マイクロホンアレーを用いた発話者方向検出によるハンズフリー音声認識

山田武志† 中村 哲† 鹿野清宏†

本論文では，実環境下でのハンズフリー音声認識を実現するための方法としてマイクロホンアレーの適用について検討する．マイクロホンアレーを音声認識に適用する際には発話者の方向に指向性ビームを安定して向けることが非常に重要となる．従来法では発話者方向を検出するために音源のパワー情報や各受音信号間の相関情報を利用しているが，低SNR環境下などでは発話者方向検出の精度が十分でなかった．本論文では，音声特有の調波構造という情報を積極的に活用し，このような状況での発話者方向検出精度を向上させることを試みる．提案法の有効性を評価するためにシミュレーションと簡易音響実験室で認識実験を行った．その結果，白色ガウス雑音と計算機雑音の両方に対して従来法より高い発話者方向検出精度が得られ，低SNR環境下での単語認識率を大幅に改善できた．

Hands-free Speech Recognition with Talker Localization by a Microphone Array

Takeshi Yamada,† Satoshi Nakamura† and Kiyohiro Shikano†

This paper shows a speech recognition system with talker localization by a microphone array to realize hands-free speech interface in real environments. In order to localize talker direction exactly in low SNR conditions, a talker localization algorithm based on extracting a pitch harmonics is used. To evaluate the performance of the proposed method, the speech recognition experiments are carried out both in simulation and in an experimental room. These results show that the proposed method attains the higher talker localization accuracy and word recognition accuracy than a conventional method.

1. はじめに

近年，隠れマルコフモデルやニューラルネットワークなどの統計的手法により，音声認識性能は著しく改善された．しかしながら，実環境下では周囲雑音や残響により音声認識性能が劣化するという問題が残されている．現在の音声認識システムでは，簡易で効果的な解決策として接話マイクロホンがよく用いられている．このように口とマイクロホンを十分接近させて発声し，信号対雑音比（SNR: Signal to Noise Ratio）を高くすることにより，音声認識性能の劣化を防ぐことができる．しかしながら，接話マイクロホンの装着には煩わしさや不快感がともない，また利用者の行動範囲を大幅に制限してしまうという欠点がある．音声インタフェースの1つの利点は，利用者がキーボードのような入力装置を意識せずに，離れた場所で自由に動き回りながら機械に指示できることである．よって，自然で使い勝手の良い音声インタフェースを実現するために，実環境下でのハンズフリー音声認識の技術は必要不可欠である．

マイクロホンから離れた位置で発声された音声は，周囲雑音や残響の影響を強く受けて著しく歪んでしまう．従来，この問題に対して主に（1）音声強調手法と（2）モデル適応手法が試みられてきた．（1）の代表的な手法としては，周囲雑音に対してスペクトルサブトラクション[1]，残響に対してケプストラム正規化法[2]が提案されている．しかしながら，これらの手法には非定常な要因に対する問題が残されている．また，周囲雑音と残響はスペクトル領域において線形でないので，これらの1チャネルの情報のみを利用する方法では周囲雑音と残響を同時に扱うことは難しい．（2）の代表的な手法としては，周囲雑音に対してモデル分割手法[3]，モデル合成手法[4]などが提案されている．また，最近では，周囲雑音と残響の両方に対して同時にモデル合成手法を適用する試みもなされている[5]．し

† 奈良先端科学技術大学院大学情報科学研究科
Graduate School of Information Science, Nara Institute of Science and Technology

かしながら，これらの手法には対象とする環境の事前知識が必要であるという問題がある．

本論文では，周囲雑音と残響の両方に対処するための手法として，マイクロホンアレーの適用について検討する．マイクロホンアレーは空間的に配置された複数のマイクロホンで構成されており，各受音信号間には音源とマイクロホンの位置関係に応じた位相差や振幅差が生じる．これらの空間的な情報を利用して，目的とする音源の方向に感度が高く，それ以外の方向に感度の低い指向性を形成することができる．この指向性ビームを発話者の方向に向けることにより，SNRを大幅に向上させることが可能となる．従来の1チャネル型の音声強調手法とは異なり空間的な情報を利用しているので，周囲雑音と残響の両方を抑圧できる．

マイクロホンアレーを音声認識に適用する際の問題点について考える．ハンズフリーの状況では，発話者はマイクロホンから離れた任意の位置で発声することができる．よって，発話者の方向に指向性ビームを安定して向けるために，発話者の方向を正確に検出する技術が非常に重要となる．指向性ビームの方向と発話者方向にずれが生じた場合や，雑音源方向を発話者方向として誤って検出した場合には音声認識性能に劣化が生じてしまう．これまでに様々な発話者方向検出法が提案されており[6)~8)]，最近では音声認識に適用した例がいくつか報告されている[9),10)]．これらの手法では，発話者方向を検出するために音源のパワー情報や各受音信号間の相関情報を利用している．しかしながら，低SNR環境下などでは発話者方向検出の精度が十分でなかった．本論文では，低SNR環境下での発話者方向検出精度を向上させるために音声特有の調波構造という情報を積極的に活用する方法を提案し，シミュレーションと簡易音響実験室で認識実験を行うことにより有効性を評価する．

2. マイクロホンアレーを用いた発話者方向検出による音声認識

本論文では，図**1**に示すように，マイクロホンアレー，発話者方向検出部，雑音抑圧部，音声認識部で構成されるシステムを使用する．発話者方向検出部では，複数の音源の方向を各々推定し，これらのうちどれが発話者方向であるのかを判定する．雑音抑圧部では，発話者方向検出部で得られた発話者方向に指向性ビームを向けて雑音と残響を抑圧する．最後に，音声認識部で音声認識を行う．本論文では低SNR環境下での発話者方向検出精度を向上させることに焦点をあてているので，以下ではマイクロホンアレー信号処理として遅延和アレー[11)]を用いることにする．

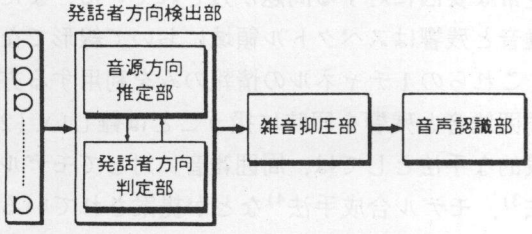

図1 システム構成
Fig. 1 Block diagram of the system.

