


Background music monitoring framework and dataset for TV broadcast audio

Hyemi Kim¹  | Junghyun Kim¹ | Jihyun Park¹ | Seongwoo Kim² | Chanjin Park³ | Wonyoung Yoo¹

¹Content Research Division, Electronics and Telecommunications Research Institute, Daejeon, Republic of Korea

²Department of Electronic Engineering, Inha University, Incheon, Republic of Korea

³Department of Computer Engineering, Yonsei University, Wonju, Republic of Korea

Correspondence

Hyemi Kim, Content Research Division, Electronics and Telecommunications Research Institute, Daejeon, Republic of Korea.

Email: miya0404@etri.re.kr

Funding information

Ministry of Culture, Sports and Tourism, Grant/Award Number: CR202104004; Ministry of Culture, Sports and Tourism, Grant/Award Number: CR202104003

Abstract

Music identification is widely regarded as a solved problem for music searching in quiet environments, but its performance tends to degrade in TV broadcast audio owing to the presence of dialogue or sound effects. In addition, constructing an accurate dataset for measuring the performance of background music monitoring in TV broadcast audio is challenging. We propose a framework for monitoring background music by automatic identification and introduce a background music cue sheet. The framework comprises three main components: music identification, music–speech separation, and music detection. In addition, we introduce the Cue-K-Drama dataset, which includes reference songs, audio tracks from 60 episodes of five Korean TV drama series, and corresponding cue sheets that provide the start and end timestamps of background music. Experimental results on the constructed and existing datasets demonstrate that the proposed framework, which incorporates music identification with music–speech separation and music detection, effectively enhances TV broadcast audio monitoring.

KEYWORDS

broadcast monitoring, cue sheet, music detection, music identification, music–speech separation

1 | INTRODUCTION

Fingerprinting for identifying background music in TV broadcast audio is crucial for transparent and accurate royalty distribution [1]. Conventional audio fingerprinting, such as landmark-based audio fingerprinting [2] and differential Hamming-distance-based binary fingerprinting [3], can be used to identify background music. To measure the robustness of music identification, various types of noise or effects, such as pitch shifting or time stretching, have been added to music clips in [2, 3]. However, when measuring the performance of background

music identification in TV broadcast audio clips, the robustness to various deformation effects should be measured, and the background music included in the broadcast audio, such as voices of actors or performers, should also be considered.

To identify background music, the features of the original music must be stored in a database for query. However, the source of background music cannot be added to a database in many cases. For example, (1) a song has not been included in the database because it was recently released; (2) music is used before its official release; and (3) a song is not officially available because

neither the music has been released nor it is expected to be released. Thus, if music not included in the database is used as background music, it cannot be identified, thus reducing the accuracy of the corresponding cue sheet.

We present a background music identification framework that addresses two major problems: (1) The performance of background music identification is hindered by overlapping sounds other than music, and (2) the original music to be searched is not in the database. The framework for identifying background music in broadcast audio comprises three parts. First, we obtain background music audio without the actors' voices by applying music–speech separation, which separates dialogue from music. The identification performance can be enhanced by using the extracted music clip as an input for fingerprinting to identify the background music. To handle music not available in the database but present in TV broadcast audio, we estimate the music segment where the music is used in the entire audio data. If music is detected within a segment, we mark it as unknown despite the segment being predicted to be music. This unknown segment presents an opportunity to improve the accuracy of the final cue sheet by reidentifying it after supplementing the database or manually labeling the music track.

To accurately evaluate the performance of background music monitoring using our framework, we must collect information on the usage of background music in TV broadcast audio, including start and end timestamps as well as metadata, including the song titles. In addition, the original music files inserted into the broadcast audio are required to build a database for searching for music. The TVSM dataset [4] provides music or speech segments and is useful for predicting such segments, being the first open-source large dataset for speech and music detection. The OpenBMAT dataset [5] provides annotations for music detection, containing over 27 h of TV broadcast audio from four countries. It includes annotations of the loudness of music in relation to other simultaneous non-music sounds. The Podcast dataset [6] can be used for music–speech separation, providing music and speech audio clips both separately and mixed. However, none of these datasets are available for background music identification because they do not include the metadata of the original music for processing in broadcast audio. The Divide and Remix dataset [7] is built by synthesizing mixtures with music from the Free Music Archive [8], speech from the LibriSpeech corpus [9], and sound effects from the Freesound Dataset 50K [10]. The Divide and Remix dataset aims to provide research support for source separation. In addition, it contains music metadata embedded in the mixture, thus enabling music identification. However, the provided audio is not sourced from real TV

broadcast recordings, and the duration of all audio clips is limited to 1 min, which substantially differs from the lengths typically found in real TV broadcast audio.

Several obstacles hinder the collection of background music to obtain a cue-sheet dataset. The cue sheet used for royalty distribution is produced such that an annotator listens to the broadcast audio, determines the title of the music track, and records the period of the music, but the annotations have limited accuracy. While it is possible to find popular songs, it is difficult to accurately identify less known songs, such as library music without vocals. Even for a popular song, annotating the exact start and end timestamps is difficult because of effects such as fade-in and fade-out. Additionally, music that has not been officially released accounts for a high proportion of background music. Unlike TV shows, which mainly use popular songs appropriate for every scene, it is common to compose background music suitable for a program and release only the music tracks with high commercial potential in TV dramas. Therefore, excluding those directly involved in production, it is challenging to search and build a database of all original background music sources contained in TV broadcast audio. The first public dataset designed for broadcast monitoring, called the BAF dataset [1], contains 57 h of TV broadcast audio recordings from 203 TV channels of 23 countries, with 2000 production music tracks composing the reference set. However, to build a comprehensive search database, they extracted only 1 min excerpts containing the embedded music rather than using the entire audio recordings from the broadcast. Additionally, multiple annotators listened to the query and reference pairs and provided annotations, resulting in label variability. In fact, inconsistencies occurred among the three annotators regarding the presence or absence of music in various segments. In addition, the start and end timestamps differed for the same music clip. Because of such inconsistencies, accurate and reliable data were difficult to obtain.

