# Deep Neural Network Based 3D Articulatory Movement Prediction Using Both Text and Audio Inputs

**Lingyun Yu, Jun Yu, Qiang Ling**
**University of Science and Technology of China**
**yuly@mail.ustc.edu.cn, (harryjun, qling)@ustc.edu.cn**

# CONTENTS

# Part One

# Introduction

- □Background
- □Application

# Background

In human speech production, it is the movements of articulators, such as the tongue, jaw, lips and velum, that generate and shape the acoustic signal.

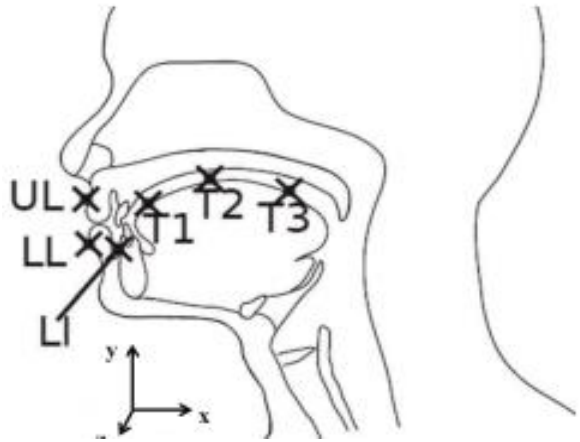
# Application

In speech recognition, articulatory features can provide additional speech production information.

In speech synthesis, articulatory features can improve the voice quality or retouch the characteristics of synthesized voice

Speech visualization

# The positions of articulators




(a)

(b)

Fig.1 Photograph of an EMA setup taken during a dataset recording session.

(b) Sensor adhering position

# Method

## Part Two

- The overall framework
- Our proposed network
- Each part of the network

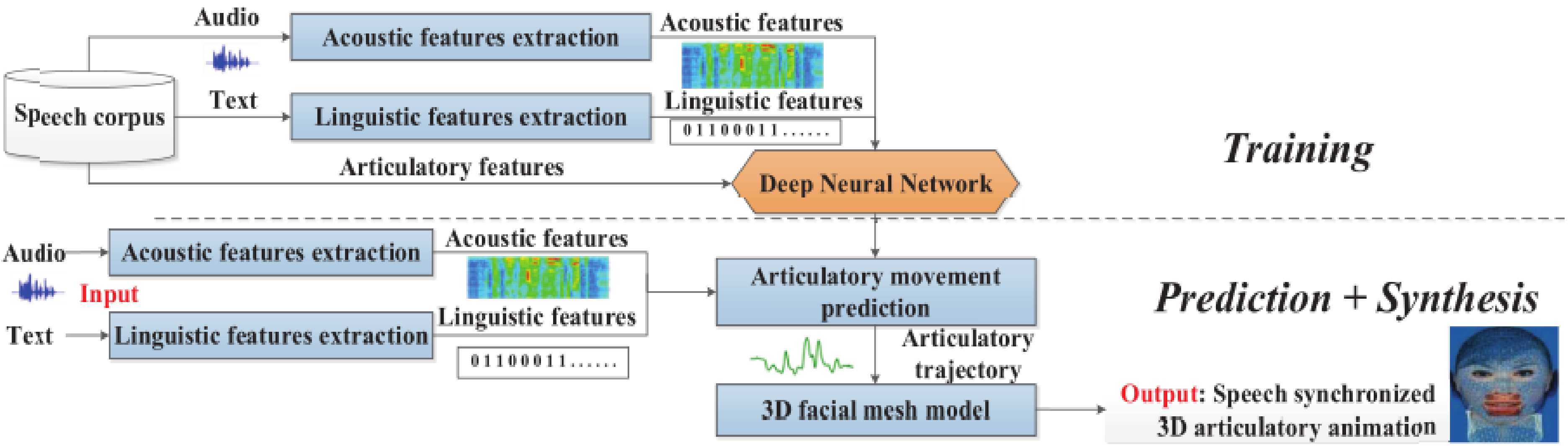



Fig.2 The overall framework.

