成蹊大学理工学研究報告
J. Fac. Sci.Tech., Seikei Univ.
Vol.58 No.2 (2021)pp.1-6

Improved System for Identification of Live Music Performances by Dynamic Time Warping

Taisuke KAWAMATA*1, Ayami HONDA*2, Yoshitatsu MATSUDA*3

ABSTRACT：The identification of a song performed at a concert, called the "live version," is not yet a popular search among users because of the song possibly having a different arrangement or other modifications, unlike audio searches for an exactly matching song. For live song identification, we examined the application of dynamic time warping to chroma feature series extracted from both studio and live versions of a song. In this paper, we especially focused on measuring the linearity of the warping path. It was evaluated using datasets collected from the internet and the results were compared to those of an existing search service, which showed that the proposed method has the highest accuracy.

Keywords：Song identification, Live recording, Chroma feature, Dynamic time warping, Correlation

(Received Nov 29, 2021)

1. Introduction

The search demand for music information has grown in recent years. Online digital music collections are on the order of millions of tracks, and personal collections can exceed the time available to listen to them. Thus, song identification has been a very active area of study within the last few years. In addition, applications such as SoundHound [1] and Shazam [2] have become widespread. They can manage not only original songs, but also cover songs and live recordings. However, the identification accuracy of live music tends to be lower than that of the original song. Although a system can recognize a faithfully played original song reproduced by the songwriter, it may misidentify a song when someone who is not the songwriter performs the song or when the artist is original, but the recording environment is different, as in studio versus live versions.

The task of live song identification is to recognize a different version or a performance of a previously recorded song, which is similar to cover song identification. Marvsik et al. [3] focused on an ambiguous definition of a cover song and demonstrated a cover song identification system by chroma features and dynamic time warping (DTW). The chroma features and DTW are frequently applied in the field of music informatics. Serra et al. [4] combined music chroma, recurrence plots, and other concepts for cover song identification. Meron and Hirose [5] applied DTW automatic alignment of a musical score to performed music. These methods, however, focused on cover songs, not live performances. Even if the task of live performance searching was similar to that of cover song searching, the restriction of recording environments is different. The recording of a live performance has limitations such as the number of microphones, instrument selection, mixing capability, and performer skill.

For live song identification, Tsai et al. [6] proposed a method identifying a live performance by recording a few seconds of the performance by cell phone. Their research was revolutionary in using GPS in addition to the sound. Doras et al. [7] applied neural networks, such as the convolutional and Siamese networks. By evaluation with concerts recording that contains 10 to 30 songs, they showed that their proposal improves results significantly for live song identification.

We assume that song identification is applied to a song

*1：Assistant Professor, Department of Computer and Information Science
kawamata@st.seikei.ac.jp

*2：Undergraduate Student, Department of Computer and Information Science

*3：Associate Professor, Department of Computer and Information Science

performed by an amateur at a live house. Because the songwriters do not perform the song in such cases, the information about the live location is not important. Thus, we focus only on the song title, not the location or artist. In addition, it is assumed that only one song is input into the system, so complex processes, such as recognizing the sequence of each song separately from a long audio recording, are not the target. It is therefore unnecessary to use a system as complex as a neural network.

Our purpose is to develop a system that identifies a song when audio data for one song recorded in a live performance is given as input. In Section 2, we provide an overview of our song identification system and explain the harmony feature, matching algorithm, and similarity calculation. Then, Section 3 discusses the results of an evaluation experiment. In Section 4, we provide conclusions and discuss future work.

2. Method

We examined live song identification with the path linearity of DTW. This section gives an overview of our system. At first, given two sound sequences, one from a live recording and one from a studio recording, we use chroma features to extract a descriptor time series representing the harmony progression. Next, to assess equivalences of states between both sequences attained at different times, we use DTW and the warping path derived from them. Finally, we measure the linearity of the path. These are shown in Fig. 1.

