適応型アレーを用いた3次元ビタビ探索に基づくハンズフリー音声認識

山田武志† 中村 哲† 鹿野清宏†

実環境下でハンズフリーな音声認識を実現するために3次元ビタビ探索法を提案している．3次元ビタビ探索法では，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，特徴ベクトルの方向・フレーム系列を計算する．そして，方向とフレームとHMMの状態からなる3次元トレリス上で最も尤度の高いパスを探索することにより，発話者の移動軌跡と音素系列を同時に推定する．これまでにマイクロホンアレー信号処理として遅延和アレーを適用し，実環境で収録したデータを用いて3次元ビタビ探索法の有効性を確認している．本論文では，3次元ビタビ探索法の性能をさらに改善するために適応型アレーに基づく方法を提案する．発話者移動と発話者位置固定という2通りの条件で収録した実環境データを用いて認識実験を行った．その結果，発話者移動時の提案法の単語認識率は，シングルマイクと比べてSNR 18 dBのとき11.1%，SNR 10 dBのとき42.6%，また遅延和アレーを適用する場合と比べてSNR 18 dBのとき6.9%，SNR 10 dBのとき28.7%改善され，適応型アレーの効果が示された．

Hands-free Speech Recognition Based on 3-D Viterbi Search Using Adaptive Beamforming

TAKESHI YAMADA,† SATOSHI NAKAMURA† and KIYOHIRO SHIKANO†

We are investigating a speech recognition algorithm based on 3-D Viterbi search using a microphone array to realize hands-free speech recognition in real environments. The 3-D Viterbi search method extracts a direction-frame sequence of parameter vectors by steering a beamform to each direction every frame. Then Viterbi search is performed in 3-dimensional trellis space composed of talker directions, input frames, and HMM states. A locus of the talker and a phoneme sequence of the speech are obtained by finding an optimal path with the highest likelihood. To improve the performance of the 3-D Viterbi search method in real environments, this paper proposes a novel method based on an adaptive beamforming technique instead of the delay-and-sum beamformer used in our previous study. Speaker-dependent isolated-word recognition experiments were carried out on real environment data to evaluate the effect of the adaptive beamformer. For a moving talker, the word recognition accuracy of the 3-D Viterbi search method with the adaptive beamformer in SNR 10 dB was 42.6% higher than that of the single microphone, and 28.7% higher than that of the 3-D Viterbi search method with the delay-and-sum beamformer. These results show that the use of the adaptive beamformer is very effective.

1. はじめに

現在の音声認識では，周囲雑音や残響による認識精度の低下を防ぐために，接話マイクロホンの使用を前提としている．このように口とマイクロホンを十分接近させて発声することにより，信号対雑音比（SNR: Signal to Noise Ratio）を高くすることができる．しかし，マイクロホンの位置をつねに意識しながら発声する必要があるので，音声インタフェースとして自然で使い勝手の良いものではない．音声インタフェースの利点を十分に生かすためには，離れた場所で自由に動き回りながら発声された音声を精度良く認識する技術，すなわちハンズフリー音声認識の技術が必要不可欠である．

実環境下でハンズフリー音声認識を実現するための1つの方法は，マイクロホンアレーを利用することである．マイクロホンアレーとは複数のマイクロホンを空間的に広く配置したものである．よって，各マイクロホンの受音信号間にはマイクロホンと音源の位置関係に応じた位相差や振幅差が生じることになる．これらの空間的な情報を利用することにより，発話者の方

† 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

向（位置）に感度が高く，それ以外の方向（位置）に感度が低い指向性を形成して周囲雑音や残響を抑圧することができる．

最近，マイクロホンアレーを用いた音声認識システムがいくつか報告されている[1)~5)]．これらのシステムでは，短時間パワーや長時間パワーなどの音声の物理的な性質に着目して発話者の移動軌跡を推定する．そして，推定した発話者の方向に指向性ビームを向けて認識のための特徴ベクトルを計算し，音声認識を行う．しかし，低SNR環境下などで発話者の移動軌跡を正確に推定することは難しく，認識精度の低下の原因となっていた．従来のシステムでは，マイクロホンアレーを音声認識の前処理としてとらえているので，マイクロホンアレー処理部で生じたエラーを音声認識部で回復することはできない．

著者らは，この問題に対処するために，マイクロホンアレーと音声認識を統合して同じ枠組みの中で扱う方法について検討しており，マイクロホンアレーを用いた3次元ビタビ探索に基づく音声認識法（以下，3次元ビタビ探索法と呼ぶ）を提案している[6),7)]．3次元ビタビ探索法では，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，特徴ベクトルの方向・フレーム系列を計算する．そして，方向とフレームとHMMの状態からなる3次元トレリス上で最も尤度の高いパスを探索することにより，発話者の移動軌跡と音素系列を同時に推定する．これまでにマイクロホンアレー信号処理として遅延和アレーを適用し，実環境で収録したデータを用いて3次元ビタビ探索法の有効性を確認している[8)]．本論文では，3次元ビタビ探索法の性能をさらに改善するために適応型アレーに基づく方法を提案し，発話者移動と発話者位置固定という2通りの条件で収録した実環境データを用いて評価する．

