

マイクアレイネットワークを用いたホームネットワークサービス向け ハンズフリー音声インタフェース

祖田 心平[†] 中村 匡秀[†] 松本 真佑[†] 松原 典行[†]
久賀田耕史[†] 和泉慎太郎[†] 川口 博[†] 吉本 雅彦[†]

[†] 神戸大学 〒657-8531 兵庫県神戸市灘区六甲台町 1-1

E-mail: [†]{soda,matsubara}@ws.cs.kobe-u.ac.jp, ^{††}{masa-n,shinsuke,shin,yosimoto}@cs.kobe-u.ac.jp,
^{†††}kugata@cs28.cs.kobe-u.ac.jp, ^{††††}kawapy@godzilla.kobe-u.ac.jp

あらまし 音声インターフェースは、ホームネットワークシステム (HNS) の操作インターフェースとして有望な技術である。しかしながら従来型のインターフェースは、ユーザが実際のマイクを使用・装着して操作しなければならず、ユーザの負担になっていた。そこで本稿では、マイクアレイネットワークを用いることで、HNS におけるハンズフリーな音声インターフェースを実現する。マイクアレイネットワークとは、複数のマイクアレイをネットワークで接続し、話者の位置推定や音声品質の向上を可能にする技術である。マイクアレイを壁や天井に設置することで、ユーザは HNS に対して、部屋のどこからでもマイクを意識することなく音声を入力できる。本研究では、16ch マイクアレイを用いたプロトタイプを作成し、実環境での音声認識率、音源定位精度、発話推定精度を評価する。応用サービスとして、ハンズフリー家電操作サービス、発話自動ログサービスに提案システムを実装し、提案システムの適用可能性を評価する。

キーワード マイクアレイネットワーク、ホームネットワークサービス、音声インタフェース、ハンズフリー

Handsfree Voice Interface for Home Network Service Using a Microphone Array Network

Shimpei SODA[†], Masahide NAKAMURA[†], Shinsuke MATSUMOTO[†],

Noriyuki MATSUBARA[†], Koji KUGATA[†], Shintaro IZUMI[†],

Hiroshi KAWAGUCHI[†], and Masahiko YOSHIMOTO[†]

[†] Kobe University Rokkoudai-cho 1-1, Nada-ku, Kobe, Hyogo, 657-8531 Japan

E-mail: [†]{soda,matsubara}@ws.cs.kobe-u.ac.jp, ^{††}{masa-n,shinsuke,shin,yosimoto}@cs.kobe-u.ac.jp,
^{†††}kugata@cs28.cs.kobe-u.ac.jp, ^{††††}kawapy@godzilla.kobe-u.ac.jp

Abstract The voice control is a promising user interface for the home network system (HNS). In our previous interface, a user had to be equipped with an actual microphone device, which imposed a burden on the user. This paper presents a hands-free voice interface using a microphone array network. The microphone array network enables voice quality enhancement, as well as sound source localization, by networking multiple microphone arrays. Attaching the arrays to the walls or ceiling, users can input voice operations to the HNS from anywhere in the room, without being aware of the microphone devices. We implement a prototype system with a 16ch microphone array, and evaluate the accuracy of speech recognition, sound source localization, and voice activity detection in a real home network environment. The hands-free operation service and automatic speech logging service are implemented as practical services.

Key words microphone array network, home network services, voice interface, hands free

1. はじめに

我々の研究グループでは、宅内の家電機器やセンサをネットワークに接続し、様々な付加価値サービスを実現するホームネットワークシステム（以下 HNS）の研究・開発を行っている [1]。HNS では、家ごとに異なる多様な環境において、様々な機器やサービスが提供される。よってユーザーにとって学習コストが低く操作しやすいユーザインタフェースが求められている。音声によって機器やサービスを操作する音声インターフェースは、HNS の操作インターフェースとして有望な技術のひとつである。多種多様な家電やサービスを「発話する」という統一的な方法で操作できるため、従来のリモコンや操作パネル等比べて学習コストが低く手軽である。我々のグループでも実際の HNS 上に混合主導型音声インターフェース [2] を構築している。

