サービスロボットのための非モノローグ HMM による音声合成

○杉浦孔明†，志賀芳則†，河井恒‡，翠輝久†，堀智織†
† 情報通信研究機構 (NICT)　　‡ KDDI 研究所

1. はじめに

ロボットが日常生活に浸透するにしたがって，人とロボットのコミュニケーションが課題となる [13]．どれほど豊富な機能を持つロボットを構築したとしても，ユーザが手軽に機能を使えなければ普及しないであろう．ロボットとの音声によるコミュニケーションは，ユーザにとって手軽であるというメリットがあるが，実現は簡単ではない [15]．

本論文では，サービスロボット向けの自然かつ親しみやすい音声合成手法を提案する．合成音声の品質は近年飛躍的に改善し，数々の製品やオープンソースソフトウェアが生まれている．しかしながら，サービスロボットの開発者にとって，現状の合成音声は満足できる品質であるとは言い難い．ロボット開発者は人-ロボット対話を用途として合成音声を用いたいというニーズがあるが，一般的な音声合成器はそのような用途に最適化されている訳ではない．そのため，ロボットの合成音声は自然さや親しみやすさに欠け，抑揚が適切でないためユーザが質問されたことに気づかないなどの事態も起こり得る．

一方，表現力の高い合成音声の研究開発は，音声合成分野を始めとして，音声合成の応用先である音声対話システムやロボティクスなどの分野で研究がなされてきた（例えば [2,3,11]）．[1] では，擬人化ロボットの合成音声に対し，感情ごとに合成パラメータを人手で変更し，被験者に感情を推定させる試みが報告されている．音声対話システム向けの合成音声を志向した研究としては，[4] が挙げられる．[4] では，数種類の感情に対して，声優に台本を読ませ別々の隠れマルコフモデル（HMM）を学習させている．対話コーパスを用いて HMM を学習させた例としては，[7,9] などがある．

本論文では [9] を拡張し，非モノローグコーパスを用いた音声合成手法を提案する．これまで，音声合成に用いるコーパス収録では話者が単独で台本を読む「モノローグ」収録が一般的であった．このような収録はナレーション向け音声合成に向き，コストが安いというメリットがあるものの，会話的特徴が失われることが多い．一方，提案手法である非モノローグ収録では，声優に掛け合いで対話を行わせて収録を行う．本論文では，このようにして収録した音声で学習させた HMM から音声合成を行った実験について報告する．

提案手法の独自性は以下である．

- 声優 2 名による収録を行いモノローグおよび非モノローグコーパスを構築し，各コーパスを用いて学習させた HMM を用いて音声合成を行う．
- 提案手法をサービスロボット（図 1）に適用し，合成音声の品質を Mean Opinion Score(MOS) により定量評価する．

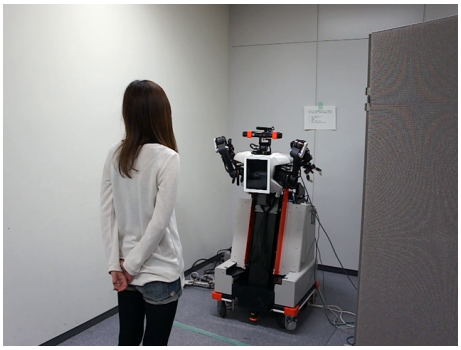

図 1 実験で用いたロボット Daia と実験環境

2. 非モノローグ音声合成

提案手法の概略を図 2 に示す．本論文では，我々が構築した京都観光案内対話コーパス [10] を用いる．本コーパスは，観光客・ガイドの間の観光プラン立案対話を収録したものである．コーパスには，328 組の観光客・ガイドによる音声対話（160 時間分）が含まれている．このうち，対話のやり取りが活発な 21 対話を抽出して書き起こしを行い，後述する収録で用いる台本を製作した．表 1 に対話例を示す．

