Multiple Sound Sources Localization under Low SNR Conditions by Automatically Finding Target Source Using NMF with Directional Clustering and HMM

Software Engineer

Signal processing/Machine learning/Acoustic engineering /Optimization problems/Blind Source Separation/Auditory Scene Analysis

Seoul National University of Science & Technology
Mechanical System Design Engineering
Sonmook Oh(Luke)

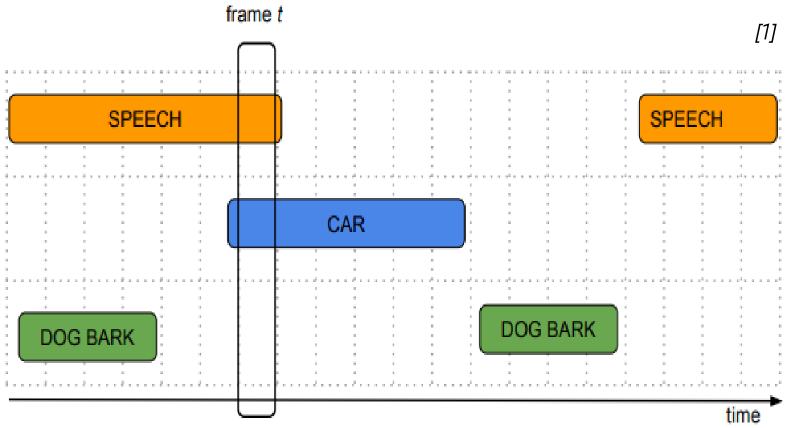
Realistic Environments & Complicated Circumstances

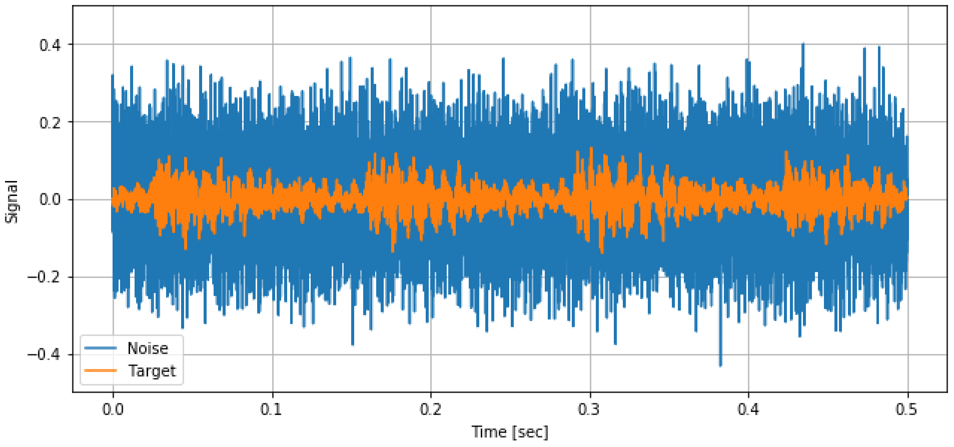
※Which source is the target? & Where is the target source? & What is the target Source?

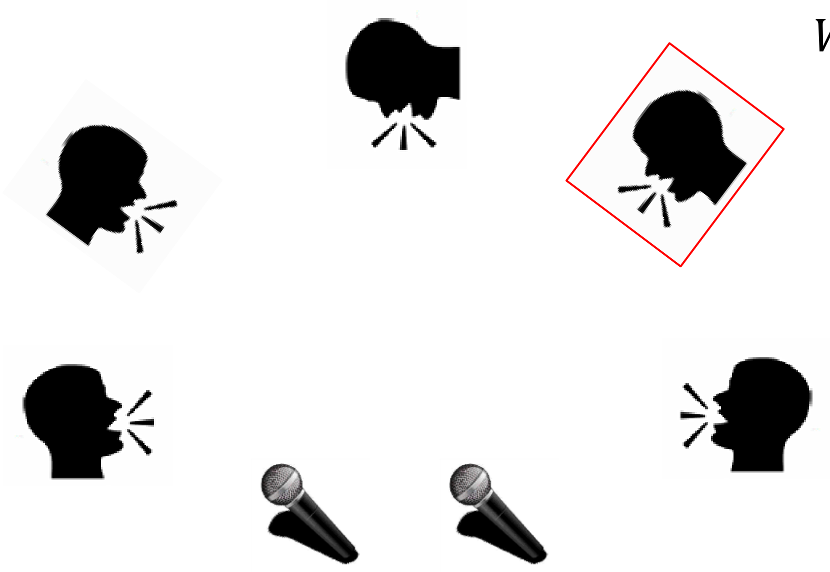
“

Hard to recognize/***localize***/separate

- ❖ Noisy environment
- ❖ Reverberant room
- ❖ Multi sources
- ❖ Same/Similar direction(depth)







$$When, \boldsymbol{a}(f,\tau_n) \cong \boldsymbol{b}(f,\tau_n) \text{ or } \boldsymbol{a}(f,\tau_n) < \boldsymbol{b}(f,\tau_n) \quad (1)$$

$$\boldsymbol{x}(t,f) = \sum_{n=1}^{N} \boldsymbol{a}(f,\tau_n) S_n(t,f) + \sum_{n=1}^{M} \boldsymbol{b}(f,\tau_n) N_n(t,f) + \boldsymbol{c}(t,f) \quad (2)$$

$$\boldsymbol{a}(f,\tau_n) S_n(t,f) \quad (3)$$

$$\boxed{\tau_n \quad (4)}$$

$When,$

$$\boldsymbol{x}(t,f) = [x_1(t,f), x_2(t,f)]^T$$

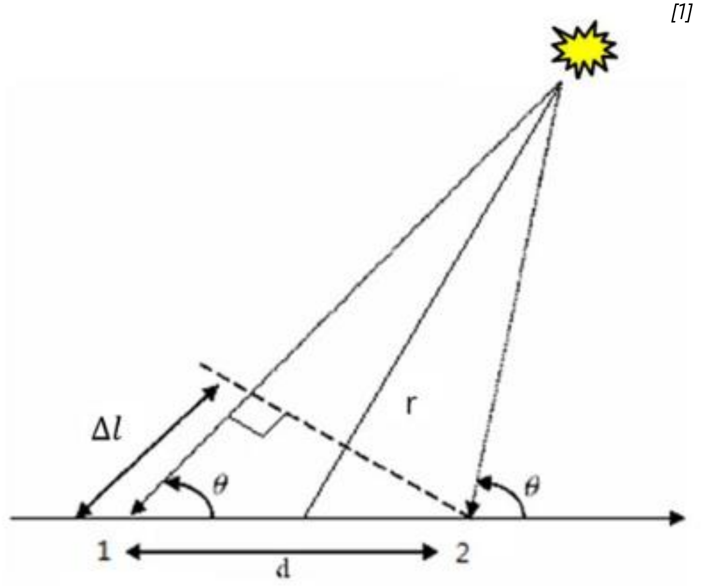
$$\boldsymbol{a}(f,\tau_n) = \left[1, a_n e^{-2i\pi f \tau_n}\right]^T$$

$$\boldsymbol{b}(f,\tau_n) = \left[1, b_n e^{-2i\pi f \tau_n}\right]^T$$

[1] Sharath Adavanne, Giambattista Parascandolo, Pasi Pertila, Toni Heittola, Tuomas Virtanen(2016). SOUND EVENT DETECTION IN MULTICHANNEL AUDIO USING SPATIAL AND HARMONIC FEATURES. Detection and Classification of Acoustic Scenes and Events 2016