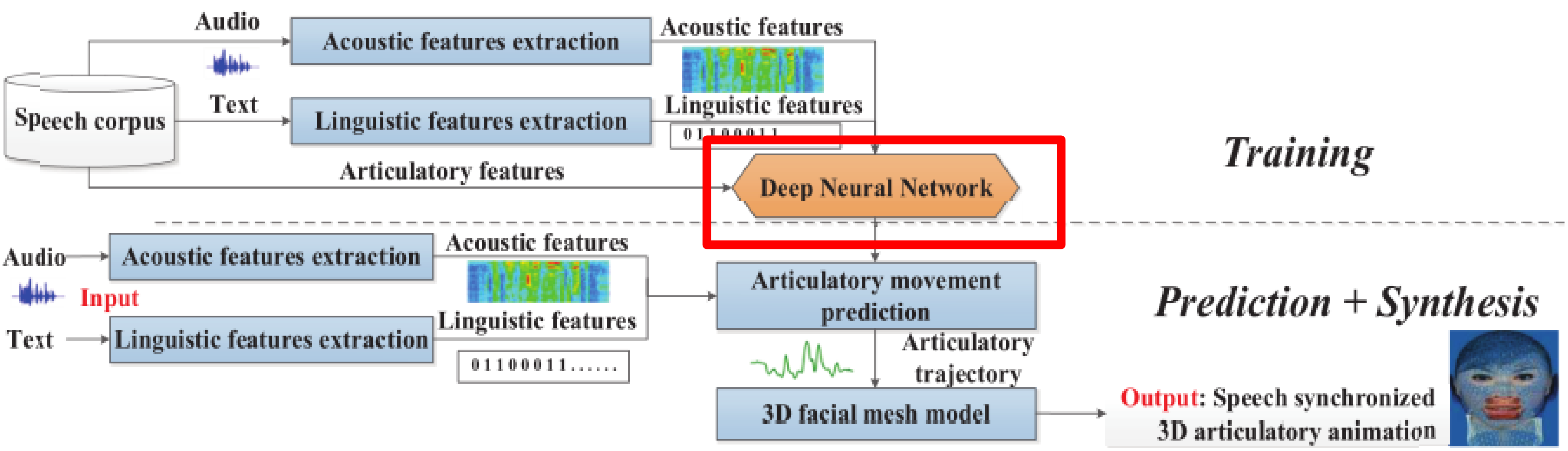



Fig.2 The overall framework.

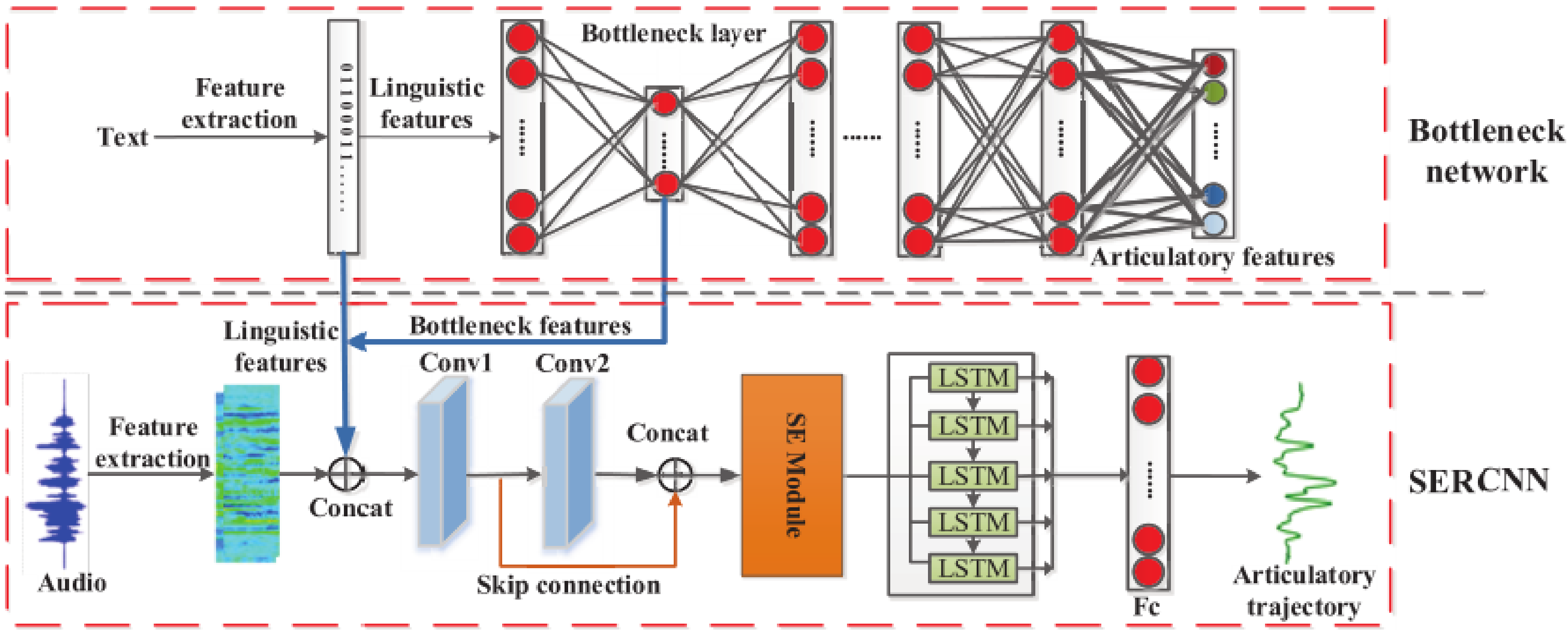



Fig.3 The proposed bottleneck squeeze-and-excitation recurrent convolutional neural network (BSERCNN) for articulatory movement prediction given both text and audio.

