# SEPARATING A FOREGROUND SINGER FROM BACKGROUND MUSIC

*Bhiksha Raj, Paris Smaragdis, Madhusudhana Shashanka*

Mitsubishi Electric Research Labs
Cambridge MA 02139

*Rita Singh*

Carnegie Mellon University
Pittsburgh PA 15213

**ABSTRACT**

In this paper we present a algorithm for separating singing voices from background music in popular songs. The algorithm is derived by modelling the magnitude spectrogram of audio signals as the outcome of draws from a discrete bi-variate random process that generates time-frequency pairs. The spectrogram of a song is assumed to have been obtained through draws from the distributions underlying the music and the vocals, respectively. The parameters of the underlying distribuiton are learnt from the observed spectrogram of the song. The spectrogram of the separated vocals is then derived by estimating the fraction of draws that were obtained from its distribution. In the paper we present the algorithm within a framework that allows personalization of popular songs, by separating out the vocals, processing them appropriately to one's own tastes, and remixing them. Our experiments reveal that we are effectively able to separate out the vocals in a song and personalize them to our tastes.

***Index Terms***— Probabilistic Latent Component Decomposition, Signal Separation

## 1. INTRODUCTION

We introduce a framework for personalizing music by changing its inherent characteristics through signal processing. In this framework, pre-recorded music, as exemplified by popular movie songs and independent albums by singers in popular genres worldwide, is first separated into its components, modified automatically and remixed to sound personally pleasing to an individual listener.

Our motivation for this was initially to make some extremely high-pitched female vocals produced in Indian movies sound more pleasing by bringing down the pitch of the singer to a softer, more natural level without affecting the overall quality of the song and background music. Note that in making this statement we neither intend to criticize Indian female singers, nor Indian listeners who find high pitched voices pleasing to the ear. We merely bring to attention the well-known fact that music is an acquired taste in human beings, and what may sound pleasing to a group of people may not sound equally pleasing to another group who may have been exposed to different strains of music altogether. We realize that in most cases, these songs are beautiful creations otherwise, and our attempt was initially to merely create the technology that would present this facet of Indian popular music to the world. In retrospect, we found that the uses of such a framework can be numerous, as we will later explain in this paper.

To understand how our framework functions, we need to first understand how the majority of studio-recorded studio music is currently produced throughout the world. A good piece of popular music, such as an Indian movie song, is usually a pleasing combination of some background music and one or more foreground singing voices. In a typical production, multiple channels of music and the singer are separately recorded. Individual channels are edited and/or corrected, their relative levels are adjusted, and the signals are mixed down to a small number of channels, typically two. The final sounds we hear are the outcome of this process.

The development of our framework begins with addressing the problem of reversal of this process. Given a segment of a song inclusive of vocals and background music, is it possible to separate these components out to extract, say, the singer in isolation? This is the topic we address in this paper. We do not attempt to completely invert the process of mixing to separate the song out into all of the component channels (although such separation is certainly not beyond the scope of the technique presented here); we are content to separate the foreground singer from the background music.

The separation of foreground vocals from background musical accompaniment is a non-trivial task that has so far not attracted much attention in the scientific community, although several parallel topics such as automatic transcription of music, separation of musical constituents from an ensemble, and separation of mixed speech signals have all garnered significant attention in recent times. Literature on the topic of separating vocals from background music is relatively sparse. Li and Wang [1] attempt to perform the separation using principles of Computational Auditory Scene Analysis (CASA). In this approach, the pitch of the foreground voice is detected, and spectro-temporal components that are presumed to belong to the voice are identified from the pitch and other auditory principles and grouped together to extract the spectrum (from which, in turn, the signal is extracted) for the voice. Similar CASA-based techniques have also been attempted by Wang [2]. Meron and Hirose [3] attempt to solve the simpler problem of separating background piano sounds from a singing voice. Sinusoidal components are learned for both the piano and the voice from training examples and are used to perform separation using a least-square approach. Alternately, the musical score for the background is used as prior information to enable the separation. Other proposals for separation of music from singing voices have also followed similar approaches, namely those of utilizing either explicitly stated harmonic relationships between spectral peaks, or through prior knowledge obtained from a musical score.