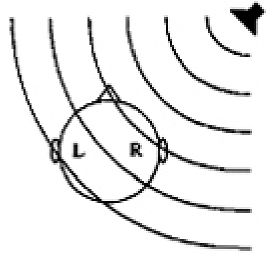
Time difference of arrival(Tdoa)

※General method of Sound direction from Multichannel signal(two microphones)



$$\cos\theta = \frac{\Delta l}{d} = \frac{c\Delta t}{d} = \frac{cN}{d f_s} \quad (1)$$

$$\theta = arcos\left(\frac{cN}{d f_s}\right) \quad (2)$$



[Preconditions]

- Plane wave
- $r \gg d$

$c = 330 + 0.6temp = 343.2\ m/s(22℃)$

N : $delay\ sample$

d : $distance\ between\ microphones = 0.1001m$

$f_s = 48000Hz$

[Constraint Conditions]

$$-1 \le \cos\theta \le 1$$

$$-1 \le \frac{c * N}{d * f_s} \le 1$$

$$-14 \le N \le 14$$

$$\Delta l < \frac{\lambda}{2}$$

$$c\Delta t < \frac{\lambda}{2} (\because c = f * \lambda)$$

$$\frac{N}{f_s} < \frac{1}{2f}$$

$$N < \frac{f_s}{2f}$$

$$d < \frac{\lambda_{min}}{2}$$

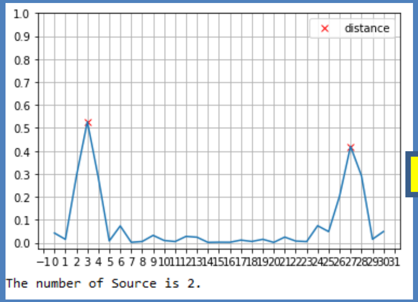
$$d < \frac{c}{2f_{max}}$$

$$f_{max} < \frac{c}{2d}$$

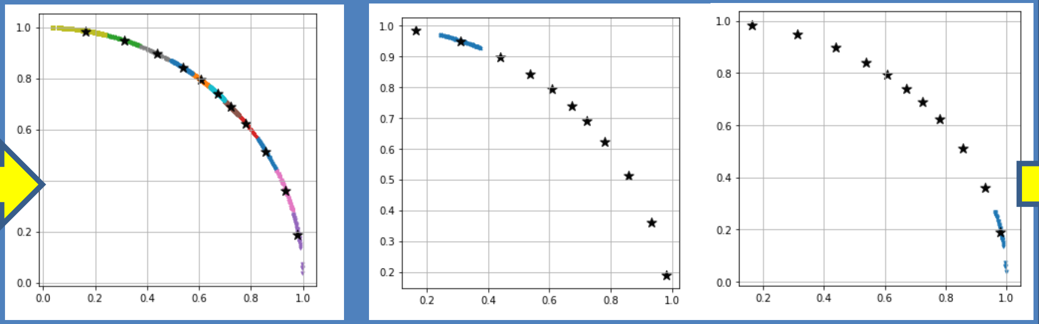
$$f_{max} < \frac{c}{2d} = 1714Hz$$

Whole Progress Flow / Idea Flow

※Whole process

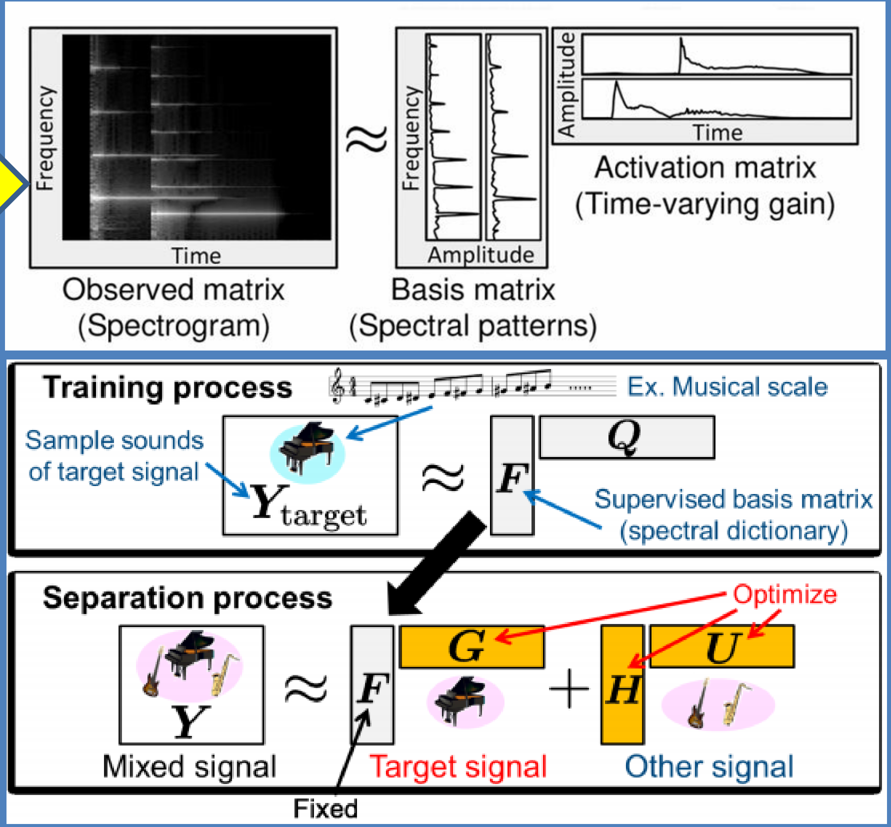


Source Counting and Rough Angle Estimation

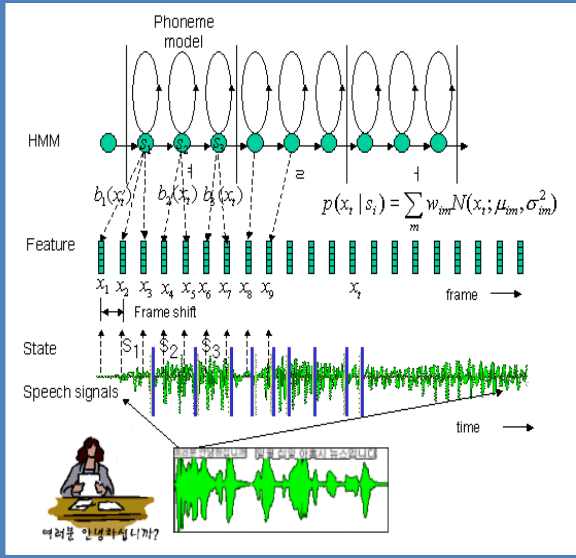


Directional Clustering and Binary Masking

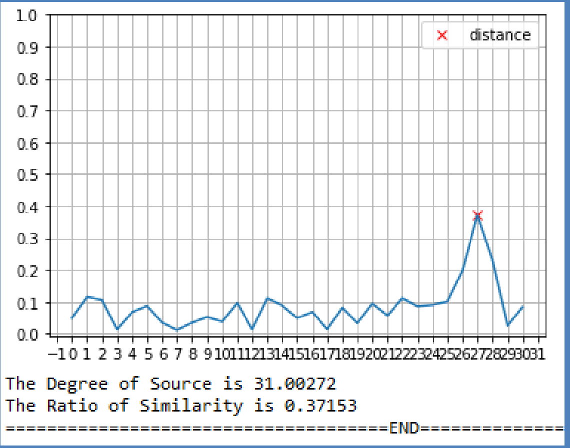
[1]



Separation by Supervised NMF



Target Selection by Supervised HMM



Final Direction Estimation

[Algorithm]

- GCC-Phat (Generalized Cross Correlation-Phase transform)
- K-means clustering
- NMF (Nonnegative Matrix Factorization)
- HMM (Hidden Markov Model)

[1] D. Kitamura, H. Saruwatari, H. Kameoka, Y. Takahashi, K. Kondo and S. Nakamura, "Multichannel Signal Separation Combining Directional Clustering and Nonnegative Matrix Factorization with Spectrogram Restoration," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 23, no. 4, pp. 654-669, April 2015.

