Articulatory copy synthesis: Acoustic performance of an MRI and X-ray based framework

Christine Ericsdotter
Department of Linguistics, Stockholm University
E-mail: ericsdotter@ling.su.se

ABSTRACT

Within the framework of refining an articulatory model, articulatory data were collected using MRI, video and X-ray. A set of two-to-three dimensional rules were derived from the axial and coronal MRI images and applied to midsagittal X-ray images. The acoustic performance of the rules was evaluated by synthesizing observed and predicted area functions.

1. INTRODUCTION

Articulatory synthesis is of ultimate interest to phoneticians, but due to the multi-facetted character of articulatory data and to the fact that articulatory patterns are mostly hidden from direct observation, the understanding of the ways in which the vocal tract changes three-dimensionally during speech production has been too scattered to form a basis for the development of effective articulatory models. The use of modern techniques such as ultrasound and MRI provide relatively non-intrusive, non-invasive methods of capturing information about vocal tract shape, and has made the body of three-dimensional speech data grow steadily. The present study is a summary of a contribution to that body of data, describing a study of Swedish vowels using MRI, digital video and sound recordings. The aim of the study is to explore the possibility of economizing three-dimensional articulatory information by defining a set of rules for prediction of the third dimension from profile vocal tract representations (“distance-to-area rules”, d2A, see e.g. [1]) for implementation in the APEX articulatory model.

The APEX articulatory model is a tool for exploring articulatory-acoustic relationships in a vocal tract, excluding the nasal cavity. The framework was presented by Lindblom and Sundberg [2] and implemented under Windows by Stark [3]. The model is designed mainly for linguistic theory research, which implies it has to meet up to the demands of physiologically realistic speech organ properties and theoretically unbiased parameters and degrees of freedom. The recent development of the model has been directed towards calibrating it so as to map out the articulatory behaviour of specific subjects onto articulatory parameters. The procedure for developing the model according to these demands has been defined in the project as (a) analyse the articulatory behaviour of two subjects; (b) calibrate the existing articulatory model with the anatomical properties of these subjects; (c) map the articulatory configurations into physiological parameters, and (d) extrapolate non-observed articulations from these parameters. The present study is presenting the work within (a) above with respect to vowels, and is in line with other work of the group [4].

Distance-to-area rules. Studies exploring distance-to-area relationships in the vocal tract conclude that these kinds of rules are highly speaker specific, and that the coefficients differ largely between places in the vocal tract. This means that a set of rules defined for one speaker can not be used for translating articulatory profiles to 3D tubes for another speaker.

Midsagittal representation of the vocal tract together with 2- to 3-dimension rules are however the most common way of controlling articulatory models, including the APEX model. It can be discussed whether this approach is phonetically sufficient to build realistic 3D vocal tract tubes, but recent data by Engwall [5] show that there is a strong correlation between midsagittal and global shape of the tongue, although this relationship was not specified explicitly.

2. DATA ACQUISITION

Subjects. One female (the author) and one male subject in their twenties were used in this study. The male subject was bilingual in Swedish and Polish and had no phonetic training, but some musical training. The female subject was monolingual in Swedish and had phonetic and musical training.

2.1 MR images

Calcified structures such as teeth or bone are displayed as dark areas on MR images due to their low concentration of hydrogen. These kinds of structures are therefore hard or impossible to distinguish from air-filled spaces, which are also displayed as dark areas. In previous investigations, external data from plaster casts have been used to estimate the volume taken up by the teeth in the oral cavity. In this study, dental casts filled with gadolinium (a contrast medium for MR images) were used to derive the shape and size of the teeth, and to account for head movements. These dental reference casts were made in the following way: Plaster casts of the subjects’ lower teeth and hard palate were made by a dentist. Plastic tubes with an outer diameter of 0.7 mm were glued along the external sides of the plaster teeth. Thin plastic casts were made from the plaster cast

ISBN 1-876346-48-5 © 2003 UAB

with the attached tubes. Tubes of the same dimension and length as those glued to the plaster casts were then filled with water and gadolinium, welded up and fit into the plastic casts, which were fitted to the subject's teeth and worn throughout the whole experiment.

