

Identifying Interludes in VOCALOID-Related Music

孟令浩，郭冠男

2023-04-19

Content

- Background
- Pipeline
- Project Structure
- About MFCC
- Feature Abstract
- Feature Selection
- Related Work
- Future Research

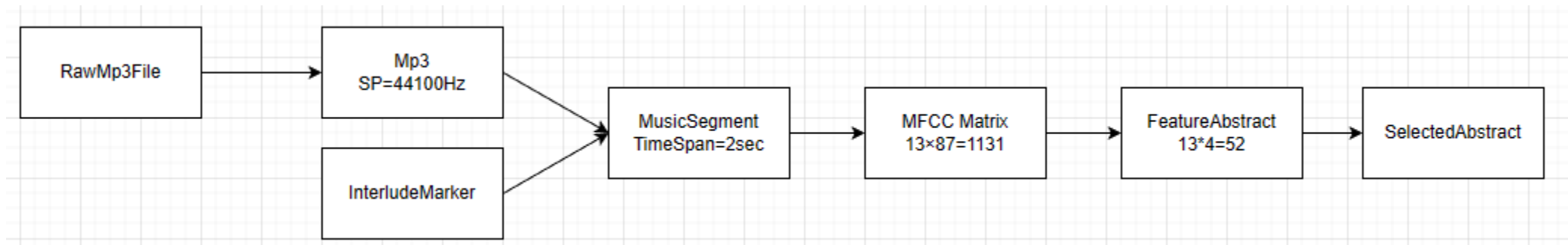
Background



- 周刊VOCALOID中文排行榜
- The QR code in the top right corner is for the latest issue of the weekly magazine (2023-04-16 #558).
- Music Segment Recognition => Interlude Identification

