30 August – 3 September 2021
INTERSPEECH 2021
BRNO | CZECHIA
Speech everywhere!

Speech Emotion Recognition with Multi-task Learning

Xingyu Cai, Jiahong Yuan, Renjie Zheng, Liang Huang, Kenneth Church

Baidu Research, USA

Speech Emotion Recognition with Multi-task Learning

- ❖ Speech emotion recognition (SER) detects the speakers' emotion from their speech signals.
- ❖ It is often treated as a classification task, with labels like *Happy*, *Angry*, *Sad* and *Neutral*.

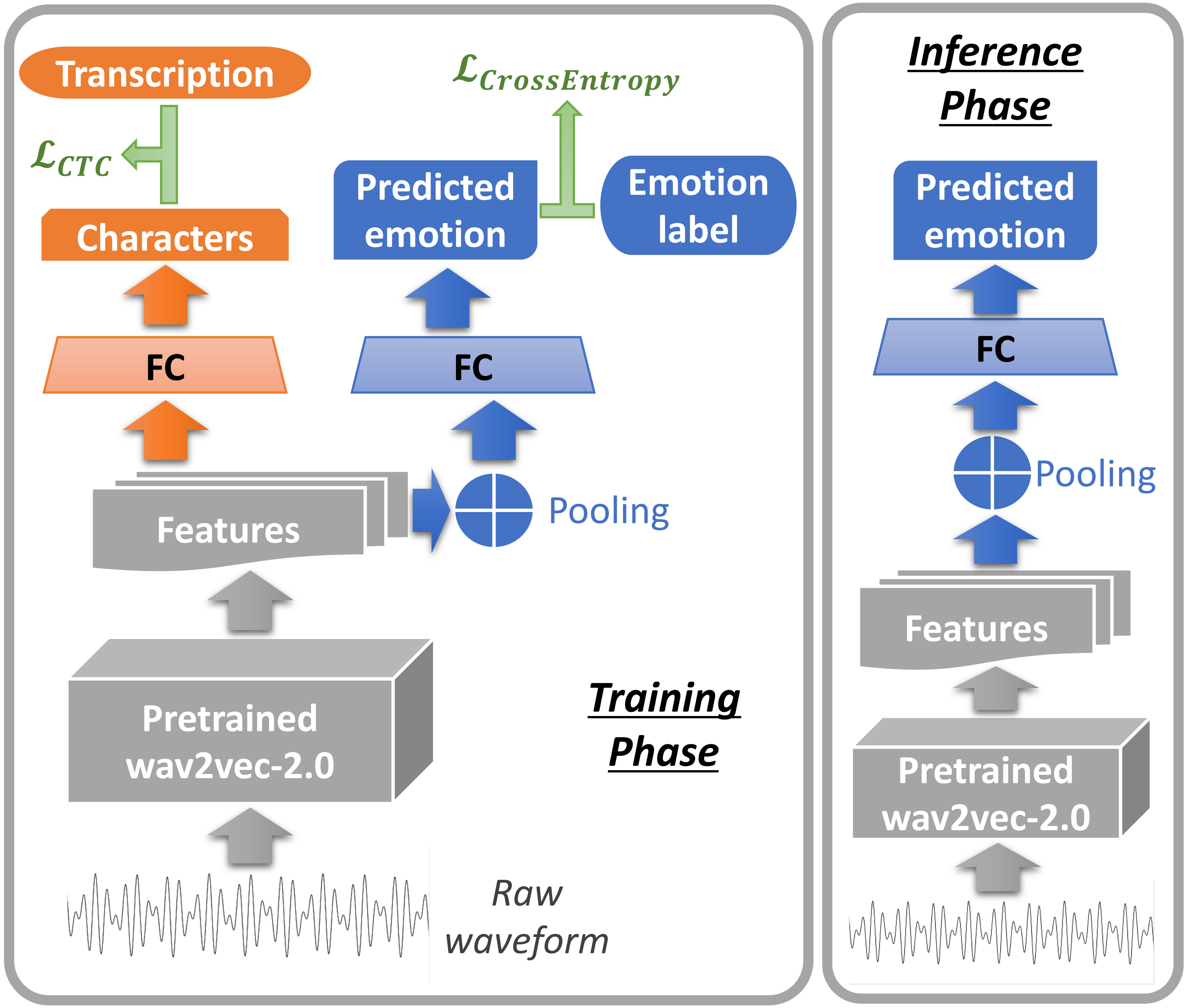
- ➢ Multi-task learning (MTL) simultaneously optimize multiple objectives in different tasks, using a shared backbone model.
- ➢ MTL is widely adopted in ASR, TTS, language model training, etc. It is also related to transfer-learning and continuous learning.

Our contributions:

1. We build an end-to-end model that achieves the state-of-the- art SER results on the standard IEMOCAP dataset.
2. We leverage the pretrained wav2vec-2.0 for speech feature extraction, and fine-tune on SER data through two tasks: SER (emotion classification) and ASR (speech recognition).
3. Ablation study verifies the effectiveness of the MTL approach, and discusses how the ASR affects the SER.
4. The speech transcription could be obtained as a byproduct.

