SJTU SpeechLab E2E ASR System for the ASRU2019 Code-Switching Challenge

卢怡宙，李豪，黄明坤，钱彦旻

{luyizhou4, lh575526, mingkunhuang, yanminqian}@sjtu.edu.cn

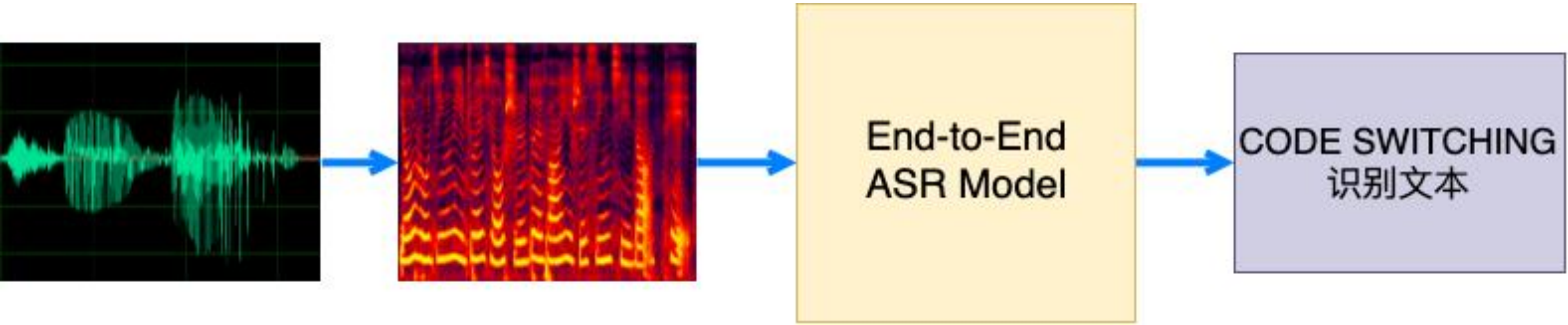
Speech Lab,

Shanghai Jiao Tong University

Outline

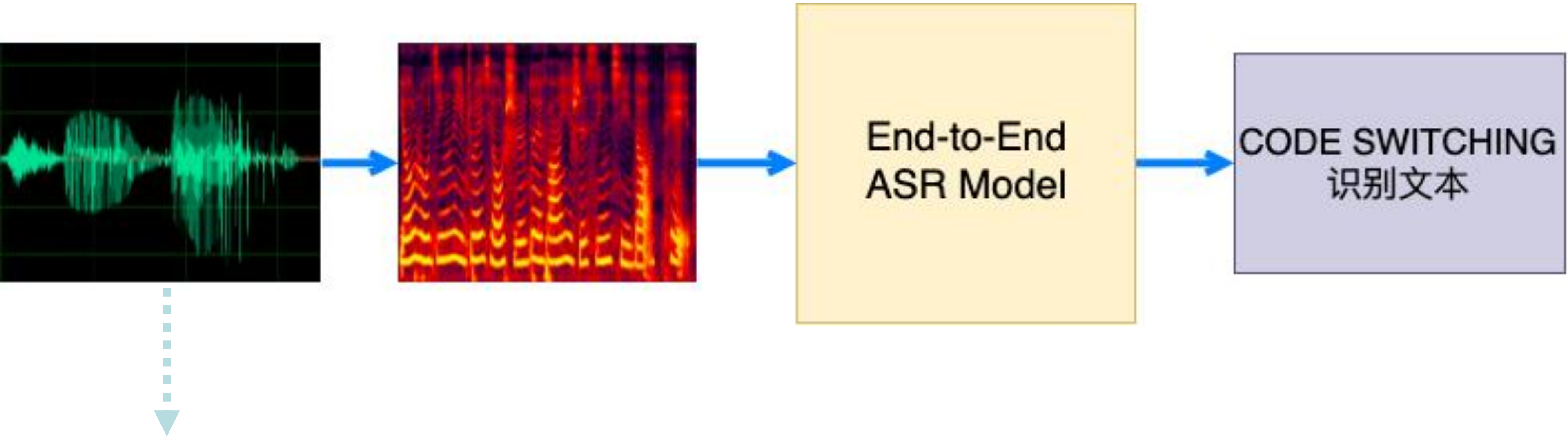
- Brief Introduction:
  - Overall System
  - Data Preparation
- E2E ASR System
  - Model Structure
  - Data Augmentation & Training Strategies
  - Leverage Monolingual & Code-Switching Data
- Experiments
  - LSTM VS Transformer
  - Leverage Monolingual & Code-Switching Data
  - Data Augmentation
  - Summary

