様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成 ２７ 年　 ６ 月 ２５ 日現在

機関番号：３４４０７
研究種目：若手研究(B)
研究期間：2012～2014
課題番号：２４７００１６９

研究課題名（和文）ロボットのための音声・環境音・背景音同時認識システムの開発

研究課題名（英文）Development of speech and environmental sounds and background sound simultaneous recognition system for robot

研究代表者
高橋　徹（Takahashi, Toru）

大阪産業大学・デザイン工学部・准教授

研究者番号：３０４１９４９４

交付決定額（研究期間全体）：（直接経費）　　3,500,000円

研究成果の概要（和文）：ロボットに装備されたマイクロホンで、音声・環境音・背景音を同時に扱うことが可能な仕組みを構築しました。複数の音の方向を検出し、各方向の音に分離します。分離した音が音声か非音声かを判定後、音声であれば音声認識、非音声であれば環境音認識する機能を実現しました。更に、方向性の有る音を全て差引き、残った音を背景音とし、どのような背景音であるかを識別する機能も実現し、3種の音を同時に扱う事が可能になりました。

研究成果の概要（英文）：A system that can recognize speech and environmental sound and background sound into a robot with a microphone-array has been developed. It detects the directions of the sound sources, and then separated into each sound for directions. The subsystem is realized by the method of selecting the recognizer depending on whether speech or environmental sounds, after the separated sound is determined whether speech or non-speech. In addition, the remaining sound subtracting all the directional sounds is handled as the background sound. A background sound recognizer is also developed. Finally, the total system has been able to simultaneously recognize speech and environmental sound and background sound.

研究分野：音情報処理

キーワード：音声認識　環境音認識　背景音認識　マイクロホンアレイ　ヒューマノイドロボット　ヒュマンコンピュータインタラクション　コミュニケーション　ロボット聴覚

１．研究開始当初の背景

本研究課題開始以前、申請者を含む多くの研究者が、ヒューマノイドロボット自身に装備された耳（マイクロホン）で聞くという「ロボット聴覚」に関する研究を行っていました。ロボット聴覚研究の重要な研究テーマの一つに、『混合音を構成するそれぞれの音を分離し、分離した音を認識する研究』がありました。このテーマの下、混合音声から個々の音声を同時に認識する「同時発話認識ロボット（通称、聖徳太子ロボット）」が開発されていました。聖徳太子ロボット開発当初は、音声を扱うロボットの研究が少なく、申請者らは、ロボットが音声を扱うことの重要性を世の中に訴えるため、オープンソースロボット聴覚ソフトウェアを開発し、同時にソフトウェア使用方法に関する講習会を開催していました。徐々に、人とロボットのインタラクションにおいて混合音を扱うことの重要性が知られるようになっていきました。

本申請課題は、ロボット聴覚の中でも、音声以外の音にも着目した研究です。音を音声、環境音、背景音に分類し、それぞれを同時に認識することを目標にしました。従来扱われていた混合音の多くは、音声と雑音の混合音、音声と音声の混合でした。実際には、聞こえてくる音には、音声、環境音、背景音が含まれます。つまり、聞える音を全て分離し、構成要素を認識する方法を開発する必要があったのです。同時認識の実現により、実環境下で人間とロボットがスムーズに音声インタラクションできると考えたのです。

２．研究の目的

ロボットが、音声・環境音・背景音を同時認識するための方法を開発すことが研究の目的でした。音声は、言語音として聞える音です。環境音と背景音の区別は、聴く者がどの音に注意を向けるかによってどの音が前景でどの音が背景かが変わってしまいます。そのため、一般に背景音を区別することが困難です。そこで、本研究では、環境音を「音が方向性を伴って聞える非言語音」とし、背景音を「方向性を伴わずに聞える非言語音」と扱おうと考えていました。従来研究では、ロボットによる音声や環境音の認識は、混合されていない音を使った研究が中心でした。これらの知見を活かし、複数音声・複数環境音・背景音を同時に扱う仕組みを検討対象として研究が開始されました。

