A METHODOLOGY FOR MODELLING AND INTERACTIVELY VISUALISING THE HUMAN VOCAL-TRACT IN 3D SPACE

Michael Barlow[1], Frantz Clermont[1] and Parham Mokhtari[2]
[1]School of Computer Science, University College, University of NSW
[2]Electrotechnical Laboratory, Tsukuba, Japan

ABSTRACT. A system is described for constructing and visualising three-dimensional (3D) images of the human vocal-tract (VT), either from directly-measured articulatory data or from acoustic measurements of the speech waveform. The system comprises the following three major components: (1) a method of inversion for mapping acoustic parameters of speech into VT area-functions, (2) a suite of algorithms which transform the VT area-function into a 3D model of the VT airway, and (3) solutions for immersing the 3D model in an interactive visual environment. The emphasis in all stages of modelling is to achieve a balance between computational simplicity as imposed by the constraint of real-time operation, and visual plausibility of the reconstructed 3D images of the human vocal-tract.

1. INTRODUCTION

Vocal-tract (VT) modelling has long received considerable research efforts, often because it is seen as a step-above the acoustic signal in the communication chain; being closer to the original communicative intent and hence being a more compact as well as a richer (or at least less convolved) source of knowledge. Application areas of VT modelling include speech synthesis, speech coding, speech recognition and speaker characterisation. Whilst previous approaches to VT-modelling have ranged from acoustically- and/or physiologically-motivated mathematical models (e.g., Mermelstein and Schroeder, 1965; Lindblom and Sundberg, 1971; Mermelstein, 1973; Coker, 1976) to parameterisation of direct measurements made by magnetic resonance imaging (MRI) or ultrasound imaging techniques (e.g., Harshman et al., 1977; Yehia et al., 1996; Story and Titze, 1998), the so-called "inverse problem" of estimating the VT geometry from the acoustic speech signal has long been regarded as a potentially revolutionary approach with far-reaching applications. However, it appears that theoretical and practical problems such as non-uniqueness and articulatory compensation have limited the use of estimated VT-shape (or area-function) data mainly to theoretical investigations of the inverse problem itself.

Our paper describes an application-driven approach to the inverse problem, with the long-term goal of plausibly realistic 3D vocal-tract models, constructed in real-time (i.e., as the user phonates). Key to our methodology is the real-time constraint with its implications for the complexity of the model that can be supported, coupled with a relaxation of the exactness/uniqueness criteria. Our system has applications in areas such as foreign-language acquisition or rehabilitation of speech pathologies, where it would be a distinct advantage for users to receive real-time visual feedback on the difference between their own VT configuration and the ideal or standard one being attempted.

The system proposed herein consists of three major components: (1) a computationally tractable VT model and method of acoustic-to-articulatory mapping (or inversion) which is used to estimate VT area-functions from the acoustics of speech; (2) a suite of 3D-modelling software used to transform an estimated area-function into a 3D polygon mesh (surface of a straight tube with varying cross-sectional areas), and then to apply such spatial transformations as to make the model conform more closely to human vocal-tract anatomy, thereby yielding a more *plausible* 3D reconstruction of the VT-shape; (3) an environment in which to present these models to the user so that they may be interactively explored. These three major components are elaborated in the remainder of this paper, and followed by a concluding discussion.

2. ACOUSTICS TO AREA-FUNCTION

The first major component of our system comprises a computationally tractable model of the VT and a method of mapping from acoustic to model parameters. As an initial step, the method selected involves a VT-shape parameterisation first introduced by Mermelstein and Schroeder (1965). In particular, the logarithmic area-function is parameterised in terms of the first few odd-indexed terms of the cosine-series, as follows:

$$\ln A(x) \approx \ln A_0 + \sum_{n=1}^{M} a_{2n-1} \cos((2n-1)\pi x / L) , \qquad (1)$$

where x is the distance along the VT airway from the glottis to the lips, L is the total length of that distance, is an area scaling factor whose value is computed to retain an overall mean logarithmic-area equal to zero, and M is the number of terms retained in the series. The elegant simplicity of this model lies in the following, quasi-linear relation which maps the n^{th} formant frequency (n^{th} resonant frequency of the vocal tract) to the corresponding model parameter a_{2n-1}:

$$a_{2n-1} \approx -2\frac{(F_n - F_{n0})}{F_{n0}} \quad , \qquad (2)$$

where $F_{n0}=(2n-1)c/4L$ is the n^{th} formant frequency of a uniform area-function of the same length L, and c=35300 cm/sec is the speed of sound in the VT airway.

