Production and perception of pseudo-V1CV2 outside the vowel triangle: Speech illusion effects

Tran Thi Anh Xuan[1], *Viet Son Nguyen*[1], *Eric Castelli*[1], *René Carré*[2]

[1] International Research Institute MICA, HUST - CNRS/UMI2954 - Grenoble INP, Hanoi University of Sciences and Technology, Hanoi, Vietnam

[2] Laboratoire Dynamique du Langage - Université de Lyon 2 - CNRS/URM5596 - Lyon, France

Anh-Xuan.Tran@mica.edu.vn, Viet-Son.Nguyen@mica.edu.vn, Eric.Castelli@mica.edu.vn, Recarre@orange.fr

Abstract

Vowels are generally described with static articulatory configurations represented by targets in the acoustic space: typically, formant frequencies in the F1-F2 and F2-F3 planes. Plosive consonants can be described in terms of places of articulation, represented by locus or locus equations in an acoustic plane. But how are a given vowel and a given consonant identified, when produced with different acoustic characteristics and in different environments? To which extent do listeners use contextual information? To which extent do they use normalization, and of which kind? These questions lead to studying both vowels and consonants from a dynamic point of view. At this level, what are the respective roles of static targets and dynamics information? Previous studies reveal that synthesized transitions situated on a F1-F2 plane but beyond the values observed in natural speech can be perceived as V1V2: that is, vowel-to-vowel transitions can be characterized simply by the direction and rate of the transitions, even when absolute frequency values are outside of the vowel triangle. The present paper extends the investigation to consonants: it reports new experiments showing that perception of pseudo-V1CV2 can also be obtained with formant transitions situated outside the vowel triangle.

Index Terms: speech dynamics, consonant perception, vowel perception, speech illusion effects.

1. Introduction

Vowels are generally characterized by the first two or three formant frequencies. Each of them can be represented as a dot in the acoustic space (F1-F2 and F2-F3 planes) [1], and specified in terms of underlying 'targets': context- and duration-independent formant values as obtained by fitting "decaying exponentials" to the data points [2]. Such static representations may come to be considered as reflecting perceptual reality: granting the status of perceptual representations to the target values.

However, formant frequencies for vowels vary considerably with the consonantal context (co-articulation) and with speaking rate/reduction phenomena [3]. This raises a central research issue: how is the perceptual representation obtained if the vowel targets depend on the speaker, and are rarely reached in spontaneous speech production?

Sensory systems have been shown experimentally to be more sensitive to changing stimulus patterns than to purely steady-state ones [4, 5]. In this light, it appears justified to look for an alternative to static targets: a specification that recognizes the true significance of the variation of the signal over time. 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 [6]. So they can be characterized by their direction.

On the topic of transition duration, we recall the results of Kent [7]: "*the duration of a transition – and not its velocity – tends to be an invariant characteristic of VC and CV combinations*". Gay [8] 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 CVs with a constant C and varying Vs, it follows that the transition rate depends on the vowel to be produced. At the very beginning of the transition and throughout the transition there is sufficient information to detect the vowel to be produced. If the perception of the following sound is based on the syllabic duration, on the transition direction and rate, then we can explain the perceptual results obtained by Strange [9] 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).

In a perceptual study [10], synthesized transitions situated outside the traditional F1-F2 vowel triangle were perceived as vowel-to-vowel transitions. In this case, there is no reference to any vowel targets in the vowel triangle. The results of this study can be summarized by saying that the region where 4 trajectories converge (acoustically closed to [a]) was perceived as [a] or [u] or [o] depending on the direction and length (i.e. rate of the transition) of the trajectories.

We intend to use the same kind of synthetic experiment leading to the perception of vocalic acoustic illusions for the study of consonant perception. We hypothesize that the transition rates are sufficient for consonant discrimination: so the consonant C is synthesized without burst. This point could be discussed [11]. This study has been carried out as part of an investigation into the production and perception of speech sounds by Vietnamese subjects [12].

