

Development of a humanoid robot mouth with text-to-speech ability

Harshani, L.K.M.D.

*Faculty of Engineering
South Eastern University of Sri Lanka
lmdh31@gmail.com*

Weerasooriya, W.M.A.S.B.

*Faculty of Engineering
South Eastern University of Sri Lanka
arunalu.myown@gmail.com*

Herath, H.M.C.S.

*Faculty of Engineering
South Eastern University of Sri Lanka
www.chanuka.sandaruwana@gmail.com*

Alahakoon, P.M.K.

*Faculty of Engineering
South Eastern University of Sri Lanka
mahinda@seu.ac.lk*

Kumara, W.G.C.W.

*Faculty of Engineering
South Eastern University of Sri Lanka
chinthakaw@seu.ac.lk*

Hinas, M.N.A.

*Faculty of Engineering
South Eastern University of Sri Lanka
ajmalhinas@seu.ac.lk*

ABSTRACT

This is a study of lip movements when speaking the English language, developed humanoid robot face with moving lips and text-to-speech ability. A humanoid robot face is a robotic face capable of mimicking an actual human being during speech. Humanoid robots are finding their way into human activities rapidly with the advancement of technology. Humanoid robots will become very close to humans since they can produce human-like speech by synchronized realistic lip movements while providing sound. However, most humanoid robots do not have very much correct lip movements at the moment. A robotic face with more realistic lip movements will apply in several situations such as news reading, teaching, storytelling, interactive robotic machines, and live advertisements. During the development, a robotic model was built that could simulate jaw movements. Additional lip features were added with driver mechanisms to move the lips in synchronization with the voice produced; audio and text analyzing technology was utilized to improve the accuracy based on phonetic symbols of the English language. Movements of the tongue are essential when reads phonetics in the English language. In this development, physical equipment does not generate vocals; therefore, we neglected tongue movements and facial expressions. We only focus on accurate synchronized lip movements using three control points. This analysis can be used to simulate lip movements in the model of a humanoid robotic face. The performance was tested, and the closeness of the robot-generated lip movements to actual human lip movement was improved by using many voice samples and digitized lip coordinates with the help of image processing techniques.

Keywords: Humanoid robots, Lip synchronization, Text to speech conversion.

situations where humanoid robots find their way into replacing human workers. Even though most of these robotic developments do not have full mobility like a human being, their presence and simulated human-like speech attract customers' attention. Today, people get their news on electronic devices instead of newspapers, television, or radio, which are traditional. Many people are turning to social media for information as well as for reading pleasure. News presenters would have to enrol with some of these media. In addition to reading news or an electronic text by oneself, many are turning towards automatic readers to which they can listen. Robots and electronically synthesized readers can do this job with high accuracy and constant energetic facial expression for more duration allowing the user to listen to the voice and the visual impression that the article is read by a human being instead of a machine. The text file of news or the document the user wishes to listen to can be entered as data, and the robot can convert it into audio and present it with voice and synchronized facial expressions as a human would read it. The user will get the feeling of togetherness and more freedom to listen to the news or the book, instead of spending time reading through pages that he or she may find tedious. This also is a valuable option for those who have reading disabilities and the elderly. This robot can be used for many applications by doing a few software level changes in the future, such as storytelling and teaching. It can be developed as a live advertisement presenter or live advisor instead of printed notices. Live advising is very effective in critical situations like pandemics or any kind of security threats. In human speech generation, vocal organs are moving to generate sounds in different patterns. Vocalization mechanisms of vocal sounds of humans are going to be an attractive research topic in the world. Voice recognition and computerized voice production technologies are essential for humanoid robot projects. In the late 1990s, a robot head called Kismet was invented with some facial expressions related to a human (Breazeal et al., 2000). After that, there were lots of robots that came up with many facial moments. In 2016, Sophia was developed with artificial intelligence as a humanoid robot (Zanotto et al., 2014). In 2019, the latest humanoid robot named *Xin Xiaomeng* was invented, and she was the first female news reading robot (Oh et al., 2006). In 1976, McGurk and MacDonald described human speech as a combination of auditory and visual signals. The McGurk effect described how

