ー学術フロンティア講義 サイバネティクス入門ー

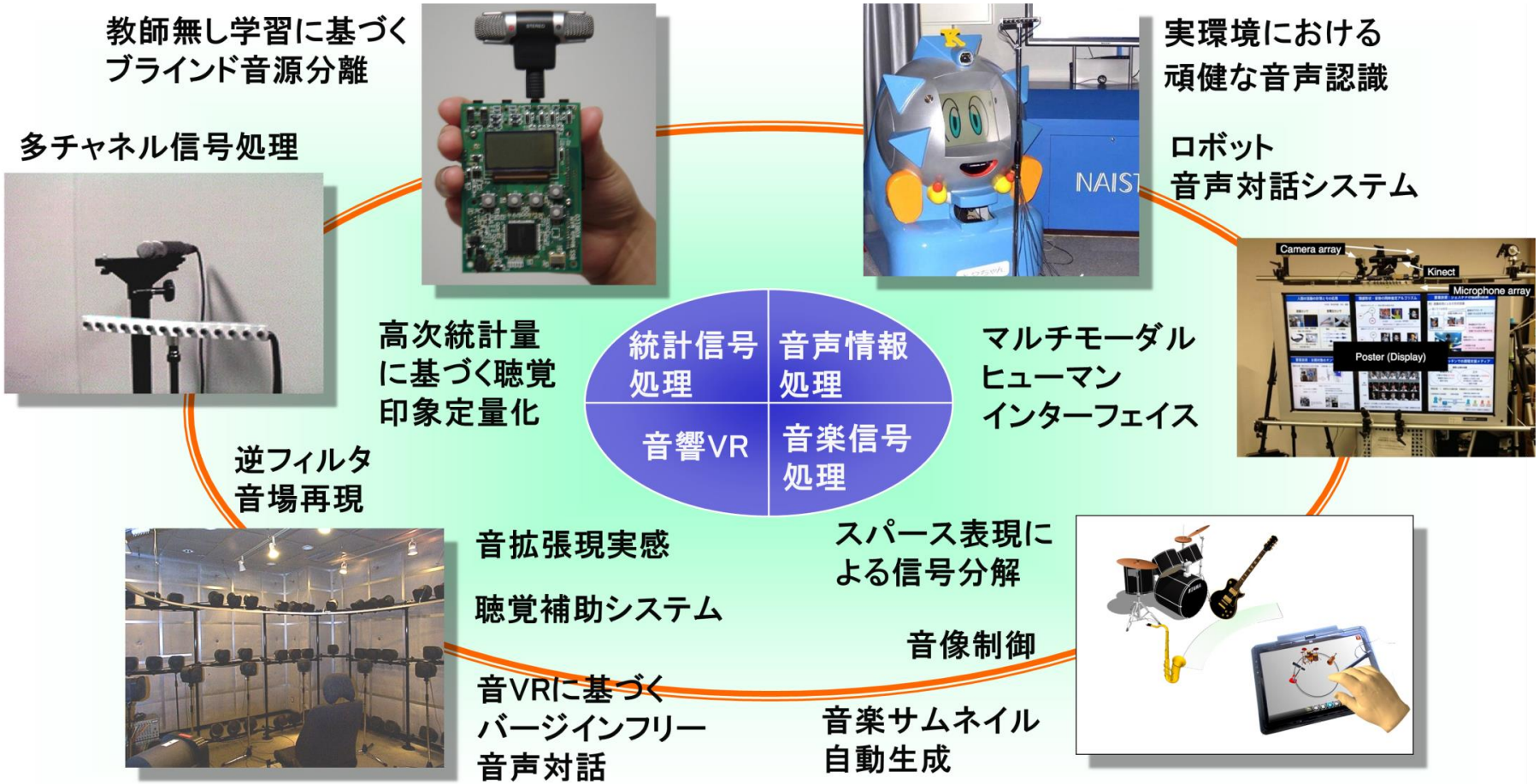
音を解析・合成する信号処理技術

東京大学 大学院情報理工学系研究科・工学部計数工学科
システム情報学専攻 第一研究室 講師
小山 翔一

計数工学科 システム情報第一研究室の紹介

- 音声・音響・音楽メディアに関する信号処理・情報処理
- ヒューマンインタフェース・コミュニケーションシステムの構築
- 統計的・機械学習論的信号処理，数理最適化問題等を研究

研究室ウェブサイト：http://www.sp.ipc.i.u-tokyo.ac.jp



信号処理とは？

- IEEE Signal Processing Societyによる信号処理の紹介動画

https://www.youtube.com/watch?v=EErkgr1MWw0

信号処理とは？

- IEEE (米国電気電子学会) は，アメリカ合衆国に本部を置く，電気・情報工学分野における世界最大規模の学術研究団体
- その中で，信号処理分野のコミュニティであるSignal Processing Societyは最初 (1948年) に設立され，現在は4番目に大きいソサイエティ（Computer, Power and Energy, Communications, Signal Processing）

https://signalprocessingsociety.org/

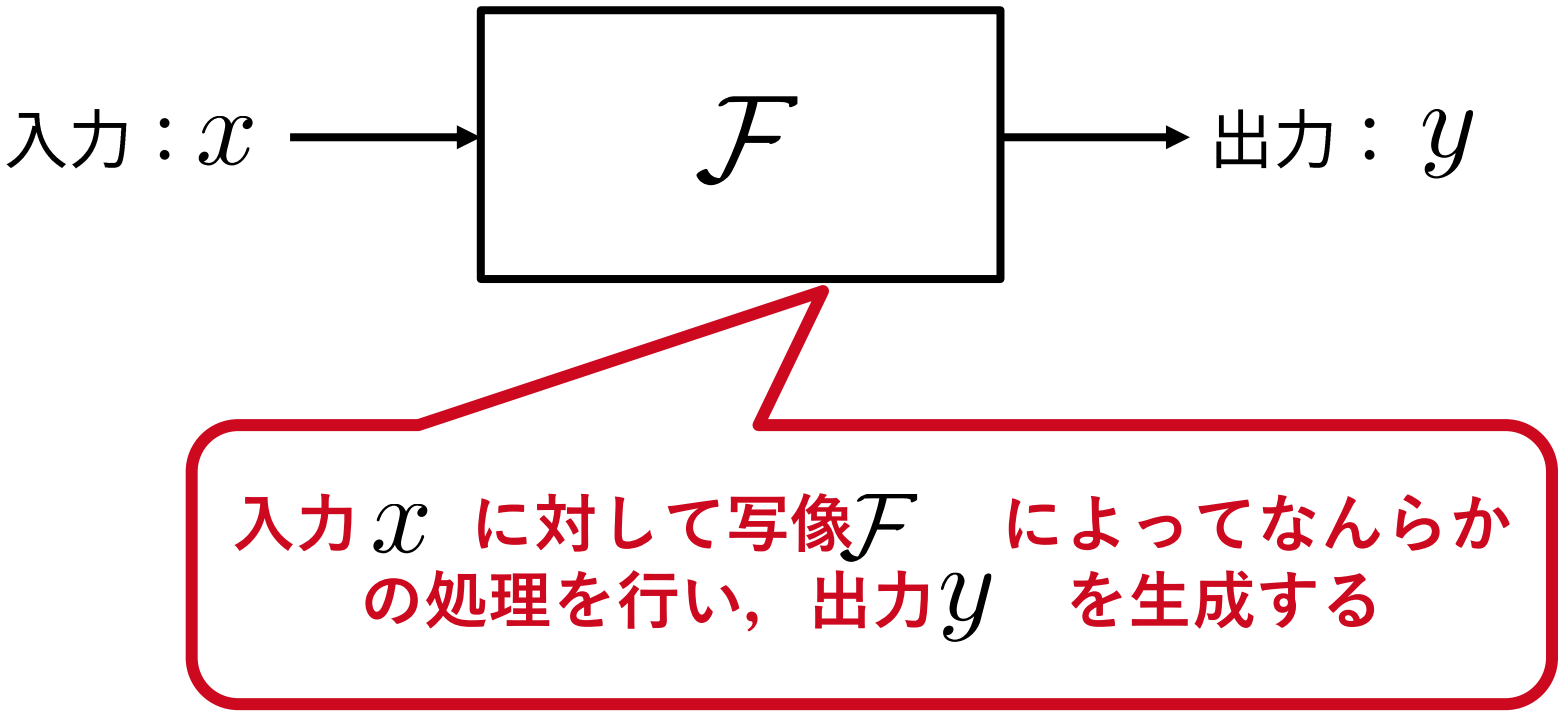
信号処理とは？

- IEEE Signal Processing Society – Technical Committees
  - Audio and Acoustic Signal Processing
  - Bio Imaging and Signal Processing
  - Computational Imaging
  - Design and Implementation of Signal Processing Systems
  - Image, Video, and Multidimensional Signal Processing
  - Industry DSP Technology
  - Information Forensics and Security
  - Machine Learning for Signal Processing
  - Multimedia Signal Processing
  - Sensor Array and Multichannel
  - Signal Processing for Communications and Networking
  - Signal Processing Theory and Methods
  - Speech and Language Processing

応用分野は多岐にわたる

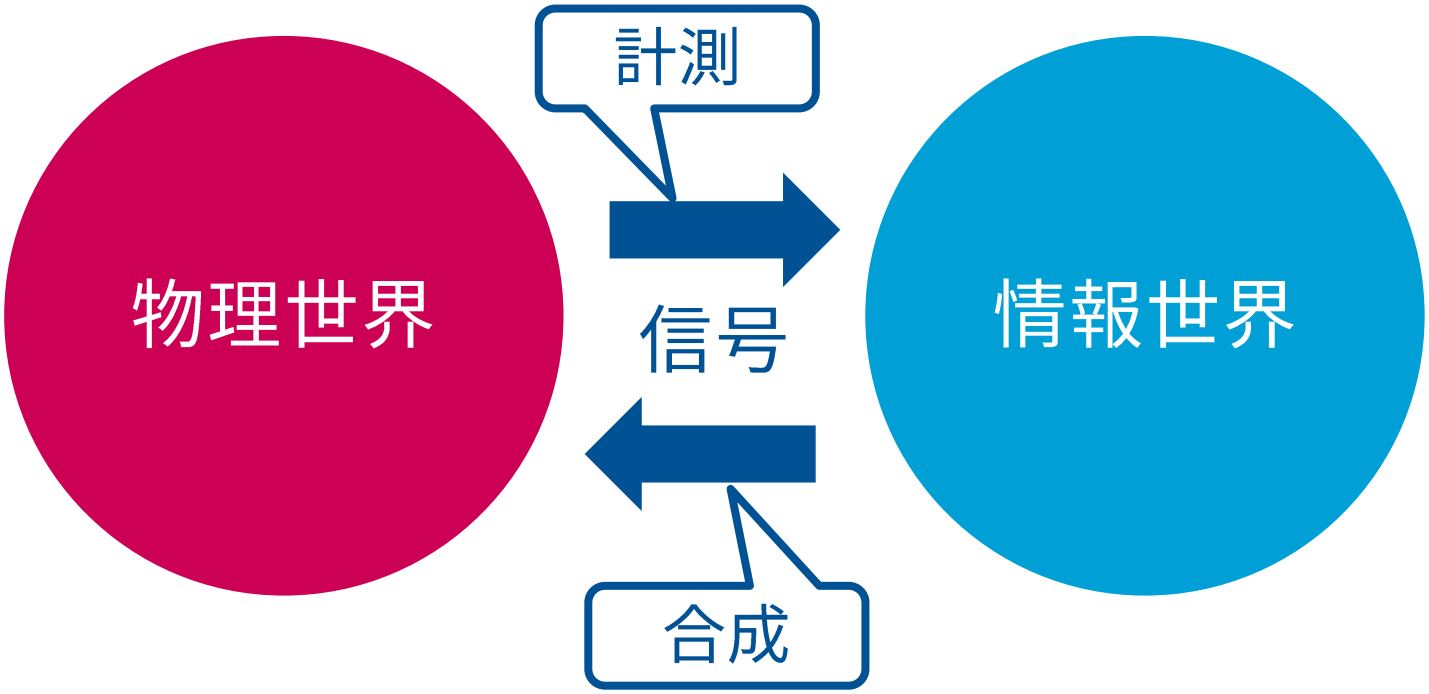
信号処理とは？

- 音声や画像などの信号を，数理的な手法で分析／加工／合成する技術



信号

- 信号とはセンサ等で計測した物理量の時間的／空間的な変動，あるいはそれを記号として表したもの
  - 音声，音楽，画像，動画，超音波ソナー，電波，脳波，筋電位，地震波，株価 etc…



音声・音響・音楽信号処理の例

- 雑音を含む音声を入力し，雑音を除去した音声を出力

➡ 雑音除去

雑音を含む
音声信号

$\mathcal{F}$

雑音が除去された
音声信号

音声や雑音の性質をうまく利用して音声のみを抽出するような処理を行う

音声・音響・音楽信号処理の例

- 音声を入力し，その内容をテキストとして出力

➡ 音声認識

音声信号 → $\mathcal{F}$ → テキスト

音声信号のパターンに対して，対応するテキストを出力するような関数を教師データを用いて学習する

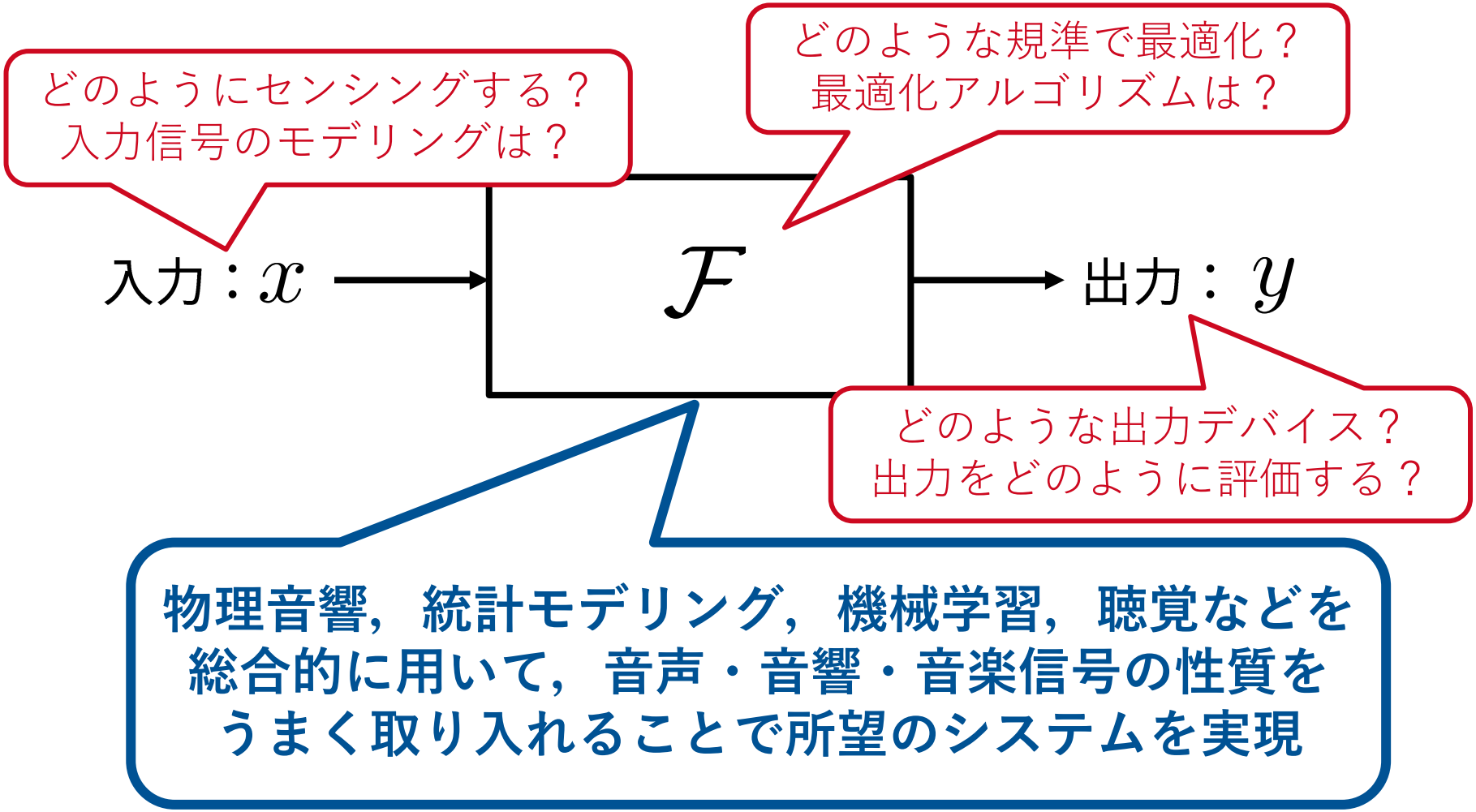
音メディアの重要性

- コミュケーション手段としての音
  - 音声は人間同士の最も原始的なコミュニケーション手段の一つ
  - 遠隔コミュニケーション手段としての電話は，形を変えながらも現在でも広く利用されている（収音・再生，符号化，エコーキャンセラ）
  - ラジオ，テレビ，インターネット放送と変遷している放送メディアとしても広く利用（収音・再生，符号化）
  - 人間とコンピュータ・ロボットがインタラクションするためのインタフェース（音声認識，合成，対話）

- 芸術表現手段としての音
  - 音楽をはじめとして，音を使った芸術表現は多様
  - 音を記録・再生する技術は時代とともに進化（デバイス，記録メディア，符号化）
  - 技術の発展が新たな芸術表現や文化を生み出す例は多数（楽音・歌声合成など）

