99-01077

マイクロホンアレーを用いた環境音の認識に関する研究

山 田 武 志　　筑波大学講師

1　研究の背景

我々が暮らしている実世界には多種多様な音が存在している。このような環境下で所望の音を抽出・認識するためには、その音に関する知識はもちろんのこと、それ以外の音に関する知識が必要不可欠である。例えば、人間の場合、その環境に存在する個々の音を（意識する・しないに関わらず）全て認識し、所望の音を知識レベルで聞き分けていると考えられる。よって、人間と同様の頑健な処理を実現するためには、その環境に存在する個々の音を全て認識する必要がある。このように周囲の音環境を認識することにより、雑音下での音声通話系の品質改善、音声認識の精度向上、自律移動ロボットにおける正確な状況把握などに貢献できると考えられる。

従来、雑音下での音声通話系の品質改善や音声認識の精度向上のために様々な研究がなされている。これらの研究のほとんどは、対象とする音声以外の音を一律的に雑音とみなすというアプローチをとっている。しかし、実世界に存在する多種多様な音を一括りにして扱うことには無理があり、極めて限られた環境でしか効果を得ることができない。また、音声と非音声を区別するための明確な基準が存在しない、複数の話者が同時に発話している場合に対象とする話者を予め判断するのは難しいという問題がある。このような問題を解決するためには、個々の音を全て認識し、所望の音の判断を上位の知識レベルに委ねるというアプローチをとる必要があると考えられる。

2　研究の目的

本研究の目的は環境音の認識を実現することであり、特に次の要素技術について研究を行う。

1.　個々の音を抽出・認識する技術

マイクロホンアレーによる音源抽出法を適用する。マイクロホンアレーを用いた多チャネルの信号処理では空間的な指向性を形成して音を抽出するので、音の種類や定常性・非定常性によらず安定した性能を得ることができる。しかし、音源がどこにあるのかは未知であり、また音源が移動するということも十分起こり得る。よって、複数の音源の位置を同時に推定し、かつ個々の 音源の移動を追尾しながら認識を行う方法について研究を行う。

2.　個々の音をモデル化する技術

音声認識で用いられている隠れマルコフモデルに基づく方法を適用する。音声の場合、いわゆる音素を単位としてモデル化する方法が確立している。しかし、音声以外の音（環境音）の場合、その多様性のためにどのような単位でモデル化すればよいのか非常に曖昧である。よって、音声認識に適した形で環境音をモデル化する方法について研究を行う。

3　研究の実施概要

3.1　個々の音を抽出・認識する技術

報告者らが提案している3次元トレリス法（3次元ビタビ法［1]、3次元N-best法［2]）を適用することを考える。本手法では、マイクロホンアレーの指向性ビームをフレーム毎に対象とする全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次元トレリス上で尤度の高いパスを探索することにより、音源の移動軌跡の推定と音源の認識を同時に行う。本手法により複数の話者の音声を同時に認識できることを示しているものの、その性能はマイクロホンアレーの指向性ビームの鋭さに強く依存するという問題がある。

マイクロホンアレーの指向性ビームの鋭さが十分ではない場合、ある方向から抽出した音源に他の方向からの音源が重畳するという問題が生じる。この問題に対処するための一つの方法は、複数の音源が重畳している区間と重畳してい

る音源の種類を事前に検出しておき、その区間に対して重畳を考慮したモデルを用いることである。

本研究では、まず音声に環境音が重畳している状況を想定し、重畳している区間と重畳している環境音の種類を検出する方法として、環境音モデルとHMM合成［3］による音声区間検出法を考案した。本手法では、音声と環境音のモデルを用いてビタビアライメントを求め、音声に重畳している環境音を予測する。そして、音声と予測した環境音の重畳モデルをHMM合成により作成し、この重畳モデルを加えて再度ビタビアライメントを求める。その結果、音声と環境音が重畳している区間、重畳している環境音とそのSN比を検出できる。種々の環境音を用いて評価実験を行った結果、本手法の有効性を確認することができた。詳細については4章で改めて述べる。以上の研究成果を2000年12月の第2回音声言語シンポジウム［4］、2001年3月の日本音響学会2001年春季研究発表会［5］、2001年4月のWorkshop on Hands-free Speech Communication［6］で発表した。

