Acoustics'08 Paris
June 29-July 4, 2008
www.acoustics08-paris.org

Production and perception of V1V2 described in terms of formant transition rates

René Carré

Laboratoire Dynamique du Langage, UMR 5596, CNRS, Université Lyon 2, 14 Avenue
Marcelin Berthelot, 69363 Lyon cedex 07, France
recarre@wanadoo.fr

Vowels can be produced with static articulatory configurations leading to stable formant frequencies (targets). Here, an algorithm computes area functions according to the criterion of minimal deformation leading to maximal acoustic variations. Within this evolutionary dynamics, the deformations of the tube are not performed to reach targets, unknown during the process, but to move in the acoustic space in order to increase acoustic contrast. The corresponding formant trajectories in the acoustic space can be described in terms of formant transition rates. For example, following this dynamic approach, to produce [ae] from [a], the transition rates of F1 and F2 are necessary and sufficient to represent [e] at the very beginning of the transition and throughout the transition there is sufficient information to detect [e]. This assertion means that the transition duration is more or less constant. Studies of V1V2 production and perception characterized by their formant transition rates are presented. Such a representation leads to new interpretations of vowel reduction, coarticulation and normalization.

1 Introduction

Vowels are generally characterized by the first two formant frequencies. Each of them can be represented in the acoustic space (F1-F2 plane) by a dot [16]. 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 consonantal context (coarticulation) and with the speaking rate (reduction phenomena [14]), and also with the speaker and the language. The point in focus here is that the vowel specification is static and, significantly, may be taken to imply that the perceptual representation corresponds to the target values.

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?

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 [15; 21; 24; 23].

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 speech time variations. One possibility is that dynamics can be characterized by the rate of the vocalic transitions.

On the topic of transition rate, we recall the results of Kent [13]: "*the duration of a transition – and not its velocity – tends to be an invariant characteristic of VC and CV combinations*". Gay [12] 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.

The perceptual results obtained by Strange [22] in 'silent center' experiments that replaced the center of the vowel by silence of equivalent duration can be explained because this manipulation preserves the rate of the transition as well as the temporal organization (syllabic rate). Also relevant are experiments by Divenyi et al. [6] showing that, in V1V2 stimuli, V2 was perceived even when V2 and the last half of the transition was removed by gating.

In this paper, at the production level, we developed a classification of V1V2 trajectories in terms of F1, F2 rates; at the perception level, identification tests of [aea] items with different transition rates were performed.

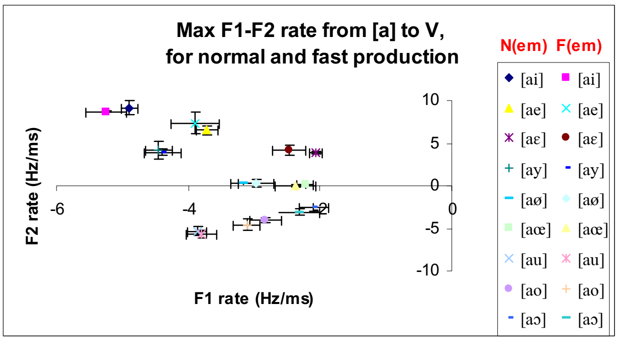
2 Production

We first recall results on the production of V1V2 transitions by French subjects [4]. Production measurements of the F1, F2 transition rates were represented in a F1 rate / F2 rate plane. In the experiments, V1 was 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.

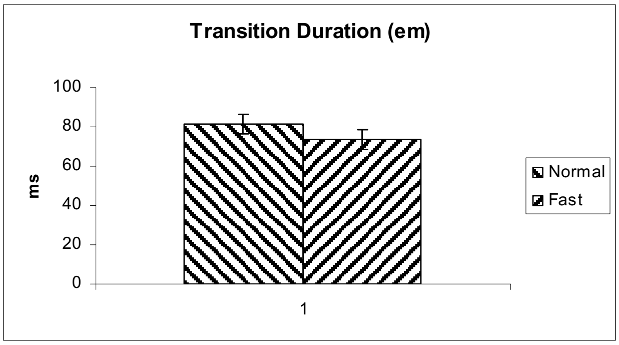
Figure 1a shows the formant transition rates (mean data and standard deviations for 5 productions) in the F1 rate/F2 rate plane for the speaker (em), for normal and fast production. The rates indicated are the maximum rates of the transitions. We do not observe large differences between normal and fast 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 characterization of the vowel V2 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].