INTRODUCTION

Robots and their applications in tasks otherwise handled by humans are becoming innumerable high. Humanoid robots are becoming increasingly popular and are considered as an alternative to human workers in several situations. Routine tasks like news reading and responding to customers in establishments like banks are

Table 1: Comparison of different TTS systems

TTS engine	Features	Languages
Google Cloud	Selection of voices & languages, customization of WaveNet voices, pitch, speaking rate, volume, audio profile	40+
Microsoft Azure	Customization of synthesized speech, text-talker voices, fine-grained audio, and flexible deployment	70
Amazon Polly	Real-time mode, custom modes, customization of pronunciation, pitch, etc.	2
IBM Watson	Customization of pronunciation, words, expressiveness, word timing	8
Bing Speech API	Customization of pronunciation, volume, pitch	78+

Table 2: Comparison of existing humanoid robots

Robot name	Degree of freedom	Number of actuators	Facial expression	Mask material
Kismet	15	4	Intentionally through facial expressions and behaviour	N/A
Albert HUBO	28	66	Neck movements, laugh, sadness, angry and surprised	Frubber
BARTHOC	10	41	Express basic human facial expressions with movements of eyebrows and forehead, jowls, and lower jaw	N/A
SAYA	3	39	Calm, Surprise, Fear, Disgust, Anger, Happiness and Sad	Urethane resin
Sophia	7	34	Joy, grief, curiosity, confusion, contemplation, sorrow and frustration	Frubber

the visual information of human speech affects auditory recognition, and it suggested that lip synchronization is an essential feature for animation production (Raheem Ali, Sulong and Kolivand, 2015). Text-to-speech (TTS) is a module used to control the real-time lip synchronization system. It takes a text file as an input and converts it into artificial human voice output; in this study, they used the *VoiceTEXT* Linux version TTS system developed by a *Voiceware* company. It can convert different languages' text into wav file format voice output (Oh et al., 2010). Recently various kinds of TTS systems are available. Table 1 shows that the comparison of different TTS systems (G2, 2021).

Tacotron 2 is a neural network architecture for speech synthesis directly from the text. *WaveNet* is a generative model of time-domain waveforms that produce audio quality that is the same as human speech, and it is already used in some complete text-to-voice synthesis systems (Hashimoto et al., 2006). It is a sequence-to-sequence for producing magnitude spectrograms from a sequence of characters, simplifies the traditional speech synthesis pipeline by replacing the production of a single neural network trained from data alone (Shen et al., 2018).

Upper (eyebrow and forehead), Center (eyes and root of the nose), and Lower (mouth and cheek) were the three areas that are independently movable by the muscles in the face (Hyung et al., 2016). Happiness can be identified by the movement of the cheek rising with the expansion of the mouth corner. Zygomatic muscle

will help to the expansion of the mouth corner according to the anatomical analysis. Anger can be identified by the movement of the upper lip raise and the mouth extension. *Upper Lip Raiser* that raises the upper lip causes the nose to be wrinkled. Chin Raiser appears characteristically when Sadness is expressed. A triangular muscle raises the lower lip according to anatomical knowledge (Hackel et al., 2005). Robot SAYA is made with six facial expressions (Hashimoto et al., 2006), and *KOBIAN-R* is made with 24 DOF heads (Trovato et al., 2012).

In speech synthesis, speech is represented by phoneme, which is a standard speech unit in the English language. The phonemes are mapped with lip motions, called *visemes*. The English language contains 44 phonemes. The same word can have several meanings according to the nuances given to the character, like facial expressions and eye movements. Lots of phonemes have their pronouncing method with specific mouth positions and tongue movement. Thus, a specific *viseme* is frequently used more than once within the exact words. When the virtual character does not alive in the foreground, some visual phonemes are preferred during the animation. To make an attractive speech, accuracy is necessary for certain characters when the speech goes long (Raheem et al., 2015).

