

2010年度 卒業論文

デジタル耳目音声抽出

2011年2月15日

大阪大学 基礎工学部
システム科学科 生物工学コース
若松孝圭

主査: _____ 日付: _____

副査: _____ 日付: _____

概要

近年、音声認識機器への入力や、雑音環境下でのコミュニケーションなどで、聞きたい音声だけを聞くという音声抽出のニーズが高まっている。音声抽出の手法としては、マイクロホンアレイを用いて目的とする音声以外の音を除去することで目的の音声を抽出するものが一般的である。マイクロホンアレイを用いた手法では、音源の位置が既知であるとするものが多い。また、移動する音源の発する音声を抽出することは困難である。

本研究では、音源の位置に応じた音声抽出を行うことを目的とする。提案するシステムでは、ステレオカメラと2chマイクロホンアレイを用いる。まず、ステレオカメラと2chマイクロホンアレイでそれぞれ画像と音声を取得する。次に、ユーザが画像上で音源を指定し、指定された音源の位置をステレオ位置計測によって取得する。取得した音源位置に応じた差分型ビームフォーマによる雑音抑圧を行い、結果をスピーカから出力する。本システムでは、ユーザが音源を指定している間、指定した音源の発する音声を出力することができる。音源が移動する場合にも、音源の移動に合わせてユーザが音源を指定することで、音声抽出が可能となる。

壁等による反響がない、マイクロホン間で感度の違いがないなどの理想的な環境と、実際の環境において、作成したシステムで音声抽出が可能であるか実験した。理想的な環境では、音声抽出が可能であることを確認した。実際の環境では、理想的な環境に比べて性能が劣化するものの、音声抽出が可能であることを確認した。高精度に感度調整が可能なマイクロホンを使用するなどのハードウェアの改良や、反響に対応した処理などができれば実際の環境でも十分な性能を発揮できるだろう。

Abstract

Recently, the need for sound extraction has been increasing for speech recognition and high-quality communications in noisy environments. A microphone array is used to extract sound, but there are some problems. The problems are that microphone array needs to know where the sound sources are, and that it is difficult to track the sound sources.

The goal of this study is to make a sound extraction system which tracks the sound source. A stereo camera and 2ch microphone array are used in this system. First, image and the sound signal are obtained with the camera and microphone respectively. Second, a user chooses a sound source on the image, and the location of the sound source is calculated with the triangulation. Final, the sound is extracted with subtraction beamformer and spectrum subtraction between the microphone array, and the user can hear the sound.

The experiment demonstrates that the proposed system enabled to extract sound in ideal environment where there is no echo nor different characteristics between microphones. The experiment also demonstrates that proposed system can extract sound in real environment. The performance of sound extraction is better in virtual environment than in real environment. To improve the proposed system in real environment, it is necessary to adjust the microphones to the same sensitivity and to cancel echo in the room.

目次

第1章	はじめに	1
第2章	三次元位置計測と音の抽出・抑圧	3
2.1	三次元位置計測法	3
2.2	音の種類と音の抑圧法	5
2.3	指向性音抑圧	6
2.4	拡散性音抑圧	8
第3章	話者の位置計測と音声抽出	10
3.1	システムの構成	10
3.2	ステレオカメラによる位置計測	11
3.3	2ch マイクロホンアレイ 音声抽出	15
第4章	システムの実装と性能の検証	17
4.1	実装システムの構成	17
4.2	位置計測	18
4.3	音声抽出	21
第5章	おわりに	32
	謝辞	33
	参考文献	34

目 次

2.1	カメラと対象物の位置関係	4
2.2	指向性音	6
2.3	拡散性音	6
3.1	システム概要	10
3.2	カメラと音源の位置関係	11
3.3	対応点探索	12
3.4	画像と実物体の距離関係	13
3.5	カメラ、マイクロホン、音源の位置関係	14
3.6	真上から見た場合	14
3.7	目的音がマイクロホンアレイに到達する様子	15
4.1	装置の外観	17
4.2	カメラ	18
4.3	マイクロホン	18
4.4	対応点探索の様子	19
4.5	音源の位置	19
4.6	誤差の距離依存性	20
4.7	音源が (1.0, 3.0) の位置にある場合	21
4.8	音源 A、B、C の配置	22
4.9	仮想環境で音源 A が音声を発生したとき	23
4.10	仮想環境で音源 B が音声を発生したとき	24
4.11	仮想環境で音源 C が音声を発生したとき	25
4.12	実験時の様子	26
4.13	音源の配置	26

4.14 実環境で音源 A が音声を発生したとき	27
4.15 実環境で音源 B が音声を発生したとき	28
4.16 実環境で音源 C が音声を発生したとき	29
4.17 反響の様子	30
4.18 インパルス応答	30
4.19 入力信号の違い	31

目 次

2.1	画像とカメラの記号の定義	4
2.2	音信号の定義	6
2.3	音抑圧技術まとめ	9
4.1	使用する機材	18
4.2	位置計測実験結果	20
4.3	実験環境	21
4.4	各パラメータまとめ	22
4.5	仮想環境と実際の環境の違い	30

第1章 はじめに

普段の生活の中で、私たちは他人とコミュニケーションを交わしながら暮らしている。会話やスキンシップなど、コミュニケーションの形は様々であるが、中でも会話は私たちが最も頻繁に行っているコミュニケーションであろう。会話によってお互いに最も伝えるべきものは、言葉としての意味を持った音、音声である。

人の身の回りにはあらゆる場面で音が存在し、人の話し声、風などの自然の音、機械の作動音など、多種多様な音が存在する。音による情報は人間にとって非常に重要であるが、音による情報は空気の振動として常に混ざり合った状態で人の耳に到達する。人は円滑に会話をするため、混ざり合った音の中から聞きたい音だけを抽出して聞くことを無意識に行っている。例えば、人は複数の話者が同時に話している状態でも、目的とする音声を聞きとり会話することができる。これはカクテルパーティ効果と呼ばれ、人が自然に有する能力である [1]。

また、人は音声抽出を行う際、聴覚によって得た情報のみを利用するのではなく、視覚によって得られる情報を有効に利用している [1-3]。視覚と聴覚の相互作用を示す例として腹話術効果がある [4]。腹話術効果とは、音源と映像の空間的位置が一致しない場合でも、音が映像の方向から聞こえるように錯覚する現象である。例えば、映画やテレビを見ているとき、スピーカが画面の隣あるいは画面から離れた位置に設置されているにも関わらず、人は音が画面の方向から聞こえるように感じる。腹話術効果が表しているように、視覚と聴覚は密接に関わりあっている。人が音声抽出を行う際にも、音声は相手が見える方向から聞こえるなど、視覚によって得た情報を有効に利用していると考えられる。

人が行っている音声抽出をコンピュータで実現することに大きな期待が寄せられている。たとえば、音声認識の分野では、目的とする音声を背景雑音から抽出できれば、騒がしい環境下での認識性能が向上すると期待されている [5]。ほかにも、防犯カメラなどに映った人物の発話内容を正確に聞き取ることができれば、早期の

犯罪防止や事件解決に貢献できるという点や、工場などの騒がしい環境下で、故障により異音を発生する機械を発見するなどの応用にも期待できる。

音声抽出の研究の歴史は古く、現在までに多くの音声抽出の手法が提案されてきた [6, 7]。中でも、マイクロホンアレイを用いる手法が効果的である [8, 9]。マイクロホンアレイを用いた方法では、マイクロホンに到達する信号の時間差を利用して、特定の方向から到来する信号を強調あるいは抑圧することができる [5]。しかし、マイクロホンアレイを用いた方法では、音源位置が既知でなければならないものが多く、また、話者の動きにより音源位置が一定でないことから、本来強調されるべき目的信号の一部が抑圧され、ひずみが生じるという問題がある [10]。コンピュータによる処理においても人と同じように、音声情報のみに頼るのではなく、画像情報も合わせて音源位置を取得することで、より正確な位置情報を得ることができ、音声抽出の助けになると考えられる。

人の視覚に基づく知覚、認識等をモデル化し、コンピュータで実現させようとする研究分野をコンピュータビジョンという [11–13]。コンピュータビジョンの分野では、人間の両眼の持つ優れた三次元認識能力をコンピュータに与えることが、大きなテーマの一つである。画像を取得する手段として、最も人間の視覚に近いデバイスはカメラである。カメラを用いて対象物体の三次元位置を取得する方法として、レンズ焦点法、ステレオ画像法、単眼視法、連続画像法など、現在までに多くの手法が提案され、実用化されている [11]。

本研究では、人の視覚、聴覚を模し、カメラとマイクロホンアレイを用いて目的音源の位置に応じた音声抽出を行うことを目的とする。提案するシステムではまず、カメラとマイクロホンアレイで画像と音を同時に取得する。次に、カメラで取得した画像上にある音源の中からユーザが音声を聞きたい人物を指定し、差分型ビームフォーマによって指定した音源の方向にマイクロホンアレイの感度の指向性を形成することで目的音声を周囲の音から抽出し、スピーカから出力することで目的を達成する。

第2章 三次元位置計測と 音の抽出・抑圧

本章では、現在までに考案されてきた三次元位置計測法と、音の抽出・抑圧法について述べる。まず、三次元位置計測法について述べ、次に音の種類、各種の音の抽出あるいは抑圧法について述べる。

2.1 三次元位置計測

三次元位置計測 [11] は能動法と、受動法に大別できる。能動法とは、対象物に対して計測の補助となる光・電波・音波等を照射し計測を行う方法であるのに対し、受動法とは、対象物に対して計測の補助となる光等を照射せず、カメラのみにより計測を行う方法である。人の視覚系は位置計測に受動法を用いており、カメラの代わりに目を使用している。本節では、受動法について述べていく。

2.1.1 単眼視法

人間は、遠近法書かれた絵画を見るときのように、物の見かけの大きさや、無限遠点の消失点等から、両眼で見なくても奥行きを知覚することができる。ただし、単眼視によって得られる奥行きは、相対的な奥行きであり、絶対的な三次元位置を得るためには、物体の大きさが既知である、一定の平面上にしか対象が存在しない等の拘束条件が必要である。

カメラを用いて単眼視による位置計測を行う場合にも、画像と物体のモデルとの対応を取る等、複雑な処理が必要となるが、そもそも未知の形状の物体を計測することを目的としているので、単眼視法が三次元画像計測に利用されることはない。

