Articulatory modeling and the definition of acoustic-perceptual targets for reference vowels

Jacqueline Vaissière

Vaissière, J., (2009), "Articulatory modeling and the definition of acoustic-perceptual targets for reference vowels ", The Chinese Phonetics Journal, Vol. 2:22-33

ARTICULATORY MODELING AND THE DEFINITION OF ACOUSTIC-PERCEPTUAL TARGETS FOR REFERENCE VOWELS

Jacqueline Vaissière

Abstract: The present paper proposes some hints to set up a number of prototypical vowels as acoustic-perceptual targets (APT). APTs are defined using theoretical considerations and the extensive use of an articulatory model for exploring the acoustic and perceptual space. This paper is a revisit of the notion of Cardinal Vowels, which were mainly described in articulatory terms. Our approach avoids the problem of the articulatory system's compensatory abilities to attain a given APT with different articulatory strategies. The availability of well-defined APTs makes it possible to characterize both acoustically and articulatorily subtle but audible deviations between the different renditions of some of the sounds. Moreover, the proposed notation has the potential to highlight the acoustic characteristics common to all phonemes (vowels, semi-vowels, consonants) that share a similar place of constriction, on the one hand, and coarticulatory effects (such as palatalization, labialization and velarization), on the other hand.
Keywords: articulatory modeling, IPA, vowels, acoustics, formant, coarticulation.

1. INTRODUCTION

The human vocal tract (VT) can produce a very large number of different sounds. In his Outline of English Phonetics, Daniel Jones (hereafter DJ) claimed "a good ear can distinguish well over fifty vowels, exclusive of nasalized vowels, vowels pronounced with retroflex modification, etc. etc" [1:29]. Each language uses only a small subset of all possibilities by selecting a limited number of phonemes to distinguish the meaning of words. Subtle variations in the sound quality between the different renditions of the same phoneme may be informative about the speaker's native language, dialect, idiolect or socio-cultural status or emotive state. Listeners from different native languages show different phoneme identification and goodness ratings when exposed to the same stimulus stimuli [2]. The present paper proposes some hints to set up a number of prototypical vowels as acoustic-perceptual targets (APT) that could eventually be used as references. It is limited to the eleven first Cardinal Vowels: [ieɛa uoɔɑ yøœ].

2. IPA AND THE CARDINAL VOWELS

2.1 IPA

Setting up references for comparison of sounds is particularly needed for language teaching. The notation proposed by the International Phonetic Association, the International Phonetic Alphabet (hereafter IPA) was established for pedagogical purposes. The IPA phonetic notation has proved to be very useful for phoneticians, linguists, socio-phoneticians, speech pathologists and therapists, in speech technologies, in foreign second language teaching, in dictionaries and for labeling speech corpora. IPA is well suited to represent the qualities of speech that are distinctive in spoken language (i.e. phonemes); its use is

more problematic for representing subtle differences between different phonetic realizations of the same phoneme. The same phonetic symbol, such as [i], may encompass several different phones recognizable by human ear. The use of diacritics to described the different phones is not standardized. For example, it is not easy to notate in a clear way the difference in timber between palatal and prepalatal [i], or to indicate whether or not the high front rounded vowel found in a language is equivalent to DJ's Cardinal [y].

The IPA vowel chart (Figure 1) is mainly based on articulatory attributes. The vowels are placed in the chart according to three articulatory dimensions: frontness-backness, openness-closedness (or height), and spread/ rounding. For a number of reasons, vowels are more difficult to describe than consonants. There is no precise place of articulation from the production point of view for the vowels which correspond to a relatively open VT. X-ray data witness compensatory phenomena that may take place between the articulators (the lips, the jaw, the tongue and, the larynx) [3, 4] and the same vowel timber can be obtained using quite different articulatory strategies (c.f. bite block experiments). Secondly, on the perception side, labeling of vowels seems more difficult than to labeling than consonants. Furthermore, the same vowel, uttered by the same a given speaker generally corresponds to quite different sounds depending on whether it is uttered in isolation, extracted from continuous speech; presented to a human transcriber with its flanking phonemes or when the listener has access to the whole message. Whenever possible, the listeners perceptually compensate for vowel target undershoot and tend to "hear" the original target. Finally, the transcriber's perception is influenced by his/her native language, his/her familiarity with the IPA symbols, his/her specific experience with second languages and non-native contrasts, etc.

There is thus no guarantee that the same IPA symbol will be used to transcribe the same timber. For example, in the UCLA Phonetics Lab Data Base, there are instances in which strikingly different sounds are transcribed with the same IPA symbol. There is a clear need to make progress towards standardization and to set up criteria for selecting a particular symbol. There is also a need to notate deviations in a principle way. To what extent is this that feasible?

2.2 The Cardinal Vowels

The set of Cardinal Vowels (hereafter CV), defined by Daniel Jones [1] in the early 20th century, are often regarded as "reference" vowels for phonetic description and transcription. Since they are well-known and still in use, they serve as a good starting point.

