

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

山田 武志 中村 哲 鹿野 清宏

奈良先端科学技術大学院大学 情報科学研究科

〒630-01 奈良県生駒市高山町 8916-5

E-Mail: {takesi-y, nakamura, shikano}@is.aist-nara.ac.jp

概要 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は、発話者方向検出の誤りが音声認識性能を劣化させるという問題がある。この問題に対処するために、発話者方向を一意に決定するのではなく、発話者方向の候補を複数残しながら音声認識を行なう方法について検討する。本稿では、その一つの実現方法として、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音声認識法を提案する。これは、マイクロホンアレーで空間を走査して得られる特徴ベクトルの空間・時間系列を入力とし、空間・時間・状態からなる3次元トレリスの探索により音声認識を行なうものである。シミュレーション実験の結果、全ての方向を対象として3次元トレリス探索を行なう場合には問題が残されているが、音声の調波構造を利用して発話者方向の候補を絞り込むことにより音声認識性能を大幅に改善できることが確認された。

Hands-Free Speech Recognition Based on 3-D Trellis Search Using a Microphone Array

Takeshi YAMADA Satoshi NAKAMURA Kiyohiro SHIKANO

Graduate School of Information Science, Nara Institute of Science and Technology

8916-5 Takayama-cho, Ikoma-shi, Nara, 630-01 JAPAN

E-Mail: {takesi-y, nakamura, shikano}@is.aist-nara.ac.jp

Abstract The performance of the conventional speech recognition systems using a microphone array will be degraded due to the errors of speaker localization. One way to solve this problem is to take account of multiple directions at the same time. This paper proposes a novel recognition algorithm based on Viterbi search on 3-D trellis space of direction, time and state. To evaluate the performance of the proposed method, the speech recognition experiments are carried out. These results show that the proposed method attains the significant improvement by introducing constraints on search space derived from pitch harmonics informations of speech.

1 はじめに

実環境下で頑健かつハンズフリーな音声認識を実現するために、マイクロホンアレーの適用が試みられている。マイクロホンアレーを用いて周囲雑音や残響などを抑圧することにより、ハンズフリーの状況でも音声認識性能の劣化を防ぐのが狙いである。

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, 2, 3]。これらのシステムでは、(1) 発話者方向を検出し、(2) その方向に感度が高い指向特性を形成する、という処理を繰り返し行なう。従って、(1) の精度が低く、雑音源を発話者として検出した場合、音声認識性能は大きく劣化してしまう。この問題に対処するための一つの方法は、発話者方向検出の精度をさらに高めることである。しかしながら、雑音源が存在する状況では、音声と雑音の性質を手がかりに発話者方向を検出せざるを得ないという問題がある。

本稿では、発話者方向を一意に決定するのではなく、発話者方向の候補を複数残しながら音声認識を行なう方法について検討する。そして、その一つの実現方法として、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音声認識法を提案する。以下、2章で提案法について詳細に説明し、3章でシミュレーション実験を行なう。最後に、4章でまとめと今後の課題について述べる。

2 提案法

2.1 アプローチ

マイクロホンアレーを用いることにより、任意の方向に感度が高い指向特性を形成することができる。従って、マイクロホンアレーで空間を走査することにより、特徴ベクトルの空間・時間系列 $\mathbf{x}(d, n)$ を得ることができる。ここで、 $\mathbf{x}(d, n)$ は、方向番号 d 、フレーム番号 n におけるメルケプストラムなどの特徴ベクトルである。

発話者と雑音源が存在する状況で、図1に示すような特徴ベクトルの空間・時間系列が得られたとする。図中の縦軸は方向番号、横軸はフレーム番号である。また、“□”は各々の方向と

