TOWARDS A MECHANICAL VOCAL APPARATUS FOR VOWEL PRODUCTION

Ian S. Howard

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

ian.howard@plymouth.ac.uk

Abstract: Here we describe preliminary results in the construction of a mechanical vocal apparatus. The design employs an elliptical tongue of fixed radii and a linear jaw and lip section that move within a mouth cavity. The dimension and geometries of the articulators and mouth were found by fitting them to published human vocal tract data. The dataset used consisted of the vocal tract area functions for a single male speaker for the production of nine American English vowels. The fit was achieved using an optimization running in Matlab, which minimized the mean-square error arising from the target vocal tract area functions with the area arising from the gap between the articulators and the roof of the mouth. The articulators and mouth roof dimension were optimised over the entire set of vowels, and the location of the articulators was found for each individual vowel. The optimized vocal tract parts were imported into a CAD package and incorporated into the design of the vocal apparatus. This was then 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. We present preliminary results from the mechanical vocal tract for several vowel sounds which were achieved by moving the articulators by hand, and make a comparison to the acoustic output achieved from a male speaker and to a control set of 3D printed tubes based directly on the published vowel cross-sectional area data.

1 Introduction

1.1 Motivation

There are several motivations for building mechanical models of speech production. An idea mechanical model would have the advantages over computer simulations because it would be a real-world physical system, which has several important implications. Kinematic and dynamic properties of the mechanism, as well as its intrinsic aerodynamic properties, would not need to be explicitly modelled since they would arise naturally. Real airflow within a mechanical model would naturally lead to phonation at the glottis and turbulence flow at constrictions within the vocal tract, with the former realizing for voiced excitation and the latter the production of fricative and plosive excitations. The human vocal apparatus exhibits non-linear source-filter interactions [1] which are either difficult to model or overlooked in many software simulations. Such complex effects would be implicitly captured in an ideal mechanical model, since, for example, closure of vocal tract would affect other behaviours, such as reducing phonation at the larynx and frication at constrictions. The incorporation of the pulmonary system in a mechanical model of speech production would also permit the largely neglected issues of speech breathing in phonetics to be addressed and investigated.

We believe the act of building models is informative because it focuses attention on aspects and basic principles of the vocal apparatus that affect speech production. These can then also be objectively evaluated, since their influence of subsequent speech can be observed and analysed. Also, insight gained from building mechanical models may also lead to further questions that can then be investigated in real speakers. Invasive measurement techniques can easily be built into a mechanical model to record pressure and palatal contact. Indeed, even simple mechanical models have been valuable tools for scientific research in speech science.

For example, frication has been investigated in this way [2]. Such invasive procedures are certainly difficult or uncomfortable to perform in real human subjects.

1.2 Basic design principles

To realize the acoustic filtering properties of the human vocal tract needed to simulate vowel production with a mechanical model relies on modelling its area function with a tube with a cross sectional areas that can be varied by the movement of the articulators (i.e. the jaw, tongue and lips). In this way an acoustic excitation from a glottal source applied at one end of the tube will be appropriately filtered, giving rise to acoustic output at the lips with a format-like structure that can simulate the production of the different vowel qualities. Fricative and other excitations are filtered in a similar fashion. For such a vocal tract model to operate effectively it is important that it is resonant and airtight, that is does not excessively leak sound though its walls and through tongue, and avoids unwanted resonances. This necessitates use of a sealed airtight construction and acoustically damping material in the construction, particularly for the tongue [3].

1.3 Previous mechanical models

Mechanical models of the speech apparatus have a long history, with one of the first models capable of reproducing a few steady state vowels [4]. This was shortly followed by a model which was capable of producing CV-like speech sounds [5]. Models that consisted of natural vocal-tract shapes were built much later [6], [7] and only more recently those having the intricate shapes of the vocal tract been constructed using rapid prototyping technology from MRI images of vowel production [8].

