VOCAL SEPARATION USING SINGER-VOWEL PRIORS OBTAINED FROM POLYPHONIC AUDIO

Shrikant Venkataramani[1], **Nagesh Nayak**[2], **Preeti Rao**[1], **and Rajbabu Velmurugan**[1]

[1]Department of Electrical Engineering , IIT Bombay , Mumbai 400076
[2]Sensibol Audio Technologies Pvt. Ltd.
[1]{vshrikant, prao, rajbabu}@ee.iitb.ac.in
[2]nageshsnayak@sensibol.com

ABSTRACT

Single-channel methods for the separation of the lead vocal from mixed audio have traditionally included harmonic-sinusoidal modeling and matrix decomposition methods, each with its own strengths and shortcomings. In this work we use a hybrid framework to incorporate prior knowledge about singer and phone identity to achieve the superior separation of the lead vocal from the instrumental background. Singer specific dictionaries learned from available polyphonic recordings provide the soft mask that effectively attenuates the bleeding-through of accompanying melodic instruments typical of purely harmonic-sinusoidal model based separation. The dictionary learning uses NMF optimization across a training set of mixed signal utterances while keeping the vocal signal bases constant across the utterances. A soft mask is determined for each test mixed utterance frame by imposing sparseness constraints in the NMF partial co-factorization. We demonstrate significant improvements in reconstructed signal quality arising from the more accurate estimation of singer-vowel spectral envelope.

1. INTRODUCTION

Source separation techniques have been widely applied in the suppression of the lead vocal in original songs to obtain the orchestral background for use in karaoke and remix creation. In stereo and multichannel recordings, spatial cues can contribute significantly to vocal separation from the original mixtures. However this separation is not complete, depending on the manner in which the multiple instruments are panned in the mix. Further, an important category of popular music recordings, dating until the 1950s in the West and even later in the rest of the world, are purely monophonic. Single-channel methods for the separation of the lead vocal from the instrumental background include harmonic sinusoidal modeling and matrix decomposition methods. Of these, harmonic sinusoidal modeling has found success in situations where no clean data is available for supervised learning [6], [10]. Based on the assumption that the vocal is dominant in the mixture, predominant pitch detection methods are applied to obtain the vocal pitch and hence the predicted vocal harmonic locations at each instant in time. Harmonic sinusoidal modeling is then applied to reconstruct the vocal component based on assigning a magnitude and phase to each reconstructed harmonic from a detected sinusoidal peak in the corresponding spectral neighborhood of the mixed signal short-time Fourier transform (STFT). The vocal signal is reconstructed by the amplitude and phase interpolation of the harmonic component tracks. The instrumental background is obtained by the subtraction of the reconstructed vocal from the original mixture. A high degree of vocal separation is obtained when the assumption of vocal dominance holds for the mixture. However some well-known artifacts remain viz. (i) “bleeding through” of some of the melodic instrumentation due to the blind assignment of the total energy in the mixed signal in the vocal harmonic location to the corresponding reconstructed harmonic; this artifact is particularly perceptible in the sustained vowel regions of singing, (ii) improper cancellation of the unvoiced consonants and breathy voice components due to the limitations of sinusoidal modeling of noise and (iii) residual of vocal reverb if present in the original [14]. To address the first shortcoming, recent methods rely on the availability of non-overlapping harmonics of the same source anywhere in the entire audio [3]. We propose to replace the binary mask (implicit in the harmonic-sinusoidal modeling) applied to the vocal harmonics before reconstruction by a soft-mask (a form of Wiener filtering). An effective soft mask would be based on an accurate estimate of the vocal signal spectrum at any time-instant [2], [14]. This would improve the reconstructed vocal signal and lead to more complete suppression in the estimated background.