We constructed a background music cue-sheet dataset for TV broadcast audio using an alternative approach to conventional annotation-based methods. TV shows primarily use released or popular songs, whereas TV dramas often incorporate unique background music, including production music tailored to the scenes, theme music for actors, and original soundtracks. Our Cue-K-Drama dataset was constructed by collecting and processing raw data accumulated through a series of operations involving the composition and editing of drama background music. The data were sourced directly from music production works in which background music was inserted into TV dramas.

In addition to the insertion of background music, another crucial postproduction process is the incorporation of sound. Although raw data from music production

are valuable for generating accurate cue sheets, music production companies typically possess audio tracks before including various sound effects. Hence, there may be some differences between the raw data and final audio for broadcasting.

To create a dataset that included TV broadcast audio with sound effects and the corresponding cue sheets, we aligned two similar audio clips. One clip was a pseudo-broadcast audio using raw data from music production, whereas the other clip was the actual TV broadcast audio with sound effects. We measured the similarity between these audio clips to perform alignment. Using the alignment results, we generated a TV broadcast audio and background music cue-sheet dataset by synchronizing the timestamps on the pseudo-cue sheet with the aligned audio. This ensured that the cue sheet accurately corresponded to specific moments in the TV broadcast audio that includes sound effects.

2 | BACKGROUND MUSIC MONITORING FRAMEWORK

The proposed framework for monitoring background music in TV broadcast audios consists of three parts. (1) Music is identified using audio fingerprinting. However, music is often combined with dialogue, being background instead of foreground music. Consequently, music cannot be identified using fingerprinting in many cases. In addition, music identification results obtained from fingerprinting requires retrieval from a database. However, if original music, which frequently occurs in broadcast audio, is not included in the music search database and used as background music, it cannot be identified. To address these limitations, (2) we use a music–speech separation method to extract the music components. The music-only audio clip can be used as an input for audio fingerprinting, thereby enhancing the performance of music identification. (3) We predict a music segment and can label a segment as “unknown” if the corresponding track cannot be identified despite containing music. This provides an opportunity to increase the accuracy of the cue sheet by reidentification after the database is expanded or the clip is manually labeled.

2.1 | Music identification

For music identification, fingerprints of the music clips are extracted and stored in a database along with metadata such as titles, composers, and singers. Then, the fingerprint of the query music clip is compared with all the fingerprints stored in the database.

Audio fingerprinting can be divided into conventional methods for extracting fingerprints from the spectrogram obtained by the frequency transform of the audio signal [2, 3, 11, 12] and neural network-based methods for extracting the embedding vectors [13–15]. Conventional methods allow to determine whether two pieces of music are the same by setting a threshold. On the other hand, in neural network-based methods, the most similar music is retrieved from a database. This leads to the inference of similar songs even in segments without background music. Consequently, these methods cannot be applied to broadcast monitoring.

We identify music by applying audio fingerprinting [16] represented as differential binary hashes, which is an improved method for music identification [3]. The input audio is divided into various frames. The input audio lasts 3 s with a 1 s stride. The results of audio fingerprinting within the initial 5 s are removed as outliers. Adjacent frames with the same music ID are concatenated. Specifically, if multiple occurrences of the same music ID within a specific time interval, such as 10 s, occur and the time intervals of both the results and position in the reference song match, we merge the frames into a single music segment. A music segment is extended by iterative merging until a different music ID or time interval occurs.

2.2 | Music–speech separation

To enhance background music identification in broadcast audio, we use an audio signal containing only background music. This is achieved by removing voices through music–speech separation and using the resulting audio as the input for music identification. In [17], dialog separation in real-world broadcast audio can be adequately solved by using transfer learning from music source-separation methods, such as Open-Unmix [18], Spleeter [19], and Demucs [20].

We use the LaSAFT model [21] for music source separation. When measuring the source-to-distortion ratio (SDR), which is widely used to evaluate music source separation, on the MUSDB18 dataset [22], which is the most representative dataset for music source separation, LaSAFT shows a lower overall performance than Demucs [20] but a much higher performance for singing voice separation. In addition, LaSAFT outperforms Demucs speech separation of a voice that is similar to a singing voice [23] from broadcast audio.

We evaluated music identification using a pretrained model [23] on a TV broadcast dataset with and without music–speech separation. LaSAFT was adopted given its robust music–speech separation. We used the mean squared error of the spectrogram for the training loss and

the l_1 loss of the waveform for the validation loss. We used the music–speech separation dataset from [24]. It contains the labels of music segments, including background music, and speech segments, including only speech without music, using the Praat tool (<https://www.fon.hum.uva.nl/praat/>) from Korean TV broadcast audio files of various genres. The dataset contains 100 h of audio data and 12 h of speech audio by concatenating all speech-only segments. Music excerpts were collected from Korean and popular foreign songs in various genres. The dataset was divided into training, validation, and test sets. Each of the splits consisted of 1823 excerpts of songs with a duration of 12 s per excerpt, 1823 speech excerpts of the same duration, and their mixtures. During training, music and speech signals were mixed with an arbitrary signal-to-noise ratio (SNR) ranging from -30 to 0 dB. The mixed signals in the validation set were composed of two cases to analyze the music identification performance according to the music-to-speech level of 0 and -10 dB. The music signal was mixed with the speech signal at the same average loudness at 0 dB, while the loudness of the music signal was relatively small at -10 dB.