Robotic models enable configurations to change in real-time, thereby realizing the production of dynamic speech sounds. One early such design was Motormouth, which was driven by 6 motors and used a single rotating mechanism to implement the tongue [9]. Other models did not emulate the physical structure of the vocal apparatus, but instead was based on a deformable silicone rubber tubular resonator with nasal cavity [10]-[13]. These also incorporated an air pump and two-layer artificial silicone rubber vocal folds. A similar design was also built by the Asada group [14], [15]. Another mechanical vocal tract model was based on a Plexiglas and resin vocal cavity, a silicone tongue that is moved with a mechanism with 5 DOF, and a velo-pharyngeal port, lips and vocal folds [16]. A computer control mechanism has been recently added to an early sliding plastic strip design, enabling is to produce dynamic sequences of vowels [17].

Anton, an animatronic model of the vocal tract, has been developed based on details of human anatomy [18]. The design takes a biomimetics approach [19] and attempts to use components that mimic biology, only resorting to functional approximations where this is not technically feasible. The tongue is central to the design and is cast from silicone rubber, which captures the hydrostatic nature of a real tongue. To simulate the behaviour of the main extrinsic tongue muscles, Dyneema filaments innervate it. To simulate the effect of muscle contraction, they move and deform the tongue in a fashion that mimics real tongue behaviour. The filaments are located in channels in the tongue body and only actually attach to the tongue at their ends using a plastic mesh. The palette is constructed from silicone on the basis of MRI images. The parts are built into a plastic skull with a movable jaw. All articulators are driven using servomotors connected to the actuating filaments and a loudspeaker was used to simulate voiced excitation. Although Anton had limited performance as a speech production device, it demonstrates the feasibility of this innovative approach.

Currently, the most sophisticated robotic model of speech production is the Waseda Talker [20]. One of the latest versions (WT-7R) models all the main functional aspects of speech production. The lungs have 1 DOF and generate a controlled airflow using motor driven

pistons that move and down in clear cylinders. The human larynx is a complex structure and the vocal folds design adopts a 5 DOF biomechanical implementation constructed using the synthetic rubber material Septon. Electric motor actuation is used to simulate the main muscle functions involved in phonation, which are to stretch and lengthen, thicken and shorten, and to abduct and adduct the vocal folds. The tongue was especially designed to capture the complex behaviour of the human tongue. This includes the intrinsic muscle functionality of tongue narrowing, lengthening and flattening, and tip movement, as well as the extrinsic tongue muscle functionality of raising the tongue front and rear, and pulling the tongue body down. The tongue was designed in three parts; the tip, the blade and the body. The tip of the tongue is driven by a 3 DOF parallel link, controlling front and back length, and rotation. Its vertical deformation is reproduced mainly by the jaw mechanism. Tongue blade and body are driven by a set of 2 DOF slider-crank links, to control length and rotation. To improve the resonance characteristics of the vocal tract, the tongue is filled with ethylene glycol. Lips play a key role in the acoustic properties of the vocal tract and were designed to mimic human lips. The lips are made of soft material and are connected by a vice mechanism to five direction links and one fixed point. They can reproduce the shapes needed for the generation of Japanese vowels using this 5 DOF rigid link actuation mechanism. They can be protruded, raised or lowered to achieve opening and closing, which is important in consonant articulation, and spread and rounded.

Recently the Asada group has developed a mechanical model of an infant vocal apparatus [21]. This is an exciting development and particularly relevant for the study of infant speech acquisition.

These current mechanical models all have some limitations in terms of speech production performance and therefore there is still useful work to do in the development of a high performance robotic vocal apparatus.

2 Methods

2.1 Design philosophy

The long-term goal of this project is to develop a realistic physical 3D model of the speech apparatus, complete with robotic actuation and control mechanisms for the speech breathing apparatus, vocal folds and vocal tract. Although modelling the jaw, lips and a vocal cavity, as well as the vocal folds and air source, are needed for a full mechanical vocal tract simulation, here we concentrate on the tongue and the fixed structures in the mouth.