しかしながら、現状の音声インターフェースはマイクの利用を想定したものがほとんどである。マイクを手にとって発話したり、耳にかけて装着したりというように、常にマイクを意識して操作しなければならない。このことは HNS 向け用途として、少なからずユーザーの負担となる。

そこで本稿では、マイクアレイネットワーク [3] を用いることで、ユーザーにマイクの位置を意識させない、ハンズフリーな音声インタフェース実現することを目指す。マイクアレイネットワークとは、複数のマイクを敷き詰めたマイクアレイをネットワークで接続し、協調動作させるシステムである。各マイクに到達する音の到達時間差を用いて、話者の位置推定や音声品質の向上、複数音源の分離などが可能となる [4] [5]。マイクアレイを壁や天井に設置することで、ユーザーは HNS に対して、部屋のどこからでもマイクを意識せずに音声を入力できる。

本稿では、16ch マイクアレイを用いたプロトタイプを開発し、実環境での音声認識率、音源定位精度、発話推定精度を評価する。また、応用サービスとして、ハンズフリー家電操作サービス、および、発話自動ログサービスを実装し、開発したマイクアレイシステムの適用可能性を評価する。

2. 準備

2.1 マイクアレイネットワーク

マイクアレイは、複数のマイクを格子状に配置したデバイスである。複数のマイクロホンの音到達時間差を利用することで、音の到来方向の推定や、指向性の制御を行うことができる。また方向推定のみならず、反射や残響の影響を抑制し、特定の音声を分離・抽出することもできる。これにより信号対雑音比（SNR）の改善が可能である。これらの効果は、使用するマイク数が多くなるほど向上することが知られている [6]。しかし、取得した音声信号をサーバへ送る際の伝送量や、マイクアレイ全体の電力消費増大の理由から、実用化されているマイクアレイのマイク数は数個から十数個に留まっている。

マイクアレイネットワークは、複数のマイクアレイをネットワークを通じて協調動作させるシステムである。図 1 に我々が開発しているマイクアレイネットワークの構成図を示す。この図では、16ch の小規模なマイクアレイ（サブアレイと呼ぶ）を

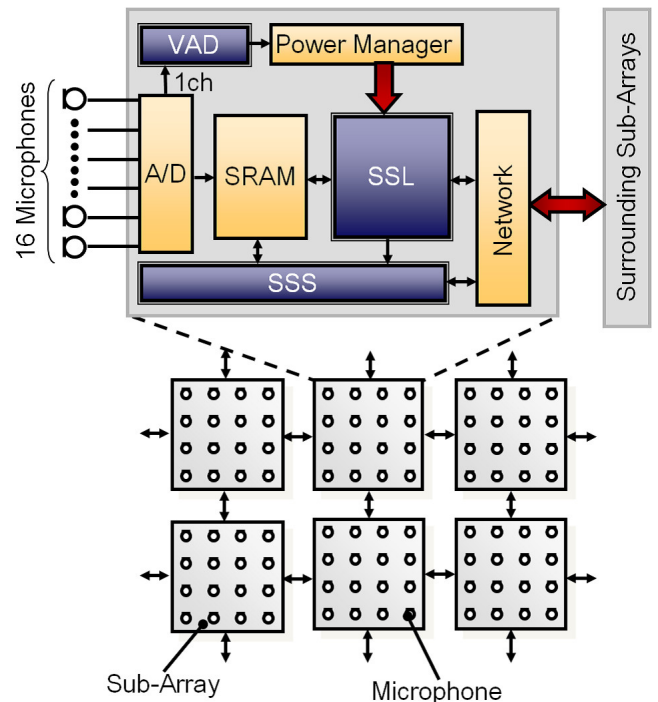


図 1 マイクアレイネットワーク

多数配置し、サブアレイ間では有線ネットワークでつなぎ、全体として大規模なマイクアレイシステムを実現している。各サブアレイが行う処理は主に以下の3つである。

- 発話推定：発話の有無を検出する。
- 音源定位：音源の位置を推定する。
- 音源分離：特定位置の音を高品質で抽出する。

これらの処理の結果として、各サブアレイから SNR の高い音声データが得られる。これらを複数ネットワーク上で集約することで、さらに高品質な音声データが得られる。

従来のマイクアレイ研究の多くは、実機を用いた動作検証 [4] や伝送量削減のための改善 [5] といった基礎研究であり、実用的なサービスへの適用や応用に関する研究は比較的少ない。