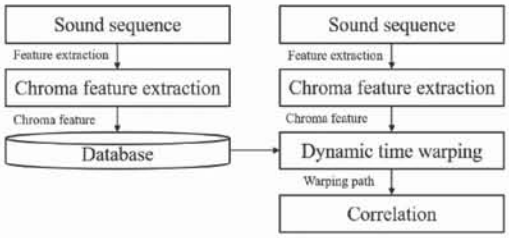

Fig. 1. System overview.

2. 2 Chroma feature

Extraction of chroma features provides a way to represent music audio signals by harmonic characteristics. With these features, the entire spectrum is projected onto 12 bins representing the 12 distinct semitones of the musical octave (C, C#, D, D#, E, F, F#, G, G#, A, A#, B). The chroma features extracted from an audio frame can be explained as a vector:

$$\boldsymbol{x} = \left(x_{\mathrm{C}},\ x_{\mathrm{C\#}},\ x_{\mathrm{D}},\ \ldots,\ x_{\mathrm{A}},\ x_{\mathrm{A\#}},\ x_{\mathrm{B}}\right). \quad (1)$$

2. 2 Dynamic time warping

Dynamic time warping is an algorithm for measuring the similarity between two sequences. DTW can calculate similarities in music, even if a live song was performed faster than the studio version, or if the tempo increased and decreased during the concert.

The $\mathrm{C}(i, j)$ indicates the cost at i and j, which is calculated as follows:

$$\mathrm{C}(i,j)=\mathrm{D}(\boldsymbol{X}_i,\boldsymbol{Y}_j)+\min\begin{pmatrix}\mathrm{C}(i,j-1)\\ \mathrm{C}(i-1,j-1)\\ \mathrm{C}(i-1,j),\end{pmatrix} \quad (2)$$

where $\boldsymbol{X}$ and $\boldsymbol{Y}$ are a sequence of chroma feature vectors $\boldsymbol{x}$ and $\boldsymbol{y}$. i and j are indexes of the audio frame ($1 \leqq i \leqq I$, $1 \leqq j \leqq J$). The parameters I and J describe the audio lengths. The distance $\mathrm{D}(\boldsymbol{x}, \boldsymbol{y})$ between $\boldsymbol{x}$ and $\boldsymbol{y}$ in the chroma feature is determined as the cosine distance:

$$\mathrm{D}(\boldsymbol{x},\boldsymbol{y}) = 1 - \frac{\boldsymbol{x}\cdot\boldsymbol{y}}{|\boldsymbol{x}||\boldsymbol{y}|}. \quad (3)$$

where $\boldsymbol{x}\cdot\boldsymbol{y}$ is the dot product of $\boldsymbol{x}$ and $\boldsymbol{y}$. $|\boldsymbol{x}|$ is the Euclidean norm of $\boldsymbol{x}$.

Fig. 2 shows a part of a cost matrix. The bold line shows the alignment, and different shades of gray indicate the costs, with darker shades indicating lower costs. The DTW tries to find an alignment that follows a route with a low cost distance. The warping path between the same song played differently becomes $j / J = i / I$. On the other hand, the path between different songs becomes a non-linear function.

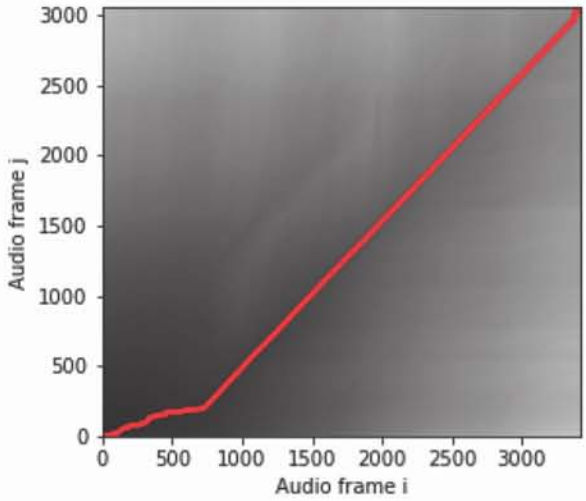

Fig. 2. Example of DTW result.