Finding the most similar point

※Generalized Cross Correlation Phase Transform(GCC-PHAT)

Source Counting and Rough Angle Estimation

↓

Directional Clustering and Binary Masking

↓

Restoration(Extrapolation)

↓

Separation by Supervised NMF

↓

Target Selection by Supervised HMM

↓

Final Direction Estimation

$$x_1(t) * x_2(t) = \int_{-\infty}^{\infty} x_1(\tau) x_2(t-\tau) d\tau \quad (1)$$

$$w(t) = x_1(t) \otimes x_2(t) = \int_{-\infty}^{\infty} x_1(\tau)^* x_2(\tau + t) d\tau \quad (2)$$

$$W(f) = X_1{}^*(f) X_2(f) \quad (3)$$

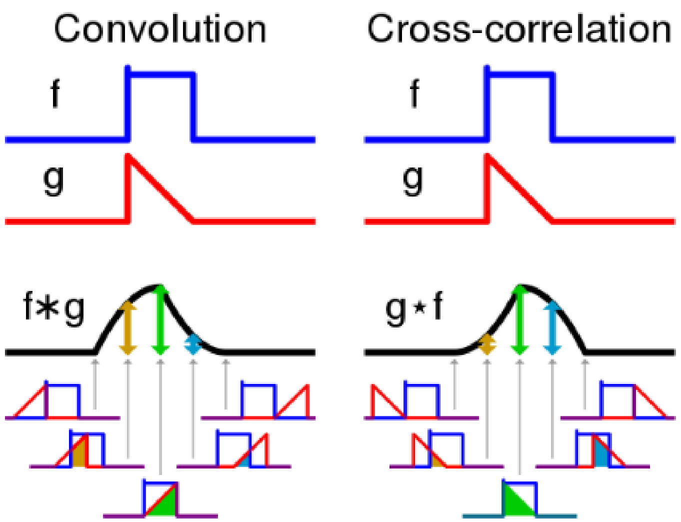
$$w(t) = \int_{-\infty}^{\infty} W(f) e^{i2\pi f t} df \quad (4)$$

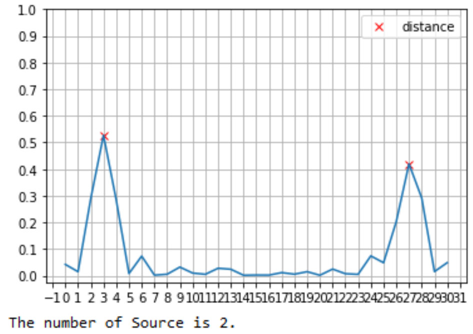
$$gcc(t) = \int_{-\infty}^{\infty} \psi(f) W(f) e^{i2\pi f t} df \quad (5)$$

$$gcc\_phat(t) = \int_{-\infty}^{\infty} \frac{W(f)}{|W(f)|} e^{i2\pi f t} df \quad (6)$$

$$TDOA \ or \ ITD = arg \max_t gcc\_phat(t) \quad (7)$$

[1]





$$\theta_i = arcos\left(\frac{c * N_i}{d * f_s}\right) \quad (8)$$

$$\text{Coordinate of ith source : } \left(\cos\frac{\theta_i}{2}, \sin\frac{\theta_i}{2}\right) \quad (9)$$

Binding the most similar things together

※K-means clustering

Source Counting and Rough Angle Estimation

↓

Directional Clustering and Binary Masking

↓

Restoration(Extrapolation)

↓

Separation by Supervised NMF

↓

Target Selection by Supervised HMM

↓

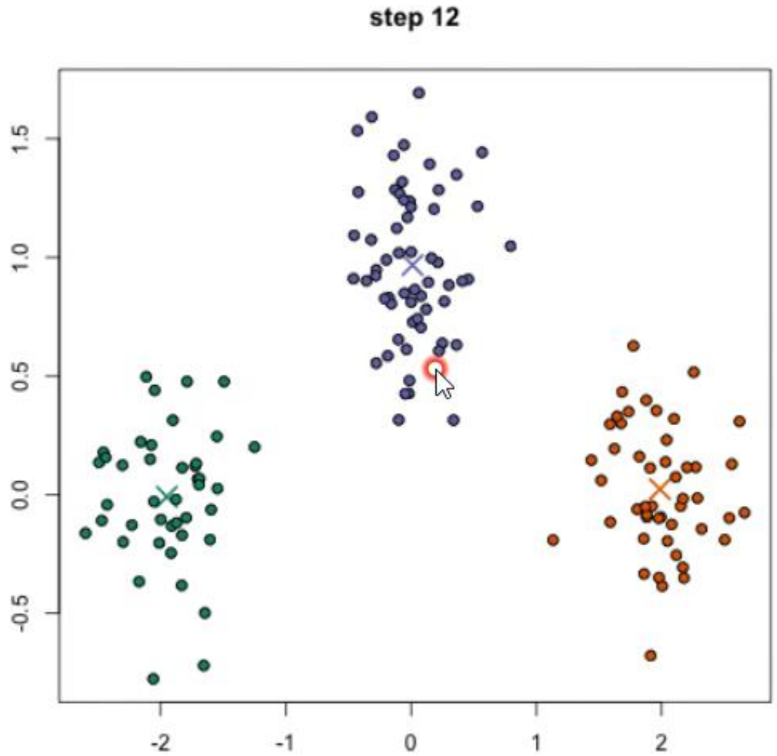
Final Direction Estimation

$$J = \sum_{i=1}^{n} \sum_{k=1}^{K} r_{ik} \|\boldsymbol{x}_i - \boldsymbol{\mu}_k\|^2 \quad (1)$$

$$r_{ik} = \begin{cases} 1 & if\ k = arg\min_j \left\|\boldsymbol{x}_i - \boldsymbol{\mu}_j\right\|^2 \\ 0 & otherwise \end{cases} \quad (2)$$

$$\boldsymbol{\mu}_k = \frac{\sum_{i=1}^{n} r_{ik} \boldsymbol{x}_i}{\sum_{i=1}^{n} r_{ik}} \quad (3)$$

