Design and Development of a Human-like Lip-Sync Robot Mechanism and Control Strategy

Yufei Zhang
Department of Mechanical Engineering
Columbia University
Email: yz4917@columbia.edu

Sylvester Zhang
Department of Mechanical Engineering
Columbia University
Email: sz3297@columbia.edu

Hod Lipson
Department of Mechanical Engineering
Columbia University
Email: hod.lipson@columbia.edu

***Abstract*—This report describes the development of a novel robotic lip mechanism designed for realistic speech synthesis and lip synchronization. By applying linguistic research, we have adopted a modified phoneme-to-viseme (P2V) mapping based on Jeffers' widely utilized model, employing 12 distinctive viseme categories. Throughout the semester, multiple mechanical designs were iteratively tested, culminating in an innovative dual-servo asymmetric parallel configuration to achieve a broader range of realistic lip motions. Future work includes third-generation design testing, algorithmic refinement using formant synthesis for more accurate speech-lip synchronization and developing simulation tools for soft materials like silicone.**

I. Introduction

Accurate lip synchronization in robotic applications is essential for realistic human-robot interactions. Linguistic studies identify 44 phonemes in English, consisting of 29 consonants and 15 vowels [2]. However, due to similar lip configurations across multiple phonemes, only 10 to 16 visemes are typically required [1]. Jeffers' phoneme-to-viseme mapping, widely adopted in animation and visual speech recognition [2], has informed the development of a 12-viseme mapping with minor adaptations for this project (Fig. 1).

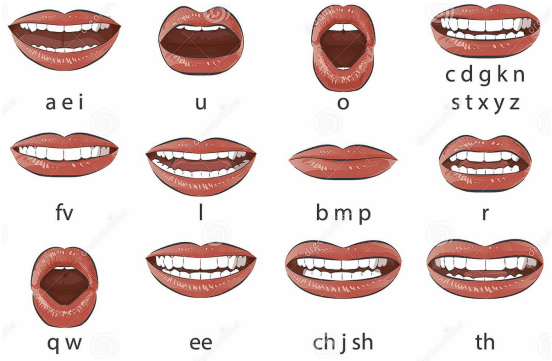


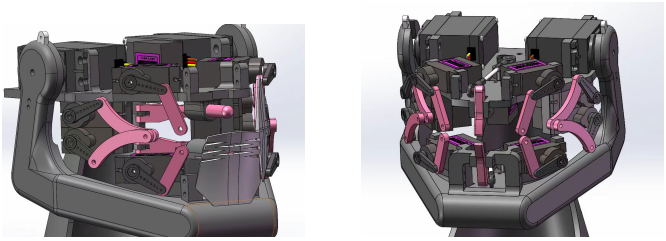
Fig. 1. 12-Viseme

II. Mechanical Design Evolution

Our objective is to construct a robotic mouth capable of imitating nuanced human lip motions. Throughout the semester, we implemented and tested several designs.

A. Single Servo Control: Versions 1 and 2

Initial iterations utilized single servo mechanisms to control each lip-driving point. Experimental evaluations revealed significant limitations, including kinematic singularities restricting the achievable lip shapes, despite the increased number of driving points (Fig. 2).



(a) Version 1 (b) Version 2

Fig. 2. Mechanical designs of Version 1 and Version 2

In order to visualize the motion effects of our mechanical model, we designed lip molds(Fig. 3) that fit our mechanical modeling and cast them with silicone to get the lip model.Our second version of modeling allows for up to 8 lip shapes(Fig. 4), and four more that cannot be simulated.

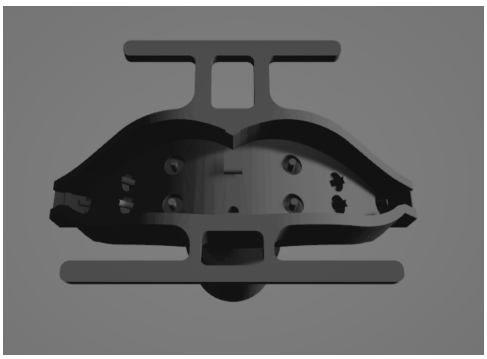


Fig. 3. Mold for Silicone Cast

B. Dual Servo Asymmetric Parallel Structure: Version 3

To resolve singularities, a third design iteration introduced dual servos in an asymmetric parallel configuration for each driving point, significantly increasing the lip mechanism's

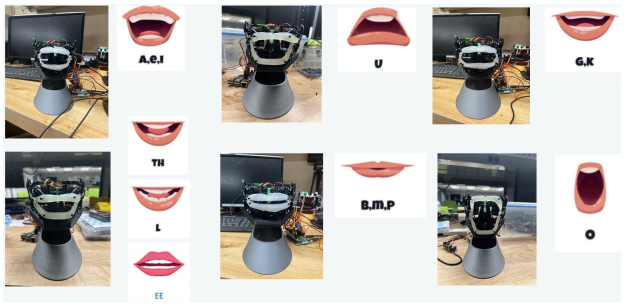


Fig. 4. 8 Lip-Shapes Achieved

degrees of freedom (Fig. 8). A Raspberry Pi 5 replaces the Raspberry Pi 4b used in earlier models, with servos controlled via a centralized PCA9685 PWM driver board. The entire system is powered by two separate 5V supplies, isolating the Raspberry Pi and servo driver.

