COMBINING MODELING OF SINGING VOICE AND BACKGROUND MUSIC FOR AUTOMATIC SEPARATION OF MUSICAL MIXTURES

Zafar Rafii[1], François G. Germain[2], Dennis L. Sun[2,3], and Gautham J. Mysore[4]
[1]Northwestern University, Department of Electrical Engineering & Computer Science
[2]Stanford University, Center for Computer Research in Music and Acoustics
[3]Stanford University, Department of Statistics
[4]Adobe Research
zafarrafii@u.northwestern.edu, {fgermain,dlsun}@stanford.edu, gmysore@adobe.com

ABSTRACT

Musical mixtures can be modeled as being composed of two characteristic sources: singing voice and background music. Many music/voice separation techniques tend to focus on modeling one source; the residual is then used to explain the other source. In such cases, separation performance is often unsatisfactory for the source that has not been explicitly modeled. In this work, we propose to combine a method that explicitly models singing voice with a method that explicitly models background music, to address separation performance from the point of view of both sources. One method learns a singer-independent model of voice from singing examples using a Non-negative Matrix Factorization (NMF) based technique, while the other method derives a model of music by identifying and extracting repeating patterns using a similarity matrix and a median filter. Since the model of voice is singer-independent and the model of music does not require training data, the proposed method does not require training data from a user, once deployed. Evaluation on a data set of 1,000 song clips showed that combining modeling of both sources can improve separation performance, when compared with modeling only one of the sources, and also compared with two other state-of-the-art methods.

1. INTRODUCTION

The ability to separate a musical mixture into singing voice and background music can be useful for many applications, e.g., query-by-humming, karaoke, audio remixing, etc. Existing methods for music/voice separation typically focus on estimating either the background music, e.g., by training a model for the accompaniment from the non-vocal segments, or the singing voice, e.g., by identifying the predominant pitch contour from the vocal segments.

Some methods estimate the background music by training a model on the non-vocal segments in the mixture, identified manually or using trained vocal/non-vocal classifiers. Ozerov et al. used Bayesian models to train a model for the background music from the non-vocal segments, which they then used to train a model for the singing voice [7]. Han et al. used Probabilistic Latent Component Analysis (PLCA) to also train a model for the background music, which they then used to estimate the singing voice [2].

Other methods estimate the background music directly, without prior vocal/non-vocal segmentation, by assuming the background to be repeating and the foreground (i.e., the singing voice) non-repeating. Rafii et al. used a beat spectrum to identify the periodically repeating patterns in the mixture, followed by median filtering the spectrogram of the mixture at period rate to estimate the background music [9]. Liutkus et al. used a beat spectrogram to further identify the varying periodically repeating patterns [6].

Other methods instead estimate the singing voice by identifying the predominant pitch contour in the mixture. Li et al. used a pitch detection algorithm on the vocal segments in the mixture to estimate the predominant pitch contour, which they then used to derive a time-frequency mask to extract the singing voice [5]. Hsu et al. also used a pitch-based method to model the singing voice, while additionally estimating the unvoiced components [3].

Other methods are based on matrix decomposition techniques. Vembu et al. used Independent Component Analysis (ICA) and Non-negative Matrix Factorization (NMF) to decompose a mixture into basic components, which they then clustered into background music and singing voice using trained classifiers such as neural networks and Support Vector Machines (SVM) [12]. Virtanen et al. used a pitch-based method to estimate the vocal segments of the singing voice, and then NMF to train a model for the background music from the remaining non-vocal segments [14].

Other methods estimate both sources concurrently. Durrieu et al. used a source-filter model to parametrize the singing voice and a NMF model to parametrize the background music, and then estimated the parameters of their models jointly using an iterative algorithm [1]. Huang et al. used Robust Principal Component Analysis (RPCA) to jointly estimate background music and singing voice, assuming the background music as a low-rank component and the singing voice as a sparse component [4].

In this work, we propose a method for modeling the

singing voice, which learns a singer-independent model of voice from singing examples using a NMF based technique. We then propose to combine this method with a method for modeling the background music, which derives a model of music by identifying and extracting repeating patterns using a similarity matrix and a median filter. Combining a method that specifically models the singing voice with a method that specifically models the background music addresses separation performance from the point of view of both sources.