The framework described in this paper, on the other hand, does not take any of the approaches mentioned above. Instead, it is built upon a purely statistically driven method, where the song is hypothesized as the combined output of two generative models, one that generates the singing voice and the other the background music. What distinguishes our approach from other statistical methods for signal separation (e.g. [4], [5]) is the nature of the statistical model used. We model individual frequencies as the outcomes of draws from a discrete random process, and magnitude spectra of the signal as the outcome of several draws from this process. The model is perfectly additive in which the spectrogram of a mixed signal is simply modeled as the cumulative histogram of the outcome of draws from the processes underlying each of its constituent signals. The problem of

separating the music from the vocals then reduces to the problem of deducing which fraction of each spectro-temporal component of the mixed signal can be attributed to each of the two, given generative models for both the music and the voice. Although the parameters of the models for the two themselves must be learnt, the nature of the algorithm is such that they can be learned on the fly from the song itself. We note that we do not attempt to automatically identify the regions of the recording that contain voice. Rather, we assume that the boundaries of these regions are either given, or are generated manually. The goal here is primarily to separate out the vocals from the song and the problem of automatically detecting exactly where the vocals lie is not (and need not be) addressed.

For the purposes of this paper we define *personalization* as "the ability to process the voice (or the music) in a manner that appeals to a user and produces a personalized version of the song for the user". In addition to personalization, separating vocals from background music can have other important uses, such as supporting automatic transcription of the background music, supporting automatic identification of the lyrics, acoustic event (or musical phrase) identification for indexing purposes etc.

The rest of this paper is arranged as follows: In Section II, we describe the basic representation of the signal used by our framework. In Section III we describe our statistical model to represent signal spectra. In Section IV we describe a supervised signal separation algorithm that forms the basis of the our algorithm for separating vocals from music, which in turn is presented in Section V. In Section VI we discussion the framework for personalization of songs. In Section VII we describe experiments evaluating the algorithm and the signals produced by it. We show that not only are we able to separate songs effectively, but are also able to modify the separated sounds to personalize a song. Finally in Section VIII we present our conclusions.

## 2. REPRESENTING THE SIGNAL

The first step in any audio processing algorithm is that of coming up with an adequate representation for the audio signal. We convert the input audio signal to a spectrogram prior to further processing. The spectrogram is obtained through the application of a short-time Fourier transform to the signal: the signal is segmented into "frames" that are 64ms long. Adjacent frames overlap by 48ms. A Hanning window is applied to each frame and a DFT is computed from it. The sequence of spectral vectors thus obtained constitutes the spectrogram for the signal. Each component of the DFT of each frame represents the contribution of a specific frequency to the signal within a specific window of time. We will refer to these components as a *time-frequency* component. The spectrogram may also be inverted to retrieve the time-domain signal through the application of the inverse DFT to each spectral vector, using the standard overlap-add method to combine the segments of the signal obtained from individual DFTs.

Each element of the spectrogram is a complex number, comprising a magnitude and a phase. The information in the signal, however, is largely encoded by the magnitude of the spectrogram. It is well known that it is possible to reconstruct perfectly intelligible signals from a spectrogram even when the phases of the time-frequency components have been completely altered. Figure 1 shows the pictorial representation of the spectrogram of a singing voice. The $X$ axis in the figure represents time (or, more accurately, the index of the spectral vectors in the spectrogram) and the $Y$ axis represents frequency. The color of each point in the figure represents the magnitude of the specific time-frequency component. Several clear spectral patterns are evident in the figure. These patterns are characteristic of the underlying sounds. They are typically different for different speakers or singers, for different musical instruments (or musical ensembles) etc. and can be treated as signatures of the singer, speaker or other processes creating each strain of an underlying sound.

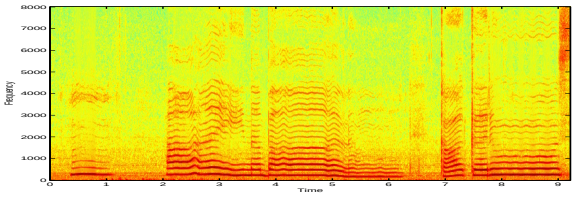


**Fig. 1**. Spectrogram of a female singing voice.

In our framework, we use the magnitude spectrogram to represent speech. The statistical models discussed below are then used to model the magnitude spectrograms. The models are used in the separation method presented in Section 4 that also separates out the magnitude spectrogram of the "component" signals. In order to obtain a separated time-domain signal, the phase of the spectrogram of the original (mixed) song is imposed on separated magnitude spectra, and the resulting complex spectrogram is inverted through an inverse short-time Fourier transform.

## 3. STATISTICAL MODEL FOR SIGNAL SPECTRA

The magnitude spectrogram for a signal is a two-dimensional data structure, comprising a sequence of magnitude spectral vectors, and can be represented as a matrix. Let $S(t, f)$ represent the $f^{\text{th}}$ frequency component of the $t^{\text{th}}$ vector in the sequence.

