Hindawi
Advances in Multimedia
Volume 2023, Article ID 9804508, 1 page
https://doi.org/10.1155/2023/9804508

Retraction

Retracted: Evaluation and Analysis of Animation Multimedia 3D Lip Synchronization considering the Comprehensive Weighted Algorithm

Advances in Multimedia

Received 15 August 2023; Accepted 15 August 2023; Published 16 August 2023

This article has been retracted by Hindawi following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of one or more of the following indicators of systematic manipulation of the publication process:

(1) Discrepancies in scope
(2) Discrepancies in the description of the research reported
(3) Discrepancies between the availability of data and the research described
(4) Inappropriate citations
(5) Incoherent, meaningless and/or irrelevant content included in the article
(6) Peer-review manipulation

The presence of these indicators undermines our confidence in the integrity of the article's content and we cannot, therefore, vouch for its reliability. Please note that this notice is intended solely to alert readers that the content of this article is unreliable. We have not investigated whether authors were aware of or involved in the systematic manipulation of the publication process.

Wiley and Hindawi regrets that the usual quality checks did not identify these issues before publication and have since put additional measures in place to safeguard research integrity.

We wish to credit our own Research Integrity and Research Publishing teams and anonymous and named external researchers and research integrity experts for contributing to this investigation.

The corresponding author, as the representative of all authors, has been given the opportunity to register their agreement or disagreement to this retraction. We have kept a record of any response received.

References

[1] Z. Xu, "Evaluation and Analysis of Animation Multimedia 3D Lip Synchronization considering the Comprehensive Weighted Algorithm," *Advances in Multimedia*, vol. 2021, Article ID 8189082, 7 pages, 2021.

Hindawi
Advances in Multimedia
Volume 2021, Article ID 8189082, 7 pages
https://doi.org/10.1155/2021/8189082

Research Article

Evaluation and Analysis of Animation Multimedia 3D Lip Synchronization considering the Comprehensive Weighted Algorithm

Zhe Xu

Academy of Fine Arts South China Normal University, Guangzhou, Guangdong, China

Correspondence should be addressed to Zhe Xu; 20150700024@m.scnu.edu.cn

Received 18 August 2021; Accepted 23 October 2021; Published 30 November 2021

Academic Editor: Zhendong Mu

The 3D lip synchronization is one of the hot topics and difficulties in the field of computer graphics. How to carry out 3D lip synchronization effectively and accurately is an important research direction in the field of multimedia. On this basis, a comprehensive weighted algorithm is introduced in this paper to sort out the related laws and the time of lip pronunciation in animation multimedia, carry out the vector weight analysis on the texts in the animation multimedia, and synthesize a matching evaluation model for 3D lip synchronization. At the same time, the goal of simultaneous evaluation can be achieved by synthesizing the transitional mouth pattern sequence between consecutive mouth patterns. The results of the simulation experiment indicate that the comprehensive weighted algorithm is effective and can support the evaluation and analysis of animation multimedia 3D lip synchronization.

1. Introduction

With the continuous advancement of social economy, the human-computer interaction has become a common operation in more and more scenes. However, how to improve the accurate recognition rate and the effectiveness of human-computer interaction has become the focus of studies [1–3]. The matching of 3D face animation with the voice is an important application direction in the fields of visual communication and multimedia teaching. In the face animation, the corresponding voice coordination is required to match, and the effect presented in this way is continuous speech. Hence, how to establish a one-to-one mapping relationship between the mouth pattern and the speech and synthesize the mouth pattern synchronously based on the 3D technology becomes a research bottle neck and hotspot, and this approach has great application prospects in film and television, publicity, display, and other aspects [4–7]. For the purpose of achieving the synchronization of voice and 3D mouth pattern, it is necessary to establish a visual model of the voice. Through the consolidation and analysis, the voice information is converted into digital information, and the mapping of voice information and digital information is established accordingly. At the same time, the mapping relationship between digital information and visual information is established; that is, the voice information is converted into the visual information, and the unified coordination of voice and vision is implemented. In this way, the virtual pronunciation can be matched with the 3D virtual animation, which can further reduce the hysteresis of the virtual accent, enhance the effect of authenticity, lower the threshold of human-computer interaction, and improve the recognition and courtesy of people [8, 9]. In the aspect of the Chinese 3D mouth pattern multimedia animation, scholars have carried out the text-driven studies, such as the introduction of the basic expressions and mouth patterns to carry out synthesis, which has implemented the transformation of 3D facial mouth pattern animation under multiple expressions. In addition, the 3D mouth patterns are also classified according to the pronunciation and expression, and relatively realistic animation models have been clustered. However, there is a lack of continuous multiframe transformation. As a result, it is impossible to implement the synchronization with speech [10, 11].

