SEPARATION OF VOCALS FROM POLYPHONIC AUDIO RECORDINGS

Shankar Vembu
German Research Centre for AI
Erwin-Schroedinger-Str. 57
67663 Kaiserslautern, Germany
shankar.vembu@dfki.de

Stephan Baumann
German Research Centre for AI
Erwin-Schroedinger-Str. 57
67663 Kaiserslautern, Germany
stephan.baumann@dfki.de

ABSTRACT

Source separation techniques like independent component analysis and the more recent non-negative matrix factorization are gaining widespread use for the monaural separation of individual tracks present in a music sample. The underlying principle behind these approaches characterises only stationary signals and fails to separate non-stationary sources like speech or vocals. In this paper, we make an attempt to solve this problem and propose solutions to the extraction of vocal tracks from polyphonic audio recordings. We also present techniques to identify vocal sections in a music sample and design a classifier to perform a vocal–nonvocal segmentation task. Finally, we describe an application wherein we try to extract the melody from the separated vocal track using existing monophonic transcription techniques. The experimental work leads us to the conclusion that the quality of vocal source separation, albeit satisfactory, is not sufficient enough for further F0 analysis to extract the melody line from the vocal track. We identify areas that need further investigation to improve the quality of vocal source separation.

Keywords: Blind source separation, independent component analysis, non-negative matrix factorization, vocal–nonvocal discrimination, melody extraction.

1 INTRODUCTION

Analysing an auditory scene and identifying the various sounds present in it has, for a long time, been the primary focus of the research field called computational auditory scene analysis (CASA). Most of the approaches in this field draw inspiration from the works of Bregman (1990) who describes a set of psychoacoustic grouping cues that could be used in the analysis and segregation of individual sources present in a mixture of sounds using signal processing techniques. A significant amount of work in the field of CASA can be found in the doctoral theses of Mellinger (1991), Cooke (1991), Brown (1992), Ellis (1996) and Martin (1999).

The works of Attneave (1954) and Barlow (1959) revealed the fact that redundancy reduction is an inherent mechanism taking place in the sensory organs and that the human brain analyses an input scene (for example, visual) by exploiting the statistical regularities present in it. In recent times, a lot of effort is being expended in sound source separation using statistical techniques like principal component analysis (PCA) and independent component analysis (ICA) (Comon, 1989) for redundancy reduction mostly inspired by the works of Casey and Westner (2000) and Smaragdis (2001). It is interesting to note that there are two parallel strands of research sharing the same goals, one of them attempting to solve the source separation problem using classical signal processing techniques and pyschoacoustic studies while the other trying to achieve the same using statistical techniques. A formal analysis and comparison of the results obtained from these research fields is yet to be done.

In this paper, we focus on a particular problem that arises when employing statistical techniques like ICA for monaural source separation, which is the inability of these models to separate non-stationary sources. Attempts to solve this problem have been made by Casey and Westner (2000) and Smaragdis (2004b). Smaragdis proposes an extension to non-negative matrix factorization (NMF) (Lee and Seung, 2001) called non-negative matrix deconvolution in which an individual non-stationary source is characterised by a set of time-dependent spectral bases. This is unlike the basic model that characterises each source using a single spectral basis and thus fails to separate non-stationary sources that necessarily should be represented using a time-varying spectral basis. Casey assumes that non-stationary sources remain stationary for small intervals of time and proceeds with the usual analysis in its basic setting. He then proposes a clustering mechanism to finally group the resulting components (spectral bases) over time. In this paper, we identify the ramifications of using statistical techniques like ICA or NMF in their basic setting while trying to separate the vocal track from polyphonic music samples with a single voice. We propose specific solutions to handle the separation of vocals from a given sound mixture.

An important application of vocal source separation is the extraction of melody. Based on the assumption that vocals carry the main melody, one could apply monophonic transcription techniques to extract the melody from the separated vocal track. This is an enticing application as monophonic transcription is much more simpler when compared to polyphonic music transcription and therefore the harder problem of extracting melody directly using polyphonic transcription techniques is bypassed. The results of this work could also be used in the design of query-by-humming systems. Existing systems try to build a database of melodies by collecting annotations in the MIDI format or by manually transcribing polyphonic music samples. This database is then used for making comparisons with the input query. Vocal source separation from polyphonic recordings and hence the extracted melody could therefore be used in the automatic creation of melody databases.

This paper is organised as follows: In Section 2, we present a vocal–nonvocal discrimination module and outline the principles behind monaural source separation using statistical techniques. In Section 3, we identify problems encountered when trying to separate non-stationary sources like vocals and propose solutions to mitigate these problems. Section 4 deals with the experimental work. We point to directions for future work in Section 5. Section 6 concludes the paper.