Pipeline

- Automated Boundaries
 - Complete mp3 music file, interlude marker file



Project Structure

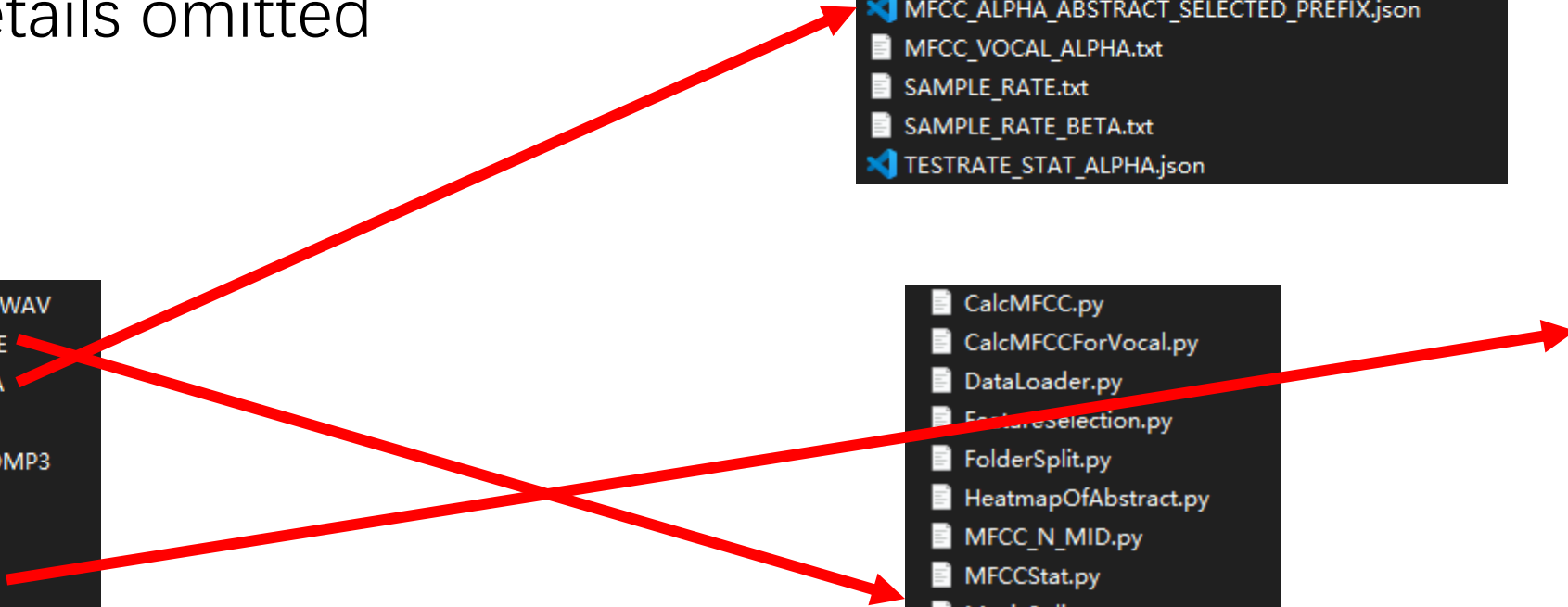
- details omitted

BGMWAV
CODE
DATA
IMG
MODMP3
MP3
SEG
TAG
TEMPDIR
TESTMP3
TMPMP3
VOCALSEG
VOCALWAV

MFCC_ABSTRACT_BEST_FEATURE_ID.json
MFCC_ABSTRACT_BEST_FEATURE_ID_AND_PVALUE.json
MFCC_ALPHA.txt
MFCC_ALPHA_ABSTRACT.txt
MFCC_ALPHA_ABSTRACT_SELECTED_PREFIX.json
MFCC_VOCAL_ALPHA.txt
SAMPLE_RATE.txt
SAMPLE_RATE_BETA.txt
TESTRATE_STAT_ALPHA.json

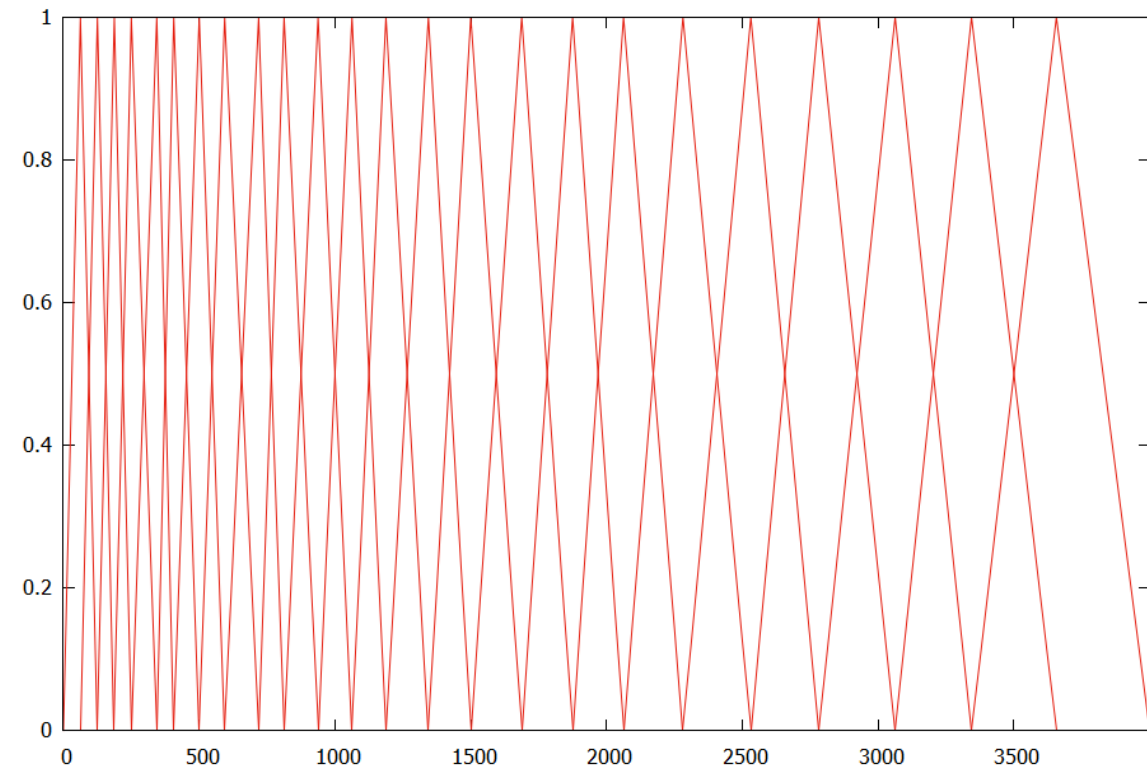
CalcMFCC.py
CalcMFCCForVocal.py
DataLoader.py
FeatureSelection.py
FolderSplit.py
HeatmapOfAbstract.py
MFCC_N_MID.py
MFCCStat.py
MusicSplitter.py
MusicSplitterForVocals.py
PlotHistogramDemo.py
PlotLine.py
RunSvmOnTestSong.py
SampleRateChecker.py
SampleRateTransformer.py
SvmOnSelectedFeature.py

