Signal dynamics in the production and perception of vowels

René Carré

In "*Approaches to Phonological Complexity*". I. Chitoran, C. Coupé, E. Marsico and F. Pellegrino, (Eds.), Mouton de Gruyter, Berlin, New York, pp 59-81, (2009).

1. Abstract

Vowels can be produced with static articulatory configurations represented by dots in acoustic space (generally by formant frequencies in the F1-F2 plane). But because vowel characteristics vary with speaker, consonantal environment (co-articulation) and production rate (reduction phenomenon), vowel formant frequencies can also be represented by their mean values and standard deviations, according to different categories (language, age and gender of speaker). The use of 'targets' means that they are generally studied from a static point of view. But, several questions can be raised: How are vowel representations set up if vowel realizations rarely reach their targets? Is representation the same from one person to another? How is a given vowel, produced several time with different acoustic characteristics and in different environments, identified? By using contextual information? By normalization? These questions lead to studying vowels from a dynamics point of view.

Here, we first propose a theoretical deductive approach to vowel-to-vowel dynamics which leads to a specification in terms of vocalic trajectories in the acoustic space characterized by their directions. Then, results on V1V2 transitions produced and perceived by subjects will be presented. In production, measurements of the F1 and F2 transition rates are represented in the F1 rate/F2 rate plane. In perception, direction and rate of synthesized transitions are studied for transitions situated outside the traditional F1/F2 vowel triangle. This situation enables the study of transitions characterized only by their directions and rates without relation to any vowel targets of the vowel triangle. Such experiments show that these transitions can be perceived as V1V2. Several issues can then be revisited in the light of this dynamic representation: vocalic reduction, hyper and hypo speech, normalization, perceptual overshoot… A fully dynamic representation of both vowels and consonants is proposed.

2. Introduction

Vowels are generally characterized by the first two or three formant frequencies. Each of them can be represented in the acoustic space (F1-F2 plane) by a dot (Peterson and Barney, 1952). This specification is static. Vowels can be produced in isolation without articulatory variations, but in natural speech such cases are atypical since their acoustic characteristics are not stable. They vary with the speaker and with the age and gender of the speaker, with the consonantal context (coarticulation), with the speaking rate (reduction phenomena), and with the language (Lindblom, 1963). So, vowels are classified into crude classes, first according to the language, and then according to speaker categories. Within each category, vowels can be specified in terms of underlying 'targets' corresponding to the context- and duration-independent values of the formants as obtained by fitting "decaying exponentials" to the data points (Moon and Lindblom, 1994). The point in focus here is that this specification is static and, significantly, may be taken to imply that the perceptual representation corresponds to the target values (Strange 1989a, 1989b)."

At this point, several questions can be raised: How is the perceptual representation obtained if the vowel targets depend on the speaker, and are rarely reached in spontaneous speech production? Are the representations the same from one person to another (Johnson, et al., 1993; Carré and Hombert, 2002; Whalen*, et al.*, 2004)? How is this perceptual representation built: by learning, or is it innate? How is the vowel perceived with its different acoustic characteristics according to the context and the speaker: by normalization (Nordström and Lindblom, 1975; Johnson, 1990; Johnson, 1997)? Why is vowel perception less categorical than consonant perception (Repp*, et al.*, 1979; Schouten and van Hessen, 1992)?

Many studies have been undertaken to answer those questions. The results are generally incomplete and contradictory. They cannot be used to set up a simple theory explaining all the results. But they help highlight the importance of the dynamics in vowel perception (Shankweiler*, et al.*, 1978; Verbrugge and Rakerd, 1980; Strange, 1989).

In view of the fact that sensory systems have been shown experimentally to be more sensitive to changing stimulus patterns than to purely steady-state ones, it appears justified to look for an alternative to static targets - a specification that recognizes the true significance of signal time variations. One possibility is that dynamics can be characterized by the direction and the rate of the vocalic transitions:

- Vowel-vowel trajectories in the F1/F2 plane are generally rectilinear (Carré and Mrayati, 1991). So they can be characterized by their direction. Moreover, privileged directions are observed in the production of vowels (in single-vowel or CV syllables) called 'vowel inherent spectral changes, VISC' (Nearey and Assmann, 1986). Moreover**,** perception experiments show the importance of VISC in improving vowel identification (Hillenbrand*, et al.*, 1995; Hillenbrand and Nearey, 1999).
- On the topic of transition rate, we recall the results of Kent (1969): "*the duration of a transition – and not its velocity – tends to be an invariant characteristic of VC and CV combinations*". Gay (1978) confirmed these observations with different speaking rates and with vowel reduction: "*the reduction in duration during fast speech is reflected primarily in the duration of the vowel,… the transition durations within each rate were relatively stable across different vowels…*". If the transition duration is invariant across a set of CV's with C the same and varying Vs, it follows that the transition rate depends on the vowel to be produced.

So the time domain could play an important role in the identification of vowels (Fowler, 1980). For example, to discriminate the sequences [ae], [aɛ], and [ai], what acoustic information does the listener need? The answer is that the second vowel V2 can be detected by using the transition rate as a cue. This parameter can be specific to the speaker and/or related to the syllabic rate. At the very beginning of the transition and throughout the transition there is sufficient information to detect V2. There are no privileged points in time (for example the middle of V2 to measure the formant frequencies) for V2 detection. The rate measure is therefore very appropriate in a noisy environment. It can also explain the perceptual results obtained by Strange (1983) in 'silent center' experiments that replaced the center of the vowel by silence of equivalent duration. This manipulation preserves the direction and the rate of the transition as well as the temporal organization (syllabic rate). Also relevant are experiments by Divenyi et al. (1995) showing that, in V1V2 stimuli, V2 was perceived even when V2 and the last half of the transition was removed by gating. Finally, it can be observed that both Arabic (Al-Tamimi*, et al.,* 2004) and Vietnamese subjects (Castelli and Carré, 2005) have difficulties in producing and perceiving isolated vowels.

In this paper, a deductive theoretical approach to the study of vowel-to-vowel dynamics is proposed. It leads to a specification of vocalic trajectories in the acoustic space characterized by their directions. Then, results on the production and the perception of V1V2 transitions by French subjects are presented. Production measurements of the F1, F2 transition rates are represented in a F1 rate / F2 rate plane. In a perceptual study, we focus on the direction and rate of synthesized transitions situated outside the traditional F1/F2 vowel triangle. This situation enables the study of transitions characterized only by their directions and rates without reference to any vowel targets in the vowel triangle.

