Music Removal by Convolutional Denoising Autoencoder in Speech Recognition

Mengyuan Zhao, Dong Wang, Zhiyong Zhang, Xuewei Zhang

Center for Speech and Language Technologies (CSLT)

Tsinghua University

2015-12-1

Contents

- Music removal based on DAE and CDAE.
- Music removal across languages.
- Experiment result.

Music Removal for ASR

- Why ?
  - Mixing music in speech usually causes significant performance reduction in ASR
- How ?
  - Traditional approaches focus on music/voice separation:
    - Robust PCA
    - non-negative matrix factorization (NMF)
    - Robust NMF

Music Removal for ASR

- Disadvantage of traditional methods
  - Rely on human-discovered music patterns and properties.
  - Have difficulty in dealing with the complexity of music signals of different genres.

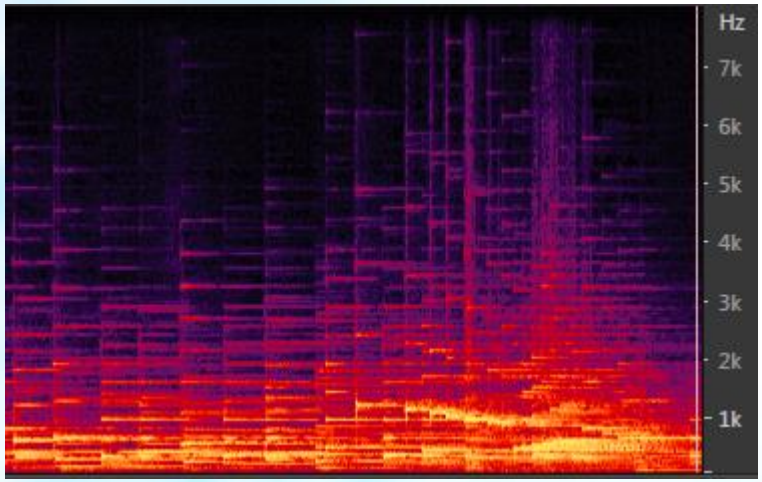

Figure 1 Spectrogram of “normal” music (Chopin Nocturne No.9 Op2)

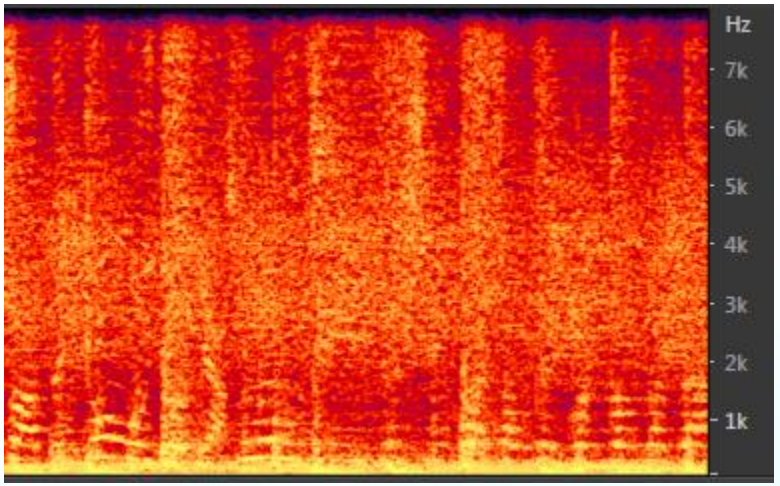

Figure 2 Spectrogram of “abnormal” music (Jay Chow Shuangjiegun)

Denoising Auto Encoder(DAE)

- Learning based approach:
  - Use Denoising Autoencoder (DAE) to learn the patterns from data.

