Dept. for Speech, Music and Hearing
Quarterly Progress and Status Report

APEX an articulatory synthesis model for experimental and computational studies of speech production

Stark, J. and Lindblom, B. and Sundberg, J.

KTH Computer Science and Communication

journal: TMH-QPSR
volume: 37
number: 2
year: 1996
pages: 045-048

http://www.speech.kth.se/qpsr

APEX an articulatory synthesis model for experimental and computational studies of speech production

Johan Stark, Björn Lindblom* & Johan Sundberg***
** Department of Linguistics, University of Stockholm*
***Department of Speech, Music and Hearing, KTH*

Abstract
This is a preliminary report of a project in progress with the purpose to create an articulatory synthesis model for studies of speech production. It is realised by a computer program which may control the lips, the shape of the tongue body and apex, and the mandible. An area function may be computed and displayed graphically and numerically. Formant values may be computed and sent to a formant synthesis model for sound production using a DSP hardware module. Automatic and systematic generation of parameters may be achieved and the results sent to a disk file. The program keeps all speaker dependent data in a disk file, enabling processing of several speakers.

Introduction

The Apex project is aimed at the creation of an articulatory speech model to be used as a tool in studying an important class of speech sounds: apical speech sounds. The need for such a model is for example found in phonology: (Browman and Goldstein 1992), speech technology: articulatory speech syntesis: (Lin 1990, Fant 1992), and in general basic phonetic research: (Hardcastle and Marchal 1990). Research in music acoustics and in singing may also be mentioned. Another goal is to gain a deeper understanding of coarticulation of speech. This knowledge may hopefully be achieved by comparing speech data from the model with data gathered by laboratory experiments, e.g. movement data using a movetrack system (Branderud) and data from signal analysis and spectrograms.

Input data

The Apex model will take input data from tracings based on X-ray images for one speaker and some selected vowels (Lindblom and Sundberg 1971). The tracings show the contours and positions of the articulators. These contours are placed in a coordinate system and sampled as x/y data points. Three different articulator tracings are sampled: The maxilla (static not vowel dependent), the tongue and the mandible. The position of the points is sampled using a precision of approximately 0.5 mm. and the spacing is chosen in order to keep the deviation from the original within +/- 0.5 mm. Some additional tracings are also added like the head, the nose, the external mandible etc mainly for aesthetic reasons. The maxilla tracing comprises the upper teeth, the palate via the rear pharynx wall and goes down to glottis. The tongue tracing is comprising the apex, the tongue body, the epiglottis the larynx all the way down to glottis. The mandible tracing comprises the lower teeth and the mouth floor. To model the complete tongue four sub models are used: one for the apex which is considered to be the first 20-40 mm of the tongue tracing, one for the body, one for the epiglottis and one for the laryngeal part. In addition there is a model for the lips. The whole synthesis model may be controlled by 8 parameters including the mandible position.

The lip model

The lip model is currently not represented by an articulatory model but rather as an area function model which will add the last area segment to the complete area function. Three basic modes are selectable: rounded, spread and neutral. The values come from tables, indexed by the mandible position.

The apex model

The apex model is created by a parabolic function. The parabola is attached by its one leg to the tongue body. The other leg's end correspond to the tip. The model is rotated in order to acheive a smooth conjunction between body and

apex. Two parameters control this model: protrusion (extension), and elevation from neutral. The protrusion is defined as the distance between the tip and the conjunction point. For the parabola this corresponds to the distance between the two end points, or if you draw the parabola as $y = ax^2$ in a local coordinate system, the difference between the x-values of the end points, see fig 1. The elevation is defined as the angle between the line from apex tip to the conjunction point and the tangent through the conjunction point. The selection of the parabolic function for this model has been made because of its great flexibility in modelling the empirically observed apex shapes.