3. Deductive approach and vowel-to-vowel trajectories

Here, we try to infer vowel properties, not from data on vowel production and perception, but by a deductive approach starting from general physical properties of an acoustic tube. If the goal is to build an optimal device for acoustic communication, i.e. a device that gives maximum acoustic contrast for minimum area shape deformation, then the tube must be structured into specific regions leading to a corresponding organization of the acoustic space (Carré, 2004; Carré, Submitted). A recursive algorithm using the calculation of the sensitivity function (Fant and Pauli, 1973) deforms “efficiently” (area shape deformation gives maximum formant variations – least effort principle) the shape of the tube in order to increase (decrease), step by step, the formant frequencies. The acoustic space automatically obtained corresponds to the vowel triangle (which is, consequently, optimal in size; it cannot be larger). [a] is obtained with a back constriction and a front cavity; [i] with a front constriction and a back cavity; [u] with a central constriction and two lateral cavities. In the two first cases, the front end of the tube is open; in the last case the front end is almost closed. The specific regions automatically obtained correspond to the main places of articulation (front, back and central) used to produce vowels. The shape deformations are obtained by deformation gestures, limited in number: three different “tongue” gestures and one “lip” gesture. The three different tongue gestures are: a transversal deformation gesture from front to back places of articulation (and vice-versa) producing [ia] (or [ai])), a longitudinal deformation gesture from front to central constriction producing [iu] and a longitudinal displacement gesture from back to central place of articulation producing [au]. The lip gesture is used to reach [u] (low F1 and F2). The deformations can easily be modelized by the Distinctive Region Model (DRM) (Mrayati*, et al.*, 1988; Carré, Submitted).

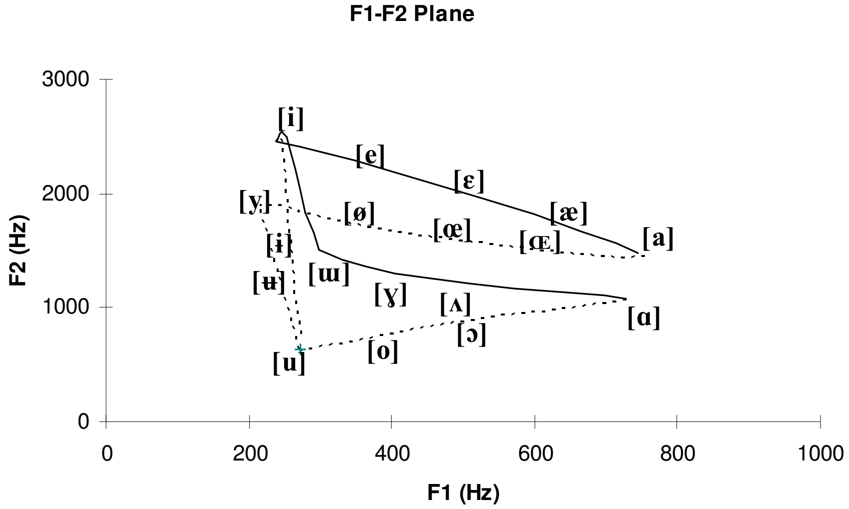


Figure 1. Vocalic trajectories obtained by deduction from acoustic characteristics of a tube and corresponding vowels. Dotted lines are labialized trajectories.

From the DRM model, eight more or less rectilinear trajectories structuring the acoustic space were obtained: [ai], [ɑu], [iu], [ay], [ɑɯ], [uy], [iɯ], [yu] (Figure 1). The maximum acoustic space obtained by this approach fits well with the vowel triangle. The use of the DRM does not lead to characterize vowels first, but rather privileged vocalic trajectories. A maximum acoustic contrast criterion would select the endpoints, and intermediate points, on the trajectories which correspond well with the vowels given for example by Catford (1988).

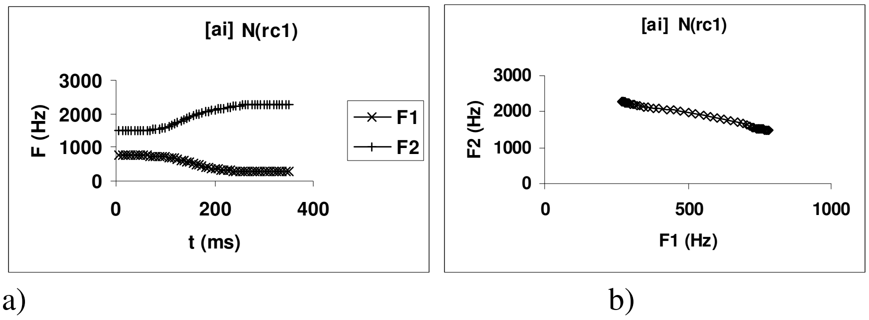
Recall that the recursive algorithm calculates, from an initial shape of the area function of a tube, a new shape according to a minimum of energy criterion (minimum deformation leads to maximum acoustic variation) (Carré, 2004). This operation is repeated until the maximum acoustic limits of the tube are reached. Thus, the algorithm simulates an evolutionary process (the goal is not pre-specified at the beginning of the process), by simply increasing acoustic contrast, step by step, according to a minimum of energy criterion. The resulting trajectories in the acoustic plane can be characterized by their directions.

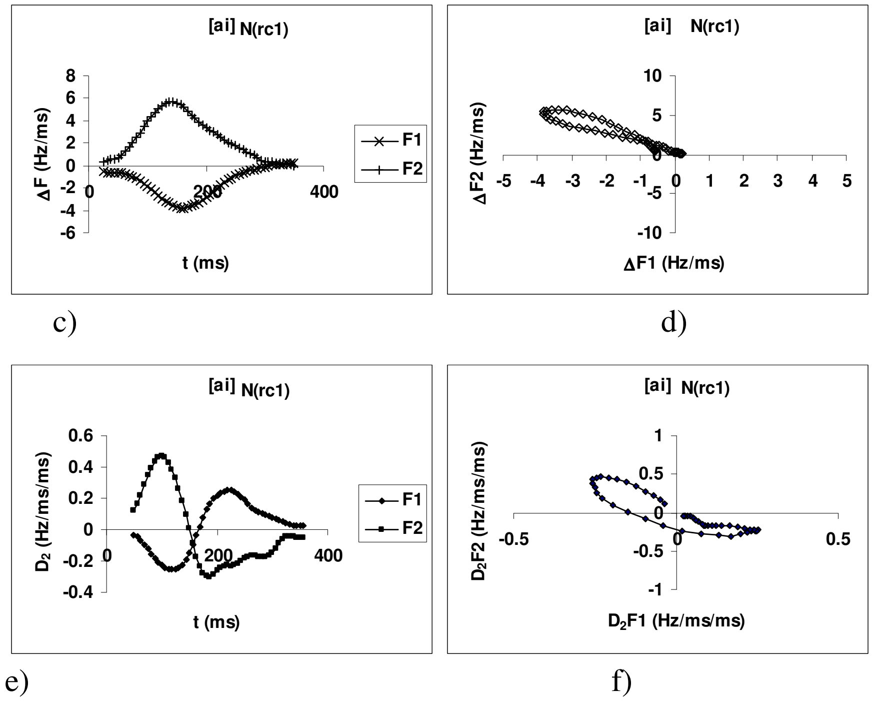
On the basis of the above discussion we hypothesize that the perception of vowels might be understood, not in terms of static targets, but in terms of a dynamic measure of the direction and rate of spectral change. We will test this hypothesis in a series of studies of vowel-to-vowel sequences.