The 18 CVs have been defined on a mixture of articulatory and auditory criteria. The three basic CVs, C1[i], C5[ɑ] and C8[u], were first defined on articulatory bases. According to DJ, C1[i] is the frontmost vowel and it is pronounced with as high a position as possible, and with spread lips. Beyond that position, a friction noise will be generated. C8[u] is produced with the tongue as far back and as high as possible, with rounded and protruded lips. C5[ɑ] is produced with the tongue as low and as far back as possible. The other five primary CVs were derived as 'auditorily equidistant' between these three 'corner-vowels', at four degrees of aperture or 'height': close-mid C2[e] and C7[o], open-mid C3[ɛ] and C6[ɔ], and open C4[a] (low tongue position). The first eight secondary CVs are obtained by using the opposite lip-rounding on each primary CV (C1[i]-C9[y], C2[e]-C10[ø], C3[ɛ]-C11[œ], etc.) The accuracy of many of DJ's statements on VT configurations when producing the CVs has increasingly come under question. It is often suggested that the labels are primarily acoustic or perceptual [5], and not articulatory. Despite such criticisms, the CVs represent an important move in the direction of much needed standardization.

How are these reference vowels accessible to transcribers? DJ claimed that it is impossible to learn how to produce the CVs without a teacher who knows them. DJ however suggests the French vowels are the closest you can get to the CVs.

Figure 1 represents the vowel system of French. French has a crowded vowel space, comprising 10/11 oral vowels and 3/4 nasal vowels, depending on the dialect or idiolect. The system includes vowels the whole set of primary CVs, C1 to C8, [i e ɛ a u o ɔ ɑ], and the three most attested secondary (rounded) CVs, C9 to C11 [y ø œ]. French uses four dimensions to contrast these vowels: 1) frontness-backness, 2) four degrees of aperture, 3) labialization and 4) nasalization.

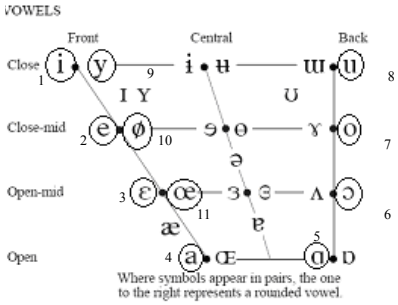

Figure 1: The vowel chart used in the IPA. The French oral vowels are circled (the nasal vowels are not represented). The numbers are related to the CV numbers.

Nowadays, it is possible for all to access the sounds corresponding to the CVs as uttered by DJ (on the Internet) and to perform an acoustic analysis. Figure 2 illustrates the spectrograms corresponding to the first 11 CVs (from C1 to C11) as produced by DJ, Peter Ladefoged (hereafter PL), and typical French vowels, uttered by a male French speaker. DJ's and PL's sets of sounds are available on the web [6]. The similarity and the discrepancy between the renditions of the CVs by experts, like DJ or PL, and the French vowels have been the key motivation for the present study.

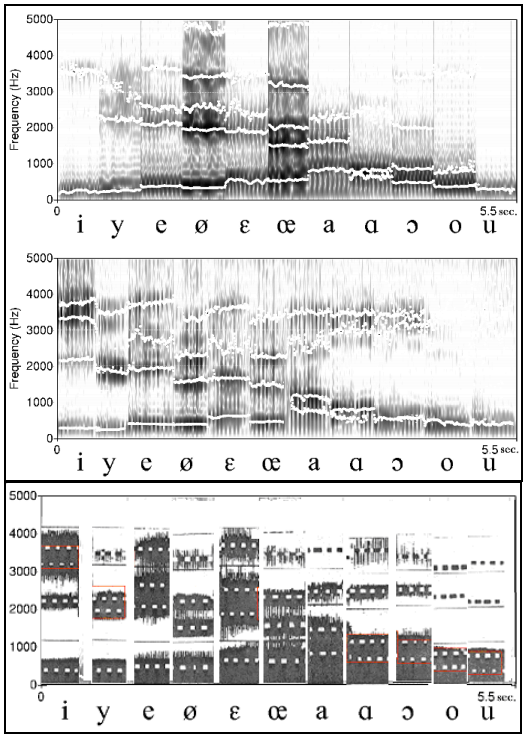

Figure 2: Spectrograms corresponding to the CVs as pronounced by DJ (top), PL (mid) and typical realizations of French vowels (bottom). All male speakers.

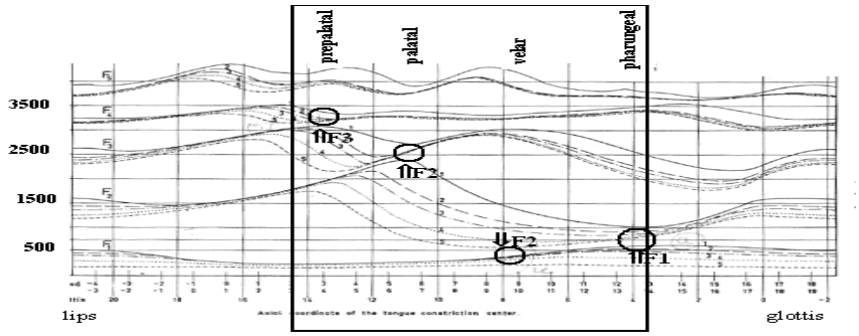

Figure 3: Fant's nomogram using the horn-shaped three-parameter VT model. The tongue constriction size is fixed at 0.65 cm2. Effect of lip configuration: no lip rounding (plain lines), and decreasing lip area: 8, 4, 2, 0.65 and 0.16cm2 (dotted lines). The glottis at the right and the lips at the left. The large square is related to Figure 4. See text. [after 7: 83]

3. BACKGROUNDS

Our description relies on a number of well-established notions outside the field of traditional articulatory phonetics.