Taking phoneme (a unit of sound that can separate words that distinguish one word from another) pairs as a unit of animation is a new technology called *Pro-Phone*. They produce the co-articulation effects via *Di-*

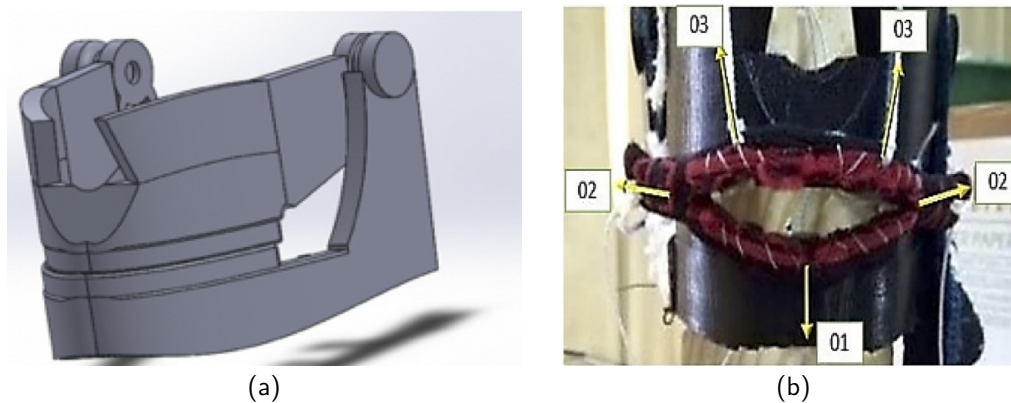


Figure 1: Robot mouth, (a) design, and (b) identified control points (01, 02, and 03: Motors)

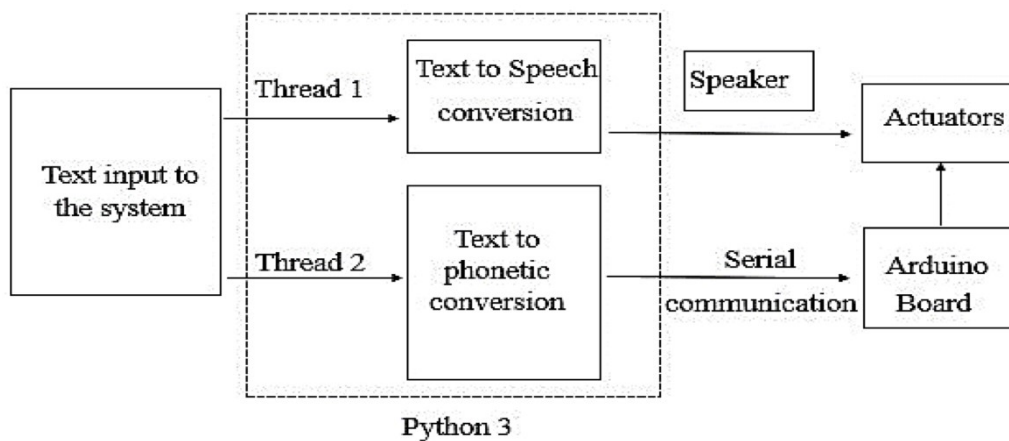


Figure 2: Method of synchronization

phones (Phoneme pair), and *Triphones* (Phoneme triple) is not possible by associating animation with individual phonemes. Machine learning techniques need the *Diphones* most of the time, representing timing between the middle of one phoneme and the next. *ProPhone* animation is intuitive, whereas the *Diphone* is nonintuitive represents (Shen et al., 2018). Statistics can be now simply calculated by counting the number of occurrences of each phoneme, phoneme pair, and phoneme triple in analyzed text, where each phoneme is just a symbol (Ziołko, Gałka, and Ziołko, 2009). In the process of pronunciation, there is a transition when different vowels are combined. The length of the upper and lower part of the mouth is very smaller at the natural state, but it increases when opening the pronunciation. The width of the mouth shape is larger when the mouth is in a natural position, but it will decrease with the opening pronunciation. The left and right corner distance, the width of the mouth, is the smallest in the stage of keeping the mouth (Liu et al., 2020). A comparison of existing humanoid robots is included in Table 2.

