ROBOTIC ACTUATION OF A 2D MECHANICAL VOCAL TRACT

Ian S. Howard

Centre for Robotics and Neural Systems, University of Plymouth, Plymouth, UK.

ian.howard@plymouth.ac.uk

Abstract: Here we present a robotic vocal apparatus incorporating a tongue body that is moved by stepper motors under computer control. The vocal tract consists of a central mouth region with a 2-dimentional tongue, a lip section and a nasal cavity. Movement of the articulators changes the vocal tract cross-sectional area, and thereby its acoustic properties, thus providing a means to generate different vocalic sounds. Numerical optimization in Matlab was used to design the mechanism, which was achieved by fitting its various dimensions to published articulatory data. The vocal tract was then built using 3D printing technology. We present some examples of vowel and nasal sound production using simulated voiced sound excitation. Finally, we demonstrate that by blowing air into the lower larynx end of the vocal tract while making appropriate constrictions, the apparatus can also generated different fricative sounds.

1 Introduction

1.1 Motivation for robotic models of speech production

The reasons for building a robotic vocal apparatus are many-fold and this is an active area of research [1]-[11]. From the author's perspective, a robotic speech apparatus will provide a good alternative to using software implementations of articulatory speech production and should therefore be a valuable tool to assist the building models of speech acquisition and speech motor control. It addition, it will provides a means to investigate the importance of the vocal tract anatomy and mechanisms that are involved in speech production.

1.2 Extending our previous work

Our former system could generate purely vocalic sounds, but in order to modulate vocal tract cross sectional area required hand actuation of the tongue to move it around the mouth cavity [12]. In this paper, we extend the design in several ways. We now incorporate robotic activation of the tongue body using linkages driven by stepper motors operated under computer control. In addition, we now include a nasal cavity. Finally, we show that such a system can generate fricative sounds by blowing air in at its base, which can lead to turbulent airflow and therefore noise generation at constrictions in the vocal tract.

2 Methods

2.1 Approach

We adopt a computational approach to mechanical design by first generating drawings in Matlab and importing them into AutoCAD Fusion 360 using Python scripts. This provides a means to incorporate sophisticated optimization procedures in the mechanical design process. The design employs an elliptical tongue of fixed radii and a linear jaw and lip sections that move within a mouth cavity. The dimension and geometries of the articulators and mouth were calculated by fitting them to published human vocal tract data. Similarly the nasal cavity was matched to published nasal area data. The vocal apparatus was built using 3D printing technology. A horn driver unit was attached to the lower end of the assembly and driven by a voice source model to simulate voiced glottal excitation.

2.2 Area Data

The acoustic properties of the mouth cavity that play an important role on speech production are determined by its cross sectional area, and changes in the cross sectional area is achieved by means of movement of the speech articulators. To operate effectively as an acoustic filter, (as in a real human vocal tract), a robotic vocal tract needs to have the appropriate cross sectional area, be airtight and keep the level of sound transmission through its walls down to low levels.

To obtain the appropriate cross sectional area, we again make use of area functions for vowel productions for a single male speaker from a published MRI study carried out by Story et al. [13]. This provides a midline path and one-dimensional area measurements of area along the midline of the vocal track. In total we used the first nine entries corresponding to the American English vowels. Their SAMPA representation [14] is as follows: /i/, /I./, /E/ /}/, /V/, /A/, /c/, /o/, /U/.

To add a nose to the model, we make use of nasal tract area data from Dang and Honda [15]. The degree of nasality is controlled by the opening of the velopharyngeal port, which we realize using a rotary flap. The nasal tract has a total length of about 11cm and a volume of about 25 cm^3. The narrowest part is at the nostrils where the area reduces to about 1 cm^3 – 2 cm^3. Its velum opening can be up to 1 cm^3, but is normally 0.2cm^3 - 0.8 cm^3. In a human nasal tract, the nasal cavity divides into two parts. However, for simplicity, here it is only modeled as a single cavity.

2.3 Mouth, tongue and tongue-tip jaw-lip design

In this vocal apparatus we employ 3 articulators that move within a fixed mouth cavity; the tongue body, tongue tip and velum. The mouth consisted of two parallel side sections separated by 2cm, sealed at the top by a curved roof and at the bottom by the tongue and the base of the mouth. This mouth geometry provides a U-section in which the articulators can move freely. To ensure the mouth is airtight, its side plates are screwed down by several bolt and grease is used to hermetically seal the cavity. In addition, grease is applied to the sides of the walls to lubricate the moving parts and also to provide an airtight seal with the articulators.

