修士論文

音響を用いた位置推定に関する研究

A Study of Acoustic-based Location Estimation

指導教員　工藤知宏 教授

東京大学大学院 工学系研究科
電気系工学専攻 融合情報コース
37-206486 鈴木優大

令和 4 年 1 月 27 日 提出

概要

近年，ロボット家電の普及や介護現場での見守り機器導入などを背景に，屋内における人間や物体の位置推定に対する需要が高まっている．位置推定には様々な手法があるが，その一つに音声を用いた位置推定法があげられる．かねてより，複数音源のそれぞれの位置推定は，主に音声認識機器やロボットへの搭載など様々な用途で研究されてきた．さらに昨今のコロナ禍においては，教室や会議室での複数の発話者の距離測定を通じた感染症対策への活用が考えられる．また，映像を用いる場合と比べ，個人のプライバシーがより守られるという点や，処理するデータ量が大幅に少ないという点が大きなメリットとして挙げられる．音源の位置推定には多様な手法が提案されているが，その中の一つに複数のマイクロフォンへの音声の到達時間差 (TDoA: Time Difference of Arrival) を測定し，マイクロフォンの座標情報と合わせて位置を推定する方法がある．TDoA の測定手法の一つとして，二つの入力信号の相関を取り，その相関が最も高くなる時間のずれを TDoA として検出する相互相関法が考案された．さらに計算量の減少による高速化や精度の向上を目的とした，入力信号を白色化した上で相関を求める CSP 法などさまざまな研究がなされてきた．本論文ではそのような CSP 法が，音源が複数ある場合や残響が存在する環境下で精度が低下するという問題に対して，精度の向上手法を提案した．まず，複数音源の場合では，収録音声に対して音源分離法である ILRMA を用いることによる CSP 法の精度向上手法を提案し，シミュレーションにより効果を確認した．実験の結果，検出精度が大きく向上することを確認した．また，音源同士のピークが近い場合では，CSP のみを用いた場合に発生するピークの消失が，音源分離を組み合わせることによって改善された．このように，音源同士の TDoA が近い場合にとくに音源分離が有効であることがわかった．次に，残響が存在する環境下での TDoA の精度の向上のための残響除去フィルタを提案した．チャープパルスなどの特定の音源を用いて RIR(Room Impulse Reaction) を推測し，それを残響成分と遅延成分にわけることで，収録音声から残響成分だけを除く残響除去フィルタを提案し，シミュレーションにより CSP 法の精度が大きく向上することを確認した．

目次

図目次

表目次

第 1 章

序論

1.1 研究背景

近年，ロボット家電の普及や介護現場での見守り機器導入などを背景に，屋内における人間や物体の位置推定に対する需要が高まっている．位置推定には様々な手法があるが，その一つに音声を用いた位置推定法があげられる．かねてより，複数音源の位置推定は，主に音声認識機器やロボットへの搭載など様々な用途で研究されてきた [1–3]．さらに昨今のコロナ禍においては，教室や会議室での複数の発話者の距離測定を通じた感染症対策への活用が考えられる．また，映像を用いる場合と比べ，個人のプライバシーがより守られるという点や，処理するデータ量が大幅に少ないという点が大きなメリットとして挙げられる．音源の位置推定には多様な手法が提案されているが [4–6]，その中の一つに複数のマイクロフォンへの音声の到達時間差 (TDoA: Time Difference of Arrival) を測定し，マイクロフォンの座標情報と合わせて位置を推定する方法がある [7, 8]．マイクロフォンのサンプリングデータから TDoA を計算する素朴なアプローチとして，一対のマイクロフォンのデータ間で相互相関関数を計算しそのピーク位置から TDoA を検出する手法があげられる．しかし相互相関法は 2 つの受信点で得られた波形の単純な積和演算であるため，特定の周波数成分の振幅が大きい場合，その周波数に大きく依存した受信時間差を得ることになるという欠点がある．また，単純な計算ゆえに計算量が多く，時間がかかるという欠点も存在する．そこで周波数スペクトルを白色化して相関関数を計算する Cross-power Spectrum Phase (CSP) 法 [9] が提案された．CSP 法では，一対のマイクロフォンペアへの入力信号に一度フーリエ変換を施し，それぞれを掛け合わせるとともにそれぞれの絶対値で除算を行なった上で逆フーリエ変換を施す．このようにして，各周波数成分を白色化（正規化）しすべての周波数成分の振幅を揃えることで，特定の周波数成分に依存しない受信時間差の推定を行うことができるため TDoA の検出法として現在主流となっている．しかし，複数の音源が存在する状況や，残響の効果が大きく影響する環境下では TDoA の検出精度が下がるという欠点がある [8, 10–12]．

1.2 研究目的

本研究では，CSP 法の精度を低下させる二つの要因である，音源が複数存在する場合と，残響が存在する場合のそれぞれについて，音源の位置推定の方法としての CSP 法によるマイクロフォンアレイへの音声の到達時間差の推定精度の向上を目的とする．

1.3 本論文の構成

第二章では，到達時間差の検出や音源分離，残響除去に関する先行研究について説明する．第三章では，研究に用いたシミュレーションの詳細や，実験環境及び実験に用いた道具などの実験環境について詳しく説明する．第四章では，実際に複数の音源や残響が存在する環境下での，CSP 法の精度向上のための手法について説明する．第五章では，シミュレーションや実際の実験の結果について述べる．第六章で，まとめや結果に対する考察を述べる．

第 2 章

関連研究

2.1 到達時間差の検出

2.1.1 相互相関法

相互相関法は，マイクロフォンペアへの二つの受信データの時間差を求める最も一般的な方法である．各受信データを $x_i(t)$，$x_j(t)$ とすると，相互相関関数は，

$$\mathrm{Corr}(t) = \mathcal{F}^{-1}\left[\mathcal{F}[x_i(t)]\overline{\mathcal{F}[x_j(t)]}\right] \tag{2.1}$$

と定義される．ここで $\mathcal{F}[]$，$\mathcal{F}^{-1}[]$ はフーリエ変換・逆変換，$\bar{c}$ は c の複素共役である．いま減衰等の効果を無視し，x_i は x_j より Δt 遅れて信号が到達するものとすると， Corr は $t = \Delta t$ にピークが立つ．従って (2.1) 式の計算から TDoA を推定することができる．

2.1.2 白色化相互相関法

前項で述べた相互相関法は, 2 つの受信点で得られた波形の単純な積和演算であるため, 特定の周波数成分の振幅が大きい場合，その周波数に大きく依存した受信時間差を得ることになる．この改善策として白色化相互相関法（CSP：Cross-power Spectrum Phase) が提案されている．この方法は GCC-PHAT 法 (Generalized Cross Correlation PHAse Trasform) とも呼ばれる．各周波数成分を白色化（正規化）することにより，すべての周波数成分の振幅を揃えることができ，特定の周波数成分に依存しない受信時間差の推定を行うことができる．CSP 法の式は以下の (2.2) 式によって定義される．

$$\mathrm{CSP}(\tau) = \mathcal{F}^{-1}\left[\frac{\mathcal{F}[x_i(t)]\overline{\mathcal{F}[x_j(t)]}}{|\mathcal{F}[x_i(t)]||\mathcal{F}[x_j(t)]|}\right] \tag{2.2}$$

CSP 法の算出結果は相互相関法と同様であり，最も高い値を示した時間が参照点とした受信点と対象とした受信点との受信時間差である．

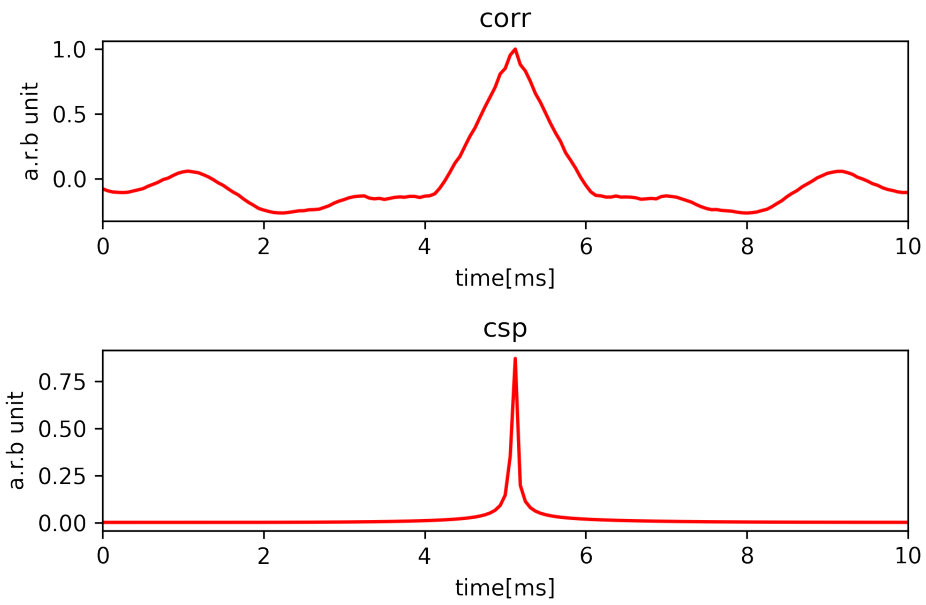

図 2.1: 相互相関法と CSP 法の比較

2.1.3 手法の比較

相互相関法と CSP 法を，音響シミュレーションを用いて比較した．無残響の屋内に音源を 1 つ，マイクロフォンを 2 つ配し，マイクロフォン間の TDoA を計算した．シミュレーションのセットアップは次章で述べる．それぞれの手法での，相関の値のグラフを図 2.1 に示す．検出された TDoA の値は変わりないが，CSP 法の方がよりシャープにピークを検出できることが確認できる．よって本研究では，到達時間差の検出に CSP 法を用いている．

2.1.4 CSP 法の課題

CSP 法では，音源が複数ある場合や残響が存在する場合は，精度が低下するという欠点がある [8]．以下に，その理由を示す．

単一音源の場合

まず，単一音源を 2 つのマイクで計測する場合について述べる．対象とする音源信号を $s(t)$，伝搬による減衰係数を α_i，伝搬による遅延時間を t_i とすると，マイクロフォン i で受信される信号 $x_i(t)$ は，

$$x_i(t) = \alpha_i s(t - t_i) \tag{2.3}$$

となる．$s(t)$ のフーリエ変換が

$$S(\omega) = b(\omega)\epsilon^{-j\psi(\omega)} \tag{2.4}$$

と表されるとき，$x_i(t)$，$x_j(t)$ のフーリエ変換は

$$X_i(\omega) = \alpha_i \epsilon^{-j\omega t_i} b(\omega) \epsilon^{-j\psi(\omega)} \tag{2.5}$$

$$X_j(\omega) = \alpha_j \epsilon^{-j\omega t_j} b(\omega) \epsilon^{-j\psi(\omega)} \tag{2.6}$$

となるので，これらの式を式 (2.2) に代入して整理すると，次式のようになる．

$$\begin{aligned} \mathrm{CSP}_{ij}(\tau) &= \mathcal{F}^{-1}[\varepsilon^{j\omega(t_j - t_i)}] \\ &= \mathcal{F}^{-1}[\varepsilon^{j\omega\tau_{ij}}] \\ &= \delta(\tau - \tau_{ij}) \end{aligned} \tag{2.7}$$

(2.7) 式は τ_{ij} でピークを持つデルタ関数となることから，CSP がピーク値を取る τ を求めることで，TDoA を推定することができる．

複数音源が存在する場合

つぎに，音源が複数存在する場合について説明する．対象とする音源信号を $s_a(t)$，$s_b(t)$ とすると，2 つの観測信号 $x_1(t)$，$x_2(t)$ は，

$$x_1(t) = \alpha_{a1} s_a(t - t_{a1}) + \alpha_{b1} s_b(t - t_{b1}) x_2(t) = \alpha_{a2} s_a(t - t_{a2}) + \alpha_{b2} s_b(t - t_{b2}) \tag{2.8}$$

と表される．ただし，α は各音源とマイクロフォン間での減衰，t_{a1}，t_{b1}，t_{a2}，t_{b2} は各音源とマイクロフォン間の音波伝搬時間である．ここで，音源信号のフーリエ変換を $S_a(\omega)$，$S_b(\omega)$ とすると，(2.2) 式で表される計算の要素の一部は次式で表される．

$$\begin{aligned} &X_1(\omega) X_2^*(\omega) \\ &\quad = (\alpha_{a1} S_a(\omega) \epsilon^{-j\omega t_{a1}} + \alpha_{b1} S_b(\omega) \epsilon^{-j\omega t_{b1}}) \times (\alpha_{a2} S_a^*(\omega) \epsilon^{-j\omega t_{a2}} + \alpha_{b2} S_b^*(\omega) \epsilon^{-j\omega t_{b2}}) \\ &\quad = \alpha_{a1}\alpha_{a2} |S_a(\omega)|^2 \epsilon^{-j\omega(t_{a1} - t_{a2})} \\ &\qquad + \alpha_{b1}\alpha_{b2} |S_b(\omega)|^2 \epsilon^{-j\omega(t_{b1} - t_{b2})} \\ &\qquad + \alpha_{a1}\alpha_{b2} \epsilon^{-j\omega(t_{a1} - t_{b2})} S_a(\omega) S_b^*(\omega) \\ &\qquad + \alpha_{a2}\alpha_{b1} \epsilon^{-j\omega(t_{a2} - t_{b1})} S_a^*(\omega) S_b(\omega) \end{aligned} \tag{2.9}$$

$S_a(\omega)$ と $S_b(\omega)$ はともに音声信号であるため，その振幅は同じような周波数特性を持つと仮定すると，(2.9) 式の前半 2 項は，逆フーリエ変換によって各音源の TDoA でピークを持つデルタ関数になる．ただし，厳密には $|S_a(\omega)| \neq |S_b(\omega)|$ デルタ関数にはならず各音源の TDoA をピークとし，その周辺に裾野を持つことになる．ここでは，デルタ関数に近似できるものとする．しかし，これらの項のほかに，$S_a(\omega)$ と $S_b(\omega)$ のクロス項が存在する．これらが完全に直交しない場合には，このクロス項によって，本来検出されるべき TDoA 以外にもピークが発生することがある．

また，残響についてもある種複数音源と見なすことができるため，同じような理由で CSP 法の精度が低下するという問題がある．実際にシミュレーション上で，CSP 法による TDoA の検出において，相関の値のグラフにおいてピークが乱立する例を，以下の図 2.2 に示す．

