Underdetermined Blind Source Separation using Normalized Spatial Covariance Matrix and Multichannel Nonnegative Matrix Factorization

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)

Sound Source Separation

※ Supervised vs Unsupervised

(1)

Sound Source Separation / Audio Source Separation

Supervised Solutions

- Prior information
- Specific
- Learning
- Low robust
- Deep Learning(input= just waveform data or additive visual data)

Unsupervised Solutions
BSS(Blind Source Separation)

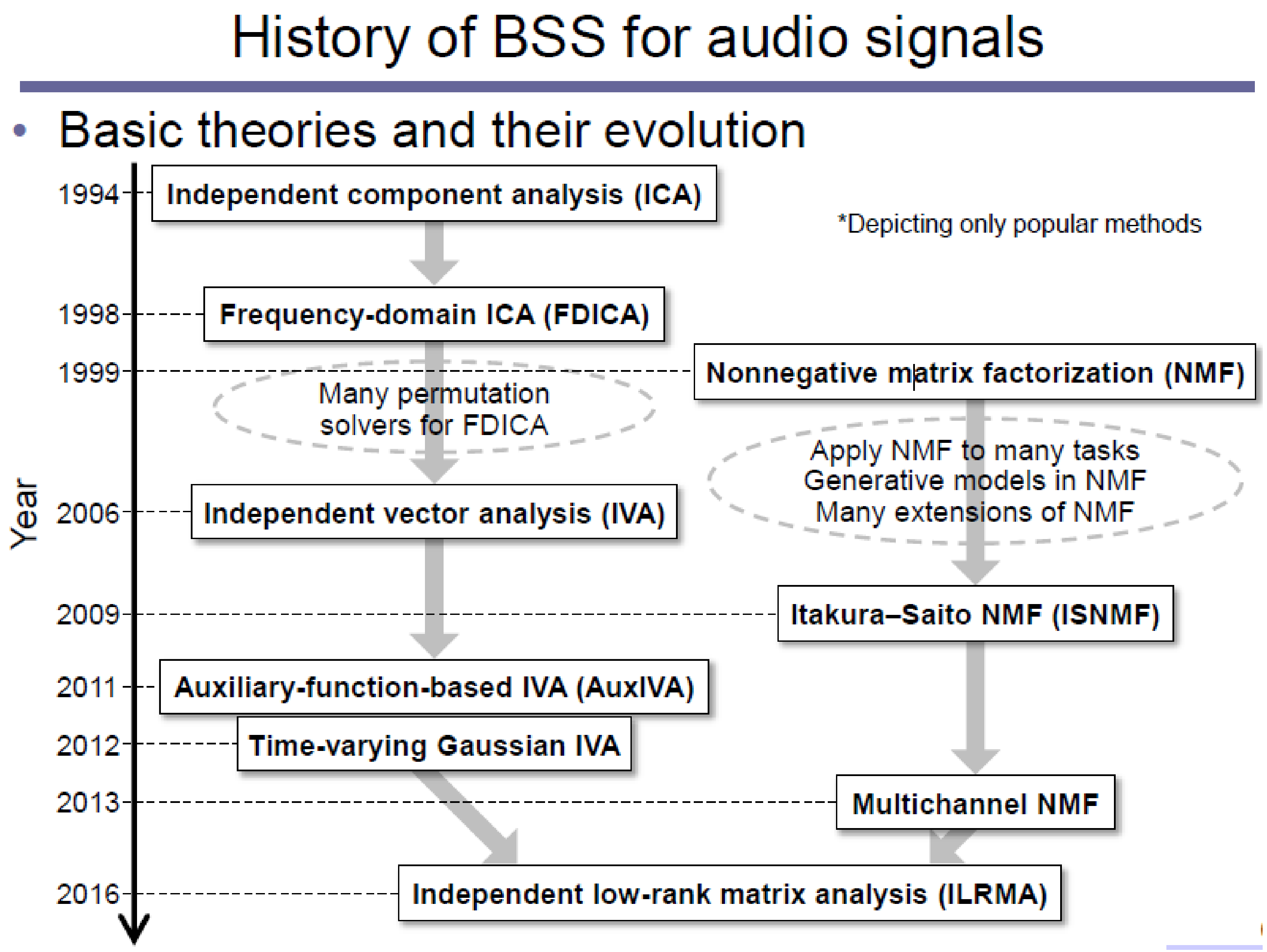
- without any information about the recording environment, mixing system, or source locations
- Realistic
- Practical

- Undetermined mixing systems
- Reverberant environments
- Presence of noise
- Non-stationary of speech

- Blind Source Separation (BSS) is a technique for separating specific sources from a recorded sound without any information about the recording environment, mixing system, or source locations.
- Blind Source Separation (BSS) is an approach for estimating source signals that uses only the mixed signal information observed at each microphone.

※ History of Blind Source Separation

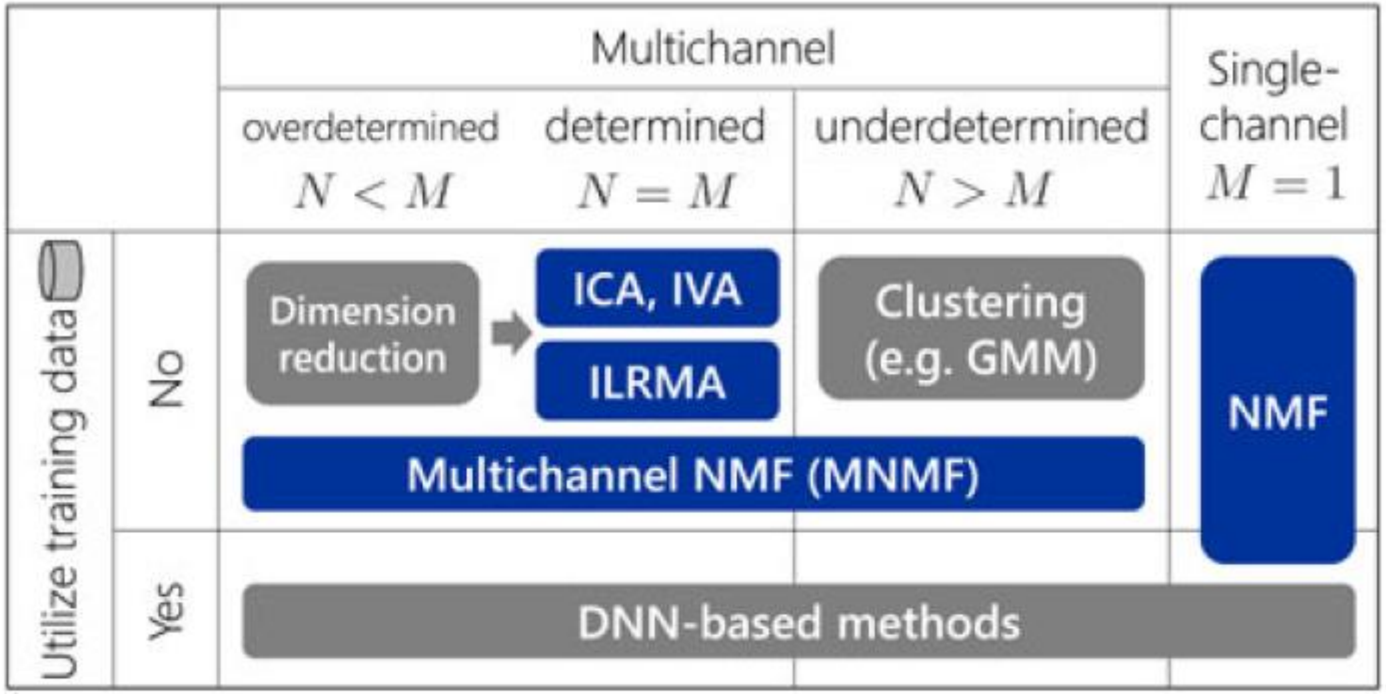
(1)