2. 3次元ビタビ探索法

従来のマイクロホンアレーを用いた音声認識システムの処理フローを**図1**に示す[1)~5)]．まず，短時間パワーや長時間パワーなどの音声の物理的な性質に着目して発話者の方向を推定する．発話者の移動にも追従できるように，この処理は短い時間ごとに行われる．そして，推定した発話者の方向に指向性ビームを向けて認識のための特徴ベクトルを計算し，音声認識を行う．従来のシステムの動作例を**図2**に示す．ここで，図中の細い実線は発話者の移動軌跡，点線は雑音源の方向，太い実線は推定された発話者の移動軌跡を表している．また，図中の□は特徴ベクトルである．音

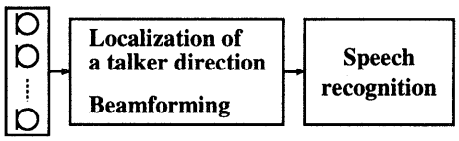


図1 従来のマイクロホンアレーを用いた音声認識システムの処理フロー

Fig. 1 A block diagram of the conventional speech recognition systems using a microphone array.

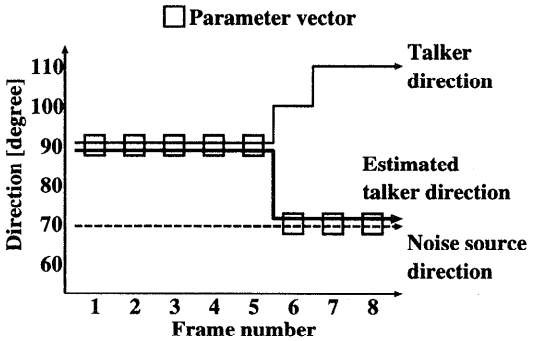


図2 従来のシステムの動作例

Fig. 2 An example of behavior of the conventional systems.

声のパワーが雑音のパワーより安定して大きい場合，1～5フレームのように発話者の方向は正しく推定される．しかし，雑音のパワーが非常に大きい場合，6～8フレームのように雑音源の方向を発話者の方向として誤って推定してしまう．このような場合，雑音源の方向に指向性ビームを向けて特徴ベクトルを計算するので，認識精度は大きく低下する．従来のシステムでは，マイクロホンアレーを音声認識の前処理としてとらえているので，マイクロホンアレー処理部で生じたエラーを音声認識部で回復することはできない．著者らは，この問題に対処するために，マイクロホンアレーと音声認識を統合して同じ枠組みの中で扱う方法について検討しており，3次元ビタビ探索法を提案している[6),7)]．

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て特徴ベクトルを計算することにより，**図3**に示すような特徴ベクトルの方向・フレーム系列が得られる．このとき，式(1)で示されるように，マイクロホンアレーと音声認識を同じ統計的な枠組みの中で扱うことが可能となる．

$$(\hat{\mathbf{q}}, \hat{\mathbf{d}}) = \underset{(\mathbf{q},\mathbf{d})}{\operatorname{argmax}} P(\mathbf{x}, \mathbf{q}, \mathbf{d} \mid \mathbf{M}) \qquad (1)$$

ここで，$(\hat{\mathbf{q}}, \hat{\mathbf{d}})$ は音素系列 $\mathbf{q}$ と発話者の移動軌跡 $\mathbf{d}$ の最適な組である．また，$\mathbf{x}$ は特徴ベクトルの方向・フレーム系列，$\mathbf{M}$ は音声のモデルである．式(1)に基

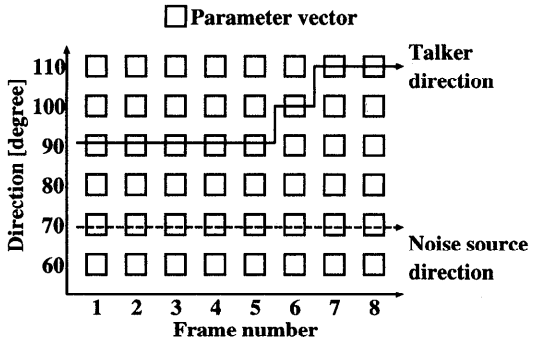


図3 特徴ベクトルの方向・フレーム系列
Fig. 3 A direction-frame sequence of parameter vectors.

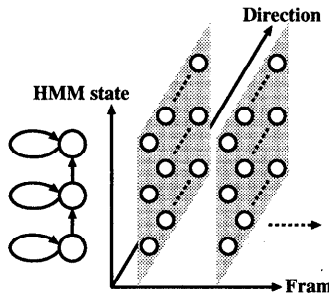


図4 方向とフレームと HMM の状態からなる 3 次元トレリス
Fig. 4 3-dimensional trellis space composed of talker directions, input frames, and HMM states.