The preceding results mean that the transition durations are more or less constant for all the [aV] produced by a same speaker. Figure 1b shows the transition durations for the speaker (em) at normal and fast production. 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 [13; 12]. This transition duration which is here around 80ms is more or less speaker dependant.



a)



b)

Figure 1. a) Vowel transition maximum rates of the transition [aV] for normal (N) and fast production (F) (speaker em); the formant frequencies F1 and F2 of each [a] vowel at the beginning of the transition were taken into account to normalize the rates; b) Transition durations for all the [aV] produced by the speaker (em) at normal and fast rate.

With such a description, it is necessary to know the starting point (here [a]) [3]. The question is: is it always necessary? If it is supposed that the transition rate is constant then a negative fast rate for F1 associated with a positive fast rate for F1 leads to [ai] and a negative fast rate for F1 associated with a negative fast rate for F2 leads to [au], then rate normalization can be applied. From this comment, a set of V1V2 transitions were classified according to their F1 and F2 transition rates. These transition rates (in Hz/ms) were calculated for the 121 V1V2 obtained with the 11 oral French vowels. The duration of the transition was fixed and equal to 100ms. Then, the corresponding Euclidian distances were calculated (module and argument). The module varies between 0 (for V1=V2) and about 15. Then an arbitrary threshold of 4 was chosen to find out the closer V1V2s. In the Table 1, the distances less than this threshold is noted as 0, above this threshold is noted 1. The argument in radian is also given.

In this case, 30 couples of V1V2 transitions on 121 can be first discriminated and if the directions of the trajectories (argument) are taken into account then only some couples are very closed such as (as expected): [ie] et [eɛ], [uo] et [oɔ], [yø] et [øœ].

We could also characterize these V1V2 transitions in terms of area function parameters: place and degree of the constriction and degree of labiality [1], then 3 main categories are obtained: without labial variation, with a positive labial variation, with a negative labial variation …

V2	[a]		[ɛ]		[e]		[i]		[ɑ]		[ɔ]		[o]		[u]		[œ]		[ø]		[y]	
V1	Mo	Arg	Mo	Arg	Mo	Arg	Mo	Arg	Mo	Arg	Mo	Arg	Mo	Arg	Mo	Arg	Mo	Arg	Mo	Arg	Mo	Arg
[a]	0	####	1	1.12	1	1.13	1	1.12	0	1.43	1	-1.2	1	-1.09	1	-0.99	0	0.71	1	0.75	1	0.76
[ɛ]	1	-1.12	0	####	0	1.13	1	1.12	1	1.26	1	1.56	1	-1.47	1	-1.35	0	-1.38	0	-0.32	0	0.14
[e]	1	-1.13	0	-1.13	0	####	0	1.11	1	1.22	1	1.44	1	1.53	1	-1.51	1	1.36	1	-1.52	0	-1.18
[i]	1	-1.12	1	-1.12	0	-1.11	0	####	1	1.2	1	1.39	1	1.47	1	1.56	1	1.29	1	1.44	1	-1.52
[ɑ]	0	-1.43	1	1.26	1	-1.22	1	-1.2	0	####	0	-0.49	1	-0.55	1	-0.58	1	1.1	1	1.03	1	0.99
[ɔ]	1	1.2	1	1.56	1	-1.44	1	-1.39	0	0.49	0	####	0	-0.66	0	-0.66	1	1.5	1	1.35	1	1.27
[o]	1	1.09	1	-1.47	1	-1.53	1	-1.47	1	0.55	0	0.66	0	####	0	-0.65	1	-1.49	1	1.49	1	1.39
[u]	1	0.99	1	1.35	1	1.51	1	-1.56	1	0.58	0	0.66	0	0.65	0	####	1	-1.34	1	-1.5	1	1.54
[œ]	0	-0.71	0	1.38	1	-1.36	1	-1.29	1	-1.1	1	-1.5	1	1.49	1	1.34	0	####	0	0.82	0	0.8
[ø]	1	-0.75	0	0.32	1	1.52	1	-1.44	1	-1.03	1	-1.35	1	-1.49	1	1.5	0	-0.82	0	####	0	0.76
[y]	1	-0.76	0	-0.14	0	1.18	1	1.52	1	-0.99	1	-1.27	1	-1.39	1	-1.54	0	-0.8	0	-0.76	0	####