The MRI experiment was carried out at the Unité de Résonance Magnétique de l'Hopital Erasme in collaboration with Université Libre de Bruxelles. The magnetic field strength of the MR machine was 1.5 T, the slice thickness 4 mm. The subject was placed in supine position on the machine's sliding stretcher, with the head inside the coil. Cushions were placed under and around the subject's head, neck and shoulders for stability and comfort. Soft earplugs were used to protect the subject from the loud MR machine noise and to give the subject feedback of vowel quality and stability. Three repetitions of an 11 Swedish vowel list were acquired with sagittal images (11 slices, acquisition time 14 s/vowel), the same materials were acquired with multi images (14 or 15 coronal and or axial slices, acquisition time 13 or 14 s/vowel). The subject started phonating the vowel 1 second prior to the imaging acquisition and sustained phonation until 1 second after image acquisition, to acquire undisturbed sound recordings from both the start and end of the vowel. A reference cube with attached gadolinium filled tubes was also recorded with multi and sagittal MR images to verify scaling and possible image distortion effects in x y pixel spacing.

Simultaneously with the MR image acquisition, digital video images were collected of the subject's lip configurations via a mirror in the receiver coil. Two video cameras were placed on each side of the sliding stretcher. The subject's lips were coloured with purple lipstick prior to the experiment. A reference grid was placed onto the lips prior to the MR imaging acquisition, to render scaling possible.

Sound recordings were also carried out simultaneously with the MR image acquisition. There have been two main problems with acquiring high quality acoustic recordings during MR image acquisition. First, since no metal is allowed within the magnetic field of the machine, microphones have had to be placed at a critically long distance from the subject. Second, the MR machine generates a loud noise during image acquisition, disturbing the sound recordings considerably. These problems were approached in the present study by the construction and use of an optical microphone connected to a DAT recorder via a fibre cable. The microphone and cable did not contain any metals and could be taped to the subject's chin, directed at the corner of his mouth. The DAT recorder was placed at a safe distance from the magnetic field.

Post-processing of the data. The MR images were upsampled and processed with a pixel sharpening filter. Cross-sections of dental contours were integrated into each image in the oral cavity (Figure 1). This integration was made possible by the information on teeth location provided by the representations of the gadolinium tube in the images (white dots). Using the results from that procedure, the subjects' jaw paths could be derived and head movements could be checked. The male subject was found to perform negligible head movements during the experiment, while the female tended to move her head backwards when producing open vowels.

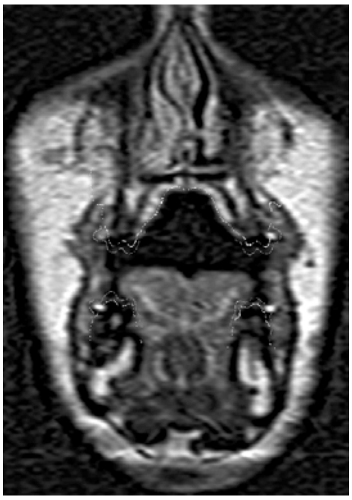

Figure 1. Cross-sections of dental contours (white) integrated in an image from the oral cavity.

Three images per vowel were extracted from the video image sequence at initial, medial and final point of the vowels. The images were upsampled and corrected for camera angles.

The sound recordings were transferred from DAT tapes to Cool Edit (http://www.syntrillium.com) using a Sound Blaster Live sound card, and downsampled to 16 kHz mono files. A copy of the material was made where the MR noise was filtered from the speech signal using noise profile filtering, a filtering method where a portion of the noise disturbance is used as a filter pattern to cancel itself out.

2.2 X-ray images

Digital lateral X-ray images were collected at Danderyd Hospital in Stockholm at an earlier stage in the project (described in [6] and [7]). The radiation dose was very low and a barium sulphate paste was used to enhance the contours of the tongue and lips. The speech materials consisted of real and nonsense Swedish words.

Post-processing of the data. In the X-ray data the subject was recorded in upright position. The head position was therefore quite different between the MR and the X-ray images. To make data more comparable, the X-ray images were rotated and translated so as to make the palate in upright position fit with the palate in supine position. The MR plane positions were then overlaid on the X-ray images, so that distances could be defined at approximately the same places along the vocal tract.

3. ANALYSIS

MR images. All articulatory analyses were made manually, using the same zooming settings and lighting conditions in the room. On the MR images, the shape of the air filled spaces in the vocal tract were outlined, and their midsagittal distance were defined. In the laryngopharynx, a great deal of effort was directed towards correctly identifying anatomical structures so as to estimate larynx height correctly in relation to the MR plane position [8]. The choice of acoustically relevant areas is not unproblematic in the larynx region and the oral region. All air filled spaces were outlined, but in this first modelling attempt, piriform and side cavities were disregarded unless they were connected to the main vocal tract tube by a non-narrow passage. Two-to-three dimension rules were derived plane by plane, by fitting a power function from distance to area. In the larynx region, different rules apply to different larynx heights.