3.1 Acoustic theory of speech production

3.1.1 Seven laws of acoustics

The acoustic theory of speech production offers a well-established method to relate a given articulatory configuration to the physical characteristics of the produced sounds. Fant's nomograms [7] illustrate the acoustic consequences of varying isolated articulatory parameters in terms of formant frequencies.

Fant's nomograms (see such a nomogram on Figure 3) concentrate on the effect of three parameters: the place of constriction, the degree of constriction and the effect of lengthening and perturbation of the tube at one of its ends (simulation of labialization)[1].

The following remarks (related to laws of acoustics) can be made.

- **Law 1: Turning point**

As the constriction moves from the back (glottis) to the front (lips), F1, F2, F3 successively reach their maximum value (⇑F1, ⇑F2, ⇑F3 in Figure 3).

- **Law 2: Focal point**

Around these turning points, where a formant Fn is maximally high, this formant Fn tends to converge with the formant Fn+1. Hence, ⇑F1, ⇑F2, ⇑F3 regions correspond to three zones of converging formants: (⇑F1⇓F2), (⇑F2⇓F3) and (⇑F3⇓F4).

- **Law 3: Zone of articulatory stability in relation to the place of constriction**

At the focal points, relatively large changes in position of the constriction will cause little change in the acoustic signal (at least for the two formants concerned).

- **Law 4: Point of highest sensibility to labialization and degree of constriction**

At the zones of articulatory stability in relation to the place of constriction (frontness/backness dimension), the formants are most sensitive to change in other dimensions (such as labialization or degree of constriction).

- **Law 5: Point of perceptual sharpness**

Due to acoustic laws, when two formants converge, the amplitudes of the two formants are mutually enhanced. The spectrum is dominated by a concentration of spectral energy in the region where the two formants converge, creating a sharp spectral salience in a well-defined frequency range. The regions where the spectrum is dominated by a spectral peak are circled in Figure 3.

- **Law 6: Effect of 'labialization'**

Simulation of labialization (rounding and protrusion) allows the creation of two more points of convergences by lowering the formant(s) that are mainly due to the front cavity. When the constriction is in the middle, labialization leads to the lowest F2 value, creating a point of convergence (⇓F1⇓F2). Note the large range of constriction positions where F1 and F2 converge (focal but not turning points): the mid and back region will be exploited to create the back (focal) CVs. At point ⇑F3, where the constriction is very fronted (say, prepalatal), labialization is maximally

[1] Note that these three parameters are far from sufficient to describe the contrasts used by languages: any part of the VT -from lips to larynx- may contribute to achieve a particular APT.

efficient in lowering F3, which becomes close to F2, creating a (↓F2⇓F3) focal point. Note also that the magnitude of the effect of labialization is larger when the constriction is fronted than when the constriction is back: the front region will be used to create the contrast between rounded and unrounded front CVs. A particular point of interest is the prepalatal region, where the unrounded and rounded renditions of the prepalatal constriction give rise to two separate focal points.

- **Law 7: Point of strong constriction**

If the formants affiliated to two different cavities are found at about the same frequency, there should be a small amount of coupling between the two cavities. For two tubes to be decoupled, the areas of the tubes have be very different: one very small and the other very large. It is possible to establish the affiliation between formant and cavity only in the cases where there is a strong constriction in the vocal tract. The formants tend to separate with increasing large constriction, and all formant values decrease, except F1, which increases [7: 84].

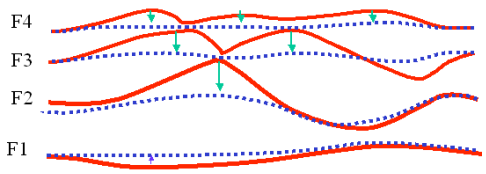

Figure 4: Effect on the formants of increasing the degree of constriction (derived from [7: 84]). Dotted lines correspond to a less constricted VT.

These remarkable regions are called points of converging formants, focal, quantal, plateau-like points or regions, "hot spots", (and also zones of"articulatory stability"), etc. Note also that vowel formants in close proximity are merged during perceptual analysis, and these are also zones of perceptual formant integration.

3.2 The QT and ADP theories

These focal regions have inspired a number of theories, mainly the Quantal Theory and the Focalization Theory, as described below.

2.1.1 The Quantal Theory (QT)

Law 3 inspired Ken Stevens to elaborate the Quantal Theory. Stevens [8] proposes the plateau-like regions of the articulation-acoustic relations (Law 3) as the correlates of the distinctive features because they correspond to zones of articulatory stability. According to the Quantal Theory, the languages of the world show a preference for these regions of articulatory stability for phonemes, independently of their inventory size (i.e. for their intrinsic qualities). For example, when there is a constriction in the rear front of the VT (the [i] region) or in the rear back (the [ɑ] region), F2 is least sensitive to local perturbations, while the central region is extremely sensitive to local perturbations (unless the lips are rounded/protruded).

3.2.2 Lindblom's Adaptive Dispersion Theory (ADT)

According to the Quantal Theory, the inventory size does not play a leading role in the choice of phonemes. However, the number of zones of articulatory stability is not sufficient to create a large number of vowels. According to ADT, the distinctive sounds of a given language tend to be positioned in phonetic space so as to maximize perceptual contrast [9] or to satisfy a functional requirement of sufficient perceptual distinctiveness [10]. The vowel quality therefore depends on the whole system of contrasts. According to this theory, the acoustic realization of the French vowels would depend on the total number of (oral) vowels in the language, in this case [11].