During validation for selecting the music–speech separation model, we considered the performance according to the SNR. When the SNR of music and speech is 0 dB, the music identification performance is higher than that at -10 dB because the average loudness of music is sufficiently large to identify music. Assuming that the overall music identification performance can be further improved by increasing the separation performance at -10 dB, we compared the results by using the models (1) with the highest overall average SDR (Avg-best) and (2) highest average SDR at -10 dB (Low-best). The music–speech separation experiments are described in Section 4.1.

2.3 | Music detection

Music detection supplements identification by extracting music clips that are not present in the reference database. During music detection, if a segment is identified as containing music but no music identification result is obtained, the segment is included in the result and marked with ID unknown.

Pretrained audio neural networks [25] have been modeled using various convolutional neural networks for audio tagging and sound event detection using weak labels from the AudioSet dataset [26], which consists of 527 sound classes extracted from YouTube videos. The convolutional neural network in [27] is used for music detection in broadcast audio.

In this study, we used a U-Net architecture with limited upsampling for music detection [28]. This architecture

is more suitable for classification than for segmentation. As the training set, we used the music–speech dataset described in Section 2.2 that includes annotations of music segments from 100 h of TV broadcast audio in a variety of genres.

The results of music identification and detection are combined by incorporating the detected music segments into a list of music segments obtained from music identification. If a detected music segment does not overlap with any of the music identification results, it is added to the list as an additional segment.

3 | DATASET FOR TV BROADCAST AUDIO

We constructed a cue-sheet dataset for monitoring background music in TV broadcast audio using raw data from music production. We randomly selected 60 episodes from five Korean dramas that aired in Korea between 2017 and 2020. The total duration of the broadcast audio data was approximately 61 h, with each episode having a duration of approximately 1 h, which is a typical broadcast length.

3.1 | Music and speech audio tracks

We collected raw data accumulated by inserting background music from a music production company in charge of composing and editing background music for TV dramas. The raw data were accumulated using Vegas, a professional video editing software widely used in the broadcasting industry. The Vegas project file contained several audio tracks, including the speech track edited from audio recorded on the location where the actors performed, a number of music or individual instrumental tracks with various effects, and the entire background music track mixed with several music tracks of the same duration as the speech track.

We extracted the background and speech tracks using the Vegas software and mixed them with various SNR values. All the tracks were rendered as monaural audio files at a sampling rate of 11,025 Hz.

3.2 | Reference music dataset for matching

To identify background music, the fingerprints of the original music used in the background were extracted and stored. Music productions contained all songs used as background music, ensuring that no songs were missing when constructing the original music search

database. However, numerous duplicate songs were encountered. If duplicate songs are included in the search database, the performance cannot be accurately evaluated. Duplicate removal techniques [29] have been applied to the SoundDesc dataset [30], which includes a collection of audio recordings and sound effects sourced from the BBC Sound Effects Archive. Similarly, we identified the factors that caused duplication in the collected original music sources and refined them to ensure that there were no duplicates according to various criteria.

First, if there were separate instrumental sources from the same song in the database, the instrumental source files were removed, and only the final mixed music was retained. When inserting background music, some of the released music is used as is. However, it is common for a composer to use only some of the tracks for each instrument. For example, even if the same song is used, only melodic instruments can be used, or a track, such as a vocal or drum, can be removed depending on the scene. To handle instrument tracks flexibly, composers store individual tracks for each instrument included in one original piece of music as separate files. However, copyright is granted to only one final piece of music with all the tracks mixed. Because a broadcast monitoring system database is intended to search for copyrighted songs, the reference music dataset was refined to include only one song in which all instruments were mixed rather than including each instrumental source track of songs.

Next, we removed additional pitch-shifted or time-stretched music files. Both pitch shifting and time stretching are commonly used to insert background music into TV broadcast audio. Pitch shifting involves changing the music key, whereas time stretching involves lengthening or shortening the music by changing the tempo. Because of these effects, the energy of the audio signal changes according to the frequency band, and a substantial change occurs in the feature vector of music. Nevertheless, listeners can easily recognize the original song. The identification of sound effects applied to the collected audio files enabled easy identification and removal of duplicate songs. Then, we included only the original songs in the reference music dataset.

Finally, instances of duplicated songs in which only a part of the original song was edited to reduce the audio length were removed. To this end, an automatic duplicated song removal method was implemented using the audio fingerprinting approach in [16]. The fingerprints of all other music files were compared for each piece of music. If a song shared more than 90% of similarity with another song of a shorter duration, it was considered as a duplicate song.

After duplicate removal, 723 songs were included in the reference music dataset.

3.3 | Pseudo-broadcast cue-sheet dataset based on music identification and detection

To generate the background music cue sheet without relying on annotations, we first applied the music identification method described in Section 2.1. Using fingerprint-based music identification and inputting a background music track instead of a mixture of music and speech, a relatively accurate initial cue sheet was obtained. The reference music dataset described in Section 3.2 was used for search.

Although using background music tracks that contain only music can be helpful, it may not always guarantee accurate music identification owing to the application of various audio effects. To obtain accurate cue sheets, all music segments must be correctly identified. Next, non-silent segments should be extracted from the background music track to establish the ground truth for the music segments. To this end, we used the `librosa.effect.split` function from the `librosa` Python package for music and audio analysis. This function enabled the accurate extraction of music segments.

