounding (manifest as a drop in F_1 and F_2) occurs later than that of the lingual constriction (manifest as a drop in F_1 and a rise in F_2). In contrast to the [iai] trajectory used in experiment 1, the one for [aya] is not a straight line in the F_1-F_2 plane[6]. The [aya] transition, therefore, should be considered as an appropriate stimulus to study the perception of curved trajectories generated by a composite gesture. In preliminary labeling tests, we observed that none of the eight stimuli shown in figure 6a yielded an unambiguous /aya/ percept.

Consequently, in the main labeling experiment (experiment 2a), we added two new tokens in which the stable portion of [y] was increased by either 10 or 20 ms, respectively. Subjects were the same native French listeners who had participated in the former experiments. They were given the task of deciding whether any of the three vowels /y/, /i/, or /ε/ or the liquid /l/ was the most appropriate label for V_2 in the token they just heard. An /l/ percept is possible because, in French, /l/ is fronted. Results of this experiment are given in figure 6c. As the figure indicates, merely reaching the target [y] was insufficient to generate a definite /aya/ percept; in order for that to happen,[y] had to be present as a *steady-state* vowel for a duration longer than 30 ms. When the duration of the steady state [y] was shorter than 30 ms, V_2 was perceived as /i/ rather than /y/. In other words, unless the transition halted at the [y] vowel and remained there for a given minimum duration (i.e. 30 ms), the lip rounding gesture was ignored. This finding strongly suggests that vowel-to-vowel transitions are governed by temporal integration of the trajectory vectors in the F_1-F_2 plane: There appears to be a mechanism that *calculates the time average* of the *length* and the *direction* of formant trajectories that are, by definition, time-varying, and that *predicts the target* of the trajectory which, in fact, may not actually be reached. Such a mechanism could explain why, in experiment 2a, there were so many /i/ responses when, in fact, the trajectory was diverging from the [i] formant values[7]. Therefore, the question arises whether an /y/ percept could be induced by presenting to the listeners an incomplete curved trajectory which, contrary to the one used in experiment 2a, would never point toward the [i] vowel.

Consequently, in experiment 2b we synthesized an F_1-F_2 trajectory that first curved downward (i.e. was consistent with a labial gesture) before pointing toward the[y] (fig. 7b). The time course of the first two formants is illustrated in figure 7a, with degrees of cutback identical to those of the tokens used in experiment 1a (fig. 2b). Because, in preliminary tests, an /aia/ percept was never obtained, we gave the subjects the task of labeling the V_2 vowel as /y/, /Ø/, or /œ/. Labeling results are shown in figure 7c for the same 5 French subjects. As a control, in experiment 2c, we asked the subjects to label, using the same three response categories (i.e. vowels /y/, /Ø/, or /œ/), 200-ms steady-state vowels that had the same formant values as those of the V_2 endpoints of the transitions in experiment 2b (fig. 7b). Results of this experiments are shown in figure 7d. Comparison of the category boundaries obtained in experiment 2b and 2c for the three intermediate vowels, shown in figures 7c and 7d, reveals a consistent perceptual overshoot for the vowels as long as the context is *dynamic*, such as in experiment 2b.

In production terms, experiments 2a and 2b used tokens which were generated with a certain time course for the labial, and a different time course for the tongue gesture. In experiment 2a, the acoustic trace of the labial gesture, signaled by a lowering of the first two formants as compared to [ai], is delayed, whereas in experiment 2b, the acoustic trace of the labial gesture is present at the very beginning of the V_1V_2 transition and is preserved until the transition is completed. In experiment 2d, we wanted to explore the perception of transitions following a trajectory that would suggest the presence of a labial gesture at the beginning and a release

[6] In natural speech, a straight trajectory is obtained due to labial anticipation [Carré and Mrayati, 1991]. Using the model, such a straight trajectory is obtained by emulati