2.1.2 ステレオ画像法

受動法の内、人間の両眼のようにカメラを二つ用いたステレオ画像法がある。ステレオ画像法の概要を図 2.1 に示す。

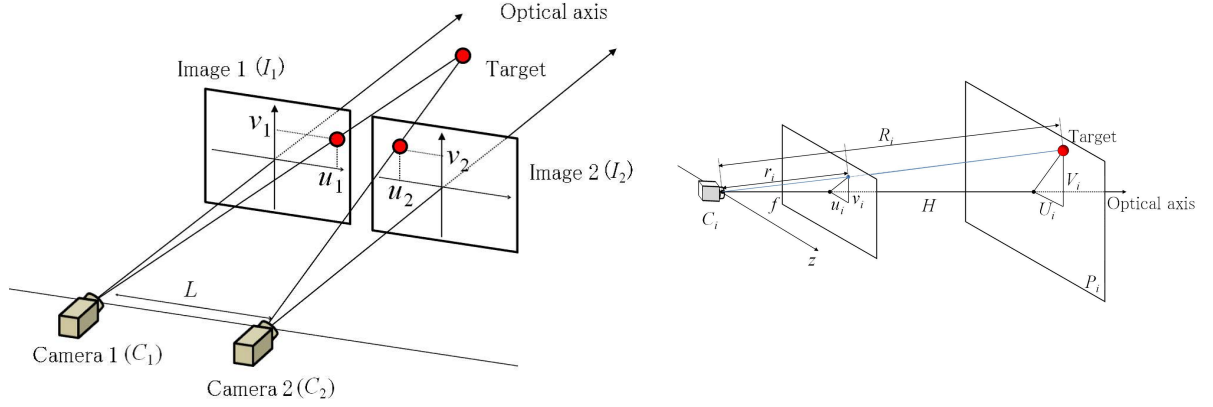


図 2.1: カメラと対象物の位置関係

左側のカメラを C_1 、右側のカメラを C_2 とする。また、 $C_i (i = 1, 2)$ が取得する画像を I_i とする。二つのカメラはある直線 z 上に設置されており、カメラの光軸は直線 z に対して垂直かつ光軸同士は平行であり、画像 I_1 、 I_2 のエピポーラ線は同一水平線上にあると仮定する。また、二つのカメラの画角、焦点距離、画素数は同じであるとする。対象物は画像と平行で、 C_i の光軸との交点を原点とする実平面 P_i に存在すると仮定する。本論文で用いる記号は表 2.1 のように定義する。

表 2.1: 画像とカメラの記号の定義

u_i	I_i 上の対象物の x 座標
v_i	I_i 上の対象物の y 座標
U_i	平面 P_i 上の対象物の x 座標
V_i	平面 P_i 上の対象物の y 座標
f	カメラの焦点距離
r_i	I_i 上の対象物と C_i までの距離
R_i	対象物と C_i までの距離
L	カメラ間の距離
H	直線 z から平面 P までの垂線の長さ

ステレオ画像法では、二つのカメラで撮影される二つの画像からカメラ間での視差を求め、視差からカメラと対象物との距離 R_i を推定する。ステレオ画像法の問題

点は、一様な表面を持った物体に対しては、三次元計測を行うことができないということである。

2.1.3 レンズ焦点法

レンズ焦点法は、一眼レフカメラの焦点合わせと同じ原理を用いる。一眼レフカメラでは、ファインダを覗いて、フォーカスリングを回してピントを合わせ、シャッターを押すが、このときのレンズの距離目盛を読めば、光学式距離計測になる。つまり、レンズの繰り出し機構と画像のボケを判定する計測機能があれば、三次元の奥行きを計測することができる。画像のボケ具合は濃淡の分散から定量することができる。鮮明な画像ほどその分散が大きくなる。入力画像の分散が最大になるようにレンズを駆動し、その時のレンズの位置から簡単に距離を知ることができる。

正確にピント位置を決めるためにはボケを大きくしなくてはならないが、ボケを大きくするためには非常に焦点深度の浅い、口径の大きなレンズが必要になる。顕微鏡のようなレンズ系では焦点深度が極めて浅くなるので、レンズ焦点法利用した微小高さ系が実用されている。しかし、数 m 以上の測定に対しては十分に浅い焦点深度を実現することは難しく、十分な計測精度を得ることはできない。

2.2 音の種類と音の抑圧法

人間が聞きたい音というのは、目的音以外の音を取り除かれた音に他ならない。聞きたい音だけを抽出するには、目的とする音以外の音をどれだけ抑圧、あるいは除去できるかが重要である。

音には大きく分けて、指向性音と拡散性音が存在する。指向性音は、点音源とみなすことができるものから発せられる音である。例えば、人が発した音声、スピーカ等から発せられる音声や音楽などである。拡散性音は、特定の到来方向を有しない音である。例えば、室内に響く空調の作動音などである。

音を抑圧するには、複数のマイクロホンを並べたマイクロホンアレイを用いた手法が効果的である [8]。マイクロホンアレイを用いた手法は、単一のマイクロホンを用いた手法に比べ、計算コスト、抑圧性能が優れている。図 2.2 に指向性音、図 2.3 に拡散性音がマイクロホンアレイに到来する様子をそれぞれ示す。

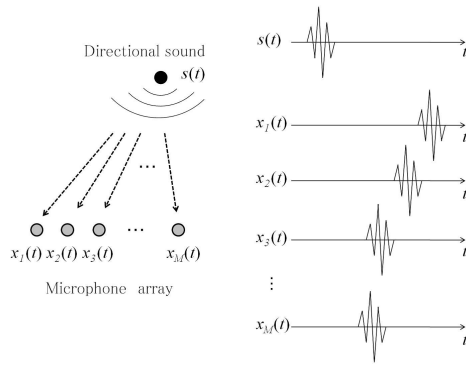


図 2.2: 指向性音

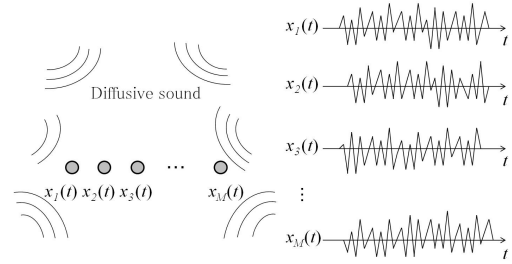


図 2.3: 拡散性音

指向性音の場合、各マイクロホンで観測される信号に遅延が生じるため、各マイクロホンに生じる時間差を有効に利用することで、指向性音を抑圧することができる。拡散性音は、定常状態に近く、各マイクロホンが受信する信号間の相関が低いことを利用した抑圧手法が存在する。

指向性音、拡散性音に対して、それぞれに有効な抑圧手法を用いて周囲の音を抑圧することで、目的音を抽出することができる。信号処理の分野において音信号は、時間領域でのたたみ込みが、周波数領域では積算となるなど計算的な利点から周波数領域で扱われることが多い [14]。

次節以降で使用する記号は表 2.2 のように定義する。なお、 t は時間、 ω は周波数、 M はマイクロホン総数を示している。

表 2.2: 音信号の定義	
$x_i(t)$	i 番目のマイクロホンの観測信号 ($i = 1, 2, \dots, M$)
$X_i(\omega)$	$x_i(t)$ の周波数領域表現
$s(t)$	目的音信号
$S(\omega)$	$s(t)$ の周波数領域表現
$\hat{s}(t)$	音抑圧処理による推定目的音信号
$\hat{S}(\omega)$	$\hat{s}(t)$ の周波数領域表現

2.3 指向性音抑圧

本節では、今までに提案されてきた指向性音抑圧技術について述べる。前述した通り、指向性音抑圧にはマイクロホン間に生じる時間差を用いた手法が効果的である。

2.3.1 DS 法

空間的な指向性を形成する技術をビームフォーミングという。DS (Delay and Sum) 法は、目的とする音源の方向にマイクロホンアレイの感度の指向性を形成する最も単純なビームフォーミング手法である。各マイクロホンでの目的信号の遅延をなくしたうえで信号の平均をとることで、目的信号は同相化して強められ、目的としない信号は位相が異なるため弱められる。

$$\hat{s}(t) = \frac{1}{M} \sum_{i=1}^M x_i(t - \tau_i) \quad (2.1)$$

式 (2.1) 中の τ_i は i 番目のマイクロホンにおける遅延量である。

DS 法は、目的とする方向に指向性を向けることができる単純な手法であるが、実用的な性能を実現するためには大量のマイクロホンが必要となる。また、目的音源位置が既知である必要がある。

2.3.2 減算型ビームフォーマ

特定方向の感度を下げる減算型ビームフォーマと呼ばれる技術がある。大量のマイクロホンが必要な DS 法に対して、減算型ビームフォーマは、二つのマイクロホンがあれば実現できる。二つのマイクロホンの観測信号を減算すれば、各マイクロホンに同位相で入力した信号は打ち消し合い、有効に除去することができる。除去すべき信号が二つのマイクロホンに、ある遅延量 τ に相当する方向から到来する場合、式 (2.2) のように、片側の観測信号の遅延をなくし位相を揃えたものを減算すれば、除去すべき信号を相殺することができる。

$$\hat{s}(t) = x_2(t - \tau) - x_1(t) \quad (2.2)$$

なお、任意の方向からの音を除去するためには、指向性音の音源の位置情報が必要である。

2.3.3 適応ビームフォーマ

小規模なマイクロホンアレイを用いた指向性雑音抑圧の技術として、適応ビームフォーマが存在する [14–16]。適応ビームフォーマは事前に目的信号が存在しない状

況において、非目的音を抑圧する指向特性を形成し、混合音から目的音のみを抽出する手法である。適応ビームフォーマは比較的少量のマイクロホンで、高い分離性能が得られるという利点があるが、非目的音学習区間と目的音位置の二つの事前情報が必要となる。実際の応用においては、二つの事前情報が得られないことが多く、適用が困難であるという問題点がある。

2.3.4 ブラインド音源分離

上述した技術はいずれも音源位置など事前情報を必要とするが、これに対して、音源に関する事前情報が得られない状況下で、観測信号のみから音源を分離する技術をブラインド音源分離という。近年、音源信号が統計的に互いに独立であるという仮定のみを用いて多チャンネルの観測信号のみから分離を行う独立成分分析に基づくブラインド音源分離が盛んに研究されている [9] [17]。