**Nework**

**Bottleneck features**

Squeeze-and-Excitation (SE) Block

# Bottleneck features

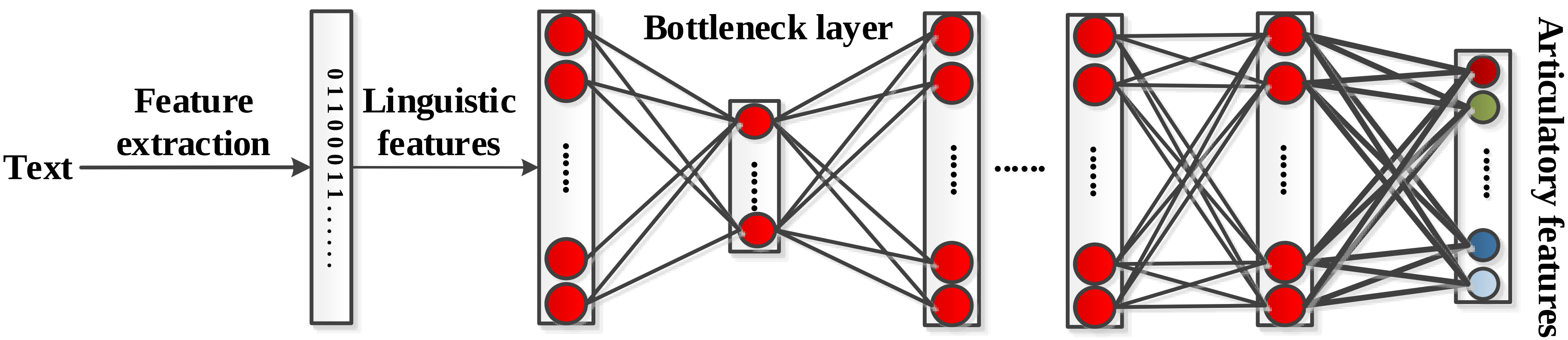
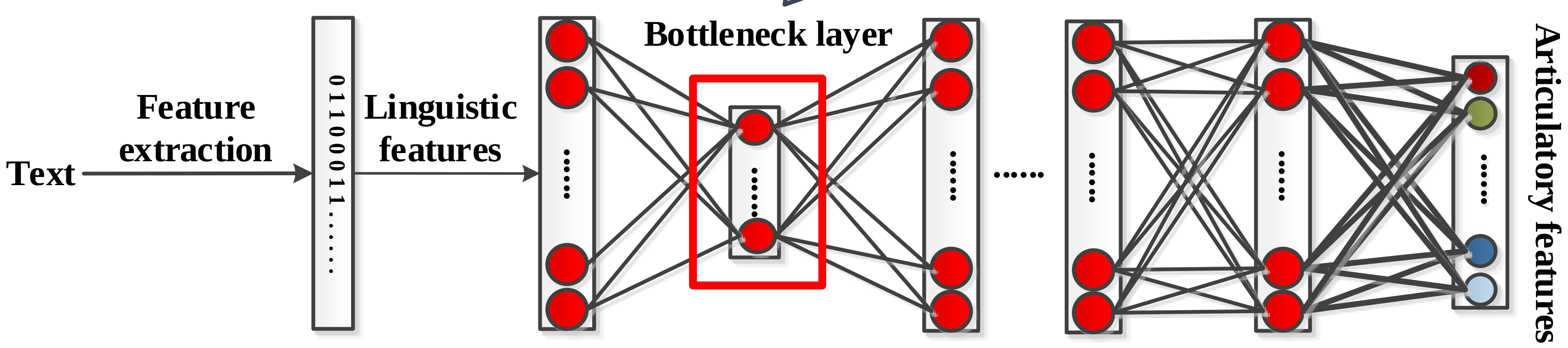



Fig.4 The bottleneck network.

# Bottleneck features

(1)Bottleneck features represent a nonlinear transform and dimensionality reduction of input features.
(2)Bottleneck features capture information that is complementary to input features




**The bottleneck features are introduced as the supplementary input features when text and audio are integrated as inputs.**

# Squeeze-and-Excitation (SE) Block

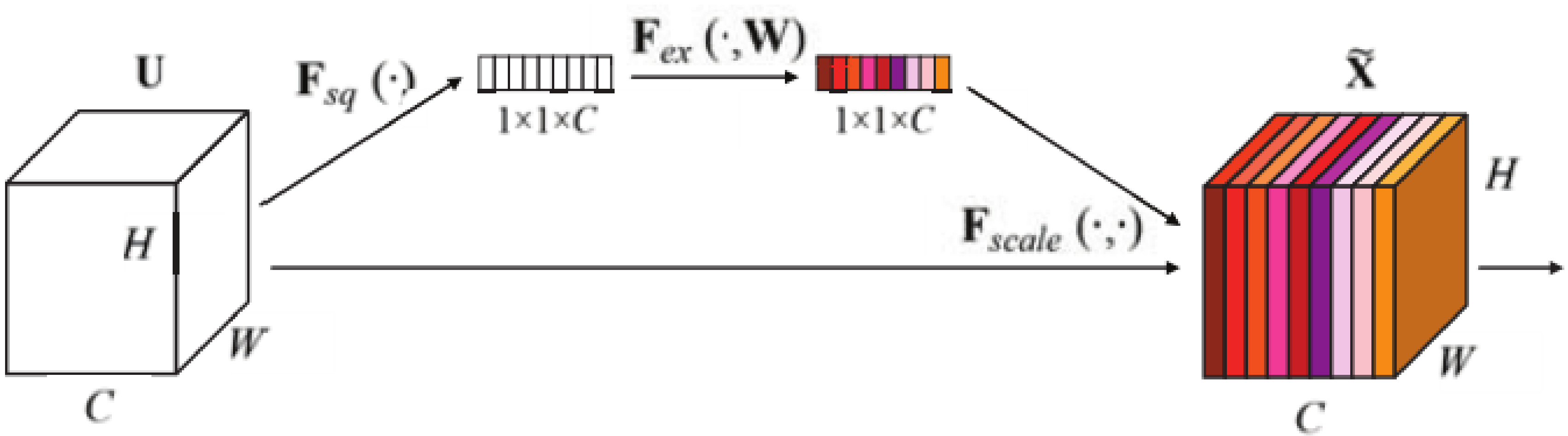
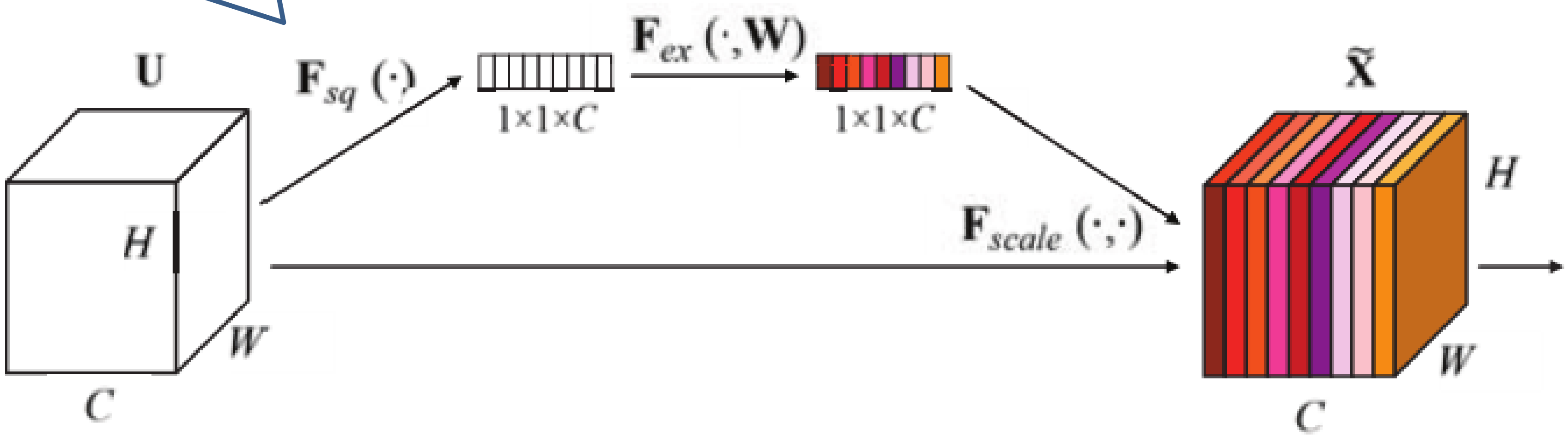