4. Vowel-vowel production

[V1V2] sequences were produced 5 times by 5 male and 5 female speakers, all French, at 2 different rates (normal and fast). In the following experiments, V1 is always /a/ and V2 is one of the French vowels situated on the [ai] ([i, ɛ, e]), [ay] ([y, œ, ø]) or [au] ([u, o, ɔ]) trajectories. A French word containing V2 appears on the computer screen with alphabetic representation to help the subject who may have no phonetic knowledge. To exemplify, the instructions were the following: at 'fast' rate: *say 'a-i, a-i-a, i' as in the word 'lit'*. The recording process was controlled by PC software that randomly presented the succession of the items to be recorded. In the case of bad pronunciation, or hesitation, the speaker had the possibility to pronounce the item again. Formant frequencies were measured using Praat software each 6.25 ms. The formant variations were smoothed with a 43.75 ms time window by calculating the mean values of the formants obtained for 7 successive frames (running mean value). Then, the derivation was taken to obtain the formant transition rate. The formant rate was also smoothed with a 43.75 ms window (running mean value).

Figure 2 shows, for [ai] as produced by speaker [rc] at normal rate, the formant transitions, the formant transition rate, the formant transition acceleration, in the time domain and in the plane defined by the F2/F1 parameters. Maxima and minima of the F1 and F2 frequencies, rates and accelerations were measured to characterize the formant transitions.





c) d)

e) f)

Figure 2. [ai] production, at normal rate, for speaker [rc], a) F1 and F2 formant transition in the time domain, b) corresponding formant trajectory in the F1-F2 plane, c) F1 and F2 rate in the time domain, d) formant rate trajectory in the F1 rate-F2 rate plane, e) F1 and F2 acceleration in the time domain, and f) formant acceleration trajectory in the F1 acceleration-F2 acceleration plane.

4.1. Vowel formants

The vowel formant frequencies for [aV] as produced by speaker (em) are represented in the F1/F2 plane (figure 3). The data points are the mean values of the 5 occurrences for normal and fast production (N and F). Standard deviations for F1 and F2 are also indicated. There are no significant differences between normal and fast productions. The vowels can be easily separated.

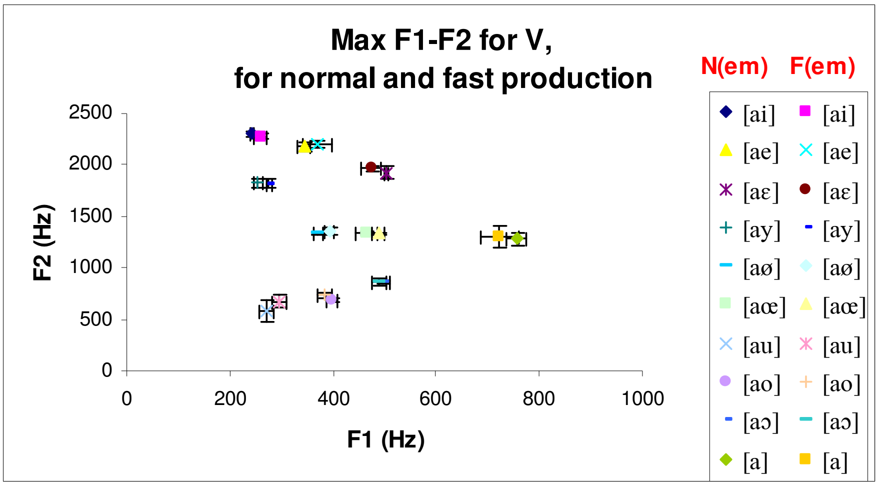


Figure 3. Mean F1 and F2 frequencies and standard deviations plotted in the F1/F2 plane for [aV] tokens produced by speaker (em), at normal (N) and fast (F) rates. Each item is produced 5 times (45 times for [a]).

4.2. [aV] transitions

4.2.1. [aV] characteristics in the F1-F2 plane

Figure 4 shows the different formant trajectories for [ai], [ae], [aɛ], [ay], [aœ], [aø] and [au], [ao], [aɔ] in the F1-F2 plane for one speaker (em). Each trajectory is a single production pronounced at normal rate. The trajectories are rather rectilinear and follow, as far as [ai] and [au] are concerned, the basic trajectories obtained by deduction (figure 1): [e], [ɛ] are situated along the formant movement of [ai]; [o] and [ɔ] on the [au] trace. Figure 4 also shows that the end parts of the trajectories corresponding to V can be characterized by small changes along the rectilinearity of the trajectory. This result corresponds to the “vowel inherent spectral changes” observed by Nearey (1986) for the final vowel in VCV sequences and by Carré et al. (2004) for isolated vowels. These characteristics are also observed for the other speakers.

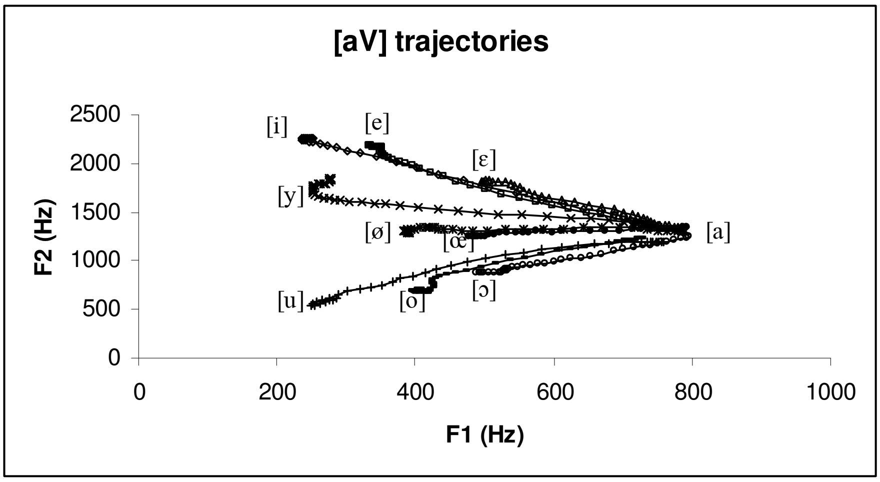
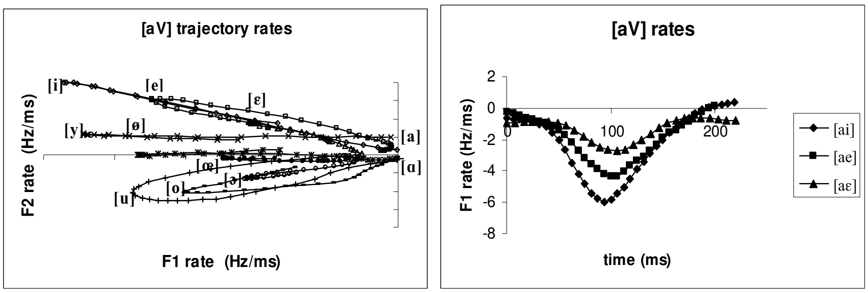


