Synthesis of 3D pronunciation trajectories of lips in Mandarin Based on HMM model

Yuanzhi Ye, Xiangzhen He*, Yihao Zhang, Yerong Hu, Jia Zeng

Key Laboratory of China's Ethnic Languages and Information Technology of Ministry of Education, Northwest Minzu University, Lanzhou, Gansu 730000, China

*Corresponding author: Xiangzhen He
*Corresponding author’s e-mail: hexz@xbmu.edu.cn

Abstract: In the era of information technology, human-computer interaction tendings to be multi-modal, and lip animation simulation driven by pronunciation actions has attracted more and more attention. Starting from Mandarin Chinese, this paper establishes a model of the trajectory of the lips during the pronunciation process. The model is based on a motion capture technology, proposes a database collection and processing method, extracts acoustic parameters and text information to establish an HMM prediction model of pronunciation action parameters, and synthesizes the pronunciation movement trajectory, and the average percentage error with the real trajectory is less than 3.42%. The results show that the predicted pronunciation action parameters can effectively synthesize the lip pronunciation movement trajectory, and the use of lip animation can enhance language understanding.

1. Introduction

In terms of speech perception, people mainly use two ways to solve it, for one thing it is hearing, for another it is vision\(^1\), which means that speech is visible\(^2\). In recent years, the human-computer interaction three-dimensional face and mouth animation simulation has been the focus of computer graphics research\(^4\), and its driving basis is the synthesis of the lip pronunciation movement trajectory. Pronunciation action parameters can describe speech features effectively, and can more intuitively show the movement of lips or other organs. Not only can it be used for speech visualization, but also can reduce the recognition error rate in the research of speech recognition and speech synthesis methods.

People use pronunciation action parameters to describe the position and movement of the vocal organs such as tongue, jaw, and lips during the process of pronunciation. The parameters can be collected through a variety of technologies, such as X-ray microbeam imaging\(^5\), magnetic resonance imaging\(^6\), electromagnetic Pronunciation instrument EMA\(^7\) and three-dimensional motion capture technology\(^8\)\(^-\)\(^9\). The methods of predicting pronunciation action parameters when given text or voice input include CM cooperative pronunciation trajectory synthesis\(^10\)\(^,\)\(^11\), prediction of pronunciation action parameters using target approximation model\(^12\), prediction of pronunciation action parameters based on HMM\(^7\)\(^,\)\(^13\)\(^-\)\(^15\), Gaussian mixture model and artificial neural network model. Due to the lack of a Chinese pronunciation action parameter database, there are few studies on Chinese pronunciation action parameters, and the acoustic parameter changes in speech are very similar to HMM, which is a double random process, so HMM is very suitable for acoustic parameter modeling\(^16\)\(^-\)\(^18\).

This paper is based on 3D motion capture technology, starting from 3D motion data collection and processing, extracting acoustic features and voice annotation, using Mandarin Chinese initials, vowels
and tones as basic units for research, establishing an HMM prediction model of pronunciation action parameters, and synthesizing Three-dimensional pronunciation movement trajectory of the lips in Mandarin Chinese.

2. Pronunciation motion capture collection system

For the collection of three-dimensional pronunciation motion parameters, this article uses the Vicon Cara expression motion capture system designed by Vicon, with the help of its helmet device and four dedicated high-definition cameras to track calculations and data with Caralive (data acquisition) and carapost (data preprocessing) The recording software collects the pronunciation movement data of the discrete points of the speaker's lips.

The collection process is as follows: According to the position points of the FDP face definition parameters of the MPEG-4 standard, 36 2mm mark points are respectively pasted on the lips and lower jaws of the speaker. At a sampling rate of 60 frames per second, the spatial positions of the facial marker points are tracked and solved from the four channels simultaneously. At the same time, a camera is used to record audio and video data simultaneously. The front view of the collection site and the marked points are shown in the figure below (Figure 1).

![Figure 1 Front view and mark point map of the collection site](image)

3. Motion capture data processing method

In the experimental data, 8 points on the lips and 1 point on the chin are selected as preselected parameters of the vocal motion trajectory. In order to reduce or eliminate the influence of habitual pronunciation actions on the three-dimensional position of the lip landmark points, a stable nose tip marking point is selected as a reference Click to eliminate the rigid body movement of the head.

