A Dual-Mode Real-Time Lip-Sync System for a Bionic Dinosaur Robot

Shuaizheng Yan
[1]*The State Key Laboratory of Management and Control for Complex Systems Institute of Automation, Chinese Academy of Sciences*
[2]*University of Chinese Academy of Sciences*
Beijing, China
yanshuaizheng2018@ia.ac.cn

Jiasheng Hao
School of Automation Engineering University of Electronic Science and Technology of China
Chengdu, China
hao@uestc.edu.cn

Zhengxing Wu
[1]*The State Key Laboratory of Management and Control for Complex Systems Institute of Automation, Chinese Academy of Sciences*
[2]*University of Chinese Academy of Sciences*
Beijing, China
zhengxing.wu@ia.ac.cn

***Abstract*—This paper provides a dual-mode real-time lip-sync system for a bionic dinosaur robot. Different from traditional mono-modality control systems, our system is constructed with different controllers and classifiers in both time domain and frequency domain. Specially, a classifier in time domain is designed to extract the sound features including pitch and intensity. Meanwhile, a nonlinear mapping relationship between time-domain feature parameters and mouth open angles is particularly designed. In time domain, an efficient algorithm consisting of original and modified short-term average amplitude difference function (AMDF) is applied for frequency measurement. With the goal of predicting the curve of mouth open, we train the audio data of dinosaurs to get a frequency classifier by Support Vector Machine with racial basis function (RBF-SVM), which has a relatively high accuracy. Finally, extensive experiments validate the effectiveness of this proposed system on a real bionic dinosaur robot.**

***Index Terms*—lip-sync, classifier, SVM, bionic dinosaur robot**

I. Introduction

With the in-depth study of artificial intelligence, the research of robots is deepened gradually and towards more intelligent development. According to the statistical data, the demand for robots that can interact deeply with humans, such as communication and even emotional communication, has greatly grown. Compared with industrial robots with poor mobility and poor human-computer interaction, service robots that are closer to human life, education, medical care, and companionship have greater market prospects in the future [1]. Generally, in the process of face to face communication, people usually focus their attention on the face, especially in the eyes and mouth. Therefore, to produce a convincing and versatile mouth-shaped change can allow people to have a near-real conversation with the machine, which can not only be a promising application in games and animations, but also become one of the key technologies for the natural interaction between humans and bionics.

Recently, many effective methods have been proposed to match lip (mouth) and human pronunciation. For instance, David Hanson produced a human-like robot named Jules with remarkably vivid emotion and lip-sync system in 2008 [2]; Wu *et al.* updated a head robot to pursue ability of lip-sync while talking with people [3]; AIST presented a new girl robot HRP-4C which can perform at least 6 kinds of expressions and give a speech almost like a real Japanese girl [4]; During 2013 to 2015, USTC continued to provide new effective algorithms on robot face control and achieved many graceful results on human-like robot named Kejia [5]; Engineered Arts developed a robot named Robo Thespian, which did well in talking with people in over 30 languages, and even performed some difficult songs with its mouth perfectly matched with the music [6]. In the field of algorithm, Tan *et al.* completed three-dimensional (3D) face and mouth modeling and data-driven text-to-visual speech conversion system [7]; Taylor *et al.* proposed a method for automatic redubbing of video that exploits the many-to-many mapping of phoneme sequences to lip movements modelled as dynamic visemes [8]; Fan and Yang provided a method that predefines a set of basic mouth movements, and then allows the designer to design mouth animation corresponding to different phonemes by defining the weight change curve of the elements in the set [9]; Xia *et al.* presents a text-speech-driven face animation generation method for redirecting two-dimensional (2D) face feature point vectors to a 3D head model [10]. These existing researches always focus on the lip-sync of human-like robots, but few take simpler mouth models of bionic animal robots into account. For the human face and lip synchronization system, the oral visual state can be divided into at least 6 basic expression states, but the classification result of animals is far less than that of humans. Therefore, the complex classification will greatly reduce the efficiency of real-time matching between lip and speech. Moreover, the animal's audio signal is entirely different from the vocal signal (whether in English or other languages), which is likely to provide many fault classification on states of animal audio signal.