2 MONAURAL SOURCE SEPARATION OF VOCALS

The different stages in the design of a source separation system for vocals are shown in Figure 1. The last module is our proposed solution to separate non-stationary sources and we defer its description until the next section. The rest of the stages are described below.

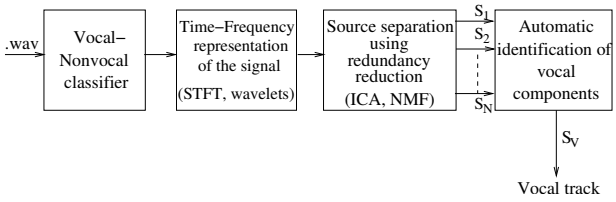

Figure 1: Building blocks of the vocal source separation system

2.1 Vocal–nonvocal Discrimination

Since we are interested only in the separation of vocals, it is essential to have a pre-processing stage that performs a vocal–nonvocal discrimination task to filter out sections that contain only the nonvocal, instrumental tracks. We identified three features as useful candidates in the design of a vocal–nonvocal classifier. Mel-frequency cepstral coefficients (MFCC) (Davis and Mermelstein, 1980) were used in the design of a classifier to identify vocals and instrumental sections in a music sample (Maddage et al., 2003). Perceptual linear predictive coefficients (PLP) introduced by Hermansky (Hermansky, 1990) were used by Berenzweig et al. (2002) to train a neural network classifier for distinguishing vocals from instrumental music. Log frequency power coefficients (LFPC) were shown to be a useful feature in discriminating vocals with instruments from pure instruments (Nwe and Wang, 2004). In this paper, we report experiments using different combinations of these features and train two classifier models; namely neural networks and support vector machines. Based on the experimental results, we arrive at the conclusion that a combination of all the aforementioned features produces the best classifier performance. The experimental results of this module also paved way to solutions for the separation of non-stationary vocal sources that will become clear in Section 3.

2.2 Monaural Source Separation: Basic Idea

The underlying principle behind these approaches (Casey and Westner, 2000; Smaragdis, 2001) is to apply redundancy reduction techniques on the time–frequency representation of signals leading to the separation of the individual sources present in the input mixture of sounds. We briefly outline the important steps.

Step 1: The first step is to project the input signal $s(t)$ into the time–frequency plane using an invertible transform ψ like short time Fourier transform or wavelets, giving rise to an $n \times k$ matrix F, where n denotes the number of frequency channels and k is the number of time frames:

$$\psi : s(t) \rightarrow F \, . \tag{1}$$

Step 2: The next step is to whiten the matrix F using PCA; this results in a matrix with uncorrelated rows. The dimension of this matrix is reduced by retaining only those rows that carry maximal information in terms of their variance contribution. The reduced dimension r is determined by using a threshold $\phi \in [0, 1]$ and the following inequality:

$$\frac{\sum_{i=1}^{r} e_i}{\sum_{i=1}^{n} e_i} \geq \phi \, , \tag{2}$$

where e_i are the eigenvalues of the covariance matrix of F. The resulting matrix F_w is of dimension $r \times k$ where $r < n$. This step also gives rise to a whitening matrix W of dimension $r \times n$ and its pseudo-inverse called the dewhitening matrix W^+.

Step 3: The next step is to exploit the higher-order statistics of the matrix F_w using ICA resulting in an $r \times k$ matrix G with independent rows. We call this matrix as the matrix with time-varying gain of the spectrum of the individual sources in the row vectors. The ICA operation gives rise to a transformation matrix I of dimension $r \times r$. Multiplying the dewhitening matrix W^+ with I gives rise to a mixing matrix B of dimension $n \times r$:

$$B = W^+ \cdot I \, . \tag{3}$$

We call this matrix the matrix with the spectral bases of the individual sources in column vectors.

The stages described so far have resulted in two important matrices that will be used for the resynthesis of the individual sources in the next step. They are the matrix B with r spectral bases and the matrix G that carry the time-varying gain of the individual spectral bases.

Step 4: This step involves the reconstruction of the sources present in the original input mixture by taking the outer product of the individual column and row vectors of B and G respectively. Inverse transform of the resulting matrices gives the individual sources in the time-domain. That is,

$$\begin{aligned} &F^i = \mathbf{b}^i * \mathbf{g}^i \text{ and} \\ &\psi^{-1} : F^i \rightarrow s^i(t) \,, \end{aligned} \tag{4}$$

where the superscript i is used to index the individual sources.