Figure 4. [aV] formant trajectories for speaker [em] (normal rate). The small changes at the ends of the trajectories corresponding to V do not deviate significantly from the rectilinearity of the trajectories.

*4.2.2. [*a*V] transition rate*

The representation of the F1-F2 transition rate as in figure 2d is used to compare the [aV] transitions for all V. Figure 5a shows the results for speaker (em) with one utterance of each [aV]. It can be observed that: if, for example, [ai], [ae], [aɛ] are compared, three distinct rate trajectories can be discriminated. The rate trajectory of [ai] is longer than that of [ae] and still longer than that of [aɛ]. In other words, the maximum rate of [ai] is greater than the maximum rates of [ae] and of [aɛ]. Figure 5b shows, for [ai], [ae], [aɛ] and for one production by speaker (em), the first formant rates in the time domain. The three vowels can be discriminated according to the maximum rates corresponding more or less to the middle of the transition ([ai] maximum rate > [ae] maximum rate > [aɛ] maximum rate). Discrimination can also be obtained throughout the transition and especially from the very beginning of the transition (from the very beginning of the production task). Figure 5b shows that the three transitions (for [ai], [ae], [aɛ]) synchronized at the beginning (corresponding to about t=50ms), stop at t=about 150ms. Because of the more or less constant duration of the transition, the 3 rates, throughout the transition, follow the inequality: [ai] rate > [ae] rate > [aɛ] rate. In principle discrimination between the three final vowels is thus possible throughout the transition and especially at its beginning.



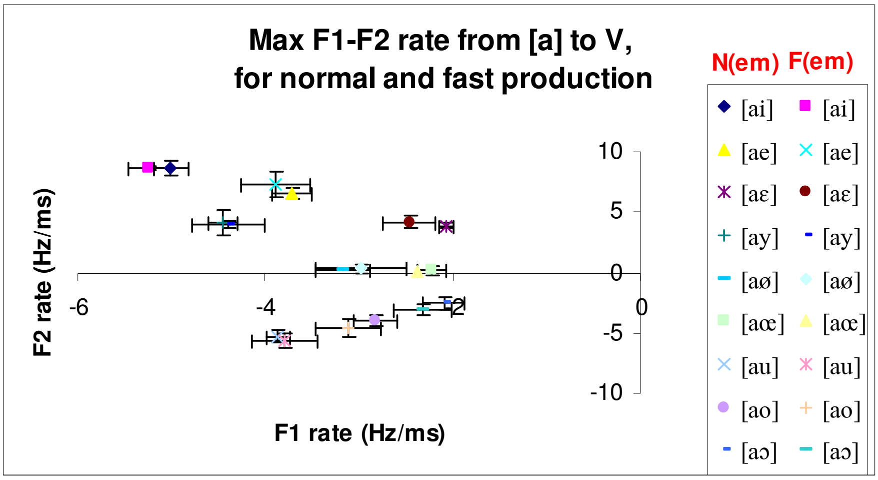
a)

b)

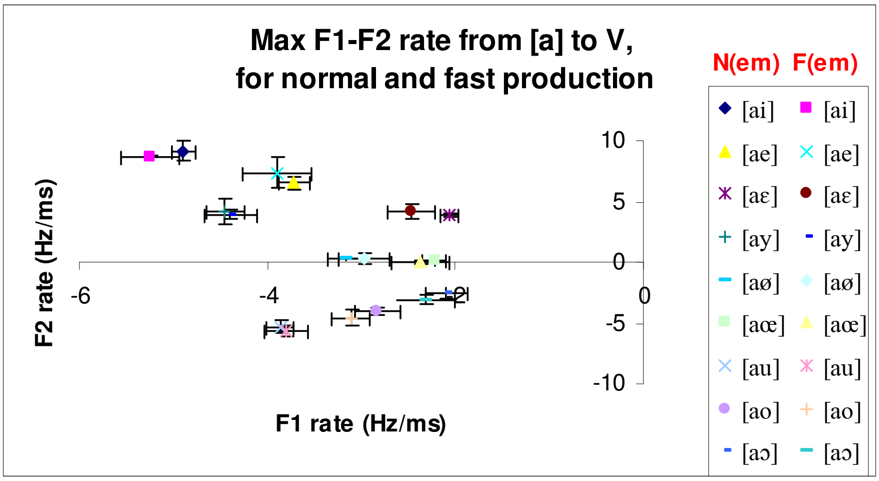
Figure 5. a) [aV] trajectory rates in the F1 rate-F2 rate plane for speaker (em) (normal rate) and b) F1 rates in the time domain for [ai], [ae], [aɛ].

Figure 6a shows the formant transition rates (mean data and standard deviations for the 5 productions) in the F1 rate/F2 rate plane for the speaker (em), for normal and fast production. The rates indicated are the maximum rates of the transitions. We do not observe large differences between normal and fast production and the vowels can be discriminated according to their rates.

According to the vowel target approach, identification would be based on formant frequency information at the end of the transition. It would not be necessary to know the characteristics of the preceding vowel (here [a]). In contrast, the dynamic approach assumes that directions and slopes of the transitions are important parameters. The identification of the vowel V would depend on the departure point in acoustic space. Standard deviations can be reduced by normalization based on the formant values of the initial [a] (figure 6b).



a)



b)

Figure 6. a) Vowel transition maximum rates of the transition [aV] for normal (N) and fast production (F) (speaker em); b) Same data but the formant frequencies F1 and F2 of each [a] vowel at the beginning of the transition are taken into account to normalize the rates.

