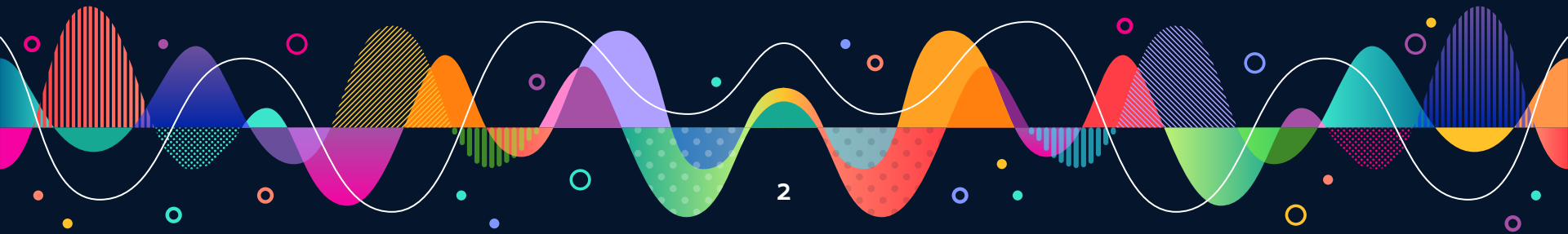


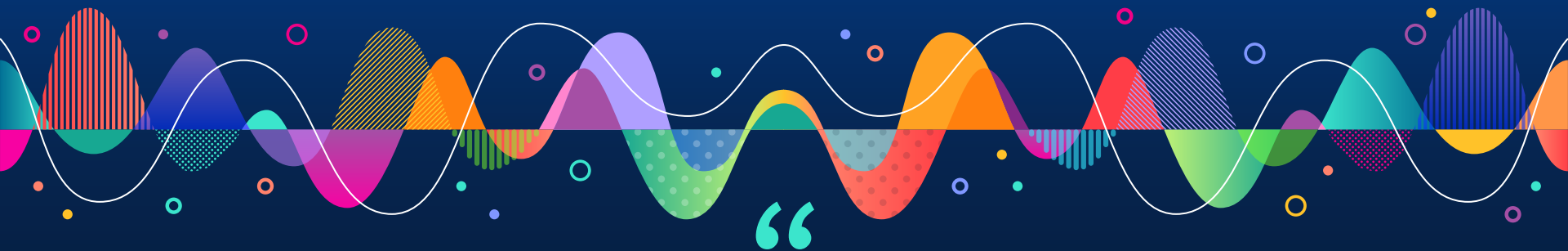
Lost in speech translation: CHN to ENG








Motivation





“

-  Explore the complications of Chinese ASR
-  Improve the quality of machine translation part of the pipeline
-  Get to learn more about Wav2Vec2.0 and ASR workflow

LET'S REVIEW SOME CONCEPTS



ASR

Sentence -> Sequence of words -> break down to words -> turn into arrays -> encoded into model -> understood and converted into texts.

Wav2Vec2

A CNN Framework that is trained with a self-supervised process, learning directly from raw-audio rather than transcribed audio.

Wav2vec 2.0 enables us to build better speech recognition systems for many more languages and domains with much less annotated data.

XLSR

“Cross-lingual speech representations”.

Benefits low resource languages, including Chinese.

XLSR Diagram



Results on the Common Voice benchmark in terms of (PER), comparing training on each language individually (XLSR-Mono) with training on all 10 languages simultaneously (XLSR-10).

PREVIOUS WORK



► Fast Speech-to-Text Modeling with FAIRSEQ

	Fr	De	Es	Zh	Tr	Ar	Sv	Lv	Sl	Ta	Ja	Id	Cy
X→En													
B-Base	23.2	15.7	20.2	4.4	2.2	2.7	1.4	1.2	1.5	0.2	1.1	1.0	1.7
+ SSL [*]	23.1	16.2	20.2	4.8	3.2	3.8	3.7	2.3	2.2	0.2	1.6	1.6	2.2
Multi. B-Big [†]	26.6	19.5	26.3	4.4	2.1	0.3	1.3	0.6	1.4	0.1	0.6	0.3	0.9
T-Sm	26.3	17.1	23.0	5.8	3.6	4.3	2.7	2.5	3.0	0.3	1.5	2.5	2.7
Multi. T-Md [‡]	26.5	17.5	27.0	5.9	2.3	0.4	0.5	0.6	0.7	0.1	0.1	0.3	1.9
En→X													
B-Base	-	12.5	-	20.0	6.7	9.1	18.1	8.7	11.6	7.4	25.6	15.2	18.9
Multi. B-Big [‡]	-	12.6	-	22.2	7.3	8.0	18.3	8.9	11.4	7.3	28.2	16.0	19.3
T-Sm	-	16.3	-	25.4	10.0	12.1	21.8	13.0	16.0	10.9	29.6	20.4	23.9
Multi. T-Md [‡]	-	15.4	-	26.5	9.5	10.8	20.9	12.2	14.6	10.3	30.5	18.9	22.0

Table 5: FAIRSEQ S2T models on CoVoST 2. Test BLEU reported (character-level BLEU for Zh and Ja targets).