Speech Emotion Recognition with Multi-task Learning

30 August – 3 September 2021
INTERSPEECH 2021 BRNO | CZECHIA
Speech everywhere!



Model Architecture:
(We use Wav2vec2.0 as the feature extractor)

1. Training phase: Two tasks are represented using orange and blue paths.
   - Orange: CTC loss training for text recognition
   - Blue: Cross-entropy loss training for emotion classification
   - $L = L_{CE} + \alpha \times L_{CTC}$
2. Inference phase: Only blue path is kept

Speech Emotion Recognition with Multi-task Learning

Method	Description	Year
Wu et al. [37]	capsule network	2019
Sajjad et al. [13]	ResNet-101 + bi-LSTM	2020
Lu et al. [35]	pretrained ASR + bi-LSTM + attention	2020
Liu et al. [38]	local + global representation learning	2020
Wang et al. [39]	Dual-Sequence LSTM	2020
Pappagari et al. [40]	ResNet based x-vector model	2020
Peng et al. [14]	3D convolution + ASRNN	2020

List of Recent Baselines

Baidu USA

Table 3: *Speech emotion recognition (SER) results.*

method	cross-validation	acc
Wu et al. [37]	10-fold	72.73%
Sajjad et al. [13]	5-fold	72.25%
Lu et al. [35]	10-fold	72.6%
Liu et al. [38]	5-fold	70.78%
Wang et al. [39]	5-fold	73.3%
Pappagari et al. [40]	5-fold	70.30%
Peng et al. [14]	5-fold	62.6%
ours	10-fold	**78.15%**

Main Result:
We achieve best classification accuracy on IEMOCAP compared to other recent baselines

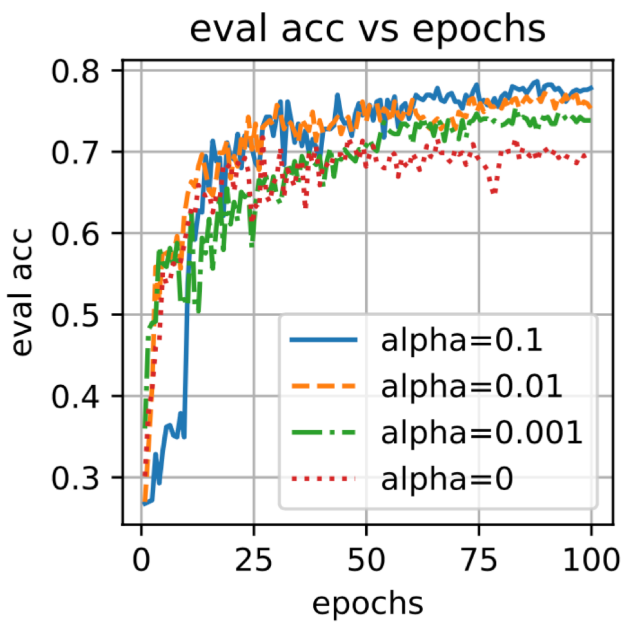
Speech Emotion Recognition with Multi-task Learning

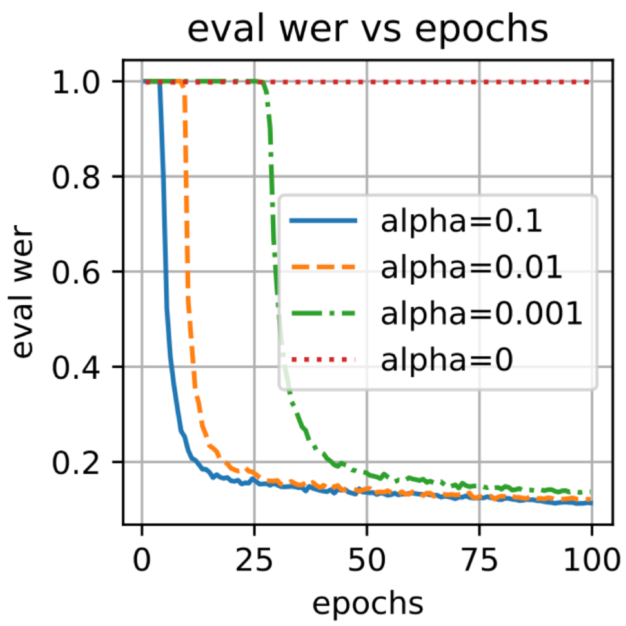
Ablation Study:

- The table shows: α plays an important role. $\alpha = 0$ means emotion classification task only. This leads to poor performance.
- The two figure shows: when α is strong (e.g. $\alpha = 0.1$), the emotion classification accuracy converges slower than others in the beginning phases (the blue curve in the left plot). However, after the word-err-rate converges (at around epoch 12), the accuracy climbs quickly and outperforms others in the end.

This verifies the effectiveness of MTL.

	acc	wer
$\alpha = 0$	71.66%	0.9981
$\alpha = 0.001$	73.97%	0.2233
$\alpha = 0.01$	76.34%	0.2007
$\alpha = 0.1$	**78.15%**	0.1929
$\alpha = 1$	77.35%	0.1877





Thank You!

Our code is available at:
https://github.com/TideDancer/interspeech21_emotion