*4.2.3. [*aV*] transition duration*

The preceding results suppose that the transition durations are more or less constant for all the [aV] produced by a same speaker. Figure 7 shows the transition durations for the speaker (em) at normal and fast production. The transition duration is defined as the interval between the maximum and minimum of the acceleration curve for F1 (see Figure 2e). The duration of the transition is around 10% smaller for faster production. The standard deviation is small for both. Our results correspond to those of (Kent and Moll, 1969; Gay, 1978), and have to be confirmed with data from more speakers.

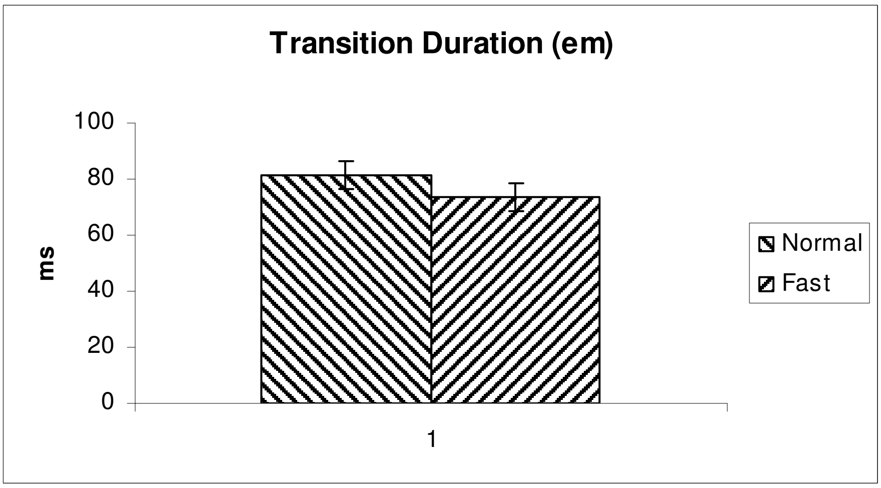


Figure 7. Transition durations for all the [aV] produced by the speaker (em) at normal and fast rate.

*4.2.4. [*aV*] transition for 10 speakers*

Our first results on the [aV] production for a single speaker lead us to hypothesize that transition rates ought to be invariant across speakers (male and female). To test the hypothesis, the first two formants of the [aV] tokens as produced by 10 speakers (5 males and 5 females) were calculated with the Praat software. The formants of the V are represented in Figure 8 in the F1-F2 plane. The standard deviations indicate that these static target representations show significant variability. However representations of transition maximum rates exhibit even more variability which is the opposite of our hypothesis (Figure 9)! These findings raise two questions: a) How accurate is the formant using classical techniques estimation, (e.g., linear prediction), especially for female voices, and b) Would it be possible to reduce variability by taking syllable rate or transition rate into account?

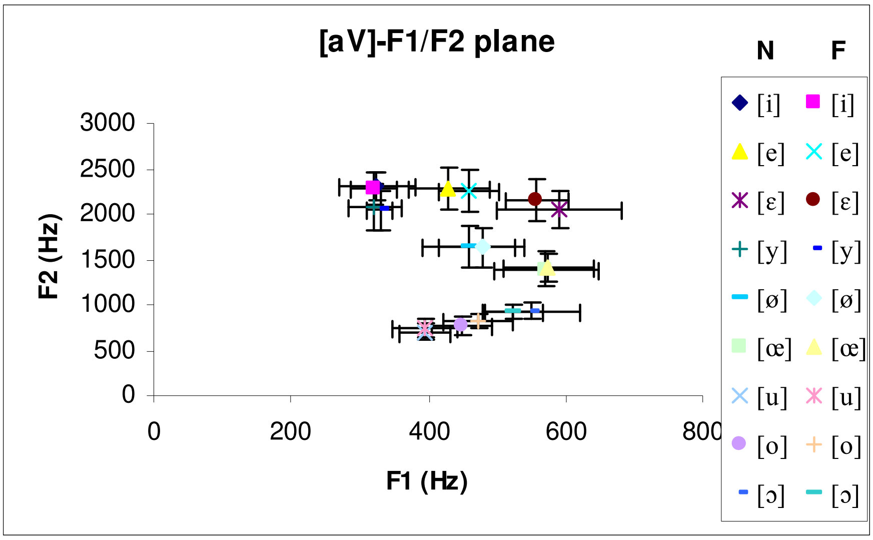


Figure 8. F1-F2 plane representation of the vowels [V] from [aV] produced at normal (N) and fast (F) rate by 10 speakers (5 males and 5 females).

First, good formant estimationtion is very difficult to obtain especially for female voices. Furthermore formant measurement errors are emphasized by the present derivative process of computing transition rates, For example, a formant frequency error of 10% can lead to an error in transition rate of 100%. Using a large time window to compute mean values can reduce these errors, but delays rate measurement. Problematic aspects of formant detection will be discussed further below.

Second, each of the speakers has his own transition rate which might also change slightly with the syllable rate (normal/fast production), for instance see the transition rate for normal and fast production for the speaker (em) figure 7.

In view of these considerations we decided to normalize the rate measurements with respect to transition duration. Transition durations were obtained from the time interval between the maximum and minimum of the acceleration curve for F1 (see Figure 2e). Figure 10 shows the new results (the reference transition duration was 100ms). The standard deviations are clearly reduced but are still greater than the ones obtained with the target formant measurements.

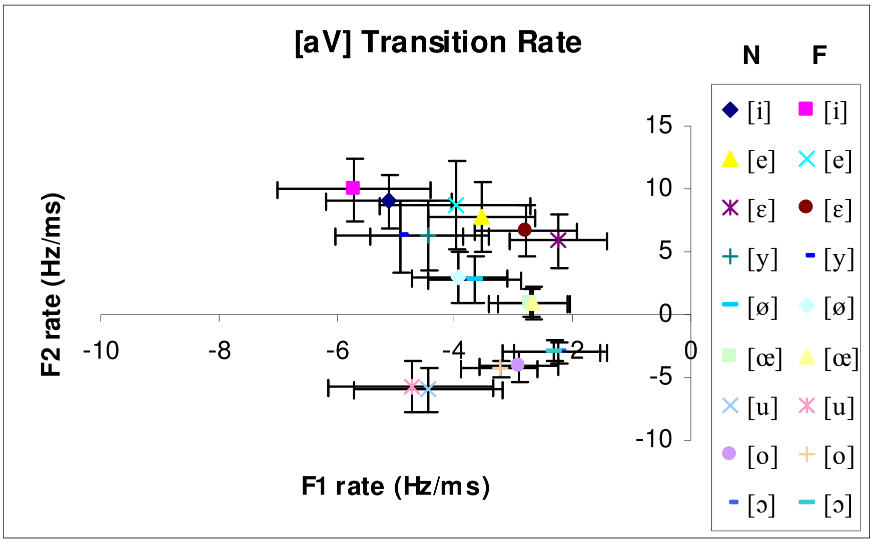


Figure 9. Maximum transition rates in the plane F1 rate versus F2 rate for [aV] uttered by 10 speakers (5 males and 5 females) for normal (N) and fast (F) production; b) same data after normalization using the transition duration.

