Multi-Scale Model for Mandarin Tone Recognition

Linkai Peng, Wang Dai, Dengfeng Ke, Jinsong Zhang

School of Information Science, Beijing Language and Culture University, China

penglinkai96@gmail.com,
daiwang_ai@163.com,
8363331@qq.com,
jinsong.zhang@blcu.edu.cn

ISCSLP 2021

Outline

- Background & Motivation
- Model Design
- Experiment
- Results
- Conclusions

Outline

- **Background & Motivation**
- Model Design
- Experiment
- Results
- Conclusions

Background & Motivation

- Previous methods
  - using **forced alignment** to get the frame/segment-level label and then predict tone.
  - CNN-CTC based model.
- Problem
  - without considering **multi-resolution** processing.

Background & Motivation

Multi-Scale

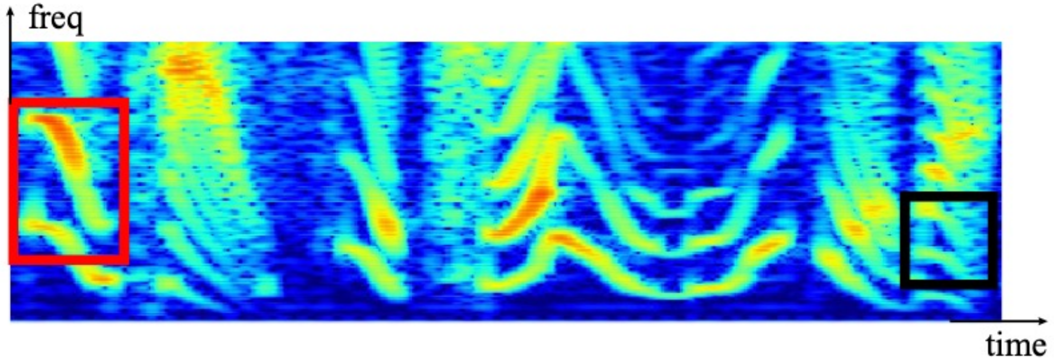

Figure 1: *Mel-spectrogram of F03A161 in dataset 863. Transcript of this part is "... Anger can be understood (... fen4 kai3 shi4 ke3 yi3 li3 jie3 de0)".*

- Different tone contours in continue speaking stream have various time and frequency range.
- Multi-scale feature representations have proven successful for many vision and speech recognition tasks compared to single-scale methods.

Outline

- Background & Motivation
- **Model Design**
- Experiment
- Results
- Conclusions

Model Design

Model Framework

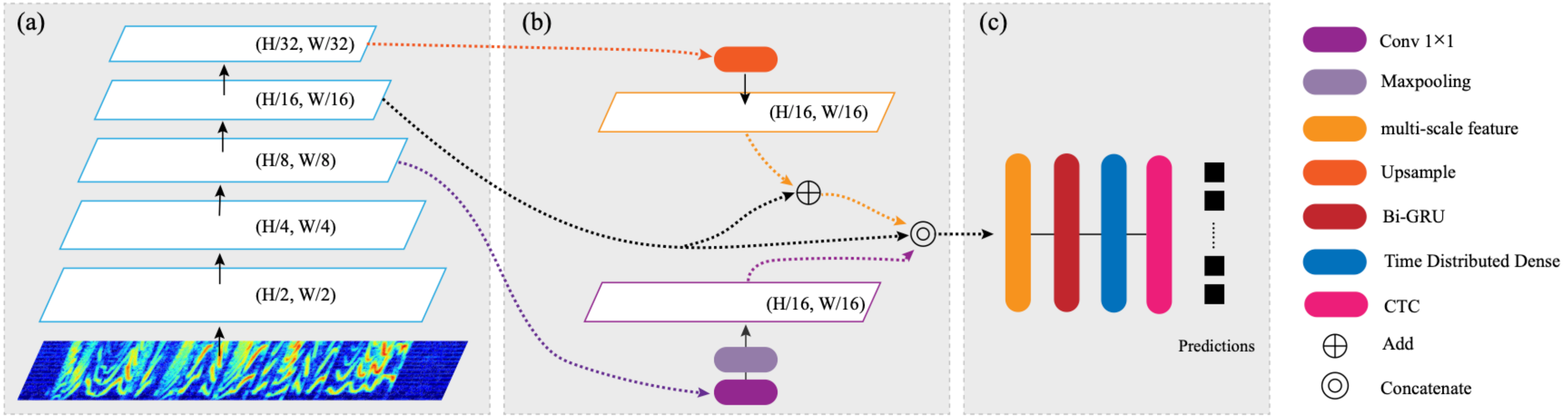

Figure 2: *Model architecture of our method. (a) Bottom-up CNN structure. (b) Merge the deeper branch (high-scale feature representation) and the shallower one (low-scale feature representation) to the standard one. (c) A bi-directional GRU is applied to capture temporal information and a fully-connected layer will predict tone sequences.*

Outline

- Background & Motivation
- Model Design
- **Experiment**
- Results
- Conclusions

Experiment

- Dataset
  - Chinese National Hi-Tech Project 863 corpus
  - Consists of 48373 utterances and total duration is 107 hours..
  - The dataset was divided into training set and test set at the ratio of 9:1. The training set and testing set have not any overlap at the speaker-level and utterance-level.
- Input Features
  - Mel-spectrogram (extracted by tool librosa; 20~8000Hz, frame length 2048(120ms), frameshift of 100 (6ms)., n-fft 2048, Mel bins 512)

Experiment

- Baselines
  - Single scale
  - LSTM, Bi-LSTM and TCN
- Training Configuration
  - learning rate was fixed 0.0001 with total 50 epochs
  - simple greedy decoding
- Evaluation Metrics
  - TER: the average Levenshtein distance between predictions and labels

Outline

- Background & Motivation
- Model Design
- Experiment
- **Results**
- Conclusions

Results

Table 2: *Performance comparison of using various CNN Modules.*

CNN module	Low	Standard	High	TER
Baseline		√		12.53%
ResNet		√		11.55%
ResNet-preact		√		11.38%
Inception-v4-A	√	√	√	16.87%
ResNet	√	√	√	**10.51%**
ResNet-preact	√	√	√	10.79%
Inception	-	-	-	11.35%

Table 3: *Ablation study of our method.*

CNN module	Low	Standard	High	TER
ResNet		√		11.55%
		√	√	10.95%
	√	√		10.74%
	√	√	√	**10.51%**

Results

Table 5: *Performance comparison between feature fusion methods.*

Operation	TER
element-wise addition	11.13%
concatenation	10.51%

Results

- Error Analysis

Table 6: *Breakdowns of errors.*

Method	Insertions	Deletions	Substitutions
Baseline	**79**	1207	7479
Standard	187	752	7211
Multi-Scale	169	**530**	**6696**

Table 7: *Pre-tone accuracy.*

Method	Tone 0	Tone 1	Tone 2	Tone 3	Tone 4
Baseline	68.8%	93.6%	90.1%	81.5%	94.4%
Proposed	**77.2%**	**94.3%**	**91.6%**	**82.0%**	**95.2%**

Outline

- Background & Motivation
- Model Design
- Experiment
- Results
- **Conclusions**

Conclusions

- Advantages
  - Both the low-resolution branch and the high-resolution one can extract more meaningful features and enrich the standard one. **multi-scale model is necessary.**
- Contribution
  - We hope this work can provide insight for researchers on adopting multi-scale method to tone-related tasks.

Thanks for Listening !