

THE APEX MODEL AS A TOOL IN THE SPECIFICATION OF SPEAKER-SPECIFIC ARTICULATORY BEHAVIOR

Johan Stark^{*}, Christine Ericsson^{*}, Peter Branderud^{*}, Johan Sundberg[‡], Hans-Jerker Lundberg[^],
Jaroslava Lander⁺

^{*}Stockholm University, [‡]Royal Institute of Technology, [^]Danderyd Hospital, Stockholm,
⁺Södersjukhuset, Stockholm

ABSTRACT

This paper presents a description of the APEX speech production model and some of its main current capabilities. We describe an ongoing X-ray project initiated to calibrate APEX according to individual speaker characteristics and to obtain data on how articulatory movements are coordinated in the coarticulation of Swedish vowels and stops. The mapping of the X-ray data onto APEX parameters is also illustrated.

1. THE APEX MODEL: DEFAULT CONFIGURATION

The APEX model is an implementation of a framework previously developed for vowels by Lindblom and Sundberg [1] and subsequently augmented with tongue tip/blade parameters.

APEX is a tool for going from articulatory positions to sound in four steps. From specifications for lips, tongue tip, tongue body, jaw opening and larynx height, APEX constructs (1) an *articulatory profile*. A coordinate system (defined with respect to fixed vocal tract landmarks) is then applied to this profile to measure (2) the *cross-distances* along the VT midline at a number of points from the glottis to the lips. The cross-distances are then converted into (3) *cross-sectional areas* using anatomically motivated and speaker-dependent rules. This area function forms the basis of (4) *formant frequency calculations* [2] and the generation of a waveform using the SENSYN speech synthesizer [3] to provide an audible illustration of the configuration under analysis.

The geometry of the APEX VT is based on X-ray data. The generation of an articulatory profile uses contours sampled as x/y points for (i) the shape of the mandible, the hard palate, the posterior pharyngeal wall and the larynx, and for (ii) the shape of articulators (tongue body, tongue blade, lips). The general philosophy is to have the model represent certain key observed configurations as faithfully as possible and then derive intermediate articulations by physiologically motivated interpolation rules.

One of the key features of the model is the mandible-based specification of lip and tongue configurations. Accordingly, a given specification of the shape and position of the tongue produces a contour that is fixed and independent of jaw opening. This is true also of the labial width and height parameters.

1.1. The articulatory parameters

The lips are described in terms of two parameters: width and height. The tongue blade is created by a parabolic function attached to the tongue body. It is controlled by two parameters: *protrusion* (extension-retraction) and *elevation* (displacement

from neutral). The tongue body is specified using two parameters: anterior-posterior *position* and *displacement* (deviation from neutral). The position dimension ranges from palatal ('i' tongue) to pharyngeal ('a' tongue) via velar ('u' tongue). Displacement regulates the size of the tongue 'hump', zero displacement corresponding the neutral contour ('rest'). The space of 'possible tongue bodies' is generated by interpolating between the reference configurations, essentially to reflect major muscular determinants of shape and position in vowels: the hyoglossus, the styloglossus and the genioglossus.

The *jaw* is one-dimensional. It ranges from 0 to 25mm along an opening-closing curvi-linear path determined empirically. The *larynx* is a fixed contour that can be translated and rotated in the x/y plane.

1.2. Acoustical characteristics

For each APEX profile, a semi-polar coordinate system is positioned in relation to fixed anatomical landmarks for determination of the VT 'midline'. Lines are drawn perpendicular to this midline so that the cross-distance between the upper and lower contours of the VT can be measured. These distances (d) are transformed into equivalent cross-sectional areas (A) according to $A=a*d^b$, where a and b are constants varying with VT location and individual speaker characteristics. The resulting area function A(x) is fed into an algorithm that computes the formant frequencies [2].

APEX allows the user to specify a sequence of articulatory targets and assign a duration to each configuration. These specifications are displayed as parameter step functions and are converted to smooth articulatory movements. APEX calculates formant tracks, creates waveforms of the sequences [3] and sends the result to a loudspeaker for perceptual evaluation.

The software is written in C++ for a PC and the Microsoft Windows environment.

2. THE X-RAY PROJECT

Recently we began making X-ray films with synchronous sound records. The project has three main objectives. (1) to calibrate APEX in accordance with individual speaker characteristics; (2) to collect data on how articulatory movements are coordinated in the coarticulation of Swedish vowels and stops, as in [4], and (3) to use the model so customized in computational experiments and as a tool in analyzing the X-ray data.