[1]



- ❖ K : *number of clusters*
- ❖ r_{ik} : *responsibility* (어디에 속해 있는지 나타내는 indicator)

[1]https://www.nextobe.com/single-post/2018/02/26/%EB%8D%B0%EC%9D%B4%ED%84%B0-%EA%B3%BC%ED%95%99%EC%9E%90%EA%B0%80-%EC%95%8C%EC%95%84%EC%95%BC-%ED%95%A0-5%EA%B0%80%EC%A7%80-%ED%81%B4%EB%9F%AC%EC%8A%A4%ED%84%B0%EB%A7%81-%EC%95%8C%EA%B3%A0%EB%A6%AC%EC%A6%98

Where and how much of source component is distributed

※Directional Clustering & Binary Mask

Source Counting and Rough Angle Estimation

↓

Directional Clustering and Binary Masking

↓

Restoration(Extrapolation)

↓

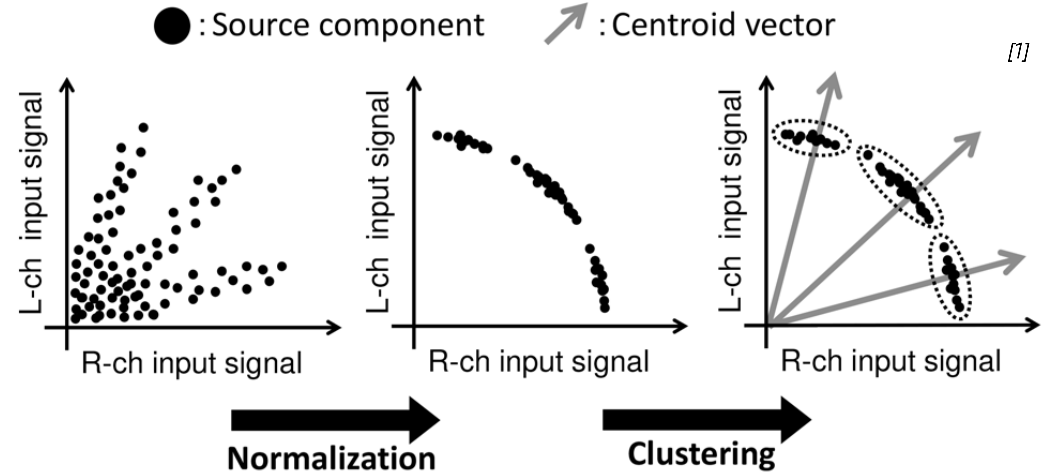
Separation by Supervised NMF

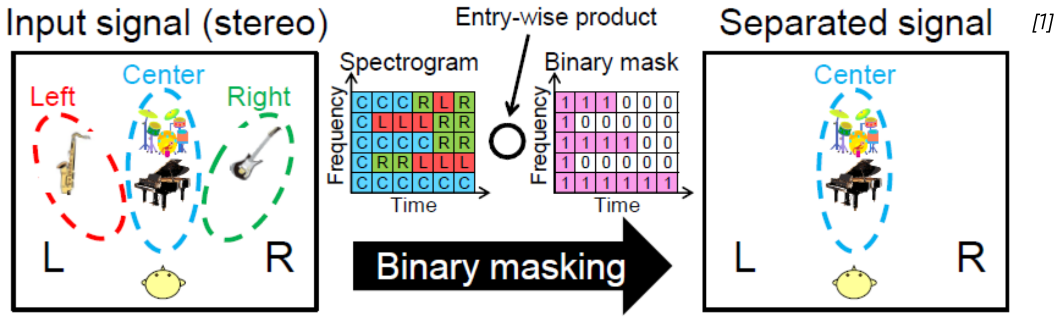
↓

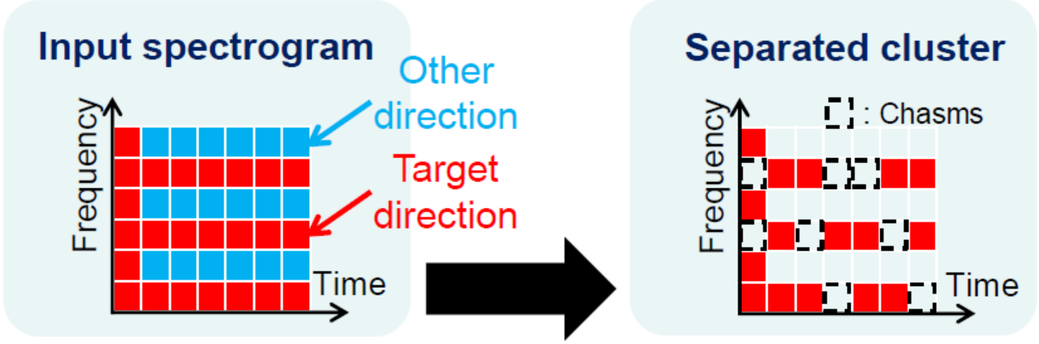
Target Selection by Supervised HMM

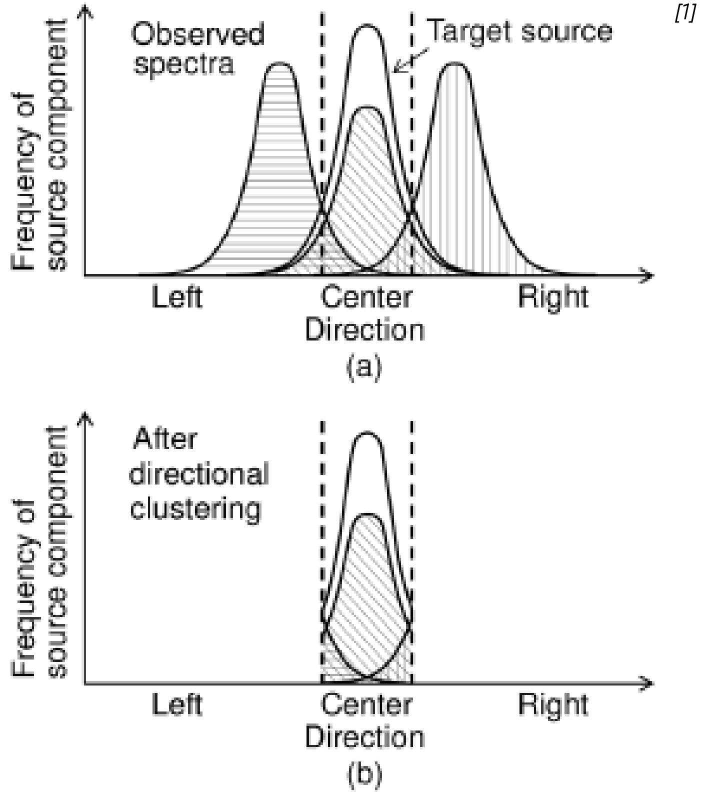
↓

Final Direction Estimation









Utilizing imaginary feature

※Hybrid Clustering combining Level difference and Phase difference

Source Counting and Rough Angle Estimation

↓

Directional Clustering and Binary Masking

↓

Restoration(Extrapolation)