Considering these problems mentioned, this paper provides a dual-mode real-time lip-sync system for a bionic dinosaur robot. Different from traditional mono-modality control systems, the proposed lip-sync system designs distinct controllers and classifiers in both time domain and frequency domain.

978-1-7281-1312-8/18/$31.00 ©2018 IEEE

Specially, in time domain, a classifier is modeled for extraction of sound intensity and pitch. An improved frequency extraction method named multiple short-term average amplitude difference function is presented to pursue a better and preciser measurement of frequency. Meanwhile, in control of intensity, the intensity curve is interpolated and the nonlinear mapping is applied to make the curve of output more stable and smooth. In frequency domain, in order to obtain a fast system response with a high accuracy, some dinosaur sounds are employed to train the classifier in frequency domain in advance, which helps this system produce an extremely vivid and precise output of mouth open angle combining with the time-domain control. Experiments are carried out to verify the effectiveness of the proposed lip-sync system. Besides, our system is not only working for a dinosaur robot, but also appropriate for nearly most kinds of animals barking with their mouths or beaks through training the different audio signals for a new classifier.

The remainder of the paper is organized as follows. The design of integrated lip-sync system is overviewed in Section II. Section III introduces the pre-processing, the Multiple AMDF and Support Vector Machine with radial basis function (RBF-SVM) classifier in detail. Experimental results are described in Section IV. Finally, Section V concludes the paper and describes an outline of future work.

II. Lip-Sync System Design

The general sound is made up of a series of complex vibrations with different frequencies and amplitudes. One of the lowest frequencies of these vibrations is called the pitch and the rest is overtone. Classified from subjective feelings, the sound consists of four basic elements: pitch, length, intensity, and tone [11]. As for the lip-sync system developed in this work, three elements of a dinosaur sound signal including intensity, pitch and tone are mainly analyzed.

For a lip-sync system, the first important work is feature extraction. For any audio signal input in time domain, it should judge the voiced and unvoiced sounds in a long audio via setting some key threshold values, e.g., the short-time zero-crossing rate and short-time energy. In phonetics, voiced sound means the sound of the vocal cord vibration at the time of pronunciation and unvoiced sound is the one that the vocal cord does not vibrate. Generally, the consonants could be unvoiced or voiced, while the vowels in most languages are voiced, and the nasal, side as well as semi-vowels are voiced. After sound judgement, some effective segments are divided to extract parameters of time-domain features. Fig. 1 illustrates the developed lip-sync system. Specially, in the process of sound feature extraction, we directly extract the relevant features such as Loudness (intensity), Frequency (pitch), and Mel Frequency Cepstral Coefficents (MFCC) (tone) by invoking a targeted algorithm. Then, the first two parameters, e.g., Loudness and Frequency, are input into the model controlled by pitch and intensity. Meanwhile, the MFCC is used as the input of the SVM trained classifier. Finally, the outputs of the two models are combined to obtain an array of the dinosaur mouth opening angle.