今後、以上の枠組みを環境音同士の重畳の場合に拡張し、3次元トレリス法に導入する方法を検討する予定である。その際には、アルゴリズムの複雑化、計算量の増加、探索自由度の増加などが懸念されるが、2段階探索の枠組みにおいて対処することを考えていく予定である。

3.2 個々の音をモデル化する技術

隠れマルコフモデル（HMM）により環境音をモデル化するにあたって、モデルの単位、モデルの構造、特徴量などについて検討する必要がある。本研究では、まずモデルの単位について検討を行った。音声の場合、音声器官や音声の意味付けなどの制約により、モデルの単位を決めるのは比較的容易である。一般には、音素、音節、単語などの単位が良く用いられている。一方、環境音の場合、その多種多様性と意味付けの曖昧さのためにモデルの単位を決めるのは非常に困難である。

環境音モデルの単位を決めるためには、何らかの基準に基づいて環境音をクラスタリングする必要がある。代表的な基準としては、ヒューリスティックな意味付けに基づくもの、特徴量の類似度に基づくもの、尤度最大化に基づくものなどがある。また、クラスタリングの方針としては、ボトムアップに統合していくアプローチとトップダウンに分割していくアプローチがある。これらの基準や方針は、音声の場合のトライフォンモデルのクラスタリングなどに適用されて一定の成果を収めているものの、環境音に対してはほとんど検討されていない。

本研究では、尤度最大化基準を用いたボトムアップクラスタリングを適用し、その有効性について検討した。まず、全ての環境音を1状態の混合分布型HMMでモデル化する。次に、尤度最大という意味において環境音モデル間の類似度を全ての組合せについて求め、最も類似しているものを統合する。以上の処理をモデル数が2になるまで繰返し行う。そして、クラスタリングの各段階での認識精度を求め、それが最大になるクラスタリング結果を環境音モデルの単位として採用する。

RWCP実環境音声・音響データベース［7］に含まれる約100種類の環境音に対して評価実験を行った。その結果、クラスリングの初期段階では音響的に似ているものが統合されているものの、それ以降の段階では認識精度が減少していく傾向があることが分かった。その主な原因は、一律的に1状態の混合分布型HMMでモデル化していることにあると考えられる。つまり、環境音によって音の時間的な構造が大きく異なるために、クラスタリングの段階が進むにつれて、モデルの精密さが不十分になったからであると考えられる。今後、さらに原因を調査すると共に、逐次状態分割法［8］などを適用することにより、モデルの構造（状態数や状態遷移など）を十分に考慮しながらクラスタリングを行う方法について検討する予定である。

4 環境音モデルとHMM合成による音声区間検出法

4.1 はじめに

音声に環境音が重畳している状況において、重畳している区間と重畳している環境音の種類を検出する方法として、環境音モデルとHMM合成による音声区間検出法を提案する。また、種々の環境音を用いて評価実験を行い、提案法の有効性を検証する。

4.2 提案法

音声と環境音の重畳パターンのうち典型的なものを図1に示す。

図中の（A）では音声区間の前後で環境音が重畳しており、（B）では音声区間を覆うように環境音が重畳している。また、（C）では音声区間の中で環境音が重畳している。ここで、（A）と（B）には環境音が単独で存在する区間があることに着目する。このような区間は音声モデルと環境音モデルを用いてビタビアライメントを求める（以下では従来法と呼ぶ）ことにより、比較的容易に検出できる。このとき、音声の直前（直後）の環境音が音声に重畳していると予測することを考える。音声と予測した環境音の重畳モデルを作成し、この重畳モデルを加えて再度ビタビアライメントを

求めることにより、音声と環境音が重畳している区間、重畳している環境音とそのSN比を検出できると考えられる。
提案法の詳細なアルゴリズムは次の通りである。

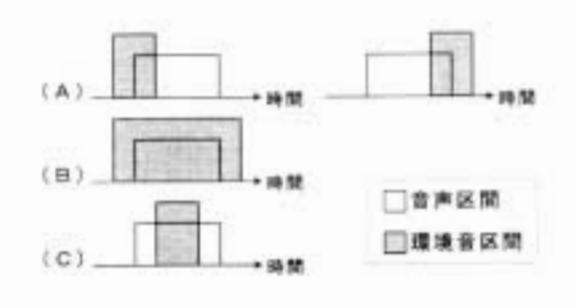

図1　音声と環境音の重畳パターン

Step 1
音声モデルと環境音モデルを図2のように連結し、入力信号のビタビアライメントを求める（従来法に相当する）。図中のN_1とN_2は環境音モデル、Sは音声モデルを表しており、簡単化のために環境音モデルの数は2、HMMの状態数は1としている。

