

Viseme Morphing and Text-to-Speech Integration for Indonesian News Broadcasting

Mirza Ardiana¹

mirzaardiana@ppns.ac.id¹

¹Politeknik Perkapalan Negeri Surabaya

Article Information

Received : 12 Apr 2025

Revised : 31 May 2025

Accepted : 16 Jun 2025

Keywords

Animation, Finite State Automata (FSA), Viseme Morphing, Text-to-Speech

Abstract

Advancements in multimedia technology and artificial intelligence have driven innovation in digital broadcasting, including virtual newsreaders. This study proposes a text-to-speech-based lip-sync animation system specifically for the Indonesian language to improve synchronization between lip movements and speech. The primary challenge in developing this system lies in generating realistic lip animations that correspond with the phonetic structure of Indonesian. The system workflow involves text input, syllable parsing using the Finite State Automata (FSA) method, viseme conversion (viseme morphing), and web-based animation output. Test results show a viseme duration accuracy of 98.5%, voice-lip movement synchronization of 94.26%, and a Mean Opinion Score (MOS) of 77.12%, indicating that the system is reasonably feasible for implementation. Despite minor delays, the system demonstrates strong potential for further development through the integration of Natural Language Processing (NLP) and deep learning, which could improve viseme mapping accuracy and enhance system flexibility across various digital broadcasting platforms.

A. Introduction

Advancements in multimedia and artificial intelligence have revolutionized the way information is delivered, particularly in digital broadcasting media. One increasingly popular innovation is the use of virtual automatic speakers capable of delivering news through text-to-speech (TTS) technology. However, to enhance the naturalness of digital communication, voice output alone is not sufficient. Synchronization between lip movements and the generated audio is also essential. This technique, known as viseme morphing, integrates TTS output with facial articulation movements, thereby creating a more realistic and easily comprehensible communication experience for the audience [1][2][3].

A viseme is the visual representation of a phoneme, the smallest unit of sound in a language that distinguishes meaning. In the domain of facial animation synthesis or speech-driven facial animation, visemes play a crucial role in ensuring that the lip movements of a digital character are accurately synchronized with the produced speech, thereby enhancing the sense of realism [4]. Proper implementation of visemes has been shown to improve the clarity of visual articulation and strengthen the audience's understanding of audio content, particularly in text-to-speech (TTS)-based systems. This becomes essential in various modern applications such as digital news broadcasting, AI-based virtual assistants, and interactive, animation-based educational media. Recent studies also indicate that the integration of dynamic and contextual visemes can significantly enhance user trust in virtual communication systems [5].

Most studies and developments in viseme morphing technology to date have been dominated by the use of English or other international languages that possess abundant linguistic resources. This has resulted in disparities in the application of natural language processing technologies for underrepresented languages, including Indonesian. One of the key aspects of creating a convincing animated character is the ability to accurately synchronize lip movements with speech [6]. In animation, voice is a core component of character expression. Speech animation has traditionally been regarded as an essential yet labor-intensive task, particularly in achieving accurate lip synchronization. This complexity arises from the dynamic interaction of facial muscles. Although various methods have been developed to simplify the creation of facial and speech animations, only a few are capable of providing fast and efficient solutions [7].

A previous study developed a visual speech synthesis system for the Indonesian language by utilizing viseme morphing and syllable merging through Finite State Automata (FSA) to enhance pronunciation learning [8]. By recording 1,029 sentences to build a viseme model synchronized with audio, the system was tested on 30 respondents and received naturalness and synchronization scores above 4.0. The results indicate that this approach is effective and realistic in supporting phoneme visualization, making it potentially beneficial for both language learning and speech therapy.

Another study addressed the challenges of phoneme-to-viseme mapping in Indonesian lip synchronization, particularly those related to the complexity of facial muscle movements and allophonic vowels [9]. The application of pre-processing through formant frequency feature extraction, followed by evaluation using Hidden Markov Models (HMM), significantly improved the accuracy of the

mapping. The results demonstrated that this approach is more reliable and relevant for handling the phonetic diversity of the Indonesian language.

Therefore, this study focuses on integrating viseme morphing and text-to-speech within the context of Indonesian-language news broadcasting. One of the main challenges in speech synchronization is generating realistic lip animations, considering that each language has distinct visual phonemes (visemes), which makes it difficult to develop a universally synchronized system. To date, no lip-sync tool has provided optimal results for the Indonesian language [10]. To address this challenge, the study proposes the development of a text-to-speech-based lip-sync animation website specifically designed for Indonesian. The animation process employs viseme morphing techniques and Finite State Automata (FSA) for word parsing, which is then integrated into a web-based platform using JavaScript. This system is implemented to support more natural and communicative virtual news broadcasting.