The vocal signal spectrum depends on several factors such as the singer’s voice, the phone being uttered, the pitch and the vocal effort. We cannot assume the availability of clean data for supervised training (i.e., unaccompanied voice of the particular singer). However popular singers typically have a large number of songs to their credit, and therefore a method for learning a dictionary of soft masks for the singer from such a training data set could be useful. The training set thus has original single-channel polyphonic songs where the vocal characteristics correspond to the singer but the background orchestration is diverse. We apply non-negative matrix factorization (NMF) methods to estimate the invariant set of basis vectors across multiple instances of the singer's phones in different songs. In the recent past, several systems have been proposed that qualify as modifications of NMF for improved performance in various scenarios where specific prior knowledge about the data are available [5] (and references therein). In the present work, we attempt to formulate a NMF approach to obtain basis elements corresponding to the singer's utterances by providing audio corresponding to a particular singer. Given the very diverse spectra of the different phones in a language, the quality of the decomposition can be improved by restricting the optimization to within a phone class [11]. We exploit the availability of song-synchronized lyrics data available in karaoke applications to achieve this. Our main contribution is to combine the advantages of harmonic-sinusoidal modeling in localizing the vocal components in time-frequency with that of soft-masking based on spectral envelope estimates from a NMF decomposition on polyphonic audio training data. Prior knowledge about singer identity and underlying phone transcription of the training and test audio are incorporated in the proposed framework. We develop and evaluate the constrained NMF optimization required for the training across instances where a common basis function set corresponds to the singer-vowel. On the test data, partial co-factorization with a sparseness constraint helps obtain the correct basis decomposition for the mixed signal at any time instant, and thus a reliable spectral envelope estimate of the vowel for use in the soft mask. Finally, the overall system is evaluated based on the achieved vocal and orchestral background separation using objective measures and informal listening. In the next sections, we present the overall system for vocal separation, followed by the proposed NMF-based singer-vowel dictionary learning, estimation of the soft mask for test mixed polyphonic utterances and experimental evaluation of system performance.

2. PROPOSED HYBRID SYSTEM

A block diagram of the proposed hybrid system for vocal separation is shown in Figure 1. The single-channel audio mixture considered for vocal separation is assumed to have the singing voice, when present, as the dominant source in the mix. We assume that the sung regions are annotated at the syllable level, as expected from music audio prepared for karaoke use. A predominant pitch tracker [9] is applied to the sung regions to detect vocal pitch at 10 ms intervals throughout the sung regions of the audio. Sinusoidal components are tracked in the computed short-time magnitude spectrum after biasing trajectory information towards the harmonic locations based on the detected pitch [8]. The pitch salience and total harmonic energy are used to locate the vowel region within the syllable. The vocal signal can be reconstructed from the harmonic sinusoidal component trajectories obtained by amplitude and phase interpolation of the frame-level estimates from the STFT. An estimate of the instantaneous spectral envelope of the singer's voice provides a soft mask to re-shape the harmonic amplitudes before vocal reconstruction. The mel-filtered spectral envelope (MFS) is computed by applying a 40-band mel-filter bank to the log-linearly interpolated envelope of the mixture harmonic amplitudes. By using the spectral envelope, we eliminate pitch dependence in the soft mask to a large extent. The phoneme dictionary consists of a set of basis vectors for each vowel, at various pitches. A linear combination of these basis vectors may be used to estimate the MFS envelope of the vocal component of the mixture, from the MFS envelope of the mixture. These spectral envelope vectors are learnt from multiple polyphonic mixtures of the phoneme as explained in Section 3. The MFS is used as a low-dimensional perceptually motivated representation. The reconstructed vocal signal is subtracted in the time-domain from the polyphonic mixture to obtain the vocal-suppressed music background.

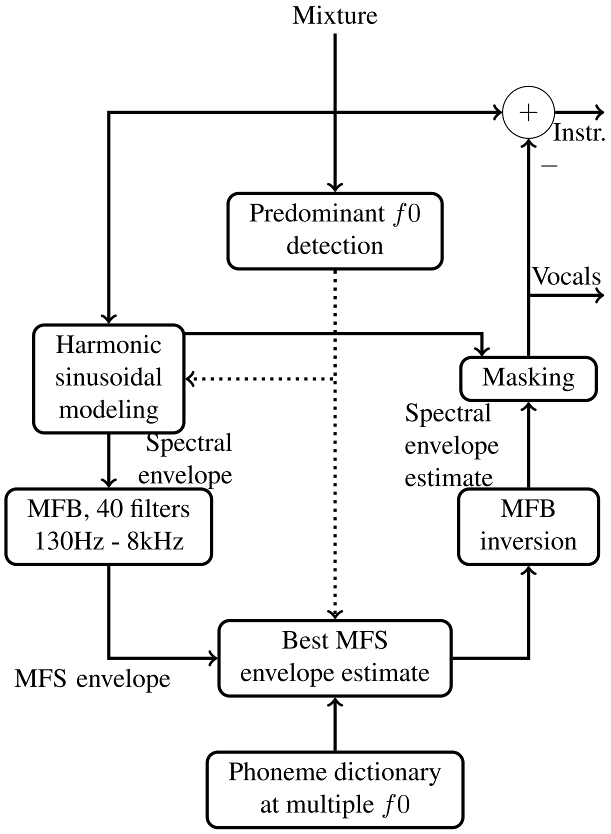
Mixture
+
Instr.
−
Predominant $f0$ detection
Vocals
Harmonic sinusoidal modeling
Masking
Spectral envelope
Spectral envelope estimate
MFB, 40 filters 130Hz - 8kHz
MFB inversion
Best MFS envelope estimate
MFS envelope
Phoneme dictionary at multiple $f0$

