Synthesizing Avatar Robot: Learning from Audio

Z.S. Fang, D.T. Feng, Y.F. Li, J.Q. Liao, Z.J. Ou, Q.H.Wang, S.H. Xie,
P.Y. Zeng, R. Zhang, Y.M. Zhao, M.C. Zhu

2019 Artificial Intelligence DeeCamp, Group 51, Guangzhou, China

Introduction

Virtual robot is an anthropomorphic robot service that combines audio and video with real human images through speech synthesis, lip synthesis, facial expression synthesis, body movement prediction and other artificial intelligence technologies.

BAIDU IFLYTEK SOGOU

Figure 1. Advanced products on the market

Product Applications Scenarios:

- Newscaster
- Online Anchorwoman
- Virtual Band
- Online Education
- Bank virtual employee

Figure 2. Newscaster

Figure 3. Online Education

Objectives

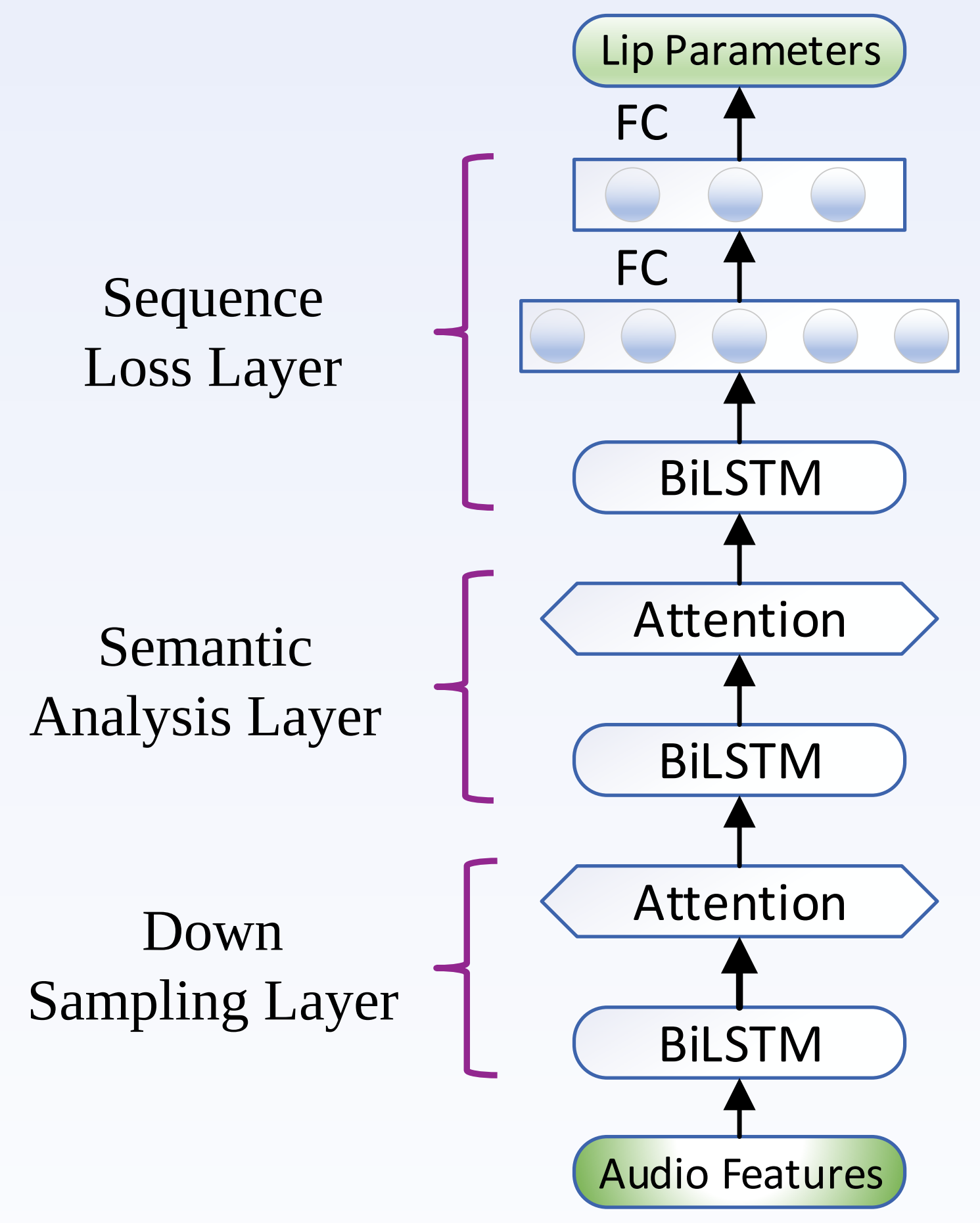
Problem 1: How to synchronize mouth movements with audio?