The approximate equality in Equation (2) becomes increasingly inexact as the formants are more distant from their neutral values. The iterative method proposed by Mermelstein (1967) is therefore used to ensure that resynthesis of the formants using a completely lossless VT acoustic model will exactly reproduce the target formant frequencies. Furthermore, the VT-length L is optimised such that the inversion procedure yields a VT-shape with minimal eccentricity from that of a uniform tube. In particular a number of potentially realistic VT-lengths are considered and the corresponding VT-shapes derived. These shapes are contrasted, as an RMS difference, with the uniform neutral tube, with the VT-shape/VT-length that yielded the minimum difference then being selected.

While the above method is preferred owing to its computational simplicity and the fact that the model parameters are directly and uniquely related to acoustic (formant) parameters, our modular approach to the system design can, in principle, accommodate any reasonable method of acoustic-to-articulatory mapping that yields a plausible estimate of the speaker's VT area-function without undue computational complexity.

3. 3D MODELLING

The second major component of our system is responsible for generating a plausible 3D model based on the area-functions estimated by the first component. Notably, this second component has been deliberately designed to be independent of any particular algorithm for estimating the area-functions, and indeed directly-measured area-functions could in principle be substituted as input. In line with this modular approach, the system therefore assumes that the area-functions correspond to a simple tube of varying cross-sectional areas, which may subsequently be transformed to a more complex shape more closely matching that of the human vocal apparatus. The process (an example of which is shown later in Figures 2 and 3) involves the three subsystems of piecewise-tube construction, imposition of VT structural scaling, and VT path of airway transformation, as described respectively in the following three subsections.

Piecewise Tube

The first subsystem of the 3D modelling component generates, for each input area-function, a 3D linear tube of concatenated, circular sections of varying diameter. As shown in the left panel in Figure 1, a number of 2D circles in the x-y plane are defined, where each circle's radius is calculated directly from the corresponding section of the area-function, and where the number of (equi-spaced) points defining the perimeter of the circle (as x-y pairs) can be adjusted by the user (the more points, the smoother the image). At this stage of modelling, the entire tube is a straightened-out version of

Figure 1: An example of the first stage in the 3D modelling process. Sets of vertices are constructed from a series of points which define circular VT segments. These are then connected in triangle strips. The model shown here was constructed from an MRI-measured VT area-function (Yang & Kasuya, 1994) of a young Japanese male phonating /a/, and consists of 340 vertices composing 640 triangles.

the vocal tract, such that all points on a given circle have a constant z-coordinate which depends only on the distance of the corresponding cross-sectional area along the VT airway from the lips.

The user may also specify a different number of segments (circles) along the VT length, compared with the number of sections provided by the input area-function. In that case, cubic spline interpolation is used to provide the intermediate area values. This approach has the distinct advantage of yielding smoothed VT-shapes, thereby transcending the artificial, step-wise values generated by some methods of inversion (e.g., by the linear-prediction method), or indeed by direct measurement methods such as MRI. However, in that context it is important to note that the Mermelstein-Schroeder model summarised in Equations (1) and (2) already provides an area-function which is smooth, and which can be sampled at any desired number of points along its length.

As shown in the right panel in Figure 1, the points defining the 2D circles then form the vertices of a set of triangles that connect successive points around each circle as well as those points on adjacent circles, in a triangle strip configuration. For a model with s segments (circles) along the length of the vocal tract and p points around each segment (circle), the entire, 3D surface of the VT-airway is defined by a total of s.p vertices and $2p(s-1)$ triangles (polygons). Typical ranges of the parameter-values yielding visually very smooth shapes are 40-60 for s and 20-40 for p, yielding a model with 1500-4500 polygons.

Structural Scaling

The second subsystem of the 3D modelling component applies transformations to the circular cross-sections, with the aim of obtaining cross-sectional shapes that more closely match human anatomy. As a first approximation (see Figures 2 and 3), Lindblom and Sundberg's (1971, p.1173) numerical values for the 2-constants power-model are used. That model assumes somewhat more realistic, *elliptical* cross-sectional shapes, and defines three regions along the length of the VT within which the cross-dimensions of the airway in the midsagittal plane can be computed from the cross-sectional area values by a power-relation involving only two constants. Similarly to the other subsystems, spline interpolation of those constants is employed to extend Lindblom and Sundberg's original 3-region model to a smoothed set of constant-pairs defined at each VT segment. Once the cross-dimension in the midsagittal plane is thereby determined at each section, the cross-sectional area can be used to compute the transverse dimension of the ellipse.