次に，図 3 に示す防音室で，声優 2 名による掛け合い収録を行った．声優には，(1) 互いに相手の発話に重ならず，(2) 対話として自然なスタイルで台本を読む，ように指示を行った．発話ごとに人手で切り出し，無音部分を除いた結果，データ量は 1 名あたり 466[分] であった．データを表 2，3 のように学習セットとテストセットに分割し，学習セットのみを HMM の学習に用いた．以降，本論文では，上記のような収録を非モノローグ収録と呼び，上記の手順で得たコーパスと HMM を，それぞれ非モノローグコーパス，非モノローグ HMM と呼ぶことにする．また，非モノローグ HMM による音声合成を非モノローグ音声合成と略記する．

本研究の新規性は，非モノローグコーパスの利用である．そこで，提案手法とモノローグコーパスを比較するために，同じ声優によるモノローグコーパスを構築した．モノローグ収録では，通常の収録と同様に，声優 1 名のみの環境で音素バランス文などの台本を読むよう指示を与えた．その結果得たモノローグコーパスのサイズは 180[分] であり，そのうち 176[分] のデータを学習セットとした．

収録した音声を 16[kHz]，16[bit] でサンプリングした．スペクトル包絡と F0 を計算するために，STRAIGHT [5] を用いた．特徴ベクトルには，静的特徴量，1 次の差分 Δ，2 次の差分 $\Delta\Delta$ が 40 次元ずつ含まれ，合計 120 次元である．音響モデルとして隠れセミマルコフモデル（HSMM）を用い，HTS [12] を利用してパラメータ

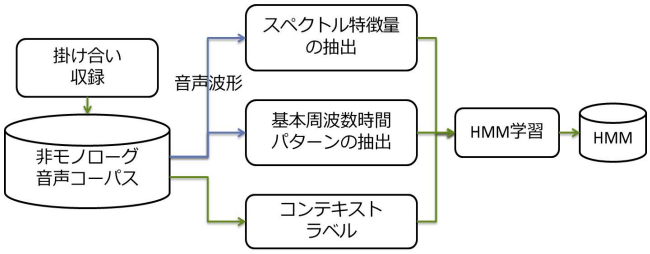

図 2 提案手法の概要

図 3 収録環境

を学習させた．HSMM の構造としては，5 状態の left-to-right model を用いた．人手による音素境界セグメンテーションは行わず，自動で行なっている．音声合成エンジンとして，NICT が開発した XIMERA [6] の後継である NX を用いた．

3. 実験設定

3·1 タスク

合成音声の品質に関して，ベースライン手法と提案手法の比較実験を行う．音声合成分野での一般的な評価法に従い，5 段階の Mean Opinion Score (MOS) 値を用いる．被験者は，20 代から 40 代までの日本語話者 24 名とし，各世代において男女同数とした．被験者数は，音声合成分野の代表的な文献と比較して十分なデータを得られるように設計した．専門家による偏った評価を避けるため，被験者に音声合成およびロボティクスを専門とする研究者や学生を含まないようにした．

実験に先立ち，分析合成音，ベースライン手法，および提案手法（3 種類）を被験者に聞かせ，合成音声の品質に慣れさせた．次に，刺激音声を被験者にランダムな順番で呈示し，自然さまたは親しみやすさを以下の 5 段階で評価させた．

5: とても良い，4: 良い，3: 普通，2: 悪い，1: とても悪い

表 1 非モノローグコーパスの対話例

ガイド	こんにちは.
観光客	こんにちは.
ガイド	京都の，観光プランを立てるお手伝いをさせていただきます，ヤマモトと申します.
観光客	はい.
ガイド	よろしくお願い致します.
観光客	こちらこそお願いします.