For the purpose of addressing the limitations described above, a comprehensive weighted algorithm is introduced in this paper. Through the simultaneous combining of the Chinese phonetic mouth patterns, in conjunction with the synthesis of continuous frames of animation, a Chinese 3D mouth pattern voice database is established accordingly. The synthesis of 3D mouth pattern animation based on the Chinese voice synchronous multimedia is implemented through the application of phonological comprehensive weighting, label comprehensive weighting analysis, and other methods, with the purpose to explore the effect of evaluation and analysis of 3D lip synchronization in animation multimedia.

2. Analysis of the Mouth Pattern Characteristics in the Pronunciation of Chinese Pinyin

In accordance with the pronunciation rules and characteristics stipulated by the current standard Mandarin pronunciation, a comprehensive classification is carried out based on the clustering of mouth patterns, initial consonant consonants, and simple or compound vowels. In this study, the mouth patterns in the pronunciation of Chinese Pinyin (Chinese Phonetic Alphabet) are classified into three categories as follows:

(1) First-level mouth pattern is mainly simple mouth-opening type pronunciation, such as open mouth pattern *a*, as well as *g*, *k*, *h*, *e*, *f*, *b*, *p*, and *m*

(2) Second-level mouth pattern is mainly the changing mouth pattern, which may include four directions: front, back, left, and right, including *i*, *y*, *j*, *q*, *x*, *z*, *c*, *s*, zh, ch, shi, *r*, *o*, *u*, *w*, and *u*.

(3) Third-level mouth pattern is mainly the relaxed mouth pattern, with the primary difference in the relative position of the tongue and the throat, which mainly includes *d*, *t*, *n*, and *l*

2.1. Rhythm in Lip Sync Animation. With respect to the rhythm, its essence is the inherent characteristics of language, which is a combination of voice and rhythm. The rhythm is highly important, especially in the context [12–14]. For example, in a speech, the modulation in tone and vocal variety are required; when delivering address, it is not necessary to have excessive emotional ups and downs in the tone. With respect to the text of the rhythm, it is implemented by adding the corresponding rhythmic tags to the text. The rhythmic tag mentioned above is a universal symbol that can be accepted by the public; that is, the parsed language grammar can be recorded by using XML [15].

2.2. Overall Framework. The input data of the whole evaluation system is a language corpus that has been designed, in which four characteristic points of the common mouth patterns in people are sorted out to represent the width and height of the mouth pattern and obtain a function curve with the time as the independent variable and the height and width of the mouth pattern as the dependent variables. The largest lip pattern frame is selected as the static positioning, and the 2D visual elements are converted into 3D visual elements accordingly.

In accordance with the existing static visual element positioning of the 3D mouth pattern in the Chinese phonetic database available, the specific frame of the lip sync animation is shown in Figure 1.

In the specific 3D dynamic synthesis module, it is first necessary to extract the corresponding characteristic points of each voice element and carry out the one-to-one mapping by using the voice and video. It should be noted that, as the texture information corresponding to the mouth pattern is relatively complicated, the flexible matching rule is applied in this paper to track the position of the initial consonant and that of the simple or compound vowel.

The characteristic points are used to calculate the distance between the images to obtain the corresponding voice characteristic curve for the lip pattern (including width and height), as shown in Figure 2.