2.1 遅延和アレー

まず，どのような音場を仮定するのかについて考える．一般に，音源とマイクロホンアレーの距離がマイクロホンアレーの全長程度しかない場合，球面波音場を仮定しないと誤差が大きくなる．本論文では，全長約37 cmのマイクロホンアレーを使用し，音源とマイクロホンアレーの距離として1 mから3 m程度を対象としている．この場合，音源とマイクロホンアレーの距離はマイクロホンアレーの全長の3倍以上となるので，以下では平面波音場を仮定している．

遅延和アレーの原理を説明する．図**2**に示すように，周波数fの複素正弦波がθ方向から到来し，マイクロホン数M，マイクロホン間隔dの等間隔直線配列マイクロホンアレーで受音される状況を考える．各マイクロホンの受音信号$x_1(t), x_2(t), \cdots, x_M(t)$間には，式(1)のような時間差が生じる．

$$x_i(t) = x_1\left(t - (i-1)\frac{d\cos\theta}{c}\right) \tag{1}$$

cは音速，iはマイクロホン番号である．このとき，遅延和アレーの出力信号$y(t)$は式(2)で表される．

$$\begin{aligned} y(t) &= \sum_{i=1}^{M} x_i\left(t + (i-1)\frac{d\cos\theta}{c}\right) \\ &= \sum_{i=1}^{M} x_i(t)\exp\left\{j2\pi f(i-1)\frac{d\cos\theta}{c}\right\} \end{aligned} \tag{2}$$

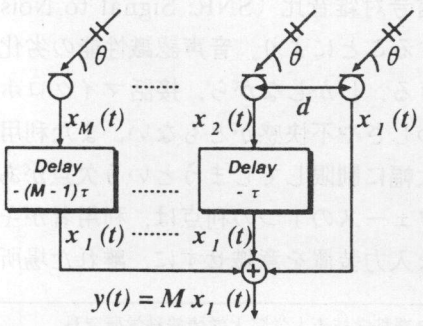

図2 遅延和アレー
Fig. 2 Delay-and-sum beamformer.

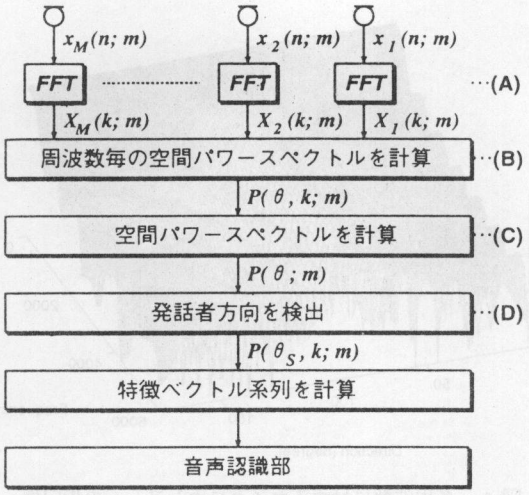

図3　処理フロー
Fig. 3　Process flow of the system.

式(2)はθ方向から到来する信号を同相化することに相当し，θ方向から到来する信号はM倍になって出力される．一方，θ方向以外から到来する信号は同相化されず，M倍にはならない．したがって，θ方向に感度が高く，それ以外の方向に感度の低い指向性が形成されることになる．

2.2　処理フロー

具体的な処理フローを図**3**に示す．以下，図中の(A)～(D)の処理について説明する．

(A)　**周波数分析部**　広帯域時間信号である音声に対して遅延和アレーを適用するために，各マイクロホンの受音信号を周波数帯域成分に分解する．図3の$x_1(n;m),\cdots,x_M(n;m)$は各マイクロホンの受音信号であり，$X_1(k;m),\cdots,X_M(k;m)$は離散フーリエ変換により得られた周波数帯域成分である．ここで，nは時刻，kは周波数帯域番号，mは短時間周波数分析におけるフレーム番号を表す．

(B)(C)　**音源方向推定部**　まず，式(3)で定義する周波数ごとの空間パワースペクトルを計算する．

$$P(\theta,k;m)=\left|\sum_{i=1}^{M}X_i(k;m)\exp\left\{j2\pi f_k(i-1)\frac{d\cos\theta}{c}\right\}\right|^2 \tag{3}$$

ここで，$\theta=0,1,\cdots,180$，f_kは周波数帯域番号kに対応する周波数である．次に，式(4)で定義する空間パワースペクトルを計算する．

$$P(\theta;m)=\sum_{k=0}^{K-1}P(\theta,k;m) \tag{4}$$

式(3)と式(4)はマイクロホンアレーの指向性ビームをフレームごとに対象とするすべての方向に順次向け，各々の方向のパワーを求めることに相当する．空間パワースペクトルは音源が存在する方向で大きな値をとる．よって，そのピークとなる方向を検出することにより，発話者と雑音源の区別なくパワーの大きい順にA個の音源方向$\theta_1,\cdots,\theta_A$が得られる．

(D)　**発話者方向判定部**　(B)(C)で推定したA個の音源方向の中から発話者方向θ_Sを判定する．発話者方向を判定した後，雑音抑圧された音声のパワースペクトルを$P(\theta_S,k;m)$，$k=0,1,\cdots,K-1$として求めることができる．ここで，Kは周波数帯域数である．具体的な発話者方向判定法については次節で述べる．

