Latent Feature Augmentation for Chorus Detection

Xingjian Du Huidong Liang
Yuan Wan Yuheng Lin Ke Chen
Bilei Zhu Zejun Ma
ByteDance AI Lab
University of California San Diego

Introduction

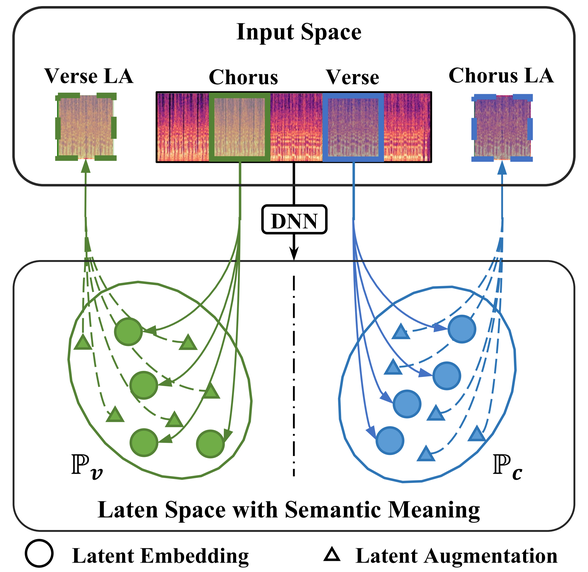
In this paper, we introduce *ByteCover2*, a chorus detection model based on **L**atent feature **A**ugmentation and ResNet-FPN architecture. We make three contributions. Firstly, we propose a method for implicitly augmenting chorus data in the latent space during the training stage. Compared to augmentations on audio surfaces such as time stretching and pitch shifting, latent augmentations indicate changes at a higher level in original audio, thereby increasing the diversity and sufficiency in training. Second, we apply Feature Pyramid Network (FPN) to generate additional embeddings from low dimension to high dimension, consequently achieving a multi-scale training paradigm. Lastly, we release , a new diversified dataset of 13 genres and 14 languages for the community of music structure analysis. In conjunction with other public datasets, we conduct comprehensive experiments to evaluate the performance of our proposed method compared to other state-of-the-art models, where *ByteCover2* outperforms other SOTAs by a considerable margin, meanwhile the proposed latent audio augmentation shows dominant advantages over traditional augmentation methods.

TL;DR: LA-Chorus improves the accuracy of chorus detection with latent augmentation trick. Moreover, we release a new music structure dataset with diversified languages and genres.

Take a picture to access the dataset

Implicit Audio Augmentation

Illustration for implicit audio augmentation in MSA with two annotation types (verse and chorus) from a constant Q transformation (CQT) spectrogram excerpt of "Smooth Criminal" in *Isophonics* [**?**]. The spectrogram is first encoded into latent embeddings by frame, where chorus embedding and verse embedding follow latent distributions $\mathbb{P}_c$ and $\mathbb{P}_v$ respectively. Then latent augmentations are sampled around embeddings of verse/chorus segments, which correspond to augmented verse/chorus segments in the input space (represented by dashed lines, meaning they are NOT shown explicitly in the input space).



ByteDance 字节跳动