International Journal of Research in Engineering and Innovation (IJREI)
journal home page: http://www.ijrei.com
ISSN (Online): 2456-6934

International Journal of Research in Engineering and Innovation
ISSN: 2456-6934
Vol-3, Issue-2
International Journal of Research in Engineering and Innovation (IJREI)

A review of vocal track separation methods for karaoke generation

Dr. Unnikrishnan. G

School of Computer Sciences, Mahatma Gandhi University, Kottayam, Kerala, India

Abstract

Generation of clean karaoke tracks is still an unsolved fundamental research problem in Digital Music Technology. Today, there is a high demand for karaoke tracks among amateur and professional singers. Even though a lot of karaoke tracks of popular songs are available online, there still exist even more songs for which the karaoke tracks are not available. This paper reviews some of the existing automatic karaoke generation methods, which can suppress or extract the vocal portion in audio music signals.

Keywords: Vocal Separation, Karaoke Generation, MIDI, HPSS.

1. Introduction

Karaoke of a song is the track of the song without the vocal part. The term karaoke stems from a popular type of Japanese entertainment, which provides pre-recorded accompaniments to popular songs so that any user can sing live like a professional singer. This entertainment medium has become very popular in many places, especially in South East Asia, and is also driving the development of many home devices and online services [13]. Karaoke is said to be invented in 1971 [1], and it is regarded as one of the earliest examples of technology-based music applications for amateur music fans. It is currently one of the major ways of enjoying music, and is considered as one of major leisure activities [14]. Songs created by amateur musicians that are typically distributed through web sites, have become considerably popular recently, especially in the last decade. Recent growth of the web-based music community prompts many amateur musicians to upload their songs, and many listeners enjoy these songs as well as the songs created by professional musicians [15].

Today there is a high demand for karaoke tracks among amateur and professional singers. Even though a lot of karaoke tracks of popular songs are available online, there still exist even more songs for which the karaoke tracks are not available. At present, there is no foolproof automatic method to generate karaoke tracks from original song tracks. The current techniques of creating karaoke for these kinds of songs is very expensive because the current karaoke systems require pre-made MIDI data, which are created manually by skilled craftspeople who have the ability of music dictation[3].

Hence research to develop techniques that can generate karaoke signals automatically from mixed music signals, such as MP3 data, have very high significance.

2. Existing Methods

In order to create karaoke data from mixed music signals, there are mainly two approaches:

2.1 Audio → MIDI → MIDI → Audio

This approach uses the following sequential processing:

- Audio to MIDI (Musical Instrument Digital Interface, a standard which provides symbolic musical information) conversion based on general music transcription techniques such as multiple F0 analysis [2].
- Estimate which parts of the MIDI data are vocal, and delete them from the MIDI data.
- MIDI → Audio synthesis, which is a trivial task in computer music technology.

However, this approach is not easy till date because there are still many difficulties to solve the first two sub-problems listed above.

2.2 Audio → Audio

Another approach is the direct audio-to-audio conversion, using signal processing techniques. A prominent work in this

Corresponding author: Unnikrishnan G
Email Address: ukgkollam@gmail.com

area is an interactive music player named *Euterpe* which generates quasi-karaoke signals from ordinary music signals [3]. It is probably the only real-time automatic karaoke system that integrates two important features: singing voice suppression and key conversion.

Euterpe does not require separately-recorded tracks of a song. It only requires the already-mixed music audio signals (such as ordinary CDs, MP3 data, etc.) The system enables users to reduce the vocal part and change the key of accompanying sounds.

Euterpe is aimed at generating a karaoke signal automatically from a wider range of music signals, especially monaural (monophonic) music signals, to which the simple center-removal-based karaoke generation techniques cannot be applied. This system has two functions: singing voice suppression and key conversion.

The singing voice suppression is based on two-stage HPSS (Harmonic/Percussive Sound Separation), which is effective as a vocal enhancer. The technique also works effectively as a vocal suppressor. Singing voice may be regarded as a stationary signal in a very short period (in milliseconds), but it may not be so in a very long period (in seconds), due to its fluctuations. The relative time-scale decides whether a singing voice appears as a harmonic (stationary) signal or percussive (non-stationary) signal. Using this property of singing voice, *Euterpe* extracts the vocal part from a music signal by applying the local convolutive operation HPSS twice [3].

HPSS separates a music signal into harmonic components and percussive components. That is:

(Input) + HPSS (L1) → (Harmonic including Vocal), (Percussive)

Where L1 is a frame length of STFT (Short-time Fourier Transform).

