SPEECH MAPS INTERACTIVE PLANT "SMIP"

L.J. Boë[1], B. Gabioud[2] and. P.Perrier[1]
[1]ICP URA CNRS n° 368 INPG/ENSERG Université Stendhal, France
[2]Institut d'Informatique, Lausanne, Suisse

ABSTRACT

The SMIP is an interactive and ergonomic software. On the basis of an articulatory model, this device delivers an output signal which is associated with geometric and acoustic informations. This articulatory model regenerates lip and vocal tract shapes, with seven articulatory parameters as input. By means of a set of coefficients, the midsagittal contour is converted into an area function with which the transfer function and/or formants of the vocal tract can be calculated. Reasonable quality sound is generated. In addition, the 37 vowel prototypes of the UPSID database are provided, and tools to compute macro-variations are implemented.

INTRODUCTION

In the frame of studies on the articulatory-acoustic relationship, it is of high interest to manipulate articulatory models which integrate morphological and articulatory constraints. Indeed, such anthropomorphic models offer the possibility to coherently vary area functions by modifying appropriate control parameters. It thus becomes possible to relate these geometric variations and associated formant changes to real behaviour in speech production systems. Such a model can be extensively used for designing vowel (Vallée, 1994) and syllable prototypes, for the prediction of speech sound systems, for articulatory-acoustic inversion, and for speech synthesis. The present paper presents briefly the different components of the SMIP, and describes in detail the various operations that can be performed with this software. The SMIP, i.e. the *Speech Maps Interactive Plant*, has been elaborated in the frame of the European ESPRIT/BR project Nr. 6975 *Speech Maps*, and constitutes the core of the *Articulotron* [1].

THE COMPONENTS

The SMIP is organised in four main modules: (1) vocal tract midsagital contours are delivered by the articulatory model from seven control parameters: lip height, lip protrusion, vertical position of the jaw, front-back position of the tongue body, vertical position of the tongue dorsum, position of the apex, and vertical position of the larynx; (2) from the midsagittal dimensions, the vocal tract area function is estimated with a set of coefficients derived from radiographic measurements; (3) vocal tract acoustic transfer functions and formants and bandwidths are calculated by means of an acoustic model; (4) sustained vowel sounds are computed using a cascade formant synthesiser excitated by an appropriately shaped glottal pulses.

Maeda's articulatory model

The core of the SMIP is Maeda's articulatory model [2] together with a variant proposed by Gabioud [3]. It was built from a thorough statistical analysis of 519 hand-drawn midsagittal contours sampled at a rate of 50 frames/sec, obtained from synchronized radiographic and front/profile labiographic films shot at the *Strasbourg Institute of Phonetics* for one subject (PB) uttering ten meaningful French sentences [4]. The midsagittal contours were described as the 28 x-y coordinates of the intersection points of the two contours describing the vocal tract with a semi-polar grid. Seven parameters enable to explain 88 % of the variance of the observed variance of tongue contours. The percentages of explanation of the variance for tongue shape and jaw are distributed as follows: 15 % for the vertical position of the jaw, 43 % for the tongue body displacement, 23 % for the tongue dorsum, 7 % for the apex position. A linear combination of the seven parameters enables the re-construction of the vocal tract midsagittal contour. The inner (intero-labial) and outer (external arches of the vermilion) contours of the lips seen from front are based on the last version of lip models developed at ICP [5].

From the midsagittal contour to the area function

The 2D midsagittal function is converted into 3D area functions using a set of (α, β) coefficients derived from general data on vowels, and from cast and scanner measurements on constriction zones obtained at the ICP [6] in collaboration with the Grenoble University Hospital Center (maxillo-facial surgery service, Dr. Lebeau; radiographic service, Pr. Crouzet). The cross-sectional area S is computed from the midsagittal distance d as $S = \alpha\, d(x)^{\beta}$, where β has a constant value of 1.5, and α depends on the region in the vocal tract (glottis, lower part of the pharynx, upper part of the pharynx, oro-pharynx region, velar region, hard palate zone, alveolar region, intero-labial lip region), and on the dimension of the cross-section width (separate sets of coefficients for dimensions less than 1 cm and greater than 2 cm, with interpolation for the intermediate values). An alternative possibility is proposed [7] where the α coefficient depends continuously on the coordinate along the vocal tract midline x, and have been optimised for vowels and and fricative sounds.

