能動的サンプリングに基づく被模倣音の発見

○三浦勝司 [1,2]　吉川雄一郎 [1]　浅田稔 [1,2]
1. JST ERATO 浅田共創知能システムプロジェクト　2. 大阪大学大学院

Realizing being imitated: vowel mapping with clearer articulation

*Katsushi Miura[1,2], Yuichiro Yoshikawa[1], Minoru Asada[1,2]
1. JST ERATO Asada Project 2. Graduate School of Engineering Osaka Univ.

Abstract— The previous approach to vowel imitation learning between a caregiver and an infant (robot) [1] has assumed that the robot can segment the caregiver's utterance into its phoneme category, where the caregiver always imitates the robot utterance. However, in real situations, the caregiver does not always imitate the robot utterance, nor the robot does have the phoneme category (no segmentation capability). This paper presents a method to solve these issues, a weakly-supervised learning along with auto-regulation, that is active selection of action and data with underdeveloped classifier. To cope with not-always imitation problem, a weakly-supervised learning method is applied that is capable to handle incompletely segmented samples (not perfectly imitated voices). Further, the regulation classifier of the imitated voices is recursively applied in order to select good vocal primitives and to segment caregiver's imitated voices that improve the performance of the classifier itself. The simulation results are shown and the future issues are given.

Key Words: Interaction, Imitated voice, Supervised learning along with auto-regulation

1. はじめに

ヒューマノイドロボットは音声やジェスチャ等により人とコミュニケーションすることが期待される．しかし，人とロボットでは身体構造が異なるため，観測した人の行動をロボットが自身の身体に当てはめて理解し再現する事は難しい．一方，人の乳児は言語獲得過程においてロボットと同様に身体構造の違いによる音声のマッピング問題を抱えているにもかかわらず，対応関係を学習し言語を獲得している．

発達心理学の知見に，乳児の母音様のクーイングが母親の模倣を促し (Masataka and Bloom [2])，母親の模倣が乳児の発話を促す (Peláez et al. [3]) という報告がある．この報告を基に，母子間で発話を相互に模倣しあうことが乳児に母親の発話の理解と模倣を可能にさせると仮定し，ロボットの発話に対する教示者の模倣によってロボットに母音を獲得させた研究がある．Yoshikawa et al. [4] は乳児の発話に対する教示者の模倣が母音獲得において重要な役割を果たすことを構成的に示した．また，人が音に対して自身の持つカテゴリにバイアスされて知覚して知られる現象 "perceptual magnet effect" [5] を考慮することによって，Miura et al. [1] は教示者の模倣が無意識のうちにロボットの発話を明瞭な母音へと導くことを示した．ただし，これらの研究 [1]，[4] において，教示者はロボットの発声する単音節の母音様の音のみを正確に模倣することが仮定されていた．つまり，ロボットは自身の発声に対する教示者の模倣を探し出す必要がなかった．しかし，実際の母子間インタラクションでは母親は必ずしも模倣せず，単音節での音声のみが使われるとは限らない．さらに，乳児の発音が未成熟であるため，親が乳児の発話を逐一模倣すると仮定すること，また模倣が可能であると仮定することは現実的であるとは言えない．

本稿ではより現実的な母子間インタラクションとして，ロボットの複数音節の発話に対して教示者が模倣しようとするが，ロボットの発話の未成熟さによって必ずしも模倣が成立しない状況を想定する．教示者が必ずしも模倣しないため，ロボットはインタラクションの経験から自律的に自身の発話に対する教示者の模倣を発見し，それをもとに発話の対応関係を獲得する必要がある．この問題設定は様々な物体が置かれた環境において母親が乳児と発話による共同注意をしようとしている状況を想定したものである．このような状況では，母親は乳児の不完全な発話から単語を推測し，その推定が正しいかを確認するために乳児に対してその単語を発話するようなインタラクションが起こると考えられる．この必ずしも模倣が成立しない問題を解く手法として，ロボット自身の能動的なサンプリング，すなわちロボット自身の判断に基づいて発話を選択すること及び教示者の応答に含まれる被模倣部分を自律的に抽出して学習データとして利用することを通じて学習する手法を提案する．次章以降では，インタラクション場面の詳細な設定，提案手法，実験内容について順に説明して行き，最後に実験結果とそれに対する考察について述べる．