づいて発話者の移動軌跡と音素系列を同時に推定するためには，**図4**に示すような方向とフレームと HMM の状態からなる 3 次元トレリス上で最も尤度の高いパスを探索すればよい．これは，各方向に対する音声認識の経過を考慮しながら発話者の方向を推定していることに相当する．累積尤度の計算式を式 (2) に示す．

$$\alpha(q,d,n) = \max_{q',d'}\{\alpha(q',d',n-1) + \log a_1(q',q) + \log a_2(d',d)\} + \log b(q,\mathbf{x}(d,n)) \quad (2)$$

ここで，α は累積尤度，q は HMM の状態番号，d は方向，n はフレーム番号である．また，$a_1(q',q)$ は HMM の状態 q' から q への遷移確率，$a_2(d',d)$ は方向 d' から d への遷移確率，b は出力確率，$\mathbf{x}$ は特徴ベクトルである．3 次元ビタビ探索法が良好に動作するためには，発話者の方向の尤度がそれ以外の方向の尤度より高いという条件が必要となる．そこで，音声らしい特徴を有する方向の尤度を高くすることを考え，式 (3) に示すような重みを式 (2) の右辺に加算している．

$$w(d,n) = \log \frac{\sum_{n'=n-(\nu-1)}^{n} \{c(d,n')\}^{\mu}}{\sum_{d'=1}^{D} \sum_{n'=n-(\nu-1)}^{n} \{c(d',n')\}^{\mu}} \quad (3)$$

ここで，$c(d,n)$ は，方向 d，フレーム番号 n でのケプストラム係数のうち，調波成分に対応する高ケフレンシ部における最大値である．この値が大きいほど調波構造が顕著に含まれることになり，音声らしいと見なすことができる．また，μ は方向間での重みの差を調節するパラメータであり，ν はどれだけ過去のフレームを考慮するのかを決めるパラメータである．

3. 適応型アレー

3 次元ビタビ探索法において特徴ベクトルの方向・フレーム系列を計算する際，どのようなマイクロホンアレー信号処理を適用するのかということが非常に重要となる．これまでにマイクロホンアレー信号処理として遅延和アレー（たとえば文献 9)）を適用し，実環境で収録したデータを用いて 3 次元ビタビ探索法の有効性を確認している[8]．

遅延和アレーの場合，複数のマイクロホンで受音した目的信号をすべて同相化して加算することにより，目的方向に超指向性を形成する．遅延和アレーは単純な処理で実現できるが，超指向性という原理上，指向性ビームの幅を鋭くしすぎると目的方向と発話者の方向のずれに非常に敏感になってしまう．また，指向性ビームの幅を鋭くするためには，マイクロホン数を増やさねばならない．一方，適応型アレーの場合，複数のマイクロホンで受音した雑音信号をすべて同相化して減算することにより，雑音源の方向に指向性の死角を形成する．適応型アレーでは周囲の環境に応じた指向性を自動的に形成するので，マイクロホン数が少ない場合でも遅延和アレーより雑音や反射音を効果的に抑圧することができる．また，発話者以外の雑音源のみを抑圧するように働くので，目的方向と発話者の方向のずれにあまり敏感ではなく，3 次元ビタビ探索法に適していると考えられる．本論文では，3 次元ビタビ探索法の性能をさらに改善するために適応型アレーに基づく方法を提案する．

適応型アレーのブロック図を**図5**に示す．図中の $S(\omega)$ は目的信号のスペクトル，$Y(\omega)$ は出力信号のスペクトルである．また，$G_m(\omega), m = 1,\cdots,M$ は音源から m 番目のマイクロホンまでの伝達特性，$H_m(\omega), m = 1,\cdots,M$ は適応フィルタの周波数特性である．ここで，M はマイクロホン数である．この

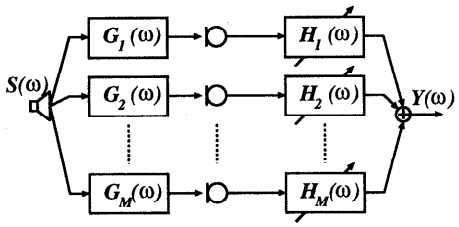


図5 適応型アレーのブロック図
Fig. 5 A block diagram of adaptive beamforming.

とき，マイクロホンアレーの目的信号に対する周波数特性 $F(\omega)$ は式 (4) で表される．

$$F(\omega) = \sum_{m=1}^{M} G_m(\omega) H_m(\omega) \tag{4}$$

適応型アレーでは $F(\omega)$ にある拘束条件を設定し，マイクロホンアレーの出力パワーを最小にするように適応フィルタ $H_m(\omega)$ を決定する．代表的な拘束条件を次に示す．

$$F(\omega) = 1 \tag{5}$$

$$D = \int |1 - F(\omega)|^2 d\omega \leq \widehat{D} \tag{6}$$

式 (5) は目的信号に劣化を許容しないという強い拘束条件である．一方，式 (6) は目的信号に $\widehat{D}$ までの劣化を許容し，式 (5) よりも大きな雑音抑圧量を得るという拘束条件である．ここで，D は目的信号の劣化量であり，レスポンス劣化量と呼ばれる．本論文では，式 (6) の拘束条件を満たす適応型アレーとして Kaneda ら[10]による AMNOR を用いることにする．