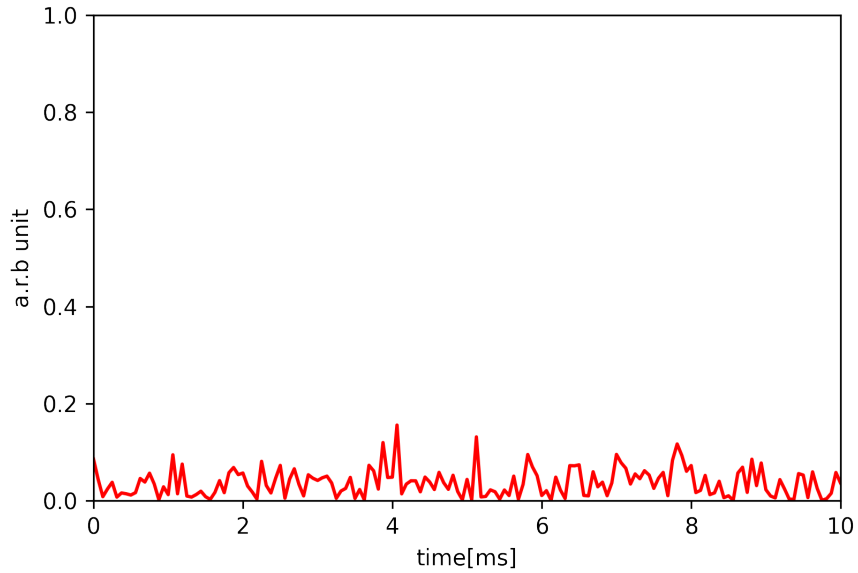

図 2.2: 相互相関法と CSP 法の比較

2.2 ブラインド音源分離

2.2.1 音源分離の概要

音源分離とは，複数音源からの音声が混合されたマイクロフォンアレイへの入力信号をもとに，それぞれの音源信号を再現する技術のことである．実生活において例えば音声認識などのシステムにおいて、複数の話者が同時に喋り出すといったような場合には，音声認識は困難になる．また，周囲の雑音や残響も不要な音となる．このような場合、目的の音のみを抽出する必要があり，多くの研究がなされている．音源分離の主流はブラインド音源分離と呼ばれる，マイクロフォン配置や目的音の到来方向に関する事前知識を使用しない音源分離法である．音源に対して統計的な過程を置くことで，最尤推定法を用いて音源モデルや空間モデルのパラメータを推定する技術が提案されている．なお，記述に当たっては [13] を参考にした．

2.2.2 入力信号のモデル化とステアリングベクトル

まず，本章で扱う入力信号 $\boldsymbol{x}_{lk}$ をモデル化する．$\boldsymbol{x}_{lk}$ は M 個のマイクロフォンの信号からなるベクトルであり，l はフレーム数，k は周波数インデックスである．$\boldsymbol{x}_{lk}$ は入力信号を短時間フーリエ変換した後の信号である．また，音源は一つと仮定すると，以下の (2.10) 式のようにモデル化できる．

$$\boldsymbol{x}_{lk} = s_{lk}\boldsymbol{a}_k + \boldsymbol{n}_{lk} \tag{2.10}$$

第一項が音源に関する項となり，第二項は雑音信号である．マイクロホンで観測された $\boldsymbol{x}_{lk}$ から s_{lk} を推定することが音源分離の目的であり，ここで第一項に存在する $\boldsymbol{a}_k$ をステアリングベクトルといい，s_{lk} の推定

にとても重要な役割を果たす．ステアリングベクトルは音源からマイクロフォンまで音が伝わる際の減衰 · 遅延を表すベクトルであり，複素ベクトルとなる．このステアリングベクトルの m 番目の要素 $a_{m,k}$ は一般的に次のように表現することができる．

$$a_{m,k} = r_m \exp(-j2\pi f_k \tau_m) \tag{2.11}$$

ここで，$j = \sqrt{-1}$，f_k は k 番目の周波数 [Hz] とする．f_k はサンプリングレートを F_s，短時間フーリエ変換のフレームサイズを N として以下のように書ける．

$$f_k = \frac{kF_s}{N} \tag{2.12}$$

k は 0 から $N/2$ までの整数とし，τ_m は音源からマイクロフォンまで音が伝搬するまでにかかる時間，r_m は減衰率となる．ブラインド音源分離においては，このステアリングベクトルを推定することが重要となる．

2.2.3 ICA

ブラインド音源分離として最も代表的なものが，独立成分分析 (ICA：Independent Component Analysis) [14] である．まずマイクロフォンの入力信号を以下の式 (2.13) のようにモデリングする．なお，ここでは雑音は一旦考慮せず議論を進める．

$$\begin{aligned} \boldsymbol{x}_{lk} &= \sum_{i=1}^{N} s_{ilk} \boldsymbol{a}_{ik} \\ &= \boldsymbol{A}_k \boldsymbol{s}_{lk} \end{aligned} \tag{2.13}$$

ここで N は音源数であり，$\boldsymbol{a}_{ik}$ は i 番目の音源のステアリングベクトルとなる．確率変数は各音源の信号 s_{ilk} で，これが最終的に推定したい変数になる．$\boldsymbol{s}_{lk}$ は，各音源の信号の推定値を要素に持つベクトルである．マイクロフォン素子数 M と音源数 N が等しいと仮定すると，式 (2.13) は次のように変換可能である．

$$\boldsymbol{s}_{lk} = \boldsymbol{W}_k \boldsymbol{x}_{lk} \tag{2.14}$$

ここで $\boldsymbol{W}_k$ を音源分離フィルタと呼び，さらに音源 s_{ilk} を推定するための分離フィルタを $\boldsymbol{w}_{ik}^H$ として，$\boldsymbol{W}_k$ は次の (2.15) のように書ける．

$$\boldsymbol{W}_k = (\boldsymbol{w}_{1k}^* \cdots \boldsymbol{w}_{Nk}^*)^T \tag{2.15}$$

$\boldsymbol{W}_k$ は $\boldsymbol{A}_k$ の逆行列であり，これを求めることで $\boldsymbol{s}_{lk}$ を推定することができる．$\boldsymbol{A}_k$ は未知のため，独立成分分析ではマイクロフォン入力 $\boldsymbol{x}_{lk}$ から $\boldsymbol{W}_k$ を求めるために，分離信号がうまく分離されているとすると，各音源の信号は統計的に独立になっているという信号源の独立性に依拠したモデルを導入する．よって分離フィルタ $\boldsymbol{W}_k$ を用いて分離された $\boldsymbol{s}_{lk}$ の確率密度関数は，(2.16) 式のように書ける．

$$p(\boldsymbol{s}_{lk}) = \prod_{i=1}^{N} p(s_{ilk}) \tag{2.16}$$

この式が成立するように，具体的には (2.16) 式の右辺と左辺の乖離が小さくなるようにパラメータ $\boldsymbol{W}_k$ を推定していく．具体的なアプローチとして，カルバックライブラ情報量という距離尺度と，尤度関数を用いる方法の二つを述べる．

カルバックライブラ情報量

カルバックライブラ情報量 (KLD：Kullback－ Leibler Divergence) は式 (2.17) のように表すことができる．

$$\begin{aligned}
\mathrm{KLD}(p\|q) &= \int_s p(s) \log \frac{p(s)}{q(s)} ds \\
&= \int_s p(s) \log p(s) ds - \int_s p(s) \log q(s) ds \\
&= \int_x p(x) \log p(x) dx - 2 \log \|\mathrm{det} W\| - \sum_i \int_x p(x) \log p(s_i = w_i^H x) dx
\end{aligned} \tag{2.17}$$

ここで第一項は分離フィルタに依存しないので，残りの項を最小化する．

$$\int_x p(x) f(x) dx = \frac{1}{L} \sum_l f(x_l) \tag{2.18}$$

よって最終的に最適化したいパラメタを含む KLD は，以下の (2.19) 式のようになる．

$$\mathrm{KLD}(p\|q) = -\frac{1}{L} \sum_{l=1}^{L} \sum_{i=1}^{N} \log p(w_i^H x_l) - 2 \log \|\mathrm{det} W\| \tag{2.19}$$

このようにして求めた KLD を最小化するようにパラメータ $\boldsymbol{W}_k$ を更新していく．

尤度関数

確率変数の変換を用いて，マイクロフォン入力信号の尤度関数を各音源信号の尤度関数に変換し，それらが向上するようにパラメータを更新する．変数が複素数であることに注意すると，確率変数の変換による対数尤度関数の変換は以下のように実行できる．

$$\sum_{l=1}^{L} \log p(\boldsymbol{x}_l|\theta) = \frac{1}{L} \sum_{l=1}^{L} \log \sum_{i=1}^{N} \log p(s_i = \boldsymbol{w}^H \boldsymbol{x}_l|\theta) + 2 \log |\mathrm{det} \boldsymbol{W}| \tag{2.20}$$

$p(s_i|\theta)$ は音源信号 s_i の尤度関数であり，音声の統計的な特徴をモデル化した音源モデルとなる．音声の音源モデルでは，ガウス分布と比べて裾野が長く小さい値を取る確率が大きいスーパーガウス分布に属するモデルを用いることが多く，特にラプラス分布がよく用いられる．また近年ではガウス分布の一種で裾野が長い分

布を表現できる時変ガウスモデルもよく用いられる．
ICA では，尤度関数の対数を取ったものを扱うため，(2.21) 式のような負の対数尤度関数がモデリングされ，これをコントラスト関数という．これが小さいほど音声らしいという事を表す．

$$\boldsymbol{G}(s_{ilk}) = -\log p(s_{ilk}|\theta_k) \tag{2.21}$$

また，音声らしさが位相成分よりも振幅成分に特徴的に現れる事を考慮し，振幅成分のみに依存する球状の関数をコントラスト関数として用いることが一般的である．ラプラス分布に基づく球状のコントラスト関数は，以下の (2.22) 式のようになる．

$$\boldsymbol{G}(s_{ilk}) = \boldsymbol{G}(|s_{ilk}|) = C|s_{ilk}| \tag{2.22}$$

$C = 2$ としてこのコントラスト関数より求められたコスト関数を用いて，パラメータ最適化問題は以下のようにかける．

$$\text{minimize} \qquad \sum_{k=1}^{K} \frac{1}{L} \sum_{l=1}^{L} \sum_{i=1}^{N} 2|\boldsymbol{w}_{i,k}^H \boldsymbol{x}_{lk}| - 2\log|\text{det}\boldsymbol{W}_k| \tag{2.23}$$

ICA では，周波数ごとに分離フィルタを更新するため，以下の議論でもそれを前提とする．周波数ごとのコスト関数を $\boldsymbol{w}_k^*$ で微分し 0 とおくと，

$$\frac{\partial \sum_{l=1}^{L} \sum_{i=1}^{N} 2|\boldsymbol{w}_{i,k}^H \boldsymbol{x}_{lk}| - 2\log|\text{det}\boldsymbol{W}_k|}{\partial \boldsymbol{w}_k^*} = \frac{\boldsymbol{x}_{lk}\boldsymbol{x}_{lk}^H \boldsymbol{w}_{i,k}}{|\boldsymbol{w}_{i,k}^H \boldsymbol{x}_{lk}|} - \boldsymbol{W}_k^{-1} e_i = 0 \tag{2.24}$$

となる．ここで e_i は i 番目の要素のみが 1 で他が全て 0 となる要素数 M のベクトルである．これは $w_{i,k}$ について解くことができないため，自然勾配法のような勾配が小さくなる方に W_k を少しずつ変化させるアプローチをとる．自然勾配法によるパラメータ更新式を以下に示す．

$$\boldsymbol{W}_k^{(t+1)} = \boldsymbol{W}_k^(t) + \mu \left(\boldsymbol{I} - \frac{1}{L} \sum_l \Phi(\boldsymbol{s}_l k) \boldsymbol{s}_l k^H \right) \boldsymbol{W}_k^(t) \tag{2.25}$$

また，この式において第二項の係数の行列の対角項は本質的に無意味であるため，その影響を無視した

$$\boldsymbol{W}_k^{(t+1)} = \boldsymbol{W}_k^(t) + \mu \left(\text{offdiag} - \frac{1}{L} \sum_l \Phi(\boldsymbol{s}_l k) \boldsymbol{s}_l k^H \right) \boldsymbol{W}_k^(t) \tag{2.26}$$

といった更新の仕方もよく用いられ，非ホロノミック拘束型の更新法と呼ぶ．ここで，offdiag は対角項を 0 にする処理とする．このようにして，分離フィルタの更新を行うことができる．これらのように，信号源の独立性を満たすような分離フィルタを求めることが，独立成分分析という名前の由来となっている．一方，このような方法では，周波数ごとに複数の音源を分離することを目的としていることから，周波数ごとにフィルタ $\boldsymbol{W}_k$ を独立に求めている．そのため，さらに二つの問題を解く必要がある．それが周波数間パーミュテーション問題と出力信号の音量の不定性問題である．

周波数間パーミュテーション問題

ICA では，周波数ごとに複数の音源が出てくるが，その順番は不定である．そのため，ある周波数で一番目の信号として出てきた信号が，他の周波数でも同様に一番目の信号として出てくるわけではない．このように，周波数間で出力される信号のインデックスが不定になるという問題を，周波数間パーミュテーション問題という．これの解法には，音源の方向推定を用いる方法などもあるが，ここでは音量の時間パターンの類似性を利用して解く方方法を述べる [15, 16]．まず，独立成分分析の出力結果を s_{ilk} とする．また，z_{ilk}=$|s_{ilk}|$ とする．これを音源方向に次のように正規化する．

$$\bar{z}_{ilk} = \frac{z_{ilk}}{\sum_{j=1}^{N} z_{jlk}} \tag{2.27}$$

さらに，周波数ごとに $\bar{z}_{ilk}$ の音源間での相関 $\boldsymbol{C}_k$ を次のように計算する．

$$\boldsymbol{C}_k = \sum_{l=1}^{L}\sum_{i=1}^{N}\sum_{j=1}^{N} \bar{z}_{ilk}\bar{z}_{jlk} \tag{2.28}$$

この相関が高いほど音源間で似ているということを示しているため，分離がうまくいっていない周波数であるということがわかる．ここで，分離がうまくいっている周波数，つまり $\boldsymbol{C}_k$ が最も小さい周波数を基準として設定し，その周波数成分の情報を用いてパーミュテーションを解いていくことを考える．$k(m)$ を m 番目の $\boldsymbol{C}_k$ が小さい周波数成分とする．基準の振幅スペクトルパターン A_{il} を以下のようにして算出する．

$$A_{il} = \bar{z}_{ilk(m=1)} \tag{2.29}$$

そして，次に分離がうまくいっている $k(m=2)$ の周波数成分について，

$$\sum_{l=1}^{L}\sum_{i=1}^{N} A_{il}\bar{z}_{p(i)lk(m=2)} \tag{2.30}$$

が最大となるようにパーミュテーション $p(i)$ を決めていく．次に，$p(i)$ を決めた後，

$$A_{il} = A_{il} + \bar{z}_{p(i)lk(m=2)} \tag{2.31}$$

で基準の振幅スペクトルパターンを更新する．上記の処理を $m = 3, 4 \ldots$ についても繰り返し行う．分離がうまくいっている周波数から順番にパーミュテーションを解きつつ，基準となる振幅スペクトルパターンを更新する．最終的に振幅スペクトルパターンが周波数間でできるだけ似るようにパーミュテーションが解かれることになる．

プロジェクションバック

分離フィルタを求める過程で，分離フィルタの大きさを考慮していないため，音源分離フィルタ $\boldsymbol{W}_k$ には大きさの不定性が存在する．そこで不定性の問題を解消するために，マイクロフォン位置で得られる信号の中

にある成分を音源ごとに分離するように，分離信号に各音源のステアリングベクトルをかけてから信号を出力することを考える．つまり，

$$\boldsymbol{x}_{lk} = \sum_i \boldsymbol{c}_{ilk} \tag{2.32}$$

として，マイクロフォン入力信号中の音源信号 $\boldsymbol{c}_{ilk}$ を得ることができると考える．このような処理をプロジェクションバックと呼ぶ．$\boldsymbol{W}_k$ の逆行列が各音源のステアリングベクトルを列ベクトルとして持つ行列 $\boldsymbol{A}_k$ になることから，分離フィルタの逆行列 $\boldsymbol{W}_k^{-1}$ を用いればよい．つまり，

$$\begin{aligned} \boldsymbol{x} &= \boldsymbol{W}_k^{-1} \underbrace{\boldsymbol{W}_k \boldsymbol{x}_{lk}}_{=\boldsymbol{s}_{lk}} \\ &= \sum_{i=1}^{N} \boldsymbol{a}_{ik} s_{ilk} \end{aligned} \tag{2.33}$$

