Characterising formant trajectories by tracking vocal tract resonances

Gérard Bailly *
Institut de la Communication Parlée, UA CNRS 368
INPG/ENSERG - Université Stendhal, 46 av. Félix Viallet, 38031 Grenoble FRANCE
e-mail: `bailly@icp.grenet.fr`

This article presents a characterisation of formant trajectories based on the tracking of each resonance of the vocal tract. Thanks to an original method called nomograms with decoupled cavities, the optimal constriction locations of the area functions of ten prototypical French vowels are given, together with the main affiliations of each formant: the formants affiliated with the back cavity are noted as R_1, R_3 ...while the formants affiliated with the front one are noted as R_2, R_4 ...The typology of the focal points (convergence of formants due to a change of affiliation between formants) is confirmed by an extensive analysis of natural vowel-vowel transitions. An original vowel space (R_1-R_2) is then proposed which maximises the distances between vowels and favours the effective tracking of the vocal tract resonances by assuming an active filtering of R_3. A simple normalisation of the formant space highlights the promising performance of a speaker-independent vowel identification system based on such an adaptive filtering. Articulatory and perceptual experiments provide evidence converging towards an effective control of an R_1-R_3 relation independent of R_2.

INTRODUCTION

In phonetic science, formants are widely used for the acoustic characterisation of sounds. However the effective tracking of formants by the human ear comes up against many obstacles. If the presence of spectral peaks in auditory representations has been attested by neurophysiology (see [11]) and psychoacoustics, the effective detection of spectral peaks as numbered formants attracts much criticism (see [6]): (a) the auditory threshold of the formants of vocalic spectra does not depend on their numbering, but only on their frequency [8]. (b) classical automatic formant trackers based on peak detection have a poor performance [19,20].

The "centre of gravity effect" [10] suggests that the formation of the auditory representation with a larger critical spacing of 3.0-3.5 Barks might not require an intermediate formant description [13,27]. However, most of the computational models for the effective formant F_2' are based on an intermediate formant detection (see [21]) and characterisation, the control of acoustic stimuli in speech perception and inverse acoustics [1,4] is based on a formant representation of speech. Besides, new families of formant trackers [16,23] have shown that formants can be estimated accurately by imposing top-down and continuity

*This work is supported by ESPRIT BR 6975 SpeechMaps.

constraints.

This paper shows that formant trajectories can be structured in the light of the acoustic theory of speech production, and also that they can be a reliable basis for the analysis of acoustic kinematics.

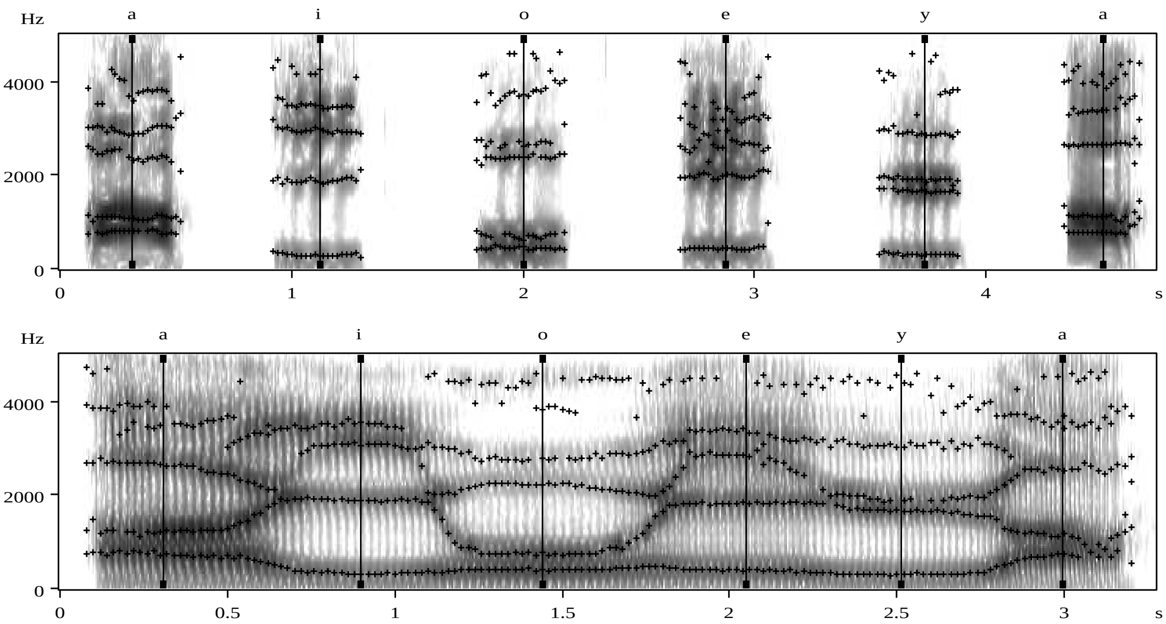

Figure 1. Sonagram of six French vowels [a], [i], [o], [e], [y] and [a] (a) pronounced in isolation and (b) as a continuum. LPC maxima are added.