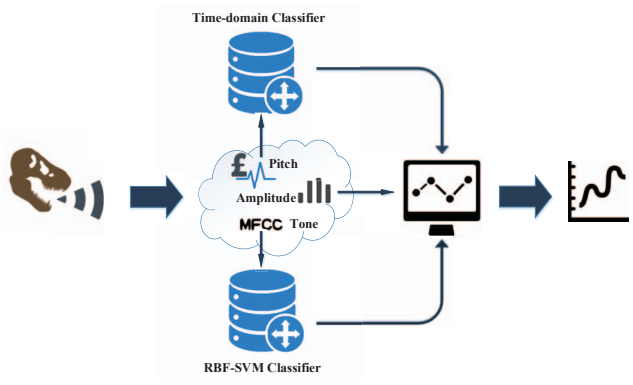


Fig. 1. Illustration of the proposed lip-sync system

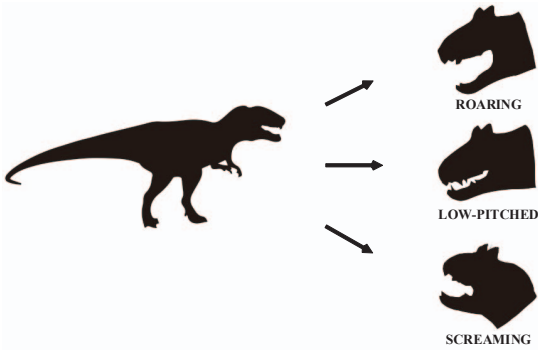


Fig. 2. Three shapes of dinosaur's mouth

In fact, the model in time domain is working as either a classifier or a prime controller. As a classifier, we know that it divides the input audio signal into unvoiced and voiced sounds. For animal sounds, most are voiced sounds, and only the silent segments between two separated audio signals and the low-noise portions are unvoiced. The dinosaur sounds in this work have the similar feature. Enlightened by the previous on dinosaur classification and modern animal vocal researches, we divide the dinosaur's voices into three categories in the time domain: low-pitched, roaring, and screaming. Fig. 2 presents the shapes of dinosaur's mouth with three sounds. As a prime controller, it employs intensity as a key factor to affect the output of dinosaurs mouth opening angles. Some interpolation algorithms are involved to optimize final output.

With regard to the frequency domain classifier, it is originally decided to improve the accuracy of the predicted mouth-engagement angles in this system. The output gained only by time-domain controller has a relatively enormous deviation compared with actual values. Therefore, a trained classifier can play an important role in this system. Taking into account only two categories of outputs we can get and the efficiency of the whole system, this frequency-domain classifier is trained in a SVM method. Besides, the RBF is also used as the kernel function of this classifier.

For the convenience of the experiment, the dinosaur sound

used in this system has a special file type, e.g., a mono wave file with a sampling frequency of 44.1 kHz. Meanwhile, in order to test the accuracy and real-time of this method, all experiments included in this system are on a dinosaur robot.

III. Software Implementation

In time domain, audio signals are pre-processed with several methods, such as central clipping, windowing operations and framing operations. The windowed and framed audio information is separately subjected to a short-time zero-crossing rate and a short-time average amplitude difference function (AMDF). The short-time zero-crossing rate curve directly divides the effective segment in the central-clipped audio signal, and then obtains a short-time energy curve, which is the sound intensity of the audio signal. Based on the AMDF, the modified short-term average amplitude difference function (MAMDF), and the cyclic short-term average amplitude difference function (CAMDF), we can obtain the pitch information of the audio signal [12]. Meanwhile, an appropriate threshold is set to form a time domain classifier. After sorting by the classifier, we can divide the mouth opening angle into three presets in the previous section: low-pitched, roaring, and screaming. According to the sound intensity, the final opening angle data array corresponding to the time is output. Fig. 3 shows the flowchart of time-domain controller.

A. Audio Signal Pretreatment

In order to ensure feature parameters got from this time-domain controller more reasonable, it is necessary to do some pre-processing work at first. Due to the short-term stability of audio signals, there are generally four steps in pretreatment, such as pre-emphasising, framing, windowing and central clipping [11]. In the model of radiation during the production of a piece of sound, the average power spectrum of the speech signal $x(n)$ is affected by the nose and mouth radiation and the glottal excitation, which causes the speech signal to have a 6 dB/Oct (octave) attenuation in the frequency domain above 800 Hz [11]. So we should pre-emphasize the input of audio signal. The formula is built as follows,

$$H(z) = 1 - \alpha Z^{-1} \tag{1}$$

where α is usually $0.95, 0.97$ or 0.98. In this system, it is 0.97.

To some degree, the whole experiment in this system is based on a principle of short-time stability. The dinosaur sounds indeed have this obvious feature. Therefore, it is easier for us to obtain feature parameters in time domain after framing and windowing the audio signals. Generally, there is always a situation that signal will reveal somehow, which is cut into pieces and applied to calculation. In this situation, we use a window function with finite length to transform and calculate the short segments of audio signal as follows [11],

$$Q_n = \sum_{m=-\infty}^{\infty} T\left[x\left(m\right)\right] w\left(n-m\right) \tag{2}$$

where $x(m)$ denotes the original audio signal input sequence; $w(n)$ is a moving window; T denotes a certain transformation of the speech signal, which can be linear or nonlinear; Q_n is the time series obtained after the transformation of each segment.

