日常支援ロボットシステムのための ケプストラム特徴量と判別分析に基づく生活音の認識

Life Sound Recognition Based on Cepstrum and Discriminant Analysis for Daily Supporting Robot System

○学 得津覚（東大）岡田慧（東大）稲葉雅幸（東大）

Satoru TOKUTSU, The Univ. of Tokyo, tokutsu@jsk.t.u-tokyo.ac.jp

Kei OKADA, The Univ. of Tokyo

Masayuki INABA, The Univ. of Tokyo

For daily assistive robots sharing household space with human beings, it is important to recognize daily life sounds and to select an appropriate behavior autonomously. To use auditory information have benefits, for example, to reason about unseeing situations. In this paper, we examine an algorithm for daily life sound recognition for using in several scenes. Through this examination, we propose a framework for using daily life sound recognition in daily assistive robot system.

Key Words: Daily Life Sound, Daily Assistive Humanoid Robot, Cepstrum, Discriminant Analysis

1 序論

ヒューマノイドロボットと人が共に生活する事を考えた時に、今のヒューマノイドにとって特に足りないのは、聴覚情報処理である。家庭における場面で人の声とそれ以外の物音を区別することも、食器を置いた時の物音から食卓の上の状況を推測したりすることも、確かな方法が存在しないのが現状である

ヒューマノイド HRP-2 を使った聴覚、および視聴覚機能に基づく支援行動は、原らによって提案されている [1]。彼らはヒューマノイドを家庭にある家電の中心的存在として捉え、人と家電とのインターフェースに必要なコミュニケーション能力、主に視聴覚情報に基づく発話認識について研究を行った。彼らが実験に用いているロボット HRP-2 はマイクロフォンアレイを有するために、そこから得られる高信頼度の音源定位情報と視覚による人間追跡情報との統合を行い、人の発話イベントの検出に関する研究を行った。

発話認識の研究分野では、奥乃らは三人の話者の同時発話の理解を実現した。彼らは、話者適応、高 S/N 比のための発話ストリームのセグリゲーション、ノイズの抑制について研究を行ってきている [2]。

ロボットの能動的センシングも、支援ロボットシステムにおける必須の機能である。中臺らは、視覚機能と自己の運動によるノイズに対する抑制を用いつつ、同時複数音源が存在する環境下での音源定位機能を提案した [3]。

本論文では日常支援ロボットの生活音の認識を対象とし、日常生活空間を人と共にするロボットにとって必要な生活音の認識についていくつか実例を挙げながら検討する。また、これらを通して支援ロボットシステムにおける生活音を利用するための枠組みを提案する。

2 日常支援ロボットの生活音認識利用場面

Fig.1 に示された複数の支援場面のように、日常支援を行うロボットシステムでは、自分が今どこにいるか、どのような状況・文脈の中にいるかというのは支援を行う上で大変重要な情報であり、実際に自己位置推定や、状

Fig. 1: Daily assistive humanoids can use daily life sound in each scene.

況推定などの研究も行われている [4, 5]。

日常支援ロボットが記憶として持つ地図内において自己位置が把握できていると仮定した場合、自己位置から周囲にどのような生活音が存在する可能性を与えることで生活音クラス数の増加にも対応できると考えられる (Fig.2 参照)。実際に本研究では、自己位置推定の部分と生活音認識を切り分けて考え、場所ごとに存在が予想される生活音候補に対する認識能力を検証する。

3 生活音認識手法

3.1 微弱音圧の生活音に対応可能なケプストラム特徴量

ケプストラム特徴量は、次の手順で求めることができる。長さ N の音データの配列 $x(n)$ ($n = 0 \sim N-1$) に対し、離散フーリエ変換 (DFT) をかける前処理として、ハミング窓をかける。

$$x'(n) = (0.54 - 0.46 \cos \frac{2\pi n}{N-1}) * x(n) \quad (1)$$

Fig. 2: Humanoids can select sound recognition model according to self position.