The background music identification results were compared with the correct music segments. For segments in which music was not identified despite being confirmed as a music segment, the song ID was directly inserted into the cue sheet by referring to the music list from the Vegas project. Most of the unidentified cases were due to severe pitch shifting or time stretching applied to the music. In addition, we detected instances in which music was correctly identified but the start or end timestamp differed owing to fade-in/fade-out effects. To improve the accuracy of the cue sheet, the timestamps in the music identification results were adjusted using the ground-truth music segments. This ensured that the cue sheet reflected the correct timing of music in the TV broadcast audio. The dataset obtained through this process was called the pseudo-broadcast cue-sheet dataset. In the next section, we describe the differences between this dataset and the final broadcast cue-sheet dataset.

3.4 | Cue-K-Drama: TV broadcast audio and cue-sheet dataset

In TV broadcast audio, in addition to background music, various sound effects are added during postprocessing. These effects include sword swings, clashes, tension building, footsteps, falling or breaking, and car engines. Similar to dialogue, the sound effects can interfere with background music identification. The cue-sheet dataset created with pseudo-broadcast audio does not include

sound effects. Therefore, it cannot be considered a dataset that accurately represents TV broadcast audio.

In terms of TV broadcast monitoring, pseudo-broadcast audio can be similar to TV broadcast audio because both have the same background music. However, some differences occur between the two types of audio clips. First, a time difference may occur in the start time of background music. Some differences were found when comparing the pseudo-broadcast audio with the corresponding TV drama series downloaded from the website. For example, the presence of age rating disclaimers or open credits may vary. Additionally, certain parts of the pseudo-broadcast audio may be deleted from the TV broadcast audio. As a result, the start and end timestamps of background music in TV broadcast audio can be shifted with respect to pseudo-broadcast audio. These differences in timing can affect the accuracy of the cue-sheet dataset, and timestamps should be aligned accordingly.

To create a cue-sheet dataset that accurately represents TV broadcast audio for monitoring, we created the Cue-K-Drama dataset based on the pseudo-broadcast cue-sheet dataset. To align the cue-sheet dataset with actual TV broadcast audio, TV broadcast audio recordings were collected. In addition, the start and end timestamps of the background music cue sheet were adjusted by aligning the TV broadcast audio recordings with the audio files in the pseudo-broadcast cue-sheet dataset.

To align the two audio signals as shown in Figure 1, the pairs of corresponding time positions were estimated using functions `librosa.chroma_stft` and `librosa.sequence.dtw`. The time differences of the corresponding time positions were then calculated to generate a time-difference graph along the time axis, as shown in the upper part of Figure 2. We obtained a histogram of the differences in intervals of 1 s. Bins with values above a threshold were selected as candidates for

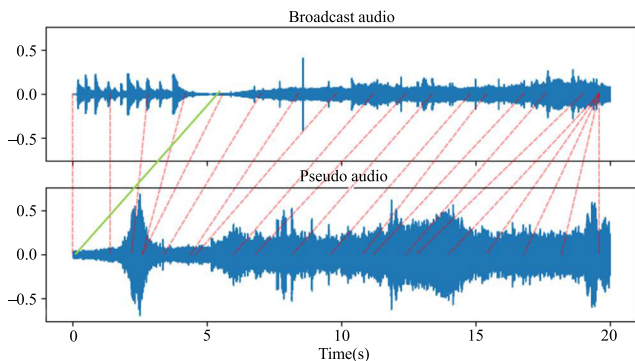


FIGURE 1 Estimated corresponding time positions in two audio signals (red dashed lines) and ground-truth start time position of first occurrence of a background music track (green line). At the beginning of the audio, additional elements are included, such as age rating disclaimers or opening credits.

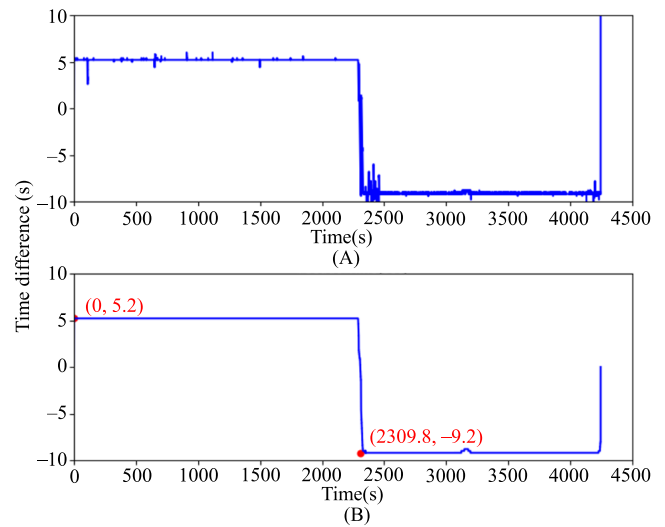


FIGURE 2 Time offset between two aligned audio clips (A) before and (B) after filtering. The red dots indicate the shifting positions.

shifting positions, and the corresponding bin was set as an integer unit shifting time. A median filter was used to mitigate noise. We divided the data into segments with a constant time difference along the time axis and calculated the start and end time positions of the segments. The time shift value was set to a floating number for averaging the time difference, as shown in the lower part of Figure 2. We compared the corresponding shifting values with the music segment results from the pseudo-broadcast cue-sheet dataset. By converting the time value into a time shift, a reference cue sheet for an actual TV broadcast audio was obtained.