We model the matrix as the histogram of outcomes of draws from a discrete bivariate distribution $P(t, f)$, per the model described in Smaragids and Raj [6]. According to the model, each draw from the distribution will produce a single *quantum* of the time-frequency pair $(t, f)$. The quantum referred to need not necessarily represent a single *instance* of $(t, f)$; rather a draw of a large number $Q$ of quanta of $(t, f)$ will result in a single instance of $(t, q)$. Drawing of less than $Q$ quanta will result in a non-integral count of observations of $(t, f)$. For the purpose of the analysis presented in this paper the value $Q$ need not be known.

Thus, the model assumes that there is a bi-variate distribution *underlying* the spectrum and that the spectrum itself is the outcome of draws from it. We note that it is not uncommon to model signal spectra as the outcome of draws from a random process. However, what distinguishes the proposed model is the description of the primary random variable. Conventional models assume that the the result of a draw from an underlying distribution is the *value* of the spectrum at a given $(t, f)$. In our model, the time-frequency pair $(t, f)$ itself is the random variable, and the value of the spectrum at $(t, f)$ equals the number of times that time-frequency pair was drawn from the underlying distribution.

The distribution $P(t, f)$ represents the *joint* distribution of the time random variable $t$ and the frequency random variable $f$. We decouple the time and frequency variables through a latent variable model as follows:

$$P(\mathbf{x}) = \sum_z P(z)P(t|z)P(f|z) \tag{1}$$

where $z$ represents a *latent* or unseen variable $z$. $z$ is a discrete

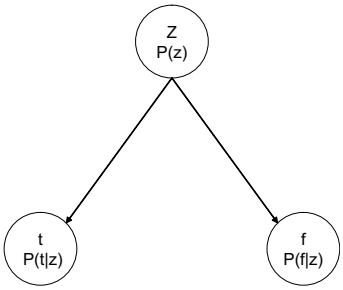


**Fig. 2**. Graphical representation of the generating process for a signal. A latent variable $z$ selects both a marginal time distribution ($P(t|z)$) and a frequency marginal distribution ($P(f|z)$). The time and frequency variables are drawn from these distributions.

RV that can take only a small set of values. Associated with each $z$ are $P(f|z)$, the marginal distribution of the frequency variable $f$, and $P(t|z)$, the marginal distribution of the time variable. The overall generating model for this process is as follows: to generate a $(t, f)$ pair the process first draws a latent variable $z$, then draws $t$ and $f$ independently from from the latent-variable-conditioned marginal distributions $P(t|z)$ and $P(f|z)$. The overall generating model is represented graphically by Figure 2.

The model represented by Equation 1 can also be represented algebraically by the following matrix expression:

$$\mathbf{P}_X = \mathbf{FZT} \tag{2}$$

where $\mathbf{P}_X$ is an $N_f \times N_t$ matrix whose elements are $P(t, f)$, where $N_f$ and $N_t$ represent the total number of frequency and time indices respectively, $\mathbf{F}$ is an $N_f \times N_z$ matrix whose entries are the probability values $P(f|z)$, where $N_z$ represents the total number of possible values for the latent variable $z$, $\mathbf{Z}$ is an $N_z \times N_z$ diagonal matrix whose columns are $P(z)$, and $\mathbf{T}$ is an $N_z \times N_t$ matrix whose entries are $P(t|z)$. Since they represent probability terms, the columns of $\mathbf{F}$, the diagonal terms of $\mathbf{Z}$ and the rows of $\mathbf{T}$ must all sum to 1.0. Equation 2 represents the columns of $\mathbf{P}$ as linear combinations of the columns of $\mathbf{F}$. If the columns of $\mathbf{F}$ are viewed as spectral basis vectors, $\mathbf{ZT}$ represents the projection of the columns of $\mathbf{P}$ onto the space spanned by the basis vectors in $\mathbf{F}$. If we represent the magnitude spectrogram for the signal generated from $\mathbf{P}$ by $\mathbf{S}$, $\mathbf{ZT}$ also represents a normalized projection of the spectral vectors onto the basis vectors in $\mathbf{F}$. Each $j^{\text{th}}$ row of $\mathbf{T}$ gives the relative contribution of the corresponding basis vector (column of $\mathbf{F}$) as a function of time.

As is clear from Equation 2, the same set of $P(f|z)$ terms are used to compose every column of $\mathbf{P}_X$, and thereby every spectral vector in $\mathbf{S}$. Thus, the $P(f|z)$ terms may be considered the *building blocks* that compose the the given sound.