2.3　調波構造検出に基づく発話者方向判定法

従来法では，複数の音源のうち最もパワーの大きい音源の方向を発話者方向として判定する．しかしながら，低SNR環境下や音声の無声子音区間などで誤判定が生じてしまう．本論文では，このような問題に対処するために，調波構造検出に基づく発話者方向判定法を導入する．調波構造は音声に特有の情報であり，声の高さに相当するピッチ成分がパワースペクトルの繰り返し構造として現れるものである．複数の発話者が同時に発声するという状況も実際には起こりうるが，本論文では最も単純な場合として1人の発話者と調波性のない雑音が存在する状況を想定している．具体的な判定アルゴリズムを以下に示す．

(1)　音源方向推定部で得られたA個の音源方向の各々に対して次の処理を実行する．a番目の音源方向θ_aのパワースペクトル$P(\theta_a,k;m)$，$k=0,1,\cdots,K-1$をケプストラム分析し，$C(\theta_a,l;m)$，$l=0,1,\cdots,L-1$を求める．ここで，Lはケプストラム分析の次数である．式(5)の$C_{\max}$が一定の閾値以上なら調波構造が含まれると判断する．

$$C_{\max}=\max_{l\in\mathbf{l}}C(\theta_a,l;m) \tag{5}$$

ここで，**l**は高ケフレンシ部（約3.3 ms以上）を表す．

(2)　A個の音源方向のうちB ($\leq A$)個の音源方向に調波構造が含まれる場合，その中から最もパワーの大きい音源の方向を発話者方向として判定する．どの音源方向にも調波構造が含まれない場合，過去の最も最近判定した発話者方向を採用する．なお，音声の開始時点から調波構造が初めて検出されるまでのフレームについては，最初に判定された発話者方向を採用する．

ステップ(1)における閾値をある程度大きくする場合，調波構造が顕著に含まれる音源方向のみで発話者方向を判定することになり，判定の信頼性は高くなる．その反面，すべてのフレームでどの音源方向にも調波構造が検出されなくなる恐れがある．また，閾値を極端に小さくする場合，ほとんどの音源方向に調波構造が含まれると判断され，従来法と同様になる．よって，閾値の値を変化させることにより，音源のパワー情報と音声の調波構造という情報のどちらを重要視するのかを調節することができる．

最後に提案法の計算時間について述べる．発話者方向検出部において最も時間がかかる処理は周波数ごとの空間パワースペクトルの計算である☆．本論文では提案法の基本性能の評価を目的としているので実時間化の検討は行っていないが，高速化のためには，たとえば周波数ごとの空間パワースペクトルの分解能を下げる，発話者の移動を制限するなどが考えられる．

3. シミュレーションデータを用いた評価実験

3.1 実験条件

マイクロホンアレーはマイクロホン数14，マイクロホン間隔2.83 cmの等間隔直線配列であり，各マイクロホンは無指向性である．発話者はATR音声データベース[12]のSetAの重要語5240単語のうち下1桁が1の500単語であり，男性話者MHTを使用する．また，雑音源として白色ガウス雑音と計算機雑音の2通りを用いる．各マイクロホンの受音信号については，平面波を仮定し，時間遅れのみを考慮して計算機シミュレーションで生成する．

サンプリング周波数は12 kHzであり，32 msec（384点）のハミング窓をかけて信号を切り出す．高域強調$(1-0.97z^{-1})$後，0づめをしてから512点でFFT分析を行う．なお，フレーム周期は8 msecである．特徴ベクトルとしてメルケプストラム16次，Δケプストラム16次，Δパワーを用いる．音声認識にはTied-mixture分布型HMM[13]を使用する．ここで，混合数はメルケプストラムとΔケプストラムについて256，Δパワーについて128である．Tied-mixture分布型HMMではすべての状態で同じ出力確率分布の集合を共有しているので，少量の学習データでも精度良くHMMを学習できるという利点がある．環境独立の54音素モデルを話者MHTの重要語5240単語のうち偶数番号の2620単語で学習している．

☆ 現在の処理では，発話者方向検出部全体で1フレームあたりのCPU時間は約4.08秒である．そのうち，音源方向推定部の計算に約4.00秒，調波構造検出に基づく発話者方向判定部で約0.08秒となっている．なお，CPU時間の測定にはDEC社製のAlphaStation500/400（CPU：Alpha21164A 400 MHz，CPU処理能力：480 SPECInt92，600 SPECfp92）を用いている．

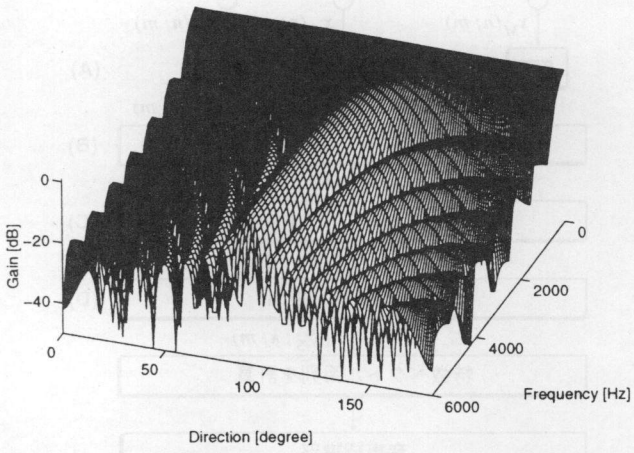

図4 各周波数に対するマイクロホンアレーの指向性（シミュレーション）
Fig. 4 Directive gain pattern for each frequency (simulation).