4 | EXPERIMENTAL RESULTS

The pseudo-broadcast cue-sheet and cue-K-Drama datasets described in Section 3 contain reference songs, audio tracks from 60 episodes of five Korean series, and the corresponding cue sheets. They share reference songs, but their audio tracks and cue sheets differ. The audio tracks of the pseudo-broadcast cue-sheet dataset consisted of speech tracks recorded onsite, background music tracks, and their mixture. The cue-sheet dataset was built from the accumulated raw data. The Cue-K-Drama dataset was constructed using TV broadcast audio recordings, and the corresponding cue sheets were built by modifying the cue sheets in the pseudo-broadcast cue-sheet dataset through alignment with pseudo-broadcast audio clips. In total, 723 distinct songs were used in the search. We evaluated the background music monitoring framework described in Section 2 on the Cue-K-Drama dataset. We conducted experiments to assess the impact of music-speech separation on the music identification

performance. We also performed experiments to detect songs that were not present in the reference dataset, demonstrating the potential for improving the music identification performance based on the music detection results.

4.1 | Performance of music–speech separation

Table 1 lists the SDR of the Avg-best and Low-best models described in Section 2.2 on the music–speech separation dataset [24]. When mixed audio at SNR values 0 and -10 dB was separated into music and speech, the model with the highest mean performance at 0 and -10 dB achieved an SDR of 8.444. On the other hand, the model with a slightly lower average performance but the highest performance at -10 dB achieved an SDR of 7.107.

The music–speech separation results for the pseudo-broadcast audio dataset using the two models are listed in Table 2. We used the common source-separation metrics of SDR, source-to-interference ratio (SIR), and source-to-artifacts ratio (SAR) [31] provided by *mir_eval* [32]. As the music SNR decreased, the SDR of the separated music audio also decreased. However, there were differences in the separation performance depending on the model. When comparing the SDR results between 0 and -10 dB, the Low-best and Avg-best models demonstrated better separation performance at -10 and 0 dB, respectively.

4.2 | Performance of music identification

To measure the music identification performance for broadcast monitoring, we calculated the average precision, recall, and F1-score across all the episodes by

TABLE 1 SDR of music for music–speech separation.

Model	0 dB	-10 dB	Average
Avg-best	10.157	6.731	8.444
Low-best	9.751	7.107	8.429

Note: The values for the best performance are marked in bold.

TABLE 2 Performance of music–speech separation on pseudo-broadcast audio dataset.

Music SNR	Model	Music			Speech		
		SDR	SIR	SAR	SDR	SIR	SAR
0 dB	Avg-best	8.08	24.90	8.28	11.15	17.02	12.78
	Low-best	7.23	26.27	7.37	10.33	15.43	12.31
-10 dB	Avg-best	4.77	17.86	5.35	18.69	29.32	19.25
	Low-best	5.29	19.60	5.77	19.10	28.79	19.86

Note: The values for the best performance are marked in bold.

comparing the estimated and reference cue sheets. We used the metrics available for the BAF dataset [1] at <https://github.com/guillemcortes/baf-dataset>.

During broadcast monitoring for automatically producing a background music cue sheet, estimating accurate start and end times is necessary to provide information on the background music duration. However, owing to the common fade-in/fade-out effects, the start or end time may not be accurately predicted. In addition, the most important aspect of distributing copyright royalties is identifying all correct songs inserted in broadcast audio. Thus, we measured the performance by considering the unique annotation to determine the correct song identification within a music segment without considering the identified period. For instance, if a song is used in two segments and identified in one segment but not in another segment, the precision becomes 100% and the recall becomes 50% regardless of the overlapping in the identification segments.

The fingerprinting method [16] described in Section 2.1 was used for music identification.

4.2.1 | Performance on Cue-K-Drama dataset

The background music identification results for the Cue-K-Drama dataset are listed in Table 3. The F1-score of background music identification after applying music–speech separation increased from 88.73% to 90.11% when the Avg-best model was applied and to 90.52% when the Low-best model was applied. The model that achieved a higher separation performance at -10 dB was more effective in improving the performance of background music identification.

When identifying background music in TV broadcast audio, the precision was 95.17%, and the recall was 83.11%. Hence, the list of predicted songs was relatively correct, but no identification occurred in many cases although music was present. Because the recall could be considerably improved, analyzing the factors that reduced it may greatly contribute to improving the music identification performance.

4.2.2 | Performance on BAF dataset

Table 4 lists the effects of music–speech separation on the performance improvement of music identification for the BAF dataset [1]. All precision values were over 96%, indicating a high performance, whereas the recall was very low at 62.17% for the broadcast audio. When applying music–speech separation, the recall substantially increased to 83.22% and 85.32%. The F1-score of 90.69% indicated the best performance when the Low-best model was applied.

The experimental results indicated that the music identification performance for both datasets was improved by applying music–speech separation. For the BAF dataset [1], the recall was very low at 62.17%, whereas it increased to 85.32% after applying music–speech separation, approaching a recall of 86.18% on the Cue-K-Drama dataset.

4.2.3 | Performance according to music SNR

Table 5 lists the music identification results according to the loudness of music relative to speech in a pseudo-

TABLE 3 Performance of background music identification on Cue-K-Drama dataset.

Input	Precision	Recall	F1-score
Mixture	0.9517	0.8311	0.8873
Separated (Avg-best)	0.9522	0.8553	0.9011
Separated (Low-best)	0.9531	0.8618	0.9052

Note: The values for the best performance are marked in bold.

TABLE 4 Performance of background music identification on BAF dataset [1].

Input	Precision	Recall	F1-score
Mixture	0.9657	0.6217	0.7564
Separated (Avg-best)	0.9662	0.8322	0.8942
Separated (Low-best)	0.9679	0.8532	0.9069

Note: The values for the best performance are marked in bold.

TABLE 5 Background music identification performance according to music SNR on pseudo-broadcast cue-sheet dataset (%) (P, precision; R, recall).