TAG_0003.txt
TAG_0004.txt
TAG_0005.txt
TAG_0006.txt
TAG_0007.txt
TAG_0008.txt
TAG_0009.txt
TAG_0010.txt
TAG_0011.txt
TAG_0012.txt
TAG_0013.txt
TAG_0014.txt
TAG_0015.txt
TAG_0016.txt
TAG_0017.txt
TAG_0018.txt
TAG_0019.txt
TAG_0020.txt
TAG_0021.txt
TAG_0022.txt
TAG_0023.txt
TAG_0024.txt
TAG_0025.txt
TAG_0026.txt
TAG_0027.txt
TAG_0028.txt
TAG_0029.txt
TAG_0030.txt
TAG_0031.txt
TAG_0032.txt



About MFCC

- $M(f) = 1125 \cdot \ln \left(1 + \frac{f}{700} \right) = 2595 \cdot \lg \left(1 + \frac{f}{700} \right)$
- <https://zhuanlan.zhihu.com/p/365714663>



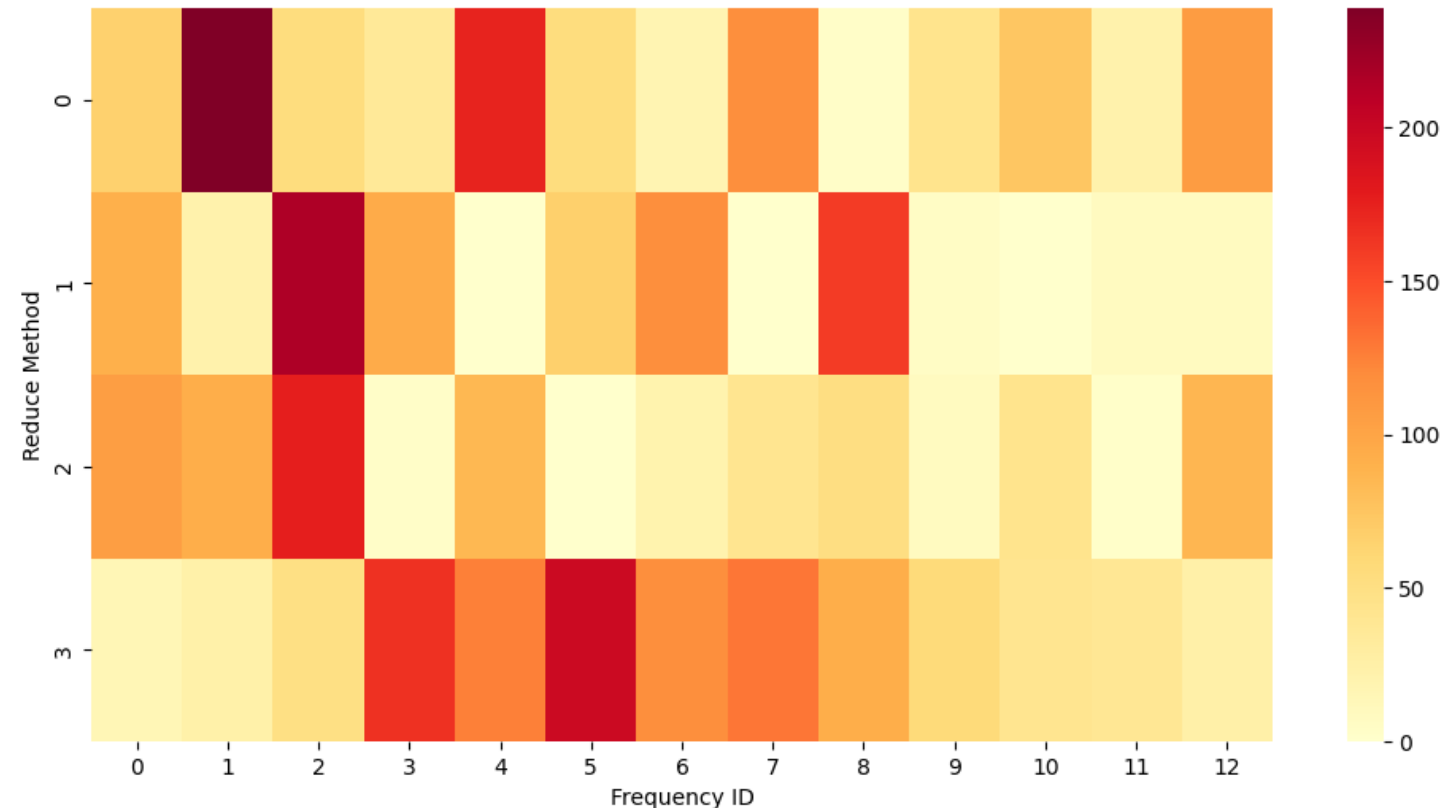
Feature Abstract

- List of $\log(E) \Rightarrow$ Min, Max, Avg, Std
- (There is a lot of professional knowledge related to music data characteristics here, and it cannot be explained fully in a short period of time.)