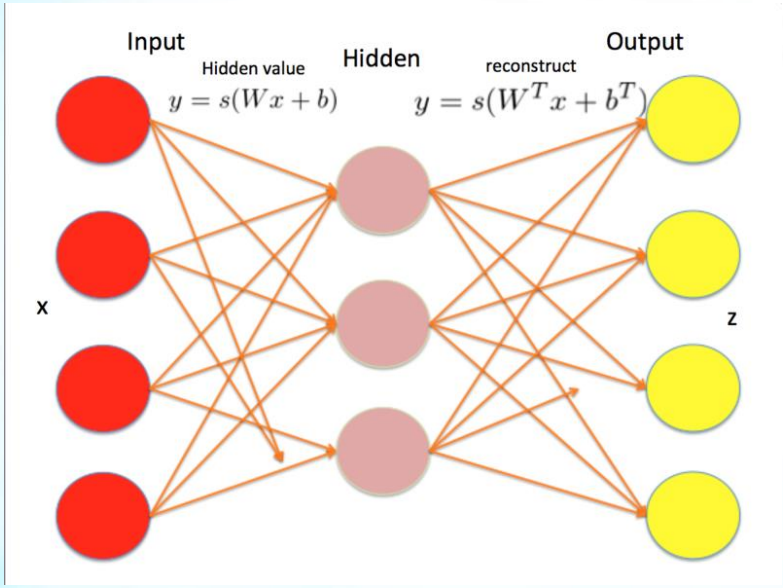

Figure 3 Structure of DAE

Speech recognition system

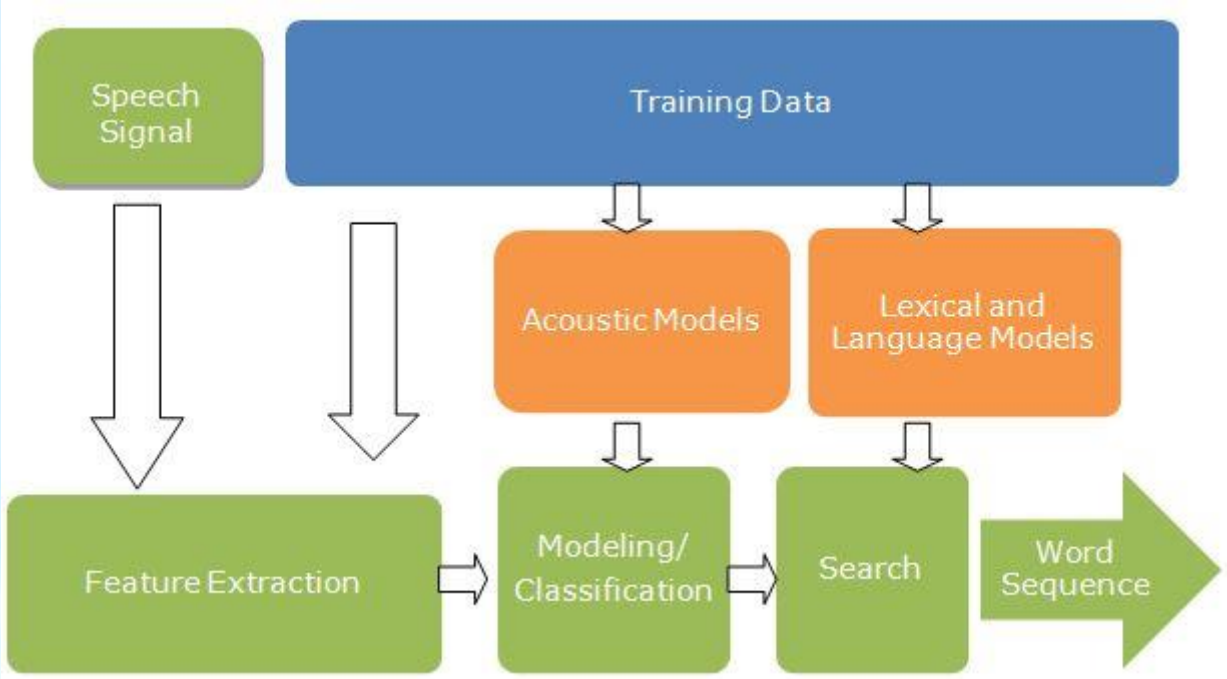

Figure 4 Flow chart of ASR system

Convolutional Denoising Auto Encoder (CDAE)