Blind Source Separation

※ Division of Blind Source Separation techniques

(1)



Blind Source Separation

※ Division of Blind Source Separation techniques

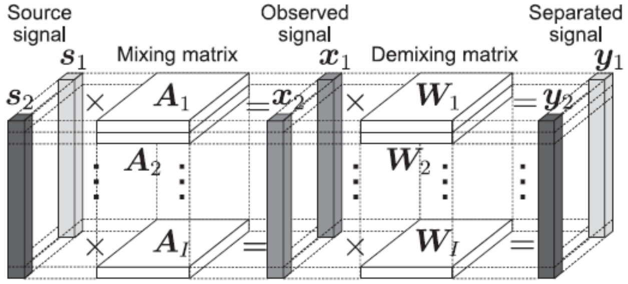


Fig. 1. Conceptual model of IVA ($N = M = 2$).

Assumption : Statistical independence between the sources to estimate a demixing matrix

ICA(Independent Component Analysis)

(Over-)determined
$N \leq M$

BSS

underdetermined
$N > M$

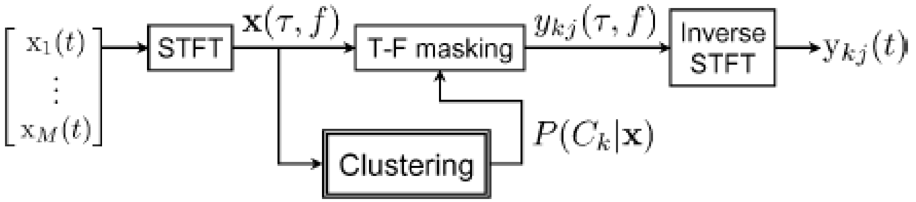
Sparseness based approaches

Assumption : Most one source is dominant at each time-frequency slot

Maximum A Posteriori(MAP) estimation

$$P(O|\theta)$$

Time-frequency binary mask



Nonnegative matrix Factorization(NMF)

Singlechannel NMF

Multichannel NMF

Spatial Covariance Model

Beamforming Algorithm

Multichannel Nonnegative Matrix Factorization Representation

※Cluster NMF bases according to the source location

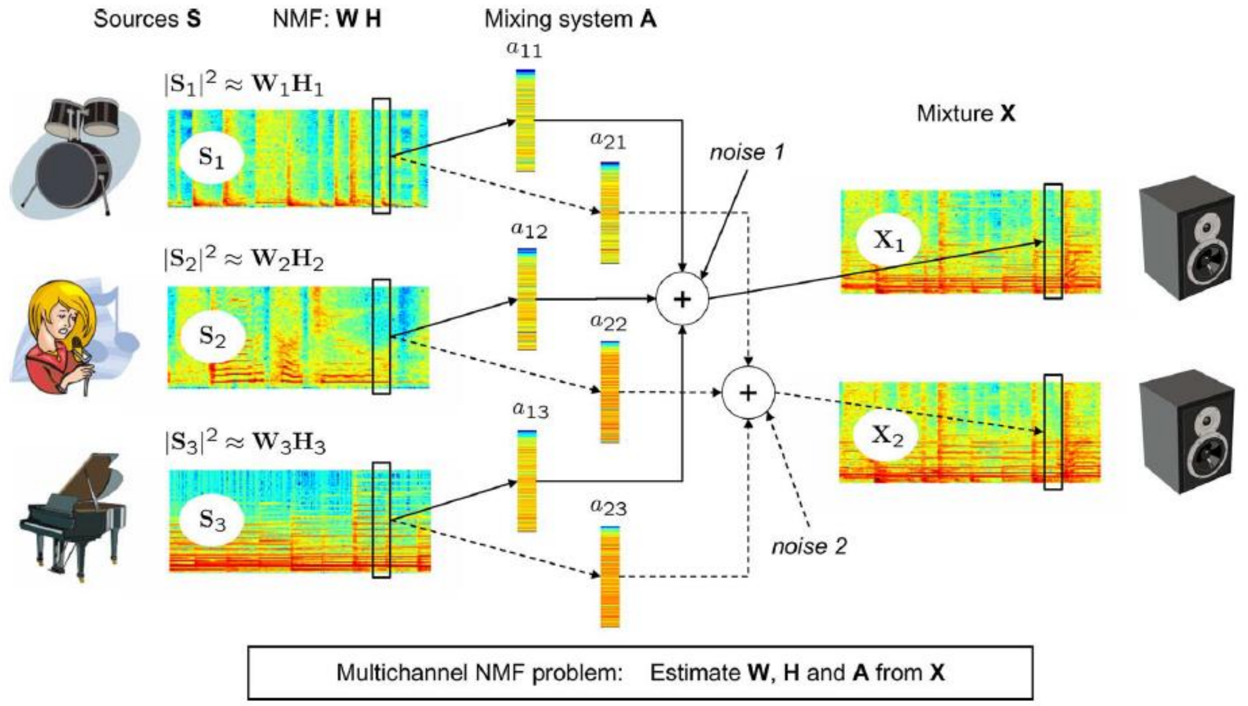


Fig. 1. Representation of convolutive mixing system and formulation of Multichannel NMF problem.