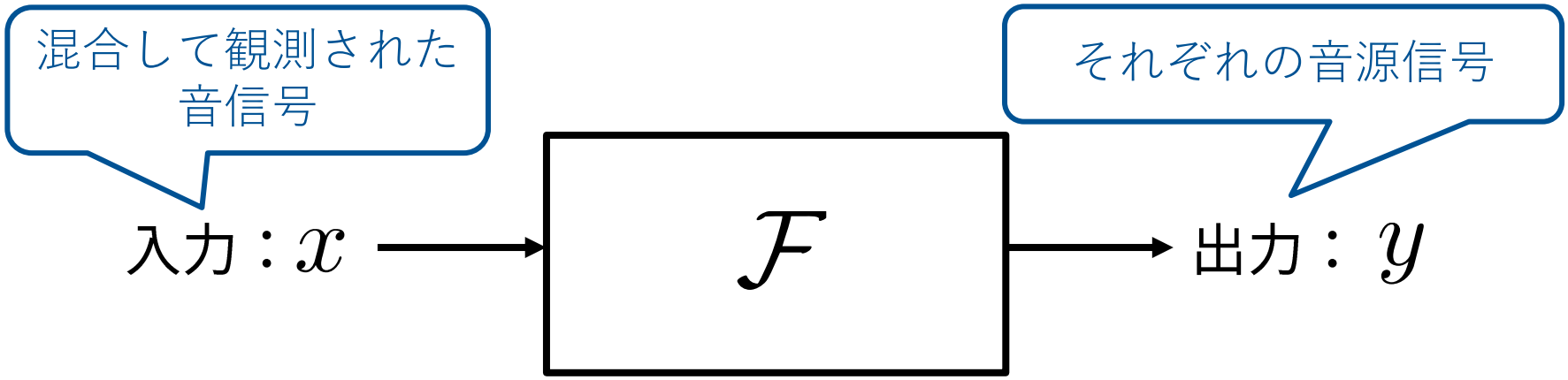
音メディアに関する信号処理研究

- 入力から所望の出力を得るために，どのようにして関数 $\mathcal{F}$ を設計するか？ ➡ 多様なアプローチ！



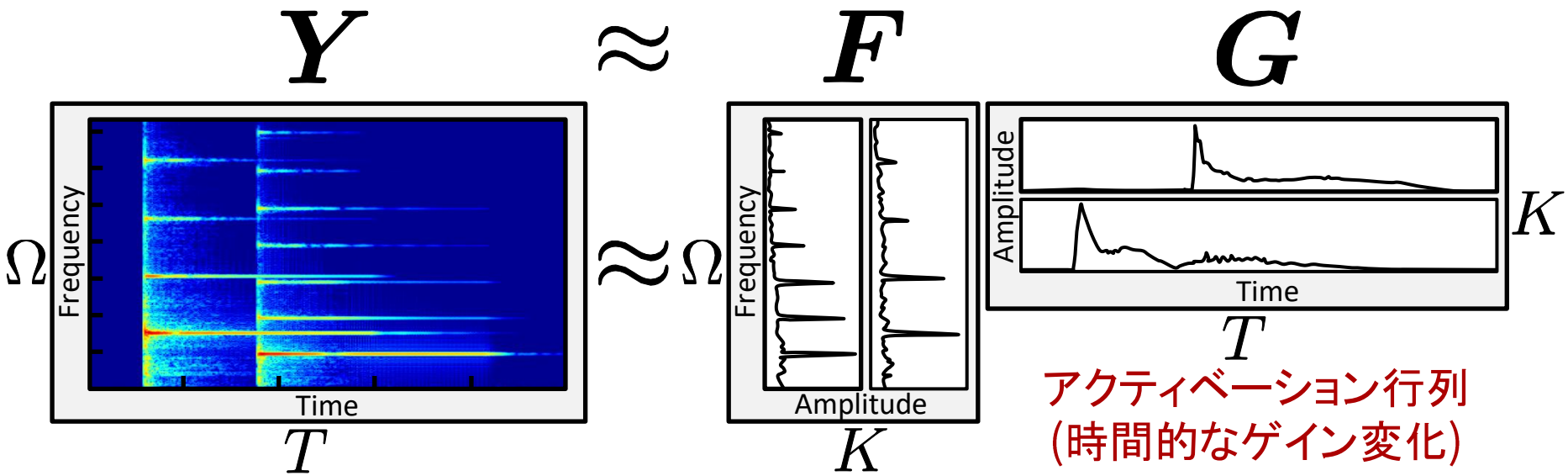
音源分離のための信号処理技術

- 音源分離
  - 混合して観測された音信号をそれぞれの音源信号に分解する
  - 例えば，複数の人が話している音声信号をそれぞれの人の音声に分離したり，複数の楽器が混ざった音をそれぞれの楽器音に分離する



非負値行列因子分解（NMF）による音源分離

- 音源分離の問題における非負値行列因子分解（nonnegative matrix factorization: NMF）
  - 観測データである非負値行列を低ランクな非負値行列の積として表現 [Lee+ 2012]
  - スペクトログラムに対してNMFを適用することで音信号を分離



非負値行列因子分解（NMF）による音源分離

- 多チャネル音楽信号を教師有りNMFで分離した例

原曲

教師１

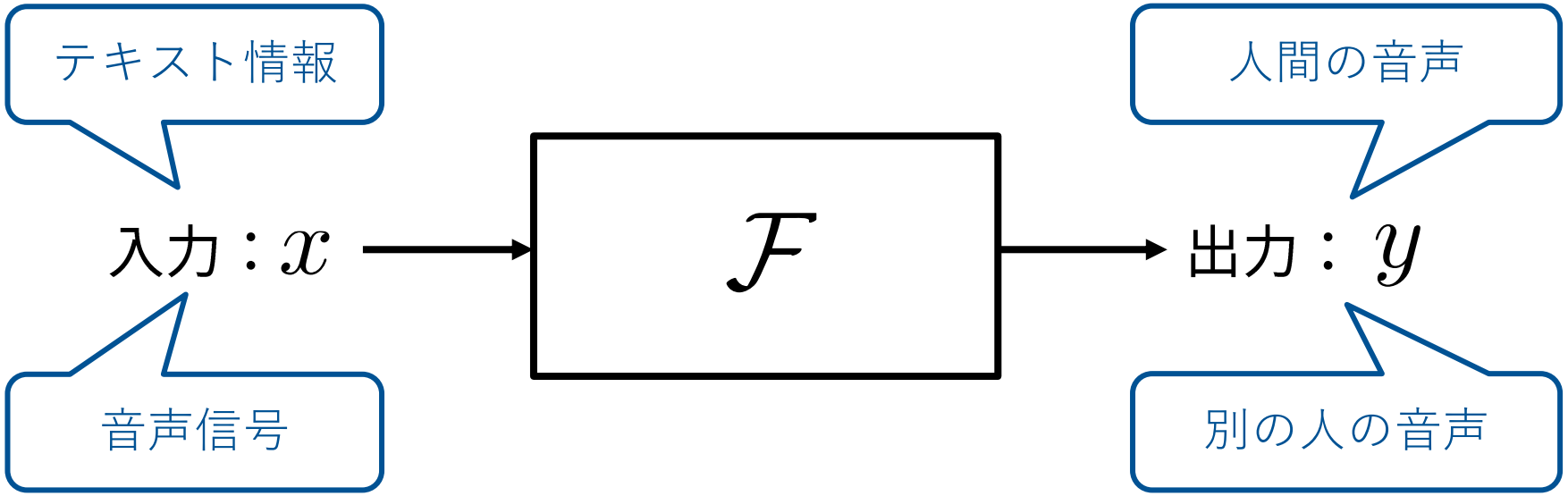
分離音１

教師２

分離音２

音声合成・変換

- 音声合成
  - テキスト情報から人間の音声を合成する
- 音声変換
  - ある人の音声を別の人の音声に変換する



リアルタイム音声変換

他の人の声にリアルタイムでなりきる技術 [Arakawa19]

https://www.youtube.com/watch?v=P9rGqoYnfCg

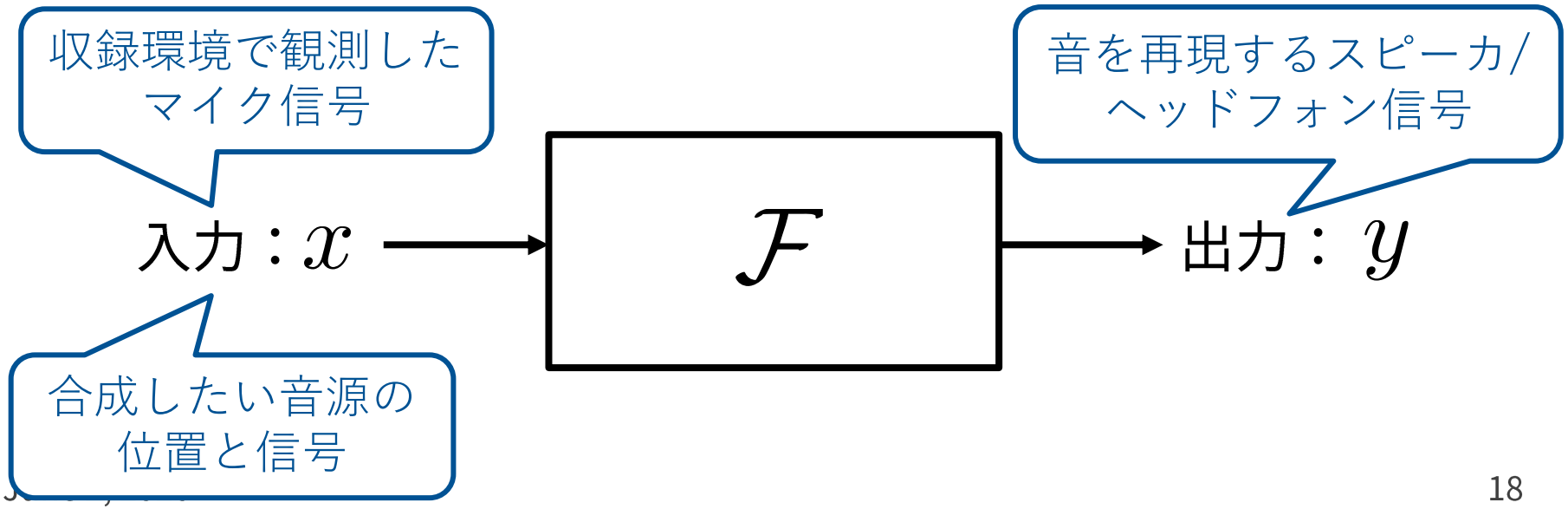
http://www.ytv.co.jp/conan/item/tai.html

音声工学的知見を活かした信号処理・深層学習により
低遅延・高品質変換を実現

空間音響のための信号処理技術

- 空間音響（spatial audio）とは？
  - 音を記録し，再生する場合に，あたかも収録した場所にいるかのように受聴者が知覚するように音を再現する技術
  - あるいは計算機上で模擬した環境にいるかのように受聴者が知覚するように音を合成する技術

➡ 音のVR/ARを実現する基礎技術



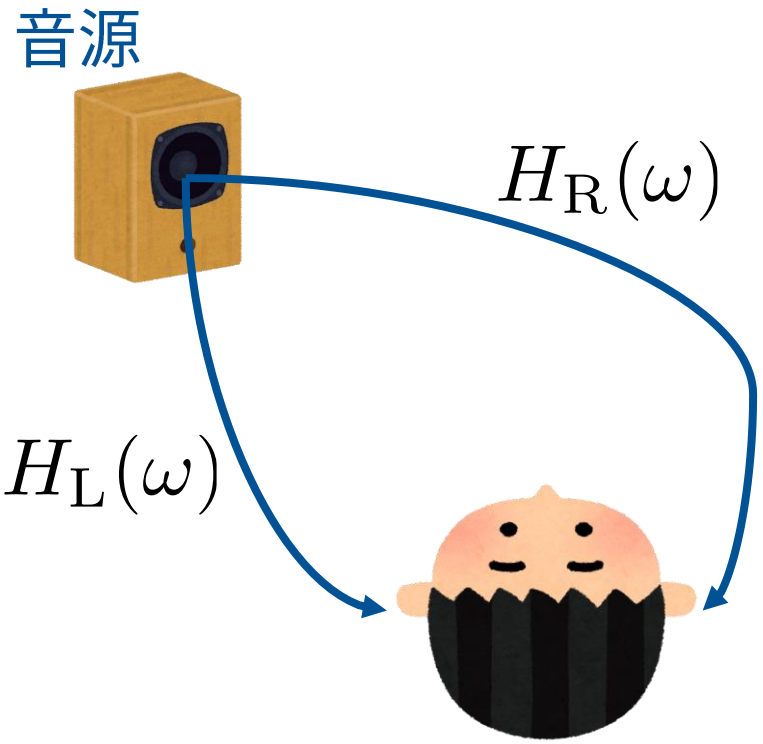
音の空間知覚と音場再現

音の空間知覚

- 人間はどのようにして音を空間的に知覚するのか？
  - 物理空間で音源から発せられた音波が人間の両耳に入力され，耳入力信号の様々な性質から心理空間における音像を知覚する [飯田+ 2010]
    1. 時間的性質：残響感，リズム感，持続感など
    2. 空間的性質：方向感，距離感，広がり感など
    3. 質的性質：大きさ，高さ，音色など
  - 音像の空間的性質（方向感，距離感，広がり感など）を知覚するための重要な役割を持つ物理特性の一つが頭部伝達関数（Head-Related Transfer Function: HRTF）

音の空間知覚

- 頭部伝達関数（HRTF）
  - 音源から放射された音がヒトの鼓膜に到達するまでの伝達特性



音源から外耳道入口/鼓膜までの伝達特性

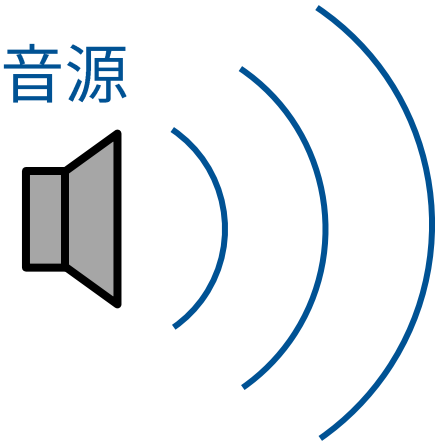
ある音源位置から右耳までのHRTF

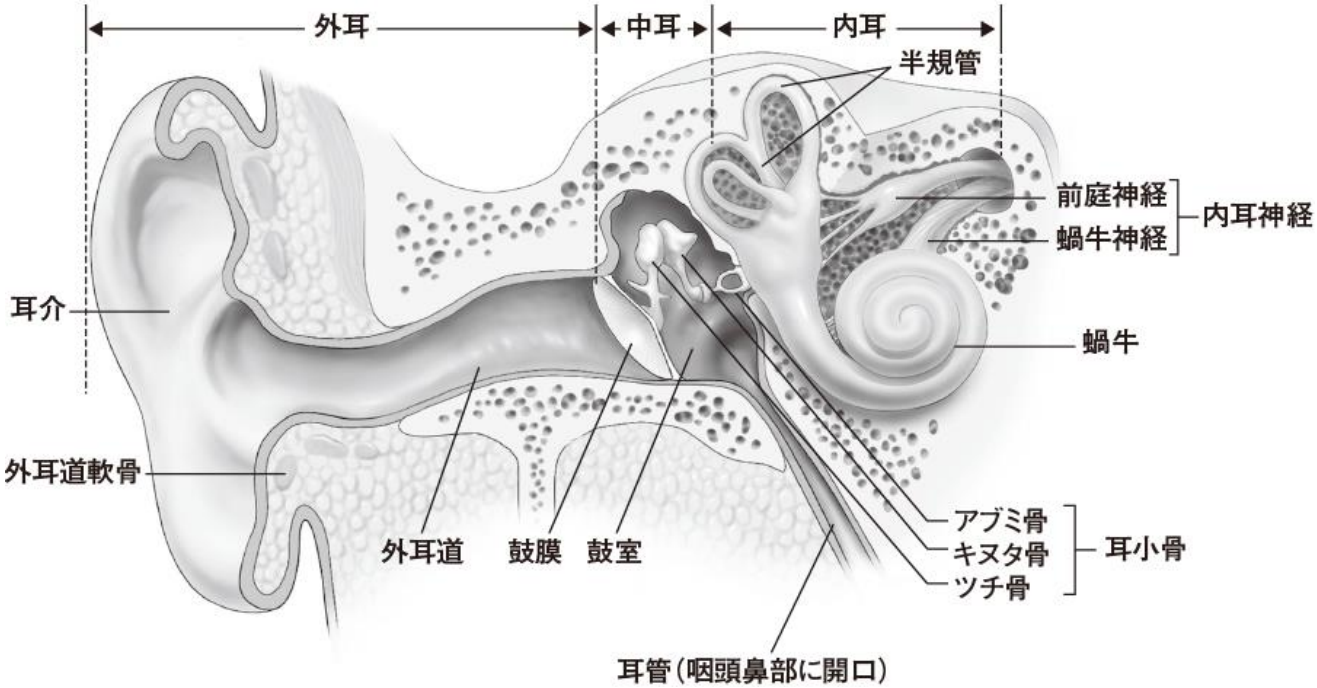
$$\mathrm{HRTF}_{\mathrm{R}}(\omega) = \frac{H_{\mathrm{R}}(\omega)}{H_0(\omega)}$$

受聴者がいない状態での頭部中心までの伝達特性

音の空間知覚

- 頭部伝達関数（HRTF）
  - 音源から放射された音がヒトの鼓膜に到達するまでの伝達特性



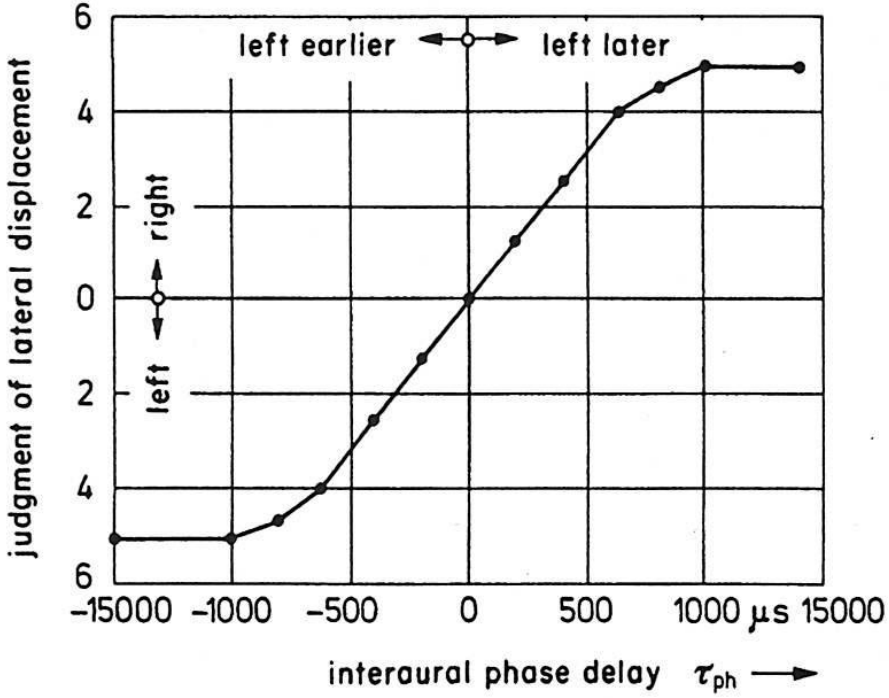


https://www.kango-roo.com/sn/k/view/1720

➡ HRTFのどの要素が音の空間知覚をもたらしている？

方向知覚

- 左右方向の知覚
  - 両耳間時間差（interaural time difference: ITD）
    - 左右の耳に到達する音の時間差

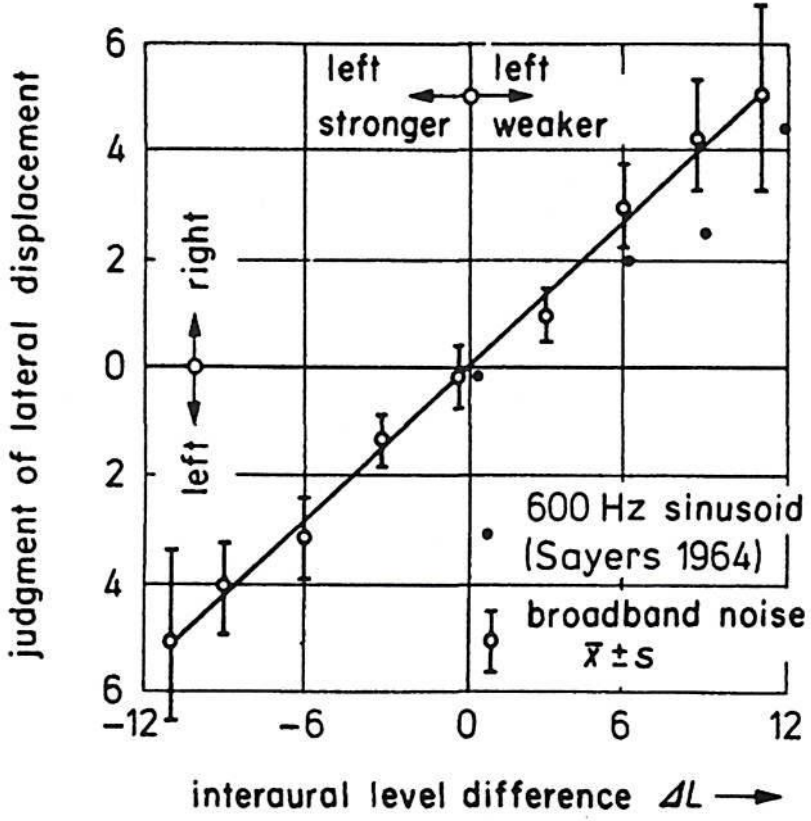


[Blauert 1997]