1. THEORETICAL BACKGROUNDS: AFFILIATIONS

1.1. Formants and resonances

When vowels are pronounced in isolation, the sonogram evidences a number of typically four or five maxima from 0 to 5 kHz (see Figure 1.a). These maxima are traditionally labelled in ascending order. The traditional methods for formant tracking come up against the problem of assigning smooth formant trajectories to continuously varying spectral shapes. One aspect of the problem is shown in Figure 1.b, which displays the same vowels but pronounced as a continuum: formants merge, split and sometimes vanish. We will present an interpretation of this phenomenon below.

The simplified model of speech production proposed by Fant [12] is controlled by four parameters: the location (X_c) and size (A_c) of the main constriction, the area (A_l) of lip constriction and the length of the vocal tract (L_t). Fig. 2 shows two nomograms (the evolution of the maxima of the transfer function of a tube with constant L_t and A_c) when X_c moves from the glottis to the lips for two different A_l corresponding to rounded (/y/, /u/, /ø/ ...) and unrounded vowels (/a/, /i/ ...). The main constriction delimits two cavities, i.e back and front, which are more or less coupled depending on A_c. The superposition of the set of resonances from these two supposedly uncoupled cavities [2]

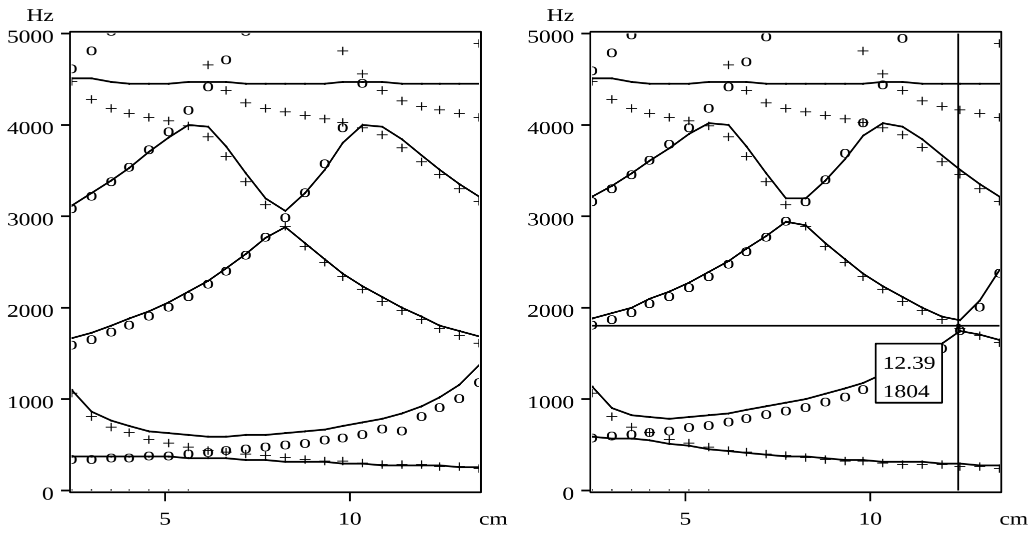

Figure 2. Superposition of nomograms and the resonances of the corresponding front and back cavities respectively indicated with circles and crosses. On the left (closed vocal tract), the nomograms are computed with $A_c = 0.65cm^2$ and $A_l = 0.5cm^2$ and on the right (open vocal tract) with $A_l = 4cm^2$. This superposition evidences some focal points such as the one indicated on the open nomogram.

on the initial nomogram shows (see Fig. 2) that the prediction of the nomogram by the set is quite satisfactory. The prediction error decreases with A_c [12].

These nomograms evidence focal points (see Fig 2), i.e. convergences of formants, which are explained by exchanges of affiliations between formants (numbered in traditional ascending order) and resonances. We note R_{2i} and R_{2i-1} (with $i > 1$), the formants affiliated with the front and back cavities respectively. The affiliation of a formant is determined by computing the minimum distance (city-block on the bark scale) between this formant and the set of resonances of the two decoupled cavities.

1.2. Nomograms with decoupled cavities

It is quite difficult to define a geometrical X_c [7] for more realistic Vocal Tract Area Functions (VTAFs), especially when the geometric constriction is large. We have developed a method to estimate an acoustic X_c : a Nomogram with Decoupled Cavities (NDC) is computed by decoupling the VTAF under consideration into a back and a front cavity. An optimal cut is determined by moving this cut from the pharyngeal to the dental area. It thus determines an *acoustic* Xc where the reconstruction error of the formants of the entire VTAF by the set of formants of each individual cavity is minimal. Moreover, the affiliation of each formant at that point is given. When applied to French phonotype VTAFs produced by a more realistic articulatory model [18], these original nomograms show that whereas the R_2 of open front vowels is close to F_4 for [i] (as shown in Fig. 4) and F_3 for [e], it is close to F_2 for the other vocalic configurations (see also [24]).

