KARAOKE SYSTEM WITH SPATIAL ACOUSTICS ESTIMATION FOR VOCAL OR INSTRUMENTAL REMOVAL

Pei Xiang and Shlomo Dubnov
Center for Research in Computing and the Arts (CRCA)
California Institute for Telecommunications and Information Technology Cal-(IT)^2
University of California, San Diego

ABSTRACT

This paper presents a new approach and prototype system for karaoke vocal or instrument removal. The algorithm estimates the singer's location and spatial transfer function from a stereo recording, then mutes components in the recording according to the estimated orientation. At the current stage, estimation is obtained by analyzing solo segments in the recording. This algorithm turns to be robust against reverb and spatial acoustics. To compensate low frequency loss that usually appears after vocal removal, the original recording is low-pass filtered around 120Hz and mixed back with the processed mono sound to generate a stereo karaoke track.

1. INTRODUCTION

In karaoke, a popular entertainment, the quality of the accompanying sound track directly affects the aesthetic experience of the "user singer". Reproduction of an accompaniment requires lots of time and labor, yet the music is probably not close enough to the original mix. Thus, removing the vocal or lead instrument track from the original recording remains an attractive topic.

Currently, the main method to achieve this goal, both in hardware and software, is to generate a mono sound track by subtracting the left and right channel of the original recording, and hope that the vocal component will disappear. This method is patented in 1996 [1]. Many karaoke machines and software like the AnalogX Vocal Remover, WinOKE, and the YoGen Vocal Remover work in this way. This job can also be done in any sound editor. The results are usually not satisfying, because it only works if the vocal component is identical at all times in both channels of the stereo recording, i.e. a mono source panned right at the center. This assumption doesn't apply to all situations especially live concert recordings where the full stereo image is picked up only by a pair of microphones. The singer isn't necessarily standing in the center, and room acoustics can make the vocal non-identical in left and right channel, which means having different amplitudes and delays in different frequency bands. Even more, artificial reverb could also have been added to the vocal component in the recording. All these conditions will cause the subtraction method to fail. Another problem is that low frequency instruments are usually mixed in the center position of a stereo mix, and this process could completely eliminate the bass, which is also undesirable. There are also other methods like applying EQs to the original recording attenuating the vocal frequency bands of the spectrum. This cannot remove the vocal completely and the spectral composition of the original material is altered a lot.

In our approach, we assume that the vocal component is stabilized in a spatial location and recorded with a pair of microphones. As soon as we can infer the transfer function associated with its location based on the stereo signal, in theory we will be able to manipulate components in that location according to different needs. The following sections will describe our model in detail and explain some experimental results with existing stereo recordings. In this paper, "vocal" could mean the lead singer or the lead instrument that is to be removed in a stereo recording.

2. METHODOLOGY

2.1. System architecture and initial assumptions

The organization of the system is shown in Figure 1. First, an algorithm tries to identify segments in the recording where the vocal is playing solo, or at least contributes most of the energy. These segments are analyzed for estimating the transfer function from the vocal's location to each of the microphones. After that, sound in the estimated location can be canceled or suppressed by projecting the stereo recording on appropriate directions in different frequency bands, resulting in a mono, vocal-suppressed sound track. Some additional spectral corrections can be done to this to further remove residual high frequency components of the vocal, such as various consonants. Finally, the bass part in the original recording, which is not overlapped with the vocal in frequency will be added back to the mono file.

We assume the recording situation is a live stereo production session where multiple instruments and the vocalist are stabilized in separate locations. Reverb and stationary room noise are assumed as well. This scene is more intuitively illustrated in Figure 2 where solid lines represent direct sound and dashed lines represent some of the early reflections. With reasonable reverb time, our model is robust to additive noise, room reverberation and other

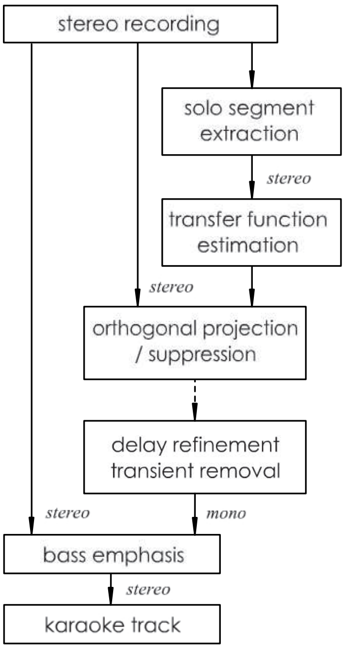
stereo recording
solo segment extraction
stereo
transfer function estimation
stereo
orthogonal projection / suppression
delay refinement transient removal
stereo
mono
bass emphasis
stereo
karaoke track

Figure 1. System diagram of our vocal removal design

factors that will be considered in the discussion section.

