CONTACT Review Meeting 1
Genova, November 14-15, 2006

From movements to sound
Contributions to building the BB speech production system

Lisa Gustavsson, Björn Lindblom, Francisco Lacerda & Elisabet Eir Cortes

Summary

In terms of anatomical geometry the infant Vocal Tract undergoes significant change during development. This research note reports an attempt to reconstruct an infant VT from adult data. Comparable landmarks were identified on the fixed structures of adult articulatory lateral profiles (obtained from X-ray images) and matching infant profiles (obtained from published data in the literature, Sobotta [Putz & Pabst 2001, and personal communication from author Prof. Dr. med. R. Pabst]. The x-coordinates of the infant landmarks could be accurately derived by a linear scaling of the adult data whereas the y-values required information on both the x- and the y-coordinates of the adult. These scaling rules were applied to about 400 adult articulatory profiles to derive a set of corresponding infant articulations. A Principal Components Analysis was performed on these shapes to compare the shapes of the infant and adult articulatory spaces. As expected from the scaling results the infant space is significantly compressed in relation to the adult space suggesting that the main articulatory degree of freedom for the child is jaw opening. This finding is in perfect agreement with published descriptions of the phonetics of early vocalizations.

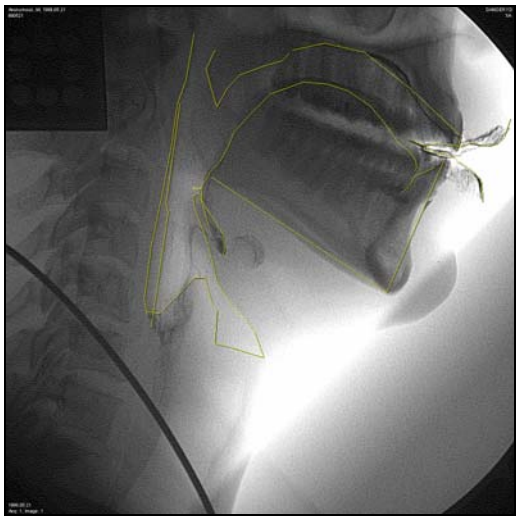

Figure 1 *Example of X-ray image with traced contours indicated. Points on the teeth, hard palate, posterior pharyngeal wall and laryngeal structures were selected for comparison with the corresponding points for the infant vocal tract.*

Analyses of adult X-ray data

Our X-ray data come from a 20-second film (figure above) of an adult Swedish male speaker [Branderud et al 1998]. The speech sample consists of about 20 test words containing consonants such as [b], [p], [d], [g], [l], [k]. [r], [j], [h], [n], [s], [t] and a representative sample of the Swedish long and short vowels. The images portray a midsagittal articulatory profile. They were sampled at 50 frames/sec. A total of about 400 frames were analyzed. Tracings of all acoustically relevant structures were made using the OSIRIS software package (University of Geneva). Using specially written software we converted the contours defined in Osiris into tables with x- and y-coordinates, calibrated in mm and corrected for head movements. For the tongue, the contours were further processed by redefining them in a jaw-based coordinate system and by resampling them at 25 equidistant 'fleshpoints'. This resampling was motivated by our choice of quantification method, Principle Components Analysis.

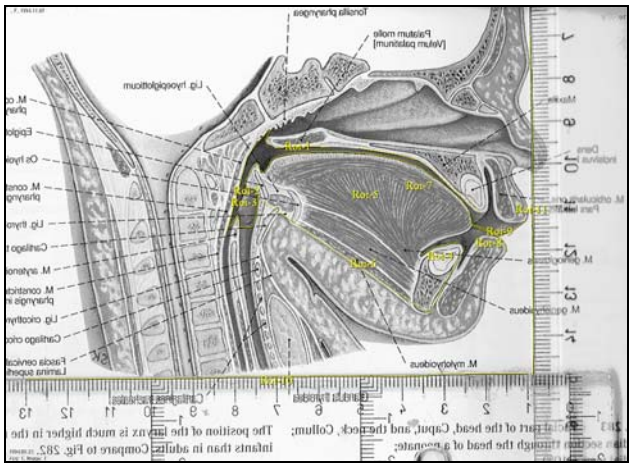

Figure 2 *Image used to determine the location of landmark points in infant anatomy. Scales indicated by rulers.* *With permission from* Putz / Pabst: Sobotta, Atlas der Anatomie des Menschen © 2005 Elsevier GmbH, Urban & Fischer Verlag München