To the contrary, if applied to a spectrogram of long frame length L2 >> L1, HPSS separates the same signal quite differently in the following way:

(Input) + HPSS (L2) → (Harmonic), (Fluctuated + Percussive)

Since singing voice often has fluctuations, vocal components tend to be separated into "percussive" component in this case. Therefore, applying HPSS twice on differently-resolved spectrograms, a signal can be separated to three components: the harmonic (e.g., guitar), the fluctuated (e.g., vocal), and the percussive (e.g., table). That is:

(Input) + Two stage HPSS → (Harmonic), (Vocal), (Percussive)

Euterpe generates the "vocal-off" signal by adding the obtained harmonic and percussive multiplying certain weights αh, αp ∈ R. Figure 1 shows a sample result of two-stage HPSS [3].

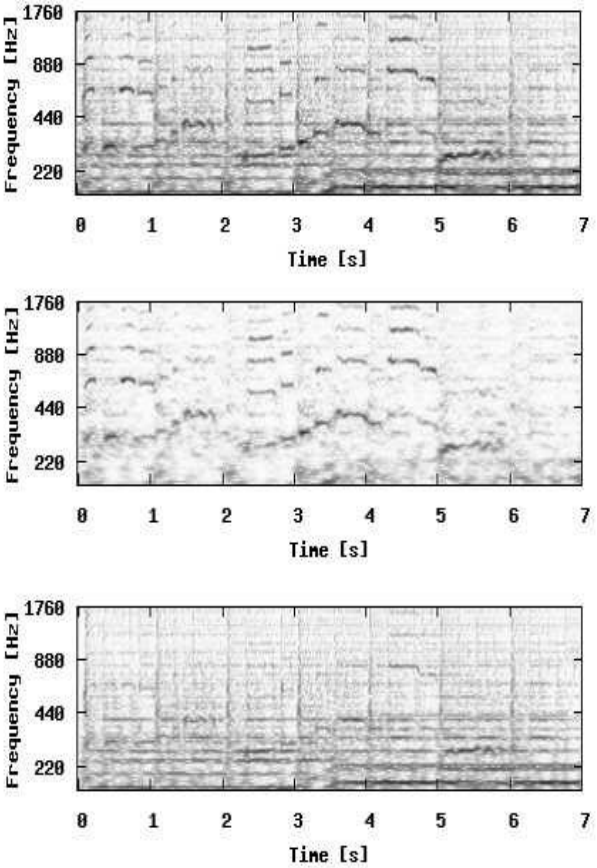

Fig.1. A sample result of two-stage HPSS. Top figure shows the spectrogram of a mixed music signal. Middle figure shows the spectrogram of extracted singing voice by two-stage HPSS. Bottom figure shows the spectrogram of the residual component, i.e., the sum of the 'harmonic' and 'percussive,' which is roughly equal to the accompaniment.

An advantage of this approach is that it does not need any explicit models, no training data of singing voice, which may be complicated or may be infeasible for real-time processing. The two-stage HPSS just need two operations: FFT and the convolutive operation (namely HPSS). Moreover, it works in real-time (on-line) since the streaming HPSS is basically a one-pass technique.

The key conversion in *Euterpe* is based on the pitch shift technique based on phase vocoder. It works in real-time and it keeps the timbre of the sound.

2.3 Vocal Suppression

For stereo music signals, there is a widely-known simple method to generate a karaoke. A quasi-karaoke signal can be obtained just by subtracting the right channel from the left channel of a stereo signal. Many free automatic karaoke generators like Center Pan Remover and some plugins for the software Audacity [4] are based on this method or it's variant. This approach is based on the common practice of the present day professional music creation procedure that all the instruments including singing voice are separately recorded, and the vocal component is placed on the center, when all the parts are mixed down by recording engineers. That is, both left and right channels contain almost exactly equal vocal components and the channel subtraction cancels the vocal component, resulting in the karaoke of the song.

A drawback of this simplest method is that it can be applied only to stereo music signals. In addition, the creation procedure should be based on the professional convention. That is, the technique may not work effectively for live recordings and cannot be applied to monaural signals. In order to cover a wider range of recordings, a technique that automatically removes singing voice based on its nature, not on the recording convention, is needed. To date, the number of academic literature that principally focused on this task seems limited. However, some automatic audio-to-audio karaoke generation techniques were proposed, such as the techniques based on Bayesian model [5], F0 estimation [6], Deep Leaning [7] and RPCA based model [8].