2.3 ICA and NMF

Non-negative matrix factorization (NMF) introduced by Lee and Seung (2001) operates on simple non-negativity constraints to arrive at reduced-rank factors of a given matrix, and has recently been used for monaural source separation of acoustic inputs (Smaragdis, 2004a,b). There is psychological and physiological evidence for parts-based representations in the brain, and NMF is one algorithm that tries to emulate this process (Lee and Seung, 1999). This makes it a suitable candidate to discover the individual objects (parts) present in an acoustic input. NMF could be seen as a replacement to the ICA step described in the previous section. An important decision lies in choosing the right value for the parameter r that determines the rank of the factorization. This could be done using the same procedure that was adopted before to perform dimensionality reduction of the uncorrelated time–frequency matrix after PCA as shown in Equation 2.

3 SEPARATION OF NON-STATIONARY SIGNALS

A major shortcoming of the presented approaches to monaural source separation is that each source is characterised by a single stationary spectral basis (column vector of B) and only its gain varies with time (row vector of G). This implies that the current setting will not be able to separate non-stationary signals that should necessarily be described by more than a single spectral basis as shown below:

$$\begin{aligned} &B^i = [\mathbf{b}^i_1, \mathbf{b}^i_2, ..., \mathbf{b}^i_j], B^i \subset B \,; \\ &G^i = [\mathbf{g}^i_1, \mathbf{g}^i_2, ..., \mathbf{g}^i_j], G^i \subset G \,; \\ &F^i = B^i * G^i \text{ and} \\ &\psi^{-1} : F^i \rightarrow s^i(t) \,, \end{aligned} \tag{5}$$

where $j < r$ indicates the number of components needed to characterise an individual non-stationary source i, $\mathbf{b}^i_{1..j}$ are j column vectors and $\mathbf{g}^i_{1..j}$ are j row vectors. Comparing the above Equation with Equation 4, we note that for non-stationary sources, we need to consider a set of column vectors of B and a set of row vectors of G for the resynthesis. The ramification of this shortcoming is that the vocal source tends to get distributed among a set of bases where each one of them contributes to the spectral composition of the entire vocal source. It is not possible to arrive at a single component describing the vocal track in its entirety unlike other stationary sources.

To illustrate the problem of separating non-stationary sources, we consider two mixtures. The first one (sample ID: a) is the drum track sequence used in Smaragdis (2001) that has three sources; namely, bass drum, snare drum and hi-hat, whose spectrum remains fairly constant over time. The second example (sample ID: b) is a mixture consisting of a vocal track with a guitar accompaniment. We sample the signal at 22.05 kHz, compute their short time Fourier transform (STFT) (512 frequency bins, Hamming window of length 128 samples, hop size of 64 samples) and perform PCA and dimensionality reduction on the resulting matrix. The results are shown in Table 1 in which the third column shows the reduced dimension of the matrix i.e., the number of retained components and the last column shows the amount of information in terms of the variance contributed by the corresponding eigenvalues. Since there are only three sources in the drum track mixture, we retained only three components after PCA and they corresponded to 94% of the information content of the STFT matrix which might be sufficient to arrive at a satisfactory separation. But if we do the same for the second example that consists of vocals and retain only two components (vocal and guitar), we observe that they correspond to only 72% of the information content. Throwing away a lot of information in this way would definitely hinder the results of separation. On the other hand, we observe that we need a total of seven components (instead of only two) to retain 95% of the information as shown in Table 1. This is due to the non-stationary nature of the vocals that resulted in the vocal spectra getting distributed among a multitude of components. When this is the case, there has to be a mechanism by which it should be possible to identify and group these distributed components to form the final individual vocal source. This is a non-trivial task and we make an attempt to solve this problem by proposing a couple of solutions:

Table 1: Results of dimensionality reduction using PCA on the STFT matrix of sound mixtures with stationary and non-stationary sources

Sample ID	**Size (in s)**	**Number of components (r)**	**Variance(%)**
a	3	3	93.89
b	3	2	72.25
b	3	7	94.9

Solution 1: One possible way to identify the distributed vocal components is to reuse the vocal–nonvocal classifier described in an earlier section. We note that the classifier's input is a feature vector consisting of a combination of MFCC, LFPC and PLP coefficients computed from the spectrum. We also note that one of the outputs from the source separation stage is the

matrix B that contains the spectra of the sources. It is therefore trivial to compute the same feature vector from the individual column vectors of the matrix B as shown below:

$$\varphi : \mathbf{b} \rightarrow \mathbf{b}_c, \mathbf{b} \in \Re^n, \mathbf{b}_c \in \Re^p \, , \qquad (6)$$

where φ is a mapping from the spectral basis $\mathbf{b}$ of dimension n (spectral dimension) to the feature vector $\mathbf{b}_c$ of dimension p (number of coefficients in MFCC, PLP and LFPC). The individual feature vectors $\mathbf{b}_c$ can be presented as inputs to the classifier to identify as being a vocal or a nonvocal component. It is now a simple task to combine all the components that were classified as vocals resulting in a grouping of all the distributed vocal components. This would eventually help us to arrive at matrices B^v and G^v (see Equation 5) that could be used to compute the individual vocal source v in the time domain.