  - ITDが手がかりになっているのは約1.6kHz以下で，それ以上では両耳入力信号の包絡線の時間差が手がかり

方向知覚

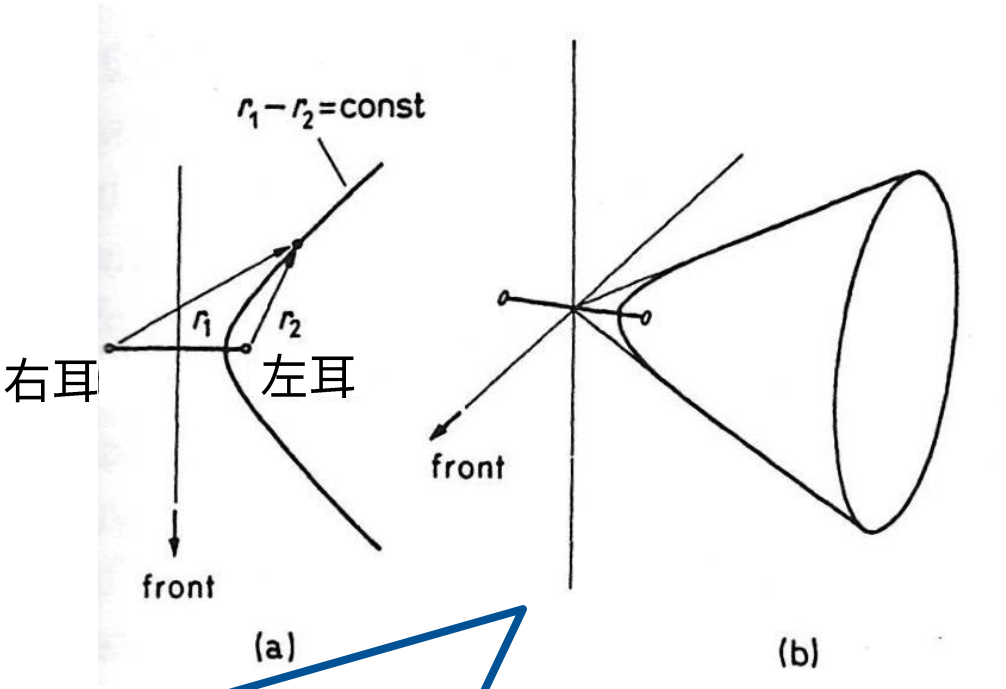
- 左右方向の知覚
  - 両耳間レベル差（interaural level difference: **ILD**）
    - 左右の耳に到達する音の音圧差



[Blauert 1997]

方向知覚

- コーン状の混同（cone of confusion）
  - 頭部や外事が前後・上下で対象な形状だと仮定すると，音源から両耳までの距離の差は円錐台上で一定となる。

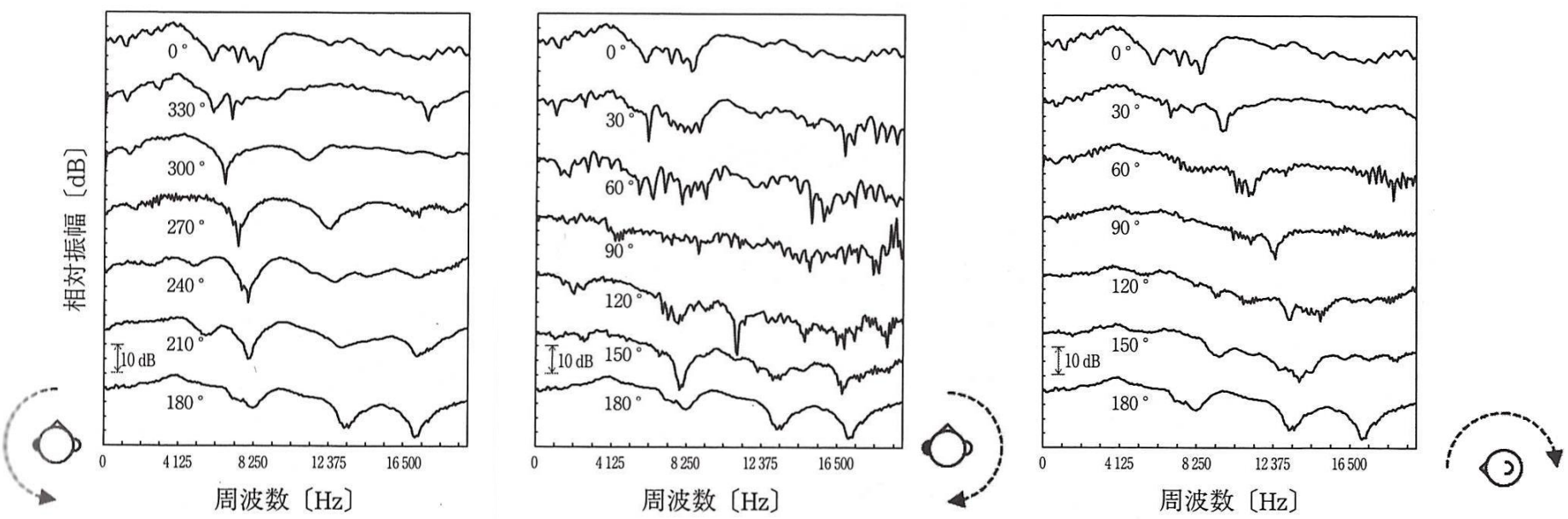


[Blauert 1997]

ITD・ILDだけでは前後・上下方向を同定することはできない

方向知覚

- 前後・上下方向の知覚
  - HRTFの振幅スペクトルが重要な役割を果たしており，これをスペクトラルキュー（spectral cue）と呼ぶ。



[飯田+ 2010]

音源の方向によってHRTFの振幅スペクトルは大きく変化

空間音響技術

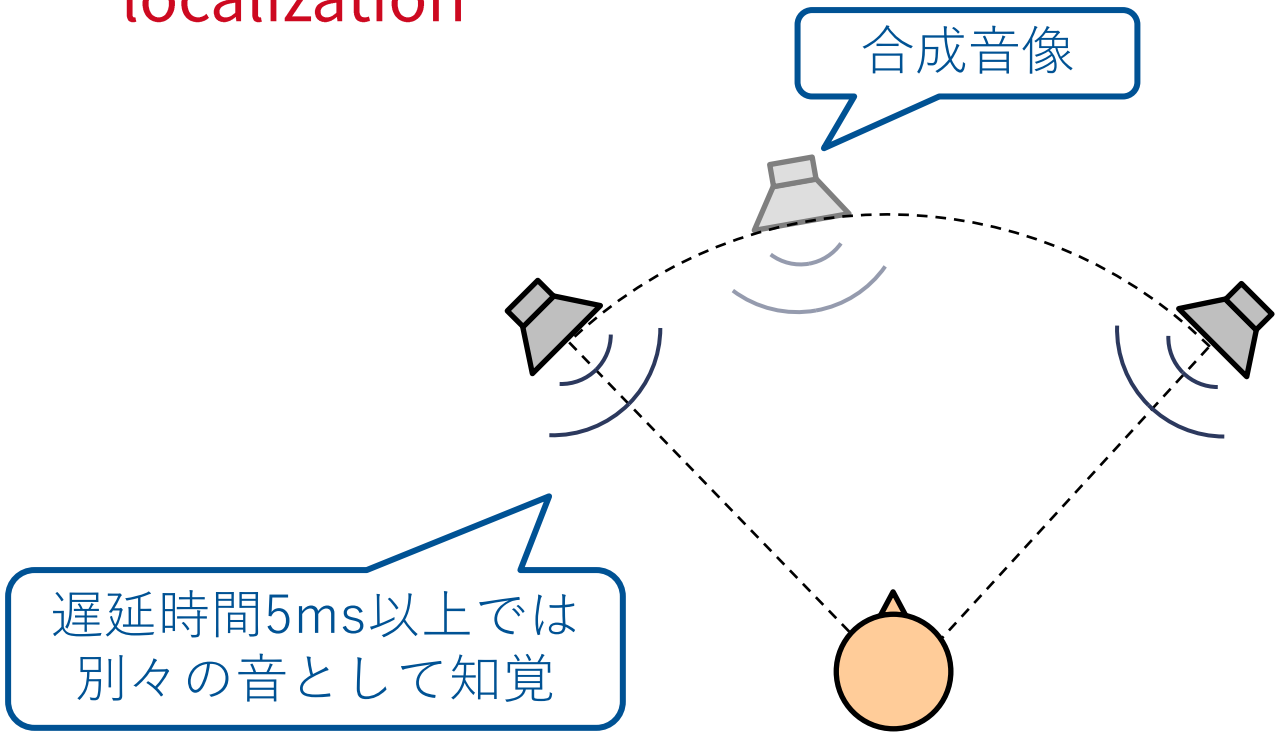
スピーカやヘッドフォンを用いて受聴者に音の空間を提示するには？

- 空間音響技術の分類
  - ステレオフォニック・サラウンド方式
    - Summing localizationと呼ばれる効果を用いて空間的な音像を提示
  - バイノーラル合成
    - 頭部伝達関数やそれを近似したものを用いて両耳の信号（バイノーラル信号）を合成し，提示する
  - 音場再現
    - 複数スピーカを用いて音空間そのものを物理的に合成

➡ それぞれの技術の概要と長所・短所（pros and cons）をまとめる

ステレオフォニック・サラウンド方式

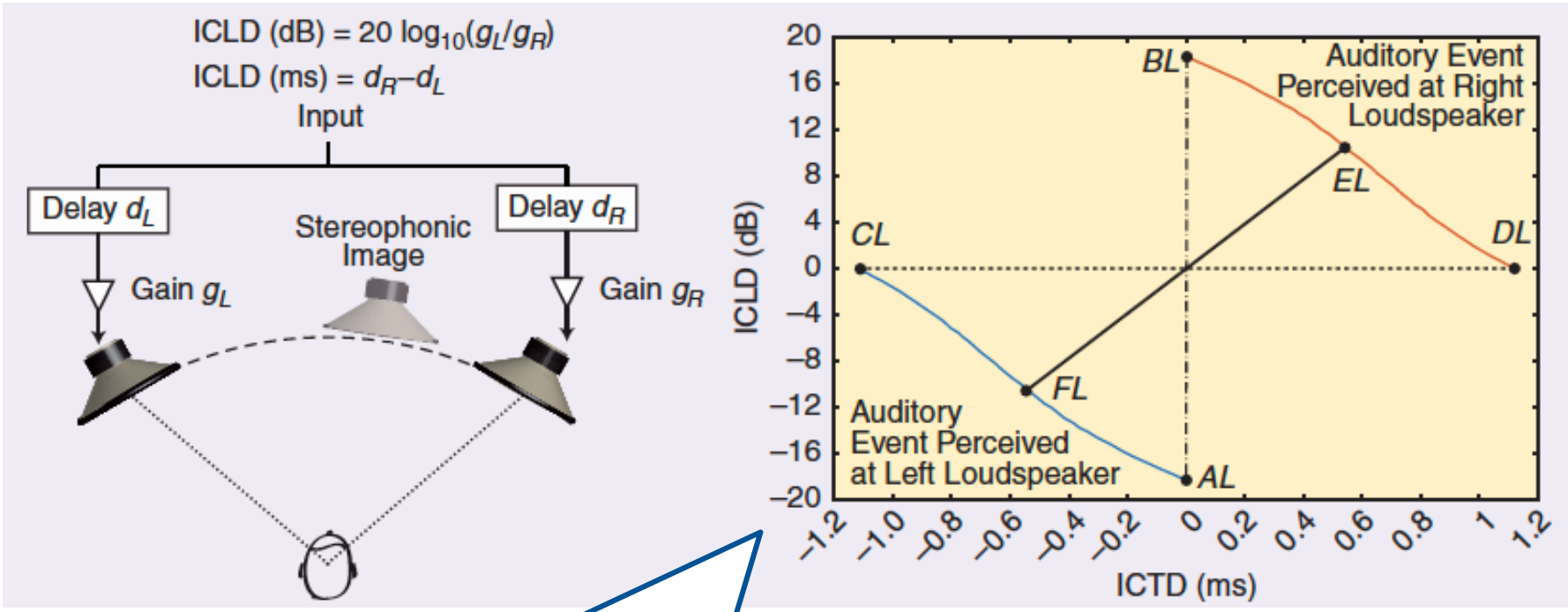
- 複数スピーカのチャネル間での時間差やレベル差によって空間的な音像を提示：2chステレオ，5.1chサラウンド，22.2chマルチチャネル音響など
- 遅延時間1ms以下で２つの音源から同一の信号が到達した場合，それらの音源の間に音像を知覚する：Summing localization



[Hacihabiboglu+ 2017]

2chステレオ

- Summing localizationを利用し，２つのスピーカのチャネル間時間差（interchannel time difference: ICTD），チャネル間レベル差（interchannel level difference: ICLD）によって音像を制御：パンニング（panning）



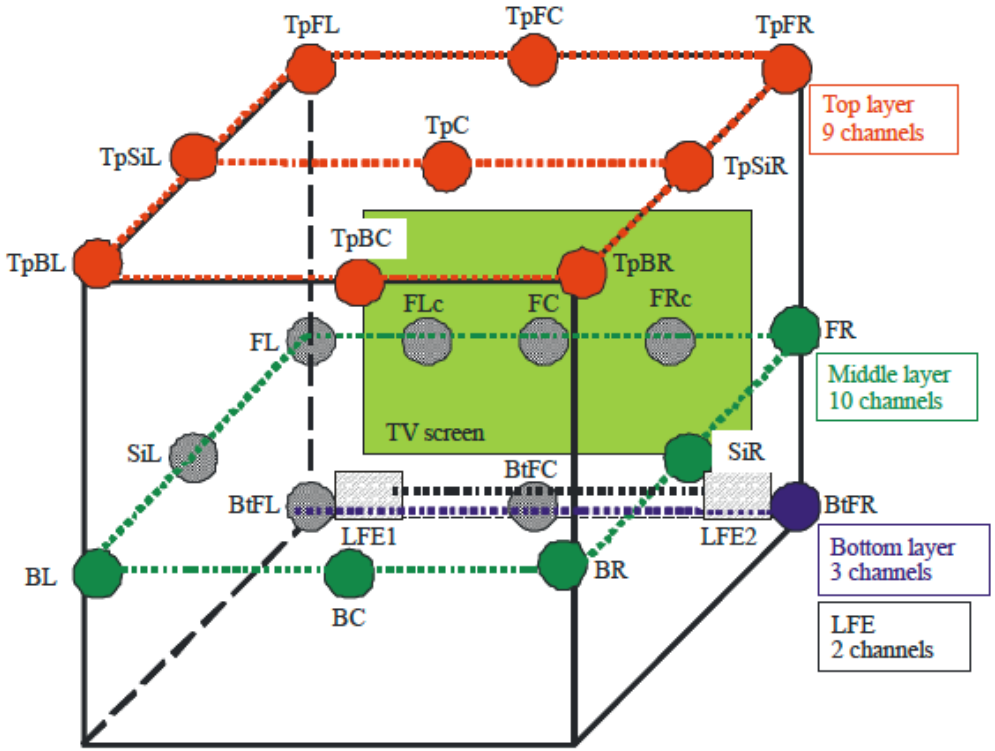
[Hacihabiboglu+ 2017]

ICTD，ICLDの様々な操作方法が存在（panning law）

サラウンド／マルチチャネルステレオ

- サラウンド，マルチチャネルステレオ
  - 受聴者を取り囲むようにスピーカを配置
- 様々な方式が存在
  - 5.1, 7.1, 10.2, 22.2, Dolby ATMOS, DTS-X, Auro-3Dなど

2chステレオと同様に
パンニングによって
音像を制御

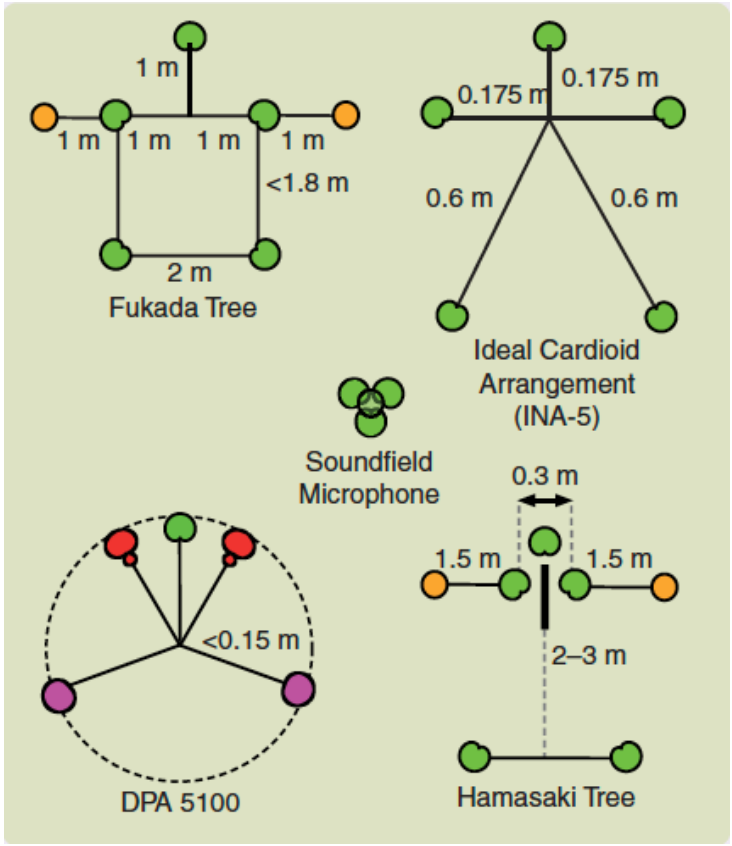


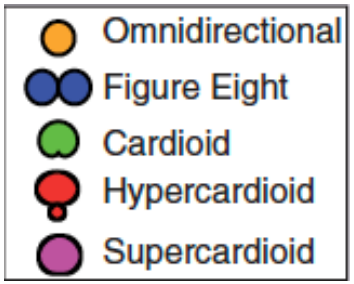
22.2chマルチチャネル音響 （ITU-R BS.2159-7）

サラウンド方式における収録

- 音像を合成する場合は各音源のICTD・ICLDを任意に操作可能。では，マイクを使って収録する場合はどうする？

➡ 複数のマイクを組み合わせたマイクアレイを用いる。
現状は経験的な設計方法によるものがほとんど。





5.1chサラウンド収録のためのマイクアレイ

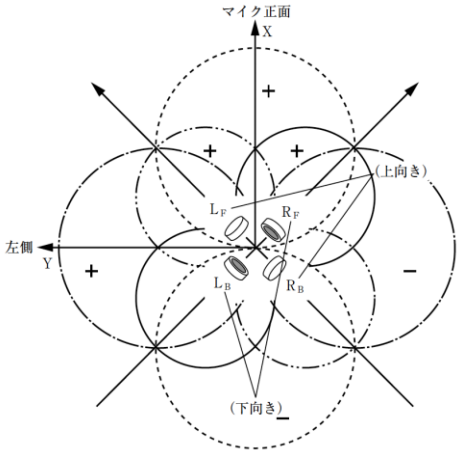
[Hacihabiboglu+ 2017]

Ambisonics

- もう少し理論的根拠のある収録・再生方式としては，Ambisonicsが知られている [Gerzon 1973]

Røde NT-SF1
(https://en.rode.com/nt-sf1)

- **エンコーディング**
  - ４つの（単一指向性）マイクを正四面体の各面に置いたアレイを用いて収音
  - 収音信号を無指向性成分とx, y, z軸方向への指向性成分に変換（B-format）



