Dynamic properties of an acoustic tube: Prediction of vowel systems

René Carré

▶ **To cite this version:**

René Carré. Dynamic properties of an acoustic tube: Prediction of vowel systems. Speech Communication, 2008, 51 (1), pp.26. 10.1016/j.specom.2008.05.015 . hal-00499221

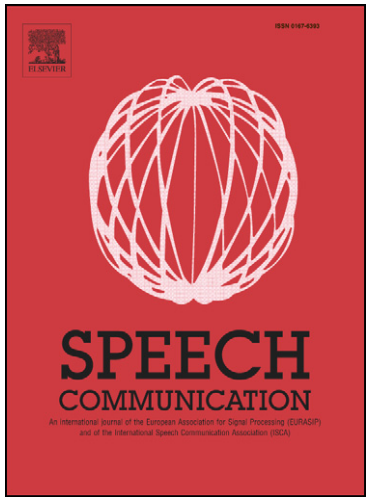
HAL Id: hal-00499221
https://hal.science/hal-00499221

Submitted on 9 Jul 2010

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Accepted Manuscript



Dynamic properties of an acoustic tube: Prediction of vowel systems

René Carré

PII: S0167-6393(08)00087-3
DOI: 10.1016/j.specom.2008.05.015
Reference: SPECOM 1729

To appear in: *Speech Communication*

Received Date: 9 January 2006
Revised Date: 27 May 2008
Accepted Date: 27 May 2008

Please cite this article as: Carré, R., Dynamic properties of an acoustic tube: Prediction of vowel systems, *Speech Communication* (2008), doi: 10.1016/j.specom.2008.05.015

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

ACCEPTED MANUSCRIPT

Dynamic properties of an acoustic tube: Prediction of vowel systems

René Carré

Laboratoire Dynamique du Langage, UMR 5596, CNRS, Université Lyon 2,

14 Avenue Marcelin Berthelot, 69363 Lyon cedex 07 France

recarre@wanadoo.fr

(33) 476879439

Abstract

Approaches to characterizing and explaining the diverse phonologies of the world's languages usually begin with data from the analysis of speech signals or from the results of speech production and perception experiments. In the present paper, the dynamic acoustic properties that arise from changing the shape of a simple acoustic tube 18cm length (without any articulatory machinery) are explored to develop a simple and efficient acoustic communication system. By efficient we mean that minimum deformations of the tube lead to maximum acoustic variations. Intrinsic characteristics of the tube are derived from these specific 'gestural' deformations associated with formant trajectories in the acoustic plane. This deductive approach (without reference to data on speech production or speech signals) leads to define an acoustic communication system characterized by its acoustic space and by several specific formant trajectories. The acoustic space fits well with the vowel triangle, and 18 oral vowels can be placed on the trajectories. From these deductive results a tentative explanation of vowel systems is proposed. The good match between deductive prediction and observation

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

results encourages to make further predictions, formulating hypotheses about a unified view of vowel and consonant production, and reconsidering the relation between phonetics and phonology.

Keywords. Speech production, acoustic tube, dynamic properties, vowel systems.

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

1. Introduction

The nature of speech sounds and their cognitive representations are still not understood. What is a phoneme? Acoustically? Phonologically? What are the links between a phoneme and its cognitive representation? What is the origin of a phoneme: Does it emerge automatically from properties of speech production and perception (Studdert-Kennedy, 1987) ? Is it learned or is it a bundle of innate features (Chomsky and Halle, 1968) ? Is speech a sequence of static spatio-temporal targets or of dynamically interwoven movements?

Vowel systems of languages have been extensively studied and data have accumulated on specific inventories (Crothers, 1978 ; Maddieson, 1984). Again numerous questions arise: How should we label and classify speech sounds? That the question is not simple is evident from the many phoneticians who participate in the description of the world's languages with different methodologies. How many segments should be included in the International Phonetic Alphabet (IPA) (Ladefoged, 1990)? Is there a limit on this number (Lindblom, 1990) ? What is the maximum possible number of vowels in a single language? With, say, one hundred possible vowels and, say, thirty vowels in a single system, many millions of different vowel systems could be obtained. Yet, no such diversity is observed in language inventories. On the contrary, they display marked regularities, and typological studies of these regularities have been developed (Crothers,

ACCEPTED MANUSCRIPT

1978 ; Maddieson, 1984; Schwartz*, et al.*, 1997). The main question then is: Why are there any regularities at all? What causal principles underlie the observed data?

To find out these underlying causes in the case of vowels and vowel systems, our research follows two methodological steps: a) a deductive approach and b) modeling.

1.1. Deductive approach

The study of speech communication, like any other scientific topic, can adopt either of two main approaches (Lindblom, 1990). The first is data-based. Observations of speech signals, area functions, vocalic systems, distinctive features, etc., are organized and processed in order to identify trends. The aim is to represent the data in a simple way in terms of a set of parameters capturing, as closely as possible, the essential properties of the phenomena under investigation. The results can be used to generate predictions, although such predictions are the fruit of inductive reasoning often fraught with inherent circularity: one specific piece of data is "explained" by the others (one vowel in a system is explained by the others) and regularities observed in one language are "explained" by regularities observed in another. Therefore, data-based approaches are inherently unable to yield strong explanatory accounts. Although data can (and should) be organized to efficiently describe findings, purely descriptive accounts cannot identify the causal and generally simple principles that presumably underlie the observed facts.

ACCEPTED MANUSCRIPT

In speech research, the construction of speech databases, — speech sound inventories across many languages complete with classification and statistics — exemplify projects that take this first approach. To cite a well-known example, although Peterson and Barney (1952) provided a useful representation of American English vowels in the F1-F2 plane, this representation alone cannot answer fundamental questions, such as what the origin of the vowel triangle is. Furthermore, by itself, a data-based approach is incapable of furnishing a lead to model the data when the phenomenon under study is complex. What may happen in such a case is that the investigator, trying to build a model, will be tempted to take into account all the data, for fear of ignoring any piece that could be essential. As time goes on, such models then tend to become more and more complex, in direct opposition to the primary raison d'être of scientific models, which is finding the simplest way to capture the essential data (Carré and Mrayati, 1990). Investigators may also be led to believe that more data need to be collected, lest they fail to uncover the underlying (and possibly general) mechanisms at the origin of disparate and seemingly incoherent bodies of data. In short, the logic of a data-based approach fuels the need for "more data". But, in the end, the resulting model is often *ad hoc*, lacking in explanatory power, and generally overly complex.

The second approach is deductive. Here, as in the data-based approach, facts are gathered but the investigator endeavors to explain the facts from general principles

ACCEPTED MANUSCRIPT

independent of the facts themselves. For example, the investigator may analyze the speech signal not only in terms of formant frequencies but in terms of the human systems that produce and perceive them, in order to explain how and why the particular signals arose in the first place. Such a deductive approach was successfully taken by Liljencrants and Lindblom (1972) and Lindblom (1986) in their attempts to predict vowel systems from an articulatory model and principles of perceptual contrast. Similarly, the quantal theory (Stevens, 1989) invokes a criterion of acoustic stability to deduce formant properties of vowels. Nevertheless, when one is successful in deriving characteristics of the speech signal from characteristics of the speech production and perception systems, a more fundamental question arises: Why do the production and perception systems possess their particular characteristics? They are biological systems which have evolved and can continue to evolve according to environmental constraints. Can we say that they have evolved to their present state to satisfy communication needs? This question is legitimate and it, too, could be deductively approached from higher-order general principles, such as maximum acoustic contrast, economy of effort, and simplicity (as opposed to complexity). In other words, the systems of speech production and perception could be the product of an evolutionary process, driven by general principles applied to a sound-producing device at the disposal of humans, that is, an acoustic tube (today called the “vocal tract” but originally perhaps not well adapted for

ACCEPTED MANUSCRIPT

acoustic communication), so as to form an efficient system. For example, on the one hand, the vowel triangle, as observed on the F1/F2 plane (Peterson and Barney, 1952), can be derived from the speech production system (by an articulatory model (Maeda, 1982) or from an area function model (Fant, 1960)). On the other hand, vowel systems as observed in inventories (Crothers, 1978), can be derived from an articulatory model and principles of perceptual contrast (Lindblom, 1986). But both the vowel triangle and vowel systems may in turn result from efficient use of an acoustic tube of 18cm, and this is the hypothesis of the present paper. The deductive approach is well suited to test this hypothesis. From an acoustic tube of 18cm, without any constraint, it is proposed to build the most efficient possible acoustic communication system characterized by its resonant frequencies. Then, the properties of such a system will be compared with those of the human system. If they fit well, as far as its resonant frequencies are concerned, we may say that the human system is acoustically efficient and the deductive approach can be used to try to explain the properties.

In what follows, an acoustic tube is viewed as an instrument for efficient acoustic communication if:

- minimum deformation of the tube leads to maximum acoustic variation (more or less equivalent to a minimum effort criterion);

ACCEPTED MANUSCRIPT

- the maximum acoustic space is used (giving maximum acoustic contrast and thus allowing communication in noisy environments);
- the coding is efficient, i.e., the number of coding units is small (for low bit rate) and fully used (each new coding unit increasing the complexity of the system must be fully used, that is, must multiply by two the number of possible new sounds);
- the sequence of coding units is not described in terms of a succession of static parameters, each characterized in absolute coordinates, but in terms of variations (as in delta modulation) to reduce the rate of information transfer;
- the rate of information transfer is increased by parallel transmission of coding units.

1.2. Modeling

Specification of an area function of an acoustic tube as a more or less continuous graph of cross-sectional area allows detailed calculations of the acoustic response, but is not practical for systematic and simple descriptions or for testing the relevance of the descriptions. Modeling deformation of the tube for efficient acoustic communication by a reduced number of parameters is a useful tool. The criteria for a “good” model are the following:

-it can capture the essential of the tube deformation/acoustic relation in a simple way,

-the parameters of the model are few and make sense for an acoustic communication system,

ACCEPTED MANUSCRIPT

-the parameters cause linear and orthogonal acoustic variations.

In short, the nomograms of the model should be simple and easily used to explain the relations between area deformation and acoustic form. The model could then explain certain aspects of the acoustic communication process and predict specific behavior.

Figure 1 summarizes our approach. First, minimal deformations of the tube must bring maximal acoustic contrast (in terms of formants): $\Delta F/\Delta A$ should be maximized, where ΔF is formant variation and ΔA corresponding variation of the area function. Then, ideally, the relation obtained between area function deformation and formant variation should be monotonic and orthogonal, i.e., two different linear area deformations should give rise to two different linear formant variations. The deformation parameters should also be simple and few in number so as to allow communication with a low bit rate.

In the paper, following this deductive approach and modeling, we set up a simple and efficient acoustic communication system which is an acoustic tube structured into regions. It is characterized by its acoustic space and by specific formant trajectories (reduced in number). The acoustic space will be compared to the vowel triangle and the formant trajectories structuring the acoustic space will be studied as supporting the production of vowels. Then, from these formant trajectories a tentative explanation of vowel systems will be proposed. According to the results, further predictions could be

ACCEPTED MANUSCRIPT

formulated as hypotheses for a unified view of vowel and consonant production and on the relation between phonetics and phonology. Our research is here limited to studies of oral vowels.