↓

Separation by Supervised NMF

↓

Target Selection by Supervised HMM

↓

Final Direction Estimation

$$\Delta\phi(n,f) \cong arg\left[\frac{x_1(n,f)}{x_2(n,f)}\right] = 2\pi f\tau(n,f) \quad (1)$$

$$\tau(n,f) = \frac{1}{2\pi f} arg\left[\frac{x_1(n,f)}{x_2(n,f)}\right] \quad (2) \qquad \theta(n,f) = arcos\left(\frac{c*\tau(n,f)}{d}\right) \quad (3)$$

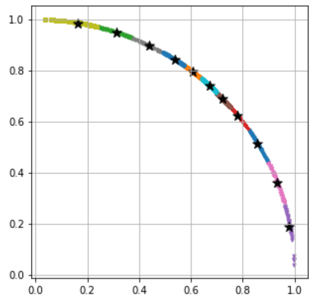
$$\text{Coordinate of each component : }\left(\cos\frac{\theta(n,f)}{2}, \sin\frac{\theta(n,f)}{2}\right) \quad (4)$$

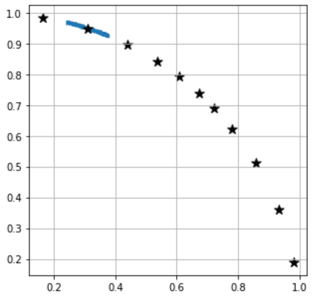
$$(x,y) = \begin{cases} \left(\cos\frac{\theta(n,f)}{2}, \sin\frac{\theta(n,f)}{2}\right) & if\ \ f < \frac{c}{2d} = 1714Hz \\ \left(\frac{|x_2(n,f)|}{\sqrt{|x_1(n,f)|^2+|x_2(n,f)|^2}}, \frac{|x_1(n,f)|}{\sqrt{|x_1(n,f)|^2+|x_2(n,f)|^2}}\right) & otherwise \end{cases} \quad (5)$$

$$\text{Coordinate of source : } \boldsymbol{x}_i = \left(\cos\frac{\theta_i}{2}, \sin\frac{\theta_i}{2}\right)(6)$$

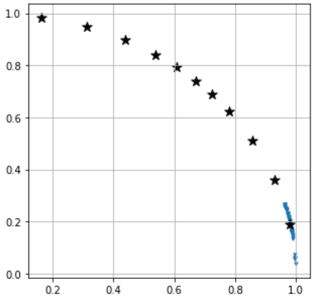
$$k = arg\min_j \left\|\boldsymbol{x}_i - \boldsymbol{\mu}_j\right\|^2 \ (7)$$

$$Binary\ mask = \begin{cases} 1 & if\ k \\ 0 & otherwise \end{cases} (8)$$





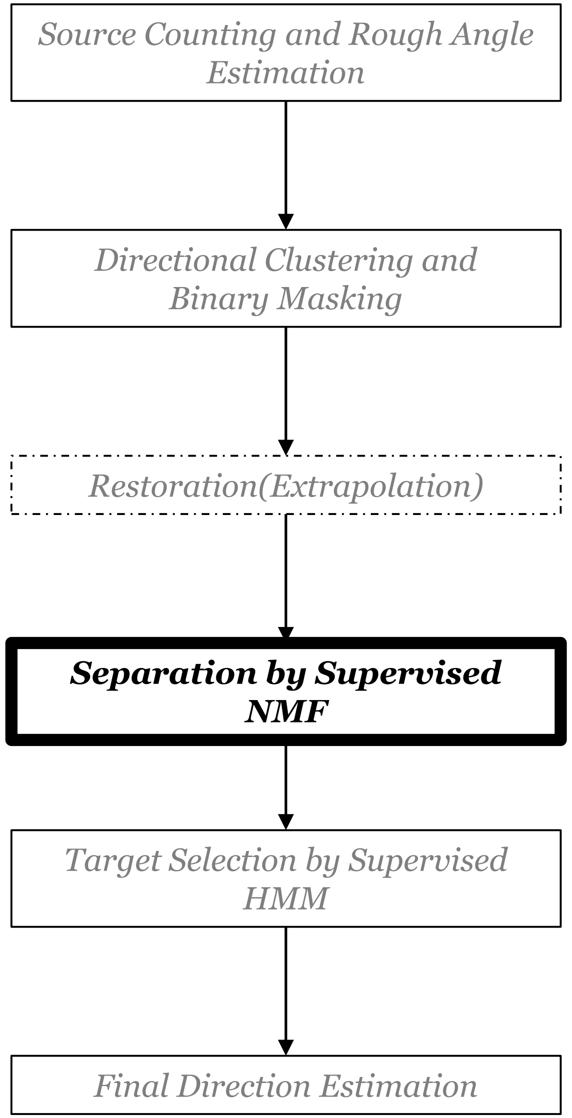
[Target]

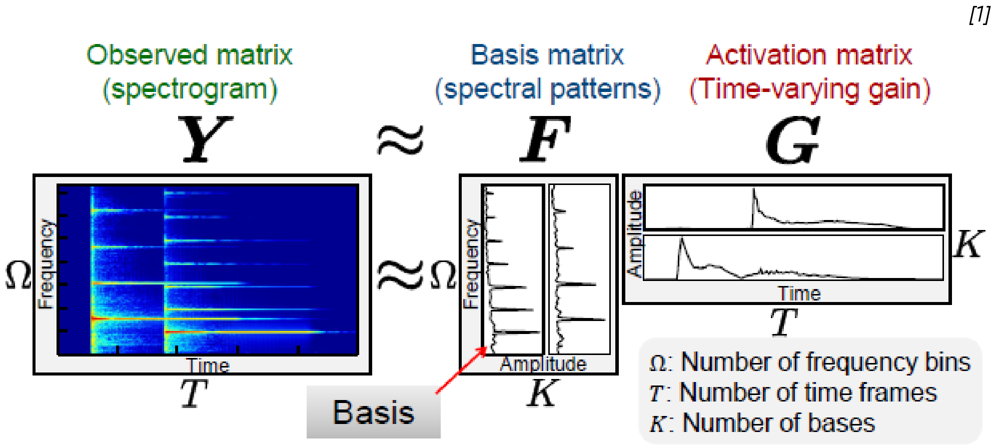


[Noise]