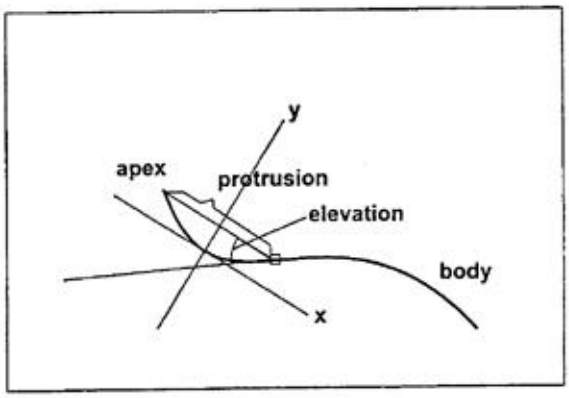

Figure 1. The apex model

The tongue body model

The tongue body may be described using two parameters (Lindblom and Sundberg 1971), position and deviation from neutral for the constriction (the hump). The position ranges from -1 ("i-tongue"), over 0 ("u-tongue") and to 1 ("a-tongue"). The deviation ranges from 0 for a neutral tongue to a maximum of 1. In order to achieve a model shape for any combination of these parameters we wish to have a two parameter model that must also closely correspond to the four input tongue shapes (for vowels i,u,a and neutral). This has been achieved by the use of a modified Gauss function, see fig. 2. The Gauss function has a hump whose size may easily be controlled both by its position and size. This is now a beginning two parameter tongue model. In order to adapt the function to real world tongue shapes some additional modifications are necessary. Firstly the model uses two halves each controlled with their own constants. The halves are joined at the summit, or at the top of the hump. Each half may be adapted to its over all length, slope of the hump, and level of its end point. The additional constants which are required to control this are stored for all the basic input tongue shapes. When an arbitrary tongue shape is requested by an arbitrary combination of position and deviation, these additional constants are interpolated from the ones of the known shapes in order to generate the requested tongue. The interpolation between known tongues is done in two steps, corresponding to the two parameters. The first step is to interpolate between the "i-tongue" and the "u-tongue" if the position parameter is between -1 and 0, or between the "u-tongue" and "a-tongue" if it is greater (>0). At the interpolation each new additional constant is created as a weighted mean value from the two known constants. The weight is controlled by the value of the position parameter. The second step interpolates between this newly achieved tongue and the neutral tongue according to the deviation parameter.

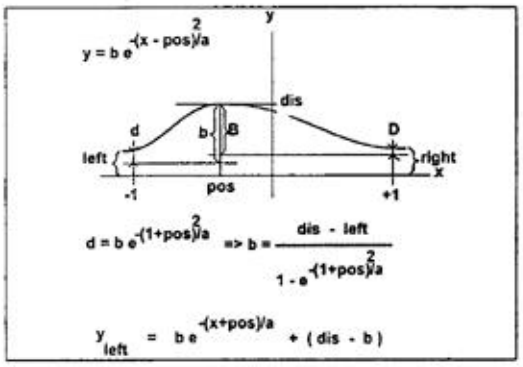

Figure 2. The tongue body model

The epiglottis and larynx

These models are static contours which only move around by translation in x/y directions or rotate around an axis perpendicular to the image plane. The larynx contour is attached to the glottis and may be translated up and down in a direction parallell to the rear pharynx wall. The parameter sets the distance between glottis and the top of the palate. The epiglottis contour is attached to the free end of the larynx and by its other end via a linear segment to the end of the tongue body model.

The mandible

The mandible comprises the lower teeth and the mouth floor. The model is merely a fixed contour like the epiglottis, which is moved according to the mandible position parameter. The mandible position which sets the distance between the teeth may be set in the range 0 to 25mm. All parts of the tongue model are fixed to an origo point defined on the mandible, and

move with it accordingly. This movement is complex and taken from the X-ray images as tracings of two reference points according to some mandible positions. From these reference point tracings a table is calculated containing translation and rotation of the origo as a function of the mandible position. Values between table entries are found through interpolation. All points fixed to the mandible system are transformed using this table if the mandible is moved.

The area function

An area function may be calculated from the glottis up to the lips. A number of equivalent cylinder segments are calculated. Each segment has an area and a length. The last segment is taken from the lip model. This segmentation is roughly achieved as follows: a help line is drawn between the tongue contour and the palate or pharynx wall (pointing to the mandible origo in the upper mouth or horizontally in pharynx), then another help line is drawn a bit further on. A line is then drawn between these two helplines' midpoints. Finally the distance is measured between the palate or pharynx wall and the tongue, perpendicular to this line, through its midpoint. This distance (D) is transformed to an equivalent area (A) using the power function $A = a * D^b$, where a and b are constants varying for different parts of the vocal tract. The length of the last help line is the equivalent length of the cylinder segment. The lengths and areas are fed into an algorithm which calculates the corresponding formant frequencies (Liljencrantz).

The computer program

The model has been realised as a computer program for a PC and the Microsoft Windows environment. The user may control the model parameters in an interactive fashion, and the corresponding model is displayed on the screen. The area function may be calculated and will be reported both graphically and numerically. The corresponding vowel sound may be played via a loud speaker. All software is written in the object oriented programming language C++ which opens the possibility to represent each articulator as a program object. This simplifies the possibility to have several user selectable models for any one articulator. As new research reports become available the program may be updated to comprise new models without requiring revision of the entire program. A comparison between the different models may then be carried out on line. All speaker data may be stored on a disk file.

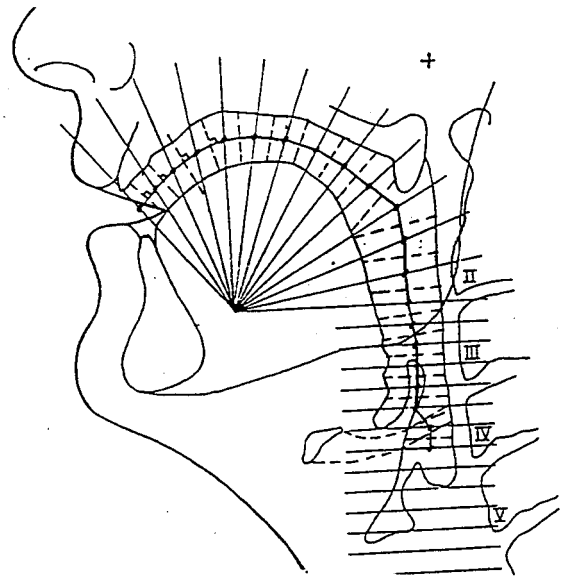

Figure 3. Calculation of area function

Sound generation

The formant values are automatically transferred to a formant synthesis model built with Aladdin (Ternström), which is a software package to control a DSP (Digital Signal Processor) hardware module. The synthesis works in real time and enables the user to listen to the sound without delay just by pressing a button.