For the purpose of distinguishing the pronunciation rules of the selected subjects from the pronunciation characteristics of other people, the continuous voice analysis is carried out, and the 2D dynamic video frames are first distinguished. By distinguishing different processes, the pronunciation phase, the holding phase, and the end phase are classified, and the specific distinction is conducted based on the quantitative exponential function, as

$$\begin{cases} \Gamma(\tau|\mu,\sigma) = \alpha e^{-(\tau-\mu)^2/\sigma_1}, \\ \Gamma(\tau|\mu,\sigma) = \alpha, \quad \mu \le \tau \le T, \\ \Gamma(\tau|\mu,\sigma) = \alpha e^{-(\tau-\mu)^2/\sigma_2}, t \le \tau \le T. \end{cases} \tag{1}$$

The initial consonant pronunciation curve is described based on the parameterized formula, and its calculation formula is as follows:

$$\begin{aligned} K'' &= \left|\frac{p_i - p_{i-1}}{p_{i+1} - p_i}\right| \quad (i = 1, 2, \ldots, N-1), \\ \mu &= \arg\left(\max\left(K''\right)\right), \\ t &= \arg\left(\max\left(K''\right)\right), \\ E_s &= \frac{1}{N}\sum_{i=0}^{N-1}\sqrt{\left(\tau|(\mu,\sigma) - p_i\right)^2}. \end{aligned} \tag{2}$$

For the purpose of reusing the curves mentioned above in other models, the data acquired from the video will be standardized based on the following equation:

$$\begin{aligned} H_{std} &= \frac{H_{original}}{H_{natural}}, \\ W_{std} &= \frac{W_{original}}{W_{natural}}. \end{aligned} \tag{3}$$

In the standard mandarin, vowels can be divided into diphthongs and monophthongs. The monophthongs are set

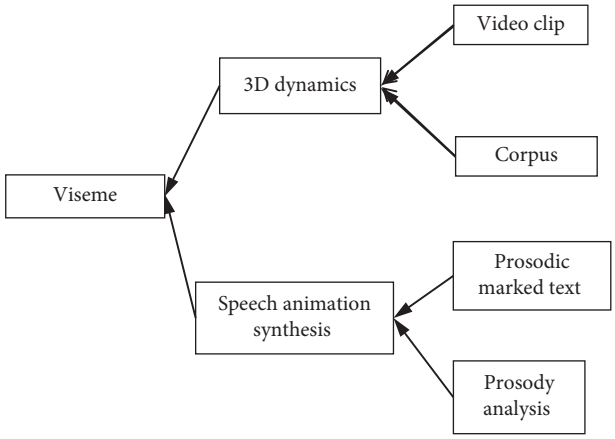

Figure 1: Frame diagram of the lip sync animation.

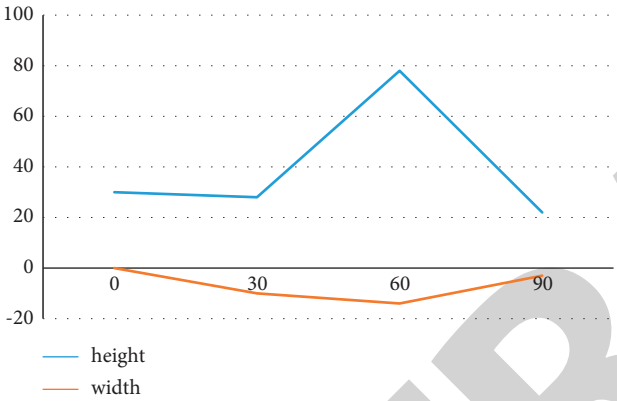

Figure 2: Pronunciation characteristic curve.

for the corresponding fusion, and the specific ω is determined by the maximum amplitude value of the [a] and [o] curves as

$$\omega_H = \frac{\left(H_{\max}^{|o|}\right)}{\left(H_{\max}^{|o|}\right)^2 + \left(H_{\max}^{\alpha}\right)^2}. \quad (4)$$

Therefore, the corresponding lip form constraint factor is introduced to resolve the issue of coordinated pronunciation as follows:

(1) Firstly, the vowel control curve of the simple or compound vowel is obtained in accordance with the order of the vowel first and then the consonant [16–18]

