7K-9

ヒューマノイドロボットによる自律的な音韻と語彙の獲得

戸塚　伸弥[†]　　鈴木　健嗣[‡]　　橋本　周司[§]

早稲田大学大学院理工学研究科[†]　筑波大学大学院システム情報工学研究科[‡]
早稲田大学理工学部応用物理学科[§]

1. はじめに

近年，人間との対話を目指したヒューマノイドロボットの研究が盛んに行われており，多くのロボットは音声認識システムを実装することで人間との対話を行っている．これら音声認識を有する対話ロボットはあらかじめ与えられた語彙と音韻列により対話を行うため，音声言語の基となる音韻を自律的に獲得したり，未知な語彙をダイナミックに獲得することは難しい．

これに対し，ロボットに実世界の事物と音声を同時に教示することによって，語彙獲得を目指した研究がある[1]．しかし，この手法ではあらかじめ用意した音韻 HMM を用いており，環境に応じて新たな音韻を形成し獲得していくことはできない．また，語彙を発話する際，音韻毎に用意された音声を使用しており，獲得した音韻によるものではない．

そこで本研究では，ロボットの情報処理機構に複数の音韻を表現できる音声主要素マップを提案する．また，人間がロボットに実世界の物体を音声と共に教示することで，ロボットによる自律的な音韻と語彙の獲得を実現した．さらに，獲得した音韻に基づく発話生成実験を行い，主観評価実験によりその有効性を検討した．

2. 音声主要素マップ

人間の音韻知覚能力は幼児期に特定の言語圏の環境下にさらされることで，その言語固有のものとして獲得される．その後，周囲の人とのコミュニケーションを通じて与えられる音を，獲得音韻と照らし合わせながら，その音が指し示す事物と対応付けて語彙を獲得していく．また，獲得した語彙を発話する際も，自ら獲得した音韻に従って発話を行っている．

このような処理を実現するため，音韻空間をモデル化した音声主要素マップを導入する(図 1)．これは，聴取した音声データの分析窓毎におけるメルケプストラム係数$c_\alpha(m)$を音声特徴量[2]とするm次元空間マップである．この空間上でウォード法によるクラスタ分析を行いK個の代表ベクトルを求める．このK個の代表ベクトルで表された音韻マップを音声主要素マップ(VME-Map: Voice Main Elements-Map)と呼ぶ．マップ

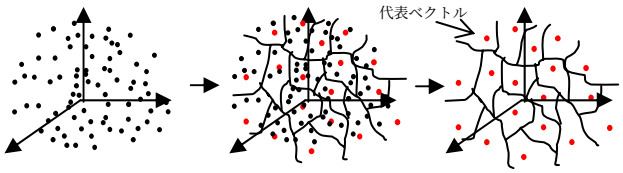
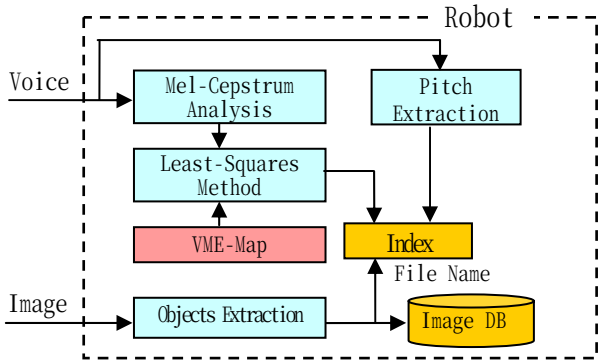


図 1．音声主要素マップ生成のイメージ

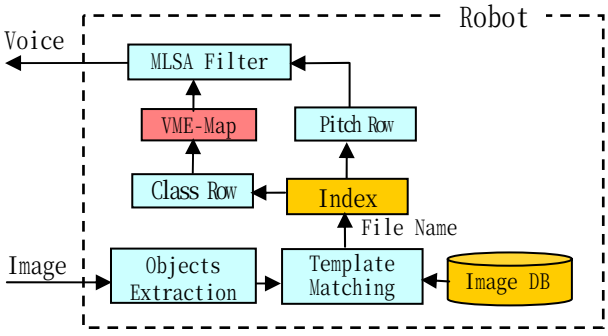
のクラス数 K は，各クラスの代表ベクトルとそのクラスに属するベクトルとのユークリッド距離を全クラスについて求め，その総和の変化が少なくなる K=250 と定め，音声主要素マップの初期クラス数とした．

3. システム構成

図 2 に，音声主要素マップに基づく語彙の獲得および獲得語彙の発話システムの構成を示す．



(a)　語彙の獲得



(b)　獲得語彙の発話

図 2 .システム構成

3. 1 語彙の獲得

教示者はロボットに物体を提示し，同時に物体を指し示す単語を発話する．ロボットは自動的に音声を検知すると音声区間のみを取得し，画像も同時に取得する．

入力された音声に対し，ロボットは音声主要素マップに基づき音韻を認識し，画像情報との関連付けを行う．同時に，入力された音声デー

“Autonomous acquisition of phonologic and lexicons by Humanoid Robot”
Shinya Tozuka, Graduate School of Science and Engineering,Waseda University, Kenji Suzuki, Graduate School of Systems and Information Engineering, University of Tsukuba, Shuji Hashimoto, Dpartment of Applied Physics, School of Science and Engineering, Waseda University.
E-mail : shinya@shalab.phys.waseda.ac.jp