The adult and infant vocal tracts: A scaling experiment

A scaling function of the vocal tract (adult-infant, infant-adult) was derived from tracings of midsagittal images of the infant-VT (figure above) and the x-ray images of the adult-VT (see Analyses adult x-ray data). Since there is a non-linear relationship between infant and adult VT:s, we needed to make accurate estimates of infant anatomical structures, rather than apply a simple linear reduction of an adult model. Using empirically determined scaling-functions we could transform a set of adult articulations (see Analyses adult x-ray data) to BabyBot articulatory settings and define its articulatory space (see Articulatory parameters: Principal components analysis of X-ray data).

The x- and y-coordinates for the BabyBot VT were derived using:

$$X_{BB}\ (horiz) = 0.765 * X_{adult}$$

$$Y_{BB}(vert) = -0.43 + 0.32 * Y_{adult} - 0.15 * X_{adult}$$

Figure 3 *Black lines pertain to the fixed structures of an adult articulatory profile. The red contours represent the corresponding structures in an infant-like vocal tract obtained by applying the two equations specified next to the diagram to the adult landmarks.*

The relationship of the front-back dimension between the infant-VT and the adult-VT was more or less linear which allowed us to use only one variable (x-coordinates) to derive BabyBots x-coordinates. In the vertical dimension however, two independent variables were needed (x- and y-coordinates) to derive BabyBots y-coordinates. This is probably because one of the most critical aspects of the growing VT is the height of larynx, the pharyngeal dimensions are close to zero in an infant while it is one of the major areas in the adult-VT.

Articulatory parameters: Principal components analyses

The input to the PCA consisted of a matrix with columns corresponding to the 25 fleshpoints and rows containing information related to the individual tongue contours. Since the specification of each fleshpoint requires two numbers (x & y), there were twice as many rows as contours. Accordingly, the data fed into the PCA was a 822-by-25 matrix. This format had the convenience of automatically

sorting the PCA output into one set of horizontal weights (for the x coordinates) and one set of vertical weights (for the y coordinates).

As earlier shown by several other phoneticians [Maeda 1990], the PCA provides considerable data reduction by quantifying input data in terms of a small set of building blocks, the PC's. Accordingly, the 25 fleshpoints of an observed shape, s(x), can be recovered by calculating

$$s(i,x,v) = w1(i,v)*PC1(x) + w2(i,v)*PC2(x) +... \qquad (1)$$

where x is fleshpoint number, i identifies the contour/image, and v chooses between x or y coordinates. The PC(x) terms are underlying, numerically derived tongue shapes which, weighted by the w coefficients and summed, generate the contour under examination. The formula expresses the idea that any observed contour is a linear combination of a set of basic shapes. The accuracy of this quantitative description depends on how many PC's are used. Any degree of accuracy is possible in principle. For the present data, PC1 was found to account for 85.7 % of the variance. Two components achieved 96.3%.