2. 3 Correlation

The best warping path (i, j) can be generally assembled by backtracking from the end ($i = I, j = J$) to the beginning ($i = 0$, $j = 0$) of a pair of songs. This method is effective when the beginning and the end of two songs match completely. However, a live version might have a special introduction (musical or spoken) that the studio does not have. Therefore, we also evaluated the warping path by trimming 30 sec off the beginning and 20 sec off the end of the song. This pre-processing removes any non-linear path at front and back, which may be talking to the audience or song editing that differs between the live and studio versions.

In identification by DTW, the distance between two sequences is measured by calculation of lowest cost $C(I, J)$. However, the distance between the songs recorded in different environments is naturally large because the value is too sensitive to the bias of the recording environments. In addition, $C(I, J)$ is the integral of cosine distance, whose calculation is affected by the arrangement or talking during the introduction.

In this paper, we focused on the correlation coefficient between two warping paths to evaluate the correspondence between studio and live versions. This coefficient can be used to calculate a matching score (a sort of similarity between two songs) using only the best warping path (i, j), regardless of the value of the cost matrix $C(I, J)$.

$$\mathrm{R}(\boldsymbol{i}, \boldsymbol{j}) = \frac{\sum_{t=s}^{e}(i_t - \bar{i})(j_t - \bar{j})}{\sqrt{\sum_{t=s}^{e}(i_t - \bar{i})^2}\sqrt{\sum_{t=s}^{e}(j_t - \bar{j})^2}}, \quad (4)$$

where s represents the start point and e represents the end point of the range of the warping path calculation for the matching score. These parameters s and e are useful for removing parts of the live version that are different arrangements or otherwise "impure." The system matches a live song to songs in the database and identifies the studio version that has the highest correlation $\mathrm{R}(\boldsymbol{i}, \boldsymbol{j})$ with the given live version.

2. 4 Experimental overview

Our proposal was evaluated by comparison with studio song identification of SoundHound [1]. We obtained 35 songs (studio version) performed by Japanese bands as mp3 files and collected a live version corresponding to each studio version from the internet. The songs were arranged by players if the length of the live version was different from the studio version.

Regarding SoundHound, we played the live version of songs from the beginning until it could be matched to the studio version. Identification failure was defined as the failure to match the live version to the studio version when the live version ended.

Regarding the proposed method, we played the songs from 10 to 130 sec because we assumed that the beginning of the song has no sound, and the latter half repeats the first half. Identification failure was defined as the failure to match the live version to the studio version.

The proposed system was implemented with Python 3.7. We adopted the librosa library [8] for extracting chroma features. We defined the chroma parameter as follows: FFT window size = 4,410, the hop size = 4,410, and audio data were expressed as 10 frames per second. The DTW parameters were given as the default librosa values.

3. Experiments and results

3. 1 Results of SoundHound

Table 1 shows the length of each song and the results of the analysis. The songs might be arranged by players if the length of a live version varied greatly from the studio version. The "Player" column identifies whether the live version is performed by the songwriter (Self) or someone else (Copy).

The last three columns show the success or failure of identification by SoundHound [1] and our proposed method. The "**S**" symbols mean the success, and the "F" symbols mean the failure. The "s" symbols mean that the identification was successful for only one part of the song, such as a hook riff or song chorus. The results show that SoundHound could identify 18 songs, including the partial identification ("s"). Moreover, the four songs performed by someone other than the songwriter could not be identified. Hence, cover songs tended to be misidentified.