Table 1. Euclidian distance (Module: Mo and Argument: Arg) between all the V1V2 transitions. The vowels are characterized by their formants F1 (in Hz) and F2 (in Hz). In yellow, the module is less than 4 (for a maximum of about 15).

3 Perception experiments

In a perception study [4], we focused 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. Results show that such transitions can be categorized as vocalic trajectories according to the direction and rate of the transition. For example, a $V_1V_2V_1$ trajectory more or less parallel and equal in size to [iui] of the vowel triangle is perceived as /iui/ though the perceived /u/ is acoustically placed at the [a] of the vowel triangle. A $V_1V_2V_1$ trajectory more or less parallel and equal in size to [aua] is perceived as /aua/ though the perceived /u/ is acoustically also placed at the [a] of the vowel triangle and a shorter in size trajectory is perceived as /aoa/ though the perceived /o/ is placed at the [a] of the vowel triangle. These results, incompatible with a static target approach, support a dynamic approach which takes for granted that humans are able to cope with these derivative – or velocity – parameters. In fact, consistent with our perceptual results, existence of velocity (and acceleration) detectors in the auditory system has been demonstrated in psychoacoustic studies [18; 7].

Recall also [2] the perception results of synthetic V1V2 with different transition durations (from 50m to 300ms). For example, in French, [ai] is perceived /ai/ with transition durations between 50 and around 200ms, then, from 200 to 300ms, /aɛi/ is perceived though there is no acoustic event during the transition (no landmark). We hypothesized that the segmentation of V-V sound streams would be performed according to the listener's phonetic experience [Johnson, 1997] of extracting language-specific attributes of speech sounds: syllabic segmentation in French. Thus, we assumed that the listener detects the vowel corresponding to the acoustic values of the signal at the expected boundary (i.e. at 200ms corresponding to the vowel /ɛ/). Here, another hypothesis is tested: if the rate of the transitions is used for vowel perception, then our preceding results could be explained by the different rates corresponding to the different durations of the transition.

To test this new hypothesis, different V1V2V3 items were synthesized with V1 = V3 = [a], V2 = [e], V1 duration = 80ms, V3 duration = 100ms, V1V2 transition duration = V2V3 transition duration = 50, 75, 100, 125, 150, 175, 200, 225, 250ms (see figure 2a with transition duration = 100ms). All the 9 items were made of the vowel targets V1= [a], V2= [e], V3= [a]. If the transition duration reference is around 100ms as observed in production, then [aea] with 50ms transition duration should be perceived as /aia/. The vowel transition duration reference could be an important factor used for transition rate normalization. In the perception experiments, to try to test this factor, each [aea] was preceded by a synthetic V1V2 with two possible different transition durations and thus with different transition rates. V1 = [a] and V2 = [e], V1 duration = 80ms, V2 duration = 100ms, and V1V2 transition duration = either 50 ms ([ae]50) or 150ms ([ae]150) (figure 2b). The inter-stimuli duration was 400ms. If the [ae] transition rate is used as a perceptual reference parameter for the identification of the following [aea], then different results according to this parameter must be observed: the identification curves must be moved. For example, in case of full use of the rate of [ae]50 as reference, we could expect to have no identification of /aia/ or /aja/; in case of full use of the rate of [ae]150 as reference, we could expect to have maximum of /aea/ identification for the item [aea] with 150ms transition duration.