In this paper, we worked on developing a robotic lip movement mechanism that was to be synchronized with the sound corresponding to a series of words given as a

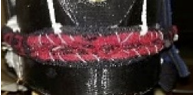






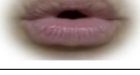
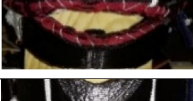
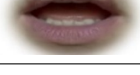
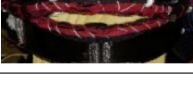

text file. Further, an analysis done to compare the lip movements of a human being to that of a robotic face is presented. The novelty of this work is the achievement of robotic lip movements very close to the human lip movements for a specific sentence with only three actuators. The material of the lips also helped to the precise movements. Moreover, the *g2p-en* python library was used to find phonetics in this work.

METHODOLOGY

Though the final objectives of the research were to design and build a humanoid robot with the ability to produce voice and synchronized lip movements to mimic a human being speaking the exact words, this paper concentrates more on the development of the jaw together with speech analyses done to derive the lip movements to achieve the text to speech ability and the synchronized lip movements.

Text to speech conversion: In various applications, machines are made to be able to communicate with humans using text-to-speech technology. When the text is input, the machine converts it into the human voice generated artificially. Recently, various kinds of TTS

Table 3: Categorized lip shapes for phonetic symbol

Lip shape		IPA symbol	Python symbol
Robot	Human		
		Silence, p, b, m	B, P, M, EH1
		æ, r, l, f, v, d, t, n, k, g, η	AA1, R, EH1, L, F, V, IY1, D, T, N, EY1, K
		a, o	OW1, AH1
		w, u	UW0, Y, UW1
		e, ε, ʊ, j, i, l, o, aʊ,ɔl, al	IY1, AH0, AY1, JH
		ɹ, h, s, Z, ʃ, tʃ, dʒ, ʒ, θ, ð	CH, S, Z, TH, T

systems have been offered on commercial platforms. In this development, we decided to use the *Pytt3* python library (Natesh, 2017) to convert text files into human voices because of its flexibility, easy integration with the software, and offline usage. The input text file is giving as a word document, and it reads word by word and converts to the voice.

Implementation of robot mouth module: Selecting appropriate control points and the directions of movements required to generate each lip shape is necessary for achieving expected lip movements. As a starting point, a 3D model of the robot mouth was designed and constructed using 3D printing technology using Polylactic Acid. Figure 1(a) shows the 3D design of the robot's mouth. The lip portions were finished with an elastic material, and actuators were fixed to obtain the jaw movements of the 3D human mouth design. The points thus identified were actuated by three servo motors controlled by the Arduino microcontroller. The mouth module was made to operate with three degrees of freedom. A mechanism is designed to obtain various lip shapes in synchronization with the sound output of the phonemes in each word to be pronounced. The identified control points are shown in Figure 1(b).

Lip synchronization: The lip synchronization mechanism was designed after several trials, as shown in Figure 2, consisting of text to speech conversion, lip movement, and lip synchronization handler stage. Identifying the foundation Phonemes of the given text file is the first step of this process. Text to speech system is not suitable to be applied directly to the synchronization process. In

the English language, there are 44 phonetic symbols. The input text file, a MsWord document or a text document is converted into phonetic symbols. *G2p-en* library was used to convert the text into phonetics. However, the symbols in the library were different from the International Phonetic Alphabet (IPA) symbols (Lenard, 2020). IPA symbols and corresponding *g2p-en* library (Pine, 2021) phonetic symbols are categorized into six groups according to the similarities in lip movement and mapped with the robot lip shapes, as shown in Table 3.

RESULTS AND DISCUSSIONS

We selected a comparison method to measure it with real human lip movement patterns. Face landmark detecting library "D-lib" (Facial landmarks with *dlib*, *OpenCV*, and *Python-PylImageSearch*, 2020) was used to identify the patterns of lips. The library can be used to identify 68 landmarks of the human face, and we used it to plot the mouth shape only. Testing was carried out in two phases as described below.

Lip movement patterns of human samples (phase I): We collected a set of videos of three males and one female pronounce the same phrase, "Hi, everyone". All the videos were recorded with the same frame rate of 30 fps and the same resolution of 640×352. Some frames of four videos are shown in Figure 3(a). Four mouth landmarks were detected in this figure for further analysis. The width (*a*) and the height (*b*) of mouth (Figure 3(b)) were measured in each of the videos, frame by frame.