Objective 1: Develop an algorithm to map audio features to lip features

Problem 2: How to achieve a good human-computer interaction?

Objective 2: Enable robots to make intelligent feedback based on user behavior

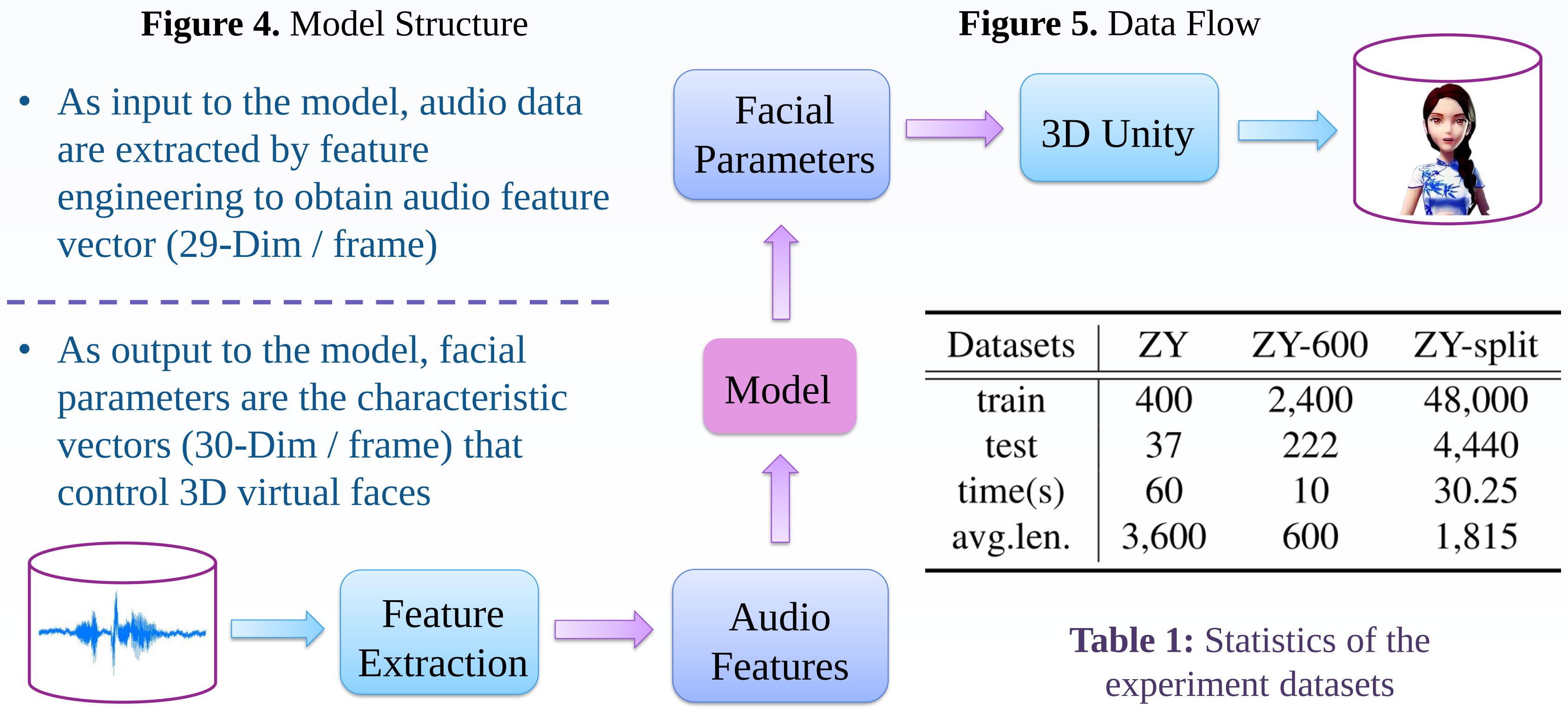
Methods – Lip Inference



- **Sequence Loss Layer**: Smooth the features learned from the first two layers of bidirectional LSTM, and connect the features from the global perspective to avoid violent shaking of expression
- **Semantic Analysis Layer**: The frame interval is further expanded to learn the relationship with longer time span
- **Down Sampling Layer**: Learn local features to increase the range of body movements