Step 2
Step 1で音声区間の直前（直後）に検出された環境音と音声の重畳モデルをHMM合成により作成する。その際、あらかじめ設定したSN比に応じて数通りの重畳モデルを作成する。

Step 3
音声モデル、環境音モデル、Step 2で作成した重畳モデルを図3のように連結し、再度入力信号のビタビアライメントを求める。図中のS'は重畳モデルを表しており、簡単化のためにSN比は一通りとしている。

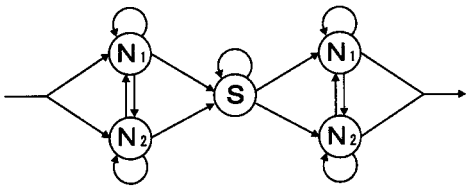

図2　音声モデルと環境音モデルの連結

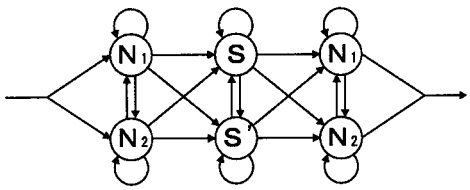

図3　音声モデル、環境音モデル、重畳モデルの連結

提案法では、各環境音に対する重畳モデルをあらかじめ複数用意するのではなく、重畳している環境音をその都度予測するので、組合せ爆発による探索の効率や精度の低下を防ぐことができると考えられる。

4.3　実験条件

評価用データは、環境音と単語音声を計算機上で加算することにより作成している。ここで、環境音はRWCP実環境音声・音響データベースから継続時間の比較的長い9つの環境音（candybowl：金属箱を金属棒で叩く音、clock1：時計のベルの音、cymbals：シンバルの音、pan：鍋を金属棒で叩く音、pipong：電子音、spray：スプレーの噴射音、toy：ぜんまいの音、trashbox：ゴミ箱を金属棒で叩く音、whistle1：ホイッスルの音）を使用し、単語音声は電総研単語音声データベースから男性1人の492個の単語データを使用している。その際、SN 比が20、10、0 dB となるように環境音の信号レベルを調整している。その結果、環境音は音声区間の前で単独で存在し、かつ音声区間の一部あるいは全てと重畳している（図1の（A）と（B）を参照）。また、サンプリング周波数を16 kHz、フレーム長を25msec（ハミング窓）、フレーム周期を10msec とし、特徴量として12次元のメルケプストラム係数を求めている。音声と無音のモデルは状態数1、混合分布数64であり、環境音のモデルは状態数1、混合分布数16 である。