Figure 1. Block diagram of the proposed vocal separation system.

3. SPECTRAL ENVELOPE DICTIONARY LEARNING USING NMF

To obtain the singer specific soft mask mentioned in the previous section, we create a dictionary of basis vectors corresponding to each of the vowels of the language. This dictionary is created from polyphonic song segments, containing the vowel, of the singer under consideration. While spectral envelope of a vowel depends on the vowel identity, there are prominent dependencies on (i) the singer, whose physiological characteristics and singing style affect the precise formant locations and bandwidths for a given vowel. This is especially true of the higher formants (4th and 5th), which depend primarily on the singer rather than on the vowel; (ii) pitch, specifically in singing where the articulation can vary with large changes in pitch due to the "formant tuning" phenomenon [12]; (iii) loudness or vocal effort. Raising the vocal effort reduces spectral tilt, increasing the relative amplitudes of the higher harmonics and consequently the brightness of the voice.

In the proposed dictionary learning, pitch dependence is accounted for by separate dictionary entries corresponding to 2 or 3 selected pitch ranges across the 2 octaves span of a singer. Since the pitch and vowel identity are known for the test song segment, the correct dictionary can be selected at any time. The basis vectors for any pitch range of the singer-vowel capture the variety of spectral envelopes that arise from varying vocal effort and vowel context. We have training data of several instances of a particular singer uttering a common vowel. These utterances have been obtained from different songs and hence, we may assume that the accompaniments in the mixtures are different. The MFS envelopes, reviewed in the previous section, are extracted for each training vowel instance in the polyphonic audio. Based on the assumption of additivity of the spectral envelopes of vocal and instrumental background, there is a common partial factor corresponding to the singer-vowel across the mixtures with changing basis vectors for the accompaniment.

We use NMF to extract common features (singer-vowel spectra) across multiple song segments. The conventional use of NMF is similar to the phoneme-dependent NMF used for speech separation in [7] where the bases are estimated from clean speech. We extend the scope of NMF further, using non-negative matrix partial co-factorization (NMPCF) [4] equivalent to NMF for multiblock data [15] techniques. NMPCF and its variants have been used in drum source separation [4], where one of the training signals is the solo drums audio. Here, we use NMPCF for multiple MFS matrices of mixed signals across segments of the polyphonic audio of the singer, without the use of clean vocal signal. This will yield a common set of bases representing the singer-vowel and other varying bases representative of the accompaniments.

We now describe the NMPCF algorithm for learning the singer-vowel basis. The MFS representation for one specific annotated segment of a polyphonic music is represented as $\mathbf{V}_l$. This section has the vowel of interest and instrumental accompaniments. We have MFS of M such mixtures for $i = 1, \ldots, M$ represented as [15],

$$\mathbf{V}_i = \mathbf{V}_{\mathrm{c},i} + \mathbf{V}_{\mathrm{a},i}, \qquad i = 1, \ldots M. \tag{1}$$

where $\mathbf{V}_{\mathrm{c},i}$ and $\mathbf{V}_{\mathrm{a},i}$ denote the MFS of the common singer-vowel and accompaniment, respectively. Using NMF decomposition for the MFS spectra we have,

$$\mathbf{V}_i = \mathbf{W}_c\mathbf{H}_{c,i} + \mathbf{W}_{a,i}\mathbf{H}_{a,i}, \qquad i = 1, \ldots M. \tag{2}$$