３．研究の方法

マイクロホンアレイ（複数のマイクロホンを同期して同時に録音可能なマイクロホン）をロボットに装備し研究プラットフォームとなるロボットシステムを製作します。ロボットに求められる機能は、混合音から各構成音の方向を検出する機能、各構成音を集音する機能です。これらは音源定位、音源分離と呼ばれる課題です。これらの機能の実現後に、各構成要素を音声か環境音か背景音かを判定し、判定結果に応じて、音声認識器、環境音認識器、背景音認識器を用いることによって、同時認識が可能になります。

次に、各認識器を構成するため、認識対象となる音声、環境音、背景音を収集する必要があります。得られた音から、統計的学習モデルを用いた認識器を構成します。

音源定位や音源分離の精度を 100%にすることは非常に困難であるため、結果として、歪を含む音が分離音として得られます。歪んだ音が、音声認識器／環境音認識器／背景音認識器の入力になるため、個々の認識精度を低下させ、結果としてシステム全体の認識精度を低下させます。認識精度の低下を補うために、ロボットに自己位置を推定するためのセンサーを取り付け、音源位置や周囲の状況を考慮して認識する方法についても検討しました。

４．研究成果

Robovie-R3 というロボットに、マイクロホンを 16 本搭載し研究プラットフォームとなるロボとを試作しました。マイクロホン 16 本はマイクロホンアレイとして機能するよう設計したことから、混合音から各構成音の到来方向を検出可能です。各構成音の到来方向を手掛かりに、各構成音を混合音から分離することも可能です。これらの実現に MUSIC アルゴリズムと GSS アルゴリズムを用いています。

分離された音の音声認識精度に関しては、従来から多くの知見が示されているため、研究事例の少ない環境音の認識について検討しました。まず環境音認識手法について説明します。

４．１　混合音のモデル

本研究では、混合音を構成する要素を、音声、環境音、背景音の３つにモデル化しています。音声は、言語音として聞える音です。背景音の区別は、聴く者がどの音に注意を向けるかによってどの音が前景でどの音が背景かが変わってしまいます。そのため背景音を厳密に区別することは、一般に困難です。そこで、本研究では、環境音を「音が方向性を伴って聞える非言語音」、背景音を「方向性を伴わずに聞える非言語音」と扱います。ロボットに備わっている音の方向を検出する機能を用いて、方向性を伴う音と、方向性を伴わない音を区別します。方向性を伴う音には、言語音と非言語音の両方が含まれます。方向性を伴わない音は、特定の方向から到来する音ではな

いため、背景音として扱います。背景音にも言語音と非言語音がありますが、ここでは背景音としての言語音は例外とし、取り扱いませんでした。

４．２　音声と環境音の識別

言語音の認識は、音声認識器が担当し、非言語音の認識は、環境音認識器が担当します。従って、方向性を伴う音を入力とし、言語音か非言語音かを識別する手法が必要です。言語音（音声）と非言語音（環境音）かを識別する手掛かりを調査した結果を文献(11)にまとめてあります。音声／環境音を音声特徴の有無に基づき識別する方法を提案し、85 % の識別率を達成しました。注目した音の特徴は、無音からはじまり音のある区間を経て無音となるまでの１区間中に現れる基本周波数の時間変化を記録し、時間方向に周辺化した特徴量です。これは、１つの音に表れる基本周波数の分布と言い換えることができます。音声であれば、人の声の高さの範囲の周波数が密な分布が得られます。環境音では、人の声の高さの範囲に関係なく広い範囲に周波数が分布します。これら分布の差を機械学習の枠組みで識別する方法を開発しました。この方法単独の識別性能は 75%でした。