ブラインド音源分離は、事前情報なしに音を分離することができるという利点があるが、演算量が多いことや、統計量の推定精度の問題により、実時間処理が困難であったり、複雑な音響環境への対応が不十分であるという欠点もある。

2.4 拡散性音抑圧

拡散性雑音を除去する技術として、スペクトルサブトラクションなどがよく知られている [8] [18]。

2.4.1 スペクトルサブトラクション

スペクトルサブトラクションは、非目的音が定常であると仮定し、非目的音区間の信号から雑音のパワースペクトルを推定し、推定した音成分を入力信号から引き去ることで目的音抽出を行う方法である。簡単な原理を用いているが、定常音抑圧には非常に効果的であることから広く用いられている。欠点としては、非定常音の抑圧が困難であることや、発話区間の推定が必要になることなどがある。

以上、指向性音と拡散性音を抽出あるいは抑圧する技術について述べた。説明した雑音抑圧技術技術の特徴を表 2.3 にまとめる。

表 2.3: 音抑圧技術まとめ

技術名	必要な情報
DS 法	音源の位置
減算型ビームフォーマ	音源の位置
適応ビームフォーマ	音源の位置、発話区間
ブラインド音源分離	なし
スペクトルサブトラクション	発話区間

上述した技術は、目的音源の位置などの事前情報を必要とするが、必要とする情報をマイクロホンから得られる信号のみを用いて推定するのは難しい。また、仮に事前情報を得ていたとしても、目的音源位置が移動した場合のように、事前情報と異なる音響環境になると効果的な雑音抑圧ができなくなるといった問題点がある。これら二つの問題に対し、本研究では、ステレオカメラによる位置計測を用いて、静止あるいは移動する目的音源位置を取得し、音源位置を利用した差分型ビームフォーマとスペクトルサブトラクションによる音声抽出システムを提案する。

第3章 話者の位置計測と音声抽出

本章では提案するシステムについて説明する。まず、システム全体の構成について述べる。次にステレオカメラによる目的音源の位置推定の方法、音声抽出の方法について述べる。

3.1 システムの構成

図 3.1 に提案するシステムの概要を示す。

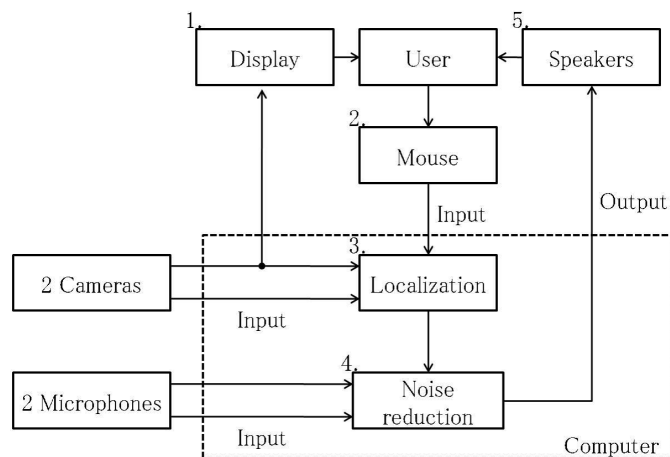


図 3.1: システム概要

1. 一方のカメラで画像を取得し、ユーザに表示する。
2. ユーザは画像を見てマウスで画像上で音源を指定する。
3. 指定された音源の位置をステレオカメラ位置計測によって推定する。
4. 音源の位置から推定されるマイクロホンへの音の到達時間差をもとに音声抽出を行う。
5. 抽出した音声をスピーカから出力する。

3.2 ステレオカメラによる位置計測

ステレオカメラにより、目的音源の位置を得る方法について述べる [19] [20]。図 3.2 はステレオカメラと目的音源の位置関係を示している。

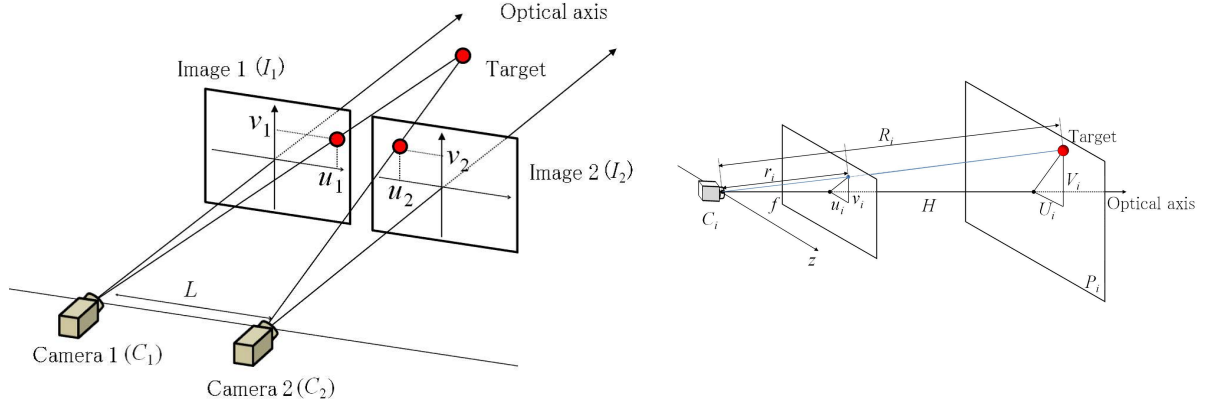


図 3.2: カメラと音源の位置関係

2.1.2 項と同様に左側のカメラを C_1 、右側のカメラを C_2 とする。また、 $C_i (i = 1, 2)$ が取得する画像を I_i とする。二つのカメラはある直線 z 上に設置されており、カメラの光軸は直線 z に対して垂直かつ光軸同士は平行であり、画像 I_1 、 I_2 のエピポーラ線は同一水平線上にあると仮定する。また、二つのカメラの画角、焦点距離、画素数は同じであるとする。目的音源は画像と平行で、 C_i の光軸との交点を原点とする実平面 P_i に存在すると仮定する。

ユーザは画像 I_1 を見て目的音源を指定し、 (u_1, v_1) を決定する。 (u_1, v_1) をもとに対応点探索を行い (u_2, v_2) を決定する。対応点探索は次のようにして行う。

I_1 、 I_2 はどちらも幅 W_I 、高さ H_I の画像であり、対応点は I_1 、 I_2 で同じ高さに存在するとする。画像の左下のピクセル位置が $(0, 0)$ 、右上のピクセル位置が $(W_I - 1, H_I - 1)$ となるような座標を考える。ユーザが I_1 のうち、 (a, b) の位置のピクセルを指定したとき、 (a, b) の位置のピクセルを左下とし、幅 W_s 、高さ H_s となる矩形領域 1 を形成する (図 3.3)。

矩形領域 1 の中のあるピクセル $(a+i, b+j)$ の画素値を RGB の順に $I_{1R}(a+i, b+j)$ 、 $I_{1G}(a+i, b+j)$ 、 $I_{1B}(a+i, b+j)$ とする。ただし、 $(i = 0, \dots, W_s - 1)$ 、 $(j = 0, \dots, H_s - 1)$ である。

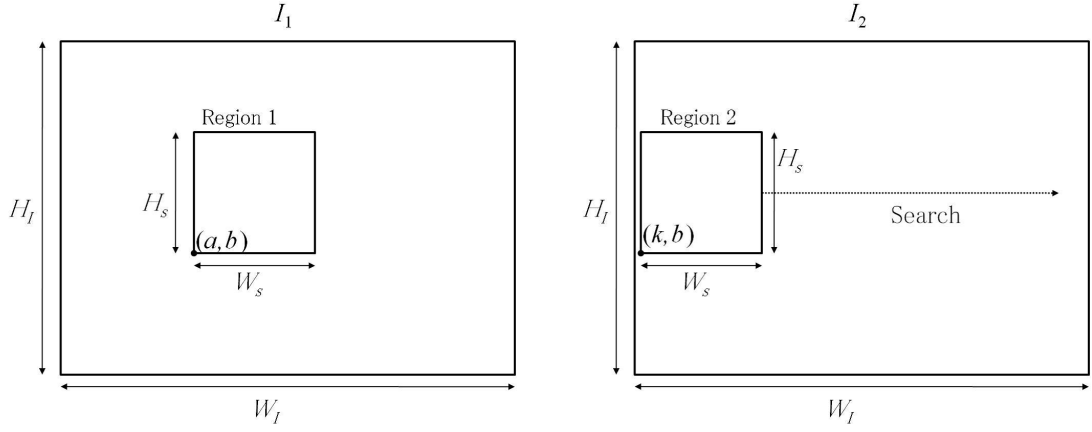


図 3.3: 対応点探索

同様に、 I_2 においても、矩形領域 1 と同じ高さで、同じ大きさをもつ矩形領域 2 を形成する。すなわち、図 3.3 の I_2 のように、矩形領域 2 は (k, b) の位置のピクセルを左下とし、幅 W_s 、高さ H_s である ($k = 0, \dots, W_I - 1$)。矩形領域 2 の中のあるピクセル $(k + i, b + j)$ の画素値を RGB の順に $I_{2R}(k + i, b + j)$ 、 $I_{2G}(k + i, b + j)$ 、 $I_{2B}(k + i, b + j)$ とする。ただし、($i = 0, \dots, W_s - 1$)、($j = 0, \dots, H_s - 1$) である。

形成した矩形領域 2 の k の値を変えながら矩形領域 1、2 を比較することで探索を行う。具体的には、矩形領域 1、2 内のピクセルの画素値の差の絶対値の和が最小になるピクセル (k, b) が I_1 中の (a, b) に対応するピクセルとなる。すなわち、矩形領域 1、2 のピクセルごとの輝度値の差の絶対値を RGB の順にそれぞれ

$$I_R(i, j) = |I_{1R}(a + i, b + j) - I_{2R}(k + i, b + j)| \quad (3.1)$$

$$I_G(i, j) = |I_{1G}(a + i, b + j) - I_{2G}(k + i, b + j)| \quad (3.2)$$

$$I_B(i, j) = |I_{1B}(a + i, b + j) - I_{2B}(k + i, b + j)| \quad (3.3)$$

としたとき

$$R(k) = \sum_{i=0}^{W_s-1} \sum_{j=0}^{H_s-1} (I_R(i, j) + I_G(i, j) + I_B(i, j)) \quad (3.4)$$