ハミング窓によって修正された音データ $x'(n)$ に離散フーリエ変換 (DFT) を行いスペクトル $X(k)$ を求める.

$$X(k) = \sum_{n=0}^{N-1} x'(n) \cdot e^{-j\frac{2\pi k}{N} \cdot n} \quad (2)$$

得られたスペクトル $X(k)$ ($k = 0 \sim N-1$) に対数を取り、逆離散フーリエ変換 (IDFT) を行うことで、スペクトルを基底 $e^{j\frac{2\pi n}{N} \cdot k}$ を使って展開する.

$$c(n) = \frac{1}{N} \sum_{k=0}^{N-1} (\log |X(k)|) \cdot e^{j\frac{2\pi n}{N} \cdot k} \quad (3)$$

得られた係数列 $c(n)$ のことをケプストラム係数, n のことを次数と呼ぶ [6].

この係数は、スペクトルの分布を表現している係数列であり、 $c(n)$ は n が小さいほどスペクトル曲線のよりおおまかな外形を表す. 一般に、音声認識でケプストラムを用いる場合には、10 数次までの係数を用いることが多く、話者認識に用いる場合にも、同じく 10 次から 20 次までの係数を用いることが多い [7]. ケプストラムはスペクトル曲線の外形を直接表現する特徴量であるため、音圧の大きさには直接影響を受けずにその分布の特徴を捉えることができる.

3.2 線形判別分析と生活音クラス距離の評価

線形判別分析とは、クラス間の分散とクラス内の分散の比を表す判別基準 ρ を最大とする線形変換 A を求めることで、特徴量をより判別しやすい空間で評価することができる. 判別基準 ρ は以下の式で定義される.

$$\rho = \text{tr}(C_W^{-1} C_B) \quad (4)$$

$$C_W = \sum_{j=1}^K \omega_j C_j, \quad \omega_j = P(C_j) = \frac{1}{K} \quad (5)$$

$$C_j = E_{\mathbf{x}_i \in C_j} \{(\mathbf{x}_i - \boldsymbol{\mu}_j)(\mathbf{x}_i - \boldsymbol{\mu}_j)'\} \quad (6)$$

$$C_B = \sum_{j=1}^K \omega_j (\boldsymbol{\mu}_j - \boldsymbol{\mu}_T)(\boldsymbol{\mu}_j - \boldsymbol{\mu}_T)' \quad (7)$$

$$\boldsymbol{\mu}_T = \sum_{j=1}^K \omega_j \boldsymbol{\mu}_j \quad (8)$$

変数に関して、 C_W はクラス内分散、 C_B はクラス間分散、 K はクラス数、 C_j は j 番目のクラス、 C_j はクラス C_j の共分散行列、 ω_j はクラス C_j の生起する確率 (ここでは等確率に生起すると仮定)、 $\mathbf{x}_i$ は i 番目の特徴量ベクトル、 $\boldsymbol{\mu}_j$ はクラス C_j の平均特徴量ベクトル、 $\boldsymbol{\mu}_T$ は全体の平均特徴量ベクトル、シングルクォートは行列の転置を表す. A は、一般化固有値問題 $C_B A = C_W A \Lambda$ ($\Lambda = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_m)$, $m = \min(K-1, n)$ (n is size of a feature vector)) を解くことで得られる.

得られた A を用いて、ケプストラム特徴を判別空間に射影し、クラスごとに各判別軸に対する平均値と分散を評価し、それらをまとめて学習データとする. 認識時には、マイクから取得できる音データをケプストラム特徴に変換後、判別空間に射影し、学習済みのクラス C_j との距離 d_j を次の式で評価する.

$$d_j = \frac{1}{n} \sum_{k=1}^n \sqrt{\frac{([A\mathbf{x} - A\boldsymbol{\mu}_j]_k)^2}{V([A\mathbf{x}]_k)}} \quad (9)$$

ただし、 $V(a)$ は a の分散を、 $[b]_i$ はベクトル $\mathbf{b}$ の i 番目の要素を表す.

