Deep Learning Approach for Singer Voice Classification of Vietnamese Popular Music

Pham Van Toan[1] Tran Ngo Quang Ngoc[1] Ta Minh Thanh[2]

[1]R&D Department
Sun-Asterisk Inc.

[2]Faculty of Computer Science
Le Quy Don Technical University

The 10th International Symposium on Information and Communication Technology, December 2019

Table of Contents

Table of Contents

Motivation

The collections of digital music is growing rapidly.

- We need an automatic audio metadata tagging system.

Motivation

The collections of digital music is growing rapidly.

- We need an automatic audio metadata tagging system.
- Specifically, we are tackling the singer problem.

Technologies

- Traditionally, the task is solved with classical models.
  for e.g., SVM, k-NN, Naive Bayes

Technologies

- Traditionally, the task is solved with classical models.
  for e.g., SVM, k-NN, Naive Bayes
  - We will be using deep learning.
    New technologies give SotA results.

Technologies

- Traditionally, the task is solved with classical models.
  for e.g., SVM, k-NN, Naive Bayes
  - We will be using deep learning.
    New technologies give SotA results.
- For audio features, we opted to use the current de facto standard.

Technologies

- Traditionally, the task is solved with classical models.
  for e.g., SVM, k-NN, Naive Bayes
  - We will be using deep learning.
    New technologies give SotA results.
- For audio features, we opted to use the current de facto standard.
  - Classical features

Technologies

- Traditionally, the task is solved with classical models.
  for e.g., SVM, k-NN, Naive Bayes
  - We will be using deep learning.
    New technologies give SotA results.
- For audio features, we opted to use the current de facto standard.
  - Classical features
    - Formant-based

Technologies

- Traditionally, the task is solved with classical models.
  for e.g., SVM, k-NN, Naive Bayes
  - We will be using deep learning.
    New technologies give SotA results.
- For audio features, we opted to use the current de facto standard.
  - Classical features
    - Formant-based
    - Frequency response

Technologies

- Traditionally, the task is solved with classical models.
  for e.g., SVM, k-NN, Naive Bayes
  - We will be using deep learning.
    New technologies give SotA results.
- For audio features, we opted to use the current de facto standard.
  - Classical features
    - Formant-based
    - Frequency response
    - Hidden Markov Model

Technologies

- Traditionally, the task is solved with classical models.
  for e.g., SVM, k-NN, Naive Bayes
  - We will be using deep learning.
    New technologies give SotA results.
- For audio features, we opted to use the current de facto standard.
  - Classical features
    - Formant-based
    - Frequency response
    - Hidden Markov Model
  - Current standard

Technologies

- Traditionally, the task is solved with classical models.
  for e.g., SVM, k-NN, Naive Bayes
  - We will be using deep learning.
    New technologies give SotA results.
- For audio features, we opted to use the current de facto standard.
  - Classical features
    - Formant-based
    - Frequency response
    - Hidden Markov Model
  - Current standard
    - Short-Time Fourier Transform (STFT)

Technologies

- Traditionally, the task is solved with classical models.
  for e.g., SVM, k-NN, Naive Bayes
  - We will be using deep learning.
    New technologies give SotA results.
- For audio features, we opted to use the current de facto standard.
  - Classical features
    - Formant-based
    - Frequency response
    - Hidden Markov Model
  - Current standard
    - Short-Time Fourier Transform (STFT)
    - Mel Frequency Cepstrum Coefficients (MFCC)

Sun*

Our model

- Vocal Segmentation

Our model

- Vocal Segmentation
- Vocal Separation

Our model

- Vocal Segmentation
- Vocal Separation
- Vocal Classification

Model visualization

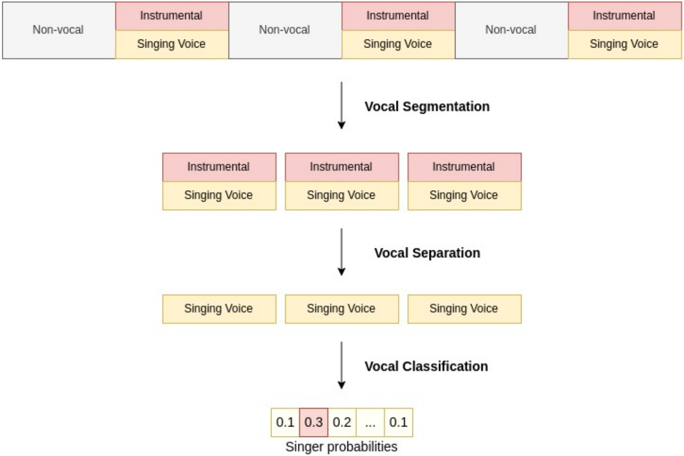
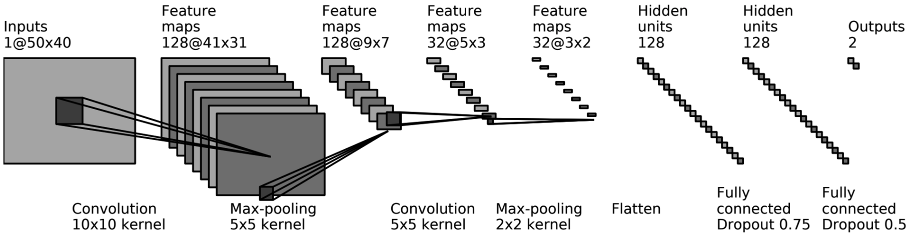