In addition, the three traditional parameters that characterize area functions are computed: the lip area (Al), the position of the intra-oral constriction (relative to the glottis Xcg, or better to the teeth Xct) and the constriction cross-sectional area .

From area functions to formants, bandwidths and transfer functions

Three possibilities are offered to compute formants, bandwidths and transfer functions: (1) line analog frequency domain simulation of the vocal tract acoustic transfer function, and extraction of the corresponding complex conjugate poles [8]; (2) estimation of the resonances by means of a variationnal method [9]; (3) direct estimation of the formants expressed as 3rd order polynomes of the seven articulatory parameters, with coefficient optimized from a codebook generated by the model [10].

A given vowel sound can be located in the Maximal Vowel Space MVS [11], i.e. an nD space (with $2 \leq n \leq 5$). The MVS of the model has been evaluated by performing an extensive exploration of combinations of input parameters. Thus, each vowel calculated by simulation can be plotted in the F1-F2 and F2-F3 projections of the MVS. If an occlusion appears in the vocal tract, a warning message is sent and no calculation is performed. The formant bandwiths needed for sound synthesis are extracted from the complex poles, when the analog model is selected. For the others two cases, the bandwidths are estimated as proposed by Båvegård et al. [12]

The transfer functions (251 frequency points over the 0-5 kHz range, amplitudes in dB) are computed using an analog simulation [8]. The boundary conditions are: (1) a closed glottis, (2) distributed wall impedances, (3) a lip radiation impedance simulated as a piston in an infinite baffle. Heat conduction and friction losses are also included.

From formants and bandwidths to acoustic signal

The vowel sounds are computed using a simplified synthesis model consisting of a cascade of five formant filters (center frequencies F1-F5 and bandwidths B1-B5), excited by an appropriate glottal waveform [13]. Particularly, each glottal pulse is pitch-synchronously damped to simulate the effect of the periodic glottis opening. In order to obtain a more natural sound, Fo excitation variations and Intensity Level envelopes have been extracted from natural isolated vowel sounds uttered by a speaker. Finally, the signal can be listen to directly through loudspeakers, and a sound file is also stored on disk. Reasonable quality sound is generated (8 bits, 22 kHz).

SOFTWARE IMPLEMENTATION

Environment and platform

Several criteria have been considered for the choice of the platform: standard configuration with no additional hardware for graphics and audio output, reasonable processing time, ease of distribution and update, and reasonable chances of longevity, considering anticipated software and hardware evolution. A Macintosh platform (Macintosh II, Powerbook, and Quadra series) has been selected.

All software has been rewritten in the C language (Think C) on the basis of original software written in Fortran (the acoustic model and the sound generation). The interface has been developed in the HyperTalk language, in a HyperCard environment. Several facilities of "object-like programming" in HyperTalk have been used: additional menus, palette; the number of XCMD (Hypercard External commands) has been limited to the strict minimum in order to ease portability. The display is automatically adapted to the size of the available screen (from 14 inches to 21 inches), and sound is output through standard Macintosh sound resources (8 bits, 22 kHz). The coprocessor (68882) is necessary for reasons of processing speed.

Macintosh platforms

All platforms with 68020-68030-68040 microprocessors, 68882 coprocessor, with at least 4 Megabytes of RAM can be used. A 14 (or more) inch screen is required, and no special sound I/O card is necessary: the SMIP generates Macintosh sound resources. The SMIP requires a system 7.x version. SMIP does not run on a PowerMac so far.