[7] These results also predict that a transition which never points to [i], such as a strictly straight line [ay] trajectory, would never lead to an /i/ percept. Tests of this predictio

from this gesture toward the end of the trajectory. Tokens having the F_1-F_2 time course illustrated in figure 8a, and the corresponding trajectory in the F_1-F_2 plane illustrated in figure 8b, fulfill this objective. It has to be stressed that such a $V_1V_2V_1$ sequence is *unrealistic*, i.e. tokens such as those shown in figures 8a and 8b are never encountered in natural speech and, consequently, cannot be obtained with a realistic production model such as the DRM which we have used so far. For this reason, stimuli for this experiment were generated using a simple formant synthesizer. The lack of realism of these tokens is manifest in that the trajectory in the F_1-F_2 plane suggests a labial gesture at the beginning (i.e. a drop in F_1 and F_2), but then moves toward, and actually reaches, [i] (fig. 8b), rather than [y] as in experiment 2b. In fact, none of the tokens generated according to this scheme were heard as natural speech. Preliminary listening tests showed that a steady-state [i] portion of 80-100 ms was necessary, in order for an /aia/ sequence to be perceived. Thus, experiment 2d was conducted with a steady-state [i] portion present in six tokens, with respective durations of 100, 80, 60, 40, 20, and 0 ms and one token not reaching [i], corresponding to a reduction with a cutback of - 20 ms. The listener's task was to label the percept as either /aia/ (i.e. with no labiality detected) or /aya/ (i.e. with labiality detected). Results for 5 listeners, shown in figure 8c, indicate that, when the trajectory reaches [i] and stays on it for between 0 and 20 ms, an /y/ V_2 percept is induced.

Discussion

In the foregoing paragraphs we described two main experiments aimed at assessing the listener's percept of vowel quality in $V_1V_2V_1$ sequences in which vowel transitions were generated by DRM gestures, i.e. by a model previously shown to be appropriate for the study of speech production [Carré and Mody, 1997]. The general finding of the experiments was that, for sequences with a single gesture (i.e. [iai], experiment 1) and, under certain conditions, also for sequences with two gestures (i.e. [aya], experiment 2), we consistently observed what can be described as a perceptual overshoot. These results, taken together with those of Carré et al. [1994] and Divenyi et al. [1995], lead to the following conclusions:

(1) In production models, a command that does not reach its target, i.e. one in which the gesture is reduced through an articulatory undershoot, leads to an instance of vowel reduction in the acoustic domain [Lindblom, 1963]. In many cases, as demonstrated in experiments 1a and 2b, the perceptual system may completely recover the target not reached either by using a mechanism that produces a perceptual overshoot [Lindblom and Studdert-Kennedy, 1967], or because of previous perceptual learning of such situations. To characterize the reduction phenomenon, the duration of the entire transition region of the reduced vowel (i.e. to and from a constant duration V_2) seems to be more important than the transition slopes (see exeriment 1c). This duration, however, is not absolute but, rather, appears (within a certain time range) to be related to the total duration of the token (experiment 1d).

(2) Reaching the target does not guarantee that it will be perceived. Results of experiment 2a suggest that the transition is evaluated by a temporal integration mechanism which computes an average transition direction used by the listener to identify the vowel perceived. This mechanism is similar to the weighted-time average process postulated to operate for the evaluation of the pitch of sinusoidal and F_0 glides [d'Alessandro and Castellengo, 1994].

(3) All the results highlight the usefulness of representing formant trajectories in the F_1-F_2 plane. This representation reveals even minute (i.e. 10- to 20-ms) temporal asynchronies between formants. Experiments on the discrimination of frequency-modulated signals suggest that asynchronies of this magnitude are easily detected by the auditory system [Moore and Sek, 1998].

(4) For reduction to occur reliably in the case of two gestures, coordination between them is needed, in order to obtain a formant trajectory shape (rectilinear or not) that can be

unambiguously interpreted by the listener. For this to happen, the timing of the movement onset as well as the synchrony of the gestures must be precisely controlled. The transition that the listener hears is a composite of the acoustic traces of two command gesture components: To perceive /aya/, in a sequence generated by two concurrent command gestures, the acoustic effect of each gesture must be present to a sufficient degree.

(5) In agreement with Strange et al. [1983] and Strange and Bohn [1998], results of the present experiments stress the importance of dynamic changes. Comparison of the results of experiments 2a and 2b serves to demonstrate this point: whereas, for a given target, one trajectory shape leads to a perceptual overshoot (experiment 2b), with another trajectory shape the target may not be perceived as such even when it is actually reached (experiment 2a).