- In order to utilize prior knowledge of music signals
  - Entropy
  - Repeating patterns
  - Harmonic structures

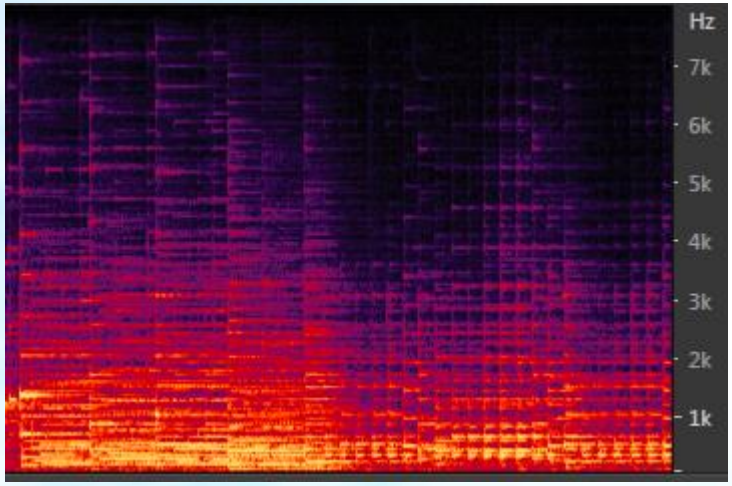

Figure 5 Spectrogram of piano solo (Beethoven Moonlight Chapter 3)

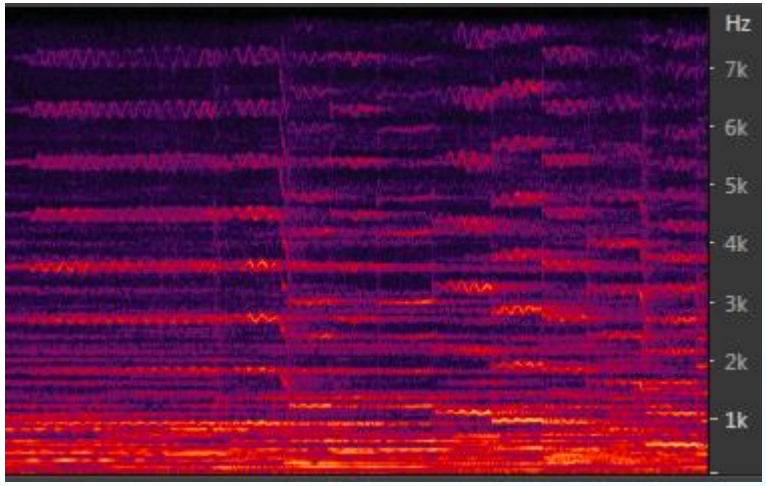

Figure 6 Spectrogram of violin solo (Theme From Schindler's List)

Convolutional Denoising Auto Encoder (CDAE)

