CINERADIOGRAPHY OF VCV SEQUENCES: ARTICULATORY-ACOUSTIC DATA FOR A SPEECH PRODUCTION MODEL

P. Badin¹, B. Gabioud², D. Beaufremp⁴, T. M. Lalouache¹, G. Baillé⁴, S. Maeda⁵, J. P. Zerling⁶, G. Brock⁴

¹ Inst. Communication Parlée, URA CNRS N°368, INPG, Univ. Stendhal, Grenoble, France
² Institut d’Informatique, Université de Lausanne, Switzerland
³ Département Signal, ENST, Paris, France
⁴ Institut de Phonétique de Strasbourg, Université de Sciences Humaines, France

Summary
At present, the need for articulatory data is motivated by the development of speech production models, in which a plant, i.e. realistic articulatory model, is driven by a controller, capable of generating trajectories to perform a linguistic task. Knowledge is still rarely needed in the domain of the plant. Thus, the present study aimed at obtaining vocal tract mid sagittal profiles and lip shapes for a subject uttering a corpus of [VCV] sequences of French voiced plosives and fricatives. The experimental set-up includes a 35 mm film camera for the vocal tract X-ray mid sagittal view, and a video camera for the lip front shape. This resulted in an articulatory-acoustic database containing, for each of the 1100 pictures, the complete sagittal profile, the description of its contour in terms of the intersections with a semi-polar grid, a number of articulatory measurements such as lip width or jaw position, and acoustic data such as hand-labeled formants. Particularly, the coherence between the articulatory and acoustic data has been verified for the complete database.

Introduction
In the last few years, the interest for speech production is increasing in the domains of inversion (Abry et al., 1994), articulatory synthesis (Cooker, 1994; Gabioud, 1994), coding or recognition (Rose et al., 1994). It seems clearly established that articulatory models are one of the most efficient means of manipulating voice shapes, and mid-sagittal profiles constitute, at present, the privileged interface between the motor control and acoustic modules of the speech production system. Complete mid-sagittal profile data associated with other relevant and complementary data are extremely scarce in the literature. The European collaborative project Speech Maps dealing with Sound-Image Inversion in Speech has thus acquired fresh mid-sagittal vocal tract profiles synchronised with lip video pictures for the reference subject PB, for whom other data such as acoustic and aerodynamic pressure and flow in the tract, electropalatography, acoustic transfer functions, lip geometric parameters, ... have already been recorded. This ensemble of data will be extremely useful for developing and improving a coherent model that has taken into account the acoustic, geometric and articulatory aspects of speech production.

Experimental set-up
The recordings have been performed at the Strasbourg Schilthiipt Hospital. The experimental set-up (see Fig. 1) was a combination of the classical set-up used by Bober et al. (1980), and of that developed by Lalouache (1990). The subject’s head was positioned at distances of 50 cm from the X-ray emitter and of 20 cm from the radiance amplifier. The vocal tract mid-sagittal images produced by the radiance amplifier were captured and recorded simultaneously by a 35 mm film camera and by a video camera. The subject’s lips, painted blue so as to allow the lip contours to be further extracted by an image processing procedure, were recorded by means of another video camera. All three cameras were operating at a rate of 50 images per second, with an exposure time of respectively 1 ms for the X-ray and of 2.5 ms for the video signal. The D signal, captured by a directive microphone placed at a distance of 10 cm from the subject’s mouth, was recorded on the left channels of the two video recorders and of a Digital Audio Tape (DAT) recorder. A counter, driven by impulses generated by the X-ray emitter, produced a digital display (optical coding) recorded simultaneously with the mid-sagittal images, and a synchronisation signal (electronic coding) recorded on the right channels of the video and DAT recorders. It was thus possible to synchronise sound, video images and 35 mm images.

Corpus
The corpus was designed as to include as many combinations of VCVs as possible, where V = [a, i, u, y] and C = [v, z, s, b, d, g], in a very limited amount of time. Six vocalic context were chosen for the consonants: aCa, aCi, aCa, aCi, iCa, iCy. Moreover, the fricative items were
preceded by a [p], in order to allow the estimation of subglottal pressure during the fricatives as the Intra-Oral Pressure during the closure of [p] (Rotenberg, 1982). In addition, a series of connected vowels [æ e i y u ø] was recorded in order to test formant/cavity affiliation hypotheses (cf. Ballay, 1993). Finally, 1100 pictures corresponding to 22 sec. of signal were obtained, with the following distribution of phonemes: 30 [a], 12 [u], 6 [y], 6 [v, z, s, b, d, g], 18 [p]. Each phoneme in the database was labelled at its centre, to help further processing.

Data pre-processing and articulatory parameter determination

For each picture, the sagittal contours were first drawn by hand from a projection of the picture onto a piece of paper in a dark room, and then digitised by a scanner. Finally 11 sub-contours were hand-edited by means of the interactive software BTsC (Maeda et al., 1993). These sub-contours correspond to different articulators or vocal tract regions: the upper and lower lips, the hard palate, the velum, the mobile lower lip, the tip of the tongue, the tongue root, the hyoid bone and the larynx (cf. Fig. 2). Note that the tip of the tongue and the sublingual cavity were paid particular attention to, as they play an important role in the production of consonants. Concerning the lips, the blue of the video front lip images was converted into absolute back by means of an analogue Kroma-key, and the inner contour was then automatically determined by thresholding (Lalouache, 1990).