The rest of the article is organized as follows. In Section 2, we present a method for modeling singing voice. In Section 3, we review an existing method for modeling background music, and we propose combining the two methods to improve music/voice separation. In Section 4, we evaluate the method for modeling the singing voice and the combined approach on a data set of 1,000 song clips, and we compare them with the method for modeling the background music alone, and two other state-of-the-art methods. Section 5 concludes this article.

2. MODELING SINGING VOICE

In this section, we present a method for modeling the singing voice. Because singer-specific training examples are generally not available for the music/voice separation methods, models for the singing voice are typically based on *a priori* assumptions, e.g., it has a sparse time-frequency representation [4], it is accurately modeled by a source-filter model [1], or it is reasonably described by pitch [5].

Recently, universal models were proposed as a method for incorporating general training examples of a sound class for source separation when specific training examples are not available [11]. We use these ideas to model the singing voice using a *universal voice model*, learned from a corpus of singing voice examples. Since the formulation of universal voice models is based on matrix factorization methods for source separation, we begin by reviewing Nonnegative Matrix Factorization (NMF).

2.1 NMF for Source Separation

The magnitude spectrogram $\boldsymbol{X}$ is a matrix of non-negative numbers. We assume that the spectrum at time t, $\boldsymbol{X}_t$, can be approximated by a linear combination of basis vectors $\boldsymbol{w}_i$, each capturing a different aspect of the sound, e.g., different pitches, transients, etc.:

$$\boldsymbol{X}_t \approx \sum_{i=1}^{K} h_{it} \boldsymbol{w}_i$$

The collection of basis vectors $\boldsymbol{W} = \begin{bmatrix} \boldsymbol{w}_1 & ... & \boldsymbol{w}_K \end{bmatrix}$ can be regarded as a model for that sound class, since all possible sounds are assumed to arise as linear combinations of these basis vectors. Likewise, $\boldsymbol{H} = (h_{it})$ can be regarded as the activations of the basis vectors over time. In matrix notation, this can be expressed as:

$$\boldsymbol{X} \approx \boldsymbol{W}\boldsymbol{H}.$$

NMF attempts to learn $\boldsymbol{W}$ and $\boldsymbol{H}$ for a given spectrogram $\boldsymbol{X}$, i.e., it solves the optimization problem:

$$\underset{\boldsymbol{W},\,\boldsymbol{H}}{\text{minimize}} \;\; D\left(\boldsymbol{X} || \boldsymbol{W}\boldsymbol{H}\right)$$

subject to the constraints that $\boldsymbol{W}$ and $\boldsymbol{H}$ are non-negative. D is a measure of divergence between $\boldsymbol{X}$ and $\boldsymbol{W}\boldsymbol{H}$.

To use NMF to separate two sources, say, singing voice and background music:

1. Learn $\boldsymbol{W}_V$ using NMF from isolated examples of the singing voice.
2. Learn $\boldsymbol{W}_B$ using NMF from isolated examples of the background music.
3. Use NMF on the mixture spectrogram $\boldsymbol{X}$, fixing $\boldsymbol{W} = \begin{bmatrix} \boldsymbol{W}_V & \boldsymbol{W}_B \end{bmatrix}$, and learning $\boldsymbol{H}_V$ and $\boldsymbol{H}_B$.
4. Estimates of the singing voice-only and background music-only spectrograms can be obtained from $\boldsymbol{W}_V\boldsymbol{H}_V$ and $\boldsymbol{W}_B\boldsymbol{H}_B$.

A more detailed description of this approach can be found in [10], although the authors use an equivalent probabilistic formulation (PLCA) instead of NMF.

Of these tasks, steps 1 and 2 pose the greatest challenge. While it may be possible to use the non-vocal segments of the background music as isolated training data of the background music, it is rare to find segments in music where the voice is isolated. Source separation is still possible in this setting where training data of only one source is available—one simply learns $\boldsymbol{W}_V$ together with $\boldsymbol{H}_V$ and $\boldsymbol{H}_B$ in step 3. This is the approach taken in [2, 7], but it requires a sufficiently accurate prior vocal/non-vocal segmentation and a sufficient amount of non-vocal segments to effectively learn a model of the background music.

2.2 Universal Voice Model

One alternative when training data of a specific singer is not available is to learn a model from a corpus of singing voice examples. The universal model is a prescription for learning a model from general training examples and incorporating the model in NMF-based source separation [11].