Reduction remaining noise

※Nonnegative Matrix Factorization





$$X \cong VW \quad (1)$$

$$\mathcal{J}_{NMF} = \mathcal{D}(X \parallel VW) \quad (2)$$

When,

$X \in \mathbb{R}_{\geq 0}^{M\times N}$

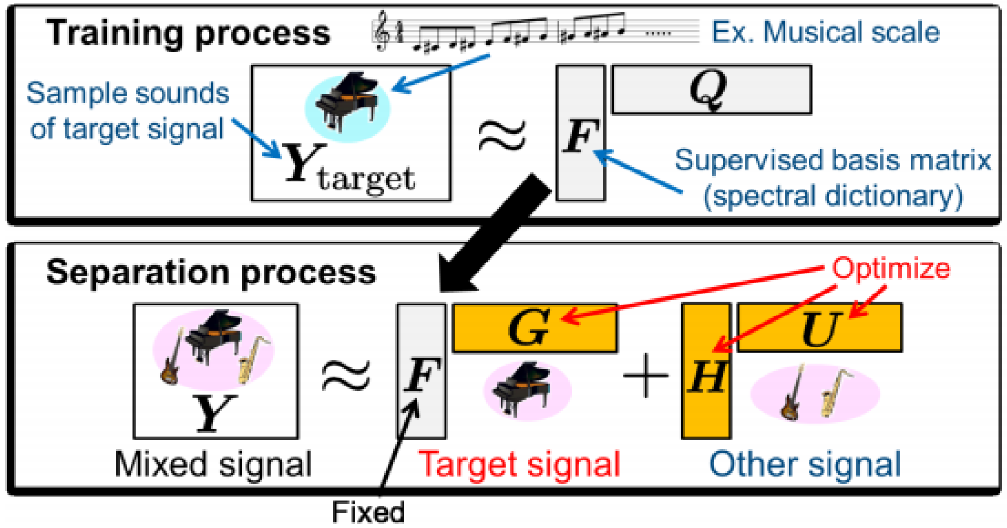
$V \in \mathbb{R}_{\geq 0}^{M\times D}$

$W \in \mathbb{R}_{\geq 0}^{D\times N}$

- β-divergence

$$d_\beta(x|y) = \frac{x^\beta}{\beta(\beta-1)} + \frac{y^\beta}{\beta} - \frac{xy^{\beta-1}}{\beta-1}$$

- $\beta = 2$ (Euclidean): $d(x|y) = \frac{1}{2}(x-y)^2$
- $\beta = 1$ (Kullback-Leibler): $d(x|y) = x\log\frac{x}{y} - x + y$
- $\beta = 0$ (Itakura-Saito): $d(x|y) = \frac{x}{y} - \log\frac{x}{y} - 1.$



$$Y_{noise} \cong FQ \quad (3)$$

$$X \cong FG + HU \quad (4)$$

$$\mathcal{J}_{NMF} = \mathcal{D}(X \parallel FG + HU) \quad (5)$$

$$h_{\omega,l} \leftarrow h_{\omega,l}\left(\frac{\sum_t y_{\omega,t}u_{l,t}z_{\omega,t}^{\beta-2}}{\sum_t u_{l,t}z_{\omega,t}^{\beta-1}}\right)^{\varphi(\beta)},$$

$$g_{k,t} \leftarrow g_{k,t}\left(\frac{\sum_\omega f_{\omega,k}y_{\omega,t}z_{\omega,t}^{\beta-2}}{\sum_\omega f_{\omega,k}z_{\omega,t}^{\beta-1}}\right)^{\varphi(\beta)},$$

$$u_{l,t} \leftarrow u_{l,t}\left(\frac{\sum_\omega h_{\omega,l}y_{\omega,t}z_{\omega,t}^{\beta-2}}{\sum_\omega h_{\omega,l}z_{\omega,t}^{\beta-1}}\right)^{\varphi(\beta)},$$

[1] D. Kitamura, H. Saruwatari, H. Kameoka, Y. Takahashi, K. Kondo and S. Nakamura, "Multichannel Signal Separation Combining Directional Clustering and Nonnegative Matrix Factorization with Spectrogram Restoration," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 23, no. 4, pp. 654-669, April 2015.

Probabilistic model suitable for time series data

※Hidden Markov Model

Source Counting and Rough Angle Estimation

↓

Directional Clustering and Binary Masking

↓

Restoration(Extrapolation)