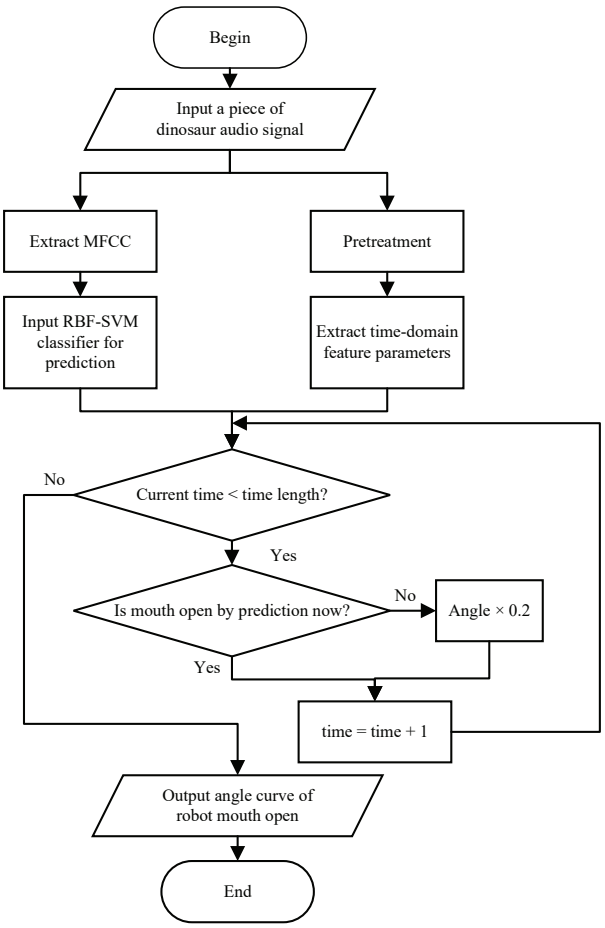


Fig. 3. Flowchart of the system design

For $w(n)$, there are many forms of window functions. Rectangular window, Hamming window and Harming window are the three most commonly used in speech recognition. Taking into account that the width of the window function will affect the smoothness of the audio signal, we finally choose Hamming window.

$$w\left(n\right) = \begin{cases} 0.54 - 0.46\left(2\pi n/\left(N-1\right)\right) & 0 < n \leq N-1 \\ 0 & others \end{cases} \tag{3}$$

After the speech signal is windowed, the slope of the start and end of each frame will be reduced. Meanwhile, the edges of the window will not change rapidly. The intercepted speech waveform will transit slowly to zero at both ends, which effectively reduces the truncation effect of the speech pause. The low amplitude part of the sound signal contains a lot of formant information, while the high amplitude part contains a lot of pitch information. The central clipping method uses a central clipping function to remove the low amplitude part of the signal. It is a nonlinear processing method [13]. The

function is defined as follows,

$$y(n) = C[x(n)] = \begin{cases} x(n) - C_L & x(n) > C_L \\ 0 & |x(n)| \le C_L \\ x(n) + C_L & x(n) < -C_L \end{cases} \quad (4)$$

where C_L is generally $60\% - 70\%$ of the maximum of signal amplitude.

B. Endpoint Detection and Pitch Estimation

For the lip-sync system, endpoint detection is very important, which has a great impact on the performance of the system. A correct goal point can remove redundant information, reduce the amount of computation, and improve the efficiency of the system. Thus, we employ a short-time zero-crossing rate function and a short-time energy function to calculate the starting and ending points of the input audio signal.

For narrowband signals, when the signal is a sine wave with a frequency of F_0, the sampling frequency is F_S, the number of samples in each cycle is $\frac{F_S}{F_0}$, and the zero-crossing number in each sine period is 2, then the average zero-crossing rate of the signal can be obtained,

$$Z = 2g\frac{F_0}{F_S} \quad (5)$$

The short-time average zero-crossing rate of speech signals is defined as follows,

$$Z_n = \sum_{m=-\infty}^{\infty} |sgn[x(m)] - sgn[x(m-1)]| w(n-m) \quad (6)$$

The short-time energy of speech signals is definited as follows,

$$E_n = \sum_{-\infty}^{\infty} [x(m) w(n-m)]^2 \quad (7)$$

where $w(n)$ is exactly Hamming window as mentioned.