(2) Through the corresponding calculation, the relevant curves are synthesized, and the splicing of the curves is carried out by matching based on the pitch length separately so as to implement the pattern of the mixed splicing point and the form of the simple or compound consonant

(3) Through the calculation of the curve for the end and the beginning of the vowel based on the corresponding mixing point, the specific process is shown in Figure 3

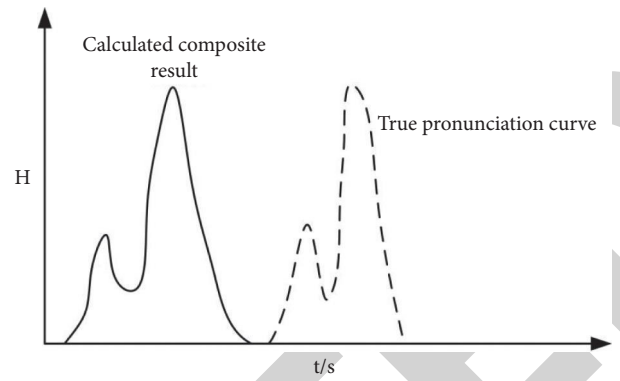

Figure 3: Synthetic result of the syllable flow.

3. 3D Lip Sync Animation Synthesis System for the Chinese Speech Synchronization

For the purpose of reclassifying the lip patterns of the pronunciation in the voice database, a comprehensive weighted algorithm is introduced in this paper to synthesize the speech synchronization in lip sync animation. With respect to the input text, the text-to-speech one-to-one mapping is first carried out. Secondly, it is transformed to the changes in the mouth patterns to implement the conversion of text to speech. The specific 3D lip sync animation synthesis framework is shown in Figure 4.

The specific steps are described as follows:

Step 1: input the Chinese texts

Step 2: convert the Chinese texts' input into standard voice

Step 3: synthesize the samples using the converted speech directly

Step 4: how to process the current voice elements through the audio operation

Step 5: calculate the current mouth pattern based on the curve trajectory of the voice response

Step 6: synthesize the voice and implement the visual display, and the process ends until there is no more voice element

For voice and graphics, the first step is to initialize the audio processing to ensure that the samples can be played continuously. The calculation of the viseme weight can be expressed by as follows:

$$D_{f-y} = \ln\left(1 - \left(1 - e^{-1}\right) \cdot R_f^{1-R_y}\right) + 1. \quad (5)$$

The relationship between the weight value and the influence on the simple or compound vowels and the relationship between the influence of successive vowels are shown in Figures 5 and 6, respectively. The weight values are negatively correlated with the influence; that is, the greater the influence is, the smaller the visual weight is.

The viseme weight affected by the vowel-vowel relationship is expressed by equations (6) and (7) as follows.

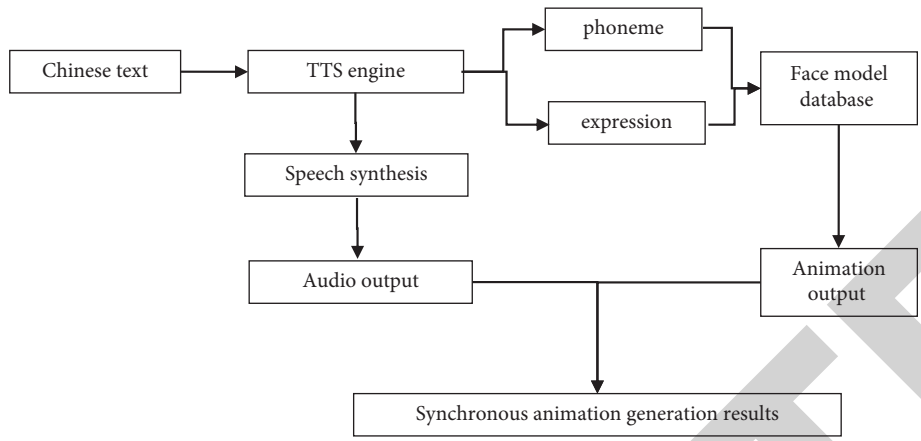

Figure 4: Flowchart of the lip sync system for the Chinese voice synchronization.