Figure 1.

2. Dynamic characteristics of an acoustic tube

Following the deductive approach described above, we set up a theoretical communication system based on efficient dynamic properties (by efficient dynamic properties, we mean: minimum shape deformation leading to maximum acoustic variation) of an acoustic tube 18 cm long, the length of a typical human male vocal tract (Carré, 2004).

2.1. Deformations from the uniform position: towards the vowel triangle

The vowel acoustic space is generally limited to the F1/F2 plane. It can be obtained from data (Peterson and Barney, 1952) or from a speech production model (Liljencrants and Lindblom, 1972; Boë*, et al.*, 1989). But how does this plane compare with the maximum acoustic F1/F2 plane of an acoustic tube? Does the speech production system exploit the whole possible acoustic space of an acoustic tube of 18cm length? In other words, is the speech production system well adapted in terms of the F1/F2 acoustic

ACCEPTED MANUSCRIPT

plane? The acoustic space can be obtained by measuring the resonant frequencies of the tube for all its possible shapes. But the acoustic properties of the tube revealed by such an approach are necessarily purely static: no information is given on the dynamic acoustical behavior of the tube, on the acoustic stability of specific shapes or on transitions from one shape to another. Yet there is an infinite number of area deformations, involving an infinite number of formant trajectories across the F1/F2 plane, in moving from one point on the plane to another (Carré*, et al.*, 2001). Our objective is both to find the maximum acoustic F1/F2 plane that can be obtained with a tube without any constraint and to determine which trajectories are best suited for acoustic communication. "Minimum effort" (or minimum tube deformation giving maximum acoustic variation) should be the selection criterion for these specific trajectories.

Figure 2.

Figure 2 recalls the general scheme of the recursive algorithm used to deform any initial shape of the tube according to the criterion (Carré*, et al.*, 1994; Carré*, et al.*, 1995; Carré, 2004). The goal is to increase or decrease F1 or F2 (or both F1 and F2) by deforming, step by step, the shape of the tube according to the appropriate sensitivity

ACCEPTED MANUSCRIPT

function (Fant and Pauli, 1974) (the sensitivity function SF1 corresponds to the first formant variations obtained for local area perturbation – for example 1% of the area – all along the tube in steps of 1cm). It has been shown (Carré, 2004) that the shape deformed according to the sensitivity function leads to maximum formant variation. This operation is repeated until the physical limits of the acoustic space are reached. In this algorithm, the goal is acoustic; the task is to deform the shape of the tube to reach the goal. Notice that the goal is not to reach static targets (which play no role in the process) but to increase (decrease) formant frequencies efficiently (i.e. so that minimum area deformation leads to maximum acoustic variation). This algorithm describes an evolutionary process: there is an initial state, a selection criterion (minimum deformation giving maximum acoustic variation) and a series of recursions or iterations in order to arrive at the maximum acoustic value. The final state is not given at the beginning: it is the end-product of the evolution. The maximum possible acoustic value and efficient deformations of the tube with the corresponding direction of trajectories in the acoustic plane are the main results obtained by the recursive algorithm.

To discover the maximum possible acoustic space, the algorithm is operated from a neutral (or uniform) initial shape of $4cm^2$. The source consists of a fixed cavity of 2cm length and $2cm^2$ area. A new shape is obtained according to the following formula (Carré, 2004):

ACCEPTED MANUSCRIPT

$A_{i+1}(n) = A_i(n)\,[1+k_1 S_i F1(n)+k_2\, S_i F2(n))]$, $1 \le i \le 10$, $3 \le n \le 18$ (1)

where n is the section number (each section of 1cm length) counted from the closed end (the sections 1 and 2 being always fixed and equal to 2cm^2, the sections between 3 and 18 being initially set equal to 4cm^2), A_i is the shape, S_iF1 is the corresponding sensitivity function for F1, A_1 is the initial shape (it has been shown that A_i can be limited between 0.5 and 10cm^2, intrinsic limits for efficiency (Carré, 2004)); the subscript i is the number of the iteration (here between 1 and 10). The algorithm is used for different values of the coefficients of the sensitivity functions k_1 and k_2. Recall that a positive k leads to an increase of the corresponding formant, and a negative k to a decrease (Carré, 2004). Here $k_1 = \cos(\alpha)$ and $k_2 = \sin(\alpha)$ for $0° < \alpha < 360°$ by steps of 5° to compute 72 different deformations from the uniform position and corresponding trajectories in the F1-F2 plane from the neutral.

Figure 3 shows the result for $k_1 = -0.702$ and $k_2 = +0.702$, a) the deformation is anti-symmetrical, b) the trajectory in the F1-F2 plane is rectilinear and c) no noticeable acoustic variation is observed after seven iterations. In the experiments described in subsequent parts of the article, formant frequencies were calculated from the area function of the tube using the algorithm proposed by Badin and Fant (1984).

Figure 4 shows the seventy two trajectories obtained from the neutral. They are more or less equally spaced in the F1/F2 plane. An acoustic triangle clearly appears, very similar

ACCEPTED MANUSCRIPT

to the familiar vowel triangle, when the seventy two trajectories are limited to their linear parts. On the trajectories, the dots correspond to the results obtained by the different iterations; the distances between successive dots are smaller towards the limits of the vowel triangle showing that the relation between a given deformation and its acoustic effects is less and less efficient towards the limits of the vowel triangle. This means that the vowel triangle is limited to the efficient parts of the articulatory-acoustic relation and cannot be larger. Strict linear relation would have given dot circles centred at the neutral. But here, the greater the distance from the centre, the more the circle is deformed. Thus the human production system is well adapted to exploit the maximum acoustic space. This maximum is obtained for an area range between 0.5 and 10 cm^2 which corresponds to the range observed in X-ray data for vowel production (Fant, 1960).

Figure 3.

Figure 4.

2.2. Deformations from any initial shape: towards efficient deformation gestures and corresponding formant trajectories

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

Recall that the algorithm was first applied from the uniform shape of the closed-open or closed-closed tube without a source cavity (Carré, 2004). Figures 5a and b show the deformation and the corresponding formant trajectory obtained with a closed-open tube: the tube is divided into four regions. The deformation is anti-symmetrical: two deformation gestures lead to two constrictions associated with two cavities. Figure 6 shows the deformation and the corresponding formant trajectory obtained with a closed-closed tube: the tube is divided into 3 regions. The deformation is symmetrical: one deformation gesture leads to one constriction associated with two cavities.

Figure 5.

Figure 6.

The algorithm can also be applied from any initial shape of the tube. Figure 7a shows the deformation obtained by the algorithm after twenty iterations for an increase of F2 (k_2 = 1) and a decrease of F1 (k_1 = -1) from a configuration with back constriction and front cavity. This deformation is rectilinear and transversal; it realizes a front constriction (associated with a back cavity – anti-symmetrical behavior). The corresponding formant trajectory (Figure 7b) is rectilinear. Toward the two ends of the formant trajectory, tube deformation is evidently no longer acoustically efficient, since

ACCEPTED MANUSCRIPT

each iteration of the recursive process (marked by dots on the trajectory) leads to smaller and smaller acoustic variations. The two trajectory endpoints [1] and [2] describing the trajectory lie at the limits of acoustic efficiency.

Figure 7.

The main results of this deductive and iterative approach are the following when F1 and F2 are taken into account (Carré, 2004):

- The gestural deformations of the tube leading to maximum acoustic space (for minimum deformation) are simple (rectilinear), few in number (two), and perpendicular to the main axis of the tube (transversal). They divide the tube into four regions in the case of a closed-open tube. The corresponding formant trajectories are also simple and rectilinear, they structure the F1/F2 plane, just as the gestural deformations structure the tube;
- The maximum acoustic space which is obtained for shape areas varying between 0.5 and 10cm^2 is a triangle and corresponds to the vowel triangle;
- The relation between gestural deformations and formant trajectories in the F1/F2 plane is monotonic and pseudo-orthogonal, i.e. more or less independent of each other;

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

- Two main tube types appear: (i) the closed-open tube (closed at the source and open at the output): its deformation is anti-symmetrical (a front constriction is associated with a back cavity and vice-versa), (ii) the closed-(quasi) closed tube: its deformation is symmetrical (a central constriction is associated with two lateral cavities);
- The complexity of the deformation in terms of the number of regions of the tube involved during the deformation process increases with the number of formants taken into account: the closed-open tube is divided into two, four and eight regions when respectively F1, F1 and F2, F1 and F2 and F3 are controlled;
- The Distinctive Region Model (DRM) (Mrayati*, et al.*, 1988) provides an adequate account of the results. The model is anti-symmetrically controlled in the case of the closed-open tube; it is symmetrically controlled in the case of the closed-closed tube. In sum, the model, intrinsically dynamic, has all the characteristics quoted above. The DRM model is deduced from acoustic theory with specific criteria such as economy of effort, not from speech production data.

Now, from the DRM model reproducing the efficient dynamic properties of an acoustic tube, we will try to predict vowel systems, acoustically distinctive in perception.

3. From the DRM model to vowel systems

ACCEPTED MANUSCRIPT

To predict vowel systems, we use the DRM model (Mrayati, et al., 1988; Carré and Mrayati, 1990; Carré and Mrayati, 1992; Carré, 2004) the characteristics of which represent the efficient dynamic properties of an acoustic tube. Recall that two main DRM models were obtained (Carré, 2004): one, anti-symmetrical, corresponding to the closed-open tube and the other, symmetrical, corresponding to the closed-closed tube. Our first objective is to produce trajectories corresponding to the borders of the "vowel triangle" in order to exploit maximum acoustic contrast.

3.1. With a closed-open DRM model

In a first step, the eight region DRM closed-open model (Figure 8a) is used but R3 and R4 are equal like R5 and R6. This scheme allows control of both F1 and F2. R1 represents the source cavity: its area is fixed at $2cm^2$. R2 is set at the mean value of R1 and R3. R7 is set at the mean value of R6 and R8. R3, R4 and R5, R6 are controlled anti-symmetrically following the results obtained by the preceding algorithm: a front constriction is associated with a back cavity and vice-versa. The model is controlled by two parameters (or deformation gestures) perpendicular to the main axis of the tube. These two deformation gestures obtained by deduction from our criteria are similar to the "tongue gesture" and the "lip gesture" of the speech production system. In the following, we will continue to call them "tongue gesture" and "lip gesture", although they are parameters of our acoustic model, and not articulatory parameters. The area

ACCEPTED MANUSCRIPT

variations are between 0.5 and 10 cm^2 (sufficient for maximal acoustic variations) and are obtained by steps proportional to the diameter of the section. The formant trajectory in the plane F1/F2 obtained by the "tongue gesture" can be described by two trajectory endpoints [1] and [2] (numbers rather than vowel symbols are used here because they are obtained by deduction from the acoustic characteristics of the tube, not from speech data), at the limits of acoustic efficiency (Figure 8b). The units [1] and [2] correspond to two peaks of the vowel triangle. The trajectory [1-2] is one of the border lines of the "vowel triangle". If the "tongue gesture" is co-produced with the "lip gesture", another trajectory is obtained [1-3]. The acoustic effects of the two control parameters are pseudo-orthogonal: the "tongue gesture" from the configuration for [1] gives an increase of F2 and a decrease of F1; the "lip gesture" closing gives a decrease of both F1 and F2. Closing the configuration for [2] gives [3]. Moreover, the trajectory [1-2] is roughly made of two elements: [1-4] parallel to the F2 axis, and [4-2].