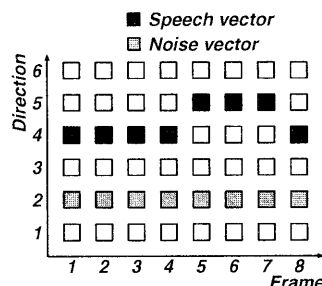


図1: 特徴ベクトルの空間・時間系列

フレームで得られた特徴ベクトルを表し、“■”は発話者の特徴ベクトル、“□”は雑音源の特徴ベクトルであることを意味する。このとき従来法では、フレーム毎に発話者方向を検出し、発話者の特徴ベクトルを一意に決定する。従って、発話者方向検出を誤った場合には発話者の特徴ベクトルが失われ、音声認識性能は大きく劣化してしまう。

本稿では、発話者方向を一意に決定するのではなく、発話者方向の候補を複数残しながら音声認識を行なう方法について検討する。そして、その一つの実現方法として、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音声認識法を提案する。通常、HMMを用いた音声認識では、特徴ベクトルの時間系列を入力とし、時間・状態からなる2次元トレリスの探索を行なう。提案法はその拡張として考えることができ、特徴ベクトルの空間・時間系列を入力とし、図2に示すような空間・時間・状態からなる3次元トレリスの探索により音声認識を行なう。提案法の最大の特徴は、発話者方向検出を明示的に行なわないことであり、全ての方向が音声認識の対象となる。

2.2 認識アルゴリズム

本稿では、3次元トレリス探索の手法としてビタビアルゴリズムを用いる。方向遷移パスをビタビアライメントにより得ることが可能なので、アルゴリズムの解析が容易になるという利点がある。具体的な認識アルゴリズムを図3に示す。ここで、 S は初期状態集合、 Q は状態数、

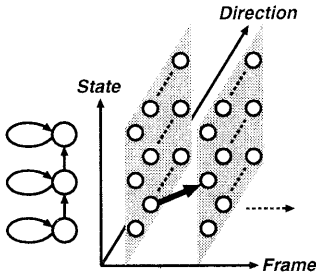


図 2: 空間・時間・状態からなる 3 次元トレリスの探索

D は方向数、 N はフレーム数である。また、 π は初期確率、 α は累積尤度、 $a(q', q)$ は状態 q' から状態 q への遷移確率、 $a(d', d)$ は方向 d' から方向 d への遷移確率、 b は出力確率を表す。

図 3 の認識アルゴリズムにおいて方向遷移確率は学習によって獲得するのが困難であるので、何らかの知識を前提にして与える必要がある。一般に、音声認識では数 msec の周期でフレーム分析を行なうので、発話者は隣り合うフレーム間でほとんど移動することはない。よって、ある一定の方向幅の中でだけ方向遷移を許すために、式 (2) に示すような方向遷移確率を設定することにする。

$$a(d', d) = \begin{cases} \frac{1}{2\Theta} & , \quad |\theta(d) - \theta(d')| \leq \Theta \\ 0 & , \quad |\theta(d) - \theta(d')| > \Theta \end{cases} \quad (2)$$

ここで、 $\theta(d)$ は方向番号 d に対応する方向を表し、 Θ は方向幅である。

図 3 の認識アルゴリズムが良好に動作する条件は、発話者方向の特徴ベクトルの尤度が他の方向より安定して高いことである。しかしながら、雑音源が存在する状況を考慮すると、この条件が満たされるという保証は得られない。このような場合には探索範囲をある程度制限することが有効であると考えられるので、音声の調波構造という情報を利用して発話者方向の候補を絞り込むことを考える。 $c(d; n)$ をフレーム番号 n で方向番号 d から得られたケプストラムのうち、高ケプレンシ部における最大値とする。この値は音声に調波構造が顕著に含まれるとき

Initialization:

```
for  $q = 1$  to  $Q$ 
  for  $d = 1$  to  $D$ 
    if  $(q, d) \in S$  then
       $\alpha(q, d, 0) = \log \pi(q, d)$ 
      , where  $\sum_{(q,d) \in S} \pi(q, d) = 1$ 
    else
       $\alpha(q, d, 0) = -\infty$ 
```