The idea is to independently learn a matrix of basis vectors for each of M singers from training data of the individual singers. This yields M matrices of basis vectors $\boldsymbol{W}_1, ..., \boldsymbol{W}_M$. The universal voice model is then simply the concatenation of the matrices of basis vectors:

$$\boldsymbol{W}_V = \begin{bmatrix} \boldsymbol{W}_1 & ... & \boldsymbol{W}_M \end{bmatrix}$$

The hope is that an unseen singer is sufficiently similar to one or a blend of a few of these singers, so that the universal voice model can act as a singer-independent surrogate for singer dependent models.

In applying the universal voice model for source separation, we make the assumption that the activation matrix for the singing voice

$$\boldsymbol{H}_V = \begin{bmatrix} \boldsymbol{H}_1 \\ \vdots \\ \boldsymbol{H}_M \end{bmatrix}$$

is block sparse, i.e., several of the $H_i \equiv 0$. This is necessary because the number of singers is typically large, and the matrix factorization problem can be underdetermined. The block sparsity is a regularization strategy that incorporates the structure of the problem; it captures the intuition that only a few voice models should be sufficient to explain any given singer. We achieve block sparsity by adding a penalty function Ω to the objective function to encourages this structure. λ controls the strength of the penalty term.

$$\underset{\boldsymbol{W},\boldsymbol{H}}{\text{minimize}} \; D\left(\boldsymbol{X}||\boldsymbol{W}\boldsymbol{H}\right) + \lambda\Omega(\boldsymbol{H}_V) \tag{1}$$

As in [11], we choose the Kullback-Leibler divergence for D:

$$D(\boldsymbol{Y}||\boldsymbol{Z}) = \sum_{i,j} Y_{ij} \log \frac{Y_{ij}}{Z_{ij}} - Y_{ij} + Z_{ij}$$

and a concave penalty on the ℓ_1 norm of the block:

$$\Omega(\boldsymbol{H}_V) = \sum_{i=1}^{M} \log(\epsilon + ||H_i||_1)$$

The algorithm for optimizing (1) is known as Block KL-NMF. Further details can be found in [11].

3. COMBINED APPROACH

In this section, we review an existing method for modeling the background music, and we propose to use it to refine the residual from the singing voice modeling.

3.1 Modeling Background Music

A number of methods have been proposed to estimate the background music, without prior vocal/non-vocal segmentation, by assuming the background to be repeating and the foreground (i.e., the singing voice) to be non-repeating. REPET-SIM is thus a generalization of the REpeating Pattern Extraction Technique (REPET) [1] , a simple approach for separating the repeating background from the non-repeating foreground in a mixture, by identification of the repeating elements and the smoothing of the non-repeating elements.

In particular, REPET-SIM uses a similarity matrix to identify the repeating elements in the mixture - which ideally correspond to the background music, followed by median filtering to smooth out the non-repeating elements - which ideally correspond to the singing voice [8]. Unlike the earlier variants of REPET that use a beat spectrum or beat spectrogram to identify the periodically repeating patterns [6, 9], REPET-SIM uses a similarity matrix and is thus able to handle backgrounds where repeating patterns can also happen non-periodically.

3.2 Combined Approach

In order to improve the music/voice separation that we obtain from using the universal voice model alone, we propose cascading the model with REPET-SIM. The idea is that the universal voice model specifically models the singing voice and, through the residual, provides a preliminary estimate of the background music, which can then be refined by feeding it to REPET-SIM. The pipeline is shown in Figure 1, and detailed below.

[1] http://music.eecs.northwestern.edu/research.php?project=repet

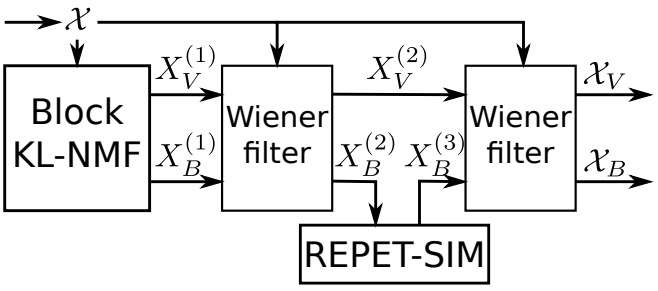

Figure 1. Combined approach which takes in the spectrogram of a mixture $\mathcal{X}$ and returns refined estimates of the spectrogram of the singing voice $\mathcal{X}_V$ and the background music $\mathcal{X}_B$