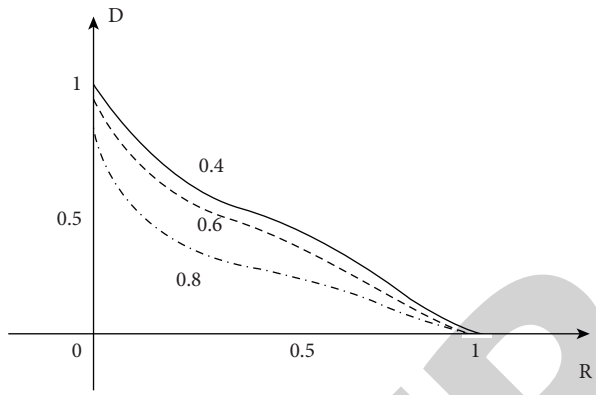

Figure 5: Relationship between the weight value and the influence on the consonants.

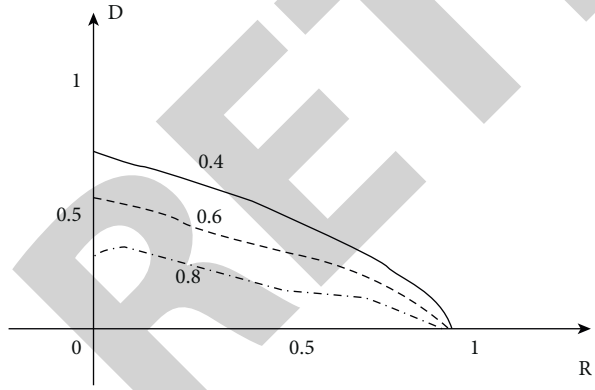

Figure 6: Relationship between the weight value and the influence on the successive vowels.

When w_s, the following can be obtained:

$$D_{y-y} = -\left(R_{y1} - R_{y2}\right) \times e^{-\left(R_{y1}-R_{y2}\right)} + 1. \quad (6)$$

When w_s, the following can be obtained:

$$D_{y-y} = 1. \quad (7)$$

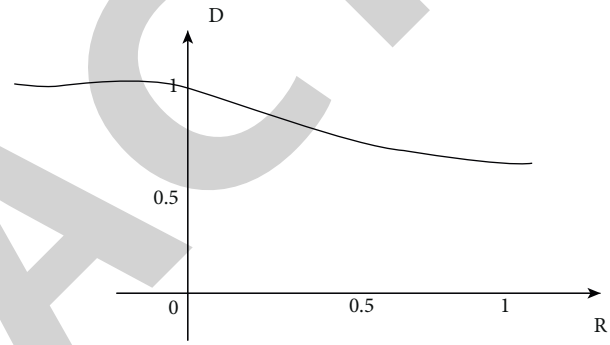

Figure 7: Relationship between the vowel-vowel visual weight value and the difference between the vowel phonons.

As shown in Figure 7, when the successive vowel factor > vowel influence, the influence of the vowel is limited. When the successive vowel factor ≤ vowel influence, the visual weight value will be negatively correlated with the difference between the above two, that is, the greater the difference between the influence level of the vowel phonon and the influence level of the successive vowel phonon, the smaller the influence weight value.

3.1. Chinese Speech Synchronization Algorithm Based on Comprehensive Weighted Algorithm

3.1.1. Pinyin Redefinition Scheme for Lip Sync Animation. The specific classifications of initial consonants and simple or compound vowels are shown in Tables 1 and 2.

For the purpose of facilitating the convenient implementation of the program, the mouth pattern marks of the initial consonant part and the simple or compound vowel part are simplified in this study. The preceding "s-" and "y-" are removed and written into one letter only. There are a total of 10 symbol letters after the simplification as follows: *a, o, e, i, b, d, f, r,* and *y*. The examples of the Pinyin conversion of some Chinese characters are shown in Table 3.

Table 1: Conversion table for the initial consonants of the standard Chinese Pinyin.