2. 問題設定

ロボットと教示者とが互いに複数の音節を発話しあうインタラクションを想定する．ここで，教示者は $/v_i/, (i = 1, \cdots, N_h)$ で示される N_h 個の母音を持っていると仮定する．一方，ロボットは N_h グループの母音のプリミティブを持っている．ここで，i 番目のグループはそれぞれ異なる確率で母音 $/v_i/$ であると教示者に認識される $N_r^{/v_i/}$ 個の発話プリミティブから構成されるものとする．ロボットの持つ i 番目のグループの j 番目のプリミティブを $/v_{i,j}/$ $(j = 1, \cdots, N_r^{/v_i/})$ と表記

する．教示者に $/v_{i,j}/$ が母音 $/v_i/$ であると正しく認識される確率を認識率と呼び，$P^{/v_{i,j}/}$ と表記する．ただし，ロボットは $P^{/v_{i,j}/}$ について知らないものとする．

インタラクションの各ターンにおいて，ロボットは自身のプリミティブから N_a 個選択し，順に発話する．このときロボットが各プリミティブを選択する確率は発話割合 Q_l^t によって定義される（発話割合 Q_l^t の決め方については 3.1.4 節を参照）．t ターン目のインタラクションにおいて k 番目にロボットが発話したプリミティブを $/R_k^t/, (k = 1, \cdots, N_a)$ とする．教示者は N_a 個のロボットの発話プリミティブそれぞれに対して模倣率 P で模倣を試みる．ここで t ターン目のインタラクションにおいて k 番目のロボットの発話に対する教示者の応答を $/H_k^t/$ とする．したがって，ロボットの k 番目の発話に対し教示者が対応する母音を返す確率は $PP^{/R_k^t/}$ であり，その他の母音を返す確率は $(1 - PP^{/R_k^t/})/(N_h - 1)$ である．

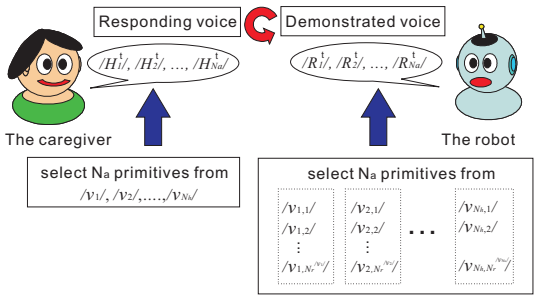


Fig.1 An overview of the interaction

ロボットは教示者の応答を音圧の系列 $S^t = \{s(1), \cdots, s(d^t F)\}$ として受け取る．ただし，$s(\tau)$ は教示者の応答開始から $\frac{\tau}{F}$ 秒後の音圧であり，d^t は t ターン目のインタラクションにおける教示者の発話持続時間，F はサンプリング周波数である．T ターンのインタラクション後，ロボットは自身が発話したプリミティブ R_k^t とそれに対する教示者の応答 S^t のセットのみから，ロボットの各プリミティブ $/v_{i,j}/$ に対応する教示者の模倣を認識する $N_r^{/v_i/}$ 個の被模倣音発見器を学習する．

3. 提案手法

教示者とのインタラクションを通じて，ロボットは被模倣音発見器 $f^{/v_{i,j}/}$ を獲得する．被模倣音発見器 $f^{/v_{i,j}/}$ はロボットの各プリミティブ $/v_{i,j}/$ 毎に定義されており，教示者の発話 S を $/v_{i,j}/$ に対応した音とそうでない音との分節化にも利用することができる．また，獲得した被模倣音発見器は教示者の模倣の有無を評価するだけでなく，模倣の成功経験に応じた能動的な発話や学習データのサンプリングにも利用される．

3·1 教師あり学習

被模倣音発見器 $f^{/v_{i,j}/}$ は教示者の応答 S^t から抽出した MFCC 特徴量の系列 $\boldsymbol{m}_n^t$（次節参照）が $/v_{i,j}/$ に対応する音の特徴を含むかどうかの判定を行う．ロボットが被模倣音発見器を獲得するための基本的なアイデアは，教示者の応答が常にロボットの発話の模倣であるとしてロボットに対応関係を学習させることである．