2. Consonant perception experiments

In order to study consonant perception, two experiments (non-illusion test and illusion test) are realized in which a V1CV2 item is synthesized by means of a formant synthesizer (a sine wave synthesizer could also be used [13]). The main difference between these two tests is the acoustic context of the consonant C (vowels V1 and V2): for the non-illusion test, it is inside the vowel triangle; for the illusion test, it is outside of it. Both experiments (non-illusion and illusion) are carried out with ten Vietnamese subjects (five men M1, M2, M3, M4, M5, and five women W1, W2, W3, W4, W5). Each subject listens to three times 60 V1CV2 stimuli presented in random

order, and chooses which consonant is perceived, among: *b, d,* and *g.* These three letters correspond, in the Vietnamese alphabet, to /ɓ/, /ɗ/ and /ɣ/: two preglottalized (injective) stops and a fricative, whose point of articulation is labial, dental and velar, respectively. The choice "NAK" (non-acknowledgment) is selected when the listener does not perceive any of /ɓ/, /ɗ/ or /ɣ/ in the stimulus.

2.1. Non-illusion experiment

In this experiment, the V1CV2 items is synthesized in which the two vowels V1 and V2 are situated inside the vowel triangle, with V1 and V2 very close to vowel [i] and vowel [a], respectively. The durations of V1 and V2 are 100ms, and 120ms, respectively. The durations of the transitions V1C and CV2 are 30ms. The consonant C (duration = 30ms) is synthesized for different formant values: F1 varies by step between 100Hz and 300Hz; F2 varies by step between 500Hz and 2500Hz; and F3 varies by step between 2000Hz and 3500Hz.

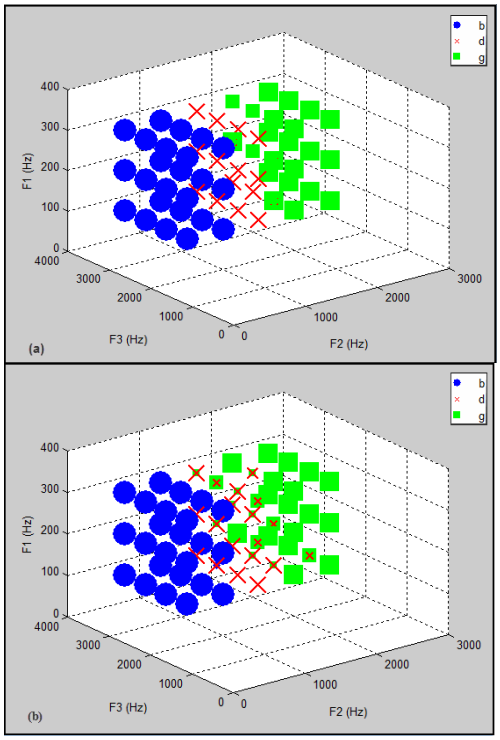

Figure 1: *The perceptive results of two subjects in the non-illusion experiment: one female subject W1 in (a), and one male M1 in (b). The blue circle corresponds to the consonant /ɓ/, the red cross corresponds to the consonant /ɗ/, and the green square corresponds to the consonant /ɣ/.*

Table 1 shows the main results of the perception test. The formant values correspond to the formant transition end points. The different items differ from the formant transition rates. The average correct recognition rates are calculated for ten subjects. It is interesting to note that: (1) all the subjects can hear easily the three consonants /ɓ, ɗ, ɣ/ (the score of NAK is very small); (2) they can recognize the three consonants /ɓ, ɗ, ɣ/ with high score; the best score of the consonants /ɓ/, /ɗ/, and /ɣ/ are 100%, 94%, and 93%, respectively; (3) the consonant /ɓ/ is perceived more easily than the others /ɗ/ and /ɣ/ (with the highest score, and without overlap with /ɗ/ and g/); the consonants /ɗ/ and /ɣ/ are less distinct with some overlaps; (4) the region of the consonant /ɗ/ is smaller than the one of the consonant /ɓ/ and /ɣ/; (5) in spite of some overlaps between the consonant /ɗ/, and the consonant /ɣ/, we still distinguish the three regions corresponding to these three consonants; (6) the F2 formant plays an important role that makes possible the discrimination of the three consonants /ɓ, ɗ, ɣ/; this agrees with the results obtained in studies of Liberman [14, 15] and Nguyen [14, 15] on F2. Figure 1 shows an example of the perceptive results of two subjects (one female subject W1 and one male M1) in the F1-F2-F3 3-D space.