Recognition:

```
for  $n = 1$  to  $N$ 
  for  $q = 1$  to  $Q$ 
    for  $d = 1$  to  $D$ 
       $\alpha(q, d, n) =$ 
         $\max_{q', d'} \{ \alpha(q', d', n-1) +$ 
           $\log a(q', q) + \log a(d', d) \}$ 
           $+ \log b(q, x(d, n)) \}$  (1)
```

図 3: 認識アルゴリズム

に大きくなるので、音声らしさを表す尺度とみなすことができる。よって、式 (3) に示すような重みを図 3 の式 (1) の右辺に加えることにする。

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

ここで、 i と j は重みの影響を制御するパラメータである。 i は方向間での重みの差を調節するためのものである。 i を大きくすることにより、方向間での重みの差を強調することができる。また、 j はどれだけ過去のフレームを考慮に入れるかを反映するものである。発話者が移動しない場合は j を大きくし、移動する場合は j を小さくするといった調節が可能となる。

3 シミュレーション実験

3.1 実験条件

実験条件を表 1 に示す。マイクロホンアレーの受信信号については、平面波音場を仮定して計算機シミュレーションにより生成した。帯域

表 1: 実験条件

標本化周波数	12 kHz
フレーム長	32 ms (ハミング窓)
フレーム周期	8 ms
高域強調	$1 - 0.97z^{-1}$
特徴ベクトル	MFCC, Δ MFCC, Δ パワー
認識方式	Tied-Mixture 型 HMM
モデル数	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 語
テストデータ	MHT 音韻バランス単語 216 語
素子数	14
素子間隔	2.83 cm
信号処理	遅延和アレー
空間分解能	$10^\circ (0^\circ, 10^\circ, \dots, 180^\circ)$

6 kHz の白色ガウス雑音に対する指向特性を図 4 に示しておく。

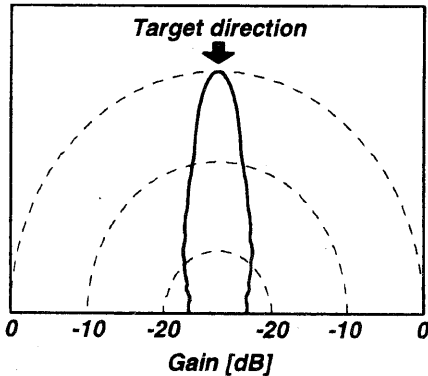


図 4: 指向特性

3.2 認識実験 1

発話者が移動しないという条件下で提案法の性能を評価する。マイクロホンアレーと音源の関係を図 5 に示す。発話者方向は真正面 90° 、雑音源方向は 40° であり、雑音源は白色ガウス雑音を使用する。

単語認識率と発話者方向検出精度を表 2 に示す。発話者方向検出精度は、発話者方向 90° を正確に検出した場合を正解とし、(正解フレーム数 / 全フレーム数) で定義する。表中の 3-D trellis search 1 ($\Theta = 10$) は全ての方向を対象として 3 次元トレリス探索を行なう場合であり、式 (2) の

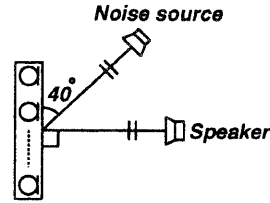


図 5: マイクロホンアレーと音源の関係

Θ を 10 としている。以下、これを提案法 1 と呼ぶ。3-D trellis search 2 ($\Theta = 10, i = 40, j = 20$) は音声の調波構造を利用して発話者方向の候補を絞り込む場合であり、式 (2) の Θ を 10、式 (3) の i と j を各々 40、20 としている。以下、これを提案法 2 と呼ぶ。

シングルマイクの場合、SNR が低下するにつれて認識率も大幅に劣化している。しかしながら、発話者方向既知という条件でマイクロホンアレーを用いることにより、90 % 以上の認識率が得られる。