Music SNR	Mixture			Separated (Avg-best)			Separated (Low-best)			Music (upperbound)		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
0 dB	98.78	87.44	92.76	98.83	88.35	93.30	98.81	89.01	93.66	98.84	91.58	95.07
−6 dB	98.78	83.24	90.35	98.87	86.53	92.29	98.85	86.98	92.53	98.85	91.35	94.95
−12 dB	98.83	77.80	87.07	98.87	80.60	88.81	98.88	82.62	90.02	98.96	91.08	94.86
−20 dB	98.92	62.74	76.78	98.68	62.41	76.46	98.83	66.60	79.57	99.11	88.43	93.47

Note: The values for the best performance are marked in bold.

broadcast cue-sheet dataset. The two signals were mixed with SNRs of 0, −6, −12, and −20 dB. SNR values below zero indicate that the loudness of the music is smaller than that of speech. The average root-mean-square (RMS) value of the entire pseudo-broadcast audio data was calculated excluding silent frames, which can distort the average value. We used function `librosa.feature.rms` to compute the RMS value for each audio frame. The SNR was calculated as follows:

$$\text{SNR} = 20 \log \frac{\text{RMS}_{\text{music}}}{\text{RMS}_{\text{speech}}}. \quad (1)$$

The precision values ranged from 98.7% to 98.9%, and no substantial difference was observed in the performance according to the SNR. When the SNR decreased, the recall and F1-score also decreased. Figure 3 shows that the F1-score for music identification increased after applying separation with the Avg-best model, and it decreased slightly at −20 dB when the loudness of music was very small. In contrast, the recall and F1-score after applying separation were the highest for the Low-best model. In addition, the F1-score increased from 76.78% to 79.57% with the Low-best model even at −20 dB, when the Avg-best model failed to improve the F1-score.

Even when identifying music using a clear music track without speech as the input, the maximum F1-score reached 95.07%, indicating the potential for performance improvement beyond speech removal. Additionally, even when music was not mixed with speech, its performance deteriorated slightly when the absolute loudness decreased.

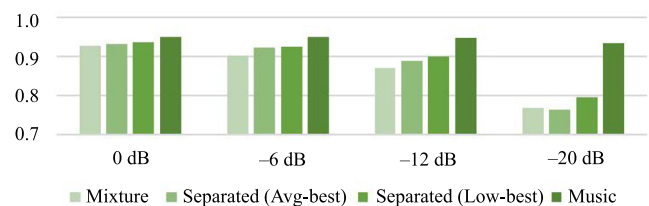


FIGURE 3 F1-score for music identification according to music SNR.

TABLE 6 Performance of background music identification with music detection on Cue-K-Drama dataset.

	Mixture			Separated (Avg-best)			Separated (Low-best)		
	P	R	F1	P	R	F1	P	R	F1
Without detection	0.9533	0.8059	0.8734	0.9523	0.827	0.8853	0.9537	0.8339	0.8898
With detection	0.9433	0.8268	0.8812	0.9432	0.8479	0.893	0.9446	0.8548	0.8975

Note: The values for the best performance are marked in bold.

4.2.4 | Performance with music detection

Music detection is necessary to obtain accurate results in monitoring TV broadcast background music by marking all segments, including those for which music identification fails. When a segment is detected as having music but music is not identified, the unknown ID is assigned to that segment. An unknown segment provides an opportunity to improve the accuracy of the cue sheet by reidentifying that segment after enlarging the database or manually annotating the corresponding music data.

To measure the performance including the unknown segments, we intentionally removed 5% of the songs from the reference song dataset and revised the corresponding IDs to unknown on the reference cue sheet. The music detection results were then combined with the music identification results from the reduced reference song dataset. If no song was identified in the music segment, it was added to the music identification result and marked as unknown. The minimum length of the music segments to be added was set to 5 s.

As listed in Table 6, the precision slightly decreased, possibly because the non-music segment was detected as a music segment and marked as unknown. In contrast, the recall and F1-score increased by detecting more music segments that were not detected because the reference song was not included in the dataset. This confirms that if music cannot be identified because the reference is not in the dataset for searching, the music identification performance can be improved by expanding the dataset.

4.3 | Performance of music activity detection

The Cue-K-Drama dataset includes the timestamps of the music segments and music IDs. Hence, it can be used to measure the music detection performance by estimating segments with background music in mixed audio of music and speech.

To measure the music detection performance, we used the segment-based metric provided by *sed_eval* [33], a library related to sound event detection. All the titles in the cue sheet were identified with the same music class, and the two results were compared on a fixed time

TABLE 7 Performance of music activity detection on Cue-K-Drama dataset.

Input	Precision	Recall	F1-score
TV broadcast audio	0.9483	0.9295	0.9384

window *time_resolution*, which was the desired segment length for evaluation and set to 1 s. The U-Net architecture with limited upsampling [28] was used for music detection. As listed in Table 7, the precision and recall of detection reached 94.83% and 92.95%, respectively.

5 | CONCLUSIONS

We introduce a monitoring framework for broadcast background music that includes separation, identification, and detection by separating music and speech, identifying the extracted music signals, and detecting music segments when music identification fails. In addition, we created the Cue-K-Drama dataset with TV broadcast audio and cue-sheet data for this framework. The dataset can be used to accurately measure the performance of monitoring background music in TV broadcast audio. The dataset was constructed by collecting TV broadcast audio recordings and generating cue sheets using raw data from a music production company. This overcomes the limitations of relying on annotations. In addition to enable monitoring of background music, the dataset can be used to measure the music detection performance. Moreover, the pseudo-broadcast music and speech audio dataset generated during the creation of the dataset can be applied to measure the music–speech separation performance.

We conducted experiments to evaluate the performance of music identification with and without music–speech separation on the Cue-K-Drama and BAF datasets [1]. It was confirmed that music–speech separation improves the music identification performance. In addition, a model with a high score at a low music SNR provides a higher performance when there is a negligible difference in the overall performance.

