How do deep speaker models treat silence and noises

王天昊

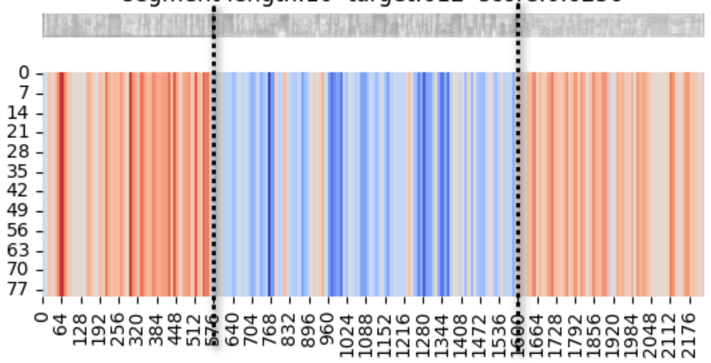
2022/10/14

Background

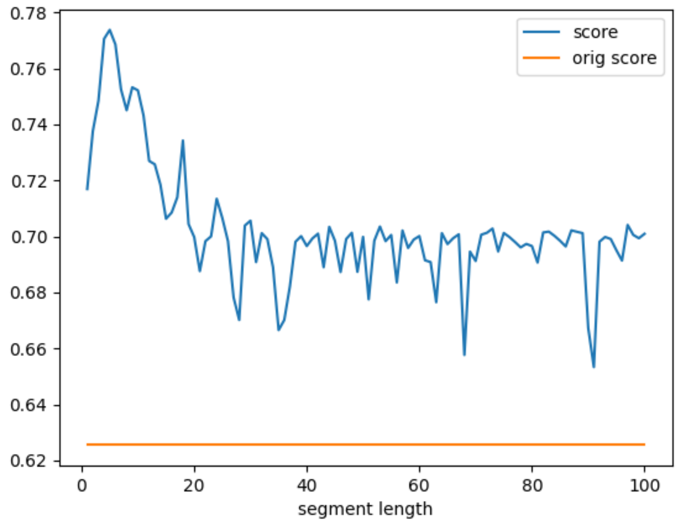
- time masking

multi-speaker(a-b-a, a is target)

segment length:10 target:612 score:0.6256

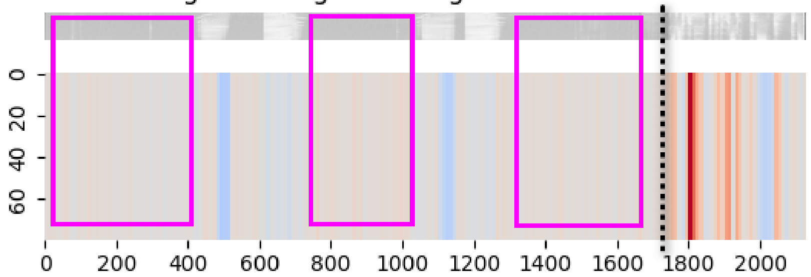


Change the segment length and multiply it to the original mel

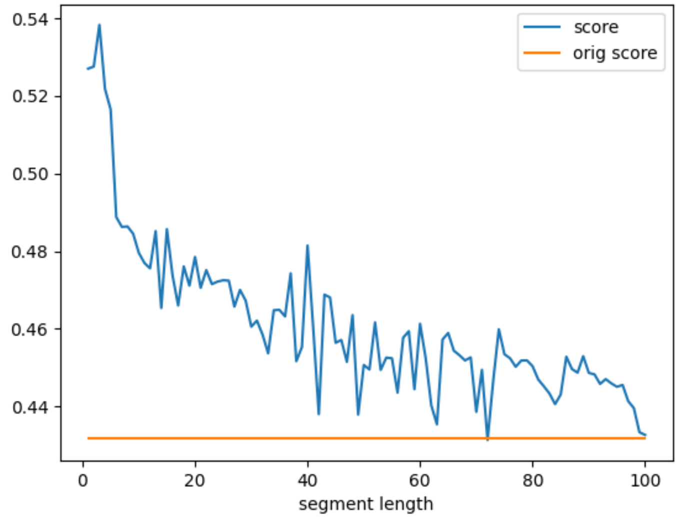


noise with silence splicing speaker's voice

segment length:10 target:0 score:0.4319



Change the segment length and multiply it to the original mel



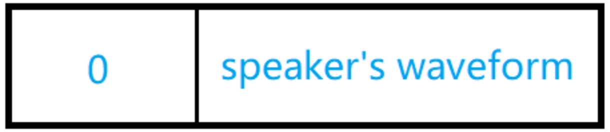
Why the silence segment also gets a positive weight ?