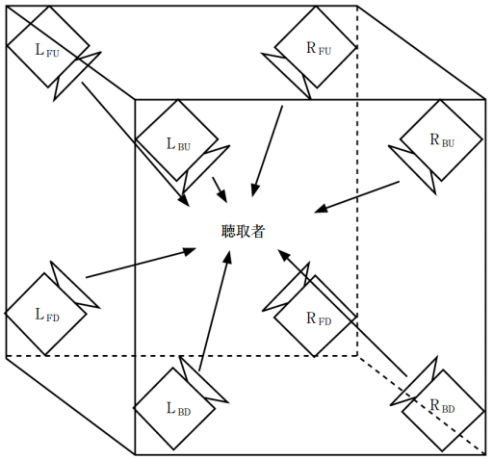
[西村 2014]

Ambisonics

- もう少し理論的根拠のある収録・再生方式としては，Ambisonicsが知られている [Gerzon 1973]

Røde NT-SF1
(https://en.rode.com/nt-sf1)

- **デコーディング**
  - 各スピーカが受聴者から十分遠方にあるとして，平面波として到来すると仮定
  - B-formatの信号をスピーカが設置してある方向の成分に変換し，各スピーカで再生



[西村 2014]

➡ 近似を少なくし，解像度を高くしていくと音場再現と等価となるため，詳細な理論は後ほど

Ambisonics

- Ambisonicsからバイノーラル信号を合成することも可能
  - YoutubeやFacebookの360°動画で採用されている

https://www.youtube.com/watch?v=vUtEW6OTQLw

ステレオフォニック・サラウンド方式

- Pros
  - 少数のスピーカとICTD・ICLDなどの簡単な処理だけで音像を操作できるため，放送・制作用途としては扱いやすい
  - 実際よりも強調した表現を行うためのデフォルメもしやすい

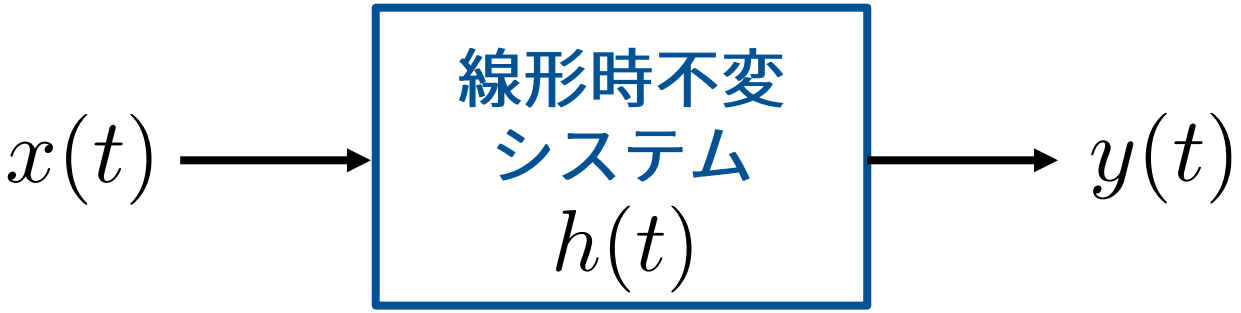
- Cons
  - Summing localizationに基づくため，物理的に音空間そのものが合成されるわけではなく，パンニングやエフェクトによって各音像をデザインすることが必要
  - 音空間を合成するのではなく，マイクで収録して再生する場合には，現状の収録方法はあくまで経験的なもの
  - スピーカの中心位置で受聴しなければ十分な効果が得られない（スィートスポット）

バイノーラル合成

- 音の空間知覚において，HRTFが重要な役割を持つ

  ➡ HRTFを直接用いて両耳の信号を合成すればよい

- HRTFを線形時不変システムとしてみなす



  - 伝達関数である$h(t)$は入力$x(t)$がデルタ関数$\delta(t)$のときの出力であり，インパルス応答とも呼ばれる
  - インパルス応答$h(t)$があらかじめ既知であれば，任意の入力信号$x(t)$に対して，$h(t)$を畳み込むことで，その出力$y(t)$を合成することができる

バイノーラル合成

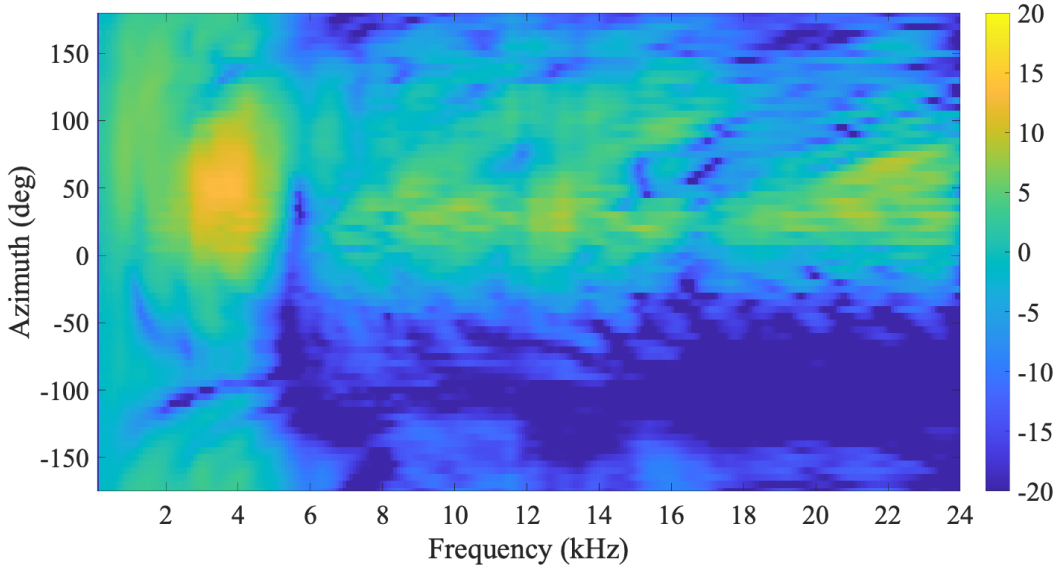
- 音の空間知覚において，HRTFが重要な役割を持つ
  - ➡ HRTFを直接用いて両耳の信号を合成すればよい
- HRTFを線形時不変システムとしてみなす

  - 任意の音源信号に対して，あらかじめ測定したHRTFを畳み込むことで，そのHRTFを測定した音源位置からのバイノーラル信号を合成することができる

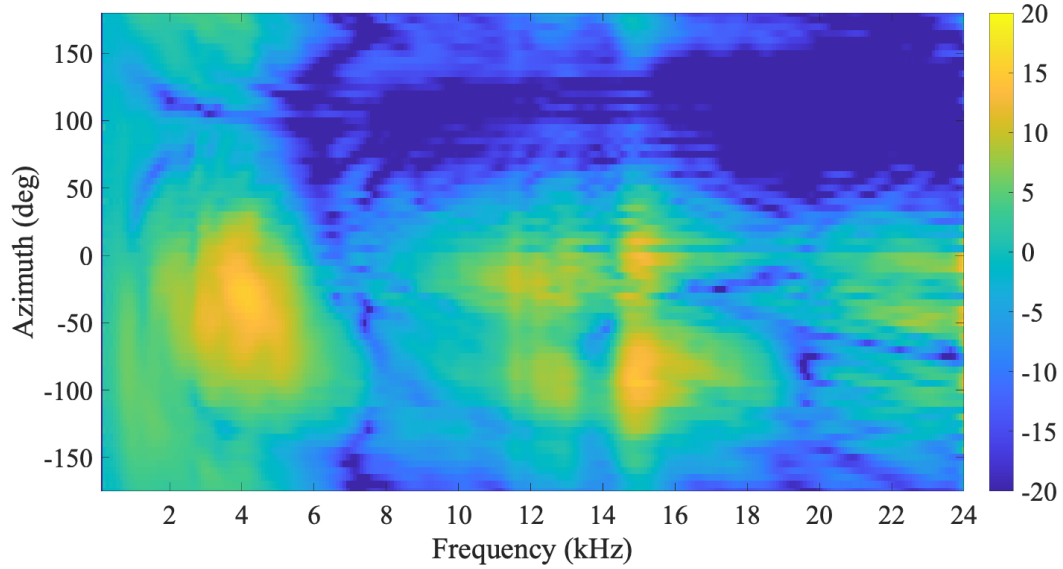
バイノーラル合成

- HRTFの例

左耳



右耳



頭部伝達関数の測定

- 任意のバイノーラル信号を合成するには，あらゆる方向からのHRTFを測定しておくことが必要

- 無響室などで外耳道入り口に設置したマイクを用いて各スピーカからのインパルス応答を計測
- 実際のスピーカでデルタ関数（パルス波）を出力することは難しいので，swept sine信号やM系列信号などを用いることが多い

東北大学のHRTF測定システム
[坂本+ 2016]

- HRTFのデータベースが様々な機関から公開されている
  - 東北大学: http://www.riec.tohoku.ac.jp/pub/hrtf/
  - UC Davis: https://www.ece.ucdavis.edu/cipic/spatial-sound/hrtf-data/

バイノーラル信号の収音

- 音像を合成するようにバイノーラル信号を生成するのではなく，マイクで収音する場合は？

  ➡ ダミーヘッドやバイノーラルマイクを用いて収音

人間の頭部を模した形状のマイク

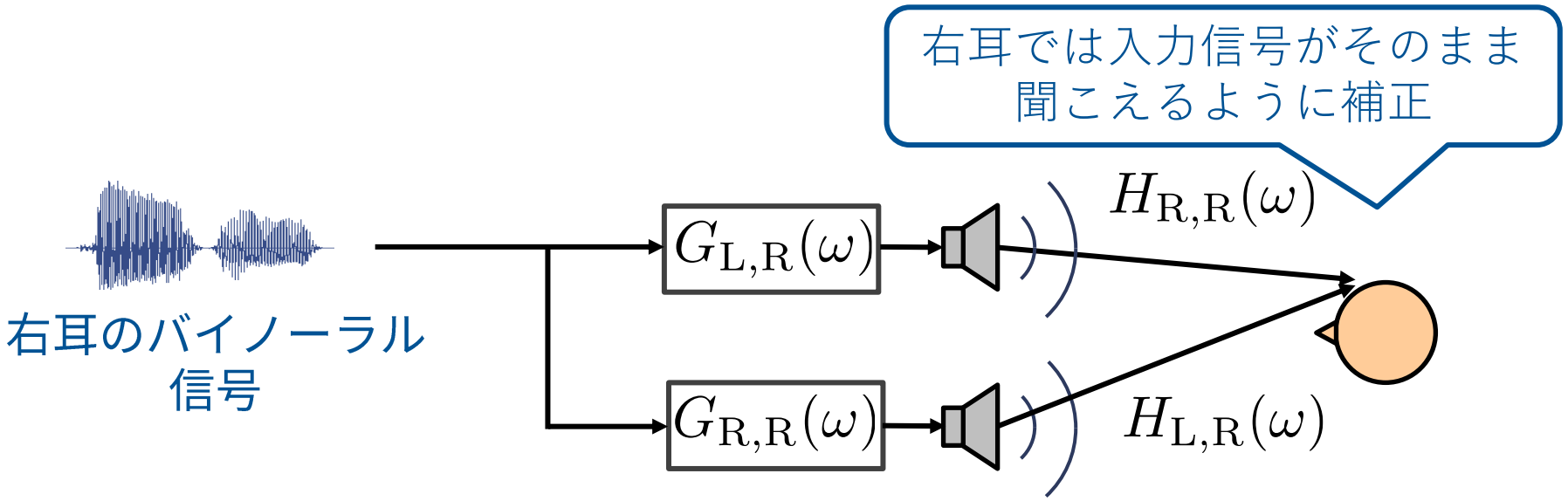
Brüel & Kjær HATS (TYPE 4128-C)
https://www.bksv.com/en/products/transducers/ear-simulators/head-and-torso/hats-type-4128c

イヤフォンのように両耳に設置するマイク

Roland CS-10EM
https://www.roland.com/jp/products/cs-10em/

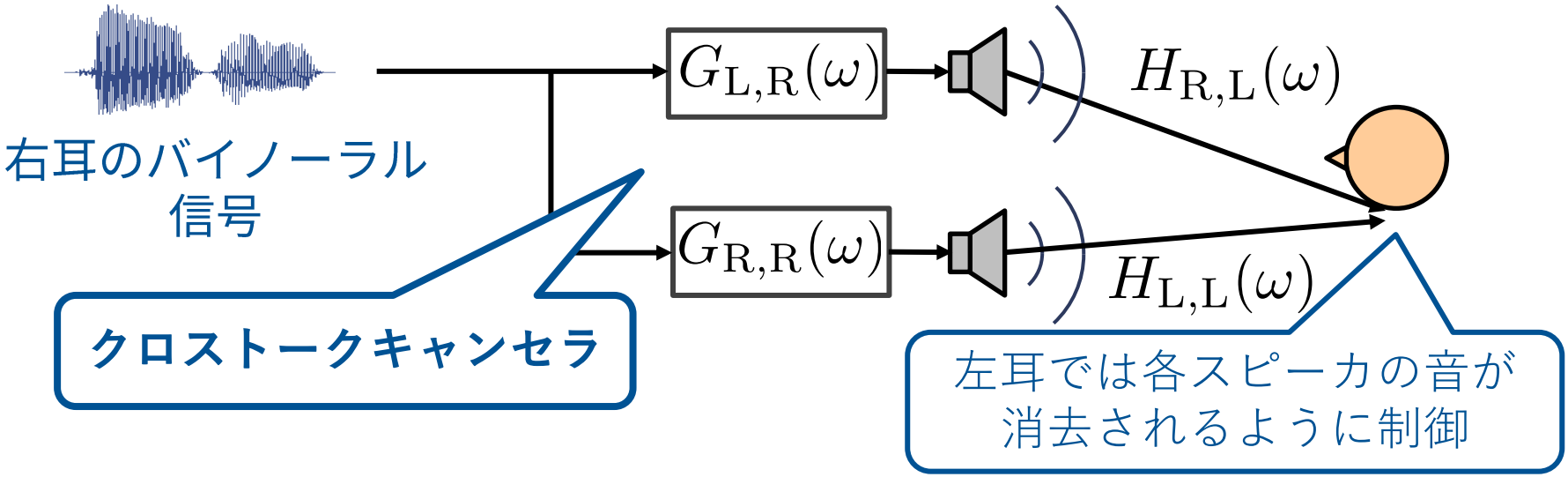
バイノーラル信号の提示

- ヘッドフォン・スピーカによるバイノーラル再生
  - バイノーラル信号はヘッドフォンを用いて受聴者に提示するのが最も簡便。ヘッドフォンから外耳道入口までの伝達特性を補正する場合もある。
  - スピーカを用いて提示する場合は，クロストークを消去・抑圧することが必要。



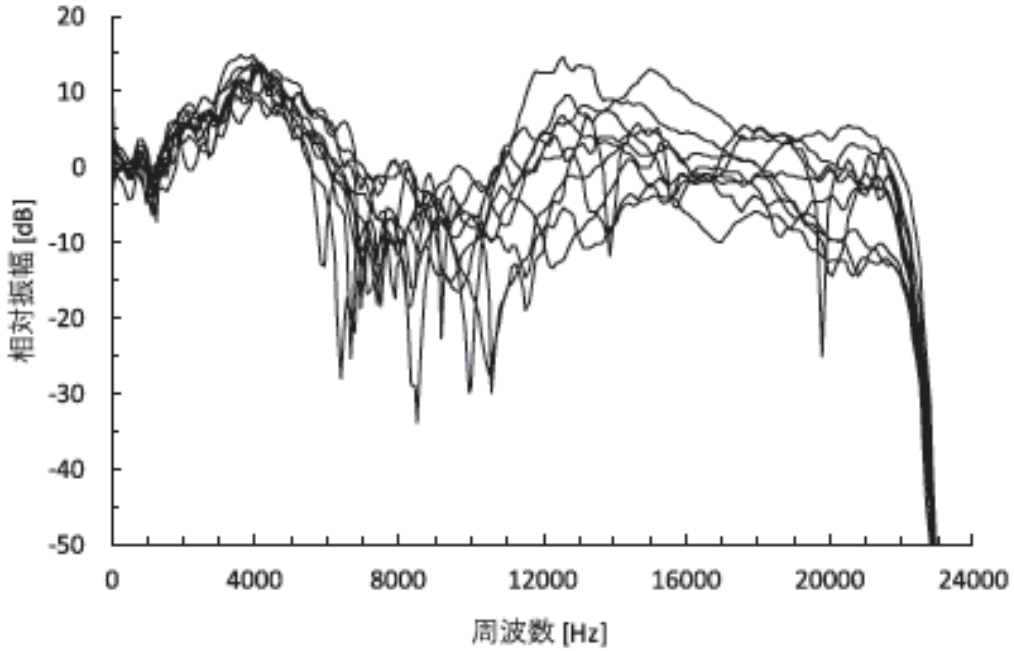
バイノーラル信号の提示

- ヘッドフォン・スピーカによるバイノーラル再生
  - バイノーラル信号はヘッドフォンを用いて受聴者に提示するのが最も簡便。ヘッドフォンから外耳道入口までの伝達特性を補正する場合もある。
  - スピーカを用いて提示する場合は，クロストークを消去・抑圧することが必要。



頭部伝達関数の個人性

- HRTFは個人差が大きい
  - 受聴者本人のHRTF以外で合成されたバイノーラル信号では，うまく音像定位ができないことが知られている



成人10名の正面方向のHRTF

[飯田 2017]

➡ バイノーラル合成ではHRTFの個人適応が大きな問題となる

バイノーラル合成

➢ Pros

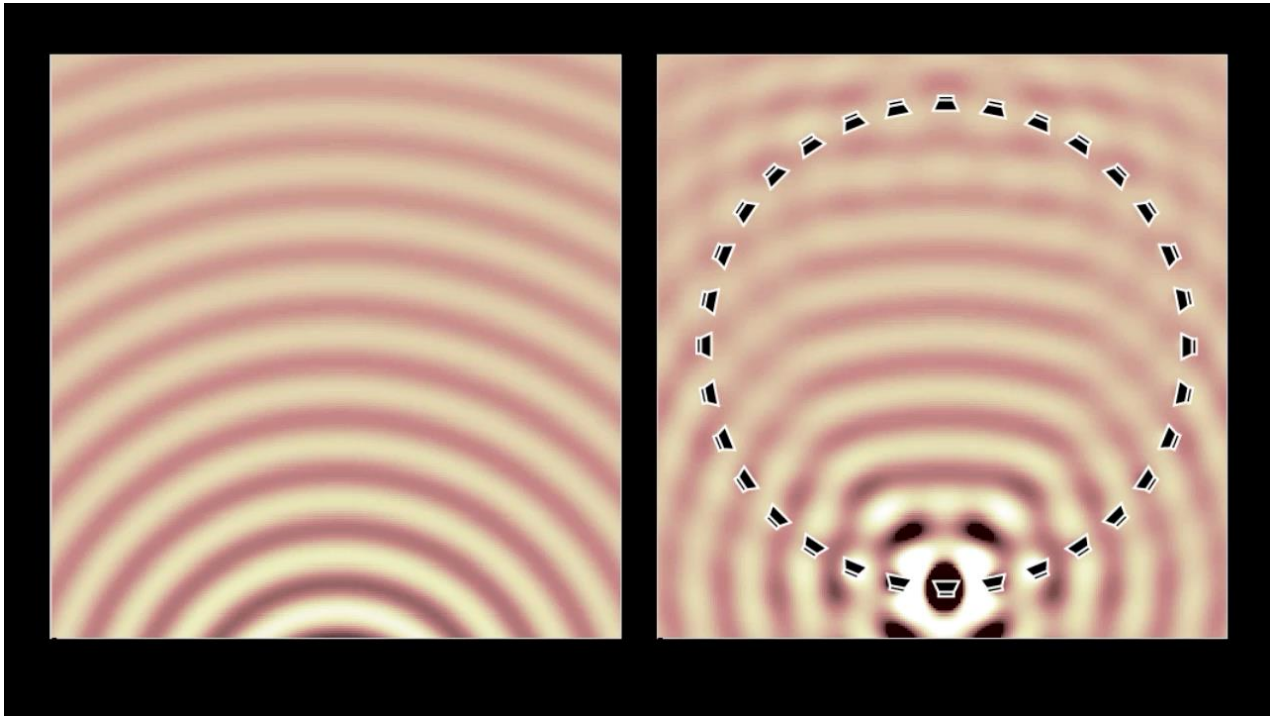
- ヘッドフォンや比較的少数のスピーカによる簡易的なシステムで実現可能
- 受聴者の頭部が動く場合には，トラッキングなどである程度対応できる