The $P(f|z)$ terms can be learned along with the $P(z)$ and $P(t|z)$ terms from the spectrogram $\mathbf{S}$ using an Expectation Maximization algorithm, which gives us the following iterative update rules:

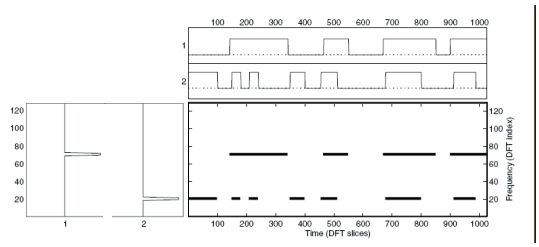


**Fig. 3**. Center panel: the spectrogram of a signal that consists of two tones turning on and off. Left panel: The two marginal frequency distributions obtained from it. One identifies the frequency of the first tone, the other the second tone. In this panel the Y axis represents frequency and the X axis represent the index of latent variable. Top panel: The marginal time distributions obtained. Each of the two distributions identifies the times at which one of the two tones occurs. Here the X axis represent time and the Y axis represent latent variable index.

$$P(z|f,t) = \frac{P(z)P(t|z)P(f|z)}{\sum_{z'} P(z')P(t|z')P(f|z')} \tag{3}$$

$$P(z) = \frac{\sum_t \sum_f P(z|t,f)S(t,f)}{\sum_{z'} \sum_t \sum_f P(z'|t,f)S(t,f)} \tag{4}$$

$$P(t|z) = \frac{\sum_f P(z|t,f)S(t,f)}{\sum_{t'} \sum_f P(z|t',f)S(t',f)} \tag{5}$$

$$P(f|z) = \frac{\sum_t P(z|t,f)S(t,f)}{\sum_{f'} \sum_t P(z|t,f')S(t,f')} \tag{6}$$

The left panel in Figure 3 shows the basis vectors derived using the above algorithm for a simple example where the signal consists simply of a mixture of two tones turning on and off. In this example we have assumed that the latent variable $z$ can only take two values. We note that the two corresponding marginal frequency distributions clearly capture the two building blocks for the signal, i.e. the two tones that compose the spectrogram. The corresponding $P(t|z)$ sequences also accurately represent the time instants at which these tones occur.

## 4. SEPARATING COMPONENT SIGNALS FROM A MIXTURE

The statistical model presented in Section 3 can be used to separate out component signals from a signal, such as the speakers from a mixed recording [7]. The set of basis vectors described by the frequency marginals $P(f|z)$ are learned for each component signal in the mixture from a separate unmixed training recording. Let $P_i(t, f)$ represent the distribution underlying the spectrogram of the $i^{\text{th}}$ component signal, and let $P_i(f|z)$ represent the frequency marginals learned for the $i^{\text{th}}$ component signal in the mixture. By the model, the spectrogram of the mixed signal is obtained through draws from the distributions of all component signals, since the spectrum of the mixed signal is obtained by addition of the spectra of the component signals. The overall distribution underlying the mixed signal, $P_{mixed}(t, f)$, is hence given as a linear combination of the distributions for the individual constituents:

$$P_{mixed}(t,f) = P(S_1)P_1(t,f) + P(S_2)P_2(t,f)\ldots \quad (7)$$

where $P(S_i)$ is the proportion of draws in the final spectrum that was drawn from the distribution of the $i^{\text{th}}$ speaker. Using the decomposition of Equation 1, this can be written as

$$\begin{aligned} P_{mixed}(t,f) &= P(S_1)\sum_z P_1(z)P_1(f|z)P_1(t|z) \\ &\quad + P(S_2)\sum_z P_2(z)P_2(f|z)P_2(t|z)\ldots (8) \end{aligned}$$

where $P_i(z)$ represents the probability distribution of the latent variable $z$ for the $i^{\text{th}}$ component signal, $P_i(t|z)$ represents the time marginal for the distribution of the component signal when the latent variable takes the value $z$, and $P_i(f|z)$ is the corresponding frequency marginal. Equation 8 is equivalent to stating that the the distribution underlying the spectrum of the mixed signal is formed by a linear combination of the marginal frequency distributions for the component signals.