Experiment

- Score based
- Concatenate special audio to speaker's voice
    - special audio:
        - silence
        - white noise
        - music
    - variable:
        - concatenation length
- Tips:
    - *torchaudio.load()* will enable normalize by default (converts the native sample type to float32, set ‘normalize=False’ to disabled it).
    - *scipy.io.wavfile.read()* will keep the sample type of the original file.

Silence

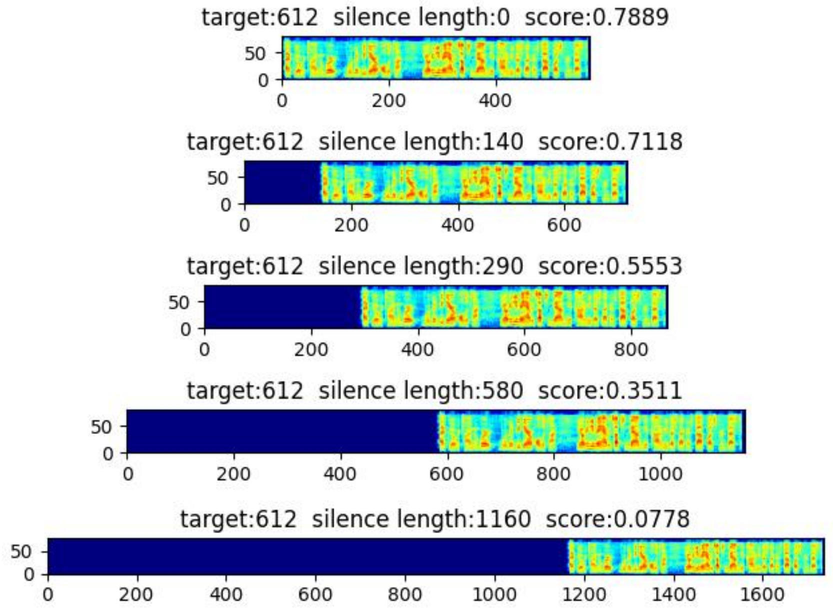
- Concatenate absolute silence with the speaker's waveform.

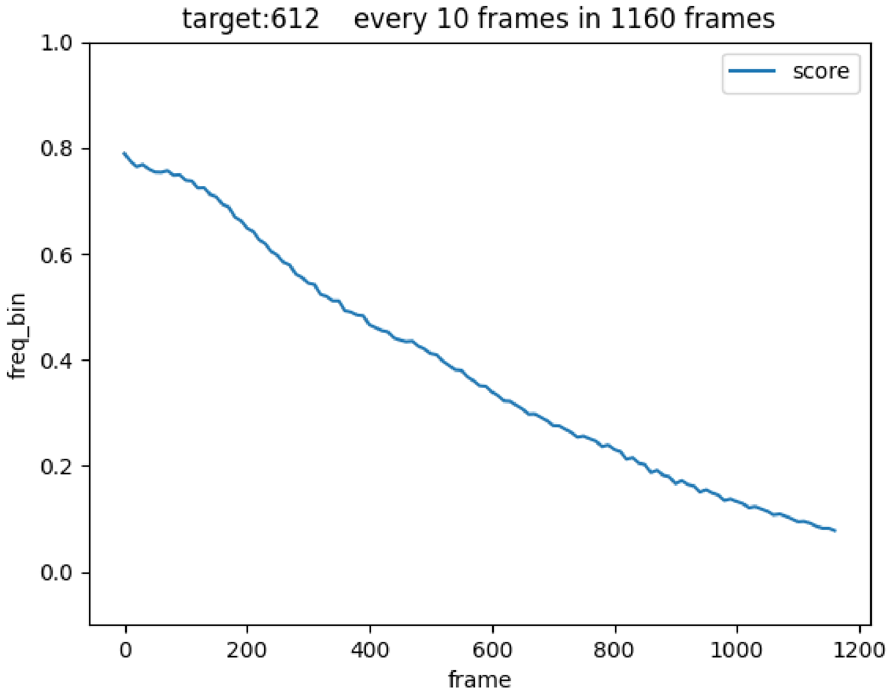


0

speaker's waveform

- Continue to increase the concatenation length, and calculate the score curve
- use ResNet34 TSP model:





- ASP model results are similar to TSP.

