RESONANCES AS POSSIBLE REPRESENTATION OF SPEECH IN THE AUDITORY-TO-ARTICULATORY TRANSFORM

Gérard Bailly

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

ABSTRACT

This article presents a characterisation of formant trajectories based on an active tracking of each resonance of the vocal tract. This tracking is necessary due to the presence of focal points i.e. changes of affiliations between F2 and F3 in most of transitions between back and front sounds. We show that this tracking can be easely implemented without any analysis-by-synthesis process by assuming a minimum elasticity principle (smoothness) and a monotonous relation between resonances associated with the same cavity. Our proposal is supported by intensive analysis of natural speech, articulatory simulations and perceptual experiments.

Keywords: Perception, Formants, Vocal Tract, Resonances.

1 INTRODUCTION

Formants are widely used for the acoustic characterization of sounds in phonetic science. Their effective tracking by the human ear faces however many obstacles. If the presence of the spectral pics in the auditory representations have been attested by the neurophysiology (cf. [11]) and psychoacoustics, the effective detection of spectral pics as numbered formants faces many critisms (see [6]):

a. Auditory threshold of formants of vocalic spectra do not depend of their numbering but only from their frequency [8].

b. Classical automatic formant trackers based on peak detection have poor performance [19, 20].

The "center of gravity effect" [10] have suggested that the formation of the auditory representation with a larger critical spacing of 3.0-3.5 Barks might not require the intermediate formant description [13, 26]. However most of the computational models for F'2 are based on formant detection (see [21]) and characterization and control of acoustic stimuli in speech perception as well as inverse acoustics [1, 5] is based on formant representation of speech. Besides new families of formant trackers [16, 23] have shown that formants can be estimated accurately by imposing top-down constraints and continuity constraints.

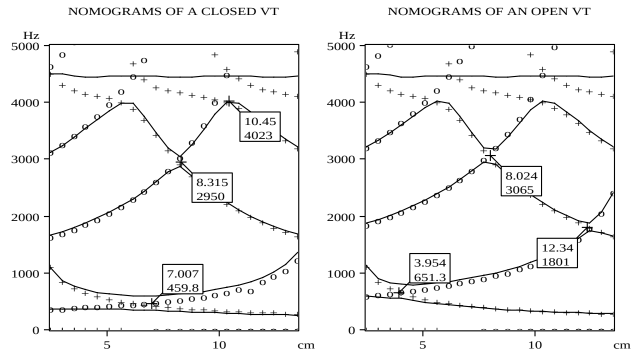

Figure 1: Superposition of nomograms and the resonances of the corresponding front and back cavities respectively figured with circles and crosses. The nomograms are computed with $Ac = 0.64cm^2$ and $Al = 0.16cm^2$ at the left (closed vocal tract) and $Al = 4cm^2$ at the right (open vocal tract). This superposition evidences some focal points such as the one indicated on the open nomogram.

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

2 THEORETICAL BACKGROUNDS: AFFILIATIONS

2.1 Formants and resonances

The simplified model of speech production proposed by Fant [12] is controlled by four parameters: the location (Xc) and size (Ac) of the main constriction, the area (Al) of the lip constriction and the length of the vocal tract (Lt). Fig. 1 shows two nomograms, i.e. the evolution of the maxima of the transfer function of a tube with constant Lt and Ac, when Xc moves from the glottis to the lips for two different Al corresponding to rounded vowels (/y/, /u/, /ø/ ...) and unrounded ones (/a/, /i/ ...). The main constriction delimits two cavities, i.e back and front cavities, which are more or less coupled depending on Ac. The superposition of the set of the resonances of these two cavities supposed uncoupled [2], with the initial nomogram shows (cf. Fig. 1) that the prediction of the nomogram by the set is quite satisfactory. The prediction error decreases with Ac.

These nomograms evidence focal points (cf. Fig 1), i.e. convergences of formants, which are explained by exchanges of affiliations between formants (numbered in classical ascending order) and resonances. We note R_{2i} and R_{2i-1} (with $i > 1$) the formants affiliated respectively to the front and back cavity.

