LATENT TIME-FREQUENCY COMPONENT ANALYSIS: A NOVEL PITCH-BASED APPROACH FOR SINGING VOICE SEPARATION

Xiu Zhang[1], *Wei Li*[1,2], *Bilei Zhu*[3]

[1]School of Computer Science, Fudan University, Shanghai, China
[2]Shanghai Key Laboratory of Intelligent Information Processing, Fudan University, Shanghai, China
[3]SAP Labs, Shanghai, China

ABSTRACT

Monaural singing voice separation has aroused considerable attention. Many pitch-based methods have been proposed to address this task, but generally have limited performance. The most crucial difficulties lie in the inaccurate judgment on voiced pitches and the failed recognition on unvoiced singing sounds. In this paper, we propose a novel algorithm based on the latent component analysis of time-frequency representation to overcome these difficulties. Specifically, the time-frequency (T-F) representations of the song are firstly decomposed into components, and each component approximately originates from a single sound source. We then construct non-overlapping T-F segments with these components, to complete the omitted useful singing voice information. Extensive experiments on the MIR-1K public dataset shows the effectiveness of the proposed algorithm.

Index Terms— Singing voice separation, Pitch-Based Inference, Latent Time-Frequency Component Analysis

1. INTRODUCTION

Separating singing voice from accompaniment has many applications in music information retrieval (MIR). To perform this separation, a number of algorithms have been proposed in recent years and many of them are within the framework of pitch-based inference [1, 2, 3, 4, 5, 6, 7]. It is known that singing voice is primarily comprised of voiced sounds, which are roughly harmonic, with frequencies of concurrent overtones being approximately integer multiples of the fundamental frequency (F0). Pitch-based inference algorithms utilize the harmonic structure of singing voice, and first extract the singing pitch as the cue for subsequent separation.

Unfortunately, pitch-based singing voice separation algorithms have several limitations. Firstly, they highly rely on the technique of singing pitch detection from polyphonic song mixtures, which, however, remains an open problem and has not been maturely solved so far [8]. As a premise, if the detected pitch is inaccurate, the harmonic structure of singing voice cannot be correctly identified. Secondly, although the majority of singing voice is voiced sounds, a small part of unvoiced sounds do exist. Having no underlying periodicity, unvoiced singing sounds cannot be characterized by pitch and further effectively separated from accompaniment using most existing pitch-based inference algorithms [4]. Both of the above factors result in poor vocal separation performance.

In this paper, we propose an algorithm for singing voice separation in monaural mixtures, based on the latent time-frequency component analysis of time-frequency representations of the original input song mixture, to overcome the limitations and complete the missing information discussed above. Firstly, each input song is expressed as a set of T-F matrices. We decompose each matrix into components by Non-negative Matrix Factorization (NMF), each of which approximately originates from a single sound source. Then we construct non-overlapping T-F segments with the independent components and complete the corresponding missing information caused by the inaccurate detection and unvoiced singing sounds which cannot be described by pitch. Admittedly there have been several attempts that combine pitch-based inference with NMF for monaural singing voice separation (e.g., [2] and [9]). In these methods, NMF is generally used to model or estimate the music accompaniment which is quite different from us. Experiments on the MIR-1K public dataset show that our proposed algorithm is rather effective.

2. PROPOSED ALGORITHM

The proposed singing voice separation algorithm follows the standard pitch-based inference framework in [1], which consists of three stages, as described in Sec. 2.1. We then illustrate the proposed algorithm which improves this framework, particularly the third stage, to complete the missing information by analyzing latent component of T-F matrix and further generating the non-overlapping T-F segments in Sec. 2.2.

2.1. Preprocessing

Singing Voice Detection The first stage is to locate singing voice portions in each input song mixture by using a hidden

This work is supported by NSFC (61171128).

Markov model (HMM)-based classification method. We use 39-dimensional MFCCs formed by 12 cepstral coefficients plus the log energy, together with their first and second order derivations. These features are extracted from Hamming-windowed frames of 40 ms with an overlap of 50%, and cepstral mean normalization is applied to reduce channel effects. The HMM for classification has two states, i.e., vocal and nonvocal, whose output distributions are represented by 32-component diagonal-covariance Gaussian mixture models (GMMs) trained from MFCCs of the vocal and the nonvocal frames respectively. The transition probabilities of the HMM are obtained by frame counting in training set. During testing, the Viterbi algorithm is used to decode given sound mixtures into vocal and nonvocal portions.