Fig.5 The SE block.

# Squeeze-and-Excitation (SE) Block

$$z_c = F_{sq}(u_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} u_c(i, j)$$



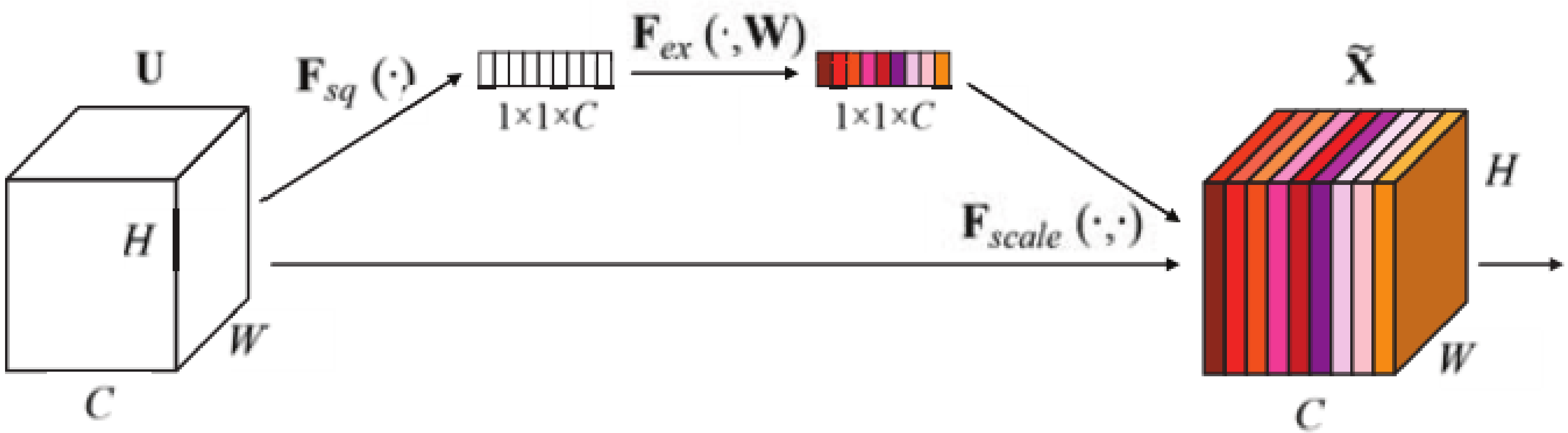

$F_{sq}(.)$ is the squeeze function

$z_c$ is the *c*-th element of the squeezed channels.