3. 2 Results of proposed method

The last two columns of Table 1 show the results of our proposed method. The column [0-120] indicates the results of the proposed method within the range of 0-120 seconds, where the proposed method could identify 20 songs. DTW could be therefore useful for song identification. Although, this method could not identify some songs which could be identify by SoundHound [1] because players added an extra phrase to the introduction.

Table. 1. Results of identification.

Song No.	Length		Live Player	Sound-Hound	Proposed	
	Studio	Live			0-120	30-100
1	04:11	04:19	Copy	F	**S**	**S**
2	04:10	04:24	Self	F	**S**	**S**
3	03:31	03:35	Copy	F	**S**	**S**
4	05:13	04:53	Copy	F	F	F
5	04:49	04:46	Self	**S**	**S**	**S**
6	03:42	04:13	Self	F	F	**S**
7	05:06	04:29	Copy	F	**S**	**S**
8	04:39	04:43	Self	F	F	**S**
9	04:25	04:32	Self	**S**	F	**S**
10	02:43	02:32	Self	F	**S**	**S**
11	04:25	04:38	Copy	F	F	**S**
12	04:13	04:30	Self	**S**	**S**	**S**
13	03:39	03:43	Self	F	**S**	**S**
14	04:51	05:27	Self	s	F	**S**
15	05:05	05:41	Self	**S**	F	**S**
16	04:00	04:05	Copy	s	**S**	**S**
17	05:01	04:30	Copy	F	F	**S**
18	03:38	03:42	Self	s	**S**	**S**
19	04:24	04:17	Copy	F	F	**S**
20	04:09	03:50	Copy	s	**S**	**S**
21	04:36	04:56	Self	F	**S**	**S**
22	05:05	05:26	Self	s	F	**S**
23	04:51	05:19	Self	**S**	F	**S**
24	05:33	06:03	Self	**S**	**S**	**S**
25	04:19	04:16	Copy	F	**S**	**S**
26	03:30	03:42	Self	**S**	F	**S**
27	04:25	04:40	Self	s	**S**	**S**
28	03:08	03:24	Self	F	F	**S**
29	04:10	05:38	Copy	F	F	F
30	04:45	04:48	Copy	s	**S**	**S**
31	05:24	05:18	Self	**S**	**S**	**S**
32	04:41	04:43	Copy	**S**	F	**S**
33	04:05	03:57	Copy	F	**S**	**S**
34	04:59	04:56	Self	**S**	**S**	**S**
35	03:59	04:03	Self	s	**S**	**S**
Successful identifications				18	20	33

In contrast, the proposed method within the range of 30-100 seconds could identify 33 songs successfully. The last column of Table 1 indicates that results. The number of successful identifications was 13 songs larger than the identification within the range of 0-120 seconds, so it suggests that the trimming technique of the front, and end of the songs is effective for live song identification.

3. 3 Discussion

Fig. 3 illustrates the cost matrix and warping path for song No. 25, which could not be identified by SoundHound [1]. This result shows that the warping path between two versions of the same song is linear in spite of that the live version is different from the studio version. On the other hand, we obtained non-linear functions from the studio version of the song No. 10 and the live one of the song No. 9. Fig. 4 shows the path of the studio and live versions of song Nos. 10 and 9, respectively.

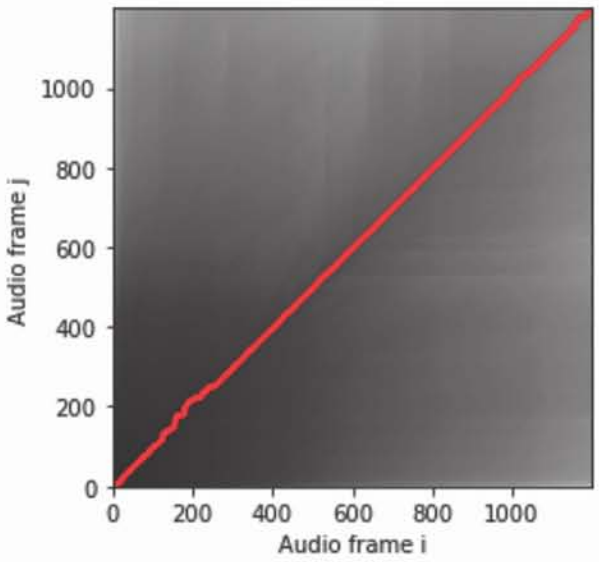

Fig. 3. Warping path (25-25, the same song)