B. Research Method

In developing the integrated system of viseme morphing and text-to-speech for news broadcasting applications, the following block diagram illustrates the main stages of the research process:

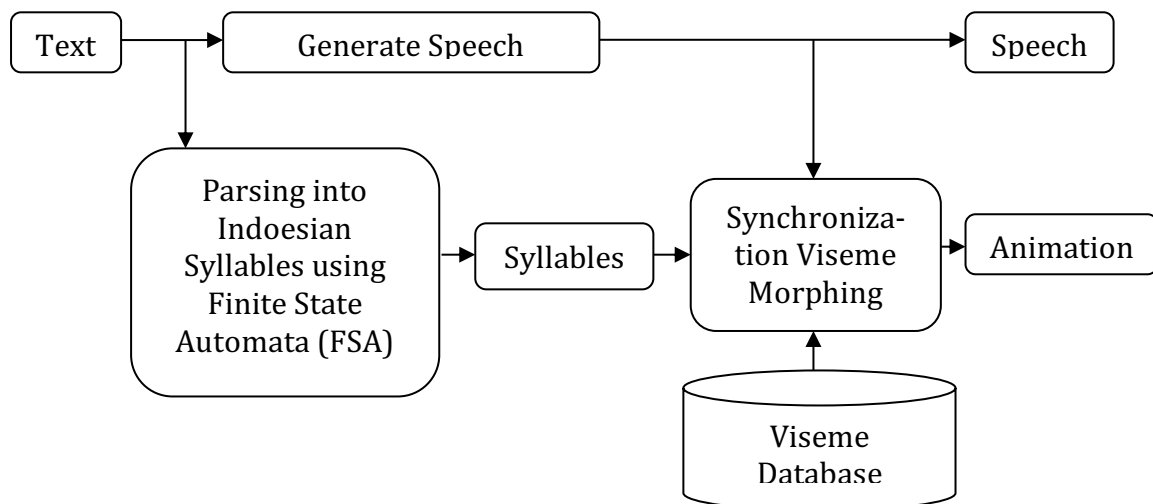


Figure1. Research Block Diagram

Based on the research block diagram presented above, the stages are explained as follows:

1. Input of Indonesian Text

The process begins with the input of Indonesian text, which may consist of letters, syllables, words, sentences, or paragraphs.

2. Generate Speech

In this stage, once the Indonesian text is entered into the system, the next step is generating speech using the Google Speech API. This technology automatically converts text into audio with natural-sounding voice quality, thereby supporting better synchronization with lip movements (visemes) in the resulting animation.

3. Parsing with Finite State Automata (FSA)

A Finite State Automata (FSA) is a mathematical model capable of receiving input and producing output. It consists of a finite number of states and transitions from one state to another based on the input and defined transition rules [11]. As it lacks memory or data storage capabilities, the FSA can only retain its current active state [12].

In this system, the FSA is used to recognize input characters and convert them into syllables. It functions as an abstract machine that identifies and segments words within a sentence. This study employs a three-level FSA model, which is illustrated in the following diagram:

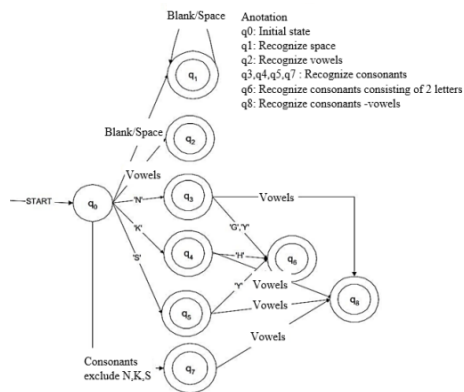


Figure 2. FSA Level 1

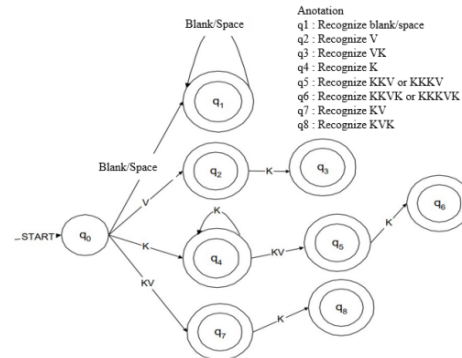


Figure3. FSA Level 2

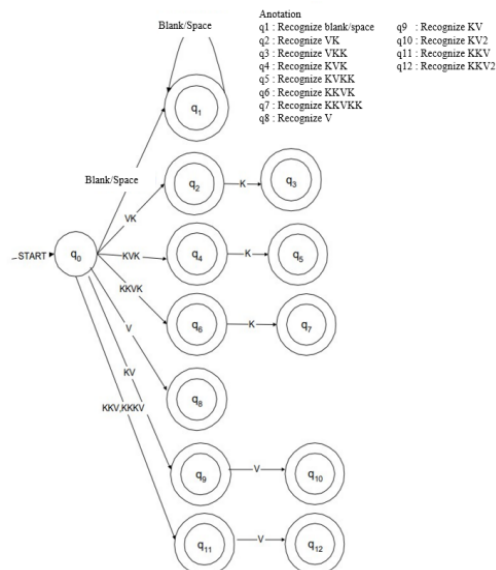


Figure 4. FSA Level 3

The Finite State Automata (FSA) is a type of machine commonly used for basic language recognition. Therefore, the FSA method can be effectively applied to the process of syllable segmentation. The FSA implemented in this study is designed with three hierarchical levels [13]. At the first level, the system identifies basic character patterns such as V (vowel), C (consonant), or VC. The output from this level then serves as the input for the next stage. At the

second level, the FSA recognizes more complex syllable structures, including patterns such as V, VC, VCC, CV, CVC, CCV, CCVC, CCCV, and CCCVC. However, at this stage, certain patterns such as VCC, CVCC, and CCVCC cannot yet be detected. Therefore, a third level FSA is incorporated to identify these remaining syllabic patterns [14][15].

4. Viseme Morphing

Viseme morphing is the process of gradually transforming the visual shape of the mouth (viseme) from one form to another, creating the illusion of smooth and natural mouth movement during speech in animation or facial synthesis [16][17]. The following flowchart illustrates the stages of the viseme morphing process:

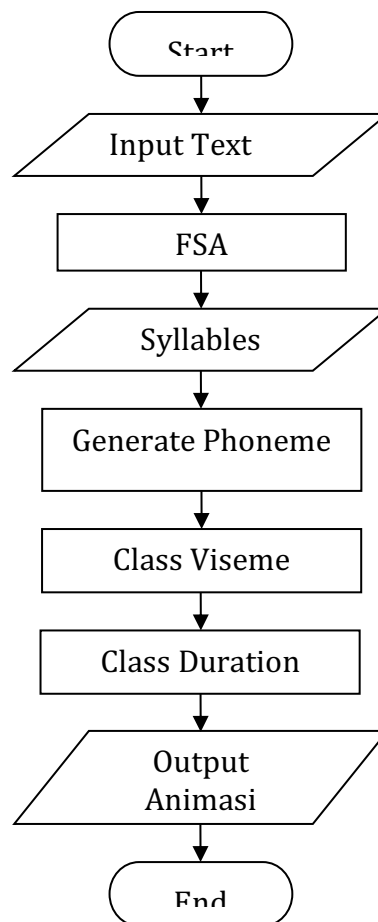


Figure 5. Flowchart Morphing Viseme

Based on the flowchart above, when the input is Indonesian text, the syllable parsing process is immediately carried out using the Finite State Automata (FSA) method. The process then continues through the following stages:

4.1. Syllable Text to Phonem

At this stage, the Indonesian text is converted into phonemes, which are the smallest units of sound[18]. This study uses a phoneme classification based on 33 Indonesian phonemic symbols, consisting of 10 vowels (including

diphthongs), 22 consonants, and 1 silent symbol, as shown in the following table.

Table1. Phoneme Classification













No.	Type of Phoneme	Phonemes
1	Consonants	p, t, l, g, j, z, m, sy, ng, w, v, b, k, d, c, f, h, kh, n, r, y, ny, silent
2	Single Vowels (Monophthongs)	a, e, E, i, o, u
3	Double Vowels (Diphthongs)	ao, au, ai, oi

















Based on the table, the input Indonesian text is classified into the corresponding phonemes. This process is essential for the subsequent stage, which involves determining the Viseme Class and Duration Class. When the user inputs a sentence, the system recognizes the text characters, converts them into phonemes (the smallest sound units that differentiate meaning), and then proceeds to the identification stage of visemes and their durations.