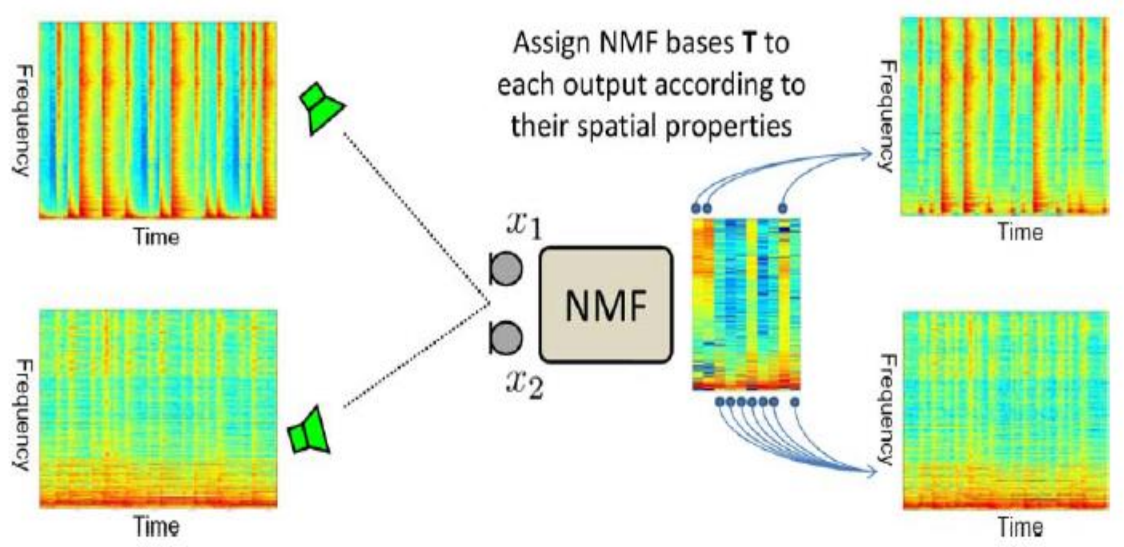
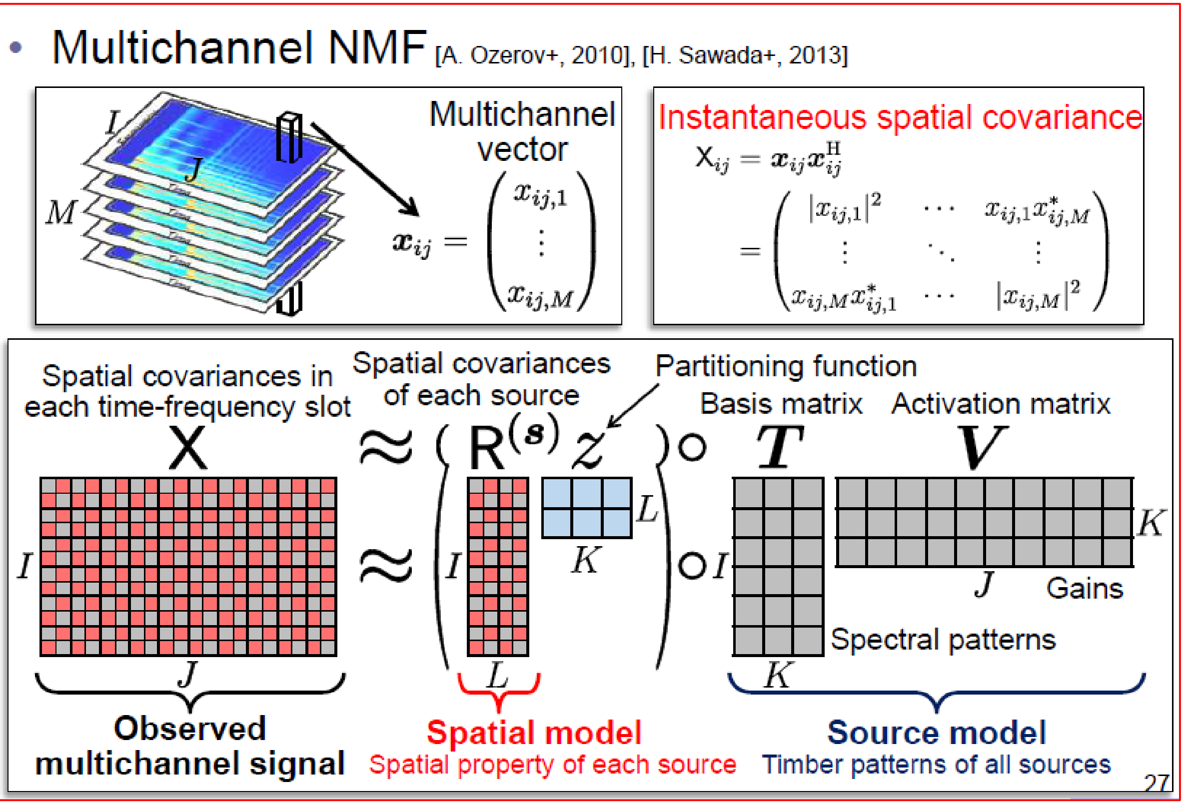


Fig. 2. Multichannel extensions of NMF associate the spatial property with each NMF basis. This enables us to cluster NMF bases according to the source location, and thus perform a source separation task.

- *Joint estimation of the source parameter and mixing coefficients*

Decomposition and Formulation of MNMF

※Multiplicative Update Spatial Model and Source Model



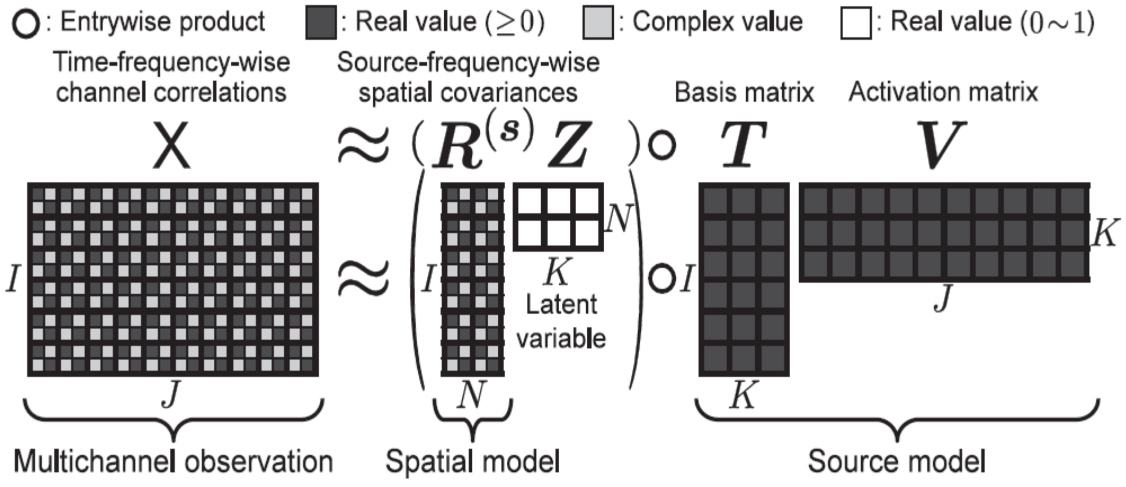


Fig. 3. Decomposition model of Sawada's MNMF ($I=6$, $J=10$, $M=N=2$, and $K=3$).