とできることに着目する．ここで，$\boldsymbol{a}_{ik}$ は $\boldsymbol{W}_k^{-1}$ の i 番目の列ベクトルとする．したがって，

$$\boldsymbol{c}_{ilk} = \boldsymbol{a}_{ik} s_{ilk} \tag{2.34}$$

とすることで，マイクロフォン位置での分離信号を得ることができる．このように，分離した信号 s_{ilk} に分離フィルタの逆行列の列ベクトルをかけることでマイクロフォン位置での分離信号を得ることができる．物理的には，分離フィルタの逆行列が仮想的な音源位置からマイクロフォン位置までの伝達関数に相当しているということを利用して，場所が不定な仮想的な音源位置の信号をマイクロフォン位置まで引き戻していると捉えることができる．

2.2.4 IVA

ICA の拡張手法として，IVA(Independent Vector Analysis) [17] がある．IVA は，全周波数の音源分離を同時に行うことで周波数間パーミュテーション問題を解く必要がないという特徴がある．IVA では，コントラスト関数を以下のように定義する．

$$G(\boldsymbol{s}_{il}) = 2\|\boldsymbol{s}_{il}\| = 2\sqrt{\sum_{k=1}^{K} |s_{ilk}^2|} \tag{2.35}$$

ここで $\boldsymbol{s}_{il}$ は

$$\boldsymbol{s}_{il} = [s_{il1} \cdots s_{ilk} \cdots s_{ilK}] \tag{2.36}$$

となるベクトルである．このようなコントラスト関数を用いることは，全ての周波数成分が同じ分散を有していると仮定することに相当し，複数の周波数成分が同時に変化する時にコストが小さくなる．つまり，異なる周波数で振幅スペクトルが高い相関を持つように分離フィルタを学習していくため，パーミュテーション問題を追加で解く必要がない．

$$\begin{aligned}\phi(s_{ilk}) &= \frac{\partial \boldsymbol{G}(s_{ilk})}{\partial s_{ilk}^*} \\ &= \frac{s_{ilk}}{\sqrt{\sum_{k=1}^{K}|s_{ilk}|^2}}\end{aligned} \tag{2.37}$$

次に，IP 法 [18] に基づくパラメータ更新法を紹介する．独立成分分析の場合はコスト関数を周波数ごとに求めたが，独立ベクトル分析の場合はコスト関数を周波数ごとに分けられないため，全周波数のコスト関数をまとめたコスト関数を最小化するように，次のような最小化問題を解いてパラメータを求める．

$$\text{minimize} \qquad \frac{1}{L}\sum_{l=1}^{L}\sum_{i=1}^{N} 2\sqrt{\sum_{k=1}^{K}|\boldsymbol{w}_{ik}^H \boldsymbol{x}_{lk}|^2} - \sum_{k=1}^{K} 2\log|\det \boldsymbol{W}_k| \tag{2.38}$$

実際に解く際には，第一項が最適化が難しい形をしているため，以下のような補助関数を用意する．

$$2\sqrt{\sum_{k=1}^{K}|\boldsymbol{w}_{ik}^H \boldsymbol{x}_{lk}|^2} \leq \frac{1}{v_{il}}\sum_{k=1}^{K}|\boldsymbol{w}_{ik}^H \boldsymbol{x}_{lk}|^2 + v_{il} \tag{2.39}$$

なお，v_{il} は補助変数ということになり，等号が成立するのは，

$$v_{il} = \sqrt{\sum_{k=1}^{K}|\boldsymbol{w}_{ik}^H \boldsymbol{x}_{lk}|^2} \tag{2.40}$$

のときである．

2.2.5 ILRMA

IVA では，全周波数成分が時間ごとに同じ分散を持ち，全ての周波数が同時に変化すると仮定した．しかし，この過程はとても大雑把であると言える．よってこの仮定を変更し，より音声の周波数構造のモデルに近くなるような音源分離法が提案されている．それが ILRMA [19] と呼ばれる手法である．ILRMA では，各音源の時間周波数ごとの分散 v_{ilk} をモデリングする．分散のモデルとして，非負行列分解 (NMF：Nonnegatie Matrix Factorization) [20] に基づくモデルを用いる．

$$v_{ilk} = \sum_{c=1}^{B} a_{icl} b_{ick} \tag{2.41}$$

ここで，a，b もそれぞれ非負の変数となる．B は同時に変化する周波数パターンの数を表すパラメータとする．b_{ick} を基底と呼び i 番目の音源の c 番目のパターンにおける k 番目の周波数の強さを表す．a_{icl} をアクティビティと呼び，パターン c がフレーム l で支配的な場合に大きい値を取る．a_{icl} の値が大きいほど，i 番目の音源についてその時間にパターン c を持つ割合が大きいということを意味している．このように周波数構

造のパターンを複数持つことにより，時間帯ごとに周波数構造のパターンが変化する音声のスペクトログラムに推定した時間周波数ごとの分散 v_{ilk} を，よりよく適合することが可能となる．

ここで，音源の分布を次式で示す時間変化する分散 v_{ilk} を持つ平均 0 のガウス分布 (時変ガウスモデル) に設定する．

$$\log p(s_{ilk}|\theta_k) = -\frac{|s_{ilk}|^2}{v_{ilk}} - \log|v_{ilk}| + \text{const.} \tag{2.42}$$

このモデルは時間ごとに音量が変化する音声信号を適切に表現しているモデルと言える．コントラスト関数は,

$$\boldsymbol{G}(s_{ilk}) = -\log p(s_{ilk}|\theta_k) = \frac{|s_{ilk}|^2}{v_{ilk}} + \log|v_{ilk}| + \text{const.} \tag{2.43}$$

となる．よって，ILRMA におけるパラメータ推定問題は，以下のような最小化問題に帰着することができる．

$$\text{minimize} \qquad \frac{1}{L}\sum_{l=1}^{L}\sum_{i=1}^{N}\sum_{k=1}^{K}\frac{|\boldsymbol{w}_{ik}^H\boldsymbol{x}_{lk}|^2}{\sum_c a_{icl}b_{ick}} + \log\sum_c a_{icl}b_{ick} - \sum_{k=1}^{K}2\log|\det\boldsymbol{W}_k| \tag{2.44}$$

$v_{ilk} = \sum_{c=1}^{B} a_{icl}b_{ick}$ としている．次にコスト関数を最小化するための補助関数を考える．逆関数が凸関数であることを利用した次式を用いて補助関数を導出する．

$$\frac{1}{\sum_i \lambda_i x_i} \le \sum_i \lambda_i \frac{1}{x_i} \tag{2.45}$$

$\sum_i \lambda_i x_i = 1$ とする．これにより，第一項の補助関数は,

$$\frac{|\boldsymbol{w}_{ik}^H\boldsymbol{x}_{lk}|^2}{\sum_c a_{icl}b_{ick}} = \frac{|\boldsymbol{w}_{ik}^H\boldsymbol{x}_{lk}|^2}{\sum_c \lambda_{iclk}\frac{a_{icl}b_{ick}}{\lambda_{iclk}}} \le \sum_c \lambda_{iclk}\frac{|\boldsymbol{w}_{ik}^H\boldsymbol{x}_{lk}|^2}{\frac{a_{icl}b_{ick}}{\lambda_{iclk}}} \tag{2.46}$$

となる．$\sum_c \lambda_{iclk} = 1$ である．等号が成立するのは,

$$\lambda_{iclk} = \frac{a_{icl}b_{ick}}{\sum_{c'} a_{ic'l}b_{ic'k}} \tag{2.47}$$

のときである．また、コスト関数の第二項 $\log\sum_c a_{icl}b_{ick}$ については，log 関数が凹関数であり，接線で上限を抑えられることを利用すると,

$$\log x \le \frac{1}{x_0}(x - x_0) + \log x_0 \tag{2.48}$$

が成立することがわかる．したがって，第二項の補助関数は,

$$\log \sum_c a_{icl} b_{ick} \leq \frac{1}{\beta_{ilk}} (\sum_c a_{icl} b_{ick} - \beta_{ilk}) + \log \beta_{ilk} \tag{2.49}$$

となる．等号が成立するのは $\beta_{ilk} = \sum_c a_{icl} b_{ick}$ のときである．したがって，最終的にコスト関数全体での補助関数は，

$$\frac{1}{L} \sum_{l=1}^{L} \sum_{i=1}^{N} \sum_{k=1}^{K} \sum_c \lambda_{iclk} \frac{|\boldsymbol{w}_{ik}^H \boldsymbol{x}_{lk}|^2}{\frac{a_{icl} b_{ick}}{\lambda_{iclk}}} + \frac{1}{\beta_{ilk}} (\sum_c a_{icl} b_{ick} - \beta_{ilk}) + \log \beta_{ilk} - \sum_{k=1}^{K} 2 \log |\det \boldsymbol{W}_k| \tag{2.50}$$

となる．補助関数を a_{icl} について，最小化すると

$$\sum_{k=1}^{K} -\lambda_{iclk}^2 \frac{|\boldsymbol{w}_{ik}^H \boldsymbol{x}_{lk}|^2}{a_{icl}^2 b_{ick}} + \sum_{k=1}^{K} \frac{b_{ick}}{\beta_{ilk}} = 0 \tag{2.51}$$

となる．したがって，

$$a_{icl}^2 = \frac{\sum_{k=1}^{K} -\lambda_{iclk}^2 \frac{|\boldsymbol{w}_{ik}^H \boldsymbol{x}_{lk}|^2}{b_{ick}}}{\sum_{k=1}^{K} \frac{b_{ick}}{\beta_{ilk}}} \tag{2.52}$$

となる．さらに，式 (2.47) を用いて λ_{iclk} を置き換えると，

$$a_{icl} \leftarrow a_{icl} \sqrt{\frac{\sum_{k=1}^{K} \frac{b_{ick}}{(\sum_{c'} a_{ic'l} b_{ic'k})^2} |\boldsymbol{w}_{ik}^H \boldsymbol{x}_{lk}|^2}{\sum_{k=1}^{K} \frac{b_{ick}}{\sum_{c'} a_{ic'l} b_{ic'k}}}} \tag{2.53}$$

となる．ここで右辺の a_{icl} は更新前の値とする．また補助関数を b_{ick} について最小化すると，

$$b_{ick} \leftarrow b_{ick} \sqrt{\frac{\sum_{l=1}^{L} \frac{a_{icl}}{(\sum_{c'} a_{ic'l} b_{ic'k})^2} |\boldsymbol{w}_{ik}^H \boldsymbol{x}_{lk}|^2}{\sum_{l=1}^{L} \frac{a_{icl}}{\sum_{c'} a_{ic'l} b_{ic'k}}}} \tag{2.54}$$

となる．フィルタの更新は IVA と同様に次式にように更新する．

$$\boldsymbol{w}_{ik} \leftarrow (\overline{\boldsymbol{W}}_k \boldsymbol{Q}_{ik})^{-1} \boldsymbol{e}_i \tag{2.55}$$

$$\boldsymbol{w}_{ik} \leftarrow \frac{\boldsymbol{w}_{ik}}{\sqrt{\boldsymbol{w}_{ik}^H \boldsymbol{Q}_{ik} \boldsymbol{w}_{ik}}} \tag{2.56}$$

ここで，

$$\boldsymbol{Q}_{ik} = \frac{1}{L} \sum_{l=1}^{L} \frac{\boldsymbol{x}_{lk} \boldsymbol{x}_{lk}^H}{\sum_{c=1}^{B} a_{icl} b_{ick}} \tag{2.57}$$

とする．

このように周波数の変化パターンを複数設定することで，より音声の構造に近いモデルとなり，高精度に分離を行うことが可能になる．さらに，近年では深層学習と組み合わせることで，さまざまな利点を持った音源分離法が提案されている [5,13,21,22]．

2.3 雑音除去

2.3.1 定義

前節では，音源からマイクロフォンまでに伝わる音として直接音だけを前提としていた．しかし，実際には直接音だけでなく壁や床，天井などで反射して生じる残響成分が存在する．この残響の影響により，CSP 法の精度が大きく下がるという問題がある．そこで，不要な残響音を除去する，残響除去技術について述べる．なお，本節も記述に当たって [13] を参考にした．残響が存在する環境での音声の入力信号は以下の (2.58) 式のようにモデリングできる．

$$\boldsymbol{x}_{lk} = \sum_{i=1}^{N} \sum_{\tau=0}^{L_\tau - 1} \boldsymbol{g}_{i,\,\tau,\,k} s_{i,l-\tau,k} \tag{2.58}$$

ここで，$\boldsymbol{g}_{i,\tau,k}$ は時間周波数領域でのインパルス応答とし，インパルス応答長を L_τ とする．ここまでの議論から，残響除去は式 (2.58) で定義される入力信号 $\boldsymbol{x}_{lk}$ から残響のない信号 s_{ilk} を推定する問題と位置付けられる．

2.3.2 空間的な逆フィルタ

残響除去の基礎理論として，multiple-input/output inverse theorem(MINT) [23] がある．これにより，複数マイクロフォンを用いた有限長のフィルタで残響を完全に除去できることが示される．まず，式 (2.58) を行列形式で簡潔に描き直すと以下の (2.59) のようになる．

$$\boldsymbol{x}_{lk} = \mathbf{G}_k \mathbf{s}_{lk} \tag{2.59}$$

ここで，$\mathbf{G}_k$ は以下の (2.60) 式のようにインパルス応答をまとめた形である．

$$\mathbf{G}_k = (\boldsymbol{G}_{\tau=0,k} \cdots \boldsymbol{G}_{\tau=L_\tau-1,k}) \tag{2.60}$$

とする．$\boldsymbol{G}_{\tau,k}$ は τ タップ目のインパルス応答であり，

$$\boldsymbol{G}_{\tau,k} = (\boldsymbol{g}_{i=1,\tau,k} \cdots \boldsymbol{g}_{i=N,\tau,k}) \tag{2.61}$$

とする．このようにインパルス応答を一つの行列 $\mathbf{G}_k$ にまとめることができる．これに対応して $\mathbf{s}_{lk}$ は残響のない音源信号をまとめたベクトルとなり，次のように定義される．

$$\mathbf{s}_{lk} = \begin{pmatrix} \boldsymbol{s}_{lk} \\ \vdots \\ \boldsymbol{s}_{l-\tau,k} \\ \vdots \\ \boldsymbol{s}_{l-L_\tau+1,k} \end{pmatrix} \tag{2.62}$$

さらに，

$$\boldsymbol{s}_{lk} = [s_{1lk} \cdots s_{Nlk}]^T \tag{2.63}$$

とする．このとき，もし $\mathbf{G}_k$ がわかっていて，かつ逆行列が存在するとすると，

$$\mathbf{s}_{lk} = \mathbf{G}_k^{-1} \boldsymbol{x}_{lk} \tag{2.64}$$

で $\mathbf{s}_{lk}$ を推定することができる．$\mathbf{G}_k$ に逆行列が存在するには，行数 M と列数 $L_\tau N$ が等しい必要があるが，インパルス応答長 L_τ が長くなると等号が成立しないおそれがある．そのとき，未知数が観測信号より多い状態となり，唯一解を求めることができない．そこで，マイクロフォン信号を以下の (2.65) 式のように書き直し，過去のマイクロフォン信号を加えることにより観測信号を増やす．

$$\overline{\boldsymbol{x}}_{lk} = \overline{\mathbf{G}}_k \overline{\mathbf{s}}_{lk} \tag{2.65}$$