AKNOWLEDGEMENTS

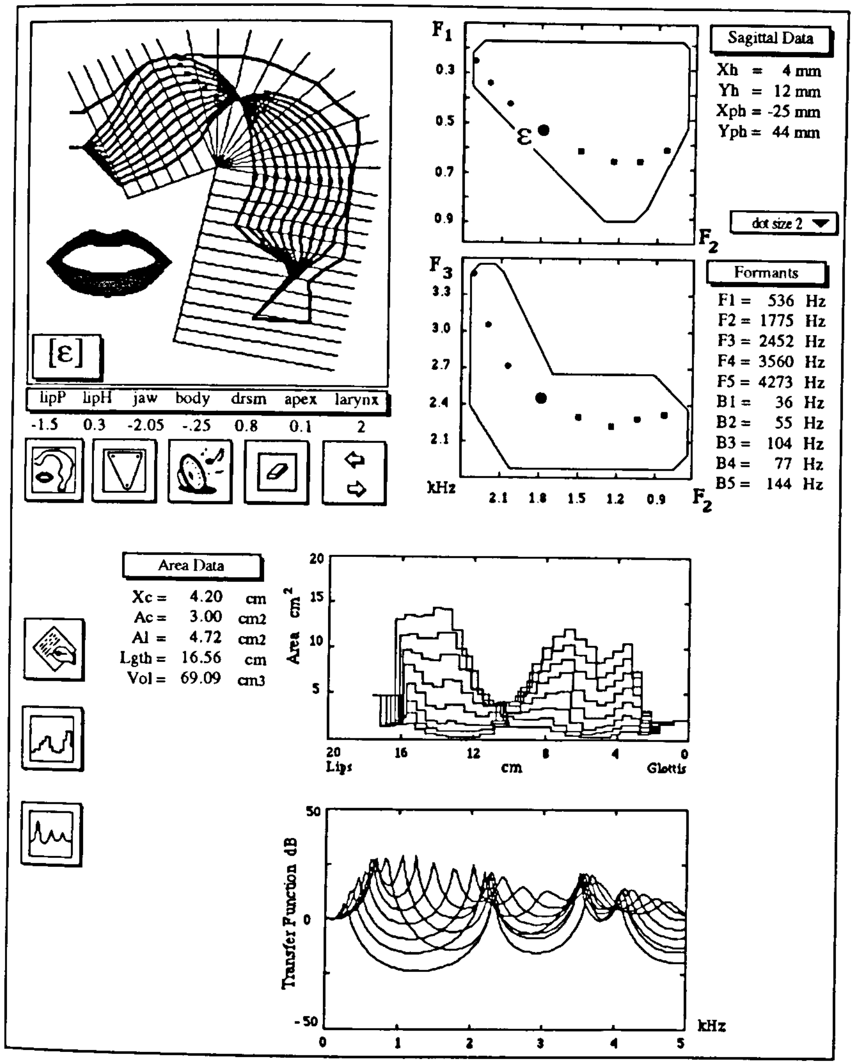
We would like to thank Paul Jospa of the institute of Phonetics in Brussels who provided us with the formant calculation program (variational method). Thanks a lot to Pierre Badin, Sonia Kandel, and Caroline Smith for improvements to the manuscript and to all the first users for their helpful criticisms

THE CONTRIBUTORS

The SMIP is the result of fruitful exchanges between Shinji Maeda, the ICP and the Institute of Informatics of Lausanne. Many direct or indirect contributors have been involved:
Adaptation of the Maeda model: P. Perrier and V. Jacquart. Translation of the sofware in C language: B. Gabioud. *Lip models:* C. Abry, L.J. Boë, P. Perrier, C. Benoît and T. Guiard-Marigny. *Coefficients from sagittal dimensions to area function*: L.J. Boë, Pascal Perrier, Rudolf Sock, D. Beautemps, P. Badin, R. Laboissière. *Harmonic simulation and formant estimation:* P. Badin and G. Fant, P. Jospa, A. Morris and E. Reynier. *Pole simulation*: G. Feng. *Translation of the sofwares in Think C:* B. Gabioud. Vowel Prototypes: N. Vallée and J. Payan. *HyperCard general conception:* L.J. Boë and B. Gabioud. HyperCard and XCMD developments: L.J. Boë, B. Gabioud, S. Bernier, P. Vacchino, F. Pinet, D. Guillem, L. Galmiche, A. Dumay, P. Déquier, M. Chaize, E. Grimont, and J. de Combret (Diadème Society).