2.1. Recordings

The digital X-ray equipment at Danderyd Hospital makes it

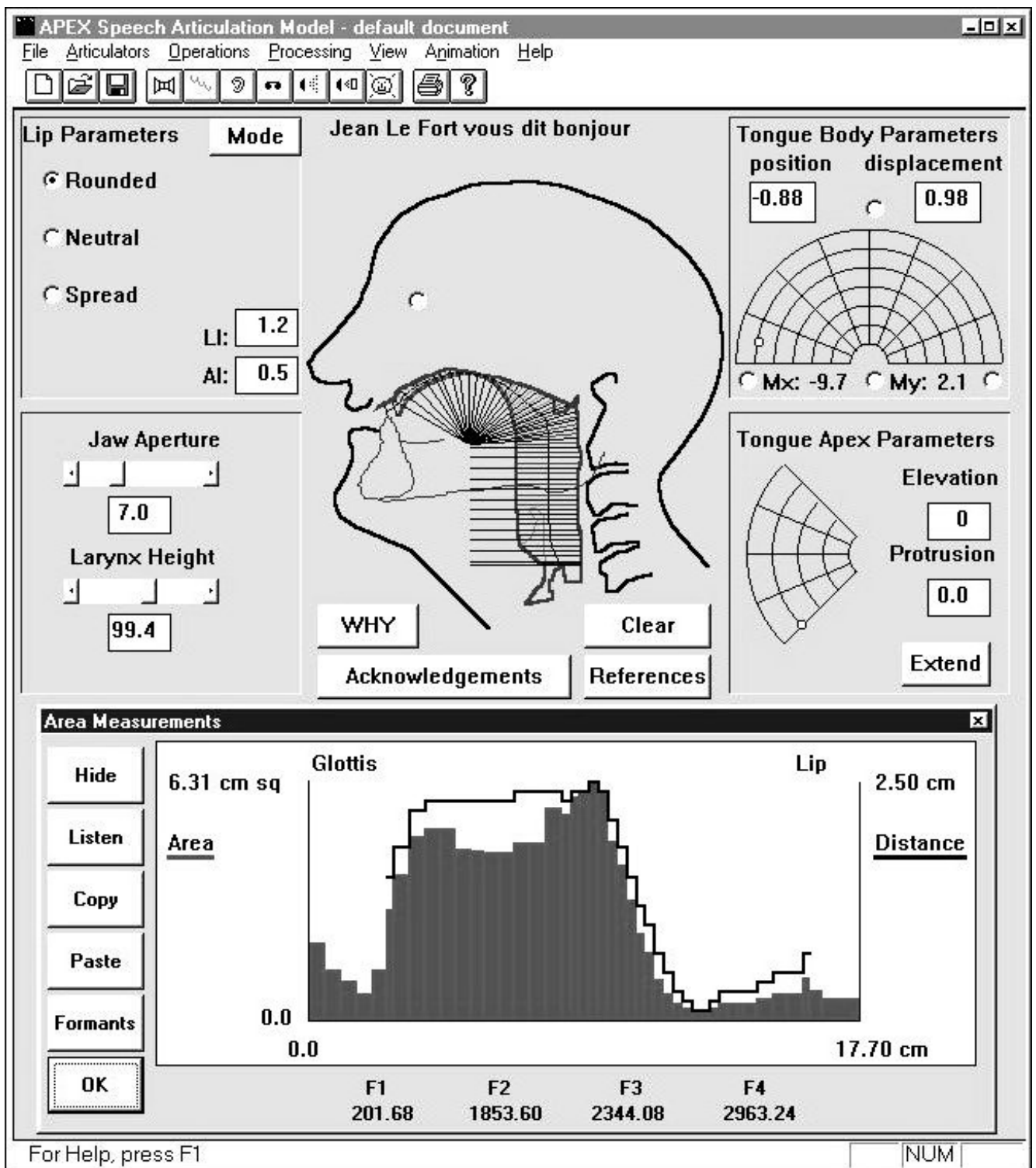


Figure 1. The APEX articulatory model. In the lower panel, the dark pattern represents the area function. The line refers to the VT cross-distances.

possible to record 50 images/second with an X-ray pulse of 3ms/image. This is short compared with articulatory speeds. Radiation is minimized by the use of a prespecified pediatric program and by removing the image intensifier scatter-radiation grid. Spatial resolution is estimated at 0.3 mm.

The source image distance (SID) was set to 120 cm. An enameled copper wire 0,5 mm in diameter was attached midsagittally along the subject's hard palate to indicate head movement and scaling effects and to identify the hard palate surface. To enhance soft tissue contrast, the subject applied barium sulfate, a contrast paste, to the lips and the mouth region. The image receptor was placed as close to the subject as possible to achieve high image quality. Dosimeters were attached to the subject's face. The imaging area was controlled by means of low-dosage fluoroscopy.