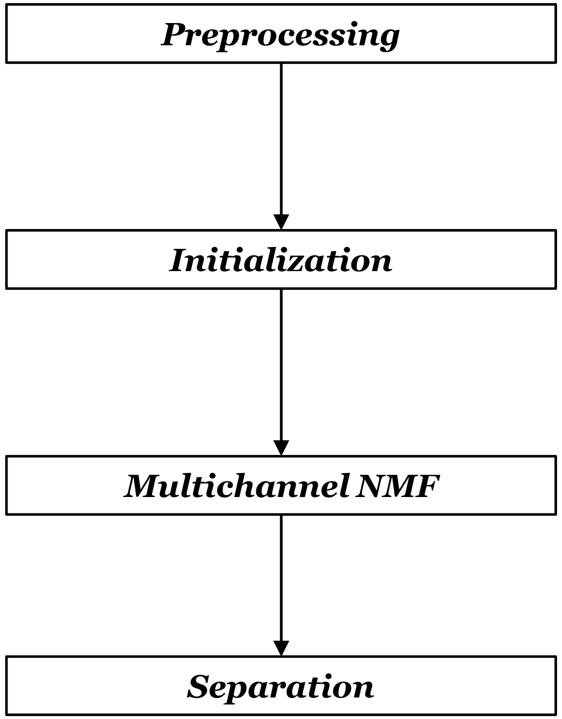
$$\mathsf{X}_{ij} = \boldsymbol{x}_{ij}\boldsymbol{x}_{ij}^{\mathrm{h}},$$

$$\mathsf{X}_{ij} \approx \hat{\mathsf{X}}_{ij} = \sum_k \left(\sum_n R_{i,n}^{(s)} z_{nk} \right) t_{ik} v_{kj},$$

$$\begin{aligned} x_{ij,1} &= a + bi = |x_{ij,1}| e^{i\theta_1} \\ x_{ij,2} &= c + di = |x_{ij,2}| e^{i\theta_2} \\ x_{ij,1} {x^*}_{ij,2} &= |x_{ij,1}|\, e^{i\theta_1} \times |x_{ij,2}| e^{-i\theta} \\ &= |x_{ij,1}|\, |x_{ij,2}|\, e^{i(\theta_1 - \theta_2)} \end{aligned}$$

Whole Process of Swada's MNMF

※Bottom-up clustering



i) 20 iterations to update $\mathbf{T}$ and $\mathbf{V}$.
ii) 200 iterations to update $\mathbf{T}$, $\mathbf{V}$, $\mathbf{H}$ and $\mathbf{Z}$ by the top-down approach with $L = L_{init} = 9$.
iii) Bottom-up clustering with $interval = 10$ until $L = 3$.
iv) 200 iterations to update $\mathbf{T}$, $\mathbf{V}$, $\mathbf{H}$ and $\mathbf{Z}$ by the top-down approach with $L = 3$.

Algorithm 1 Multichannel NMF with bottom-up clustering

Procedure MchNMF_BottomUpClustering
  $iteration \leftarrow 0$
  While $L > finalClusterSize$ **do**
    $iteration \leftarrow iteration + 1$
    update $\mathbf{T}$ by (42) or (48)
    update $\mathbf{V}$ by (43) or (49)
    If $mod(iteration, interval) = 1$ **then**
      $(\mathbf{H}, \mathbf{Z}) \leftarrow Pairwisemerge(\mathbf{H}, \mathbf{Z})$
      $L \leftarrow L - 1$
    end if
    update $\mathbf{H}$ by (45) or (51)
    update $\mathbf{Z}$ by (44) or (50)
  end while
end procedure

Procedure $PairwiseMerge(\mathbf{H}, \mathbf{Z})$
  $(l_1, l_2) \leftarrow findPair(\mathbf{H})$
  $w_1 \leftarrow \sum_k z_{l_1 k}$
  $w_2 \leftarrow \sum_k z_{l_2 k}$
  $\{\mathsf{H}_1, \ldots, \mathsf{H}_I\} \leftarrow weightedMean(\mathbf{H}, l_1, l_2, w_1, w_2)$
  $\mathbf{H} \leftarrow removeAdd(\mathbf{H}, l_1, l_2, \{\mathsf{H}_1, \ldots, \mathsf{H}_I\})$
  $\mathbf{Z} \leftarrow merge(\mathbf{Z}, l_1, l_2)$
end procedure

Spatial Covariance Matrix Formulation

※Example of H matrix

$$X_{ij} = \tilde{\boldsymbol{x}}_{ij}\,\tilde{\boldsymbol{x}}^{H}{}_{ij} = \begin{pmatrix} |\tilde{x}_{ij,1}|^2 & \tilde{x}_{ij,1}\tilde{x}^*{}_{ij,2} \\ \tilde{x}_{ij,2}\tilde{x}^*{}_{ij,1} & |\tilde{x}_{ij,2}|^2 \end{pmatrix} = \begin{pmatrix} |\tilde{x}_{ij,1}|^2 & |x_{ij,1}|\,|x_{ij,2}|\,e^{i(\theta_1-\theta_2)} \\ |x_{ij,1}|\,|x_{ij,2}|\,e^{i(\theta_2-\theta_1)} & |\tilde{x}_{ij,2}|^2 \end{pmatrix} \approx \hat{X}_{ij} = \sum_k \left(\sum_l H_{il} z_{lk}\right) t_{ik} v_{kj}$$

$$\boxed{H_{il} = \begin{pmatrix} A & Ce^{i(\theta_1-\theta_2)} \\ Ce^{i(\theta_2-\theta_1)} & B \end{pmatrix}}$$