3.2.3 The Dispersion-focalization theory (DFT)

DFT states that sound systems are shaped by ADT and Focalization driving auditory spectra towards patterns with close neighboring formants. Focalization is based on Law 5. Close neighboring formants produce vowel spectra with marked peaks,

and because of their acoustic qualities, these spectra are easier to process and memorize in the auditory system. Such salience is likely to provide the ground for anchor vowels, by increasing their perceptual salience (see the DFT [11, 12]).

3.3 Articulatory modeling

Table 1: parenthesis notates focalization. ⇑ and ⇓ mean maximally high and low, respectively. ↑ and ↓ correspond to higher and lower, respectively.

Place of constriction	Focal points= points of strong constriction	
	Spread lips	Rounded lips
Back	(⇑F1↓F2) $^{F2=1000Hz}$ *(C5[a])*	
Mid		(↓F1⇓F2)$^{F2=700Hz}$ *(C8[u])*
Palatal	(⇑F2↓F3) $^{F3=2500Hz}$	
Prepalatal	(⇑F3↓F4)F3$^{F2=3200Hz}$ *(C1[i])*	(↓F2⇓F3)$^{F2=1900Hz}$ *(C9[y])*

Table 1 summarizes some of the preceding remarks (for more details, see [15]).

Fant's nomograms are sufficient to comprehend the basic acoustic principles underlying the production of contrasts in the human VT. However, there is no guaranty that the formant patterns produced by such modeling mirror the capabilities of a human VT. It is also necessary to be able to perform an auditory analysis of the sounds produce by the models and to compare them with the real sounds to be transcribed.

In fact, human speakers can produce vowels which cover only less than half of the range represented in Fant's nomograms [14]. Bladon [13] was not able to entirely reproduce Fant's nomogram; it was impossible for a constriction location at less than 5 cm from the glottis, and at more than 13 cm from the glottis.

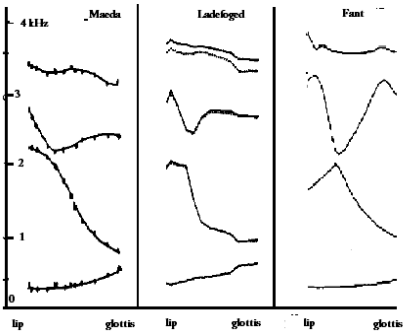

Figure 5: Nomogram produced by Fant's four-tube model (right), by Maeda's model (left) when the constriction location is varied from 5 to 13 cm from the glottis. Mid: attempt to imitate Fant's nomograms by Bladon (adapted from [14]).

Figure 5 illustrates three nomograms: one produced by Fant's second model (right) [7], the attempt by Bladon [13] to imitate Fant's nomogram (middle) and the one by Maeda's articulatory model (left) [14]. The output of the articulatory model is closer to Bladon's production than Fant's nomogram and gives credit to Maeda's widely used model as a tool in speech research [15].

Maeda's articulatory model [3, 4] is used here to (re)explore the space of articulatorily possible vowels with a more realistic model. Seven parameters are controllable: three parameters for the tongue, two for the lips (rounding and protrusion), one for larynx height, and one for the jaw. The acoustic-perceptual consequences of any minor change in one of the 7 parameters can be calculated and heard.

The effect of retroflexion and of change in the tongue shape (such as bunching) can be modeled by a direct modification of the area function. Such parameters are not controlled by the model, since they are not used in French: Maeda's anthropomorphic model is based on statistical analysis of X-ray data of native speakers of French.

Table 2 summarizes the observations made on the position of the constriction, the lip configuration and the shape of the tongue to obtain maximum or minimum F1, F2 and F3 values. Roughly speaking, there is a single large movement to manipulate F1, two for

F2, and three for F3. Retroflexion, secondary constriction, larynx height, or the shape of the tongue may be used to reinforce the movement of a formant in one language (as indicated in Table 1), or to create new contrasts in another.

Table 2: Tongue constriction position, lip configuration, tongue shape and most extreme F-pattern: a summary.

Location of the constriction			
Lowest possible formant		**Highest possible formant**	
☐F1	Anterior part of the VT	☐F1	Posterior part of the VT
☐F2	Middle (velar) region + Lip rounding	☐F2	Mid-palatal region + Glottal region
☐F3	Back (pharyngeal) region + Bunching of the tongue, retroflexion + Lip rounding and lip protrusion	☐F3	Front (apical and prepalatal) regions + Lip spreading (larynx lowering)

4. DJs, PLs AND FRENCH VOWEL COMPARISON

Let us now compare the CVs (Figure 3) and the focal points derived from modeling. For more details, see [15].

4.1 Cardinal C1[i] = (⇑F3↓F4)F3$_{F2=3200Hz}$

As exemplified in Figure 2, the renditions of [i] by DJ, PL and French [i] share the same acoustic characteristics: the spectrum is dominated by a concentration of energy above 3000 Hz, due to converging F3 and F4 formants. The distance between F3 and F4 depends on the speaker's F4; a few speakers realize [i] with converging F4 and F5.