![Fig 2: Various articulatory parameters](image)

![Fig 3: Lip height (B: solid; lips: dashed) for the sequence [paazupipip]![](image)

From these contours, a number of articulatory parameters was automatically determined (cf. Fig. 2): lip height (liphei, also B from front views), lip width (A, from front views), and the intra-labial lip area (S); upper (protip) and lower (protip) lip protrusions, upper lip elevation (lipsop), jaw height (jawhei), jaw advancement (jawadv), advancement of the tongue tip, tongue root height (trigho), height of the tongue (tongue), and velum height (velhei), height of the highest point of the tongue (tongue) and advancement of the backest point of the tongue (toxrig). The trajectories of the articulators are expected to be related to smooth, due to their long time response (e.g. the jaw has a characteristic resonance frequency of 5-6 Hz). The inspection of the measured parameters has revealed that the noise added by the whole chain of acquisition and processing to the smooth trajectories was in the range of 1 mm peak to peak. The distance between the upper and lower lips can be determined either from the midsagittal profile (lipsop) or from the front view (B): the two measures are very close to each other (cf. i.e. Fig. 3). liphei being less accurate when the lip opening is close to zero, particularly for the rounded vowels [u, y], due to the fact that the subject’s upper lip tend to hide the opening in such cases. A semi-polar grid is also used to describe the midsagittal profiles. Two parts of this grid are dynamically adjustable in order to follow the movements of the larynx and of the tongue tip. The intersections between the midsagittal contours and the grid were thus determined in order to provide the data needed to compute the midsagittal functions. These intersections will also be used to elaborate an articulatory model for this subject.

Measured and re-constructed formant spaces

The speech signal was digitised at 16 kHz, and the first four formants were determined by LPC, for each 20 ms window centred around the instants where the midsagittal views were taken. Because of the background noise due to the X-ray emitter machinery, the signal-to-noise ratio was not very high (about 25 dB), and thus some of the formants had to be hand-edited. This was done in reference with another version of the same corpus recorded by the subject in good recording conditions, thus allowing the resolution of the uncertainties. The maximal F1/F2 and F1/F3 spaces for the vowels and consonants pooled are shown in Fig. 4 (left).

In order to test the coherence of the measurements, formants have been computed for each midsagittal profile. The midsagittal function was first derived from the intersections with the grid by a method based on the decomposition of the vocal tract into quadrilateral sections. Two different methods were then used to determine the area functions A(x) using the relation A(x) = α(x) d(x) ln(x). In the first method the α coefficients are constant for each of seven regions in the vocal tract (Perrier et al., 1992). In the other method, the relation between α and the coordinate y along the vocal tract midline and the midsagittal distance d has been optimised for the same subject on a corpus of nine still X-ray pictures previously acquired (Beaumont et al., 1995). The maximal formant spaces obtained (cf. Fig. 4, middle and right) are comparable with the measured ones. However, the computed formant F1 of [a] is about 50 Hz too low for both methods. The F1 obtained in the [i, y] region with the Beaumont et al.’s method is about 200 Hz too high. The F3 of [i] is about 150 Hz too low with the Beaumont et al.’s method.

- 350 -
as the Intra-Oral e i u o a] was
0 1100 pictures
0 [a, i, i] 32 [u], 6 per processing.
re onto a piece of
the pharynx, the larynx, vocal cavity were
firming the lips, the
orna-key, and the

-351-
Conclusions and perspectives
We have built a database of articulatory and acoustic parameters measured for a subject uttering series of [VCV] sequences of French vowels, voiced plosives and fricatives. Beside the obvious interest of these articulatory data in the domain of speech production, a major advantage of this new database is the presence of reliable acoustic parameters coherent with the detailed articulatory description of the vocal tract. Indeed we have verified that the formants measured on the speech signal and those re-computed from the articulatory data are in fairly good agreement, at least for F1, F2 and F3, on a frame by frame basis. This coherence is an essential feature, if one attempts to study the articulatory-to-acoustic relationship at a static as well as at a cinematic level.
A first version of an articulatory model based on these data has been developed. This model allows to manipulate the mid sagittal profiles of the database in a compact manner with a reduced number of articulatory parameters. This model will serve as reference for further articulatory studies. Particularly, in the frame of the sound-to-gesture inversion, it will be possible to assess inversion strategies by comparing the results of inversion with the measured articulatory parameters. Further, it will be possible to generate new articulatory data by inversion from acoustic signals. The fact that the model is based on a single subject offers also the possibility to acquire supplementary articulatory data on that subject and to be sure that they are coherent with the model. Finally, the model, in conjunction with acoustic vocal tract simulations, offers an efficient basis to gain more insight in articulatory strategies and to develop articulatory synthesis.

Acknowledgements
This study has been partially funded by the European ESPRIT/BRA project Speech Maps. We are very much indebted to P. Simon (Strasbourg Phonetic Inst.), to Agnès (Strasbourg Schültze Hosp.), and to T. Guiraud-Marigny and C. Benoit (ICP, Grenoble), for their valuable help in the data gathering and processing.

References