For the perception tests, 5 subjects took part in the experiments. The [aea] sequence had to be identified among 8 possible choices: /aja, aia, aea, aeja, aeɛa, aɛeɛa, aɛiɛa/ and no answer.

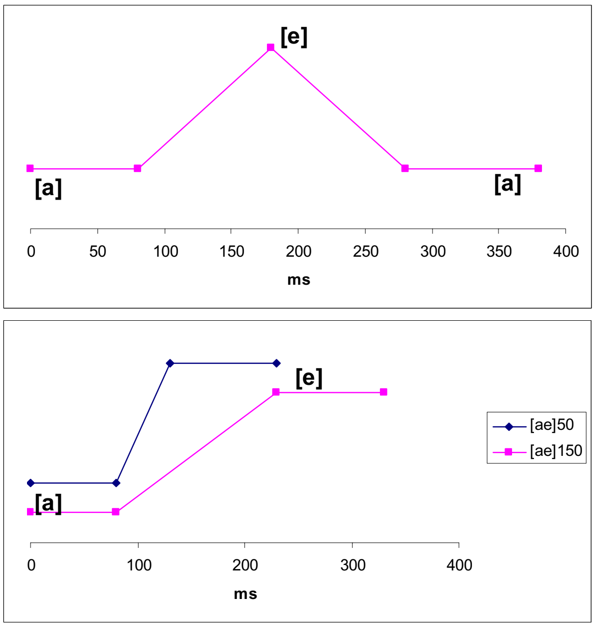
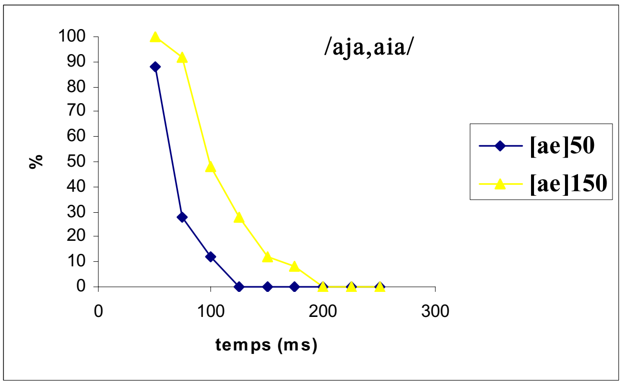
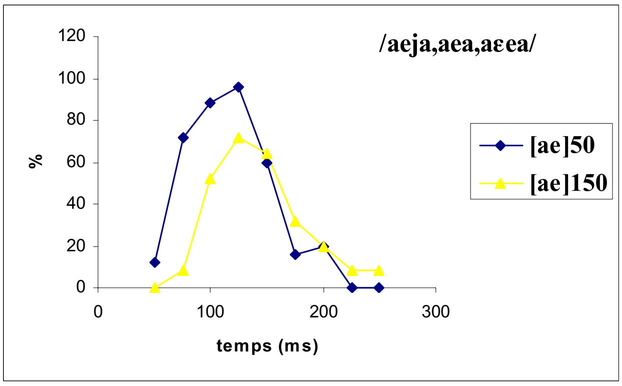


Figure 2. a) Characteristics in the time domain of the [aea] succession to be tested (the transition duration is here 50ms); b) Characteristics in the time domain of the two preceding [ae] signals. The inter-stimuli distance was 400ms.

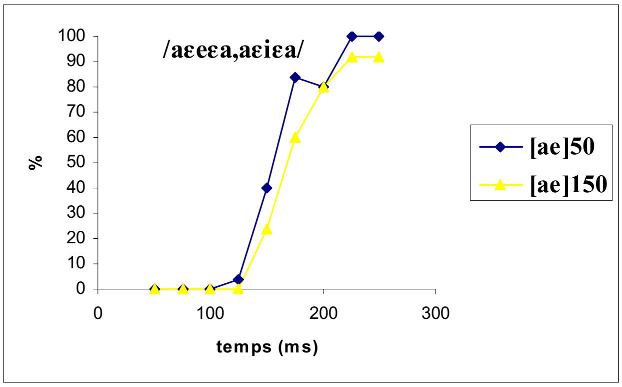
Figure 3 shows the results obtained with the 5 subjects. The different responses were grouped into 3 main classes: /aja/ and /aia/; /aeja/, /aea/ and /aeɛa/; /aɛeɛa/ and /aɛiɛa/.