Diagonal term：Power Gain

Off − Diagonal term：Phase difference

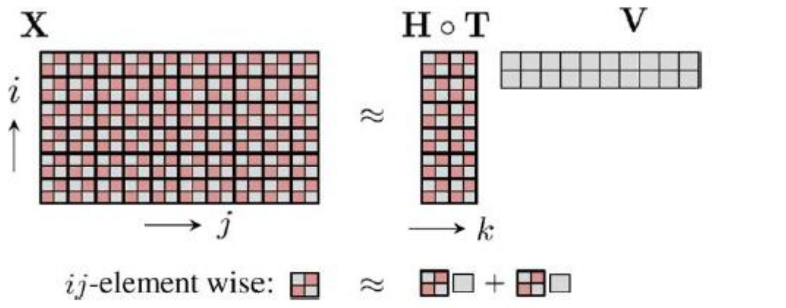


Fig. 5. Illustrative example of multichannel NMF: $I = 6$, $J = 10$, $K = 2$, $M = 2$. Non-negative values are shown in gray and complex values are shown in red.

$$\hat{x}_{ij} = |\tilde{x}_{ij,1}|^2 = \sum_k t_{ik} v_{kj}$$

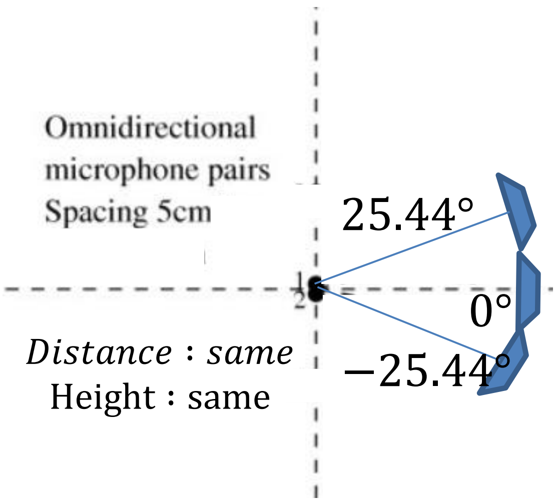
where t_{ik} and v_{kj} are non-negative scalars as in the single-channel case. To solve the scaling ambiguity between H_{ik} and t_{ik}, let H_{ik} have a unit trace $\mathrm{tr}(\mathsf{H}_{ik}) = 1$.

$$\boxed{H_{il} = \begin{pmatrix} 1 & \dfrac{|x_{ij,2}|}{|x_{ij,1}|} e^{i(\theta_1-\theta_2)} \\ \dfrac{|x_{ij,2}|}{|x_{ij,1}|} e^{i(\theta_2-\theta_1)} & \dfrac{|\tilde{x}_{ij,2}|^2}{|\tilde{x}_{ij,1}|^2} \end{pmatrix}}$$

Simulation

※Single Frequency Sine Wave Generation

Conditions	Value
Sampling Frequency	16kHz
Frame length	1024 samples(64ms)
Shifting length	256samples(16ms)
Window	Hanning
Delay sample	1sample$(62.5\mu s)$
Sound velocity	343.7 $\mathrm{m}/s(20℃)$



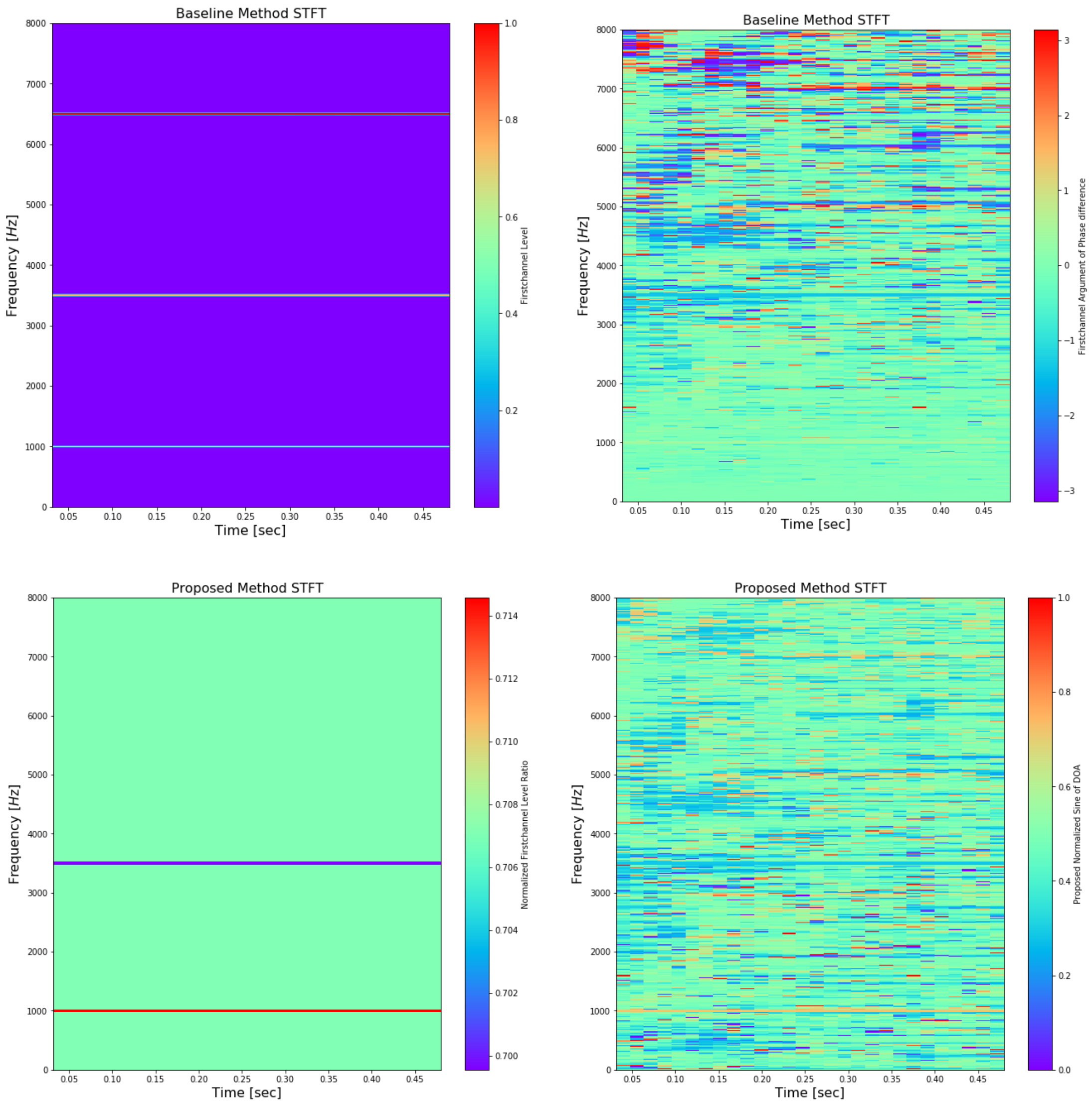
Simulation

※Input source Formulation

$$S1 = 5 \times \sin(2\pi \times 1000 \times t)$$
$$S2 = 6 \times \sin(2\pi \times 3500 \times t)$$
$$S3 = 7 \times \sin(2\pi \times 6500 \times t)$$