が最小となる k が (a, b) の対応点 (k, b) の x 成分となる。

求めた (a, b) 、 (k, b) から (u_1, v_1) 、 (u_2, v_2) を求めるには、

$$u_1 = a - \frac{W_I}{2} + 0.5 \quad (3.5)$$

$$v_1 = b - \frac{\bar{H}_I}{2} + 0.5 \quad (3.6)$$

$$u_2 = k - \frac{\bar{W}_I}{2} + 0.5 \quad (3.7)$$

$$v_2 = b - \frac{\bar{H}_I}{2} + 0.5 \quad (3.8)$$

とすればよい。

以上で指定した音源の画像上の位置を求めることができる。次に、指定した音源の実際の三次元空間での位置を求める。図 3.2 において、 u_1 と u_2 の差を

$$d = u_1 - u_2 \tag{3.9}$$

とすると、 H は

$$H = \frac{L}{d}f \quad (3.10)$$

となる。図 3.4 に画像と実物体の距離の関係を示す。

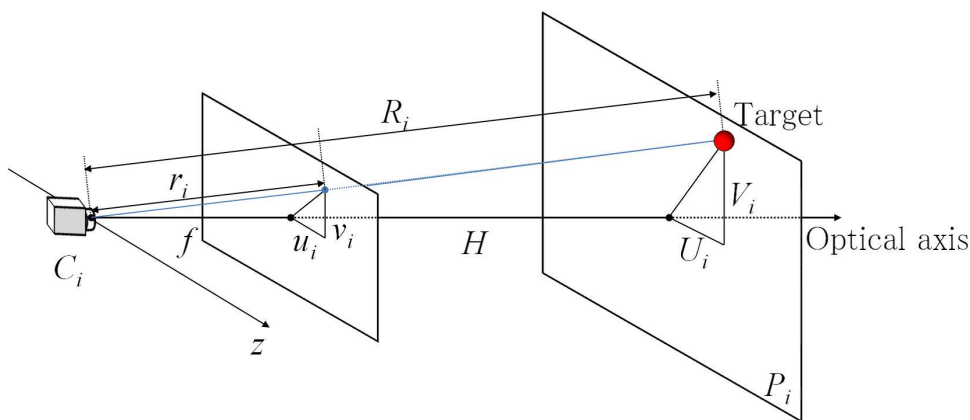


図 3.4: 画像と実物体の距離関係

図 3.4 において、 $r_i : R_i = f : H$ であることから

$$R_i = r_i \frac{H}{f} = \sqrt{u_i^2 + v_i^2 + f^2} \frac{H}{f} \quad (3.11)$$

となり、式 (3.10) を用いると

$$R_1 = \frac{L}{d} \sqrt{u_1^2 + v_1^2 + f^2} \quad (3.12)$$

$$R_2 = \frac{L}{d} \sqrt{u_2^2 + v_2^2 + f^2} \quad (3.13)$$

となる。

図 3.5 に目的音源とステレオカメラカメラ、2ch マイクロホンアレイの位置関係を示す。図 3.6 は図 3.5 を真上から見た場合である。左側のマイクロホンを M_1 、右側のマイクロホンを M_2 とする。二つのマイクロホンは直線 z 上にあり、目的音源と M_i の距離を R'_i 、 C_i と M_i の距離を l とする。

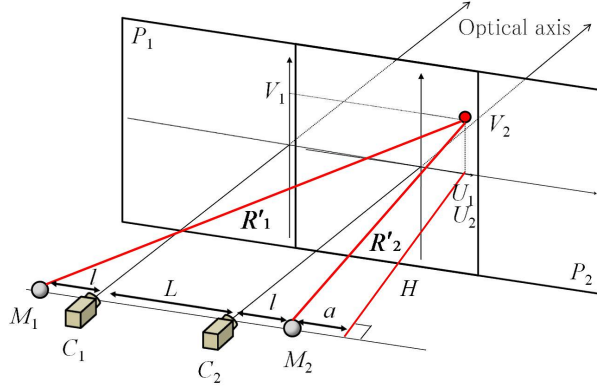


図 3.5: カメラ、マイクロホン、音源の位置関係

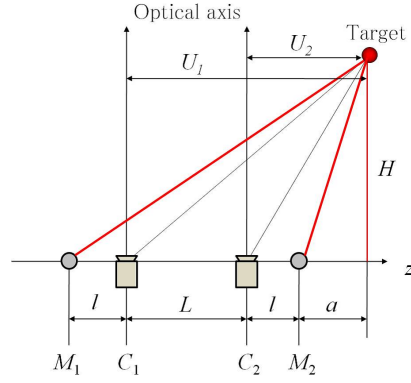


図 3.6: 真上から見た場合

目的音源のある平面から直線 z に垂直に下した点と M_2 との距離を a とすると、次の関係が成り立つ。

$$R'_2 = \sqrt{a^2 + H^2 + V_2^2} \quad (3.14)$$

$$a = U_2 - l \quad (3.15)$$

式 (3.14)、(3.15) から

$$U_i = \frac{H}{f} u_i \quad (3.16)$$

$$V_i = \frac{H}{f} v_i \quad (3.17)$$

であることに注意して a を消去すると

$$R'_2 = \sqrt{\left(\frac{L}{d}\right)^2 (u_2^2 + v_2^2 + f^2) + l^2 - 2\frac{L}{d}lu_2} \quad (3.18)$$

$$= \sqrt{R_2^2 + l^2 - 2\frac{L}{d}lu_2} \quad (3.19)$$

として R'_2 が求められる。同様にして R'_1 を求めると、

$$R'_1 = \sqrt{R_1^2 + l^2 + 2\frac{L}{d}lu_1} \quad (3.20)$$

となる。

3.3 2ch マイクロホンアレイ音声抽出

図 3.7 に、目的音が遅延 τ_s を伴って M_1 、 M_2 に到達する様子を示す。

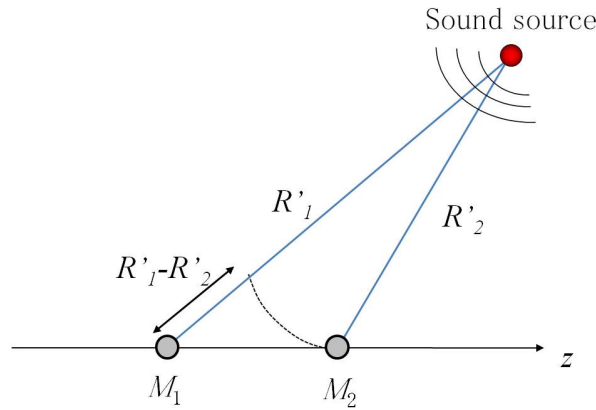


図 3.7: 目的音がマイクロホンアレイに到達する様子

τ_s は、音速を c とすると

$$\tau_s = \frac{R'_1 - R'_2}{c} \quad (3.21)$$

として求められる。

マイクロホン M_i が受け取る信号を $x_i[n]$ とする。 n はデータ番号である。目的音のマイクロホン間での遅延データ数を n_s 、サンプリング周波数を f とすると

$$n_s = \tau_s f \quad (3.22)$$

となる。ただし、式 (3.22) の左辺は整数に限られるため、右辺の小数点以下を四捨五入する。

目的音源方向に死角を向ける差分型ビームフォーマを用いる。目的音成分を除去した差分フィルタ $y[n]$ は

$$y[n] = x_1[n] - x_2[n - n_s] \quad (3.23)$$

となる。 $y[n]$ は雑音成分のみを含んでいる。窓関数

$$w_m[n] = \begin{cases} 1 & (mF \leq n < (m+1)F, m = 0, 1, 2, \dots) \\ 0 & (\text{otherwise}) \end{cases} \quad (3.24)$$

を入力信号 $x_i[n]$ に乗じた信号を

$$x_{i,m}[n] = x_i[n]w_m[n] \quad (3.25)$$

とし、 $x_{i,m}[n]$ をデータ数 F で離散フーリエ変換したものを

$$X_i[k, m] = \sum_{n=mF}^{(m+1)F-1} x_{i,m}[n] \exp\left(-j\frac{2\pi nk}{F}\right) \quad (3.26)$$

とすると、式 (3.23) の短時間フーリエ変換は

$$Y[k, m] = X_1[k, m] - \exp(-j2\pi f\tau_s)X_2[k, m] \quad (3.27)$$

となる。推定目的音の周波数領域の表現である $\hat{S}[k, m]$ は原信号から雑音成分をスペクトルサブトラクションすることで得られる。すなわち

$$|\hat{S}[k, m]| = \begin{cases} \sqrt{|X_1[k, m]|^2 - |Y[k, m]|^2} & (\text{if } |X_1[k, m]| > |Y[k, m]|) \\ 0 & (\text{otherwise}) \end{cases} \quad (3.28)$$

$$\hat{S}[k, m] = |\hat{S}[k, m]| \exp(j\theta) \quad (3.29)$$

$$\theta = \tan^{-1} \frac{\text{Im}(X_1[k, m])}{\text{Re}(X_1[k, m])} \quad (3.30)$$

として得られる。 $\hat{S}[k, m]$ を逆離散フーリエ変換すると

$$\hat{s}_m[n] = \hat{s}[n]w_m[n] = \frac{1}{F} \sum_{k=mF}^{(m+1)F-1} \hat{S}[k, m] \exp\left(j\frac{2\pi nk}{F}\right) \quad (3.31)$$

となり、

$$\hat{s}[n] = \sum_{m=0}^{\infty} \hat{s}_m[n] \quad (3.32)$$

として推定目的信号が得られる。

第4章 システムの実装と性能の検証

前章で述べた理論をもとに、静止する音源が発する音声を抽出するシステムを構築する。構築したシステムを用いて、雑音のある環境下で、位置の異なる音源の音声を抽出する実験を行う。本章では、システムの実装と行った実験について述べる。