As the most important parameter in time domain, pitch frequency needs to be extremely precise. Thus, we employ a kind of innovative and efficient algorithm, which includes a AMDF, a MAMDF and a CAMDF [12] [15].

Similar to ACF method, the AMDF is employed to enlarge periodic points. Through finding the distance between the first local minimum point and the starting point, AMDF can obtain the pitch period. This function is defined as follows,

$$D(k) = \frac{1}{N}\sum_{n} |s(n) - s(n-k)| \quad (8)$$

We can obtain the peak point in AMDF curve, which is (n_{max}, R_{max}). By connecting this point with each point $(k, 0)$, for k from 0 to N (except n_{max}), we can get N lines. Every line is subtracted from the AMDF curve, and the absolute value of the difference is averaged. Finally, the maximum five values are obtained by comparison. The formula is defined as follows,

$$L(n) = \frac{R_{max}}{|n_{max}-k|} \times n \quad n = 0, 1, 2, ..., |n_{max} - k| \quad (9)$$

$$M(k) = \begin{cases} \frac{1}{|n_{max}-k|}\sum_{n=0}^{|n_{max}-k|} |L(n) - R(k+n)| & k < n_{max} \\ \frac{1}{|n_{max}-k|}\sum_{n=0}^{|n_{max}-k|} |L(n) - R(k-n)| & k > n_{max} \end{cases} \quad (10)$$

where $L(n)$ denotes each line function connecting (n_{max}, R_{max}) and $(k, 0)$; $M(k)$ denotes an array of MAMDF values; $R(n)$ denotes the values of AMDF for each point.

After the last step, we input the top 7 values of k into CAMDF, and regard the minimum value as the accurate pitch frequency by comparison. The CAMDF is defined as follows,

$$D(k) = \frac{1}{N}\sum_{n=1}^{N} |s_\omega(mod(n+k, N)) - s_\omega(n)| \quad k = 1, 2, ..., N \quad (11)$$

C. SVM Classifier

Because of the complexity of the audio input, the real-time performance of the time domain controller and the accuracy of the time domain classifier are not perfect. For an accurate classification, we adopt the frequency domain feature parameters to train the SVM classifier with the RBF kernel function, which can effectively assist the time domain controller [14]. On the other hand, in most researches on speech recognition, MFCC is widely used as the input in different sorts of classifier-training. Mel scale is a nonlinear frequency scale based on the human ear's sensory judgment of pitch changes. The relationship with the frequency is as follows,

$$m = 2595\log_{10}\left(1 + \frac{f}{700}\right) \quad (12)$$

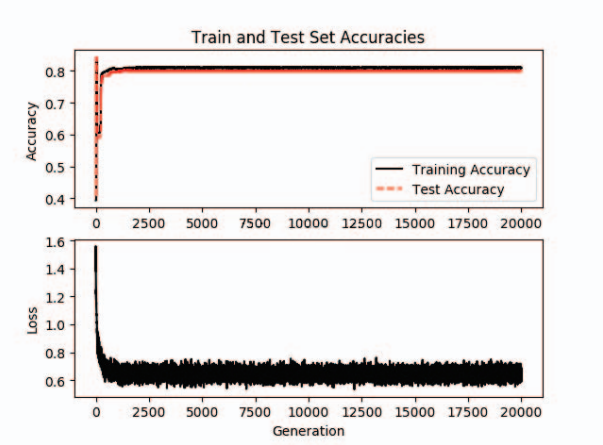
Based on SVM with a linear kernel function, we enter 3289 groups of datasets, where each dataset is a 20-dimensional MFCC feature vector, corresponding to a certain 1 or -1 label. In this paper, a cross-validation method of 80% training set and 20% test set is used. After 5000 cycles of training, the effect is good with the conditions that the penalty coefficient is 0.01 and the learning rate is 0.1.