Although we do not wish to engage in a protracted discussion of the theoretical aspects of vowel production and perception, the present results present a puzzle as to the processes that putatively gave rise to them. To solve the puzzle, we propose two logical possibilities to be considered. The first, phonetic, possibility is that a given global formant trajectory pattern is identified by way of comparing its acoustic-perceptual characteristics to those of previously learned and stored tokens. Since, during the learning process the stored trajectory tokens become associated with corresponding production gestures, it is plausible that the listener will attempt to associate any new trajectory he/she hears with a gesture previously associated with one of the stored trajectories. According to this possibility, identification of the gesture does not occur directly but, rather, it is the consequence of the mapping process. The second possibility is that there is a mechanism whose function is to dissociate the acoustic traces generated by different gestures and to map these traces directly into a repertoire of learned gestures, i.e. the perceptual representation would no longer be phonetic but phonological. By relying on a small phonological rather than a large phonetic repertoire for the decoding of speech, this mechanism would appreciably reduce the information load on the perceptual system.

But would information reduction alone compel us to adopt the second possibility over the first, in an attempt to provide a general explanation for the perception of transitions? In fact, if we had terminated our investigation after experiment 1, both possibilities could account for the results obtained. On the other hand, results of experiment 2 (and especially experiment 2d) can be explained only by the second.[8] As far as the traces of the two gestures are concerned, our synthesis methods were able to manipulate them independently and generate trajectories with various degrees of labiality present. A glance at the trajectory in the F_1-F_2 plane promptly reveals that labiality is signaled by a northeast-to-southwest vector that displaces a (nominally [ai]) southeast-to-northwest trajectory toward the origin of the F_1-F_2 coordinate system. In contrast, this southwestward displacement occurred too late in the trajectory used in experiment 2a and, had we failed to add tokens with a steady-state [y] portion, the labiality information available to the listeners would have been clearly insufficient. One the other hand, in experiment 2b, such a displacement was introduced early on (at the onset of the transition) and, as the results showed, it was unambiguously interpreted as a sign of labiality. This was true to an even greater extent for the trajectory of experiment 2d — a trajectory not producible by the human vocal tract — in which the labial information was present apparently to such a high degree that, when [i] was the target V_2, it was simply not perceived as such at all. Considered as a whole, results of the study suggest that the gestures had to be separately recognized because the listeners appear to have effectively dissociated the acoustic traces of the gestures from one another and recognized them independently. Such perceptual independence may be looked upon as analogous to the relative autonomy of gestures in production. Since relative autonomy is a sine-qua-non requirement of multiple command gestures, the independence of the

[8] In experiment 2d, the trajectories were fully artificial: a nearly-realistic labial gesture was followed by a fully artificial ‘unlabial’ one, the sequence occurring concurren

perceptual effects of gestures observed in our experiments could open the door to mapping articulatory units into perceptual units — a possibility much discussed in the past 50 years and on which the dynamic approach we adopted sheds a new light.

Clearly, gestures take place in time and so do the trajectories defined in the F_1-F_2 plane. Consequently, the vectors described above must reflect displacement over a certain period of time inside this coordinate system. Thus, the lenght and the direction of trajectory vectors must be the integral of instantaneous displacements in time. The existence of such an integrator is implied, for example, by results of experiments 2a and 2d that showed an /aia/ percept for a curved [aya] trajectory and an /aya/ percept for a curved [aia] trajectory. It is also clear that the temporal integrator must use a window function that, as suggested by results of experiments 2a and 2b, would be skewed. A formal model specifying the temporal integration mechanism more closely will be proposed in a further report.

Acknowledgment

The authors are particularly grateful to their colleagues Michael Studdert-Kennedy and Maria Mody for their many insightful comments on the present article. The research benefited from support by the NATO Scientific Office, the Department of Veterans Affairs Medical Research, and the National Institute on Aging.

References

d'Alessandro C, Castellengo M. The pitch of short duration vibrato tones. J. acoust. Soc. Am 1994;95:1617–1630.