Brief Introduction: Overall System



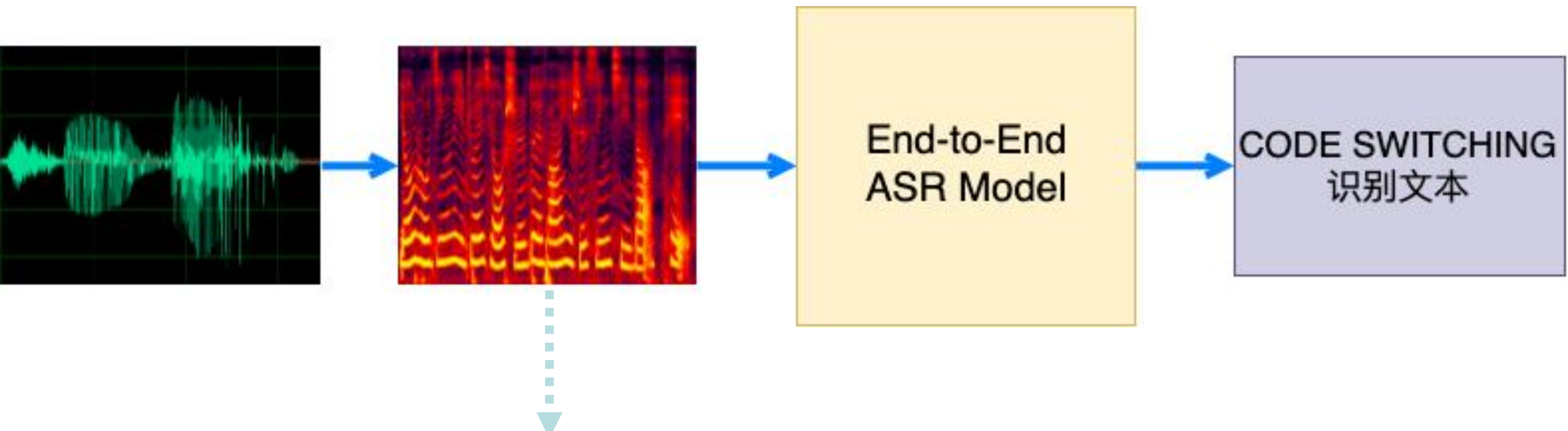
Track 3 E2E-ASR: Overall System

Data Preparation



16kHz wav文件：
500h 中文数据、960h Librispeech数据、 200h 中英混合数据
20h 中英混合的dev集(新、旧)
20h 中英混合的测试集