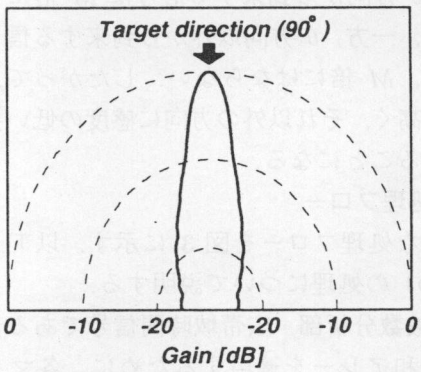

図5 全帯域に対するマイクロホンアレーの指向性（シミュレーション）
Fig. 5 Directive gain pattern (simulation).

3.2 マイクロホンアレーの指向性

90° 方向から帯域6 kHzの白色ガウス雑音を放射させて求めた各周波数に対するマイクロホンアレーの指向性を図4に示す．また，全帯域に対するマイクロホンアレーの指向性を図5に示す．図5は図4の各周波数に対する指向性を全帯域にわたって加算することにより得られる．図5から，たとえば40° 方向に対するゲインは約 −20 dBであることが分かる．

一般に，マイクロホンアレーの指向性はマイクロホン数を増やすほど鋭くなる．ただし，空間的折返し現象を防ぐためには，空間のサンプリング周波数 $1/d$（dはマイクロホン間隔）が最大空間周波数 $f_{\max}$ の2倍以上であるという条件が加わる[14]．この場合，低周波数帯域については，割り当てられるマイクロホン数が

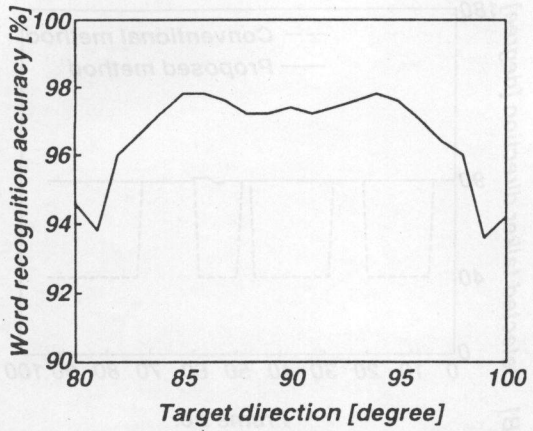

図6 目的方向を変化させたときの単語認識率（発話者方向は 90°）
Fig. 6 Sensitivity of word recognition accuracy to difference between target direction and talker direction.

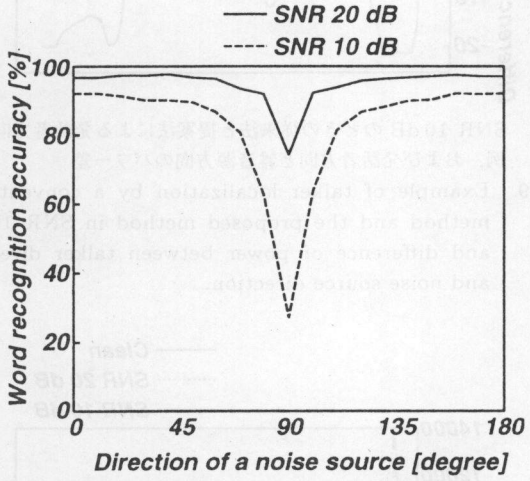

図7 雑音源方向を変化させたときの単語認識率（発話者方向と目的方向は 90°）
Fig. 7 Sensitivity of word recognition accuracy to difference between talker direction and noise source direction.

少なくなるので鋭い指向性を得ることができない．その結果，図4に示すようにマイクロホンアレーの指向性の鋭さには周波数依存性が生じる．したがって，マイクロホンアレーの指向性ビームの方向（目的方向）と発話者方向にずれが生じた場合，音声に周波数歪みがかかることになる．発話者方向を真正面 90° とし，目的方向を 80° から 100° の間で変化させたときの単語認識率を**図6**に示す．図6から約 ±5° 以内のずれであれば単語認識率にほとんど劣化が生じないことが分かる．次に，発話者方向と目的方向を 90° とし，白色ガウス雑音源の方向を 0° から 180° の間で変化させたときの単語認識率を**図7**に示す．SNR 20 dB のとき，SNR 10 dB のときともに，発話者と雑音源の方向差が約 50° 以内の場合には，指向性の鋭さが十分ではないので単語認識率に劣化が見られる．

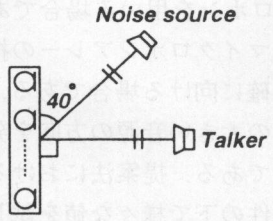

図8 音源とマイクロホンアレーの配置
Fig. 8 Sound sources and a microphone array.

表1 単語認識率（WA）[%] と発話者方向検出精度（TLA）[%]（シミュレーション，白色ガウス雑音）
Table 1 Word recognition accuracy (WA) [%] and talker localization accuracy (TLA) [%] (simulation, white Gaussian noise).

		SNR [dB]		
		Clean	20	10
シングルマイク	WA	97.2	74.8	27.4
	TLA	—	—	—
発話者方向既知	WA	97.4	97.6	90.0
	TLA	100.0	100.0	100.0
従来法	WA	97.4	83.0	29.8
	TLA	100.0	56.6	24.1
提案法	WA	97.2	97.6	90.0
	TLA	98.6	99.8	99.8

表2 単語認識率（WA）[%] と発話者方向検出精度（TLA）[%]（シミュレーション，計算機雑音）
Table 2 Word recognition accuracy (WA) [%] and talker localization accuracy (TLA) [%] (simulation, computer noise).