Table 1. *Main perceptive results (%) in the non-illusion experiment. The average correct recognition rates are calculated for ten subjects.*

Formant			Correct recognition rate			
F1	F2	F3	/ɓ/	/ɗ/	/ɣ/	NAK
100	500	2000	100	0	0	0
200	500	2000	100	0	0	0
300	500	2000	100	0	0	0
100	500	2500	100	0	0	0
200	500	2500	100	0	0	0
300	500	2500	100	0	0	0
100	500	3000	100	0	0	0
200	500	3000	100	0	0	0
300	500	3000	100	0	0	0
100	500	3500	100	0	0	0
200	500	3500	100	0	0	0
300	500	3500	100	0	0	0
100	1000	2000	100	0	0	0
200	1000	2000	100	0	0	0
300	1000	2000	100	0	0	0
100	1000	2500	100	0	0	0
200	1000	2500	100	0	0	0
300	1000	2500	100	0	0	0
100	1000	3000	100	0	0	0
200	1000	3000	100	0	0	0
300	1000	3000	100	0	0	0
100	1000	3500	100	0	0	0
200	1000	3500	100	0	0	0
300	1000	3500	100	0	0	0
100	1500	2000	10	84	3	3
100	1500	2500	17	80	3	0
300	1500	2500	7	87	7	0
200	1500	3000	0	89	11	0
300	1500	3000	0	94	6	0
100	1500	3500	3	91	6	0
200	1500	3500	3	70	27	0
300	1500	3500	3	85	12	0
200	2000	2000	0	16	84	0
300	2000	2000	0	15	85	0
300	2000	2500	0	30	70	0
100	2500	2000	6	23	71	0
200	2500	2000	0	10	90	0
300	2500	2000	3	3	94	0
200	2500	2500	0	17	83	0
300	2500	2500	0	18	82	0
200	2500	3000	3	12	85	0
300	2500	3000	0	25	75	0
200	2500	3500	3	16	81	0
300	2500	3500	0	18	82	0

The blue circle corresponds to the consonant /ɓ/, the red cross

corresponds to the consonant /ɗ/, and the green square corresponds to the consonant /ɣ/. A sign (circle, cross, or square) will be marked in the 3-D space if the correct recognition rate of the consonant (/ɓ/, /ɗ/, or /ɣ/) is higher than 50%. The sign dimension is also proportional to the correct recognition rate value. We can observe that: (1) both subjects can recognize easily the three consonants /ɓ, ɗ, ɣ/ (the dimension of circles, crosses and squares are great); (2) both subjects discriminate the three consonants by more or less the same values of F1, F2, and F3 (though there is a small overlap between the consonant /ɗ/ with /ɓ/ and /ɣ/ in the perceptive results of M1, the three regions corresponding to the three consonants /ɓ, ɗ, ɣ/ are distinct).

2.2. Illusion experiment

The illusion experiment is performed to find out the region of the consonants /ɓ, ɗ, ɣ/ in the F1-F2 and F2-F3 planes in V1CV2 context in which the pseudo-vowels V1, V2 are situated outside the vocalic triangle. The V1CV2 item is synthesized from the first three formants in which the trajectory of the sequence V1V2 is more or less parallel to the trajectory [i-a] in the vowel triangle: the F1, F2 values of pseudo-vowel V1 are 420Hz, 2680Hz, respectively, and the ones of pseudo-vowel V2 are 1000Hz, 1490Hz, respectively (see Figure 2).