where $\mathbf{W}_c \in \mathbb{R}_+^{F \times N_c}$ denotes the basis vectors corresponding to the common vowel shared by the M mixtures and $\mathbf{W}_{a,i} \in \mathbb{R}_+^{F \times N_a}$ are the basis vectors corresponding to the accompaniments. Here F is the number of mel-filters (40) used, N_c and N_a are the numer of basis vectors for the vowel and accompaniments, respectively. The matrices $\mathbf{H}_{c,i}$ and $\mathbf{H}_{a,i}$ are the activation matrices for the vowel and accompaniment basis vectors, respectively. Our objective is to obtain the basis vectors $\mathbf{W_c}$ corresponding to the common vowel across these M mixtures. We achieve this by minimizing the Frobenius norm $\parallel . \parallel_F^2$ of the discrepancy between the given mixtures and their factorizations, simultaneously. Accordingly, the cost function,

$$D = \sum_{i=1}^{M} \frac{1}{2} \parallel \mathbf{V}_i - \mathbf{W}_c\mathbf{H}_{c,i} - \mathbf{W}_{a,i}\mathbf{H}_{a,i} \parallel_F^2 + \frac{\lambda_1}{2} \parallel \mathbf{W}_{a,i} \parallel_F^2, \tag{3}$$

is to be minimized with respect to $\mathbf{W}_c$, $\mathbf{W}_{a,i}$, $\mathbf{H}_{c,i}$, and $\mathbf{H}_{a,i}$. The regularizer $\parallel \mathbf{W}_{a,i} \parallel_F^2$ and λ_1 the Lagrange multiplier lead to dense $\mathbf{W}_{a,i}$ and $\mathbf{W}_c$ matrices [15]. The basis vectors thus obtained are a good representation of both the common vowel and the accompaniments, across the mixture. In this work, we choose $\lambda_1 = 10$ for our experimentation as it was found to result in the sparsest $\mathbf{H}_{c,i}$ matrix for varying values of λ_1. We solve (3) using the multiplicative update algorithm. The multiplicative update for a parameter $\mathbf{P}$ in solving the NMF problem takes the general form,

$$\mathbf{P} = \mathbf{P} \odot \frac{\nabla_{\mathbf{P}}^{-}(D)}{\nabla_{\mathbf{P}}^{+}(D)}, \tag{4}$$

where $\nabla_{\mathbf{X}}^{-}(D)$ and $\nabla_{\mathbf{X}}^{+}(D)$ represent the negative and positive parts of the derivative of the cost D w.r.t. the parameter $\mathbf{X}$, respectively, $\odot$ represents the Hadamard (element-wise) product and the division is also element-wise. Correspondingly, the multiplicative update for the parameter $\mathbf{W}_c$ in (3) is,

$$\mathbf{W}_c = \mathbf{W}_c \odot \frac{\nabla_{\mathbf{W}_c}^{-}(D)}{\nabla_{\mathbf{W}_c}^{+}(D)}. \tag{5}$$

where,

$$\nabla_{\mathbf{W}_c}(D) = \sum_{i=1}^{M} \left(\mathbf{W}_c\mathbf{H}_{c,i} + \mathbf{W}_{a,i}\mathbf{H}_{a,i} - \mathbf{V}_i\right)\mathbf{H}_{c,i}^T. \tag{6}$$

Similarly, the update equation for other terms in (3) are,

$$\mathbf{H}_{c,i} = \mathbf{H}_{c,i} \odot \frac{\mathbf{W}_c^T\mathbf{V}_i}{\mathbf{W}_c^T\mathbf{W}_c\mathbf{H}_{c,i} + \mathbf{W}_c^T\mathbf{W}_{a,i}\mathbf{H}_{a,i}}, \tag{7}$$

$$\mathbf{H}_{a,i} = \mathbf{H}_{a,i} \odot \frac{\mathbf{W}_{a,i}^T\mathbf{V}_i}{\mathbf{W}_{a,i}^T\mathbf{W}_c\mathbf{H}_{c,i} + \mathbf{W}_{a,i}^T\mathbf{W}_{a,i}\mathbf{H}_{a,i}}, \tag{8}$$