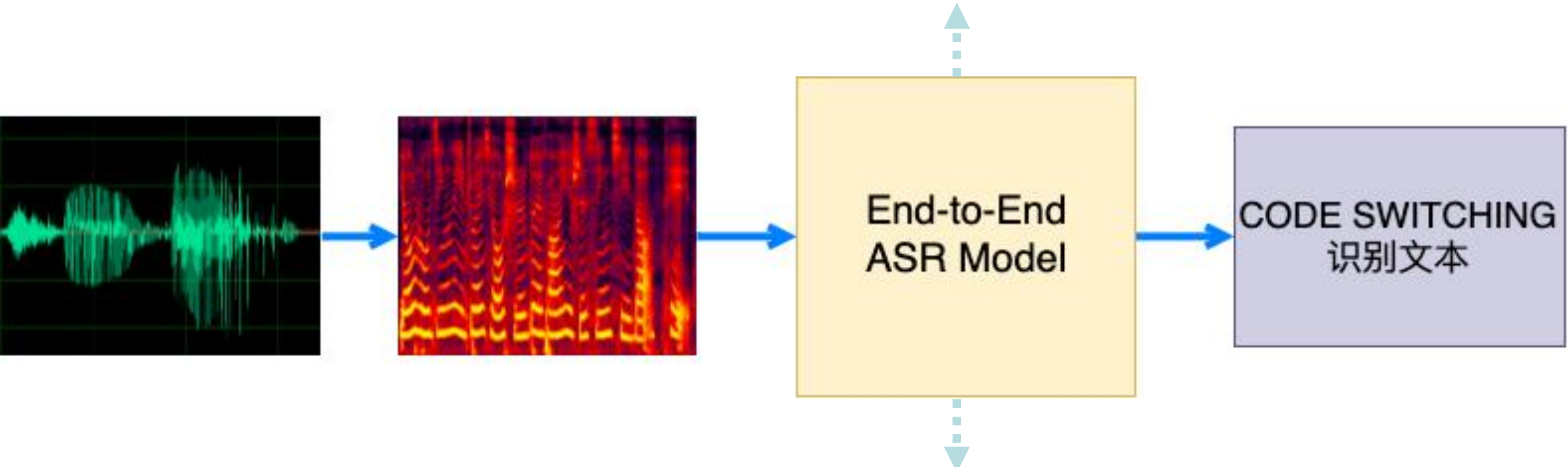
Data Preparation



80维logmel filter bank特征,
Per Speaker CMVN / Per Utterance CMVN

E2E-ASR Model

Sequence-to-Sequence Modeling Task:
Alignment & Classification

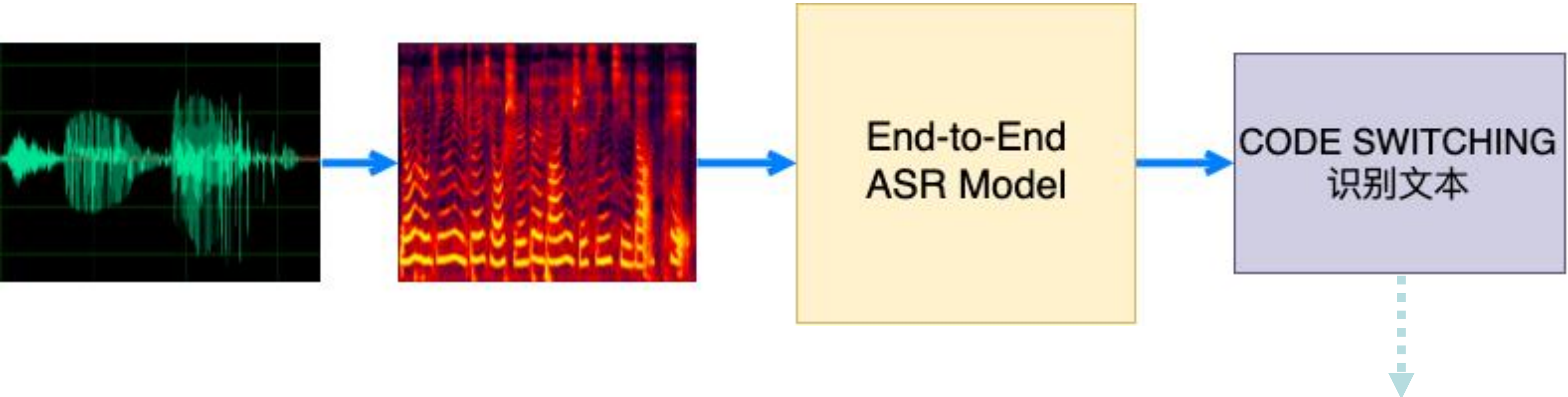


SAN based Joint CTC/Attention Model[1]

$$L_{MTL} = -\lambda log P_{ctc}(C|X) - (1-\lambda) log P_{att}(C|X)$$

$$\hat{C} = \arg\max_{C \in U^*} \{\lambda \, log P_{ctc}(C|X) + (1-\lambda) log P_{att}(C|X)\}$$

Data Preparation



中文用单字建模，词频>=25，3006个字+UNK；
英文用BPE建模，1k个建模单元；
Universal character set；
例：_CO DE _SW IT CH ING 识 别 文 本