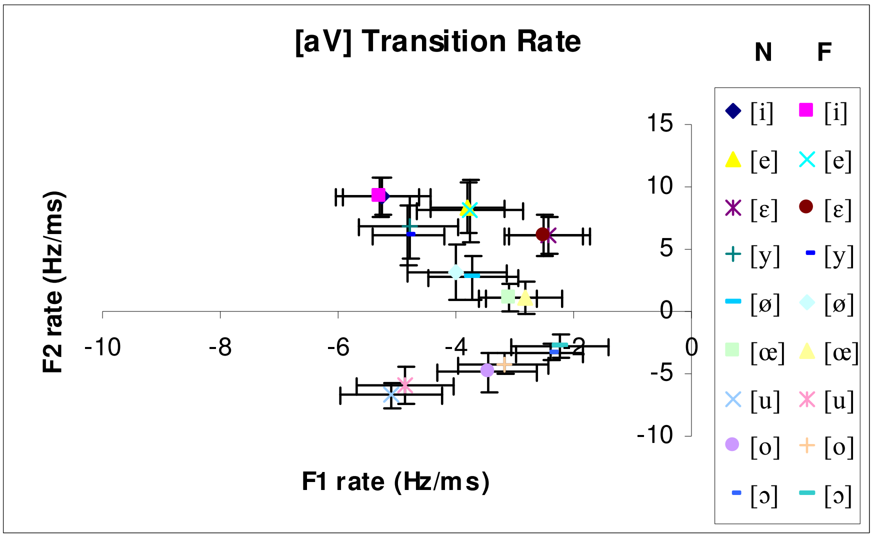


Figure 10. Maximum F2 and F1 transition rates for [aV] uttered by 10 speakers (5 males and 5 females) for normal (N) and fast (F) production after normalization based on transition durations.

5. Transition perception

Since direction and rate of transitions provide discriminating acoustic information on the vowel identities in vowel to vowel sequences, it seems possible that these two attributes could be used in perception. To test this hypothesis, trajectory stimuli outside the vowel triangle were chosen. So, the use of normal target values for the vowels was abandoned but typical rates and directions in the acoustic space were retained. Four different stimuli (A, B, C, D) were synthesized with 2 formants. The trajectories of these sequences are shown in the F1/F2 plane Figure 10, and in the time domain Figure 11. F0 is 300 Hz at the beginning of each sequence, held constant during the first quarter of the total duration decreasing to 180 Hz at the end. Possible responses for identification tests were chosen during a pre-test, i.e. [iai], [ɛuɛ], [aua], [aoa]. A fifth case, "????" was offered in case of impossible identification (no response). Twelve subjects took part in the perception tests.

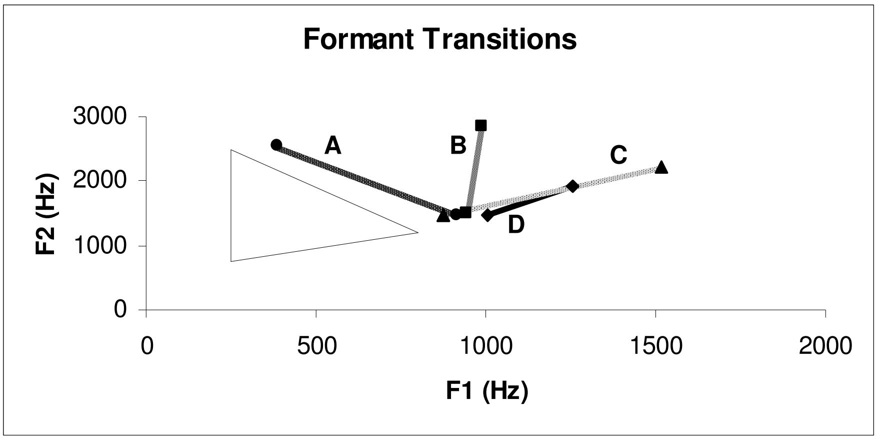


Figure 11. The four trajectories (A, B, C, D) in the plane F1-F2 and the vowel triangle. The trajectories are outside the vowel triangle. Their directions and sizes in the acoustic plane vary.

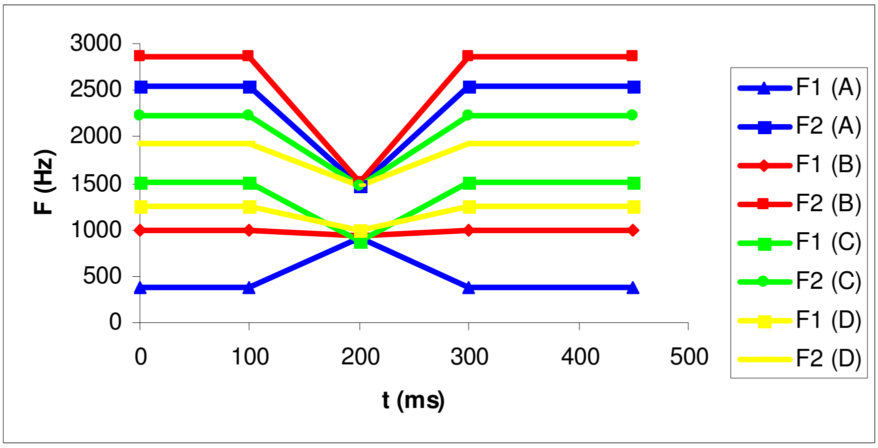


Figure 12. F1 and F2 in the time domain for the four sequences (A, B, C, D). The duration of the first part of each sequence was 100 ms, the duration of the transition was constant and equal to 100 ms. The duration of the last part was 150 ms. The rates of the transitions in Hz/ms vary. The first and last parts of each sequence were stable and equal in formant frequency. The transitions of the four sequences reached more or less the same point in the acoustic plane.

The responses (in %) are given in Figure 12. The sequence A is identified 71% of the times as [iai]. B is identified 87% as [ɛuɛ]. C is identified 95% and 96% as [aua] or [aoa] and the long trajectory corresponding to a faster rate of transition is more [aua] than the short one which is more [aoa]. The option of “no response” is generally avoided. The sequence A which has the same direction in the acoustic plane and transition rate as [ai] is perceived as /iai/; B which has more or less the same direction and rate as [uɛ] is perceived as /uɛu/; C which has the same direction and rate as [au] is perceived as /aua/ and D which has the same direction the [au] but at lower rate is more often perceived as /aoa/.