2.2 ホームネットワークシステム

ホームネットワークシステムは、照明やテレビなど家庭内における様々な家電機器と、温度計や湿度計などセンサをネットワークに接続することで構築される。HNS 内における機器はユーザーや外部エージェントがネットワーク越しに制御できるように、制御 API を備えている。HNS は次世代のスマートホーム技術として期待されており、音声による機器操作インタフェースや、外部環境やユーザーの状況といったコンテキストに応じた機器の自律制御などの付加価値サービスを提供可能である。我々の研究室で開発している CS27-HNS [1] は、サービス指向アーキテクチャ(SOA) [7] の考えを取り入れ、すべての制御 API を、機種や実行環境に依存しない標準的な Web サービスとして公開している。

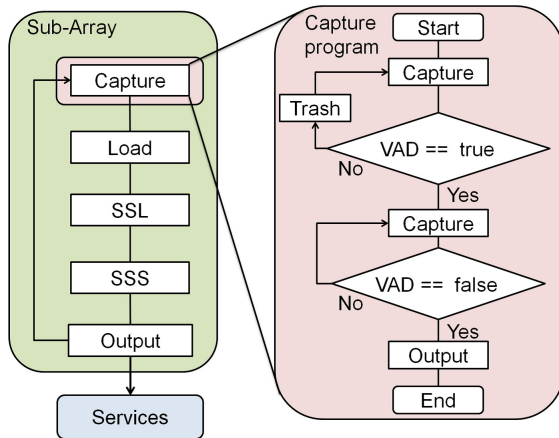


図2 プロトタイプシステムのフローチャート

3. マイクアレイネットワークを用いたホームネットワークサービスインタフェース

3.1 システム要求

まず HNS 向け音声インタフェースへの要求事項を述べる．

R1: ユーザに負担をかけないこと 従来の音声インタフェースでは，単一マイクを口元に近付けて集音する手法が一般的だが，日常生活の中では常にマイクに向かって話す煩わしさが問題となる．HNS 向けの音声インタフェースとしては，ユーザがマイク等のデバイスを身につけなくても利用できるシステムが要求される．

R2: 雑音環境下でも使用できること 取得した音声を生認識等に使用する場合，高精度な認識率を保つには残響やノイズのないクリアな音声が必要．HNS が実装されている環境下ではテレビの音やエアコンの動作音等の雑音が予想される．提案システムには，これらの雑音を抑えてユーザの音声を高品質に抽出する必要がある．

R3: 部屋のどこからでも音声を入力できること 日常生活において HNS の機器やサービス进行操作しようとする場所は，ドアの前やソファに座っているとき，キッチンなど様々な場所が考えられる．また，場所に応じた (location-aware) サービスを考える場合には，部屋の中で会話の起こりうる場所をなるべく広くカバーする必要がある．

3.2 プロトタイプシステムの概要

前節で述べた要求を満たすため，マイクアレイネットワークを用いた音声インターフェースを試作した．要求 R3 はサブアレイを増やすことで対処できるため，今回は，要求 R1, R2 を満たす 16ch の単一サブアレイを用いたシステムを作成した．

図2にサブアレイが行う処理のフローチャートを示す．発話推定 (Voice Activity Detection, VAD) 及びキャプチャプログラムは C++ で実装されており，音源定位 (Sound Source Localization, SSL)，音源分離 (Sound Source Separation, SSS) は Matlab で実装されている．まずキャプチャプログラム内で，サブアレイ内の 16 個のマイクで集音した音声信号から発話の有無を検出する．発話があれば ch 毎に録音，出力する．次に，録音されたデータは Matlab 上にロードされ，音源定位によって