↓

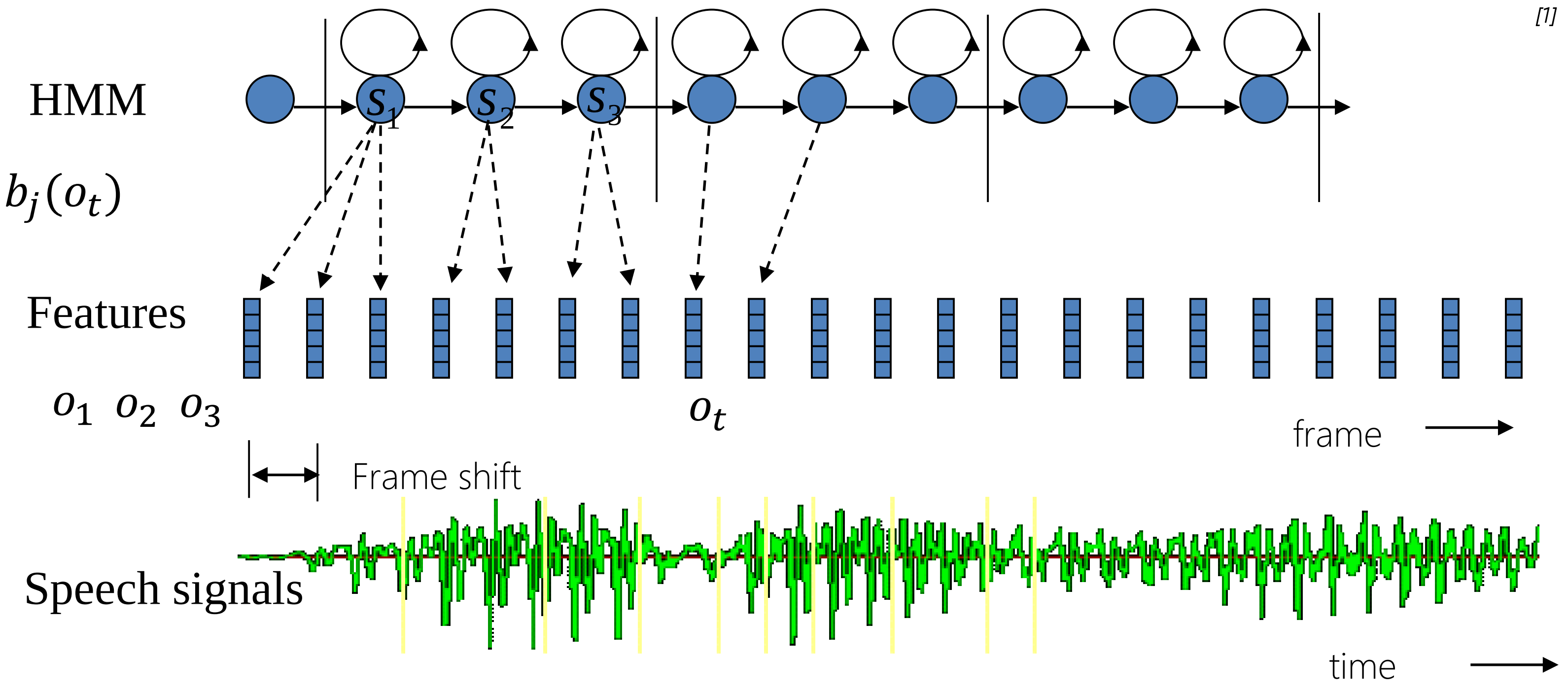
Separation by Supervised NMF

↓

Target Selection by Supervised HMM

↓

Final Direction Estimation



$$Target = arg\max_{O_k} P(O_k|\lambda)$$

$$O = \{o_1,\ o_2,\ o_3,\ o_4,\ \cdots\}$$

$$\lambda = (A, B, \pi)$$

$$P(O|\lambda)$$

A : *state transition probability* $\{a_{ij}|a_{ij} = p(q_{t+1} = j|q_t = i)\}$

B : *symbol output&observation probability* $\{b_j(o_t)|b_j(o_t) = p(x = o_t|q_t = j)\}$

π: *initial state distribution probability* $\{\pi_i|\pi_i = p(q_1 = i)\}$

Automatically Finding Target and Estimating Target Direction

※Maximum likelihood Estimation & GCC-PHAT

Source Counting and Rough Angle Estimation

↓

Directional Clustering and Binary Masking

↓

Restoration(Extrapolation)

↓

Separation by Supervised NMF

↓

Target Selection by Supervised HMM

↓

Final Direction Estimation

$$Target_1 = arg \max_{x_1} p(x_1|\lambda) \quad (1)$$

$$Target_2 = arg \max_{x_2} p(x_2|\lambda) \quad (2)$$

$$w(t) = x_1(t) \otimes x_2(t) = \int_{-\infty}^{\infty} x_1(\tau)^* x_2(\tau + t) d\tau \quad (3)$$

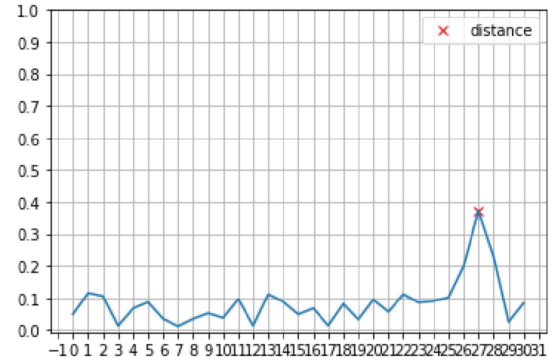
$$W(f) = {X_1}^*(f) X_2(f) \quad (4)$$

$$w(t) = \int_{-\infty}^{\infty} W(f) e^{i2\pi ft} df \quad (5)$$

$$gcc(t) = \int_{-\infty}^{\infty} \psi(f) W(f) e^{i2\pi ft} df \quad (6)$$

$$gcc\ phat(t) = \int_{-\infty}^{\infty} \frac{W(f)}{|W(f)|} e^{i2\pi ft} df \quad (7)$$

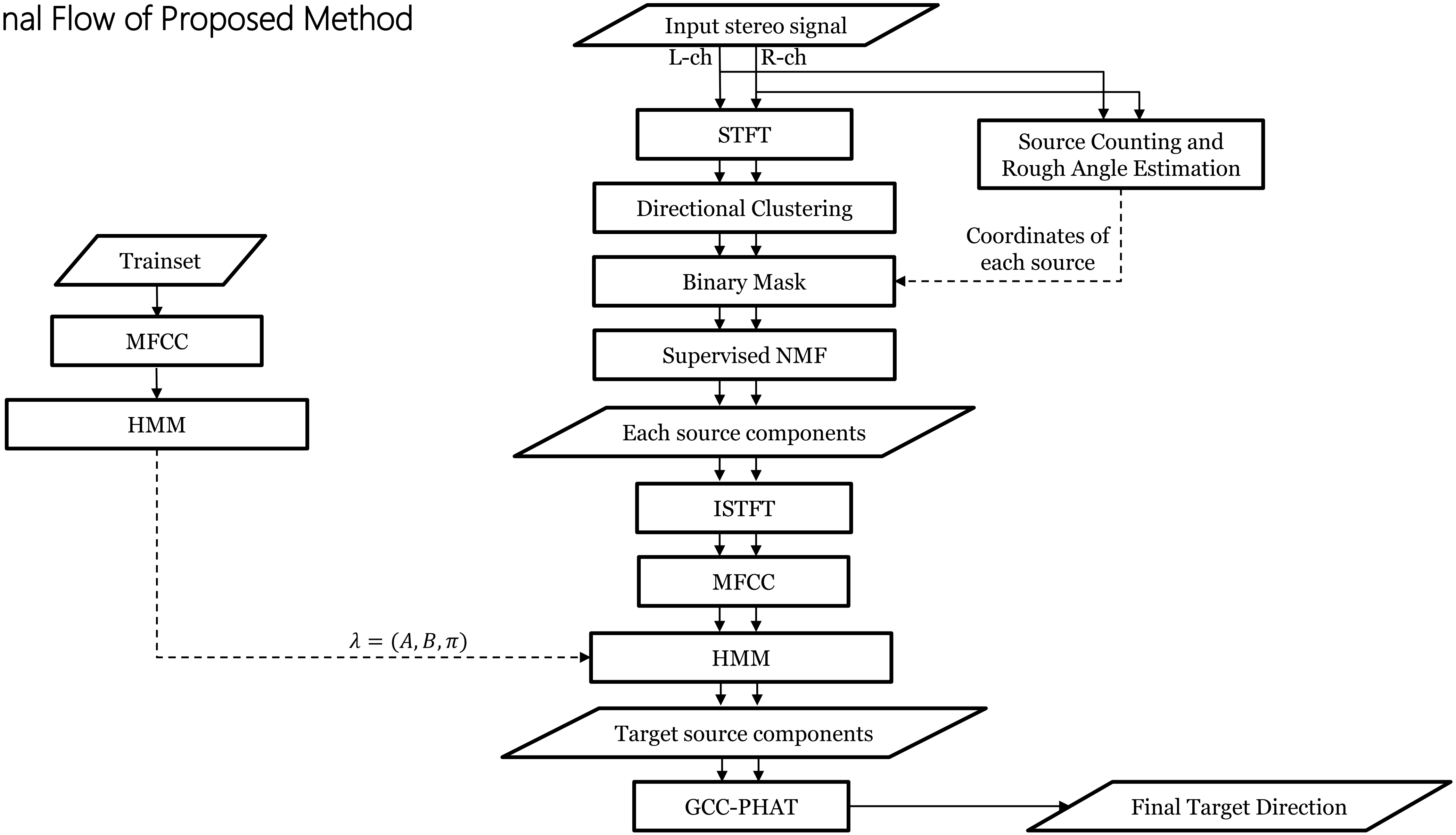
$$TDOA\ or\ ITD = arg \max_{t} gcc\_phat(t) \quad (8)$$




```
The Degree of Source is 31.00272
The Ratio of Similarity is 0.37153
====================================END==============
```