Given a new mixed signal and the marginal frequency distributions for all its component signals (i.e. assuming that all $P_i(f|z)$ terms are known, having been learned from some training corpus for component $S_i$), the parameters of $P_{mixed}(t,f)$ that remain unknown are the terms $P(S_i)$, $P_i(z)$ and $P_i(t|z)$. Alternately stated, the distribution underlying the mixed signal is a linear combination of the known marginal frequency distributions for all component signals; however the proportions to which they must be mixed to obtain the final distribution are unknown. The unknown terms are easily determined using an EM algorithm that involves iterative updates of the following equations:

$$\begin{aligned} P(S_i|f,t) &= \frac{P(S_i)\sum_z P_i(z)P_i(f|z)P_i(t|z)}{P(S_j)\sum_z P_j(z)P_j(f|z)P_j(t|z)} \\ P(z|S_i,f,t) &= \frac{P_i(z)P_i(f|z)P_i(t|z)}{\sum_{z'} P_i(z')P_i(f|z')P_i(t|z')} \\ P(S_i) &= \frac{\sum_t\sum_f P(S_i|t,f)S(t,f)}{\sum_j\sum_t\sum_f P(S_j|t,f)S(t,f)} \\ P_i(z) &= \frac{\sum_t\sum_f P(z|S_i,t,f)S(t,f)}{\sum_j\sum_t\sum_f P(z|S_j,t,f)S(t,f)} \quad (9) \\ P_i(t|z) &= \frac{\sum_f P(z|S_i,t,f)S(t,f)}{\sum_{t'}\sum_f P(z|S_i,t',f)S(t',f)} \quad (10) \end{aligned}$$

We are now set to separate out the component signals from a mixture. Given the frequency marginals $P_i(f|z)$ for all component signals and the magnitude spectrum for the mixed signal, all unknown terms in Equation 8 are obtained through iterations of Equation 10. Once derived, the partial distribution that represents the contribution of the $i^{\text{th}}$ component to the spectrum of the mixed signal is given by $\sum_z P_i(z)P_i(f|z)P_i(t|z)$. Figure 4 shows a graphical representation of the statistical framework used for separation and the components of this framework that must be estimated.

We recall that the value of the spectrum $S(t,f)$ of the mixed signal at any time-frequency location $(t,f)$ is the outcome of several draws from the distribution of the mixed signal. The corresponding spectral component of the $i^{\text{th}}$ component signal is obtained by estimating the number of these draws that were obtained from the partial distribution for that component. Thus, the overall separated spectrum for the $i^{\text{th}}$ component is obtained by estimating individual components of this spectrum through the expressions

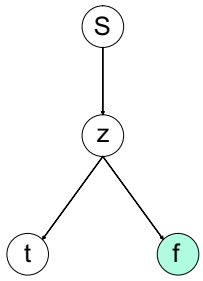


**Fig. 4**. Graphical representation of the generating process for a mixed signal. A first latent variable $S$ selects the speaker; a second level latent variable signal $z$ selects the marginal time and frequency distributions (that are specific to the speaker and latent variable). Only the marginal distributions for the frequency variable $f$ (shown shaded) are known; all other parameters must be estimated.

$$\hat{S}_i(t,f) = S(t,f)\frac{P(S_i)\sum_z P_i(z)P_i(f|z)P_i(t|z)}{\sum_j P(S_j)\sum_z P_j(z)P_j(f|z)P_j(t|z)} \quad (11)$$

The above equation only reconstructs the magnitude spectrum for the $i^{\text{th}}$ component. To reconstruct the time domain signal, the phase of the original mixed signal is imposed on the spectrogram and the resulting complex spectrogram is inverted through an inverse short-time Fourier transform.

## 5. SEPARATING A SINGING VOICE FROM BACKGROUND MUSIC

Vocals can be separated from background music using a variant of the procedure described in Section 4. One drawback with the procedure from Section 4 is that the frequency marginals must be learnt separately from unmixed training data for each component source. Such training data are often not available for a song, since the background music for most songs is unique. Instead, we use an *adaptive* version of the algorithm for separating the vocals.

Most songs contain music-only sections sans voices. To effectively separate out the singing from the music, it is important to identify the regions of the song where the voice(s) are actually active and to selectively apply the separation algorithm to only these regions. In this paper we assume that these regions are marked *a priori*, either by some automated technique or by hand. We do not explicitly address the problem of marking these regions automatically.

In a first step we learn frequency marginals $P_{music}(f|z)$ for the music from a typical segment of music-only recording using Equations 6. Although the equations also give us time marginals $P_{music}(t|z)$ and the latent variable probabilities $P_{music}(t|z)$, we do not utilize those since they are specific to the training segments. It is only the frequency marginals that are expected to generalize and effectively model the music in the segments that have both voice and music.