2.2 Story et al. dataset

In this work we make use of area functions for vowel productions for a single male speaker from a published MRI study carried out by Story et al. [22] This study provides midline and one-dimensional area measurements of area along the midline of the vocal track. The area data is presented in tabular form. In this work we used the first nine entries corresponding to the American English vowels (represented here in SAMPA [23]): /i/, /I./, /E/ /}/, /V/, /A/, /c/, /o/, /U/. We did not use the vowel /u/ since this involved considerable extension of the lips, which was not the focus of our current project

Story et al. also plot the vocal tract the midline location for the single male speaker. We digitized this plot to recover these distance measurements.

2.3 Rectangular vocal tract control sections

A set of fixed configuration rectangular vocal tracts were built to act as measurement controls for three selected vowels, /i/, /A/ and /U/. To achieve this, the area functions for the respective vowels were used to calculate the height of the vocal tract aligned to the midline given a 2cm

vocal tract width. The upper and lower extents of the roof and base of the rectangular section were then set from the calculation of the normal vectors of length ±height/2 at each midline sample point. The vocal tract surfaces were then plotted out in Matlab as sealed tubes with wall thickness 10 mm and saved in STL file format. The printer files were prepared with 100% infill to minimize acoustic transmission though their walls using Simplyfy3D, and the mechanical parts were manufactured in PLA using a Flashforge Creator Pro 3D printer. The resulting tube shapes are shown in Fig. 1. They are shown here for 5mm wall thickness for clarity (since wider wall thickness obscures the tube shapes).

2.4 Mouth, tongue and jaw-lip design

Here we adopt a simple design that employs articulators that moves within a fixed mouth cavity. Unlike in software models which can use sophisticated geometries (e.g. [24]), we were careful to ensure the geometries are easy to realize using physical mechanical components. The mouth thus consisted of two parallel side sections separated by 2cm, sealed at the top by a curved roof and at the bottom by the tongue. This mouth geometry provides a U-section in which the tongue can move freely. Such a mouth construction with an arbitrary shaped mouth roof is easy to build using 3D printing techniques. The tongue itself consists of a 2-dimensional elliptical structure with thickness such that it exactly fits into the U-section of the mouth, providing an airtight seal with its sides – which can be enhanced further by the application of a small quantity of lubricant between the articulator and the U-section (such as water or thin oil). Changes in the cross sectional areas of the vocal tract are achieved by moving the articulators around within the mouth channel by means of mechanical actuation. In this initial work, only a simple tongue was investigated since this ensured that it could be easily realized mechanically and fabricated using 3D printing. It consisted of an ellipse specified by its two radii, and these dimensions remained fixed for all articulations. The tongue had 3 degrees of freedom (2 translational and 1 rotational), so its location in the mouth and orientation could change to articulate different vowels. Thus although the tongue could not change shape to articulate different vowel sounds, it could move around in the mouth cavity to change its effective cross sectional area. A second articulator was used to model the area function that arises from the front jaw and lips. This was modelled as an arbitrary curve that could rotate around its stating position and also translate in 2 dimensions.

2.5 Fitting articulator geometries to the Story et al. dataset

The shape of the roof of the mouth, the radii that specified the elliptical tongue and shape of jaw-lip 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 arise from the gap between the articulators to the mouth roof. The fixed articulator and mouth roof dimension were optimised over the entire set of vowels, and the translation and rotation of the articulators (i.e. tongue and jaw-lip sections) was found for each individual vowel configuration. This process was constrained to yield a fixed mouth roof contour and 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. The discovered articulator geometries were then imported into AutoCAD Fusion 360, which was subsequently used to finalize the design of the mechanical vocal tract apparatus. Examples of the fitted mouth and articulators for three of the nine fitted vowels are shown in Fig. 2A-C. In addition the different individual tongue and jaw-lip locations for all nine vowels in the dataset is shown on 2D.