REFERENCES

[1] Abry C., Badin P. & Scully C. (1994) *Sound-to-gesture Inversion in Speech:* The Speech Maps *Approach.* ESPRIT Research Rept. 6975. In *Advanced Speech Applications.* Varghese K., Pfleger S. & Lefèvre J.P. (Eds.), 182-196. Springer Verlag, Berlin.
[2] Maeda S. (1989) *Compensatory Articulation during Speech: Evidence from the Analysis and Synthesis of Vocal-Tract Shapes using an Articulatory Model.* In *Speech Production and Modelling*, 131-149. W.J. Hardcastle & A. Marchal (Eds.), Academic Publishers, Kluwer.
[3] Gabioud B. (1994). *Articulatory Models in Speech Synthesis.* In E. Keller (Ed.), *Fundamentals of Speech Synthesis and Recognition:*, 215-230, Chichester, John Willey.
[4] Bothorel A., Simon P., Wioland F., & Zerling J.-P. (1986). *Cinéradiographie des voyelles et des consonnes du français.* Institut de Phonétique, Strasbourg.
[5] Guiard-Marigny T. (1992). *Modélisation des lèvres.* DEA, Institut National Polytechnique de Grenoble.
[6] Perrier P., Boë L.-J. & Sock R. (1992) Vocal Tract Area Functions Estimation from Midsagittal Dimensions with CT Scans and a Vocal Tract Cast: Modelling the Transition with two Sets of Coefficients. *J. of Speech and Hearing Research*, 35, 53-67
[7] Beautemps D., Badin P., & Laboissière R. (1995). Deriving vocal-tract area functions from midsagittal profiles and formant frequencies. *Speech Communication*, 16, 27-47.
[8] Badin P., & Fant G. (1984). Notes on vocal tract computations. *STL Quaterly Progress Status Report,* 2 -3, 53-108.
[9] Jospa P. (1992). Caractérisation variationnelle des modes de résonance dans le conduit vocal. *Rapport d'Activité de l'Institut de Phonétique de Bruxelles*, 28, 13-30.
[10] Morris A. (1992) Least-squares fit to Maeda model dictionary. *Technical report, ICP*, Grenoble, 8 p.
[11] Boë L.-J., Perrier P., Guérin B., Schwartz J.-L. (1989) Maximal Vowel Space. *EuroSpeech 89*, 2, 281-284.
[13] Feng G. (1983). Vers une synthèse par la méthode des pôles et des zéros. *13^{e}J* EP (GFCP, SFA) 155-157.
[12] Båvegård M., Fant G., Gauffin J., & Liljencrants J. (1994). Vocal tract sweeptone data and interpretation. In S. Maeda (Ed.) *From Speech Signal to Vocal Tract Geometry.* PPR 2, European ESPRIT/BR N° 6975 *Speech Maps* project. Vol. III.



lipP	lipH	jaw	body	drsm	apex	larynx
-1.5	0.3	-2.05	-.25	0.8	0.1	2

Sagittal Data
Xh = 4 mm
Yh = 12 mm
Xph = -25 mm
Yph = 44 mm

Formants
F1 = 536 Hz
F2 = 1775 Hz
F3 = 2452 Hz
F4 = 3560 Hz
F5 = 4273 Hz
B1 = 36 Hz
B2 = 55 Hz
B3 = 104 Hz
B4 = 77 Hz
B5 = 144 Hz

Area Data
Xc = 4.20 cm
Ac = 3.00 cm2
Al = 4.72 cm2
Lgth = 16.56 cm
Vol = 69.09 cm3

For prototypic [ɛ]:
- The sagittal contour of the vocal tract generated by the Maeda model, and macro-variations calcultaded for ± 3σ variations of the tongue body.
Lip shape contours are calculated by using the ICP models. The upper point of the tongue body (Xh, Yh), and the furthest point of the tonge root in the pharynx (Xph, Yph) are indicated by dots (▪).
- Area function derived from contour In addition, the three traditional parameters that characterize area functions are computed: the lip area (Al), the position of the intra-oral constriction (Xc) relative to the teeth, the constriction cross-sectional area (Ac), the length and the volume of the vocal tract.
- The transfer function derived from the area function, the associated F1-F4 formants and B1-B4 bandwidths, and the the location of [ɛ] in the Maximal Vowel Space of the model.