Modeling shows that the convergence of F3 and F4 formants requires the constriction to be located in the prepalatal region, where F3 is maximally high (⇑F3 represented in Figure 3), and that the constriction has to be very narrow (Figure 4). F3 is a half-wave resonance affiliated with the front cavity (underlined formant — as "F3" — denotes the affiliation of a formant mainly with the front cavity), and is thus sensitive to lip configuration: lip spreading allows raising F3 further, diminishing the distance between F3 and F4. By contrast, when the constriction is palatal (⇑F2 represented in Figure 3), F2 reaches its maximum value and F3 and F4 diverge, resulting in a F2-F3 convergence.

The next figure illustrates two types of [i], proposed by PL as typical of British English and of American English (AE).

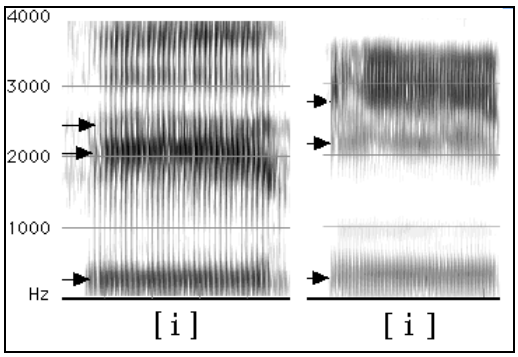

Figure 6: Spectrograms of British (left) and American English [i] (right), proposed as typical by PL.

At first sight, British English [i] is a palatal [i] (converging F2-F3), and does not correspond to a CV (at least produced by DJ, PL or by a French speaker). AE [i] seems to be close to (C1[i]) prepalatal vowel (converging F3-F4). This does not, however, seem to be the case. Gendrot et al. (2008) [16] calculated the formant frequencies of vowels in continuous speech (radio news) in eight languages, including English and French. Their data confirm that French [i] has the highest F3, the lowest F1 and the smallest (F3-F4) distance among the eight languages studied. AE is, on the contrary, one of the two languages with the largest F3-F4 distance for [i], and a much higher F1 and slightly higher F2 than French. (Note that English [i] is diphthongized unlike French [i]: the data found in Gendrot et al. are mean values).

4.2 Cardinal C9[y]= (↓F2⇓F3)$^{F2=1900Hz}$

The renditions of the vowel [y] by DJ, PL and the French [y] differ. The vowel [y] proposed by DJ does not sound like a French [y] to the French ear, but PL's does. In DJ's [y], F3 is located at mid distance between F2 and F3, while on PL's and French [y], F2 and F3 are very close (a single peak is detected in PL's [y]).

PL's and French [y] are focal vowels. They are characterized by a concentration of energy slightly below 2000 Hz due to converging F2 and F3 (F2F3) vowel). There is an exchange of affiliation of the third formant between focal/French [i] and [y]. Due to lip rounding and protrusion, the front cavity becomes longer than the back cavity and resonates at a lower frequency than the (originally) second resonance which is due to the back cavity (F1 is a Helmholtz resonance) [7].

The contrast [i]-[y] exploits the high sensibility of F3 to lip configuration when the constriction is prepalatal. Woods [17] argues that languages which contrast [i] and [y] will prefer prepalatal [i] to palatal [i], which is the point marked as ⇑F3 in Figure 3.

4.3 Cardinal C5[ɑ], C6[ɔ], C7[o], C8[u]: (F1F2) focal vowels with low F2

All back CVs, C8[u], C7[o], C6[ɔ] and C5[ɑ] are focal vowels. They share a low F2, close to F1. They are characterized by a concentration of energy increasing in frequency, from C8[u], the lowest concentration to C5[ɑ], around 1000 Hz.

The four back CVs form a natural class: they differ from the other vowels in that F3 and higher formants are not perceived (because of too low an amplitude): F2' (F2 prime) is equal to F2 for the back cardinal vowels. F2' is obtained by successful matchings of the vowel timber by an approximation of a single formant, F1 being fixed. F2' ir related to the perceptual contribution of upper formants (see [18] for details). F2' is higher than F2 for the non back vowels (see[19] for Swedish). The back vowels can even be synthesized using a single intermediate formant located between F1 and F2 [20], while the other vowels need a minimum of two formants.

Modelization shows that C8[u] is the vowel with the lowest possible concentration of energy that a human VT is capable of producing; and C5[ɑ] is the highest possible concentration of energy created by F1 and F2 convergence.

C7[o] and C6[ɔ] seem to be derivable by acoustic and auditory criteria. They are focal (but no turning points) and they are defined by an equal distance in terms of F1 and F2 between C8[u] and C5[ɑ], and thus are most likely to maintain perceptual distinctiveness between the four back vowels (in conformity with DJ's description and ADT).

When the constriction is in the back section of the VT, jaw opening has a much lesser effect on the formants than when it is in the front. In order to keep F1 and F2 close together in moving from C8[u] (two Helmholtz resonances) to C5[ɑ] (two quarter-wave resonances), there is a synchronization between the lips and the tongue movement: the degree of rounding has to decrease as the constriction is realized further back in the VT (see for example X-ray data [21]. For the back vowels, F2 frequency depends as much on the tongue constriction place as on the lip configuration. Rounding/protrusion and backing form a single functional entity, which has a single acoustic correlate: a low F2. The acoustic manifestation of backness and that of rounding are linked through the basic laws of acoustics and should not be considered separately.

4.4 Cardinal C2[e], C3[ɛ], C4[a], C9[y], C10[ø], C11[œ]: non focal vowels with high F2

PL's and French renditions of the front (rounded and unrounded) vowels are perceptually and spectrographically very close (Figure 3).