We make use of more complex tongue design than that used previously. The body of the tongue again consists of an ellipse specified by its two radii, and these dimensions remained fixed for all articulations. The tongue now has an adjustable hinged tip, which is used to model the jaw-lip area function that occurs at the front of the mouth (See Figs. 1 & 2.). The lower part of the tongue now has a sliding section tensioned by a spring to close off the vocal tract at the tongue base. This ensures there is a continuous airtight passage through the vocal tract from the larynx section to the tip of the tongue. The tongue body has 3 degrees of freedom (2 translational and 1 rotational), so its location in the mouth and orientation can change to articulate different vowels. The tongue tip has a single rotational degree of freedom. Similarly the velum has a single rotational degree of freedom.

2.4 Fitting articulator geometries

The shape of the roof of the mouth, the radii that specified the elliptical tongue and shape of jaw-lip tip section were found using non-linear optimisation. This was achieved using the Matlab function fmincon to minimize the mean-square error arising from the target vocal tract area functions with the area arising from the gap between the articulators to the mouth roof. The fixed articulator and mouth roof geometries were optimised over the entire set of vowels, and the translation and rotation of the articulators (i.e. tongue and tip jaw-lip sections) were found for each individual vowel configuration. This process was thereby constrained to yield a fixed mouth roof and palate contour with articulators that could be moved within the mouth

to achieve the cross sectional areas needed to realize each of the individual target vowel cross sectional areas. Fig. 1 A-C shows examples of three fitted configurations. Fig. 1 D shows the location of the centre of the tongue body. Note these locations are similar to those seem on the classical vowel quadrilateral, except the mouth is on the right side and not on the left.