ここで，

$$\overline{\boldsymbol{x}}_{lk} = \begin{pmatrix} \boldsymbol{x}_{lk} \\ \vdots \\ \boldsymbol{x}_{l-\tau,k} \\ \vdots \\ \boldsymbol{x}_{l-L_f+1,k} \end{pmatrix} \tag{2.66}$$

$$\overline{\mathbf{G}}_k = \begin{pmatrix} \mathbf{G}_k & 0 & \cdots & & \\ 0 & \mathbf{G}_k & 0 & \cdots & \\ \vdots & \ddots & & & \vdots \\ 0 & \cdots & & \mathbf{G}_k & 0 \\ 0 & \cdots & & & \mathbf{G}_k \end{pmatrix} \tag{2.67}$$

$$\overline{\mathbf{s}}_{lk} = \begin{pmatrix} \boldsymbol{s}_{lk} \\ \vdots \\ \boldsymbol{s}_{l-\tau,k} \\ \vdots \\ \boldsymbol{s}_{l-L\tau-L_f+1,k} \end{pmatrix} \tag{2.68}$$

とする．このような変形を施すことによって，$\overline{\mathbf{G}}_k$ の逆行列が計算できれば，

$$\overline{\boldsymbol{s}}_{lk} = \overline{\mathbf{G}}_k^{-1}\overline{\mathbf{x}}_{lk} \tag{2.69}$$

で音源信号を推定できる．$\overline{\mathbf{G}}_k$ は行数が ML_f，列数が $(L_\tau + L_f - 1)N$ の行列なので，

$$ML_f = (L_\tau + L_f - 1)N \tag{2.70}$$

が成立すれば $\overline{\mathbf{G}}_k$ が正方行列となり，逆行列が計算できる．このような逆行列 $\overline{\mathbf{G}}_k^{-1}$ を空間的な逆フィルタは，L_f を次のように調整することで L_τ によらず算出可能である．

$$L_f = \frac{(L_\tau - 1)N}{M - N} \tag{2.71}$$

よって，マイクロフォン数 M が音源数 N よりも大きければ，(2.71) 式を満たす L_f を求めることができ，$\overline{\mathbf{G}}_k$ が正方行列となるため逆行列が存在することとなる．これは L_τ の値によらず成立する．このとき，残響成分を完全に除去することが可能になる．

2.3.3 線形予測に基づく残響除去

MINT の欠点として，前項では音源からマイクロフォンまでのインパルス応答が既知である事を前提にしていたが，実際にインパルス応答を事前に知ることは困難であるという点がある．よってインパルス応答を推定することなく残響を除去できる方法として，WPE(Weighted Prediction Error) [24] が考案された．WPE では，(2.72) 式の $\boldsymbol{W}_k$ を，インパルス応答の逆行列を求めるというアプローチ以外の方法で求める必要がある．

$$\boldsymbol{s}_{lk} = \boldsymbol{W}_k\overline{x}_{lk} \tag{2.72}$$

概要

まず，線形予測に基づく残響除去法の概要を述べる．音源数 N を 1 と仮定し，$mathbfs_{lk}$ の一つ目の成分だけを推定する事を考えてみます．この時，$boldsymbols_{i=1,l,k}$ を推定すれば良いため，$\boldsymbol{W}_k$ の 1 行目である w_{1k} を推定できれば良いということになる．出力信号の大きさの不定性を考え，最初の要素を 1 に固定すると，w_{1k} は以下のようにかける．

$$w_{1k} = (1 \quad \boldsymbol{h}_{k,1}^H \cdots \boldsymbol{h}_{k,L_h}^H) \tag{2.73}$$

これは次のような指揮に基づき残響除去後の信号 $\widetilde{d}_{lk}$ を推定することに相当する．

$$\widetilde{d}_{lk} = x_{1lk} - \sum_{\tau=1^{L_h}} \boldsymbol{h}_{k,\tau}^H x_{l-\tau,k} \tag{2.74}$$

ここで，$\boldsymbol{h}_k, \tau$ が残響除去フィルタということになる．これを少し簡略化して

$$\widetilde{d}_{lk} = x_{1lk} - \boldsymbol{h}_k^H \overline{x}_{l,k} \tag{2.75}$$

と表す．ここで，

$$\boldsymbol{h}_k = [\boldsymbol{h}_{k,1}^T \cdots \boldsymbol{h}_{k,\tau}^T \cdots \boldsymbol{h}_{k,L_h}^T]^T \tag{2.76}$$

$$\overline{\boldsymbol{x}}_{lk} = [\boldsymbol{x}_{l-1,k}^T \cdots \boldsymbol{x}_{l-L_h,k}^T]^T \tag{2.77}$$

とする．ここで (2.75) 式の右辺の第一項が現在のマイクロフォン入力であり，これから第二項で推定した残響成分を除去するような式となっている．x_{1lk} の中には，次の (2.78) 式のように直接音成分 d_{lk} と残響成分 r_{lk} の二つの成分が混ざっていると考えられる．

$$x_{1lk} = d_{lk} + r_{lk} \tag{2.78}$$

(2.75) 式は残響成分を第二項で $\widetilde{r}_{lk} = \boldsymbol{h}_k^H \overline{\boldsymbol{x}}_{lk}$ という形で推定し，x_{1lk} から日ことで直接音成分を求める．つまり，推定した直接音成分は，

$$\widetilde{d}_{lk} = d_{lk} + (r_{lk} - \widetilde{r}_{lk}) \tag{2.79}$$

となる．過去の入力信号から線形フィルタ $\boldsymbol{h}$ を用いて残響成分を推定していることから，線形予測に基づく残響除去法と呼ばれている．しかし，この手法では，二つの問題がある．一つ目は，直接音と初期反射音が高い相関を持つことから，分離精度があまり良くないという点である．二つ目は，コスト関数に直接音が入っているため，直接音が大きい時間帯ほど直接音が除去されてしまうという点である．これらを解決したのが WPE である．

WPE

WPE では，上記の問題を解決し，高精度に線形予測に基づく残響除去を可能にしている．まず直接音と初期反射音が相関を持つことの解決策として，マルチステップ線形予測法 [25] を用いている．この方法では，$\boldsymbol{x}_{lk}$ 中の直接音声文と相関が高い初期反射音が含まれる過去のマイクロフォン入力信号 $\boldsymbol{x}_{l-1,k} \cdots \boldsymbol{x}_{l-D,k}$ を $\overline{\boldsymbol{x}}_{lk}$ から取り除き，残響成分を次のように推測する．

$$\widetilde{r}_{lk} = \boldsymbol{h}_k^H \overline{\boldsymbol{x}}_{l,k} = \sum_{\tau=D+1}^{D+L_h} \boldsymbol{h}_{k,\tau}^H \boldsymbol{x}_{l-\tau,k} \tag{2.80}$$

また，コスト関数として，残響除去後の信号は音声信号であるため，時変ガウスモデル分布で表すのが適当である．これにより，二つ目の問題も解決することができる．

$$p(d_{lk}|\theta) = \mathcal{N}(0, v_{lk}) \tag{2.81}$$

ここで v_{lk} は時間周波数ごとに変化する音声の分散である．この分布を用いた最尤推定は，以下の最小化問題を解くことに相当する．

$$\text{minimize} \qquad \sum_{l=1}^{L}\sum_{k=1}^{K} \frac{|x_{1lk} - \boldsymbol{h}_k^H \overline{\boldsymbol{x}}_{lk}|^2}{v_{lk}} + \log v_{lk} \tag{2.82}$$

このコスト関数が単調減少するように最適化していく．WPE では，このようにして残響除去を行なっている．また，音源分離である ILRMA と WPE を組み合わせた手法も考案されている [26, 27]．さらに, 分離行列の推定に，IP 法よりも計算量が少ない ISS [28] を使用した ILRMA-ISS なども提案されている．

第 3 章

複数音源の場合の音源分離による CSP 法の精度の向上

3.1 背景

第 2 章で述べたように，CSP 法による TDoA の検出において音源が複数個存在する場合では，検出精度が低下するという問題がある [8]．そこで，まずはシミュレーションを用いて音源が複数の場合のマイクロフォンへの入力信号を収録し，その入力信号を用いて ILRMA による音源分離を施した場合と音源分離をしない場合の二種類の CSP 法による到達時間差の検出を行った．

3.2 実験環境

本論文での実験は全て，pyroomacoustics [29] という，室内音響を再現する python で記述されたライブラリを用いておこなった．pyroomacoustics では，部屋の音響を再現した上で音声の収録シミュレーションをすることができ，部屋の大きさやマイクや音源の座標，壁や天井での音の反射率などを非常に細かく設定することができる．音源を設定し，諸条件を決定した上でシミュレーションを行うと，鏡像法による壁や天井，床への反射を考慮した各音源から各マイクロフォンへの室内伝達関数を計算し，音源との畳み込みによりマイクへの収録音声を生成することができる．

3.3 実験内容

本実験での，シミュレーションのセットアップの詳細を示す．5m 四方で高さが 2.25m の部屋を作成し，マイクと音源の高さは 1.5m で統一した．よって，位置関係は二次元平面上で考える．本シミュレーションではマイクと音源を 3 つずつ用意し，図 3.1 に示すように設置した．マイクの座標は m1，m2，m3 それぞれ [1.0, 2.0, 1.5]，[4.0, 2.0, 1.5]，[2.5, 3.5, 1.5] となっており，音源の座標は s1，s2，s3 それぞれ [4.5, 4.5, 1.5]，[2.0, 4.0, 1.5]、[1.0, 3.0, 1.5] とした．また，m1 と m2 のペアを paer1 とし，m1 と m3 のペアを pair2 とし，m2 と m3 のペアを pair3 とした．なお，サンプリング周波数を 16kHz，音声と雑音の比率である SNR を 90dB に設定している．また考慮する壁や天井，床による反射回数 0 としており，無残響を想定したシミュレーションを行っている．音源には 1，s2，s3 それぞれ音声コーパスである CMU ARCTIC Corpus [30] を用いている．なお，サンプリング周波数を 16kHz，音声と雑音の比率である SNR を 90dB に設定している．このよう

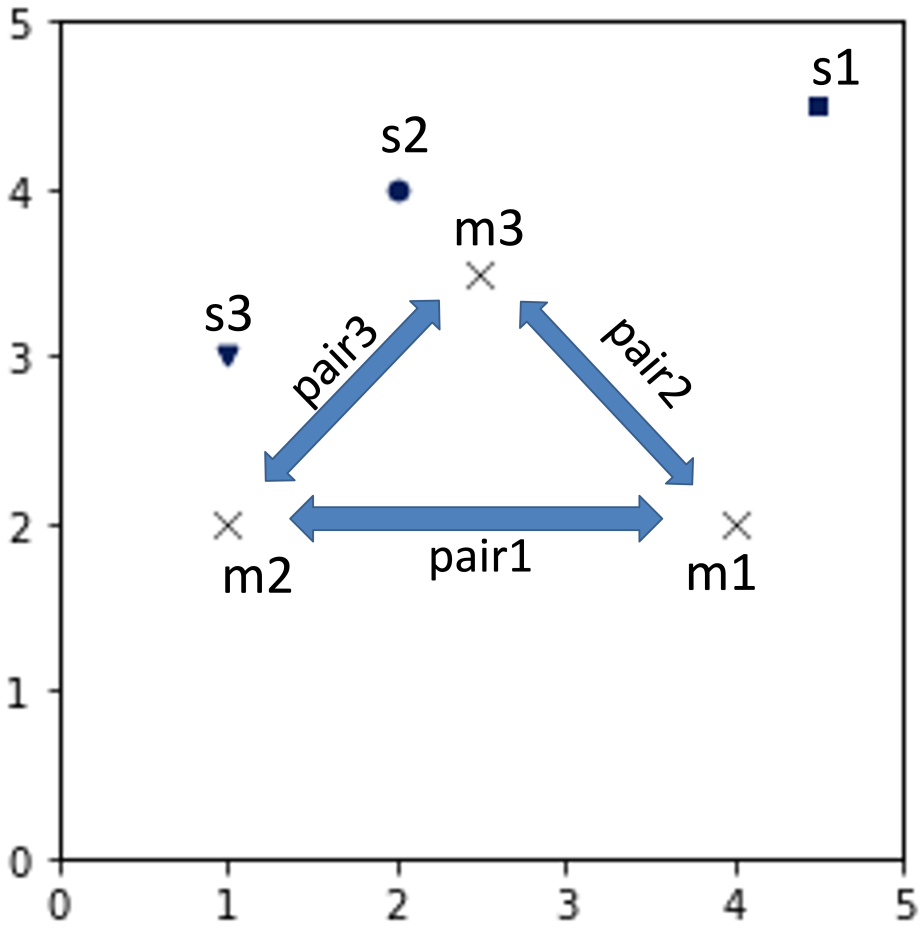

図 3.1: シミュレーションのセットアップ

な設定をした上で収録シミュレーションを行い，得られた音声をもとに検討を行なった．

3.4 結果と考察

音源分離を用いない場合

まず，音源分離をせずにそれぞれのマイクペアへの入力信号に対して，CSP 法を適用した．その結果を図 3.2 に示す．横軸は時間で縦軸は式によって算出された CSP の値である．ピークが生じている場所が，ある音源の TDoA を表しており，今回は音源を三つ用いているため，理論的にはピークが三つ生じるはずである．どのマイクペアでも音源によってピークの立ち上がりに差があり，相関の値も非常に小さく，ノイズと区別が難しい．実際にマイクペアによっては一つの音源の立ち上がりが消失してしまっているように見える．これらのように，検出精度は良いとは言えない．

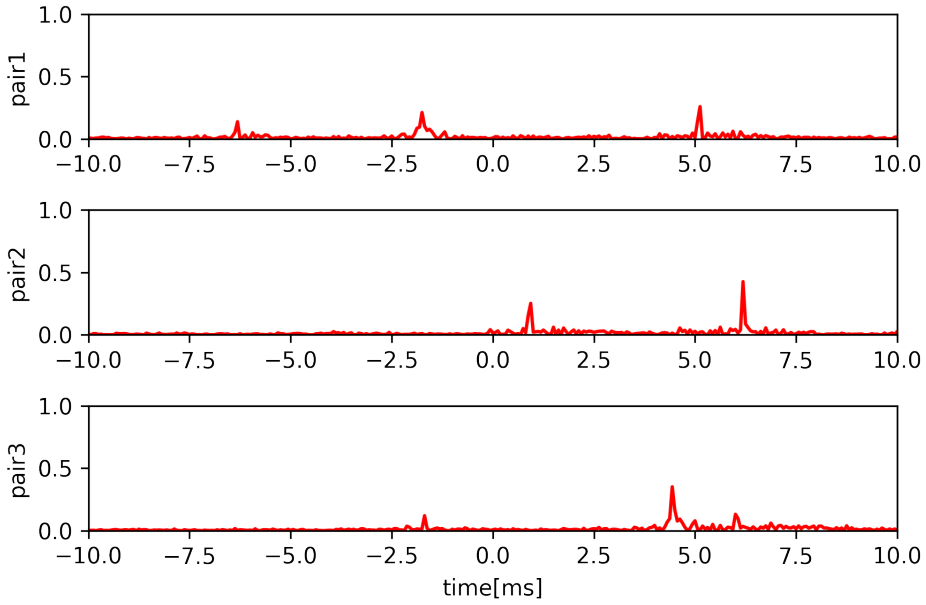

図 3.2: 音源分離無しでの CSP 法

ILRMA による音源分離を行なった場合

次に，それぞれのマイクペアへの入力音源に対して ILRMA による音源分離を施した上で，CSP 法を適用した．その結果を図 3.3 に示す．横軸は時間で縦軸は式によって算出された CSP の値である．それぞれ青色が s1，緑色が s2，黄色が s3 を表している．図 3.2 と比較して，消えていたピークも検出できており，かつすべてのマイクペアにおいて全ての音源のピークの立ち上がりが高い値となっており，検出精度が向上していることがわかる．