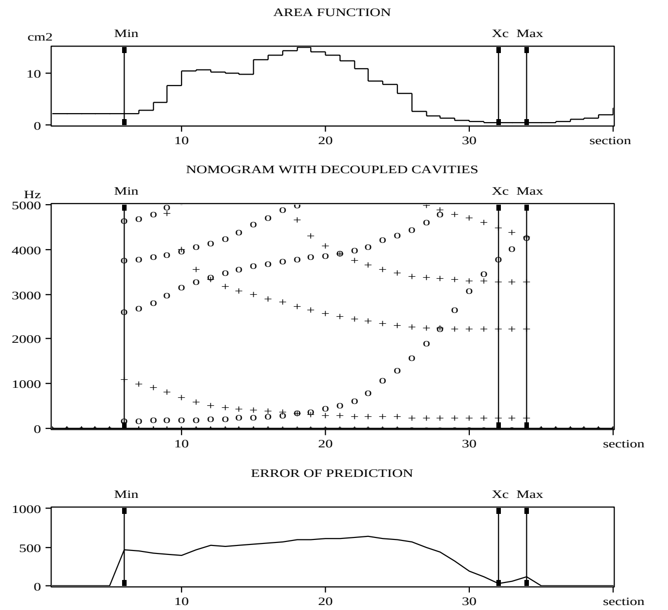

Figure 2: NDC for a /i/ vowel. The captions are from top to bottom: the area function, the NDC and the evolution of the prediction error. In the NDC, resonances of the "front cavities" and "back cavities" are figured respectively with crosses and circles. Xc is evidenced by a vertical bar. Values of the R_i at Xc give the presumed affiliation of the formants: F1, F2, F3 are affilated with the back cavity and F4 with the front one, i.e. for /i/, R_2=F4 $\simeq$ 4000 Hz.

2.2 Nomograms with decoupled cavities

It is quite difficult to define a geometrical Xc [7] for more realistic Vocal Tract Area Functions (VTAFs). We developed a method to estimate an acoustic Xc : a Nomogram with Decoupled Cavities (NDC) is computed by decoupling the considered VTAF into a back and front cavity. An optimal cut is considered by moving this cut from the pharyngeal area to the teeth area, when the reconstruction error of the formants of the entire VTAF by the set of the formants of each individual cavity reaches a minimum [3]. When applied to VTAFs of French phonotypes produced by a more realistic articulatory model [18], these original nomograms show that the R_2 of open front vowels are close to F4 for /i/ (as shown in Fig. 2) and F3 for /e/ whereas it is close to F2 for other vocalic configurations (see also [24]).

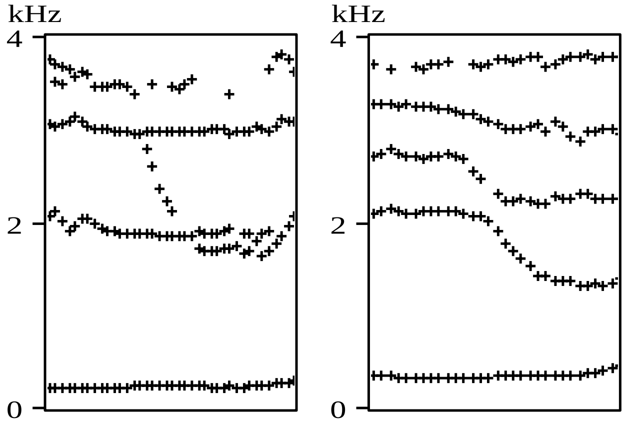

Figure 3: Spectral peaks of two vocalic transitions pronounced by a male speaker showing the acoustic correlate R_2 of the lip closing-protrusion gesture. At the left, the /i/ vs /y/ contrast. At the right, the /e/ vs /ø/ contrast. Please note the stable resonance at 2000 Hz for /iy/ and the slightly rising resonance at 2100 Hz for /eø/. For /e/ and /ø/, F3 is at equal distance between F2 and F4. This phenomenon gives unstable F'2 when computed after "large scale" integration.

3 ANALYSIS OF NATURAL SPEECH

Analysis of natural vocalic transitions between the ten French vowels was then conducted to verify this theoretical prediction. Twenty speakers produce VV syllables in a carrier phrase format *"c'est VV ça?"*. Maxima of the LPC spectra were superposed on the wideband spectrogram display. Time markers were placed on the centers of realization of each vowel. Resonance targets were determined using mainly continuity constraints on affiliations (smoothness to get across focal points) and coherence of the targets (one target per vowel). An example of such a transition is figured in Fig. 3.

	males				*females*			
	R1	R2	R3	$\overline{\tau}$	R1	R2	R3	$\overline{\tau}$
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 the target Ri of seven French vowels across 15 male and 5 female speakers. Means of standard deviation of R3 for each vowel is presented.