Outline

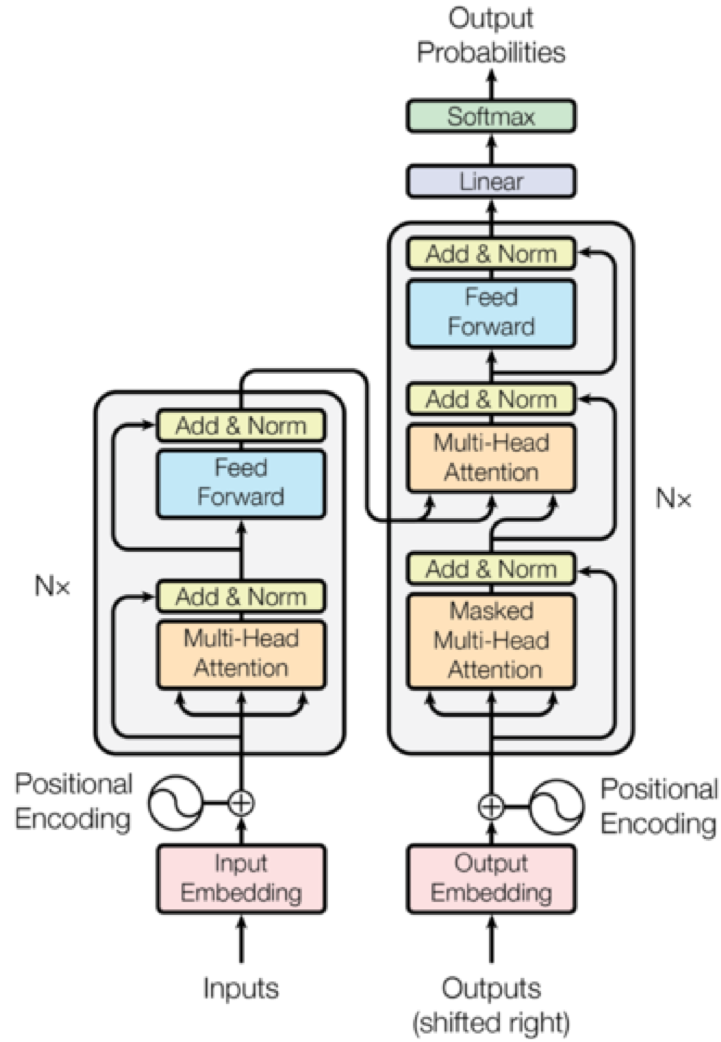
- Brief Introduction:
  - Overall System
  - Data Preparation
- E2E ASR System
  - Model Structure
  - Data Augmentation & Training Strategies
  - Leverage Monolingual & Code-Switching Data
- Experiments
  - LSTM VS Transformer
  - Leverage Monolingual & Code-Switching Data
  - Data Augmentation
  - Summary

E2E ASR System

- Transformer[2]结构
  - 12层Encoder，6层Decoder
  - Convolutional DownSampling
  - Pre Layer Normalization