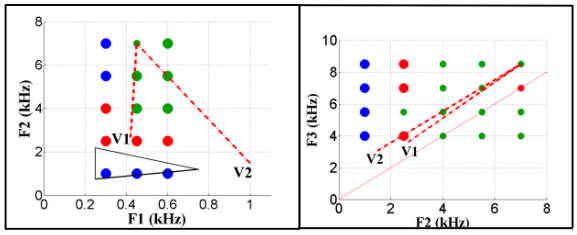

Figure 2: *The V1CV2 stimuli in the illusion experiment: the pseudo-vowels V1, V2 are situated outside the vocalic triangle; the consonant C is synthesized without burst for different formant values of F1, F2, F3.*

The durations of V1 and V2 are 100ms, and 120ms, respectively. The durations of V1C and CV2 are 30ms. The consonant C is synthesized without burst for different formant values yielding 60 stimuli of V1CV2 items (the circle points in Figure 2): F1 varies by steps between 300 Hz and 600 Hz; F2 varies by step between 1000 Hz and 7000 Hz; and F3 varies by step between 4000 Hz and 8500 Hz. In all above figures, the dot size is proportional to the recognition rate for a specific consonant. Almost all the subjects still recognize the three consonants /ɓ, ɗ, ɣ/, but there is one subject (M5) who cannot recognize the consonants /ɗ, ɣ/: his responses are either /ɓ/ or NAK. The results of M5 call for a separate analysis. Also we decide to calculate the average correct recognition rates of nine subjects (without M5). The main results are shown in Table 2. It can be noted that: (i) almost all the subjects still recognize the three consonants /ɓ, ɗ, ɣ/, but with a worse score than the one in the non-illusion test (recall that the perceptive tests are performed in an abnormal situation); (ii) the region of the consonant /ɗ/ is still smaller than the one of the consonants /ɓ/ and /ɣ/; (iii) in spite of some overlaps among the three consonants /ɓ/, /ɗ/, and /ɣ/, we still find out three distinct regions corresponding to the three consonants; (iv) although the test is carried out in a context where the pseudo-vowels V1, V2 are situated outside the vocalic triangle, F2 plays again an important role to discriminate the three consonants /ɓ, ɗ, ɣ/. Figure 3 presents the perceptive results of two subjects (one female subject W1, and one male M1) in a F1-F2-F3 3D space. We can note that: (i) almost V1CV2 can be perceived easily by both subjects (the dimension of almost circles, crosses and squares are great); (ii) there exists some overlap between /ɗ/ with /ɓ/ and /ɣ/; in particular, for the male subject M1, there is an overlap between the consonant /ɓ/ with the consonant /ɣ/ where the consonant /ɓ/ is recognized with very high values of F2 and small one of F3.

Table 2. *Main perceptive results (%) in the illusion experiment. The average correct recognition rates are calculated for nine subjects (without the results of the subject M5).*

Formant			Correct recognition rate			
F1	F2	F3	/ɓ/	/ɗ/	/ɣ/	NAK
300	1000	4000	85	13	2	0
450	1000	4000	93	4	3	0
600	1000	4000	78	0	6	16
300	1000	5500	92	8	0	0
450	1000	5500	97	3	0	0
600	1000	5500	74	4	6	16
300	1000	7000	92	8	0	0
450	1000	7000	97	0	0	0
300	1000	8500	92	8	0	0
450	1000	8500	96	0	4	0
300	2500	4000	19	74	7	0
450	2500	4000	0	88	12	0
300	2500	5500	11	71	18	0
450	2500	5500	0	92	8	0
450	2500	7000	0	93	7	0
450	2500	8500	6	92	2	0
600	4000	4000	4	10	86	0
600	4000	5500	7	13	75	5
600	4000	7000	3	18	79	0
600	4000	8500	4	6	82	8
600	5500	4000	14	14	72	0
600	5500	5500	8	7	85	0
600	5500	7000	12	7	81	0
600	5500	8500	7	4	89	0
600	7000	4000	4	7	81	8
600	7000	5500	10	16	74	0
600	7000	7000	15	4	81	0
600	7000	8500	6	12	82	0

3. Discussion

The dynamic approach is attractive because it potentially allows for the integration of consonants and vowels 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, acoustic measurements indicate 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.

In this study, the results of both experiments (non-illusion and illusion) show that the subjects can recognize and discriminate the three consonants /ɓ, ɗ, ɣ/ in the two different vowel contexts (inside and outside the vocalic triangle). The corresponding items differ from the transition rates. In the illusion experiment, the three consonants /ɓ, ɗ, ɣ/ are perceived with more difficulty (the correct recognition rates are smaller) than they are done in the non-illusion one. This

can be explained by the fact that the perceptive tests in the illusion experiment are carried out in an abnormal situation.