Figure 8.

Figure 9 shows the "lip gesture" effect from the configuration [2] for ten steps of equal diameter variation. There is almost no acoustic effect between 10 and 2 cm^2 "lip" areas; the main acoustic effect is between 2 and 0.5 cm^2. This is not the case for the "tongue

ACCEPTED MANUSCRIPT

gesture" (see the equal acoustic variation between marks on [4-2] in Figure 8b). This means that if [1-3] is produced with strict co-production of the "tongue gesture" and the "lip gesture" (i.e. if the two gestures are identically phased in time and in area), the acoustic effect of the "lip gesture" is not rectilinear: From [1], the trajectory points initially to [2] and only at the end to [3]. For an expected rectilinear trajectory (Carré and Mrayati, 1991), with progressive acoustic effects of the "lip gesture", we must either anticipate the "lip gesture" in the time domain or (as in Figure 8b) apply a logarithmic scale to the realization of the gesture. The trajectory [1-3] is then rectilinear and the acoustic effects for each gesture are equal during their realization in the time domain (Carré and Divenyi, 2000).

Figure 9.

3.2. With a closed-closed DRM model

In a second step, the closed/quasi-closed DRM model is exploited. The recursive algorithm using the sensitivity function, with a fixed "lip" opening equal to 0.5 cm^2 (Carré, 2004), yields a three region model, each region of equal length, taking into account only the first two formants (Figure 6). Practically, to represent the results of the algorithm, the preceding eight region closed-open model is used, but the R4 and R5

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

region areas are equal. Figure 10a shows the model: a central constriction is automatically obtained associated with two lateral cavities. The unit [5] is reached with its corresponding unit [6] for "lip" opening equal to 10cm^2 (Figure 10b). The unit [5] corresponds to the third corner of the "vowel triangle".

Figure 10.

3.3. From a closed-open to a closed-closed DRM model

In a third step, we are interested in finding how to pass from the closed-open model to the closed-closed in order to obtain the trajectories [1-5] and [2-5] ([1] and [2] from Figure 8b, [5] from Figure 10b) corresponding to the two other border lines of the "vowel triangle". To determine these transitions, the preceding algorithm was again used with various fixed "lip" openings ranging from the largest to the narrowest opening between 16 and 0.01 cm^2 (16, 4, 2, 1, .5, .25, .1, .01 cm^2), so that situations intermediate between "lip" opening and "lip" closing could be evaluated.

Figure 11.

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

From the uniform tube, increasing F1 and decreasing F2 lead to the results shown in Figure 11 after ten iterations. Acoustic effects are small for lip openings between 16 and 2 cm^2, as already remarked. Saturation effects are observed near F1=F2.

Figure 11a shows that a rectilinear trajectory [1-5] can be obtained with a constriction area equal to 0.5 cm^2 moving from the back to the center of the tube (Figure 11b) (Carré and Mrayati, 1995).

Then, the trajectory [1-5] obtained is obtained from a closed-open to a closed-closed DRM model by a "longitudinal tongue gesture" (displacement of the constriction -- area equal to 0.5 cm^2) from the back to the center of the tube co-produced with the "lip gesture" (figure 12). The trajectory [1-6] is obtained without "lip" closure.

Figure 12.

The same approach can be used to study the trajectory [2-5] ([2] from Figure 8b, [5] from Figure 10b). The constriction moves longitudinally from the front to the center of the tube. The "tongue gesture" is co-produced with the "lip gesture". For the trajectory [6-2], the "lip" opening is 10 cm^2.

3.4. Vowels: sub-products of formant trajectories

ACCEPTED MANUSCRIPT

The deductive approach used to study the properties of an acoustic tube leads to:

- an acoustic space similar to the vowel triangle;
- three main different "tongue gestures" similar to those observed in production: the transversal displacement of the tongue constriction from front to back (and vice-versa) and the two longitudinal displacements from front to center and back to center (and vice-versa) which are similar to those proposed by Gunnilstam (1974) from articulatory data;
- a range of area variation between 0.5 and 10cm^2 (intrinsic limits of efficiency) similar to that observed by X-ray in vowel production (Fant, 1960);
- a structuring of the tube into specific regions ("distinctive regions") corresponding to the main places of articulation observed in speech production (Mrayati, et al., 1988);
- a dynamic structuring of the tube by a finite number of specific deformations of its shape, leading to a structuring of the acoustic plane into a finite number of trajectories.

Figure 13.

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

Full use of the three main different “tongue gestures” and of the “lip gesture” (one of the three “tongue gestures” and/or one “lip gesture”) leads to the formant trajectories shown in Figure 13. These gestures and their associated acoustic trajectories are the dynamic coding units of our acoustic communication system. On each trajectory sub-units can be positioned, first at both ends (maximum acoustic contrast), then, in the middle (acoustically equidistant from both ends), then again in the middle but new sub-units acoustically too close to be discriminated are not retained. The sub-units obtained by deduction and their positions on the trajectories are similar to the vowels generally proposed by phoneticians and represented in the F1/F2 plane (see for example Catford (1988)). So, in the following, we denote them by their corresponding phonetic symbols. Figure 13 gives the possible basic vowels limited in number to eighteen. Maddieson (1984) observed a maximum of fifteen basic vowels. For more vowels, another gesture is needed (nasal, advanced tongue root (ATR), long/short vowels…).

The three main tongue gestures lead to three acoustic trajectories (solid lines in Figure 13). Associated with the lip gesture, they lead to complementary trajectories (dotted lines in Figure 13) increasing the acoustic space. So, Figure 13 shows the phonetic capacity of the acoustic tube that results from efficient and simple deformations, i.e., from minimum gestural deformations sufficient to obtain maximum acoustic variations. The capacity is in terms of trajectories, not in terms of static positions. Each formant

ACCEPTED MANUSCRIPT

trajectory is obtained by means of only one or two co-produced parameters of the DRM model: one corresponding to the tongue constriction displacement (associated with cavities) and the other to the lip opening.

Here, let us recall the characteristics proposed above to define an efficient acoustic communication system, i.e.,: a) minimum deformation of the tube leading to maximum acoustic variation, b) maximum acoustic space used, c) small and fully used number of coding units, d) dynamic coding (variations are coded), e) use of coding units in parallel (co-produced units). The number of coding units is small (three main different "tongue gestures" and one "lip gesture", producing formant trajectories), and they can be used in parallel (co-production of the tongue and lip gestures). These acoustically efficient gestures lead to maximum acoustic variations (dynamic coding). One coding unit can be fully used in terms of number of sub-units (as we will see later).

To summarize, three main trajectories are produced by the DRM model [ai], [au], and [iu] with four complementary ones: [ay] labialized, [aɯ], [iɯ] non-labialized and [uy] fully labialized. Our approach emphasizes the structuring role of formant trajectories produced by two elementary gestures ("tongue gesture and lip gesture"). Vowels lying on the trajectories are consequences of these trajectories. This approach emphasizes the dynamic aspect of vowels, frequently described as static events.

3.5. Prediction of vowel systems from vowel trajectories

ACCEPTED MANUSCRIPT

The preceding trajectories deductively obtained can now be used to set up, from the simple to the complex, an acoustic communication system with coding elements (vowels). To select the coding elements, three criteria are taken into account: maximum acoustic dispersion, maximum use of each of the trajectories (i.e. maximum number of coding elements on one trajectory), and low complexity (i.e. low number of trajectories). Use of the criteria leads to compromises: maximum acoustic contrast and increasing the number of coding elements can lead to either an increase in the number of trajectories, or an increase in the number of coding elements on one trajectory. Increasing the number of trajectories can lead to a choice between different trajectories with more or less equally good acoustic dispersion. At this level our approach is not fully deductive: to expand the number of coding elements in a system, we cannot predict whether a system will prefer to increase the number of coding elements on one trajectory or to add a new trajectory (and, if the latter, which one). But, generally, once a new trajectory is chosen, then “maximum” use of the trajectory develops. Our approach proposing a general framework that may lead to several different possible solutions cannot be considered as a limitation of our deductive approach. On the contrary, this result could explain the diversity of the vowel systems. For example, from a specific system with n coding elements, two solutions with n+1 coding elements following two different branches of a tree could be proposed being more or less equivalent in terms of

ACCEPTED MANUSCRIPT

acoustic dispersion and complexity. Here, choices appear as a result of the deductive approach. This situation is consistent with classifications proposed from vowel inventories (Crothers, 1978; Schwartz, et al., 1997) as for example in Crother, Figure 10, (1978), with the two branches, one with /ɨ/ and the other without. A deductive approach without choice would lead to only one solution (no tree and branches) (for example (Lindblom, 1986)).

To predict vowel systems, from the simple to the complex, the [ai] trajectory can be the first best choice: the corresponding /ai/ phonological gesture to produce this trajectory involves only one phonetic gesture (the tongue gesture); then to get maximum acoustic contrast, the [au] trajectory can be the second best choice ([iy], involving only one phonetic gesture, could have been the second best choice but this solution is not retained because of too small acoustic contrast): the corresponding /au/ phonological gesture involves two phonetic co-produced gestures (the tongue and lip gestures). With these two main trajectories, the following systems can be successively obtained by adding vowels following the criterion of maximum acoustic contrast between one new vowel and the preceding ones (all of them lying on the trajectories):

[a, i, u], [a, i, u, ɛ,], [a, i, u, ɛ, ɔ], [a, i, u, ɛ, ɔ, e], [a, i, u, ɛ, ɔ, e, o].

ACCEPTED MANUSCRIPT

The three, five and seven vowel systems are acoustically well balanced and likely to be more frequent because more stable. A system is acoustically well balanced (Maddieson, 1984, p. 138) and stable if the basic trajectories at the origin of vowels are symmetrically and fully used. In the case of one trajectory (able to produce several vowels) used to produce only one vowel, then, either this trajectory will disappear because it is too costly to keep it, or more vowels on this trajectory will be produced – principle of maximum use of a feature (Ohala, 1979). In a five vowel system, the vowels [ɛ] and [ɔ] are placed in the acoustic middle of the [ai] and [au] trajectories. Then for the seven vowel system, [e] and [o] are placed in the acoustic middle of the [ɛi] and [ɔu]. Vowels close to [a] are not chosen because their relative acoustic proximity makes them difficult to discriminate.