The results of the simulations with the prototypical articulatory configurations are

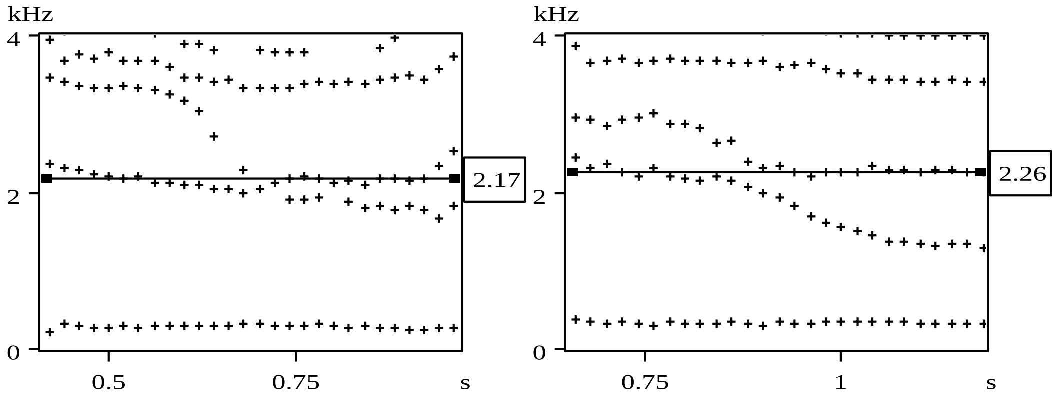

Figure 3. The spectral peaks of two vocalic transitions pronounced by a male speaker. These formant transitions evidence the acoustic correlate R_2 of the lip closing-protrusion gesture. An example of the [i] vs [y] contrast is shown on the left and the [e] vs [ø] contrast on the right. Please note the stable resonance at 2170 Hz for [iy] and the slightly rising resonance at 2260 Hz for [eø]. For [e] and [ø], F_3 is equidistant from F_2 and F_4. This phenomenon gives an unstable F_2' (the distance between F_2 and F_3 is close to the critical distance) and a poorer discrimination between [e] and [ø] using F_2' than using a simple city-block distance on formants.

shown in Fig. 5. This diagram shows that most of the prototypes for back vowels are focal in R_1-R_2. Since A_c is quite large for these vowels, the coupling strongly affects the determination of the precise location of the constriction both geometrically and acoustically (see for example the nomogram of a closed vocal tract in Fig. 2). In what follows we will consider the affiliation of the F_2 of high front vowels with the back cavity (see the exact superposition of F_2 and R_3 for these vowels in Fig. 5), thus keeping $R_i = F_i$ for all the other cases.

2. ANALYSIS OF NATURAL SPEECH

An analysis of natural vocalic transitions between the ten French vowels was carried out to verify the above predictions obtained by simulation. Twenty speakers produced VV syllables in a carrier phrase format *"c'est VV ça?"*. The maxima of the LPC spectra were superposed on the wide-band spectrogram display. Using the temporal decomposition technique [5], time markers were placed on the centre of gravity of each vowel. Resonance targets were determined, using mainly continuity constraints on affiliations (i.e. smoothness to get across focal points) and target coherence (one target per vowel). An example of such a transition is illustrated in Fig. 3.

The figure 7 shows the projection of our vocalic targets in the R_1-R_2 plane. It demonstrates that the contributions of the front and back cavities to the transfer function of the entire vocal tract can be separated. The R_1-R_2 plane has a double advantage: (a) it provides a good separation of vocalic targets, whereas F_3 dimension is required to separate high-front vowels in the classical F_1-F_2 plane (see how the problem of the separation of