比較に用いた手法を表 2 に示す．収録した 2 名の声優のうち，一方の音声のみを一貫して実験で使用した．ベースライン手法は，モノローグコーパスを用いて HMM を学習させたものである．提案手法とベースラインに対して同じ量の学習データを用いた場合の結果を比較するために，Mono-176（モノローグ）と NonM-176（非モノローグ）を用意した．

応用分野を想定し，以下の 3 種類のタスク（読み上げ，対話，サービスロボット）を用意した．表 3 に各タスクにおける刺激の比較を示す．

- 読み上げタスク
  ATR503 [8] J セット（音素バランス文）に含まれる文を刺激（30 文）とした．これら 30 文は学習セットに含まれていないため，未知テキストに対する性能を調査することができる．
- 対話タスク
  音声対話システムへの適用を想定した合成音声に対する品質評価を行う．テストセットの各対話は，非モノローグコーパスのうち京都観光案内音声対話に関する対話を用いた．コーパスから学習に用いていない 12 対話を抽出したのち，一方の声優の発話を 5 種類の合成音声で置き換えた．他方の声優の発話は元の音声をそのまま用いた．
- サービスロボットタスク
  サービスロボットへの適用を想定した合成音声に対する品質評価を行う．そのために，ロボカップ@ホームタスク [14] で代表的に使用されるロボット発話から 12 文を抽出した．表 4 に刺激文を示す．被験者には，合成音声の親しみやすさについて 5 段階で評価させた．

3·2 実験環境

読み上げおよび対話タスクでは，受聴環境としてヘッドホンとノート PC を用いた．被験者に，各自の作業速度で評価作業を行わせた．このように，音声合成分野で一般的な環境と評価方法を用いた．

表 2 学習セットの比較

システム	収録スタイル	学習セットサイズ
(0) AS	分析合成音	-
(1) Mono-176 (ベースライン)	モノローグ	176 分 (2359 文)
(2) NonM-176 (提案手法)	非モノローグ	176 分 (4485 文)
(3) NonM-325 (提案手法)	非モノローグ	325 分 (8861 文)
(4) NonM-433 (提案手法)	非モノローグ	433 分 (14179 文)

表 3 刺激の比較

タスク	テストセットサイズ	被験者数
読み上げ	150 文 (30 × 5) (ATR503 J-set)	24
対話	60 対話 (12 × 5) (京都観光ガイド)	24
サービスロボット	60 文 (12 × 5) (ロボカップ@ホーム)	24

サービスロボットタスクでは，ロボットプラットフォーム Daia と被験者を対面させて実験を行った（図 1）．Daia は，上半身ヒューマノイド（川田工業製 HIRO），全方位移動台車（Neobotix 製 Omni-Drive-Module），レーザレンジファインダ（北陽電機製 UTM-30LX）2 台，RGB-D カメラ（Microsoft 製 Kinect），指向性マイクロホン（三研マイクロホン製 CS-3e，スピーカ（ヤマハ製 NX-U10）からなる．

サービスロボットタスクでは，実験を Wizard-of-Oz(WOZ) 形式で行った．すなわち，被験者にはロボットが自律的に動作すると伝えるが，実際には実験者 2 名が発話タイミングと頭部の制御を遠隔から行った．ロボットと被験者との距離は 1[m] とした．被験者の位置での等価騒音レベル L_{Aeq} は 66.7 [dB] であり，信号対雑音比（SNR）は 3.9 [dB] であった．これは，ロボカップ@ホーム環境とほぼ同等の騒音レベルであるといえる．

表 4 サービスロボットタスクでの刺激文

これから綿菓子を作ります，準備が整ったら合図して下さい．
すみません，物体をつかめません．
すみません．飲み物が見つかりません．
はい，分かりました．
ペットボトルを覚えました．
ペットボトルを把持します．
リビングルームまで移動します，いいですか？
時間になりました．
次の人は正面に立って下さい．
他の所で探してみます．
物体を取ってきます，いいですか？
綿菓子が完成しました．

