librispeech-clean-100 raw dataset 정보

In [1]:

```
import pickle
import numpy as np
import matplotlib.pyplot as plt
```

먼저 mapping.pkl 파일에는 character mapping 정보가 들어가 있습니다.

In [2]:

```
with open('mapping.pkl', 'rb') as f:
    mapping = pickle.load(f)
    print(mapping)
```

```
{'<eos>': 1, 'D': 12, 'Q': 28, 'A': 5, 'S': 10, 'O': 6, 'H': 9, 'Z': 29, 'K': 24, 'E': 3, 'V': 23, 'P': 21,
'C': 16, '<sos>': 0, 'I': 8, 'T': 4, 'L': 13, 'M': 15, 'G': 19, ' ': 2, 'X': 26, 'W': 17, 'R': 11, 'U': 14,
'B': 22, 'J': 27, 'Y': 20, "'": 25, 'F': 18, 'N': 7}
```

In [3]:

```
key_list = list(mapping.keys())
key_list.sort()
```

In [4]:

```
for key in key_list:
    print('{}: {}'.format(key, mapping[key]))
```

```
 : 2
': 25
<eos>: 1
<sos>: 0
A: 5
B: 22
C: 16
D: 12
E: 3
F: 18
G: 19
H: 9
I: 8
J: 27
K: 24
L: 13
M: 15
N: 7
O: 6
P: 21
Q: 28
R: 11
S: 10
T: 4
U: 14
V: 23
W: 17
X: 26
Y: 20
Z: 29
```

In [5]:

```
sample_data = np.load('train-clean-100/7859-102521-0017.npy')
```

주어진 data은 librispeech-clean-100 raw 데이터를 spectrogram으로 변환한 data 입니다.

train dataset의 가장 첫 번째 data를 뽑아 보았습니다.

이 데이터는 2453개의 timestep과, 각 timestep 마다 40차원의 filter bank spectra feature 로 이루어져 있습니다.

In [6]:

```
sample_data.shape
```

Out[6]:

(2453, 40)

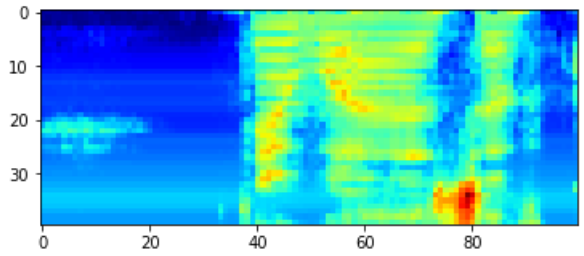
처음 100 timestep의 spectrogram을 그려보면 다음과 같이 나옵니다.

In [7]:

```
plt.imshow(np.transpose(sample_data[:100]), cmap='jet')
```

Out[7]:

<matplotlib.image.AxesImage at 0x7fc78d43f5c0>



각 csv 파일에는 train dataset의 정보와 test dataset의 정보가 들어가 있습니다.

csv 파일의 각 row는 하나의 spectrogram 형태의 음성 파일과 그에 해당하는 transcription 문장의 정보가 나와 있습니다.

지금은 모든 데이터가 따로따로 분리되어 있으므로, LAS 모델 학습을 위해서는 먼저 batch를 만들어주는 작업 부터 하셔야 합니다.

(아시다피시, 길이가 모두 다르므로 batch를 만드실 때는 padding을 해 주어야 합니다.)