But, before extending the number of vowels on [ai] and [au], the trajectory [iu] could have been retained leading to a new class of vowel systems with the central vowel [ɨ] situated in the acoustic middle of [iu]:

[a, i, u], [a, i, u, ɨ], [a, i, u, ɨ, ɛ], [a, i, u, ɨ, ɛ, ɔ], [a, i, u, ɨ, ɛ, ɔ, e],

[a, i, u, ɨ, ɛ, ɔ, e, o]•

In this case, the six vowel system is acoustically well balanced.

ACCEPTED MANUSCRIPT

Then, to follow the criterion of maximum acoustic contrast, a new gesture may be needed to increase the number of vowels (compromise in complexity between more vowels on one trajectory and adding a new trajectory). With a labial gesture co-produced with the /ai/ gesture, [ay] is obtained. The complementary systems are obtained:

[a, i, u, ɛ, ɔ, e, o, y], [a, i, u, ɛ, ɔ, e, o, y, œ], [a, i, u, ɛ, ɔ, e, o, y, œ, ø].

This last system uses all the possibilities of the labial gesture.

Instead of [ay], the [aɯ] trajectory could have been chosen.

Sociolinguistic studies may explain why such or such solution is retained (Labov, 1972).

With the basic tongue and lip gestures, other complementary gestures, such as nasal/non-nasal, ATR/non-ATR, and so on, can be added leading to complementary sets of vowels.

The systems obtained according to our deductive approach are similar to the more frequent systems observed in the Crothers' inventory (1978) (also used by Lindblom to test his predictions (Lindblom, 1986, p. 16)) and in UPSID (UCLA Phonological Segment Inventory Database) (Maddieson, 1984) (used by Schwartz et al. (Schwartz*, et al.*, 1997, p.273) to test their predictions). In the Crothers' inventory (209 languages),

ACCEPTED MANUSCRIPT

there can be observed (Figure 14): a) two main classes of vowel systems: one without [ɨ], and one with [ɨ], this vowel appearing after the three vowel /a, i, u/ system; b) one system with n+1 vowels is obtained from the preceding one with n vowels by addition of a new vowel, c) the acoustically well balanced systems are the most common (the five vowel system in systems without [ɨ] and the six vowel system with [ɨ]), d) a trend to full use of the possible vowels on a trajectory. The same results can be observed in UPSID. With or without [ɨ], our predictions are good.

Figure 14.

4. Discussion

In the past, several accounts have been proposed to predict and so to explain vowel systems. These approaches are deductive (Liljencrants and Lindblom, 1972; Lindblom, 1986; ten Bosch*, et al.*, 1986; ten Bosch, 1991; Schwartz, et al., 1997; de Boer, 1997; de Boer, 1999; Diehl*, et al.*, 2003; Oudeyer, 2005): they explain vowel systems not from the formal properties of, for example, distinctive feature theory, but from substantive, functional properties of production and perception systems. Researchers typically first define a maximum acoustic space in the F1/F2 plane from an articulatory model giving

ACCEPTED MANUSCRIPT

rise to the vowel triangle, then they use a perceptual criterion (maximum perceptual dispersion, for example) to predict the distribution of vowels within this space. Among the discrepancies between predicted and observed systems (reduced to a minimum in (Diehl, et al., 2003)), are the difficulties in predicting central vowels, and the excess of vowels predicted on the [iu] axis. They also fail in the prediction of several possible different systems: for a given number of vowels in a system, only one solution is generally proposed.

The main comments concerning these different approaches to prediction of vowel systems are the following:

- Taking into account only the maximum acoustic space to predict vowel systems does not seem to be sufficient because then all points in the vowel triangle would have the same status and would be reached with the same ease. Lindblom (1986, p.36) noted commenting on his own approach to deriving vowel systems, that "One of the more striking features of language, including its phonological aspect, is its structuring in terms of *discrete* and hierarchically organized units. We have nevertheless made the present predictions with the aid of a *continuous* space". Each point in the vowel triangle can, in fact, be reached by the articulatory machinery, but the transition from one vowel to another, i.e., the dynamics, results from strategies favoured by geometrical and acoustical

ACCEPTED MANUSCRIPT

properties of the tube. For example, the lip gesture, closing/opening of the tube at one end (corresponding to the labial/non-labial contrast), is easily realized and is used in various contexts (note that in a continuous space, the vowel obtained cannot be characterized as labial or non-labial).

- The lip gesture can be co-produced with the tongue gesture corresponding to the [ai] trajectory in the F1/F2 plane, leading to [ay]. But [y] is not situated at the acoustic or perceptual midpoint of the [iu] trajectory: it is close to [i]. The maximum acoustic dispersion criterion within an unstructured vowel triangle predicts a new vowel between [i] and [u] which is always situated at the acoustic or perceptual midpoint of the [iu] trajectory.
- Two central vowels can have more or less the same F-pattern (for example /ɯ/ and /ø/) and quite different production characteristics: one is obtained by a central tongue constriction without labialization, the other by a front tongue constriction with labialization. Even if their static F-patterns are more or less the same, the formant trajectories in the F1-F2 plane, from other vowels to these central vowels, are different (Carré and Mrayati, 1995).
- The hierarchy of vowel complexity cannot be accounted for with only a perceptual approach (even if it can be assumed that the perceptual system is

ACCEPTED MANUSCRIPT

well adapted to perceive characteristics of the acoustic tube). In a perceptual approach, the 'complexity' of the systems is only proportional to the number of vowels. But, the more gestures involved in producing a vowel, the more complex it is to produce (for example [ay], two gestures, is more complex to produce than [ai], one gesture) (Lindblom, 1990).

According to the present account, the acoustic properties of the tube structure and dynamically 'discretize' the F1/F2 plane. The more vowels we need in a system, the more gestures are needed, thus increasing the complexity of the system. And a new gesture (chosen among others) must be fully used to set up an efficient system. A well balanced system is more stable and should be more frequent. Maximum utilization of gestures corresponds to the "maximum utilization of the available distinctive features" of Ohala (1979, p. 185), as observed by Clements (2003)). The use of 'gestures' as deformations of the area function in production instead of 'features' is an answer to Lindblom (1986, p. 41): "A major difficulty, though, is to give a substantive, *deductive account* of the features".

Our deductive evolutionary approach leads to vocalic trajectories structuring the acoustic space with corresponding specific gestural deformations of the vocal tract. Following our approach, vowels as stable states cannot be the goals of the evolution process because they are unknown at the beginning. If the human communication

ACCEPTED MANUSCRIPT

system is indeed the result of an evolutionary process, then the static nature of vowels can be discussed. The quantal nature of vowels (Stevens, 1972; Stevens, 1989) must therefore be redefined: vowels are quantal only at the borders of the vowel triangle as observed in Figure 4), and the acoustic space is quantized into a finite number of quantal vocalic trajectories corresponding to efficient gestural deformations of the vocal tract. The relations between these efficient "articulatory" deformations and corresponding "acoustic" trajectories are specific examples of the famous curve (region II) proposed by Stevens (1972). Regions I and III are not useful: in these regions, "articulatory" deformations do not lead to acoustic variation and so to acoustic information.

Our first results deductively obtained, matching with observed data lead us to formulate several hypotheses as perspectives to be studied.

5. Perspectives

Applied to an acoustic tube, our deductive evolutionary approach leads to a maximum acoustic space similar to the vowel triangle and to distinctive deformation gestures that structure the tube into regions and the acoustic space into vocalic trajectories. Vowel systems can then be correctly predicted. The good match between observation and deductive prediction encourages us to make further predictions, formulating a unified view of vowel and consonant production and reconsidering the relation between phonetics and phonology.

ACCEPTED MANUSCRIPT

(i) First, recall Ohala (1979, p185) comment on the deductive approach to predicting vowel systems: "…It would be most satisfying if we could apply the same principles to predict the arrangement of consonants, i.e., to posit an acoustic-auditory space and show how the consonants position themselves so as to maximize the inter-consonantal distance…". In fact, the loci of the regions deductively obtained and represented by the closed-open DRM model taking into account F1, F2, and F3 correspond to the standard places of articulation for plosive consonants (Mrayati, et al., 1988; Carré and Chennoukh, 1995; Carré and Mody, 1997). All the VCV studied by Öhman (1966) were reproduced by means of the DRM model, the Region R8 (see Figure 8a) being used to produce the [b] closure, the region R6 to produce the [d] closure and the region R5 to produce the [g] closure (Carré and Chennoukh, 1995). Moreover, good simulations of the locus equation characteristics were obtained with the DRM model (Chennoukh*, et al.*, 1997; Carré, 1998).

(ii) Towards the same theoretical framework for vowel and plosive consonant production.

ACCEPTED MANUSCRIPT

a) The DRM model, deductively obtained, structures the acoustic tube into regions that correspond to standard places of articulation in both vowels and plosive consonants. It produces vocalic trajectories in the linear part of the relation between area function and acoustic form, and consonants both in the linear part and in those nonlinear parts corresponding to occlusions. The vocalic gestures are obtained by the 4 region DRM model (taking into account the first two formants) within linear relations (area between about 0.5 and 10 cm2); the plosive consonant gestures are obtained by the 8 region DRM model (taking into account the first three formants) (Mrayati, et al., 1988; Carré and Chennoukh, 1995; Carré and Mody, 1997) within nonlinear and linear relations (closing-opening area between 0 and value from .5 to 10 cm2). Thus, places of articulation of both vowels and plosive consonants can be derived from the same theoretical approach. They are deduced from acoustic characteristics of the tube, reflecting its inherent phonetic properties. So both vowels and plosive consonants can be obtained from specific gestural deformations of the tube leading to specific formant trajectories in the acoustic plane.

ACCEPTED MANUSCRIPT

b) How should we describe these deformations and formant trajectories? First, besides static characteristics, by their directions in articulatory space (structured in terms of distinctive regions) or by formant trajectories in the corresponding acoustic space; second, by their "strength" or rate of deformation or of formant frequency change over time (the rates of both F1 and F2 give the direction in the F1/F2 plane). High rate leading to fast transitions in the time domain would correspond to consonant production, intermediate rate to diphthongs, low rate to vowel-to-vowel. Different degrees of low rate could lead to different degrees of constriction corresponding to different vowels. No static prototypical targets would be necessary to describe vowels. This description of gestures in terms of direction and rate of formant frequency change may be more or less invariant in vowel reduction (Lindblom, 1963), consonant reduction (Duez, 1995), and hyper or hypo speech (Lindblom, 1990). If it is well known that transitions are essential to characterize plosive consonants (Delattre*, et al.*, 1955; Dorman*, et al.*, 1977; Kewley-Port, 1982), it is also admitted that transitions bear information in the case of vowel perception (Lindblom and Studdert-Kennedy, 1967; Strange*, et al.*, 1976; Verbrugge and

ACCEPTED MANUSCRIPT

Rakerd, 1980; Nearey and Assmann, 1986; Nearey, 1989; Strange, 1989; Strange, 1989; Di Benedetto, 1989; Di Benedetto, 1989). For example, experiments have shown the importance of initial and final transitions in CVC syllables for vowel identification (silent center experiments, (Strange*, et al.*, 1983; Strange, 1989)); they contribute to the debate on how to characterize vowel transitions: the dual target hypothesis (initial target plus final target), the initial target plus slope hypothesis, and the initial target plus direction hypothesis (Nearey and Assmann, 1986; Pols and van Son, 1993).