Since it is rare that songs will contain pure voice-only regions with no background music (and even when they do, such segments are rarely of sufficient length to learn the marginal frequency distributions for the voice from), it is assumed that the marginal frequency

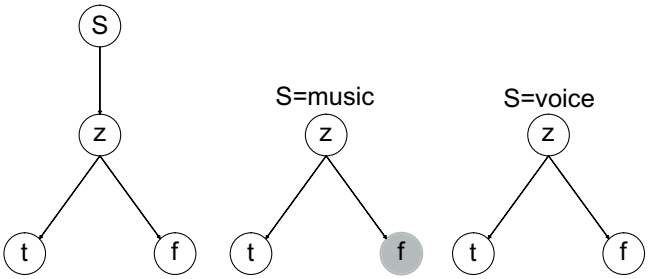


**Fig. 5**. First panel: Graphical representation of the generating process for a song, used to separate background music from singing voices. Second panel: sub graph when the first-level latent variable selected represents the music. The marginal distributions for the frequency variable $f$ are known (shown shaded). Third panel: sub graph when the first-level latent variable selected is the voice. None of the parameters are known. All unknown (unshaded) parameters must be estimated.

distributions for the voice in the song are *not* known *a priori* and must be learnt.

The probability distribution underlying the voice+music segments of the song is given by

$$P_{song}(t,f) = P(music)\sum_z P_{music}(z)P_{music}(f|z)P_{music}(t|z) + P(voice)\sum_z P_{voice}(z)P_{voice}(f|z)P_{voice}(t|z) \quad (12)$$

In the above equation, $P(music)$ (the fraction of all spectral magnitudes that are attributable to music), $P(voice)$, $P_{music}(z)$, $P_{music}(t|z)$, $P_{voice}(t|z)$, $P_{voice}(f|z)$ and $P_{voice}(z)$ are all unknown; only $P_{music}(f|z)$, the marginal frequency distributions for the music are known. We estimate all unknown components using Equation 10. Figure 5 shows a graphical representation of the statistical framework used for separation in this case and the components of this framework that must be estimated.

Once all components of the distribution are known, the spectrograms for the voice-only and music-only components of the mixed recordings are obtained using Equation 11. Time-domain signals are finally obtained by imposing the phase of the mixed song on the separated magnitude spectrograms and inverting the resultant complex spectrograms through an inverse short-time Fourier transform.

## 6. PERSONALIZATION OF SONGS

As mentioned earlier, music is a very acquired taste. The sound of altos, tenors and sopranos singing classical western Operas at unnatural pitches, while producing an extra formant as learned from many years of training, may sound extremely pleasing to a classically minded person from the western world, and yet sound grating to an untrained ear from a different part of the world.

A similar phenomenon may also be observed in popular Indian music. Ever since the advent of the immensely talented singer Lata Mangeshkar on the music scene in India in the 1950s, it has been fashionable for female playback singers in Indian movies to sing at an unnaturally high pitch. In particular, the authors have observed that the pitch of female playback singers in Indian movies has shown an increasing trend over the decades. Shamshad Begum maintains a pitch of around 200Hz in the song "Mere Piya Gaye Rangoon" sung in 1949. In 1956, Geeta Dutt also maintains a pitch just above 200Hz in the song "Jaata kahan hai deewane". The upward trend in female pitch begins with the arrival of Lata Mangeshkar who hit a pitch of about 380Hz in the song "Tumko piya dil diya" sung in 1963. The song "Dayya dayya" sung in 2003 hits a peak pitch of over 760Hz in parts. Figure 6 shows the pitch track for a segment of the vocals in "dayya dayya" demonstrating the high pitch employed by the singer.

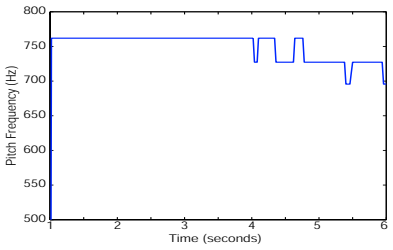


**Fig. 6**. Pitch track for a segment of the song "dayya dayya".

These high pitches are not always pleasant to everyone, although the underlying song itself may be very melodious. We note that the average pitch range of a human adult female voice is between 150 and 250Hz, throughout the world. When songs are rendered in pitches outside this range, they sound good until the deviation from the average pitch becomes extreme. While these pitches are clearly appreciated by a majority of Indian listeners, to the unaccustomed ear they sound screechy. The high pitch of Indian female playback singers (in pop music) has, in fact, been commented upon, both in informal blogs and in popular literature. For instance, on Page 24 ofher Book "Holy Cow", published by Bantam in 2002, Sarah McDonald cites an encounter with the voice of a female playback singer thus: *"..and the driver and his friend sing along to a tape featuring the high-pitched wail of a woman obviously being tortured."* Similar statements abound in blogged travelogues of visitors to India as well.