2.2. Source separation model with transfer function

A typical model for blind source separation of an N-channel sensor signal $\mathbf{x}(t)$ arise from M unknow scalar source signals $\mathbf{s}(t)$, in the case of instantaneous mixture, is described by

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t) + \mathbf{v}(t) \tag{1}$$

where

$$\mathbf{x}(t) = \begin{bmatrix} x_1(t) \\ \vdots \\ x_N(t) \end{bmatrix}, \mathbf{s}(t) = \begin{bmatrix} s_1(t) \\ \vdots \\ s_M(t) \end{bmatrix} \tag{2}$$

and $\mathbf{A}$ is a $N \times M$ linear mixing matrix and $\mathbf{v}(t)$ is a zero-mean, white additive noise. In convolutive environment such as the recording session situation, signals at different locations will have different transfer functions:

$$x_n(t) = \sum_{m=1}^{M} \int a_{nm}(\tau) s_m(t-\tau) d\tau + v_n(t)$$
$$n = 1, \ldots, N \tag{3}$$

where $a_{mn}(\tau)$ is the impulse response of the transfer function from the mth source signal to the nth microphone. Short Time Fourier Transform (STFT) of (3) turns this into instantaneous mixture separately for every frequency, giving

$$X_n(t,\omega) = \sum_{m=1}^{M} A_{nm}(\omega) S_m(t,\omega) + V_n(t,\omega)$$
$$n = 1, \ldots, N \tag{4}$$

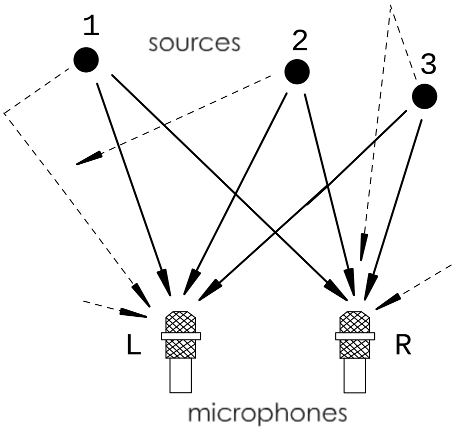

Figure 2. Schematic of a live stereo recording session

where $S_m(t,\omega)$ and $V_n(t,\omega)$ are the STFTs of $s_m(t)$ and $v_n(t)$, respectively. Here t denotes the STFT window position. The temporal transfer function of the mth source signal to the microphone n is defined as

$$A_{nm}(\omega) = \int a_{nm}(\tau) e^{-j\omega\tau} d\tau \doteq \widehat{a}_{nm}(\omega) e^{-j\omega\widehat{\delta}_{nm}(\omega)} \tag{5}$$

where we define

$$\widehat{a}_{nm}(\omega) = \|A_{nm}(\omega)\|; \; \widehat{\delta}_{nm}(\omega) = \angle A_{nm}(\omega) \tag{6}$$

In matrix notation, the model (1) can be written as

$$\mathbf{X}(t,\omega) = \mathbf{A}(\omega)\mathbf{S}(t,\omega) + \mathbf{V}(t,\omega) \tag{7}$$

In our case, $N = 2$, so that (7) in detail looks like

$$\begin{bmatrix} X_1(t,\omega) \\ X_2(t,\omega) \end{bmatrix} = \begin{bmatrix} \widehat{a}_{11}(\omega)e^{-j\omega\widehat{\delta}_{11}(\omega)} & \cdots & \widehat{a}_{1M}(\omega)e^{-j\omega\widehat{\delta}_{1M}(\omega)} \\ \widehat{a}_{21}(\omega)e^{-j\omega\widehat{\delta}_{21}(\omega)} & \cdots & \widehat{a}_{2M}(\omega)e^{-j\omega\widehat{\delta}_{2M}(\omega)} \end{bmatrix} \cdot \begin{bmatrix} S_1(t,\omega) \\ \vdots \\ S_M(t,\omega) \end{bmatrix} + \begin{bmatrix} V_1(t,\omega) \\ V_2(t,\omega) \end{bmatrix} \tag{8}$$