$$\mathbf{W}_{a,i} = \mathbf{W}_{a,i} \odot \frac{\mathbf{V}_i \mathbf{H}_{a,i}^T}{\mathbf{W}_c \mathbf{H}_{c,i} \mathbf{H}_{a,i}^T + \mathbf{W}_{a,i} \mathbf{H}_{a,i} \mathbf{H}_{a,i}^T + \lambda_1 \times \mathbf{W}_{a,i}}. \quad (9)$$

for $i = 1, \ldots, M$. The basis vectors $\mathbf{W_c}$ for the various phonemes form the dictionary and act as a prior in the spectral envelope estimation. Each dictionary entry is associated with a vowel and pitch range. We denote each entry in the dictionary as $\mathbf{W}_c(/p/, f0)$ for each vowel $/p/$ at pitch $f0$.

4. SOFT MASK ESTIMATION USING SINGER-VOWEL DICTIONARY

In this section, we describe the approach to estimate the frame-wise soft mask for a test polyphonic vowel mixture segment. We first obtain the MFS envelope for the mixture as mentioned in Section 2. With the vowel label and pitch range known, we obtain the corresponding set of basis vectors $\mathbf{W}_c(/p/, f0)$ from the dictionary. Given this MFS representation of the mixture and the basis vectors, our objective is to separate the vocal component from the mixture. We do this by minimizing the cost function

$$D_T = \frac{1}{2} \parallel \mathbf{V}_T - \mathbf{W}_c \mathbf{H}_{c,T} - \mathbf{W}_{a,T} \mathbf{H}_{a,T} \parallel_F^2 + \frac{\lambda_2}{2} \parallel \mathbf{H}_{c,T} \parallel_F^2, \quad (10)$$

where the subscript T refers to the test case. The minimization is done with the dictionary bases $\mathbf{W}_c$ kept fixed and using multiplicative updates for $\mathbf{H}_{c,T}$, $\mathbf{W}_{a,T}$ and $\mathbf{H}_{a,T}$. The sparsity constraint on $\mathbf{H}_{c,T}$ in (10) accounts for the fact that the best set of bases representing the vowel would result in the sparsest temporal matrix $\mathbf{H}_{c,T}$. Under this formulation, $\mathbf{W}_c \mathbf{H}_{c,T}$ will give an estimate of the vowel's MFS envelope $\mathbf{V}_c$ (as in (1)) for the mixture. An alternate way is to use Wiener filtering to estimate $\mathbf{V}_c$ as,

$$\widehat{\mathbf{V}}_c = \frac{\mathbf{W}_c \mathbf{H}_{c,T}}{\mathbf{W}_c \mathbf{H}_{c,T} + \mathbf{W}_{a,T} \mathbf{H}_{a,T}} \odot \mathbf{V}_T. \quad (11)$$

This estimated vowel MFS can be used to reconstruct the spectral envelope of the vowel $\hat{\mathbf{C}}$. This is done by multiplying $\widehat{\mathbf{V}}_c$ with the pseudoinverse of the DFT matrix $\mathbf{M}$ of the mel filter bank [1] as $\hat{\mathbf{C}} = \mathbf{M}^\dagger \widehat{\mathbf{V}}_c$. A soft mask corresponding to this spectral envelope can be obtained using the Gaussian radial basis function [2],

$$\mathbf{G}_b(f, t) = \exp\left(-\frac{\left(\log \mathbf{X}(f, t) - \log \hat{\mathbf{C}}(f, t)\right)^2}{2\sigma^2} \right) \quad (12)$$

where, σ is the Gaussian spread, $\mathbf{X}$ is the magnitude spectrum of the mixed signal. The soft mask (12) is evaluated with $\sigma = 1$, in a 50 Hz band around the pitch ($f0$) and its harmonics [14].

Having obtained the soft mask, the vocals track is reconstructed by multiplying the soft mask with the harmonic amplitudes of the sinusoidally modeled signal. The resynthesized signal then corresponds to the reconstructed vocals. The accompaniment can be obtained by performing a time-domain subtraction of the reconstructed vocals from the original mixture.