$u_c$ is the c-th channel of the input

# Squeeze-and-Excitation (SE) Block

$$s = F_{ex}(z, W) = \sigma(g(z, W)) = \sigma(W_2 \sigma(W_1 z))$$

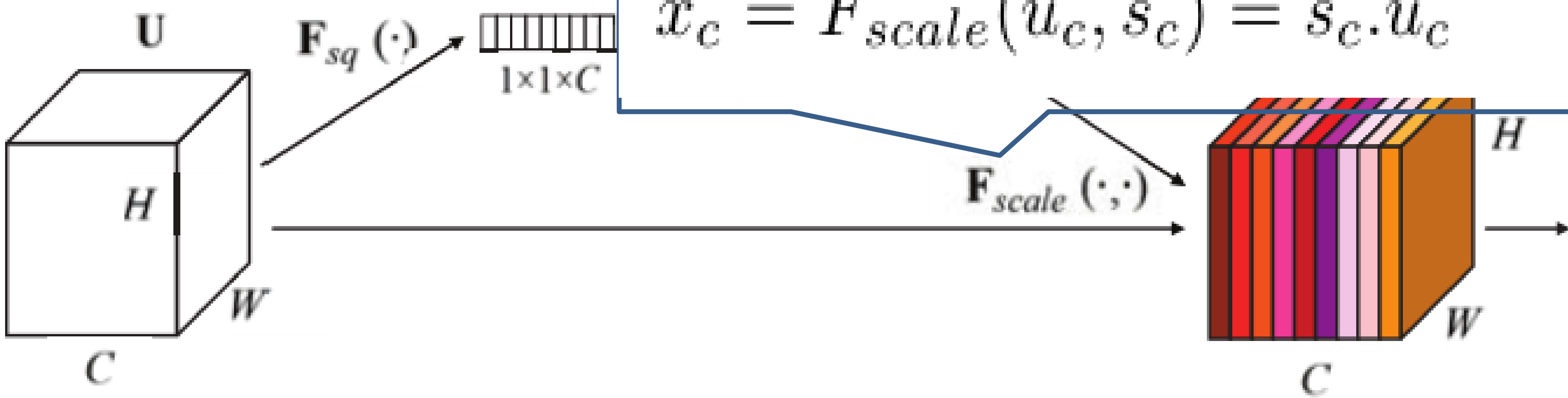



$F_{ex}(.)$ is the excitation function.

$\sigma$ denotes the Sigmoid function.

$W_1$ and $W_2$ denote the 1x1 convolutional layer.

# Squeeze-and-Excitation (SE) Block




$$\tilde{x}_c = F_{scale}(u_c, s_c) = s_c . u_c$$

$F_{scale}(u_c, s_c)$ denotes the channel-wise multiplication between the feature map $u_c$ and the scale $s_c$ .

# Part Three

## Experiment

- Audio input alone
- Text input alone
- Both text and audio

# Dataset

**Dataset &Input Features**

01

## Dataset

MNGU0 dataset with 1263 English utterances

02

## Acoustic features

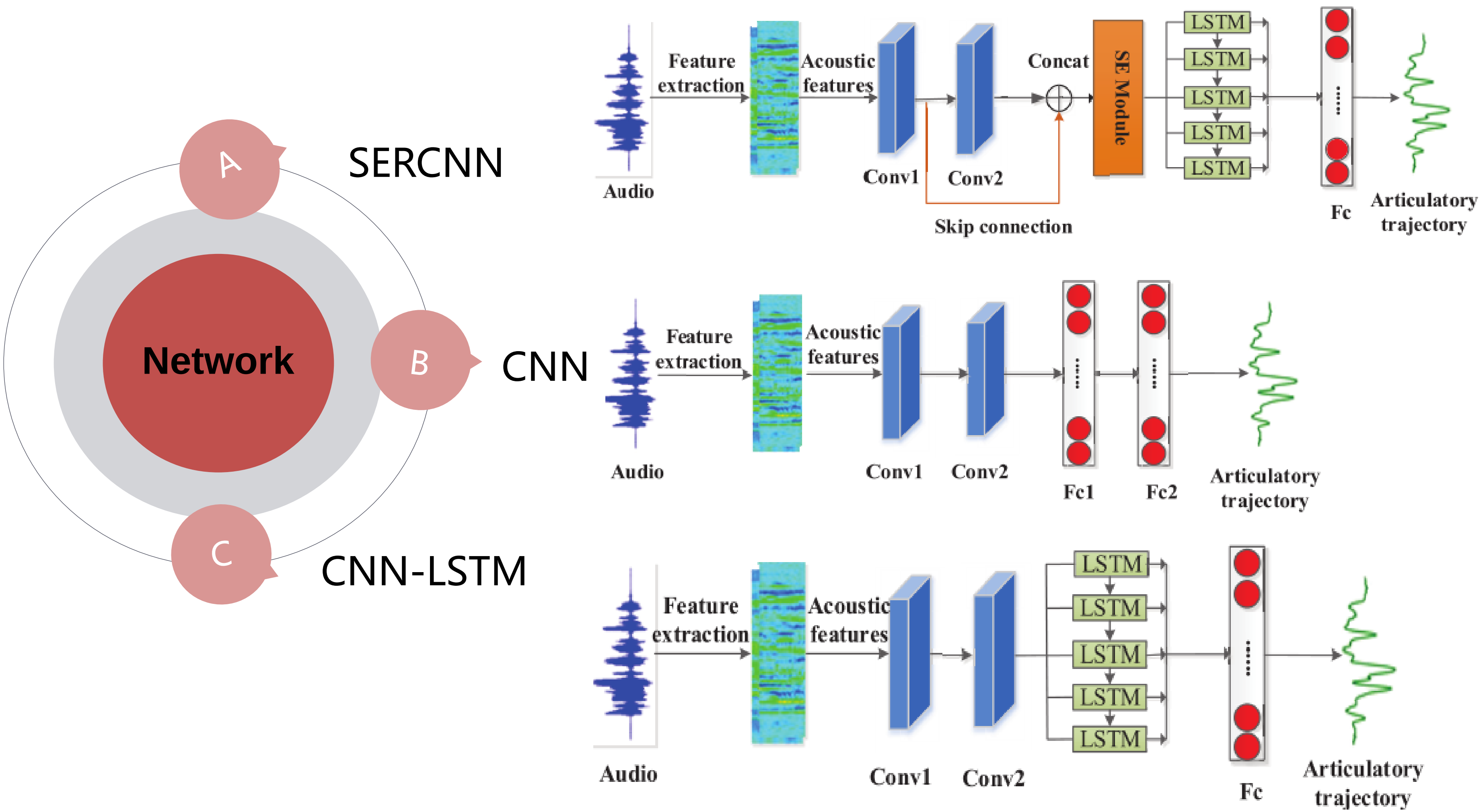
Including 60-dimensional Mel Cepstral Coefficients (MCCs), 1-dimensional band aperiodicities (BAPs) and 1-dimensional log-scale fundamental frequency (logF0) with deltas and delta-delta features.