Without loss of generality, we can absorb the attenuation and delay parameters for each frequency bin in the first microphone signal $x_1(t)$ into the definition of the source. In this way, (8) can be rewritten as

$$\begin{bmatrix} X_1(t,\omega) \\ X_2(t,\omega) \end{bmatrix} = \begin{bmatrix} 1 & \cdots & 1 \\ a_1(\omega)e^{-j\omega\delta_1(\omega)} & \cdots & a_M(\omega)e^{-j\omega\delta_M(\omega)} \end{bmatrix} \cdot \begin{bmatrix} S_1(t,\omega) \\ \vdots \\ S_M(t,\omega) \end{bmatrix} + \begin{bmatrix} V_1(t,\omega) \\ V_2(t,\omega) \end{bmatrix} \tag{9}$$

In (9), suppose the vocal component is the kth source, which is associated with STFT $S_k(t,\omega)$, if we can estimate the kth vector $\begin{bmatrix} 1 \\ a_k(\omega)e^{-j\omega\delta_k(\omega)} \end{bmatrix}$ in $\mathbf{A}$, then left multiplying by a vector that is orthogonal to it will completely remove the kth source, achieving the initial goal.

2.3. Estimating the vocal component location

With solo segments of the vocal extracted and their STFT time-frequency cells available, it is possible to estimate the unstructured spatial transfer function for each frequency [2]. Variety of methods to estimate the transfer function are being explored. Considering equation (9) in case of a single source allows estimation of the transfer function parameters by division of the STFT's of the left and right channels, giving

$$a_k(\omega) = \left\| \frac{X_2(\omega)}{X_1(\omega)} \right\| \tag{10}$$

and

$$\delta_k(\omega) = -\frac{1}{\omega}\Im(\log(\frac{X_2(\omega)}{X_1(\omega)})) \tag{11}$$

This simple method of STFT division [3] is not robust to noise and gives unsatisfactory results. Here we describe in detail an autocorrelation method that is robust to uncorrelated additive noise. The autocorrelation matrix of (7) at a given time-frequency cell will be obtained by averaging the following equation:

$$\begin{aligned}\mathbf{X}(t,\omega)\mathbf{X}(t,\omega)^H &= \mathbf{A}(\omega)\mathbf{S}(t,\omega)\mathbf{S}(t,\omega)^H\mathbf{A}(\omega)^H \\ &+ \mathbf{A}(\omega)\mathbf{S}(t,\omega)\mathbf{V}(t,\omega)^H \\ &+ \mathbf{V}(t,\omega)\mathbf{S}(t,\omega)^H\mathbf{A}(\omega)^H \\ &+ \mathbf{V}(t,\omega)\mathbf{V}(t,\omega)^H \end{aligned} \tag{12}$$

We have assumed that signals and noise are uncorrelated. Denoting $\mathbf{R}_x$, $\mathbf{R}_s$ and $\mathbf{R}_v$ as the correlation matrices of the microphones, source signals and noise respectively, and considering signals as stationary, after averaging over time window, we obtain

$$\mathbf{R}_x(\omega) = \mathbf{A}(\omega)\mathbf{R}_s(\omega)\mathbf{A}(\omega)^H + \mathbf{R}_v(\omega) \tag{13}$$

Here we assume that the noise is white, so that matrix $\mathbf{R}_v(\omega)$ is diagonal and $\mathbf{R}_v(\omega) = \sigma_v^2\mathbf{I}_N$, $\forall\omega$ where $\mathbf{I}_N$ is an identity matrix of size N. In segments where only one component s_k exists, (13) becomes

$$\mathbf{R}_x(\omega) = \sigma_{s_k}^2\vec{\mathbf{A}}_k(\omega)\vec{\mathbf{A}}_k(\omega)^H + \sigma_v^2\mathbf{I_N} \tag{14}$$

where

$$\vec{\mathbf{A}}_k(\omega) = \begin{bmatrix} 1 \\ a_k(\omega)e^{-j\omega\delta_k(\omega)} \end{bmatrix} \tag{15}$$

