WenetSpeech

A 10000+ hours multi-domain Mandarin corpus for ASR

Chen Chen

2021.10.27

Links

- Homepage

  https://wenet-e2e.github.io/WenetSpeech/

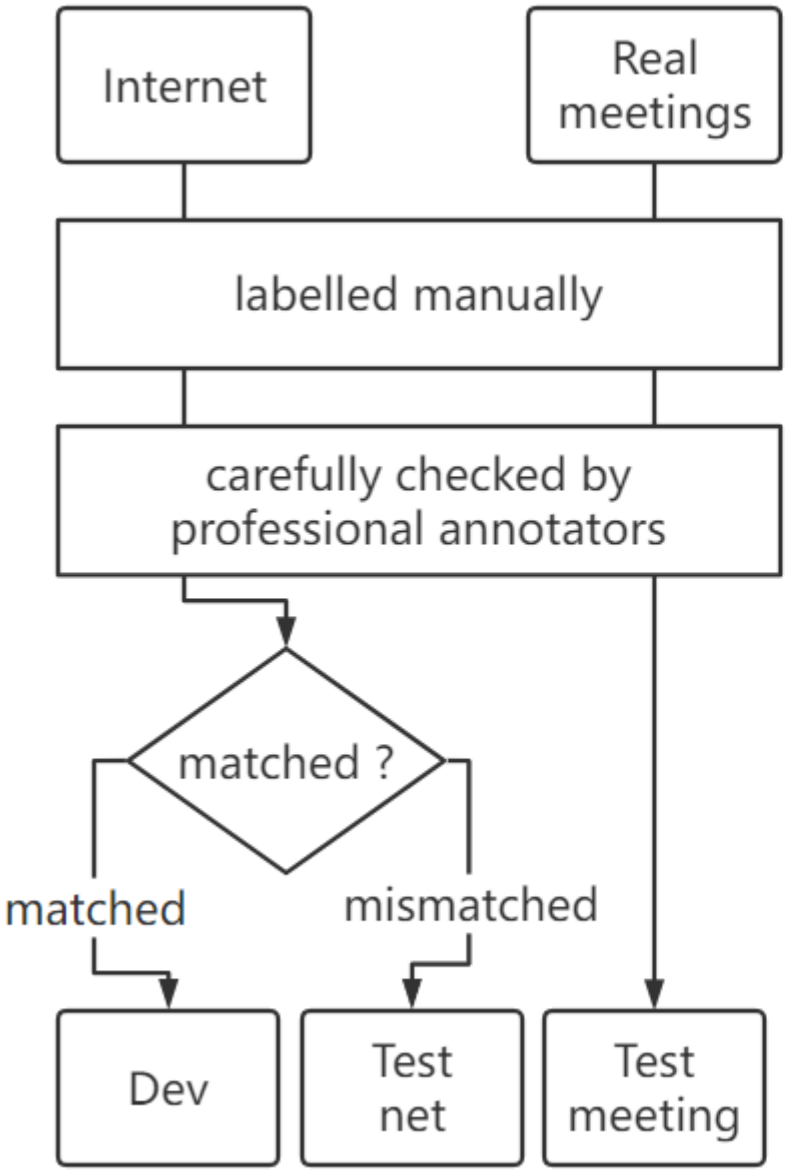
- Code

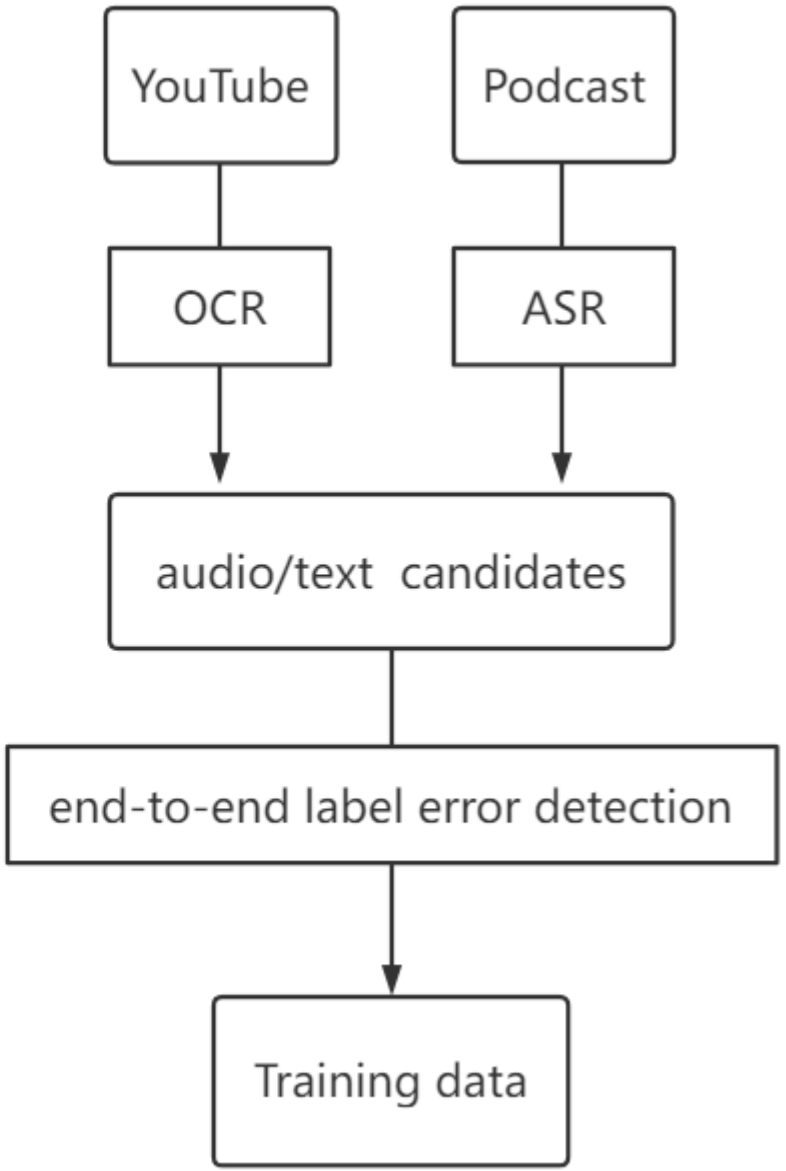
  https://github.com/wenet-e2e/WenetSpeech

- Paper

  https://arxiv.org/pdf/2110.03370.pdf

Source


Internet
Real meetings
labelled manually
carefully checked by professional annotators
matched ?
matched
mismatched
Dev
Test net
Test meeting


YouTube
Podcast
OCR
ASR
audio/text candidates
end-to-end label error detection
Training data

Components

Table 1. WenetSpeech partition

Set	Confidence	Hours
Strong Label	[0.95, 1.00]	10005
Weak Label	[0.60, 0.95)	2478
Others	/	9952
Total(hrs)	/	22435

Table 2. Training data in different domains with duration (hrs)

Domain	Youtube	Podcast	Total
audiobook	0	250.9	250.9
commentary	112.6	135.7	248.3
documentary	386.7	90.5	477.2
drama	4338.2	0	4338.2
interview	324.2	614	938.2
news	0	868	868
reading	0	1110.2	1110.2
talk	204	90.7	294.7
variety	603.3	224.5	827.8
others	144	507.5	651.5
Total	6113	3892	10005

Table 4. The WenetSpeech evaluation sets

Evaluation Sets	Hours	Source
Dev	20	Internet
Test_Net	23	Internet
Test_Meeting	15	Real meeting

contained in Others

Training data

Table 1. WenetSpeech partition

Set	Confidence	Hours
Strong Label	[0.95, 1.00]	10005
Weak Label	[0.60, 0.95)	2478
Others	/	9952
Total(hrs)	/	22435