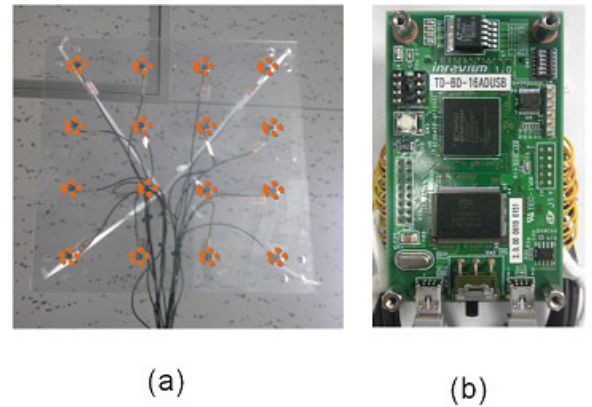


図3 (a) サブアレイ (b) キャプチャモジュール TD-BD-16USB

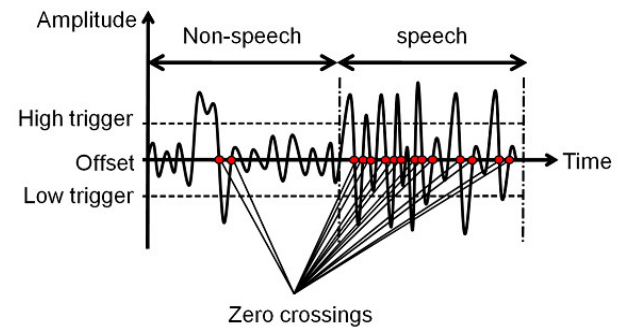


図4 ゼロ交差数を用いた発話推定

音源の位置を推定する．最後に，推定された位置に基づいて，音源分離を行い，集約された 1ch 分の高品質な音声データを出力する．出力された音声データは HNS の各サービスに利用される．

3.3 サブアレイとキャプチャモジュール

図3(a)にサブアレイのマイク配置を示す．サブアレイの寸法は 30cm 四方で，16 個のマイクを格子状に配置している．各マイクにより集音された音声は，図3(b)に示すキャプチャモジュール (TD-BD-16USB 東京エレクトロン社製) を通して PC に転送される．PC 上ではキャプチャプログラムが動作しており，発話推定によって録音の開始/停止を制御する．

3.4 発話推定 (VAD)

発話推定のアルゴリズムとして，本研究では演算量の少ないゼロ交差数を用いる手法 [8] を使用する．直感的には，音声信号の波形が単位時間あたりにどれだけ 0 と交差するかをカウントすることで，発話の有無を推定するものである (図4)．無音時の雑音と発話を区別するため，トリガ振幅を超えた直後のゼロ交差のみをカウントする工夫がなされている．この方法で，ある時間フレームごとにゼロ交差数を求め，得られたゼロ交差数と無音区間での平均ゼロ交差数を比較する．その差が指定した閾値を越えた場合，そのフレームを発話区間とみなす．

音声信号は 64 サンプルを 1 フレームとして，待機中でも一定のフレーム数を保持する．発話推定にはそのうちの 1ch を使用する．保持されているフレームが全て発話区間と判定された時に録音を開始する．また，録音中に保持しているフレームが全て無音区間になった時に録音を停止する．

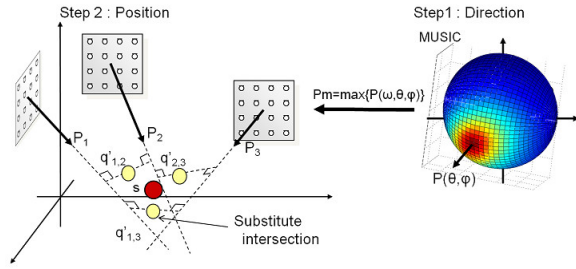


図 5 3次元音源定位

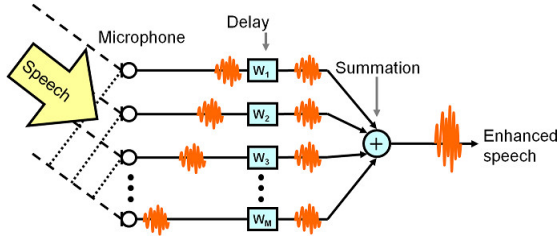


図 6 遅延和ビームフォーミングを用いた音源分離

3.5 音源定位 (SSL)