^{*} Replaced mel-filter bank features with wav2vec ones (Schneider et al., 2019; Wu et al., 2020). [†] Trained jointly on all 21 X-En directions with temperature-based (T=2) resampling (Arivazhagan et al., 2019a). [‡] Trained jointly on all 15 En-X directions.

PREVIOUS WORK



1. Cantonese CER
2. Chinese Grammatical Errors on CER with BERT
3. Chinese NER character embedding
4. Fine-tuned with Common Voice dataset only
5. MT Zh-En, BLEU 33.52

A black and white photograph of a young boy in profile, shouting or singing into a professional microphone. The microphone is mounted on a stand and has a pop filter. The word "Data" is superimposed in white text over the boy's face. At the bottom of the image, there is a decorative horizontal band featuring various colorful waveforms, including sine waves and complex patterns, with small colored circles scattered around them.

Data

78

Hours of spoken text

3,000+

People from different regions in
China

38%

Of participants between 19-29

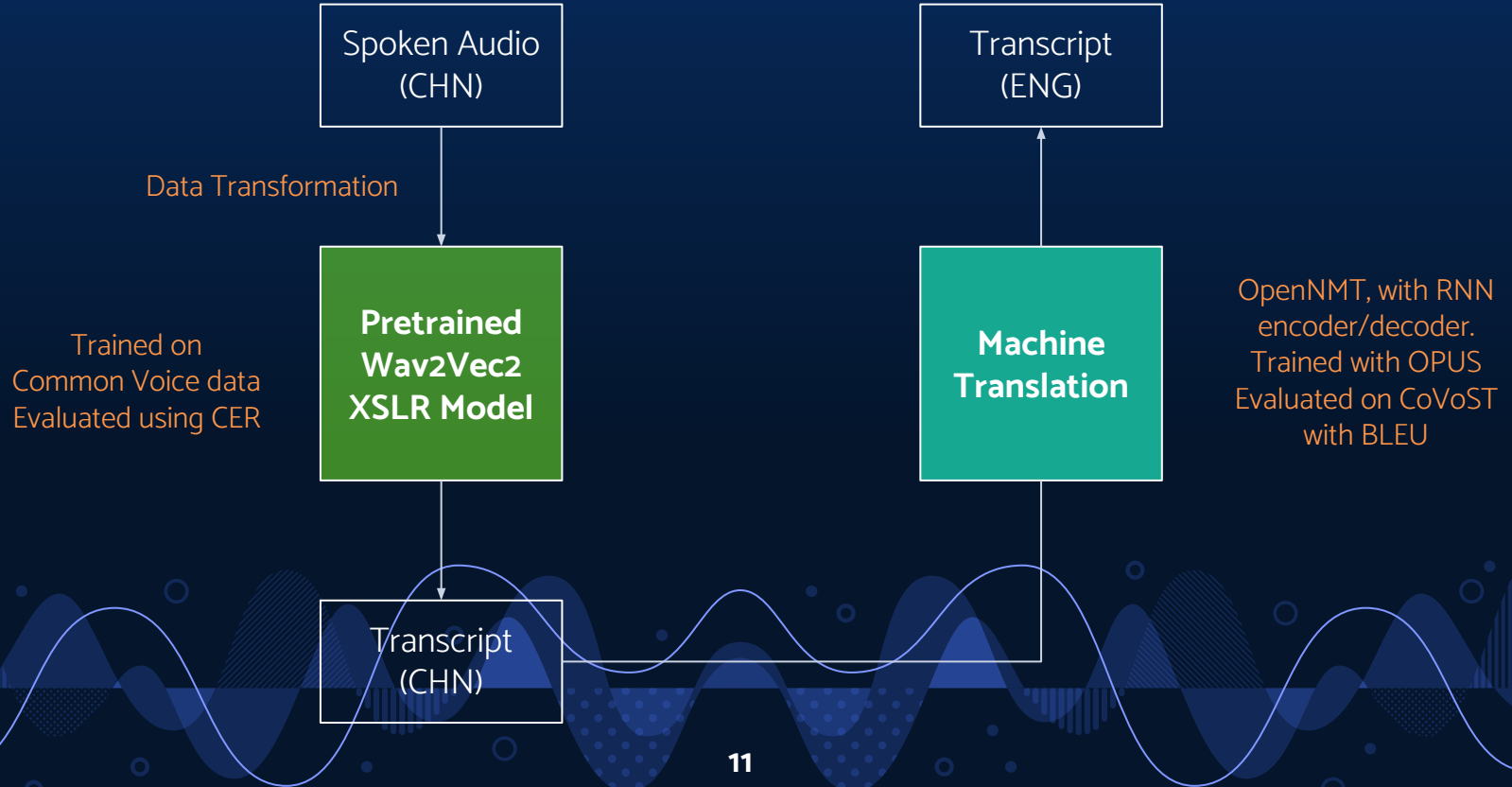
Accent

5% 出生地: 32 江苏省,
5% 出生地: 31 上海市,
4% 出生地: 41 河南省,
4% 出生地: 23 黑龙江省,
4% 出生地: 11 北京市,
3% 出生地: 51 四川省,
3% 出生地: 44 广东省,
3% 出生地: 37 山东省,
3% 出生地: 33 浙江省,
2% 出生地: 50 重庆市,
2% 出生地: 45 广西壮族自治区,
2% 出生地: 42 湖北省,
2% 出生地: 36 江西省,
2% 出生地: 34 安徽省,
2% 出生地: 13 河北省,
1% 出生地: 61 陕西省,
1% 出生地: 46 海南省,
1% 出生地: 43 湖南省,
1% 出生地: 35 福建省,
1% 出生地: 21 辽宁省,
1% 出生地: 14 山西省