Figure 4. Model Structure

Figure 5. Data Flow



- As input to the model, audio data are extracted by feature engineering to obtain audio feature vector (29-Dim / frame)
- As output to the model, facial parameters are the characteristic vectors (30-Dim / frame) that control 3D virtual faces

Datasets	ZY	ZY-600	ZY-split
train	400	2,400	48,000
test	37	222	4,440
time(s)	60	10	30.25
avg.len.	3,600	600	1,815

Table 1: Statistics of the experiment datasets

Methods – Dialog Management

Based on the user voice processing technology, we realize the interactive effect of situational reasoning through the engineering system design. Our Virbot achieved the same result as Baidu.

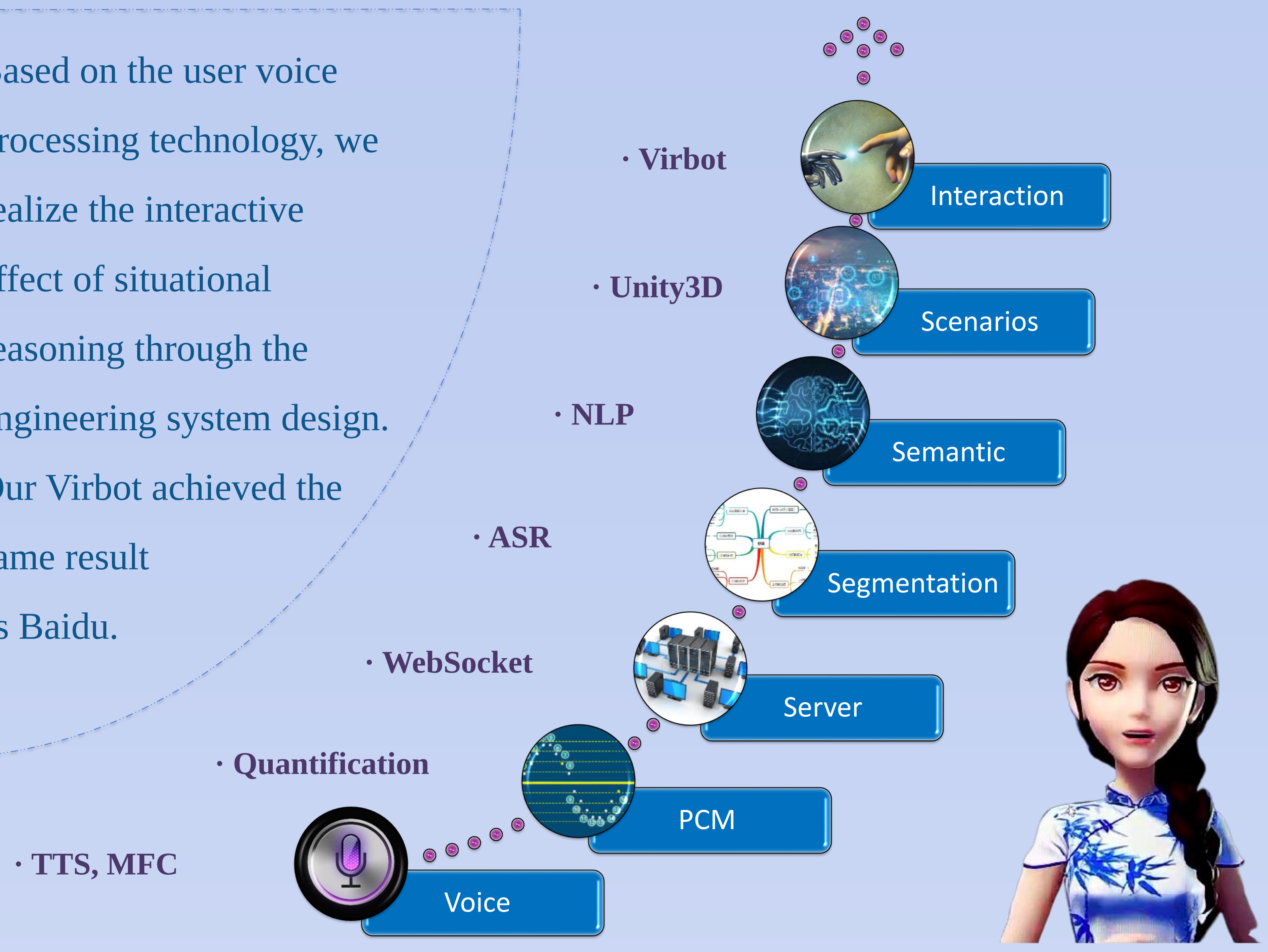
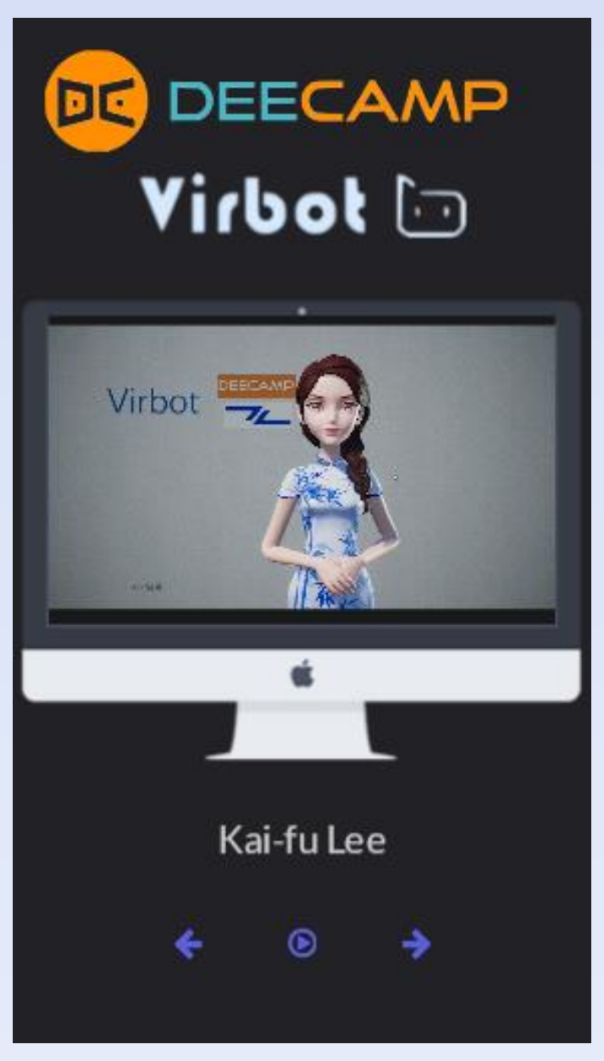