各マイクへの到達時間や音量の差を分析することで、その音声はどの方向から到来したかを推定することができる。これを複数のサブアレイを用いて行い、各アレイが推定した方向の交わる先を音源の位置と見積もる。図 5 に概念図を示す。各サブアレイにおける方向推定には、少ないマイク数でも高い精度が得られる MUSIC 法 [9] を用いる。この手法は、以下の式によりパワースペクトル $P(\omega, \theta, \phi)$ を計算し音源の方向 (θ, ϕ) を求める。

$$P(\omega, \theta, \phi) = \frac{|v(\omega, \theta, \phi)|^2}{|v^H(\omega, \theta, \phi)E_n|^2} \quad (1)$$

ここで、 $v(\omega, \theta, \phi)$ は仮想的な音源の位置ベクトル、 E_n は入力相関行列の雑音に由来する固有ベクトルである。 θ と ϕ を走査し、 $v(\omega, \theta, \phi)$ が実際の音源の位置ベクトルと一致した時、 $P(\omega, \theta, \phi)$ は極大値をとる。こうして推定した音源方向を集約し、音源の位置を計算する。図 5 に示すとおり、3次元では音源方向ベクトル P_m を基に引いた直線同士が必ずしも交点をもつとは限らない。そこで、直線間を最短で結ぶ線分上に仮想的な交点を取り、それらの重心を最終的な音源位置とする。

3.6 音源分離 (SSS)

音源定位で音源の位置が推定できると、その位置の音声に特に耳をすますことで高品質な音声を得られる。これを音源分離という。本研究では、遅延和ビームフォーミング [10] によって音源分離を行う (図 6)。この方法は、音源から時間差で到着した音声を、位相を揃えて足し合わせる。Delay の重みは推定された位置から計算される。重ね合わせの原理により、特定位置の音声のみを強調することができる。

4. ホームネットワークサービスの例

プロトタイプシステムを評価するにあたり、HNS において想定している応用サービスの例を 2 つ紹介する。

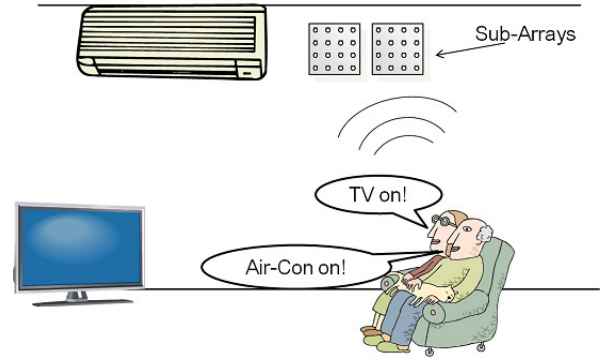


図 7 ハンズフリー家電操作サービス

Sound	Date	Location
(a)	2011/9/20/15:20:45	(320, 256, 55)
	2011/9/20/15:21:03	(238, 448, 129)
	2011/9/20/15:21:54	(171, 66, 60)
	2011/9/20/15:22:36	(194, 71, 95)
	2011/9/20/15:22:58	(101, 321, 57)
	2011/9/20/15:23:42	(250, 227, 133)
	2011/9/20/15:24:04	(97, 146, 87)

(a)

(b)

図 8 発話自動ログサービス

4.1 ハンズフリー家電操作サービス

ハンズフリー家電操作サービスは、従来マイクを用いて行っていた HNS 機器やサービスの音声制御 [2] を、提案システムを用いてハンズフリーで行うサービスである。図 7 にサービスイメージを示す。この例では、ユーザがソファに座ったまま、テレビやエアコンの電源をつけている。このように、ユーザは物理的なマイクを意識せずに、部屋の様々な場所から音声入力が行える。そのためユーザの負担が劇的に軽減される。また、マイクアレイネットワークでは、推定した音源位置に指向性を向け、雑音を分離して SNR の高い音声を集音できる。よって、精度の高い音声コマンドの入力が期待できる。また、リビングや寝室等比較的静かな環境下に限らず、キッチンなどの雑音環境下でも精度の高い音声入力が可能である。

4.2 発話自動ログサービス

発話自動ログサービスは、取得した音声と位置情報をログとして蓄積していき、いつ、どこで、何を話したかという発言記録を作成するサービスである。図 8 にサービスイメージを示す。