On the video images, the whole lip shape could not be traced because of the presence of a receiver coil bar in the left part of the lip area. The height and width of the lip opening could however be reliably defined, and they were used to derive the area, innermost point of contact of the lip corners and tangent of the lip plane, using equations derived from a reference video session recording the lips both from a profile and frontal projection. Two-to-three dimension rules were derived for rounded and spread lip configurations separately, by fitting a power function from distance to area as above. 12 out of the 16 d2A-functions derived from MR and video images showed an r^2 >0.90.

The acoustic recordings were analysed to form reference materials for the modelling, and to quantify articulatory stability throughout the vowel. Short term energy measurements, auditory impression of voice quality and timbre, manual formant measurements on DFT and LPC spectrum sections and automatic measurements of formant movements throughout the vowel were carried out. The f0 of both subjects were found to be affected by the pitch of the MR machine noise. The male subject showed greater general stability throughout the vowels than the female subject.

X-ray images. Distances were outlined along the overlaid MR plane positions. Formant measurements were carried out on the vowels.

3. APPLICATION

For the purpose of illustration, 3 different vowels [e], [a], [o] from the male subject will now be discussed regarding the accuracy of the distance-to-area rules, the application of the rules on the same set of data they were derived from, and the application of the rules on another set of data collected from the same subject.

Tubes from MR images. The centre of gravity of the observed areas of the vocal tract was used to define the length coordinate (x) of a vowel tube. The first section of each vowel tube were set to a fixed tube simulating the lowest part of the larynx [8]. The length coordinate of the last tube was set to the tangent value of the lips.

The first set of tubes (MR obs) was put together from the observed (x) and areas A(x) of the MR materials. The second set of tubes (MR pred) was put together from the observed (x) and by deriving A(x) from observed distances in the MR materials.

Tubes from X-ray images. To test the goodness of the d2A-functions on another set of midsagittal data, three frames of the chosen vowel qualities were chosen for analysis: initial [e] in [e:be:pe:k], third [o] in [o:bo:po:l] and the [a] in [ɛˈdas:]. The midline was outlined connecting the midpoints on the distance tracings.

Formant predictions. The area functions were upsampled to 1 mm tube lengths to avoid abrupt area changes. They were then run through a formant prediction program [9], and the F1 values were corrected so as to account for impedance of the vocal tract walls [10, page 159, equ.3.47]. The results are summarized in Table 1. The table contains observed formant values (obs. F) for the MR and X-ray vowel tokens. These are for the MR vowel the mean of the initial and final formant value. The next column shows the formant values (synt. F) for the synthesized vowels (MR pred, Xray) together with percentage difference from the observed values.

As can be seen from Table 1, two thirds of the predicted F values differ less than 10% from the observed values, and in more than half of the cases the discrepancy is less than or equal to 5%. Given inaccuracies in both the measurement of the formants in the original signals as well as indeterminacies in the tracings of articulatory data, 5% discrepancies can be regarded as negligible. However, there are some stark discrepancies, most notably F2 predictions in [a]. Given the very good estimation of F2 from the observed MR areas in [a] (2%) it can be concluded that the d2A-functions require refinement to account more adequately for all vowel types.

4. DISCUSSION

Defining and applying a first set of distance-to-area rules on midsagittal articulatory configurations from three articulatorily different vowels and two different sets of data generated generally good acoustic results. However, they can and will be improved by further analysis.

The first point to be improved is the definition of the vocal tract midline, since the position and length of constrictions in the vocal tract are of utmost importance for the acoustic outcome. Different midline derivation strategies produce quite different sets of length coordinates (x) and hence tube lengths [11]. The methods for deriving the midline in this study differed between data sets, which might explain some of the differences in the acoustic performance as the d2A-rules depend heavily on (x).