BadinPFantG198453107Notes on the vocal tract computations. Q. Prog. Status Rep., Speech Transm. Lab., R. Inst. Technol., Stockh., No. 2/3

BeautempsD1993Récupération des gestes de la parole à partir de trajectoires formantiques: identification de cibles vocaliques non atteintes et modèles pour les profils sagittaux des consonnes fricatives; thèse Institut National Polytechnique, Grenoble

Browman, C.; Goldstein, L. Ewan, Anderson, Phonol. Yb. Cambridge University Press; Cambridge: 1986. Towards an articulatory phonology; p. 219-252.

Brownlee, SA. The role of sentence stress in vowel reduction and formant undershoot: a study of lab speech and informal spontaneous speech. University of Texas; Austin: 1996. PhD thesis

Carré R, Chennoukh S, Divenyi P, Lindblom B. On the perceptual characteristics of 'speech gestures'. J. acoust. Soc. Am 1994;96:S3326.

Carré, R.; Mody, M. Prediction of Vowel and Consonant Place of Articulation. Proc. 3rd Meet; ACL Special Interest Group in Computational Phonol; SIGPHON 97, Madrid. 1997; p. 26-32.

Carré R, Mrayati M. Vowel-vowel trajectories and region modeling. J. Phonet 1991;19:433–443.

Chiba, T.; Kajiyama, M. The vowel: its nature and structure. Tokyo-Kaiseikan Publishing Company; Tokyo: 1941.

Di Benedetto MG. Frequency and time variations of the first formant: properties relevant to the perception of vowel height. J. acoust. Soc. Am 1989;86:67–77.

Divenyi, P.; Lindblom, B.; Carré, R. Proc. 13th Int. Congr. Phonet. Sci. Stockholm: 1995. The role of transition velocity in the perception of V_1 V_2 complexes; p. 258-261.

Fant, G. Speech sounds and features. MIT Press; Cambridge: 1973.

FantG1975114Vocal tract area and length perturbations. Q. Prog. Status Rep., Speech Transm. Lab., R. Inst. Technol., Stockh., No. 4

Fowler, CA. Haskins Lab. Status Rep. Speech Res. Haskins Laboratories; New Haven: 1992. Phonological and articulatory characteristics of spoken language; p. 1-12.

Gay T. Effect of speaking rate on vowel formant movements. J. acoust. Soc. Am 1978;63:223–230. [PubMed: 632415]

House AS, Fairbanks G. The influence of consonant environment upon the secondary acoustical characteristics of vowels. J. acoust. Soc. Am 1953;25:105–113.

Huang, CB. Int. Congr. on Phonet. Sci. Tallinn: 1987. Perception of first and second formant frequency trajectories in vowels; p. 194-197.

Johnson, K. An exemplar model; in Johnson, Mullennix, Talker variability in speech processing. Academic Press; New York: 1997. Speaker perception without speaker normalization; p. 145-165.

Kozhevnikov, VA.; Chistovich, LA. NTIS. US Department of Commerce; 1965. Speech, articulation, and perception.

Kuehn DP, Moll KL. A cineradiographic study of VC and CV articulatory velocities. J. Phonet 1976;4:303–320.

Kuhl, P. Proc. ICSLP '92. Banff: 1992. Infants' perception and representation of speech: development of a new theory; p. 449-456.

Kuwabara H. An approach to normalization of coarticulation effects for vowels in connected speech. J. acoust. Soc. Am 1985;77:686–694. [PubMed: 3973240]

Lindblom B. Spectrographic study of vowel reduction. J. acoust. Soc. Am 1963;35:1773–1781.

Lindblom, B. Hardcastle, Speech production and speech modelling. Kluwer Academic Publishers; Dordrecht: 1990. Explaining phonetic variation: a sketch of the H and H theory; in Marchal; p. 403-439.

Lindblom B, Studdert-Kennedy M. On the role of formant transitions in vowel perception. J. acoust. Soc. Am 1967;42:830–843. [PubMed: 6075568]

LindgrenRLindblomB199614Reduction of vowel chaos. Q. Prog. Status Rep., Speech Transm. Lab., R. Inst. Technol., Stockh., No. 2

Mattingly IG. The global character of phonetic gesture. J. Phonet 1990;18:445–452.

Moore BCJ, Sek A. Discrimination of frequency glides with superimposed random glides in level. J. acoust. Soc. Am 1998;104:411–421. [PubMed: 9670533]

Mrayati M, Carré R, Guérin B. Distinctive region and modes: a new theory of speech production. Speech Commun 1988;7:257–286.