Our own hypothesis is that direction of the trajectory in the F1-F2 plane, from the acoustic starting point which has to be known, and rate of formant frequency change can suffice for vowel identification. Consider, for example, the production of [ae], starting from [a]. At the very beginning of the transition, the direction of the trajectory suggests that the following vowel is situated on the [ai] trajectory and the rate of formant frequency change depends on whether final vowel is [ɛ], [e] or [i]. This assertion is supported by the constant duration of the transition observed by Kent (1969) and Gay (1978). This means that the vowel

ACCEPTED MANUSCRIPT

can be identified at the very beginning of the transition (and also throughout the length of the transition). This approach which could readily accommodate speaker variations, vowel reduction, noisy environments, and so on, must be experimentally tested.

First results on V1V2 transition characteristics produced and perceived by subjects were presented in Carré (2007). In [aV] production, measurements of the F1 and F2 transition rates were represented in the F1 rate/F2 rate plane. V can be discriminated. In perception, direction and rate of synthesized transitions were studied for transitions situated outside the traditional F1/F2 vowel triangle. This situation enables the study of transitions characterized only by their directions and rates independent of any vowel targets in the vowel triangle. These transitions are perceived as different V1V2 vocalic trajectories according to their directions in the acoustic plane and their rates.

c) Normalization. Considerable speaker variability is observed in vowel production (Peterson and Barney, 1952), especially if vowel target formant frequencies of male adults and children are compared. Generally, for speech recognition purposes, speech parameters are

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

normalized according to the fundamental frequency, a bark scale representation,… (Syrdal and Gopal, 1986; Miller, 1989; Nearey, 1989; Johnson, 1997). Speech described dynamically, in terms of direction of the trajectory in the acoustic plane and rate of change along the trajectory, may be closer to invariance than a succession of static targets, so that normalization may not even be necessary (Verbrugge and Rakerd, 1980). It is also well known that speech recognition is improved with dynamic cepstral coefficients (Furui, 1986). Studies on this point are also needed.

d) The role of the neutral position in the F1/F2/F3 space (corresponding to the uniform configuration). Speech production being gestural deformation around a uniform ("undeformed") configuration (Mrayati, *et al.*, 1990), the neutral position could be a possible reference position. Vowels could then be specified in polar coordinates from the neutral. A negative angle indicates [ai] or [ay], a positive angle [au] or [aw]; in the former case, decreasing the length of the vector indicates labialization, in the latter, the opposite. For plosive consonants, place of articulation can be easily derived from the falling or rising formant frequency

ACCEPTED MANUSCRIPT

transitions from neutral (Carré, et al., 2002). This simple formant behavior around the neutral position must be studied for all vocalic contexts.

e) Complexity. The properties of the DRM model suggest a way to describe vowel and consonant complexity: increasing complexity means increasing the number of formants to be controlled and so the number of gestures (in increasing the number of regions) needed to control the formant frequencies (and, consequently, the number of possible vowels and consonants). In vowel production, F3 could be controlled (for example with an anti-symmetrical command of R5 and R6, see Figure 8a) to get a complementary set of vowels; in consonant production, F4 could be controlled (for example, by dividing the region R6, see Figure 8a, into two parts – one place of articulation -> two places) to get also a complementary set of consonants.

(iii) Phonology/phonetics relation. The good correspondence between predicted and observed vocalic systems, and between predicted and observed plosive consonant places of articulation in speech production leads us to discuss the symbol/area-function or phonology/phonetics relation. This relation is

ACCEPTED MANUSCRIPT

generally considered weak and mediated cognitively by some process of translation. O'Shaughnessy (1996, p) for example noted the "lack of a simple relationship between many phonemes and their acoustic realizations". Nearey (1997) considered speech as weakly constrained by characteristics of the production and perception apparatus. He developed a "double weak" concept. And Lindblom (1996, p. 1689) questioned the Liberman and Mattingly assertion concerning gestural invariance (1985, p22): "the gestures have a virtue that the acoustic cues lack: instances of a particular gesture always have certain topological properties not shared by any other gesture", and replied "the current evidence does not favour such a position…the prospect of finding articulatory invariance, or of showing that articulatory representations are richer and more distinctive than acoustic patterns, appears utterly remote". He also questioned Fowler's claim (1986) asserting that "Both the phonetically structured vocal-tract activity and the linguistic information… are directly perceived (by hypothesis) by the extraction of invariant information from the acoustic signal… the signal is transparent to the phonetic segments…".

ACCEPTED MANUSCRIPT

The present approach is compatible with Fowler's claim, because it deduces deformation gestures from inherent acoustic properties of the tube. These gestures can be directly linked to the phonetic code. They can be considered as symbolic primitives of the system, vowels and consonants being products of the gestures. They are similar to those derived from data by Browman and Goldstein in articulatory phonology (1992). Thus, as reasonably proposed by Fowler et al. (1980), there is no mediating translation from symbol to articulation and acoustics. If the speech communication system is explained (as far as formant frequencies are concerned) by physical properties of the acoustic tube, then this system could be supported by the concept of a "single-strong theory" (around the acoustic tube) instead of accepting the premises of a "double-weak theory" (Nearey, 1997).

Our algorithm (Carré, 2004) leads, step by step, to minimal deformations of an acoustic tube yielding maximal acoustic change as in an evolutionary process. If the human speech production system evolved for communication needs, each evolutionary step must have carried information on the goal. Recall that the goal in the algorithm is a dynamic goal, such as increasing F1 and decreasing F2 at the same time; the goal is not to reach a static target that only becomes known at the end of the process. Therefore, throughout the gestural transition, information on the goal can be available. These last

ACCEPTED MANUSCRIPT

comments may seem to invite discussion of how speech (and language) could have evolved (recall that, here, only speech sounds characterized by formant trajectories are taken into account). Many studies will have to be undertaken to gauge the validity of our deductions and to assess their possible importance for an evolutionary account of speech.

6. Conclusions

In the paper, we tried to explore the power of the deductive approach as a scientific methodology. From acoustic properties of a tube, with criteria such as the minimum of energy, we deduced that phonetic properties are inherent characteristics of the tube. We were able to automatically obtain the vowel triangle and specific formant trajectories where almost all the oral vowels can be placed. Then, a tentative explanation of vowel systems was proposed. Further research on hypotheses leading to a same theoretical framework for the vowel and consonant production and on the phonetics/phonology relation must be undertaken.

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

7. Acknowledgments

This research was supported by the French Ministère de la recherche: ACI “Systèmes complexes en SHS, 2003”. The author thanks Michael Studdert-Kennedy for his very helpful comments on an early draft and Pierre Divenyi, Björn Lindblom, Egidio Marsico François Pellegrino, Willy Serniclaes for stimulating discussions.

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

8. References

Badin, P. and Fant, G., 1984. Notes on the vocal tract computations. KTH, STL-QPSR 2-3, 53-107.

Boë, L. J., Perrier, P., Guérin, B. and Schwartz, J. L., 1989. Maximal vowel space. In: Proc. of the First European Conf. on Speech Communication and Technology, Paris, 2, pp. 281-284.

Browman, C. P. and Goldstein, L., 1992. Articulatory phonology: An overview. Phonetica 49, 155-180.

Carré, R., 1998. Linear correlates in the speech signal: Consequences of the specific use of an acoustic tube? Behavioral and Brain Sciences 21, 261-262.

Carré, R., 2004. From acoustic tube to speech production. Speech Communication 42, 227-240.

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

Carré, R. and Chennoukh, S., 1995. Vowel-consonant-vowel modeling by superposition of consonant closure on vowel-to-vowel gesture. J. of Phonetics 23, 231-241.

Carré, R. and Divenyi, P., 2000. Modeling and perception of "gesture reduction". Phonetica 57, 152-169.

Carré, R., Liénard, J. S., Marsico, E. and Serniclaes, W., 2002. On the role of the “schwa" in the perception of plosive consonants. In: Proc. of the Int. Conf. on Speech and Language Processing, Denver, pp. 1681-1684.

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

Carré, R., Lindblom, B. and MacNeilage, P., 1994. Acoustic contrast and the origin of the human vowel space. J. Acoust. Soc. Am. 95, S2924.

Carré, R., Lindblom, B. and MacNeilage, P., 1995. Rôle de l'acoustique dans l'évolution du conduit vocal humain (Acoustic factor in the evolution of the human vocal tract). Comptes Rendus de l'Académie des Sciences, Paris t. 30, série IIb, 471-476.

Carré, R. and Mody, M., 1997. Prediction of Vowel and Consonant Place of Articulation. In: Proceeding of the Third Meeting of the ACL Special Interest Group in Computational Phonology, SIGPHON 97, Madrid, pp. 26-32.

Carré, R. and Mrayati, M., 1990. Articulatory-acoustic-phonetic relations and modeling, regions and modes. In A. Marchal and W. J. Hardcastle, (Eds.), *Speech Production and Speech Modelling*, NATO ASI Series, Kluwer Academic Publishers, Dordrecht.

Carré, R. and Mrayati, M., 1991. Vowel-vowel trajectories and region modeling. J. of Phonetics 19, 433-443.

Carré, R. and Mrayati, M., 1992. Distinctive regions in acoustic tubes. Speech production modeling. J. d'Acoustique 5, 141-159.

Carré, R. and Mrayati, M., 1995. Vowel transitions, vowel systems, and the Distinctive Region Model. In C. Sorin, J. Mariani, H. Méloni and J. Schoetgen, (Eds.), *Levels in Speech Communication: Relations and Interactions*, Elsevier, Amsterdam, pp. 73-89.

Carré, R., Pellegrino, F. and Divenyi, P., 2007. Speech dynamics: epistemological aspects. In: Proc. of the ICPhS, Saarbrücken, pp. 569-572.

Catford, J. C., 1988. A practical introduction to phonetics. Clarendon Press, Oxford.

Chennoukh, S., Carré, R. and Lindblom, B., 1997. Locus equations in the light of articulatory modeling. J. Acoust. Soc. Am. 102, 2380-2389.

Chomsky, N. and Halle, M., 1968. The sound pattern of English. Harper & Row, New York.

ACCEPTED MANUSCRIPT

Clements, G. N., 2003. Feature economy as a phonological universal. In: 15th ICPhS, Barcelona, pp. 371-374.

Crothers, J., 1978. Typology and universals of vowel systems. In J. H. Greenberg, C. A. Ferguson and E. A. Moravcsik, (Eds.), *Universals of human language. Vol. 2: Phonology*, Stanford University Press, Stanford, pp. 93-152.

de Boer, B., 1997. Emergent vowel systems in a population of agents. In: Proceedings ECAL 97.

de Boer, B. (1999). Self-organized in vowel systems. PhD, University of Brussels (VUB), Brussels.

Delattre, P. C., Liberman, A. M. and Cooper, F. S., 1955. Acoustic loci and transitional cues for consonants. J. Acoust. Soc. Am. 27, 769-773.