The universal voice model first outputs an estimate for the magnitude spectrogram of the singing voice $X_V^{(1)}$, and a residual $X_B^{(1)}$ corresponding to the background estimate, which are initially filtered into $X_V^{(2)}$ and $X_B^{(2)}$ by using Wiener filtering, as follows:

$$X_V^{(2)} = \left| \frac{X_V^{(1)}}{X_V^{(1)} + X_B^{(1)}} \odot \mathcal{X} \right|$$

$$X_B^{(2)} = \left| \frac{X_B^{(1)}}{X_V^{(1)} + X_B^{(1)}} \odot \mathcal{X} \right|$$

where $\mathcal{X}$ denotes the complex spectrogram of the mixture and $\odot$ the Hadamard (component-wise) product. Wiener filtering is used here to reduce separation artifacts. Note that we only use the magnitudes of the estimates here.

The background estimate from the universal voice model $X_B^{(2)}$ is then fed to REPET-SIM which refines it into $X_B^{(3)}$.

The estimates for the complex spectrogram of the singing voice $\mathcal{X}_V$ and the background music $\mathcal{X}_B$ are finally obtained by filtering $X_V^{(2)}$ and $X_B^{(3)}$ using Wiener filtering, as follows:

$$\mathcal{X}_V = \frac{X_V^{(2)}}{X_V^{(2)} + X_B^{(3)}} \odot \mathcal{X}$$

$$\mathcal{X}_B = \frac{X_B^{(3)}}{X_V^{(2)} + X_B^{(3)}} \odot \mathcal{X}$$

4. EVALUATION

In this section, we evaluate the method for modeling the singing voice and the combined approach on a data set of 1,000 song clips. We also compare them with the method for modeling the background music alone, as well as two other state-of-the-art methods.

4.1 Data Set

The MIR-1K data set [2] consists of 1,000 song clips in the form of split stereo WAV files sampled at 16 kHz, with the background music and the singing voice recorded on the left and right channels, respectively. The song clips were extracted from 110 karaoke Chinese pop songs performed by 8 female and 11 male singers. The durations of the clips range from 4 to 13 seconds [3].

We created a set of 1,000 mixtures by summing, for each song clip, the left (background music) and right (singing voice) channels into a monaural mixture

4.2 Performance Measures

The BSS Eval toolbox [3] consists of a set of measures that intend to quantify the quality of the separation between a source and its estimate. The principle is to decompose an estimate into a number of contributions corresponding to the target source, the interference from unwanted sources, and the artifacts such as "musical noise."

Based on this principle, the following measures were then defined (in dB): Sources to Interferences Ratio (SIR), Sources to Artifacts Ratio (SAR), and Sources to Distortion Ratio (SDR) which measures the overall error [13].

4.3 Competitive Methods

Durrieu et al. proposed a method [4] based on the modeling of a mixture as an instantaneous sum of a signal of interest (i.e., the singing voice) and a residual (i.e., the background music), where the singing voice is parametrized as a source-filter model, and the background music as an unconstrained NMF model [1]. The parameters of the models are then estimated using an iterative algorithm in a formalism similar to NMF. A white noise spectrum is added to the singing voice model to better capture the unvoiced components. We used an analysis window of 64 milliseconds, a window size of 1024 samples, a step size of 32 milliseconds, and 30 iterations.

Huang et al. proposed a method [5] based on Robust Principal Component Analysis (RPCA) [4]. RPCA is a method for decomposing a data matrix into a low-rank component and a sparse component, by solving a convex optimization problem that aims to minimize a weighted combination of the nuclear norm and the L_1 norm. The method assumes that the background music typically corresponds to the low-rank component and the singing voice typically corresponds to the sparse component.

4.4 Training Universal Models

Our experiments used a leave-one-out cross validation approach. For each of the 19 singers, we learned a universal model using NMF on the other 18 singers, with different choices for the number of basis vectors per singer: $K = 5, 10, 20, 30, 40, 50, 60$.

[2] http://sites.google.com/site/unvoicedsoundseparation/mir-1k

[3] http://bass-db.gforge.inria.fr/bss_eval/

[4] http://www.durrieu.ch/research/jstsp2010.html

[5] https://sites.google.com/site/singingvoiceseparationrpca/

4.5 Parameters

We used a Hamming window of 1024 samples, corresponding to a duration of 64 milliseconds at a sampling frequency of 16 kHz, with 50% overlap.

For REPET-SIM [6], pilot experiments showed that a minimal threshold of 0, a maximal order of 50, and a minimal distance of 0.1 second gave good separation results.