4.4 実験結果と考察

図4に従来法と提案法による音声区間検出例を示す。

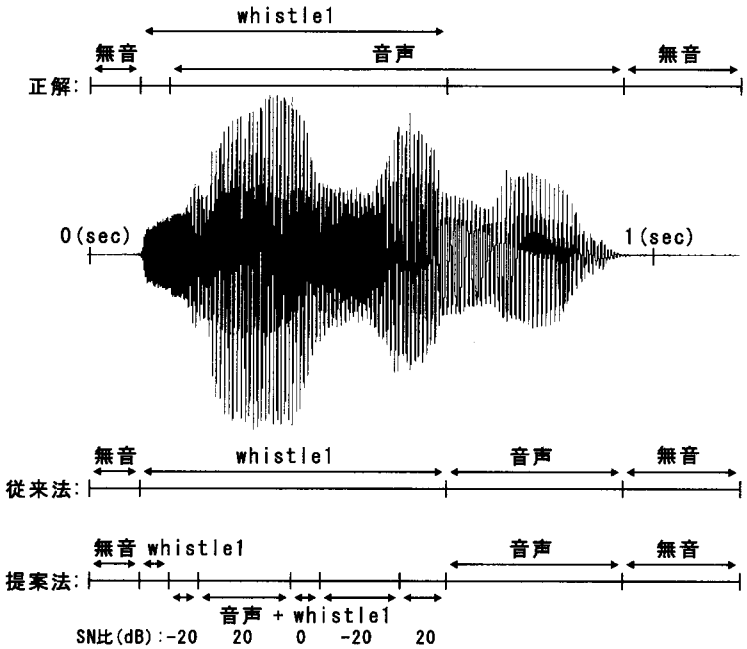

図4　従来法と提案法による音声区間検出例

図中の波形は「るいぎご」という音声に環境音（whistle1）がSN比0dBで重畳しているときのものである。波形の上部には正解区間を示しており、音声開始フレームから音声の中ほどまでwhistle1が重畳していることが分かる。波形の下部には従来法と提案法により検出された区間を示している。従来法では音声と環境音が重畳している区間をWhistle1と誤って検出しているが、提案法では音声とWhistle1の重畳区間として検出できていることが分かる。

表1に音声区間検出率（括弧内は重畳環境音検出率）を示す。

表1　音声区間検出率［%］（括弧内は重畳環境音検出率［%］）

重畳環境音	SN比20 dB			SN比10 dB			SN比0 dB		
	従来法	提案法	提案法(既知)	従来法	提案法	提案法(既知)	従来法	提案法	提案法(既知)
candybowl	86.8	87.4 (0.0)	87.2 (100.0)	80.5	82.5 (0.0)	80.7 (100.0)	67.9	75.2 (0.0)	67.9 (100.0)
clock1	72.5	88.2 (73.8)	93.0 (100.0)	42.0	73.9 (78.6)	78.1 (100.0)	17.5	42.1 (63.4)	60.8 (100.0)
cymbals	38.8	32.5 (0.0)	30.9 (100.0)	6.9	5.9 (0.0)	5.9 (100.0)	1.2	1.2 (0.0)	1.8 (100.0)
pan	40.4	51.4 (0.0)	40.4 (100.0)	47.0	56.6 (0.0)	47.2 (100.0)	69.1	70.9 (0.0)	69.1 (100.0)
pipong	75.4	87.2 (57.5)	95.9 (100.0)	57.6	81.0 (59.1)	91.6 (100.0)	25.0	60.7 (61.2)	74.6 (100.0)
spray	87.8	86.8 (3.4)	24.8 (100.0)	49.3	46.5 (14.6)	0.4 (100.0)	1.6	1.6 (75.4)	0.2 (100.0)
toy	53.2	66.1 (68.3)	63.2 (100.0)	35.9	67.2 (73.3)	83.5 (100.0)	27.2	69.7 (87.4)	75.2 (100.0)
trashbox	75.0	77.4 (76.0)	85.6 (100.0)	61.3	72.9 (78.6)	85.6 (100.0)	37.0	62.4 (86.2)	74.0 (100.0)
whistle1	84.9	79.8 (94.1)	81.3 (100.0)	71.3	92.5 (99.0)	93.1 (100.0)	33.3	86.1 (97.8)	87.6 (100.0)

ここで、提案法（既知）は提案法において音声に重畳している環境音が既知である場合を示しており、提案法のStep 2において正しい環境音を用いることに相当する。音声区間検出率は音声開始フレームの検出に成功した単語数が全単語数に占める割合であり、音声データベースに添付されている音声区間ラベルを正解とし、±5フレームの誤差を許容している。重畳環境音検出率は音声に重畳している環境音を正しく検出した単語数が全単語数に占める割合である。

表1より、提案法の音声区間検出率は従来法と比べて数%から最大で50%程度改善していることが分かる。特に、SN比が低いときほど従来法に対する改善量は大きくなっている。また、提案法の重畳環境音検出率は環境音により大きな開きがあることが分かる。よって、音声に重畳している環境音の予測方法に何らかの改良が必要であると考えられる。しかし、提案法と提案法（既知）の音声区間検出率を比べると、提案法（既知）の方が低い場合がある。これは、重畳環境音検出率の改善が必ずしも音声区間検出率の改善につながっていないことを意味している。このことを詳しく調べるために、図5に音声区間検出率と重畳環境音検出率の関係を示す。