➢ Cons

- HRTFの測定は特殊な設備が必要であり，時間もかかる
- 残響環境下でのバイノーラル信号を合成する場合は，反射音があらゆる方向から到来するため，工夫が必要
- 測定できるHRTFの数は有限であるため，動きのある音源を合成するには補間が必要（特に距離方向の補間は難しい）
- HRTFの個人適応については今のところ効果的な方法が確立されていない

音場再現

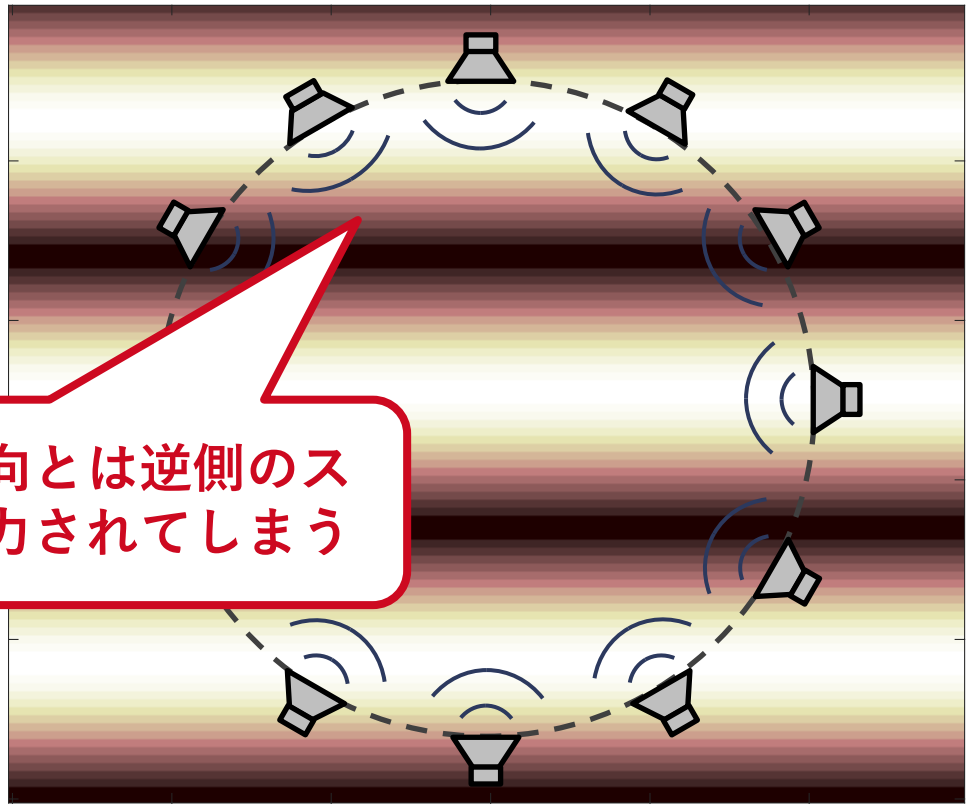
- 複数スピーカを用いて音空間そのものを物理的に合成
  - 広い受聴領域が実現可能であり，受聴者が複数いる場合や，受聴者が動く場合にも適応可能



概念としては1990年頃からあったが，
理論的な整理がなされたのはここ10年ほど

音場再現

- ホイヘンスの原理で考えると，スピーカを連続的に並べ，
  それぞれのスピーカ位置での音圧を出力すればよい？



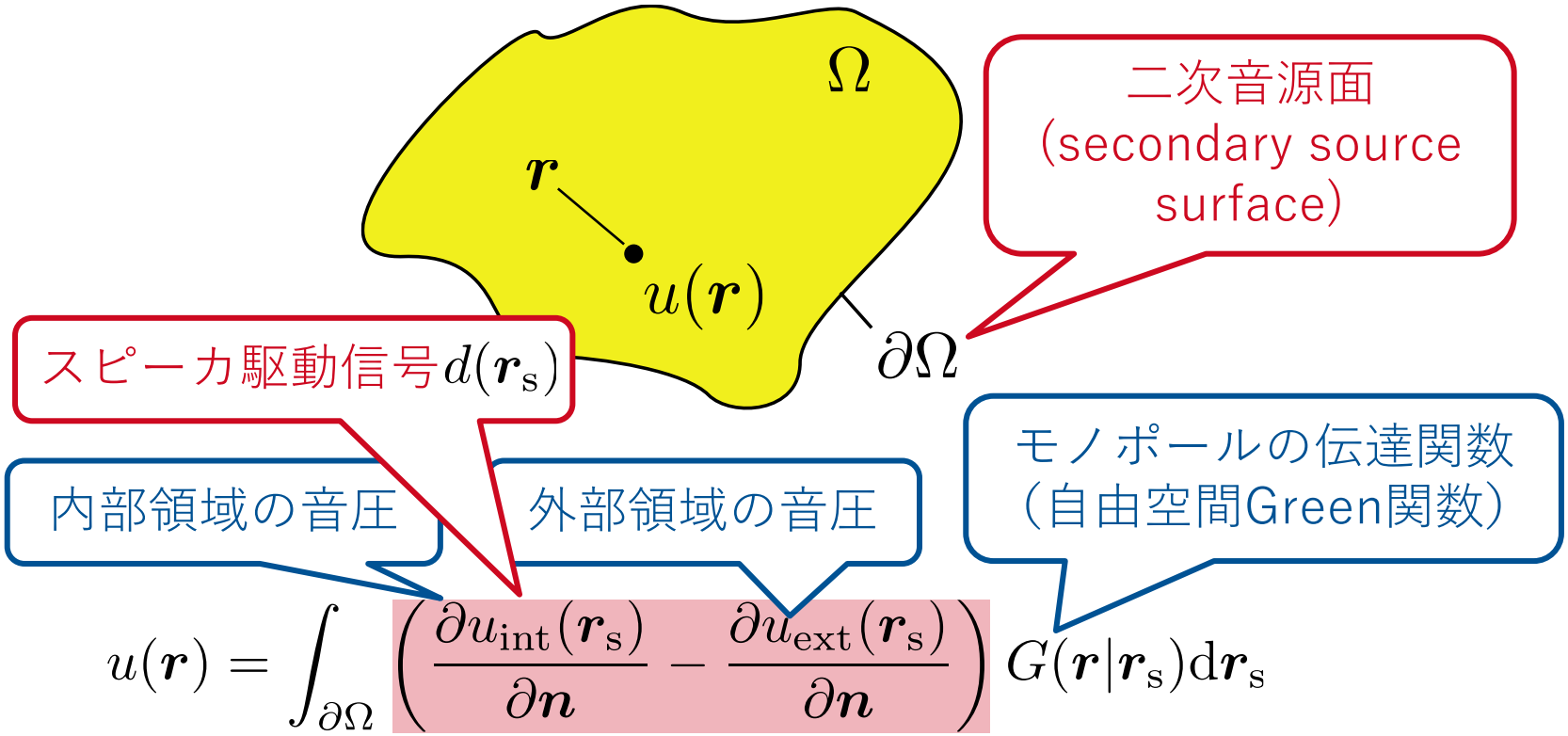
平面波の進行方向とは逆側のスピーカからも出力されてしまう

平面波の進行方向

各スピーカ位置での音圧を直接出力しても
音場を再現することはできない

音場再現

- Single layer potentialに基づく定式化
  - ある音源を含まない領域Ω内の任意の音場は，境界面$\partial\Omega$上に連続的に分布するモノポール（点音源）によって表現可能

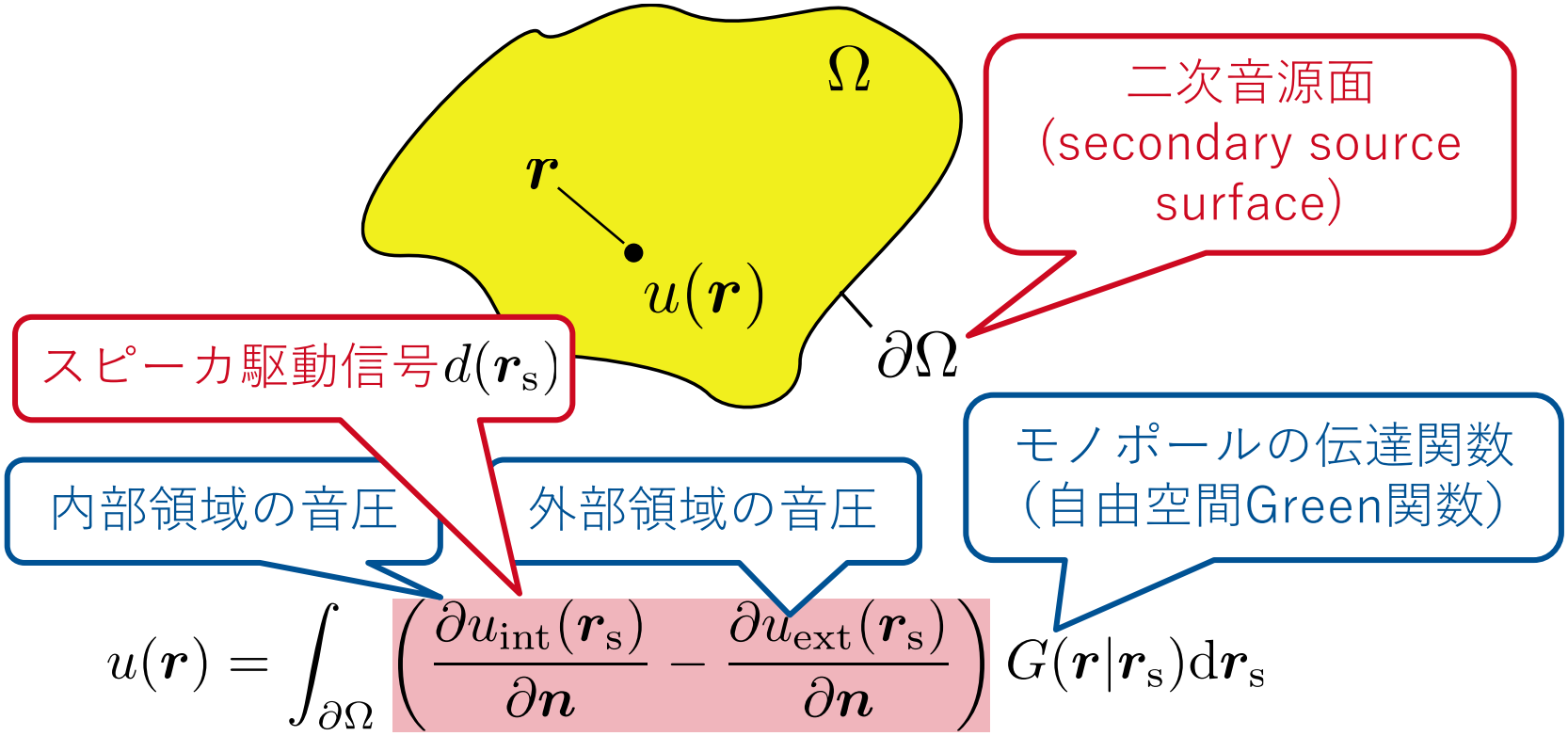


$$u(\boldsymbol{r}) = \int_{\partial\Omega} \left(\frac{\partial u_\mathrm{int}(\boldsymbol{r}_\mathrm{s})}{\partial \boldsymbol{n}} - \frac{\partial u_\mathrm{ext}(\boldsymbol{r}_\mathrm{s})}{\partial \boldsymbol{n}} \right) G(\boldsymbol{r}|\boldsymbol{r}_\mathrm{s}) \mathrm{d}\boldsymbol{r}_\mathrm{s}$$

➡ モノポール特性を持つようなスピーカを連続的に配置することで任意の音場を合成できる

音場再現

- Single layer potentialに基づく定式化
  - ある音源を含まない領域Ω内の任意の音場は，境界面$\partial\Omega$上に連続的に分布するモノポール（点音源）によって表現可能



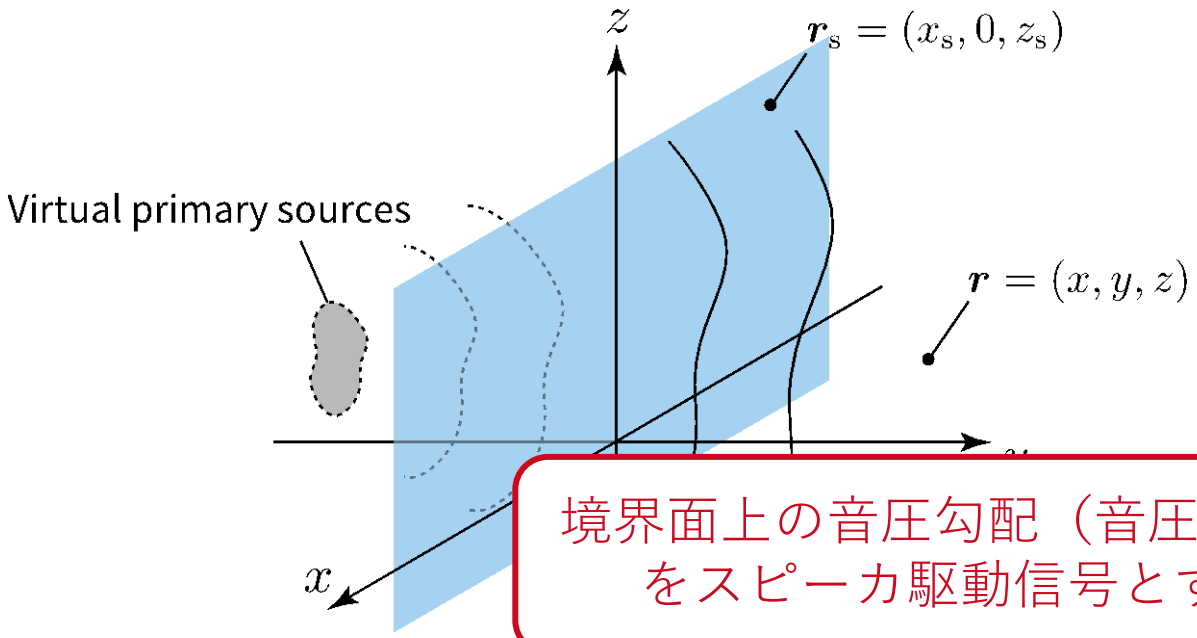
$$u(\boldsymbol{r}) = \int_{\partial\Omega} \left(\frac{\partial u_\mathrm{int}(\boldsymbol{r}_\mathrm{s})}{\partial \boldsymbol{n}} - \frac{\partial u_\mathrm{ext}(\boldsymbol{r}_\mathrm{s})}{\partial \boldsymbol{n}} \right) G(\boldsymbol{r}|\boldsymbol{r}_\mathrm{s}) \mathrm{d}\boldsymbol{r}_\mathrm{s}$$

領域Ωが単純な形状の場合は解析的にスピーカ駆動信号を導出できる

波面合成法：Wave Field Synthesis

- 波面合成法（Wave Field Synthesis: WFS）
  - Single layer potentialにおいて境界面を無限大の平面とした場合

[Spors+ 2008]



境界面上の音圧勾配（音圧の空間微分）
をスピーカ駆動信号とすればよい

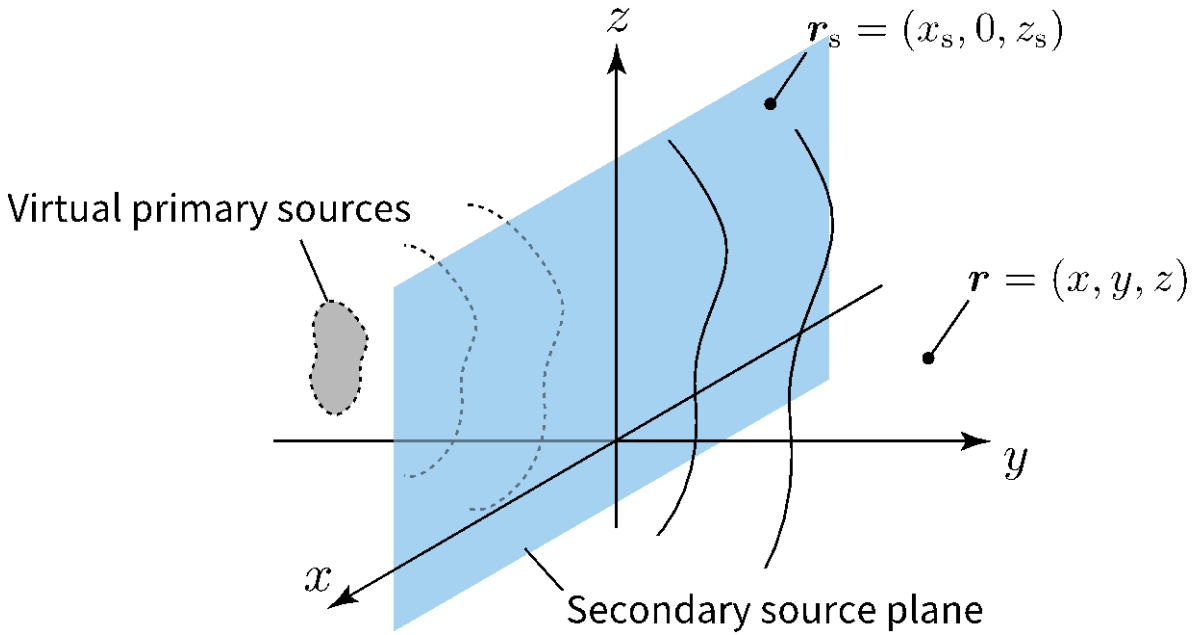
第１種Rayleigh積分

$$u(\boldsymbol{r}) = -2 \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \frac{\partial u_\mathrm{int}(\boldsymbol{r}_\mathrm{s})}{\partial y_\mathrm{s}} G(\boldsymbol{r}|\boldsymbol{r}_\mathrm{s}) \mathrm{d}x_\mathrm{s} \mathrm{d}z_\mathrm{s}$$

境界面上のマイクで収音した信号（音圧）を
そのまま出力すればよいわけではない

波面再構成フィルタ法

- 収音した音場を再現するにはどうする？
  - 波数領域で音圧分布から音圧勾配の分布を推定 [Koyama+ 2013]