For the universal voice model, pilot experiments showed that different settings of K, K_B (number of background music basis vectors), and λ yielded optimal results for different measures (see Section 4.2) of the separation quality of singing voice and background music. We considered $K = 5, 10, 20, ..., 60$, $K_B = 5, 10, 20, 30, 50, 80$, and a logarithmic grid of λ values.

4.6 Comparative Results

Figures 2, 3, and 4 show the boxplots of the distributions for the SDR, SIR, and SAR (in dB) for the background music (left plot) and the singing voice (right plot) estimates, for the method of Durrieu et al. (*Durrieu*), the method of Huang et al. (*Huang*), REPET-SIM alone (*REPET*), universal voice model alone (*UVM*), and the combination of universal voice model and REPET-SIM (*combo*). The horizontal line in each box represent the median of the distribution, whose value is displayed above the box. Outliers are not shown. Higher values are better.

We used two parameter settings for the universal voice model: one that gave the best SDR for the background music estimates ($K = 20$, $K_B = 5$, and $\lambda = 1448$), and one that gave the best SDR for the singing voice estimates ($K = 10$, $K_B = 5$, and $\lambda = 2896$). The boxplots then show the results for the background music estimates (left plots) and the singing voice estimates (right plots) for the parameter settings that gave the best SDR, for the universal voice model (*UVM*) and the combination (*combo*).

The plots show that the universal voice model alone, for the right parameter settings, achieves higher SDR than REPET-SIM and the other state-of-the-art methods, for both the background music and the singing voice estimates. Combining the universal voice model with REPET-SIM typically yields further improvement.

If we focus on SIR, for the background music estimates, the universal voice model alone achieves higher SIR than REPET-SIM and the other competitive methods; the combination further increases the SIR. For the singing voice estimates, the universal voice model alone achieves higher SIR than REPET-SIM and the method of Huang et al., but the combination does no better than the universal voice model alone.

On the other hand, if we focus on SAR, for the background music estimates, the universal voice model alone has slightly lower SAR than REPET-SIM and the other competitive methods; the combination further decreases the SAR. For the singing voice estimates, the universal voice model alone has higher SAR than the method of Durrieu et al.; the combination further improves the results.

[6] http://music.eecs.northwestern.edu/includes/projects/repet/codes/repet_sim.m

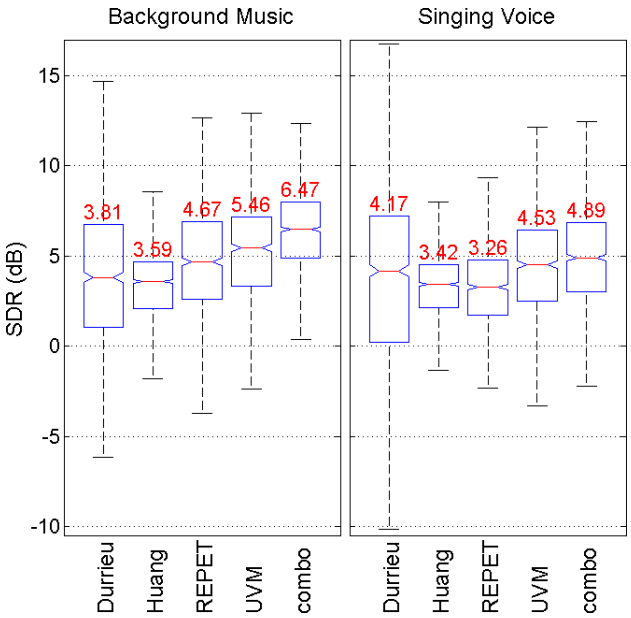

Figure 2. Box plots of the distributions for the SDR (dB).