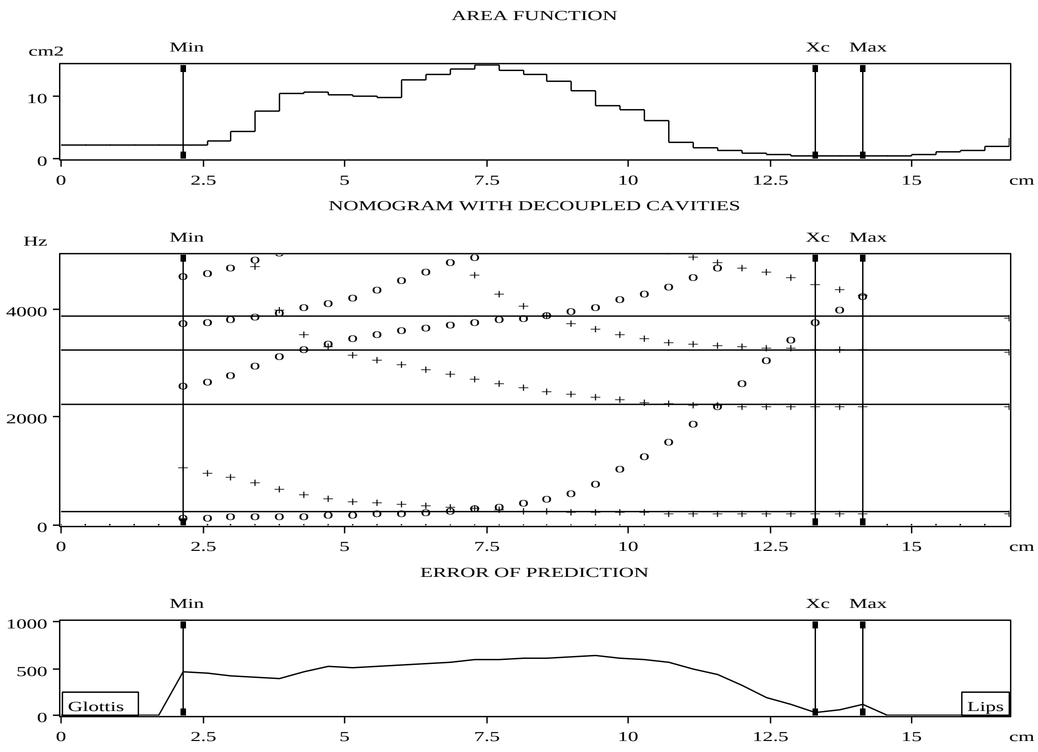

Figure 4. Nomogram with Decoupled Cavities for an [i] vowel. The captions are from top to bottom: the area function, the NDC and the evolution of the prediction error. In the NDC, the resonances of the "front cavities" and "back cavities" are drawn with circles and crosses respectively. The NDC is computed between Min and Max. X_c is evidenced by a vertical bar. Horizontal bars give the formant values which are, of course, the asymptotes for the resonances of the back cavity when X_c increases. Values of the R_i at Xc give the presumed affiliation of the formants: F_1, F_2, F_3 are affiliated with the back cavity and F_4 with the front one, i.e. for [i], $R_2 = F_4 \simeq 4000$ Hz.

the pairs [i] vs [y] and [e] vs [ø] is solved by eliminating R_3: in both cases, the distance between their R_2 is near 3 barks); (b) a close analysis of the kinematics of the trajectories was confirmed by mimicking experiments (see section 4): vowel-to-vowel transitions are characterised by linear resonance trajectories (see figure 6). Such a constraint on acoustic trajectories could explain how listeners can cope with vocalic undershoot: during the whole of its course, the trajectory is in the direction of the target aimed at and goes through any other phonetic targets (for example in Figure 6, the trajectory between [a] and [i] goes through [ɛ] and [e]).

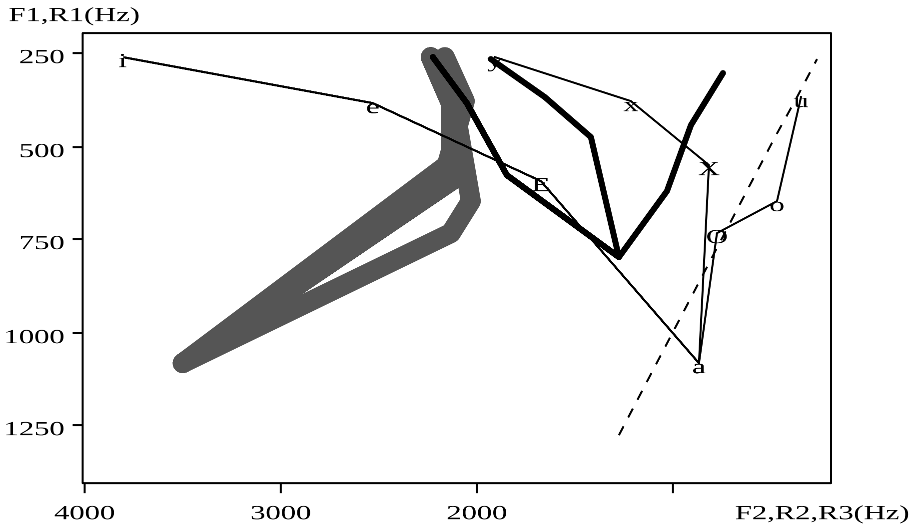

Figure 5. The vocalic triangle reconsidered. The standard F_1-F_2 plane is shown by thick dark lines. The new locations of the vowels in the R_1-R_2 plane are cued by symbols from a Computer Phonetic Alphabet connected by thin lines. R_3 as a function of R_1 is also shown by very thick lines. This diagram shows the extensive movement of R_2 towards higher frequencies for high front vowels and the R_1-R_2 focalisation of back vowels [25] (the dashed line is for $R_1 = R_2$).