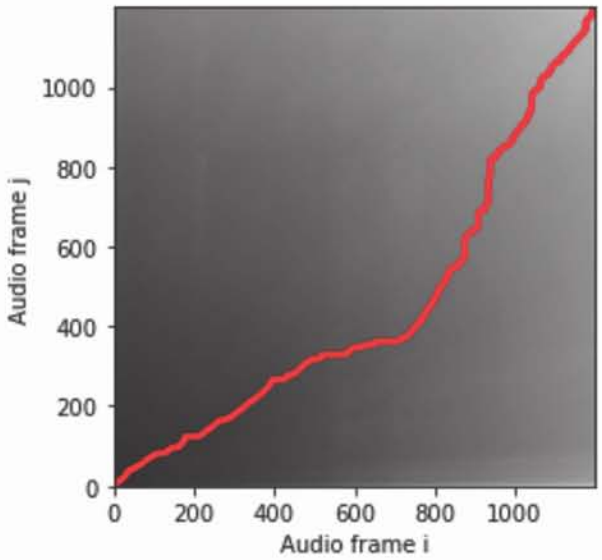

Fig. 4. Warping path (10-9, the different songs)

They show that it is quite effective for live song identification to measure the linearity of the warping path by the correlation.

In spite of that SoundHound is not able to identify Nos. 3, 7, and 33, the proposed system could identify these songs because the warping paths of these songs were linear. Especially, the sound of No. 7 had low quality because the performance and recording were not professional. Hence, the proposed system could identify a song even when performed by an amateur. A linear path exists somewhere in the cost matrix when a live song is the same as the studio version. Fig. 5 shows the warping path of songs, where an extra phrase is added to the beginning and end of the live version. If a song is edited, the edited parts are horizontal lines, vertical lines, or non-linear curves.

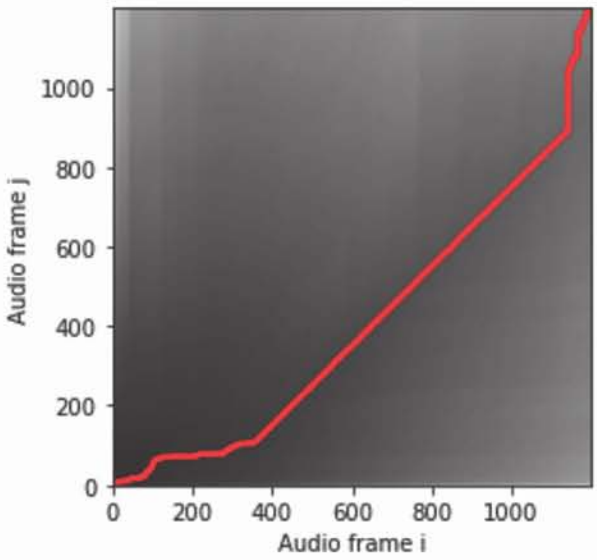

Fig. 5. Warping path (6-6, the same song)