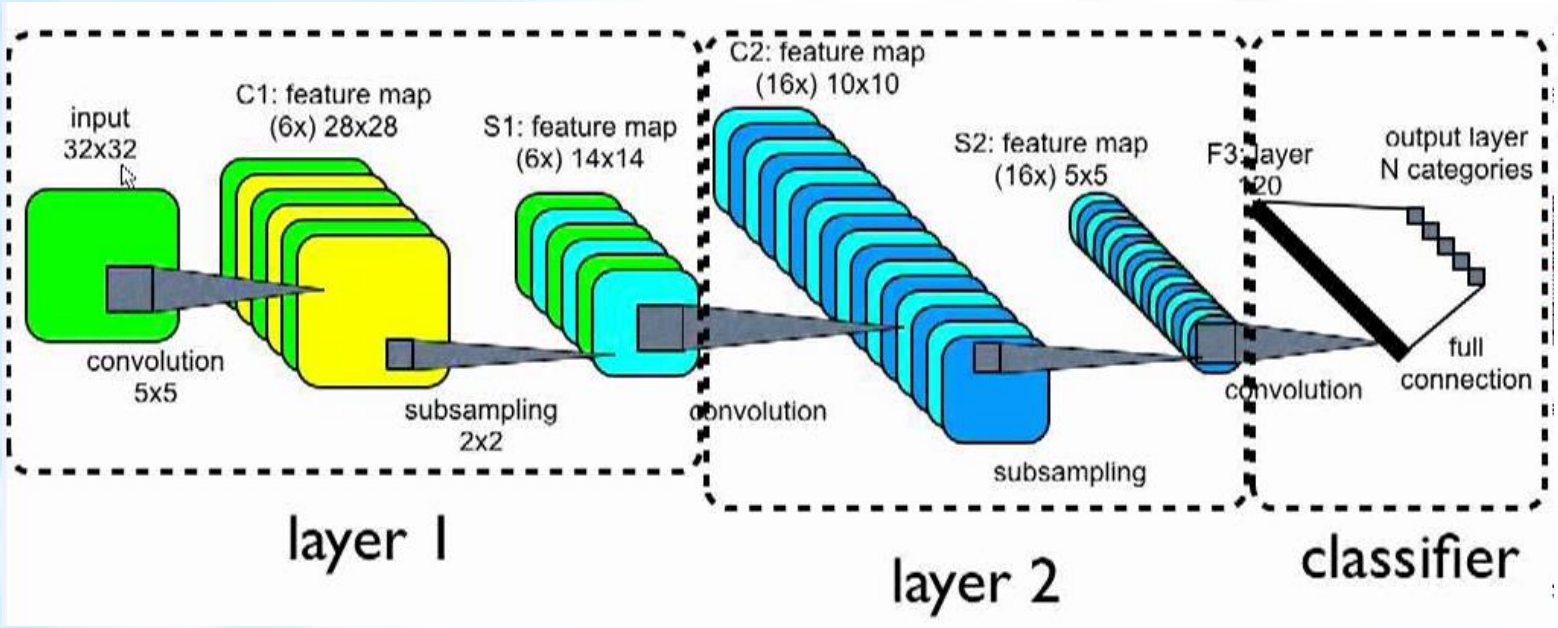

Figure 7 Structure of convolutional neural network (CNN)