As a remedy, we have created a framework where, given a song, a person can (for the effort of manually tagging the locations of voice regions of the song) create modified *personalized* versions of the song that are better suited to their listening tastes. Given a track of voice-only recording, the vocals and the music are separated using the procedure from Section 5. It then becomes possible to modify the pitch or the perceived gender of the voice through pitch and frequency modification algorithms such as PSOLA [8]. Harmonics may be introduced by blending multiple modified versions of the voice and remixing them with the music. Similarly, it now becomes possible to add in new music to the ensemble, or to modify the existing music in the song through signal processing techniques without affecting the quality of the voice.

## 7. EXPERIMENTAL EVALUATION

In this section we report experiments evaluating the separation algorithm proposed in Section 5, as well as the personalization framework described in Section 6. In the first experiment, we demonstrate that the algorithm is able to separate the voices from the background from a monoaural recording of a popular hindi song.

In a second experiment we show that the separated signals produced by our algorithm can be personalized through pitch modification without recognizable artifacts (except those attributable to the pitch modification algorithm itself).

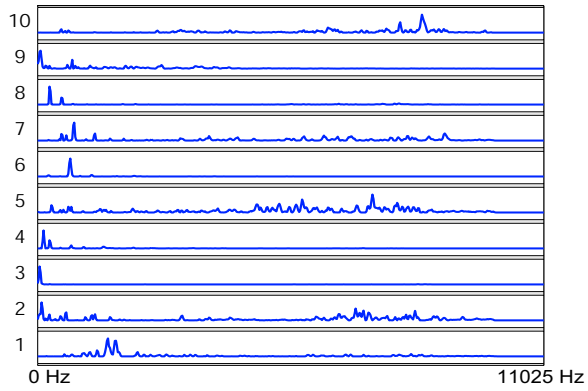


**Fig. 7**. Marginal frequency distributions learnt for the music in "dayya dayya", for 10 values of the latent variable $z$.

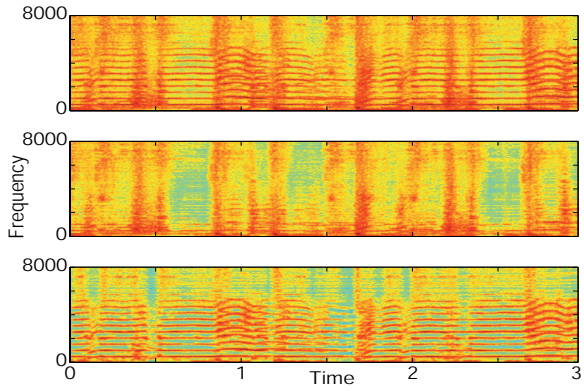


**Fig. 8**. Top panel (a): Spectrogram for the mixed voice and music in a segment of the song "dayya dayya". Middle panel (b): Separated spectrogram obtained for the music in the same segment of the song. Bottom panel (c): Separated spectrogram obtained for the voice in the same segment of the song.

### 7.1. Separating Vocals from Music

For this experiment we selected the song "Dayya dayya" from the 2003 Hindi movie "Dil ka Rishta", sung by Alka Yagnik, to largely percussive background music. The song was ripped from a legally obtained CD from a retail shop and was sampled at 44100 Hz. The entire signal was converted to a spectrogram as described in Section II.

We hand-segmented the song to mark the boundaries of the regions that included voice. The music-only segments of the recording were used to compute the distribution underlying the music spectra. The distribution was modelled through a mixture of 100 products of marginals (*i.e.* $z$ could take 100 values), resulting in 100 sets of marginal frequency distributions $P(f|z)$ characterizing the music. Alternately viewed, a set of 100 basis vectors were learnt to represent the music. Figure 7 shows some of the basis vectors learnt for the music.

The algorithm described in Section 5 was then used to separate out the singing voice from voice regions of the song. A set of 100 basis vectors were learnt for the voice from the song itself, in addition to the 100 vectors learnt separately for the music. These were then used to separate the music and the voice.

Figure 8a shows the spectrogram of the mixed song and music. Figures 8b and 8c show the spectrograms of the separated music and voice. We note that while the separated music shows minimal residue from the voice, the spectrogram of the voice primarily shows the harmonic voice activity of singing with minimal residue from the music.