距離差の検出

また，検出した到達時間差とその誤差として適当を考えられる半値幅を用いて，音源からのマイクペアまでの距離差を推定し，幾何学的に計算することのできる実際の距離差との比較を行ったものが以下の表である．それぞれのセルの上の数字が推定された距離差の値と半値幅による許容誤差であり，下の数字が幾何学的に算出される距離差である．どの場合においても，半値幅によって考えられる推定誤差の範囲内に実際の値が収まっており，非常に精度よく距離差の推定ができていると言える．

3.5 考察

全てのマイクロフォンペアと全ての音源の組み合わせに対して，精度の大きな向上が確認された．これは音源分離によってある音源に対するマイクロフォンペアへの入力の相関が高まったためと考えられる．また，図 3.2 と図 3.3 を比較すると，音源同士の TDoA が近い場合では，特に CSP 法単体でのピーク検出精度が小さくなることがわかる．このことから，近距離にある二音源に対して，本手法が有用であることがわかる．また，今回の ILRMA による音源分離では遅延成分を残したままの分離が可能となっている．これはある一定時

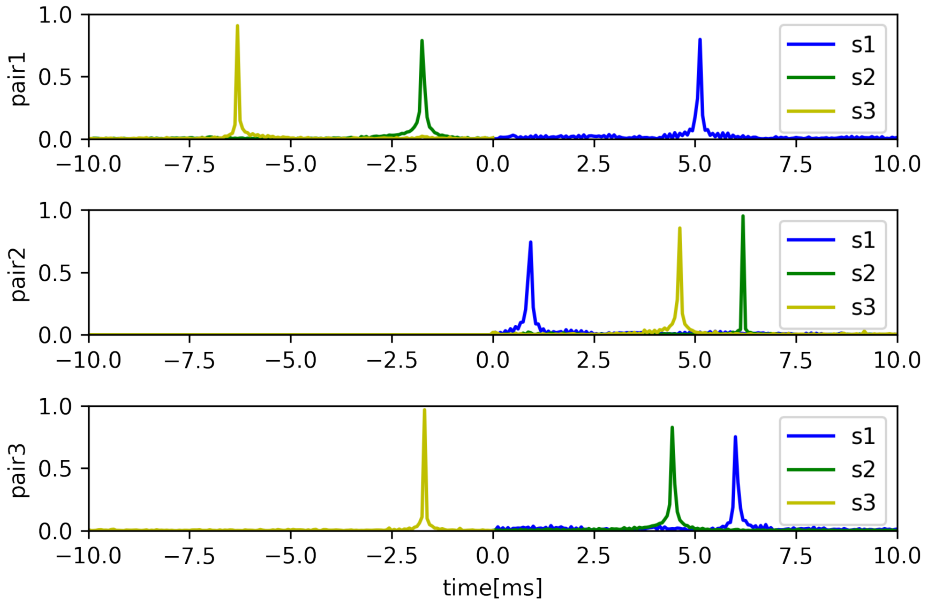

図 3.3: ILRMA による音源分離後の CSP 法

表 3.1: 推定した距離差と実際の距離差の比較

	s1	s2	s3
pair1	1.76 ± 0.02	0.60 ± 0.02	2.17 ± 0.01
	1.75	0.59	2.16
pair2	0.32 ± 0.02	2.12 ± 0.01	1.59 ± 0.02
	0.31	2.12	1.58
pair3	2.06 ± 0.02	1.52 ± 0.02	0.58 ± 0.01
	2.07	1.53	0.58

間収録した音声に対し，短時間フーリエ変換 (STFT:Short Time Fourier Transform) を用いて，ある短時間のフレームにおける信号を周波数ごとに分離しているため，時間成分に手を加えることなく分離ができるためだと考えられる．これは音源の時間ごとの周波数と信号成分の強さを表すグラフである，スペクトログラムからも推察できる．音源のスペクトログラムを図 3.4 から図 3.6 に，マイクへの入力信号のスペクトログラムを図 3.7 に，音源分離後の信号のスペクトログラムを図 3.8 から図 3.10 にそれぞれ示した．それぞれ，横軸が時間，縦軸が周波数を表しており，各点の色で強さを表している．

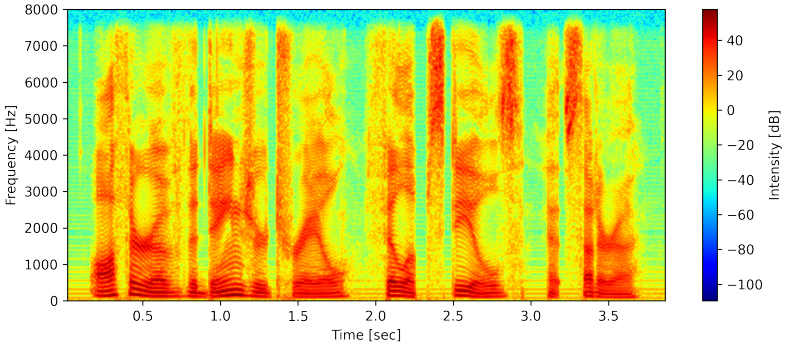

図 3.4: s1 のスペクトログラム

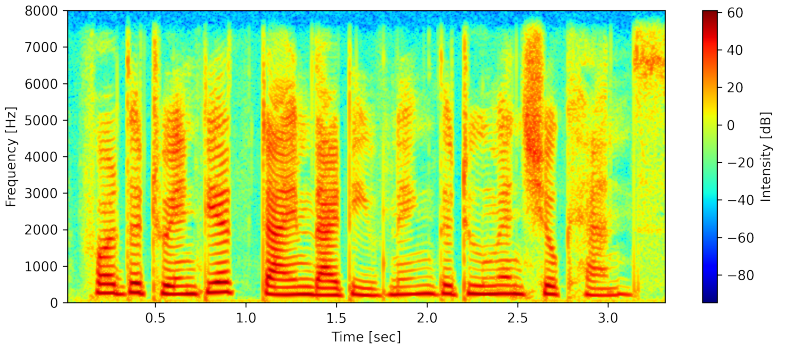

図 3.5: s2 のスペクトログラム

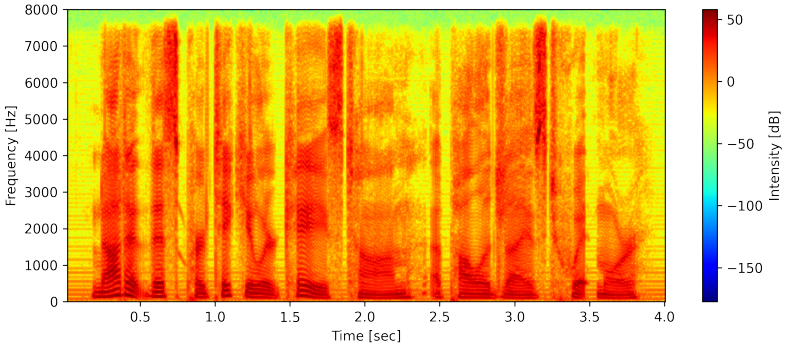

図 3.6: s3 のスペクトログラム

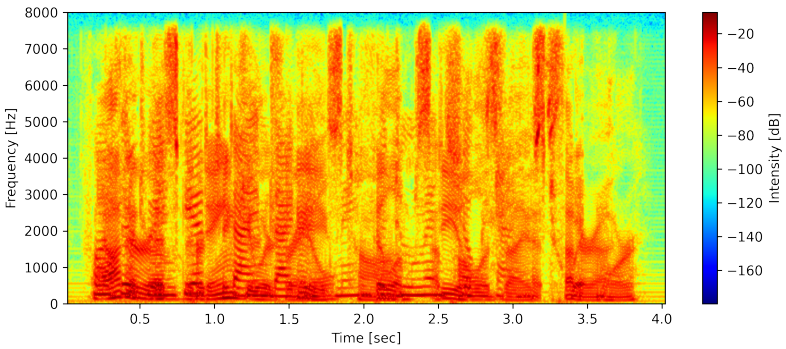

図 3.7: マイクロフォンへの入力信号のスペクトログラム

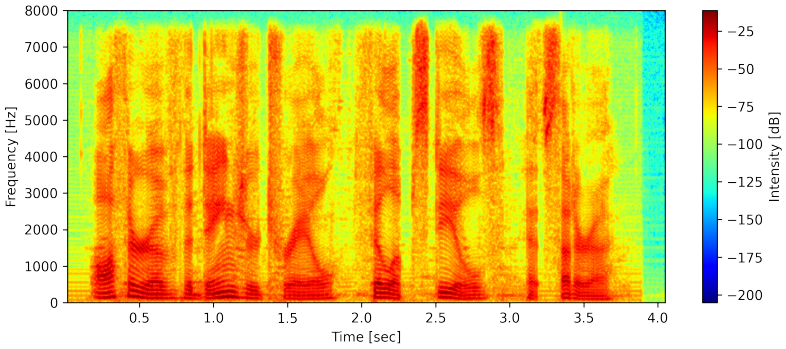

図 3.8: 音源分離後の s1 のスペクトログラム

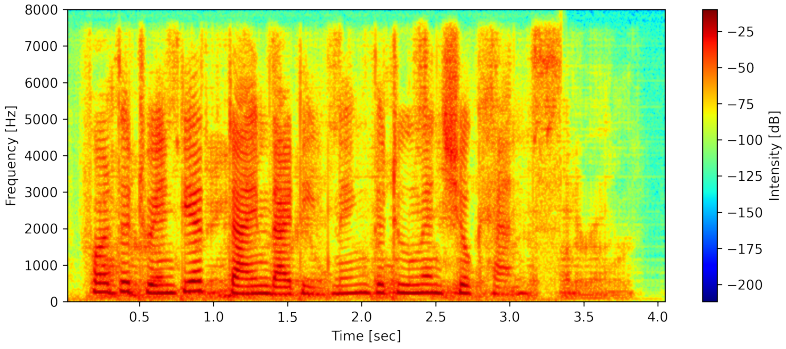

図 3.9: 音源分離後の s2 のスペクトログラム

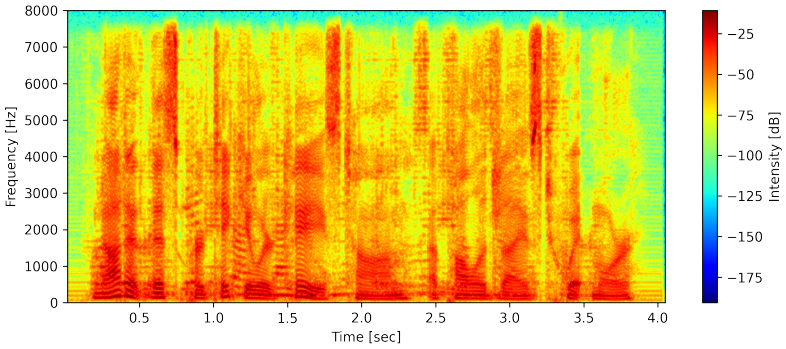

図 3.10: 音源分離後の s3 のスペクトログラム

第 4 章

残響が存在する場合の残響除去による CSP 法の精度の向上

4.1 背景

次に，残響が存在する環境下での CSP 法について検討を行う．既存研究では，残響下において CSP 法の精度が低下するという問題が指摘されている [8]．単音源に対してマイクロフォンを二つ配し，残響が存在するような環境でのシミュレーションを行い，収録された音声信号による CSP 法の結果を図 4.1 に示す．このように残響が存在する環境では CSP 法のシグナルがノイズに埋もれてしまい，正確な TDoA の検出が困難になることがわかる．これは部屋内で音源が壁や床，天井で反響し，その音声もマイクロフォンが拾うことにより，音源から直接マイクロフォンに届く音同士の TDoA 以外のピークが生じてしまうためである．よって本章では，残響を除去することで，CSP 法の精度の向上を目指す．