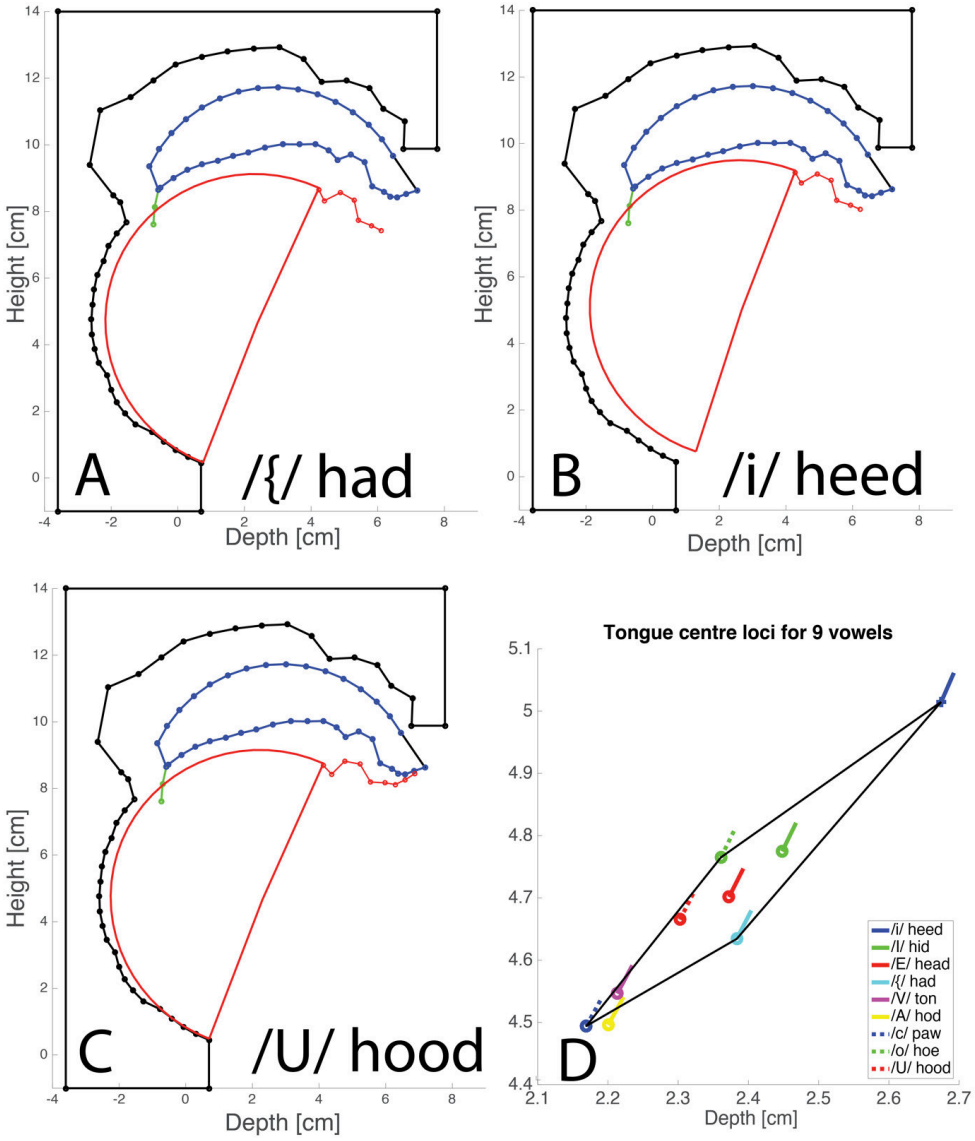

Figure - 1 A-C: Examples of 3 vocal tract and nasal area functions generated in Matlab. The tongue section is shown in red with the single line indicating the upper tongue tip contour. The roof of the mouth and nasal cavity are shown in black. The palate section is shown in blue. The velopharyngeal port is shown open and plotted in green. **D**: Tongue center location for vowel production. The short line indicates orientation. It can be seen the range of movement across vowel qualities is quite small.

The computed vocal tract geometries were imported into AutoCAD Fusion 360, which was subsequently used to finalize the design of the mechanical vocal tract apparatus. The optimized mouth and articulator geometries were used to generate STL format files and the

mechanical parts were built in PLA using a Flashforge Creator Pro 3D printer. The AutoCAD model of the mouth roof and articulators are shown in Fig. 2.

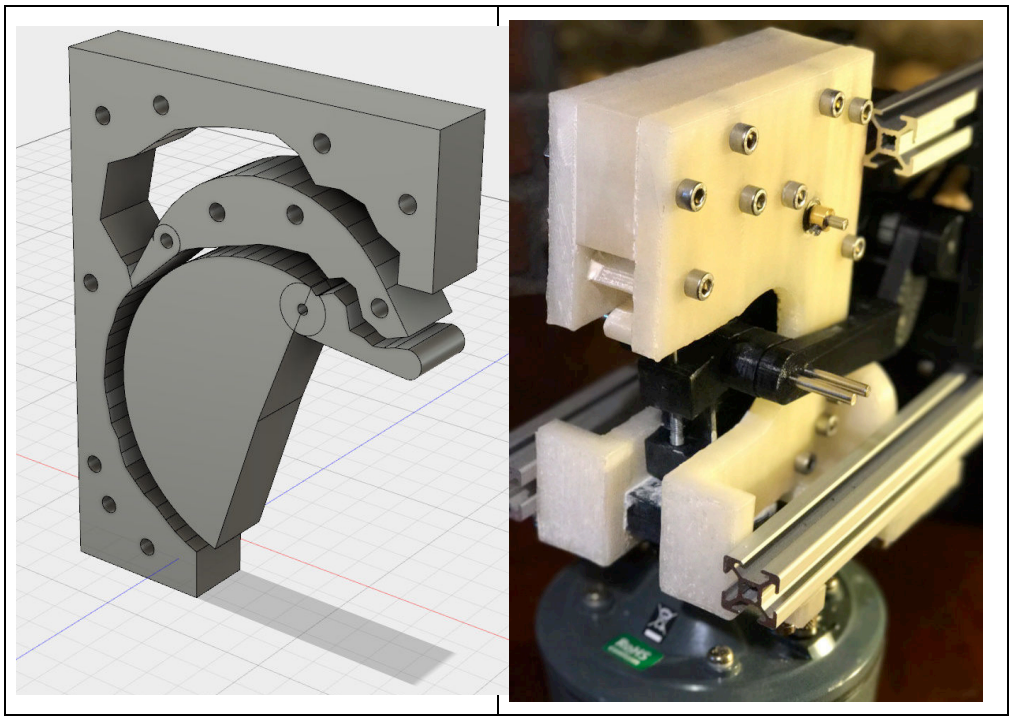

Figure - 2 LHS: AutoCAD Fusion design of the central section of vocal apparatus. Here the sides are not shown so that the tongue mechanism vocal and nasal cavities and the velar flap can be seen. **RHS:** Oblique front view of the 3D printed vocal apparatus. The black movable tongue body and white tongue tip can be seen, as can the white side plates of the apparatus.