Singing Pitch Detection In this stage, the well-known autocorrelation-based F0 estimator YIN algorithm [10] is applied on each input song mixture to extract the singing pitch contour. The reason that we us YIN instead of a multi-pitch detection algorithm is twofold. First, the performance of current multi-pitch detection algorithms is rather limited. Second, singing voice is usually the most predominant part in popular songs, approximating the case of monophonic music. We set the frame length 40 ms with an overlap of 50%, and the F0 range 80$\sim$500 Hz. Since the pitch is detected as the cue for vocal separation, only the pitch contours of the vocal portions are retained for further processing and evaluation.

Singing Voice Separation In this stage, the singing voice is separated from each song mixture by using the detected singing pitch. This starts with an auditory peripheral model for T-F decomposition. First, the input mixture is passed through a 128-channel gammatone filterbank, whose center frequencies are equally distributed on the equivalent rectangular bandwidth (ERB) scale between 80 Hz and 5 kHz. The output signal of each filter is then divided into half-overlapped 40-ms frames. In this way, the song mixture is decomposed into a collection of T-F units. Each unit is denoted as u_{cm}, where c and m are indexes of the filter channel and the time frame respectively.

Then, the next step is to estimate the ideal binary mask (IBM). The IBM is a binary matrix, where 1 means that the energy of singing voice is stronger than that of accompaniment within the corresponding T-F unit and 0 indicates weaker [11]. To estimate this mask, [1] and [4] use a periodicity criterion, where a T-F unit is identified as singing dominant and labeled with 1 if it is located in a vocal frame and its local periodicity matches the detected pitch of the frame, otherwise the T-F unit is deemed as accompaniment dominant and labeled with 0. Specifically, for a T-F unit u_{cm} corresponding to the filter channel c at the time frame m, it is identified as singing dominant if the frame m is classified as vocal and u_{cm} satisfies $\frac{\mathrm{acf}_{cm}(\tau_m)}{\max_{2\le\tau\le 12.5(\mathrm{ms})}\mathrm{acf}_{cm}(\tau)} \ge \theta$, where $\mathrm{acf}_{cm}(\tau)$ is the autocorrelation function of u_{cm}, $\tau \in [0, 200]$ ([0, 12.5 ms]) is the time delay, τ_m is the time delay corresponding to the detected pitch at frame m, and θ is a threshold. In [4], θ is set to 0.99. We follow this setting in our implementation.

The singing voice is finally resynthesized from the masked T-F representation of the original song mixture. This is performed by applying the inverse of gammatone filterbank and the technique of overlap and addition.

2.2. Latent Component Analysis of T-F matrix

Unfortunately, since the section above uses pitch as the only cue, it essentially cannot recognize time-frequency units dominated by unvoiced singing sounds. Besides, due to the inaccuracy of the detected pitch, the proposed approach incorrectly identifies many voiced singing-dominant time-frequency units as accompaniment dominant. However, these two types of errors influence the step of using binary mask. The corresponding matrix elements are marked 0 (the accompaniment sound) instead of the right 1 (the singing voice). In this paper we employ the latent component of T-F matrix to refine the inner properties of relationship with the units labeled 1 and labeled 0. With the refined information, we can construct T-F segments that indicate possible pitch areas in T-F matrix and further relabel the mask matrix to recover the neglected information mentioned above to provide final singing voice separation results.

We obtain the latent components of T-F matrices by the Non-negative Matrix Factorization (NMF) [12] technique. To the best of our knowledge, there is no prior work investigating the latent component of T-F matrix to recover the neglected information like the proposed algorithm in singing voice separation.

Given a non-negative matrix $\mathbf{X}$ of dimensions $C \times M$ and a positive integer R, NMF finds an approximate factorization

$$\mathbf{X} \approx \mathbf{WH} \tag{1}$$

where $\mathbf{W}$ and $\mathbf{H}$ are non-negative matrices of dimensions $C \times R$ and $R \times M$ respectively.

Recently, NMF and its extensions have been successfully applied for monaural sound source separation [13, 14]. In this case, the observation matrix $\mathbf{X}$ is typically a phase-invariant T-F representation of the input sound mixture, where C is the number of frequency bins and M is the number of time frames. The model matrices, $\mathbf{W}$ and $\mathbf{H}$, are basis matrix and gain matrix respectively, where the columns of $\mathbf{W}$ are spectral bases and the rows of $\mathbf{H}$ are their gains in each frame. Each spectral basis and its time-varying gain are referred to as a component and there are thus R components in total.