Table 3. The training data subsets

Training Subsets	Confidence	Hours
L	[0.95, 1.0]	10005
M	1.0	1000
S	1.0	100

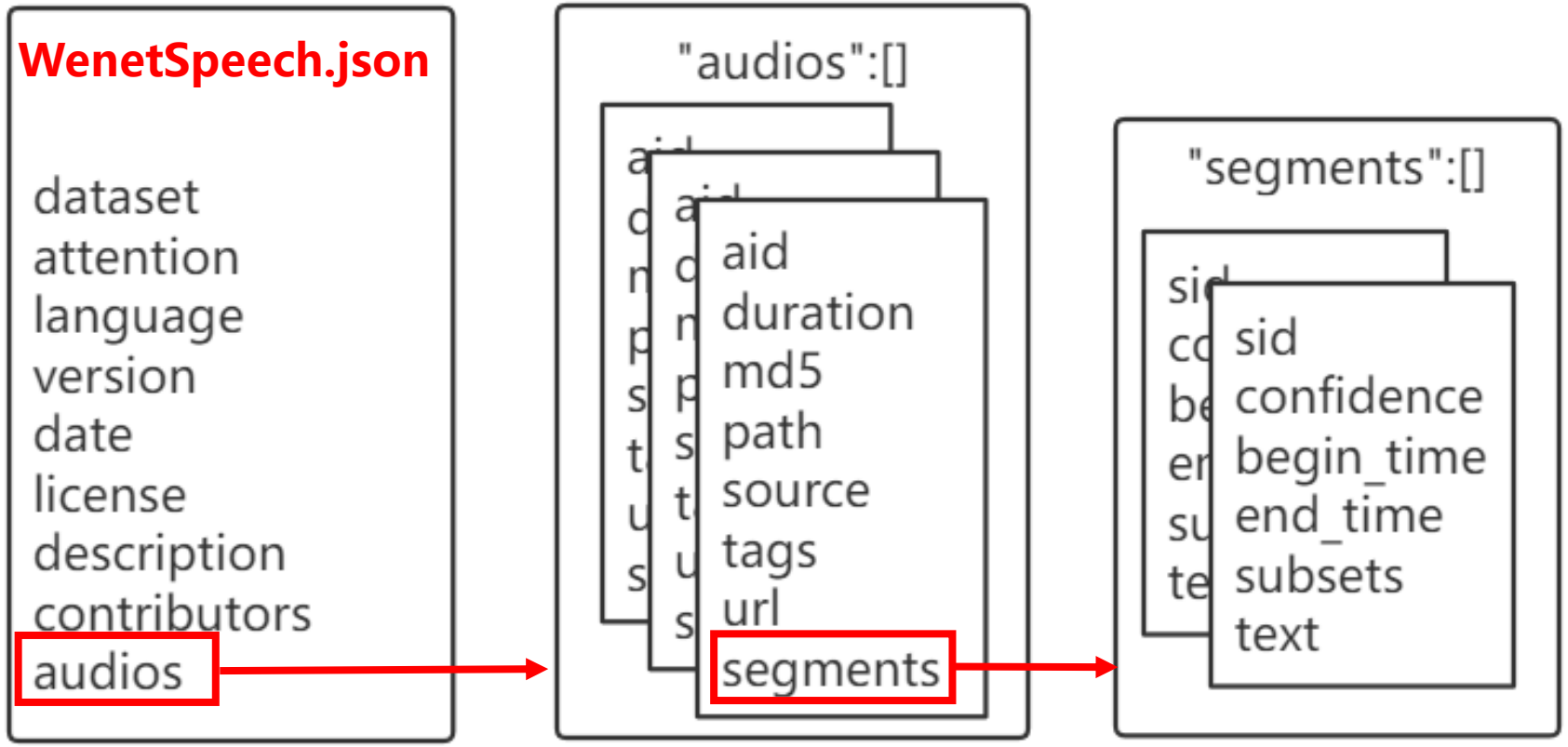
Strong label for supervised

Weak Label for semi-supervised

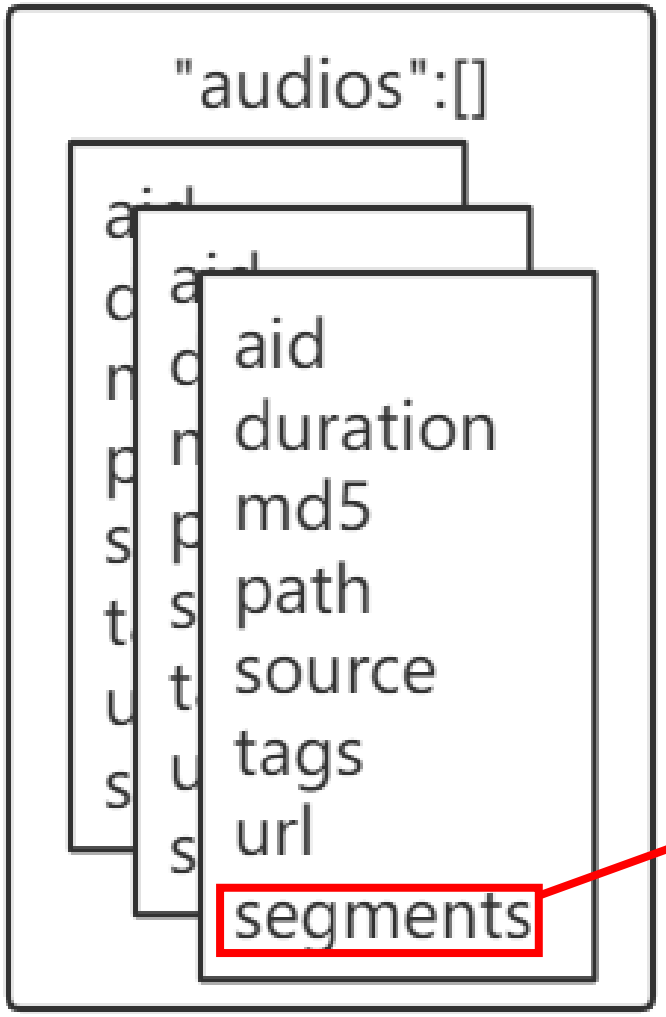
Unlabel for unsupervised

Metedata

A single json file with 0.18 billion lines



Metedata--audios

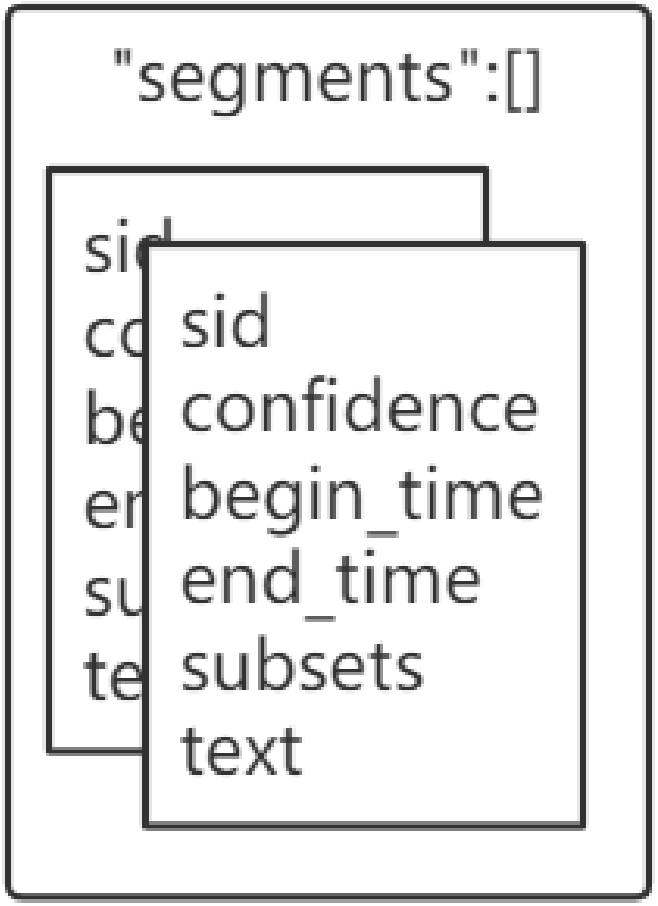


"aid": "Y0000000000_--5llN02F84",
"duration": 2494.57,
"md5": "48af998ec7dab6964386c3522386fa4b",
"path": "audio/train/youtube/B00000/Y0000000000_--5llN02F84.opus",
"source": "youtube",
"tags": [
"drama"
],
"url": "https://www.youtube.com/watch?v=--5llN02F84",
"segments": [

Labelled data segments

Unlabel = audios - segments

Metedata--segments




```
{
    "sid": "Y0000000002_--s1SMM6PBU_S00485",
    "confidence": 1.0,
    "begin_time": 1173.76,
    "end_time": 1175.12,
    "subsets": [
        "L",
        "M",
        "S"
    ],
    "text": "这些日子"
},
```


L
M
S
W
DEV
TEST_NET
TEST_MEETING

Format

- Format

  Opus, a lossy audio coding format

- According to the paper

  16k sampling rate, single-channel, and 16-bit signed-integer format

- Informations from ffprobe

  Bitrate: 33 kb/s

  Audio: opus, 48000 Hz, mono

Benchmark Performance

diversity and reliability challenging

Table 5. Results (MER%) on different test sets for baseline systems trained using WenetSpeech training subset L

Toolkit	Dev	Test_Net	Test_Meeting	AIShell-1
Kaldi	9.07	12.83	24.72	5.41
ESPNet	9.70	8.90	15.90	3.90
WeNet	8.88	9.70	15.59	4.61

Performance ∝ data amount

Table 6. Kaldi baseline results (MER%) for different WenetSpeech training subsets

SubSet	Dev	Test_Net	Test_Meeting	AIShell-1
L	9.07	12.83	24.72	5.41
M	9.81	14.19	28.22	5.93
S	11.70	17.47	37.27	7.66