IDENTIFYING COVER SONGS USING DEEP NEURAL NETWORKS

COMPUTER AUDITION ECE 477 FINAL PROJECT

MARKO STAMENOVIC

UNIVERSITY of ROCHESTER

HAJIM SCHOOL OF ENGINEERING & APPLIED SCIENCES
UNIVERSITY of ROCHESTER

INTRODUCTION

A cover song, cover version, or simply cover, by definition, is a new performance or recording of a previously recorded, commercially released song by someone other than the original artist or composer. Automatic cover song detection has been an active research area in the field of Computer Audition for the past decade.

In this paper, we propose a novel method for cover song detection using automatic extraction of audio features with a stacked auto-encoder (SAE) combined with beat tracking in order to

OVERVIEW

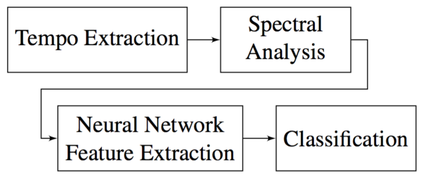

Figure 1. System block diagram

Tempo Extraction

Tempo is extracted using the open source Music Audio Tempo Estimation and Beat Tracking tool by D. Ellis and LabROSA. The audio is truncated to begin from the first beat point in order to remove any introduction clapping or speach, which are a common feature of live cover songs.

Spectral Analysis

The beat information is used to set window and hop size for two spectral analyses: a CQT and chroma features. Each window is set to length of 1/4 note of the audio and hop size is set to the length of 1/8 note of the audio. The audio is segmented according to tempo in order to encourage the neural network to learn tempo features of the music.

Feature Extraction

Features are extracted using a two-layer stacked auto-encoder (SAE) with 1440 input neurons, 500 first hidden layer neurons and 100 2nd layer hidden neurons. Each CQT input patch consists 8 frames with 180 frequency bins or one measure of the song and each chroma feature input consists of 144 frequency bins and 10 frames with 50% overlap. The SAE is then trained on a set of concatenated spectral inputs from 80 original songs.

Classification

Dynamic time warping is used to determing the distance of the output features of each song in the 165-song dataset to each cover seed song. If a cover-seed and it's cover are closest.