3.1. Data optimization and extraction

After the data collection is completed, the 4 channels of the two-dimensional image data captured by Logger in .Pico format are downloaded to the local computer, and then opened in the carapost software, and the data is subjected to three-dimensional construction and subsequent data optimization processing. First drag the timeline to view the recording effect of the data, then locate the frame with the best data effect, track the marker points of each channel, and then import the recorded calibration file, and compare the two-dimensional marker points of each channel with the half moon The mark points of the board are fitted, seed points are established, and the three-dimensional space position of the mark points is automatically calculated by the system to construct a three-dimensional environment. Merge all the mark points of the face into the three-dimensional environment, face topology, and initialize the 3D face model. Finally, through the backward tracking algorithm, the data of each subsequent frame is tracked as a whole, and the corresponding modification, completion, and deletion operations are performed according to the color information and continuity of the tracking bar. Finally, select three stable points of the face to achieve data alignment, and finally output a .BVH file.

The .BVH file records the running information of the human skeleton in consecutive frames with the node as the center. This experiment calculates the bit space position of each coordinate point, so it is necessary to convert the bone information to its corresponding three-dimensional coordinate point.
The .bvh file mainly includes two parts of information: skeleton information, data block information. The skeleton information defines the position and rotation components of the parent joint and the child joint according to the hierarchical relationship, thereby forming an overall skeleton. The data block corresponds to the offset of 3 or 6 degrees of freedom of each joint from the initial coordinate system (and global coordinate system), which is the three-dimensional data coordinate. Write a reading program in Python, read the total number of frames of motion in the motion data file, create a motion or state array to save the motion data, and use recursive operation. If a coordinate system P is rotated and a coordinate system q is obtained, this rotation process The rotation matrix of can be expressed as $R_{pq}$, and there is a certain point A in this space. Its coordinates in the two coordinate systems are $A_p$ and $A_q$, so there are:

$$A_q = R_{pq}A_p$$

Finally, analyze the three-dimensional coordinates of these 10 points, keeping two decimal places.

### 3.2 Audio annotation

The video data collected by the camera is segmented and formatted using the editing software Adobe primere. During segmentation, the lips of the speaker are guaranteed to return to their original natural state from a static state to the end of the pronunciation. After the segmentation is completed, the segmented video file is converted into a .WAV audio file. Then the primitives are intercepted in the Praat software. Figure 2 shows the result of voice annotation for the paragraph "It is difficult to give up". This experiment uses the finals as the unit for analysis. Therefore, the audio file is divided into initials and finals to reflect the unitization of the label information. The process of segmenting finals is not only cumbersome and time-consuming, but also Attention should be paid to retaining the transitional segment of the synonym when intercepting, to ensure the co-pronunciation phenomenon of Jiyuan. Praaa loads the .WAV file to manually generate its corresponding TextGrid format file. The labeling information is mainly divided into two layers. The first layer is the information layer, which mainly labels all audio information including voice and noise. The second layer is mainly labeled with vowels as the unit. Syllables and pronunciation tones are represented by Arabic numerals 1, 2, 3, and 4. Finally, save the annotation result of the TextGrid file, which contains the start time information corresponding to each vowel.

![Figure 2](image.png)

**Figure 2** Schematic diagram of voice annotation

### 4. HMM pronunciation motion trajectory model

Taking contextual text and acoustic features (LSP) as input, and lip marking point coordinate data as output, modeling training is performed based on the hidden Markov model to generate the corresponding viseme sequence, and the maximum likelihood algorithm generates the optimal smooth motion trajectory. Predict the three-dimensional motion trajectory characteristics of the lips with the pronunciation, and the results can be used to drive the control points on the lips to achieve the simulation of the pronunciation action.
4.1 Model construction

4.1.1 Model parameter definition
Hidden Markov model is a double random process. Its parameters are determined by the observation sequence \((X_1,...,X_M)\), namely, the pronunciation action parameter and the state sequence \((Z_1,...,Z_N)\) (\(M\) and \(N\) respectively represent the number of observed states and the number of hidden states), often consider the first-order Markov chain in applications (Figure 3). There are two important basic assumptions: ① Homogeneous Markov chain hypothesis, the hidden state at any time only depends on its previous hidden state; ② The observation independence hypothesis, that is, the observed value at any time only depends on the hidden state at the current time.