Figure 6. Dialog management module data flow

Result

We applied web development and Unity3D modeling technology to facilitate the display of project effects



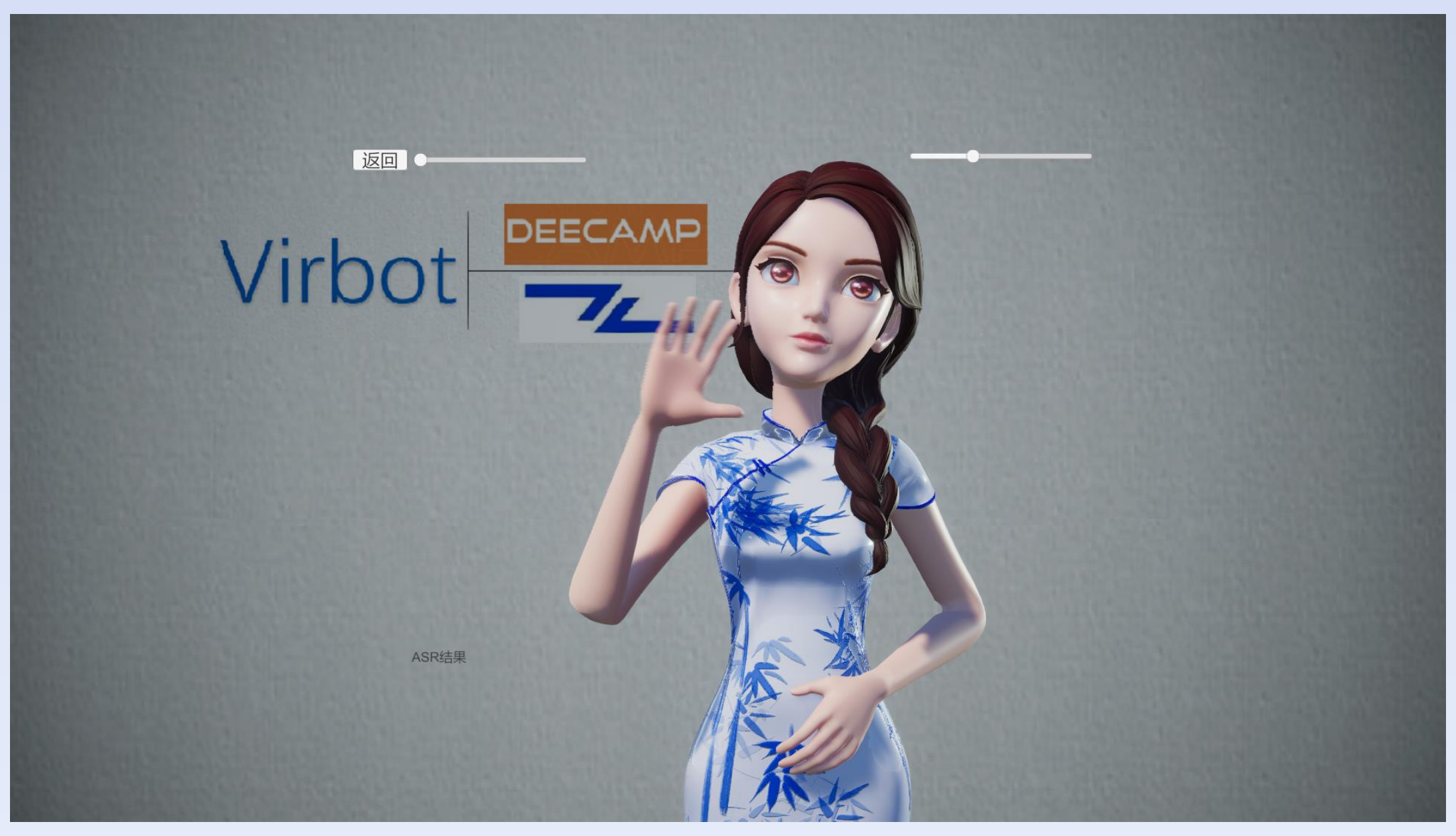


Figure 7. Web Interface

Figure 8. Unity3D Interface

Performance of two different Loss (RRMSE, MSE) on three different data sets

	ZY		ZY-600		ZY-split	
	RRMSE%	MSE%	RRMSE%	MSE%	RRMSE%	MSE%
VBOT-d	22.52	21.71	20.31	19.78	18.30	17.56
VBOT-s	19.44	18.19	18.32	18.28	15.43	14.11
VBOT	16.32	13.66	15.12	13.33	12.60	10.72
VBOT-pca	**12.56**	**11.21**	**10.94**	**9.43**	**9.36**	**8.54**

$$\mathrm{L}_{\mathrm{RRMSE}}=\frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2}}{\sum_{i=1}^{n}y_i}\times 100$$

$$\mathrm{L}_{MSE}=\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2\times 100$$

Table 2: The overall performance

Conclusions

- The fitting of lip shape and audio has always been a major technical problem for the existing virtual robots on the market. In this project, we use three-layer BiLSTM architecture to smooth the relationship between audio frames and improve the predictive effect of mouth shape and motion. It makes our products more competitive in the market
- At the same time, we design a complete engineering process for intelligent voice conversations to improve the usability of our virtual robot. Intelligent expression and dialogue will make our products more attractive to customers
- The technology and effect of our robot is comparable to Baidu in the field of 3D virtual robot

Acknowledgements

This project is supported by ZHUIYI