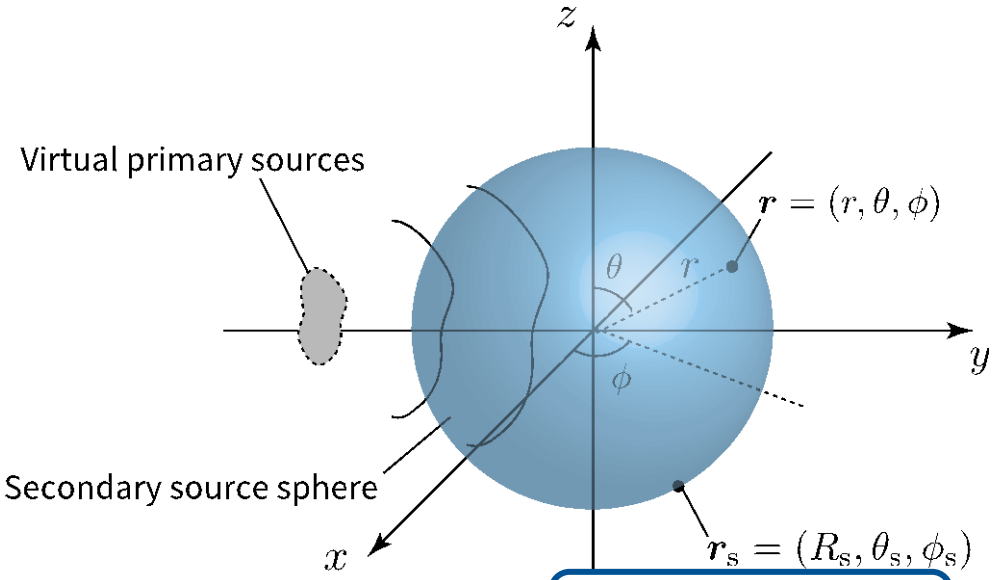
第１種Rayleigh積分の波数領域表現

$$\tilde{u}(k_x, y, k_z) = -2i\sqrt{k^2 - k_x^2 - k_z^2}\,\tilde{u}_\mathrm{int}(k_x, 0, k_z)\tilde{G}(k_x, y, k_z)$$

- 平面状のマイク・スピーカアレイを用いた音場収音・再現
- 直線状アレイを用いる形に近似することも可能

高次アンビソニックス：Higher Order Ambisonics

- 高次アンビソニックス（Higher Order Ambisonics: HOA）
  - Single layer potentialにおいて境界面を球状とした場合
  - Ambisonicsからより近似を小さくした場合と等価



音場の球波動関数展開表現

球Bessel関数

球面調和関数

$$u_{\mathrm{int}}(\boldsymbol{r}) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} A_{nm} j_n(kr) Y_n^m(\theta, \phi)$$

球Hankel関数

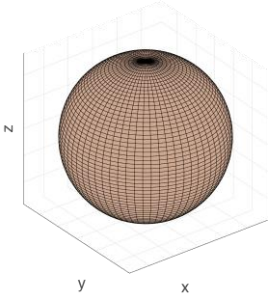
$$u_{\mathrm{ext}}(\boldsymbol{r}) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} B_{nm} h_n^{(1)}(kr) Y_n^m(\theta, \phi)$$

球面調和関数

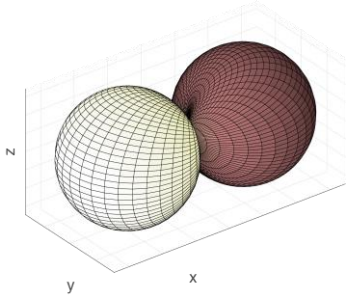
球面調和関数：

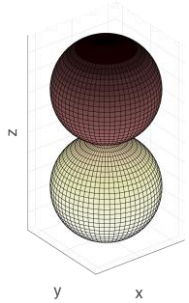
$$Y_n^m(\theta,\phi) = \sqrt{\frac{(2n+1)}{4\pi}\frac{(n-m)!}{(n+m)!}} P_n^m(\cos\theta) e^{im\phi}$$

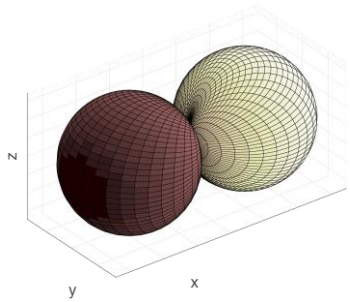
$n = 0$



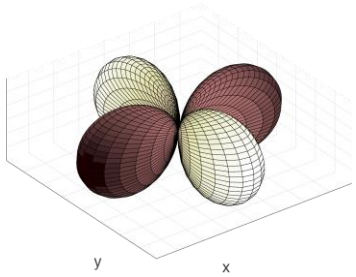
$n = 1$

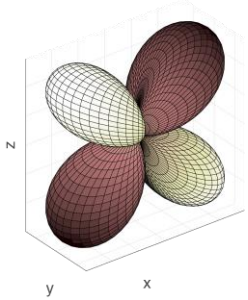


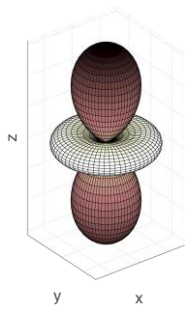


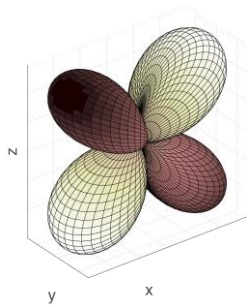


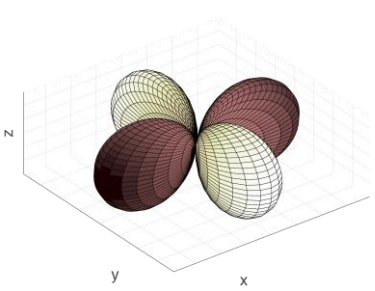
$n = 2$











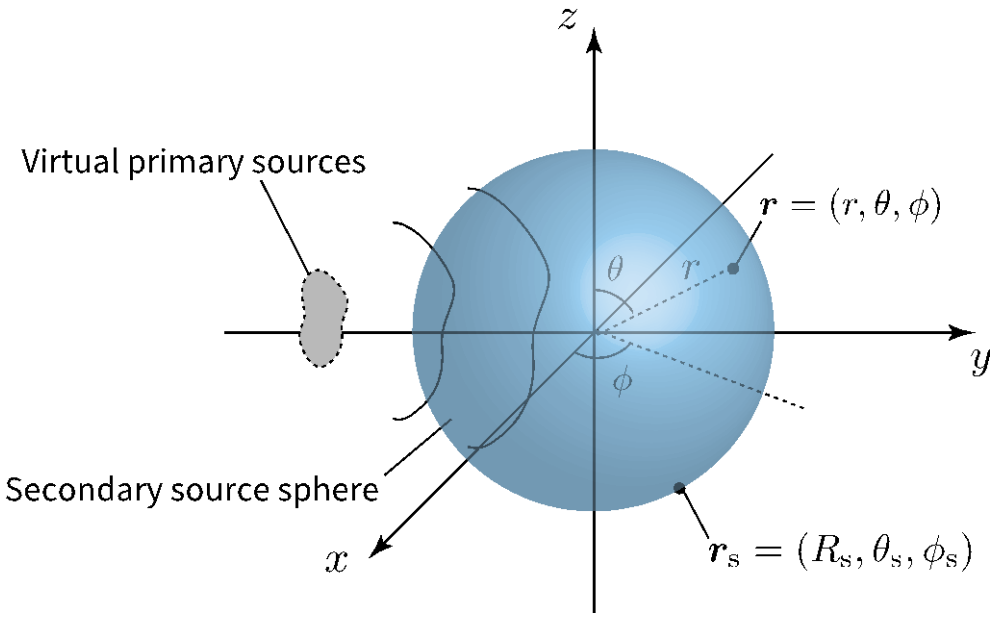
$m = -2$ $m = -1$ $m = 0$ $m = 1$ $m = 2$

高次アンビソニックス：Higher Order Ambisonics

- 球波動関数展開表現をSingle layer potentialに代入
  - 球の半径をR_sとすると，スピーカ駆動$d(\boldsymbol{r}_\mathrm{s})$は，

所望音場の展開係数

$$d(\boldsymbol{r}_\mathrm{s}) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} \frac{A_{nm}}{ikR_\mathrm{s}^2 h_n^{(1)}(kR_\mathrm{s})} Y_n^m(\theta, \phi)$$



[Poletti+ 2005]

- 球表面上に配置したスピーカによって内部の音場を再現
- 第1次までの展開係数を用い，各スピーカを平面波近似したものがAmbisonicsに相当

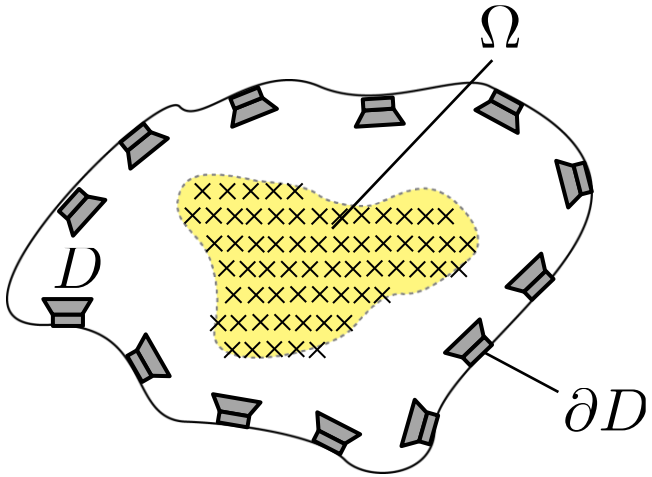
高次アンビソニックス：Higher Order Ambisonics

- 収音した音場を再現するにはどうする？
  - 収音場の球波動関数による展開係数A_{nm}を推定することが必要
  - 球状マイクアレイを用いることで推定可能だが，ある特定の周波数で推定ができない禁止周波数問題が知られている
  - 音響的に剛体のバッフル上にマイクを設置する，指向性マイクによるアレイを用いる，多重の球状アレイ配置を用いる，などの方法で禁止周波数問題を回避 [Poletti+ 2005, Koyama+ 2016]

mh acoustics Eigenmike microphone
https://mhacoustics.com/

音圧制御法：Pressure Matching

- 任意形状の領域内に音場を再現するには？
  - 制御対象領域を離散化し，離散的な制御点上での音圧が所望音場と一致するようにスピーカ駆動信号を数値的に求める



所望音圧分布のベクトル

二次音源と制御点間の伝達関数行列

スピーカ駆動信号のベクトル

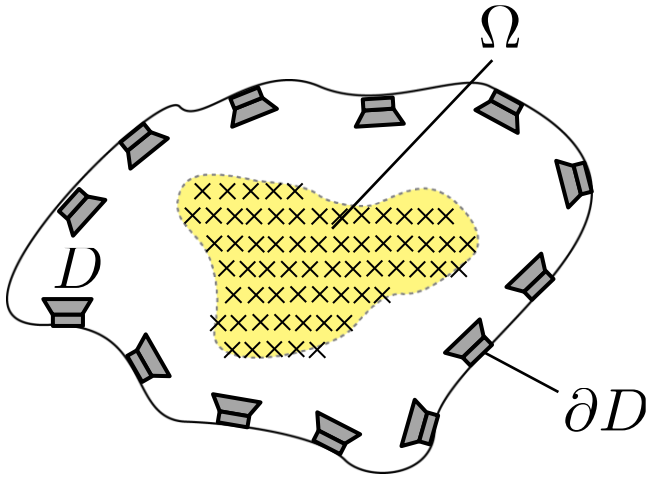
$$\boldsymbol{u}^{\mathrm{des}} = \boldsymbol{G}\boldsymbol{d}$$

$$\Rightarrow \boldsymbol{d} = \boldsymbol{G}^{\dagger}\boldsymbol{u}^{\mathrm{des}}$$

Moore-Penrose 型擬似逆行列

音圧制御法：Pressure Matching

- 任意形状の領域内に音場を再現するには？
  - 制御対象領域を離散化し，離散的な制御点上での音圧が所望音場と一致するようにスピーカ駆動信号を数値的に求める



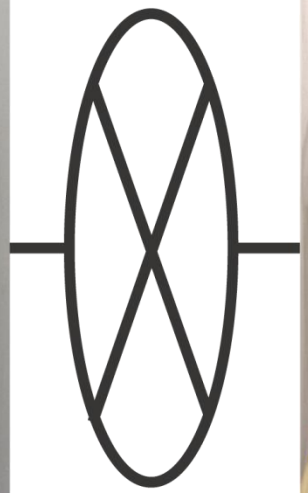
  - 残響環境下で収音した音場を再現する場合に，対象領域内の音圧分布を密に計測することは難しい
  - Kirchhoff-Helmholtz積分方程式によれば，境界面上の音圧と音圧勾配を一致させればよいが，音圧勾配の計測は困難
  - できるだけ少数の素子でPressure matchingを実現するためのスピーカ・制御点配置最適化法 [Koyama+ IEEE/ACM TASLP 2020]

音場再現システムの実現例

音場の収音・伝送・再現をリアルタイムで実現 [Koyama+ IEICE Trans 2014]

Yokosuka

Network

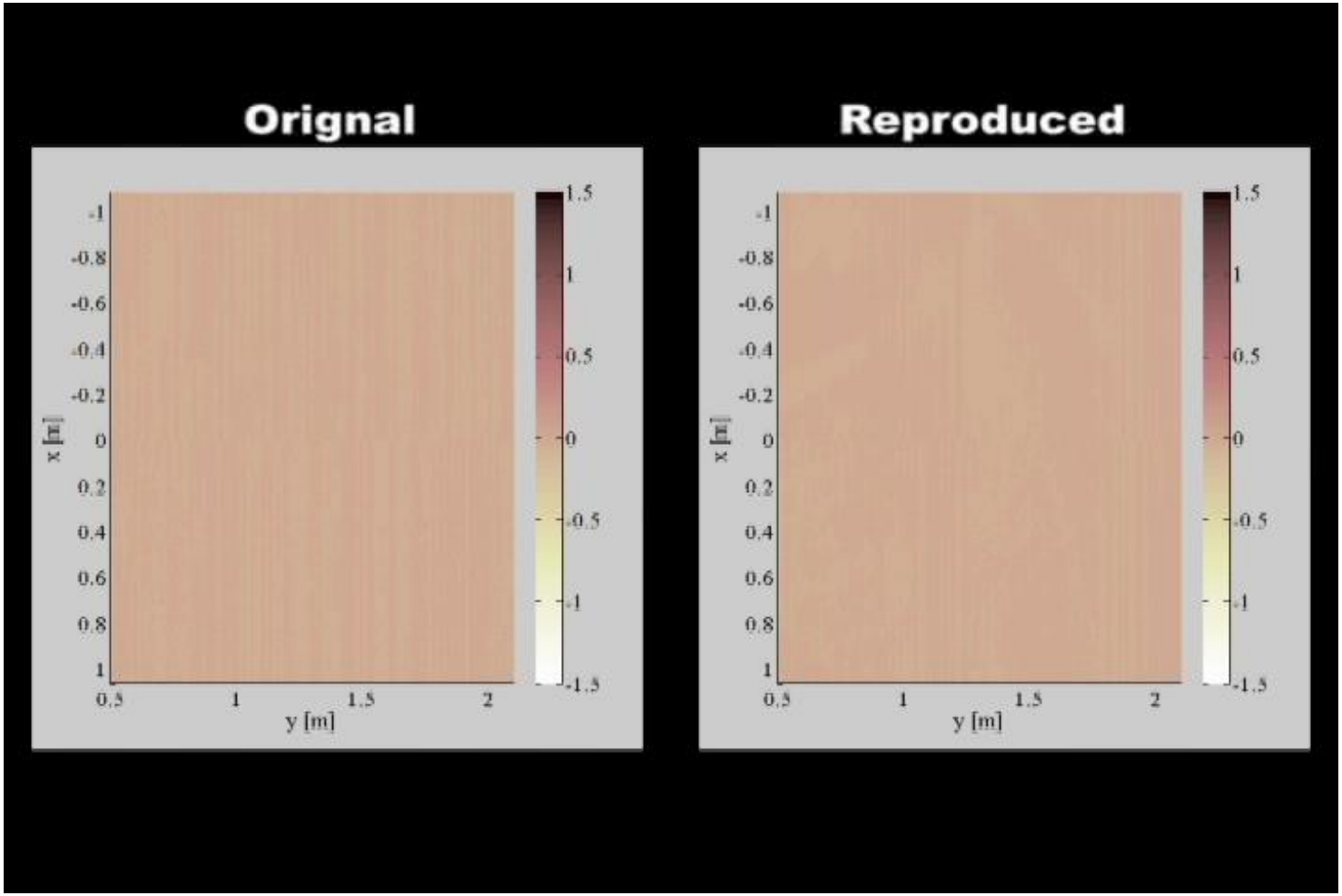


Musashino

- Loudspeakers (for high freq.): 64, 6cm intervals
- Loudspeakers (for low freq.): 32, 12cm intervals
- Microphones: 64, 6cm intervals
- Array size: 3.84 m
- Sampling freq.: 48 kHz, Delay: 152 ms

再現音場の可視化

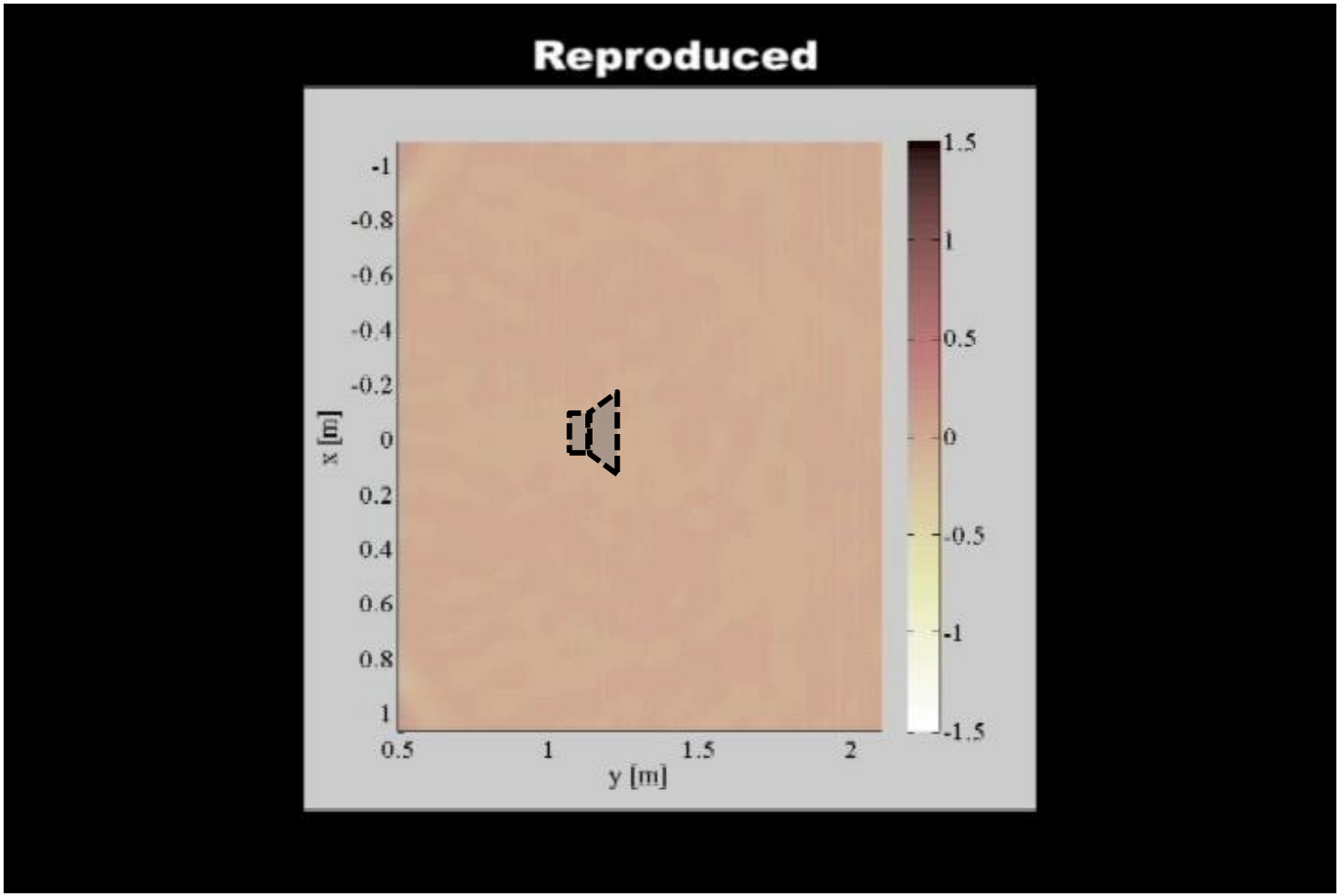
– Source signal: Low-passed pulse (0 – 2.6kHz)
– Source: Loudspeaker, Position: (-1.0, -1.0, 0.0) m