		SNR [dB]		
		Clean	20	10
シングルマイク	WA	97.2	88.6	56.4
	TLA	—	—	—
発話者方向既知	WA	97.4	91.6	71.4
	TLA	100.0	100.0	100.0
従来法	WA	97.4	91.6	66.6
	TLA	100.0	85.0	67.9
提案法	WA	97.2	91.4	71.0
	TLA	98.6	99.3	99.5

3.3 認識実験

音源とマイクロホンアレーの配置を**図8**に示す．発話者方向は 90°，雑音源方向は 40° である．

雑音源として白色ガウス雑音源を用いたときの単語認識率（WA: Word recognition Accuracy）と発話者方向検出精度（TLA: Talker Localization Accuracy）を**表1**に示す．同様に，雑音源として計算機雑音を用いたときの結果を**表2**に示す．ここで，発話者方向検出精度は，発話者方向を ±3° 以内の誤差で検出した場合を正解とし，(正解フレーム数/全フレーム数) × 100 で定義される．表中の Clean は雑音源なしの場合を表す．シングルマイクはマイクロホンアレーの

8番目のマイクロホンを用いる場合である．また，発話者方向既知はマイクロホンアレーの指向性ビームを発話者方向に正確に向ける場合である．さらに，従来法は最もパワーの大きい音源の方向を発話者方向として判定する場合である．提案法における閾値については，同じ実験条件の下で様々な値を試し，それらのうちで最も性能の高かったものを用いている．閾値の影響については後述する．実験結果を以下にまとめる．

- 発話者方向既知の単語認識率は，SNR 20 dB のとき，SNR 10 dB のときともに，シングルマイクと比べて大幅に改善されている．これは，マイクロホンアレーの指向性ビームを発話者方向に向けることにより図5に示すようにSNRを約20 dB向上させていることによる．
- 計算機雑音に対する発話者方向既知の単語認識率は，白色ガウス雑音に対する場合よりも低くなっている．この原因は，計算機雑音のエネルギーは低周波数帯域に集中しているが，3.2節で述べたようにマイクロホンアレーの指向性は低周波数帯域で鋭くないことにある．
- 計算機雑音に対する従来法の単語認識率は，SNR 20 dB のとき，SNR 10 dB のときともに白色ガウス雑音に対する場合よりも高くなっている．この原因は，計算機雑音に対する発話者方向検出精度が白色ガウス雑音に対する場合よりも高くなっていることにある．従来法では発話者方向を検出するために空間パワースペクトルを計算する．しかしながら，音源のSNRが同一であっても空間パワースペクトルの形状は，マイクロホンアレーの指向性における周波数依存性，音声のスペクトル形状，雑音のスペクトル形状など様々な要因により影響を受ける．これらの影響のために，本実験では計算機雑音に対する発話者方向検出精度の方が高くなっているものと考えられる．
- 従来法の発話者方向検出精度は，SNRの低下につれて大きく劣化している．一方，提案法では，SNR 20 dB のとき，SNR 10 dB のときともにほとんど完全に発話者方向を検出している．その結果，提案法では発話者方向既知と同等の単語認識率が得られている．

SNR 10 dB のときの従来法と提案法による発話者方向検出例，および発話者方向と雑音源方向のパワー差を**図9**に示す．ここで，雑音源は白色ガウス雑音である．従来法では音声のパワーが小さいフレームで雑音源方向を発話者方向として誤判定しているが，提案法では発話者方向を良好に検出しているのが分かる．

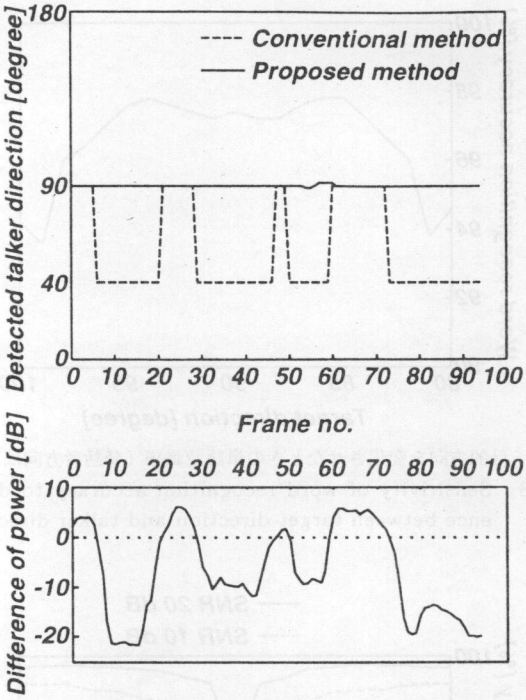

図9 SNR 10 dB のときの従来法と提案法による発話者方向検出例，および発話者方向と雑音源方向のパワー差

Fig. 9 Example of talker localization by a conventional method and the proposed method in SNR 10 dB, and difference of power between talker direction and noise source direction.

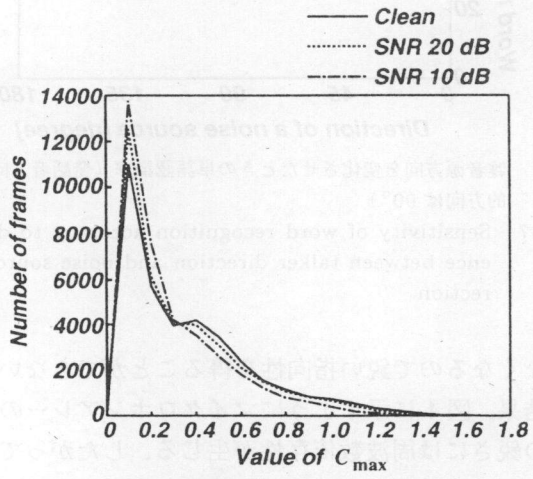

図10 発話者方向における $C_{\max}$ の頻度分布

Fig. 10 Histogram of $C_{\max}$ in talker direction.