$$x_{l+1} = LN(x_l + \mathcal{F}(x_l;\theta_l))$$

$$x_{l+1} = x_l + \mathcal{F}(LN(x_l);\theta_l)$$



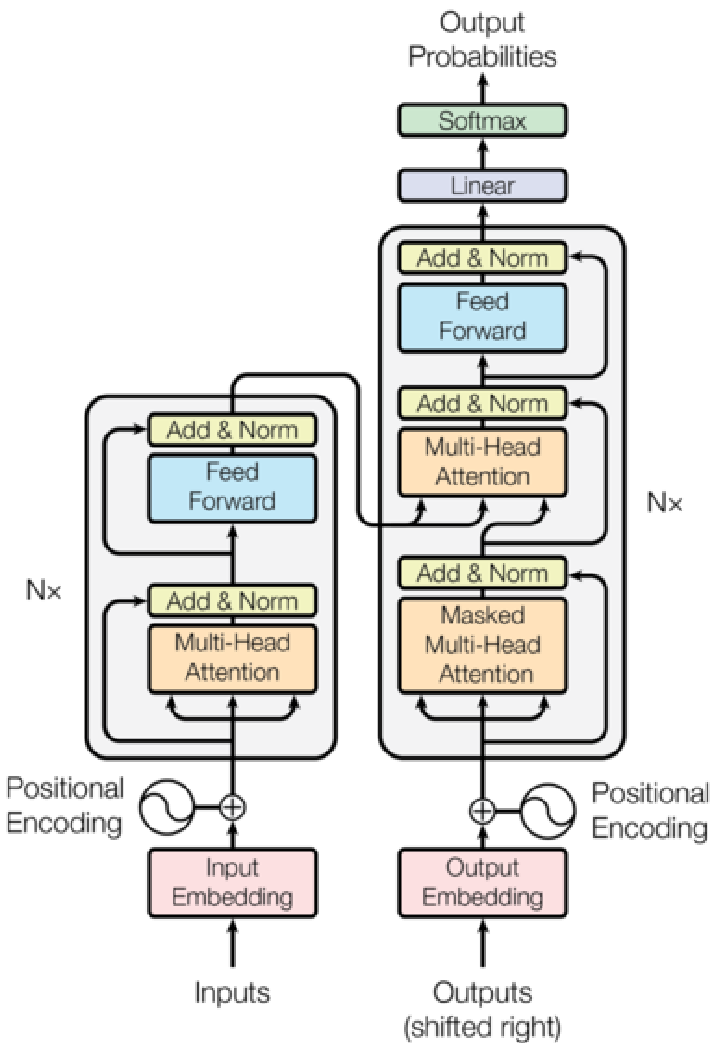
Transformer Structure in [2]

E2E ASR System

- Joint CTC/Attention Model[1]
  - Training λ=0.3
  - Decoding λ=0.4

$$L_{MTL} = -\lambda logP_{ctc}(C|X) - (1-\lambda)logP_{att}(C|X)$$

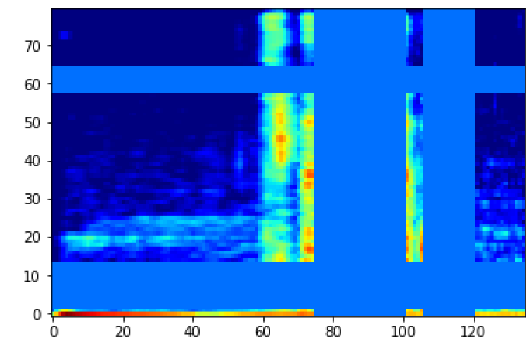
$$\hat{C} = \arg\max_{C\in U^*}\{\lambda\, logP_{ctc}(C|X) + (1-\lambda)logP_{att}(C|X)\}$$



Transformer Structure in [2]

Data Augmentation

- Task-Independent augmentation
  - **Specaugmentation[3]**
  - Speed Perturbation (0.9x, 1.0x, 1.1x speed)
  - ~~混响~~
- Task-Related augmentation
  - ~~**code-switching**数据风格数据（裁剪、拼接）~~



Specaugmentation[3] example

Training Strategies

- Training Strategies
  - Uniform Label Smoothing
  - Gradient Clipping
  - Large Batch Size to Stabilize Training
  - Average Checkpoint
  - Learning Rate Schedule (warmup and decay)

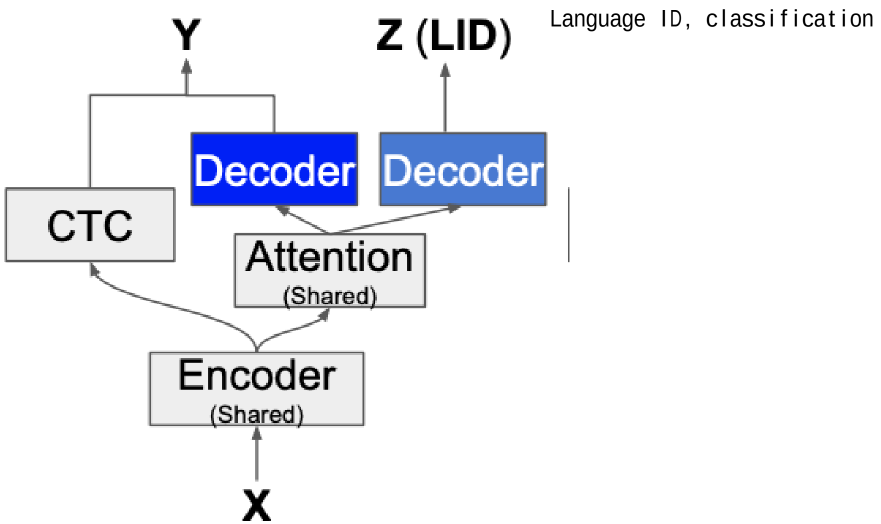
$$lrate = d_{\text{model}}^{-0.5} \cdot \min(step\_num^{-0.5}, step\_num \cdot warmup\_steps^{-1.5})$$