4.1 実装システムの構成

図 4.1 に装置の外観を示す。

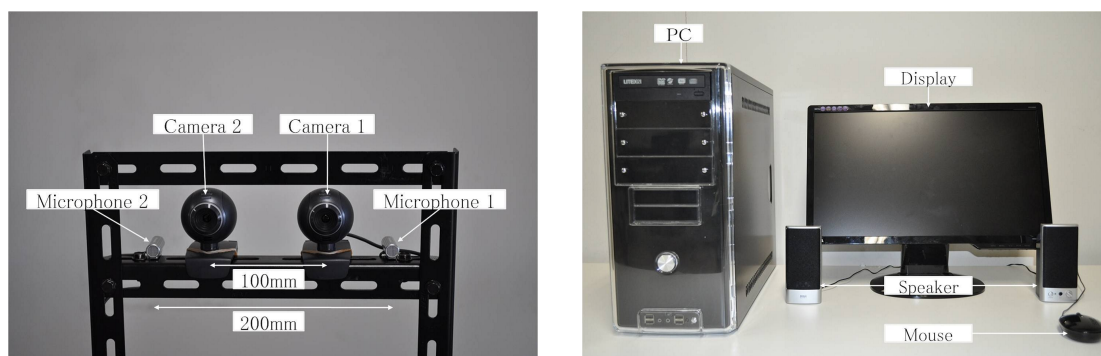


図 4.1: 装置の外観

カメラとマイクロホンは左右対称になるように配置する。カメラ間の距離は 100 mm、マイクロホン間の距離は 200mm となっている。本システムは、画像と音、画像上の音源の座標を取り込む入力部と、音源位置推定と音声抽出を行う演算部、画像と音を出力する出力部に分かれている。図 4.2、図 4.3 に使用したカメラとマイクロホンをそれぞれ示す。表 4.1 に使用した機材の仕様を示す。



図 4.2: カメラ

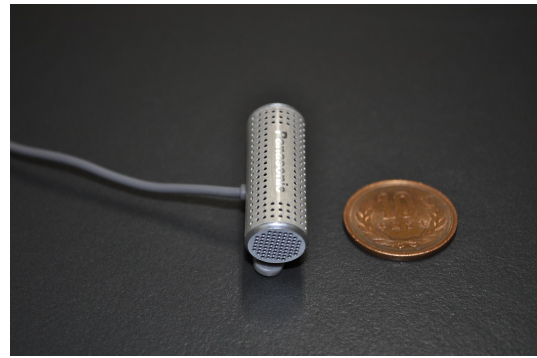


図 4.3: マイクロホン

表 4.1: 使用する機材

機材	詳細	
PC	OS CPU メモリ	Ubuntu Linux 10.10 Intel(R) Core(TM)2 Duo 3.00GHz 2.0GB
カメラ	製品名 画素数 画角	Logicoool webcam C300 1280 × 1024 60deg
マイクロホン	製品名 指向性 周波数帯域	パナソニック RP-VC151 モノラルマイクロホン 無指向性 100～20000Hz

4.2 位置計測

4.2.1 実験

ステレオカメラによる位置計測を用いて音源の位置を推定する実験を行った。図 4.4 は音源をユーザが指定するときの様子である。ユーザは左の画像を見て音源を矢印（カーソル）で指定し、指定した点の対応点が右の画像中に白点として表示されている。

図 4.5 の点 A～L の位置に音源を置き、それぞれの点について距離計測を行った。結果を表 4.2 にまとめる。

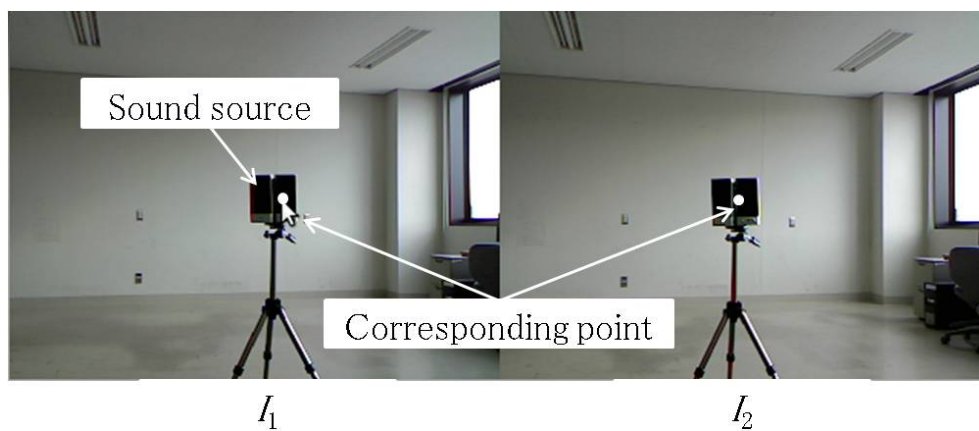


図 4.4: 対応点探索の様子

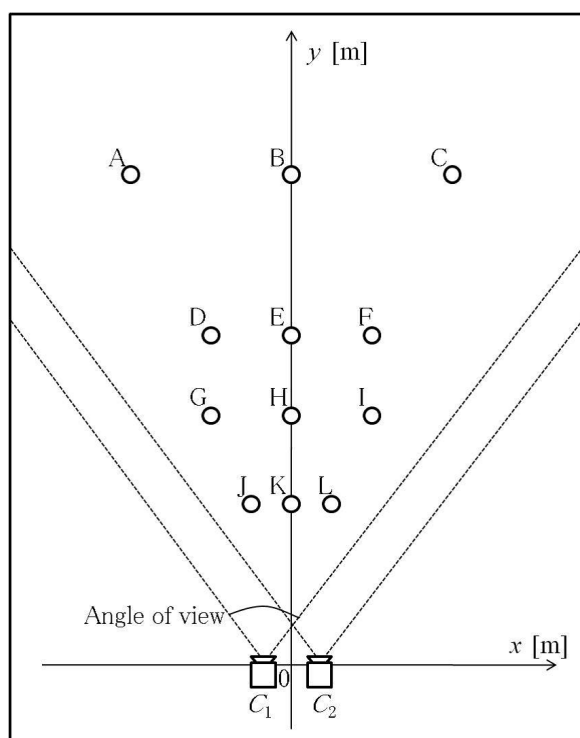


図 4.5: 音源の位置

表 4.2: 位置計測実験結果

	実際の座標 (x, y) [m]	実際の距離 (R'_1, R'_2) [m]	推定距離 (R'_1, R'_2) [m]
A	(-1.00, 3.00)	(3.13, 3.20)	(4.06, 4.13)
B	(0.00, 3.00)	(3.00, 3.00)	(3.95, 3.95)
C	(1.00, 3.00)	(3.20, 3.13)	(3.79, 3.73)
D	(-0.50, 2.00)	(2.04, 2.09)	(2.29, 2.34)
E	(0.00, 2.00)	(2.00, 2.00)	(2.42, 2.42)
F	(0.50, 2.00)	(2.09, 2.04)	(2.33, 2.28)
G	(-0.50, 1.50)	(1.55, 1.62)	(1.74, 1.80)
H	(0.00, 1.50)	(1.50, 1.50)	(1.65, 1.65)
I	(0.50, 1.50)	(1.62, 1.55)	(1.66, 1.59)
J	(-0.25, 1.00)	(1.01, 1.06)	(1.01, 1.06)
K	(0.00, 1.00)	(1.00, 1.00)	(1.00, 1.00)
L	(0.25, 1.00)	(1.06, 1.01)	(1.05, 1.00)

実際の距離と推定した距離との誤差について、 R'_1 、 R'_2 の内、代表して R'_1 について考える。図 4.6 に、実際の R'_1 と推定される R'_1 の誤差と、実際の R'_1 の関係を示す。

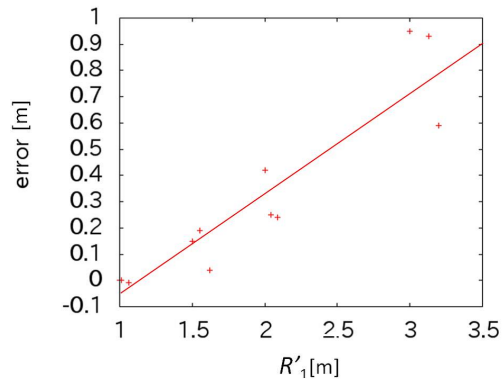


図 4.6: 誤差の距離依存性

R'_1 が大きくなるにつれて、実際の距離と推定距離の誤差が大きくなることが分かる。カメラの設置、キャリブレーションを手動で行ったため、位置推定の際、式 (3.20) 中の f 、 L 、 d の内、 f 、 L を実際より大きく設定してしまっている、あるいは二つのカメラの光軸が厳密に平行になっていないためだと考えられる。より正確に位置計測を行うためには、カメラの内外部のパラメータを正確に設定する必要がある。また、視差 d はサブピクセルの表現はできないことも誤差が生じる原因であるといえる。

4.3 音声抽出

作成したシステムで、音声抽出が行えるかどうか実験を行った。行った実験の環境を表 4.3 に示す。

表 4.3: 実験環境

サンプリング周波数	44100Hz
量子化 bit 数	16bit
短時間フーリエ変換における 1 フレームのデータ数	1024
拡散性音	空調音、PC のファンノイズ

以下に実験の内容を示す。まず、音源となる音声信号 $s[n]$ を用意する。 M_1 の入力信号 $x_1[n]$ として $s[n]$ を、 M_2 の入力信号 $x_2[n]$ として、ある遅延量 n_s を持った $s[n - n_s]$ を代入することで、遅延量 n_s を生成する位置に音源がある状態を仮想的に作り出す。仮想的な音源の位置を変えながら、それぞれの位置で音声抽出ができているかどうか実験した。

図 4.7 に音源の配置の一例を示す。マイクロホンアレイの中心を原点とし、正面の方向に y 軸、右方向に x 軸をとる。単位は m である。音源は $(x, y) = (1.0, 3.0)$ の位置にあるとする。このとき、 $R'_1 = 3.20\text{m}$ 、 $R'_2 = 3.13\text{m}$ となるので音速を $c = 344.6\text{m/sec}$ として

$$\tau_s = \frac{R'_1 - R'_2}{c} = 0.000183\text{s} \quad (4.1)$$

$$n_s = 44100\tau_s = 8.07 \quad 8 \quad (4.2)$$

となる。

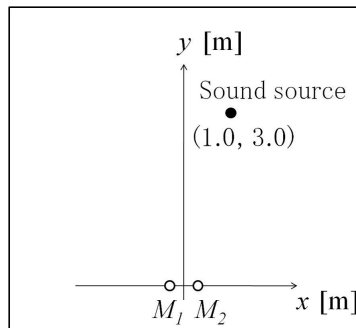


図 4.7: 音源が $(1.0, 3.0)$ の位置にある場合

4.3.1 仮想空間での発話

図 4.8 のように $(-1.0, 3.0)$ 、 $(0.0, 3.0)$ 、 $(1.0, 3.0)$ の位置にそれぞれ指向性音を発生する音源 A、B、C を仮想的に設置し、それぞれが単独で音声を発生したとき、提案したシステムで音声抽出が可能かどうか実験した。音源 A、B、C から発生する音声は、いずれも同じ成人男性が「こんにちは」と発音したものである。