次に，提案法における閾値の影響について調べる．まず，発話者方向における $C_{\max}$ の頻度分布を**図10**に示す．ここで，横軸は式(5)の $C_{\max}$ の値，縦軸はその値をとるフレームの数を表す．図10から0.5以上の値をとるフレーム数が激減しているのが分かる．実際，Cleanのときは閾値を0.7以上に設定するとすべてのフレームで調波構造が検出されない単語が現れる．同様に，SNR 20 dB のときは0.6以上，SNR 10 dB のときは0.5以上に閾値を設定するとすべての

フレームで調波構造が検出されない単語が現れる．調波構造が検出されない単語については発話者方向を検出できないので，認識を行うことができなくなる．本論文では，すべての単語が認識可能であるという条件で閾値を設定している．次に，提案法における閾値の値を変化させたときの単語認識率と発話者方向検出精度を図**11**に示す．ただし，雑音源として白色ガウス雑音源を用いている．図11から，閾値が0.4のときに最も高い性能となっているのが分かる．また，このときの単語認識率は発話者方向既知と同等となっており，提案法で達成できる上限である．雑音源として計算機雑音を用いる場合にも同様である．以上から，提案法の閾値として約0.3～0.4の値を設定すればよいと考えられる．

表**3**に女性話者FTKの場合の単語認識率と発話者方向検出精度を示しておく．ここで，雑音源として白色ガウス雑音源を用いている．表3から，女性話者の場合にも，提案法では良好に発話者方向を検出しており，従来法よりも高い単語認識率が得られていることが分かる．

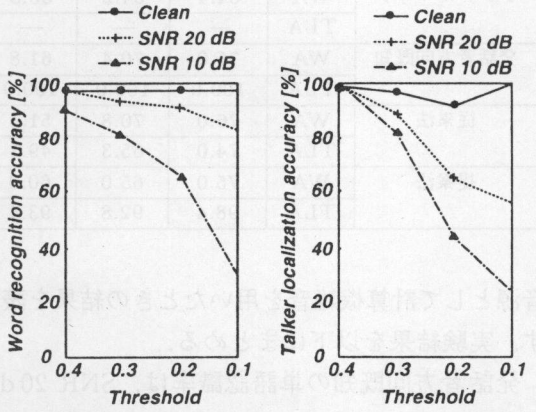

図**11** 閾値を変化させたときの単語認識率［%］と発話者方向検出精度［%］（シミュレーション，白色ガウス雑音）

Fig. 11 Word recognition accuracy [%] and talker localization accuracy [%] for various values of threshold (simulation, white Gaussian noise).

表**3** 女性話者FTKの場合の単語認識率（WA）［%］と発話者方向検出精度（TLA）［%］（シミュレーション，白色ガウス雑音）

Table 3 Word recognition accuracy (WA) [%] and talker localization accuracy (TLA) [%] for a woman talker FTK (simulation, white Gaussian noise).

		SNR [dB]		
		Clean	20	10
シングルマイク	WA	92.0	72.4	21.6
	TLA	—	—	—
発話者方向既知	WA	92.0	93.0	85.8
	TLA	100.0	100.0	100.0
従来法	WA	92.0	76.2	27.4
	TLA	100.0	55.2	21.7
提案法	WA	91.8	93.0	80.8
	TLA	98.8	99.5	91.7

4. 実環境データを用いた評価実験

4.1 実験条件

実環境データの収録状況を図**12**に示す．図12は残響時間約0.18秒の簡易音響実験室である．簡易音響実験室における残響時間は通常のオフィスなどの環境と比べて多少短い．マイクロホンアレーはマイクロホン数14，マイクロホン間隔2.83 cmの等間隔直線配列であり，各マイクロホンは無指向性である．また，発話者方向は真正面90°，雑音源方向は40°とし，マイクロホンアレーから2 mの距離に設置している．なお，音源としてスピーカを使用している．その他の実験条件は3.1節と同様である．

4.2 マイクロホンアレーの指向性

90°方向から帯域6 kHzの白色ガウス雑音を放射させて求めた各周波数に対するマイクロホンアレーの指

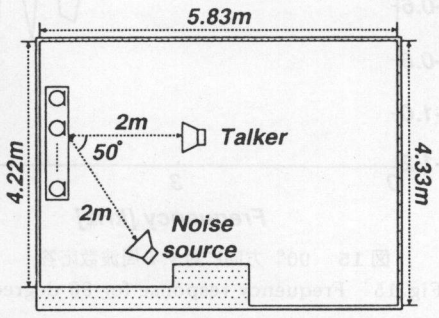

図**12** 簡易音響実験室における音源とマイクロホンアレーの配置

Fig. 12 Sound sources and a microphone array in an experimental room.

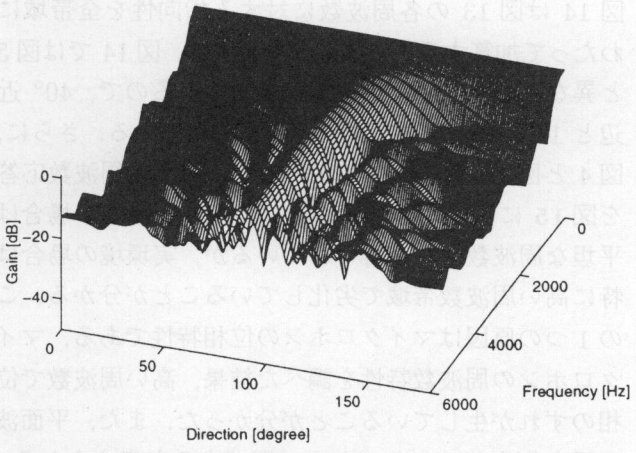

図**13** 各周波数に対するマイクロホンアレーの指向性（簡易音響実験室）

Fig. 13 Directive gain pattern for each frequency (experimental room).