4. 実験結果

4·1 類似度と合成音声の品質

前述したように，提案手法の新規性は非モノローグコーパスを学習セットとして使用することであった．ここで，そのようなコーパスとテストセットが類似していれば，合成品質が高くなるのは明らかである．仮に，刺激文とほぼ同じ文が学習セットに大量に含まれていたとする．その場合，学習されたモデルは声優の発話を再生していることとほぼ等しいので，合成音声の品質が高くなることは当然である．そのため，学習セットとテストセットの類似度を定量的に評価することが必要である．

本論文では，学習セットとテストセットの類似度としてテストセット対数尤度 $\mathcal{L}$ を用いる．$\mathcal{L}$ は，以下のように定義される．

$$\mathcal{L} = \frac{1}{N}\sum_{i}^{N}\frac{1}{T_i}\log p(\boldsymbol{O}_i|\boldsymbol{\lambda}_i), \tag{1}$$

ここに，N はテストセットサイズを表し，$T_i, \boldsymbol{O}_i, \boldsymbol{\lambda}_i$ は，i 番目のサンプルの，フレーム数，観測特徴量，単語列から得られた HMM，をそれぞれ表す．つまり $\mathcal{L}$ は，対数尤度をフレーム数で正規化したものの平均値である．$\boldsymbol{O}_i$ は声優が読み上げた音声から得られる．$\boldsymbol{\lambda}_i$ は，書き起こしと学習済みの HMM から得られる．

4·2 読み上げタスク

図 4 左図に読み上げタスクにおける類似度の結果を示す．縦軸はフレームごとの対数尤度を表す．図より，ベースラインは提案手法より $\mathcal{L}$ が高いため，ベースラインの HMM のほうがテストセットに類似しているといえる．これは，テストセットの文が小説などに見られる内容であったので，モノローグ的な発話になったものと考えられる．

以上の結果から，合成音声の MOS 値評価について，ベースラインの方が提案手法より優れると予想される．しかし実際には，両者に統計的に有意な差がないことを示す．

図 5 に提案手法とベースラインの MOS 値を示す．読み上げ，対話，サービスロボットの各タスクは，それぞれ左図，中図，右図に対応する．図において，(0) は分析合成音（理論上の上限）を示す．エラーバーは 95%信頼区間を示す．2 つの信頼区間が重なっていなければ，統計的有意差があるといえる．図中では，有意差を視覚的に示すため**($P < .01$) および***($P < .001$) を補助的に示している．

図 5 左図より，(1)Mono-176 と (2)NonM-176 の差は，統計的有意でないことがわかる．さらに，提案手法（NonM-325，NonM-433）は，Mono-176 より高い MOS 値を示すことがわかる．以上より，本実験条件に限定すれば，テストセットの文がモノローグ的であっても，非モノローグコーパスを使用することに品質上のデメリットはないといえる．

4·3 対話タスク

図 4 中図に，類似度 $\mathcal{L}$ を示す．図より，提案手法（NonM-176，NonM-325，NonM-433）の $\mathcal{L}$ は，ベースラインに比べて明確に高いことがわかる．

このことから，非モノローグ HMM とテストセットが類似していると考えられるので，提案手法の品質がベースラインと比較して優れると推測できる．実際に，図 5 中図において，提案手法（NonM-176，NonM-325，NonM-433）の MOS 値は，ベースラインより大きい．

4·4 サービスロボットタスク

図 4 右図に，類似度 $\mathcal{L}$ を示す．Mono-176 より NonM-176 の方がテストセットに類似しているが，その差は読み上げタスクや対話タスクの場合と比較して小さい．

図 5 右図に MOS 値による比較結果を示す．図より，提案手法（NonM-176，NonM-325，NonM-433）の MOS 値はベースラインと比べて高く，分析合成音（理論上の上限）に近い値であることがわかる．

以上の 3 つのタスクの結果を総合すると，提案手法は対話システムやロボットでは特に有効であると考えられる．また，読み上げタスクであっても，提案手法はベースラインと同等の品質を得られることが示唆される．

