BSTC: A Large-Scale Chinese-English Speech Translation Dataset

Ruiqing Zhang, Xiyang Wang, Chuanqiang Zhang, Zhongjun He, Hua Wu, Zhi Li, Haifeng Wang, Ying Chen, Qinfei Li

Baidu Inc.
Automatic Speech Translation

Wireless Speech Translator

Simultaneous Interpretation
# Speech Translation Datasets

| Speech Translation             | Languages | Hours |
|-------------------------------|-----------|-------|
| F-C (2013)                    | Es→En     | 38    |
| KIT-Disfluency (2014)         | De→En     | 13    |
| BTEC (2016)                   | En→Fr     | 17    |
| MSLT V1.0 (2016)              | En↔Fr/De  | 23    |
|                               | En→Zh/Jp  | 6     |
| MSLT V1.1 (2017)              | Zh→En     | 5     |
| Travel (2017)                 | Jp→En     | 9     |
| Aug-LibriSpeech (2018)        | Am→En     | 8     |
| MuST-C (2019)                 | En→8 Euro langs | 3617 |
| Europarl-ST (2020)            | 9 Euro langs | 1642 |
| Covost (2020a; 2020b)         | En↔21 langs | 2880 |

| Simultaneous Translation      | Languages   | Hours  |
|-------------------------------|-------------|--------|
| CIAIR (2004)                  | En↔Jp       | 182    |
| EPPS (2009)                   | En↔Es       | 217    |
| Simul-Trans (2014)            | En↔Jp       | 22     |
| BSTC (ours)                   | Zh→En       | 68     |

30 hours in total
BSTC Datasets

- The first large-scale (68 hours) Chinese-English Speech Translation corpus.

- We organized two simultaneous interpretation tasks in the 1st and 2nd workshop on Automatic Simultaneous Translation.

- BSTC can also be taken as:
  - A Chinese Spelling Correction (CSC) corpus <ASR, Transcription>
  - A Zh->En Document Translation corpus <Document, Translation>
BSTC Training set

(a). Training set

- **Audio**
- **Sentence Splitting**
- **Transcript**: 大家下午好！
  - Duration: From:00:00:00 To:00:00:03
- **5-best ASR**: 大家效果好。大家下午好。大家项目好。大家那个效果好。
- **Translation (1 reference)**

(b). Development set and Test set

| Dataset | Talks | Utterances | Transcription (characters) | Translation (tokens) | Audio (hours) | WER (1-best) |
|---------|-------|------------|----------------------------|----------------------|---------------|--------------|
| Train   | 215   | 37,901     | 1,028,538                  | 606,584              | 64.57         | 27.90%       |
| Dev     | 16    | 956        | 26,059                     | 75,074               | 1.58          | 15.21%       |
| Test    | 6     | 975        | 25,832                     | 70,503               | 1.46          | 10.32%       |
BSTC Dataset

|     | BLEU | AP  | Omissions |
|-----|------|-----|-----------|
| A   | 24.20| 83.0% | 53%       |
| B   | 17.14| 62.8% | 47%       |
| C   | 25.18| 76.5% | 53%       |

Table 4: Comparison of the simultaneous interpretation results of three interpreters (A, B, and C) on the BSTC test set. “AP” is the Acceptability and the “Omissions” indicates the proportion of missing translation in all translation errors.

"translation": "In fact, every one of you has multiple digital devices, handheld devices and mobile phones."

"interpreter A": "But actually you own several devices, mobile devices, mobile phones."

"interpreter B": "But every of you have multiple equipments with you hand held equipment like phone, smartphone."

"interpreter C": "But every one of you have multi devices, we have mobile phones."
Benchmark for Speech Translation

ASR system: Baidu ASR results
MT: Big Transformer

| Systems                        | Test on Transcript |           | Test on ASR   |           |
|-------------------------------|--------------------|-----------|---------------|-----------|
|                               |                    | Dev       | Test          | Dev       | Test      |
| pre-train on WMT              |                    | 20.78     | 35.13         | 18.22     | 33.32     |
| Finetune on <transcript, translation> |                | 23.47 (2.69↑) | 41.14 (6.01↑) | 19.68 (1.46↑) | 35.71 (2.39↑) |
| Finetune on <ASR, translation> |                | 22.53 (1.75↑) | 39.23 (4.1↑)  | 19.82 (1.6↑)  | 36.89 (3.57↑) |
Benchmark for Speech Translation

ASR system: Baidu ASR results
MT: Big Transformer
Policy: Meaningful Unit

Performance on the BSTC testset.
Thank you.