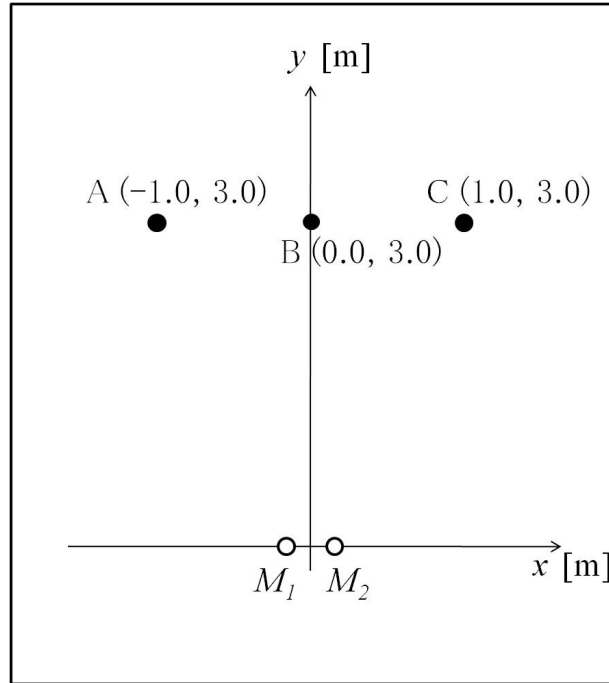


図 4.8: 音源 A、B、C の配置

M_i から点 A、B、C までの距離 R'_i 、生じる遅延時間 τ_s 、遅延データ数 n_s を表 4.4 にまとめる。

表 4.4: 各パラメータまとめ

音源	$R'_1[\text{m}]$	$R'_2[\text{m}]$	$\tau_s[\text{s}]$	n_s
A	3.13	3.20	-0.000183	-8
B	3.00	3.00	0	0
C	3.20	3.13	0.000183	8

実験結果を図 4.9、図 4.10、図 4.11 に示す。入力信号 (赤) の上に出力信号 (緑) を重ねて表示している。縦軸は信号の振幅、横軸は時間となっている。

図 4.9 は、音源 A が単独で音声を発生したときに、上からそれぞれユーザが音源 A、B、C の位置を指定した場合の結果である。

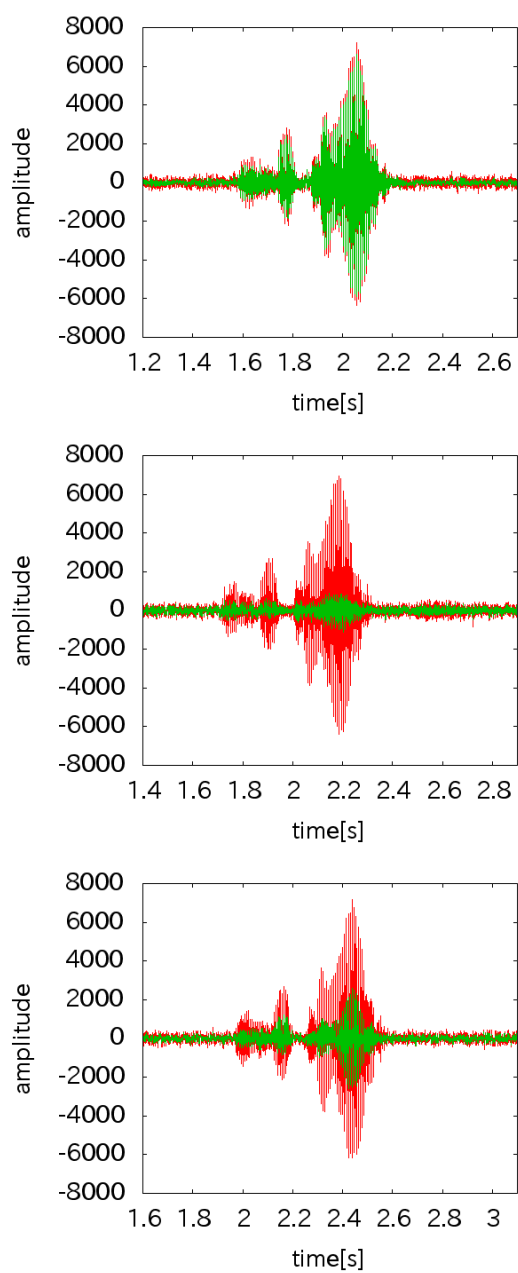


図 4.9: 仮想環境で音源 A が音声を発生したとき

図 4.10 は、音源 B が単独で音声を発生したときに、上からそれぞれユーザが音源 A、B、C のを位置指定した場合の結果である。

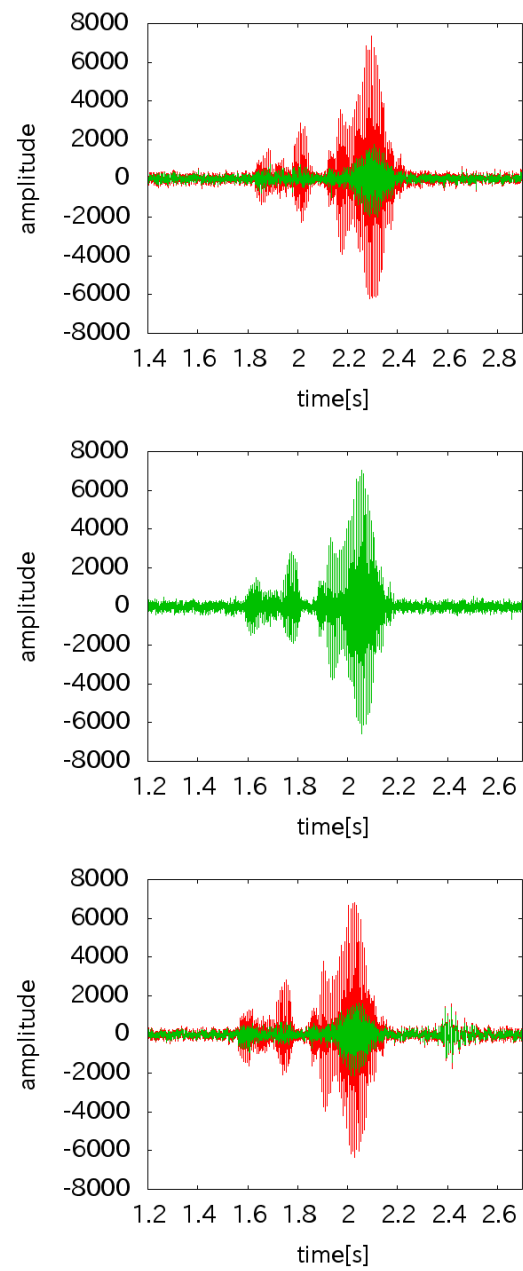


図 4.10: 仮想環境で音源 B が音声を発生したとき

図 4.11 は、音源 C が単独で音声が発生したときに、上からそれぞれユーザが音源 A、B、C の位置を指定した場合の結果である。

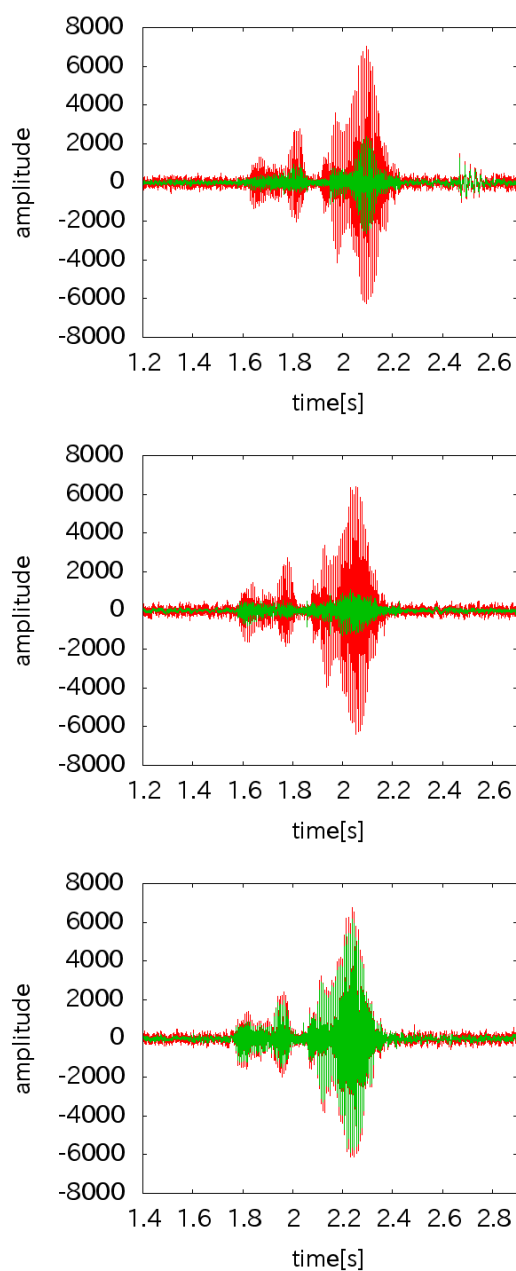


図 4.11: 仮想環境で音源 C が音声が発生したとき

図 4.9、図 4.10、図 4.11 を見ると、ユーザが指定した位置の音源が発する音声は、ほぼ等倍で出力されて、その他の音声は入力信号の 1/2 以下に抑圧されて出力されているのが分かる。作成したシステムで効果的に音声抽出ができるといえる。