4. 実環境データを用いた認識実験

4.1 実環境データの収録

発話者位置固定と発話者移動という2通りの条件で実環境データを収録した．実環境データの収録に用いた部屋は残響時間約 0.18 秒の音響実験室である．実験室内には計算機や空調による周囲雑音が存在している．発話者位置固定時と発話者移動時の発話者，白色ガウス雑音源，マイクロホンアレーの配置を図6と図7に示す．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る．ここで，各マイクロホンは無指向性である．発話者と白色ガウス雑音源としてスピーカ（JBL Control 5 Plus）を用いており，各スピーカからマイクロホンアレーまでの距離は約 2 m である．発話者と白色ガウス雑音源の各音源位置でのパワーの比を Clean，20 dB，10 dB となるようにゲインを調節している．実際の受音信号の SNR は，計算機や空調による周囲雑音の影響を受けて各々21 dB，18 dB，10 dB となっている．発話者位置固定の場合，発話者の方向は真正面 90°，白色ガウス雑音源の方向は 40° である．発話者移動の場合，発話者は1単語を発声する間に 70° 近辺から移動を開始し，140° 近辺で停止する．また，白色ガウス雑音源の方向は 40° である．なお，発話者移動の際，スピーカの中心を人間の口の高さになるようにし，スピーカの放射面をマイクロホンアレーに向けながら，人間の歩行とともに移動させている．

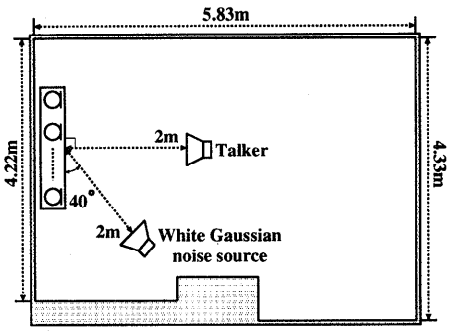


図6 発話者位置固定時の発話者，白色ガウス雑音源，マイクロホンアレーの配置
Fig. 6 Arrangement of a fixed-position talker, a white Gaussian noise source, and a microphone array.

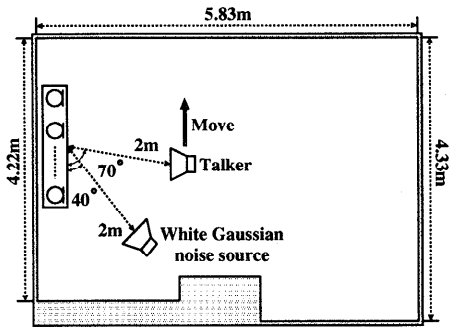


図7 発話者移動時の発話者，白色ガウス雑音源，マイクロホンアレーの配置
Fig. 7 Arrangement of a moving talker, a white Gaussian noise source, and a microphone array.

4.2 AMNOR の基本性能の評価

本節では，発話者の方向が既知であるという条件で AMNOR を図1の音声認識システムに適用する．そして，4.1 節で述べた実環境データを用いて認識実験を行うことにより，その基本性能を評価する．

4.2.1 実験条件

音声認識部の仕様を**表1**に示す．標本化周波数は 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．フレーム周期は 8 msec である．そして，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．特徴ベクトルとしてメ

表1 音声認識部の仕様
Table 1 Experiment conditions.

標本化周波数	12 kHz
フレーム長	32 msec
フレーム周期	8 msec
高域強調	$1-0.97z^{-1}$
特徴ベクトル	MFCC16次，Δ MFCC16次 Δパワー1次
HMM	Tied-mixture型HMM (混合数256, 256, 128)
モデル	音素環境独立54モデル
データベース	ATR音声データベースSetA
学習データ	MHT重要語から2620単語
テストデータ	MHT音韻バランス216単語

ルケプストラム係数（MFCC: Mel Frequency Cepstrum Coefficients）16次，Δ MFCC 16次，Δパワー1次を計算する．音声認識にはTied-mixture分布型HMM[11]を用いる．ここで，混合数はMFCCとΔ MFCCについて256，Δパワーについて128である．音素環境独立の54音素モデルをATR音声データベースSetAの話者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．テストデータは話者MHTの音韻バランス216単語である．

本論文では，事前に収録した雑音信号を用いてオフラインでAMNORのフィルタ係数を計算している．ここで，目的信号としては，平面波音場において白色ガウス雑音が目的方向から到来する状況を考え，計算機シミュレーションで生成したマイクロホンアレー受音信号を用いている．また，雑音信号としては，