Generally, each component represents parts of a single sound source and each sound source in the mixture is modeled as a sum of one or more components. We employ the very point as the criteria for cluster. The specific procedures are as follows. Consequently, sound source separation is done by first decomposing the sound mixture into NMF components and then grouping these components to sound sources.

The factorization of NMF, Eq. (1), is usually sought by minimizing a chosen cost function between $\mathbf{X}$ and $\mathbf{WH}$ while restricting their elements to non-negative values. In this paper, we choose the Kullback-Leibler divergence as the cost function and apply the multiplicative update rules proposed in [15] to solve the minimization problem.

1. *Construct an energy matrix $\mathbf{X}$ of T-F units.* The element $\mathbf{X}_{cm}$ corresponding to u_{cm} is calculated as

$$\mathbf{X}_{cm} = \sum_{n=1}^{N} u_{cm}^2(n) \tag{2}$$

   where $u_{cm}(n)$ is the n^{th} sample in u_{cm}, N is the frame length in samples. Obviously, $\mathbf{X}$ is a non-negative matrix of dimensions $C \times M$, where $C = 128$ is the number of frequency bins and M is the number of time frames.

2. *Perform NMF on the obtained matrix $\mathbf{X}$ to decompose it into a set of components.* Given the factorization $\mathbf{X} \approx \mathbf{WH}$ and the number of components R ($R = 60$ in our implementation), a component here is denoted as $\mathbf{X}^r$, $r = 1, \ldots, R$ being the component index, and represented as a T-F matrix. The T-F matrix representation of $\mathbf{X}^r$ is calculated from its spectral basis (i.e., the r^{th} column of $\mathbf{W}$) and the temporal gain (i.e., the r^{th} row of $\mathbf{H}$). Specifically, the matrix element at position (c, m) is computed as

$$\mathbf{X}^r = \mathbf{w}_r \mathbf{h}_r. \tag{3}$$

   where $\mathbf{w}_r$ is the r^{th} column of $\mathbf{W}$, $\mathbf{h}_r$ is the r^{th} row of $\mathbf{H}$). Based on the property of NMF, each component approximately originates from a single sound source.

3. *Generate a T-F segment from each component obtained above.* Specifically, for a given component $\mathbf{X}^r$, its T-F representation is compared with those of other components, with the elements satisfying Eq. (4) selected.

$$\mathbf{X}_{cm}^r = \max_{i=1}^{R} \mathbf{X}_{cm}^i. \tag{4}$$

   In general, each selected element $\mathbf{X}_{cm}^r$ corresponds to a T-F unit u_{cm}, and all these T-F units form a segment S^r corresponding to $\mathbf{X}^r$.

Fig. 1 illustrates how to generate T-F segments from NMF components. For a given component, red elements in its T-F representation are those satisfying Eq. (4), meaning that they are larger than all the green elements in the same positions of other T-F representations. Typically, each red element corresponds to a T-F unit with the same index, and all these units form the segment corresponding to the given component.

As a result of the above procedure, the input song mixture is decomposed into a set of T-F segments, each of which is indivisible, with energy primarily originating from a single sound source. With the constraint of Eq. (4), these segments are non-overlapping, i.e., a T-F unit only belongs to a single segment. In other words, segments are disjoint clusters of T-F units, and all the T-F units included in a segment are dominated by the same sound source. Given this property, a T-F unit can be labeled based on not only the periodicity information provided by the singing pitch, but also the origination of the segment that it belongs to. Given the T-F segments, we now describe how to estimate the IBM by using pitch and segment as two complementary cues. First, let $\mathbf{M}^0$ be the mask computed using the conventional pitch-based inference method. Then, the segment cue is considered to get additional masking information. To be specific, for each segment, if more than 10% of its belonging T-F units have been identified as singing dominant by the pitch-based method, the whole segment is deemed as originating from vocals and all its T-F units are labeled with 1. This forms a new masking matrix, denoted as $\mathbf{M}^1$. The final estimation of IBM, denoted as $\mathbf{M}$, is the combination of $\mathbf{M}^0$ and $\mathbf{M}^1$, i.e.,

$$\mathbf{M} = \mathbf{M}^0 \| \mathbf{M}^1 \tag{5}$$

where $\mathbf{A} \| \mathbf{B}$ is the element-wise *logical OR* operation of matrices $\mathbf{A}$ and $\mathbf{B}$.