2.5 Robotic actuation

Control of articulator position plays a critical role in speech production. As can be seen from the central location of the tongue (Fig. 1D) the actual movement range for the tongue across vowel qualities is quite small, changing only by about 5mm over the production of the range of vowels modeled in this study. We adopted a simple 3-link revolute arm construction to move the main tongue section in two dimensions and also to control its orientation (Fig. 3). It was operated using GT2 timing belt drive. Miniature ball bearing were used at its joints to ensure all moving parts could rotate with high precision and with little frictional resistance despite high belt tension. All parts were designed in AutoCAD Fusion 360 and subsequently 3D printed in PLA.

To drive the tongue body we use 3 high-resolution 400 PPR Nema17 stepper motors connected to them by means of GT2 timing belts. Currently the tip of the tongue and the velum are moved by hand but will soon be similarly actuated. The vocal assembly and actuation are mounted on 20mm aluminium profile and attached by means of bolts and T-nuts. This greatly facilitates adjustment of the apparatus. It also provides an elegant means to tension the timing belts, since the motors can simply be slid backwards and forwards until appropriate tension is achieved. The motors are driven from a Kuman 3D Printer Controller kit for Arduino RAMPS 1.4 Controller Board, which can operate up to 5 stepper motors. It was operated from an Arduino Mega 2560 R3 Microcontroller programmed in C++.

2.6 Adding airflow

Speech breathing and the role of the respiratory system play an important role in speech production, and we previously briefly touched this theme in simple speech sound learning simulations [16]. In short, we believe that incorporating breathing and airflow would make a robotic vocal apparatus more realistic and useful for modeling the learning of speech production, and therefore consider some of its characteristic below.

During normal adult speech, inhalation is performed under muscular control. Exhalation also takes advantage of the elastic properties of the chest walls, although in order to hold sub-glottal pressure constant during speech production, muscular control is also involved. In contrast, in young infants this chest wall elasticity is almost absent [17]. Although the normal adult male lungs have a vital capacity between 3000-5000 cm^3, speech production tidal volume is typically only 500 cm^3. An air pressure of about 20 cm H_2O is required for normal phonation, although the lungs can generate up to 10 times this pressure during coughing. Air pressure during speech is typically around 5-10cm H_2O. Airflow during phonation is normally at around 500 cm^3-1000 cm^3 per second displacement. Phonation threshold pressure is generally between 1 and 3 cm H_2O, and it increases with fundamental frequency [18]. For conversational speech pressure is typically about 5 cm H_2O for loud speech it can be 10 to 15 cm H_2O [19]. We note that atmospheric pressure is typically 1030 cm H_2O. Thus compared to most commercial air pumps that typically develop many times atmospheric pressure, the human respiratory system operates at very low pressures with only moderate airflow volumes.

Airflow within the human vocal tract naturally leads to phonation at the glottis and turbulence flow at constrictions, with the latter leading to the production of fricatives and plosives. The location of the frication noise source within the vocal apparatus leads to acoustic filtering, thereby giving rise to a fricative spectral balance that is dependent on the place of frication. Ideally to achieve the right airflow conditions, it would be necessity to model the vocal tract cross section for fricative production, which can again achieved on the basis of published data. In addition, obstacles to airflow, such as the teeth and alveolar ridge, affect the generation of turbulence and need to be taken into account.

2.7 Sound sources for vocal excitation

As before to simulate voiced excitation, a Monacor die cast horn driver (model KU-516) was attached to the lower end of vocal apparatus and driven from a Mac computer via a Lepai LP-2020A Audio Mini Amplifier using a simulated glottal excitation [12].

Alternative airflow-based means of excitations were also investigated. Although to generate a suitable airflow, a compressor or pump could be used (and this approach will be investigated in the future), a simple approach was adopted here. Since the human respiratory system is the perfect source for such airflow, this air source was adopted. Thus in the current study, activation involved a single participant (the author) manually blowing down the air tube.

To examine airflow generated voiced excitation, a makeshift artificial larynx consisting of a duck call was attached at the lower end of the vocal tract (see supplementary material for results).