[Koyama+ IEEE TASLP2013]

再現音場の可視化

– Source signal: Low-passed pulse (0 – 2.6kHz)
– Source: Loudspeaker, Position: (0.0, -1.0, 0.0) m, 2.0 m forward shift



[Koyama+ IEEE TASLP2012, ICASSP2012]

音場再現

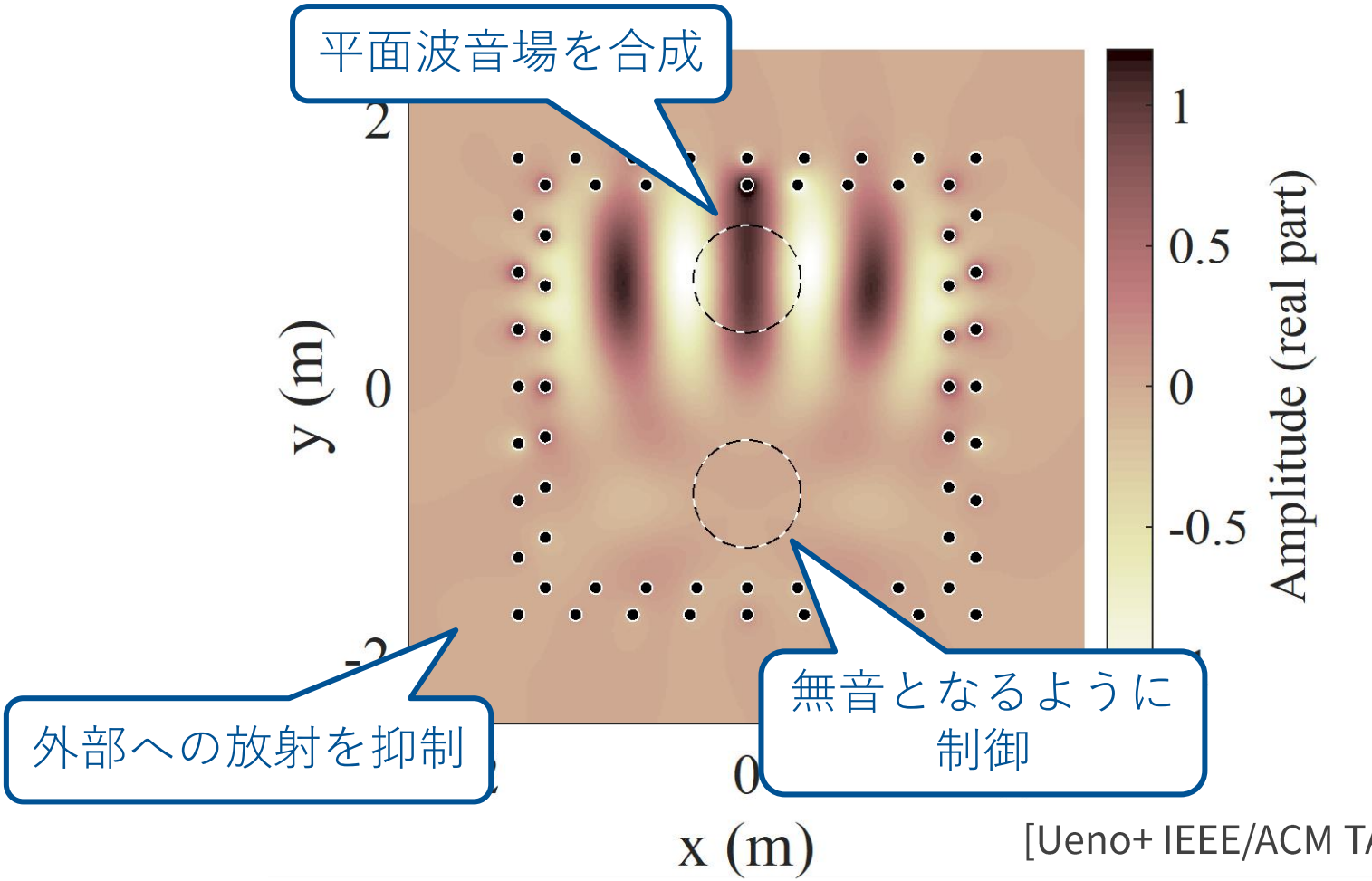
- Pros
  - 物理的に音場そのものを再現するため，広い受聴領域が実現可能
  - 複数の受聴者や受聴者の動きにもそのまま対応可能であり，空間知覚の観点からも有利
  - 収音場の再現も可能であり，人手での操作やデザインを介することなく残響のある音場を再現することも可能

- Cons
  - 再現可能な周波数や範囲がスピーカ・マイクの数によって決まり，それ以上では空間エイリアシングと呼ばれる誤差が生じる
  - 多数のスピーカ・マイクが必要となるため，システムの規模が大きくなる

    ➡ 素子の小型化・低コスト化，A/D・D/A変換器の多チャネル化による実現可能性の高まり

空間音響以外での音場再現

- 音場再現は単純に空間音響としてだけではなく，複数の領域に異なる音場を合成することや，空間的な騒音制御などにも応用できる



[Ueno+ IEEE/ACM TASLP 2019]

空間音響のための信号処理技術

- 人間の音の知覚において，空間知覚は重要な要素の一つ
- 空間音響技術として，ステレオフォニック・サラウンド方式，バイノーラル合成，音場再現を紹介
- 計算機上で模擬した音場を合成する場合，収録した音場を再現する場合とで，それぞれの技術にpros and consがある
- 音場再現の技術によって，より精細かつ自由度の高いVR/AR音響を実現できる可能性

ブラインド音源分離

ブラインド音源分離とは？

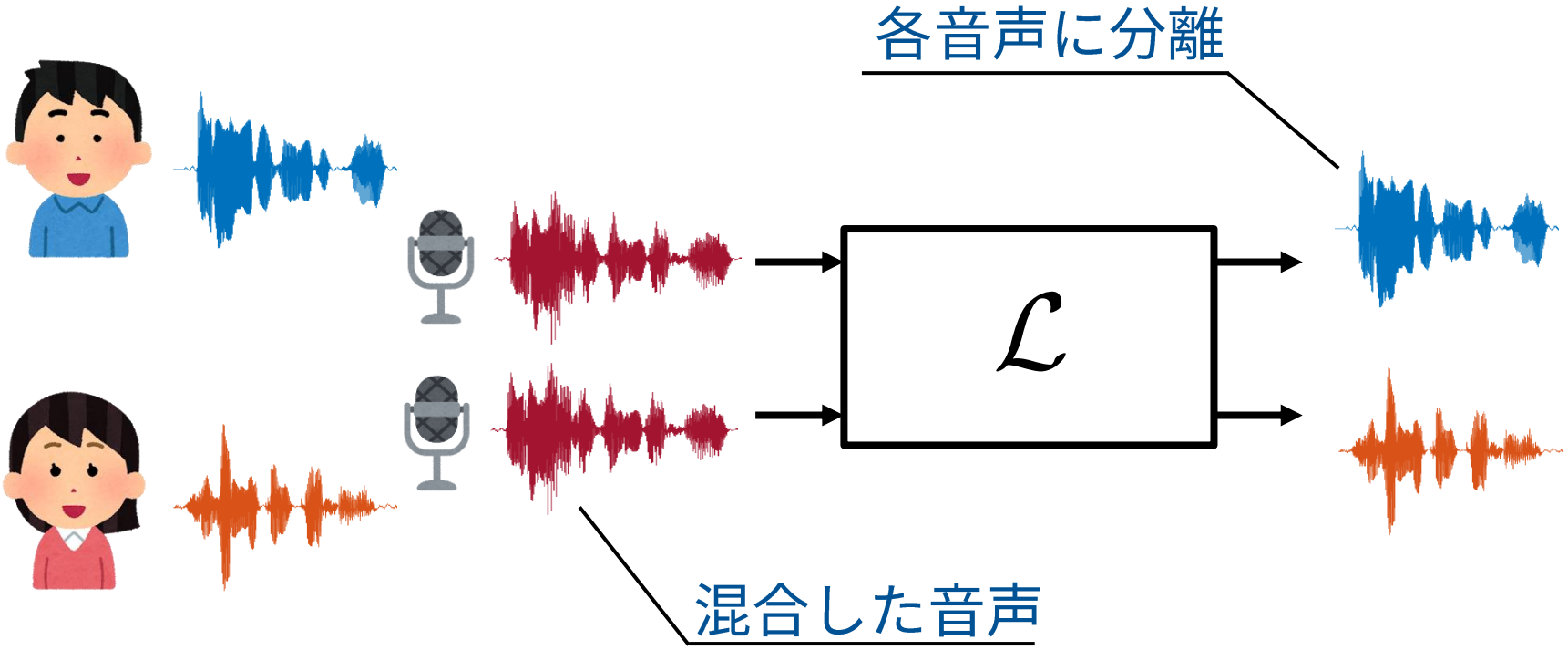
- **カクテルパーティー効果**
  - カクテルパーティのような騒がしい人混みの中でも，自分の名前や自分が興味のある会話は自然と聞き取ることができる現象。

人間はたくさんの音が混ざり合っている状況でも音声を選択的に聴取できる

ブラインド音源分離とは？

- 音の選択的な聴取をコンピュータで実現したい！

➡ ブラインド音源分離（Blind Source Separation: BSS）



音源やマイクの位置関係などの情報が未知の状態で，
（複数の）マイク信号のみから各音源の信号を分離

問題設定

- J 個の音源信号とM 個のマイクによる観測信号が，時間周波数領域で以下のように関係付けられるとする。

$$x_{ft,m} = \sum_{j=1}^{J} a_{f,mj} s_{ft,j}$$

音源信号

観測信号

伝達関数

$$\begin{cases} m & : \text{マイク} \\ j & : \text{音源} \\ t & : \text{時間} \\ f & : \text{周波数} \end{cases}$$

- 行列形式で書けば，

$$\boldsymbol{x}_{ft} = \boldsymbol{A}_f \boldsymbol{s}_{ft}$$

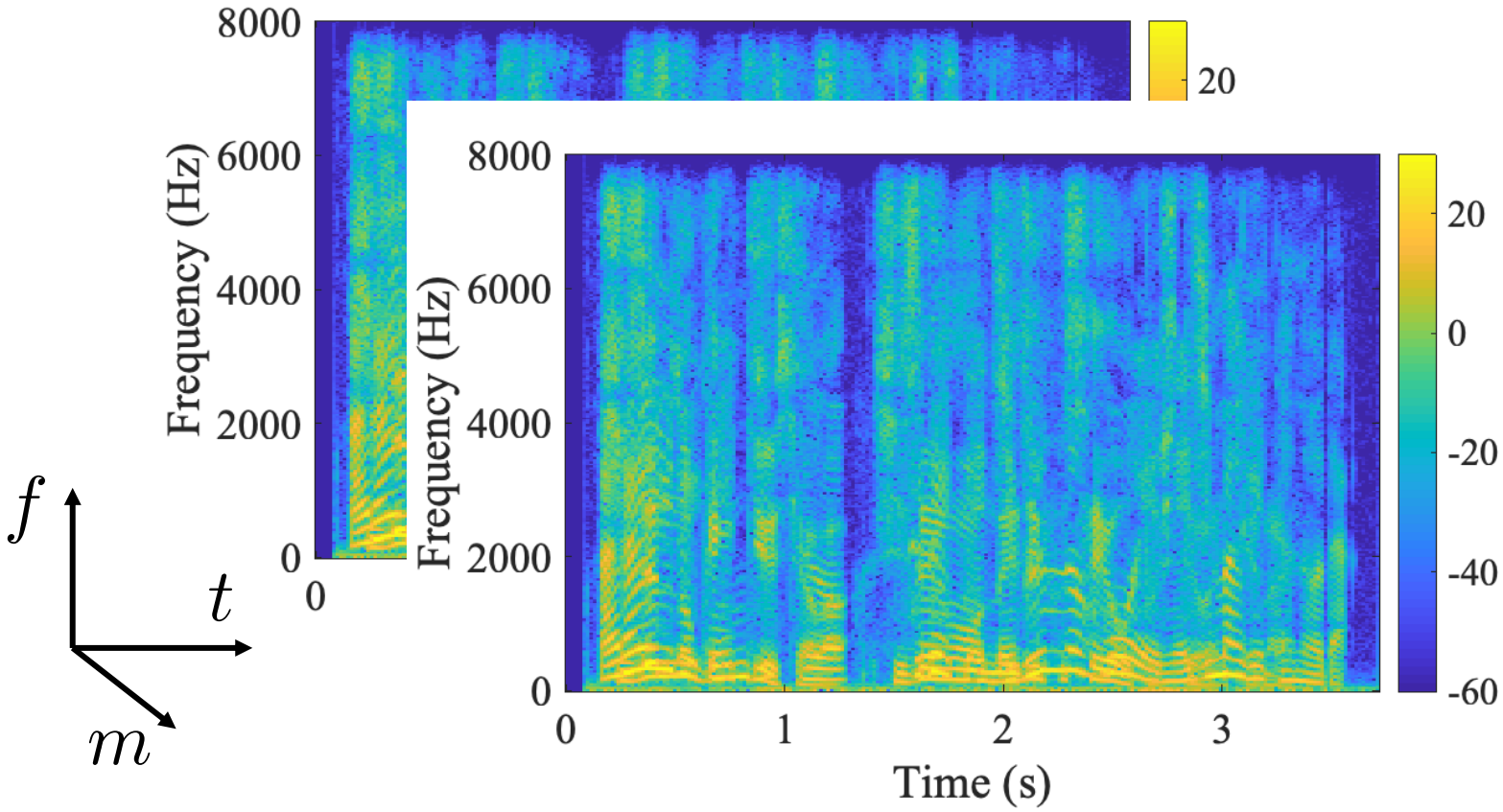
伝達関数行列（混合行列）

観測信号ベクトル

音源信号ベクトル

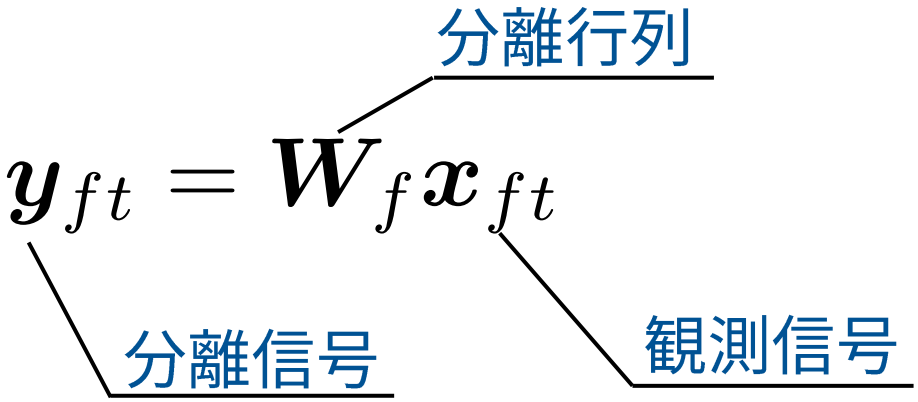
時間周波数領域表現

- 時間信号を短い時間フレームに区切ってフーリエ変換を行うことで，時間変化する信号の周波数分析を行う。
  - ➡ 短時間フーリエ変換：Short-Time Fourier Transform (STFT)



問題設定

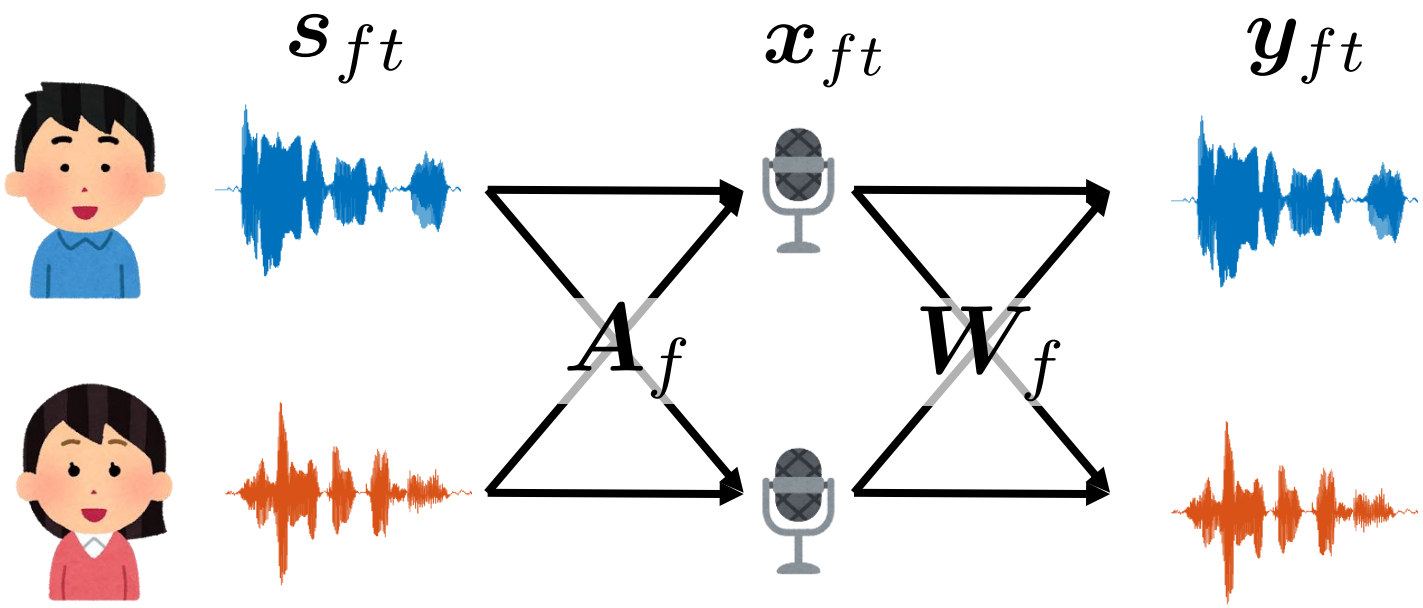
- 混合した観測信号 $\boldsymbol{x}_{ft} \in \mathbb{C}^M$ を各音源の信号 $\boldsymbol{y}_{ft} \in \mathbb{C}^J$ に分離するための，分離行列 $\boldsymbol{W}_f \in \mathbb{C}^{J \times M}$ を求めたい。



- 簡単のため，$J = M$の場合のみを考えることとし，分離行列 $\boldsymbol{W}_f$は正方行列とする。

問題設定

- 混合行列 $\boldsymbol{A}_f$ が未知の状況で，マイク信号 $\boldsymbol{x}_{ft}$ のみから各音源を分離する分離行列 $\boldsymbol{W}_f$ を推定する。



独立成分分析（ICA）

- 独立成分分析（Independent Component Analysis: ICA）では，混合された音源が**統計的に独立**であるという仮定の下で分離行列を推定する。