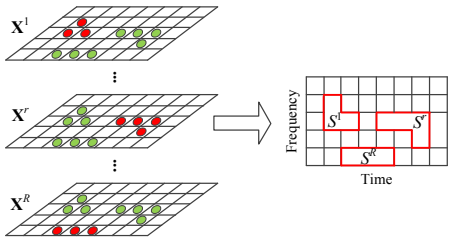

Fig. 1. The generation of T-F segments from latent components. Please refer to Eq. 4 and the related illustrations for details. (best viewed in color)

3. EVALUATION

The evaluation is carried out on the MIR-1K public dataset [4], which contains 1000 song clips sampled at 16 kHz, with durations ranging from 4 to 13 s. These clips are extracted from 110 karaoke Chinese pop songs performed by male and female amateurs, with accompaniment and vocals recorded in the left and right channels, respectively. On the basis of the 1000 song clips, we create three sets of monaural mixtures at different qualities for evaluation. To be exact, for each original song clip in MIR-1K, the singing voice and music accompaniment are mixed at three different signal-to-noise ratios (SNRs), i.e., -5 dB (accompaniment is louder), 0 dB (same level), and 5 dB (singing voice is louder). Note that in this circumstance, signal refers to the singing voice, while the accompaniment is deemed as noise.

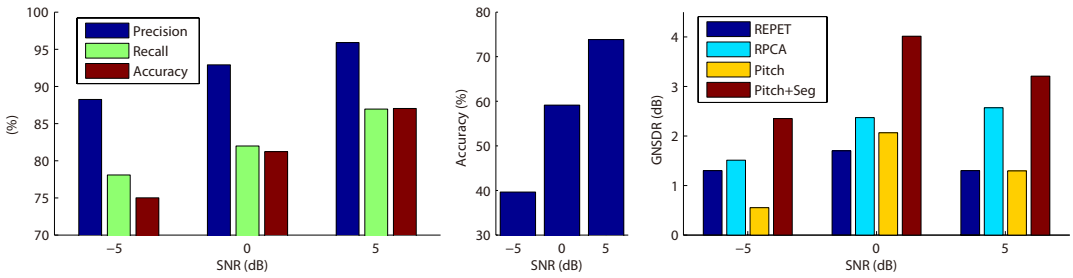

Fig. 2. Performance of (a) singing voice detection, (b) singing pitch detection, (c) singing voice separation.

3.1. Evaluation for Singing Voice Detection

1) Dataset Description: All 1000 song clips of MIR-1K are used to evaluate the performance of singing voice detection. Since the detection is based on supervised classification, the whole dataset is further divided into two non-intersecting subsets with nearly equal size (483 vs. 517) for training and testing. The final results are given through two-fold cross validation.

2) Performance Measure: The performance of singing voice detection is measured by frame-level precision, recall and overall accuracy. Specifically, the precision is the percentage of correctly detected vocal frames over all the detected vocal frames; the recall is the percentage of correctly detected vocal frames over all the vocal frames. The overall accuracy is the percentage of all the frames (both vocal and nonvocal) that are correctly classified.

3) Experimental Results: As shown in Fig. 2 (a), by using the HMM-based classification method, vocal and nonvocal portions in the song mixtures can be accurately partitioned. Especially, the precision of singing voice detection is very high for all three SNRs.

3.2. Evaluation for Singing Pitch Detection

1) Dataset Description: All 1000 song clips of MIR-1K are used to evaluate the performance of singing pitch detection.

2) Performance Measure: The performance of singing pitch detection is measured by the overall accuracy, which is defined as the percentage of the frames satisfying the following criteria: (a) if the frame is a nonvocal frame, it is classified as nonvocal; (b) if the frame is a vocal frame, it is classified as vocal and the absolute difference between the detected pitch and the ground truth is less than 1 semitone.

3) Experimental Results: Fig. 2 (b) illustrates the performance of singing pitch detection. As can be expected, when the SNR increases, i.e., the energy of singing voice becomes more prominent, the overall detection accuracy gets higher.

3.3. Evaluation for Singing Voice Separation

1) Dataset Description: All 1000 song clips of MIR-1K are used to evaluate the performance of singing voice separation.

2) Performance Measure: Given the resynthesized singing voice $\hat{v}$ and the reference clean vocal signal v, the signal-to-distortion ratio (SDR) is calculated using the *BSS_EVAL toolbox* [16] to measure the separation quality between them.