2.4 Vocal Extraction

In the music signal processing community, the opposite side of the karaoke generation task, namely vocal extraction and enhancement, has been attracting more interest. The aim here is not to remove the vocal, but to separate it from the orchestra in order to enhance it. There have been many studies including Hsu & Jang's unvoiced singing voice enhancement [16], techniques based on Nonnegative Matrix Factorization [10], two-stage HPSS [11] etc.

Of these, two-stage HPSS [11] has an advantage that it performs well in terms of GNSDR (Generalized Normalized Signal to Distortion Ratio [12]), it is efficient, and it works in real-time with a little latency. This technique is effective also as a vocal suppression technique, and is used as a technical component of automatic audio-to-audio karaoke systems.

3. Conclusion

An introduction to the research problem of karaoke generation, and a review of some of the existing automatic karaoke generation methods which can suppress or extract the vocal portion in audio music signals are presented in this paper. The task of generating clean karaoke tracks remains to be an unsolved fundamental research problem in Digital Music Technology and offers scope for rigorous research. Vocal extraction for enhancement, the other side of the karaoke generation task, also demands extensive research.

References

[1] Xun, Z. and Tarocco, F. "Karaoke: The Global Phenomenon", Reaktion Books (2007).

[2] Klapuri, A. "Multiple Fundamental Frequency Estimation based on Harmonicity and Spectral Smoothness", IEEE Transactions on Speech Audio Process, Vol.11, No.6, pp.804–816 (2003).

[3] Tachibana,H., Mizuno,Y., Ono, N. and Sagayama, S. "A Real-time Audio-to-audio Karaoke Generation System for Monaural Recordings Based on Singing Voice Suppression and Key Conversion Techniques" Journal of Information Processing Vol.24 No.3 470–482 (May 2016

[4] Audacity Vocal Removal Plug-ins, available from http://wiki.audacityteam.org/wiki/Vocal_Removal_Plug-ins〉

[5] Ozerov, A., Philippe, P., Bimbot, F. and Gribonval, R. "Adaptation of Bayesian Models for Single-Channel Source Separation and Its Application to Voice/Music Separation in Popular Songs", IEEE Transactions on Audio, Speech, and Language Processing, Vol.15, No.5, pp.1564–1578 (2007).

[6] Ryynanen, M., Virtanen, T., Paulus, J. and Klapuri, A. "Accompaniment separation and karaoke application based on automatic melody transcription", Proc. 2008 IEEE International Conference on Multimedia and Expo(ICME 2008) pp.1417–1420 (2008)

[7] Simpson, A.J.R., Roma, G. and Plumbley, M.D. "Deep Karaoke: Extracting Vocals from Musical Mixtures using a Convolutional Deep Neural Network (2015)". arXiv:1504.04658.

[8] Ikemiya, Y., Yoshii, K. and Itoyama, K. "Singing Voice Analysis and Editing based on Mutually Dependent F0 Estimation and Source Separation", Proc. ICASSP (2015).

[9] Hsu, C.L. and Jang, J.S.R. "On The Improvement of Singing Voice Separation for Monaural Recordings using The MIR-1K Dataset", IEEE Transactions on Audio, Speech, and Language Processing (2010).

[10] Zhu, B., Li, W., Li, R. and Xue, X. "Multi-stage Non-negative Matrix Factorization for Monaural Singing Voice Separation", IEEE Transactions on ASLP, Vol.21, No.10, pp.2096–2107 (2013)

[11] Tachibana, H., Ono, N. and Sagayama, S. "Singing Voice Enhancement in Monaural Music Signals Based on Two-stage Harmonic/Percussive Sound Separation on Multiple Resolution Spectrograms", IEEE Transactions on ASLP, Vol.22, No.1, pp.228–237 (2014).

[12] Ozerov, A., Philippe, P., Gribonval, R. and Bimbot, F. "One Microphone Singing Voice Separation using Source-Adapted Models", Procedings of 2005 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA) pp.90–93 (2005)

[13] Zhu, Y. and Gao,S. "Extracting vocal melody from karaoke music audio," Proceedings of IEEE International Conference on Multimedia & Expo, 2005.

[14] Japan Productivity Center: White Paper of Leisure 2011 (2011).

[15] Hamasaki, M., Takeda, H., Hope, T. and Nishimura, T. "Network Analysis of an Emergent massively Collaborative Creation Community how Can People Create Videos Collaboratively without Collaboration?", Procedings of 3rd International ICWSM Conference, AAAI, pp.222–225 (2009).

Cite this article as: Unnikrishnan G, A review of vocal track separation methods for karaoke generation, International Journal of Research in Engineering and Innovation Vol-3, Issue-2 (2019), 141-143.