03

## Linguistic features

Consisting of 416-dimension binary features and 9-dimension numerical features . The 416-dimensional features, derived from the fully-context labels, represent the context information. The 9-dimensional numerical features represent the frame position information.

# Articulatory movement prediction from audio input alone



# Articulatory movement prediction from audio input alone

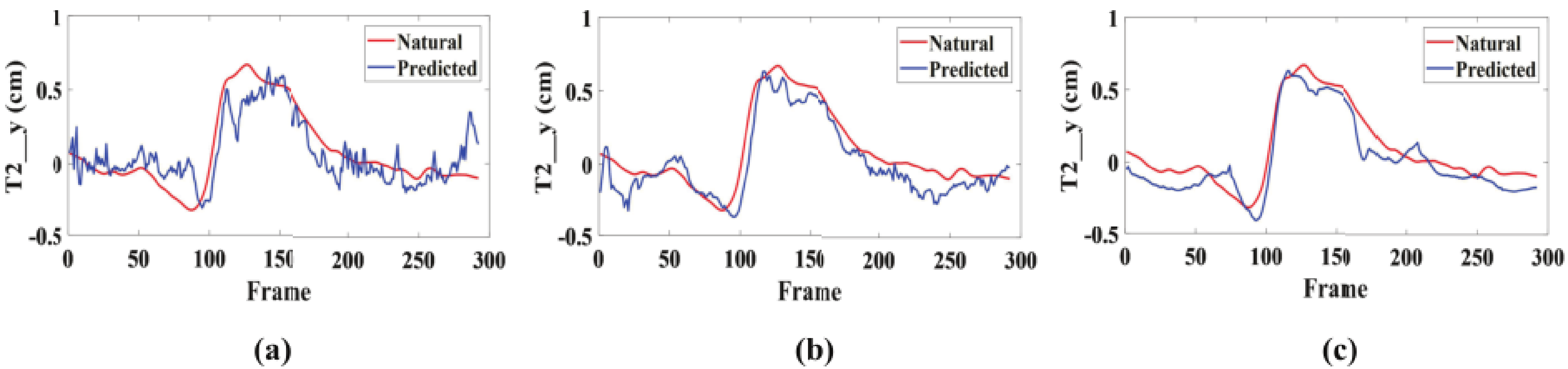



Fig.5 The comparisons for articulatory trajectories predicted from (a) CNN, (b) CNN-LSTM and (c) SERCNN with only audio input.

# Articulatory movement prediction from audio input alone

**Table 1.** The comparison of the RMSE and the correlation coefficient for different network architectures with audio input alone.

| | RMSE | Correlation coefficient |
|---|---|---|
| CNN | 1.191mm | 0.822 |
| CNN-LSTM | 1.001mm | 0.883 |
| SERCNN | **0.747mm** | **0.924** |

# Articulatory movement prediction from text input alone

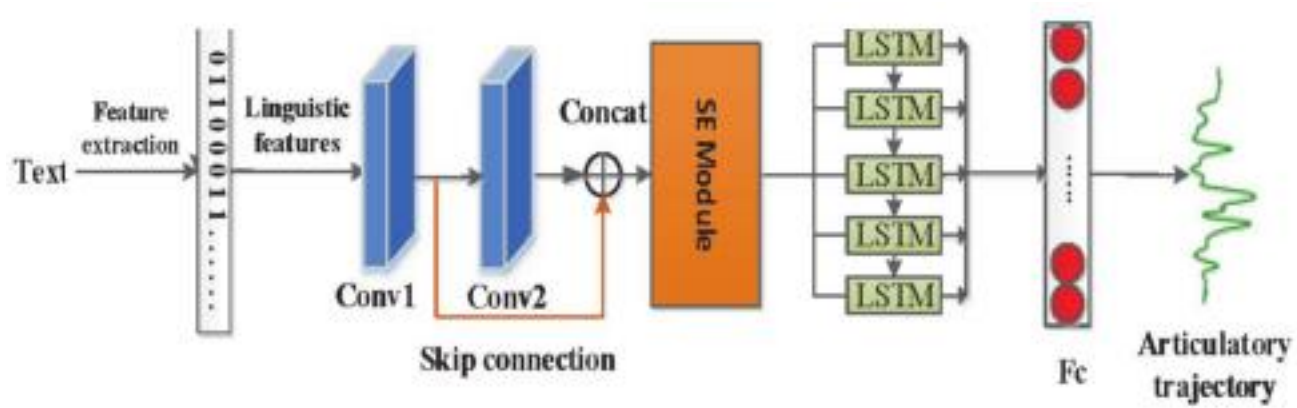



Fig.6 The network architecture of SERCNN for articulatory movement prediction with text input alone.

