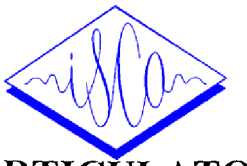

Models and Analysis of Vocal
Emissions for Biomedical Applications
(MAVEBA 1999)
Firenze, Italy, September 1-3, 1999

APEX – AN ARTICULATORY MODEL FOR SPEECH AND SINGING

Christine Ericsdotter*, ericsdotter@ling.su.se
Johan Sundberg**, pjohan@speech.kth.se
Björn Lindblom*, lindblom@ling.su.se
Johan Stark*, johan@ling.su.se

* Department of Linguistics, Stockholm University, 106 91 Stockholm, Sweden
** KTH Voice Research Centre, Department of Speech, Music and Hearing, KTH (Royal Institute of Technology),100 44 Stockholm

ABSTRACT

The APEX articulatory synthesis model is being developed as a joint project at the Department of Speech, Music and Hearing at the Royal Institute of Technology and at the Department of Linguistics at Stockholm University. It is a direct development of an earlier vowel model [1], implemented as a computer program under Windows [2]. It calculates formants and produces sound according to articulatory profiles from a virtual vocal tract, it generates possible articulatory configurations within a specified articulatory space and it also parameterizes and animates series of articulatory configurations. The default vocal tract is based on lateral X-ray data from a male adult speaker complemented with frontal and mid-sagittal measures from a standard vocal tract. However, the model can be calibrated with and run on vocal tract data from any individual. The APEX model is used for testing and shedding light on theories of speech and singing production, in general as well as for specific speakers or singers. It is primarily a research instrument, continually developed according to new findings and the needs of its users.

AIMS

Articulatory models ideally allow control of the positions of the articulators, the lower jaw, the lips, the root, the body and the tip of the tongue, the velum, and the larynx. If accurate, such models will produce realistic area functions. Articulatory models are similar to but not equivalent with area function models, where the control parameters are location and degree of the tongue constriction plus the cross-sectional area of the lip opening; such models may produce also area function unavailable to a real vocal tract.

Our work with an articulatory model started many years ago, and was originally based on tracings of X-ray profiles of a Swedish subject who produced a dozen sustained vowel sounds [1]. We have now updated and expanded this model. In its present form, called APEX, it runs as a PC program and produces sounds by means of a conventional sound card.

The ultimate goal of our APEX model is to contribute to a better understanding of the function and potentials of the human voice organ in speech and singing. A physiologically realistic model should offer an efficient tool for translating acoustic characteristics into articulatory gestures and settings, and vice versa. An articulatory model would also offer powerful pedagogical means.

APPEARANCE

The APEX program runs under Windows and is written in C++ [2]. The display (Figure 1, upper panel) shows a virtual vocal tract (VT) profile, articulatory parameter regulators and function toolbars allowing variation of the position and shape of the model's lips, tongue tip, tongue body, jaw opening and larynx height. A coordinate system template defined with respect to fixed VT landmarks is then applied to this profile. This template is used for measuring sagittal distances along the VT midline at a number of points from the glottis to the lips. These distances are then converted into cross-sectional areas (Figure 1, lower panel) using anatomically motivated and speaker-dependent rules. This area function is used for calculating the formant frequencies. Using the SENSYN speech synthesizer sound examples of articulatory configurations can be obtained.