This analysis confirms the theoretical prediction : transitions between unrounded front vowels and others exhibit formant convergences which are always predicted by an exchange of affiliation (evidenced by kinematics of formants and exchange of bandwiths across focal points). Another major result of this analysis is the apparent correlation between R_1 and R_3 : the function R_3 = 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 set of resonances inside which monotonous relations are maintained and thus gains robustness.

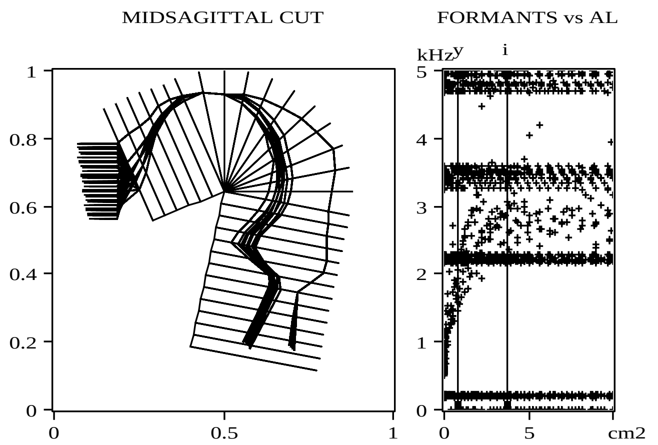

Figure 4: Acoustic and geometrical characteristics of articulatory configurations occluded at 12 cm from the glottis using the Maeda's model [18]. From left to right: superposition of midsagittal sections of selected configurations, scatterplot of the corresponding formants according to Al: it shows the exchange of affiliation at (Al=200 cm2, F2=F3=2100 Hz). Usual locations of the front vowels /y/ and /i/ are given in the scatterplot.

4 LOCUS EQUATIONS

The acoustic characterization of place of articulation for consonants has promoted 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 analysis investigations (see [25]) on the relation between loci and subsequent vocalic target face the problem of the nonlinear behavior of /g/ : the F2 loci of velar /g/ are linearly dependent of F2 vocalic target whereas the F2 loci of palatal /g/ is almost constant. In Klatt [14], we can see (page 754) that the relation is opposite for F3. Such a behavior is easely understandable if we think in terms of resonances (see Fig. 4) : whereas the tongue position is highly constrained by the place of articulation, the speaker can anticipe the Al of the subsequent vowel. The acceptable place of articulation for /g/ will produce a R_3 (halfwave-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 quaterwave-length of the front cavity for /i/ and /a/) spans the interval 600-3500 Hz. By superposing locus equations for F2 and F3 in Klatt's paper, it is easely found that the acoustic correlate for place of articulation for /g/ is F2 for open front vowels and F3 for the others whereas the acoustic correlate for lip opening, thus highly dependent on vocalic context is F3 for open front vowels and F2 for the others. The loci equations expressed in R_2 will thus exhibit more linear behavior.

This shows that a relecture of formant loci in terms of resonances can simplify memorization of coarticulation models and open the use of continuity constraints to infer virtual consonantal targets [9].

5 PERCEPTION EXPERIMENTS

To show the importance of maintaining the R_1 vs R_3 relation, we replicate Repp and William's experiment [22]. The subjects were asked to replicate twelve sustained synthetic vocalic stimuli (500 ms) produced by sampling a linear interpolation between formants of two target vowels at regular intervals. They originaly move between the high front vowel /i/ and the low back vowels /u/ and /æ/. We asked 12 French male speakers to imitate such stimuli while replacing /æ/ by /a/. In case of interpolation between resonances, "plateaus" which appear in formant interpolation disappear and speakers have no difficulty in imitating the stimuli (see Fig. 5).

6 CONCLUSIONS

This study feeds the debate on the physical invariants of the speaker/listener interaction with a new quantitative model. We take the objects of perception to be acoustic events related to properties of the sound producing system. We have shown 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 two front and back resonators. These resonances may be tracted without an analysis-by-synthesis approach involving an internal forward model [17], using two simple constraints : a/ a limited elasticity constraint as introduced by Kluender *et al* [15] and b/ a R_1-R_3 relation which could only explained by an active control of the speaker.

We claim that 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 characterization of the sounds which linearizes the articulatory-to-acoustic transformation. This latter property could be of most importance for acoustic-to-articulatory inversion [5] i.e. for building "forward" models of the articulatory-to-acoustic transformation as well as for building description of the kinematics of the acoustic trajectories [4].

ACKNOWLEDGMENTS: This work was supported by ESPRIT BR 6975 SpeechMaps.