Di Benedetto, M. G., 1989. Frequency and time variations of the first formant: Properties relevant to the perception of vowel height. J. Acoust. Soc. Am. 86, 67-77.

Di Benedetto, M. G., 1989. Vowel representation: Some observations on temporal and spectral properties of the first formant frequency. J. Acoust. Soc. Am. 86, 55-66.

Diehl, R. L., Lindblom, B. and Creeger, C. P., 2003. Increasing realism of auditory representations yields further insights into vowel phonetics. In: Proc. of the 15th ICPhS, Barcelona, pp. 1381-1384.

Dorman, M. F., Studdert-Kennedy, M. and Raphael, L. J., 1977. Stop-consonant recognition: Release bursts and formant transitions as functionally equivalent, context-dependent cues. Perception & Psychophysics 22, 109-122.

Duez, D., 1995. On spontaneous French speech aspects of the reduction and contextual assimilation of voiced stops. J. of Phonetics 23, 407-427.

Fant, G., 1960. Acoustic theory of speech production. Mouton, The Hague.

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

Fant, G. and Pauli, S., 1974. Spatial characteristics of vocal tract resonance modes. In: Proceedings of the Speech Communication Seminar, Stockholm, pp. 121-132.

Fowler, C., 1986. An event approach to the study of speech perception from a direct-realist perspective. J. of Phonetics 14, 3-28.

Fowler, C. A., Rubin, P., Remez, R. and Turvey, M. T., 1980. Implications for speech production of the general theory of action. In B. Butterworth, (Eds.), *Speech Production I: Speech and talk*, Academic Press, London, pp. 373-420.

Furui, S., 1986. On the role of spectral transition for speech recognition. J. Acoust. Soc. Am. 80, 1016-1025.

Gay, T., 1978. Effect of speaking rate on vowel formant movements. J. Acoust. Soc. Am. 63, 223-230.

Gunnilstam, O., 1974. The theory of local linearity. J. of Phonetics 2, 91-108.

Johnson, K., 1997. Speaker perception without speaker normalization. An exemplar model. In K. Johnson and J. W. Mullennix, (Eds.), *Talker Variability in Speech Processing*, Academic Press, New York, pp. 145-165.

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

Kewley-Port, D., 1982. Measurement of formant transitions in naturally produced stop consonant-vowel syllables. J. Acoust. Soc. Am. 73, 379-389.

Labov, W., 1972. Sociolinguistics Patterns. University of Pennsylvania, Philadelphia.

Ladefoged, P., 1990. Some reflections on the IPA. J. of Phonetics 18, 335-346.

Liberman, A. M. and Mattingly, I. G., 1985. The motor theory of speech perception revisited. Cognition 21, 1-36.

Liljencrants, J. and Lindblom, B., 1972. Numerical simulation of vowel quality systems: The role of perceptual contrast. Language 48, 839-862.

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

Lindblom, B. (1963) "On vowel reduction," Report N° 29. Stockholm: The Royal Institute of Technology, Speech Transmission Laboratory.

Lindblom, B., 1986. Phonetic Universal in Vowel Systems. In J. J. Ohala and J. J. Jaeger, (Eds.), *Experimental Phonology*, Academic Press, Orlando, pp. 13-43.

Lindblom, B., 1990. Explaining phonetic variation: a sketch of the H and H theory. In A. Marchal and W. J. Hardcastle, (Eds.), *Speech Production and Speech Modelling, NATO ASI Series*, Kluwer Academic Publishers, Dordrecht, pp. 403-439.

Lindblom, B., 1990. On the Notion of "Possible Speech Sound". Journal of Phonetics 18, 135-152.

Lindblom, B., 1990. Phonetic contents in phonology. Phonetic Experimental Research at the Institute of Linguistics, University of Stockholm (PERILUS) XI, 101-118.

Lindblom, B., 1996. Role of articulation in speech perception: Clues from production. J. Acoust. Soc. Am. 99, 1683-1692.

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

Maddieson, I., 1984. Patterns of sounds. Cambridge University Press, Cambridge.

Maeda, S., 1982. A digital simulation method of vocal tract system. Speech Communication 1, 199-229.

Miller, J. D., 1989. Auditory-perceptual interpretation of the vowel. J. Acoust. Soc. Am. 85, 2114-2134.

Mrayati, M., Carré, R. and Guérin, B., 1988. Distinctive regions and modes: A new theory of speech production. Speech Communication 7, 257-286.

Mrayati, M., Carré, R. and Guérin, B., 1990. Distinctive regions and modes: articulatory-acoustic-phonetic aspects. A reply to Boë and Perrier comments. Speech Communication 9, 231-238.

ACCEPTED MANUSCRIPT

Nearey, T. and Assmann, P., 1986. Modeling the role of inherent spectral change in vowel identification. J. Acoust. Soc. Am. 80, 1297-1308.

Nearey, T. M., 1989. Static, dynamic, and relational properties in vowel perception. J. Acoust. Soc. Am. 85, 2088-2113.

Nearey, T. M., 1997. Speech perception as pattern recognition. J. Acoust. Soc. Am. 101, 3241-3254.

O'Shaughnessy, D., 1996. Critique: Speech perception: Acoustic or articulatory. J. Acoust. Soc. Am. 99, 1726-1729.

Ohala, J., 1979. Moderator introduction to symposium on phonetic universals in phonological systems and their explanations. In: Proceedings of the IXth Int. Congress of Phonetic Sciences, Copenhagen, Vol. III, pp. 181-185.

Öhman, S., 1966. Coarticulation in VCV utterances: spectrographic measurements. J. Acoust. Soc. Am. 39, 151-168.

Oudeyer, P.-Y., 2005. The Self-Organization of Speech Sounds. Journal of Theoretical Biology 233, 435-449.

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

Pols, L. C. W. and van Son, R. J., 1993. Acoustics and perception of dynamic vowel segments. Speech Communication 13, 135-147.

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

Schwartz, J. L., Boë, L. J., Vallée, N. and Abry, C., 1997. Major trends in vowel system inventories. Journal of Phonetics 25, 233-253.

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

Stevens, K. N., 1972. The quantal nature of speech: evidence from articulatory-acoustic data. In E. E. David and P. B. Denes, (Eds.), *Human Communication: a unified view*, Mac Graw-Hill, New York, pp. 51-66.

Stevens, K. N., 1989. On the Quantal Nature of Speech. J of Phonetics 17, 3-45.

Strange, W., 1989. Dynamic specification of coarticulated vowels spoken in sentence context. J. Acoust. Soc. Am. 85, 2135-2153.

Strange, W., 1989. Evolving theories of vowel perception. J. Acoust. Soc. Am. 85, 2081-2087.

Strange, W., Jenkins, J. J. and Johnson, T. L., 1983. Dynamic specification of coarticulated vowel. J. Acoust. Soc. Am. 74, 695-705.

Strange, W., Verbrugge, R. R., Shankweiller, D. P. and Edman, T. R., 1976. Consonant environment specifies vowel identity. J. Acoust. Soc. Am. 60, 213-224.

Studdert-Kennedy, M., 1987. The phoneme as a perceptuomotor structure. In A. Allport, D. G. Mackay, W. Prinz and E. Scheerer, (Eds.), *Language Perception and Production: Relationships between Listening, Speaking, Reading and Writing*, Academic Press, London, pp. 67-84.

Syrdal, A. K. and Gopal, H. S., 1986. A perceptual model of vowel recognition based on the auditory representation of American English vowels. J. Acoust. Soc. Am. 79, 1086-1100.

ten Bosch, L. (1991). On the structure of vowel systems; aspects of an extended vowel model using effort and contrast. Dissertation Thesis, University of Amsterdam, Amsterdam.

ten Bosch, L. F. M., Bonder, L. J. and Pols, L. C. W., 1986. Static and dynamic structure of vowel systems. In: Proceedings of the XIth International Congress on Phonetic Sciences, Tallinn, pp. 235-238.

ACCEPTED MANUSCRIPT

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

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

Figure captions

Figure 1. Overview of the deductive approach. The acoustic tube produces a signal with maximal acoustic contrast by minimal deformation of the area function (ΔF/ΔA, ratio of formant variation to corresponding area function variation is maximal – equivalent here to the minimum effort criterion). It is hypothesized that the speech production and perception system is efficient (as far as the formants are concerned) for acoustic communication, i.e., exploits efficiently the dynamic characteristics of the acoustic tube.

Figure 2. Overview of the recursive algorithm used to automatically find the maximum acoustic space with minimum deformation of the shape of the tube. The goal is to increase or decrease F1 or F2 (or both F1 and F2) by deforming, step by step, the shape of the tube. This deformation is efficient: minimum deformation leads to maximum formant variation. This operation is repeated until the intrinsic limits of the acoustic space are reached.

Figure 3. a) Deformations of the area function obtained by the algorithm for a decrease of F1 (k_1 = -0.702; k_1 is the coefficient of the sensitivity function S_iF1 in the formula

ACCEPTED MANUSCRIPT

(1)) and an increase of F2 (k_2 = +0.702; k_2 is the coefficient of the sensitivity function S_iF2 in the formula *(1)*). b) Corresponding formant trajectory in the F1/F2 plane.

Figure 4. Formant trajectories in the F1/F2 plane obtained by the algorithm for different proportions of k_1 and k_2 in the formula *(1)*. On the trajectories, the dots correspond to the results obtained by the different iterations. The trajectories cover the whole plane. An acoustic triangle appears limited by the linear parts of the trajectories. This triangle corresponds to the vowel triangle.

Figure 5. a) Deformation of the tube automatically obtained by the algorithm for an increase of F2 (k_1 = 0, k_2 = +1 in the formula *(1)*). The initial shape is a uniform closed-open tube. The area range is between 0.5 and 16cm2. The deformation is anti-symmetrical: two deformation gestures lead to two constrictions associated with two cavities. b) Corresponding trajectory in the F1-F2 plane.

Figure 6. a) Deformation of the tube automatically obtained by the algorithm for a decrease of F2 (k_1 = 0, k_2 = -1 in the formula *(1)*). The initial shape is a uniform closed-closed tube. The area range is between 0.5 and 16cm2. The deformation is symmetrical:

ACCEPTED MANUSCRIPT

one deformation gesture leads to one constriction associated with two cavities. b) Corresponding trajectory in the F1-F2 plane.

Figure 7. a) Deformation of the tube automatically obtained by the algorithm for a decrease of F1 associated with an increase of F2 ($k_1 = -1$, $k_2 = +1$ in the formula *(1)*) in order to obtain maximal acoustic variation. The initial shape is a schematic representation of the vowel [a]. The area range is between 0.5 and 16cm2. b) Corresponding trajectory in the F1-F2 plane.

Figure 8. a) The two black arrows correspond to the two command gestures of the closed-open DRM model: the "tongue gesture" and the "lip gesture" to control F1 and F2; b) Corresponding formant trajectories in the F1/F2 plane: the trajectory [1-2] is obtained with the "tongue gesture", the trajectory [1-3] is obtained with co-produced "tongue and lip gestures".