The optimized mouth and articulator geometries were used to generate STL format files, which were again manufactured in PLA following the 3D printing procedure used for the control tubes. The CAD models of the mouth roof and articulators are shown on the LHS of Fig. 3.

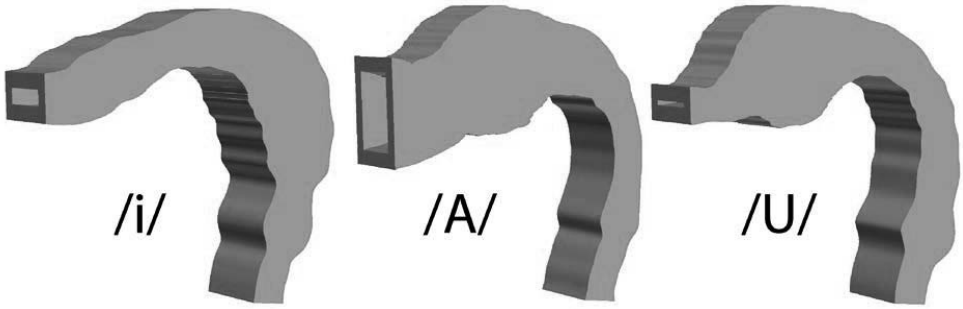

Figure 1 – Rectangular 3D printed vocal tract sections for the vowels /i/, /A/ and /U/ generated around the vocal tract midline. Control tubes shown here with 5mm wall thickness for clarity.