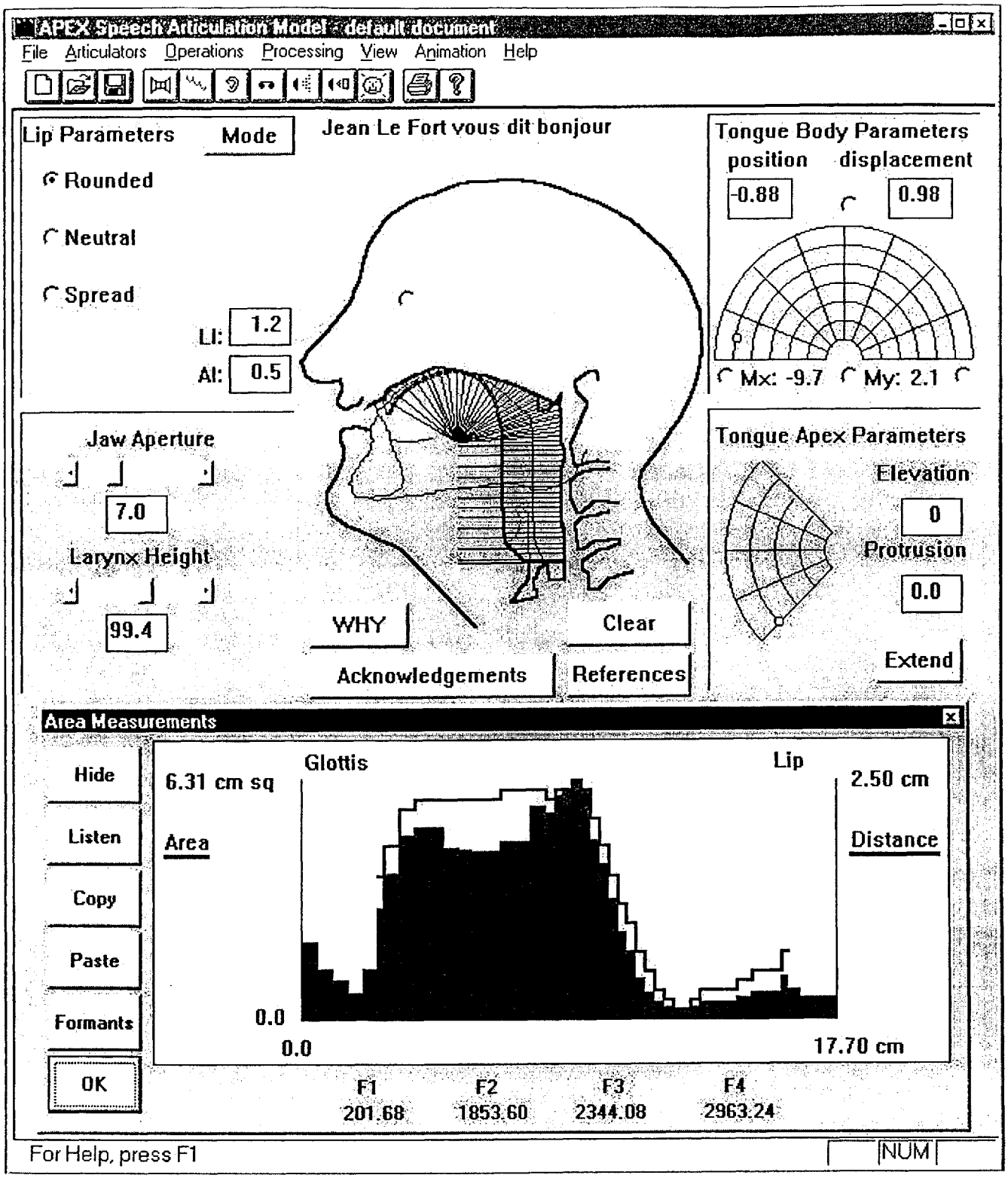

Figure 1. The APEX articulatory model. In the lower panel, the dark pattern represents the area function. The line refers to the VT cross-distances.

ARTICULATORY PARAMETERS AND CALIBRATION

To generate an articulatory profile, the shape and position of the fixed contours (mandible, maxilla, posterior pharyngeal wall and larynx) and the variable contours (tongue body, tongue blade and lips) are combined. The adjustable articulatory parameters are the following: the one-dimensional mandible ranging from 0 to 25 mm along a curvi-linear, empirically determined path; the larynx which can be translated and rotated in the x/y plane; the lips described in terms of two parameters: width and height; the tongue blade created by a parabolic function

attached to the tongue body and controlled by two parameters: protrusion (extension-retraction) and elevation (displacement from neutral); and the tongue body specified in terms of two parameters: anterior-posterior position and displacement (deviation from neutral).