Next, as done in [4], the normalized SDR (NSDR) is defined in Eq. (6). It is the improvement of SDR between the original mixture x and the separated singing voice $\hat{v}$, and used to measure the separation performance for each mixture.

$$\mathrm{NSDR}(\hat{v}, v, x) = \mathrm{SDR}(\hat{v}, v) - \mathrm{SDR}(x, v). \qquad (6)$$

Finally, for the overall performance measure, the global NSDR (GNSDR) is calculated by taking the mean of NSDRs over all the mixtures of each set, weighted by their length. Generally, higher values of GNSDR suggest better separation.

3) Experimental Results: Fig. 2 (c) demonstrates the performance of pitch-based vocal separation without and with the proposed NMF-based segmentation method (denoted as *Pitch* and *Pitch+Seg* respectively). For comparison, the GNSDRs provided by two state-of-the-art singing voice separation methods outside the pitch-based inference framework, i.e., *REPET* and *RPCA*, are also presented. *REPET* refers to the repeating pattern extraction technique (REPET) proposed by Rafii and Pardo [17], which exploits the repeating musical structure for voice/music separation. *RPCA* refers to the robust principal component analysis (RPCA)-based method devised by Huang *et al.* [18], which performs singing voice separation by applying RPCA on the mixture spectrogram.

As shown in the figure, the GNSDR of *Pitch+Seg* is about 2 dB larger than that of *Pitch* for all the three SNRs, indicating the effectiveness of the proposed segmentation method. Compared with *REPET* and *RPCA*, *Pitch* provides smaller or similar GNSDRs, while *Pitch+Seg* outperforms the two methods significantly.

4. CONCLUSION

In this paper, we present a new pitch-based approach based on latent time-frequency component analysis for singing voice separation. This algorithm manages to complete the missing information caused by the inaccurate detected pitch as well as the unvoiced sounds in existing pitch-based methods. Quantitative evaluation on 1000 song clips demonstrates the effectiveness of the proposed algorithm.

5. REFERENCES

[1] Y. Li and D. L. Wang, "Separation of singing voice from music accompaniment for monaural recordings," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1475–1487, 2007.

[2] T. Virtanen, A. Mesaros, and M. Ryynänen, "Combining pitch-based inference and non-negative spectrogram factorization in separating vocals from polyphonic music," in *ITRW on SARP*, 2008.

[3] M. Ryynänen, T. Virtanen, J. Paulus, and A. Klapuri, "Accompaniment separation and karaoke application based on automatic melody transcription," in *ICME*, 2008.

[4] C. L. Hsu and J. S. R. Jang, "On the improvement of singing voice separation for monaural recordings using the MIR-1K dataset," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 2, pp. 310–319, 2010.

[5] C. L. Hsu, D. L. Wang, J. S. R. Jang, and K. Hu, "A tandem algorithm for singing pitch extraction and voice separation from music accompaniment," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 5, pp. 1482–1491, 2012.

[6] H. Fujihara, M. Goto, T. Kitahara, and H. G. Okuno, "A modeling of singing voice robust to accompaniment sounds and its application to singer identification and vocal-timbre-similarity-based music information retrieval," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 638–648, 2010.

[7] E. Cano, C. Dittmar, and G. Schuller, "Efficient implementation of a system for solo and accompaniment separation in polyphonic music," in *EUSIPCO*, 2012.

[8] J. Salamon and E. Gómez, "Melody extraction from polyphonic music signals using pitch contour characteristics," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 6, pp. 1759–1770, 2012.

[9] J. L. Durrieu, G. Richard, B. David, and C. Févotte, "Source/filter model for unsupervised main melody extraction from polyphonic audio signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 564–575, 2010.

[10] A. de Cheveigné and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1917–1930, 2002.

[11] D. L. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*, P. Divenyi, Ed., pp. 181–197. Kluwer, Norwell MA, 2005.

[12] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[13] S. Vembu and S. Baumann, "Separation of vocals from polyphonic audio recordings," in *ISMIR*, 2005.

[14] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 1066–1074, 2007.

[15] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *NIPS*, 2001.

[16] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[17] Z. Rafii and B. Pardo, "Repeating pattern extraction technique (REPET): a simple method for music/voice separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 1, pp. 71–82, 2013.

[18] P. S. Huang, S. D. Chen, P. Smaragdis, and M. Hasegawa-Johnson, "Singing-voice separation from monaural recordings using robust principal component analysis," in *ICASSP*, 2012.