この仮定により，ロボットの発話 $/R_1^t/, \cdots,$ or $/R_{N_a}^t/$ のラベルを教師信号に，教示者の発話 S^t を入力として教師あり学習を行うことができる．ただし，教示者は必ずしも模倣しないため，S^t に対するラベル $/R_1^t/, \cdots,$ or $/R_{N_a}^t/$ には間違いも含まれる．

Fig. 2 は被模倣音発見器の更新システムの概要を示している．ロボットは学習途中の発見器を使って MFCC 特徴量 $\boldsymbol{m}_n^t$ がロボットの発話したプリミティブ $/R_k^t/$ のうちのどれか 1 つに対応しているかを判定する．ここで，対応していると判定されたロボットの発話プリミティブを $/\hat{v}_{i,j}/_n^t$ とする．そして，対応していると判定された $/\hat{v}_{i,j}/_n^t$ と $\boldsymbol{m}_n^t$ をセットにして記憶する．T ターンのインタラクション終了後，記憶されたすべての $/\hat{v}_{i,j}/_n^t$ と $\boldsymbol{m}_n^t$ の組み合わせをもとに被模倣音発見器の更新を行う．

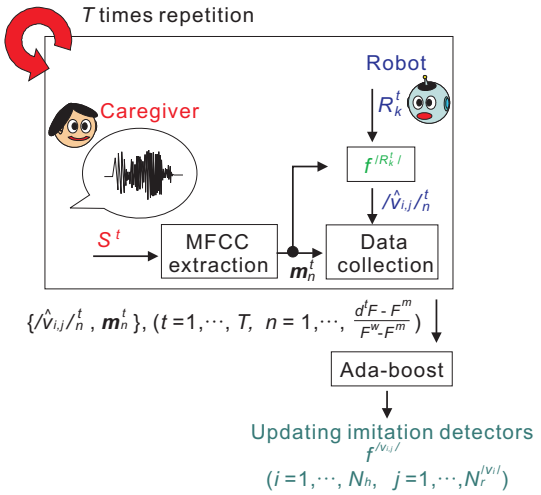


Fig.2 An overview of updating imitation detectors

3·1.1 MFCC (Mel-Frequency Cepstral Coefficient)

$\boldsymbol{m}_n^t$ は音響特徴量としてよく知られる MFCC を用いた 25 次元で表現される $(\boldsymbol{m}_n^t = [m_n^t(1), \cdots, m_n^t(25)]^T, (n = 1, \cdots, \frac{d^t F - F^m}{F^w - F^m}))$．ここで F^w はサンプリング窓，F^m はサンプリング窓の移動幅である．また，$\boldsymbol{m}_n^t$ の 1 次元目の要素は音の大きさ，2〜13 次元目の要素はメルスケール上でのスペクトルの強さ，14〜25 次元目の要素はスペクトルの強さの微分を示している．

3·1.2 被模倣音発見器

被模倣音発見器によってロボットは $\frac{F^w}{F}$ 秒毎に模倣の有無を判定する．ただし，$\boldsymbol{m}_n^t$ に対応するプリミティブ $/\hat{v}_{i,j}/_n^t$ は発話したプリミティブ $/R_k^t/$ のどれか 1 つから選択される．

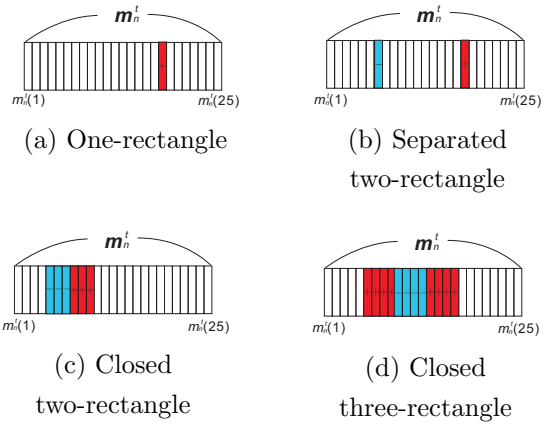
$$/\hat{v}_{i,j}/_n^t = \underset{/R_k^t/}{\arg\max} f^{/R_k^t/}(\boldsymbol{m}_n^t) \tag{1}$$