In contrast to the other vowels, [e ɛ a ø œ] are non focal vowels. [e ɛ a ø œ] are (FvF2vF3vF4) vowels. The notation

(F1vF2vF3vF4) denotes a non focal vowel. In the case of non focal vowels, formants depend on the whole VT. [e] and [ɛ] are characterized by a F3 at about mid distance between F2 and F4 (a criterion that the author uses in spectrogram readings to differentiate between [i], [y] and [e], which may have similar F2 — and sometimes F1 — values in continuous speech).

The two members of the pairs [e ɛ], [ø œ] differ by their F1. F1 is heard as a separate entity of F2, unlike in [u], [o] and [ɑ]. [e]- [ø] and [ɛ]-[œ] differ by lowering of the upper formants: [e] and [ø] have a lower (F2vF3vF4) than [ɛ] and [œ]. Unlike for the pair [i]-[y], where only F3 is lowered, all higher formants F2, F3 and F4 are lowered as a whole.

DJ's and French [a]s have the same quality and are unique in the sense that [a] has the highest possible F1 (higher than back [ɑ]). For a male speaker, F1[a] is usually around 700 Hz, and F1[ɑ] around 600 Hz. PL's [a] has the first two formants too close to be considered a good representative of a front vowel.

Unlike back vowels, where rounding and backing are necessary to keep them focal, lowering the jaw is sufficient to reproduce the different degrees of aperture of the front vowel.

5. FROM VOWELS TO CONSONANTS

As Roman Jakobson often pointed out, humans have only one VT, with which they produce both vowels and consonants. All phonemes sharing a similar sagittal profile and lip configuration will share similar acoustic characteristics.

The principal aim of this paper is to deal with the problem of defining references for vowels. In the present paragraph, we will limit ourselves to some illustrations to show the potential of a symbolic representation based on acoustics. The aim is to enhance the common acoustic characteristics between all the phonemes (vowels, glides, consonants) which share a similar place of constriction and tongue shape but a different degree of constriction (Figure 4) on the one hand, and between the coarticulatory effects (such as palatalization, labialization and velarization) on the other (for more information, readers are invited to look up the author's website).

The F-pattern of any phoneme (for the notion of "F-pattern", see [7]) is always partly determined by the F-pattern of the surrounding phonemes. The F-pattern is not always visible (during stops, for example), or it is partially visible (during fricatives). The resonances that can be seen on spectrograms of fricative are mainly the ones that due to the front cavity, but the whole F-pattern of stops and fricatives can always be calculated from sagittal profiles, see [7, 15].

Figure 7 illustrates the similarities in terms of F-pattern for four phonemes with the same fronted position of the tongue and global tongue shape (for the corresponding X-ray, see [21]). These similarities (such (as close F3 and F4) are due to their identity (such as palatal [j], which is intrinsically a (F3-F4) phone) or to coarticulatory influences of the following context vowel.[g] and [l] are palatalized due to the following [i], and the F-pattern of these two consonants in that context is characterized by close F3 and F4. As can be seen in this figure, the consonants have a resonance (when visible) of around 2000 Hz (corresponding to F2) and a concentration of energy around 3000 Hz or above due to F3 and F4 clustering. The onset of the following vowel (here [a], [i] and [u]) is characterized by a F2 at around 2000 Hz, which reflects the F-pattern of the consonants.

All sounds produced with an [i]-like tongue shape and position (such as those in Figure 7), i.e. all palatal and palatalized sounds (such as the palatalized consonants in Russian [7]) share a high F2 (around 2000 Hz), visible at the following vowel onset.

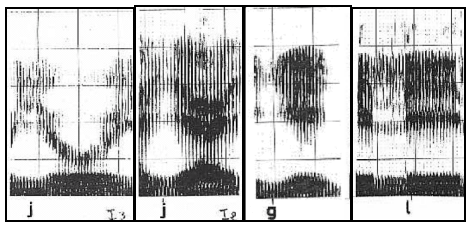

Figure 7: Spectrograms of [juj], [jaj], [gi] and [li].

Similarly, non palatal sounds in palatal context (such as [u] in [tut] or [juj], or [a] in [jaj]) will have a higher F2 and a lower F1 than their expected target values. Palatal context has a (↓F1↑F2) effect. If the lips are spread, the distance between F3 and F4 will diminish, and the effect of coarticulation will increase as the speech rate increases.

All sounds produced with an [u]-like tongue shape and position, and labialization, i.e. all labio-velar and labio-velarized sounds share a low F2, visible for the consonants at the following vowel onset. Similarly, the phonemes surrounded by labio-velar sounds will display a lower than expected F2. For example, in Figure 7, the [i]-like cluster (F3F4) that characterizes palatal [j] is lowered in the [u] context as compared to the [a] context.

All sounds produced with an [ɑ]-like tongue shape and position, i.e. all pharyngeal and pharyngealized sounds, will share a low or lower than expected F2 and a high or a higher than expected F1, and therefore converging (F1F2) formants, visible at the following vowel onset (due to the continuity of the F-pattern in speech). The effect can be described as a (↑F1↓F2) effect.