中国科学技术大学
University of Science and Technology of China

# Articulatory movement prediction from text input alone

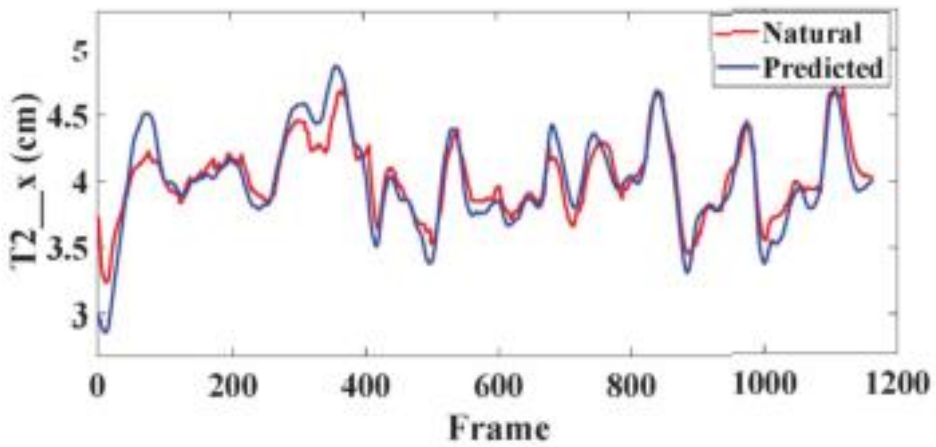



Fig.7 Articulatory trajectories predicted from SERCNN with only text input for T2_x.

# Articulatory movement prediction from text and audio inputs

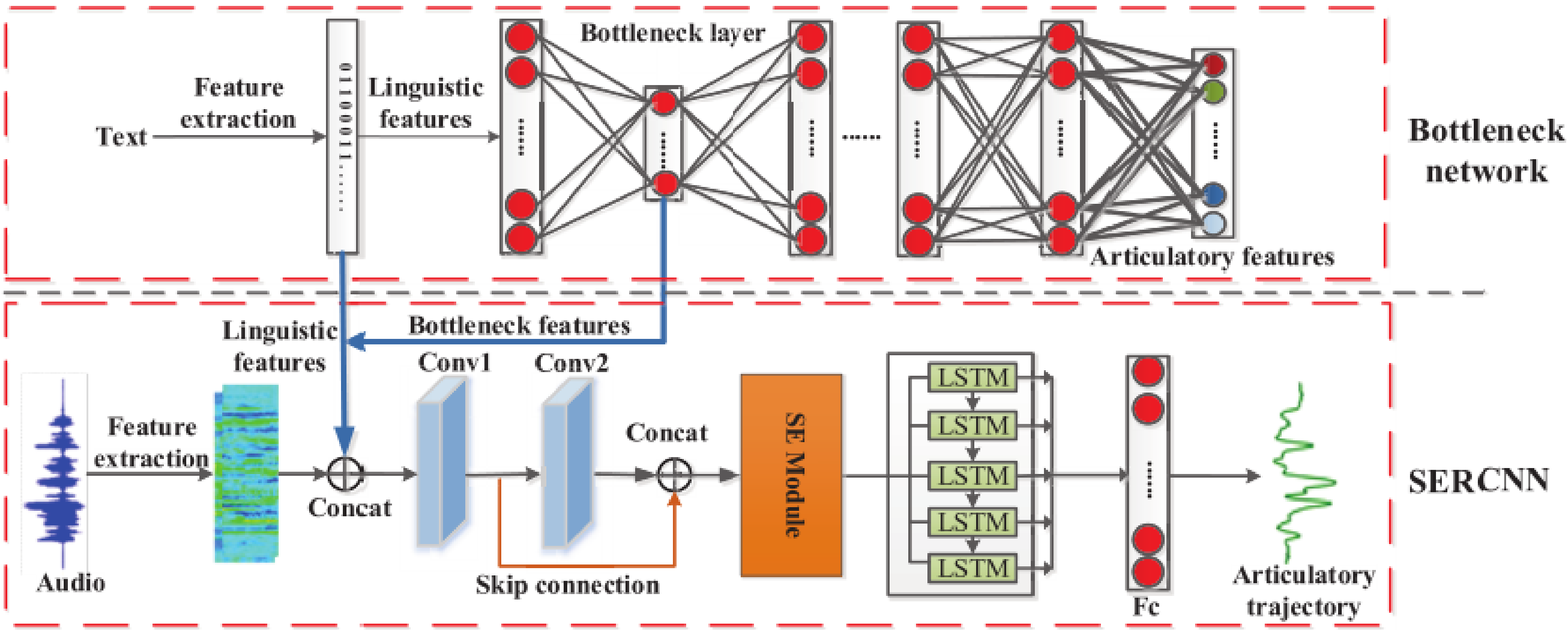



Fig.3 The proposed bottleneck squeeze-and-excitation recurrent convolutional neural network (BSERCNN) for articulatory movement prediction given both text and audio.