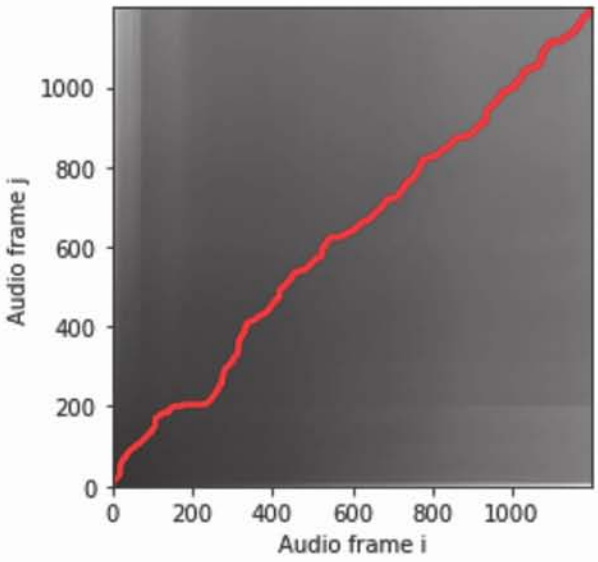

Fig. 6. Warping path (4-4, the same song).

Nevertheless, the intermediate part of a sequence is still near to linear. Therefore, as the correlation of that path takes a low value within all the range, the proposed method within the range of 0-120 seconds could not identify this song. In contrast, the proposed method within the range of 30-100 seconds could identify it due to removal of the non-linear path at the beginning and end of the live version.

Song No. 29 could not be identified, in spite of the live version being performed by one of the most popular bands in Japan and mixed by professional engineers. This was caused by the system being unable to extract the chroma features because the studio version of No. 29 had no chord stroke during the 1st verse, and the bass line between the studio and live versions was also different. To identify a song containing not enough chords, such as No. 29, it is necessary to use the voice melody.

Song No. 4 could not be identified either. Fig. 6 shows that the warping path of No. 4 was non-linear. The key of song No. 4 was changed in the live version because the vocalist in the studio version was male but the live vocalist was female. That change affected the chroma features. The proposed system assumed the same key between the studio and live versions, so No. 4 could not be identified. To solve this problem, it may be beneficial to express the chroma features as a ring buffer rotated through an octave before DTW. However, changes in song key are not frequent because it demands the performer to rearrange the parts for all instruments used in the song.

4. Conclusion

The identification of a song performed at a concert is a task that has not yet become popular among users due to different arrangements or improvisation, unlike audio searches for an exactly matching song. Most research has focused on the identification of cover songs, which is similar to live song identification, except that live songs have issues specific to the performance time or the number of instruments. We examined an identification method that applied DTW to chroma feature series extracted from live songs and measured the linearity of the warping path.

Our test set was smaller than previous studies on cover songs by several orders of magnitude, and future work will collect additional samples.

Acknowledgement

This work was supported by the Japan Society for the Promotion of Science, KAKENHI Grant Number JP20K22193.

References

[1] SoundHound http://www.soundhound.com (Accessed by 1 Dec. 2021)

[2] Shazam https://www.shazam.com (Accessed by 1 Dec. 2021)

[3] L. Marsik, M. Rusek, K. Slaninova, J. Martinovic and J. Pokorny, “Evaluation of Chord and Chroma Features and

Dynamic Time Warping Scores on Cover Song Identification Task," Proc. IFIP International Conference on Computer Information Systems and Industrial Management, Vol.61, pp.205-217, May 2017

[4] J. Serra, X. Serra and R. G. Andrzejak, "Cross recurrence quantification for cover song identification," New Journal of Physics, Vol.10244, No.9, pp.1-18, Sept 2009

[5] Y. Meron, K. Hirose, "Automatic alignment of a musical score to performed music," Vol.22, No.3, pp.189-198, Dec 2001

[6] T. Tsai, T. Pratzlich, M. Muller and J. Martinovic, "Known-Artist Live Song Identification Using Audio Hashprints," IEEE Transactions on Multimedia, Vol.19, No.7, pp. 1569-1582, Feb 2017

[7] G. Doras, G. Peeters, "A Prototypical Triplet Loss for Cover Detection," Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 3797-3801, Apr 2020

[8] librosa, https://librosa.org (Accessed by 1 Apr. 2021)