OUR DATA

CATEGORY	TRAIN	TEST
ASR - Common Voice	18451 clips	8067 clips
MT - OPUS	119,999 utterances	34,580 utterances

MODEL PIPELINE



EXPERIMENTS & RESULTS

ASR Model

- Since the vocab size for Chinese is very large (4502 in total for train + test sentences), we were facing severe ram & memory issue. Therefore, we had to limit the length of each train clip and try different combinations of hyper-parameters.

Steps	Training Loss	CER
4000	1.482400	0.578819

- We trained our model on about 25% of the original train data, which contains clips under 4 seconds only with batch_size = 2 and gradient accumulation step = 16. The reported CER is 57.88%

Predication	二十一年去世
Reference	二十一年去世

Machine Translation Model

- OpenNMT, with rnn encoder/decoder
- Dropout: 0.2, global_attention: mlp, optim: adam
- Early stopping at 21,500 steps, Validation perplexity: 22.2187, Validation accuracy: 43.5361

```
Step 19600/40000; acc: 49.81; ppl: 11.06;  
Step 19700/40000; acc: 48.61; ppl: 11.72;  
Step 19800/40000; acc: 49.41; ppl: 11.32;  
Step 19900/40000; acc: 49.04; ppl: 11.46;  
Step 20000/40000; acc: 50.62; ppl: 10.57;
```

- Initially tried different steps, dropout rates, and different train/dev splits (90/10 instead of 70/30).
- BLEU: 11.63, 19000 step model.

Predication	So here's my explanation on the left.
Reference	所以在左下角就是我的解释

CONCLUSION & FUTURE WORK



- ▷ ASR Model conclusions
 - ▶ Lots of improvement still could be made
- ▷ Zh-En specific conclusions
 - ▶ Different syntax and grammatical structures is difficult to work with, especially since low amount of Chinese data
- ▷ MT Model conclusions
 - ▶ Room to grow still, especially since MT needs a lot of data to train
- ▷ What we learned from this project, and possible improvements

THANKS!



Any questions?

REFERENCE - Papers

Jia, Bingjing, et al. “Enhanced Character Embedding for Chinese Named Entity Recognition.”

Measurement and Control, vol. 53, no. 9-10, 21 Sept. 2020, pp. 1669–1681,

10.1177/0020294020952456. Accessed 20 Apr. 2021.

Wang, Changhan, et al. *FAIRSEQ S2T: Fast Speech-To-Text Modeling with FAIRSEQ*. , 2020.

Wang, Hongfei, et al. *Chinese Grammatical Correction Using BERT-Based Pre-Trained Model*. , 2020.

Yao, Kaisheng & Cohn, Trevor & Vylomova, Katerina & Duh, Kevin & Dyer, Chris. (2015).

Depth-Gated Recurrent Neural Networks.

Zhou, Shiyu, et al. *A Comparison of Modeling Units in Sequence-To-Sequence Speech Recognition with the Transformer on Mandarin Chinese*. , 18 May 2018.

REFERENCE - Online Resources

“Ctl/Wav2vec2-Large-Xlsr-Cantonese · Hugging Face.” *Huggingface.co*, huggingface.co/ctl/wav2vec2-large-xlsr-cantonese. Accessed 20 Apr. 2021.

“Fine-Tune XLSR-Wav2Vec2 on Turkish ASR with Transformers.ipynb.” *Colab.research.google.com*, 11 Dec. 2020, colab.research.google.com/github/patrickvonplaten/notebooks/blob/master/Fine_Tune_XLSR_Wav2Vec2_on_Turkish_ASR_with_%F0%9F%A4%97_Transformers.ipynb. Accessed 20 Apr. 2021.

“Huggingface/Datasets.” *GitHub*, 20 Apr. 2021, github.com/huggingface/datasets. Accessed 20 Apr. 2021.

“Librosa — Librosa 0.8.0 Documentation.” *Librosa.org*, librosa.org/doc/latest/index.html.

“README.md · Ydshieh/Wav2vec2-Large-Xlsr-53-Chinese-Zh-Cn-Gpt at Main.” *Huggingface.co*, huggingface.co/ydshieh/wav2vec2-large-xlsr-53-chinese-zh-cn-gpt/blob/main/README.md. Accessed 20 Apr. 2021.

“Wav2vec 2.0: Learning the Structure of Speech from Raw Audio.” *Ai.facebook.com*, ai.facebook.com/blog/wav2vec-20-learning-the-structure-of-speech-from-raw-audio/. Accessed 20 Apr. 2021.