提案法 1 の場合、Clean のときの発話者方向検出精度は 99.3 % であり、発話者方向既知と同等の認識率が得られる。一方、SNR が低下するにつれて発話者方向検出精度の低下が顕著となり、音声認識性能が劣化している。Clean と SNR 20 dB のときの単語/iroiro/に対する発話者方向検出結果を図 6 に示す。ここで、縦軸は

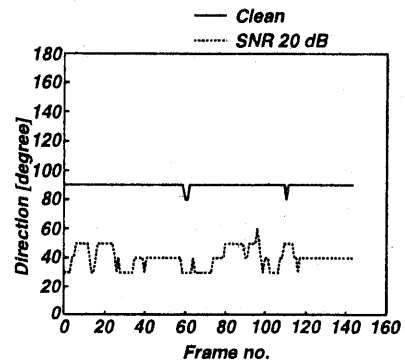


図 6: 発話者方向検出結果 (提案法 1)

発話者方向検出結果、横軸はフレーム番号である。Clean のときは発話者方向にほぼ追従して

表 2: 単語認識率 (WA) [%] と発話者方向検出精度 (SLA) [%]

	Clean		SNR 20 dB		SNR 10 dB	
	WA	SLA	WA	SLA	WA	SLA
Single microphone	96.2	—	80.0	—	25.9	—
Delay-and-sum beamformer	96.2	100.0	94.9	100.0	90.7	100.0
3-D trellis search 1 ($\Theta = 10$)	96.2	99.3	72.6	24.9	28.2	10.8
3-D trellis search 2 ($\Theta = 10, i = 40, j = 20$)	96.2	99.4	94.9	50.4	88.4	45.7

いるが、SNR 20 dB のときは雑音源方向を発話者方向として検出しているのが分かる。以上から、全ての方向を対象にして3次元トレリス探索を行なうには、尤度という情報だけでは不十分であると考えられる。

提案法2の場合、Clean と SNR が低いときの両方で発話者方向既知と同等の認識率が得られる。図7に SNR 20 dB における発話者方向検出の頻度分布を示す。ここで、横軸は方向、縦

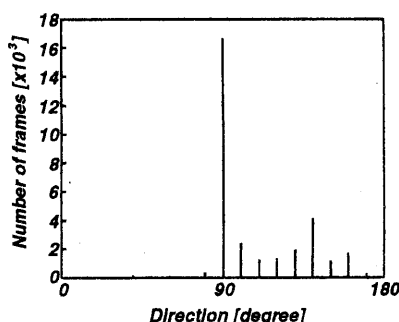


図 7: 発話者方向検出の頻度分布 (提案法2)

軸は発話者方向として検出したフレーム数を表す。図7から分かるように、発話者方向検出の誤りは $90^\circ \sim 180^\circ$ の方向に集中しており、雑音源方向への遷移は抑えられている。その結果、提案法1よりも高い音声認識性能を達成していると考えられる。

3.3 認識実験2

発話者が移動するという条件下で提案法の性能を評価する。マイクロホンアレーと音源の関係を図8に示す。発話者は1単語を発声する間に 0° 方向から 180° 方向に移動する。雑音源方向は 40° であり、白色ガウス雑音を使用する。

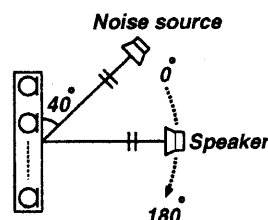


図 8: マイクロホンアレーと音源の関係

単語認識率と発話者方向検出精度を表3に示す。発話者方向検出精度は、実際の発話者方向と最も近い方向を検出した場合を正解とし、(正解フレーム数/全フレーム数)で定義する。表中の3-D trellis search 2 ($\Theta = 10, i = 40, j = 10$)では、式(2)の Θ を10、式(3)の i と j を各々40、10としている。