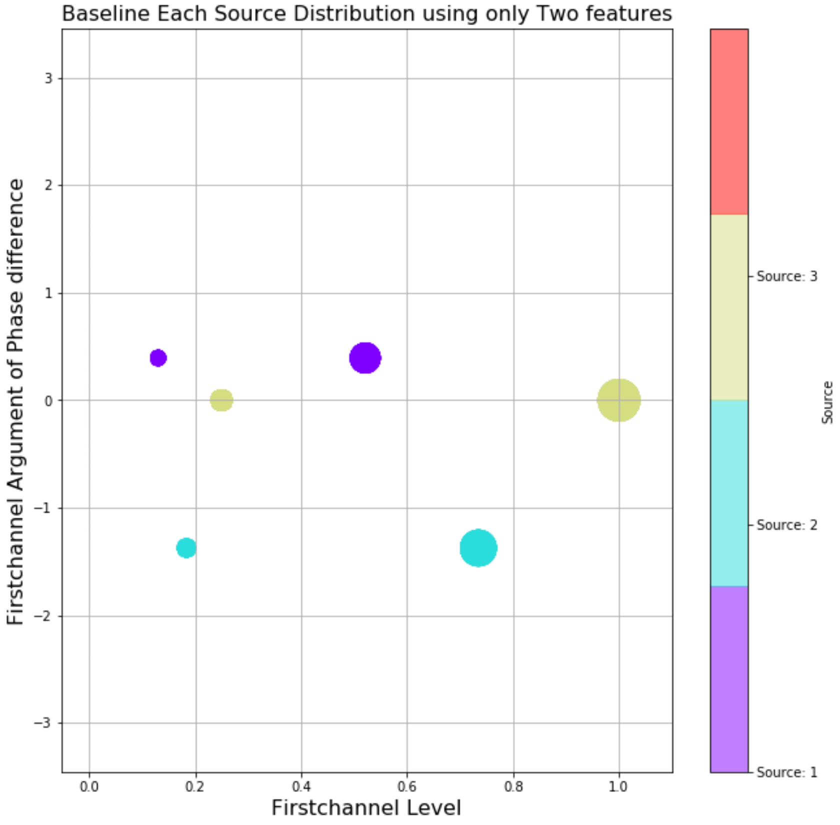
$$\sin\theta = \frac{\Delta l}{d} = \frac{c \times \Delta t}{d} = \frac{c \times N}{d \times f_s}$$
$$\theta = arcsin\left(\frac{c \times N}{d \times f_s}\right)$$

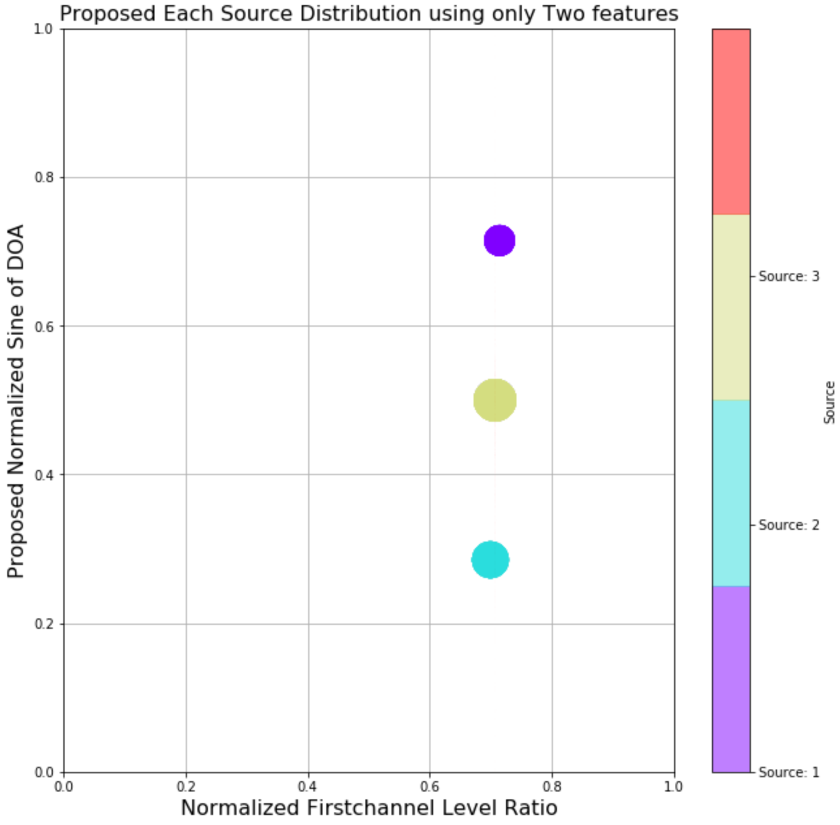
$$\theta_{S1} = arcsin\left(\frac{343.7 \times 1}{0.05 \times 16000}\right) \cong 25.44°$$
$$\theta_{S2} = arcsin\left(\frac{343.7 \times (-1)}{0.05 \times 16000}\right) \cong -25.44°$$
$$\theta_{S3} = arcsin\left(\frac{343.7 \times 0}{0.05 \times 16000}\right) \cong 0°$$