Initial consonants of the standard Chinese Pinyin	Definition of the initial part of the mouth pattern
b, p, m	*s-b*
F	*s-f*
d, t, n, l	*s-d*
zh, ch, sh, *r*	*s-r*
y, j, q, x, z, c, s	*s-y*
g, k, h	*s-g*

Table 2: Conversion table for the simple or compound vowels of the standard Chinese Pinyin.

Simple or compound vowels of the standard Chinese Pinyin	Definition of the simple or compound vowel part of the mouth pattern
ia, ai, ao, an, ua, iao, ian uai, uan, van, ang, iang, uang	y-a
o, u, ii, ue, ou, uo, ui, un, iu, iou, ong, iong, w	y-o
e, er, ei, en, uei, uen, ueng, eng	y-e
i, ie, in, ing	y-i

Table 3: Examples of the conversion of some Chinese characters.

Standard Chinese Pinyin	Definition of the initial part and the simple or compound vowel part of the mouth pattern	Mouth pattern of the Chinese Pinyin after simplification
Dong	s-d—y-o	do
Ren	s-r—y-e	re
A	& —y-a	&a

4. Synthesis of Lip Sync Animation

4.1. Comprehensive Weighted Algorithm. In this study, M sets of data are taken, with N samples in each set, and the average treatment is carried out on the samples [19]. The calculation formula is shown as follows:

$$\overline{t} = \frac{1}{N}\sum_{i=1}^{N} t_i. \tag{8}$$

The time variance D_t can be obtained from the average time data $\overline{t}$, which is calculated based on

$$D_t = \sum_{i=1}^{N} t_i (\Delta p_i)^2, \Delta p_i = t_i - \overline{t}. \tag{9}$$

Based on the consonant weight value, the consonant segmentation ratio obtained by the consonant segmentation algorithm based on the distance between segments and the ear speech consonant segmentation method based on the entropy function can be analyzed comprehensively, and the consonant time control ratio can be obtained accordingly. Let $w_s = 1 - w_y$, and the calculation is shown in equations:

$$t_s = w_s \overline{t}_p, \tag{10}$$

$$t_y = w_y \overline{t}_p. \tag{11}$$

4.2. Analysis of the Label Weight Vector. In this study, mainly 7 types of dots with long pauses in or at the end of a sentence are taken into consideration, such as period, exclamation mark, question mark, pause, comma, semicolon, and colon, and their pause time in the sentence is weighted, as shown in Figure 8.

Similarly, different weights are assigned to the labels in Chinese to generate more realistic, vivid, and continuous lip sync animation. The calculation formula is shown as follows:

$$t'_s = w_s \overline{t}_p w_{bi}; t'_y = w_y \overline{t}_p w_{bi}. \tag{12}$$

In the above equation, w_{bi}stands for the weight value of the ith label in the labels.

4.3. Treatment of the Mouth Pattern Transition. The position $X(s)$of each node with the inset type in the middle can be calculated based on the position of the sum of the initial viseme node $X^0$and the ending viseme nodeX^1, as follows:

$$X(s) = \left[uX_0^0 + sX_0^1, uX_1^0 + sX_1^1, \ldots, uX_n^0 + sX_n^1\right]. \tag{13}$$

The cosine function is used to improve this action:

$$s' = s * \frac{(1 - \cos(\pi * (s_0 - s)))}{2}. \tag{14}$$

With regard to the dynamic calculation of node displacement, it is based on the physical movement of the lips, that is, if the position, mass, and velocity of the initial consonant structure of the node $X_i(t)$are specified, it can be calculated as follows: $X_i = [m_i, V(t); i = 1, 2, \ldots, n]$. Once the geometric structure is determined, it can be calculated through the application of the Newtonian physics:

$$\frac{d_{X_i}}{d_t} = V_i, \tag{15}$$

$$m\frac{d_{V_i}}{d_t} = f_i - \gamma V_i. \tag{16}$$

The motion equation is a function of the system time t, and t stands for the driving time from the audio server, which can be used to calculate the new velocity and node position based on the following equations:

$$V_i = V_i^0 + \frac{f_i}{m_i}\Delta t, \tag{17}$$

$$X_i = X_i^0 + V_i \Delta t. \tag{18}$$

In the above equations, the velocity speed V_i^0 and the position X_i^0 of the previous step are used to calculate the new position of the node.