IV. Experiments and Results

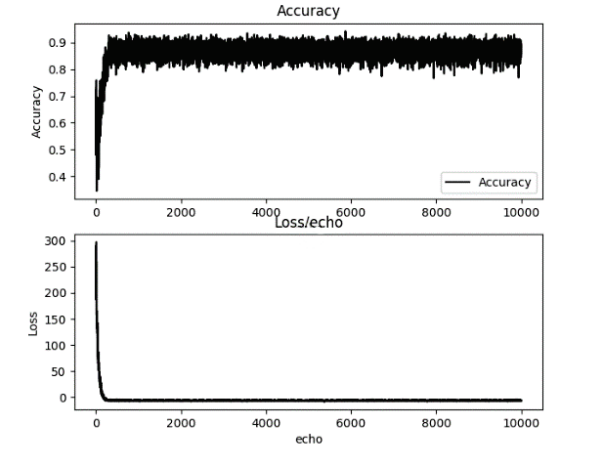
In order to test the proposed lip-sync system, extensive experiments were carried out.

A. Classification Effects of Linear-SVM and RBF-SVM

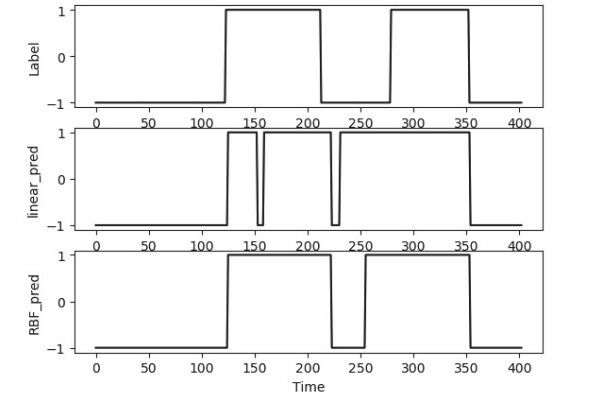
The first experiment focuses on the classification effects of linear-SVM and RBF-SVM. Fig. 4 shows the experimental results about the mouth openshut prediction based on these two different SVM classifiers. Firstly, we separaetly trained and tested SVM classifiers with different kernel function by effective audio signals of dinosaurs, as shown in Figs. 4(a) and 4(b). For the linear-SVM, the accuracy of the classification of test set is steadily at 80%, and the convergence of loss function is relatively steady. Due to a large number of input feature vectors, the trained classifier with the linear kernel function dose not perform as well as we predicted. By contrast, the RBF-SVM mode improves the classification accuracy of the test set. Specially, the average value is about 87%, and the highest is 92.7%. It is obvious that for high-latitude vector



(a) Classification results of dinosaur audio data by the linear-SVM



(b) Classification results of dinosaur audio data by the RBF-SVM



(c) Comparison result of classification of a random audio between linear-SVM and RBF-SVM

Fig. 4. Experimental results in SVM classifier with linear and RBF kernel function.

input, a nonlinear classifier can generally achieve better results than a linear one. Then we captured a piece of dinosaur audio signal from a database as the input of these two classifiers,