タにより，音声主要素マップを更新する．まず，抽出した音声区間から，分析窓毎にメルケプストラム係数とピッチを求める．これより，音声主要素マップを参照して MSE(Mean Square Error)が最小となる代表ベクトル（クラス番号）を分析窓毎に検索し，クラス列として入力画像に関連付けるためのインデックスファイルに記録する．また，発話時に使用するためにピッチも逐次インデックスファイルに記憶する．

この処理によりi番目の単語w_iについて式(1)，(2)で表される獲得語彙に関する情報(=クラス列，ピッチ列)がインデックスファイルに記憶される．

CR(クラス列)= C(0),C(1),… ,C(L)　　(1)
P(ピッチ列) = p(0),p(1), …,p(L)　　(2)
L:フレーム数

画像に対しては，濃度分布の分散比が最大になる点で２値化，ラベリング処理を行い，最大面積以外をクリアすることで自動的に物体抽出処理を行う．処理後の画像を取得時刻，および上記のクラス列とともに記録し，画像データベースに保存する．

3．2 獲得語彙の発話

教示者がロボットに対し物体のみを提示し，一定時間音声が検知されない場合，ロボットは発話生成を行う．画像に対し上述の物体抽出処理を行い，獲得した画像データベース中から最大相関値をとる画像を検索する．次に，この画像に関連付けられているインデックスより，対応する音韻クラス列とピッチ情報を読み込み，メル対数スペクトル近似(MLSA)フィルタにより音声合成を行う[3].

3．3 音声主要素マップ更新

新たに入力された音韻データがすでに獲得済みかを判別する閾値を定めるため，母音のみの学習データで生成した音声主要素マップを用いて，(a)母音のみの音声データと(b)母音以外の音声データについて MSE を求めた結果を図 3 に示す．これは，分析窓毎に MSE が最小となる代表ベクトルを用い，その値を全データについて平均した値である．実験条件は，サンプリング周波数 11kHz，窓長 46ms，窓間隔 15ms，メルケプストラム次元数 25 である．実験では，母音のみで合計 20 秒，母音以外で合計 65 秒の音声データを用いた．母音以外，つまり未獲得音韻を含むデータについては，MSE>3.0，既に獲得済みの音韻である母音のみのデータでは，MSE<2.0 となった．そのため，入力データ中の音韻が未獲得であるかの基準として中間値である MSE=2.5 を閾値とし，1 単語毎にその単語中で MSE>2.5 となるメルケプストラム係数を新規音韻として音声主要素マップへ追加する．また，MSE>2.5 となるフレームがない場合，3 回以上同じクラスを参照するメルケプストラムを現在のマップに加え再び同じクラス数となるようクラスタリング処理を行う．以上により，語彙の獲得に伴った音韻マップ更新を行うこととした．

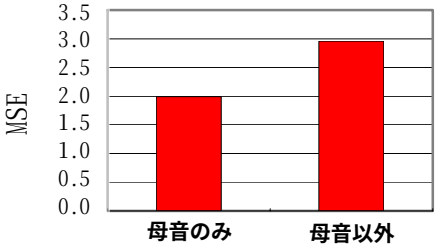


図 3．MSE の比較結果

4．評価実験

提案手法の場合，その性質上音韻認識結果を明確に確認することができない．そこで，学習前(母音のみのマップ)と学習後（子音も含めたマップ）で獲得語彙の発話結果に対し，知覚精度に差が出るか比較実験を行った．実験条件は，3.3 と同じ条件とし，獲得させた語彙数は一般的な日本語の単語 20 個とした．また，被験者には語彙集合をあらかじめ知らせないこととした．図 4 に比較実験の結果を示す．

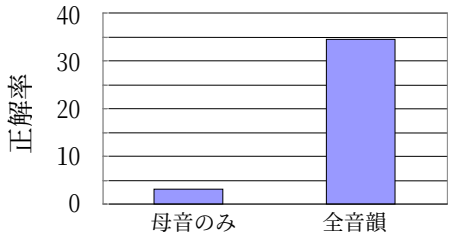


図４　比較実験結果

母音のみのマップによる生成では語彙をうまく獲得できず，発話の知覚精度は非常に低いものとなった．一方母音以外を含む語彙を獲得させた場合，知覚精度の向上が見られた．

5．まとめ

本稿では音声主要素マップを提案し，ロボットによる自律的な語彙の獲得，及び獲得語彙の発話を実現した．また，入力された音声に応じて，音韻を新規に獲得するだけでなく，獲得した音韻を徐々に変化させることも可能となった．

今後は，音声主要素マップの更新で用いた閾値設定を検討するとともに，ロボットの動作と結びつけた自律的な語彙獲得システムの構築を行いたい．

参考文献

[1] Iwahashi,N: “Active and unsupervised learning of spoken words through a multimodal interface“, Proc.13th IEEE Workshop Robot and Human Interactive Communication, 437-442, (2004)

[2] 徳田恵一, 小林隆夫, 深田俊明, 斎藤博徳, 今井聖: “メルケプストラムをパラメータとする音声のスペクトル推定,” 信学論(A), Vol.J74-A, No.8, pp.1240–1248, Aug. 1991.

[3]今井聖, 住田一男, 古市千枝子: “音声合成のためのメル対数スペクトル近似(MLSA)フィルタ,” 信学論(A), vol.J66-A, no.2, pp.122.129, Feb. 1983.