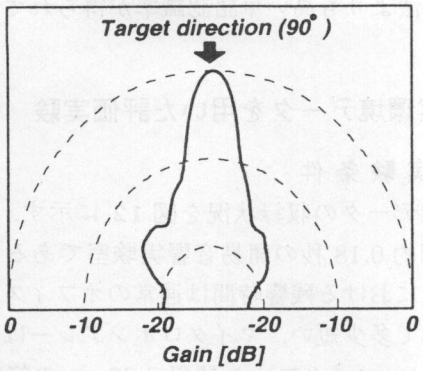

図 14 全帯域に対するマイクロホンアレーの指向性（簡易音響実験室）
Fig. 14 Directive gain pattern (experimental room).

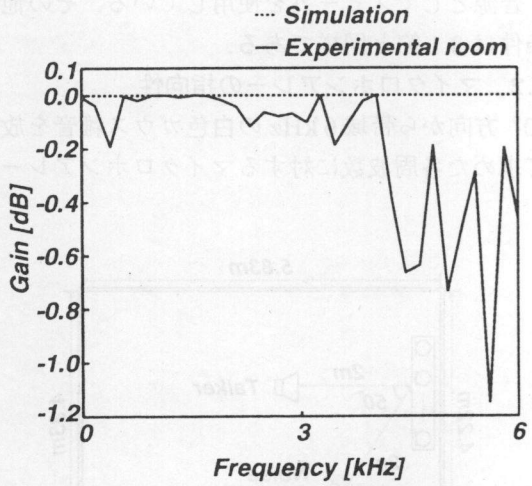

図 15 90° 方向に対する周波数応答
Fig. 15 Frequency response for 90 degree.

向性を図 **13** に示す．図 13 では図 4 と比べて全体的に減衰量が低下しているのが分かる．また，全帯域に対するマイクロホンアレーの指向性を図 **14** に示す．図 14 は図 13 の各周波数に対する指向性を全帯域にわたって加算することにより得られる．図 14 では図 5 と異なり壁からの反射波を受音しているので，40° 近辺と 140° 近辺で減衰量が特に低下している．さらに，図 4 と図 13 から求めた 90° 方向に対する周波数応答を図 **15** に示す．図 15 からシミュレーションの場合は平坦な周波数応答が得られているが，実環境の場合は特に高い周波数帯域で劣化していることが分かる．この 1 つの原因はマイクロホンの位相特性である．マイクロホンの周波数特性を調べた結果，高い周波数で位相のずれが生じていることが分かった．また，平面波音場を仮定していることも一因であると考えられる．

4.3 認識実験

雑音源として白色ガウス雑音源を用いたときの単語認識率と発話者方向検出精度を表 **4** に示す．同様に，

表 4 単語認識率（WA）[%] と発話者方向検出精度（TLA）[%]（簡易音響実験室，白色ガウス雑音）
Table 4 Word recognition accuracy (WA) [%] and talker localization accuracy (TLA) [%] (experimental room, white Gaussian noise).

		SNR [dB]		
		Clean	20	10
シングルマイク	WA	64.4	32.8	7.8
	TLA	—	—	—
発話者方向既知	WA	75.0	67.6	48.4
	TLA	100.0	100.0	100.0
従来法	WA	76.0	52.6	18.6
	TLA	74.0	44.9	21.7
提案法	WA	75.0	62.6	42.8
	TLA	98.4	86.3	73.4

表 5 単語認識率（WA）[%] と発話者方向検出精度（TLA）[%]（簡易音響実験室，計算機雑音）
Table 5 Word recognition accuracy (WA) [%] and talker localization accuracy (TLA) [%] (experimental room, computer noise).

		SNR [dB]		
		Clean	20	10
シングルマイク	WA	64.4	54.2	33.8
	TLA	—	—	—
発話者方向既知	WA	75.0	70.4	61.8
	TLA	100.0	100.0	100.0
従来法	WA	76.0	70.8	51.4
	TLA	74.0	65.3	49.1
提案法	WA	75.0	65.0	60.0
	TLA	98.4	92.8	93.4

雑音源として計算機雑音を用いたときの結果を表 **5** に示す．実験結果を以下にまとめる．

- 発話者方向既知の単語認識率は，SNR 20 dB のとき，SNR 10 dB のときともに，シングルマイクと比べて大幅に改善されている．しかしながら，シミュレーションの場合と比較すると劣化が見られる．この主な原因は，マイクロホンアレーの指向性が鋭くないために残響を十分抑圧できないことにある．今回の実験では直線配列のマイクロホンアレーを床に並行に設置しているので，天井や床からの反射音を抑圧することができない．よって，2 次元配列のマイクロホンアレーを使用し，残響の抑圧量を向上させる必要がある．
- 計算機雑音に対する発話者方向既知の単語認識率は白色ガウス雑音に対する場合よりも高くなっており，シミュレーションの結果と逆になっている．この原因は，4.2 節で述べたように減衰量が高周波数帯域で低下していることにある．
- 従来法の発話者方向検出精度は，SNR の低下につれて大きく劣化している．一方，提案法では，SNR

20 dB のとき，SNR 10 dB のときともに従来法より良好に発話者方向を検出している．しかしながら，単語認識率に関しては従来法よりも低下する場合がある．これは，残響の影響により音源が存在しない方向で調波構造が検出され，このような誤りが長時間にわたって伝搬した結果ではないかと考えられる．