4 生活場面ごとの生活音認識検証実験

以下の検討では、ヒューマノイドロボット HRP-2 を改良した HRP2V の頭部に搭載されたマイクロフォンを使用し、サンプリング周期 44.1[kHz]、分析窓長 30[msec]、分析窓シフト長 10[msec] で解析を行った.

4.1 寝具周辺にいる場合

日常支援ロボットが、人の睡眠する場所である寝具周辺にいる場合に、想定される生活音に対して検討する. 生活音クラスは、1. 目覚まし時計のアラーム音、2. 携帯電話の呼び出し音 (一定周波数による断続的な音)、3. パソコンのキーボードの打鍵音 (エンターキーの強めの打鍵音)、4. ヒューマノイド HRP-2 の歩行音、5. 背景音とし、ケプストラム特徴量と線形判別分析を用いて解析を行った. 結果として、各軸の方向を示すベクトルを Fig.3 (a) に、各軸の判別具合を表す固有値を Fig.3 (b) に、空間内における特徴量の分布の様子を Fig.4 に示す.

背景音と目覚まし時計のアラーム音のスペクトルは Fig.5 に示す通りで、目覚まし時計のアラーム音は、2[kHz] 周期 (2kHz, 4kHz, 6kHz, ...) に上に凸なグラフであり、マイクのサンプリング周期が 44.1[kHz] で窓長が 30[msec] であるため、その周波数分布の特徴がケプストラム係数の 22 次に見れることが分かる. よって、Fig.4 では 30 次までのケプストラム係数を用いて解析を行っているため、目覚まし時計の音を判別するのに十分な情報量を用いている. 図中からも各クラスがほぼ判別できていることが分かる.

4.2 部屋の出入口周辺にいる場合

日常支援ロボットが、人の往来する場所である出入口周辺にいる場合に、想定される生活音に対して検討する. 生活音クラスは、1. ドアを閉める音、2. ドアの鍵をかける音、3. 蛍光灯のスイッチを押す音、4. 給湯用の水道の水が流れる音、5. 背景音とし、ケプストラム特徴量と線形判別分析を用いて解析を行った. 結果として、各軸の方向を示すベクトルを Fig.8 (a) に、各軸の判別具合を表す固有値を Fig.8 (b) に、空間内における特徴量の分布の様子を Fig.9 に示す.

4.3 調理場周辺にいる場合

日常支援ロボットが、人およびロボットが調理を行う場所である調理場周辺にいる場合に、想定される生活音に対して検討する。生活音クラスは、1. 水道の水が流れる音、2. 冷蔵庫の引き出しをしまう音、3. 冷蔵庫のドアを閉める音、4. 電子レンジの扉を閉める音、5. 電子レンジの電子音、6. 包丁をまな板に当てる音、7. キッチンの引き出しをしまう音、8. フライパンをIHに置く音、9. 皿をキッチン台に置く音、10. 皿を皿の上に重ねる音、11. 背景音とし、ケプストラム特徴量と線形判別分析を用いて解析を行った。結果として、各軸の方向を示すベクトルを Fig.8 (a) に、各軸の判別具合を表す固有値を Fig.8 (b) に、空間内における特徴量の分布の様子を Fig.9 に示す。

特徴量の分布を見てほぼ判別できそうなクラスは、水道、冷蔵庫ドア、包丁、電子レンジの電子音、背景音の5クラスが挙げられる。固有値のグラフ (Fig.8 (b)) を見ると、8クラスまでは認識できるという結果が得られている。ただし、全体的にサンプル数が少ないという過学習による判別しやすい条件で実験を行っているため、もっとサンプル数を増やした場合の実験が必要である。

5 結論と今後の展望

本論文では、日常支援ロボットシステムが場所ごとに生活音認識候補を変えていくことで、クラス数の増加にも原理的に対応できる枠組みを提案した。また、ケプストラム特徴量と線形判別分析に基づき、屋内を想定した場合の生活音認識能力を確認し、合計で13クラスの生活音がほぼ判別可能であることを確認した。