The geometry of the APEX vocal tract is based on articulatory X-ray data [3] selected from an articulatory database that contains 12 subjects. Using a digital X-ray technique the subjects were recorded for 20 seconds each at a speed of 50 images/second. Audio signals were registered synchronously.

To calibrate the APEX model with data from an individual speaker, the vocal tract contours from the vowels /i, ᵉ₁, u/ and a neutral ᴺʟ-like vowel were traced on X-ray images, using the Osiris Imaging Software (Figure 2). The general strategy was to have the model represent these vowel configurations as faithfully as possible and then to derive intermediate articulations by physiologically motivated interpolation rules. All X-ray tracings were specified in the program by x/y coordinates and extracted as labeled lists, using a specially developed tool (Papex). The tracings were then rescaled to real mm, and transferred to the origin of the coordinate system in APEX.

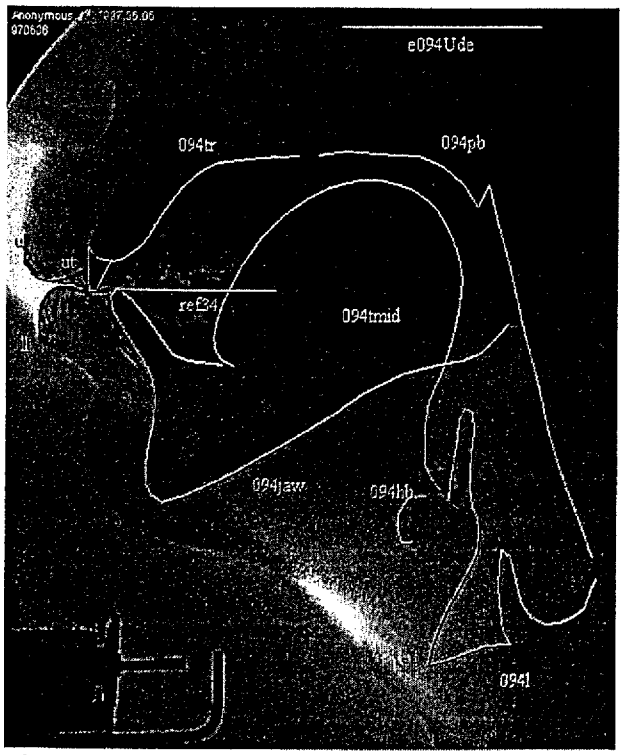

Figure 2. X-ray image with contour tracings of an articulation of the vowel [u:].

To allow for control of the jaw aperture, a path was derived by plotting the positions of mandible reference points versus the maxilla from selected articulations (/s, u, y, ᴺʟ, o, ᵉ₁/). Using the equation of this plotted jaw function, the path was expanded by extrapolation.

The contours were stored as lists of plain x/y mm coordinates in a calibration text file, written in a syntax readable by APEX. This file contained information about scaling, origin, subject identity, templates for fixed and variable contours, the four selected reference vowels specified in terms of their tongue blade, tongue body, larynx position and jaw opening, jaw path described.

The input file is loaded in the APEX program, which opens three mathematical models for calibration of tongue body, tongue blade and larynx, respectively. Calibration is achieved by adjusting the control constants of these models so as to simulate the vocal tract configurations for each reference vowel. This is realized in a graphic interface, showing the observed and the modeled curve (Figure 3). The goodness of fit is displayed as the root mean square error for every adjustment. The calibration process implies modeling of 12 tracings (four vowels with three parts each). The class of all possible tongue bodies of the modeled speaker is generated by interpolating between the reference configurations. In essence, this interpolation reflects the action of the hyoglossus, the styloglossus and the genioglossus, the major muscular determinants of shape and position in vowel articulation. A more detailed account for the calibration of the APEX can be found in Stark et al. [4] [5].