The (voice portions of) the original song, and the separated music and voice can be heard at:
http://www.cs.cmu.edu/ bhiksha/audio/songseparation

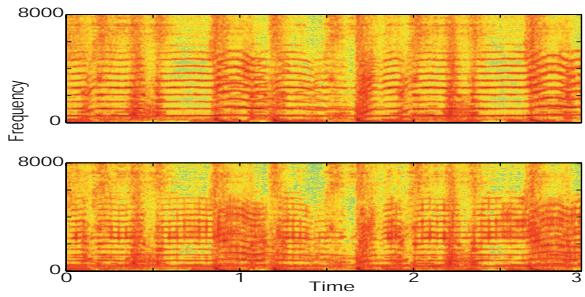


**Fig. 9**. Upper panel (a): Spectrogram of a segment of the song "dayya dayya" including both voice and music. Middle panel (b): Spectrogram of the same segment when the pitch of the voice has been lowered by 4 semitones. The harmonic frequencies are observed to occur much closer together. The vertical artifacts in the lower panel are a result of deficiencies in the overlap-add mechanism used in our version of time-domain PSOLA.

### 7.2. Personalization by Pitch Modification

By separating the vocals out from the music, we are able to reduce the pitch of the vocals to more acceptable levels, remix the music and the song to produce pleasanter sound. In particular, in order to demonstrate the effectiveness of our separation algorithm we used time-domain PSOLA [8] for pitch modification. Time-domain PSOLA requires two operations that are critically dependent on the fidelity and cleanliness of the time-domain waveform: pitch detection and pitch-period compression. A filter-bank based pitch detection algorithm was used to detect pitch [9]. For this experiment we reduced the pitch of the voice uniformly by four semi-tones, or roughly by 20%, and remixed the song with the music. Ideally, one would reduce the pitch of the music by 4 semi-tones as well; however since pitch modification of complex music is significantly more difficult than that for voice, this was not attempted. The result is therefore slightly different than what might have been intended (musically speaking) by the musical directors of the song.

Figure 9a shows the spectrogram of the original signal including both music and voice. Figure 9b shows the spectrogram of the processed signal that we eventually obtained. The original and pitch-reduced (and remixed) signals can be heard at:
http://www.cs.cmu.edu/ bhiksha/audio/songseparation

It is clear from the example that our processing is successfully able to produce a pitch modified version of the song, without significant artifacts. It is the opinion of the authors that the pitch modified version of the song is also more pleasant to hear than the original song itself.

## 8. CONCLUSIONS

We have presented an algorithm for separating forground vocals from background music in songs. The proposed algorithm is observed to be very effective at separating the two. Although the current algorithm requires hand-marking of the boundaries of voiced segments in the songs, we do not expect this to be a problem – methods such as those proposed by Li and Wang [1] can be utilized to detect voice boundaries automatically.

The proposed algorithm has been presented within a framework of personalization of songs. We believe that such personalization is not only eminently possible, it is also a very attractive commercial proposition. We envision a system that will allow a user to modify vocals by changing the pitch, gender, adding choruses, harmonization etc., modifying the music by changing its timbre etc., and remixing the vocals and the music to produce versions of songs that are to their liking. While most of the algorithms required for such personalization exist, some technical challenges still remain. These and other related topics will be the focus of future research.

## 9. REFERENCES


[1] Li Y. and Wang D. L. (2006). Separation of singing voice from music accompaniment for monaural recordings, *IEEE Transactions on Audio, Speech, and Language Processing*, in press.

[2] Wang, A. L.-C. (1994). Instantaneous and frequency-warped signal processing techniques for auditory source separation. *Ph.D dissertation, Stanford University, Department of Electrical Engineering*.

[3] Meron, Y. and Hirose, K. (1998). Separation of Singing and Piano Sounds. *Proc. 5th International Conference on Spoken Language Processing (ICSLP98)*.

[4] ROWEIS01) S. T. Roweis. (2001). One Microphone Source Separation, *Advances in Neural Information Processing Systems*, 13:793–799, 2001

[5] Reddy, A.M. and Raj, B. (2006). Soft Mask Methods for Single-Channel Speaker Separation. *IEEE Transactions on Audio, Speech and Language Processing*. To Appear.

[6] Smaragdis, P. and Raj, B. (2006). Shift-Invariant Probabilistic Latent Component Analysis. *Submitted to the Journal of Machine Learning Research*.

[7] Raj, B. and Smaragdis, P. (2005). Latent Variable Decomposition of Spectrograms for Single Channel Speaker Separation, *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 17-20, October 2005.

[8] Moulines, E. and Charpentier, F. (1990). Pitch-Synchronous Waveform Processing Techniques for Text-to-Speech Synthesis using Diphones. *Speech Communication*, Vol. 9, No. 5, pp: 453-467.

[9] Seltzer, M. (2000). Automatic Detection of Corrupt Spectrographic Features for Robust Speech Recognition. *Master's Thesis, Carnegie Mellon University, Dept. of Electrical and Computer Engg.*, Chapter 4.