Solution 2: Another solution is to rely on unsupervised learning algorithms to cluster the spectral bases in the matrix B into two groups (vocal and nonvocal). Instead of using the spectra directly as feature vectors to the clustering algorithm, we could once again use the mapping φ in Equation 6 to compute features with MFCC, LFPC and PLP coefficients. Since these coefficients are able to distinguish well between vocals and nonvocals, they provide a better parameterization of the spectrum. The resulting feature vector could also be augmented using additional information available in the matrix G that consists of the time-varying gain of the spectra, to provide more discriminatory power. Virtanen (2003) mentions a similar approach to group multiple components per source using the independence of the time-varying gain.

We validate the presented solutions in the experimental section of this paper.

4 EXPERIMENTS AND RESULTS

4.1 Vocal–nonvocal Discrimination

A random collection of 40 minutes of popular music distributed uniformly between pure instrumental and vocals (with accompaniment) was used for the experiments. The entire database comprised of 240 audio files of around 5 seconds each with vocals and 80 files of around 15 seconds each with pure instrumentals. The samples were carefully hand-picked to be representative of both male as well as female playback singers from Eastern and Western music. All the audio files were 16-bit mono and sampled at the rate of 22.05 kHz. The first step was to compute the short time Fourier transform (STFT) of the signal using a window function. The window size was set to $\approx$ 23 ms and the amount of overlap was half of the window size. Candidate features like PLP, MFCC and LFPC were calculated from the STFT of the signal on a frame-by-frame basis resulting in a matrix representation of the signal with feature vectors in the columns. Finally, an average of 15 analysis frames was computed to reduce the computational load. Each resultant feature vector, thus, represented an analysis frame of length ≈ 184 ms.

4.1.1 Using Neural Networks

We trained a neural network using different combinations of the features namely MFCC, PLP and LFPC. The architectural details of the network are given in Table 2 and the network's performance for various combinations of features is given in Table 3.

Table 2: Neural network architectural details

No. of inputs	13 MFCC and/or 12 LFPC and/or 39 PLP
No. of outputs	2
No. of hidden layers	1
Training algorithm	Resilient backpropagation with early-stopping
Activation function	Sigmoidal
Evaluation	10-fold cross-validation

Table 3: Results of the vocal–nonvocal classifier using neural networks

	Frame-based classification efficiency (%)	
Feature	**Before smoothing**	**After smoothing**
PLP	69.25	82.02
MFCC	67.68	74.94
LFPC	68.12	73.09
PLP+MFCC	73.11	81.55
PLP+LFPC	75.25	82.35
MFCC+LFPC	71.69	78.65
PLP+LFPC+MFCC	77.24	84.87

It can be seen from the results that combinations of features give better performance when compared to individual features. The best performance of 77.24% efficiency resulted when all the features were used as inputs to the network. The last column in Table 3 is the result of a simple smoothing operation on the network output using an autoregressive low-pass filter.

4.1.2 Using Support Vector Machines

Experiments using neural networks led to the conclusion that the network's best performance resulted when using a combination of PLP, MFCC and LFPC features. Therefore, the same features were used to train an SVM with an RBF kernel. An appropriate combination of the kernel parameter σ and the penalty parameter C (Burges, 1998) should be made. The optimal values were found using cross-validation by repeating the experiments on various combinations of σ and C. We first performed a coarse grid search in the region $C = 2^{16}, 2^{14}, ..., 2^{-4}$ and $\sigma = 2^4, 2^2, ..., 2^{-10}$. The classifier's frame-based error rate was computed using 5-fold cross-validation. This

was followed by a fine grid search in regions of best performance to arrive at the final values for the parameters σ and C. The grid search is a computationally intensive operation, and therefore only 30% of the original database comprising of 40 minutes of audio recordings was used for the model selection. The optimal choice of C and σ was finally fixed at 2^8 and 2^2 respectively (see Table 4). The

Table 4: Results of the vocal–nonvocal classifier using SVM

Features	PLP + LFPC + MFCC
Kernel	RBF
Parameter: C	2^8
Parameter: σ	2^2
Frame-based classification efficiency (%)	93.47

classifier was trained again with a larger database comprising 30 minutes of audio recordings. We arrived at a generalization error of 6.53% that was computed using 5-fold cross-validation.