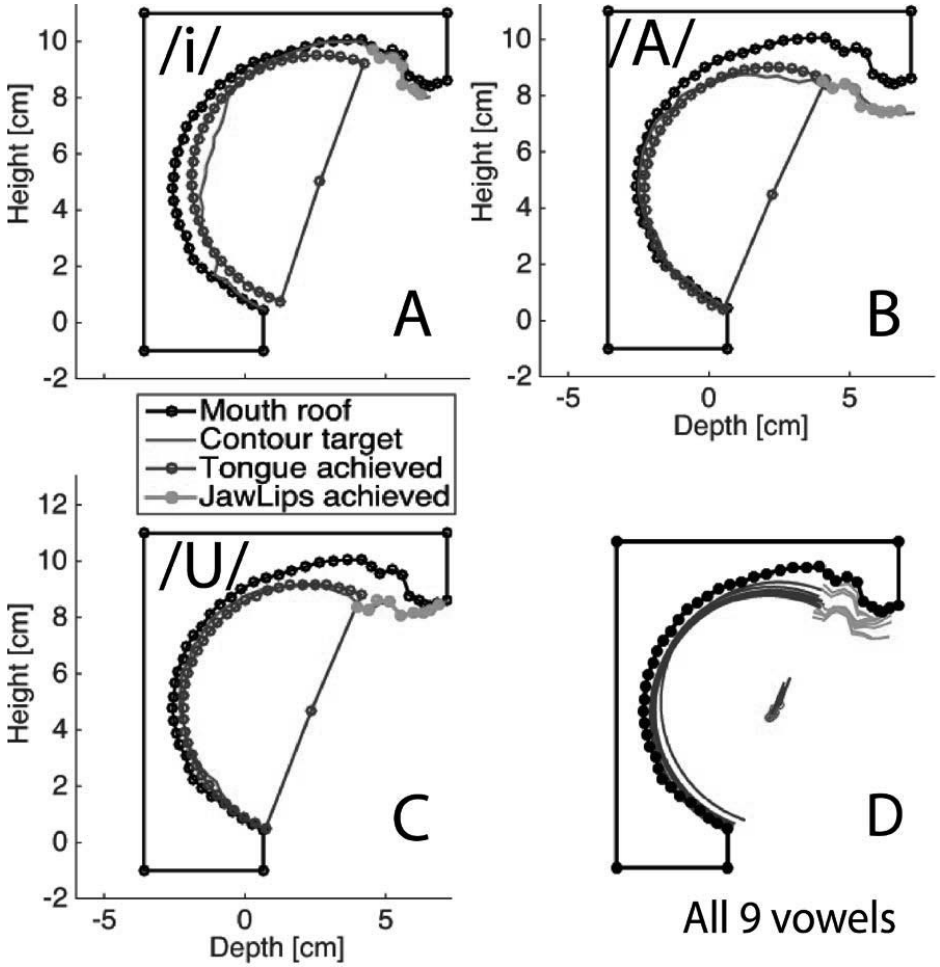

Figure 2 – Vocal tract area functions fitted by optimization with fixed geometry roof, an elliptical tongue and arbitrary curved jaw-lip section. A-C Configurations for vowels /i/, /A/ and /U/. Mouth roof plotted in black, vocal tract articulator target in red, fitted elliptical tongue shown in blue and fitted jaw-lip section shown in green. D Variations in the overall tongue locations across the nine vowel dataset. The centre and movement directions of the ellipse are also shown.

2.6 Voice source

A Monacor die cast horn driver (model KU-516) was attached to the lower end of the 3D printed mechanical vocal apparatus assembly (see Fig. 3 RHS). This horn driver has a respectable frequency range of 160-6500 Hz, although to simulate male speech a unit with a lower cut-off would have been more desirable. It was driven from the sound output from a Mac computer via a Lepai LP-2020A Audio Mini Amplifier using a signal generated in software by a Rosenberg voice source model to simulate acoustic glottal excitation. To test the mechanical vocal apparatus, simple single second long linearly falling (140Hz-120Hz) intonation pitch contours were used.

2.7 Analysing acoustic output

Speech production was recorded at the lips of the apparatus using a Podcaster USB microphone and spectrographic analysis was carried out using Matlab.

3 Results

Preliminary results for the static vowel sounds /i/, /A/ and /U/, and shown in Fig. 4. This illustrates spectrographic comparison of the acoustic outputs generated by a single adult male speaker that achieved with the control 3D printed tubes and also the mechanical vocal apparatus. Good agreement for the lower formats can be seen across the conditions. The mechanical vocal apparatus generated respectable vowel qualities, which could easily be recognized. Subjectively it performed almost as well as the control tubes. These sounds can be listened to using the online supplementary material, which can be found at: Howardlab.com/publications/ESSV2016LeipzigSupMat.pptx.

4 Discussion

4.1 Future work

Here we used the data from Story et al. [22] to demonstrate the principles of our design approach. Currently this is limited to a one-dimensional model of area although in the future using more sophisticated datasets this could easily be used to build a 3D vocal tract.

Articulator actuation of the tongue is currently achieved by moving it around the mouth by hand. We are in the process of constructing a mechanism using linkages driven by small servomotors, which operated under computer control.

A more complex tongue design would give a better fit to the area functions. In addition, it would also potentially be much more effective in generating the constructions needed for frication and for plosive sounds. Mathematically it is easy in increase tongue complexity. For example tongue elliptical radii can also be fitted on a vowel-by-vowel basis for give a better fit to the dataset. However will still be important to limit tongue designs to those that can be mechanically realized in practise.

Currently we use a compression driver to generate the glottal waveform, which generates an acoustic excitation with no net airflow. Further work will involve implementing a speech breathing apparatus and vocal folds, thereby enabling the inclusion of these important aspects of speech production.

4.2 Conclusions

Here we formulated a mechanical vocal tract design process as an optimization problem. This involved fitting articulators in a mechanical model to the real measured vocal area functions. We presented some preliminary results from the mechanical vocal tract for several vowel sounds, and showed that even using a simple tongue geometry it is still possible to generate acoustic output that compares reasonably well with more accurate vocal tract geometries.

Overall we believe this work will form basis for further developments in mechanical speech synthesis.