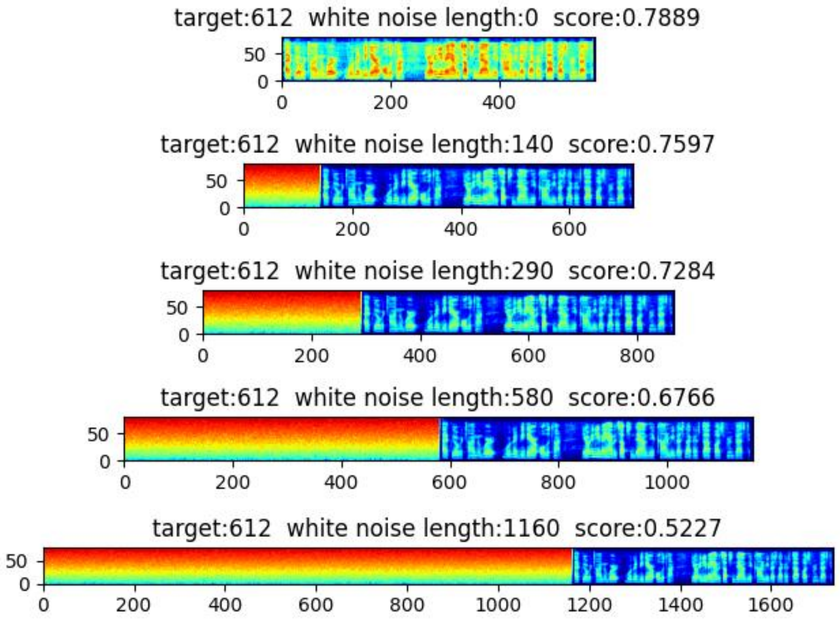
Silence

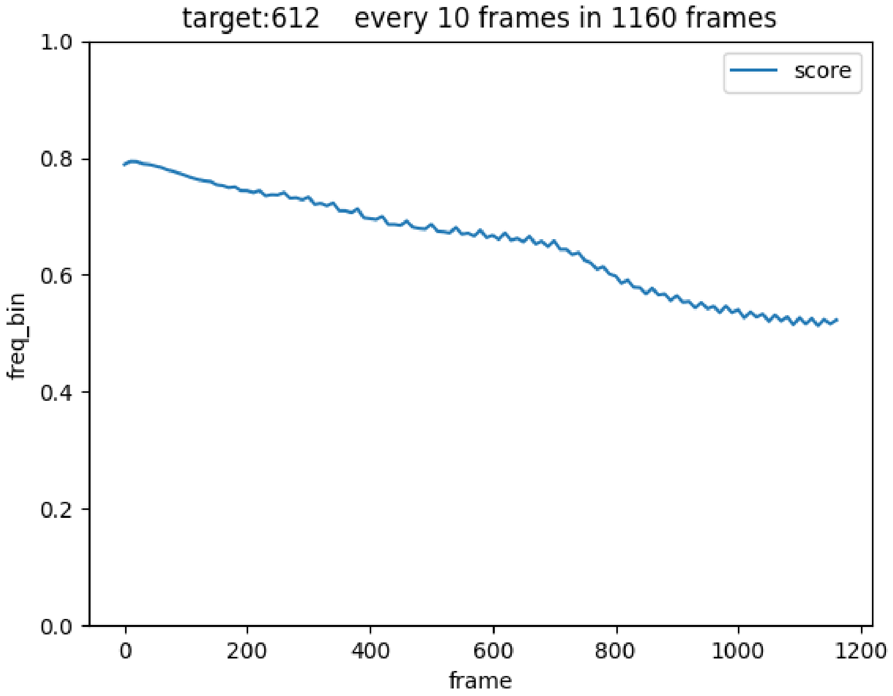
- IDR
- DataSet: 2000 utterances from VoxCeleb2, including 100 speakers

	TSP model IDR	ASP model IDR
original	98.35%	98.20%
concatenate 1/4 length silence	95.70%	94.10%
concatenate 1/2 length silence	55.50%	44.00%
concatenate same length silence	**3.15%**	**0.40%**

White Noise

- Concatenate white noise’s waveform with the speaker's waveform.
- use ResNet34 TSP model:





- ASP model results are similar to TSP.