The overall length of the tract is also a crucial point for the acoustic outcome. Determination of the vocal tract end points is not unproblematic. A physiological experiment is planned to gain further insights into systematic effects of

ISBN 1-876346-48-5 © 2003 UAB

	Tube	Obs. F,	Synt. F	Δ obs
		(mean)		–pred.
F1	[e] MR obs	327	350	7%
	[e] MR pred		343	5%
	[e] X-ray	360	354	-2%
	[a] MR obs	694	670	-3%
	[a] MR pred		653	-6%
	[a] X-ray	635	569	-10%
	[o] MR obs	347	396	14%
	[o] MR pred		343	-1%
	[o] X-ray	390	377	-3%
F2	[e] MR obs	2154	2218	3%
	[e] MR pred		2025	-6%
	[e] X-ray	2060	1982	-4%
	[a] MR obs	1051	1069	2%
	[a] MR pred		1382	31%
	[a] X-ray	1175	1487	27%
	[o] MR obs	684	702	3%
	[o] MR pred		728	6%
	[o] X-ray	665	747	12%
F3	[e] MR obs	2521	2499	-1%
	[e] MR pred		2443	-3%
	[e] X-ray	2580	2633	2%
	[a] MR obs	2623	2765	5%
	[a] MR pred		2368	-10%
	[a] X-ray	2520	2499	-1%
	[o] MR obs	2531	2898	15%
	[o] MR pred		2347	-7%
	[o] X-ray	2620	2329	-11%

Table 1. Comparison of observed and predicted formant values. See text for details.

different lip configurations.

Experimentation with the area function for [o] (MR obs.) showed that an increase of the slice area directly anterior of the dorso-pharyngeal stricture brought about a striking decrease in the value of F3 (about 200 Hz). This is a reminder that regions in the vocal tract may be particularly sensitive to abrupt area changes in area along midline [12, page 120, Table 2.33-4]. In fact, when predicting the formants of the non-upsampled tubes and hence forcing more abrupt area changes onto the tract, the F3 predicted for [o] (MR obs.) was much lower and closer to the observed value. This implies that linear upsampling of the tube might be an oversimplification and worth looking into in more detail.

ACKNOWLEDGEMENTS

This research was funded by a National Institute of Health grant (R01 DC02014) and by a Stockholm University research student's grant (SU 617-0230-01). The author is grateful to Thierry Metens, Didier Demolin and Alain Soquet for making possible the acquisition of MR images, to Hans-Jerker Lundberg and Jaroslava Lander for the acquisition of X-ray images, and to Johan Stark and Peter Branderud for collaboration on the teeth fitting procedure.

REFERENCES

[1] J. M. Heinz and K. N. Stevens, “On the derivation of area functions and acoustic spectra from cinéradiographic films of speech”, *Journal of the Acoustical Society of America, Sixty-seventh meeting,* pp. 1037-1038, 1964.

[2] B. Lindblom and J. Sundberg, “Acoustical consequences of lip, tongue, jaw and larynx movement”, *Journal of the Acoustical Society of America, vol 50,* pp. 1166-1179, 1971.

[3] J. Stark, B. Lindblom and J. Sundberg, “APEX: An articulatory synthesis model for experimental and computational studies of speech production”, *KTH STL-QPSR*, vol 2, pp. 45-48, 1996.

[4] J. Stark, C. Ericsdotter, P. Branderud, J. Sundberg, H.-J. Lundberg and J. Lander, “The APEX model as a tool in the specification of speaker-specific articulatory behaviour”, *The XIVth ICPhS, San Francisco, California,* 1999.

[5] O. Engwall, *Tongue Talking. Studies in Intraoral Speech Synthesis.*, Doctoral Dissertation, Speech, Music and Hearing, Royal Institute of Technology, 2002.

[6] P. Branderud, H.-J. Lundberg, J. Lander, H. Djamshidpey, I. Wäneland, D. Krull and B. Lindblom, “X-ray analyses of speech: Methodological aspects”, *FONETIK 98, Papers presented at the annual Swedish phonetics conference, Dept of Linguistics, Stockholm University,* 1998.

[7] C. Ericsdotter, J. Stark and B. Lindblom, “Articulatory coordination in coronal stops: Implications for theories of coarticulation”, *Proceedings of the XIVth ICPhS, San Francisco, California,* 1999.

[8] C. Ericsdotter, “Determining vertical larynx height on axial MR images”, *AQL 2003 - Advances in Quantitative Laryngology, Voice and Speech Research, VIth international conference, Hamburg,* (forthcoming).

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

[10] K.N. Stevens, *Acoustic Phonetics*, Cambridge, Massachusetts, The M.I.T, 1998.

[11] U. G. Goldstein, *An articulatory model for the vocal tracts of growing children*, Doctoral Dissertation, Department of Electrical Engineering and Computer Science, Massachusetts Institute for Technology, 1980.

[12] G. Fant, *Acoustic Theory of Speech Production*. The Hague, Mouton, 1960.