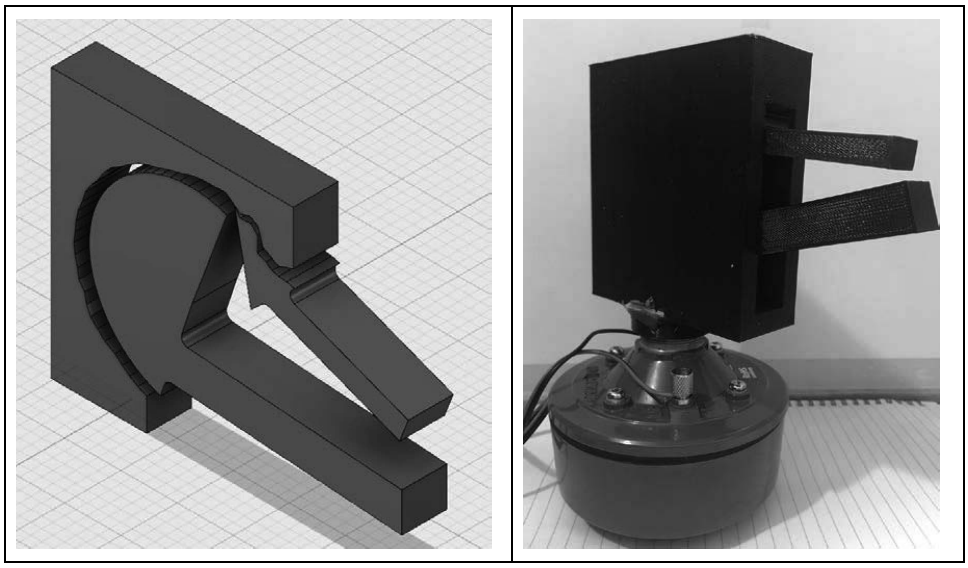

Figure 3 – LHS: 3D CAD assembly consisting of the fixed mouth roof and the movable tongue and jaw-lip sections, shown here without the side plates that seal the mouth cavity. RHS: Assembled mechanical vocal tract attached to horn driver to provide voiced excitation.

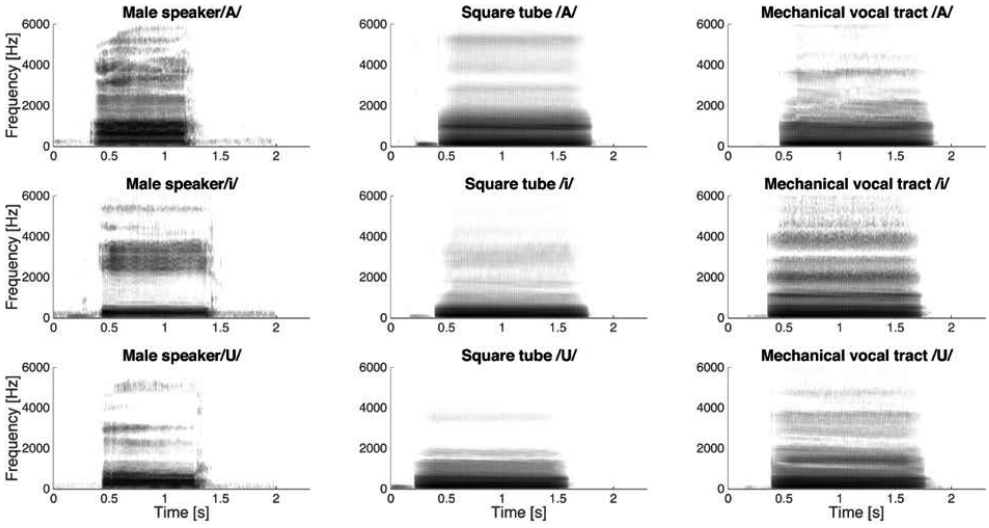

Figure 4 – Spectrographic analysis for the three vowel sounds /A/, /i/ and /U/. Human vowels from a male subject are shown in the left column, vowel simulations from control tubes set to the exact published area functions are shown in the middle column and the mechanical vocal tract simulations are shown in the right column.