識別性能を改善するため、スペクトルの傾斜を特徴量化し、識別に用いました。無音から無音までの１つの音の間に、基本周波数が検出される区間と、基本周波数が検出されない区間があります。音声であれば、声帯が振動している区間と、振動していない区間に相当します。環境音であれば、音のエネルギー供給源である振動が存在する区間と存在しない区間に相当します。いずれの音の場合にも、基本周波数が検出される区間では、その音固有のスペクトルが観測されます。音声のスペクトルは、声道の形状を表す特徴が見られます。スペクトル傾斜が-6dB/Oct 程度であることから、スペクトル傾斜は、音声らしさを表す特徴量とみなすことができます。この特徴量による識別性能は、15%でした。開発した２つの特徴量を統合すると 85%の識別率を達成できました。

４．３　環境音の認識

環境音は、方向性を伴う非言語音であると定義しました。まず、非言語音を認識するということは、どういうことかを考える必要があります。狭義の言語音認識（音声認識）は、音声からその音声を構成する音素記号を機械的に書き起こすことを指します。これに倣い、非言語音の認識（環境音認識）を、音からその音を音素記号列に機械的に書き起こすことと定義します。つまり、音から擬音語を生成することを環境音認識と呼ぶことにします。

環境音の認識は、文献(14)にまとめています。音声認識器の構成と同様に、環境音の認識器を構成しました。環境音を人間が聴取し、どのように聞こえたかを書き起こします。特定の１名により聴取した大量の環境音に聞こえた通りの音素系列を割り当てていきます。同一音素が割り当てられた音の特徴をガウス混合分布でモデル化し、音の時間変化をマルコフモデルでモデル化しました。具体的には隠れマルコフモデル(HMM)で記述します。学習データを用い、モデルのパラメータを最尤推定により決定したものが環境音認識器です。

環境音認識器に、音を入力すると、その音に対応する音素系列が出力されます。この音素系列出力が、ロボットの聞えを表しています。この仕組みを応用し、人と機械の協調に擬音語を用いる方法を検討もしました（文献(13)）。

環境音認識に用いる音の特徴（音響特徴）量について検討しました(文献(12))。環境音認識は、音声認識に倣い構成したため、音響特徴量も音声認識で広く用いられる MFCC を用いていました。一方、音源分離や分離音声認識では、MSLS を用いることで MFCC を上回る認識精度が達成されていました。MFCC と MSLS で特徴量による識性能比較を行い、同等の性能を示すことが確認されました。そのため、その後の研究では、音源分離処理と相性の良い MSLS 特徴量を使うこととしました。

４．４　背景音の認識

背景音は、方向性を伴わない音と定義しました。混合音から方向性を伴う音をすべて引き去った後に残る音が背景音です。背景音を認識することは、ロボットが置かれた聴覚的環境を認識することに相当します。ここでは、4 ヶ所の背景音を区別することについて検討しました。注意すべき点は、聴覚的環境を区別するという点です。場所が異なったとしても、聴覚的環境が異なる保証がないため、場所を区別するためには、音以外の情報が必要になります。

オフィス、廊下、エレベータホール、階段の踊り場の 4 ヶ所の背景音の識別実験を行いました(文献(8))。スペクトルの時間的分散を指標とすることで背景音を識別する方法を開発しました。オフィスは、100%、廊下は、96%、エレベータホールは、88%、階段の踊り場は、83% 識別可能でした。4 ヶ所の識別という比較的簡単な識別課題であるにも関わらず、エレベータホールと階段の踊り場の区別が困難であることが確認できます。エレベータホールを階段の踊り場と誤識別される率は、12%、階段の踊り場をエレベータホールと誤識別される率は、10% であることから、階段の踊り場とエレベータホールの聴覚的