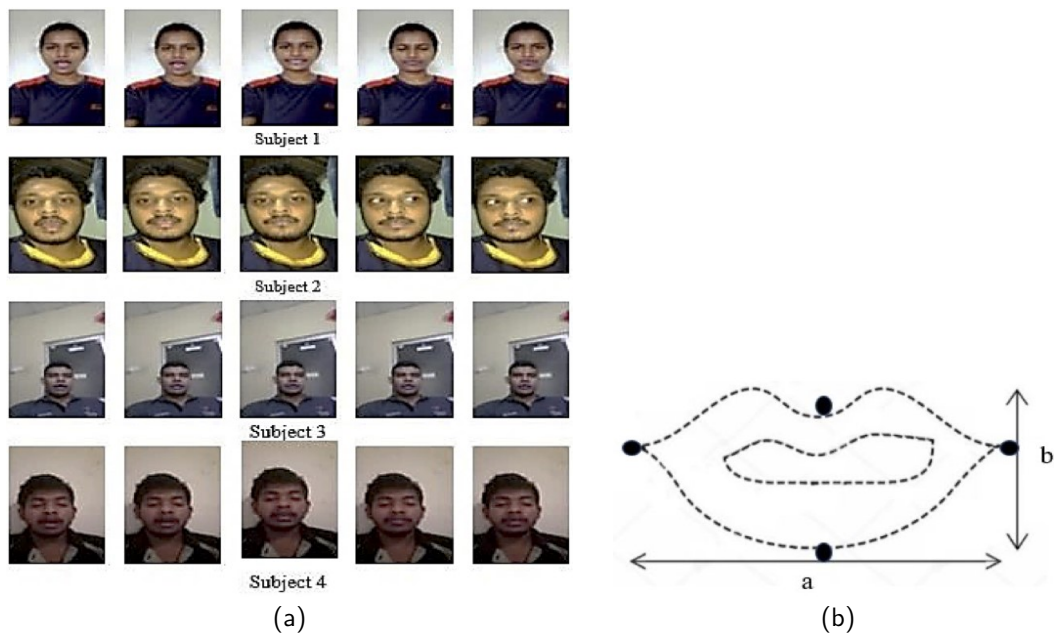


Figure 3: (a) Five sample frames extracted from videos (four subjects), and (b) mouth shape parameters