Results

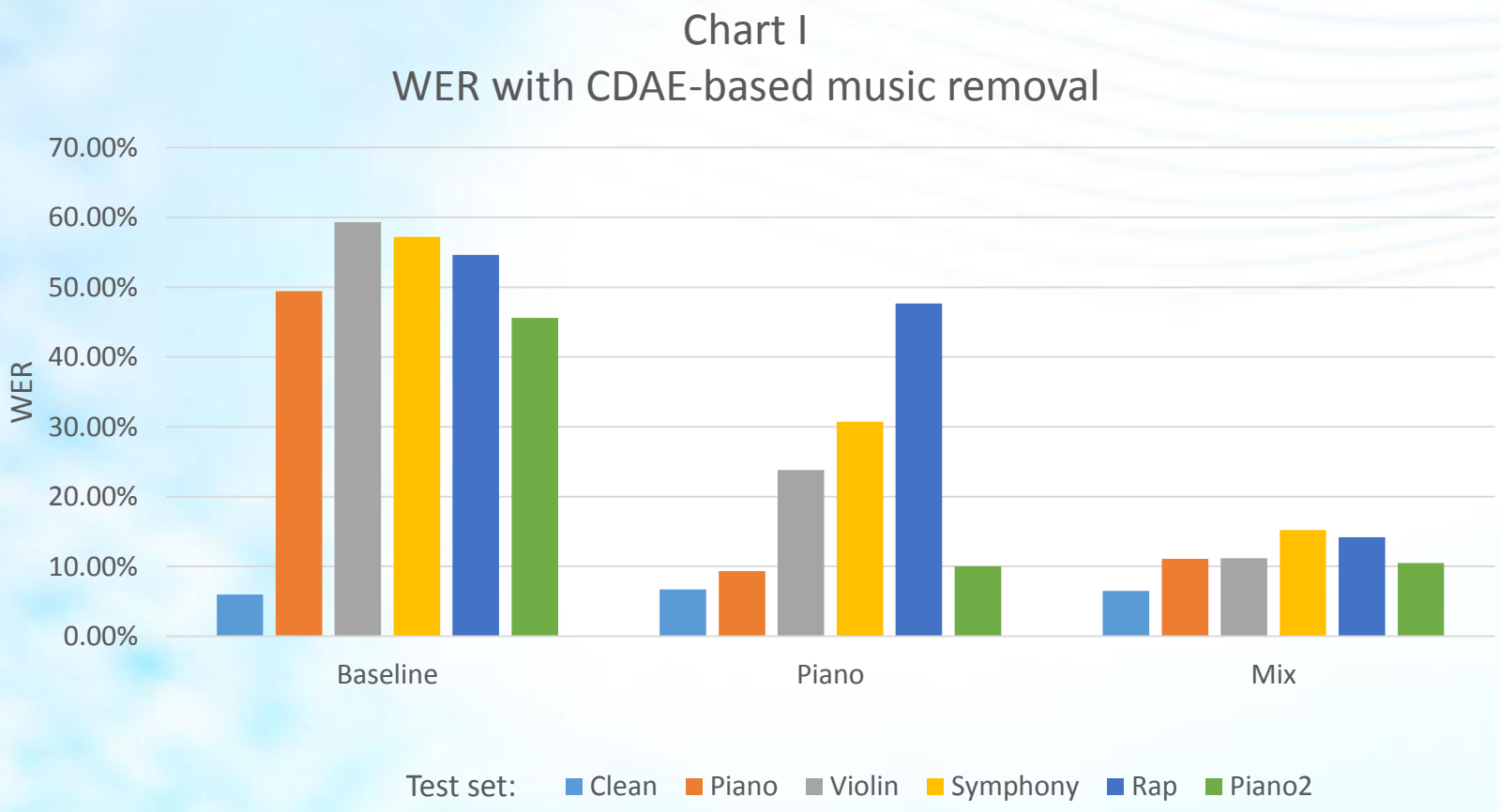
Chart I
WER with CDAE-based music removal
70.00%
60.00%
50.00%
40.00%
30.00%
20.00%
10.00%
0.00%
WER
Baseline
Piano
Mix
Test set:
Clean
Piano
Violin
Symphony
Rap
Piano2