5. EXPERIMENTS AND PARAMETER CHOICES

Given a polyphonic vowel segment, the vocal is separated by applying the generated soft mask corresponding to the given mixture. We compare the separated vocal with the ground truth to evaluate the performance. The performance evaluation of the proposed system is carried out in two steps. The first step is to choose the parameters of the system using the distance in the MFS space between the estimated and ground-truth MFS vectors obtained from the clean utterance. The second step is the computation of signal-to-distortion (SDR) measure (in dB) on the separated vocal and instrumental time-domain signals which will be given in Section 6. We present the training and test data used in the experiments next.

5.1 Description of the Dataset

The training dataset comprised of nine instances of three vowels viz., /a/, /i/, /o/ at two average pitches of 200 Hz and 300 Hz and sung by a male singer over three different songs with their accompaniments, annotated at the phoneme level. The training data was chosen so as to have different accompaniments across all the instances of a vowel. The training audios thus contained the vowel utterances throughout in the presence of background accompaniments. The training mixtures were pre-emphasised using a filter with a zero located at 0.7 to better represent the higher formants. A dictionary of bases was created for all the vowels for the two pitch ranges using the NMCPF optimization procedure discussed in Section 3. The performance was evaluated over a testing dataset of 45 test mixtures with 15 mixtures for each vowel over the two pitch ranges. The mixtures used for testing were distinct from the training mixtures. Since the audios were obtained directly from full songs, there was a significant variation in terms of the pitch of the vowel utterances around the average pitch ranges and in terms of coarticulation. The training and testing mixtures were created in a karaoke singing context and hence, we had available, the separate vocal and accompaniment tracks to be used as ground truth in the performance evaluation. All the mixtures had durations in the range of 400 ms - 2.2 s and were sampled at a frequency of 16 kHz. The window size and hop size used for the 1024 point STFT were 40 ms and 10 ms, respectively.

5.2 Choice of Parameters

There are several training parameters likely to influence the performance of the system. These include the ranks of the matrices in the decomposition ($\mathbf{W}_c$, $\mathbf{W}_a$) and the number of mixtures M. We obtain these parameters experimentally using a goodness-of-fit measure. The goodness-of-fit is taken to be the normalised Frobenius norm of the

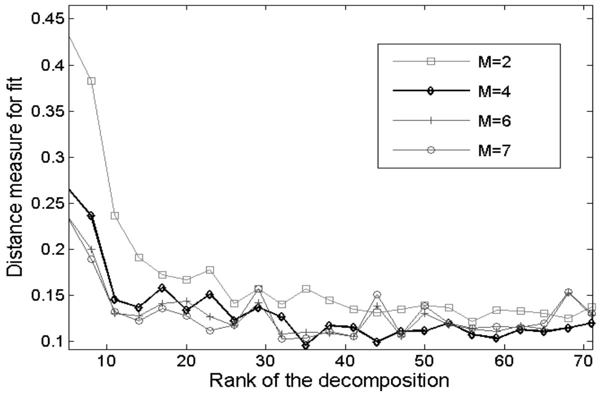

Figure 2. Goodness-of-fit, averaged over all phoneme bases, with increasing number of mixtures (M) used in the parallel optimization procedure for different decomposition ranks. The distance measure decreases indicating an improving fit as the number of mixtures and rank increase.

difference between the ideal envelope of a vowel in the MFS domain $\mathbf{V}_c$ and its best estimate $\widehat{\mathbf{V}}_c$ obtained as a linear combination of the bases, for a mixture containing the vowel and accompaniment. This estimate can be calculated as explained in Section 4. Lower the value of this distance measure, closer is the envelope estimate to the ideal estimate. The various bases may be compared by calculating D_i for different bases $\mathbf{W}_{c_i}$ using

$$\begin{aligned} D_i &= \frac{\| \mathbf{V}_c - \widehat{\mathbf{V}}_{c_i} \|_F^2}{\| \mathbf{V}_c \|_F^2} \\ &= \frac{\| \mathbf{V}_c - \frac{\mathbf{W}_{c_i}\mathbf{H}_{c_i}}{\mathbf{W}_{c_i}\mathbf{H}_{c_i}+\mathbf{W}_{a_i}\mathbf{H}_{a_i}} \odot \mathbf{V}_T \|_F^2}{\| \mathbf{V}_c \|_F^2}, \quad (13) \end{aligned}$$

and comparing the same. To account for variabilities in the evaluation mixtures, the distance measure is evaluated and averaged over a number of mixtures and combinations, for each set of bases $\mathbf{W}_{c_i}$. The goodness-of-fit is used only to choose the appropriate parameter values for the system. The performance of the overall system, however, is evaluated in terms of SDR.