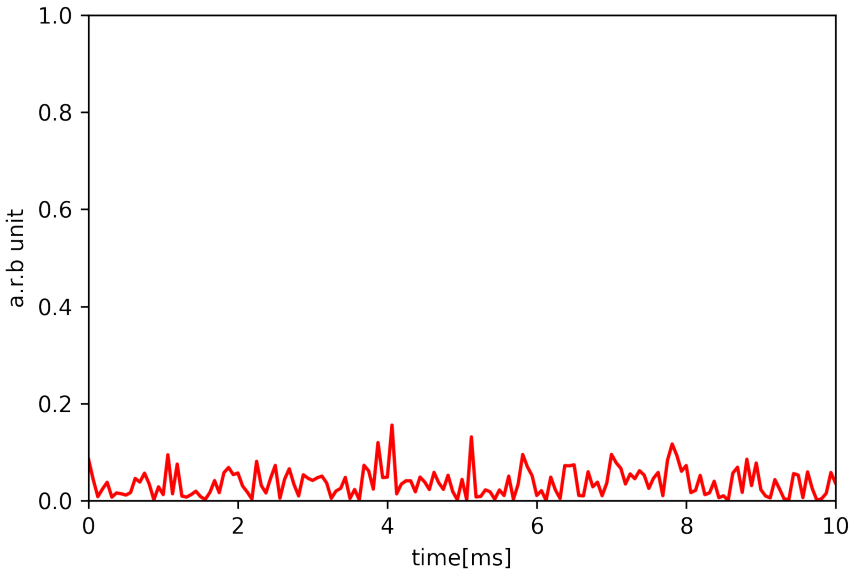

図 4.1: 残響環境下での CSP 法

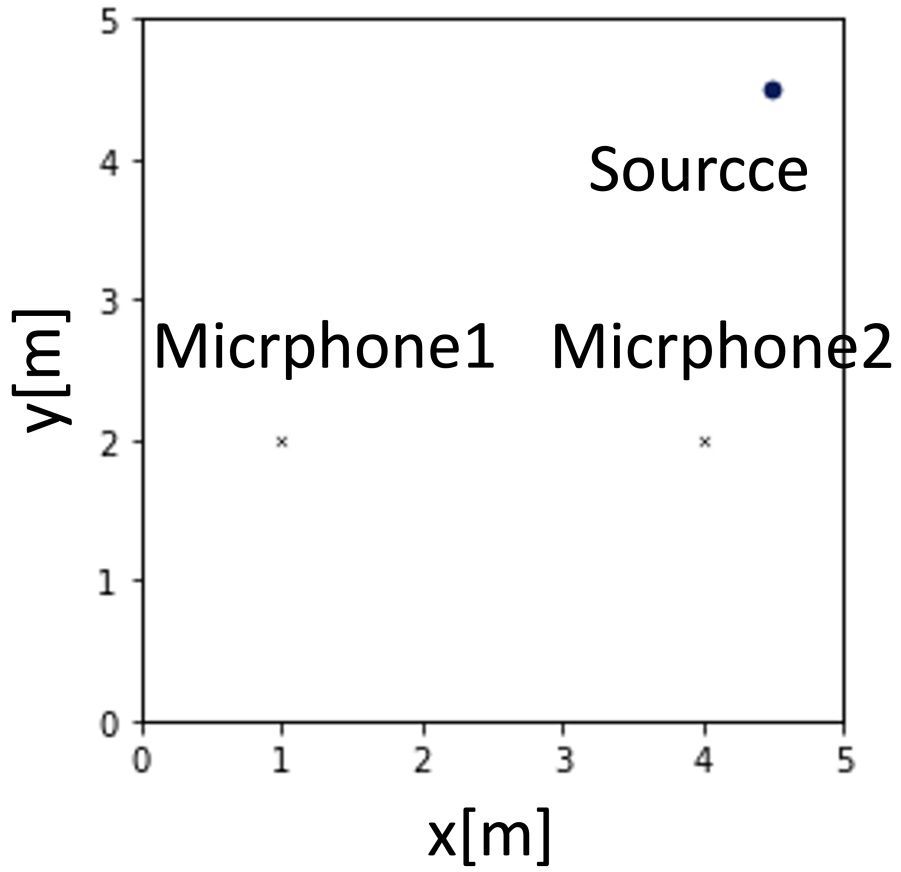

図 4.2: シミュレーションのセットアップ

4.2 実験内容

シミュレーションは pyroomacoustics [29] を用いて実施した．直方体の部屋に一つの音源と二つのマイクロフォンを設置し，壁や床・天井の材質は残響条件に応じて適宜，変更した．部屋のサイズは縦 5m 横 5m 高さ 2.25m の直方体にとった．部屋の左下手前を原点に取り，音源の座標は (4.5, 4.5, 1.5) に，マイクロフォン 1 を (1.0, 2.0, 1.5)，マイクロフォン 2 を (4.0, 2.0, 1.5) に配した．マイクロフォンと音源は全て同じ高さにあるため，マイクロフォンと音源との距離差は同一平面上で考えることができる．Z=1.5 での xy 平面を図 4.2 に示す．音源には CMU ARCTIC Corpus [30] という音声コーパスの中の男性の肉声音源を用いている．またサンプリングレートは 16kHz としている．無残響のシミュレーションでは，部屋 6 面の材質の吸音率を 100 パーセントに設定した．残響下でのシミュレーションは，部屋 6 面の材質の吸音率を 35 パーセント，考慮する壁や床，天井での音声の反射回数を 17 回として実施した．

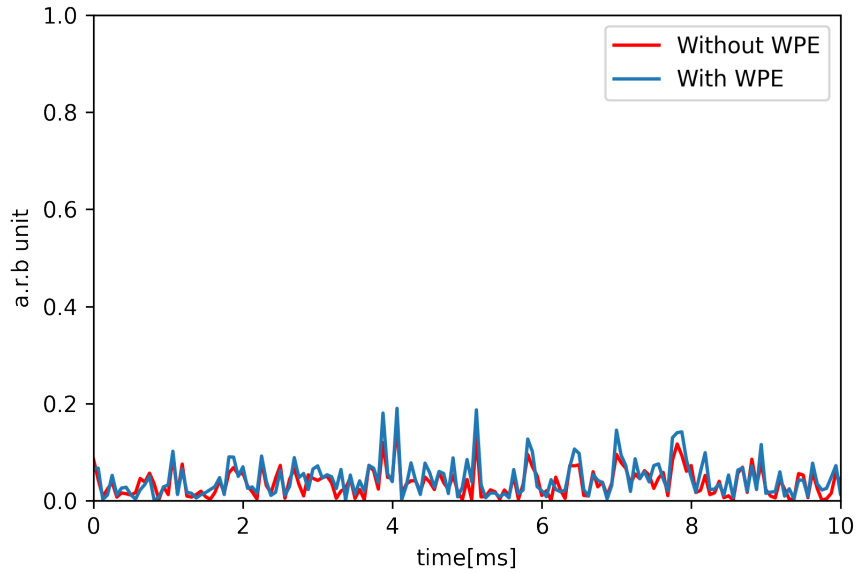

図 4.3: WPE による残響除去前後での CSP 法の変化

4.3 WPE による残響除去

二章で述べた通り，残響除去の方法として WPE [24] がある．これを用いてマイクロフォンの入力信号に対して残響除去を行った上で CSP 法を適用した．図 4.3 に WPE による残響除去前後の CSP 法の比較を示す．図からわかる通り，多少ピークの高さが改善されているものの，ピークの乱立は軽減できておらず，精度の向上は達成できていないと言える．そこで，インパルス応答を推測するアプローチで残響を低減する手法を提案する．

4.4 残響除去型 CSP 法の提案

概要

音源からマイクロフォンに到達する音声は，マイクロフォンに直接到達する成分の他に，壁や床で反射や減衰を繰り返し，様々な経路を経た上で到達する成分が重畳する．こうした残響の効果は音源とマイクロフォンごとのペアごとに Room Impulse Response (RIR) 関数として表現される．もし RIR を遅延成分と残響成分に分解し，マイクロフォンの収録音声から遅延成分だけを残して残響成分を除去することができれば，TDoA を精密に計算できると考えられる．

原理

音源からマイクロフォンまでの信号伝達遅延を Δt とすると，無残響環境下で音源とマイクロフォンの間に定義される RIR は，

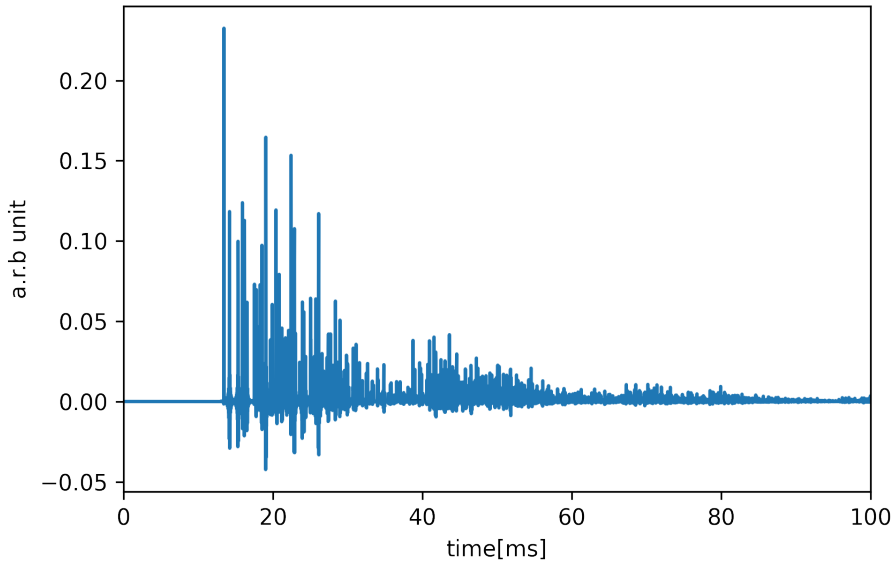

図 4.4: 通常の RIR のグラフ

$$r_0 = \delta(t - \Delta t) \tag{4.1}$$

となる．残響環境における RIR は，残響成分を $v(t)$ とすると遅延成分 $r_0(t)$ との畳み込みとして，

$$r(t) = \int r_0(s-t)v(s)ds \tag{4.2}$$

のように表せる．従って，$r(t)$ から $r_0(t)$ をデコンボリューションすることで残響成分を取り出すことができる．図 4.4 には，実際にシミュレーション上で計算された伝達関数である RIR を表している．図 4.4 に示すように，RIR にははじめにシグナルのない領域が一定時間あり，これが遅延成分 $r_0(t)$ を反映している．これを取り除いたものを $v(t)$ とし，図 4.5 に示した．

マイクロフォンの信号 $x(t)$ に対し，

$$X = \mathcal{F}[x(t)] \tag{4.3}$$
$$V = \mathcal{F}[v(t)] \tag{4.4}$$

として，

$$x_{nr}(t) = \mathcal{F}^{-1}\left[\frac{X}{V}\right] \tag{4.5}$$

と表される信号 $x_{nr}(t)$ は，マイクロフォンへの入力信号から残響のみが除去された信号となる．これを各音源-マイクロフォンペアに施し，CSP 法を用いることで TDoA の検出精度を高めることができる．

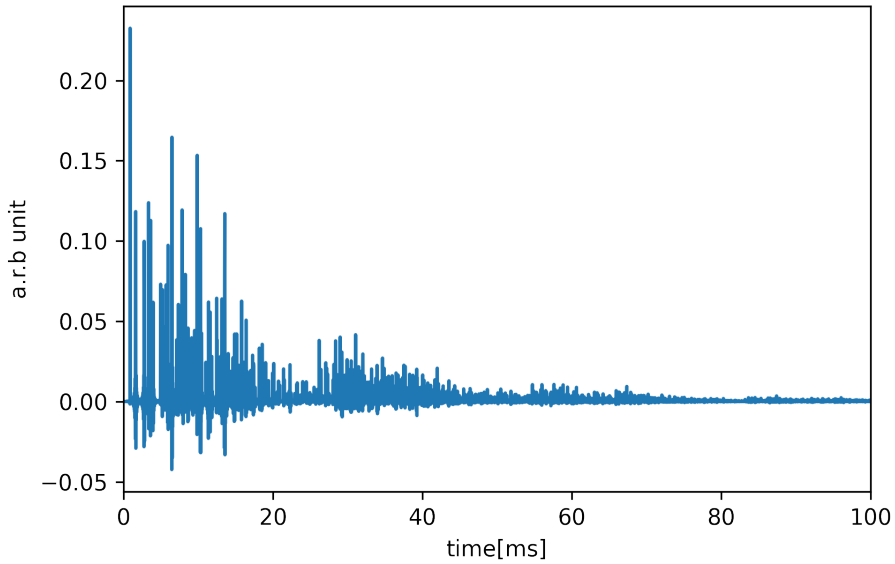

図 4.5: 遅延成分を取り除いた RIR のグラフ

手法

まずシミュレーションを用いて，本手法の妥当性を確認する．音源としてインパルス波を用い，RIR を取得する．インパルス波はデルタ関数で表現され，その波形は図 4.6 のようになる．マイクロフォンで音声を収録する際，信号の立ち上がりをトリガーとしてそれ以降の信号を抽出することで，マイクロフォンごとに実効的な $v(t)$ を得た．次いで音源に肉声を用いてシミュレーションを実行し，マイクロフォンの収録音声に対して前述の残響除去処理を施した上で CSP 法を適用した結果が図 4.7 である．この図より，TDoA は 5.13ms と検出され，無残響下での結果と一致した．従って，提案手法により，RIR から遅延成分だけを効果的に取り除き，残響除去フィルタを構築できることがわかる．

現実的な音源を用いた残響除去フィルタの構築

前項で取り扱ったインパルス音源は仮想的なものであり，現実の環境で実験を行うことは難しい．そこでより現実に即した音源を用意し，残響除去フィルタの構築が可能か調査した．残響除去フィルタの構築に用いる音源の信号を $s(t)$，マイクロフォンの収録音声を $x(t)$ とする．$x(t)$ から最初の無音部分を取り除いて $x_{\mathrm{trim}}(t)$ とし，