Web カメラ等と連携すれば、推定した音源方向に自動でカメラを向けて、発言者の音声と映像を繋げた議事録を作成することもできる。また、録音された議事録を見返すことで、参加者の発言回数や発言内容から意欲評価にも繋げられる。さらに、音声認識モジュールと連携すれば、会議中の発言内容を自動的に書き起こすディクテーションサービスや、インタビューのプロトコル分析を行うサービスにも応用可能である。

音声データを常に蓄積していくとデータ量が膨大になるという問題が生じるが、発話推定機能を用いて録音期間を制御することで、無駄なデータ取得を抑えることができる。

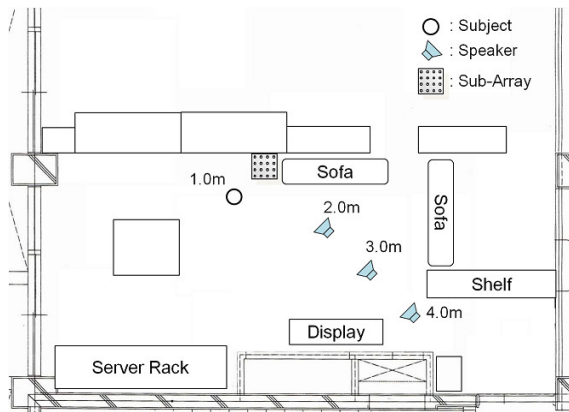


図 9 音声認識評価実験 被験者、スピーカの配置

表 1 音声認識評価実験 1 被験者毎の認識率

Subject No.	1	2	3	4	5
Age	twenties	twenties	twenties	twenties	forties
Sex	Male	Male	Male	Male	Male
Recognition ratio (%)	90	80	82	94	90

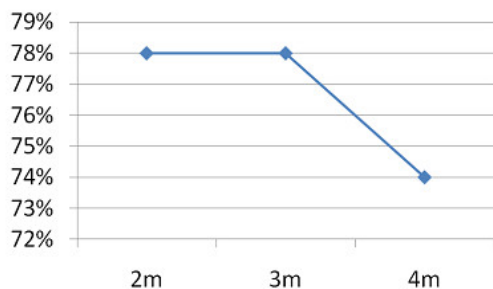


図 10 音声認識評価実験 2 距離による認識率の変化

5. 評価実験

提案システムで取得した音声データを用いて、音声認識率、音源定位精度、発話推定精度を評価する実験を行った。

5.1 音声認識率の評価

まず音声認識率の評価実験を説明する。4.1 のハンズフリー家電操作サービスでは、ユーザが発話する音声コマンドを高い精度で HNS に伝える必要がある。そこで、実際の HNS においてユーザに操作コマンドを発話してもらい、提案システムで録音、その認識率を評価した。図 9 に実験室のレイアウトを示す。

実験 1：5 名の被験者それぞれに、サブアレイから 1.0m の位置から HNS の音声コマンド 50 語を発話してもらう。

実験 2：サブアレイから 2.0m, 3.0m, 4.0m の位置にそれぞれスピーカを配置し (図 9)、実験 1 での被験者 1 名の発話を録音したものを再生して、距離による認識率の変化を測定する。

表 1 に実験 1 の結果を示す。1 行目から順に、被験者番号、年齢、性別、認識率を表している。被験者毎に滑舌等の違いはあるが、80% ~ 94% の高い認識率が確認できた。この結果から、提案システムはハンズフリー家電操作サービスに十分適用できると考える。被験者 2 と被験者 3 の認識率が低いのは、発話の 1 文字目が録音されていなかったり、最後の文字までは

