Automatic music separation (separation of vocals and accompaniment)

Name: Kiro Chen

SID: 530337094

Github ID: kirrot5

Github link: https://github.com/kirrot5/elec5305-project-530337094

Project Overview

This project aims to develop a system that can automatically separate mixed music into vocals and accompaniment. This task has significant application value in scenarios such as karaoke, music production, and audio editing. The project will initially utilize short-time Fourier transform (STFT) and spectral masking methods to achieve baseline separation, and further explore deep learning methods based on U-Net and Demucs to improve the separation quality. By comparing traditional and deep learning methods and conducting experiments and evaluations on the MUSDB18 dataset, the project will develop a functional prototype for automatic music separation and demonstrate its potential value in accompaniment generation and intelligent audio applications.

Background

Music signal separation is an important research direction in audio signal processing, aiming to extract different sound sources from mixed audio, such as human voices and accompaniment. This issue has practical application value in karaoke, music production, and information retrieval. Early methods mainly relied on traditional signal processing and matrix decomposition, such as non-negative matrix factorization (Lee & Seung, 1999; Virtanen, 2007) and independent component analysis. However, their performance was limited in complex music scenarios.

In recent years, deep learning methods have significantly improved the separation performance. The U-Net network achieved excellent results in human voice separation (Jansson et al., 2017), while the Demucs model performed outstandingly in international evaluations through time-domain modeling (Defossez et al., 2019). Nevertheless, existing models still face challenges such as insufficient generalization ability, high computational cost, and artifact problems.

Proposed Methodology

This project will adopt a research approach that combines traditional time-frequency analysis methods with deep learning methods. Firstly, the input audio signal will be converted to the time-frequency domain through the short-time Fourier transform to obtain the time resolution and frequency resolution of the signal (Oppenheim & Schafer, 2009). On this basis, the spectral masking method will be used to initially separate the mixed signal, that is, based on the difference in spectral energy distribution between the human voice and the accompaniment, an appropriate masking function will be designed to enhance the human voice components and suppress the accompaniment.

In the deep learning part, this project will focus on studying the U-Net structure based on convolutional neural networks and the Demucs network based on time-domain modeling. U-Net performs well in the time-frequency spectrum and can capture multi-level feature information using the encoder-decoder structure, thus achieving good results in the human voice separation task (Jansson et al., 2017). The Demucs model adopts the time-domain modeling approach, avoiding the information loss caused by Fourier transformation, and realizes high-fidelity separation through convolution and deconvolution modules (Defossez et al., 2019). This project will separately implement the masking method based on STFT and the deep learning method, and conduct performance comparisons between the two.

In terms of data, this project will use the public MUSDB18 dataset, which contains 150 pieces of music with independent source tracks and has become the standard benchmark for international source separation evaluation. In the experiments, some data will be used for training, and some data will be used for validation and testing.

The experimental platform of this project will be based on Python and MATLAB. Python will be mainly used for the implementation of the deep learning part, using the PyTorch framework for network training and inference; MATLAB will be used for modules such as STFT, spectral masking, and signal visualization to facilitate intuitive understanding of the separation process and results. Performance evaluation will be conducted using standardized indicators, including signal distortion rate, SNR improvement, and perceptual audio quality assessment, to comprehensively evaluate the separation effect.

Expected Outcomes

This project is expected to develop a prototype system for automatic music separation based on public datasets. Firstly, baseline separation is achieved using short-time Fourier transform and spectral masking, which is done in an intuitive and computationally efficient manner as a comparison. Subsequently, a deep learning model based on U-Net and Demucs will be implemented and optimized. It is anticipated that better results for separating vocals and accompaniment will be achieved on the MUSDB18 dataset. Previous studies have shown that such methods excel in preserving vocal details and suppressing accompaniment (Jansson et al., 2017).

The project will evaluate the model using indicators such as signal distortion rate (SDR), signal-to-noise ratio improvement (SNRi), and perceptual audio quality assessment (PESQ). These indicators have been widely used in source separation research. Additionally, subjective scores will be collected through listening tests to complement the objective results. The final outcome will be released in the form of a GitHub repository, including source code, data processing scripts, and demonstration examples, providing an experimental platform for academic research and practical applications (such as karaoke and music production).

Timeline

The implementation period of this project is from the sixth to the thirteenth week, divided into four stages. The first six to seventh weeks mainly involve literature review and data preparation, investigation of traditional spectral masking and deep learning methods, summarization of their advantages and disadvantages, and downloading and organizing the MUSDB18 dataset for preprocessing. The eighth to ninth weeks will implement and test the baseline model based on STFT and spectral masking to verify the data processing procedure and provide a control for subsequent experiments. The tenth to eleventh weeks focus on the implementation and optimization of deep learning models, training U-Net and Demucs using PyTorch, and evaluating the indicators through the validation set. The twelfth to thirteenth weeks complete the final work, including conducting final evaluation on the test set, writing the report and publishing the GitHub repository, presenting the source code, documentation and demonstrations, and comparing the effects of different methods, summarizing the conclusions and improvement directions.

References

Défossez, A., Usunier, N., Bottou, L., & Bach, F. (2019). Music source separation in the waveform domain. *arXiv preprint arXiv:1911.13254.*

Jansson, A., Humphrey, E. J., Montecchio, N., Bittner, R., Kumar, A., & Weyde, T. (2017). Singing voice separation with deep U-Net convolutional networks. In *Proceedings of the 18th International Society for Music Information Retrieval Conference (ISMIR)* (pp. 745–751).

Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature, 401*(6755), 788–791. https://doi.org/10.1038/44565

Oppenheim, A. V., & Schafer, R. W. (2009). *Discrete-time signal processing* (3rd ed.). Pearson

Virtanen, T. (2007). Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. *IEEE Transactions on Audio, Speech, and Language Processing, 15*(3), 1066–1074. https://doi.org/10.1109/TASL.2006.885253