When specific songs are not included in the reference music dataset, the added music detection allows to increase the accuracy of the estimated cue sheet by

reidentifying it after enlarging the database with missing song data.

There was a considerable difference between precision and recall. When music identification with separated music audio achieved an F1-score of 90.52%, the precision was very high at 95.31%, and the recall was low at 86.18% on the Cue-K-Drama dataset. In future work, we will analyze failure cases and investigate methods to improve the recall.

ACKNOWLEDGEMENTS

This research was supported by the Culture, Sports, and Tourism R&D Program through a Korea Creative Content Agency grant funded by the Ministry of Culture, Sports and Tourism in 2023 (Project: Development of high-speed music search technology using deeplearning, No. CR202104004, Contribution Rate: 50%, Project: Development of artificial intelligence-based copyright infringement suspicious element detection and alternative material content recommendation technology for educational content, No. CR202104003, Contribution Rate: 50%).

CONFLICT OF INTEREST STATEMENT

The authors declare that there are no conflicts of interest.

ORCID

Hyemi Kim  <https://orcid.org/0000-0002-3446-3498>

REFERENCES

- G. C. Sebasti, Ciurana, E. Molina, M. Miron, O. Meyers, J. Six, and X. Serra, *BAF: an audio fingerprinting dataset for broadcast monitoring*, (Proc. 23rd Int. Soc. Music Inf. Retr. Conf., Bengaluru, India), 2022, pp. 908–916.
- A. Wang, *An industrial-strength audio search algorithm*, (Proc. Int. Conf. Music Inf. Retr., Baltimore, USA), 2003, pp. 7–13.
- J. Haitsma and T. Kalker, *A highly robust audio fingerprinting system*, (Proc. Int. Soc. Music Inf. Retr. Conf., Paris, France), 2002, pp. 107–115.
- Y.-N. Hung, C.-W. Wu, I. Orife, A. Hipple, W. Wolcott, and A. Lerch, *A large TV dataset for speech and music activity detection*, *EURASIP J. Audio Speech Music Process.* **2022** (2022), no. 21, 1–12.
- B. Melendez-Cataln, E. Molina, and E. Gomez, *Open broadcast media audio from TV: a dataset of TV broadcast audio with relative music loudness annotations*, *Trans. Int. Soc. Music Inform. Retrieval* **2** (2019), no. 1, 43–51.
- N. Schmidt, J. Pons, and M. Miron, *PodcastMix: a dataset for separating music and speech in podcasts*, (Proc. Interspeech, Incheon, Republic of Korea), 2022, pp. 231–235.
- D. Petermann, G. Wichern, Z.-Q. Wang, and J. L. Roux, *The cocktail fork problem: three-stem audio separation for real-world soundtracks*, (IEEE Int. Conf. Acoust. Speech Signal Process. IEEE, Singapore), 2022, pp. 526–530.
- M. Defferrard, K. Benzi, P. Vandergheynst, and X. Bresson, *FMA: a dataset for music analysis*, (18th Int. Soc. Music Inf. Retr. Conf. (ISMIR), Suzhou, China), 2017, pp. 316–323.
- V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, *Librispeech: an ASR corpus based on public domain audio books*, (IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP), IEEE, South Brisbane, Australia), 2015, pp. 5206–5210.
- E. Fonseca, X. Favory, J. Pons, F. Font, and X. Serra, *FSD50K: an open dataset of human-labeled sound events*, *IEEE/ACM Trans. Audio, Speech, Lang. Process.* **30** (2022), 829–852.
- D. Ellis, *The 2014 labrosa audio fingerprint system*, (Int. Soc. Music Inf. Retr. Conf., Taipei, Taiwan), 2014.
- J. Six, *Panako: a scalable audio search system*, *J. Open Source Softw.* **7** (2022), no. 78, 4554.
- B. A. Arcas, B. Gfeller, R. Guo, K. Kilgour, S. Kumar, J. Lyon, J. Odell, M. Ritter, D. Roblek, M. Sharifi, and M. Velimirovic, *Now playing: continuous low-power music recognition*, (Proc. NeurIPS 2017 Workshop Mach. Learn. Phone Other Consum. Devices, Long Beach, CA, USA), 2017, pp. 1–6.
- A. Bez-Surez, N. Shah, J. A. Nolasco-Flores, S.-H. S. Huang, O. Gnawali, and W. Shi, *SAMAF: sequence-to-sequence autoencoder model for audio fingerprinting*, *IEEE Int. Conf. Acoust. Speech Sig. Process.* **16** (2021), no. 2, 1–23.
- S. Chang, D. Lee, J. Park, H. Lim, K. Lee, K. Ko, and Y. Han, *Neural audio fingerprint for high-specific audio retrieval based on contrastive learning*, (Proc. IEEE Int. Conf. Acoust. Speech Signal Process. IEEE, Toronto, Canada), 2021, pp. 3025–3029.
- J. S. Seo, J. Kim, and H. Kim, *Audio fingerprint matching based on a power weight*, *J. Acoust. Soc. Korea* **38** (2019), no. 6, 716–723.
- M. Strauss, J. Paulus, M. Torcoli, and B. Edler, *A hands-on comparison of DNNs for dialog separation using transfer learning from music source separation*, (Proc. Interspeech 2021, Brno, Czech Republic), 2021, pp. 3900–3904.
- F.-R. Stöter, S. Uhlich, A. Liutkus, and Y. Mitsufuji, *Open-Unmix—a reference implementation for music source separation*, *J. Open Source Softw.* **4** (2019), no. 41, 1667.
- R. Hennequin, A. Khlif, F. Voituret, and M. Moussallam, *Spleeter: a fast and efficient music source separation tool with pre-trained models*, *J. Open Source Softw.* **5** (2020), no. 50, 2154.
- A. Defossez, N. Usunier, L. Bottou, and F. Bach, *Music source separation in the waveform domain*, arXiv preprint, 2019, DOI [10.48550/arXiv.1911.13254](https://doi.org/10.48550/arXiv.1911.13254)
- W. Choi, M. Kim, J. Chung, and S. Jung, *LaSAFT: latent source attentive frequency transformation for conditioned source separation*, (Proc. IEEE Int. Conf. Acoust. Speech Signal Process. IEEE, Toronto, Canada), 2021, pp. 171–175.
- Z. Rafii, A. Liutkus, F.-R. Stoter, S. I. Mimilakis, and R. Bittner, *MUSDB18—a corpus for music separation*, 2017.
- H. Kim, J. Kim, and J. Park, *Performance analysis for background music identification in TV contents according to state-of-the-art music source separation methods*, (Proc. Korea Multimedia Society, Seoul, Korea), 2021, pp. 30–32.
- H. Kim, W.-H. Heo, J. Kim, and J. Park, *Monaural music-speech source separation based on convolutional neural network for background music identification in TV shows*, *J. Korean Inst. Commun. Inform. Sci.* **45** (2020), no. 5, 855–866.
- Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, *PANNs: large-scale pretrained audio neural networks for audio pattern recognition*, *IEEE/ACM Trans. Audio, Speech, Lang. Process.* **28** (2020), 2880–2894.
- J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, *Audio set: an ontology and human-labeled dataset for audio events*, (IEEE