環境が類似しているとも言えます。

４．５　GPS センサーによる自己位置推定

背景音から場所を区別することが困難であることがわかってきました。ロボットの自己位置を推定するための手がかりを他のセンサーから得る方法を検討しました。GPS は、屋外の様に人工衛星とセンサーの間に遮るものの無い環境が必要です。そこで、ロボット上で評価する代わりに、センサーを単独で屋外に持ち出し実験を行いました。手軽にロボットに搭載できるよう汎用の GPS センサー（パソコンなどに USB 接続可能な機器）GPS を用い、1〜2 秒毎に位置を確認することで 20m 以内の誤差で位置を推定できることを確認できました。位置情報取得失敗時への対応などの機能を備えたソフトウェアを開発しました。この成果は、堺市の低床式路面電車の位置情報通知サービスへ利用され、公共交通機関の位置情報サービス実証試験を行うまで発展しました(文献(2,3,7))。

４．６　身の回りの混合音

私達の生活の中で、そもそも混合音はどのように現れているかという疑問があります。生活音を記録し、分析することはプライバシーの問題を伴い、直接的に実施することは困難です。そこで、テレビ音声に含まれる混合音を身の回りの混合音ととらえ、実験を行いました。テレビ音を対象に、音声、効果音、ＢＧＭなどの出現区間を調査しました(文献(9))。番組のジャンルや出演者数などによって混合音の出現する割合に何らかの傾向がみられることを期待したが、そのような傾向は確認できませんでした。調査開始当初、バラエティ番組では、出演者同士が同時に発話するケースが多発しているのではないかという期待がありました。しかし、発話を発話で遮るケースはほとんど見られず、発話に笑い声が重なるケースが多いことを確認しました。

４．７　音楽への対応

音楽が流れている部屋を考えると、音環境の取り扱いの難しさが浮かび上がります。音楽は、スピーカーから再生されます。ロボットがスピーカー付近に位置すると、音楽は方向性を伴った音となり、環境音と扱われます。スピーカーから離れた位置に居ると、音楽は方向性を伴わないことがあり、背景音と扱われます。環境音と扱われる場合は、他の方向性を伴う音と同様に、音楽は分離処理が施されます。背景音と扱われる場合、方向性のある音を引き去ることから、音楽信号は歪んでしまいます。従って、環境音、背景音のいずれに取り扱われても、歪んでしまいます。歪んだ音楽から、その曲の情報を得るためには、歪を含む信号の楽曲検索課題を解く必要があります。

音楽だけを分離することは困難であるため、分離が不十分な場合の楽曲検索を実現する必要があります。そもそも分離をすれば分離歪が発生するため、分離処理により歪を発生させるより、分離せずに楽曲を含む混合音から直接楽曲を検索できないかと考えました。楽曲と音声の混合音を検索キーとして楽曲を検索する課題を検討しました(文献(1,4,5,6,10))。

４．８　音声・環境音・背景音の同時認識

本研究課題によって得られた最終的な認識システムの構成は文献(4)にまとめられています。各サブシステムは、当初想定していた範囲では動作可能なシステムに組み上がりました。

当初の計画に沿って、個々のサブシステムを開発していく中で、音楽の取り扱いに課題が残りました。音楽を扱うためには、与えられた音が音楽であるかどうかを判定する必要があります。また、開発したシステムは、無音から入力音がありまた無音になる音を想定しているため、音楽の様に、無音区間という切れ目の無い音（無音から無音までが数十秒以上の音）を扱えないという課題があります。真のロボット聴覚を実現するためには、多くの課題を解決する必要があります。特に、音の切れ目を仮定しない連続処理可能な手法を確立する重要性が明らかになってきたと言えます。この問題は、今後の研究活動で解決していきたいと考えております。