These results can be summarized by saying that the region where the 4 trajectories converge (acoustically closed to [a]) is perceived as /a/ or /u/ or /o/ depending on the direction and length (i.e. rate of the transition) of the trajectories.

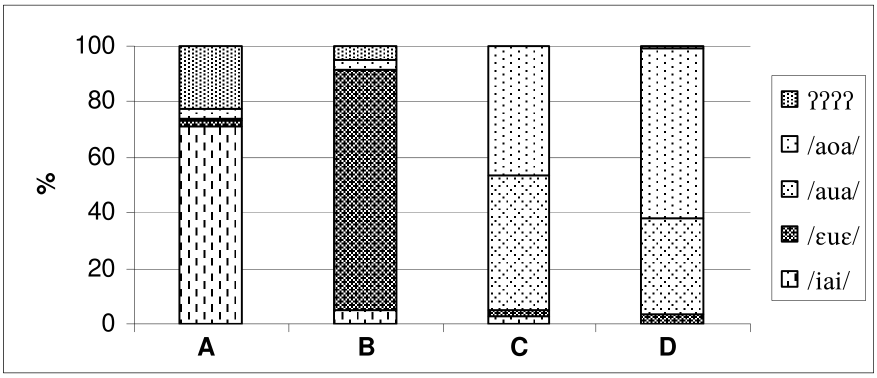


Figure 13. Results of the perception tests. The sequence A is mainly perceived as [iai], B as [iui], C as [aua] (long trajectory), D as [aoa] (short trajectory).

Discussion

At different levels our preliminary results raise several problems and questions about the dynamic approach and its consequences for the theory of speech production and perception. Also the findings motivate a closer examination of current speech analysis techniques and the methodology of perception tests.

The dynamic approach is very attractive because it may permit consonants and vowels to be integrated within a single theory. Conceivably, using the parameter of transition rate, one might propose that fast transitions tend to produce consonants, whereas slow transitions produce vowels.

In the case of perceiving V1V2 sequences, we have reported acoustic measurements indicating that signal information on V2 is available throughout the transition and especially at its very beginning. This strategy presupposes that the identity of the previous V1 has been determined. The question is: How is this information to be obtained? According to a target theory of speech perception, V2 can only be identified on the basis of its target, the goal being to reach the target irrespective of the starting point.

One of the aims of the present study has been to suggest that dynamic parameters such as direction of spectral change in acoustic space and transition rate could be more invariant across males, females and children than vowel targets. This hypothesis would make normalization in terms of static targets unnecessary. However, normalization of transition rate with respect to the different transition durations observed in production would seem necessary. Such normalization could be readily available perceptually, thanks to temporal coding and the sensitivity of the auditory system to rate (derivatives) and acceleration (Pollack, 1968; Divenyi, 2005).

The hyper-hypo speech, reduction phenomena (Lindblom, 1963; Lindblom, 1990) of fast and normal speech (Kent and Moll, 1969; Gay, 1978; van Son and Pols, 1990; van Son and Pols, 1992; Pols and van Son, 1993) should be further studied with respect to the parameters of transition direction and rate. The results obtained from experiments with 'silent center' (Strange, et al., 1983) can be explained in terms of 'dynamic specification' in the sense that is not necessary to compensate for 'undershoot' at the production level (target not reached because of coarticulation, fast speech or hypo speech) by perceptual 'overshoot' (Lindblom and Studdert-Kennedy, 1967) calculated '*not solely by the formant-frequency pattern at the point of closest approach to target, but also by the direction and rate of adjacent formant transitions*'. This finding is compatible with the assumption that, given a specification of their point of origin in phonetic (acoustic) space, direction and rate of formant transitions could be sufficient to specify the following vowel.

Our preliminary results on vowel production represent a first few steps in support of a full dynamic approch. More studies on the normalization process must also be undertaken.

It is well known that predictive coding based on a model of speech production is not well adapted for analyzing speech signals with high fundamental frequencies or with noise. Furthermore, such a technique is ill-suited to measuring spectral variations. A dynamic approach necessitates a reconsideration of analysis techniques in light of our knowledge of the auditory system. The spikes observed in auditory nerve fibers are statistically synchronized by the time domain shape of the basilar membrane excitation around the characteristic frequencies (Sachs*, et al.*, 1982).So they can give information not only on the amplitudes of spectral components but also on the shape in the time domain of the components and thus on the phases. The phase variation (-180° around formant frequencies for second order filters describing the transfer function of the vocal tract (Fant, 1960)) could be used to measure the rate of the transitions.

To attain some of these goals new tools would be needed. For example Chistovich (1982) described a model of the auditory system which detects spectral transitions without specific formant detection.

These considerations make it evident that in order to test the hypothesis of 'greater invariance in transition rates than in formant targets', it would be necessary both to improve current analysis techniques and to study more deeply the normalization of transition durations.

Perception tests of formant transitions outside the vowel triangle encourage us to study general dynamic properties of the auditory system that may be used in speech. Formant transitions can be converted into sine-waves: preliminary tests have shown results close to those obtained with formants. Differences have to be explained. Many experiments must be undertaken with such a tool creating speech illusions.

6. Conclusions

This paper follows up results previously published on the deductive approach (Carré, 2004) proposing a dynamic view of speech production, on acoustic modelling (Mrayati, et al., 1988), on structuring an acoustic tube into regions corresponding to the main places of articulation, and on the prediction of vocalic systems (Carré, 2009). The preliminary results presented here on vowels must be extended to consonants, many of which are intrinsically dynamic. At the same time, the evident importance of dynamic characteristics does not mean that static targets are not used in perception. The limits of the dynamic approach and the balance between the use of static and dynamic parameters in perception must be known. But the dynamic approach needs to develop new ways of thinking and new tools. Formant transitions cannot be obtained from a succession of static values but from directions and slopes. It means that a new tool able to measure directly these characteristics has to be developed. The dynamic approach is not a static approach plus dynamic parameters taken into account, it must be an approach intrinsically dynamic. It calls for an epistemologic study of the dynamic nature of speech (Carré, 2007).

Acknowledgements

The author thanks for their very helpful comments and stimulating discussions Eric Castelli, Pierre Divenyi, Björn Lindblom, Egidio Marsico, François Pellegrino, Michael Studdert-Kennedy, Willy Serniclaes. He thanks also Claire Grataloup for the efficient management of the perception tests.

7. References

Al-Tamimi, J., Carré, R. and Marsico, E., 2004. The status of vowels in Jordanian and Moroccan Arabic: insights from production and perception. J. Acoust. Soc. Am. 116, S2629.