4.2 Vocal Source Separation

A major problem encountered while performing source separation experiments is evaluation of the quality of separation of the individual sources. It is difficult to come up with a good measure that describes the quality of source separation and that adheres well to the auditory perception due to information loss in the analysis process. In our case, the evaluation problem is mitigated to some extent with an application that tries to extract the melody from the separated monophonic vocal track. This is because we decided to perform experiments on excerpts from the ISMIR 2004 melody extraction contest[1] that had music samples with annotated vocals. This helped us to make comparisons of the extracted melody with the existing annotations that served as a ground-truth. But for the vocal source separation, we only present an analytical description of the experimental results.

Setup: The source separation was performed using ICA[2] and NMF in the MATLAB environment. Some of the experimental details are shown in Table 5. The

Table 5: Experimental details of vocal source separation

Sample ID	**Size (in s)**	**Number of components** (r)	**Variance (%)**
1	6.2	10	96.17
2	3.5	10	98.19
3	5.3	9	98.19
4	5	9	98.12
5	4.2	8	98.36
6	5.1	9	98.17

third column refers to the number of components that were retained after the dimensionality reduction using PCA and were further analysed using ICA. In case of NMF, this value determined the rank r of the matrix factorization. The value of r was chosen such that the PCA decomposition is overcomplete resulting in maximal retained information (Uhle et al., 2003). For the experiments, r was determined by

$$r = \min\left\{r_{\max}, \min\left\{r | \frac{\sum_{i=1}^{r} e_i}{\sum_{i=1}^{n} e_i} \geq \phi\right\}\right\}, \qquad (7)$$

where ϕ was set to 0.98 (i.e., 98% variance) and r_{max} was set to a value of 10 to make sure that we do not end up with too many components that might affect the identification and grouping of vocal components in the later stage. The next column in Table 5 shows the information content of these components in terms of the variance contribution of the corresponding eigenvalues. All the excerpts were sampled at a rate of 22.05 kHz and the short time Fourier transform (512 frequency bins, Hamming window of length 128 samples, hop size of 64 samples) was used as the time–frequency representation. We refrained from using wavelets even though this is a better time–frequency representation of the signal. This is because wavelet coefficients are not always non-negative and therefore cannot be used for source separation using NMF.

Validation of proposed solutions: Unfortunately, reusing the vocal–nonvocal classifier did not yield satisfactory results. The reason could be that this classifier was trained to operate on inputs that were computed from spectra of clean signals. By this we mean spectra that provided a holistic parameterisation of the inputs. But we note that the output of the source separation algorithm results in vocal source components with distributed and noisy spectra, and therefore the classifier was unable to operate on these inputs.

Clustering the parameterised spectra produced perceptually satisfactory results. In most of the test cases, we observed that the two output components from the clustering algorithm had, in one of them, the vocals as predominant source. We also noted that in a few examples, the quality of separation was superior when we augmented the feature vector $\mathbf{b}_c$ with the information present in the corresponding time-varying gain $\mathbf{g}$. In our experiments, we presented this information directly without extracting other information — for example, that uses the independence of the time-varying gains present in $\mathbf{G}$ as was done in Virtanen (2003). In most of the examples, we used only the parameterised spectral bases obtained from the matrix $\mathbf{B}$ and we plan to investigate on extracting any other relevant information from the matrix $\mathbf{G}$ in the future.

Observations and analysis: We analyse a particular example in detail. The music sample is a 3.5 s excerpt consisting of a vocal track accompanied with guitar. A total of 10 components were retained for analysis that carried 98.19% of the information. It is not surprising to arrive at 10 components despite the fact that only two tracks were present in the music sample. This is

[1] http://ismir2004.ismir.net/melody_contest/results.html
[2] http://www.cis.hut.fi/projects/ica/fastica/

due to the non-stationary nature of the vocal track, the spectra of which were distributed among a multitude of components. We performed a source separation using NMF ($r = 10$) and arrived at two matrices that had 10 spectral bases and their corresponding time-varying gains. We parameterised the spectra using MFCC, LFPC, PLP coefficients and the time-varying gain and clustered the resulting feature vectors. This resulted into two groups with 7 and 3 components. The individual spectrograms were determined using Equation 4. Inverse transform of these spectrograms gave rise to both the individual sources in the time domain. From Figure 2, we observe that the guitar track (spikes) is clearly separated from the vocal track. The same can be observed from Figure 4 in which the formants are clearly visible. In Figure 5, we observe the spectra of the guitar in the low frequency range that is clearly missing from Figure 4 after careful inspection. The vocal track was constructed from the