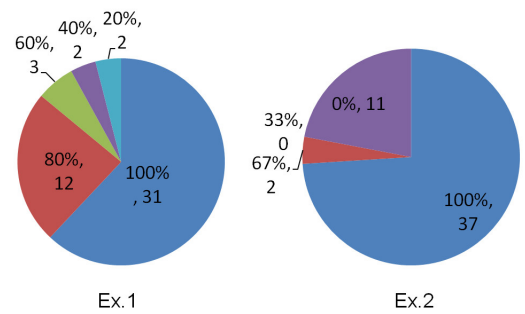


図 11 音声認識評価実験 単語ごとの認識率の割合

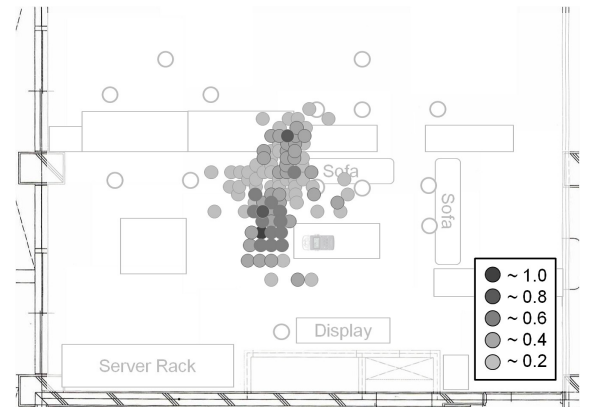


図 12 音源定位評価結果 発話箇所の分布

きり発話しなかったコマンドが含まれていたことが原因である。

図 10 に実験 2 の結果を示す。横軸がサブアレイとスピーカとの距離、縦軸が認識精度を表している。スピーカの再生ノイズ等の影響から、実験 1 と比較して全体的に認識率が下がっているが、4m 離れた位置でも 70% 以上の認識率が達成できた。

図 11 に実験 1、実験 2 における単語毎の認識率の割合を示す。認識率の右の数字はコマンド数を表している。実験 1 では、9 割のコマンドが認識率 80% 以上 (5 人中 4 人以上で認識された) だが、実験 2 では、距離にかかわらず認識率が 0% のコマンドが 11 個あった。認識率を上げるためには、ユーザのクリアな発話や、音源分離によるノイズ除去機能の強化が挙げられる。本研究では単一サブアレイによる実験だったが、今後サブアレイを増やすことでカバー範囲を拡大し、音源分離の性能向上を図る予定である。

5.2 音源定位の精度評価

4.2 で述べた音声ログサービスでは、発話のタイミングと音源の位置を高い精度で検出することが求められる。そこで、実際のミーティングを対象とした評価実験を行った。具体的には、我々の研究室で行われた定例ミーティングを、プロトタイプシステムと IC レコーダー (PCM-D50 sony 社製) で録音した。IC レコーダーで録音された音声、プロトタイプシステムの発話推定によって実際に録音されているかを調べる。また、発話した参加者の席の位置と音源定位した位置がおおむね一致しているかを調べた。ミーティングの参加者は准教授 1 名、助教 1 名、学生 15 名の計 17 名である。

実験室のレイアウトは図 9 に示したものと同じである。発話

表 2 発話推定評価実験 Clipping ratio 及び Sleeping ratio

Target interval [s]	120
Utterance interval [s]	61.32
Noise interval [s]	58.68
Recorded utterance interval [s]	17.06
Recorded noise interval [s]	4.65
Clipping ratio [%]	72
Sleeping ratio [%]	92

が特に多いのは教員と進捗報告を行う学生であり、教員はディスプレイの左側、進捗報告者はソファ周辺に位置している。ICレコーダーはディスプレイ前のテーブルの上に設置している。録音時間は約 16 分で、プロトタイプシステムは発話検知時に発話時刻と音源方向を音声データと共に出力する。ICレコーダーは、雑音時間も含めミーティングの全発話内容を録音する。

図 12 に、プロトタイプシステムによって推定されたミーティング中の発話位置の分布を示す。図中の丸印は、方向推定によって得られた発話の座標をプロットしており、発話があった回数が多い箇所ほど濃く表示される。丸印の濃さは、一番発話回数の多かった箇所を 1.0 として正規化し、5 段階で変化させている。ミーティングでは教員が頻繁に発話しており、ディスプレイの左側の方向へ発話箇所の分布が伸びているのがわかる。しかし、教員の位置がサブアレイのカバー領域外であり定位の精度が悪い。精度向上にはサブアレイを増やす必要がある。

5.3 発話推定の精度評価

発話推定の精度を評価するため以下の尺度を定義する。

- **Clipping ratio:** 実際に発話した時間に対して、提案システムが誤って雑音と認識して録音しなかった時間の割合。発話推定の失敗割合を意味するため、小さい値をとる方が良い。
- **Sleeping ratio:** 発話されていない時間に対して、提案システムが正しく発話がないと認識して録音しなかった時間の割合。雑音区間で期待通りに待機動作を行っている時間割合であるため、大きい値をとる方が良い。