a)



b)



c)

Figure 3. Percentage of responses for two different [ae] primes: one with 50ms transition duration, the other with 150ms. The responses are grouped into three main classes a) /aja/ and /aia/ b) /aeja/, /aea/ and /aeɛa/ c) /aɛeɛa/ and /aɛiɛa/

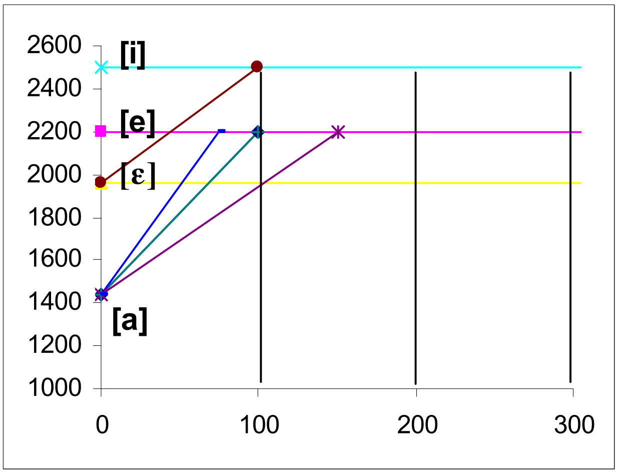


Figure 4. Three different transition rates from [a] to [e] and one transition from [ɛ] to [i]. The violet transition rate from [a] is around the same as the transition rate from [ɛ]. If the reference transition rate is 100ms, the green transition should be perceived as /ae/, the blue as /ai/ and the violet as /aɛ/, then /ɛi/.

As expected, /aja/ is more often perceived with the preceding reference [ae]150 than for [ae]50 but the percentage of identification is not zero for [ae]50. It means that such preceding primes cannot move too much the boundary. /aea/ is perceived for transition duration around 100ms as expected with an influence of the prime. And /aɛiɛa/ is perceived for transition duration more than 150ms without influence of the prime. The question is why the vowel /i/ is perceived here? Figure 4 can help for an explanation. If the reference transition duration is 100ms, then [aea], with a transition duration of 75ms, should be perceived as /aia/. With transition duration of 150ms, because of the rate of the transition, first /aɛ/ is perceived, and then the rate from /ɛ/ corresponding to the [ɛi] rate, /i/ is perceived. Such an explanation needs to consider an integration windowing of the rate of about 100ms. The perception of /aja/ or /aia/ for small transition duration could be explained by an overshoot effect [15]. But, the perception of /aɛiɛa/ for long transition duration cannot be explained by the overshoot hypothesis; the role of the transition rate should be emphasized.

4 Discussions and conclusions

Our results point out the role of the transition rate in V1V2V1 perception. Therefore, the time domain could play an important role in the identification of vowels [10]. But our experiments must be extended to study the validity of this assertion with more subjects (it seems that some of them do not perceive differences between the two different primes), with other languages such as English (recall that, for the perception of V1V2 with different transition durations, the results were different comparing French and English [2]). Our experiments must also be tested with children. Such a representation leads to new interpretations of normalization, invariance, vowel reduction, perceptual overshoot, co-articulation.

Without saying that static information is not used in speech perception, the dynamic approach needs to revisit different questions [3]. With the static approach, normalization is generally proposed in the frequency domain. With the dynamic approach, normalization should be in the time domain (because, for example, the duration of the vowel-to-vowel transition is speaker dependant). If important information is in between static targets then the production undershoot (reduction) and corresponding perceptual overshoot must be revisited. Is it necessary to compute the target to be reached? The co-articulation phenomena should also be reconsidered [5]. With the static approach, intrinsic characteristics of vowels correspond to target formant frequencies; with the dynamic approach, they are extrinsic characteristics!