Whole Progress Flow / Idea Flow

※Signal Flow of Proposed Method



※Experiment Parameters and Circumstance

Parameter	Notation	Value
Number of microphones	M	2
Distance between microphones	d	0.1001 m
Sampling frequency	f_s	48000 Hz
Sound speed	c	343.2 m/s(22℃)
Speaker distance		0.5 m
History length (block size)		10 frames (0.5 second)
Frame size		0.05 second
Overlapping in time		50 %
STFT size		256 samples
Source count threshold	ϵ	0.2
Window type	hamming	

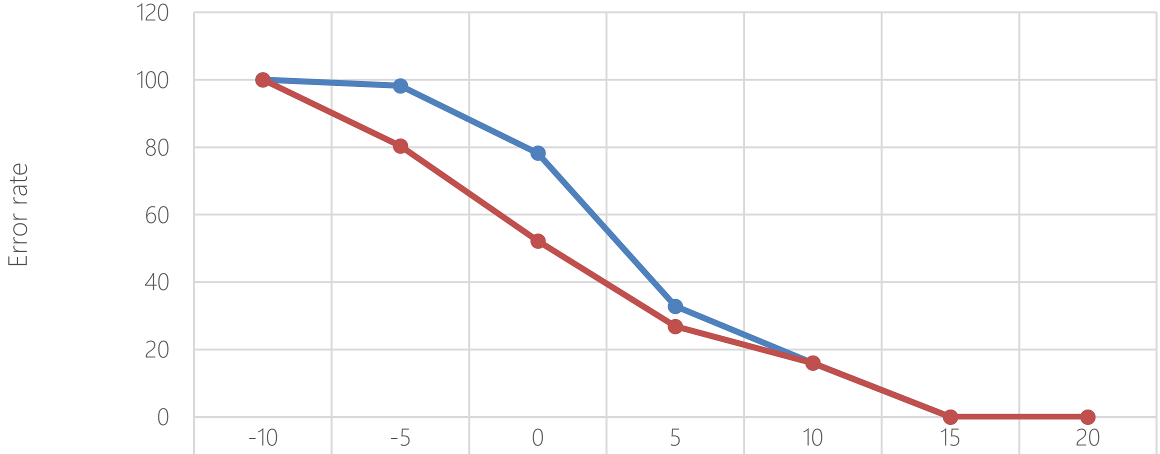
[Ideal Circumstance & Conditions]

- ❖ 평면파
- ❖ 초소형 고성능 마이크
- ❖ 잔향 최소화
- ❖ 노이즈 최소화

Experiment

※Experiment Result

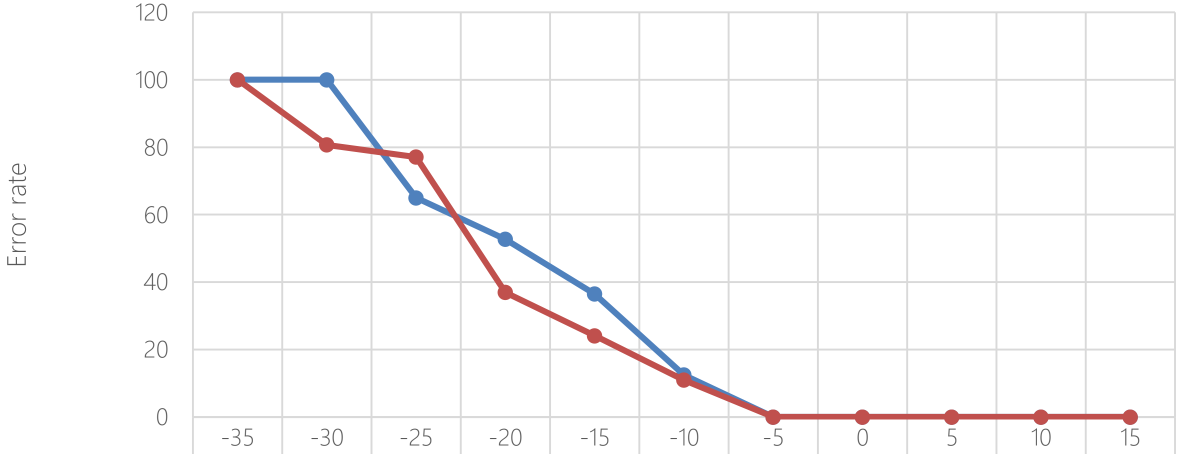
Error rate in white noise



SNR[dB]	-10	-5	0	5	10	15	20
GCC-PHAT	100	98.21428571	78.26086957	32.8358209	16	0	0
HMSL	100	80.35714286	52.17391304	26.86567164	16	0	0

SNR[dB]

Error rate in speech noise



SNR[dB]	-35	-30	-25	-20	-15	-10	-5	0	5	10	15
GCC-PHAT	100	100	65	52.7174	36.4964	12.5	0	0	0	0	0
HMSL	100	80.7018	77.0833	36.9565	24.0876	10.9375	0	0	0	0	0

SNR[dB]

※Discussion and Future Research Plan

[정성적 결과]

- ❖ Automatically Binary Masking(Source Coordinate)
- ❖ Automatically selecting source for learning and Automatically selecting source for separation
- ❖ Automatically finding Target Source

→ Blind Source Localization

[정량적 결과]

- ❖ SNR이 0dB이하에서 최대 약 20% 정확도 증가

Future Research Plan

Common

- ❖ Algorithm Parameter Optimizing
- ❖ Same / Similar Direction(depth) Idea
- ❖ 다양한 실험환경 및 파라메터에서의 테스트(거리, 실내/실외, 각도)
- ❖ 다양한 testset test를 통한 알고리즘 성능 검증 및 일반화
- ❖ 다른 Clustering(GMM, SVDD, etc.)기법과 Separation(ICA)기법들과의 비교분석 ; 최적의 조합 test
- ❖ GCC-PHAT뿐만 아니라 최근 SCI논문에서의 state of art method들과의 비교분석

Directional Clustering and Binary Masking	*Nonnegative Matrix Factorization*
❖ Target component 손실 최소화를 위한 개선된 Mask, Restoration Idea	❖ Complex number로 구성된 행렬을 분해할 수 있는 알고리즘 구현 ❖ 행렬 분해시 penalty term등 bias와 variance 조절