Results

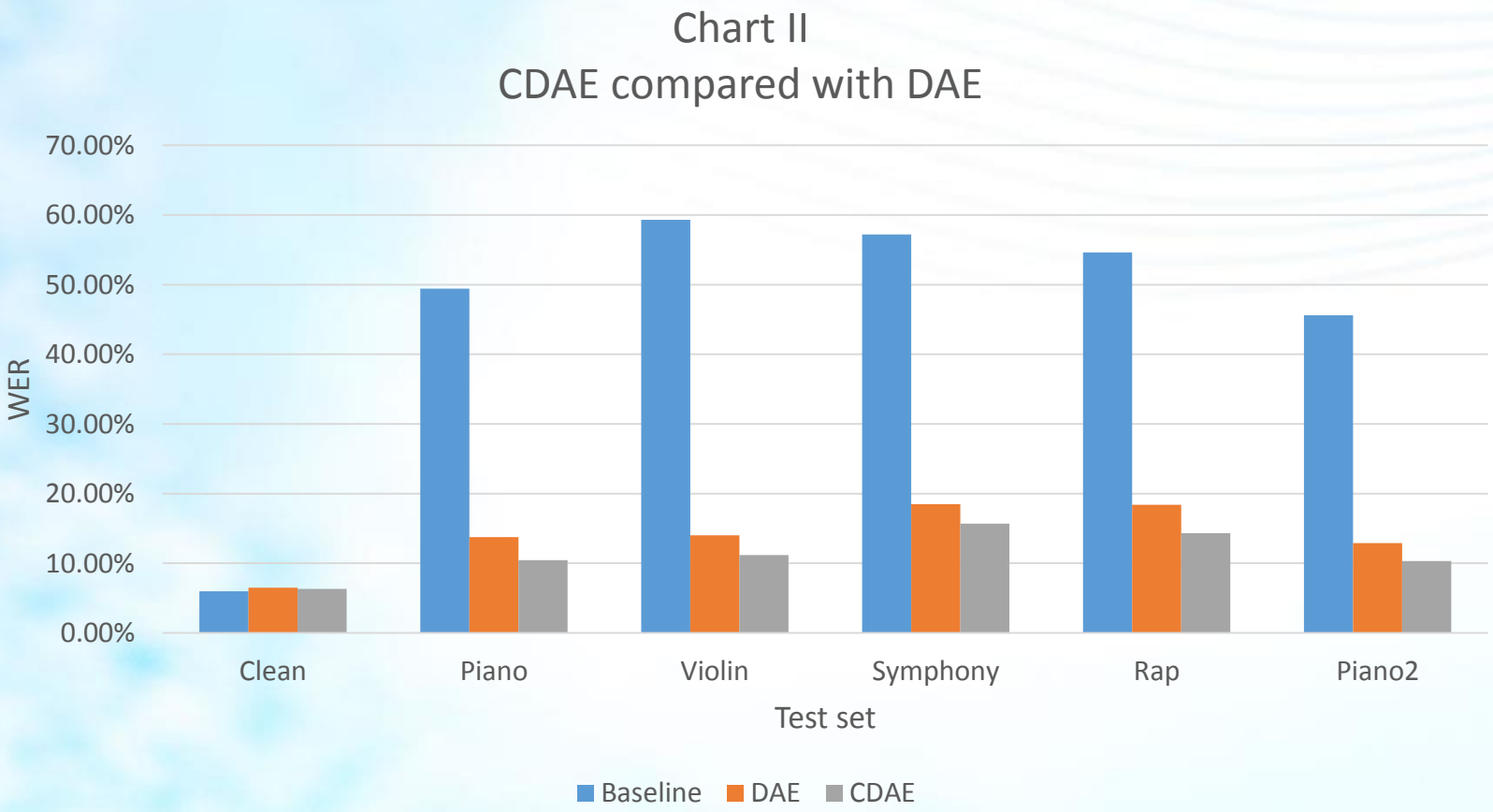
Chart II
CDAE compared with DAE
70.00%
60.00%
50.00%
40.00%
30.00%
20.00%
10.00%
0.00%
WER
Clean
Piano
Violin
Symphony
Rap
Piano2
Test set
Baseline
DAE
CDAE

Music removal across languages

- Music is assumed to be language-independent.
- We assume that music-removal model can be trained and applied across languages.