Furthermore, it would be necessary to reconsider analysis techniques: with the static approach, analysis techniques in the frequency domain are well adapted: formant measurements at a specific time can give information on targets, onsets.... With the dynamic approach, the measurements of the rates of the formant transitions need to consider analysis techniques in the time domain as in our 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 [19]. 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 [9]) could be used to measure the rate of the transitions. Perception experiments with synthetic signals characterized by flat amplitude spectrum and specific phase variations show that glides can be obtained [20].

From this dynamic epistemological approach, we have to reconsider many old results. For example, Furui [11] demonstrated the importance of the dynamic cepstral coefficients in speech recognition. See also Dusan [8]. It is also possible to explain the asymmetry in vowel perception observed by Polka [17]. For example, from [a] to [e], the transition rate must be between two values: the one to [ɛ] and the one to [i]. But from [e] to [a], the rate has to be very fast because there is no vowel after [a].

All these considerations and results lead to test the hypothesis of 'greater invariance in transition rates than in formant targets'.

Acknowledgments

The author thanks Emmanuel Ferragne, Egidio Marsico François Pellegrino for stimulating discussions. This work is supported by the French Agence Nationale de la Recherche (project NT05-3_43182 CL^2, P.I. F. Pellegrino).

References

[1] Carré, R., Bourdeau, M. and Tubach, J. P. "Vowel-vowel production: the distinctive region model (DRM) and vocalic harmony," Phonetica 52, 205-214, (1995).

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

[3] Carré, R., Pellegrino, F. and Divenyi, P. "Speech dynamics: epistemological aspects," Proc. of the ICPhS, (Saarbrücken), pp. 569-572 (2007).

[4] Carré, R. (Submitted). "Signal dynamics in the production and perception of vowels," in Approaches to Phonological Complexity, edited by I. Chitoran, C. Coupé, E. Marsico and F. Pellegrino (Mouton de Gruyter, Berlin, New York).

[5] Carré, R. (submitted). "Coarticulation en production de parole: aspects acoustiques," in La Coarticulation : Indices, Direction et Représentation, edited by M. Embarki and C. Dodane (L'Harmattan, collection Langue & Parole, Paris).

[6] Divenyi, P., Lindblom, B. and Carré, R. "The role of transition velocity in the perception of V1V2 complexes," Proceedings of the XIIIth Int. Congress of Phonetic Sciences, (Stockholm), pp. 258-261 (1995).

[7] 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. Cheveigné and S. McAdams (Springer-Verlag, New York) pp. 176-184.

[8] Dusan, S. "On the relevance of some spectral and temporal patterns for vowel classification," Speech Communication 49, 71-82, (2007).

[9] Fant, G. Acoustic theory of speech production (Mouton, The Hague) (1960).

[10] Fowler, C. "Coarticulation and theories of extrinsic timing," J. of Phonetics 8, 113-133, (1980).

[11] Furui, S. "On the role of spectral transition for speech recognition," J. Acoust. Soc. Am. 80, 1016-1025, (1986).

[12] Gay, T. "Effect of speaking rate on vowel formant movements," J. Acoust. Soc. Am. 63, 223-230, (1978).

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

[14] Lindblom, B. "Spectrographic study of vowel reduction," J. Acoust. Soc. Am. 35, 1773-1781, (1963).

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

[16] Peterson, G. E. and Barney, H. L. "Control methods used in the study of the vowels," J. Acoust. Soc. Am. 24, 175-184, (1952).

[17] Polka, L. and Bohn, O.-S. "Asymmetries in vowel perception," Speech Communication 41, 221-231, (2003).

[18] Pollack, I. "Detection of rate of change of auditory frequency," J. Exp. Psychol. 77, 535-541, (1968).

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

[20] Schroeder, M. R. and Strube, H. W. "Flat-spectrum speech," J Acoust Soc Am 79, 1580-1583, (1986).

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

[22] Strange, W., Jenkins, J. J. and Johnson, T. L. "Dynamic specification of coarticulated vowel," J. Acoust. Soc. Am. 74, 695-705, (1983).

[23] Strange, W. "Evolving theories of vowel perception," J. Acoust. Soc. Am. 85, 2081-2087, (1989).

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