STFT Components Distribution by Frequencies and Sources

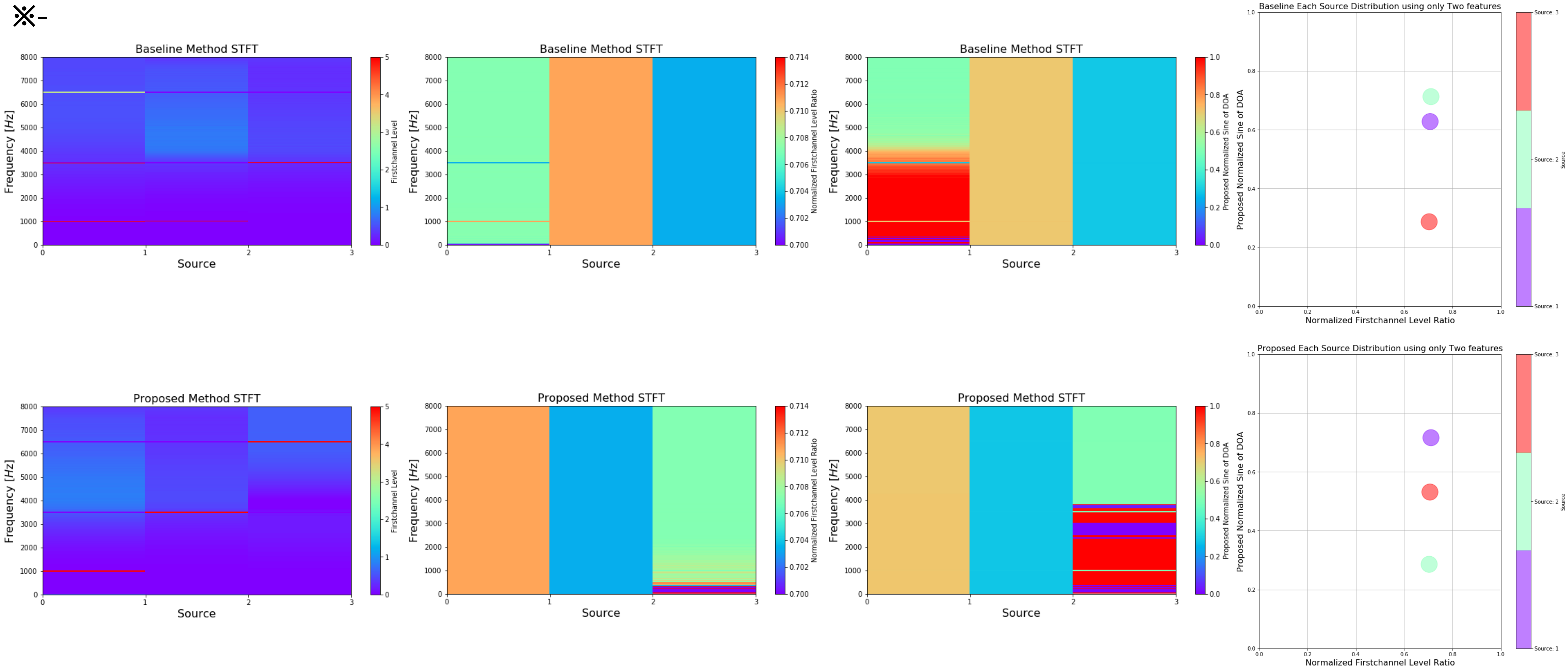
※Using Only Two Features for Visualizing Source Distribution





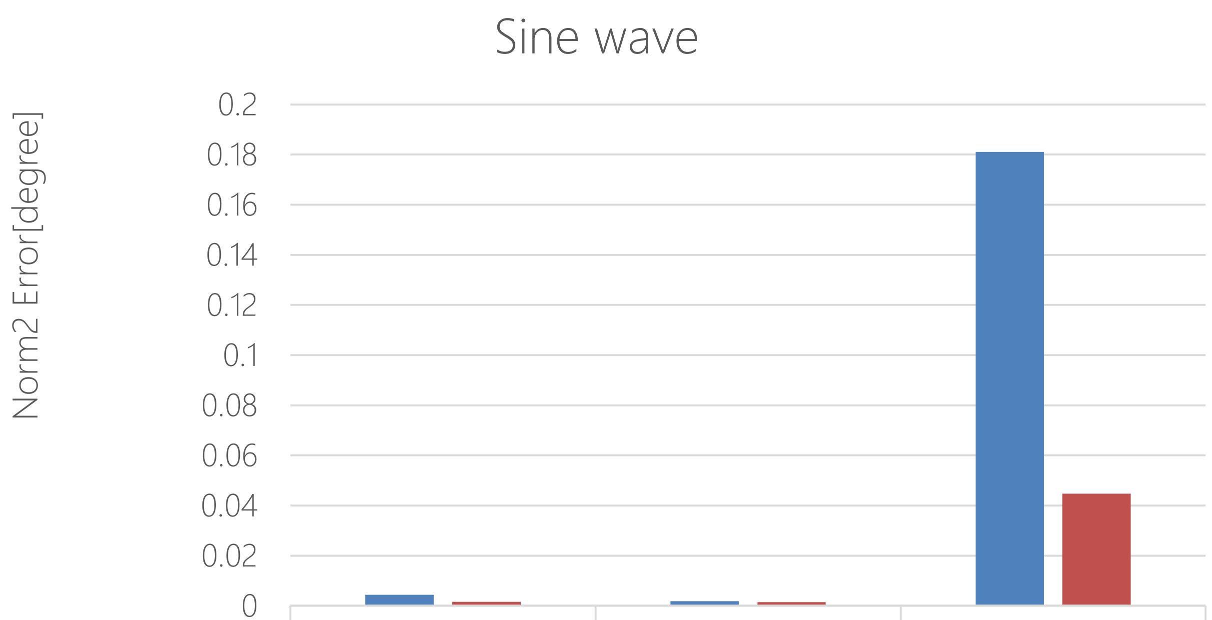
APPENDIX

※-


Baseline Method STFT
Frequency [Hz]
Source
Firstchannel Level
Normalized Firstchannel Level Ratio
Proposed Normalized Sine of DOA
Baseline Each Source Distribution using only Two features
Normalized Firstchannel Level Ratio
Source: 1
Source: 2
Source: 3
Proposed Method STFT
Proposed Each Source Distribution using only Two features

Simulation Result

※Spatial Covariance Matrix for Distance



	Source 1	Source 2	Source 3
Baseline	0.004364181	0.001799721	0.181073968
Proposed method	0.001555022	0.001430689	0.044645296

SiSEC dataset Analysis

※데이터 개요 및 분석

- Under-determined speech and music mixtures
- Determined and over-determined speech and music mixtures
- Head-geometry mixtures of two speech sources in real environments, impinging from many directions
- Professionally produced music recordings