4.3.2 実際の空間での発話

実験 1 と同様の実験を実空間で行った。実験時の様子を図 4.12 に示す。

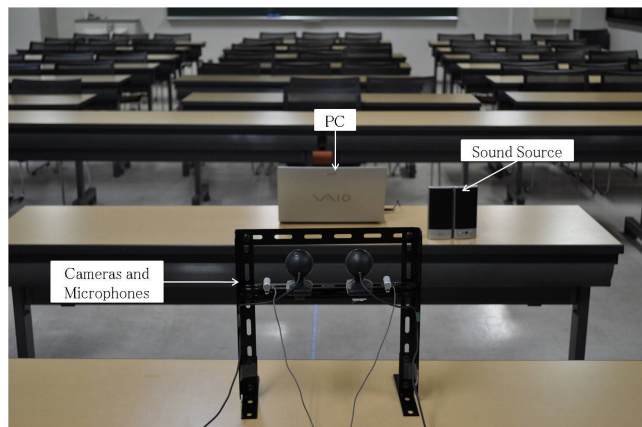


図 4.12: 実験時の様子

音源としてスピーカを用いた。PC からあらかじめ録音しておいた音声をスピーカに出力する。音源の配置を図 4.13 に示す。

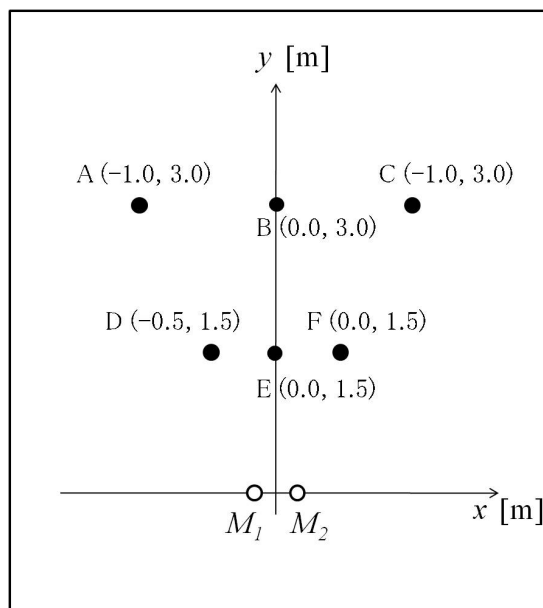


図 4.13: 音源の配置

仮想空間での実験と比較するため、音源 A、B、C の音声抽出の結果を図 4.14、図 4.15、図 4.16 に示す。赤の信号が入力信号、緑の信号が出力信号である。縦軸は信号の振幅、横軸は時間となっている。

図 4.14 は、音源 A が単独で音声を発生したときに、上から順にそれぞれユーザが音源 A、B、C の位置を指定した場合の結果である。

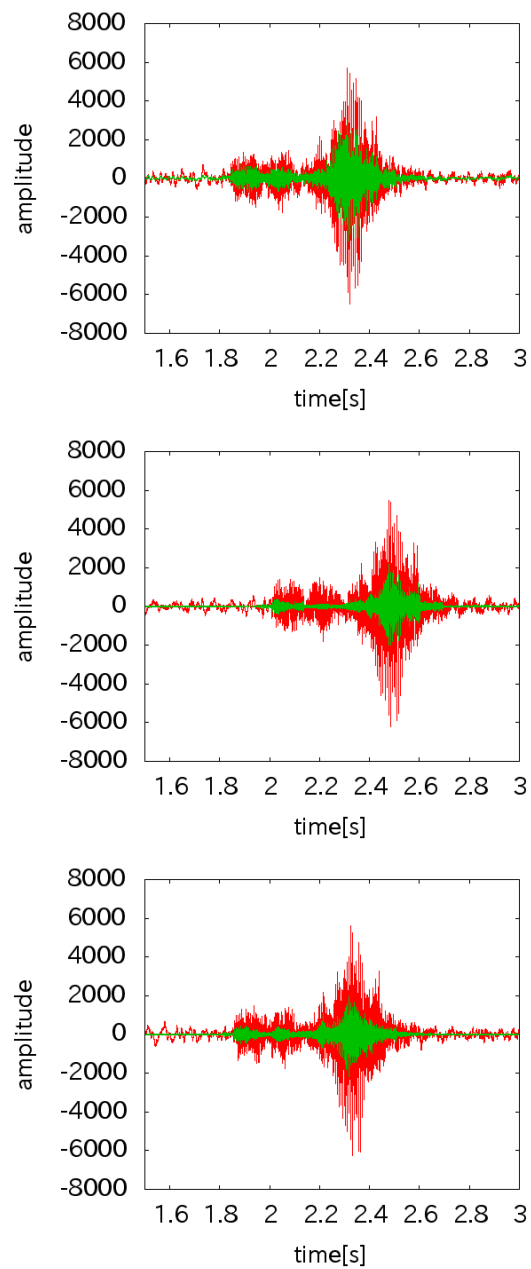


図 4.14: 実環境で音源 A が音声を発生したとき

図 4.14 を見ると、上の画像では出力信号は入力信号の 1/2 程度に、中央、下の画像は出力信号は入力信号の 1/3 以下に抑圧されていることが分かる。

図 4.15 は、音源 B が単独で音声を発生したときに、上から順にそれぞれユーザが音源 A、B、C の位置を指定した場合の結果である。

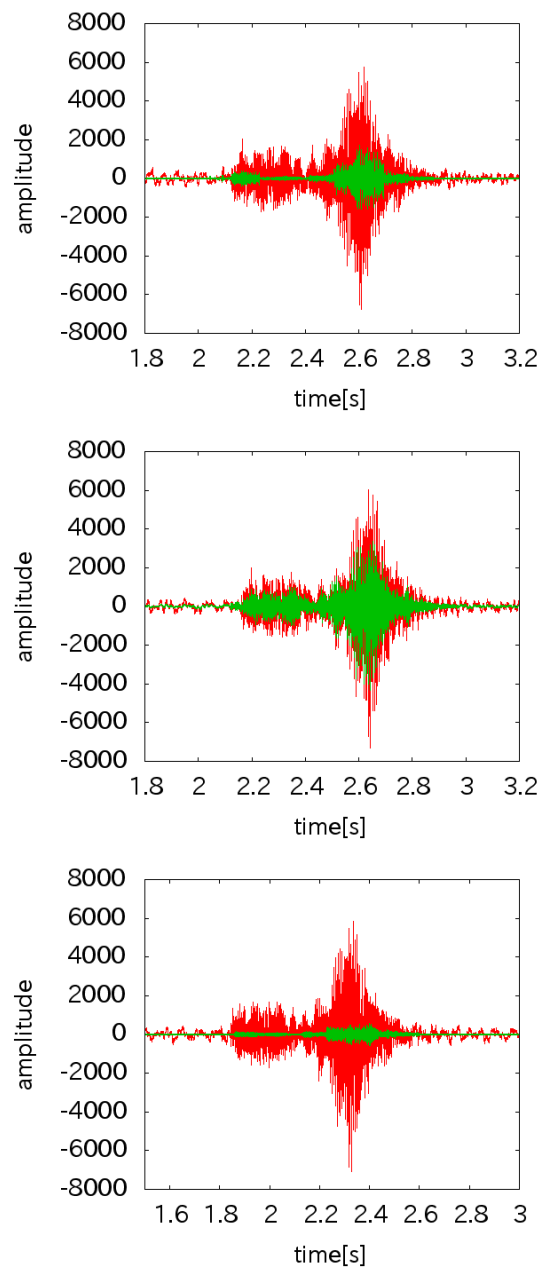


図 4.15: 実環境で音源 B が音声を発生したとき

図 4.15 を見ると、中央の画像では出力信号は入力信号の $2/3$ 程度に、上と下の画像では出力が入力の $1/3$ 以下にまで抑圧されていることが分かる。

図 4.16 は、音源 C が単独で音声を発生したときに、上から順にそれぞれユーザが音源 A、B、C の位置を指定した場合の結果である。

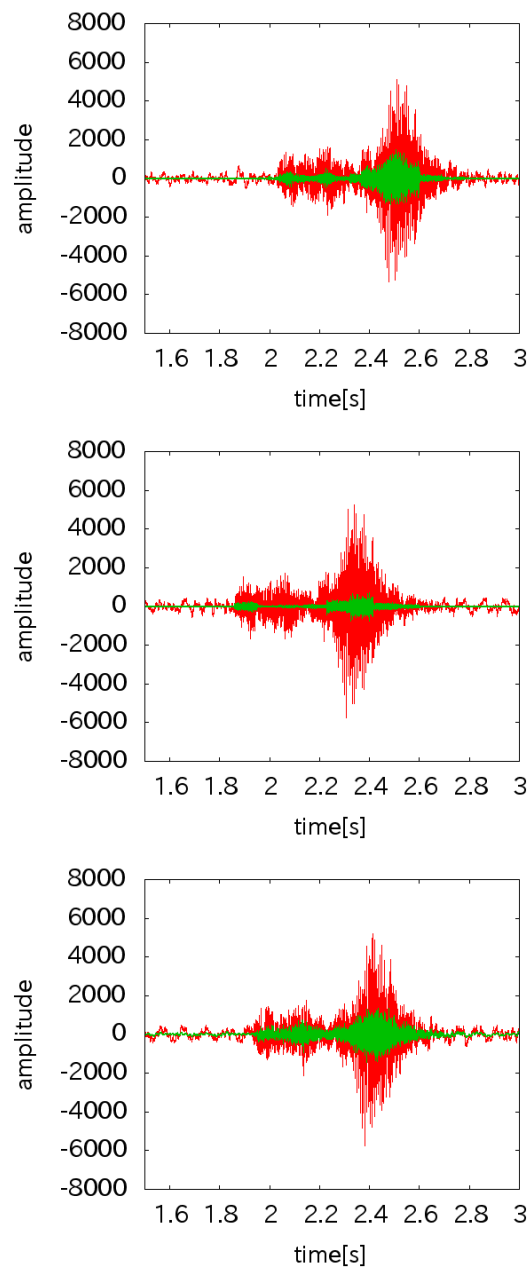


図 4.16: 実環境で音源 C が音声を発生したとき

図 4.16 を見ると、中央の画像では出力信号は $1/4$ 以下に、上と下の画像では出力信号は入力信号の $1/4$ 程度に抑圧されていることがわかる。

図 4.14、図 4.15、図 4.16 のいずれの場合も、抑圧すべきときに抑圧できていなかったり、抽出すべきときに抑圧していたりと、仮想環境での実験に比べて効果的には音声抽出できていない。仮想環境に比べて、実際の環境では音声抽出の性能が悪くなる原因について考える。仮想環境と実際の環境の違いについて表 4.5 にまとめる。