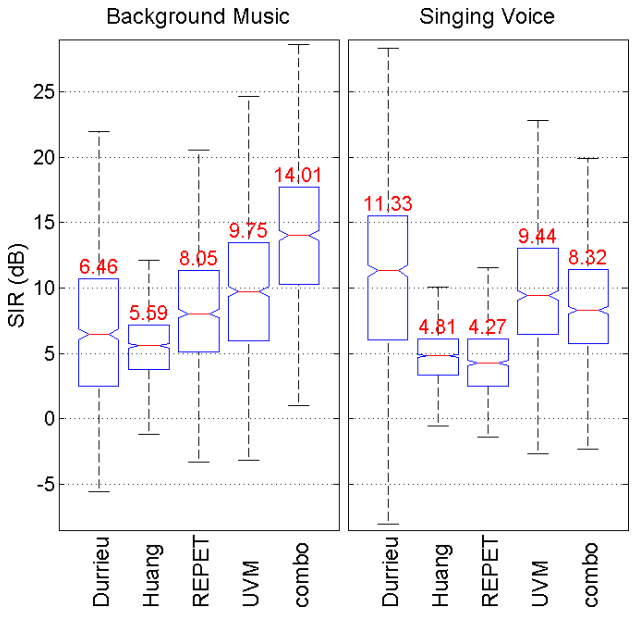

Figure 3. Box plots of the distributions for the SIR (dB).

These results show that, given the right parameter settings, the universal voice model is particularly good at reducing in one source the interference of the other source, however at the expense of adding some artifacts in the estimates. This is related to the SIR/SAR performance trade-off commonly seen in source separation.

The results also show that combining the universal voice model with REPET-SIM helps to increase the SIR for the background music estimates and the SAR for the singing voice estimates, but at the expense of decreasing the SAR for the background music estimates and the SIR for the singing voice estimates. This is related to the music/voice performance trade-off commonly seen in music/voice separation. In other words, the combination helps to reduce in the background music estimates the interference from the singing voice but at the expense of introducing some artifacts in the estimates. On the other hand, it helps to reduce artifacts in the singing voice estimates, at the expense of introducing interference from the background music.

4.7 Statistical Analysis

We compared the SDR of the background music and singing voice estimates across the different methods using a two-sided paired t-test. The universal voice model alone achieved a significantly higher SDR on the background music than the three state-of-the-art methods: the closest competitor was REPET-SIM ($t = 3.92$, $p < .0001$). The combination represented a significant improvement over the universal model alone ($t = 19.4$, $p \approx 0$). A similar story is true for the SDR of the singing voice estimates: the universal voice model alone is significantly better than any of the existing methods, with the method of Durrieu et al. the closest competitor ($t = 6.13$, $p \approx 0$), and the combination represents a significant improvement over it ($t = 13.8$, $p \approx 0$).

In terms of the SIR of the background music estimates, the combination is significantly better than the universal voice model alone ($t = 37.7$, $p \approx 0$), which is significantly better than any of the existing methods, with the closest competitor being REPET-SIM ($t = 7.75$, $p \approx 0$). For the SIR of the singing voice estimates, the universal voice model is not significantly different from the method of Durrieu et al. ($t = -0.29$, $p = .77$), but significantly better than the other existing methods, and also the combination ($t = 20.1$, $p \approx 0$).

Finally, for the SAR of the background music estimates, the universal voice model is competitive with REPET-SIM ($t = -1.26$, $p = 0.21$), but significantly worse than the other competitive methods ($t = 9.61$ and $t = 13.2$). On the other hand, in terms of the SAR of the singing voice estimates, the combination performs significantly better than the universal voice model ($t = 50.1$), which in turn is significantly better than the method of Durrieu et al. ($t = 9.69$). However, both are significantly worse than the other two competitors, the closest being the method of Huang et al. ($t = -15.6$).

Note that there are 7 tests for each configuration of the three measures (SDR, SIR, SAR) and the two sources (background music and singing voice): comparing the universal voice model and the combination to each of the three competitors, and then comparing the universal voice model to the combination. Therefore, we are implicitly conducting a total of $3 \times 2 \times 7 = 42$ tests. All of the findings above remain significant at the $\alpha = .05$ level if we use a Bonferroni correction to adjust for the 42 tests, corresponding to a rejection region of $|t| > 3.25$. These results confirm the findings in Figures 2, 3, and 4.

5. CONCLUSION

In this work, we proposed a method for modeling the singing voice. The method can learn a singer-independent model from singing examples using a NMF based technique. We