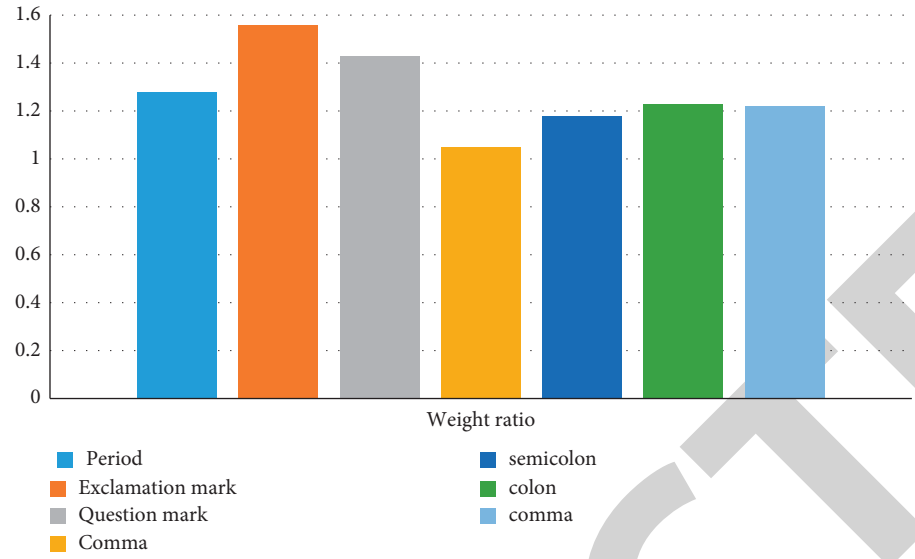

Figure 8: Label weight.

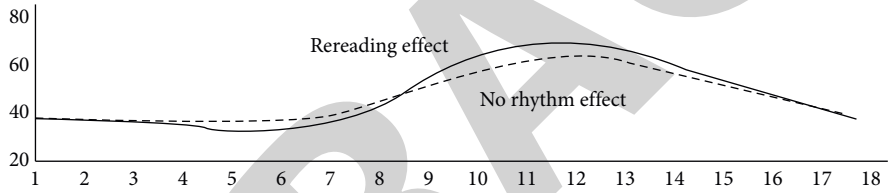

Figure 9: 3D lip sync animation.

4.4. Analysis of the Test Results. In the simulation experiment, the 3D dynamic visual position of the animation multimedia is acquired through the video drive to implement the conversion from 2D to 3D. Hence, from an individual perspective, the synthesized lip pattern is similar to the original video. The details are shown in Figure 9. The corresponding error was calculated by comparing the features to measure the similarity. The average composite error was only 0.02, and the maximum composite error was about 0.04. Therefore, it can be considered that the comprehensive weighting algorithm can ensure that the changing trends of human pronunciation and mouth shape are similar, and the comprehensive weighting algorithm can ensure that the three-dimensional mouth shape synchronization is more efficient, more realistic, and more vivid.

5. Conclusions

Animation multimedia 3D lip synchronization has highly important applications in AI hosting, video conferencing, and so on. However, how to ensure its effectiveness and efficiency is the focus and difficult points in the research. Based on the comprehensive weighted algorithm, the Chinese voice multimedia database is sorted out to implement the continuous changes in the mouth patterns through the voice lip sync animation matching, in conjunction with the expression animation to ensure that the 3D lip synchronization is more realistic and vivid. The simulation experiment has indicated that the comprehensive weighted algorithm put forward in this paper is effective.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] L. Wei and Z. Deng, "A practical model for live speech-driven lip-sync," *IEEE Computer Graphics and Applications*, vol. 35, no. 2, pp. 70–78, 2015.