(1) 計算機や空調による周囲雑音のみが存在する場合
(2) 白色ガウス雑音源と(1)の周囲雑音がともに存在する場合

に収録した2通りのマイクロホンアレー受音信号を用いている．なお，チャネルごとのフィルタ長は52である．

実環境で収録した音声信号にはスピーカやマイクロホンの周波数特性，部屋の伝達関数などによる乗法性歪みがかかるので，音声認識の精度は著しく低下してしまう．マイクロホンアレーの指向性ビームが十分鋭ければ，残響についてはある程度抑圧することができる．しかし，抑圧しきれない残響成分，スピーカやマイクロホンの周波数特性については何らかの対処が必要である．本論文では，この問題に対処するために，ケプストラム平均正規化法（CMN: Cepstrum Mean Normalization）[12]を用いる．一般に，CMNは式(7)で表される．

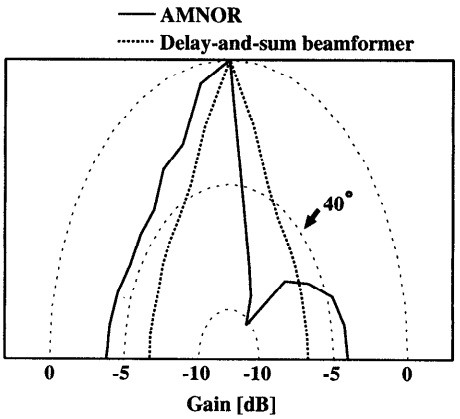


図8 AMNORと遅延和アレーの指向性
Fig. 8 Directive gain patterns obtained by AMNOR and the delay-and-sum beamformer.

$$\hat{\mathbf{x}}(n)=\mathbf{x}(n)-\frac{1}{L}\sum_{l=0}^{L-1}\mathbf{x}(l) \tag{7}$$

ここで，$\mathbf{x}$はメルケプストラム係数ベクトル，$\hat{\mathbf{x}}$は正規化された係数ベクトル，nはフレーム番号，Lは1単語内のフレーム数である．本論文では，音声と非音声の区別をせずに単語ごとにすべてのフレームを用いて平均ベクトルを計算している．

4.2.2 AMNORの指向性

AMNORと遅延和アレーの指向性を**図8**に示す．ここで，AMNORのフィルタ係数は4.2.1項で述べた(2)の場合に計算したものであり，レスポンス劣化量を0.0としている．遅延和アレーの場合，90°にピークを持つ超指向性を形成している．一方，AMNORの場合，白色ガウス雑音源を設置している40°方向に死角を形成しており，遅延和アレーよりも減衰量が大きいことが分かる．

4.2.3 認識実験

発話者位置固定時にCMNを用いない場合とCMNを用いる場合の単語認識率を**表2**と**表3**に示す．表中の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．ここで，表3ではシングルマイクにもCMNを用いている．遅延和アレーとAMNORは，発話者の方向が既知であるという条件で正しい発話者方向に指向性ビームを向ける場合である．SNR 21 dBのときには白色ガウス雑音源を設置しておらず，計算機や空調による周囲雑音のみが存在している．SNR 18 dB，SNR 10 dBのとき，白色ガウス雑音源と周囲雑音がともに存在している．AMNORのフィルタ係数は，SNR 21 dBのとき4.2.1項で述べた(1)の場合，SNR 18 dBとSNR 10 dBの

表2 発話者位置固定時の単語認識率 [%]（CMNを用いない場合）
Table 2 Word recognition accuracy [%] for the fixed-position talker (without CMN).

	SNR [dB]		
	21	18	10
シングルマイク	81.0	66.6	17.5
遅延和アレー	81.0	81.9	66.6
AMNOR	89.3	87.9	83.7

表3 発話者位置固定時の単語認識率 [%]（CMNを用いる場合）
Table 3 Word recognition accuracy [%] for the fixed-position talker (with CMN).

	SNR [dB]		
	21	18	10
シングルマイク	89.8	76.8	37.0
遅延和アレー	92.1	86.5	75.0
AMNOR	94.4	91.2	89.3

とき(2)の場合に計算したものであり，レスポンス劣化量を0.0としている．実験結果を以下にまとめる．

- CMNを用いない場合のAMNORの単語認識率は，遅延和アレーと比べてSNR 21 dBのとき8.3%，SNR 18 dBのとき6.0%，SNR 10 dBのとき17.1%改善されている．特に，SNR 21 dBのときの単語認識率は，遅延和アレーではシングルマイクと同じであるが，AMNORでは周囲雑音や反射音などを効果的に抑圧しているのでさらに改善がみられる．
- CMNを用いる場合のAMNORの単語認識率は，CMNを用いない場合と比べてSNR 21 dBのとき5.1%，SNR 18 dBのとき3.3%，SNR 10 dBのとき5.6%改善されている．