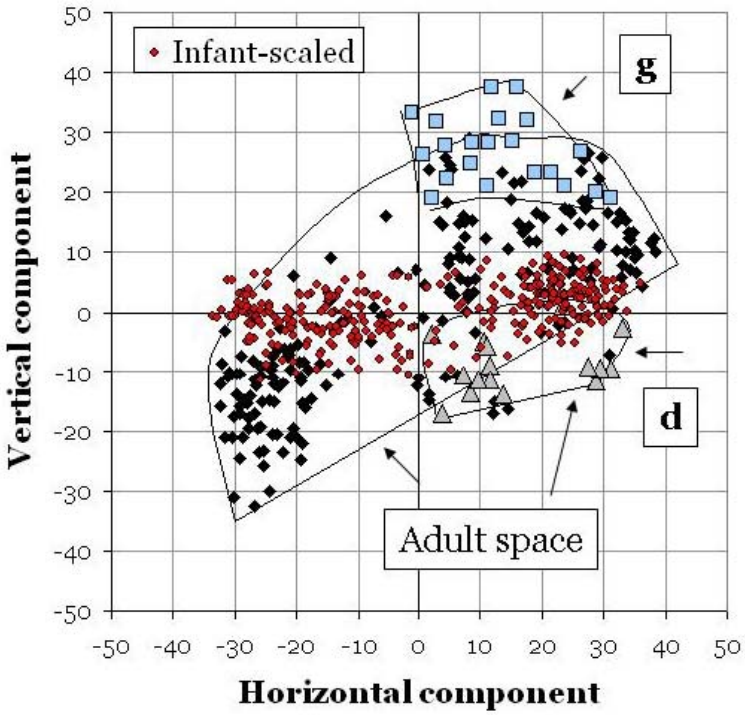

Figure 4 *The above figure presents the results of a PC analysis of adult and infant tongue shapes. Along the ordinate: the vertical component of PC1 (value of weight for y coordinate). Along the abscissa: the horizontal component of PC1 (value of weight for y coordinate). For the adult data the locations of [d] and [g] tokens and vowels are indicated. For comparison the infant-scaled version of the entire database is shown with red dots. As can be seen there is considerable compression particularly along the vertical dimension.*

A short-cut method of formant frequency derivation

The first step of deriving the acoustic consequences of articulatory movements is to quantify the vocal tract shape in terms of an cross-sectional area function. This is done by measuring the cross-distance along the VT profile and then converting those distances into cross-sectional areas using empirically based rules. (see figure below). From the area function the formant frequencies of the articulation are then calculated. This requires a certain amount of computation.

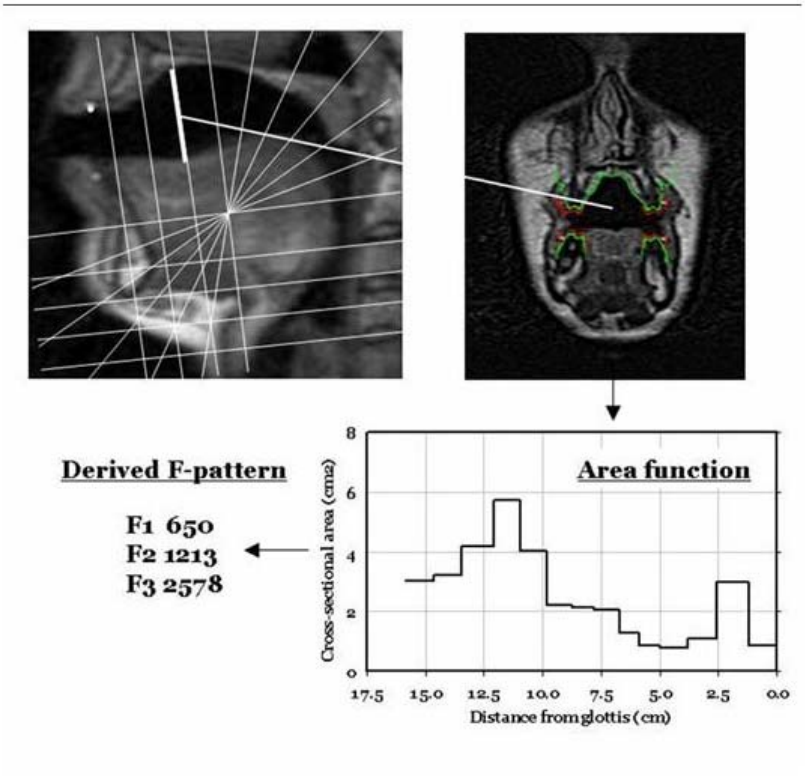

Figure 5 *The standard method of calculating the formant frequencies for an arbitrary articulatory configuration. The articulatory profile (top left) is first analyzed with respect to the cross-distances along the vocal tract. A cross-sectional area (A) corresponding to a given distance (d) can be described in terms of power functions of the form* $A=\alpha*d^{\beta}$*. The area variations along the vocal tract, the so-called "area function", is then used to compute the formant frequencies of the articulation. This fairly cumbersome procedure is the standard way of going from movement to sound.*

Using our X-ray data we have done some pilot work exploring the possibility of using empirical mapping rules to speed up and simplify the step from articulation to sound.