今後は、生活音認識の精度を上げることと、屋外の生活音への拡張、実際に生活音認識が支援ロボットの支援行動につながることを示し、場所だけでなく状況に応じて生活音候補を変えていく方法論について研究を行っている。

参考文献

- [1] Isao Hara, Futoshi Asano, Hideki Asoh, Jun Ogata, Naoyuki Ichimura, Yoshihisa Kawai, Fumio Kanehiro, Hirohisa Hirukawa, and Kiyoshi Yamamoto. Robust Speech Interface Based on Audio and Video Information Fusion for Humanoid HRP-2. In *Proceedings of 2004 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 2404–2410, 2004.
- [2] Hiroshi G. Okuno, Tomohiro Nakatani, and Takashi Kawabata. Challenge problem for computational auditory scene analysis: Understanding three simultaneous speeches. In *Proceedings of 15th International Joint Conference on Artificial Intelligence (IJCAI-97)*, Vol. 1, pp. 30–35, 1997.
- [3] Kazuhiro Nakadai, Tino Lourens, Hiroshi G. Okuno, and Hiroaki Kitano. Active Audition for Humanoid. In *Proceedings of 17th National Conference on Artificial Intelligence (AAAI-2000)*, pp. 832–839, 2000.
- [4] 畑尾直孝, 岡田慧, 稲邑哲也, 稲葉雅幸. 屋内支援ロボットのための室内と廊下に対する自律的地図作成機能の実現. 第25回日本ロボット学会学術講演会講演論文集, p. 3L17, 9 2007.
- [5] 得津寛, 岡田慧, 稲葉雅幸. 日常支援ヒューマノイドの生活音と人の認識履歴に基づく視聴覚状況認識・推論. 日本機械学会ロボティクス・メカトロニクス講演会'08 講演論文集, June 2008.
- [6] B.P. Bogert, M.J.R. Healy, and J.W. Tukey. The frequency analysis of time series for echoes: Cepstrum, pseudo-autocovariance, cross-cepstrum and saphe cracking. In *Proceedings of the Symposium on Time Series Analysis*, pp. 209–243, 1963.
- [7] 松井知子, 黒岩眞吾. 音声による個人認証技術の現状と展望 - 今、なすべきことは何か! -. 電子情報通信学会誌, Vol. 87, No. 4, pp. 314–321, 2004.

Fig. 3: Discriminant axes and eigen values for sounds around bedclothes (5 classes), and a color scale. (a) shows direction of discriminant axes (horizontal: order of discriminant axis, vertical: order of cepstrum, color represents value: [-1.0, 1.0]). (b) shows color scale (darker blue is smaller and more light yellow is bigger. Tick marks shows 1/10 of the whole range). (c) shows eigen value for each axis (horizontal: order of discriminant axis, vertical: eigen value. Red line shows 1.0 which is discriminable threshold).

Fig. 4: Feature vectors distribution in discriminant space based on sounds around bedclothes (5 classes)

Fig. 5: Spectra of background noise and an alarm clock (horizontal: [dB], vertical: [Hz]. Red line shows spectral envelope).

Fig. 6: Discriminant axes and eigen values for sounds around an entrance door (5 classes), and a color scale. Description of graph (a), (b) and (c) are same as shown in Fig.3

(a) Discriminant space based on the 1st, 2nd and 3rd discriminant axes

(b) Discriminant space based on the 1st and 4th discriminant axes

Fig. 7: Feature vectors distribution in discriminant space based on sounds around an entrance door (5 classes)

Fig. 8: Discriminant axes and eigen values for sounds in a cooking place (11 classes), and a color scale. Description of graph (a), (b) and (c) are same as shown in Fig.3

(a) Discriminant space based on the 1st, 2nd and 3rd discriminant axes

(b) Discriminant space based on the 1st and 4th discriminant axes

Fig. 9: Feature vectors distribution in discriminant space based on sounds in a cooking place (11 classes)