Results

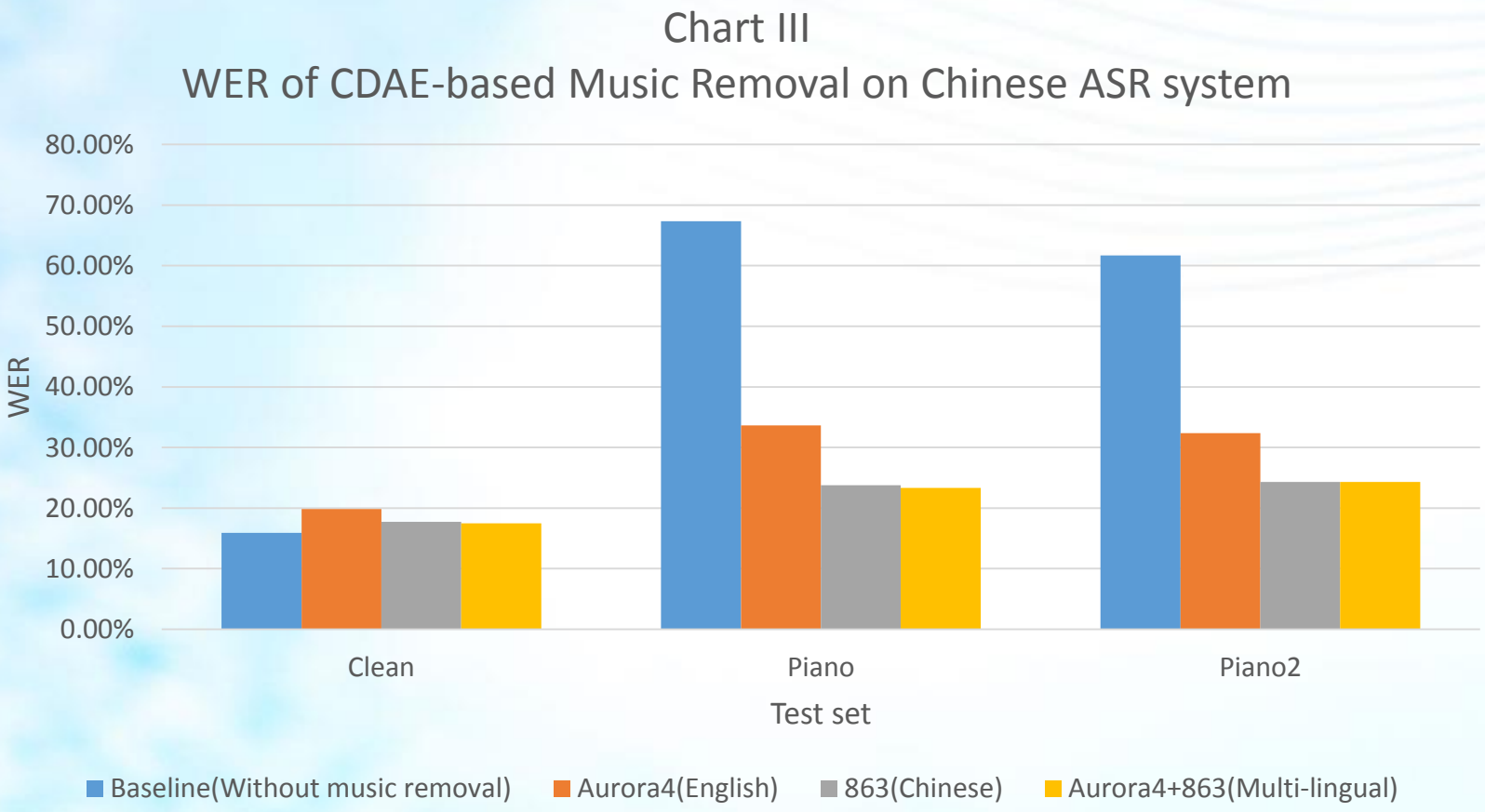
Chart III
WER of CDAE-based Music Removal on Chinese ASR system
80.00%
70.00%
60.00%
50.00%
40.00%
30.00%
20.00%
10.00%
0.00%
WER
Clean
Piano
Piano2
Test set
Baseline(Without music removal)
Aurora4(English)
863(Chinese)
Aurora4+863(Multi-lingual)

Conclusions

- CDAE can learn music patterns and remove them from music-embedded speech signals.
- CDAE model is more powerful than DAE model.
- Music removal model can be applied across languages.
- A general music-removal model is possible by learning with multilingual data embedded with multiple music.

Future work

- Investigate more complex music types.
- Study the multiple music embedding which involves several music signals in the same speech segment.

Thanks