提案法1の場合、Clean のときの認識率は96.2%である。一方、SNR が低下するにつれて発話者方向検出精度の低下が顕著となり、音声認識性能が劣化している。Clean と SNR 20 dB のときの単語/iroiro/に対する発話者方向検出結果を図9に示す。Clean のときは発話者方向によく追従しているが、SNR 20 dB のときは雑音源方向を発話者方向として検出しているのが分かる。

提案法2の場合、Clean のときの認識率は96.7%であり、SNR 20 dB と SNR 10 dB のときの認識率は提案法1と比べて各々19.4%、57.9%改善されている。SNR 20 dB のときの単語/iroiro/に対する発話者方向検出結果を図10に示す。提案法2では、音声区間(フレーム番号33~109)において提案法1よりも発話者方向にうまく追従しており、雑音源方向への遷移が抑えられて

表 3: 単語認識率 (WA) [%] と発話者方向検出精度 (SLA) [%]

	Clean		SNR 20 dB		SNR 10 dB	
	WA	SLA	WA	SLA	WA	SLA
3-D trellis search 1 ($\Theta = 10$)	96.2	76.4	74.5	32.3	26.8	15.5
3-D trellis search 2 ($\Theta = 10, i = 40, j = 10$)	96.7	72.0	93.9	41.5	84.7	33.7

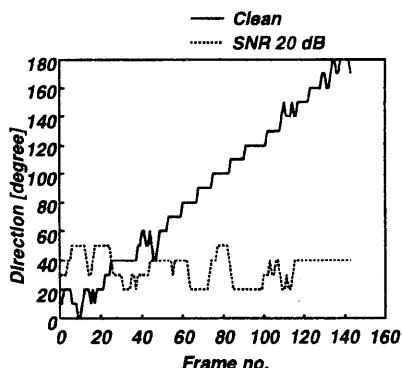


図 9: 発話者方向検出結果 (提案法 1)

いるのが分かる。その結果、提案法 1 よりも高い音声認識性能を達成していると考えられる。

4 おわりに

本稿では、発話者方向を一意に決定するのではなく、発話者方向の候補を複数残しながら音声認識を行なう方法について検討した。そして、その一つの実現方法として、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音声認識法を提案し、シミュレーション実験により性能を評価した。その結果、全ての方向を対象として 3 次元トレリス探索を行なう場合には問題が残されているが、音声の調波構造を利用して発話者方向の候補を絞り込むことにより音声認識性能を大幅に改善できることが確認された。

今後は、提案法の挙動をさらに詳細に分析するために、方向分解能の向上、様々な雑音に対する実験などを予定している。また、今回は 3 次元トレリス探索の手法としてビタビアルゴリズムを用いたが、N-Best 探索アルゴリズムを用いて複数の話者が存在する状況で同時に音声認識を行なう方法についても検討していきたい。

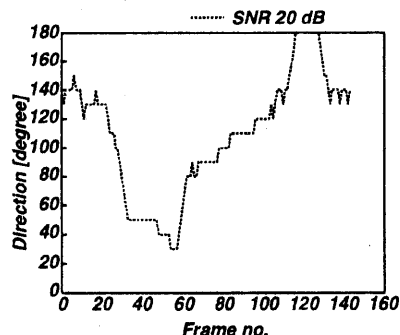


図 10: 発話者方向検出結果 (提案法 2)

参考文献

- [1] D. Giuliani, M. Omologo, and P. Svaizer, "Talker Localization and Speech Recognition Using a Microphone Array and a Cross-Powerspectrum Phase Analysis", Proc. ICSLP94, S22-1, pp. 1243-1246, Sep. 1994.
- [2] Qiguang Lin, Ea-Ee Jan, ChiWei Che, Bert de Vries, "System of Microphone Arrays and Neural Networks for Robust Speech Recognition in Multimedia Environment", Proc. ICSLP94, S22-2, pp. 1247-1250, Sep. 1994.
- [3] T. Yamada, S. Nakamura, K. Shikano, "Robust Speech Recognition with Speaker Localization by a Microphone Array", Proc. ICSLP96, pp. 1317-1320, Oct. 1996.