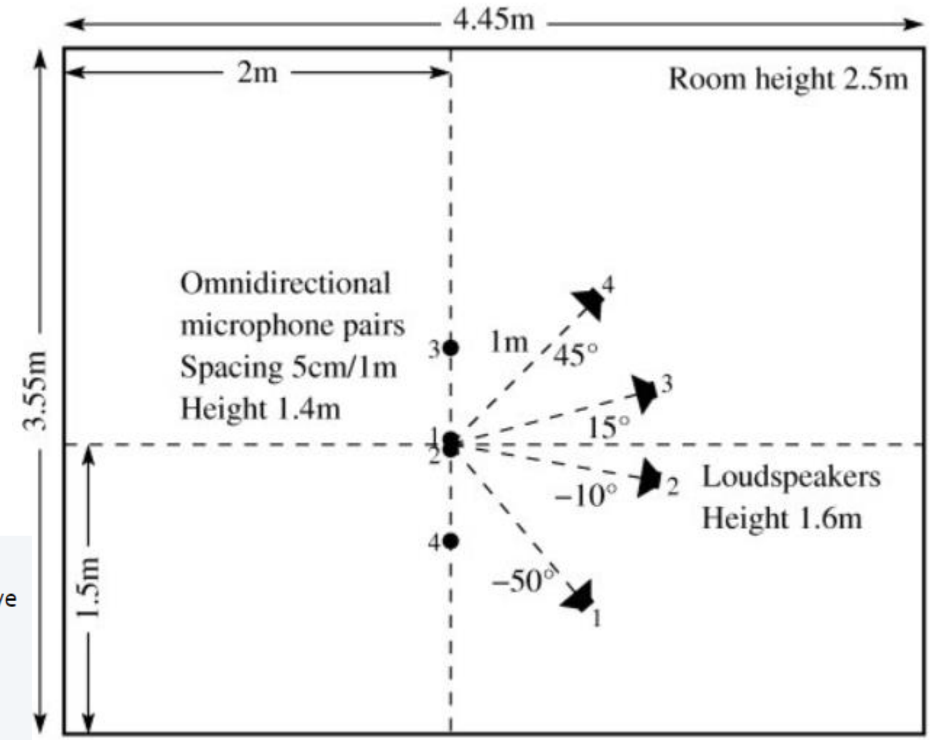
(1)

- **instantaneous mixtures** (static sources scaled by positive gains)
- **synthetic convolutive mixtures** (static sources filtered by synthetic room impulse responses simulating a pair of omnidirectional microphones via the Roomsim toolbox)
- **live recordings** (static sources played through loudspeakers in a meeting room, recorded one at a time by a pair of omnidirectional microphones and subsequently added together)

- 4 male speech sources
- 4 female speech sources
- 3 male speech sources
- 3 female speech sources
- 3 non-percussive music sources
- 3 music sources including drums

TABLE I
EXPERIMENTAL CONDITIONS

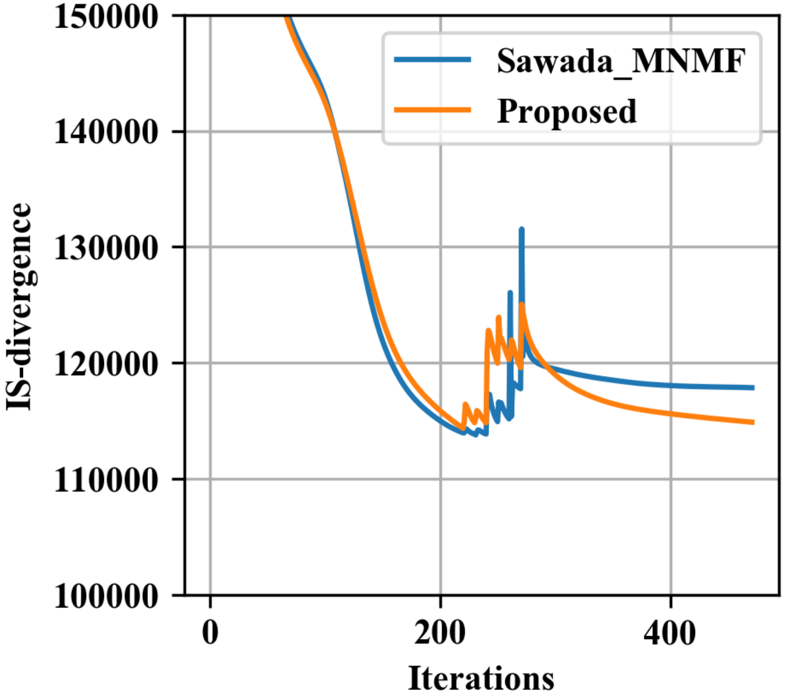
Sampling rate	16 kHz
Frame length	
Frame shift	
Window function	Hanning
Signal length	10 s
Mixture signal ($N = 2$)	
Reverberation time (RT_{60})	130 ms/250 ms
Microphone spacing	5 cm/1 m

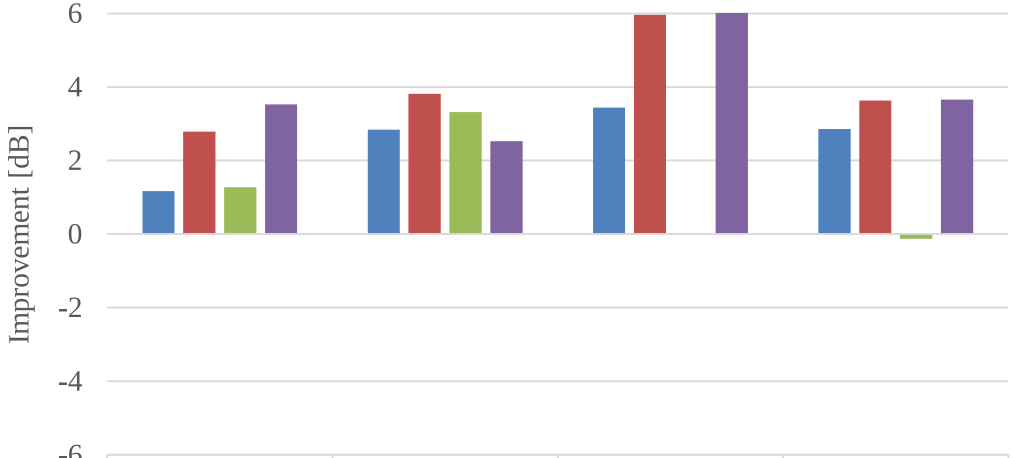


1. **source counting** (estimate the number of sources)
2. **mixing system estimation** (estimate the mixing matrix for instantantaneous mixtures or the frequency-dependent mixing matrix for convolutive mixtures)
3. **source signal estimation** (estimate the mono source signals)
4. **source spatial image estimation** (estimate the stereo contribution of each source to the two mixture channels)

Experiment Result

※SiSEC 2008 Dev1 liverec 250msec





	Male	Female	nodrum	wdrum
SDR	1.170228604	2.83994793	3.437905294	2.853895946
ISR	2.785891205	3.813666789	5.962582168	3.634412361
SIR	1.274801204	3.31387163	0.014030228	-0.125796678
SAR	3.527648419	2.523068138	6.159634155	3.655837049