The magnitude of the effect on the surrounding phonemes depends on the language, and on the prosodic status of the phoneme, but the direction of the shift can always be predicted. Depending on its position within the syllable, on the word boundaries and on word stress, a consonant will be more consonant-like (closer constriction somewhere in the VT, larger contact, higher velum, tenser vocal folds, less voicing), or more vowel-like (looser constriction, lower velum, laxer vocal folds, greater tendency to voicing). Similarly, depending on some factors, a phoneme (or a whole syllable) will be more or less influenced by the surrounding context, or influence it. Such tendencies can be modeled using articulatory modeling (see [14] for a few examples). Articulatory modeling thus allows us to make step-by-step simulations of any sound change and to infer about the causes which can have induced the segmental change.

6. CONCLUSIONS

It is necessary to establish fixed references for the phonetic description of vowels. Such references may correspond to well-defined acoustic criteria. For a complete study of the realisation of the vowel system in a given language, it may be useful to set up five types of analyses: 1) acoustic analysis of vowels produced in isolation by speakers, 2) acoustic characteristics of a collection of the stimuli selected by the same speakers/listeners as the best prototype of each vowel (see for example the method used in [2]), 3) statistics on the vowel formants in continuous speech and studies of the vowels in different contexts, 4) detailed analysis of the time course of the formants in sequences comprising pairs of close vowels or vowels in "adverse" phonetic contexts and 5) analysis of the speaker maximal acoustic space. There are more and facilities to create large data bases and to exploit them (for the third type of analysis). Analysis of pairs of close vowels such as [iyiyiy], or [uyuyuy], or sequences like [tututu] or [ririri] (the fourth type of analysis) highlight the acoustic means of maximizing contrasts: the symbol C9[y] should only be used when there is a clear exchange of cavity affiliation when the speaker is asked to utter [iyiy], and when the formants F2 and F3 are as close as possible (Swedish [y] does not seem to be a C9[y] [15]). The availability of a real time speech analyser helps a speaker explore his/her maximum vocalic space by providing visual feed-back (the fifth type of analysis). For

example, the symbol C1[i] should only be used if the measured F3 of the vowel to be transcribed is the highest F3 that the speaker can produce. How far such references can be used for describing the sounds of a new undescribed language is currently under progress [22]. Real time computer displays (F1-F2 displays are not sufficient) and written description of the acoustic characteristics of the sounds may also help teaching reference vowels in second language teaching and the teaching of pronunciations to hearing-impaired and normal-listening students [22].

Articulatory modeling provides students of speech sciences with access to almost all subtleties of the acoustic theory of speech [7] in a rather enjoyable manner, without requiring a high level background in mathematics or physics. Maeda's model is appropriate as a teaching tool (there are other valuable articulatory models available on the Web). It has been used in France and abroad for a very large number of applications (see [14] for references). A very small number of formulae based on the resonance properties of tubes holistically explain the common acoustic features of phonemes, the effects of phonemes on their surrounding contexts, and the effects of secondary articulation. Maeda's model can be used realistically to evaluate the acoustic consequences of any perturbation in the VT and any compensatory phenomena. It still needs to be amended in order to be usable for contrasts that are not used in French. The 7 parameters do not provide an independent control of tongue blade (retroflexion), pharynx width, and the form of the cross-sectional area. The effects are evaluated by a manual change in the area functions produced by the model. It cannot simulate laterals, and the supraglottal friction sources are located in an ad hoc manner (for strong constrictions, more detailed aerodynamical aspects have to be introduced).

The whole spectrum — from F1 to F4 — has to be taken into account for the calculations of acoustic distance between vowels as is the case for the calculations of F2'. Using only F1 and F2 (in Hz, Bark, Mel, etc.) should be avoided. Moreover, the relative amplitude of the formants is extremely important, at least in languages like French which has distinctive nasalization: reduction of F1 amplitude leads to the perception of a nasal(ized) vowel [17].

QT insists on articulatory stability, ADT on perceptual distinctiveness and DFT tends to make a compromise between inter- and intraspecificities of the vowels. The plateau-like regions are those of *articulatory* stability, but for no more than two formants, Fn and Fn+1. These are regions of *perceptual* salience of the converging formants, whose salience contributes to perceptual distinctiveness. Furthermore, these regions are favorable for creating new *contrasts* (if needed) for languages with crowded systems by manipulating the degree of constriction and/or labialization. Anatomy further constrains the range of possible contrasts: it would be much more difficult to manipulate larynx height and width than lip protrusion and rounding. These three theories are not contradictory, and principles of different nature may be at work.

The definition of the CVs by DJ may have been influenced by the French system (DJ learned French phonetics from the French phonetician Paul Passy (the inventor of the IPA) and he was teacher of French). Spectral sharpness seems to play a key role in the vowel system of French, and X-ray data show a surprisingly strong constriction along the VT for all focal vowels (French [i] even tends to be fricatized). The sensibility of the F3 to labialization when the constriction is prepalatal and the possibility to create a focal vowel such as [y] seem to play a key role for the selection of C1[i] and C9[y]. There is no reason for a given language to choose C1[i] as the acoustic realization of the phoneme /i/, if labialization is not contrastive in the language. The mid front vowels-which seem to be defined by perceptual distinctiveness are not focal, and therefore less stable. They are more likely to be confounded, or they

tend to alternate in French [e/ø] and [ɛ/œ] depending to the syllable structure. The mid back vowels are focal, but less distinctive between them: there are not turning points (Law 1). [a] and [ɑ] are confusable, because large jaw opening for the front vowel [a] leads to a decrease in the cross-sectional area in the back cavity, which characterizes [ɑ].