- Int. Conf. Acoust. Speech Signal Process. (ICASSP), IEEE, New Orleans, USA), 2017, pp. 776–780.
27. B.-Y. Jang, W.-H. Heo, J. Kim, and O.-W. Kwon, *Music detection from broadcast contents using convolutional neural networks with a Mel-scale kernel*, EURASIP J. Audio Speech Music Process. **2019** (2019), no. 11, 1–12.
 28. S. Lee, H. Kim, and G.-J. Jang, *Weakly supervised u-net with limited upsampling for sound event detection*, Appl. Sci. **13** (2023), no. 11.
 29. B. Weck and X. Serra, *Data leakage in cross-modal retrieval training: a case study*, (ICASSP 2023—2023 IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP), Rhodes Island, Greece), 2023, pp. 1–5.
 30. A. S. Koepke, A.-M. Oncescu, J. Henriques, Z. Akata, and S. Albanie, *Audio retrieval with natural language queries: a benchmark study*, IEEE Trans. Multimed. **25** (2022), 2675–2685.
 31. E. Vincent, R. Gribonval, and C. Fevotte, *Performance measurement in blind audio source separation*, IEEE Trans. Audio Speech Lang. Process. **14** (2006), no. 4, 1462–1469.
 32. C. Raffel, B. McFee, E. J. Humphrey, J. Salamon, O. Nieto, D. Liang, and D. P. W. Ellis, *mir_eval: a transparent implementation of common MIR metrics*, (Proc. Int. Soc. Music Inf. Retr. Conf., Taipei, Taiwan), 2014, pp. 367–372.
 33. A. Mesaros, T. Heittola, and T. Virtanen, *Metrics for polyphonic sound event detection*, Appl. Sci. **6** (2016), no. 6, 1–17.

AUTHOR BIOGRAPHIES



Hyemi Kim received the BS degree in Electrical Engineering from Pusan National University, Busan, Republic of Korea, in 2004, and the MS degree in Electrical Engineering from the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Republic of Korea, in 2006. She is a principal researcher at the Content Research Division, Electronics and Telecommunications Research Institute (ETRI), Daejeon, Republic of Korea, and a PhD candidate at the Kim Jaechul Graduate School of Artificial Intelligence, KAIST. Her research interests include music information retrieval technologies such as audio fingerprinting, music identification, source separation, and automatic music transcription.



Junghyun Kim received the BS and MS degrees in Computer Science from Chonnam National University, Gwangju, Republic of Korea, in 1999 and 2001, respectively. Since 2001, she has been a member of the research staff at ETRI, Daejeon,

Republic of Korea. Her research interests include music similarity and retrieval for copyright protection.



Jihyun Park received the BS and MS degrees in Computer Science from Sogang University, Seoul, Republic of Korea, in 1997 and 1999, respectively, and PhD degree in Computer Engineering from Chungnam National University in 2010.

Since 1999, he has been with ETRI, where he is a principal researcher in the Content Research Division. His research interests include music information retrieval and copyright technologies for digital content.



Seongwoo Kim received the BS degree in Electronics Engineering from Inha University, Incheon, Republic of Korea, in 2023. His research interests include audio signal processing and music information retrieval.



Chanjin Park is expected to receive the BS degree in Computer Engineering from Yonsei University, Wonju, Republic of Korea, in February 2024. His research interest include speech and audio signal processing.



Wonyoung Yoo received the BS, MS, and PhD degrees in electronics engineering from Jeonbuk National University, Republic of Korea, in 1996, 1998, and 2003, respectively. In 2001, he joined ETRI and is currently working as the team managing director. His research interests include image/video watermarking, fingerprinting, and signal processing of video and audio.

How to cite this article: H. Kim, J. Kim, J. Park, S. Kim, C. Park, and W. Yoo, *Background music monitoring framework and dataset for TV broadcast audio*, ETRI Journal **46** (2024), 697–707, DOI [10.4218/etrij.2023-0249](https://doi.org/10.4218/etrij.2023-0249)