また，すべての $f^{/R_k^t/}(\boldsymbol{m}_n^t)$ が閾値（$= 0$）よりも小さい場合，$\boldsymbol{m}_n^t$ はどのプリミティブにも対応しない雑音として認識される．ロボットは T ターンのインタラクショ

ンを通じて $\boldsymbol{m}_n^t$ と $/\hat{v}_{i,j}/_n^t$ の組み合わせを獲得し，これらを用いて Ada-boost による被模倣音発見器 $f^{/v_{i,j}/}$ の更新を行う．さらに，T ターンのインタラクションを L 回繰り返し，繰り返しのたびに教師あり学習を行うことで被模倣音発見器の能力を向上させる．

3·1.3 弱学習機の設計

Viola and Jones [6] の示した Ada-boost を用いた顔認識を参考に，教示者の応答の MFCC 特徴量 $\boldsymbol{m}_n^t$ を入力とする 534 個の弱学習機を用意した．そして，こららの弱学習機をもとに Ada-boost による被模倣音発見器の獲得を行った．用意した弱学習機は Figs. 3 に示す 4 種類あり，$\boldsymbol{m}_n^t$ 中の赤色で表示された次元の特徴量の和から青色で表示された時限の特徴量の和を引いた値で表現される．各種類ごとの学習機の数はそれぞれ (a) が 25 通り，(b) が 253 通り，(c) が 156 通り，(d) が 100 通りである．



(a) One-rectangle　(b) Separated two-rectangle

(c) Closed two-rectangle　(d) Closed three-rectangle

Fig.3 Example of rectangle features

3·1.4 ロボットによる能動的な発話のサンプリング

ロボットの各発話プリミティブの発話割合 $Q_l^{/v_{i,j}/}$ $(0 \leq Q_l^{/v_{i,j}/} \leq 1,\ l = 1, \cdots, L)$ は，l 回目の教師あり学習において各 $/v_{i,j}/$ に対して教示者の模倣が起きたと被模倣音発見器によって判定された回数によって定義される．本研究において模倣回数は各母音 $/v_i/$ 毎に分類されており，ロボットが母音 $/v_i/$ を発声するときに $/v_{i,j}/$ が選択される割合 $Q_l^{/v_{i,j}/}$ は以下の式で与えられる．

$$Q_l^{/v_{i,j}/} = \frac{I^{/v_{i,j}/}}{\sum_j I^{/v_{i,j}/}} \tag{2}$$

ここで $I^{/v_{i,j}/}$ は $(l-1)$ 回目の教師あり学習における $/v_{i,j}/$ に対する模倣の回数である．ただし，$l = 1$ における $Q_1^{/v_{i,j}/}$ の値はすべて同じ値とする．

3·1.5 初期学習

ロボットに被模倣音発見器の初期値を与えていないため，1 回目の教師あり学習では被模倣音発見器による模倣の判定を用いた学習を行うことができない．そこで，t ターン目のインタラクションで教示者の応答 S^t のどこかにロボットの発話したプリミティブ $/R_1^t/, \cdots,$ $/R_{N_a}^t/$ が含まれているとした弱い教師あり学習によって被模倣音発見器の初期値の獲得を行う．Fig 4 は 1 回目の弱い教師あり学習による被模倣音発見器の初期値獲得の概要を示している．この 1 回目の weakly-supervised learning では教示者の発話を分節化することができないため，ロボットは S から抽出した $\boldsymbol{m}_n^t$ の平均値 $\boldsymbol{m}^t$ を使用している．

$$\boldsymbol{m}^t = \frac{\sum_n \boldsymbol{m}_n^t}{\frac{d^t \cdot F - F^m}{F^w - F^m}} \tag{3}$$

そして，t ターン目のインタラクションでロボットが発話したすべてのプリミティブ $/R_1^t/, \cdots,$ $/R_{N_a}^t/$ とその時の教示者の応答 $\boldsymbol{m}^t$ の T セットの組み合わせを用いて学習を行う．