By adjusting the model parameters the virtual subject's articulatory possibilities can be investigated and compared with those of the real speaker.

ACOUSTICAL CHARACTERISTICS

To obtain the acoustic correlate of a given articulatory profile, the first step is to determine the vocal tract 'midline'. A coordinate system template is positioned relative to anatomical landmarks (see Figure 1, upper panel). Lines perpendicular to this midline are used for measuring sagittal VT distances (d). These

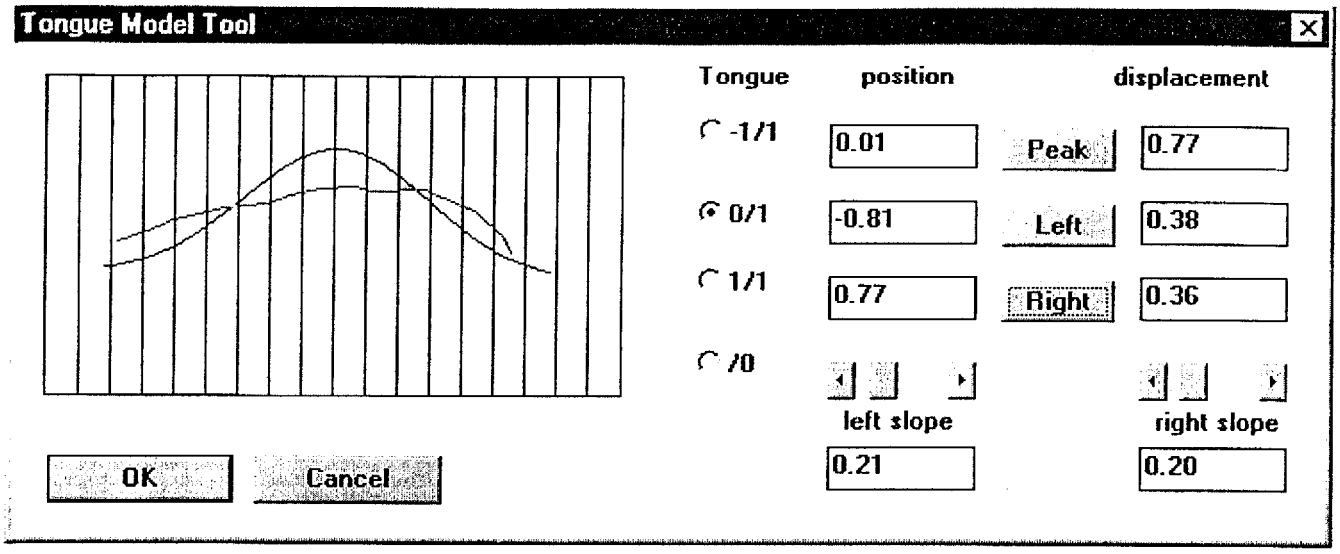

Figure 3. Tongue model tool. The shape of the Gaussian curve in the left panel is defined by the parameters specified in the windows to the right. The other curve represents an observed tongue contour.

distances are then transformed into cross-sectional areas (A) according to

$$A=a*d^b$$

where a and b are constants varying with location along the VT and the characteristics of individual speakers. The resulting area function A(x) is fed into an algorithm that computes the formant frequencies by means of a software developed by Liljencrants.

The user can specify a sequence of articulatory targets and assign a duration to each configuration. These specifications are displayed as parameter step functions and converted to smooth articulatory movements. APEX calculates formant trajectories, creates waveforms of the sequences (SENSYN) and sends the result to a loudspeaker for perceptual evaluation.