![Figure 3 Experimental Hidden Markov Chain](image)

There are three important probability matrices in the middle of the model process, namely the matrix parameters \(\Theta(A, B, \pi)\) of the model, the hidden state transition probability matrix \(A\), the observed state transition probability matrix \(B\), and the initial state probability matrix \(\pi\). The algorithm used in its realization process includes forward and backward algorithm, dimension bit algorithm, Baum-Welch algorithm based on EM algorithm.

4.1.2 Model data preparation
(1) Context attribute triphone format
Since co-articulation has a great influence on the pronunciation of lips in continuous speech flow, monophonic vowels cannot be used directly as the modeling unit. The pronunciation action parameter prediction system in this paper uses the consonants and vowels as the modeling unit, and the context attribute set is used in the HMM training process. The context attributes of the triphone model include the current phoneme and one phoneme before and after it, that is, the three phoneme text is formed by adding the units before and vowel units to form a three-phone text. Figure 4 shows the triphone text of "Rain Tonight".

![Figure 4 The triphone text of "Rain Tonight"](image)

After the single-phoneme model is expanded into the three-phoneme model, the number of models increases sharply, but the training data is very limited. Therefore, similar models should be clustered through decision tree clustering to make them share the training data and avoid the overfitting problem caused by data sparseness.

(2) Set up problem sets and clustering
Set up the question set of each model unit, that is, ask what the attributes of the current or before and after the initials and vowels are. The coverage of the question set should have a certain breadth and include all the combined features as much as possible. Then according to the HTS settings, the question set is converted into the corresponding question set format to determine the current and the type of phoneme combination before and after. Decision tree clustering analysis is performed from the problem...
set. The clustering uses the HHEd module of the HTK tool. The system will traverse all the set problem sets, cluster the models according to the classification of the problems, and the clustered models will share the training data. After the decision tree clustering, the model needs to be re-evaluated and trained again, and the model parameters are optimized with the shared training data. In this step, the HERest tool is still used. After this step of training, a decision tree file and a trained context-sensitive HMM model file are obtained for subsequent prediction of pronunciation action parameters.

4.1.3 Model establishment
The trajectory hidden Markov model (HMM) can understand the motion trajectory distribution of each speech unit. It is the latest time series random model and can model the dynamic changes of the signal. Using the parameter generation of the trajectory HMM, a smooth output can be synthesized by considering the first and second derivatives of the data, and its parameter generation algorithm can generate a smooth trajectory from a random model.

Use context-sensitive HMM to simultaneously model pronunciation actions, speech and text (Figure 5). The pronunciation action parameters are modeled by using 8 mark points of the lip movement and 1 mark point of the chin. In order to better smooth the output, the first and second derivatives of the movement characteristics of the mark points are used to construct the dynamics of the movement trajectory. mold. For speech, the acoustic parameter LSP is extracted, and the text input is in a processed context triphone format. HMM uses a mixture of Gaussian distributions in each state.

4.2 Model training methods and predictions
Based on phoneme system analysis, the topological structure of the model is a left-right model with no jump, and the specific parameters of the model (mainly including mean, variance, and transition matrix parameters, which will be continuously updated through the subsequent training process).

(1) Modeling unit and topological structure: Chinese vowels are selected as the modeling unit, and the topological structure is from left to right.
(2) Number of states: Generally, 10 states of HMM are used to describe a syllable, and 5 states of HMM are used to describe consonants and vowels.
(3) Parameter distribution: choose multi-dimensional Gaussian distribution.
(4) Feature vector length: each feature vector length is 27, that is, 9 pronunciation parts, each part has xyz three-dimensional data, a total of 27 dimensions.

In the training process, a set of context-sensitive HMM models are obtained by maximizing the joint distributed likelihood function \( P(X, Y, Z | \lambda) \) of the acoustic and pronunciation action parameters. Then carry out context-sensitive HMM training initialization. Train a decision tree with the minimum description length (MDL) criterion and the context attribute problem set, and use the decision tree to cluster HMM.