The feature "height" is best modeled by a jaw lowering in the case of the front CVs, and by a synchronization between delabialization and constriction backing in the case of the back CVs. The feature "labial" mainly affects F3 in the case of [i], it effects the upper formants in the case of the other front vowels and F2 in the case of back vowels. Height, backness/frontness and rounding are therefore not orthogonal on the acoustic-perceptual side, and the axes used in IPA are somewhat misleading, at least for a use in phonetic transcription. The present paper has tried to propose some hints for refining the notion of Cardinal Vowels, formally proposed by DJ, as acoustic-perceptual targets (APT), based on nomograms and a number of current theories.

7. ACKNOWLEDGMENTS

We would like to thank Shinji Maeda for his helpful comments, Nick Clements and Alexis Michaud for fruitful stimulation, Takeki Kamiyama and Cedric Gendrot for their careful reading of the final version.

8. REFERENCES

[1] Jones, D. 1918. *An Outline of English Phonetics*. Cambridge University Press. Ninth edition, 1967.

[2] Willerman, R., Kuhl, P. 1996. Cross-language speech perception of front rounded vowels: Swedish, English, and Spanish speakers. *Proceedings of the International Conference on Spoken Language Processing* ICSLP, 96.1: 442-445.

[3] Maeda, S. 1989. Compensatory articulation in speech: analysis of x-ray data with an articulatory model. *EUROSPEECH-1989*, 2441-2445.

[4] Maeda, S. 1990. Compensatory articulation during speech: evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model. In: Hardcastle, W., Marchal, A. (eds). *Speech Production and Speech Modelling*. Dordrecht: Kluwer Academic, 131-149.

[5] Ladefoged, P. 1982. *A course in Phonetics*. Second Edition. Harcourt Brace Jonanovich: New York.

[6] Daniel Jones's and Peter Ladefoged's renditions of the cardinal vowels: http://www.phonetics.ucla.edu/course/chapter9 / cardinal/cardinal.html

[7] Fant, G. 1960. *Acoustic theory of speech production, with calculations based on X-ray studies of Russian articulations*. The Hague: Mouton.

[8] Stevens, K.N. 1989. On the quantal nature of speech. *J. of Phonetics* 17.1/2, 3-45.

[9] Liljencrants L., Lindblom, B. 1972. Numerical simulations of vowel quality systems: The role of perceptual contrast. *Language*, 48, 839-862.

[10] Lindblom, B. 1986. Phonetic Universals in Vowel Systems. In: J. Ohala et al. (eds), *Experimental Phonology*. Academic Press, 13-44.

[11] Schwartz, J.L., Abry, C., Boé, L-J, Ménard, L., Vallée, N. 2005. Asymmetries in vowel perception, in the context of the Dispersion-focalization Theory. *Speech Communication* 45, 425-434.

[12] Schwartz, J.L., Boë, L.J, Vallée, N., Abry, C. 1997. The dispersion-focalization theory of vowel systems. *J. of* Phonetics 25, 255-286.

[13] Ladefoged, P., Bladon, A. 1982. Attempts by human speakers to reproduce Fant's nomograms. *Speech Communication* 9, 231-298.

[14] Badin, P., Perrier, P., Boë, L-J. 1990. Vocalic nomograms: Acoustic and articulatory considerations upon formant convergences. *J. Acoust. Soc. Am.* 87, 1290-1300.

[15] Vaissière, J. 2007. 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. In: Beddor, P.S., Solé, M.J., Ohala M. (eds), *Experimental Approaches to Phonology*. Oxford: Oxford University Press, 54-72.

[16] Gendrot, C., Adda-Decker, M., Vaissière, J. 2008. Les voyelles /i/ et /y/ du français: aspects quantiques et variations formantiques.

Proceedings of Journées d'Etude de la Parole, Avignon, France, 205-208.

[17] Woods, S. 1986. The acoustical significance of tongue, lip and larynx maneuvers in rounded palatal vowels. *J. Acoust. Soc. Am. 80*. 391-401.

[18] Carlson, R., Fant, G., and Granstrom, B. 1975. Two-formant models, pitch and vowel perception. In: Fant, G. and Tatham, M. (eds). *Auditory and Analysis and Perception of Speech*, Academic Press, London.

[19] Delattre, P. C., Liberman, A. M., Cooper, F. S. 1951. Voyelles synthétiques a deux formants et voyelles cardinales. *Le Maître Phonétique* 96, 30-36.

[20] Carlson, R., Granström, B., Fant, G. 1970. Some studies concerning perception of isolated vowels. *Speech Transmission Laboratory Quarterly Progress Status Report: Stockholm*, 2/3: 19-35.

[21] Bothorel, A., Simon, P., Wioland, F., Zerling, J-P. 1986. *Cinéradiographie des voyelles et consonnes du français*. Publications de l'Institut de Phonétique de Strasbourg : Strasbourg.

[22] Lindbom, B., Sundberg, J. 1969. A quantitative theory of cardinal vowels and the teaching of pronunciation. *Speech Transmission Laboratory Quarterly Progress Status Report*: Stockholm, 10, 19-25.

Jacqueline Vaissière, Laboratoire de Phonétique et
Phonologie (LPP), CNRS/Univ. Paris 3.
jacqueline.vaissiere@univ-paris3.fr