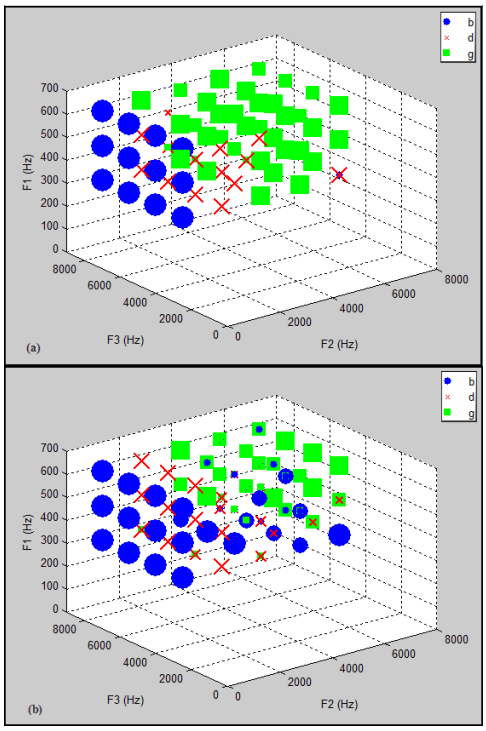

Figure 3: *The perceptive results of two subjects in the illusion experiment: one female subject W1 in (a), and one male M1 in (b). The blue circle corresponds to the consonant /ɓ/, the red cross corresponds to the consonant /ɗ/, and the green square corresponds to the consonant /ɣ/.*

However, the two experiments emphasize the important role of F2 in distinguishing the three consonants /ɓ, ɗ, ɣ/. Comparing the results of the both experiments, we can realize that as F2 increases (from 500Hz to 2500Hz in the non-illusion experiment, or from 1000Hz to 7000Hz in the illusion one), the order of the three distinct regions corresponding to the three consonants /ɓ, ɗ, ɣ/ does not change: the consonant /ɓ/ is perceived with the lowest F2 values, and the consonant /ɣ/ is recognized with the highest ones. Though there is an overlap between the F2 values in the non-illusion experiment and the illusion one (from 1000Hz to 2500Hz), the positions of the consonants /ɓ, ɗ, ɣ/ in both experiments are different. This confirms that the aspect of static F2 value is not important, but the relative F2 variation in relation to the pseudo-vowels and the dynamic value (i.e. rate of the F2 transition) of the trajectories V1C and/or CV2 play a significant role.

We interpret these results as suggesting 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 (and depending on the speaker) 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 [4, 5].

Our preliminary results on consonant perception with transitions outside the vowel triangle represent another step (after the perception of vowel-vowel transitions [10]) in support of a full dynamic approach. Further experiments with other combinations of vowels (/iu/, /au/) must be undertaken. More studies on the normalization process must also be undertaken.

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 [16]. 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.

To attain some of these goals new tools would be needed. For example: Chistovich [17] described a model of the auditory system which detects spectral transitions without specific formant detection. For the notation of vowels, Vaissière [18, 19] also develops new tools that reflect the nature of vowels more adequately than the articulatory description in IPA format.

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.

4. Conclusions

This paper follows up results previously published on the deductive approach [20] proposing a dynamic view of speech production, and on the prediction of vocalic systems [21]. A static approach fails to properly acknowledge the fundamentally dynamic aspect of formant evolution and the intrinsic temporal characteristics of speech sound [18]. In short, the fact that most speech theories can still be qualified as static, makes it imperative to stress the necessity of dynamic studies. The preliminary results presented here on consonants following the ones on vowels [10] clearly show the importance of dynamic characteristics – which 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 epistemological study of the dynamic nature of speech [22]. With such an approach, in syllabic co-production, traditional static targets are extrinsic values whereas transition parameters become intrinsic values.

5. Acknowledgements

Financial support from Agence Nationale de la Recherche (contract ANR-2010-BLAN-1916-04: Phonetic and Phonological Asymetries in Syllable) is gratefully acknowledged. This work is also supported by the Vietnamese government project KC.03.07/11-15.

6. References

[1] Peterson, G. E. and Barney, H. L., "Control methods used in the study of the vowels," *Journal of the Acoustical Society of America,* vol. 24, pp. 175-184, 1952.