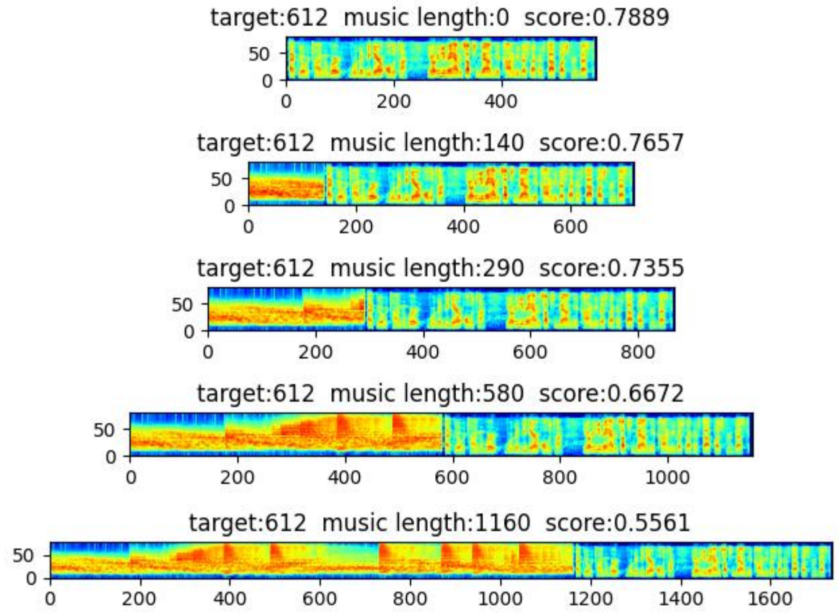
White Noise

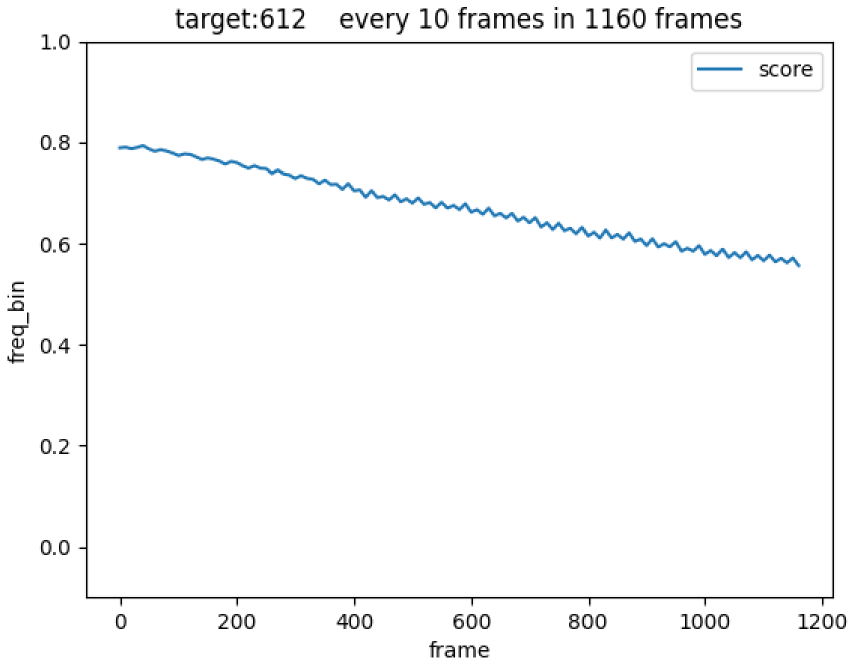
- IDR
- DataSet: 2000 utterances from VoxCeleb2, including 100 speakers

	TSP model IDR	ASP model IDR
original	98.35%	98.20%
concatenate 1/4 length white noise	96.80%	96.35%
concatenate 1/2 length white noise	87.80%	89.30%
concatenate same length white noise	**59.10%**	**73.50%**

Music

- Music DataSet: 5 music selected from the musan dataset
- use ResNet34 TSP model:





- ASP model results are similar to TSP.
- When the concatenation length is relatively long, the difference between the scores of different music is about 0.1.

Music

- IDR
- DataSet: 2000 utterances from VoxCeleb2, including 100 speakers

	TSP model IDR	ASP model IDR
original	98.35%	98.20%
concatenate 1/4 length music	98.15%	97.80%
concatenate 1/2 length music	97.15%	97.25%
concatenate same length music	93.10%	93.90%

Music

- Other style music

	TSP model IDR	ASP model IDR
original	98.35%	98.20%
concatenate 1/4 length music	97.65%	97.50%
concatenate 1/2 length music	93.35%	93.85%
concatenate same length music	75.65%	80.20%

Conclusion

- Influence on model's recognition ability: silence > white noise > music.

- The influence of silence on both the TSP model and the ASP model is very large.

- The ASP model can reduce the influence of white noise and music on the model compared to the TSP model.

More

- Why does silence affect the model so much?