表 4.5: 仮想環境と実際の環境の違い

項目	仮想環境	実際の環境
反響	なし	あり
マイクロホン間での感度の違い	なし	あり

仮想環境と違い、実際環境には部屋の壁等による音の反響が存在する。図 4.17 に反響の様子を示す。

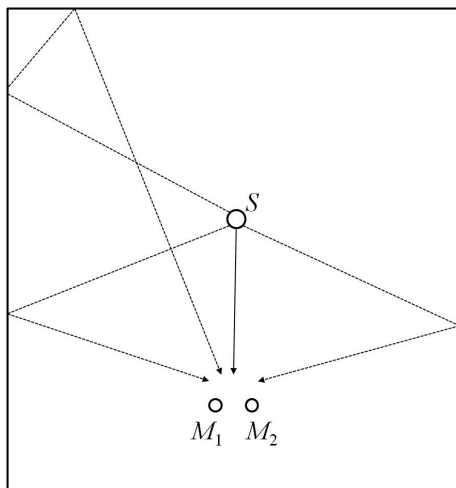


図 4.17: 反響の様子

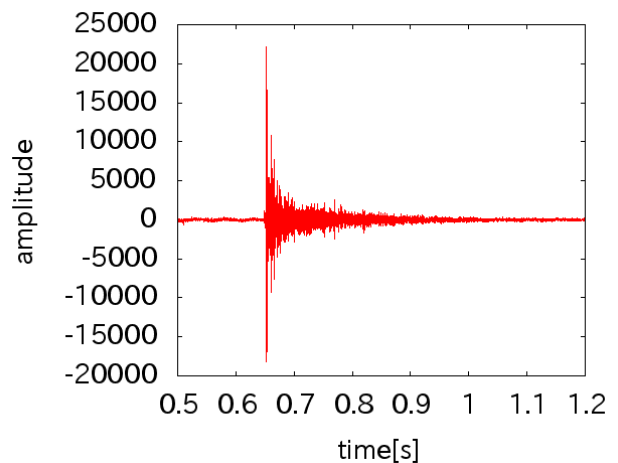


図 4.18: インパルス応答

図 4.17 において、 S は音源を表している。 S の発する音声を抽出しようとする場合について考える。反響により、 S から発せられた音声は、破線のような経路を通して様々な方向からマイクロホンアレイに到来するため、式 (3.27) における $Y[k, m]$ は目的信号の成分も含まれていることになる。すなわち、式 (3.28) によって、入力信号から目的信号の成分まで差し引いていることになる。この場合、実際の環境での実験における図 4.16 下のように、抑圧されるべきでないときも信号が抑圧されると考えられる。実験を行った部屋でのインパルス応答を図 4.18 に示す。図 4.18 におい

て、時間が経つにつれて信号が徐々に減衰していくのは反響が存在することが原因の一つであり、図 4.18 は部屋に反響があることを示している。

また、仮想環境と実際の環境ではマイクロホン間の感度の違いが存在する。図 4.19 は実際の環境での実験において、音源 B の音声を抽出するときの入力信号 $x_1(t)$ (赤)、 $x_2(t)$ (緑) である。

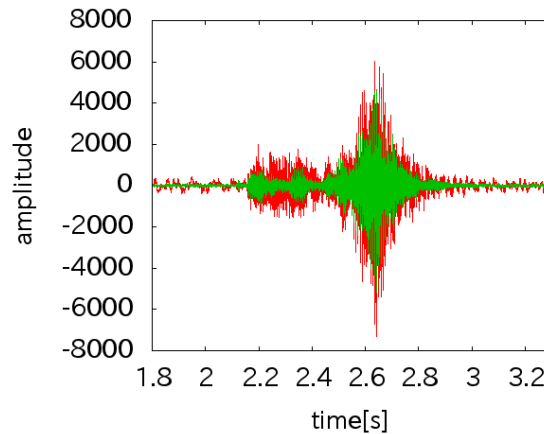


図 4.19: 入力信号の違い

本来、この場合の $x_1(t)$ と $x_2(t)$ は振幅が同じであるはずだが、図 4.19 から、 $x_2(t)$ の方が $x_1(t)$ に比べて大きいことが分かる。これはマイクロホンの製造過程等で各マイクロホン固有の感度特性が形成されるためと考えられる。ゲイン補正を行いマイクロホンの感度が同じになるように調整できればよいが、調整は困難である。入力信号の大きさが異なると式 (3.27) から正確な $Y[k, m]$ が得られないので、効果的な音声抽出は行えない。したがって、反響やマイクロホンの感度の違いが音声抽出の性能を劣化させているのではないかと考える。

第5章 おわりに

音声認識の分野などで音声抽出の技術に期待が寄せられている。これまでに提案されてきた方法では、音源の位置が不明であったり、音源が移動するなどすると、音声抽出の性能が劣化する方法が多い。本研究では、ステレオカメラとマイクロホンアレイを用いて、目的音源の位置に応じた音声抽出を行うことを目的とした。ステレオカメラを用いて音源の位置を計測し、計測した位置をもとにマイクロホンアレイを用いて差分型ビームフォーマとスペクトルサブトラクションによって音源の位置に応じた音声抽出システムを構築した。壁による反響がない、マイクロホン間で感度の違いがないなどの理想的な環境と、実際の環境で実験を行い、提案したシステムで音声抽出が可能であることを確認した。実際の環境では、理想的な環境に比べて音声抽出の性能が劣化するものの、高精度に感度調整が可能なマイクロホンを使用するなどのハードウェアの改良や、反響に対応した処理などを行えば、実際の環境でも十分な性能を発揮できるだろう。

謝 辞

本研究は大阪大学基礎工学部で行ったものです。

本研究を実施するにあたり、多くの方々にご指導、ご協力いただきました。大阪大学大学院基礎工学研究科 大城 理 教授には、研究テーマを決定するにあたり、多数の御指導、御助言を頂きました。進捗報告会での鋭い御指摘、御指導から、工学者としての研究に対する姿勢を学ぶことができました。深く感謝するとともに、篤く御礼申し上げます。本研究を完成させるにあたり、副査として御指導頂きました大阪大学大学院基礎工学研究科 野村 泰伸 教授に深く感謝致します。本研究を進めるにあたり、大阪大学大学院基礎工学研究科 井村 誠孝 准教授には、日頃から研究テーマについての相談に乗って頂き、また、研究の手法、論文の書き方に至るまで多大な御指導を頂きました。心から御礼申し上げます。大阪大学大学院基礎工学研究科 黒田 嘉宏 助教には、研究テーマについて非常に多数の提案、アドバイスをはじめ、多数の御指導を頂きました。心から感謝致します。山梨大学大学院医学工学総合研究部 鍵山 善之 助教には様々な面で御助言を頂きました。篤く御礼申し上げます。

本研究を進めるにあたり、日頃から研究テーマについて親身になって相談に乗って頂き、些細な疑問にも懇切丁寧に御指導して頂きました大城研究室の吉元 俊輔 氏、奥村 陽介 氏、重枝 慧 氏、末竹 哲也 氏、田中 翔太 氏、徳井 隆博 氏、濱田 友貴 氏、加藤 雄樹 氏、繁田 悠 氏、増田 拓 氏、森口 裕樹 氏に深く感謝致します。最後に、唯一の同期として日頃から研究生活での苦楽を共にし、お互いに切磋琢磨してきた同研究室の八木 明日華 氏に感謝致します。

参考文献

- [1] 内川恵二. 聴覚・触覚・前庭感覚. 朝倉書店, 東京, 2008.
- [2] 斎藤秀昭, 森晃徳. 視覚認知と聴覚認知. オーム社, 東京, 1999.
- [3] 川人光男, 行場次朗, 藤田一郎, 乾俊郎, 力丸裕. 岩波講座 認知科学 3 視覚と聴覚. 岩波書店, 東京, 1994.
- [4] 佐藤純生, 福江一智, 木下雄一郎, 小澤賢司. 鉛直方向の腹話術効果に及ぼす音響再生系の影響. 映像情報メディア学会誌, Vol. 63, No. 1, pp. 110–113, 2009.
- [5] 矢頭隆, 森戸誠, 山田圭, 小川哲司. 正方形マイクロホンアレイによる音源分離技術. 情報処理, Vol. 51, No. 11, pp. 1410–1416, 2010.
- [6] 斎藤収三, 中田和男. 音声情報処理の基礎. オーム社, 東京, 1981.
- [7] 中田和男. 音声. コロナ社, 東京, 1977.
- [8] 荒木昭二, 向井良, 荒木章子, 片桐滋. 混じりあった声を解く -遠隔発話の認識を目指して-. 東芝レビュー, Vol. 59, No. 12, pp. 42–44, 2004.
- [9] 澤田宏, 荒木章子, 牧野昭二. 音源分離技術の最新動向. 電子情報通信学会誌, Vol. 91, No. 4, pp. 292–296, 2008.
- [10] T.Amada and H.Kanazawa. Microphone array technique for speech recognition. NTT R D, Vol. 59, No. 9, pp. 937–944, 2004.
- [11] 井口征士, 佐藤宏介. 三次元画像計測. 昭晃堂, 東京, 1990.
- [12] 日本機械学会. 画像処理とパターン認識. 朝倉書店, 東京, 1986.

- [13] 岸野文郎, 佐藤隆夫, 横矢直和, 相沢清晴, 有川正俊. 画像と空間の情報処理. 岩波書店, 東京, 2000.
- [14] 辻井重雄. デジタル信号処理の基礎. 電子情報通信学会, 東京, 1988.
- [15] L.J. Griffiths and C.W. Jim. An alternative approach to linearly constrained adaptive beamforming. IEEE Trans. Antennas Propag, Vol. 30, No. 1, pp. 27–34, 1982.
- [16] 飯國洋二. 基礎から学ぶ信号処理. 培風館, 東京, 2004.
- [17] 村田昇. 入門 独立成分分析. 東京電気大学出版局, 東京, 2004.
- [18] 水町光徳, 赤木正人. マイクロホン対を用いたスペクトルサブトラクションによる雑音除去法. 電子情報通信学会論文誌, Vol. J82-A, No. 4, pp. 292–296, 1999.
- [19] 徐剛, 辻三郎. 3次元ビジョン. 共立出版株式会社, 東京, 1998.
- [20] David A. Forsyth, Jean Ponce, 大北剛. コンピュータビジョン. 共立出版株式会社, 東京, 2007.