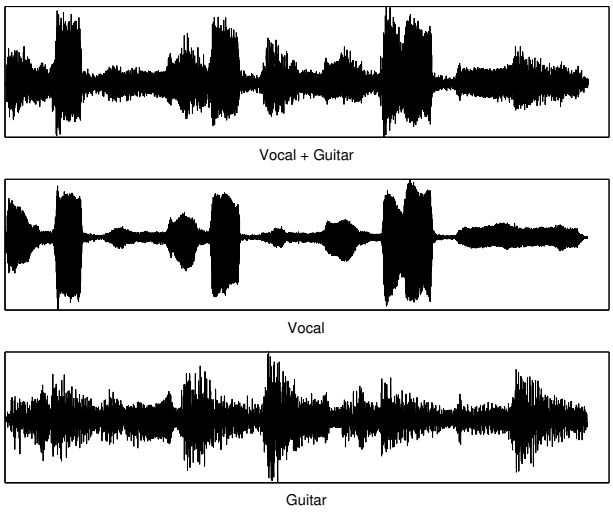

Figure 2: Vocal source separation results. The waveforms depict the energy of the signals in the time domain and scaled to fall in the range $[-1, +1]$

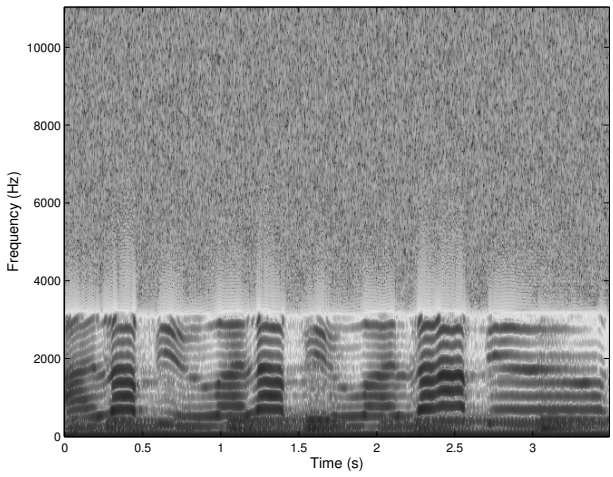

Figure 3: Spectrogram of the mixture shown in Figure 2 with vocal and guitar sources

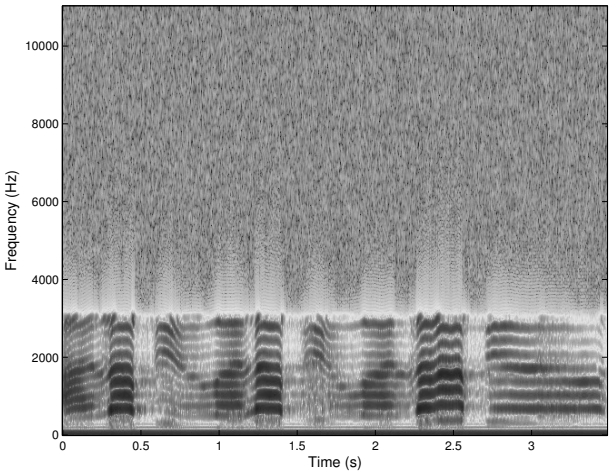

Figure 4: Spectrogram of the extracted vocal from the mixture shown in Figure 2

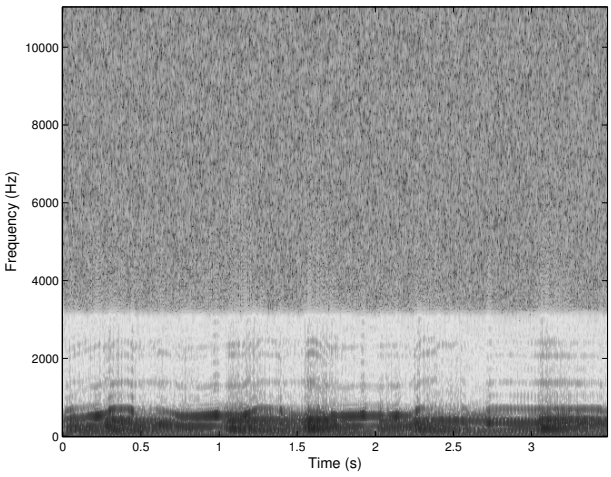

Figure 5: Spectrogram of the extracted guitar from the mixture shown in Figure 2

cluster that had 7 components and the guitar track from the one that had 3 components. An interesting observation is that the guitar track had remnants of vocals present in it (seen obscurely in Figure 5) but the instrument was clearly dominant. This is also the reason why the cluster with the guitar track had 3 components instead of just 1. Nevertheless, the separation was satisfactorily *clean* (the word *clean* is used subjectively as a result of listening tests) and one could easily distinguish the vocal track from the guitar track by listening to these tracks.