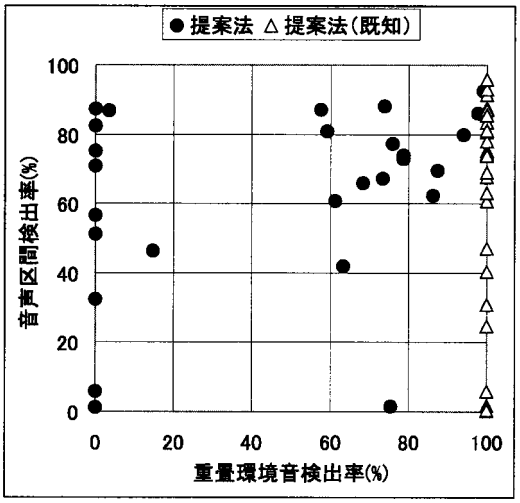

図5　音声区間検出率と重畳環境音検出率の関係

図中の●は提案法、△は提案法（既知）である。図5より、重畳環境音検出率が高くなるほど音声区間検出率も高くなる傾向が見られる。しかし、重畳環境音検出率が高い場合に音声区間検出率が低くなっていたり、重畳環境音検出率が低い場合に音声区間検出率が高くなっていることがある。前者の主な理由は、環境音が単独で存在している区間を音声と環境音が重畳している区間として誤って検出していることにある。よって、HMM合成の際のSN比の設定方法などについて検討が必要である。一方、後者の主な理由は、音声に重畳していると誤って予測した環境音が実際に重畳している環境音と音響的に似ていることにある。よって、環境音を精密にモデル化する方法（モデルの構造や単位）などについて検討が必要である。

4.5　まとめと今後の課題

音声に環境音が重畳している状況において、重畳している区間と重畳している環境音の種類を検出する方法として、環境音モデルとHMM合成による音声区間検出法を提案し、シミュレーション実験によりその有効性を示した。今後、音声に重畳している環境音の予測方法、環境音の精密なモデル化、3次元トレリス法への導入などについて研究を進める予定である。

参考文献

[1] T. Yamada, S. Nakamura, K. Shikano, “ Hands-free speech recognition based on 3-D Viterbi search using a microphone array, ” Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing,1998.

[2] P. Heracleous, T. Yamada, S. Nakamura, K. Shikano, “ Simultaneous recognition of multiple sound sources based on 3-D N-best search using microphone array, ” Proc. European Conference on Speech Communication and Technology,1999.

[3] F. Martin, K. Shikano, Y. Minami, “ Recognition of noisy speech by composition of speech and noise, ” Proc.

European Conference on Speech Communication and Technology,1993.
[4] 渡部生聖, 山田武志, 浅野太, 北脇信彦, “ 環境音モデルとHMM合成を用いた音声区間検出の検討, ” 第2回音声言語シンポジウム,2000.
[5] 渡部生聖, 山田武志, 浅野太, 北脇信彦, “ 環境音モデルとHMM合成による音声区間検出法, ” 日本音響学会2001年春季研究発表会,2001.
[6] T. Yamada, N. Watanabe, F. Asano, N. Kitawaki, “ Voice activity detection using non-speech models and HMM composition, ” Proc. Workshop on Hands-free Speech Communication,2001.
[7] S. Nakamura, K. Hiyane, F. Asano, T. Nishiura, T. Yamada, “ Acoustical sound database in real environments for sound scene understanding and hands-free speech recognition, ” Proc. International Conference on Language Resources and Evaluation,2000.
[8] 鷹見淳一, 嵯峨山茂樹, “ 逐次状態分割法による隠れマルコフ網の自動生成, ” 信学論, Vol. J76-D-II, No. 10,1993.

＜ 発　表　資　料 ＞

題　　名	掲載誌・学会名等	発表年月
Voice activity detection using non-speech models and HMMcomposition	Proc.Workshop on Hands-free Speech Communication	2001年４月
環境音モデルとHMM合成による音声区間検出法	日本音響学会2001年春季研究発表会	2001年３月
環境音モデルとHMM合成を用いた音声区間検出の検討	第2回音声言語シンポジウム	2000年12月