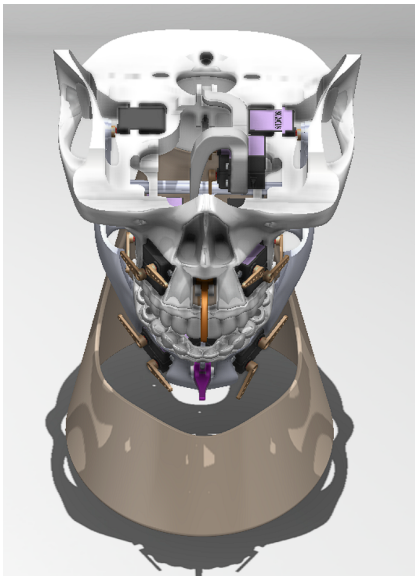


Fig. 5. Front views of Version 3 lip mechanism

III. Control Algorithm Development

Current robotic lip synchronization approaches often map speech frames directly to static lip shapes, resulting in mechanical and unnatural speech animations. Our approach instead decomposes textual inputs into phoneme-level sequences, utilizing the adapted P2V mapping for viseme generation. To address natural speech complexities, such as linking and weakening phenomena, we conceive a Formant Synthesis-Based Lip Control method (Fig. 6), inspired by eSpeak's formant-based speech synthesis technology. This method synthesizes speech by simulating human vocal tract resonances, enhancing naturalness in lip movements through nuanced audio-lip synchronization.

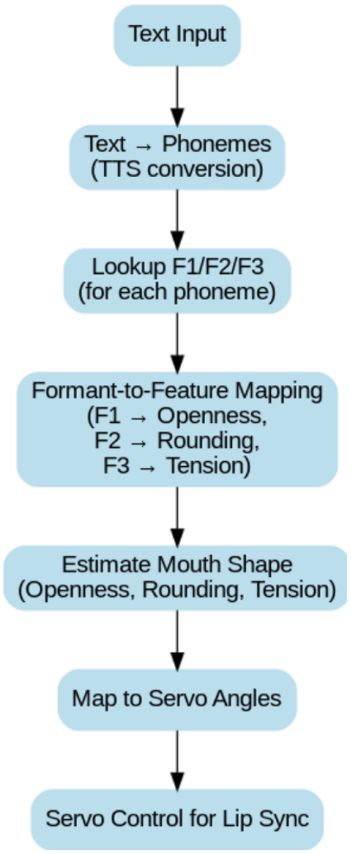


Fig. 6. Formant Synthesis-Based Lip Control approach

IV. Future Work

Upcoming efforts will focus on testing the third-generation design thoroughly to validate its capability to realize all 12 viseme-based lip movements. Post-validation, the control algorithm will undergo further refinements for improved real-time speech synchronization. After realizing the basic 12 mouth movements, we will try to incorporate the tongue design in our modeling, which is conceived as follows(Fig. 7), and the inclusion of the tongue will make our robot more natural in mimicking human speech. Additionally, recognizing the lack of simulation tools for soft robotic materials, we intend to develop a dedicated simulation platform for accurately modeling silicone-based lip mechanisms, addressing a significant gap in current simulation technology.

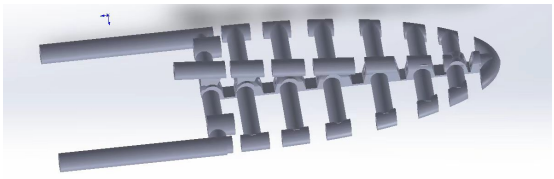


Fig. 7. Tongue Model Design

REFERENCES

[1] L. Cappelletta and N. Harte, *"Viseme definitions comparison for visual-only speech recognition,"* in *Proc. 19th European Signal Processing Conf. (EUSIPCO)*, Barcelona, Spain, 2011, pp. 2109–2113.

[2] H. L. Bear and R. Harvey, *"Phoneme-to-viseme mappings: The good, the bad, and the ugly,"* Speech Communication, vol. 95, pp. 40–67, 2017.

APPENDIX

Viseme	Visibility Rank	Occurrence [%]	TIMIT Phonemes
/A	1	3.15	/f/ /v/
/B	2	15.49	/er/ /ow/ /r/ /q/ /w/ /uh/ /uw/ /axr/ /ux/
/C	3	5.88	/b/ /p/ /m/ /em/
/D	4	.70	/aw/
/E	5	2.90	/dh/ /th/
/F	6	1.20	/ch/ /jh/ /sh/ /zh/
/G	7	1.81	/oy/ /ao/
/H	8	4.36	/s/ /z/
/I	9	31.46	/aa/ /ah/ /ay/ /eh/ /ey/ /ih/ /iy/ /y/ /ae/ /ax-h/ /ax/ /ix/
/J	10	21.10	/d/ /l/ /n/ /t/ /el/ /nx/ /en/ /dx/
/K	11	4.84	/g/ /k/ /ng/ /eng/
/S	-	-	/sil/ /pcl/ /tcl/ /kcl/ /bcl/ /dcl/ /gcl/ /h#/ /#h/ /pau/ /epi/

Fig. 8. Jeffer P2V Mapping