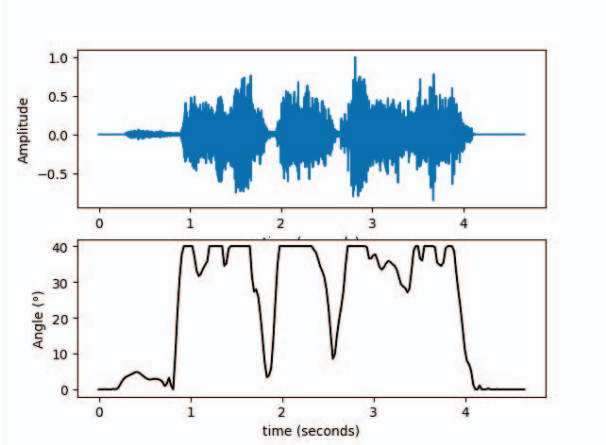


Fig. 5. Experimental results on the time-domain controller

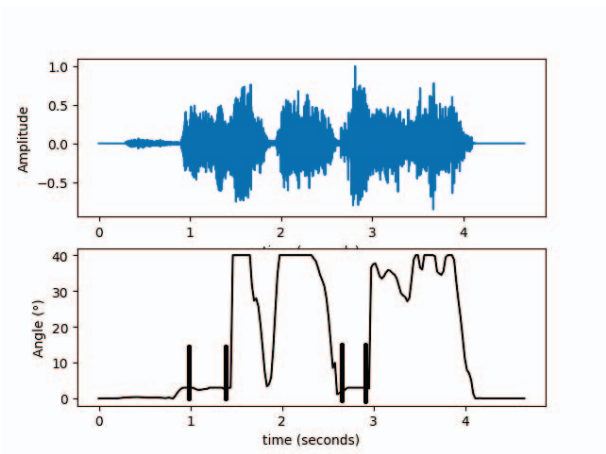


Fig. 6. Experimental resluts on the dual-domain controller

as shown in Fig. 4(c). The two figures at the bottom of Fig. 4(c) represent separately the classification results of linear-SVM and RBF-SVM. As for the linear-SVM, we can see that some large deviation appears when $t = 155$ s to 155 s and $t = 210$ s to 280 s. By contrast, the RBF-SVM has a great improvement in accuracy, although some disagreement also appears owing to some unavoidable deviation of training data. However, it is sure that RBF-SVM classifier works better on high dimensional input. Finally, the prediction accuracy of the system among the 16454 groups of data reached 92.1%.

B. Experiments on Different Controllers

The second experiments were carried out to explore the performance of two different controllers for a random sound of dinosaurs, i.e., a time-domain controller and a dual-mode controller. As show in Figs. 5 and 6, the mouth-open curves can effectively respond to the sound of dinosaurs under the control of these two controllers. Obviously, the dual-mode controller has a better performance. For example, it successfully identifies two fragments belonging to the low pitch type, benefiting from the detection of RBF-SVM classifier, see the

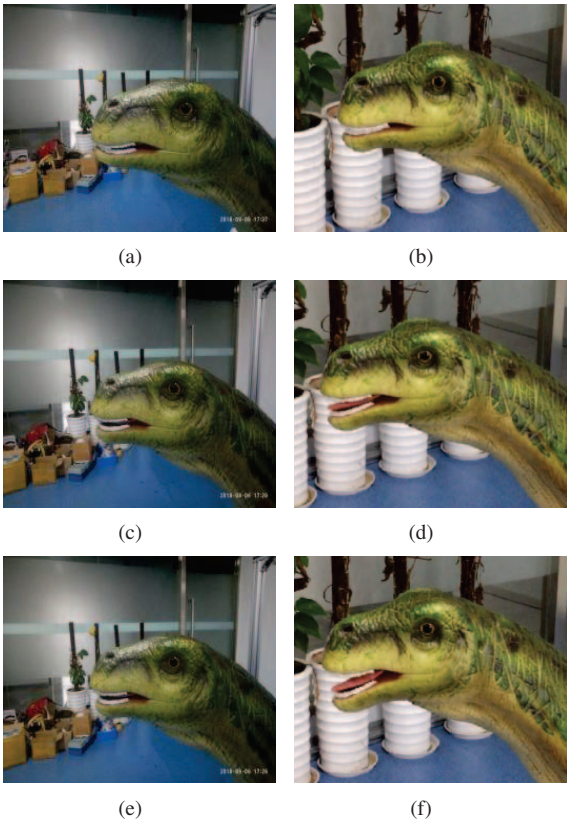


Fig. 7. Experimental snapshots of the dinosaur robot

short black lines in Fig. 6. In experiments, 95.2% of 189 volunteers thought the dinosaur mouth animation controlled by this dual-mode lip-sync system was reasonable, realistic and natural enough, although 2.6% thought the animation was stiff and 2.2% thought it was totally fake.

C. Experiments on a Dinosaur Robot

The last experiment was performed based on a dinosaur robot. Two types of audio files were adopted, including a low pitch sound and a roaring one. In experiment, we can find that the dinosaur robot can show different mouth shapes with the audio files. As show in Figs. 7(a), 7(c) and 7(e) the robot slightly opened its mouth when a low pitch sound was played. By contrast, a large mouth was opened when a roaring sound was heard, see Figs. 7(b), 7(d) and 7(f).