F1-F2 diagram

Automatic generation of parameters is possible through specification of range and step. The area function is calculated and accumulatively marked on a F1-F2 diagram (see fig. 4), as well as being numerically transferred to a disk file. More advanced functions are planned. for future versions. One such function is to find all tongue shapes compatible with a certain apical target and plot a line in a position deviation diagram.

Acknowledgements

This research was supported by HSFR of Sweden (project APEX).

References

Branderud P (1985). Movetrack, *Perilus IV,* University of Stckholm.

Browman CP & Goldstein L (1992). Articulatory phonology: An overview, *Phonetica 49.*

Fant G (1992). Vocal tract area functions of Swedish vowels and a new three-parametermodel. *Proc. of*

ICSLP 92, International meeting for speech research, Banff, Canada.

Hardcastle WJ & Marchal A (eds., 1990): *Speech Production and Speech Modeling*, Dordrecht: Kluwer Publishers.

Liljencrantz J. Formf.c, c-program for calculation of formant frequencies from an area function. TMH, KTH.

Lindblom B & Sundberg J (1971/1991). Acoustical consequences of lip, tongue, jaw and larynx movement. In: Kent RD, Atal BS & Miller JL, eds. *Papers in Speech Communication: Speech Production*, Acoust Soc Am, New York, 329-342.

Ternström S. Aladdin, A DSP processing system for PC. TMH, KTH.

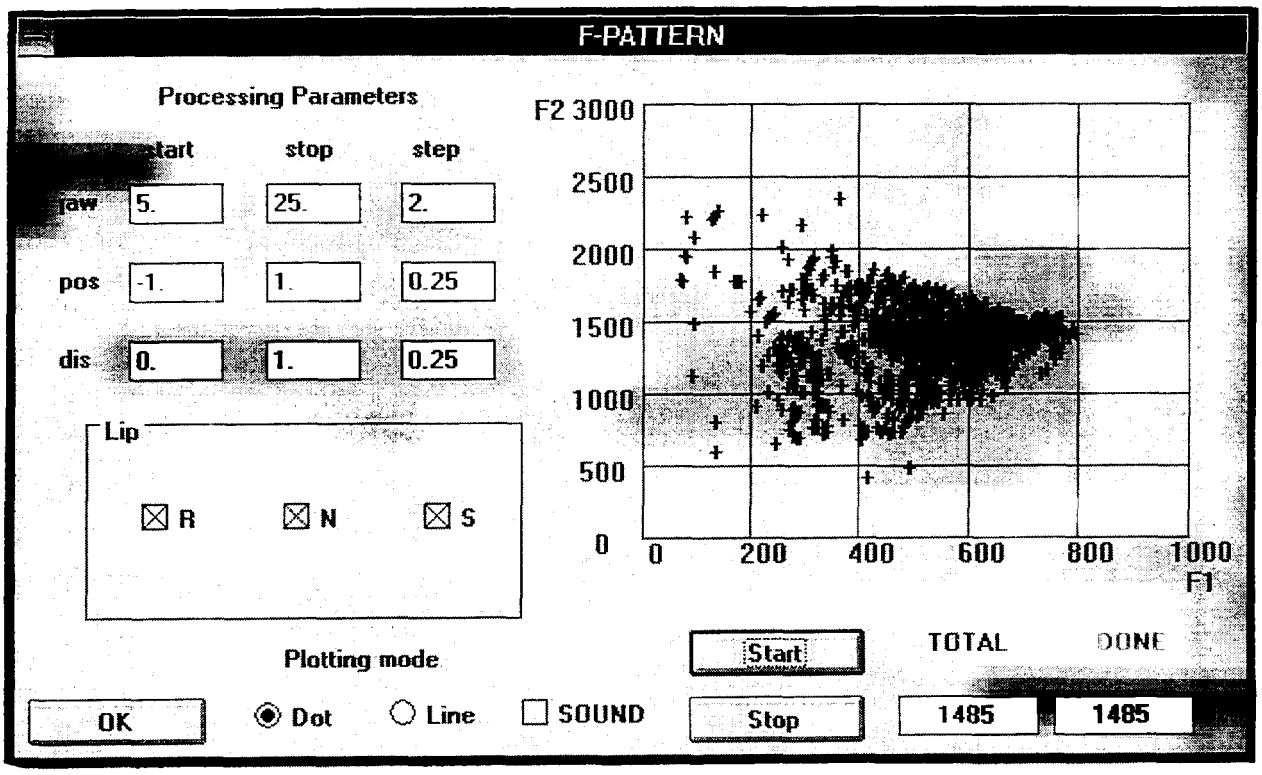

Figure 4. Screen dump of F1-F2 pattern generation.