[2] Moon, J. S. and Lindblom, B., "Interaction between duration, context and speaking style in English stressed vowels," *Journal of the Acoustical Society of America,* vol. 96, pp. 40-55, 1994.

[3] Lindblom, B., "Spectrographic study of vowel reduction," *Journal of the Acoustical Society of America,* vol. 35, pp. 1773-1781, 1963.

[4] Pollack, I., "Detection of rate of change of auditory frequency," *J. Exp. Psychol.,* vol. 77, pp. 535-541, 1968.

[5] Divenyi, P. L., "Frequency change velocity detector: A bird or a red herring?," in *Auditory Signal Processing: Physiology, Psychology and Models*, D. Pressnitzer*, et al.*, Eds., ed New York: Springer-Verlag, 2005, pp. 176-184.

[6] Carré, R. and Mrayati, M., "Vowel-vowel trajectories and region modeling," *Journal of Phonetics,* vol. 19, pp. 433-443, 1991.

[7] Kent, R. D. and Moll, K. L., "Vocal-tract characteristics of the stop cognates," *Journal of the Acoustical Society of America,* vol. 46, pp. 1549-1555, 1969.

[8] Gay, T., "Effect of speaking rate on vowel formant movements," *Journal of the Acoustical Society of America,* vol. 63, pp. 223-230, 1978.

[9] Strange, W., Jenkins, J. J., and Johnson, T. L., "Dynamic specification of coarticulated vowel," *Journal of the Acoustical Society of America,* vol. 74, pp. 695-705, 1983.

[10] Carré, R., "Signal dynamics in the production and perception of vowels," in *Approaches to Phonological Complexity*, F. Pellegrino*, et al.*, Eds., ed Berlin: Mouton de Gruyter, 2009, pp. 59-81.

[11] Kewley-Port, D., Pisoni, D. B., and Studdert-Kennedy, M., "Perception of static and dynamic acoustic cues to place of articulation in initial stop consonants," *Journal of the Acoustical Society of America,* vol. 73, pp. 1779-1793, 1983.

[12] Tran, T. A. X., Carré, R., Castelli, E., and Vallée, N., "Perception of the Vietnamese short vowel /ɤ̆, ɔ̆, ă/ in /bVk/ produced by female voice," in *International conference on Asian language processing*, Hanoi, Vietnam, 2012, pp. 197-200.

[13] Remez, R. E., Rubin, P. E., Pisoni, D. B., and Carrell, T. D., "Speech perception without traditional speech cues," *Science,* vol. 212, pp. 947-950, 1981.

[14] Liberman, A. M., Delattre, P. C., Cooper, F. S., and Gerstman, L. J., "The role of consonant vowel transitions in the perception of the stop and nasal consonants," *Psychological Monographs,* vol. 68, pp. 1-13, 1954.

[15] Nguyen, V. S., Castelli, E., and Carré, R., "Production and perception of Vietnamese final stop consonants /p, t, k/," in *The second International workshop on spoken languages technologies for under-resourced languages SLTU'10*, Penang, Malaysia, 2010, pp. 136-141.

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

[17] Chistovich, L. A*., et al.*, "Temporal processing of peripheral auditory patterns of speech," in *The representation of speech in the peripheral auditory system*, R. Carlson and B. Grandström, Eds., ed Amsterdam: Elsevier Biomedical Press, 1982, pp. 165-180.

[18] Vaissière, J., "Area functions and articulatory modeling as a tool for investigating the articulatory, acoustic and perceptual properties of the contrast between the sounds in a language," Oxford University Press ed Oxford, 2007, pp. 54-71.

[19] Vaissière, J., "On the acoustic and perceptual characterization of reference vowels in a cross-language perspective," in *Proceedings of ICPhS XVII*, Hong Kong, 2011, pp. 52-59.

[20] Carré, R., "From acoustic tube to speech production," *Speech Communication,* vol. 42, pp. 227-240, 2004.

[21] Carré, R., "Dynamic properties of an acoustic tube: Prediction of vowel systems," *Speech Communication,* vol. 51, pp. 26-41, 2009.

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