APPLICATIONS

APEX modeling has already been used in speech, singing and voice research. Engstrand et al. (forthcoming [6]) used APEX for finding articulatory interpretations of dialectic variants of a Swedish vowel. By translating coarticulation data to APEX parameters, Ericsdotter et al. [7] proposed an improved description of principles underlying coarticulation. Likewise, the model has been used to investigate the articulatory possibilities that are available to singers in order to increase the first formant frequency; such increases are needed to keep this frequency higher than the fundamental in cases of high-pitched singing [8].

In the future, work will be spent on improving the calibration of the model. Another important goal is to interpret formant frequency patterns in connected speech and in singing in terms of articulatory gestures.

ACKNOWLEDGEMENTS

This work is supported by the Bank of Sweden Tercentenary Foundation (RJ 95-5173:03) and the Swedish Council for Research in the Humanities and Social Sciences (HSFR F 0707/97).

REFERENCES

[1] Lindblom, B. et Sundberg, J. (1971): "Acoustical consequences of lip, tongue, jaw, and larynx movement." *J. Acoust. Soc. Am.* 50: 1166-1179.

[2] Stark, J., Lindblom, B et Sundberg, J. (1996): "APEX an articulatory synthesis model for experimental and computation studies of speech production." *TMH-QPSR* 2/1996: 45-48.

[3] Branderud, P., Lundberg, H.-J., Lander, J., Djamshidpey, H., Wäneland, I., Krull, D., Lindblom, B. (1998): "X-ray analyses of speech: Methodological aspects." Proceedings from *FONETIK 98, the eleventh Swedish Phonetics Conference, Stockholm University, May 27-29, 1998:* 168-171.

[4] Stark, J., Ericsdotter, C., Lindblom, B., Sundberg, J. (1998) "Using X-ray data to calibrate the APEX articulatory synthesis model." In *FONETIK 98, the eleventh Swedish Phonetics Conference, Stockholm University, May 27-29, 199:* 184-187.
[5] Stark, J., Ericsdotter, C., Branderud, P., Sundberg, J., Lundberg, H.-J., Lander, J. (1999): "The APEX model as a tool in the specification of speaker-specific articulatory behavior." To appear in the *Proceedings of the XIVth ICPhS, San Francisco, California, 1-7 August 1999.*
[6] Engstrand, O., Björsten, S., Lindblom, B., Bruce, G. et Eriksson, A. (forthcoming): "Hur udda är Viby-i? Experimentella och typologiska observationer." To appear in Folkmålsstudier 39, 1999, published by Forskningscentralen för de inhemska språken, Helsinki, Finland.
[7] Ericsdotter, C., Stark, J. et Lindblom, B. (1999): "Articulatory coordination in coronal stops: Implications for theories of coarticulation." To appear in the *Proceedings of the XIVth ICPhS, San Francisco, California, 1-7 August 1999.*
[8] Ericsdotter, C., Sundberg, J., Lindblom, B. et Stark, J. (forthcoming): "Widening Jaw opening or tongue constriction for raising F1 in high pitch singing?" *Paper to be presented at PEVOC III (3rd Pan European Voice Conference), University of Utrecht, August 26-29, 1999.*

SOFTWARE

PapEx, Papyrus to Excel text file (csv) converter, version 1.0. 1998. Roberto Bresin, Department of Speech, Music and Hearing, KTH, SE-100 44 Stockholm, Sweden. http://www.speech.kth.se/~roberto/

Liljencrants J. *Formf.c*, C-program for the calculation of formant frequencies from area functions. TMH, KTH, Stockholm.

Osiris Medical Imaging software, version 3.5. University Hospital of Geneva, Digital Imaging Unit. 24, rue Micheli-du-Crest, 1211 Genève 14, Switzerland. http://www.expasy.ch/UIN/html1/projects/osiris/osiris.html

SENSYN Speech synthesizer. Formant synthesizer that produces speech waveform files based on the (Klatt) KLSYN88 synthesizer. Sensimetrics Corporation Sidney Street, Cambridge MA 02139. http://www.sens.com/