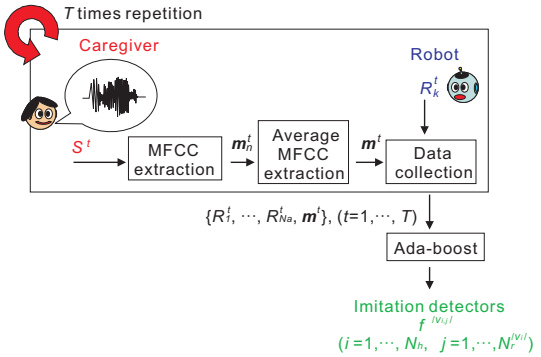


Fig.4 An overview of initial weakly-supervised learning

4. 実験

計算機シミュレーションで，提案手法により以下の 2 つの内容を実現できるかを確認する．1 つは教示者による模倣を発見し自身のプリミティブと教示者の母音との対応関係を獲得することである．もう 1 つはロボットの発話 $/v_{i,j}/$ に対して教示者が $/v_i/$ であると認識できる確率 (認識率)$P^{/v_{i,j}/}$ が高いプリミティブの発話頻度を増加させていくことである．

4·1 実験設定

シミュレーションでは，初めにあらかじめ録音された教示者の発話がランダムに選択される．そして，選択された発話につけられた母音ラベルに対し模倣率 P と認識率 $P^{/v_{i,j}/}$ をもとにロボットの発話するプリミティブ $/R_k^t/$ を選択する．インタラクションにおいて発声されるプリミティブの数 $N_a = 3$ であり，教示者は日本語の母音を発声するため，$N_h = 5$（$/v_i/ = /a/, /i/, /u/, /e/,$ or $/o/$）である．ロボットが持つ各母音グループ $/v_i/$ に属するプリミティブの数 $N_r^{/v_i/} = 3$ であり，ロボットは全部で 15 個のプリミティブを持つ．また，認識率はそれぞれ $P^{/v_{i,1}/} = 0.4$，$P^{/v_{i,2}/} = 0.7$，$P^{/v_{i,3}/} = 1.0$ とした．

発話を録音した教示者は日本人男性であり，日本語の 5 母音から重複を含めて 3 音を選び（125 通り），連続で発声させた．これを 8 セット行い全部で 1000 通りの教示者の発話を準備した．録音時のサンプリング周波数は 16000[Hz] である．また，日本語 5 母音を単

音でそれぞれ 10 回ずつ発声したときの録音データをロボットの学習結果に対する評価に使用した．

教師あり学習は最初の 1 回を学習した後，9 回繰り返し学習 ($L = 10$) を行う．最初の 1 回の学習回数は 1000 回 ($T = 1000$)，その後は各 100 回 ($T = 100$) ずつとした．そして，学習毎に獲得された被模倣音発見器を用いて，評価用の 50 通りの単母音がロボットのどのプリミティブに対応するかを確認した．評価用の音声 S はサンプリング窓 512 点，移動幅 256 点毎に被模倣音発見器によってロボットのどの発話プリミティブに対応するか判定され，全発話区間を通して最も多く対応すると判定されたプリミティブに対応する音声として判定される．その後，プリミティブの母音ラベル $/v_i/$ と実際の評価音声の母音ラベルが一致する割合（正答率）を評価する．

4·2　実験結果

模倣率を $P = 0.2, 0.5, 0.8$ に設定し，10 回のシミュレーションを行った．そのときの正答率の平均値と標準偏差を Fig. 5 に示す．Fig. 5 における赤，緑，青の線はそれぞれ模倣率 $P = 0.2, 0.5, 0.8$ に対応する結果を示している．

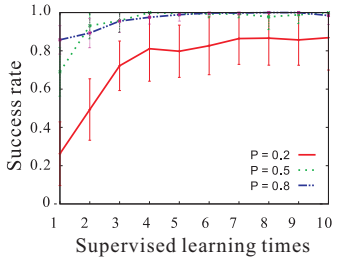
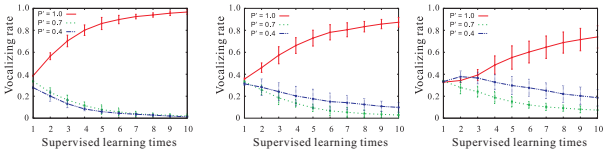