Feature Selection

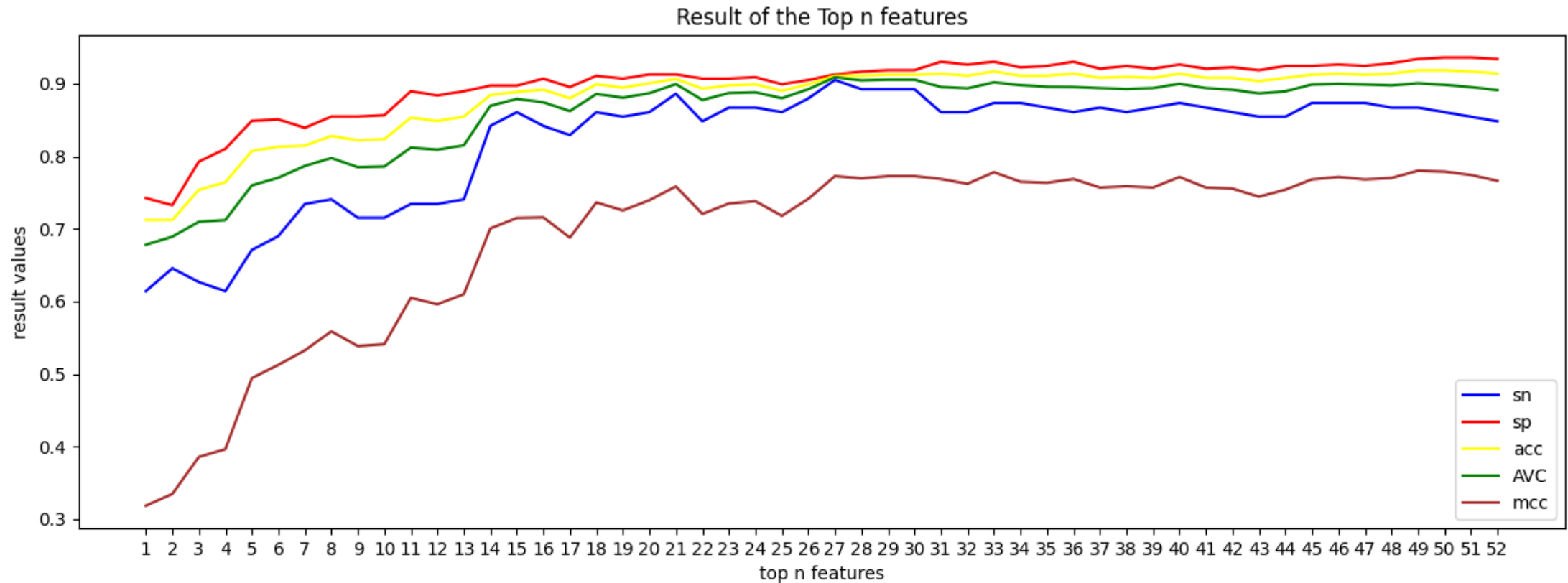
- Ttest: 0-Min, 1-Max, 2-Avg, 3-Std
- based on: $-\ln(\text{pvalue})$

1:	0.00(Hz),	55.40(Hz),	115.19(Hz)
2:	55.40(Hz),	115.19(Hz),	179.71(Hz)
3:	115.19(Hz),	179.71(Hz),	249.33(Hz)
4:	179.71(Hz),	249.33(Hz),	324.47(Hz)
5:	249.33(Hz),	324.47(Hz),	405.55(Hz)
6:	324.47(Hz),	405.55(Hz),	493.05(Hz)
7:	405.55(Hz),	493.05(Hz),	587.47(Hz)
8:	493.05(Hz),	587.47(Hz),	689.37(Hz)
9:	587.47(Hz),	689.37(Hz),	799.33(Hz)
10:	689.37(Hz),	799.33(Hz),	918.00(Hz)
11:	799.33(Hz),	918.00(Hz),	1046.06(Hz)
12:	918.00(Hz),	1046.06(Hz),	1184.25(Hz)



Feature Selection (cont.)