Figure 9. The trajectory [2-3] obtained by "lip" closure. The weak acoustic effect of the first part of the "lip gesture" is shown.

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

Figure 10. a) The closed-closed DRM model and the "lip" gesture from a "labialized" configuration (corresponding to the minimum possible of F1 and F2) to a "non-labialized" configuration. Three regions are observed: two cavities and one constriction (R3 and R6 are equal); b) Corresponding formant trajectory from [5] to [6].

Figure 11. Results obtained with the algorithm for different fixed lip openings (between 0.01 and 16cm2): a) Trajectories in the F1-F2 plane; b) Corresponding area functions after 10 iterations.

Figure 12. a) From back constriction (obtained with a closed-open DRM model) to central constriction (obtained with a closed-closed DRM model). The displacement of the constriction is longitudinal and coproduced (or not) with a lip closing gesture and b) corresponding [1-5] and [1-6] formant trajectories.

Figure 13. Trajectories obtained with the DRM model and the possible vowels. The dotted lines are labialized trajectories.

Figure 14. Vowel systems, with or without [ɨ], observed in the Crothers inventory (209 languages). For example, without [ɨ], 55 languages have a 5 vowel system.

ACCEPTED MANUSCRIPT

Figures

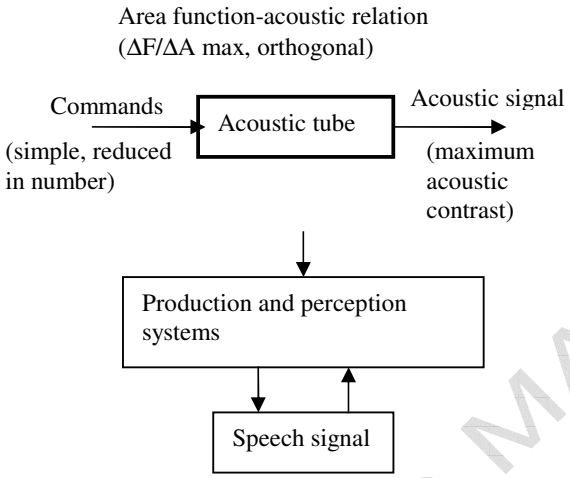


Figure 1.

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

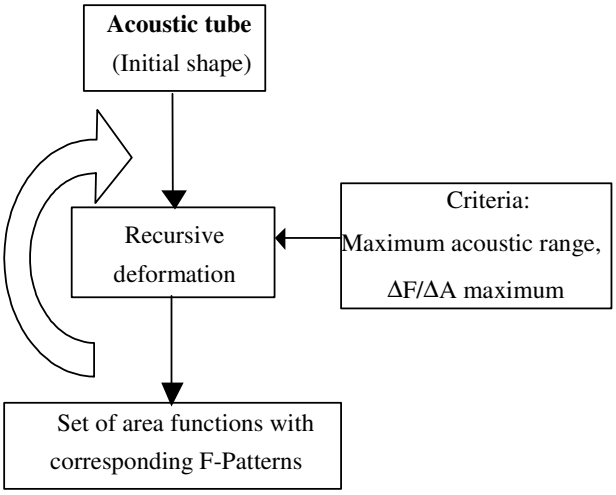
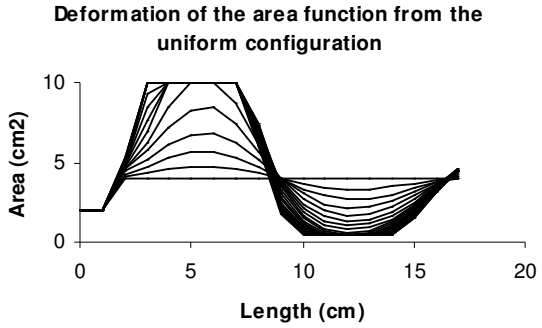


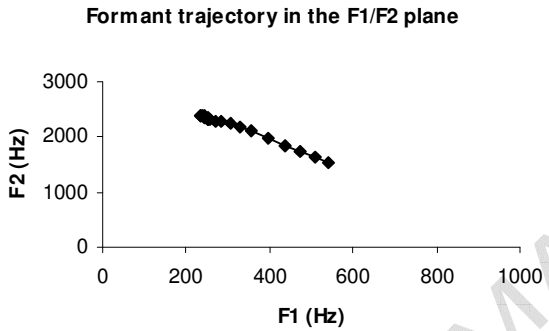
Figure 2.

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT



a)



b)

Figure 3.

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

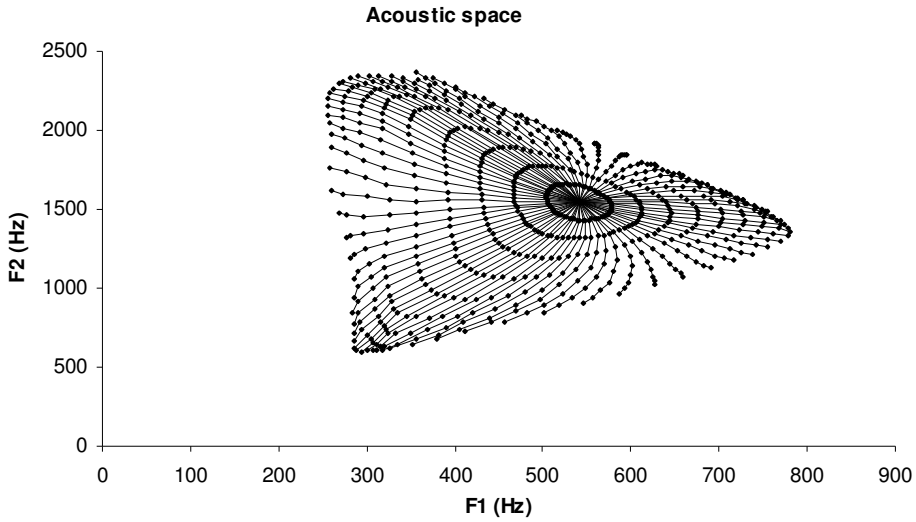
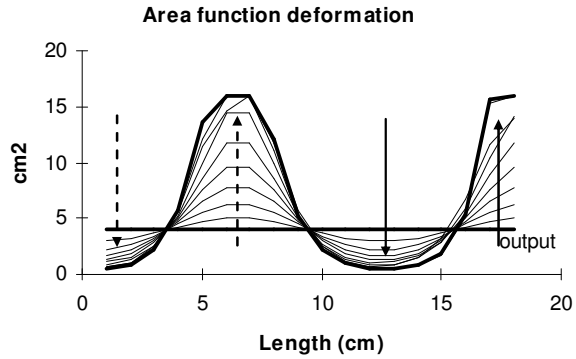


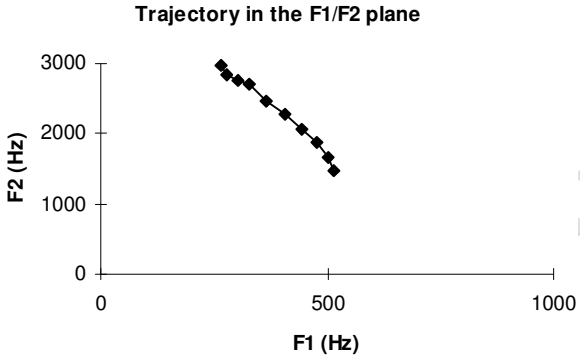
Figure 4.

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT



a)

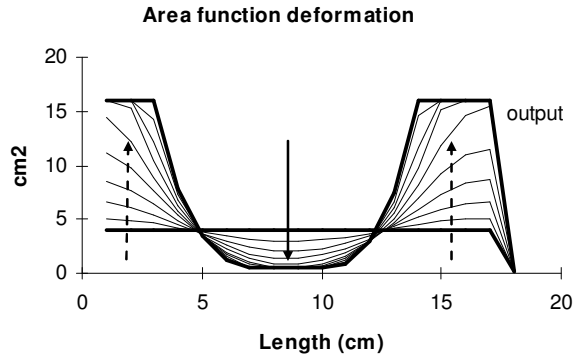


b)

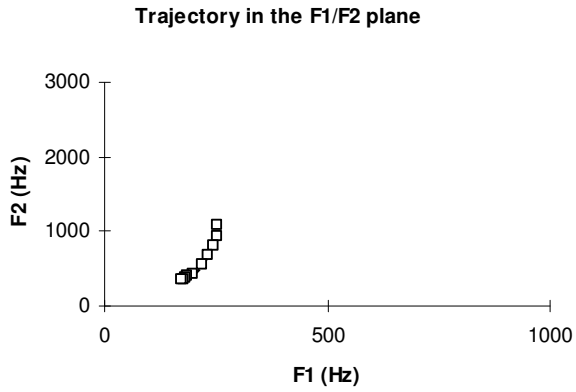
Figure 5a and 5b.

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT



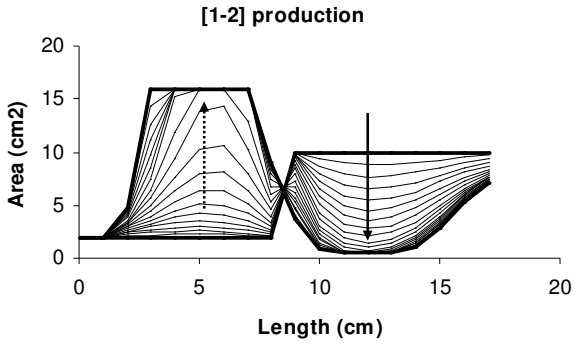
a)



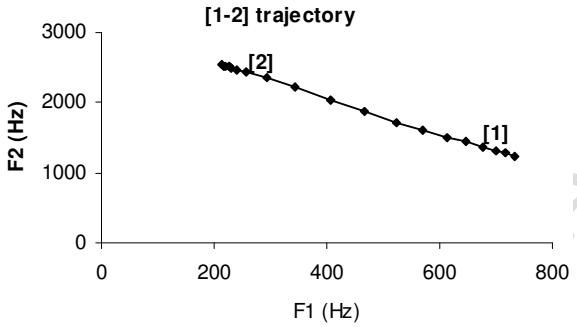
b)

Figure 6a and 6b.

ACCEPTED MANUSCRIPT



a)

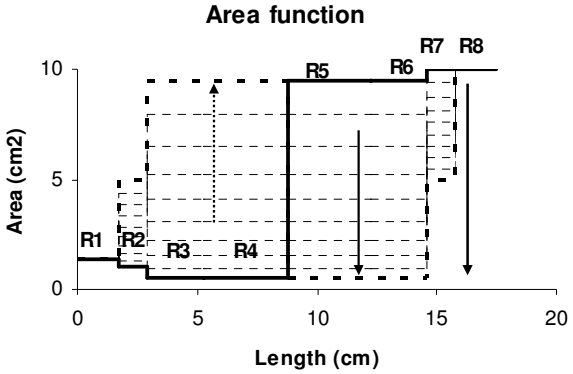


b)

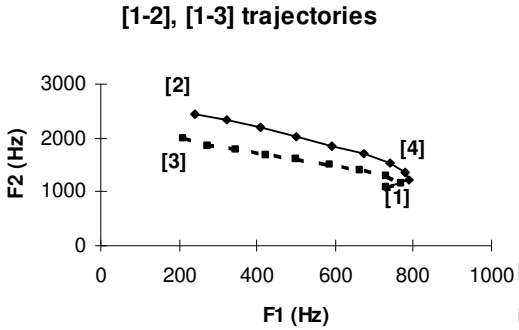
Figure 7a and 7b.