以上から，AMNORとCMNの併用が非常に有効であることが分かる．

AMNORのレスポンス劣化量が単語認識率に与える影響について調べる．CMNを用いない場合のレスポンス劣化量と単語認識率の関係を**図9**に示す．レスポンス劣化量が0.01のとき単語認識率に多少の改善がみられる．よって，式(6)の拘束条件が音声認識においても有効であることが分かる．一方，レスポンス劣化量を大きくする，すなわち音声の劣化が大きくなるにつれて，単語認識率は著しく低下することが分かる．レスポンス劣化量とフィルタの周波数特性の関係を**図10**に示す．レスポンス劣化量を大きくするにつれて，0～1 kHzと3 kHz近辺での歪みが大きくなることが分かる．この歪みが単語認識率の低下の主な原因である．レスポンス劣化量を大きくするときに生じるフィルタの周波数特性の歪みは，マイクロホンやスピーカの周波数特性と同様にCMNにより補正できると考えられる．CMNを用いる場合のレスポンス劣化量と単語認識率の関係を**図11**に示す．CMNを用いる場合の単語認識率は，CMNを用いない場合と比べて全体的に高くなっている．また，レスポンス劣化量が0.01のときの単語認識率に明確な改善はみられなくなっているものの，レスポンス劣化量を大きくしても単語認識率の低下はかなり緩やかになっていることが分かる．

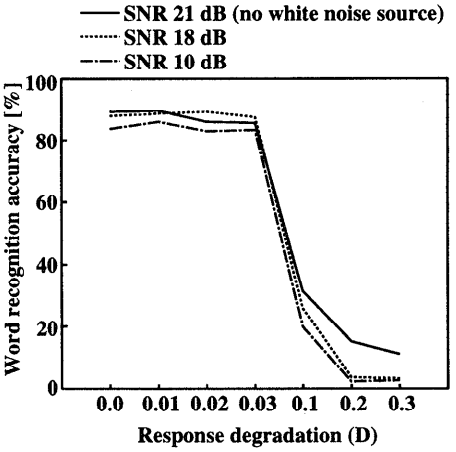


図9 レスポンス劣化量と単語認識率の関係（CMNを用いない場合）
Fig. 9 A relationship between response degradation and word recognition accuracy (without CMN).

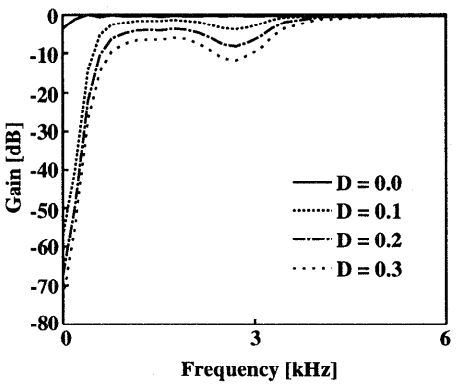


図10 レスポンス劣化量とフィルタの周波数特性の関係
Fig. 10 A relationship between response degradation and frequency response of AMNOR.

4.3 AMNORの3次元ビタビ探索法への適用

本節では，発話者の方向が未知であるという条件でAMNORを3次元ビタビ探索法に適用する．そして，4.1節で述べた実環境データを用いて認識実験を行うことにより，その性能を評価する．

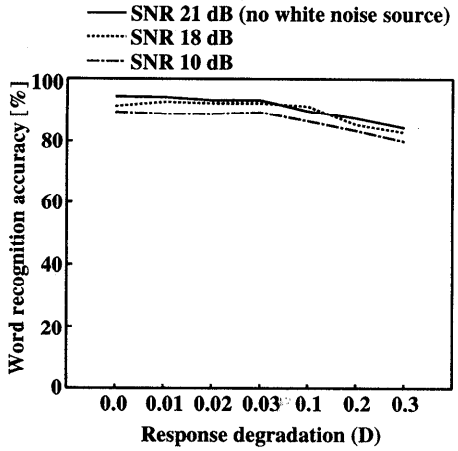


図11 レスポンス劣化量と単語認識率の関係（CMNを用いる場合）

Fig. 11 A relationship between response degradation and word recognition accuracy (with CMN).

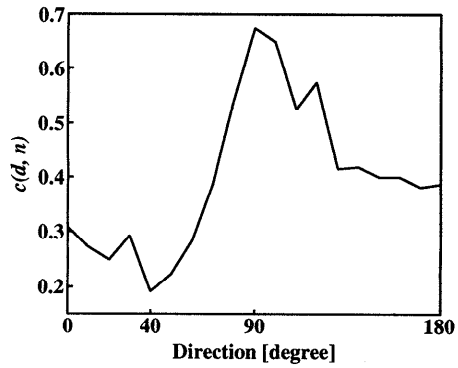


図12 音素/i/の1フレームにおける $c(d,n)$ の計算例

Fig. 12 An example of $c(d,n)$ in a frame of the phone /i/.

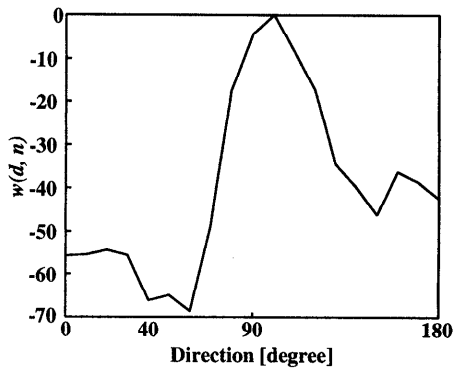


図13 音素/i/の1フレームにおける $w(d,n)$ の計算例（$\mu = 80$, $\nu = 10$）

Fig. 13 An example of $w(d,n)$ in a frame of the phone /i/, where $\mu = 80$ and $\nu = 10$.