$$V_s = \frac{\mathcal{F}[x_{\mathrm{trim}}(t)]}{\mathcal{F}[(s(t)]} \tag{4.6}$$

を求める．この $V(s)$ を式 (4.5) の V として用いることで残響除去フィルタを構築する．これが実際に機能するか検証した．

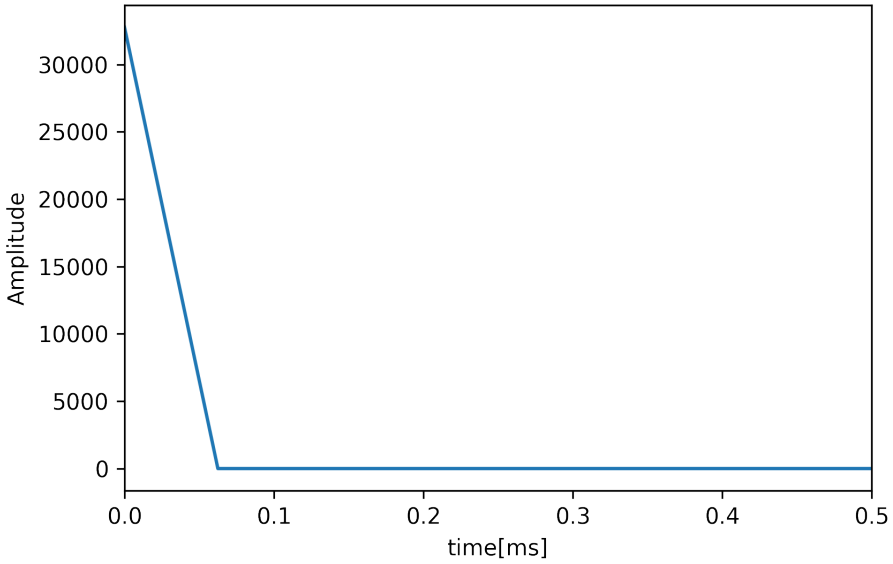

図 4.6: インパルス波の波形

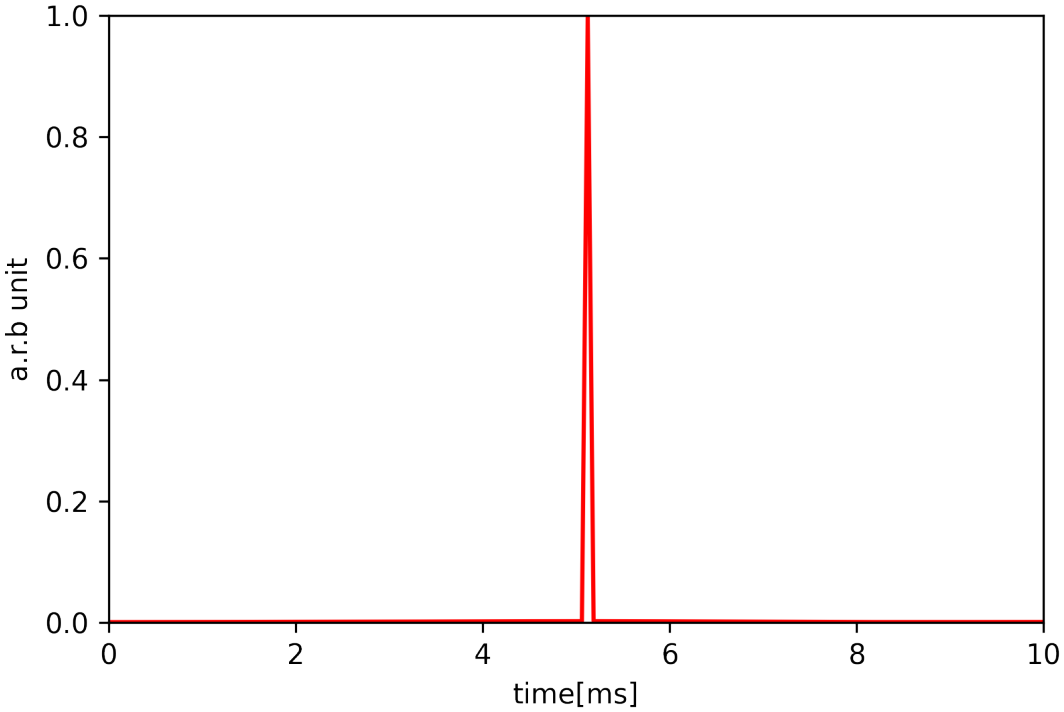

図 4.7: 残響除去後の CSP 法

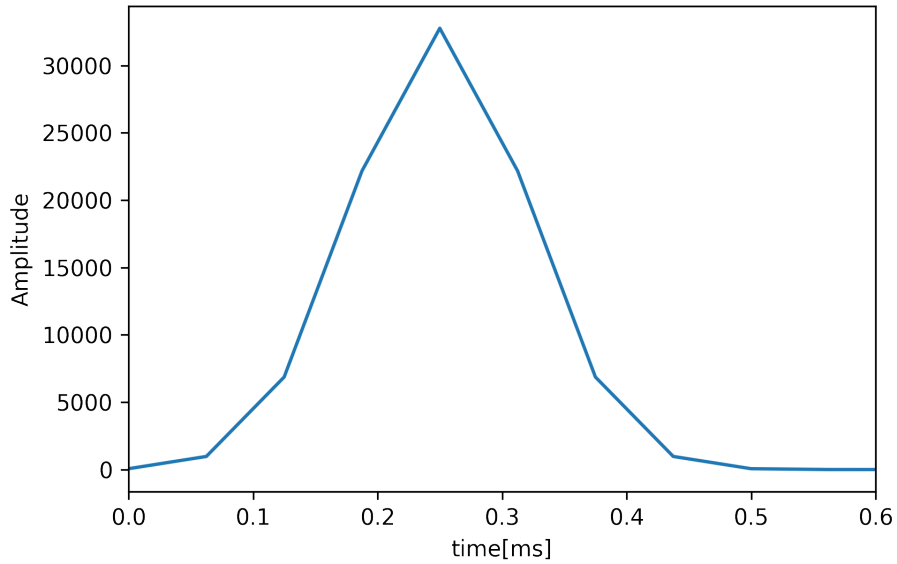

図 4.8: ガウシアンインパルス波の波形

ガウシアンインパルス波の場合

まず，インパルス波よりももう少し現実的な音源として，ガウシアンインパルス波を検討した．ガウシアンインパルス波の波形は，

$$g(t) = A \exp\left(\frac{-(t-\mu)^2}{\sigma^2}\right) \tag{4.7}$$

のように定義される．パラメータとして μ=0.25 ms, σ=0.1 ms を用いた．波形を図 4.8 に示す．

チャープパルスの場合

次に，ガウシアンインパルス波でも現実の実験に用いる場合には難しいと考えられるため，チャープパルスを検討した．チャープパルスは開始周波数 f_0 と終了周波数 f_1，終了周波数に到達する時間 t_1 に対し，

$$c(t) = A\cos(2\pi f(t)) \tag{4.8}$$

$$f(t) = f_0 + \frac{(f_1 - f_0)t}{t_1} \tag{4.9}$$

のように定義される．実験では，それぞれのパラメータを f_0=100Hz, f_1=1kHz, t_1=1 s とした．このチャープパルスの，周波数の変化がわかりやすい 0〜400ms までの波形を図 4.9 に示す．

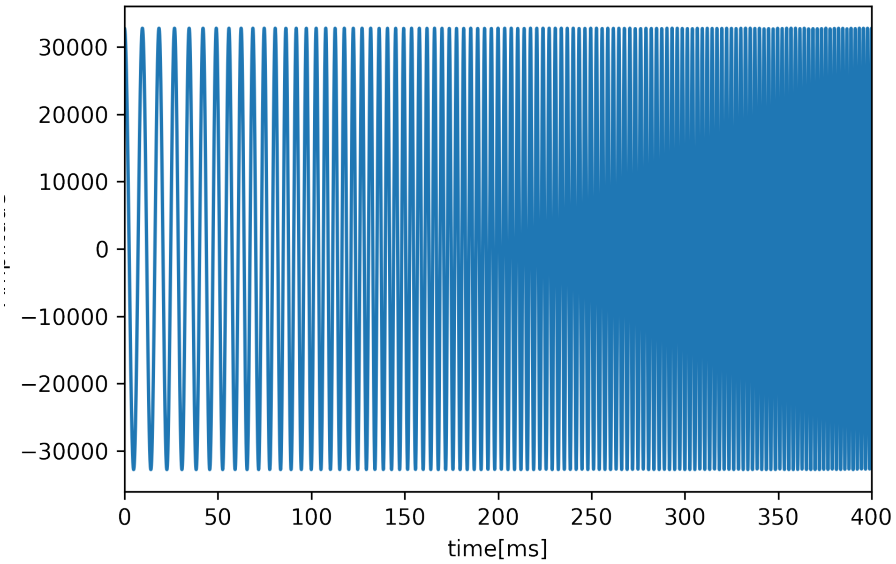

図 4.9: チャープパルスの波形

4.5 結果

ガウシアンインパルス波，チャープパルスそれぞれを音源としてシミュレーションを行い，各マイクロフォンごとに残響除去フィルタを構築した．次いで男性肉声を音源とするシミュレーションを行い，残響除去フィルタつきの CSP 法を適用した．ガウシアンインパルス波，チャープパルスによる残響除去フィルタを適用した後の，CSP 法による結果をそれぞれ図 4.10，図 4.11 に示す．どちらのケースでも CSP シグナルには綺麗なピークが立ち，TDoA はそれぞれ 4.56ms, 4.62ms と検出された．無残響下での結果と比較すると，両ケースとも TDoA を過小評価する傾向が見て取れる．この傾向は肉声音源の種類によらない系統的なものであり，その原因については今後考察を重ねていく必要があると考える．

4.6 考察

結果から，残響除去を行なった結果ピークの乱立は大幅に抑えることができたが，TDoA を過小評価する傾向が見てとれた．この原因として考えられる点はいくつかある．まず一つ目に遅延成分としてゼロをトリムした際の精度に問題があるという点である．誤差が系統的であることから，反響成分と遅延成分を区別するための RIR の立ち上がりをうまく検出できていないことが考えられる．遅延成分を切り分ける際に，遅延成分として考慮するべき基準を，小数点以下の範囲でいくつか検討したが，結果は変化しなかった．また，チャープパルスにおいては，高周波から低周波へと変化するチャープパルスも用いたが，実験結果は変化がなかった．付録に実環境での実験の結果について記載しているが，実環境での実験の際はさらにノイズが乗り，立ち上がりの検出はより難しくなると考えられるため，大きな課題であると思われる．次に，ガウシアンインパルス波とチャープパルスで検出した TDoA に差がある点について考察する．これらの周波数領域でのグラフを図

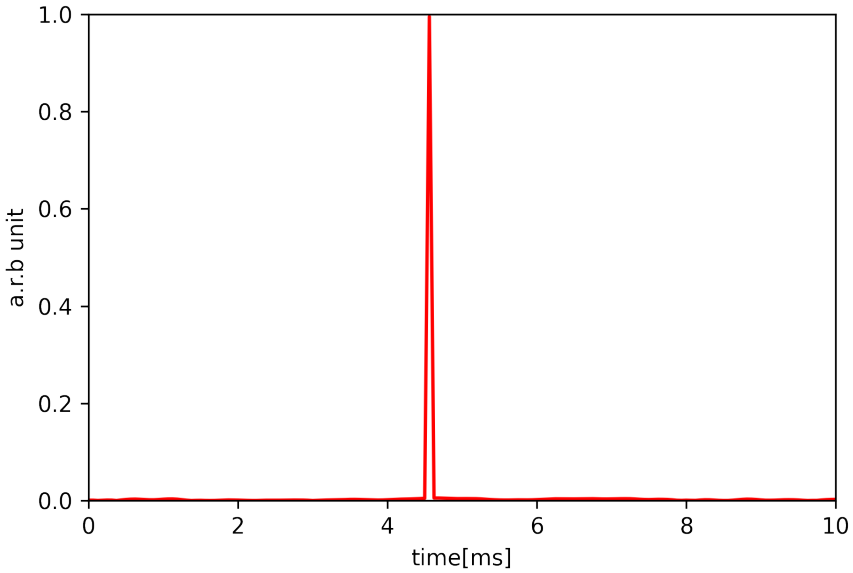

図 4.10: ガウシアンインパルス波による CSP 法

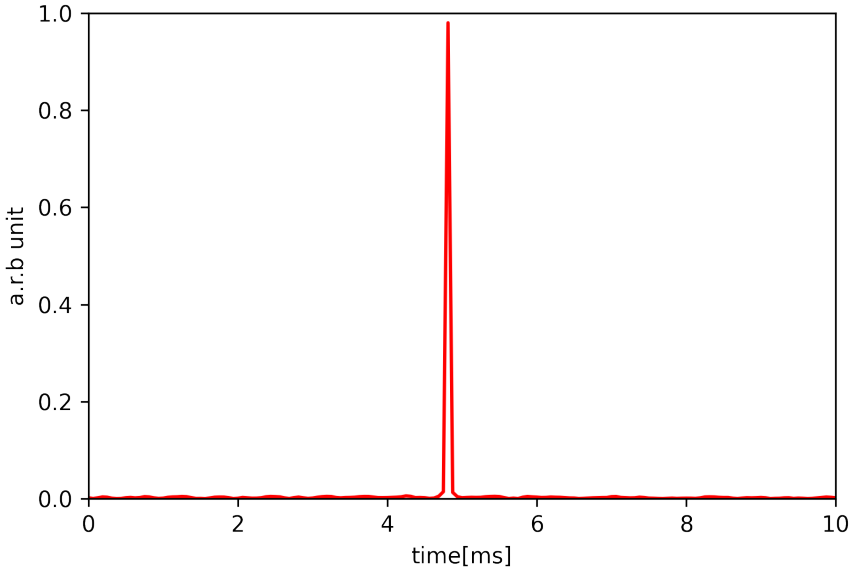

図 4.11: チャープパルスによる CSP 法

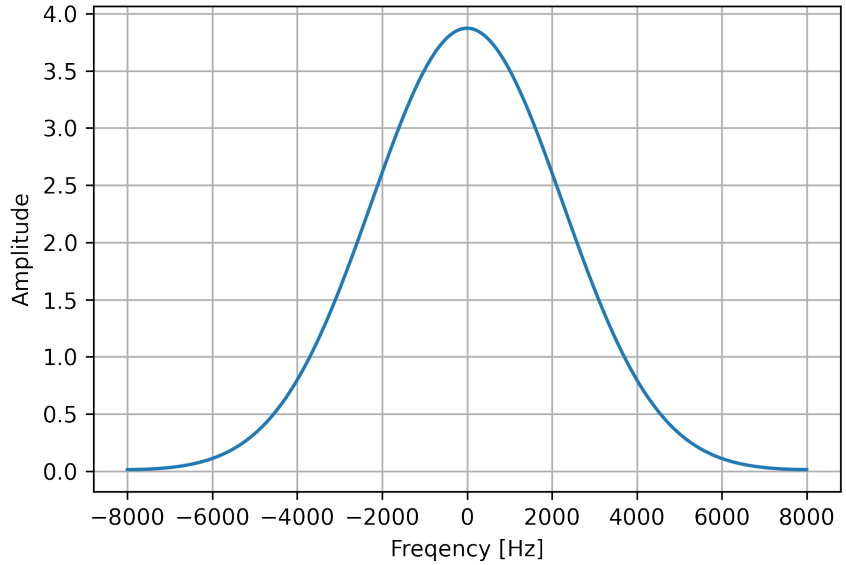

図 4.12: ガウシアンインパルス波の周波数グラフ

4.12，図 4.13 にそれぞれ示す．一般的に，人間の声の周波数は 50〜1000Hz 程度と言われているが，チャープパルスの方がその範囲をより強力にカバーしていることがグラフからわかる．またガウシアンインパルス波は立ち上がりを滑らかに作るために，理想的なインパルス波形と比べてピークが後ろにずれてしまう．そのようなピークのずれも理由の一つだと考えられる．この手法では，本来なら考慮すべきマイクやスピーカーごとのユニークな特性をまとめて関数として考えて除去できる点が大きな利点だと思われる．

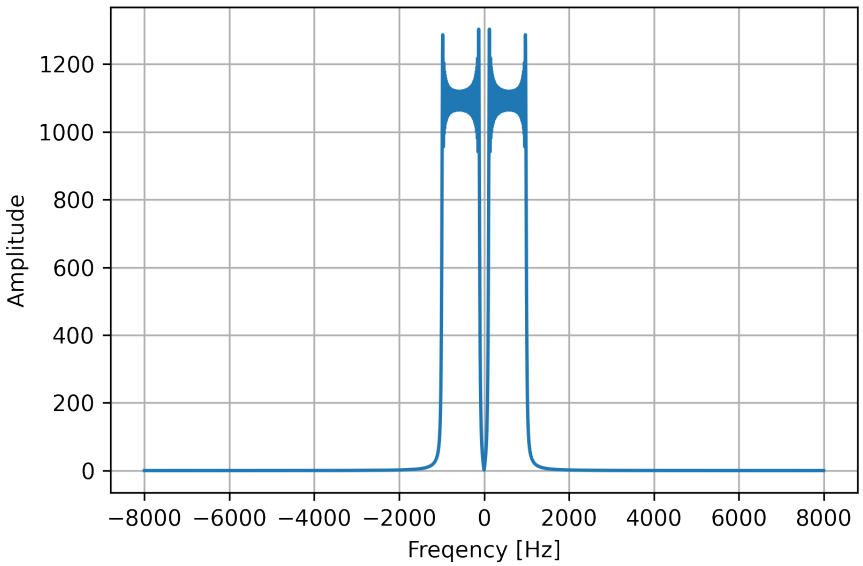

図 4.13: チャープパルスの周波数グラフ

第 5 章

結論と今後の課題

5.1　結論

マイクロフォンと音源の相対的な位置を図るための手段の一つとして，マイクロフォオンペアへの音声の到達時間差を測定し，マイクロフォンと音源のどちらかの座標情報を用いることで，もう片方の座標を推定することができる．TDoA の推定としては，CSP 法と呼ばれる入力信号を周波数領域で白色化した上で相関を求める方法が主流となっている．しかし，音源が複数の場合や，残響が存在する環境下では，CSP 法の精度が低下するという問題があった．そこで本研究では，音源分離を組み合わせることで複数音源の際の精度の向上を図り，特定の音源によるインパルス応答の推測によって残響環境での精度の向上をはかった．そのために pyroomacoustics による室内音響シミュレーションによって音声収録を行い，そのシミュレーションによって得られたマイクロフォンへの入力信号をもとにそれぞれ実験を行なった．その結果，どちらの状況においても精度の向上を達成できることを確認した．

5.2　今後の課題

今後の課題として以下のようなことが挙げられる．まず，実環境での実験の必要性である．本論文の実験は全てシミュレーション上で収録された音声を用いて行なったが，実環境ではシミュレーションでは考慮できないさまざまな雑音が生じ，より精度の向上は難しくなると推察される．特に第四章では到達時間差の推定誤差について，遅延成分と残響成分の RIR の立ち上がりによる分離の精度が問題として考えられ，これは実環境でより大きな問題になると考えられる．次に，複数音源かつ残響環境下での精度の向上である．本論文ではそれぞれ場合を分けて実験を行なったが，実際にはそれらを組み合わせた状態が主に目的とする環境であり，その状況における CSP 法の精度の向上は必要不可欠であると考えられる．こちらについてもシミュレーションによる実験を十分に行なった上で，実環境による実験も行なっていく必要があると考える．

謝辞

本研究を進めるにあたり，指導教員の工藤知宏教授には多大なるご指導ご鞭撻を賜りました．コロナ禍で研究室に行くことができず，新しい分野での研究を自分一人で行なっていくことに不安を抱いていたり，不慣れなことも多かったのですが，貴重なお時間を割いてご指導をしていただきました．深く感謝致します．産業技術総合研究所の池上努さんには，研究に関する技術的なことから，心構えなどの精神的なことまで，技術者としてのさまざまなことを教えていただきました．また研究を進める上でも，助言を多数いただき，多大なるご支援を賜りました．深く感謝致します．家族には，常に自分を応援していただき，さまざまな形で支えていただき，非常に励みになりました．修士２年間で自分を支えてくださった全ての方に感謝致します．

発表文献

[1]鈴木優大，池上努，工藤知宏,”残響のある環境での音声の到達時間差の推定に関する研究”， 信学技報, vol. 121, no. 311, EA2021-58, pp. 7-12, 2021 年 12 月.