Path of Airway Transformation

The first two stages of the 3D modelling component yield a tube which is horizontally straight (in the z-dimension), and which therefore is visually still considerably different from the human vocal-tract. In particular, the human vocal-tract does not remain horizontal throughout its length but follows a path from lips to glottis in which it rotates through more than 90 degrees as well as experiencing varying horizontal and vertical offsets in the sagittal plane.

In order to model the VT centreline, and hence the gross shape of the vocal-tract itself, a set of direct measurements is required, which should correspond to the orientation and displacement (from the lips) of the VT at regular intervals. These were obtained by hand measurement of published midsagittal profiles for neutral vowels in an MRI-based study (Story et al., 1996, p.542, Fig.2). Measurements of the VT centreline were encoded as a set of value triplets taken at equi-spaced intervals along the length of the vocal-tract. For each such point the horizontal offset from the lips, vertical offset from the lips, and angle of rotation relative to the horizontal were recorded.

These values were then used to define a set of transformations to the 3D model's vertices, such that the new set of vertices are morphed (moved and rotated) to follow the centreline. As a first step, spline interpolation was again employed to ensure that there were as many centreline triplet values as there were segments in the VT model. All vertices corresponding to a particular segment of the VT were then transformed as per the corresponding centreline triplet, as follows:

$$v' = T_{Z\text{-}offset} T_{Y\text{-}offset} T_{X\text{-}rotation} T_{Z\text{-}zero} v \; . \qquad (3)$$

Equation 3 specifies the transformation as a chained set of homogeneous matrix operations (transformations) on the original vertex to derive the new vertex v'. In particular, each vertex is first translated in the z-direction onto the same plane as the lips ($T_{Z\text{-zero}}$), then rotated about the horizontal axis ($T_{X\text{-rotation}}$), and finally translated (or offset) in the vertical ($T_{Y\text{-offset}}$) and horizontal ($T_{Z\text{-offset}}$) directions (in the sagittal plane). The bottom panel in Figures 2 and 3 illustrates the result of these transformations, which yield a visually more realistic 3D image of the supralaryngeal VT airway.

4. MODEL VIEWING AND INTERACTION

In order to make full use of the 3D models constructed by the processes described above, it was felt that users should also have the opportunity to dynamically interact with and explore them. The purpose of the third major component of our system is therefore to enable visual interaction, whereby the user can view, interact with, move through, and manipulate the 3D model created by the first two components. To achieve this goal, two complementary approaches were taken: one employing an existing Web 3D standard and emphasising

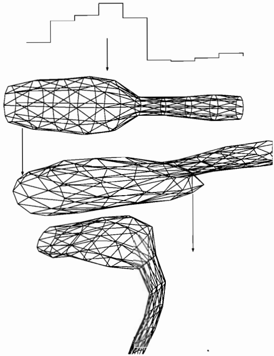

Figure 2: The three stages of the 3D modelling system, shown in wireframe form, from top to bottom. A VT area-function is first transformed into a piecewise tube with circular cross-sections. VT-structural scaling is then imposed which transforms the circular cross-sections to more plausible ones while retaining the area value at each section. Finally, a model of the VT airway is used to transform the straightened tube into a more natural, bent VT-shape.

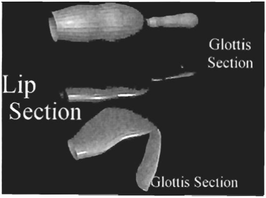

Figure 3: The three stages of the 3D modelling system, shown in solid-shaded mode, taken as a snapshot of a VRML world. The three models are the result of all three components of the system: acoustics-to-area function, 3D modelling (at various levels of complexity), and model viewing. The input data for the models consisted of the first 4 formant frequencies {669, 1241, 2736, 3356}Hz of an /a/ vowel from an adult Japanese speaker.

accessibility, the other employing immersive virtual reality technology.

The first, widely accessible solution, generates the model as VRML (Virtual Reality Modelling Language), a World Wide Web standard for 3D (ISO/IEC, 1997). Hence the models are made available on the Web (ref. URL-1) and can be viewed via any one of a number of free plug-ins for browsers such as Netscape Navigator or Microsoft's Internet Explorer. This provides access to anyone with a modern PC and internet connection, and, combined with the features of VRML, provides new opportunities for research and education in this area of speech processing (Barlow & Clermont, 2000). Figure 3 is an example of one such VRML model.

The second solution employs an immersive projection theatre known as the WEDGE (Gardner et. al., 1999). Images are back-projected in stereo onto twin-screens (each 2.7 metres wide by 1.5 metres high) meeting at right-angles, providing a semi-immersive sense of presence. Within the Vee formed by the screens, the images, viewed through stereo goggles, are perceived as having true depth and are seen to 'float' off the screens. As for the VRML solution, users may manipulate and interact with the WEDGE-projected 3D VT model. However, the WEDGE system does not currently support viewing of VRML data, and hence the 3D models are saved in LightWave .obj format, for which a viewer is available.