5. おわりに

ロボットのコミュニケーション機能の開発においては自然な音声合成が求められているが，一般的な音声合成器は人-ロボット対話に最適化されている訳ではない．そこで本論文では，ロボットの音声対話機能を志

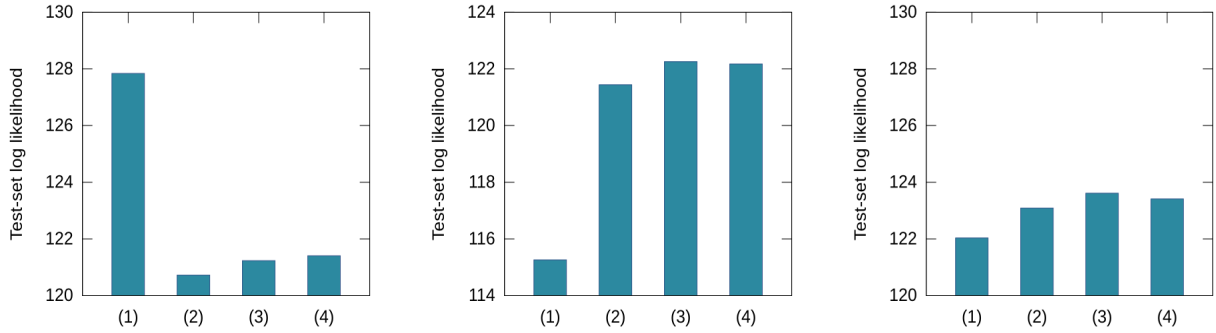

図 4 (1) Mono-176 (ベースライン)，(2) NonM-176，(3) NonM-325，(4) NonM-433 に対するテストセット対数尤度の比較．左図: 読み上げタスク．中図: 対話タスク．右図: サービスロボットタスク．

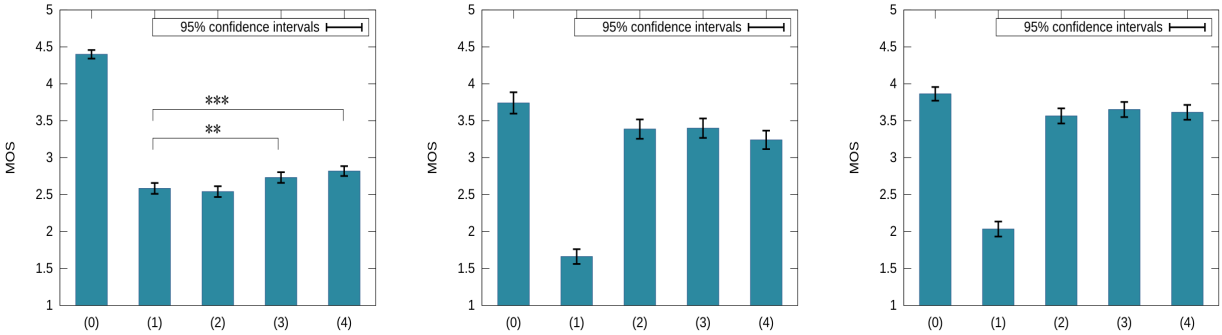

図 5 (0) 分析合成音 (理論上の上限)，(1) Mono-176 (ベースライン)，(2) NonM-176，(3) NonM-325，(4) NonM-433 に対する MOS 値．左図: 読み上げタスク．中図: 対話タスク．右図: サービスロボットタスク．**および***は，$P < .01$ および $P < .001$ を示す．

向した非モノローグ音声合成手法を提案した．サービスロボットにロボットへの応用を想定した被験者実験を行い，標準的な評価尺度である MOS 値による評価を行った．実験の結果，ベースライン手法に比べて品質が優れるという結果を得た．

謝辞

本研究の一部は，科研費（若手 (B)24700188，および新学術領域「人ロボット共生学」公募課題 24118710）の助成を受けて実施されたものである．

参考文献

[1] Breazeal, C.: Emotive Qualities in Lip-Synchronized Robot Speech, *Advanced Robotics*, Vol. 17, No. 2, pp. 97–113 (2003).