To examine fricative generation, an air tube was directly attached at the same location. Currently we only used the vocal tract optimized for vowels, but still were able to show that frication can be achieved if the articulators are first positioned to generate a point of partial closure of the vocal tract. Then when air is then blown into the vocal tract at its base, it results in air turbulence at the constriction, leading to the production of a fricative.

3 Results

The vocal apparatus was used to first generate a set of static vowel sounds and then dynamic sounds. Speech production was recorded at the lips of the apparatus using a Podcaster USB

microphone and spectrographic analysis was carried out using Matlab. Preliminary results presented as spectrographic comparisons between those generated by a single adult male speaker and those from the robotic vocal apparatus are shown in Fig. 4 for vowels and a CVCV. These indicate some similarity in their formant structure. To generate fricatives the vocal tract was suitably constricted and air blown into its base. Spectral comparisons between fricatives for the adult male speaker and robot are shown in Fig. 5, which show similar spectral characteristics. These figures and their corresponding WAV sound files, together with other additional material, is available online in supplementary material, so readers can make their own hearing evaluations. The link is located at:

http://www.Howardlab.com/publications/ESSV2017SaarbrueckenSupMat.pptx.

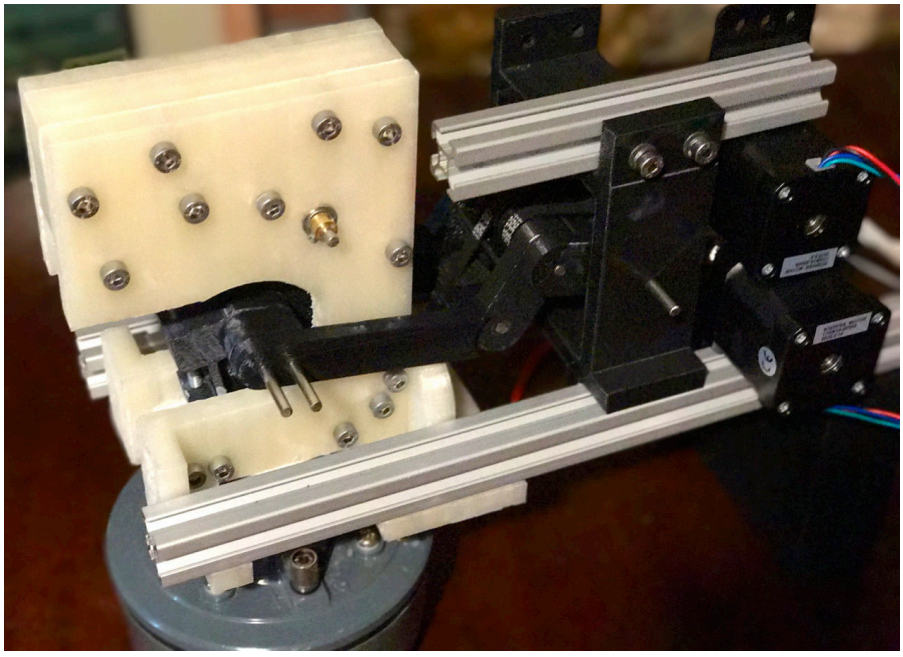

Figure 3 - Robotic vocal apparatus with stepper motor actuation mechanism. On the left is the PLA vocal tract. A horn driver is attached below to provide simulated voiced excitation for vowel production. The stepper motor actuated revolute arm mechanism can be seen on the right.

4 Discussion

4.1 Summary

Here we presented a preliminary design and results from a computer controlled robotic vocal tract that is able to generate vocalic, nasal and fricative sounds. The current system is still a work in progress, and motor actuation of the tongue tip and the velum in particular need to be implemented. Overall the focus here has been to adopt an agile approach to development and concentrate first on the main design principle, and then improve the design in future iterations.