Carré, R., 2004. From acoustic tube to speech production. Speech Communication 42, 227-240.

Carré, R., Pellegrino, F., Divenyi, P. 2007. Speech dynamics: epistemological aspects. In: Proc. of the ICPhS, Saarbrücken, pp. 569-572.

Carré, R., 2009. Dynamic properties of an acoustic tube: Prediction of vowel systems. Speech Communication 51, 26-51.

Carré, R. and Hombert, J. M., 2002. Variabilité phonétique en production et perception de parole : stratégies individuelles (Phonetic variabilities in speech production and perception: individual strategies). In J. Lautrey, B. Mazoyer and P. van Geert, (Eds.), *Invariants et Variabilité dans les Sciences Cognitives*, Presses de la Maison des Sciences de l'Homme, Paris.

Carré, R. and Mrayati, M., 1991. Vowel-vowel trajectories and region modeling. J. of Phonetics 19, 433-443.

Carré, R., Serniclaes, W. and Marsico, E., 2004. Production and perception of vowel categories. In: Proc. of the From Sound to Sense Conference, MIT, Cambridge.

Castelli, E. and Carré, R., 2005. Production and perception of Vietnamese vowels. In: ICSLP, Lisbon, pp. 2881-2884.

Catford, J. C., 1988. A practical introduction to phonetics. Clarendon Press, Oxford.

Chistovich, L. A., Lublinskaja, V. V., Malinikova, T. G., Ogorodnikova, E. A., Stoljarova, E. I. and Zhukov, S. J., 1982. Temporal processing of peripheral auditory patterns of speech. In R. Carlson and B. Grandström, (Eds.), *The representation of speech in the peripheral auditory system*, Elsevier Biomedical Press, Amsterdam, pp. 165-180.

Divenyi, P., Lindblom, B. and Carré, R., 1995. The role of transition velocity in the perception of V1V2 complexes. In: Proceedings of the XIIIth Int. Congress of Phonetic Sciences, Stockholm, pp. 258-261.

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

Fant, G., 1960. Acoustic theory of speech production. Mouton, The Hague.

Fowler, C., 1980. Coarticulation and theories of extrinsic timing. J. of Phonetics 8, 113-133.

Gay, T., 1978. Effect of speaking rate on vowel formant movements. J. Acoust. Soc. Am. 63, 223-230.

Hillenbrand, J. M., Getty, L. A., Clark, M. J. and Wheeler, K., 1995. Acoustic characteristics of American English vowels. J. Acoust. Soc. Am. 97, 3099-3111.

Hillenbrand, J. M. and Nearey, T. M., 1999. Identification of resynthesized /hVd/ utterances: Effects of formant contour. J. Acoust. Soc. Am. 105, 3509-3523.

Johnson, K., 1990. Contrast and normalization in vowel perception. J. of Phonetics 18, 229-254.

Johnson, K., 1997. Speaker perception without speaker normalization. An exemplar model. In K. Johnson and J. W. Mullennix, (Eds.), *Talker Variability in Speech Processing*, Academic Press, New York, pp. 145-165.

Johnson, K., Flemming, E. and Wright, R., 1993. The hyperspace effect: Phonetic targets are hyperarticulated. Language 69, 505-528.

Kent, R. D. and Moll, K. L., 1969. Vocal-tract characteristics of the stop cognates. J. Acoust. Soc. Am. 46, 1549-1555.

Lindblom, B., 1963. Spectrographic study of vowel reduction. J. Acoust. Soc. Am. 35, 1773-1781.

Lindblom, B., 1990. Explaining phonetic variation: a sketch of the H and H theory. In A. Marchal and W. J. Hardcastle, (Eds.), *Speech Production and Speech Modelling, NATO ASI Series*, Kluwer Academic Publishers, Dordrecht, pp. 403-439.

Lindblom, B. and Studdert-Kennedy, M., 1967. On the role of formant transitions in vowel perception. J. Acoust. Soc. Am. 42, 830-843.

Moon, J. S. and Lindblom, B., 1994. Interaction between duration, context and speaking style in English stressed vowels. J. Acoust. Soc. Am. 96, 40-55.

Mrayati, M., Carré, R. and Guérin, B., 1988. Distinctive regions and modes: A new theory of speech production. Speech Communication 7, 257-286.

Nearey, T. and Assmann, P., 1986. Modeling the role of inherent spectral change in vowel identification. J. Acoust. Soc. Am. 80, 1297-1308.

Nordström, P. E. and Lindblom, B., 1975. A normalization procedure for vowel formant data. In: 8th International Congress of Phonetic Science, Leeds.

Peterson, G. E. and Barney, H. L., 1952. Control methods used in the 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., 1993. Acoustics and perception of dynamic vowel segments. Speech Communication 13, 135-147.

Repp, B., Healy, A. F. and Crowder, R. G., 1979. Categories and context in the perception of isolated steady-state vowels. Journal of Experimental Psychology: Human Perception and Performance 5, 129-145.

Sachs, M., Young, E. and Miller, M., 1982. Encoding of speech features in the auditory nerve. In C. R. and G. B., (Eds.), *The Representation of Speech in the Peripheral Auditory System*, Elsevier Biomedical, Amsterdam, pp. 115-130.

Schouten, M. and van Hessen, A., 1992. Modeling phoneme perception. I: Categorical perception. J Acoust Soc Am 92, 1841-1855.

Shankweiler, D., Verbrugge, R. R. and Studdert-Kennedy, M., 1978. Insufficiency of the target for vowel perception. J. Acoust. Soc. Am. 63, S4.

Strange, W., 1989. Evolving theories of vowel perception. J. Acoust. Soc. Am. 85, 2081-2087.

Strange, W., Jenkins, J. J. and Johnson, T. L., 1983. Dynamic specification of coarticulated vowel. J. Acoust. Soc. Am. 74, 695-705.

van Son, R. J. and Pols, L. C. W., 1990. Formant frequencies of Dutch vowels in a text, read at normal and fast rate. J. Acoust. Soc. Am. 88, 1683-1693.

van Son, R. J. J. H. and Pols, L. C. W., 1992. Formant movements of Dutch vowels in a text, read at normal and fast rate. J. Acoust. Soc. Am. 92, 121-127.

Verbrugge, R. R. and Rakerd, B., 1980. Talker-independent information for vowel identity. Haskins Laboratory Status Report on Speech Research SR-62, 205-215.

Whalen, D. H., Magen, H. S., Pouplier, M. and Kang, A. M., 2004. Vowel production and perception: hyperarticulation without a hyperspace effect. Language and Speech 47, 155-174.