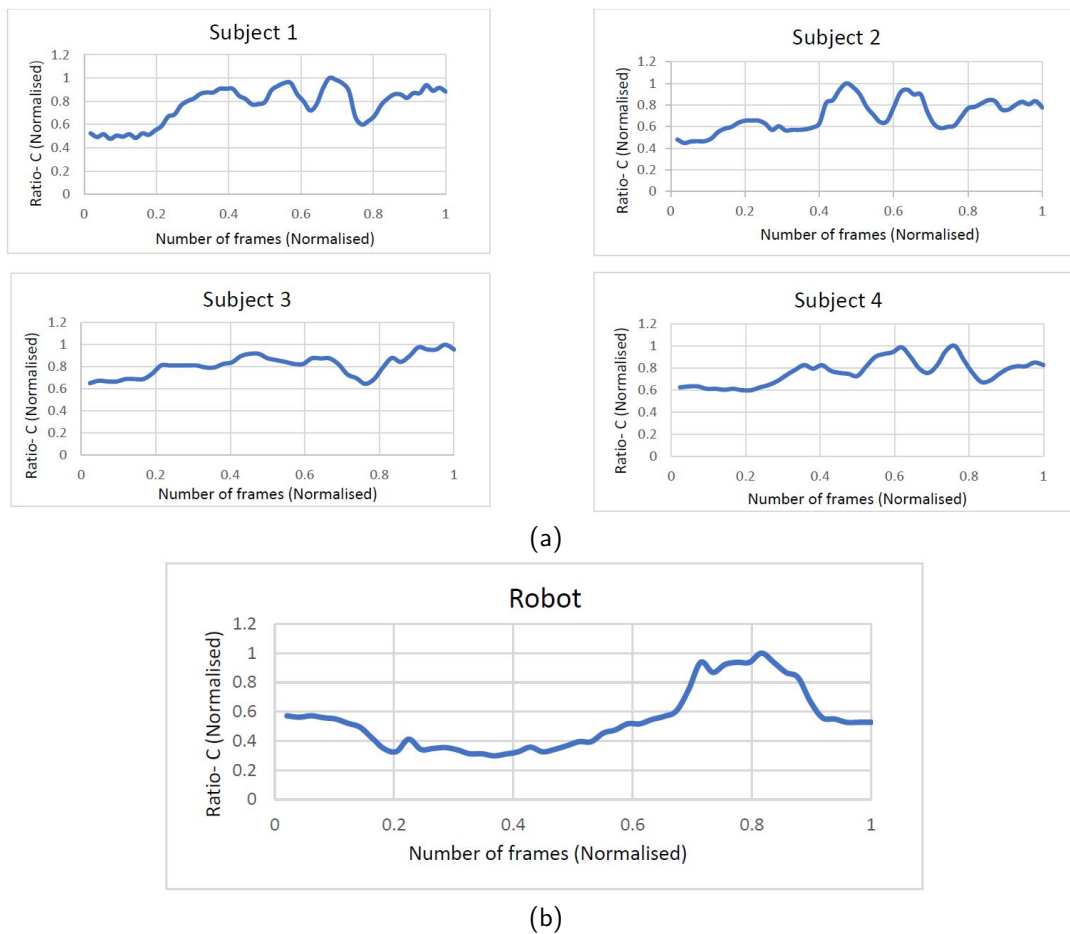


Figure 4: Variation of normalized width to height ratio c to the normalized frame number, (a) for all four subjects, (b) for robot

Then the ratio (c) between the two measures was calculated using equation 1,

$$c = \frac{a}{b} \quad (1)$$

The duration of the videos and total frame numbers were not the same among the human speakers. This could be one factor that gives variation in lip ratio patterns, and also it hindered performing direct comparisons between the variations. To minimize this unwanted effect, both the x and y axes of Figure 4 graphs were normalized using the relationship given in (equation 2), So that all data are scaled to the same time frame.

$$n_i = \frac{r_i}{r_{max}} \quad (2)$$

where, n_i represents the normalized values, r_i represents the real values, r_{max} represents the maximum real values. All these values are in matrix form.

Lip movement pattern of robot vs. human (phase II): To achieve human-like lip movements, it was necessary to compare and tune the lip movement of the robotic simulation. We used the manual length measuring technique of the captured frames from the mimicking video of the robot. MATLAB was used to measure the distances.

Figure 4(a) shows the variation of the calculated parameter c derived from Subject 1 to 4, plotted against the frame number. It could be seen that the patterns are similar even though there were additional noise and time variations due to the differences in the speaking style of human subjects. Several mathematical operations and parameters thus derived could be used to do a shape comparison among the graphs. One such approach is described herein. As can be seen in Figure 4(b) the pattern of the robot mouth is not exactly equal to the other patterns of the human subjects. But we can deduce that the robot design at the simulation stage at the moment is capable of mimicking the human speech patterns by its lips.

CONCLUSION

Lip synchronization is a critical factor in producing voice outputs of humanoid robots that are supposed to replace humans in performing specific tasks. Precise and natural lip synchronization with the voice output of robots would enable them to make favourable impressions on users, and thus is likely to have a beneficial influence on human-robot interaction. The main challenges were designing precise lip shapes with actuators, syncing them with the audio speech, finding the appropriate material for the lip design, and maintaining smooth movements. The proposed lip synchronization method is enough to generate synchronized lip movements and audio generated

through a text-to-speech system. Future developments would include a complete robot face and producing more accurate facial expressions along with speech. The used image processing and comparison technique will be developed further by increasing the number of video samples with equal quality and very accurate length measurements.