Mrayati M, Carré R, Guérin B. Distinctive regions and modes: articulatory-acoustic-phonetic aspects. A reply to Boë and Perrier comments. Speech Commun 1990;9:231–238.

NordL19861936Acoustic studies of vowel reduction in Swedish. Q. Prog. Status Rep., Speech Transm. Lab., R. Inst. Technol., Stockh., No. 4, pp.

Peterson GE, Barney HL. Control methods used in the study of the vowels. J. acoust. Soc. Am 1952;24:175–184.

Son, R.J.J.H. van Vowel perception: a closer look at the literature. Proc. Inst. Phonet. Sci., Univ. Amsterdam 1993;17:33–64.

Son, R.J.J.H. van; Pols, LCW. Formant movements of Dutch vowels in a text, read at normal and fast rate. J. acoust. Soc. Am 1992;92:121–127. [PubMed: 1512318]

Son, R.J.J.H. van; Pols, LCW. Proc. Eurospeech '93. Berlin: 1993. Vowel identification as influenced by vowel duration and formant track shape; p. 285-288.

Strange W. Dynamic specifications of coarticulated vowels spoken in sentence context. J. acoust. Soc. Am 1989;85:2135–2153. [PubMed: 2732388]

Strange W, Bohn OS. Dynamic specification of coarticulated German vowels: perceptual and acoustical studies. J. acoust. Soc. Am 1998;104:488–504. [PubMed: 9670540]

Strange W, Jenkins JJ, Johnson TL. Dynamic specification of coarticulated vowel. J. acoust. Soc. Am 1983;74:695–705. [PubMed: 6630725]

Fig. 1.

Schematic diagram of the vocal tract deformation gestures for an [iV_2i] sequence. The vocal tract shape of [i] is shown as the solid line, whereas the one for the (arbitrary) V_2 is shown as the broken line. The gesture for [iV_2] is indicated by the solide arrow (**1** and associated 1) and that for the [V_2i] by the broken arrow (**2** and associated 2).

Fig. 2.

Schematic representation of the eight [iV_2i] sequences with different degrees of constriction, synthesized to be used as stimuli in experiment 1a. **a** DRM command amplitude (arbitrary units) as a function of time. **b** F_1 and F_2 transitions corresponding to the gesture in **a**, shown in a spectrogram plot. The temporal representation of F_3 is also shown (dotted line) for the case in which the [a] target is reached. **c** F_1-F_2 plot of the eight formant trajectories in **b**. Note that all V_2 vowels (shown as points) fall on the [ia] trajectory and can be interpreted as incomplete [iai] transitions, except for the rightmost point. The time labels on some of the points refer to the time of return to the final [i] vowel, before completing the [ia] transition, that is, the time value of vowel reduction.

Fig. 3.

Experiments 1a and 1b. **a** Average labeling results (and standard deviation) of 5 listeners in experiment 1a for V_2 in a [iV_2i] context. Ordinate: percent responses; V_2 = /a/ (filled diamonds), V_2 =/ε/ (squares), V_2 = /e/ (triangles). Abscissa: point of return of the transition to [i] prior to reaching [a]. **b** Average labeling results (and standard deviation) of 1 listener in experiment 1a for V_2 in a [iV_2i] context. **c** Average labeling results (and standard deviation) of 5 listeners in experiment 1b for a steady-state V_2 vowel. Ordinate: percent responses; V_2 = /a/ (filled diamonds), V_2 = /ε/ (squares), V_2 = /e/ (triangles). Abscissa: F_1-F_2 value of the V_2 vowel (fig. 2c) shown as the point of return of the transition to [i] prior to reaching [a], had the vowel been the midpoint of an [iV_2i] transition.

Fig. 4.

Experiment 1c. **a** Spectrogram representation of the five different temporal patterns of the formant transitions; note that the transition for all tokens reaches the same V_2 and that the total duration of the transition is always constant at 150 ms. **b** F_1-F_2 plane representation of the transitions: note that the extreme value, i.e. the V_2 vowel corresponds to the formant values of the vowel [ε]. **c** Average labeling results (and standard deviation) of 5 listeners in experiment 1c for V_2 in a [iai] context with changing transition durations. Ordinate: percent responses; V_2 = /a/ (filled diamonds), V_2 = /ε/ (squares), V_2 = /e/ (triangles). Abscissa: Token numbers corresponding to the five different transition slopes (**a**).