4.3.1 実験条件

音声認識部は4.2.1項で述べたものと同様である．3次元ビタビ探索法における特徴ベクトルの方向・フレーム系列の計算では10°ごとに0°～180°を対象としており，$0°, 10°, \cdots, 180°$ から到来する目的信号を各々生成してAMNORのフィルタ係数を計算している．

3次元ビタビ探索法では特徴ベクトルの方向・フレーム系列を入力としているので，CMNにおける平均ベクトルの求め方として様々な方法が考えられる．本論文では，式(8)に示すように方向ごとに平均ベクトルを計算することにする．

$$\hat{\mathbf{x}}(d,n) = \mathbf{x}(d,n) - \frac{1}{L}\sum_{l=0}^{L-1}\mathbf{x}(d,l) \qquad (8)$$

ここで，$\mathbf{x}$ はメルケプストラム係数ベクトル，$\hat{\mathbf{x}}$ は正規化された係数ベクトル，d は方向，n はフレーム番号，L は1単語内のフレーム数である．

最後に，式(3)の重み関数について説明する．μ と ν の値については，様々な値の組合せで認識実験を行い，単語認識率が最も高かった組合せを採用している．発話者位置固定時に，音素/i/の1フレームにおいて $\mu = 80$, $\nu = 10$ として計算した $c(d,n)$ と $w(d,n)$ の例を**図12**と**図13**に示す．ここで，SNRは18 dBであり，遅延和アレーを適用している．図12と図13より，発話者の方向である90°で $c(d,n)$ が大きくなり，その結果 $w(d,n)$ においても90°に重みがかかることが分かる．

4.3.2 認識実験

発話者位置固定時と発話者移動時にCMNを用いる場合の3次元ビタビ探索法の単語認識率を**表4**と**表5**に示す．表中の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．ここで，表4と表5ともにシングルマイクでもCMNを用いている．3次元ビタビ探索法（遅延和アレー）は，特徴ベクトルの方向・フレーム系列の計算に遅延和アレーを適用していることを表す．同様に，3次元ビタビ探索法（AMNOR）はAMNORを適用していることを表す．重み関数は，式(3)の重み関数の有無とパラメータを表している．AMNORのフィルタ係数はレスポンス劣化量を0.0として計算している．発話者位置固定時と発話者移動時の実験結果には同様の傾向がみられるので，ここでは表5の発話者移動時の実験結果を以下にまとめる．

- 3次元ビタビ探索法（AMNOR）の重み関数を用いる場合の単語認識率は，シングルマイクと比べてSNR 18 dBのとき11.1%，SNR 10 dBのとき

表4 発話者位置固定時の3次元ビタビ探索法の単語認識率 [%]（CMNを用いる場合）
Table 4 Word recognition accuracy [%] for the fixed-position talker obtained by the 3-D Viterbi search method (with CMN).

	重み関数	SNR [dB]		
		21	18	10
シングルマイク		89.8	76.8	37.0
3次元ビタビ探索法（遅延和アレー）	なし	89.3	52.7	13.8
3次元ビタビ探索法（遅延和アレー）	あり（$\mu = 80, \nu = 10$）	92.5	79.1	53.2
3次元ビタビ探索法（AMNOR）	なし	93.5	88.8	81.0
3次元ビタビ探索法（AMNOR）	あり（$\mu = 80, \nu = 10$）	93.9	89.8	83.3

表5 発話者移動時の3次元ビタビ探索法の単語認識率 [%]（CMNを用いる場合）
Table 5 Word recognition accuracy [%] for the moving talker obtained by the 3-D Viterbi search method (with CMN).

	重み関数	SNR [dB]		
		21	18	10
シングルマイク		92.5	77.7	38.4
3次元ビタビ探索法（遅延和アレー）	なし	89.8	60.6	23.1
3次元ビタビ探索法（遅延和アレー）	あり（$\mu = 80, \nu = 10$）	89.3	81.9	52.3
3次元ビタビ探索法（AMNOR）	なし	93.0	87.5	77.3
3次元ビタビ探索法（AMNOR）	あり（$\mu = 80, \nu = 10$）	92.5	88.8	81.0

42.6% 改善されている．

- 3次元ビタビ探索法（AMNOR）の重み関数を用いる場合の単語認識率は，3次元ビタビ探索法（遅延和アレー）の重み関数を用いる場合と比べてSNR 21 dBのとき3.2%，SNR 18 dBのとき6.9%，SNR 10 dBのとき28.7%改善されている．この主な要因は，適応型アレーでは遅延和アレーよりも特に雑音源の方向に対して雑音を抑圧していることにある．
- 3次元ビタビ探索法（遅延和アレー）の重み関数を用いる場合の単語認識率は，重み関数を用いない場合と比べてSNR 18 dBのとき21.3%，SNR 10 dBのとき29.2%改善されている．この結果は，遅延和アレーでは指向性の鋭さが不十分なために，調波構造などによる発話者の方向への重み付けが必要であることを示している．
- 3次元ビタビ探索法（AMNOR）の重み関数を用いる場合の単語認識率は，重み関数を用いない場合と比べてわずかではあるが改善されている．