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., Bo, L.J., Perrier, P., and Abry, C. Acoustic considerations upon formant convergence. *Journal of*

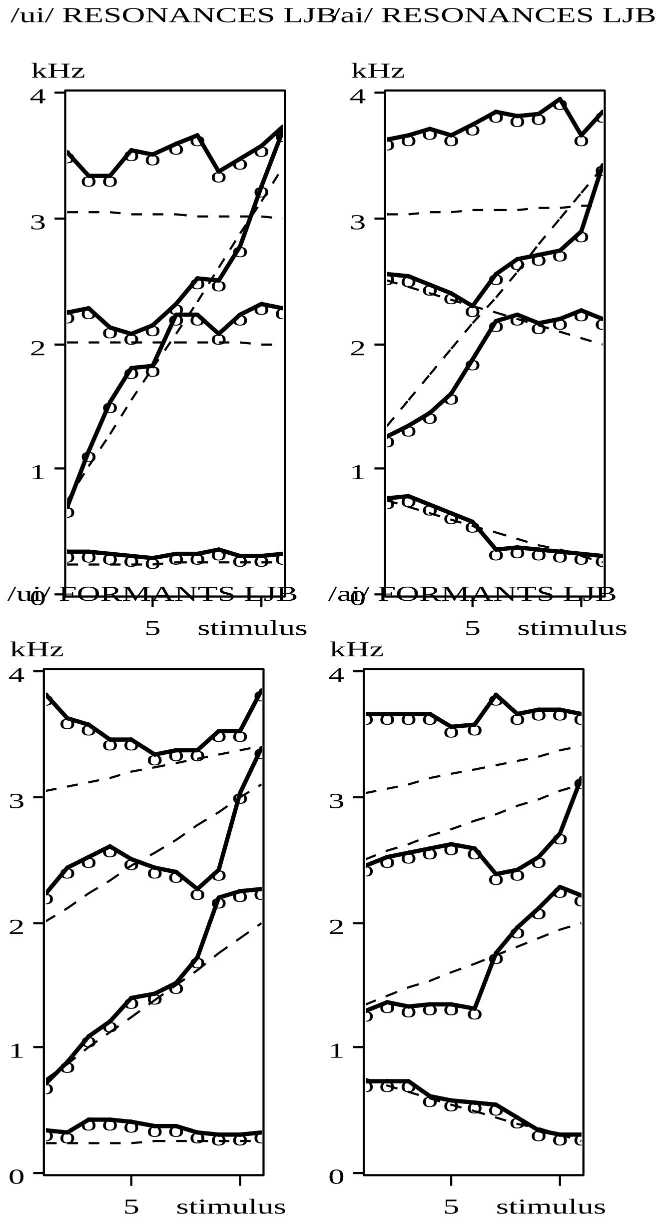

Figure 5: Replication of Repp & William's experiment [22] using formant vs resonance interpolation. The speaker's formants (crosses) is superposed with the stimuli (straight lines). Stimuli have been concatenated for lisibility. Presumed focal point are evidenced: resonance stimuli are in complete accordance with realizations.

the Acoustical Society of America, 87(3):1290–1300, 1990.

/3/ Bailly, G. Caracterisation of formant trajectories by tracking vocal tract resonances. In Sorin, C., Mariani, J., Mloni, H., and Schoentgen, J., editors, *Levels in speech communication : relations and interactions*, pages 91–102. Elsevier, Amsterdam, 1995.

/4/ Bailly, G. and Guerti, M. Synthesis-by-rule for french. In *12th International Congress of Phonetic Sciences*, pages 506–509, Aix-en-Provence, France, August 1991.

/5/ Bailly, G., Laboissire, R., and Schwartz, J.L. Formant trajectories as audible gestures: An alternative for speech synthesis. *Journal of Phonetics*, 19(1):9–23, 1991.

/6/ Bladon, A. Arguments against formants in the auditory representation of speech. In Carlson, R. and Granstrm, B., editors, *The Representation of Speech in the Peripheral Auditory System*, pages 95–102. Elsevier Biomedical, Amsterdam, 1982.

/7/ Bo, 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. The klattalk text-to-speech conversion system. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1589–1592, 1982.

/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. *TrASSP*, 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. *TrASSP*, 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. *TrASSP*, 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/ 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.

/26/ Tuffelli, D. and Y, H. Can we predict f'2 by means of a similarity measure? *11th International Congress of Phonetic Sciences*, 4:74–77, 1987.