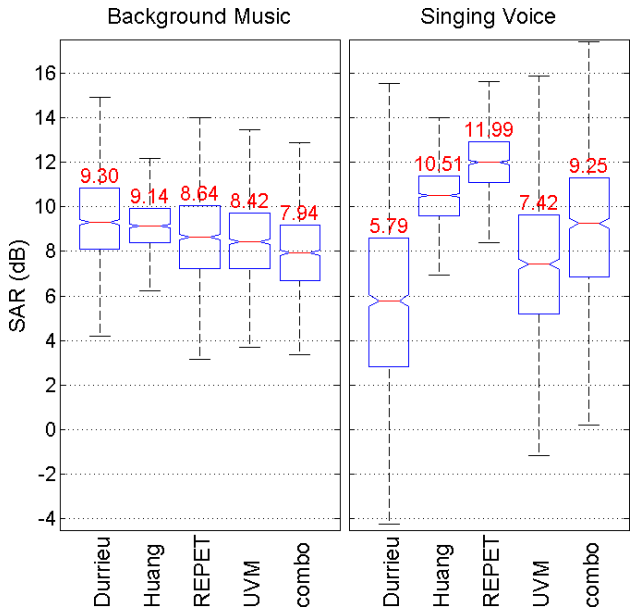

Figure 4. Box plots of the distributions for the SAR (dB).

then proposed to combine this method with a method that models the background music. Combining a method that specifically models the singing voice with a method that specifically models the background music addresses separation performance from the point of view of both sources.

Evaluation on a data set of 1,000 song clips showed that, when using the right parameter settings, the universal voice model can outperform different state-of-the-art methods. Combining modeling of both sources can further improve separation performance, when compared with modeling only one of the sources.

This work was supported in part by NSF grant number IIS-0812314.

6. REFERENCES

[1] Jean-Louis Durrieu, Bertrand David, and Gaël Richard. A musically motivated mid-level representation for pitch estimation and musical audio source separation. *IEEE Journal on Selected Topics on Signal Processing*, 5(6):1180–1191, October 2011.

[2] Jinyu Han and Ching-Wei Chen. Improving melody extraction using probabilistic latent component analysis. In *36th International Conference on Acoustics, Speech and Signal Processing*, Prague, Czech Republic, May 22-27 2011.

[3] Chao-Ling Hsu and Jyh-Shing Roger Jang. On the improvement of singing voice separation for monaural recordings using the MIR-1K dataset. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(2):310–319, February 2010.

[4] Po-Sen Huang, Scott Deeann Chen, Paris Smaragdis, and Mark Hasegawa-Johnson. Singing-voice separation from monaural recordings using robust principal component analysis. In *37th International Conference on Acoustics, Speech and Signal Processing*, Kyoto, Japan, March 25-30 2012.

[5] Yipeng Li and DeLiang Wang. Separation of singing voice from music accompaniment for monaural recordings. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(4):1475–1487, May 2007.

[6] Antoine Liutkus, Zafar Rafii, Roland Badeau, Bryan Pardo, and Gaël Richard. Adaptive filtering for music/voice separation exploiting the repeating musical structure. In *37th International Conference on Acoustics, Speech and Signal Processing*, Kyoto, Japan, March 25-30 2012.

[7] Alexey Ozerov, Pierrick Philippe, Frédéric Bimbot, and Rémi Gribonval. 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*, 15(5):1564–1578, July 2007.

[8] Zafar Rafii and Bryan Pardo. Music/voice separation using the similarity matrix. In *13th International Society for Music Information Retrieval*, Porto, Portugal, October 8-12 2012.

[9] Zafar Rafii and Bryan Pardo. REpeating Pattern Extraction Technique (REPET): A simple method for music/voice separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(1):71–82, January 2013.

[10] Paris Smaragdis, Bhiksha Raj, and Madhusudana Shashanka. Supervised and semi-supervised separation of sounds from single-channel mixtures. In *Independent Component Analysis and Signal Separation*, pages 414–421. Springer, 2007.

[11] Dennis L. Sun and Gautham J. Mysore. Universal speech models for speaker independent single channel source separation. In *38th International Conference on Acoustics, Speech and Signal Processing*, Vancouver, BC, Canada, May 26-31 2013.

[12] Shankar Vembu and Stephan Baumann. Separation of vocals from polyphonic audio recordings. In *6th International Conference on Music Information Retrieval*, pages 337–344, London, UK, September 11-15 2005.

[13] Emmanuel Vincent, Rémi Gribonval, and Cedric Févotte. Performance measurement in blind audio source separation. *IEEE Transactions on Audio, Speech and Language Processing*, 14(4):1462–1469, July 2006.

[14] Tuomas Virtanen, Annamaria Mesaros, and Matti Ryynänen. Combining pitch-based inference and non-negative spectrogram factorization in separating vocals from polyphonic music. In *ISCA Tutorial and Research Workshop on Statistical and Perceptual Audition*, pages 17–20, Brisbane, Australia, 21 September 2008.