V. Conclusion and Future Work

This paper has proposed a dual-mode real-time lip-sync system for a dinosaur robot, which consists of different controllers and classifiers in both time and frequency domains. In time domain, a classifier is designed to extract the sound features including pitch and intensity. Meanwhile, considering the direct influence of input, a nonlinear mapping relationship between time-domain feature parameters and mouth open angles is designed instead of a common linear one. In frequency domain, an improved AMDF is employed in frequency measurement to obtain a real-time and high accuracy output of mouth open angles. Furthermore, in order to ensure the precision of classification, the SVM with RBF kernel function is adopted to train audio data and generate a frequency classifier with high accuracy in testing. Combining these two models together, the proposed lip-sync system realizes an efficient and real-like curve of mouth open control.

The future work will focus on improving the precision of angles of simulating animal robots' mouth open by changing complexity of labels. Furthermore, this system can be extended to other bionic animal robots in ways of trained by different kinds of animal audio data, which will be necessary in social public education, child accompany and other fields requesting bionic service robots.

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China under Grant 61603388, Grant 61633020, and Grant 61633017, and in part by the Key Project of Frontier Science Research of Chinese Academy of Sciences (Grant No. QYZDJ-SSW-JSC004).

References

[1] T. Wang, Y. Tao, and Y. Chen, "Research status and development trend of service robot technology," *Science in China*, 42(9):1049–1066, 2012.

[2] D. Hanson, D. Mazzei, C. Garver, D. D. Rossi, and M. Stevenson, "Realistic humanlike robots for treatment of ASD, social training, and research; Shown to appeal to youths with ASD, cause physiological arousal, and increase human-to-human social engagement," In *Proceedings of International Conference on Pervasive Technologies Related To Assistive Environments Petra*, Nov. 2012.

[3] W. G. Wu, C. Song, and Q. M. Meng, "Development and experimentation on speech sounds and degree of lip rounding system for 'H&Frobot-III' humanoid head portrait robot," *Journal of Machine Design*, 25(1):15–19, 2008.

[4] K. Kaneko, F. Kanehiro, M. Morisawa, and K. Miura, "Cybernetic human HRP-4C," In *Proceedings of IEEE-RAS International Conference on Humanoid Robots*, Paris, France, Dec. 2009, pp. 7–14.

[5] Y. Chen, F. Wu, W. Shuai, N. Wang, and R. Chen, "KeJia robot-an attractive shopping mall guider," *Social Robotics. ICSR 2015. Lecture Notes in Computer Science*, 9388:145–154, 2015.

[6] I.M. Verner, A. Polishuk, and N. Krayner, "Science class with RoboThespian: Using a robot teacher to make science fun and engage students," *IEEE Robotics & Automation Magazine*, 23(2):74–80, 2016.

[7] Z. M. Wang, L. H. Cai, and H. Z. Ai, "Text-to-visual speech in chinese based on data-driven approach," *Journal of Software*, 16(6):1054–1063, 2005.

[8] S. Taylor, B. J. Theobald, and I. Matthews, "A mouth full of words: Visually consistent acoustic redubbing," In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, Brisbane, QLD, Austrialia, April 2015, pp. 4904–4908.

[9] X. Fan, X. Yang, and S. O. Software, "A speech-driven lip synchronization method," *Journal of Donghua University*, 4:466–471, 2017.

[10] Institute of Computing Technology, Chinese Academy of Sciences, "A Text-Speech-Driven Face Animation Generation Method and System," CN 201510876078.4, May 4 2016.

[11] L. Zhao, Speech Signal Processing, Mechanical Industry Press, 2016.

[12] Y. Gao, W. Chen, G. Yan, and J. Du, "An efficient pitch estimation algorithm," *Electronic Design Engineering*, 18(1):24–25, 2010.

[13] S. Zhang, J. Fan, and C. Wu, "Improvement of centric clipping for chinese sound tone recognition with adaptive speaking speed," In *Proceedings of International Conference on Signal Processing*, Beijing, China, Nov. 2006, pp. 1–4.

[14] S. Zhou, J. Liao, and X. Shi, "Kernel parameter selection of RBM-SVM and its application in fault diagnosis," *Journal of Electronic Measurement and Instrumentation*, 9:69–74, 2014.

[15] J. Sun, "Estimation of precession period based on improved circular AMDF," *Science Technology and Engineering*, 15(27):152–158, 2015.