5. DISCUSSION

We have proposed a methodology for modelling and interactively displaying a 3D representation of the human vocal-tract. While more sophisticated models of the vocal-tract which include the nasal cavities, sinus piriformus, and other physiological and articulatory structures have previously been proposed in the literature, the aim of our simplified modelling approach is to present a 3D representation which is at once visually plausible and computationally inexpensive to construct. Specific real-time applications envisaged for the model include foreign-language learning and training of individuals with speech pathologies, both of which would certainly benefit from a real-time computer display of the 3D VT-shapes produced by the user during phonation of certain speech sounds.

A number of avenues for further research remain and are currently being explored. Within the 3D modelling component a structural scaling system is being developed that employs MRI data to accurately model the fixed structure of the upper-palate. This will further improve the plausibility of the 3D model, which at times appears stretched and narrow due to the elliptical transformation currently employed to perform structural scaling. Efforts are also underway to reduce the complexity, in terms of number of polygons, of the 3D model by reducing the number of vertices in areas of near-linear shape. For the acoustic to area-function subsystem, alternates to formant frequency inputs are being explored. Perhaps most importantly, FEM (Finite Element Method) and perceptual experiments are planned to validate the models generated, with particular emphasis in the perceptual trials on the utility of the model for language acquisition.

Finally, a MATLAB implementation of the software for the 3D modelling and visualisation of the vocal tracts based on area-function data is freely available at URL-2. It consists of a suite of functions for carrying out the various stages in modelling outlined above, i.e., transforming area-functions to a variable-width model with circular cross-sections, morphing on the basis of VT-structure and -centreline, and viewing.

REFERENCES

Barlow M, and Clermont F. (2000) "Seeing is Believing: Beyond a Static 2D-View of Formant Space for Research and Education", *Proc. Eighth Australian Int. Conf. on Speech Science and Tech.*, Canberra, Australia, 118-123.

Coker, C. H. (1976). "A Model of Articulatory Dynamics and Control", *Proc. IEEE* **64**, 452-460.

Gardner, H.J., Boswell, R.W., and Whitehouse, D. (1999). "The WEDGE Emmersive Projection Theatre", *Proc. 4th International SimTecT Conf.* 383-385.

Harshman, R., Ladefoged, P. and Goldstein, L. (1977). "Factor analysis of tongue shapes", *J. Acoust. Soc. Am.* **62**, 693-707.

ISO/IEC (1997) "VRML 97", International Specification ISO/IEC IS 14772-1, www.vrml.org.

Lindblom, B. E. F. and Sundberg, J. E. F. (1971). "Acoustical Consequences of Lip, Tongue, Jaw, and Larynx Movement", J. Acoust. Soc. Am. 50, 1166-1179.

Mermelstein, P. (1967) "Determination of vocal-tract shape from measured formant frequencies", *J. Acoust. Soc. Am.* **41**, 1283-1294.

Mermelstein, P. (1973). "Articulatory model for the study of speech production", *J. Acoust. Soc. Am.* **53**, 1070-1082.

Mermelstein, P., and Schroeder, M. R. (1965). "Determination of smoothed cross-sectional area functions of the vocal tract from formant frequencies", *Proc. 5th Int. Cong. Ac.*, Liège, Paper A24.

Story, B.H., Titze, I.R., Hoffman, E.A. (1996). "Vocal Tract Area Functions from Magnetic Resonance Imaging", *J. Acoust. Soc. Am.* **100**, 735-554.

Story, B. H. and Titze, I. R. (1998). "Parameterization of vocal tract area functions by empirical orthogonal modes", *J. Phonetics* **26**, 223-260.

URL-1:http://www.cs.adfa.edu.au/~spike/Research/VisualSpeech/index.html#VT

URL-2: ftp://ftp.cs.adfa.edu.au/pub/users/spike/matlabVT.zip.

Yang, C.-S. and Kasuya, H. (1994). "Accurate measurement of vocal-tract shapes from magnetic resonance images of child, female and male subjects", *Proc. Int. Conf. Spoken Lang. Process.* Yokohama, Japan, 623-626.

Yehia, H. C., Takeda, K. and Itakura, F. (1996). "An Acoustically Oriented Vocal-Tract Model", *IEICE Trans. Inf. & Syst.* **E79-D**(8), 1198-1208.