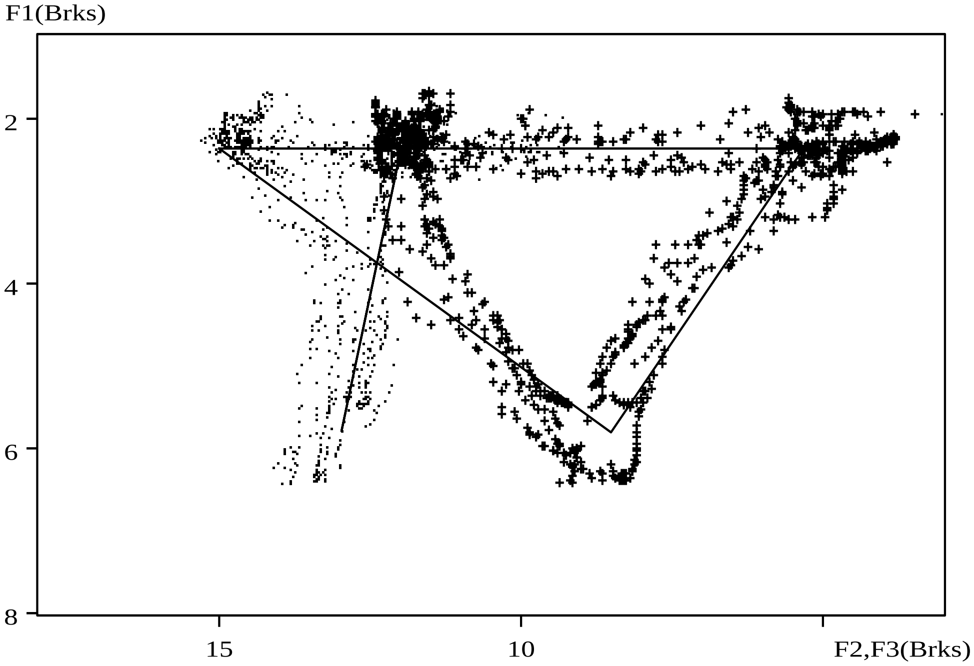

Figure 6. Using a Bark scale, superposition of the formant trajectories between the vowels [a], [i] and [u] on the F_1-F_2 plane.

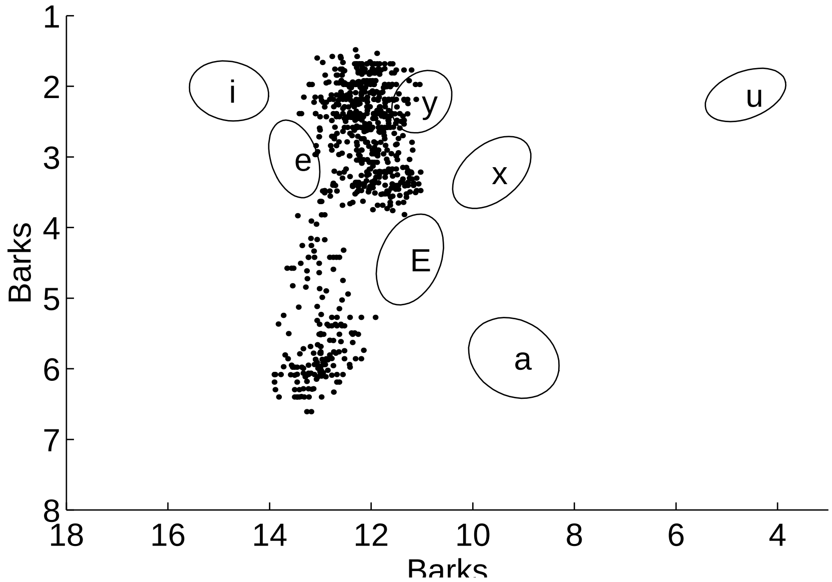

Figure 7. Using a Bark scale, projection of 326 vowels from our 20 speakers in the R_1-R_2 plane after using a simple normalisation procedure using the speakers' register F_0: if $R_i < 1000, R_i = R_i * (1 - F_0/1000)$ else $R_i = R_i - F_0$. Only dispersion ellipses are shown for R_2 target values for clarity. R_3 values are plotted with dots.

This analysis thus confirms our theoretical prediction : transitions between unrounded front vowels and others exhibit formant convergences which are always predicted by an exchange of affiliation (evidenced by formant kinematics and bandwidth exchanges across focal points). Another major result of this analysis is the apparent correlation between R_1 and R_3 : function $R_3 = \mathrm{f}(R_1)$ is monotonous and ascending, i.e. as $R_1([u]) \simeq R_1([y]) \simeq R_1([i])$, $R_3([u]) \simeq R_3([y]) \simeq R_3([i])$ (see Tab. 1). This implication is not justified by articulatory constraints, but can ease the tracking of the two independent front and back resonance "families": the tracking system faces two sets of resonances inside of which monotonous relations are maintained: it thus gains in robustness.

	males				*females*			
	R_1	R_2	R_3	$\overline{\tau}$	R_1	R_2	R_3	$\overline{\tau}$
a	744	1345	2553	80	911	1594	2879	72
ɛ	526	1927	2570	42	673	2250	3087	67
ø	329	1360	2247	62	450	1653	2623	44
e	350	2711	2105	64	434	3133	2429	66
y	256	1810	2138	55	255	2179	2596	70
i	242	3484	2172	61	250	3796	2589	66
u	257	647	2218	63	259	708	2551	47

Table 1
Means of target R_i of seven French vowels across 15 male and 5 female speakers. The average standard deviations of R_3 is presented for each vowel.

3. LOCUS EQUATIONS

The acoustic characterisation of the place of articulation for consonants has encouraged the concept of "locus" [9], defined as "... a place on the frequency scale at which a transition begins or to which it may be assumed to 'point' (p. 769)."