Leverage Monolingual & Code-Switching Data

- Transfer Learning[5]：
  - Step1：预训练Librispeech模型M1
  - Step2：M1模型作为初始化，拿全部中文+**部分英文数据**+全部中英文混合数据训练模型M2
  - Step3：用M2模型在200h中英文混合数据上fine tune

Leverage Monolingual & Code-Switching Data

- LID Multitask[5,6]
  - Language ID Multitask Training Using Attention Context Vector



LID-MTL[6] example

Outline

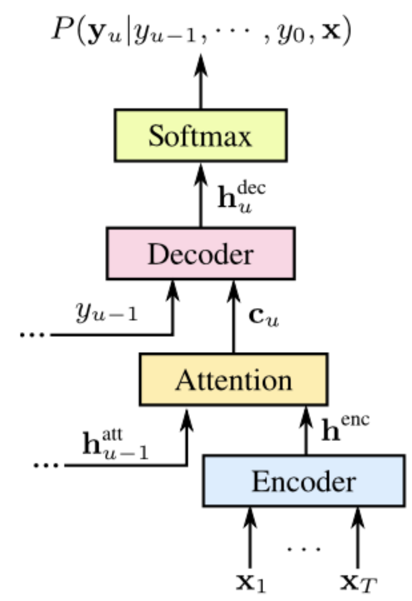
- Brief Introduction:
  - Overall System
  - Data Preparation
- E2E ASR System
  - Model Structure
  - Data Augmentation & Training Strategies
  - Leverage Monolingual & Code-Switching Data
- Experiments
  - LSTM VS Transformer
  - Leverage Monolingual & Code-Switching Data
  - Data Augmentation
  - Summary

Experiments: LSTM VS Transformer

我们在比赛中尝试了两种结构：LSTM、Transformer…

- LSTM Based LAS[4]结构
  - Encoder：VGGBN + 5层512BLSTM (4x sampling)
  - Decoder：2层1024LSTM
  - Attention: Location-Aware Attention

$$P(\mathbf{y}|\mathbf{x}) = \prod_u P(y_u|\mathbf{h}, \mathbf{y}_{<u})$$



S2S Model Structure[7]

- 300h Switchboard

Model	SWBD/CALLHM
Our VGGBLSTM*	9.2/18.2
ESPnet Transformer	9.0/18.1

Experiments: LSTM VS Transformer

Code-Switching Task (mix200h raw, specaug, speed perturb):

Model	**DevSet_prev**
VGGBLSTM	6.22/18.96~(8.17)
+ large bs, Noam schedule	**5.68/17.01~(7.41)**
Transformer Base	**4.76/15.34~(6.38)**

Transformer wins， but…（model size、hyper-parameters）

Leverage Monolingual & Code-Switching data

- 我们发现把全部中文、英文、中英混合数据放一起训练，并没有得到很好的效果 (在新的dev集合测试)

Model	Data	Dev MER
Transformer Big	Mix200	9.90/30.86~(12.21)
Transformer Big	all data	9.53/29.10~(11.68)

可能的原因：

a) Librispeech数据都是长语音(10几秒)，而中文/中英文混合数据都是短语音

b) 中式英语和美式英语的差异

Leverage Monolingual & Code-Switching data

Tricks in Our Transfer Learning experiments:

- Librispeech数据没利用上 ==> 丢掉过长的librispeech数据
- 丢掉过长的句子又没法全部利用上lib数据 ==> librispeech pretrain后初始化
- 同样的结论在LSTM结构下也适用
- Adaptation？

Model	Dev MER
Transformer Big	9.53/29.10~(11.68)
+ 丢掉过长的lib数据	**7.56**/27.20~(9.73)
++ librispeech pre-train	7.68/**25.60~(9.66)**

上海交通大学 SHANGHAI JIAO TONG UNIVERSITY SJTU SPEECH LAB 上海交通大学智能语音实验室

Leverage Monolingual & Code-Switching data

i-vector (mix200h fine tune)

Model	DevSet_Prev
VGGBLSTM	**4.60/15.32~(6.24)**
+ ivector	4.62/15.01~(6.21)
++ FiLM ivector	**4.61/15.03~(6.20)**

Speaker Adaptation：只在mix200h上测试i-vector实验，当时没做进一步实验（全部数据上的speaker adaptation）
Dataset Adaptation：Domain Adversarial Training？

Leverage Monolingual & Code-Switching data

LID-MTL (mix200h fine tune)

Model	DevSet_Prev
VGGBLSTM	**4.79/16.08~(6.52)**
+ LID-MTL	**4.73/15.81~(6.43)**

LID-MTL只有很小的提升

Experiments: Data Augmentation

- Specaugment提升很大；
- 做完specaug后，Speed perturb改进有限，加混响没效果；
- 造code-switching数据在mix200训时有效，但是fine tune没做出来

造Code-Switching风格数据

Model	DevSet_New
MSS	10.16/30.93~(12.45)
+ 1x sync_CS	**9.60/29.85~(11.83)**

Experiments: Summary

- Experiment Summary:

Model	Data	Init.	Dev_new MER	Dev MER(Old Dev)
VGGBLSTM	All data*	LSTM librispeech	N/A	4.99/17.86~(6.96)
VGGBLSTM	mix200h	-	N/A	6.22/18.96~(8.17)
+ All data Init.	mix200h	LSTM all_data*	N/A	4.59/15.87~(6.31)
++ LID Multitask	mix200h	LSTM all_data*	**8.96/28.45~(11.11)**	**4.53/15.21~(6.16)**
Transformer Big	mix200h	-	9.90/30.86~(12.21)	4.44/14.77~(6.02)
Transformer Big	All data*	TF_Big librispeech	7.68/25.60~(9.66)	3.86/15.41~(5.63)
Transformer Big	mix200h	TF_Big all_data*	**7.34/27.16~(9.52)**	**3.66/14.91~(5.38)**

测试集结果: 6.93/24.35~(8.82), Track 3第二名:

- Transformer实验由于时间原因没训完，fine tune没调好;
- Language Model (rescore、LM fusion、spell correction)
- Others：
  - data augmentation，LID-MTL，i-vector
  - mWER，DAT…

Reference

[1] Hori T, Watanabe S, Zhang Y, et al. Advances in joint CTC-attention based end-to-end speech recognition with a deep CNN encoder and RNN-LM[J]. arXiv preprint arXiv:1706.02737, 2017.
[2] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Advances in neural information processing systems. 2017: 5998-6008.
[3] Park D S, Chan W, Zhang Y, et al. Specaugment: A simple data augmentation method for automatic speech recognition[J]. arXiv preprint arXiv:1904.08779, 2019.
[4] Chan W, Jaitly N, Le Q V, et al. Listen, attend and spell[J]. arXiv preprint arXiv:1508.01211, 2015.
[5] Shan C, Weng C, Wang G, et al. Investigating End-to-end Speech Recognition for Mandarin-english Code-switching[C]//ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019: 6056-6060.
[6] Zeng Z, Khassanov Y, Pham V T, et al. On the end-to-end solution to mandarin-english code-switching speech recognition[J]. arXiv preprint arXiv:1811.00241, 2018.
[7] Prabhavalkar R, Rao K, Sainath T N, et al. A Comparison of Sequence-to-Sequence Models for Speech Recognition[C]//Interspeech. 2017: 939-943.

Thank you！

Q&A