REFERENCES

- Breazeal, C., Edsinger, A., Fitzpatrick, P., Scasselati, B. and Varchavskaia, P. (2000) 'Social constraints on animate vision', IEEE Intelligent Systems and their Applications, 15(4), pp. 32–37. doi: 10.1109/5254.867910.
- Facial landmarks with dlib, OpenCV, and Python - PyImageSearch (2020). Available at: <https://www.pyimagesearch.com/2017/04/03/facial-landmarks-dlib-opencv-python/> (Accessed: 17 March 2021).
- G2 (2021) Compare Text to Speech Software. Available at: <https://www.g2.com/categories/text-to-speech>.
- Hackel, M., Schwoppe, S., Fritsch, J., Wrede, B. and Sagerer, G. (2005) 'Humanoid robot platform suitable for studying embodied interaction', in 2005 IEEE/RSJ International Conference on Intelligent Robots and Systems. 2005 IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 2443–2448. doi: 10.1109/IROS.2005.1544959.
- Hashimoto, T., Hitramatsu, S., Tsuji, T. and Kobayashi, H. (2006) 'Development of the Face Robot SAYA for Rich Facial Expressions', in 2006 SICE-ICASE International Joint Conference. 2006 SICE-ICASE International Joint Conference, pp. 5423–5428. doi: 10.1109/SICE.2006.315537.
- Hyung, H., Ahn, B., Choi, D. and Lee, D. (2016) 'Evaluation of a Korean Lip-sync system for an android robot', in 2016 13th International Conference on Ubiquitous Robots and Ambient Intelligence (URAI). 2016 13th International Conference on Ubiquitous Robots and Ambient Intelligence (URAI), pp. 78–82. doi: 10.1109/URAI.2016.7734025.
- Liu, Z., Kang, X., Nishide, S. and Ren, F. (2020) 'Vowel priority lip matching scheme and similarity evaluation model based on humanoid robot Ren-Xin', Journal of Ambient Intelligence and Humanized Computing. doi: 10.1007/s12652-020-02175-9.
- Natesh M Bhat (2017) Using pyttsx3 — pyttsx3 2.6 documentation. Available at: <https://pyttsx3.readthedocs.io/en/latest/engine.html> (Accessed: 1 September 2020).
- Oh, J., Hanson, D., Kim, W., Han, Y., Kim, J. and Park, I. (2006) 'Design of Android type Humanoid Robot Albert HUBO', in 2006 IEEE/RSJ International Conference on Intelligent Robots and Systems. 2006 IEEE/RSJ International Conference on

- Intelligent Robots and Systems, pp. 1428–1433. doi: 10.1109/IROS.2006.281935.
- Oh, K., Jung, C., Lee, Y. and Kim, S. (2010) 'Real-time lip synchronization between text-to-speech (TTS) system and robot mouth', in 19th International Symposium in Robot and Human Interactive Communication. 19th International Symposium in Robot and Human Interactive Communication, pp. 620–625. doi: 10.1109/ROMAN.2010.5598656.
- Pine, A. (2021) g2p: Module for creating context-aware, rule-based G2P mappings that preserve indices. Available at: <https://github.com/roedoejet/g2p> (Accessed: 17 March 2021).
- Raheem Ali, I., Sulong, G. and Kolivand, H. (2015) 'Realistic Lip Syncing for Virtual Character Using Common Viseme Set', *Computer and Information Science*, 8(3), p. p71. doi: 10.5539/cis.v8n3p71.
- Shen, J., Pang, R., Weiss, R.J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerrv-Ryan, R., Saurous, R.A., Agiomvrgiannakis, Y. and Wu, Y. (2018) 'Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions', in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4779–4783. doi: 10.1109/ICASSP.2018.8461368.
- Trovato, G., Kishi, T., Endo, N., Hashimoto, K. and Takanishi, A. (2012) 'Development of facial expressions generator for emotion expressive humanoid robot', in 2012 12th IEEE-RAS International Conference on Humanoid Robots (Humanoids 2012). 2012 12th IEEE-RAS International Conference on Humanoid Robots (Humanoids 2012), pp. 303–308. doi: 10.1109/HUMANOIDS.2012.6651536.
- Zanotto, D., Rosati, G., Minto, S. and Rossi, A. (2014) 'Sophia-3: A Semiadaptive Cable-Driven Rehabilitation Device With a Tilting Working Plane', *IEEE Transactions on Robotics*, 30(4), pp. 974–979. doi: 10.1109/TRO.2014.2301532.
- Ziołko, B., Gałka, J. and Ziołko, M. (2009) 'Phone, diphone and triphone statistics for Polish language', *St. Petersburg*, p. 5.