4.2 Future work

We currently make no attempt to model dynamics of speech production and instead use open loop feed-forward stepper motor position control to simplify the task. In engineering systems, feedback control mechanisms are used to achieve operating goals in the presence of noise and to compensate for changes in plant characteristics. In the in future we will be investigate using

force and compliant control to better take advantages of the dynamics of the system and made use of state space and optimal feedback control strategies using somatosensory feedback. This will include investigating state feedback control methods (SFC), as outlined by Houde et al. [20]. In further designs of the apparatus, we will also try to increase acoustic performance by adopting better area functions built on the basis of better datasets. The area functions that

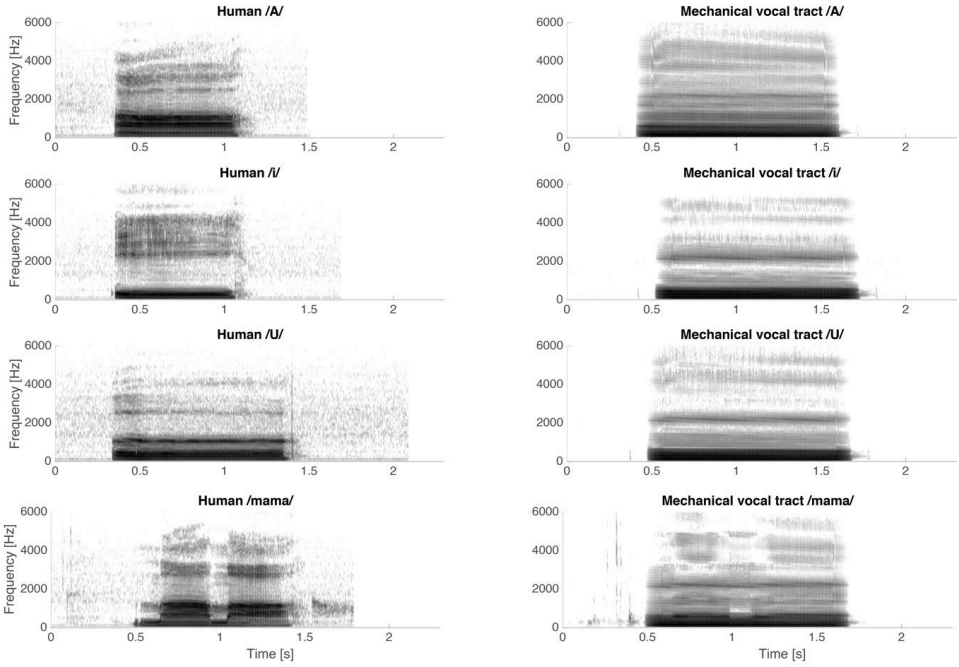

Figure 4 - Spectrographic analysis for the three vowel sounds /A/, /i/ /U/ and the CVCV /mama/. The latter were made with the velopharyngeal port open to nasalize the sounds. Data from a male participant and the vocal apparatus are shown on the left and right respectively.

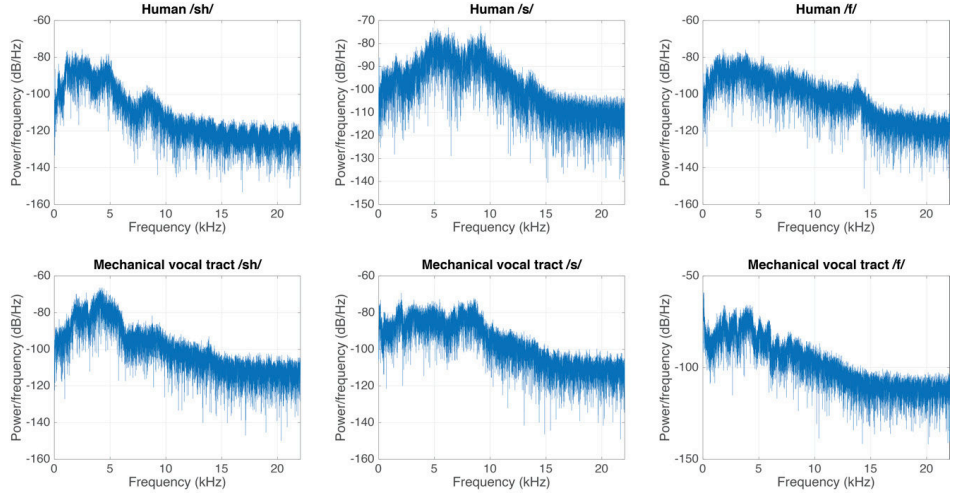

Figure 5 - Spectrographic analyses for fricative from a male participant are shown in the upper row and those from the robotic vocal apparatus are shown in the lower row.

were used here represent early research in the fields and more sophisticated and accurate datasets now exist that could easily be used to generate an improved version of the vocal tract in 3D. Although we took initial steps to utilize real airflow in the model generated by a participant blowing down a tube, future work will involve implementing a computer controllable speech breathing apparatus. In the future the incorporations of a model of the folds will also be tackled. Currently we use PLA to print the vocal tract. Utilised 3D printing of soft materials (e.g. Ninjaflex) also has great potential to improve many aspects of the design, including the construction of airtight flexible joints and a deformable tongue and lips.