参考文献

[1] Tania Habib Muhammad Umair Khan. Concurrent speakers localization using blind source separation and microphone array geometry. *Multidimensional Systems and Signal Processing*, Vol. 32, No. 4, pp. 1159–1184, 2021.

[2] Luiz C. F. Nogueira and Mariane R. Petraglia. Robust localization of multiple sound sources based on bss algorithms. In *2015 IEEE 24th International Symposium on Industrial Electronics (ISIE)*, pp. 579–583, 2015.

[3] H. Buchner, R. Aichner, J. Stenglein, H. Teutsch, and W. Kellermann. Simultaneous localization of multiple sound sources using blind adaptive mimo filtering. In *Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, Vol. 3, pp. iii/97–iii100 Vol. 3, 2005.

[4] Herbert Buchner, Robert Aichner, and Walter Kellermann. *TRINICON-based Blind System Identification with Application to Multiple-Source Localization and Separation*, pp. 101–147. Springer Netherlands, Dordrecht, 2007.

[5] Yoshiki Masuyama, Masahito Togami, and Tatsuya Komatsu. Multichannel loss function for supervised speech source separation by mask-based beamforming. *CoRR*, Vol. abs/1907.04984, , 2019.

[6] Mehdi Azadi and Hamid Reza Abutalebi. Modified state coherence transform to reduce spatial aliasing in tdoa estimation of multiple sound sources. In *7'th International Symposium on Telecommunications (IST'2014)*, pp. 492–496, 2014.

[7] 門倉丈, 川喜田佑介, 五百蔵重典, 田中博. 白色化相互相関法を用いた受信時間差検出による屋内各種音源の測位手法とその実験的評価. 測位航法学会論文誌, Vol. 10, No. 3, pp. 23–32, 2019.

[8] 綱沢駿, 大山真司. 複数フレームの csp 係数を用いた tdoa 推定による複数人物定位・追跡. 計測自動制御学会論文集, Vol. 53, No. 12, pp. 644–653, 2017.

[9] Xinwang Wan and Zhenyang Wu. Sound source localization based on discrimination of cross-correlation functions. *Applied Acoustics*, Vol. 74, No. 1, pp. 28–37, 2013.

[10] Y. Huang, J. Benesty, and G.W. Elko. Adaptive eigenvalue decomposition algorithm for real time acoustic source localization system. In *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No.99CH36258)*, Vol. 2, pp. 937–940 vol.2, 1999.

[11] S. Bedard, B. Champagne, and A. Stephenne. Effects of room reverberation on time-delay estimation performance. In *Proceedings of ICASSP '94. IEEE International Conference on Acoustics, Speech and Signal Processing*, Vol. ii, pp. II/261–II/264 vol.2, 1994.

[12] B. Champagne, S. Bedard, and A. Stephenne. Performance of time-delay estimation in the presence of room reverberation. *IEEE Transactions on Speech and Audio Processing*, Vol. 4, No. 2, pp. 148–152, 1996.

[13] 真人戸上. Python で学ぶ音源分離. 機械学習実践シリーズ. インプレス, 2020.

[14] Pierre Comon. Independent Component Analysis, a new concept? *Signal Processing*, Vol. 36, pp. 287–314, April 1994.

[15] Noboru Murata, Shiro Ikeda, and Andreas Ziehe. An approach to blind source separation based on temporal structure of speech signals. *Neurocomputing*, Vol. 41, No. 1, pp. 1–24, 2001.

[16] Vaninirappuputhenpurayil Gopalan Reju, Soo Ngee Koh, and Ing Yann Soon. Underdetermined convolutive blind source separation via time–frequency masking. *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 18, No. 1, pp. 101–116, 2010.

[17] Taesu Kim, Torbjørn Eltoft, and Te-Won Lee. Independent vector analysis: An extension of ica to multivariate components. In Justinian Rosca, Deniz Erdogmus, José C. Príncipe, and Simon Haykin, editors, *Independent Component Analysis and Blind Signal Separation*, pp. 165–172, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg.

[18] Nobutaka Ono. Stable and fast update rules for independent vector analysis based on auxiliary function technique. In *2011 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 189–192, 2011.

[19] Daichi Kitamura, Nobutaka Ono, Hiroshi Sawada, Hirokazu Kameoka, and Hiroshi Saruwatari. Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 24, No. 9, pp. 1626–1641, 2016.

[20] Daniel D. Lee and H. Sebastian Seung. Learning the parts of objects by nonnegative matrix factorization. *Nature*, Vol. 401, pp. 788–791, 1999.

[21] 亀岡弘和. 深層学習に基づく音源分離. 日本音響学会誌, Vol. 75, No. 9, pp. 525–531, 2019.

[22] 中臺一博. 音響信号処理の変遷と最先端. 日本音響学会誌, Vol. 74, No. 7, pp. 394–400, 2018.

[23] M. Miyoshi and Y. Kaneda. Inverse filtering of room acoustics. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. 36, No. 2, pp. 145–152, 1988.

[24] Tomohiro Nakatani, Takuya Yoshioka, Keisuke Kinoshita, Masato Miyoshi, and Biing-Hwang Juang. Speech dereverberation based on variance-normalized delayed linear prediction. *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 18, No. 7, pp. 1717–1731, 2010.

[25] Keisuke Kinoshita, Marc Delcroix, Tomohiro Nakatani, and Masato Miyoshi. Suppression of late reverberation effect on speech signal using long-term multiple-step linear prediction. *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 17, No. 4, pp. 534–545, 2009.

[26] Hideaki Kagami, Hirokazu Kameoka, and Masahiro Yukawa. Joint separation and dereverberation of reverberant mixtures with determined multichannel non-negative matrix factorization. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 31–35, 2018.

[27] Rintaro Ikeshita, Nobutaka Ito, Tomohiro Nakatani, and Hiroshi Sawada. A unifying framework for blind source separation based on a joint diagonalizability constraint. In *2019 27th European Signal*

Processing Conference (EUSIPCO), pp. 1–5, 2019.

[28] Robin Scheibler and Nobutaka Ono. Fast and stable blind source separation with rank-1 updates. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 236–240, 2020.

[29] Robin Scheibler, Eric Bezzam, and Ivan Dokmanic. Pyroomacoustics: A python package for audio room simulation and array processing algorithms. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr 2018.

[30] John Kominek and Alan Black. The cmu arctic speech databases. *SSW5-2004*, 01 2004.

付録

時間の都合上結果を得ることができなかったが，実験を並行して行なっていた．実験設備として，マイクを第四章と同じように配し，さらに音源近くにもマイクを一つ設置した．それぞれのマイクはオーディオインターフェースによって同期を保ちながら PC へと接続している．また，スピーカーとして今回は PC である Macbook Air を用いた．それらを図 5.1，図 5.2，図 5.3，図 5.4，図 5.5 に示す．

このようなセットアップの上で，音源としてチャープパルスとガウシアンインパルス波を実際に再生し，スピーカー近傍のマイク，マイク 1，マイク 2 の三つのマイクで収録した．それぞれの収録音声の波形を図 5.6，図 5.7 に示す．

第四章の考察で述べたように，立ち上がりがさまざまな雑音によって非常にわかりづらくなっている．立ち上がりの部分を拡大したものを図 5.8，図 5.9 に示す．

実験結果からわかるように，実環境の実験では，より正しく遅延成分と残響成分を切り離すことの難しさが再確認された．

図 5.1: スピーカーから見たマイクの配置

図 5.2: スピーカーから見たマイクの配置

図 5.3: オーディオインターフェース

図 5.4: マイク 1

図 5.5: マイク 2

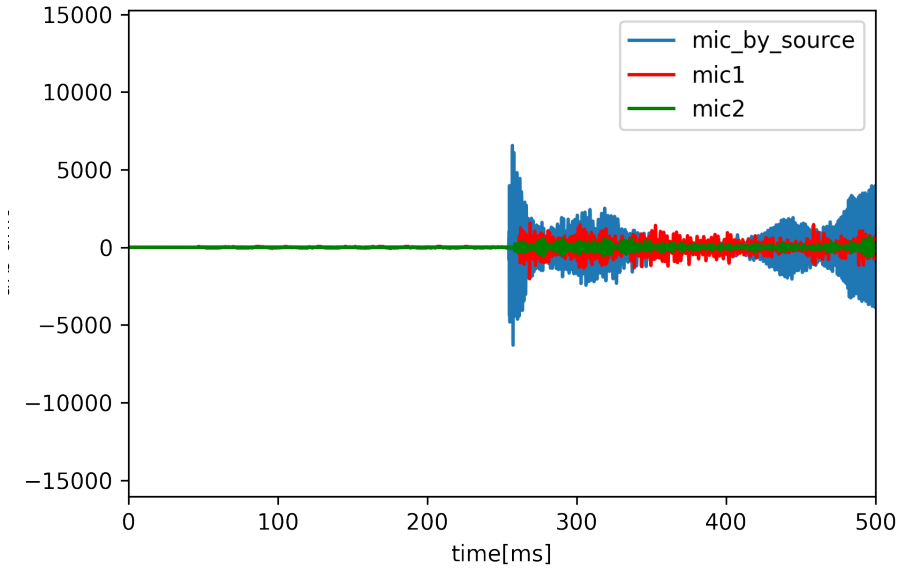

図 5.6: 実験でのチャープパルスを音源とした時のマイクへの入力

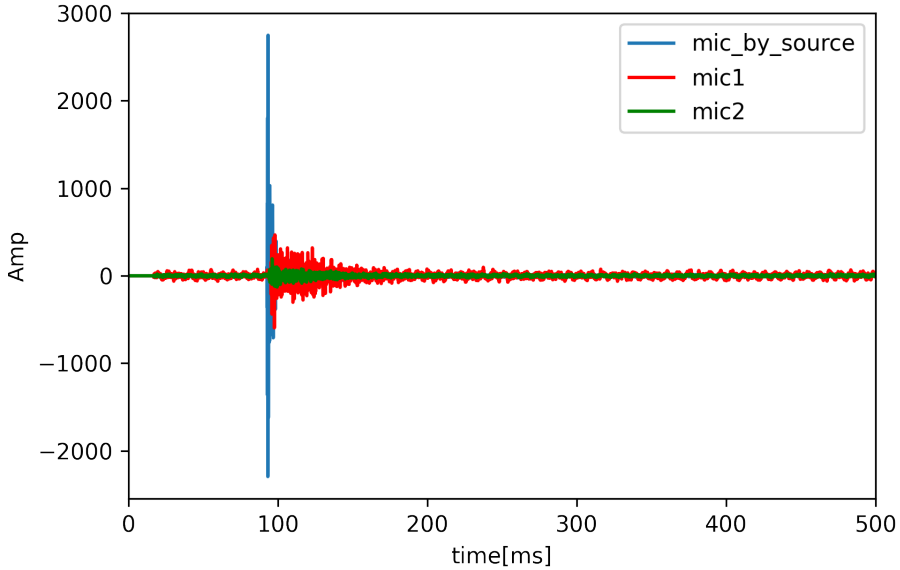

図 5.7: 実験でのガウシアンインパルス波を音源とした時のマイクへの入力

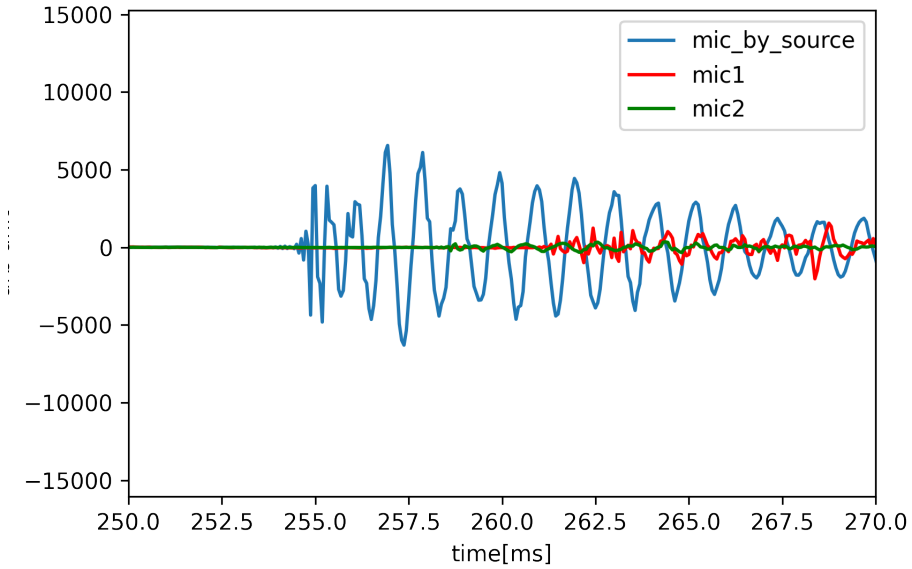

図 5.8: 実験でのチャープパルスを音源とした時のマイクへの入力の立ち上がり付近

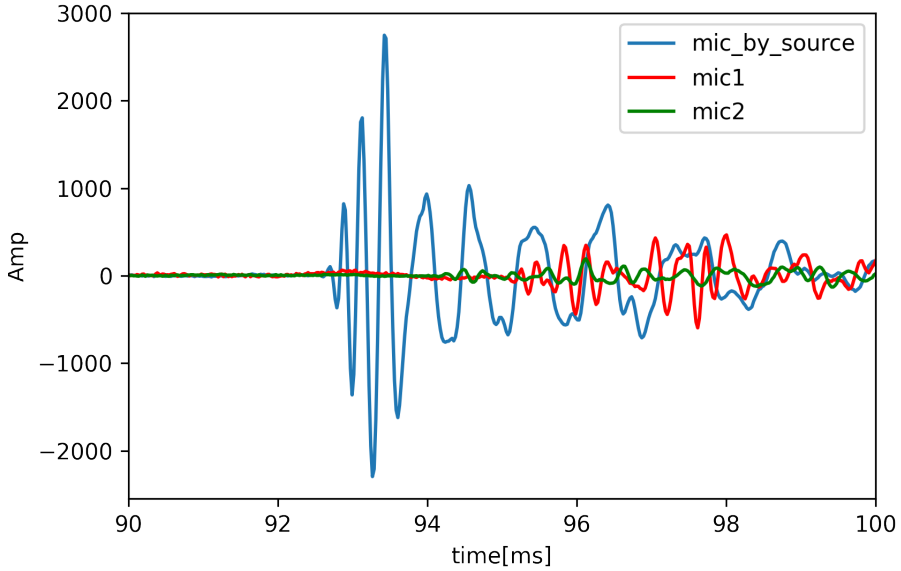

図 5.9: 実験でのガウシアンインパルス波を音源とした時のマイクへの入力の立ち上がり付近