As shown in Figure 2, the goodness-of-fit measure decreases with increasing rank of the decomposition (number of vocal basis vectors) for a given M. The decreasing trend flattens out and then shows a slight increase beyond rank 35. For a fixed rank, the goodness-of-fit improves with increasing number of mixtures. Of the configurations tried, the distance measure is minimum when four mixtures ($M = 4$) are used in the NMPCF optimization to obtain the dictionary. Thus, a rank 35 decomposition with $M = 4$ is chosen for each singer-vowel dictionary for system performance evaluation.

As for the rank of the accompaniment basis, it is observed that the regularization term in the joint optimization (3) seems to make the algorithm robust to choice of number of basis vectors for the accompaniment. Eight basis vectors were chosen for each mixture term in the joint optimization. Although the number of accompaniment bases seems to be comparatively low, eight bases were sufficient to reconstruct the accompaniment signals from the mixtures. A high value for λ_2 in the test optimization problem of (10) results in a decomposition involving a linear combination of the least number of bases per time frame. This sparse decomposition may not necessarily lead to the best reconstruction in more challenging scenarios involving articulation variations. Thus a small value of $\lambda_2 = 0.1$ was chosen.

Separated track	Binary mask	Soft mask	
		Original singer	Alternate singer
Vocal	8.43	9.06	8.66
Instrumental	13.63	14.16	14.10

Table 1. SDR values (in dB) for separated vocals and instruments obtained using a binary mask, soft mask from the original singer's training mixtures and soft mask from an alternate singer's vocals.

6. RESULTS AND DISCUSSION

We evaluate the performance of the system using the SDR. The SDR is evaluated using the BSS_eval toolbox [13]. The SDR values averaged across 45 vowel test mixtures, separately for the reconstructed vocals and instrumental background are given in Table 1. To appreciate the improvement, if any, the SDR is also computed for the harmonic sinusoidal model without soft masking (i.e., binary masking only). While the proposed soft masking shows an increase in SDR, closer examination revealed that the improvements were particularly marked for those mixtures with overlapping vocal and instrumental harmonics (accompaniments) in some spectral regions. This was also borne out by informal listening. When we isolated these samples, we observed SDR improvements of up to 4 dB in several instances. This is where the selective attenuation of the harmonic amplitudes in accordance with the estimated vowel spectral envelope is expected to help most. The harmonics in the non-formant regions are retained in the instrumental background rather than being canceled out as in binary masking, contributing to higher SDR [1] .

To understand the singer dependence of the dictionary, we carried out soft mask estimation from the polyphonic test mixtures using the basis vectors of an alternate singer. This basis was a set of clean vowel spectral envelopes obtained from another male singer's audio with the same vowels and pitches corresponding to our training dataset. We observe from Table 1 that the alternate singer soft mask does better than the binary mask, since it brings in the vowel dependence of the soft mask. However, it does not perform as well as the original singer's soft mask even though the latter is obtained from clean vowel utterances. As depicted in Figure 3 (for a sample case), the envelope obtained using the original singer's data closely follows the

[1] Audio examples are available at `http://www.ee.iitb.ac.in/student/~daplab/ISMIR_webpage/webpageISMIR.html`.

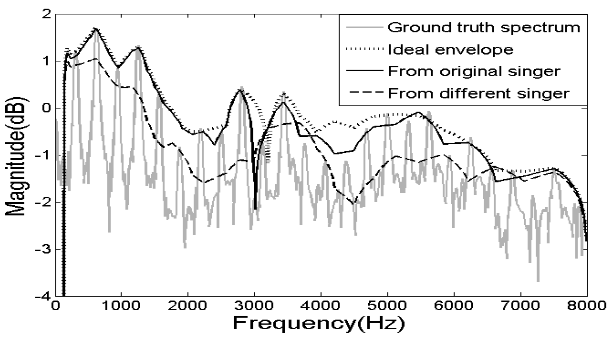

Figure 3. Comparison of the reconstructed phoneme envelope of the phoneme /a/ obtained from training mixtures of the same singer and with that obtained from pure vocals of a different singer.

ideal envelope of the phoneme.