5 References

[1] K. KLAEDEFABRIK, Martin Riches - Maskinerne / the Machines. Kehrer Verlag: Klaedefabrik, KB, 2005.

[2] H. SAWADA AND S. HASHIMOTO, "Mechanical Model of Human Vocal System and Its Control with Auditory Feedback," JSME International Journal Series C, vol. 43, no. 3, pp. 645–652, 2000.

[3] H. SAWADA AND S. HASHIMOTO, "Mechanical construction of a human vocal system for singing voice production," vol. 13, no. 7, pp. 647–661, Jan. 1998.

[4] T. HIGASHIMOTO, "A mechanical voice system: construction of vocal cords and its pitch control," Robotics and Automation, 2002. Proceedings. ICRA '02. IEEE, 2003.

[5] H. SAWADA, M. KITANI, AND Y. HAYASHI, "A Robotic Voice Simulator and the Interactive Training for Hearing-Impaired People," J. Biomed. Biotech., vol. 2008, pp. 1–8, 2008.

[6] K. MIURA, Y. YOSHIKAWA, AND M. ASADA, "Unconscious anchoring in maternal imitation that helps find the correspondence of a caregiver's vowel categories," 2007.

[7] Y. YOSHIKAWA, M. ASADA, K. HOSODA, AND J. KOGA, "A constructivist approach to infants' vowel acquisition through mother–infant interaction," Connection Science, vol. 15, no. 4, pp. 245–258, Dec. 2003.

[8] M. C. BRADY, "Prosodic Timing Analysis for Articulatory Re-Synthesis Using a Bank of Resonators with an Adaptive Oscillator," presented at the Eleventh Annual Conference of the International Speech Communication Association, 2010, pp. 1029–1032.

[9] T. ARAI, "Mechanical Vocal-Tract Models for Speech Dynamics," Interspeech, 2010.

[10] K. FUKUI, K. NISHIKAWA, AND S. IKEO, "Development of a Talking Robot with Vocal Cords and Lips Having Human-like Biological Structures," Intelligent Robots and Systems, 2005. (IROS 2005)., 2005.

[11] N. ENDO, T. KOJIMA, Y. SASAMOTO, H. ISHIHARA, T. HORII, AND M. ASADA, "Design of an Articulation Mechanism for an Infant-like Vocal Robot 'Lingua'," in Biomimetic and Biohybrid Systems, vol. 8608, no. 39, Cham: Springer International Publishing, 2014, pp. 389–391.

[12] I. S. HOWARD, "Towards a mechanical vocal apparatus for vowel production," presented at the ESSV Leipzig Germany, 2016.

[13] B. H. STORY, I. R. TITZE, AND E. A. HOFFMAN, "Vocal tract area functions from magnetic resonance imaging," The Journal of the Acoustical Society of America, vol. 100, no. 1, pp. 537–554, Jul. 1996.

[14] J. C. WELLS, "SAMPA computer readable phonetic alphabet." Handbook of standards and resources for spoken language systems 4, 1997.

[15] J. DANG, K. HONDA, AND H. SUZUKI, "Morphological and acoustical analysis of the nasal and the paranasal cavities.," The Journal of the Acoustical Society of America, vol. 96, no. 4, pp. 2088–2100, Oct. 1994.

[16] I. HOWARD AND P. MESSUM, "Modeling motor pattern generation in the development of infant speech production," International Seminar on Speech Production, 2008.

[17] R. Netsell, W. K. Lotz, J. E. Peters, and L. Schulte, "Developmental patterns of laryngeal and respiratory function for speech production.," J Voice, vol. 8, no. 2, pp. 123–131, Jun. 1994.

[18] I. R. Titze, "Principles of Voice Production," National Center for Voice and Speech, 2000.

[19] T. J. Hixon, G. Weismer, and J. D. Hoit, Preclinical speech science. Plural Pub Inc, 2008.

[20] J. F. HOUDE AND S. S. NAGARAJAN, "Speech production as state feedback control.," Front Hum Neurosci, vol. 5, p. 82, 2011.