All the investigations (see [26]) into the relation between loci and the subsequent vocalic targets face the problem of the nonlinear behaviour of [g] : the F_2 loci of velar [g] are linearly dependent on vocalic target F_2, whereas the F_2 loci of palatal [g] are almost constant. Klatt [14] states (page 754) that the relation is the inverse for F_3. Such behaviour is easily understandable if we think in terms of resonances (see Fig. 9) : while the tongue position is highly constrained by the place of articulation, the speaker can anticipate the A_l of the subsequent vowel. The acceptable place of articulation for [g] will produce an R_3 (half wave-length of the back cavity) comprised between 2000 Hz for [u] and [a], and 2150 Hz for [i], whereas the R_2 of the front cavity (from a Helmholtz frequency for [u] to a quarter wave-length of the front cavity for [i] and [a]) spans 600-3500 Hz. The locus equations for F_2 and F_3 in Klatt's paper are superposed in Fig. 8. It soon emerges that the acoustic correlate for the place of articulation for [g] is F_2 for open front vowels and F_3 for the others while the acoustic correlate for lip opening, thus highly dependent on vocalic context, is F_3 for open front vowels and F_2 for the others. The loci equations expressed in R_2 will thus exhibit a more linear behaviour. This shows that a reinterpretation of formant loci in terms of resonances can simplify the memorisation of

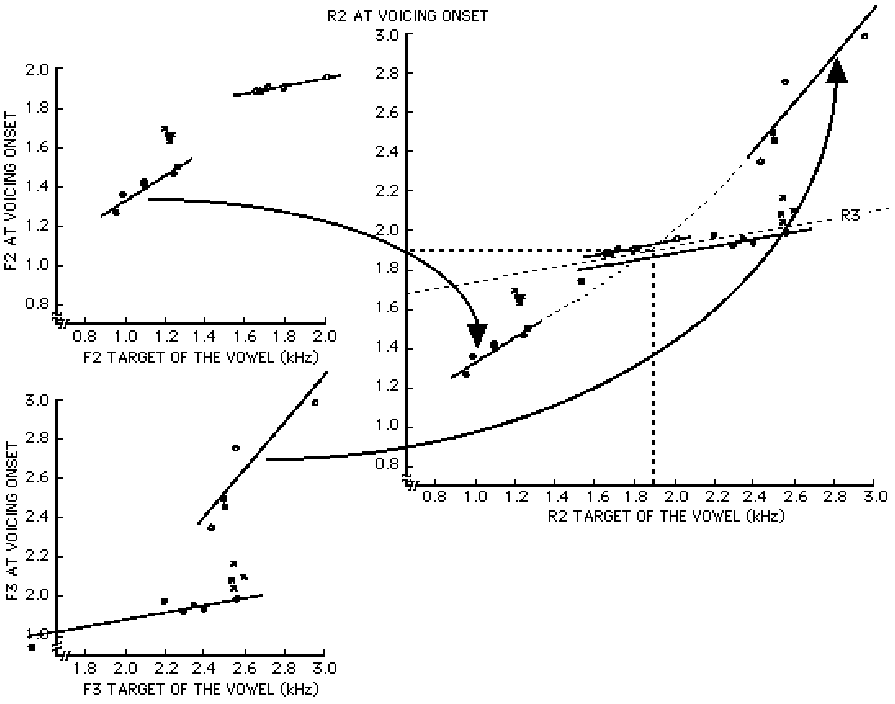

Figure 8. Superposition of the coarticulation functions (CFs) for F_2 and F_3 given by Klatt [14]. It shows the perfect superposition of the CF(F_2) for front vowels and CF(F_3) for back vowels.

coarticulation models (see for example the model proposed by Sussman and coll. in [26]) and extends the use of continuity constraints to infer virtual consonantal targets [9].

4. PERCEPTION EXPERIMENTS

We replicated Repp and William's experiment [22] to show the importance of maintaining the R_1 vs R_3 relation. In this earlier experiment, subjects were asked to imitate twelve sustained synthetic vocalic stimuli (500 ms) produced by sampling at regular intervals a linear interpolation between the formants of two target vowels. They originally moved between high front vowel [i] and low back vowels [u] and [æ]. As summarised by the authors: *"The formant frequency data offer ample evidence for nonlinearities in stimulus-response mapping, but little evidence for response categories in the strict sense."* [22, p.113]. We prepared two other sets of stimuli interpolated between resonances, i.e. where the F_2 of the low back vowels connects with the F_4 of [i]. We asked 12 male French speakers to imitate such stimuli while replacing [æ] by [a]. In cases of interpolation between resonances, "plateaus" appearing in formant interpolation disappear, and speakers have no difficulty in imitating the stimuli (see Fig. 10): the stimuli originally put to the speakers did not obey the typology of normal production constraints. Moreover, the [ui] continuum may be obtained by a release of the constriction: none of our speakers did this. This could be explained by an articulatory control in terms of vocal tract variables (keeping A_c small) which aims at keeping R_1 and R_3 constant [4].