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT



a)



b)

Figure 8a and 8b.

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

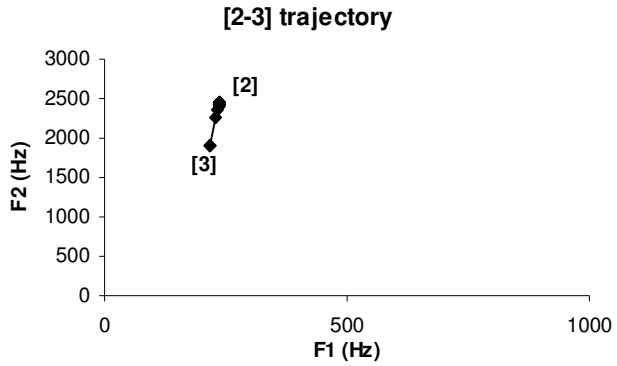
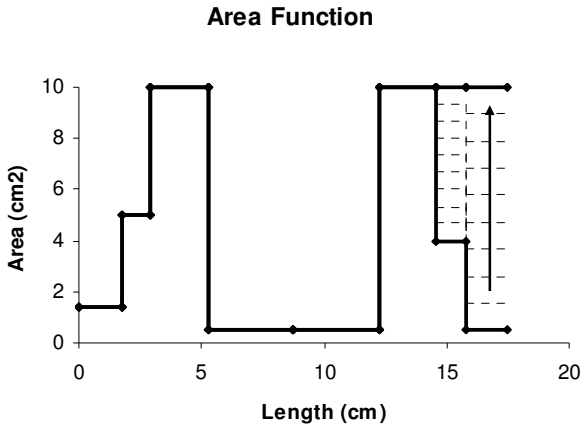


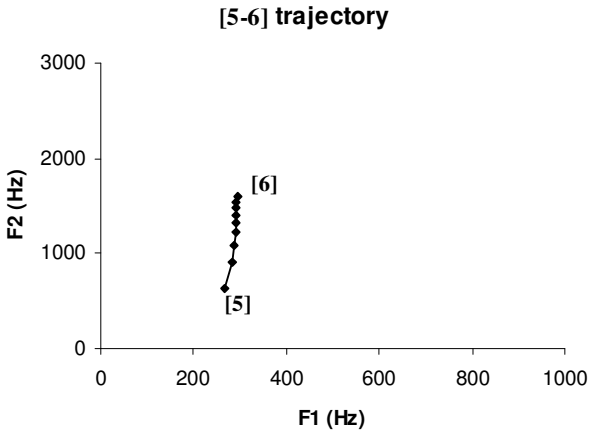
Figure 9.

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT



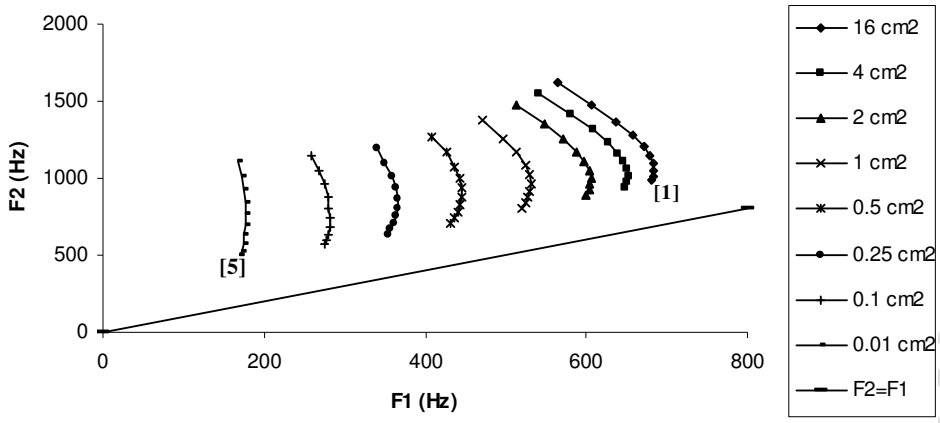
a)



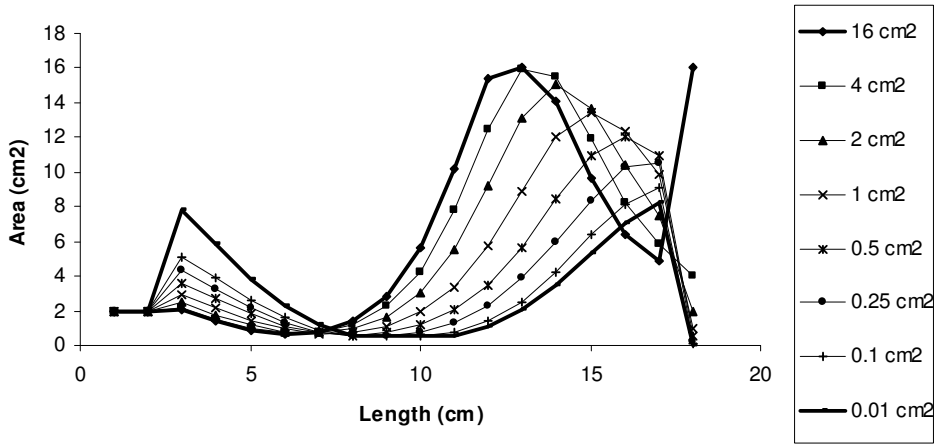
b)

Figure 10a and 10b.

ACCEPTED MANUSCRIPT



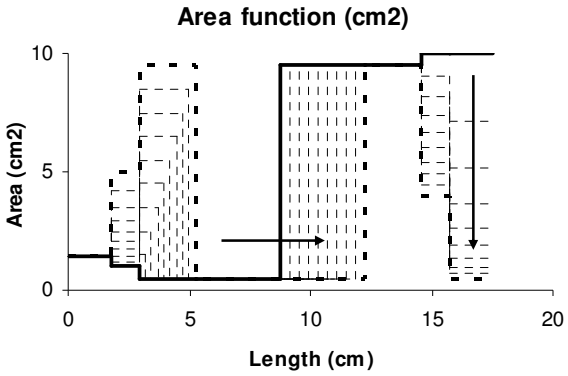
a)



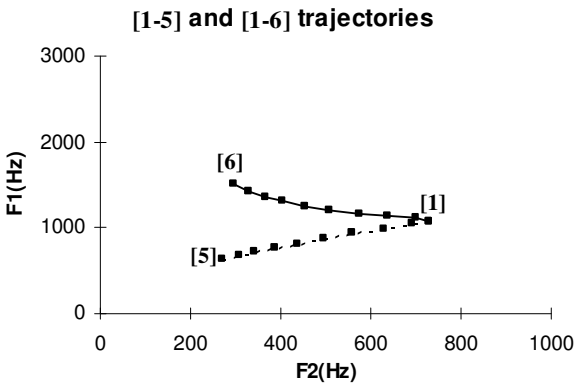
b)

Figure 11a and 11b.

ACCEPTED MANUSCRIPT



a)



b)

Figure 12a and 12b.

ACCEPTED MANUSCRIPT

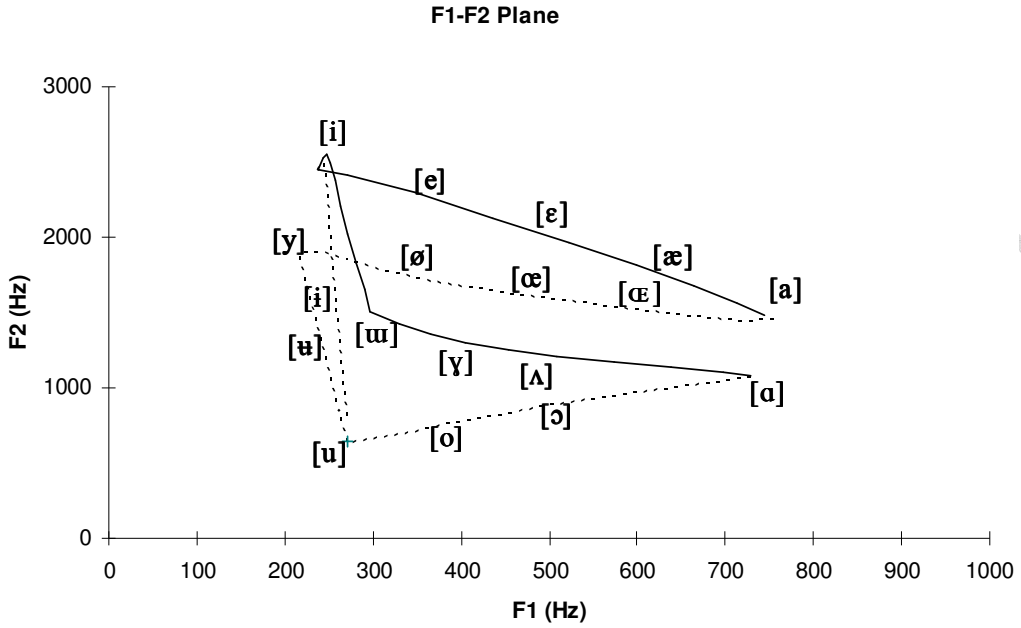
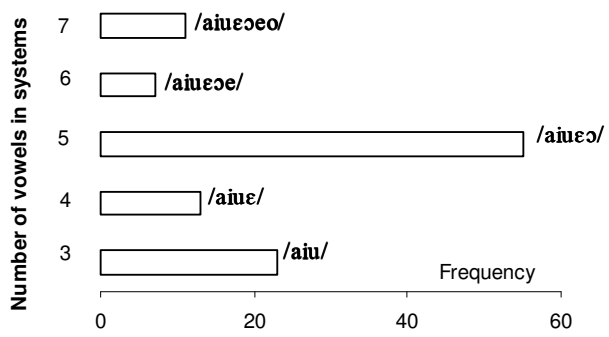
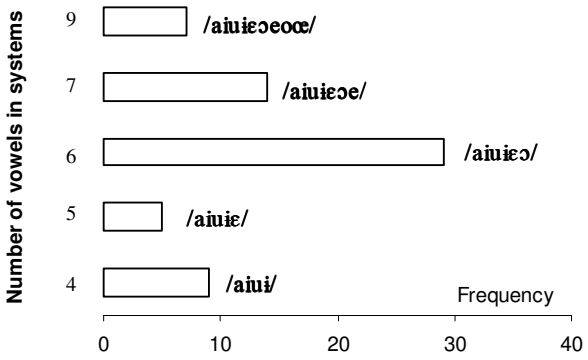


Figure 13.

ACCEPTED MANUSCRIPT



a) without [ɨ]



b) with [ɨ]

Figure 14a and 14b.