4.2. Class Viseme

A viseme is the visual representation of a group of phonemes that produce similar mouth movements when spoken [18][19]. Therefore, it is essential to design mouth animations that correspond to the predefined viseme classifications. Accurate phoneme-to-viseme mapping is crucial for improving the quality of automatic lip movement animation, particularly in synchronizing speech and mouth movements in the Indonesian language [9]. Vowels play a dominant role in speech recognition and lip animation, as they have the highest energy and longest duration in speech signals [20]. The following are the viseme classes used in this study:

Table2. Viseme Class Classification

Phoneme Class	Viseme Design	Animation Output	Phoneme Class	Viseme Design	Animation Output
Silent			/d/,/n/,/t /,/ny/,/n g/		
/e/			/b/,/m/,/p/		
/i/,/y/			/a/,/au/,a i/		

/u/,/w/			/h/,/g/,/ k/,/q/,/x /,/kh/		
/s/,/z/			/l/		
/c/,/j/			/o/,/oi/		
/f/,/v/			/r/		

4.3. Class Duration

The Duration Class stage aims to determine the duration of each viseme based on predefined parameters. This duration plays a crucial role in maintaining the alignment of the animation during information delivery. In this study, the duration was analyzed by measuring the audio signals of several sentences over time. The parameters analyzed include vowels (a, i, u, e, o), initial word pauses, inter-word pauses, and the duration of each consonant. Based on the analysis and average duration calculations from a number of sample sentences, the following duration classes were obtained:

Table3. Class Durations

Phoneme	Durations (ms)	n Frame	Phoneme	Durations (ms)	n Frame
'a'	125	25	'k'	130	26
'i'	100	20	'l'	95	19
'u'	105	21	'm'	100	20
'e'	85	17	'n'	90	18
'o'	105	21	'p'	120	24
Initial Silence '-'	60	12	'q'	145	29
Inter-word Silence '-'	5	1	'r'	80	16
'b'	115	23	's'	115	23
'c'	130	26	't'	115	23
'd'	105	21	'v'	95	19

'f'	95	19	'w'	110	22
'g'	125	25	'x'	120	24
'h'	100	20	'y'	85	17
'j'	145	29	'z'	135	27

Based on the data in the table above, once the input text has been classified into phonemes, the system determines the appropriate viseme along with its display duration. For example, if the system detects the phoneme 'a', the viseme [a].png will be displayed for 125 ms. With a frame rate of 200 fps (where one frame lasts 5 ms), the viseme will be shown for 25 frames to match the specified duration. This principle is applied to all phoneme classifications listed in the table.

In the Indonesian language, there are also diphthongs and consonant clusters, which are combinations of two vowels or two consonants pronounced together. In such cases, the viseme duration is calculated based on the average duration of the individual component letters. The detailed data are presented as follows:

Table 4. Duration Class Data for Diphthongs and Consonant Clusters

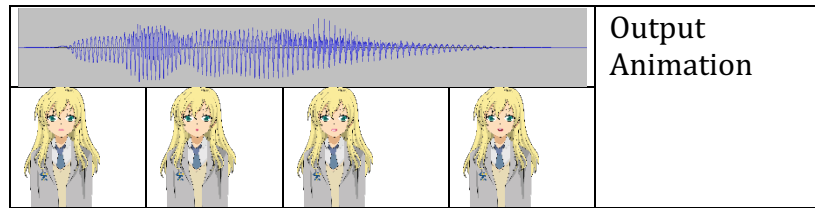
Phoneme	Average Duration (ms)	Programmed Duration (ms)	n Frames
KH	117	115	23
NG	110.2	100	22
NY	87.5	85	17
SY	102.1	100	20
AI	114.29	110	22
AU	117.59	115	23
AO	118.45	115	23
EI	95.08	90	18
OI	105.63	105	22

4.4. Output Animation

This stage is the final step in the viseme mapping process, aiming to synchronize all previous stages to produce an animated output accompanied by audio. The audio is generated using the Google Speech API and synchronized with the animated mouth movements based on their respective durations. The following is an illustration of the animation output, using the word "Bunga" as an example:

Table 5. Viseme Mapping Implementation

/b/	/u/	/ng/	/a/	Text-To-Phonem
Mulut [B].png	Mulut [U].png	Mulut [NG].png	Mulut [A].png	Class Viseme
80 ms	100 ms	80 ms	125 ms	Class Duration



C. Result and Discussion

This section presents the results of the website interface development and the integration of viseme morphing with text-to-speech for news broadcasting media. The website is designed as an interactive platform that allows users to input Indonesian text, which is then processed into an automated talking-face animation. The animation displays mouth movements that are adjusted to match the sequence of phonemes and visemes, and is synchronized with the synthesized voice generated by Google Speech. The main features and interface of the developed website are as follows:

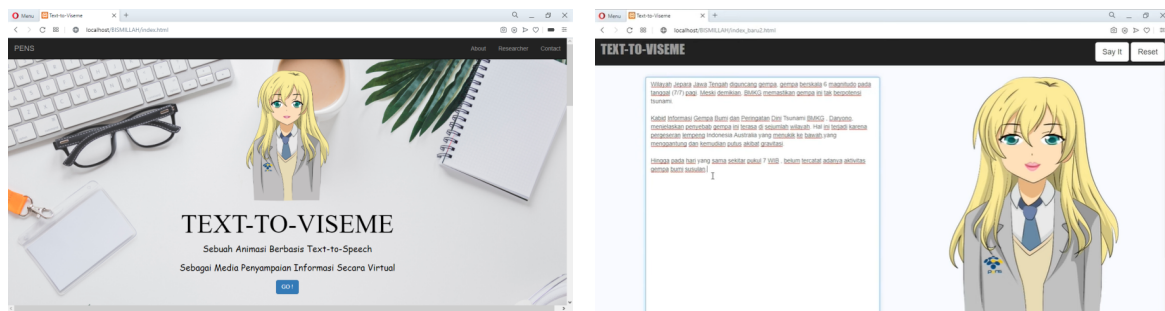


Figure 6. Website Interface Display

To enhance the realism of facial animation in this study, an animated eye element was added to simulate natural movement and blinking. The design utilizes four variations of eye images: Fully Open Eyes, Slightly Closed Eyes, Moderately Closed Eyes, and Fully Closed Eyes. Transitions between these images are displayed sequentially to create a smooth eye movement effect, thereby supporting the overall facial expression.

Table 6. Eye Variations in Animation

Eye Type	Eye Shape	Eye Type	Eye Shape
Fully Open Eyes		Moderately Closed Eyes	
Slightly Closed Eyes		Fully Closed Eyes	

The testing results in this study cover three main aspects aimed at evaluating the performance and quality of the viseme-based lip movement animation system. These tests include: (1) Synchronization of Viseme Display Duration, (2) Synchronization between Audio and Viseme, and (3) Subjective Evaluation using

the Mean Opinion Score (MOS) method. Each test was conducted to assess how effectively the system presents animations that are synchronized with the audio and deliver a realistic visual experience for users.

1. Viseme Display Duration Synchronization Testing

The purpose of this test is to evaluate the difference between the theoretical viseme display duration and the actual duration during program execution. The duration in question refers to the total display time of visemes based on Indonesian text input, as presented through the web-based system. The test was conducted by comparing the theoretical duration calculated using the viseme class duration formula with the actual duration measured during execution. The difference between the two reflects the system's accuracy in displaying visemes according to the specified duration. The following are the test results obtained using sample words and sentences:

Table7. Viseme Display Duration Synchronization Test Data

No.	Sample	Trial	Experimental Duration (ms)	Theoretical Duration (ms)	Error (%)
1	<i>Yakin</i>	1	532.15	530	0.41
		2	537.59		1.43
		3	531.36		0.26
2	<i>Ibu sedang memasak</i>	1	1716.5	1700	0.97
		2	1722.31		1.31
		3	1728.4		1.67
3	<i>Ia menghitung 2 % dari uang nya</i>	1	3052.18	2980	2.42
		2	3124.34		4.84
		3	3162.57		6.12

The test results indicate a delay in both word and sentence samples during program execution. This delay is observed from the discrepancy between the theoretical viseme display duration and the actual duration measured during runtime, using the **performance.now** method. The likely cause of the delay is the additional computation time required by the system, such as looping processes and interactions with the database, which increase the computational load and impact the viseme display duration. The following shows the percentage of synchronization success in the viseme display duration test, based on samples of 26 words, 15 sentences, and 10 acronym sentences:

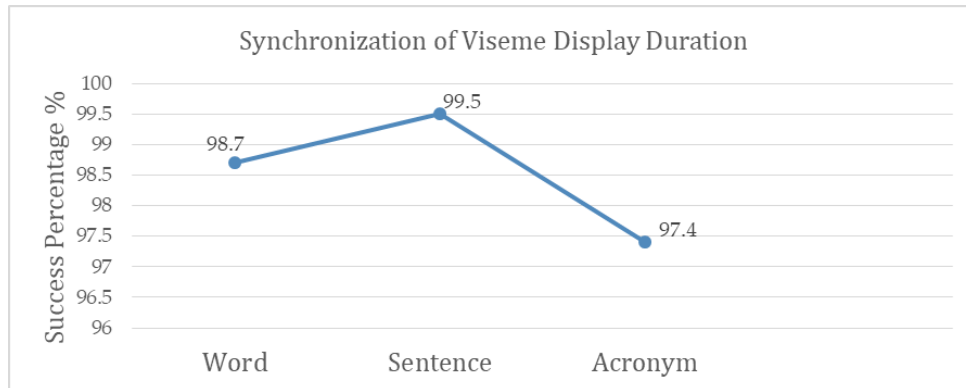


Figure 7. Viseme Display Duration Synchronization Test Results

2. Audio and Viseme Synchronization Test

The purpose of the audio and viseme synchronization test is to measure the alignment between the duration of the audio produced by the lip-sync animation and the duration of the corresponding mouth movements (visemes). The test results reveal three possible outcomes: perfect synchronization, delayed mouth movements compared to the audio, or mouth movements occurring faster than the audio. The audio duration is measured using the Audacity application to determine the length of each audio sample. Meanwhile, the duration of the viseme is measured using the **performance.now** function, which records the viseme display time in milliseconds, starting from the moment the “Say It” button on the web system is activated. The comparison between these two durations determines the level of synchronization and identifies any significant discrepancies, which is the main focus of this test. The following shows the success rate of the audio and viseme synchronization test based on samples consisting of 26 words, 15 sentences, and 10 acronym-based sentences:

Table 8. Audio and Viseme Synchronization Test Data

No.	Sample	Trial	Experimental Duration (ms)	Theoretical Duration (ms)	Error (%)
1	<i>Abjad</i>	1	546.01	555	1.62
		2	536.42		3.35
		3	537.3		3.19
2	<i>Dia sedang membaca</i>	1	1731.41	1650	4.93
		2	1733.28		5.05
		3	1734.98		5.15
3	<i>Dia mempunyai \$50</i>	1	2713.32	2383	13.86
		2	2697.48		13.2
		3	2707.21		13.6

The results of the synchronization test between audio and viseme in the lip-sync animation indicate a time discrepancy between the audio and the animated mouth movements. One of the main causes is the delay introduced by computational processes during program execution. Additionally, the audio system in this study relies on the Google Speech service, which requires an internet

connection. The synchronization delay is likely influenced by unstable internet connectivity. Therefore, more accurate synchronization could be achieved by ensuring a stable internet connection, allowing the audio and viseme to run in precise alignment. The diagram below shows the percentage of success in the audio and viseme synchronization test for words, sentences, and acronyms.

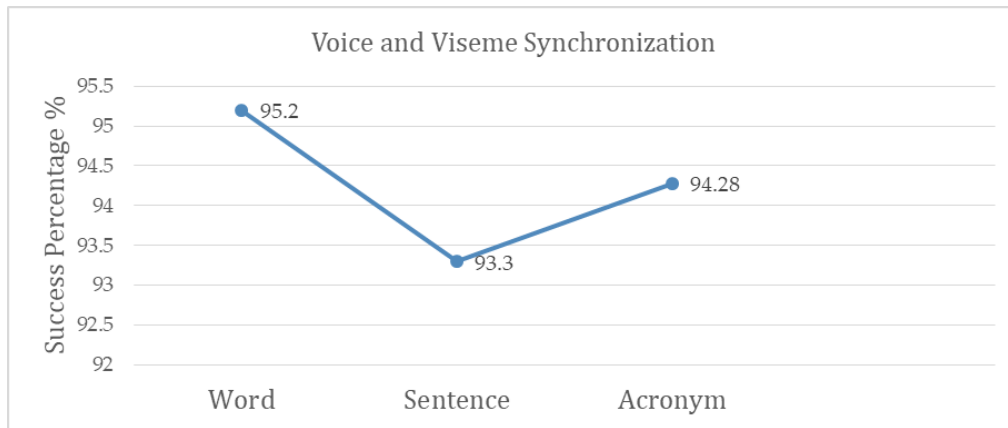


Figure 8. Results of Audio and Viseme Synchronization Test

3. Mean Opinion Score (MOS) Evaluation

The Mean Opinion Score (MOS) evaluation was conducted to assess the system subjectively through a questionnaire. The questionnaire focused on the synchronization between audio and mouth movements (visemes), with respondents rating the accuracy of the animation in matching these two elements. The evaluation employed the Absolute Category Rating (ACR) method, which classifies quality on an ordinal scale from highest to lowest. The results of the evaluation are presented as follows:

Table 9. Mean Opinion Score (MOS) Evaluation Results

No.	Evaluation of lip-sync animation pronunciation accuracy based on:	Result MOS
1	A single word:	<p>35 tanggapan</p> <p> 1 (Buruk / Bad) 2 (Kurang Bagus / Poor) 3 (Cukup / Fair) 4 (Bagus / Good) 5 (Sangat Bagus / Excellent) </p>
2	Two or more words:	<p>35 tanggapan</p> <p> 1 (Buruk / Bad) 2 (Kurang Bagus / Poor) 3 (Cukup / Fair) 4 (Bagus / Good) 5 (Sangat Bagus / Excellent) </p>

3	A full sentence:	<p>35 tanggapan</p> <p> 1 (Buruk / Bad) 2 (Kurang Bagus / Poor) 3 (Cukup / Fair) 4 (Bagus / Good) 5 (Sangat Bagus / Excellent) </p>
4	A sentence containing symbols, acronyms, and abbreviations:	<p>35 tanggapan</p> <p> 1 (Buruk / Bad) 2 (Kurang Bagus / Poor) 3 (Cukup / Fair) 4 (Bagus / Good) 5 (Sangat Bagus / Excellent) </p>
5	A paragraph:	<p>35 tanggapan</p> <p> 1 (Buruk / Bad) 2 (Kurang Bagus / Poor) 3 (Cukup / Fair) 4 (Bagus / Good) 5 (Sangat Bagus / Excellent) </p>

The results of the Mean Opinion Score (MOS) test indicate that the text-to-speech-based lip-sync animation web system is considered fairly feasible for implementation, with an average score of 77.12%. However, respondents suggested further development to improve several aspects, such as enhancing the realism of the animation, refining the synchronization between voice and mouth movements, and improving the web interface to be more flexible and compatible across various platforms.

D. Conclusion

This study successfully designed a virtual newsreader animation system based on text-to-speech for the Indonesian language, achieving a relatively high level of synchronization between speech and mouth movements. The implementation of the viseme mapping method proved effective in supporting the automatic broadcasting of information. The test results demonstrate that the system is capable of delivering content in the form of words, sentences, and paragraphs in a fairly synchronized and natural manner. Although minor delays still occur due to viseme duration settings, this presents opportunities for further development, particularly in refining the class duration parameters to improve lip-sync accuracy. For future improvements, the system could be enhanced through the integration of Natural Language Processing (NLP) technology to better recognize sentence context, intonation, and expressions that align with the news content. Additionally, the application of deep learning models such as LSTM could be used to automatically learn the phoneme-to-viseme mapping from audiovisual data, further improving the realism and accuracy of mouth movements.

Consequently, this system has the potential to evolve into a more flexible web-based solution that can be integrated across various digital broadcasting

platforms, including education, public services, and other AI-powered media applications.