Below we exemplify this method with the test utterance Johan spoken at loud voice. Pulse by pulse measurements were made of formant frequencies. From the X-ray analyses synchronized data are available on the time variations of the first two PC:s of the tongue contour, the degree of jaw opening, larynx height, vertical separation and protrusion of the lips. Using a multiple regression technique we derived predictions of each formant individually from linear combinations of the

above articulatory parameters. Very high numerical accuracy was obtained with absolute error scores below 2%.

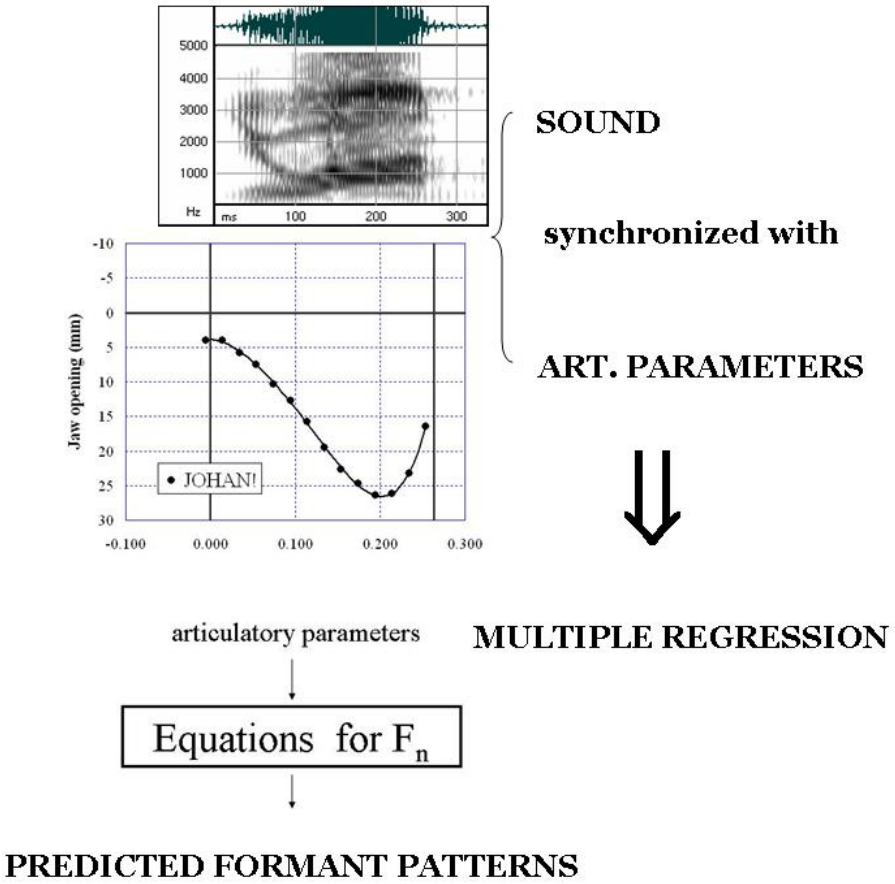

Figure 6 *A short-cut method. Deriving formants from articulatory data using empirically established equations describing the relationship between formant frequencies and articulatory parametric tracks.*

References

Branderud P, Lundberg, H-J Lander J, Djamshidpey H, Wäneland I, Krull D & Lindblom B (1998): "X-ray analyses of speech: methodological aspects", in *Fonetik 98: Papers presented at the Swedish Phonetics Conference*, Stockholm University, 1998.

Lindblom B (2003): "A numerical model of coarticulation based on a Principal Components analysis of tongue shapes", *XVth International Congress of Phonetic Sciences*, Barcelona, Spain.

Maeda S (1990): "Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model", in *Speech Production and Speech Modeling*, W Hardcastle & A Marchal (eds), pp 131-149, Dordrecht:Kluwer. 1990.

Putz R & Pabst R (2001) *Sobotta – Atlas of Human Anatomy – Head, Neck, Upper Limb*, Putz R & Pabst R (eds), Volume 1, 13th edition, Urban&Fisher.