Through the above training process, the final trained model includes the clustering HMM of spectrum, fundamental frequency, and pronunciation action parameters and their respective decision trees. In the
prediction process, first use the results of front-end text analysis and the decision tree to determine the HMM sequence, and then use the MLPG algorithm to generate the optimal pronunciation action parameters.

\[
X^*_i = \arg \max P \left( \frac{X}{\lambda} \right) = \arg \max P \left( \frac{W_X X_S}{\lambda} \right) = \arg \max \sum y_q P(W_X X_S, q|\lambda)
\]

where \( X = W_X X_S, X_S = [x_{s1}^T, \Delta x_{s2}^T,\ldots,x_{SN}^T]^T \) is the static parameter sequence of the pronunciation action, \( W_X \in R^{30D \times ND_X} \) is an extended matrix of pronunciation actions, \( q = \{q_1, q_2, q_3,\ldots,q_N\} \) represents the state sequence, and finds the best sequence through second-order difference and approximation algorithm optimization.

The pronunciation action parameters and the state sequence are alternately updated using an iterative update method, which synthesizes the optimal actual pronunciation movement trajectory from the optimal HMM sequence, and aligns the predicted pronunciation action trajectory (three-dimensional continuous function of X, Y, Z) with the input speech in time.

### 4.3 Trajectory Synthesis

There are two types of co-articulation effects. Among them, the former phoneme is affected by the pronunciation situation of the latter phoneme. It is called Anticipatory Coarticulation. The latter phoneme is affected by the pronunciation situation of the previous phoneme. It is called Carry Over. Coarticulation, scholars generally believe that the adverse effect is greater than the forward effect.

Since the training data text is in the triphone format, the output prediction parameters also correspond to the triphone format. Any phoneme unit in the middle of a sentence will be repeated three times, such as the "night" initial y phoneme in "tonight (ye)" j-in-y, in-ye, and y-e-y exist (Figure 4), namely forward, forward and backward, and reverse modes. Simply align and synthesize according to time, there will be overlapping parts, which is not in line with reality. The importance coefficient matrix is established by analytic hierarchy process to calculate the weights of the three, which are 0.27, 0.41, and 0.31. The predicted value of each frame in the monophone is multiplied by the corresponding weight to accumulate the actual displacement of the frame \( \tilde{X}_w \):

\[
\tilde{X}_w(t) = \sum_{i=1}^{3} \alpha_i X_i(t)
\]

Connecting all frames of the phoneme is a monophone prediction trajectory, and connecting all the monophone trajectories in the prediction sentence is a smooth total pronunciation trajectory.

### 4.4 Experimental Results and Analysis

The experiment uses a three-dimensional motion capture database of Chinese female speakers' continuous speech stream, which contains both speech waveforms and pronunciation action parameters and text. In this paper, 40-order line spectral pair (LSP) is used as the spectral acoustic parameters, and text annotation is used as the basis of clustering analysis. The 27-dimensional features of 9 mark points on lips and chin (9 reflective balls, each containing X, Y, Z three-dimensional) as the pronunciation action parameter. The experiment is designed to collect 500 sentences of Mandarin standard recording, 490 sentences are used as training data, and the remaining 10 sentences are used as test data.

The experiment synthesized the three-dimensional motion trajectory of the lips during the pronunciation process, and used the RMS error and relative percentage error of the synthesized pronunciation trajectory coordinate points to calibrate the reliability of the prediction result. Table 1 shows the RMS error (mm) of the X, Y, and Z coordinates of all the frames of the 9 marked points in the segment "It's raining tonight". The average value of the lip 3D pronunciation motion trajectory is 0.668mm, which is relatively small.
Table 1 Experimental RMS error data table

|   | X   | Y   | Z   |
|---|-----|-----|-----|
| L1| 0.714 | 0.577 | 0.525 |
| L2| 0.662 | 0.580 | 0.772 |
| L3| 0.615 | 0.669 | 0.767 |
| L4| 0.712 | 0.551 | 0.761 |
| L5| 0.712 | 0.648 | 0.505 |
| L6| 0.532 | 0.695 | 0.757 |
| L7| 0.580 | 0.781 | 0.629 |
| L8| 0.693 | 0.650 | 0.729 |
| J | 0.626 | 0.690 | 0.619 |
| average | 0.668 |

Figure 6 shows the Z coordinate composite curve of the center mark points of the upper and lower lips of the "toy" syllable. The relative percentage errors are 2.89% and 3.42%, respectively. The results show that the parameters synthesized by the HMM prediction model are in good agreement with the measured mouth shape parameters and have good results.