E. References

- [1] E. Setyati, S. Sumpeno, M. H. Purnomo, K. Mikami, M. Kakimoto, and K. Kondo, "Phoneme-viseme mapping for Indonesian language based on blend shape animation," *IAENG Int. J. Comput. Sci.*, vol. 42, no. 3, pp. 1–12, 2015.
- [2] Arifin, S. Sumpeno, M. Hariadi, and A. M. Syarif, "Development of indonesian text-to-audiovisual synthesis system using syllable concatenation approach to support indonesian learning," *Int. J. Emerg. Technol. Learn.*, vol. 12, no. 2, pp. 166–184, 2017, doi: 10.3991/ijet.v12i02.6384.
- [3] B. Hao *et al.*, "LipGen: Viseme-Guided Lip Video Generation for Enhancing Visual Speech Recognition," pp. 2–6, 2025, [Online]. Available: <http://arxiv.org/abs/2501.04204>
- [4] C. Cao, Q. Hou, and K. Zhou, "Displaced dynamic expression regression for real-time facial tracking and animation," *ACM Trans. Graph.*, vol. 33, no. 4, 2014, doi: 10.1145/2601097.2601204.
- [5] S. Suwajanakorn, S. M. Seitz, and I. Kemelmacher-Shlizerman, "Synthesizing Obama," *ACM Trans. Graph.*, vol. 36, no. 4, pp. 1–13, 2017, doi: 10.1145/3072959.3073640.
- [6] G. Llorach, A. Evans, J. Blat, G. Grimm, and V. Hohmann, "Web-based live speech-driven lip-sync," *2016 8th Int. Conf. Games Virtual Worlds Serious Appl. VS-Games 2016*, pp. 1–4, 2016, doi: 10.1109/VS-GAMES.2016.7590381.
- [7] Y.-M. Chen *et al.*, "Animating Lip-Sync Characters," 2010.
- [8] Aripin, H. Haryanto, and S. Sumpeno, "A realistic visual speech synthesis for Indonesian using a combination of morphing viseme and syllable concatenation approach to support pronunciation learning," *Int. J. Emerg. Technol. Learn.*, vol. 13, no. 8, pp. 19–37, 2018, doi: 10.3991/ijet.v13i08.8084.
- [9] A. Rachman, R. Hidayat, and H. A. Nugroho, "Improving Phoneme to Viseme Mapping for Indonesian Language," *IJITEE (International J. Inf. Technol. Electr. Eng.)*, vol. 4, no. 1, p. 1, 2020, doi: 10.22146/ijitee.47577.
- [10] M. Liyanthy, H. Nugroho, and W. Maharani, "Realistic facial animation of speech synchronization for Indonesian language," *2015 3rd Int. Conf. Inf. Commun. Technol. ICoICT 2015*, pp. 563–567, 2015, doi: 10.1109/ICoICT.2015.7231486.
- [11] S. Chatterjee, K. Paul, R. Roy, and A. Nath, "A Pilot Study on Natural Language Processing--Applications of Finite State Automation," vol. 06, no. November, pp. 580–586, 2015.
- [12] F. H. Rachman, Qudsiyah, and F. Solihin, "Finite State Automata Approach for Text to Speech Translation System in Indonesian-Madurese Language," *J. Phys. Conf. Ser.*, vol. 1569, no. 2, 2020, doi: 10.1088/1742-6596/1569/2/022091.
- [13] R. Meiyanti and C. L. M. Sandy, "Implementation of Finite State Automata Method in Text to Speech Conversion System," *J. Adv. Comput. Knowl. Algorithms*, vol. 1, no. 3, p. 59, 2024, doi: 10.29103/jacka.v1i3.16917.
- [14] R. A. W, H. Tolle, and O. Setyawati, "Pengembangan Aplikasi Text-to -,"

- Pengemb. Apl. Text-to-Speech Bhs. Indones. Menggunakan Metod. Finite State Autom. Berbas. Android*, vol. 5, no. 1, 2016, [Online]. Available: https://www.academia.edu/61451518/Pengembangan_Aplikasi_Text_to_Speech_Bahasa_Indonesia_Menggunakan_Metode_Finite_State_Automata_Berbasis_Android
- [15] I. Indrianto, A. Abdurrasyid, M. Nur Indah Susanti, G. Fairus Ferdiansyah Deu, and A. Ramadhan, "Text to Speech Using Finite State Automata on Health Data," *J. EECCIS (Electrics, Electron. Commun. Control. Informatics, Syst.*, vol. 17, no. 1, pp. 1–7, 2023, doi: 10.21776/jeccis.v17i1.1631.
 - [16] I. Raheem Ali, G. Sulong, and H. Kolivand, "Realistic Lip Syncing for Virtual Character Using Common Viseme Set," *Comput. Inf. Sci.*, vol. 8, no. 3, 2015, doi: 10.5539/cis.v8n3p71.
 - [17] N. Dave and N. M. Patel, "Phoneme and Viseme based Approach for Lip Synchronization," *Int. J. Signal Process. Image Process. Pattern Recognit.*, vol. 7, no. 3, pp. 385–394, 2014, doi: 10.14257/ijcip.2014.7.3.31.
 - [18] Arifin, Muljono, S. Sumpeno, and M. Hariadi, "Towards building Indonesian viseme: A clustering-based approach," *Proceeding - IEEE Cybern. 2013 IEEE Int. Conf. Comput. Intell. Cybern.*, pp. 57–61, 2013, doi: 10.1109/CyberneticsCom.2013.6865781.
 - [19] Arifin, S. Sumpeno, Muljono, and M. Hariadi, "A model of Indonesian dynamic visemes from facial motion capture database using a clustering-based approach," *IAENG Int. J. Comput. Sci.*, vol. 44, no. 1, pp. 41–51, 2017.
 - [20] S. M. Hwang, B. H. Song, and H. K. Yun, "Korean speech recognition using phonemics for lip-sync animation," *Proc. - 2014 Int. Conf. Inf. Sci. Electron. Electr. Eng. ISEEE 2014*, vol. 2, pp. 1011–1014, 2014, doi: 10.1109/InfoSEEE.2014.6947821.