[redacted]

Fig. 5.

Average labeling results (and standard deviation) in experiment 1d using a V_2 value corresponding to a steady-state [ε]. Ordinate: percent responses; V_2 = /a/ (filled diamonds), V_2 =/ ε / (squares). Abscissa: total duration of the token.

Fig. 6.

Experiment 2a. **a** Temporal representation of the F_1 and F_2 transitions of eight of the ten tokens used in the [aV_2a] experiment where only two tokens completed the transition to the vowel [y]. The other six had their transitions progressively cut back, i.e. their formant frequencies returned toward those of [a] after reaching various V_2 endpoint vowels on the trajectory. The temporal representations of F_3 is also shown (dotted line) for the case in which the [y] target is reached. **b** F_1-F_2 plane representation of the [aya] transition for the eight tokens shown above. Note that, in contrast to the [iai] trajectory (fig. 2c), the [aya] trajectory is curved. **c** Results (average and standard deviation) of experiment 2a: Percent /aya/ (filled diamonds), /aia/ (squares), /ala/ (triangles) responses as a function of the duration of [y] or the transition cutback point (i.e. the point of return to [a]) where 0 ms refers to the condition in which the vowel [y] is reached but the transition immediately returns toward [a]. Note that 100% /y/ responses were obtained only for the 30-ms ‘positive cutback’ condition, i.e. for the condition in which there was a 30-ms steady-state [y] before the transition actually took a turn back to [a]. Also, note that the intersubject variability is much larger than the one observed in the subtests of experiment 1.

Fig. 7.

Experiments 2b and 2c. **a** Temporal representation of the F_1 and F_2 transitions of the eight tokens used in the [aV_2a] experiment where only two tokens completed the transition to the vowel [y]. The other six had their transitions progressively cut back,i.e. their formant frequencies returned toward those of [a] after reaching various V_2 endpoint vowels on the trajectory. The temporal representation of F_3 is also shown (dotted line) for the case in which the [y] target iy reached. **b** F_1-F_2 plane representation of the [aya] transition for the eight tokens used in experiment 2b (solid line). Note that, compared with the [aya] trajectory (broken line) shown in figure 6b, the trajectory is symmetrically curved. **c** Results of experiment 2b: Percent /aya/ (filled diamonds), /aØa/ (squares), /aœa/ (triangles) responses as a function of the duration of [y] or the transition cutback point (i.e. the point of return to [a]) where 0 ms refers to the condition in which the vowel [y] is reached but the transition immediately returns toward [a]. Note that /aya/ responses were obtained for the -20 ms ‘negative cutback’ condition, i.e. for the condition in which there was a 20 ms before reaching the target [y]. Average results of 5 subjects. **d** Results in experiment 2c: Labeling of steady-state V_2 vowels. Ordinate: percent responses V_2 = /y/ (filled diamonds), V_2 = /Ø/ (squares), V_2 = /œ/ (triangles). Abscissa: F_1-F_2 value of the V_2 vowel shown as the point of return of the transition to [a] prior to reaching [y], had the vowel been the midpoint of an [aV_2a] transition.

Fig. 8.

Experiment 2d. **a** Temporal representation of the F_1 and F_2 transitions of the seven tokens used in the [aV_2a] experiment where six tokens completed the transition to the vowel [i]. The last one had its transition cut back, i.e. its formant frequencies returned toward those of [a] after reaching V_2 endpoint vowel on the trajectory. The temporal representation of F_3 is also shown (dotted line) for the case in which the [i] target is reached. **b** F_1-F_2 plane representation of the [aia] transition for the seven tokens used in experiment 2d (solid line). Note that, compared with the [aya] trajectory (broken line) shown in figure 7b, the trajectory is also curved but reaches [i]. **c** Results of experiment 2d: Percent /aia/ (filled diamonds) and /aya/ (squares) responses as a function of the duration of [i] or the transition cutback point (i.e. the point of return to [a]) where 0 ms refers to the condition in which the vowel [y] is reached but the transition immediately returns toward [a].