Right multiply (14) with $\vec{\mathbf{A}}_k(\omega)$, we get

$$\mathbf{R}_x(\omega)\vec{\mathbf{A}}_k(\omega) = (\sigma_{s_k}^2\vec{\mathbf{A}}_k(\omega)^H\vec{\mathbf{A}}_k(\omega) + \sigma_v^2)\vec{\mathbf{A}}_k(\omega) \tag{16}$$

This means that $\vec{\mathbf{A}}_k(\omega)$ can be estimated from eigenvalues of the rank-1 matrix $\mathbf{R}_x(\omega)$. The corresponding eigenvalue is

$$\lambda = \sigma_{s_k}^2\vec{\mathbf{A}}_k(\omega)^H\vec{\mathbf{A}}_k(\omega) + \sigma_v^2 \tag{17}$$

One can note that in principle the transfer function can be estimated by dividing the two components of the eigenvector. As will be discussed later, an advantage of our method is that it is robust to added noise. The correlation method can be easily extended for the case of higher order statistics [4] by generalizing equation (12). For example, for the case of 4th order cumulants, we construct a matrix $X(\omega)X^3(\omega)^H - 3X(\omega)X(\omega)^H$. It can be shown that for the case of Gaussian signal, this matrix equals zero since 4th cumulant of a Gaussian signal equals three times the 2nd cumulant (correlation). This effectively eliminates the additional noise matrix from equation (13). The eigenvectors of the resulting matrix are same as for the correlation case.

2.4. Vocal removal

After obtaining $\vec{\mathbf{A}}_k(\omega)$, we find a vector orthogonal to it

$$\vec{\mathbf{A}}_k^\perp = \begin{bmatrix} -a_k(\omega)e^{-j\omega\delta_k(\omega)} & 1 \end{bmatrix} \tag{18}$$

and then left multiply the microphone signal in (9) with $\vec{\mathbf{A}}_k^\perp$ to remove the kth component:

$$\begin{aligned}&\vec{\mathbf{A}}_k^\perp \begin{bmatrix} X_1(t,\omega) \\ X_2(t,\omega) \end{bmatrix} = \\ &\begin{bmatrix} \vec{\mathbf{A}}_k^\perp\vec{\mathbf{A}}_1 & \cdots & 0_{(kth)} & \cdots & \vec{\mathbf{A}}_k^\perp\vec{\mathbf{A}}_M \end{bmatrix} \\ &\cdot \begin{bmatrix} S_1(t,\omega) \\ \vdots \\ S_M(t,\omega) \end{bmatrix} + \vec{\mathbf{A}}_k^\perp \begin{bmatrix} V_1(t,\omega) \\ V_2(t,\omega) \end{bmatrix}\end{aligned} \tag{19}$$

One special case of this is when only two components exist in the stereo recording, in which we can extract the two sources individually. Both sides of (19) are one dimensional, so that the resulting sound is mono.

2.5. Solo segment extraction

As the starting point of the estimation, solo segments in a song carry all the information of the unknown transfer functions. In order to use the algorithm described in 2.2, it's important to have a good extraction algorithm for the solo segments. There are several approaches available. We can use the Gaussian Mixture Model in [2] to find time-frequency cells with only one source, or take the W-disjoint assumption as indicated in [3] and and look for clusters near the center of the stereo field. Voice activity detection (VAD) algorithms can be utilized as well [5]. Implementation for solo segment extraction is currently manual.

2.6. Delay refinement and bass emphasis

Estimation accuracy will start to decrease in high frequency bins as the wavelength of the signal becomes so short that transfer functions become sensitive to movements of the singer or instrument performer. This will result in some residual transients (consonant attacks) of the vocal in the processed sound. A method for delay refinement is being used on a frame by frame basis in order to search for a value of optimal delay correction that might happen due to minor movements of the sound source. By searching over range of possible delay parameters around the theoretical value we are looking for minimum energy of the resulting signal, which will occur if the source is more precisely removed.