- Top **27** Features, SVM(rbf), $sn = 90.5\%$, $sp = 91.3\%$



Feature Selection (cont.): Baseline

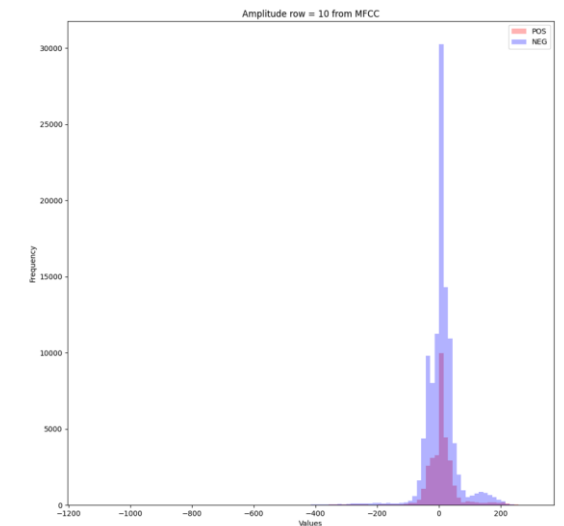
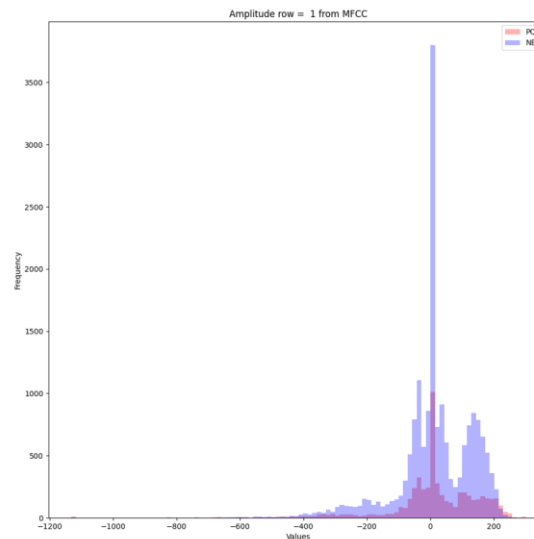
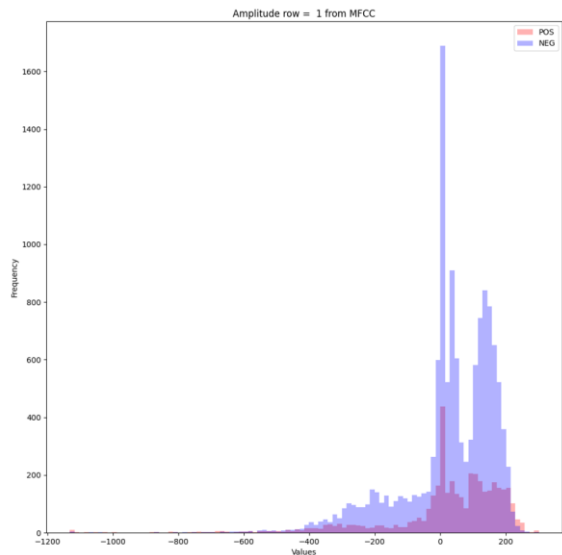
- SVM on total **1131** Features: $sn = 72.8\%$, $sp = 93.2\%$
- SVM on total **52** Abstracts: $sn = 84.8\%$, $sp = 93.4\%$
- Feature selection provides a classification method with better interpretability and accuracy.

Related Work

- In addition to the aforementioned feature selection, we also attempted to use deep learning models (CNN) for voice extraction, but the optimal classification performance, as measured by **SN**, was around **84**.
- After voice extraction, a significant amount of information is lost, which is not sufficient for accurate detection of interludes.

Related Work (cont.)

- Regarding the distribution of data, on the Abstract dataset, the logarithmic energy in most frequency intervals follows a normal distribution, and the distributions of POS and NEG are the same.
- In the low-frequency range (Index=1), the (logarithm) energy seldom follows a normal distribution.



Future Research

- Due to time constraints, here are some further research works that can be explored.
- Combining clustering and binary classification methods.
- Analyze the correlation between spectral summary items.
- Enumerate and select better feature combinations.

That's All

- Thank you for your listening.
- 2023-04-19
- Version=01