We present a few general observations from all the experiments. In most of the cases NMF was found to produce qualitatively better separation of the vocal source when compared to ICA. One possible reason for ICA resulting in poor results when compared to NMF could be the independence assumptions that might not be precisely true for the application at hand (Virtanen, 2004). On the other hand, NMF imposes only the less stringent non-negativity constraints.

In all the experiments, the input signal was band-pass filtered in the range of 100 Hz and 3000 Hz as most of

the energy in the singing voice lies in this range. Due to this, we were able to remove the high frequency instruments and only those instruments whose spectra were falling in the singing voice range remained unseparated. This helped a lot in further stages of the analysis. Most importantly, we observed that the number of components (r) retained for analysis after the PCA stage went down and this proved to be very useful when we were trying to group the components into vocals and nonvocals.

As already stated, the quality of the separation was assessed subjectively through listening tests. In most of the cases, both the components that resulted from the clustering stage had remnants of the nonvocal sections of the input mixture thereby not resulting in a perfect separation of vocal and nonvocal tracks. But only one of them had the vocal track predominant in it whose energy was sufficient enough for perceiving it as a vocal track.

4.3 Monophonic Transcription using MAMI

We supplement the source separation results by trying to extract the melody line from the separated monophonic vocal track. The vocal tracks obtained from the previous stage were transcribed using MAMI, which is a system designed for the monophonic transcription of singing voices (Mulder et al., 2003). The output from this stage is a sequence of notes with their F0 estimates and onset/offset values or a MIDI file with the melody. For comparing the results with the annotated melodies, we performed simple computation of melodic similarities using the MIDI toolbox (Eerola and Toiviainen, 2004). The toolbox provides functions to calculate the distance (or similarity) between two melodies using a user-defined representation (distribution of pitch classes and note durations, or melodic contour) and a distance measure (taxicab, Euclidean, cosine). The similarity can be scaled to range between 0 and 1, with 1 indicating perfect similarity. The results are shown in Table 6 where the numbers indicate the similarities on a scale of 0 to 1 using taxicab norm as the distance measure.

Table 6: Results of the melodic similarity computations

Sample ID	Similarity measure		
	Pitch distribution	Durational distribution	Contour
1	0.16	0.5727	0.325
2	0.2371	0.9048	0.6188
3	0.3781	0.3333	0.355
4	0.6105	0.7	0.4
5	0.3096	0.3333	0.5050
6	0.4927	0.6825	0.3563

Observations and analysis: The results are admittedly hazy. The melodic similarity comparisons based on the durational distribution of notes produced the best results with an average of 58.7% whereas the comparisons based on pitch distribution and contour produced poor results with an average of 36.5% and 42.7% respectively. This should not come as a surprise owing to the fact that there is information loss mostly the pitch information during the analysis process of the source separation stage. We performed the melody extraction experiments only to get a quantitative indication of the quality of the separated vocal track. There is still a long way to go before one is really able to extract the melody line using these approaches and whose results could be compared to the existing melody extraction techniques that uses complex F0 estimation procedures.

5 FUTURE WORK

Our experiments provide a nice starting point to investigate more complex approaches to grouping the various components that arise from the source separation stage into vocals and non-vocals. Our primary intention was to focus on techniques that allow us to identify and group the vocal track that appears distributed from the source separation algorithm. It might be difficult in the current setting of monaural separation to arrive at the vocal source in a single component. Trying to influence the separation algorithm (ICA or NMF) by the usage of prior information like vocal source models might not be of much help in this case, as anyhow we cannot obtain time-dependent spectra at the output. An interesting step forward would be to use techniques like non-stationary ICA (Everson and Roberts, 1999) wherein the mixing matrix evolves over time to give rise to time-dependent spectra that would suffice to characterise non-stationary signals. Another interesting direction would be to revisit our proposed solution to use a vocal–nonvocal classifier as a means to identify and group vocal component spectra. Reusing the model of the classifier described in this paper did not work, but one could possibly design a robust classifier that is an accurate model of instruments *only* and that treats any other inputs as *don't-care*. This classifier would then classify the distributed, noisy vocal spectra as negative inputs and classify positively only the spectra of the instruments. We could finally group all the negatively classified inputs to determine the vocal source.