Figure 6 The true and synthetic curves of the Z coordinates of the marked points in the "toy" syllable

5. Conclusion
This paper introduces 3D motion capture software and motion capture data collection and processing methods. Based on the established Chinese Mandarin database, the HMM-based pronunciation action parameter prediction experiment is carried out. Acoustic parameters and text annotation segmentation are extracted as input, and pronunciation action parameters are used as output. The context triphone model is used, combined with cluster analysis, to obtain the HMM continuous regression training model, and the maximum likelihood algorithm is used to generate the optimal pronunciation action parameters. Experiments show that the three-phone context HMM prediction model can effectively synthesize the lip 3D pronunciation motion trajectory. In the later stage, it can be combined with a three-dimensional face model to generate a mouth animation to realize a talking 3D human head, that is, voice visualization, and improve the speech recognition rate when speaking.

Acknowledgments
This work was jointly supported by Science and Technology program of Gansu Province (20JR5RA507), the research projects of the social science and humanity on Young Fund of the ministry of Education (NO.17YJCZHO57), the Fundamental Research Funds for the Central Universities (NO.31920200116), and the Fundamental Research Funds for the Central Universities (NO. Yxm2019123).

References
[1] Wang, A, H.. A Preliminary Study on the Viseme System of Mandarin Chinese[D]. Beijing Language and Culture University, Beijing Language and Culture University, 2000.
[2] Wu C, J.. Research on Visualized Cooperative Pronunciation Synthesis of Uyghur Language[D]. Xinjiang University, 2014.
[3] Li, R.. 3D visualization of pronunciation[D]. University of Science and Technology of China, 2016.

[4] Wang, Y.. Research on text-driven 3D mouth animation synthesis based on Chinese collaborative pronunciation model[D]. Shandong University of Finance and Economics, 2014.

[5] Kiritani S.. X-ray microbeam method for the measurement of articulatory dynamics: Technique and results [J]. Speech Communication, 1986, 45: 119-140.

[6] Bare T., Gore J C., Boyce S.. Application of MRI to the analysis of speech production [J]. Magnetic Resonance Imaging, 1987, 5: 1-7.

[7] Cai M, Q., Ling Z, D., Li, R.. Prediction method of Chinese pronunciation action parameters based on hidden Markov model[J]. Data Acquisition and Processing, 2014, 29(02): 204-210.

[8] Zhu K.. Research on 3D facial expression synthesis based on motion capture[D]. Northwest University for Nationalities, 2019.

[9] Li S.. Chinese Mandarin Phonetics Learning Software Based on 3D Motion Capture[D]. Northwest University for Nationalities, 2019.

[10] Zheng H, N., Zhu, Y., Wang, L., Chen, H.. Synthesis and dynamic simulation of Chinese three-dimensional pronunciation action[J]. Integration Technology, 2013, 2(01): 23-28.

[11] Cohen M., Massaro, D, W.. Modeling coarticulation in synthetic visual speech [J]. Models Technique in Computer Animation, 1993: 1 39-156.

[12] Birkholz, P., Kroger, B., Neuschaefer-Rube C. Model based reproduction of articulatory trajectories for consonant-vowel sequences[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2011, 10(5): 1422-1433.

[13] Yin, Q.. Research and implementation of Chinese speech synthesis system based on pronunciation action characteristics[D]. Chongqing University of Posts and Telecommunications, 2019.

[14] Ling, Z. H., Richmond, K., & Yamagishi,. An Analysis of HMM-based prediction of articulatory movements. [J]Speech Communication, 2010, 52(10), 834-846.

[15] Hofer G, Yamagishi, J, & Shimodaira,H.. Speech-driven Lip Motion Generation with a Trajectory HMM. [C]Centre for Speech Technology Research, University of Edinburgh, UK, 2008.

[16] Latorre, J., Iwano, K., & Furui, S.. New approach to the polyglot speech generation by means of an HMM-based speaker adaptable synthesizer. [J] Speech Communication, 2006, 48(10), 1227-1242.

[17] Wu, Y., J.. Research on Speech Synthesis Technology Based on Hidden Markov Model[D]. University of Science and Technology of China, 2006.

[18] Biagetti, G., Crippa, P., Falaschetti, L., Orcioni, S., & Turchetti, C.. Learning HMM State Sequences from Phonemes for Speech Synthesis. Procedia Computer Science, 2016, Volume 96, 1589-1596.