Single-Voice Separation From Monaural Recordings using robust principal component analysis

report by :

Mohsen Nabian

April 2017

- goal: separating voices from music accompaintiment

- assumption of this paper:

Music accompainiment $\longrightarrow$ low rank subspace

Singing voices $\longrightarrow$ relatively sparse within songs.

- Technique :

we use RPCA (Robust principal component Analysis) which is a matrix factorization algorithm for solving underlying low-rank and sparse matrices
(1) (2)

- mathematical formolulation :

$M \in \mathbb{R}^{n_1 \times n_2} \longrightarrow$ original song

$L \in \mathbb{R}^{n_1 \times n_2} \longrightarrow$ accompanied music

$S \in \mathbb{R}^{n_1 \times n_2} \longrightarrow$ singing voice

$$\boxed{\begin{array}{l} \text{minimize } \|L\|_* + \lambda \|S\|_1 \\ \text{s.t.} \quad L + S = M \end{array}}$$

use $\lambda_k = \dfrac{k}{\sqrt{\max(n_1, n_2)}}$ with different values of k and test which k is better.

- First we compute spectogram of music signals as matrix M, calculated from short-time-fourier-transform (STFT)

- Second, use ALM, an efficient algorithm for solving SPCA problems: $L+S=|M|$

- then we have found L and S seperately.

- also to obtain waveforms of L and S back, we record the phase of original signal: $P = \text{phase}(M)$

then append the phase to matrix L and S.

$$\begin{cases} L(m,n) = L\, e^{jP(m,n)} \\ S(m,n) = S\, e^{jP(m,n)} \end{cases}$$

Time-Frequency Masking

after finding L and S, we apply binary time-frequency masking for better separation.

binary-time frequency masking M_b as follows:

$$M_b(m,n) = \begin{cases} 1 & |S(m,n)| > \text{gain} * |L(m,n)| \\ 0 & \text{otherwise} \end{cases}$$

for all $m = 1, \dots, n_1$
$n = 1, \dots, n_2$

So

with mask $\Rightarrow$ $$\begin{cases} X_{singing}(m,n) = M_b(m,n)\, M(m,n) \\ X_{music}(m,n) = (1 - M_b(m,n))\, M(m,n) \end{cases}$$

without mask $$\begin{cases} X_{singing}(m,n) = S \\ X_{music}(m,n) = L \end{cases}$$

3/

Experiment

~~1) The effect of λ_k~~

- MIR-1K dataset
- 1000 song clips   Karaoke + amature singers
- sample rate 16KHz   duration 4 to 13 sec.

3 sets of mixtures:

for each clip, the singing voice and the music accompanied were mixed at -5, 0, and 5 dB SNRs.

energy music is larger   Same energy level   energy singing is larger

evaluations: Source to inference ratio (SIR)
Source to artifacts " (SAR)
Source to distortion " (SDR)

$$NSDR(\hat{v}, v, x) = SDR(\hat{v}, v) - SDR(x, v)$$

singing voice   original clean singing voice   Preprocessed mixture

length of nth song.

$$GNSDR(\hat{v}, v, x) = \frac{\sum_{n=1}^{N} w_n \, NSDR(\hat{v}_n, v_n, x_n)}{\sum_{n=1}^{N} w_n}$$

4/

1) effect of λ_K :

Fig 3 : $\left\{ \begin{array}{l} \text{with no mask} \\ \text{with } \cancel{\text{no}} \text{ binary mask} \end{array} \right.$

$\lambda_K = \{0.1, 0.5, 1, 1.5, 2, 2.5\}$

$SNR = \{-5, 0, 5\}$

in all these cases $\left\{ \begin{array}{l} SDR \\ SAR \\ SIR \end{array} \right.$ is reported.

2) The effect of gain factor with a binary mask

Fig 4 : Cases : $\left\{ \begin{array}{l} \lambda_1 = \{0.1, 0.5, 1, 1.5, 2\} \\ SNR = \{-5, 0, 5\} \end{array} \right.$

in all these cases $\left\{ \begin{array}{l} SDR \\ SAR \\ SIR \end{array} \right.$ are reported

Conclusion: higher λ_1, results in lower power sparse matrix S. $\Rightarrow$ Larger interference
fewer artifacts

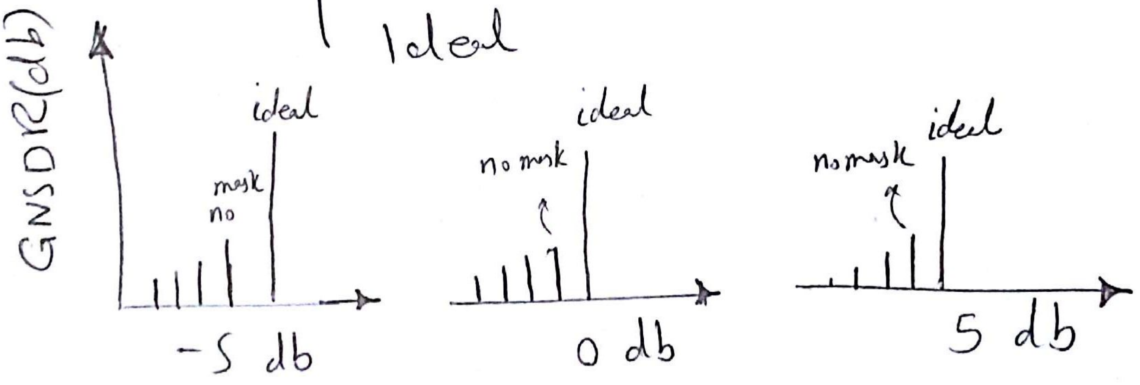
51

3) Comparison with prev. systems:

5 Cases are Compared for each 3 Conditions

Conditions: SNR = {-5, 0, 5} → voice to music ratio

Cases:
- Hsu algorithm
- Rafii ~
- binary mask λ_1, gain=7
- no mask
- Ideal



no mask provided the best result out of others.