Table of Contents

Sun*

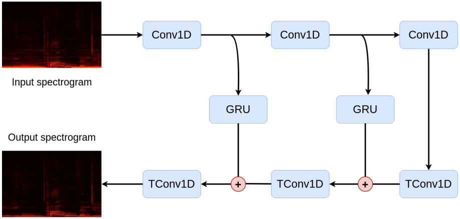
Vocal Segmentation

We use Convolutional Neural Network on the audio spectrogram.



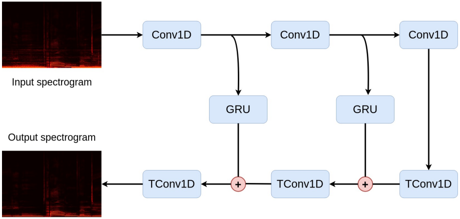
Vocal Separation

- The model is a derivative of U-Net



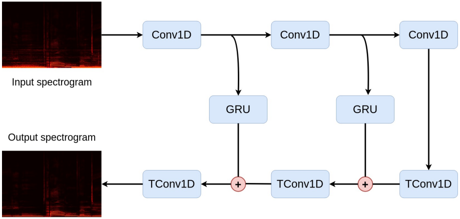
Vocal Separation

- The model is a derivative of U-Net
- Skip connections are passed through GRU first



Vocal Separation

- The model is a derivative of U-Net
- Skip connections are passed through GRU first
- Input and output are spectrograms.



Vocal Classification

Very standard audio classification settings:

- 13 MFCCs with 26 filter bands
- 3 stacked bidirectional LSTMs
- Softmax loss

Table of Contents

Vocal Segmentation

Table: Vocal and non-vocal segmentation result

Song genre	CNN Precision			CNN + Viterbi Precision		
	Vocal	Non vocal	Mean	Vocal	Non vocal	Mean
Country	91.30	97.20	94.25	97.82	99.64	98.73
Balad	92.85	94.24	93.55	98.65	99.86	99.26
Bolero	94.32	90.24	92.28	96.30	98.12	97.21
Rock	88.23	97.15	90.69	90.64	90.67	97.10

Vocal Separation

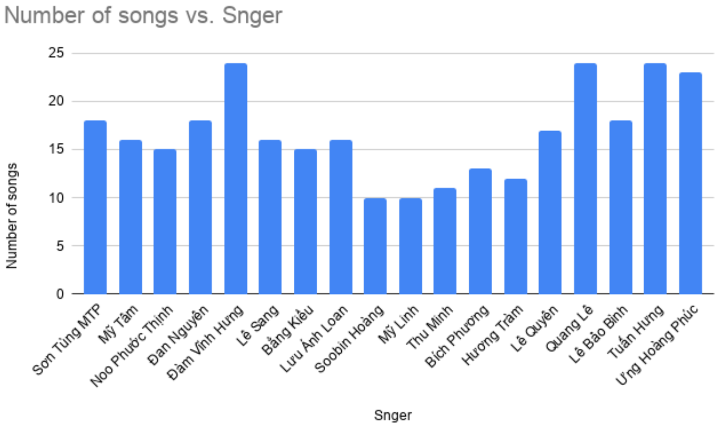
Table: The result of vocal separation

	DSD100	MUSDB18
GRU Skip connection	5.92	5.84
LSTM Skip connection	5.82	5.78

Sun*

Vocal Classification Dataset

Distribution of the dataset:



Vocal Classification Result

Table: The result of vocal classification with two audio signal

	Mean precision	Mean recall	Mean F1 score
Raw signal	85.4	82.6	83.96
Separated signal	93.94	91.78	92.84

Table of Contents

Future works

- Comparisons to be done

Future works

- Comparisons to be done
- Improvements to be made

Thank you for listening!