# Articulatory movement prediction from text and audio inputs

## ✓ Effect of bottleneck network

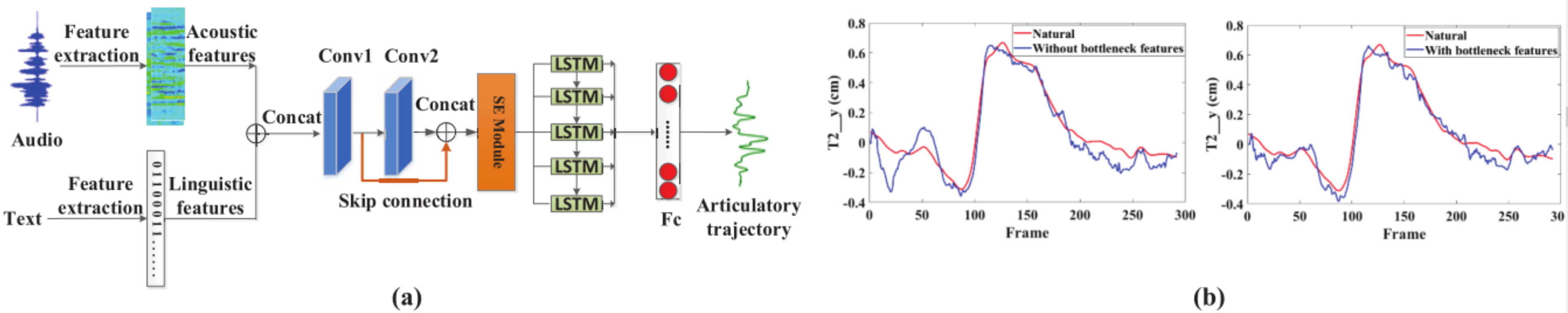



Fig.8 The network of SERCNN for articulatory movement prediction by concatenating linguistic features and acoustic features directly as inputs. (b) The articulatory trajectories for T2 y with or without bottleneck features as input.

# Articulatory movement prediction from text and audio inputs

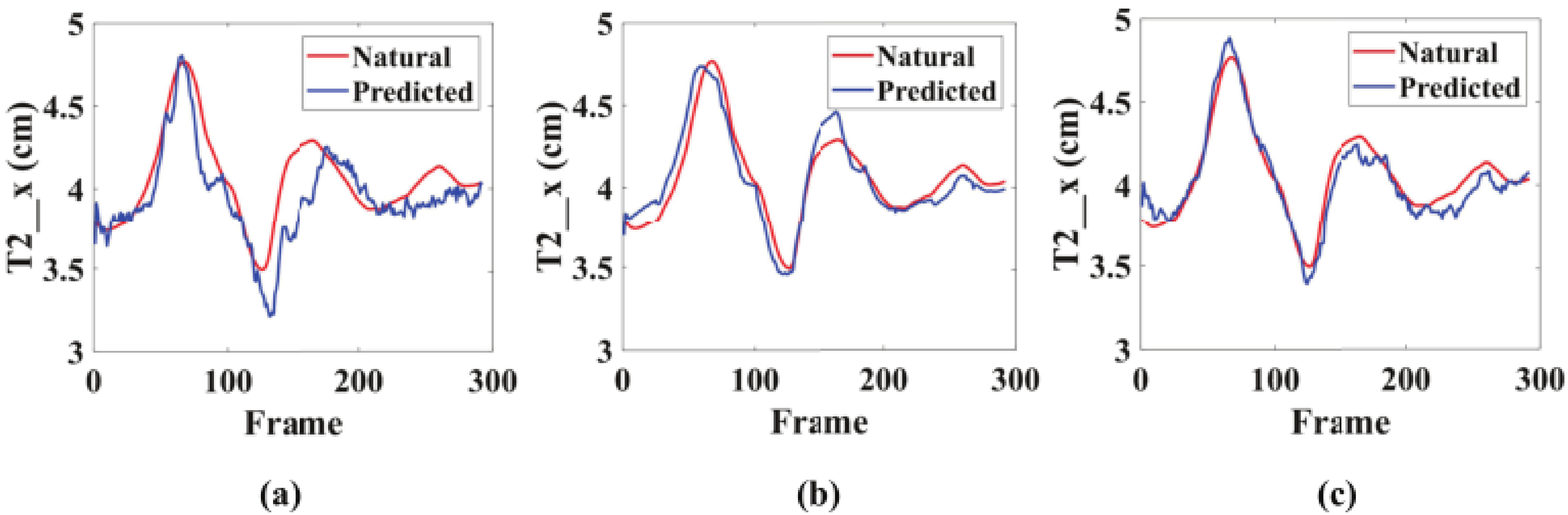



Fig.9 The comparison of the real and predicted articulatory trajectories for T2_x with (a) only audio input, (b) only text input, (c) both text and audio inputs.

# Articulatory movement prediction from text and audio inputs

Tab.2 The RMSE and the correlation coefficient predicted from different methods on MNGU0 dataset.

| | TMDN [26] | HMM [14] | DRMDN [15] | BLSTM [15] | DNN [13] | BLSTM [12] | **BSERCNN** |
|---|---|---|---|---|---|---|---|
| RMSE | 0.99mm | 0.90mm | 0.832mm | 0.816mm | 0.737mm | 0.565mm | **0.563mm** |
| Correlation coefficient | \ | 0.812 | 0.914 | 0.921 | \ | \ | **0.954** |

# Part Four

## Conclusion

- Conclusion
- Future Work

# Conclusion

In this paper, the overall network architecture BSERCNN, combining CNN, LSTM, a skip connection and bottleneck network, is proposed for articulatory movement prediction with both text and audio inputs. Our BSERCNN achieves the state-of-the-art results with the RMSE 0.563mm and the correlation coefficient 0.954. Besides, we also analyze the performance when the input is text alone and audio alone, respectively. Our network also achieves the lowest RMSE 0.695mm in text-to-articulatory mapping. Comprehensive experimental results further prove that both text and audio are essential for this prediction.

# Future Work

In the future, the visualization method of predicted articulatory movements will be improved to increase the realism of the developed system.

Thanks!

Q&A