5. おわりに

本論文では，実環境下でのハンズフリー音声認識を実現するための方法としてマイクロホンアレーの適用について検討した．マイクロホンアレーを音声認識に適用する際には，発話者の方向に指向性ビームを安定して向けることが非常に重要となる．従来法では，発話者方向を検出するために音源のパワー情報や各受音信号間の相関情報を利用しているが，低 SNR 環境下などでは発話者方向検出の精度が十分でなかった．本論文では，低 SNR 環境下での発話者方向検出精度を向上させるために音声特有の調波構造という情報を積極的に活用する方法を提案し，シミュレーションと簡易音響実験室で認識実験を行うことにより有効性を評価した．その結果，白色ガウス雑音と計算機雑音の両方に対して従来法より高い発話者方向検出精度が得られ，低 SNR 環境下での単語認識率を大幅に改善できた．しかしながら，実環境下では提案法の単語認識率が従来法よりも低下する場合があった．これは，残響の影響により音源が存在しない方向で調波構造が検出され，このような誤りが長時間にわたって伝搬した結果ではないかと考えられる．提案法の閾値を自動で設定する方法などを含めてさらに検討が必要である．

本論文では 1 人の発話者と調波性のない雑音が存在する状況を想定したが，今後調波性のある雑音が存在する状況について評価を行う予定である．また，複数の発話者が存在するという状況や高残響下などでは，現在の発話者方向を一意に決定するという枠組みでは対処することができない．今後は，発話者方向の候補を複数同時に考慮しながら音声認識を行う方法について検討していく予定である．

参考文献

1) Boll, S.F.: Suppression of acoustic noise in speech using spectral subtraction, *IEEE Trans. Acoustics, Speech, and Signal Processing*, Vol.27, No.4, pp.113–120 (1979).
2) Atal, B.: Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification, *J. Acoust. Soc. Am.*, Vol.55, pp.1304–1312 (1974).
3) Varga, A.P. and Moore, R.K.: Hidden markov model decomposition of speech and noise, *ICASSP 90*, pp.845–848 (1990).
4) Martin, F., Shikano, K. and Minami, Y.: Rocognition of noisy speech by composition of speech and noise, *EUROSPEECH*, pp.1031–1034 (1993).
5) 滝口哲也，中村 哲，鹿野清宏：雑音と残響のある環境下での HMM 合成によるハンズフリー音声認識法，電子情報通信学会（D-II），No.12, pp.2047–2053 (1996).
6) Flanagan, J.L., Johnston, J.D., Zahn, R. and Elko, G.W.: Computer-steered microphone arrays for sound transduction in large rooms, *J. Acoust. Soc. Am.*, Vol.78, No.5, pp.1508–1518 (1985).
7) Silverman, H.F. and Kirtman, S.E.: A two-stage algorithm for determining talker location from linear microphone array data, *Computer Speech and Language*, Vol.6, pp.129–152 (1992).
8) Giuliani, D., Omologo, M. and Svaizer, P.: Talker localization and speech recognition using a microphone array and a cross-powerspectrum phase analysis, *ICSLP 94*, pp.1243–1246 (1994).
9) Lin, Q., Jan, E., Che, C. and Vries, B.: System of microphone arrays and neural networks for robust speech recognition in multimedia environment, *ICSLP 94*, pp.1247–1250 (1994).
10) Giuliani, D., Omologo, M. and Svaizer, P.: Experiments of speech recognition in a noisy and reverberant environment using a microphone array and HMM adaptation, *ICSLP 96*, pp.1329–1332 (1996).
11) Pillai, S.U.: *Array signal processing*, Springer-Verlag, New York (1989).
12) 武田一哉，匂坂芳典，片桐 滋，桑原尚夫：研究用日本語音声データベースの構築，日本音響学会誌，Vol.44, No.10, pp.747–754 (1988).
13) Bellegarda, J.R. and Nahamoo, D.: Tied mixture continuous parameter models for large vocabulary isolated speech recognition, *ICASSP 89*, pp.13–16 (1989).
14) 大賀寿郎，山崎芳男，金田 豊：音響システムとディジタル処理，コロナ社 (1995).

(平成 9 年 6 月 30 日受付)
(平成 10 年 1 月 16 日採録)

山田　武志（学生会員）

昭和46年生．平成6年大阪市立大学工学部情報工学科卒業．平成8年奈良先端科学技術大学院大学情報科学研究科情報処理学専攻博士前期課程修了．現在，同博士後期課程在学中．マイクロホンアレー，音声認識の研究に従事．音響学会会員．

中村　哲（正会員）

昭和33年生．昭和56年京都工芸繊維大学工芸学部電子工学科卒業．昭和56～平成6年シャープ（株）中央研究所および情報技術研究所に勤務．昭和61～平成元年ATR自動翻訳電話研究所に出向．平成6年4月より奈良先端科学技術大学院大学情報科学研究科助教授．平成8年3～8月Rutgers University・CAIP Center客員教授．音声情報処理，主として音声認識の研究に従事．京都大学博士（工学）．平成4年日本音響学会粟屋学術奨励賞受賞．IEEE，電子情報通信学会，日本音響学会，人工知能学会各会員．

鹿野　清宏（正会員）

昭和22年生．昭和45年名古屋大学工学部電気学科卒業．昭和47年同大学院工学研究科修士課程修了．同年電電公社武蔵野電気通信研究所入所．昭和59～61年カーネギーメロン大客員研究員．昭和61～平成2年ATR自動翻訳電話研究所音声情報処理研究室長．平成4年NTTヒューマンインタフェース研究所主席研究員．平成6年4月より奈良先端科学技術大学院大学情報科学研究科教授．音情報処理学講座を担当，主として音声・音情報処理の研究および研究指導に従事．工学博士．昭和50年電子情報通信学会米沢賞，平成3年IEEE SP 1990 Senior Award，平成6年日本音響学会技術開発賞受賞．IEEE，音響学会各会員．