Fig.5 Comparison of success rate of detecting being imitated when the caregiver replied with vowels and the propagation rate $P^{/v_{i,j}/} = 1.0$

Fig. 5 より，模倣率 P が高ければ正答率は初めから高い値を示すことがわかる（青線を参照）．一方で模倣率が低い場合，最初の 1 回の弱い教師あり学習だけでは正答率は 0.3 程度と低い数値を示している．しかし，教師あり学習を繰り返すことにより正答率が 0.8 程度まで回復しており，ロボットによる能動的な発話・学習データのサンプリングの効果が確認できる（赤線を参照）．

次に，ロボットの発話プリミティブの発話割合が能動的な発話・学習データのサンプリングによってどのように変化していくかを分析した．Figs. 6 は認識率毎に発話プリミティブの発話割合の変化の平均値と標準偏差を示したものである．これら Figs. 6 の (a)，(b)，(c) はそれぞれ模倣率 $P = 0.8, 0.5, 0.2$ の実験結果を示している．

Figs. 6 の結果より，認識率の最も高い発話プリミティブのみが学習を繰り返すことにより発話割合を増加させていくことがわかる．この最も認識率の高い発話プリミティブはロボットの持つ明瞭な母音としてみなすことができる．そのため，Fig. 5，Figs. 6 の結果から，インタラクションを通じてロボットが自身の発話プリミティブと教示者の母音との対応関係を獲得し，明瞭な母音を頻繁に発話するようになることを示すことができたと考えられる．



(a) P=0.8　(b) P=0.5　(c) P=0.2

Fig.6 The average changes in utterance frequency of each primitive when when the caregiver replied with vowels

5.　考察

実験結果より，ロボットが教示者の母音と自身の発話との対応関係を獲得し，教示者にとって明瞭な母音の発話を増やせることを確認した．これらの結果は Fig. 5 の模倣率 $P = 0.2$ の条件や Fig. 6 (c) においても同様である．この模倣率 $P = 0.2$ はロボットと教示者が互いに日本語 5 母音の中からランダムに 1 音選んだときに偶然一致する確率と同じである．このことから，乳児が親の発話に偶然含まれた模倣をきっかけに親との母音の対応関係を獲得し，インタラクションにおいて母音様の発話を増加させることができる可能性があるのではないかと考えられる．

本実験において教示者は 1 人であったため，ロボットがそれぞれに異なる音響特徴を持つ複数の話者との間で母音の対応関係を同時に獲得できるかは不明である．この問題を解くためには，教示者の複数音節を発話する時の音節間の音節音響特徴の変化など個人に強く依存しない音響特徴を用いることを考える必要がある．

参考文献

[1] Katsushi Miura, Minoru Asada and Yuichiro Yoshikawa, "Unconscious Anchoring in Maternal Imitation that Helps Finding the Correspondence of Caregiver's Vowel Categories." RSJ Advanced Robotics Special Issue on Imitative Robots, vol. 21, no. 13, pp. 1583-1600, September 2007.

[2] N. Masataka and K. Bloom, "Acoustic Properties That Determine Adult's Preference for 3-Month-Old Infant Vocalization." Infant Behavior and Development, vol. 17, pp. 461-464, 1994.

[3] M. Peláez-Nogueras, J. L. Gewirtz, and M. M. Markham, "Infant vocalizations are conditioned both by maternal imitation and motherese speech." Infant behavior and development, vol. 19, pp. 670, 1996.

[4] Yuichiro Yoshikawa and Minoru Asada and Koh Hosoda and Junpei Koga, "A Constructivist approach to infants' vowel acquisition through mother-infant interaction." Connection Science, vol. 15, no. 4, pp. 245-258, December 2003.

[5] Patricia K. Kuhl, "Plasticity of development", chapter 5 Perception, cognition, and the ontogenetic and phylogenetic emergence of human speech., pp. 73-106, MIT Press, 1991.

[6] Paul Viola and Michael Jones, "Rapid Object Detection using a Boosted Cascade of Simple Features." IEEE Computer Vision and Pattern Recognition, 2001.