A microphone was placed about 10 cm in front of the mouth of the subject and connected to a tape recorder. Synchronizing sound and images was done by placing an X-ray detector within the view of the camera and by recording the analog output of this device on a second channel of the tape recorder.

In the most exposed organ (*parotis sin*), the absorbed radiation dose was less than 4 mGy, the effective dose not exceeding 0.1 mSv

To check for optical distortion and to convert measurements to real mm-distances, a reference grid was recorded in the position of the subject's midsagittal plane.

2.2. Data extraction

The films are saved on CD's for subsequent processing by means of the Osiris software [5]. Speed, size, filtering, color settings and contrast can be altered to facilitate analyses. Tools are available for drawing and measuring contours.

The upper and lower lips were traced along their profiles produced by the contrast medium. The tongue contour followed the midsagittal outline of the blade and the tongue body down to the tongue root. The sides of the tongues were marked where they differed from the midsagittal. A maxillar contour was drawn from the upper incisors and continuing with the hard and soft palate and the post pharyngeal wall. It ended at the entrance to the esophagus. The lower jaw component comprised the floor of the mouth, the lower incisors and the mandible up to *angulus mandibulae*. The larynx object ran from the epiglottis to the glottis. To register head movements, the wire attached to the maxilla was traced as reference. The path of the jaw was defined in terms of two mandible reference points. The most anterior part of hyoid bone was included. The accuracy of the tracings is estimated to be 0.5-1 mm.

The analysis results were then extracted from the image files using a specially developed program [6] yielding labeled lists of the xy-coordinates of the VT contours. The contours were then rescaled and corrected for optical distortion and head movements.

3. SPEAKER-SPECIFIC RECALIBRATION

Like other production models APEX is useful for studying the acoustic consequences of articulatory movement and for investigating possible articulatory origins of observed acoustic patterns. The model has been used in a number of applications.

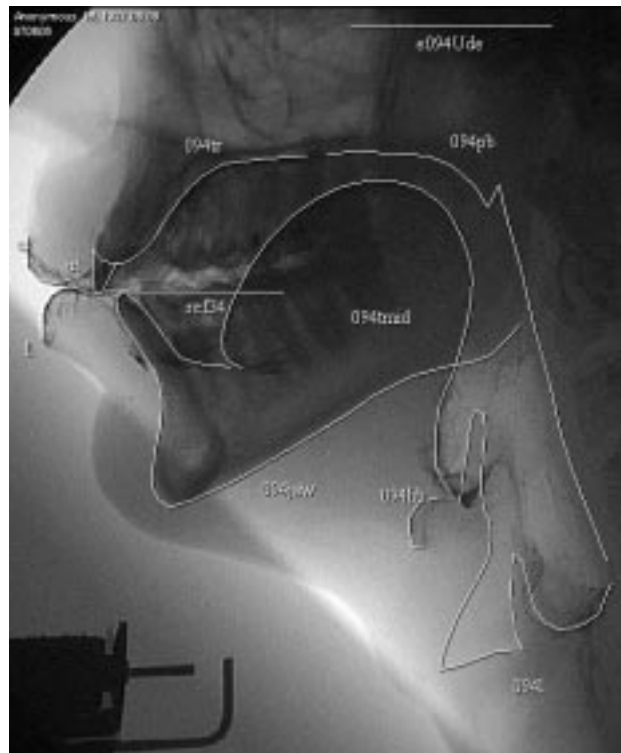


Figure 2. Traced X-ray profile of an [u]-articulation.

In a 'learning' experiment [7] we made a comprehensive search through the total APEX parameter space for the formant patterns of Swedish vowels. It was concluded that the default version of APEX is capable of producing those patterns in both articulatorily and acoustically realistic ways. The same conclusion was reached for the ability of APEX to produce the 'locus' values of Swedish dental and retroflex stops. We take these findings to indicate that both the choice of control parameters and the APEX interpolation philosophy are workable approximations.

In analyzing the X-ray films we have set ourselves the goal of mapping the articulatory motions onto APEX parameters and to try to recreate both the articulatory and the acoustic observations from those specifications in a realistic manner. In this effort APEX is used in a matching mode.

Figure 3 illustrates this use. The diagram shows a comparison of the tongue observed at the closure of the stop in [u:dε] (in white). The dark curve represents the best match produced by APEX.

The insert shows the jaw-based tongue body space. APEX tongue contours are described in terms of position (angle of radius) and displacement from neutral (length of radius). The time course of the articulatory movement during the word is depicted as a trajectory in this semi-circular space. It begins in the 'velar' region of the space and then moves into more palatal-neutral territory.

In an accompanying paper [4] we present some results on using APEX in this manner.