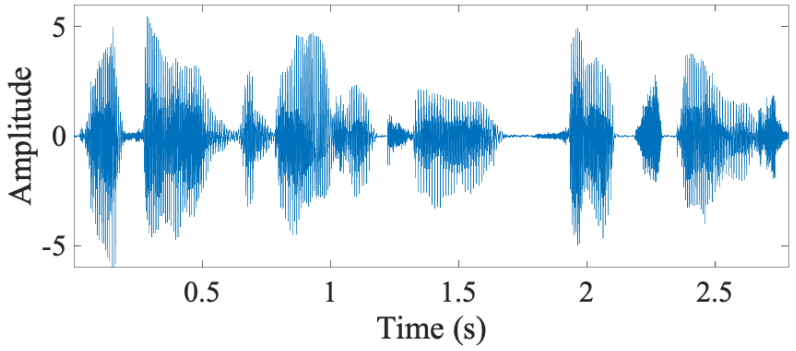
$$p(\boldsymbol{y}_{ft}) = p(y_{ft,1}, \ldots, y_{ft,J}) = \prod_{j=1}^{J} p(y_{ft,j})$$

- 統計的な独立性は無相関性/白色性よりも強い仮定であり，優ガウスな分布に従う信号源の混合を分離することを可能にする。
- ICAに基づくBSSは様々な評価尺度を用いて実現されているが，ここでは音源分離においてよく用いられる最尤推定によるICAを紹介する。

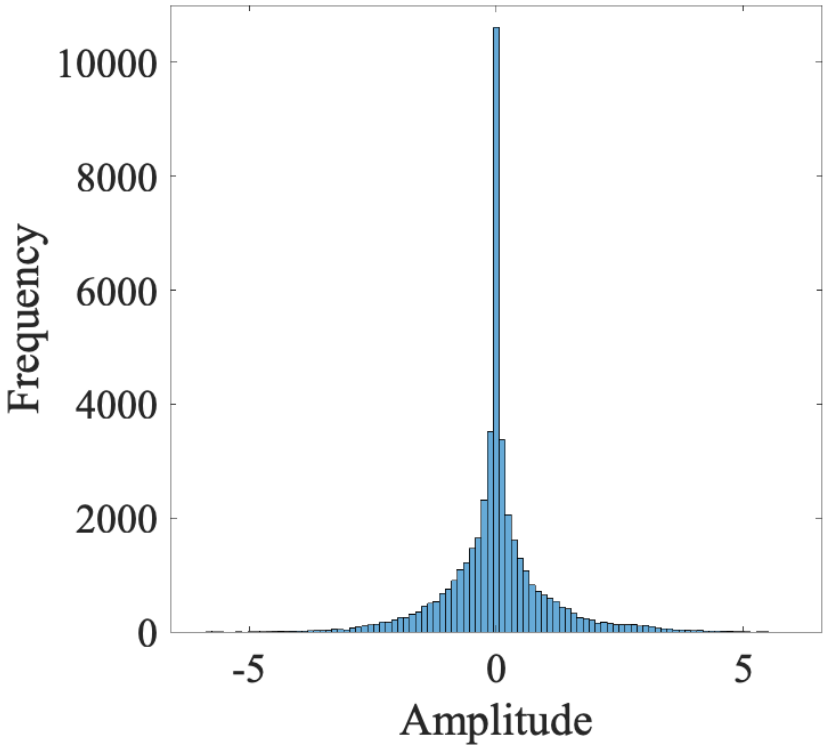
なぜ統計的独立性を使う？

- 音声の統計的性質

音声波形



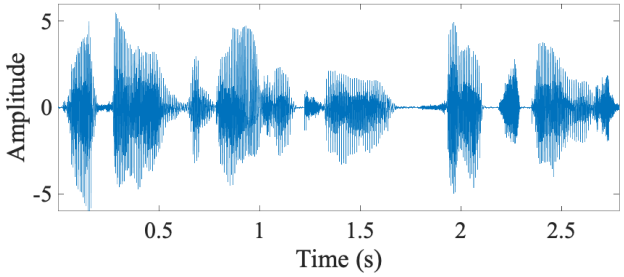
振幅値のヒストグラム



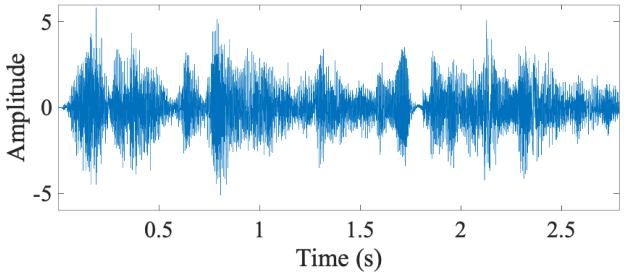
なぜ統計的独立性を使う？

- 複数の音声を混合すると・・・

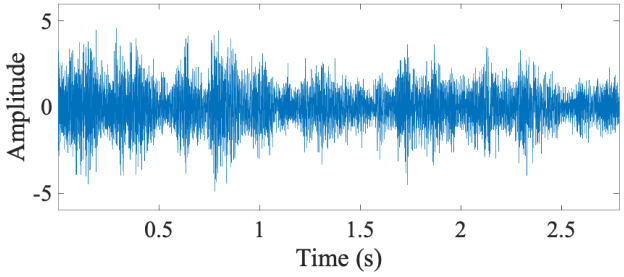
音源数 1:



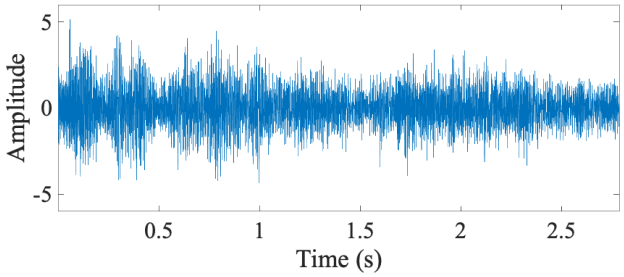
音源数 4:



音源数 8:

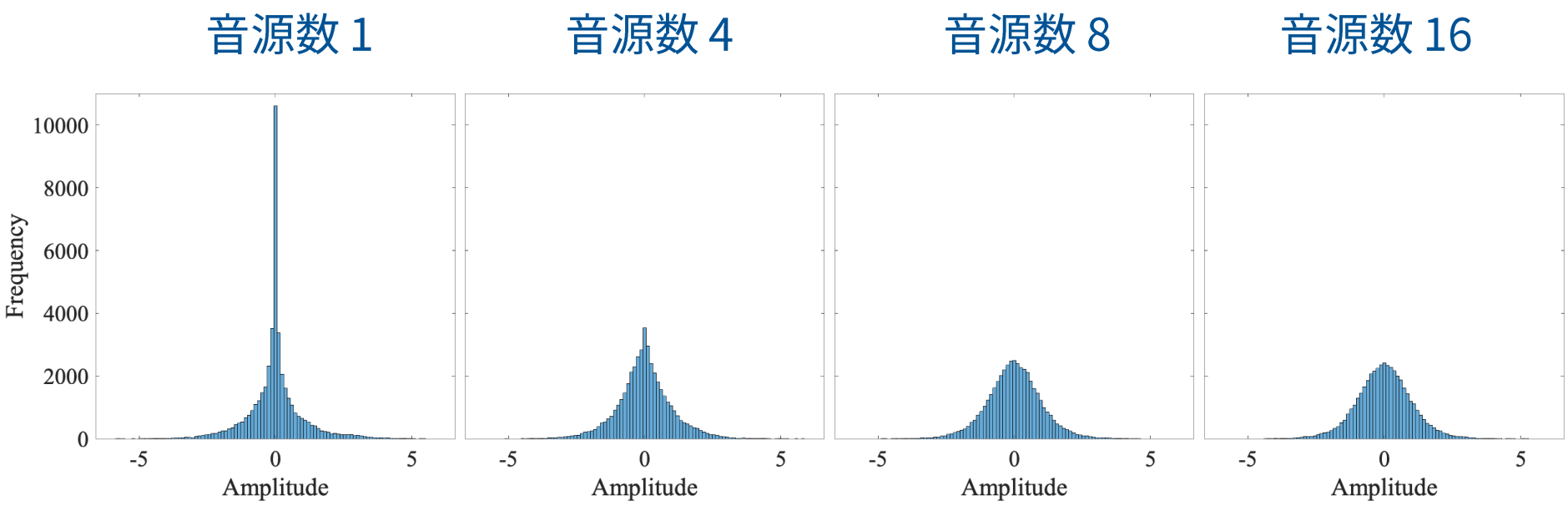


音源数 16:



なぜ統計的独立性を使う？

- 複数の音声を混合すると・・・



混合数を増やすと正規分布に近づく

分離を実現するにはこの逆を行うことが必要

最尤推定によるICA

- 分離行列 $\boldsymbol{W}_f$ の尤度関数を考えると，

$$\begin{aligned}\mathcal{L}(\boldsymbol{W}_f) &= p(\boldsymbol{x}_{f,1}, \ldots, \boldsymbol{x}_{f,T} | \boldsymbol{W}_f) \\ &= \prod_{t=1}^{T} p(\boldsymbol{x}_{ft} | \boldsymbol{W}_f) \\ &= \prod_{t=1}^{T} |\det \boldsymbol{W}_f|^2 \, p(\boldsymbol{y}_{ft})\end{aligned}$$

- ここで，線形変換 $\boldsymbol{y}_{ft} = \boldsymbol{W}_f \boldsymbol{x}_{ft}$ の確率密度に関する以下の関係式を用いた。

$$p(\boldsymbol{x}_{ft} | \boldsymbol{W}_f) = |\det \boldsymbol{W}_f|^2 \, p(\boldsymbol{y}_{ft})$$

最尤推定によるICA

- 音源信号が互いに独立であると仮定すれば，

$$p(\boldsymbol{y}_{ft}) = \prod_{j=1}^{J} p(y_{ft,j})$$

- $\boldsymbol{W}_f$ の負の対数尤度を考えると，

$$\begin{aligned}\mathcal{J}(\boldsymbol{W}_f) &= -\log \mathcal{L}(\boldsymbol{W}_f) \\ &= -\log \left\{ \prod_{t=1}^{T} |\det \boldsymbol{W}_f|^2 \prod_{j=1}^{J} p(y_{ft,j}) \right\} \\ &= \sum_{t=1}^{T} \sum_{j=1}^{J} G(y_{ft,j}) - 2T \log |\det \boldsymbol{W}_f|\end{aligned}$$

ここで，$G(y_{ft,j}) = -\log p(y_{ft,j})$ とした。

最尤推定によるICA

- $G(y_{ft,j})$はコントラスト関数と呼ばれ，音源信号が従うと仮定できる確率密度関数に基づいて設定する必要がある。
- 音声・音響信号では，優ガウスな分布として以下のような分布を用いる場合が多い。

$$p(y_{ft,j}) \propto \exp\left(-\frac{\sqrt{|y_{ft,j}|^2+\alpha}}{\beta}\right)$$

ここで，α，β は非負のパラメータ。

- コスト関数$\mathcal{J}(\boldsymbol{W}_f)$を最小化する$\boldsymbol{W}_f$を求めることで分離が達成できると考えられるが，分離行列 $\boldsymbol{W}_f$は周波数 f ごとに別々に求まるため，音源の順序に関する任意性が残ることになる。（パーミュテーション問題）

独立ベクトル分析（IVA）

- パーミュテーション問題を解決する方法の一つとして，音源信号の各要素 $y_{ft,j}$ ではなく，全周波数の要素を並べたベクトル $\boldsymbol{y}_{t,j} = [y_{1t,j}, \ldots, y_{Ft,j}]^{\mathsf{T}}$ の独立性を仮定する，独立ベクトル分析（Independent Vector Analysis: IVA）が知られている。[Hiroe+ 2006, Kim+ 2006]

- 音源信号間の独立性の仮定より，$\{\boldsymbol{y}_{t,j}\}_{j=1}^{J}$ の確率密度関数は，以下のように書ける。

$$p(\boldsymbol{y}_{t,1}, \ldots, \boldsymbol{y}_{t,J}) = \prod_{j=1}^{J} p(\boldsymbol{y}_{t,j})$$

独立ベクトル分析（IVA）

- 分離行列 $\boldsymbol{W}_f$ の尤度関数を考えると，ICAの場合と同様の定式化により，

$$\mathcal{L}(\boldsymbol{W}_1, \ldots, \boldsymbol{W}_F) = \prod_{t=1}^{T} p(\boldsymbol{x}_{t,1}, \ldots, \boldsymbol{x}_{t,M} | \boldsymbol{W}_1, \ldots, \boldsymbol{W}_F)$$

$$= \prod_{t=1}^{T} p(\boldsymbol{y}_{t,1}, \ldots, \boldsymbol{y}_{t,J}) \prod_{f=1}^{F} |\det \boldsymbol{W}_f|^2$$

$$= \prod_{t=1}^{T} \left\{ \prod_{j=1}^{J} p(\boldsymbol{y}_{t,j}) \right\} \prod_{f=1}^{F} |\det \boldsymbol{W}_f|^2$$

独立性の仮定

独立ベクトル分析（IVA）

- 分離行列 $\boldsymbol{W}_f$ の負の対数尤度は，

$$\mathcal{J}(\boldsymbol{W}_1, \ldots, \boldsymbol{W}_F) = -\log \mathcal{L}(\boldsymbol{W}_1, \ldots, \boldsymbol{W}_F)$$
$$= \sum_{t=1}^{T} \sum_{j=1}^{J} G(\boldsymbol{y}_{t,j}) - 2T \sum_{f=1}^{F} \log |\det \boldsymbol{W}_f|$$

ここで，$G(\boldsymbol{y}_{t,j}) = -\log p(\boldsymbol{y}_{t,j})$ はコントラスト関数である。

- IVAにおいても，音源信号の確率密度関数には，優ガウスな分布として以下のような分布を用いる場合が多い。

$$p(\boldsymbol{y}_{t,j}) \propto \exp\left(-\frac{\sqrt{\sum_{f=1}^{F} |y_{ft,j}|^2 + \alpha}}{\beta}\right)$$

独立ベクトル分析（IVA）

- コスト関数 $\mathcal{J}(\boldsymbol{W}_1, \ldots, \boldsymbol{W}_F)$ を最小化する $\{\boldsymbol{W}_f\}_{f=1}^{F}$ を求めればよい。

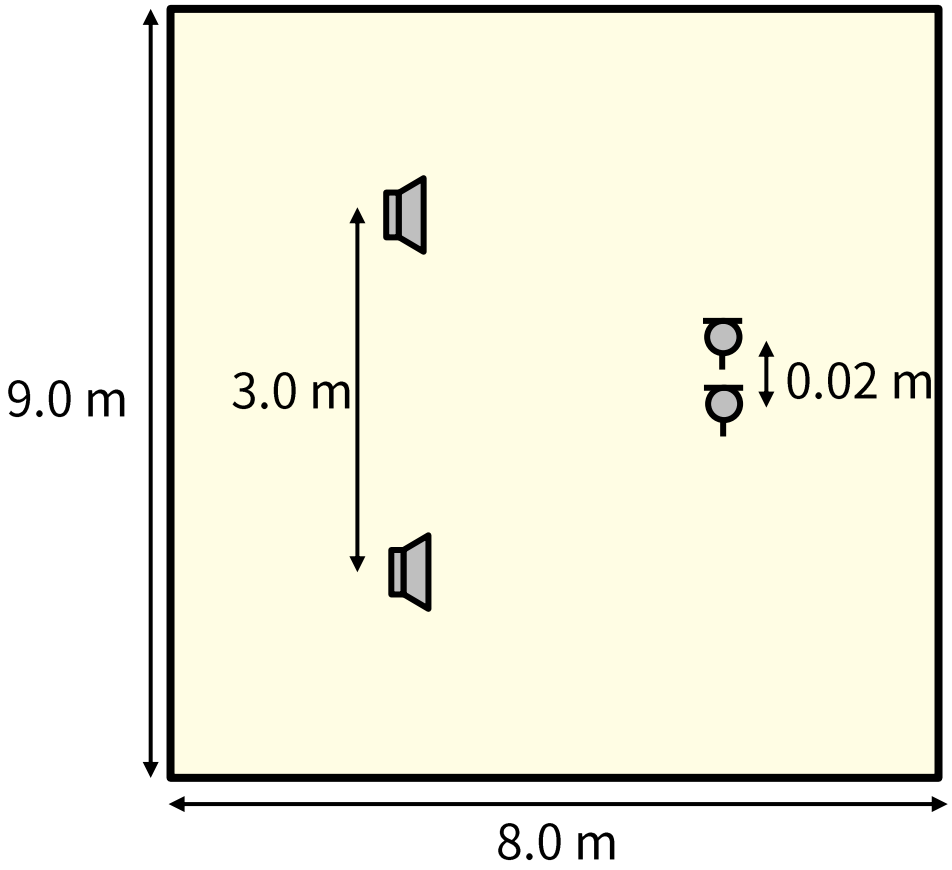
$$\underset{\boldsymbol{W}_1,\ldots,\boldsymbol{W}_F}{\text{minimize}}\, \mathcal{J}(\boldsymbol{W}_1, \ldots, \boldsymbol{W}_F)$$

$$= \sum_{t=1}^{T}\sum_{j=1}^{J} G(\boldsymbol{y}_{t,j}) - 2T\sum_{f=1}^{F} \log|\det \boldsymbol{W}_f|$$

- この最適化問題を解くための様々な最適化アルゴリズムが提案されているが，Majorization-minimization (MM) アルゴリズムと呼ばれる，非線形最適化問題を効率的に解くためのアルゴリズムがよく用いられる。 [Ono+ 2011]

シミュレーション実験例

- 残響環境下でのIVAによる２音源の分離



混合音声：

分離音声１：

分離音声２：

ブラインド音源分離

- カクテルパーティー効果のような，人間の音の選択的聴取をコンピュータで実現する
- 複数の音源からの音信号が混じり合って観測されたとして，音源やマイクなどの位置関係が未知の状態でこれらを分離する
- 各音源の統計的な独立性を手がかりに，混合信号を各音源信号に分離するフィルタを推定する

これからの音メディアとその信号処理

- 音メディアとビッグデータ・IoT
  - 深層学習は教師データが十分にあれば高い性能を発揮する
  - 音声・音響・音楽信号処理では，データ量を十分に用意できない応用先も多い（e.g. 音空間の計測）
  - 一方でセンサ・トランスデューサの数は爆発的に増え，安価になってきている（cf. トリリオン・センサ）
  - 多数のセンサによって観測される音情報をうまく活用する信号処理とは？
  - 物理音響や聴覚などの音声・音響・音楽信号における知識/方法論と，統計的機械学習などの数理的モデリング/アルゴリズムとのより高度な融合が必須

- 講義に関する質問や感想はこちらまで
  - Mail: shoichi_koyama@ipc.i.u-tokyo.ac.jp
  - Twitter: @sh01
- 講義資料はこちら
  - http://www.sh01.org/ja/teaching/

参考文献

1. 日本音響学会編, “音響学入門ペディア,” コロナ社, 2017.
2. Blauert, “Spatial Hearing: The Psychophysics of Human Sound Localization,” MIT Press, 1997.
3. 飯田, 森本, “空間音響学,” コロナ社, 2010.
4. Hacihabiboglu, et al., “Perceptual spatial audio recording, simulation, and rendering,” IEEE Signal Processing Magazine, 2017.
5. 小山, “音場再現技術における数理問題：波面合成・高次アンビソニックスの数理,” 日本音響学会誌, 2012.
6. 小山, “未来の音の収録・再生・編集技術の実現に向けて,” 電子情報通信学会誌, 2017.
7. Hyvärinen, et al., “独立成分分析 – 信号解析の新しい世界,” 東京電機大学出版局, 2007.
8. Sawada, et al. “A review of blind source separation methods: two converging routes to ILRMA originating from ICA and NMF,” *APSIPA Trans. SIP*, 2019.