5 References

[1] I. TITZE, T. RIEDE, AND P. POPOLO, "Nonlinear source–filter coupling in phonation: Vocal exercises." The Journal of the Acoustical Society of America 123.4 (2008): 1902-1915.

[2] A. BARNEY, C. H. SHADLE, AND P. O. A. L. DAVIES, "Fluid flow in a dynamic mechanical model of the vocal folds and tract. I. Measurements and theory." The Journal of the Acoustical Society of America 105.1 (1999): 444-455.

[3] K. NISHIKAWA, H. TAKANOBU, T. MOCHIDA, M. HONDA, AND A. TAKANISHI, "Speech production of an advanced talking robot based on human acoustic theory." Robotics and Automation, 2004. Proceedings. ICRA'04. 2004 IEEE International Conference on. Vol. 4. IEEE, 2004.

[4] C. KRATZENSTEIN, "Sur la raissance de la formation des voyelles." J. phys 21 (1782): 358-380.

[5] W. Von Kempelen, "Mechanismus der menschlichen Sprache." Degen, 1791.

[6] N. UMEDA AND R. TERANISHI, "Phonemic feature and vocal feature: Synthesis of speech sounds, using an acoustic model of vocal tract." J. Acoust. Soc. Jpn 22.4 (1966): 195-203.

[7] R. R. RIESZ, "Description and demonstration of an artificial larynx." The Journal of the Acoustical Society of America 1.2A (1930): 273-279.

[8] S. FUJITA AND K. HONDA, "An experimental study of acoustic characteristics of hypopharyngeal cavities using vocal tract solid models."Acoustical science and technology 26.4 (2005): 353-357.

[9] K. KLAEDEFABRIK, "Martin Riches-Maskinerne/The Machines." (2005): 10-13.

[10] H. SAWADA AND S. HASHIMOTO, "Mechanical Model of Human Vocal System and Its Control with Auditory Feedback." JSME International Journal Series C 43.3 (2000): 645-652.

[11] 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.

[12] T. HIGASHIMOTO, "A mechanical voice system: construction of vocal cords and its pitch control." International Conference on Intelligent Technologies. Vol. 7624768. 2003.

[13] H. SAWADA, M. KITANI, AND Y. HAYASHI, "A robotic voice simulator and the interactive training for hearing-impaired people." BioMed Research International 2008 (2008).

[14] K. MIURA, Y. YOSHIKAWA, AND M. ASADA, "Unconscious anchoring in maternal imitation that helps find the correspondence of a caregiver's vowel categories." Advanced Robotics 21.13 (2007): 1583-1600.

[15] Y. YOSHIKAWA, M. ASADA, K. HOSODA, AND J. KOGA, "A constructivist approach to infants' vowel acquisition through mother–infant interaction." Connection Science 15.4 (2003): 245-258.

[16] M. C. BRADY, "Prosodic timing analysis for articulatory re-synthesis using a bank of resonators with an adaptive oscillator." INTERSPEECH. 2010.

[17] T. ARAI, "Mechanical vocal-tract models for speech dynamics."INTERSPEECH. 2010.

[18] R. HOFE AND R. MOORE, "Towards an investigation of speech energetics using 'AnTon': an animatronic model of a human tongue and vocal tract." Connection Science 20.4 (2008): 319-336.

[19] Y. BAR-COHEN, "Biomimetics—using nature to inspire human innovation." Bioinspiration & Biomimetics 1.1 (2006): P1.

[20] 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 IEEE/RSJ International Conference on. IEEE, 2005.

[21] 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"." Biomimetic and Biohybrid Systems. Springer International Publishing, 2014. 389-391.

[22] 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 100.1 (1996): 537-554.

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

[24] P. BIRKHOLZ, D. JACKEL, AND B. J. KRÖGER, "Construction and control of a three-dimensional vocal tract model." Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on. Vol. 1. IEEE, 2006.