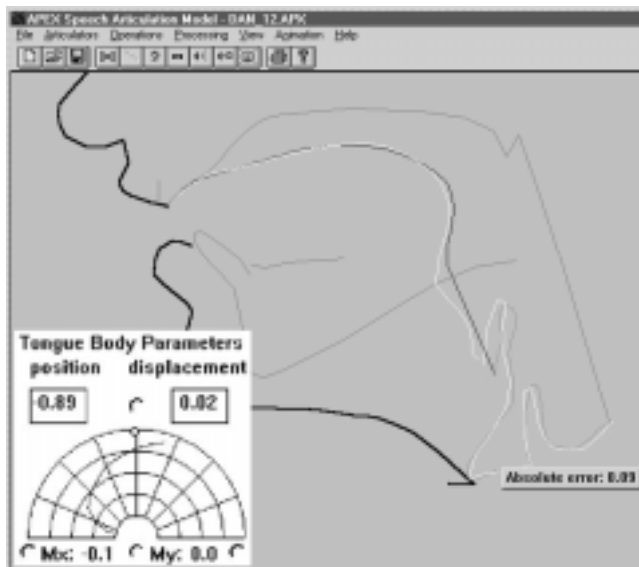


Figure 3. The APEX model imitating a [d] articulation.

The goal of APEX-based synthesis creates a demand for physiological realism and more detailed speaker-specific information than available in the default version of APEX. To a large extent the X-ray films are capable of catering to this need. A key point however is the quality of the rules describing the conversion from cross-distances to cross-sectional areas, the $d(x)$ -to- $A(x)$ rules. The pharyngeal region presents a particularly difficult problem.

To address this problem, we have made dental casts of the individual subjects' anterior VT's. For a given film frame, together with the tracings of the tongue contours (mid-sagittal and edges) they provide accurate information on the cross-sectional areas in the front. On a frame-by-frame basis we are currently exploring combining these area function estimates with a computer algorithm that automatically estimates the areas in the posterior region from the subject's formant values for the frame under analysis. The algorithm takes the anterior part of the area function as given, divides the posterior region into a number of sections and then, under constraints on spatial continuity, searches through combinations of pharyngeal area values for a large number of articulations and until a minimum formant error has been achieved. A report on this project will be presented at the ICPhS 99 meeting.

4. CONCLUDING REMARKS

Production models are useful research tools for gaining insights into the articulatory origins and perceptual function of acoustic speech patterns. They provide a link between direct articulatory observation and acoustics. They supplement techniques such as X-ray films, which are necessarily limited in scope because of health-related radiation risks. They force the investigator to consider *all* the articulatory information acoustically relevant, not just selectively sampled parameters. They can serve these supplementary functions especially if calibrated to represent phonetically significant characteristics of a given individual speaker. They can then be used in an analysis-by-synthesis fashion to describe and interpret the acoustic output of that

speaker in articulatory terms. When confronted with articulatory facts, they can be used heuristically to shed light on how to best model the physiological control dimensions of the VT. In learning experiments they highlight the 'degrees of freedom' problem by revealing the non-unique mapping between articulatory parameters and acoustic result. In so doing they are helpful in suggesting hypotheses about the natural constraints that bootstrap the acquisition of speech production. Our current work on the APEX model is oriented towards achieving those goals.

ACKNOWLEDGMENTS

This research was supported by grants from the Swedish Council for Research in the Humanities and Social Sciences (HSFR F0707/97) and the Bank of Sweden Tercentenary Foundation (RJ 95-5173:03).

REFERENCES

- [1] Lindblom B and Sundberg J. 1971/1991. Acoustical consequences of lip, tongue, jaw and larynx movement. Kent R D, Atal B S and Miller J L (eds): *Papers in Speech Communication: Speech Production*, 329-342, Acoust Soc Am:New York.
- [2] Liljencrants J. *Formf.c*, C-program for the calculation of formant frequencies from area functions. TMH, KTH, Stockholm.
- [3] *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/>
- [4] Ericsson, C., Stark, J. & Lindblom, B. 1999. Articulatory coordination in coronal stops: Implications for theories of coarticulation. *Proceedings from the XIVth ICPhS, San Francisco*, August 1-7 1999.
- [5] *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/UIN.html>
- [6] Bresin, R. 1998. *PapEx, Papyrus to Excel file converter, version 1.0*. Roberto Bresin, Department of Speech, Music and Hearing, KTH, SE-100 44 Stockholm, Sweden.
- [7] Lindblom B, Stark J & Sundberg J. 1997. From sound to vocal gesture: learning to (co)-articulate with APEX. In *Fonetik 97: Papers presented at the Swedish Phonetics Conference*. Phonum, Umeå University.