Although the NMF optimization converges slowly, the number of iterations to be carried out to obtain the envelope is low, for both training and testing procedures. It is observed that the bases and envelopes attain their final structure after 4000 and 1000 iterations, respectively.

7. CONCLUSION

Soft masks derived from a dictionary of singer-vowel spectra are used to improve upon the vocal-instrumental music separation achieved by harmonic sinusoidal modeling for polyphonic music of the particular singer. The main contribution of this work is an NMF based framework that exploits the amply available original polyphonic audios of the singer as training data for learning the dictionary of singer spectral envelopes. Appropriate constraints are introduced in the NMF optimization for training and test contexts. The availability of lyric-aligned audio (and therefore phone labels) helps to improve the homogeneity of the training data and have a better model with fewer basis vectors. Significant improvements in reconstructed signal quality are obtained over binary masking. Further it is demonstrated that a vowel-dependent soft mask obtained from clean data of a different available singer is not as good as the singer-vowel dependent soft mask even if the latter is extracted from polyphonic audio.

8. ACKNOWLEDGEMENT

Part of the work was supported by Bharti Centre for Communication in IIT Bombay.

9. REFERENCES

[1] L. Boucheron and P. De Leon. On the inversion of mel-frequency cepstral coefficients for speech enhancement applications. In *Int. Conf. Signals Electronic Systems, 2008.*, pages 485–488, 2008.

[2] D. Fitzgerald. Vocal separation using nearest neighbours and median filtering. In *Irish Signals Systems Conf.*, 2012.

[3] J. Han and B. Pardo. Reconstructing completely overlapped notes from musical mixtures. In *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Process., 2011 (ICASSP '11.)*, pages 249 – 252, 2011.

[4] M. Kim, J. Yoo, K. Kang, and S. Choi. Nonnegative matrix partial co-factorization for spectral and temporal drum source separation. *IEEE Journal Selected Topics Signal Process.*, 5(6):1192 – 1204, 2011.

[5] A. Lefevre, F. Bach, and C. Fevotte. Semi-supervised NMF with time-frequency annotations for single-channel source separation. In *Proc. Int. Soc. Music Information Retrieval (ISMIR 2012)*, pages 115–120, 2012.

[6] Y. Li and D. Wang. Separation of singing voice from music accompaniment for monaural recordings. *IEEE Trans. Audio, Speech, Lang. Process.*, 15(4):1475–1487, 2007.

[7] B. Raj, R. Singh, and T. Virtanen. Phoneme-dependent NMF for speech enhancement in monaural mixtures. In *Proc. Interspeech*, pages 1217–1220, 2011.

[8] V. Rao, C. Gupta, and P. Rao. Context-aware features for singing voice detection in polyphonic music. In *Proc. Adaptive Multimedia Retrieval*, Barcelona, Spain, 2011.

[9] V. Rao and P. Rao. Vocal melody extraction in the presence of pitched accompaniment in polyphonic music. *IEEE Trans. Audio, Speech, Lang. Process.*, 18(8):2145–2154, 2010.

[10] M. Ryynanen, T. Virtanen, J. Paulus, and A. Klapuri. Accompaniment separation and karaoke application based on automatic melody transcription. In *IEEE Int. Conf. Multimedia Expo*, pages 1417–1420, 2008.

[11] M. Schmidt and R. Olsson. Single-channel speech separation using sparse non-negative matrix factorization. In *Proc. of Interspeech*, pages 2617–2614, 2006.

[12] J. Sundberg. The science of the singing voice. *Music Perception: An Interdisciplinary Journal*, 7(2):187 – 195, 1989.

[13] E. Vincent, R. Gribonval, and C. Fevotte. Performance measurement in blind audio source separation. *IEEE Trans. Audio, Speech, Lang. Process.*, 14(4):1462–1469, 2006.

[14] T. Virtanen, A. Mesaros, and M. Ryynänen. Combining pitch-based inference and non-negative spectrogram factorization in separating vocals from polyphonic music. In *ICSA Tutorial and Research Workshop on Statistical and Perceptual Audition*, Brisbane, Australia, 2008.

[15] G. Zhou, A. Cichocki, Q. Zhao, and S. Xie. Nonnegative matrix and tensor factorizations: An algorithmic perspective. *IEEE Signal Process. Magazine*, pages 54–65, 2014.