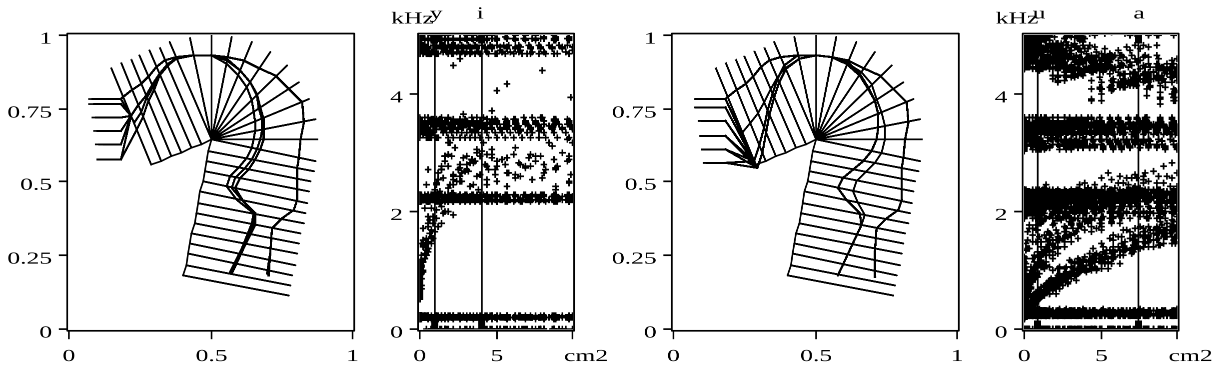

Figure 9. Acoustic and geometrical characteristics of articulatory configurations occluded at 12 cms (left) and 11 cms (right) from the glottis by means of the Maeda model [18]. For each X_c value we show the superposition of midsagittal sections of selected configurations and the scatter-plot of the corresponding formants according to A_l. The usual locations of the vowels are given in the scatter-plots. This evidences the stable location of R_3 around 2100 Hz. The families of R_2 "exponential" equations (especially evident for 11 cms) are due to the position of the apex, which was not constrained in the database search: a low apex corresponds to a lower locus!!

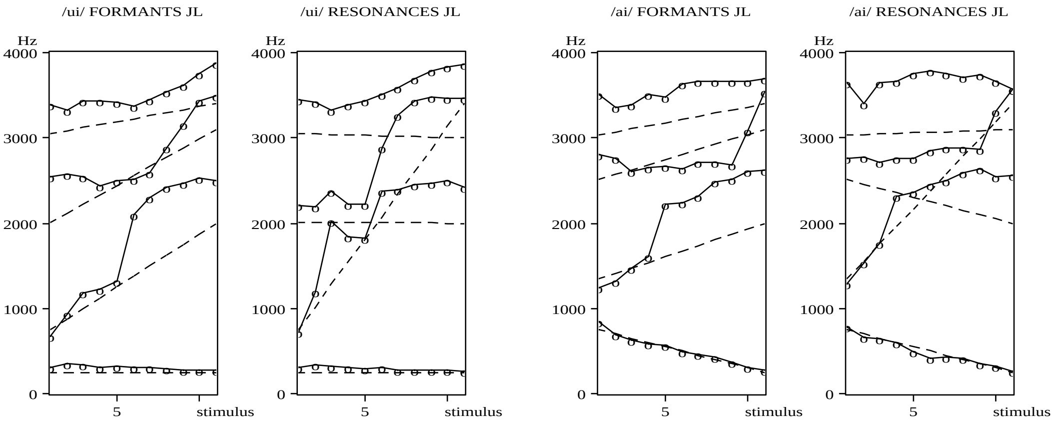

Figure 10. Replication of Repp & William's experiment [22] using formant vs resonance interpolation. Speaker's formants (straight lines) are superposed with the stimuli (dashed lines). The stimuli along [ui] and [ai] are shown on the left and right hand side respectively. Please note the focal points in all captions, and also that the speaker fulfils the task quite well in R_2 despite an overshoot of R_3 by 500 Hz.

5. CONCLUSIONS

This study adds a new quantitative model to the debate on the physical invariants of the speaker/listener interaction. We take the objects of perception to be acoustic events related to properties of the sound producing system. We show that the kinematics of the resonances of the vocal tract may be better understood if we consider their relation to the underlying shapes of the front and back resonators. These resonances may be tracked without an analysis-by-synthesis approach involving an internal forward model [17]. Two simple constraints are used: (a) a limited elasticity constraint as introduced by Kluender *et al* [15] and (b), a R_1-R_3 relation which can only be explained by active articulatory control. We claim that the effective tracking of the resonances of the vocal tract produces (a) an adequate representation space to ensure sufficient distinctiveness (see for example the acoustic contrast for [i] vs [y] and [e] vs [ø]) and (b), an acoustic characterisation of the sounds which linearizes the articulatory-to-acoustic transformation. This latter property could be of prime importance for acoustic-to-articulatory inversion [4] i.e. for constructing predictive models of the direct transform as well as for establishing descriptions of the kinematics of acoustic trajectories [3].