6 CONCLUSIONS

Monaural source separation using statistical techniques for redundancy reduction is gaining widespread use in the research community. The drawbacks of these approaches in separating non-stationary signals from sound mixtures were identified and we proposed solutions to handle the non-trivial problem of separating vocal tracks from polyphonic music samples. Subjective evaluation of the experimental work indicated that the results are promising. We also presented an application wherein we made an attempt to extract the melody from the separated monophonic vocal track that also served as a quantitative indicator of the quality of the source separation. We also presented experiment work on discriminating vocal and nonvocal segments present in a music sample and arrived at encouraging results.

REFERENCES

F. Attneave. Some informational aspects of visual perception. *Psychological Review*, 61(3):183–193, 1954.

H. B. Barlow. Sensory mechanisms, the reduction of redundancy, and intelligence. In *The Mechanisation of Thought Processes*, pages 535–539. London: Her Majesty's Stationery Office, 1959.

A. Berenzweig, D. Ellis, and S. Lawrence. Using voice segments to improve artist classification of music. In *Proc. AES-22 International Conference on Virtual, Synthetic and Entertainment Audio*, Espoo, Finland, June 2002.

A. S. Bregman. *Auditory Scene Analysis: The Perceptual Organization of sound*. MIT Press, Cambridge, MA, 1990.

G. J. Brown. *Computational auditory scene analysis: A representational approach*. PhD thesis, Department of Computer Science, University of Sheffield, 1992.

C. J. C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167, June 1998.

M. Casey and A. Westner. Separation of mixed audio sources by independent subspace analysis. In *Proceedings of the International Computer Music Conference*, Berlin, August 2000.

P. Comon. Independent component analysis – a new concept? *Signal Processing*, 36:287–314, 1989.

M. P. Cooke. *Modeling auditory processing and organisation*. PhD thesis, Department of Computer Science, University of Sheffield, 1991.

S. B. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(4):357–366, 1980.

T. Eerola and P. Toiviainen. MIR in Matlab: The MIDI toolbox. In *Proc. 5th International Conference on Music Information Retrieval*, Barcelona, Spain, October 10–14 2004.

D. Ellis. *Prediction-driven computational auditory scene analysis*. PhD thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, 1996.

R. M. Everson and S. J. Roberts. Non-stationary independent components analysis. In *Proc. International Conference on Artificial Neural Networks*, pages 503–508, Edinburgh, 1999.

H. Hermansky. Perceptual linear predictive (PLP) analysis for speech. *Journal of Acoustic Society of America*, 87 (4):1738–1752, 1990.

D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788–791, 1999.

D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. In *Advances in Neural Information Processing Systems 13*, pages 556–562. MIT Press, Cambridge, MA, 2001.

N. C. Maddage, C. Xu, and Y. Wang. A svm-based classification approach to musical audio. In *Proc. 4th International Conference on Music Information Retrieval*, USA, October 26–30 2003.

K. D. Martin. *Sound-Source Recognition: A Theory and Computational Model*. PhD thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, 1999.

D. K. Mellinger. *Event formation and separation in musical sound*. PhD thesis, Department of Music, Stanford University, 1991.

T. D. Mulder, J. P. Martens, M. Lesaffre, M. Leman, B. D. Baets, and H. D. Meyer. An auditory model based transcriber of vocal queries. In *Proc. 4th International Conference on Music Information Retrieval*, USA, October 26–30 2003.

T. L. Nwe and Y. Wang. Automatic detection of vocal segments in popular songs. In *Proc. 5th International Conference on Music Information Retrieval*, Barcelona, Spain, October 10–14 2004.

P. Smaragdis. *Redundancy Reduction for Computational Audition, a Unifying Approach*. PhD thesis, Media Laboratory, Massachusetts Institute of Technology, May 2001.

P. Smaragdis. Discovering auditory objects through non-negativity constraints. In *Proc. Workshop on Statistical and Perceptual Audio Processing*, Jeju, Korea, October 2004a.

P. Smaragdis. Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs. In *Proc. 5th International Conference on Independent Component Analysis and Blind Signal Separation*, Granada, Spain, September 22–24 2004b.

C. Uhle, C. Dittmar, and T. Sporer. Extraction of drum tracks from polyphonic music using independent subspace analysis. In *Proc. 4th International Symposium on Independent Component Analysis and Blind Signal Separation*, Nara, Japan, April 2003.

T. Virtanen. Sound source separation using sparse coding with temporal continuity objective. In *Proc. International Computer Music Conference*, Singapore, 2003.

T. Virtanen. Separation of sound sources by convolutive sparse coding. In *Proc. Workshop on Statistical and Perceptual Audio Processing*, Jeju, Korea, October 2004.