[2] Iida, A., Campbell, N., Higuchi, F. and Yasumura, M.: A Corpus-Based Speech Synthesis System with Emotion, *Speech Communication*, Vol. 40, No. 1, pp. 161–187 (2003).

[3] Ishi, C. T., Ishiguro, H. and Hagita, N.: Analysis of Acoustic-Prosodic Features Related to Paralinguistic Information Carried by Interjections in Dialogue Speech, *Proc. INTERSPEECH* (2011).

[4] Iwata, K. and Kobayashi, T.: Conversational Speech Synthesis System with Communication Situation Dependent HMMs, *Proc. of the Paralinguistic Information and its Integration in Spoken Dialogue Systems Workshop*, pp. 113–123 (2011).

[5] Kawahara, H., Masuda-Katsuse, I. and de Cheveigné, A.: Restructuring Speech Representations Using a Pitch-Adaptive Time–Frequency Smoothing and an Instantaneous-Frequency-Based F0 Extraction: Possible Role of a Repetitive Structure in Sounds, *Speech communication*, Vol. 27, No. 3, pp. 187–207 (1999).

[6] Kawai, H., Toda, T., Ni, J., Tsuzaki, M. and Tokuda, K.: XIMERA: A New TTS from ATR Based on Corpus-Based Technologies, *Proceedings of Fifth ISCA Workshop on Speech Synthesis*, pp. 179–184 (2004).

[7] Koriyama, T., Nose, T. and Kobayashi, T.: On the Use of Extended Context for HMM-Based Spontaneous Conversational Speech Synthesis, *Proc. INTERSPEECH*, pp. 2657–2660 (2011).

[8] Kurematsu, A., Takeda, K., Sagisaka, Y., Katagiri, S., Kuwabara, H. and Shikano, K.: ATR Japanese Speech Database as a Tool of Speech Recognition and Synthesis, *Speech Communication*, Vol. 9, No. 4, pp. 357–363 (1990).

[9] Misu, T., Mizukami, E., Shiga, Y., Kawamoto, S., Kawai, H. and Nakamura, S.: Analysis on Effects of Text-to-Speech and Avatar Agent in Evoking Users' Spontaneous Listener's Reactions, *Proc. of the Paralinguistic Information and its Integration in Spoken Dialogue Systems Workshop*, pp. 77–89 (2011).

[10] Misu, T., Ohtake, K., Hori, C., Kashioka, H. and Nakamura, S.: Annotating Communicative Function and Semantic Content in Dialogue Act for Construction of Consulting Dialogue Systems, *Proc. Interspeech*, pp. 1843–1846 (2009).

[11] Tsuzuki, R., Zen, H., Tokuda, K., Kitamura, T., Bulut, M. and Narayanan, S.: Constructing Emotional Speech Synthesizers with Limited Speech Database, *Proc. ICSLP*, pp. 1185–1188 (2004).

[12] Zen, H., Nose, T., Yamagishi, J., Sako, S., Masuko, T., Black, A. and Tokuda, K.: The Hmm-Based Speech Synthesis System (HTS) Version 2.0, *Proc. of Sixth ISCA Workshop on Speech Synthesis*, pp. 294–299 (2007).

[13] 西田豊明: 人とロボットの意思疎通, 情報処理, Vol. 44, No. 12 (2003).

[14] 杉浦孔明: ロボカップ@ホームリーグ, 情報処理, Vol. 53, No. 3, pp. 250–261 (2012).

[15] 杉浦孔明: ロボット対話 -実世界情報を用いたコミュニケーションの学習-, 人工知能学会誌, Vol. 27, No. 6, pp. 580–586 (2012).