SYSTEM

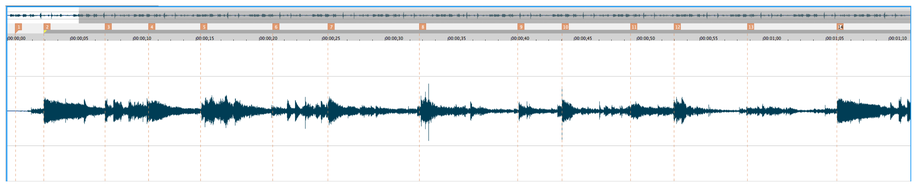

Figure 2. Beat extraction using onset detection

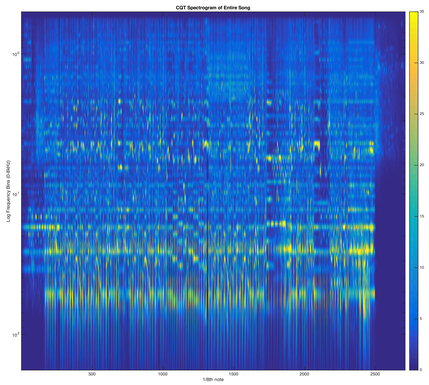

Figure 3. Typical CQT Spectrogram

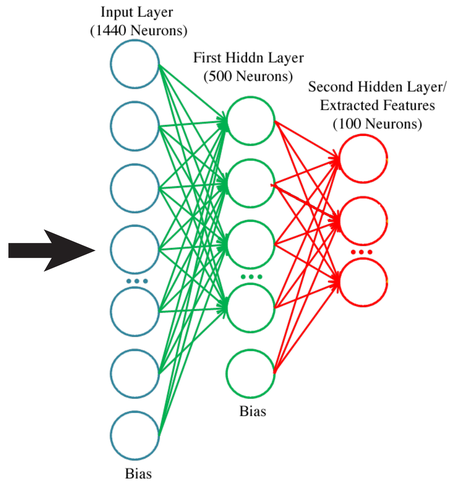

Figure 4. Typical two layer SAE implementation

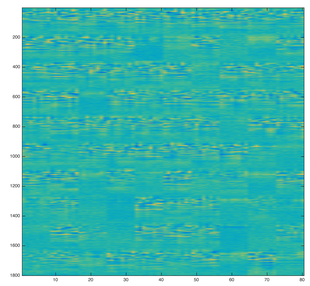

Figure 5. Visualization of the first 100 features of the first hidden layer

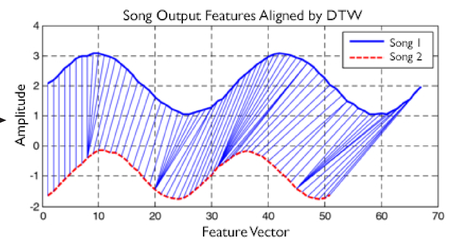

Figure 6. Example of DTW alignment and distance calculation

Figure 2 is a visual representation of beat extraction using onset identification.

Figure 3 shows an example CQT spectrogram for one of the seed songs. The CQT groups frequency energy into bins based on musical octaves. Our CQT groups them into 15 bins/octave (180 total) in a frequency range of 20Hz to 8000Hz. The CQT is used insetad of more common frequency transforms such as the DFT due to the fact that that the log-frequency scale corresponds better to human auditory perception.

The chroma features are a spectrogram binned to reflect the 12 semitone of the the western musical scale. For our chroma features we created a "wrapped" set of chroma features by creating a transposition for each semitone corresponding to a potential shift in key. This corresponds to 144 frequency bins. Thus our spectral input covered every possible key change a cover of a cover song.

Figure 4 shows a typical two layer SAE implementation. The first layer intakes a patch of 8 CQT frames each containing 180 frequency bins or 10 chroma frames each containing 144 frequency bins (8X180=1440 for CQT or 10X144 for chroma). The 8 frames correspond to one musical measure for CQT. We hope that training the neural network in correspondence to music tempo will encourage it to learn tempo features.

Figure 5 shows a visualization of thefirst 100 features of the first hidden layer of the SAE.

Figure 6 shows a typical example of dynamic time warping in measuring distance between two signals. Dynamic time warping is used in order to overcome any tempo errors that may have occured during the beat extraction.

EVALUATION

The system was evaluated using the *covers80* dataset, a dataset commonly used for evaluating cover song detection systems.

We evaluated twosystems in total:
CQT : The system described using the CQT as spectral input. We used a concatenated matrix of cover seed patches to train the SAE and test all songs against.

Chroma Features: The system described using the CQT as spectral input. We used a concatenated matrix of cover seed patches to train the SAE and test all songs against.

The results are shown below in table 1 compared against random guessing.

1/2 Note Patch Overlap Accuracy		
	DTW	Bag of Features
CQT	13.75%	7.5%
Chroma Features	13.75%	10%
Random Guess	1.25%	1.25%

Table 1. Cover system detection results

As shown in table 1, both systems are better than random guessing, indicating that they are indeed classifying songs based on our system. In addition, the table indicates that the beat tracking implemented in System 1 gives better performance than a similar system with no beat-tracking implementation.

However, the results also show that the system has a long way to go before it can match the performance of current state of the art cover song detection systems.

FUTURE WORK

There are various avenues planned for future improvements of the system. One is implementing a part-of-song detection system before the input to the cover song detection system in order to isolate certain parts of songs such as the chorus, which are often more similar in cover seed/covers than other parts of the system.

Another is to experiment with different and perhaps more granular beat-matched window sizes for the CQT as well as different patch sizes for the SAE to see how that will affect results.

We would also like to try an unwrapped set of chroma features with a longer time domain length to see how this will affect results.