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計　０件）

〔学会発表〕（計１４件）

(1) 高橋 徹, "混合音を検索キーとした音楽検索のための高速特徴量比較手法の検討", 日本音響学会 2015 年春季研究発表会, 3 日本大学（東京都千代田区）, March, 16—18, 2015, (発表日 3/18).
(2) 高橋 徹, 能勢 和夫, 塚本 直幸, 吉川 耕司, "Twitter を用いた路面電車の位置通知システムの検討", 福祉情報工学研究会, 筑波技術大学（茨城県つくば市）, March, 13—14, 2015, (発表日 3/13).
(3) 高橋 徹, 能勢 和夫, 塚本 直幸, 吉川 耕司, "路面電車の位置通知システムの設計と実装", 福祉情報工学研究会, 産業技術総合研究所（東京都江東区）, 臨海副都心センター, December, 11, 2014, (発表日 12/11).
(4) 高橋 徹, "ロボットのための音シーン理解技術の実装例", 日本音響学会 2014 年秋季

研究発表会，北海学園大学（北海道札幌市），September，3－5，2014，(発表日 9/5)．
(5) 樋口 颯，高橋 徹，"混合信号を検索キーとした音楽検索のための特徴量帯域幅に関する考察"，日本音響学会 2014 年秋季研究発表会，北海学園大学（北海道札幌市），September，3－5，2014，(発表日 9/4)．
(6) 樋口 颯，高橋 徹，"特徴量間の累積距離を用いた混合音からの音源検索システムの評価"，電子情報通信学会，信号処理研究会 立命館大学大阪梅田キャンパス（大阪府大阪市），August，28，2014，(発表日 8/28)．
(7) 谷口 哲也，高橋 徹，"GPS を使った堺市低床式車両位置情報通知サービスの開発"，情報処理学会第 76 回全国大会，東京電機大学（東京都足立区），March，11－13，2014，(発表日 3/11)．
(8) 高橋 徹，"マイクロホンアレイを用いた音声・環境音・背景音の識別"，日本音響学会 2014 年春季研究発表会，日本大学（東京都千代田区），March，10－12，2014，(発表日 3/11)．
(9) 赤塚 俊洋，高橋 徹，"テレビ番組のジャンル別音声の混合・非混合シーンの調査"，日本音響学会 2014 年春季研究発表会，日本大学（東京都千代田区），March，10－12，2014，(発表日 3/12)
(10) 樋口 颯，高橋 徹，"音楽と音声の混合音からの楽曲同定に関する一考察"，日本音響学会 2013 年秋季研究発表会，豊橋技術科学大学（愛知県豊橋市），September，25－27，2013，(発表日 9/26)．
(11) 高橋 徹，"ロボットのための音声と環境音の識別手法"，日本音響学会 2013 年秋季研究発表会，豊橋技術科学大学（愛知県豊橋市），September，25－27，2013，(発表日 9/25)．
(12) 高橋 徹，"環境音の特徴量調査"，日本音響学会 2013 年春季研究発表会，東京工科大学（東京都八王子市），March，13－15，2013，(発表日 3/13)．
(13) Yusuke Yamamura，Toru Takahashi，Tetsuya Ogata，Hiroshi G. Okuno，“Sound Source Selection System Using Onomatopoeic Queries from Multiple Sound Source”，Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems，Portugal(Vilamoura，Algarve)，October，7-12，2012，(発表日 10/9)．
(14) 高橋 徹，"自動擬音語生成法の検討"，日本音響学会 2012 年秋季研究発表会，信州大学（長野県長野市），September，19－21，2012，(発表日 9/21)．

〔図書〕（計　0件）

〔産業財産権〕
○出願状況（計　0件）

名称：
発明者：
権利者：
種類：
番号：
出願年月日：
国内外の別：

○取得状況（計　0件）

名称：
発明者：
権利者：
種類：
番号：
出願年月日：
取得年月日：
国内外の別：

〔その他〕
ホームページ等
http://www.ise.osaka-sandai.ac.jp/~takahashi/index.html

６．研究組織
(1)研究代表者
高橋 徹（Takahashi，Toru）
大阪産業大学・デザイン工学部・准教授
研究者番号：30419494

(2)研究分担者
（　　）

研究者番号：

(3)連携研究者
（　　）

研究者番号：