Removing the vocal component in the manner described in this paper usually remove a certain amount of bass from the recording at the same time, since low frequency instruments are often not very localized sources and their locations could coincide with the vocal. Another fact is that vocal usually doesn't occupy the same frequency band as the bass. The fundamental frequency of voiced speech varies according to a lot of factors like emotion and age. Literature [6] shows that the typical range is 85-155Hz for adult male and 165-255 for female. Taking this as a reference, we choose a cut-off frequency of 120Hz to lowpass filter the original stereo material, and mix it with the mono vocal-free track. The filtered material won't contain much of the vocal and still will contain most of the bass information as well as preserve the stereo field in the low frequencies. Simulated results shows that this method is very effective for the compensation of low frequencies and it increases some stereo feeling of the accompanying track.

3. IMPLEMENTATION AND PROTOTYPE

This system design is implemented in Matlab. Several CD recordings of different genres have been tested and compared to the existing method (left-right subtraction). We chose hamming window for the STFT and window lengths of 1024 and 2048. For songs with heavy post productions where the vocal is recorded mono and panned to the center with few reverb, our system works similarly to the existing method of vocal removal, but the bass emphasis enhances the final sound quality noticeably. In recording situations more similar to our initial assumptions as are mentioned in 2.1, for example a concert recording of Billie Holiday with a band, the existing method fails to remove the vocal at all, while our system is as robust as the in the previous situation. Some sound examples are available online. [1]

4. DISCUSSION AND FUTURE RESEARCH

The prototype system turns out to be well functioning, especially when the real situation is close to the assumptions. As can be seen from equation (16), added noise doesn't compromise the accuracy of estimating $\vec{\mathbf{A}}_k(\omega)$ from the eigenvector of $\mathbf{R}_x(\omega)$. So, this algorithm is robust in a noisy environment. The model in the system consider different attenuations and delays for different frequencies, and transfer function is assumed, so it is also robust to reverberant vocal sound, even artificial reverb, if it's linear. For a typical live recording session, there are usually close mics for individual instruments. Usually, these mic signals are later added to the stereo recording for the whole ensemble to boost certain instruments. This is still a linear operation which well fits in our assumptions for our model to work well in theory. We also have considered delay refinement in order to compensate for small movements of the source. However, many more detailed properties of the acoustics are yet to be investigated, such as the geometrical conditions and spatial sampling distance between the microphones. STFT window types and window sizes also matter. The window should at least be suitable for perfect reconstruction with proper hop length - a hamming window for example. The window length should be longer than the longest possible reverb time so that the transfer function model holds for most portion of a window, instead of letting to much convoluted tails fall into neighboring windows. On the other hand, lengthened window will result in fewer STFT time slices for the autocorrelation time average. As a result, the decision of window length and shape should consider all these tradeoffs. Further more, solo segment extraction algorithms are to be compared and tested to automate the extraction process, and delay refinement and transient removal algorithms are under development.

5. REFERENCES

[1] Nomura, Takashi, Denki, M., Kaisha, K. *Voice canceler with simulated stereo output* 1996. U.S. Pat. 05550920.

[2] Dubnov, S., Tabrikian, J., Arnon-Targan, M. "A Method for Directionally-Disjoint Source Separation in Convolutive Environment", *ICASSP*, Vol. 5 , pp. 489 C 492, Montreal, Canada, 2004

[3] Jourjine, A., Rickard, S., and Yilmaz, O. "Blind Separation of Disjoint Orthogonal Signals: Demixing N Sources from 2 Mixtures", *ICASSP*, Vol. 5, pp. 2985-2988, Istanbul, Turkey, June 2000

[4] Mendel, J.M. "Tutorial on higher-order statistics (spectra) in signal processing and system theory: theoretical results and some applications", *Proceedings of the IEEE*, Vol. 79, Issue 3, pp. 278 - 305, March 1991

[5] Fisher, E., Tabrikian, J., Dubnov, S. "Generalized likelihood ratio test for voiced/unvoiced decision using the harmonic plus noise model" *ICASSP*, vol.1, pp. 440-443, 2003

[6] Baken, R. J. *Clinical Measurement of Speech and Voice*. London: Taylor and Francis Ltd., 1987.

[1] http://crca.ucsd.edu/˜pxiang/research/vocalremoval.html