5.2 の実験で録音したある 2 分間について、上記の 2 つの尺度を評価した。表 2 に結果を示す。計算した区間では Clipping ratio が 72 %、Sleeping ratio が 92 % となった。この結果から、実験においては雑音はほぼ正しく除去できたが、約 7 割の発話を取得できなかったということになる。この理由について、3.4 で述べた発話区間を識別する閾値の設定が厳しすぎたことが挙げられる。Clipping ratio と Sleeping ratio は互いにトレードオフの関係にあり、閾値を厳しくすると、Sleeping ratio が向上するが、Clipping ratio も同時に上がってしまう。発話自動ログサービスでは、発話の取りこぼしをより少なくしなければならないので、より発話を拾いやすく閾値を下げるのが好ましい。逆に、ハンズフリー家電操作サービスでは、ユーザの操作コマンド以外は反応してはならず、閾値を高く設定する方が良い。このように応用サービスに応じて、柔軟に閾値を設定可能にすることが今後の課題である。

6. おわりに

本研究では、HNS のためのハンズフリーな音声インターフェースを実現するために、マイクアレイネットワークを用い

たシステムを開発した。具体的には、16ch 単一サブアレイを用いたプロトタイプシステムを実装し、音声認識率、音源定位精度、発話推定精度の評価を行った。音声認識率の評価では、近距離では 80 % ~ 94 %、4m 離れた位置では 74 % の認識率を確認した。今後の課題は、サブアレイを複数台に増やし、音声入力のカバー範囲を拡大し、発話推定・音源定位の精度を向上したい。また、マイクアレイネットワークを用いた他の HNS サービスも考案し、実証実験を通してさらなる有効性の評価を行いたい。

7. 謝 辞

本研究の一部は、株式会社半導体理工学研究センター (STARC)、科学技術研究費 (基盤研究 B 23300009, 若手研究 B 21700077, 研究活動スタート支援 22800042)、および、ひょうご科学技術協会の助成を受けて行われている。

文 献

- [1] M.Nakamura, A.Tanaka, H.Igaki, H.Tamada, and K.Matsumoto, "Constructing home network systems and integrated services using legacy home appliances and web services," International Journal of Web Services Research, vol.5, no.1, pp.82-98, 2008.
- [2] 松原典行, 江上公一, 井垣 宏, "暗黙的なユーザ要求を抽出・推定するホームネットワークのための対話型音声インターフェース," 電子情報通信学会技術研究報告, 第 109 巻, pp.61-66, 2010.
- [3] T. Takagi, H. Noguchi, K. Kugata, M. Yoshimoto, and H. Kawaguchi, "Microphone array network for ubiquitous sound acquisition," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.1474-1477, 2010.
- [4] 祖田心平, 久賀田耕史, 高木智也, 和泉慎太郎, 野口紘希, 吉本雅彦, 川口 博, "分散処理を用いた超低消費電力ネットワーク型マイクロホンアレーの研究," 日本音響学会 2010 年秋季研究発表会, pp.585-588, 2010.
- [5] S. Izumi, H. Noguchi, T. Takagi, K. Kugata, S.S. and M. Yoshimoto, and H. Kawaguchi, "Data aggregation protocol for multiple sound sources acquisition with microphone array network," 20th International Conference on Computer Communications and Networks (ICCCN), pp.1-6, 2011.
- [6] C. Australia and J. Glass, "Loud: A 1020-node microphone array and acoustic," 2007.
- [7] M.P.Papazoglou and D.Georgakopoulos, "Service-oriented computing," Communication of the ACM, vol.46, no.10, pp.25-28, 2003.
- [8] M.M.S. J. Benesty and Y. Huang, Springer handbook of speech processing, Springer-Verlag, 2008.
- [9] R. Schmidt, "Multiple emitter location and signal parameter estimation," Antennas and Propagation, IEEE Transactions on, vol.34, pp.276-280, 1986.
- [10] K.B. Van Veen, "Beamforming: a versatile approach to spatial filtering," ASSP Magazine, IEEE, vol.5, pp.4-24, 1988.