ACKNOWLEDGEMENTS

I would like to thank my colleagues, particularly Christian Abry, Pierre Badin, Eric Castelli and Jean-Luc Schwartz, for their valuable support and comments on this work. This article has gained from a review by René Carré and another reviewer.

REFERENCES

1. Atal, B.S., Chang, J.J., Mathews, M.V., and Tukey, J.W. Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer sorting technique. *Journal of the Acoustical Society of America*, 63:1535–1555, 1978.
2. Badin, P., Boë, L.J., Perrier, P., and Abry, C. Acoustic considerations upon formant convergence. *Journal of the Acoustical Society of America*, 87(3):1290–1300, 1990.
3. Bailly, G. and Guerti, M. Synthesis-by-rule for French. In *12th International Congress of Phonetic Sciences*, pages 506–509, 1991.
4. Bailly, G., Laboissière, R., and Schwartz, J.L. Formant trajectories as audible gestures: An alternative for speech synthesis. *Journal of Phonetics*, 19(1):9–23, 1991.
5. Bailly, G., Marteau, P.F., and Abry, C. A new algorithm for temporal decomposition of speech. application to a numerical model of coarticulation. *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 508–511, 1989.
6. Bladon, A. Arguments against formants in the auditory representation of speech. In Carlson, R. and Granström, B., editors, *The Representation of Speech in the Peripheral Auditory System*, pages 95–102. Elsevier Biomedical, Amsterdam, 1982.
7. Boë, L.J., Perrier, P., and Bailly, G. The geometric vocal tract variables controlled for vowel production: Proposals for constraining acoustic-to-articulatory inversion. *Journal of Phonetics*, 20(1):27–38, 1992.
8. Chistovich, L.A. Auditory processing of speech. *Language and Speech*, 23(1):67–73, 1980.

9. Delattre, P.C., Liberman, A.M., and Cooper, F. Acoustic loci and transitional cues for consonants. *Journal of the Acoustical Society of America*, 27:769–773, 1955.
10. Delattre, P.C., Liberman, A.M., Cooper, F., and Gerstman, L.J. An experimental study of the acoustical determinants of vowel colour. *Word*, 8:195–210, 1952.
11. Delgutte, B. and Kiang, N.Y.S. Speech coding in the auditory nerve: I. vowel-like sounds. *Journal of the Acoustical Society of America*, 75(3):865–878, 1984.
12. Fant, G. *Acoustic theory of speech production.* The Hague: Mouton, 1960.
13. Hermansky, H. and Broad, D.J. The effective second formant F'2 and the vocal tract front-cavity. *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 480–483, 1989.
14. Klatt, D.H. Review of text-to-speech conversion for English. *Journal of the Acoustical Society of America*, 82:737–793, 1987.
15. Kluender, K.R. Psychoacoustic complementarity and the dynamics of speech perception and production. *PERILUS XIV - Publication of the Department of Linguistics, University of Stockholm, Sweden*, pages 131–135, 1991.
16. Kopec, G.E. Formant tracking using hidden Markov models and vector quantization. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 34(4):709–729, 1986.
17. Liberman, A.M. and Mattingly, I.G. The motor theory of speech perception revisited. *Cognition*, 21:1–36, 1985.
18. Maeda, S. Improved articulatory model. *Journal of the Acoustical Society of America*, 81(S1):S146, 1988.
19. McCandless, S. An algorithm for automatic formant extraction using linear prediction analysis. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 22:135–141, 1974.
20. Niederjohn, R. and Lahat, M. A zero-crossing consistency method for formant tracking of voiced speech in high noise levels. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 33(2):349–355, 1985.
21. Paliwal, K.K., Ainsworth, W.A., and Lindsay, D. A study of two-formant models for vowel identification. *Speech Communication*, 2:295–303, 1983.
22. Repp, B. and Williams, D.R. Categorical trends in vowel imitation: preliminary observations from a replication experiment. *Speech Communication*, 4:105–120, 1985.
23. Rigoll, G. Formant tracking with quasilinearization. *IEEE International Conference on Acoustics, Speech and Signal Processing*, 1:307–310, 1988.
24. Schwartz, J.L., Beautemps, D., Abry, C., and Escudier, P. Inter-individual and cross-linguistic strategies for the production of the [i] vs [y] contrast. *Journal of Phonetics*, 21:411–425, 1993.
25. Stevens, K.N. The quantal nature of speech: Evidence from articulatory-acoustic data. In David Jr., E.E. and Denes, P.B., editors, *Human Communication: A unified view*, pages 51–66. McGraw-Hill, New York, 1972.
26. Sussman, H.M., McCaffrey, H.A., and Matthews, S.A. An investigation of locus equations as a source of relational invariance for stop place categorization. *Journal of the Acoustical Society of America*, 90(3):1309–1325, 1991.
27. Tuffelli, D. and Yé, H. Can we predict F'2 by means of a similarity measure? *11th International Congress of Phonetic Sciences*, 4:74–77, 1987.