以上から，AMNORを3次元ビタビ探索法に適用することにより，遅延和アレーを適用する場合と比べて単語認識率を大幅に改善できることが分かる．

本論文では実験の再現性を確保するために発話者としてスピーカを使用したが，今後は実際に人間が発話しているデータを収録して評価する必要がある．また，発話者の正確な移動軌跡を測定し，発話者移動時にも発話者方向既知の実験を行いたいと考えている．さらに，様々な環境，たとえば雑音の種類や性質（定常・非定常），残響時間の長い部屋などで実験を行う必要がある．

5. おわりに

本論文では，3次元ビタビ探索法の性能をさらに改善するために適応型アレーに基づく方法を提案し，発話者移動と発話者位置固定という2通りの条件で収録した実環境データを用いて評価した．

まず，発話者の方向が既知であるという条件でAMNORを従来のマイクロホンアレーを用いた音声認識システムに適用し，発話者位置固定の実環境データを用いてその基本性能を評価した．その結果，AMNORとCMNを併用することにより，SNR 21 dBのとき94.4%，SNR 18 dBのとき91.2%，SNR 10 dBのとき89.3%の単語認識率が得られた．次に，発話者の方向が未知であるという条件でAMNORを3次元ビタビ探索法に適用し，発話者位置固定と発話者移動の実環境データを用いてその性能を調べた．その結果，発話者移動時にAMNORとCMNを併用する場合の単語認識率は，シングルマイクと比べてSNR 18 dBのとき11.1%，SNR 10 dBのとき42.6%，また遅延和アレーとCMNを併用する場合と比べてSNR 18 dBのとき6.9%，SNR 10 dBのとき28.7%改善された．また，発話者位置固定時についても同程度の単語認識率が得られた．よって，AMNORを3次元ビタビ探索法に適用することにより，遅延和アレーを適用する場合と比べて単語認識率を大幅に改善できることが分かった．

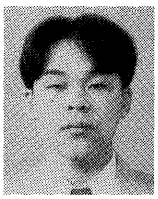
本論文では，AMNORのフィルタ係数をオフライ

ンで計算している．今後，環境の変動に対応するために，フィルタ係数をオンラインで更新する方法について検討する予定である．また，複数の発話者により同時に発声された音声を認識するために，3次元トレリスをN-bestアルゴリズムにより探索する方法について検討する予定である．

参考文献

1) Lin, Q., Jan, E., Che, C. and Vries, B.: System of microphone arrays and neural networks for robust speech recognition in multimedia environment, *ICSLP94*, pp.1247–1250 (1994).
2) Giuliani, D., Omologo, M. and Svaizer, P.: Experiments of speech recognition in a noisy and reverberant environment using a microphone array and HMM adaptation, *ICSLP96*, pp.1329–1332 (1996).
3) Yamada, T., Nakamura, S. and Shikano, K.: Robust speech recognition with speaker localization by a microphone array, *ICSLP96*, pp.1317–1320 (1996).
4) Kiyohara, K., Kaneda, Y., Takahashi, S., Nomura, H. and Kojima, J.: A microphone array system for speech recognition, *ICASSP97*, pp.215–218 (1997).
5) Hughes, T., Kim, H., DiBiase, J. and Silverman, H.: Using a real time, tracking microphone array as input to an HMM speech recognizer, *ICASSP98*, pp.249–252 (1998).
6) 山田武志，中村　哲，鹿野清宏：マイクロホンアレーによる3次元トレリス探索に基づく移動話者の音声認識，情報処理学会音声言語情報処理研究会，97-SLP-15-6, pp.35–40 (1997).
7) 山田武志，中村　哲，鹿野清宏：マイクロホンアレーによる3次元ビタビ探索に基づく移動話者の音声認識，電子情報通信学会音声研究会，SP97-22, pp.31–38 (1997).
8) Yamada, T., Nakamura, S. and Shikano, K.: Hands-free Speech Recognition Based on 3-D Viterbi Search Using a Microphone Array, *ICASSP98*, pp.245–248 (1998).
9) 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).
10) Kaneda, Y. and Ohga, J.: Adaptive microphone array system for noise reduction, *IEEE Trans. Acoustics, Speech, and Signal Processing*, Vol.34, No.6, pp.1391–1400 (1986).
11) Bellegarda, J.R. and Nahamoo, D.: Tied mixture continuous parameter models for large vocabulary isolated speech recognition, *ICASSP89*, pp.13–16 (1989).
12) Atal, B.: Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification, *J. Acoust. Soc. Am.*, Vol.55, No.6, pp.1304–1312 (1974).

(平成10年6月1日受付)
(平成10年12月7日採録)

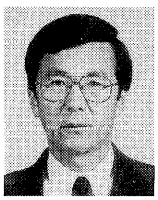


山田　武志（学生会員）

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．