[2] G. Bunting, "Lip-sync editing and mixing of nonperforated magnetic tape using the new synchrolock tape system," *Smpte Journal*, vol. 86, no. 7, pp. 482–486, 2015.

[3] L. Peter, "Lip-sync patent eligibility won't get full federal circuit hearing," *Trademark & Copyright Journal*, vol. 93, no. 2292, pp. 2701-2702, 2017.

[4] S. Highfill, "Exclusive promo for lip sync battle's holiday special puts joseph gordon-levitt against anthony mackie," *Entertainment Weekly Com*, vol. 4, no. 2, pp. 1–8, 2015.

[5] M. E. Watch, "Giannis antetokounmpo lip-syncs to justin bieber," *Sports Illustrated Com*, vol. 2, no. 4, pp. 76–79, 2015.

[6] M. Gonzalez-Franco, A. Steed, and S. Hoogendyk, "Using facial animation to increase the enfacement illusion and avatar self-identification," *IEEE Transactions on Visualization and Computer Graphics*, vol. 4, no. 1, pp. 58–64, 2020.

[7] R. Hoffner, "Audio-video synchronization across DTV transport interfaces: the impossible dream?" *Smpte Journal*, vol. 109, no. 11, pp. 881–884, 2015.

[8] S. Schreitmüller, M. Frenken, and L. Bentz, "Validating a method to assess lipreading, audiovisual gain, and integration during speech reception with cochlear-implanted and normal-hearing subjects using a talking head," *Ear and Hearing*, vol. 39, no. 3, pp. 1–9, 2017.

[9] S. E. Eskimez, R. K. Maddox, and C. Xu, "Noise-resilient training method for face landmark generation from speech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 3, no. 99, pp. 1–10, 2019.

[10] I. Varela-Benavides and R. Peña-Santiago, "Metaxonchium toroense n. sp. (Nematoda, Dorylaimida, Belondiridae) from Costa Rica, with the first molecular study of a representative of the genus," *Journal of Helminthology*, vol. 4, no. 3, pp. 1–9, 2017.

[11] C. J. Daly, J. M. Bulloch, M. Ma, and D. Aidulis, "A comparison of animated versus static images in an instructional multimedia presentation," *Advances in Physiology Education*, vol. 40, no. 2, pp. 201–205, 2016.

[12] S. Doukianou, D. Daylamani-Zad, and K. O'Loingsigh, "Implementing an augmented reality and animated infographics application for presentations: effect on audience engagement and efficacy of communication," *Multimedia Tools and Applications*, vol. 4, no. 5, pp. 1–23, 2021.

[13] N. S. Narayanan and R. J. Dileone, "Lip sync: gamma rhythms orchestrate top-down control of feeding circuits," *Cell Metabolism*, vol. 25, no. 3, pp. 497-498, 2017.

[14] C. L. Fiorella, "Effects of observing the instructor draw diagrams on learning from multimedia lessons," *Dissertations & Theses-Gradworks*, vol. 5, no. 8, pp. 109–116, 2015.

[15] S. Szeszak, R. Man, and A. Love, "Animated educational video to prepare children for MRI without sedation: evaluation of the appeal and value," *Pediatric Radiology*, vol. 46, no. 12, pp. 1–7, 2016.

[16] C. Baggott, J. Baird, and P. Hinds, "Evaluation of Sisom: a computer-based animated tool to elicit symptoms and psychosocial concerns from children with cancer," *European Journal of Oncology Nursing*, vol. 9, no. 8, pp. 16–20, 2015.

[17] M. Heo and N. Toomey, "Learning with multimedia: the effects of gender, type of multimedia learning resources, and spatial ability," *Computers & Education*, vol. 146, no. 3, pp. 103–112, 2019.

[18] A. Zbigniew and B. Marcin, "Interactive multimedia learning environment for geometrical specification indication & verification rules-ScienceDirect," *Procedia CIRP*, vol. 75, no. 5, pp. 161–166, 2018.

[19] P. Szczuko, "Simple gait parameterization and 3D animation for anonymous visual monitoring based on augmented reality," *Multimedia Tools and Applications*, vol. 75, no. 17, pp. 1–21, 2016.