重回帰 HSMM を用いたパラ言語情報を制御可能な対話音声合成の検討*

☆永田智洋, 森大毅 (宇都宮大), 能勢隆 (東工大)

1 はじめに

表情豊かな対話音声合成を実現するためには，パラ言語情報をいかに表現するかが重要となる。これまで，パラ言語情報は「怒り」や「悲しみ」といった範疇的な項目で表現されることが一般的であったが，対話音声に含まれるパラ言語情報は多様であるため，このような範疇的な項目のみで表現することは困難である。そのため，宇都宮大学パラ言語情報研究向け音声対話データベース (UUDB)[1] では，感情の次元説に基づく方法によってパラ言語情報が表現されている。また，合成音声のパラ言語情報を制御する手法に重回帰 HSMM を用いた手法がある [2]。重回帰 HSMM では，合成音声のパラ言語情報を低次元のベクトルを用いて制御している。そこで本研究では，UUDB に記述されているパラ言語情報と，重回帰 HSMM に基づく手法によって，対話音声合成におけるパラ言語情報を制御することを目的とする。

2 合成音声のパラ言語情報制御

UUDB では収録されている全ての発話に対して，音声から知覚される話者の感情状態を記述したパラ言語情報ラベルが付与されている。付与されているパラ言語情報ラベルは，話者の感情状態を表す「快-不快」,「覚醒-睡眠」，話者間の対人関係を表す「支配-服従」,「信頼-不信」，話者の態度を表す「関心-無関心」,「肯定的-否定的」という 6 つの次元を用い，各次元について 7 段階の評価を施すことによって記述されている。

重回帰 HSMM は HSMM における出力確率分布および状態継続長分布の平均ベクトルが重回帰モデルによって表現されると仮定する [2]。文献 [2] では，HSMM の状態 i における出力確率分布と継続長分布に単一のガウス分布を仮定したとき，それぞれの分布の平均ベクトル $\boldsymbol{\mu}_i$，m_i を次のように仮定する。

$$\boldsymbol{\mu}_i = \boldsymbol{H}_{bi}\boldsymbol{\xi}, \quad m_i = \boldsymbol{H}_{pi}\boldsymbol{\xi} \tag{1}$$

$$\boldsymbol{\xi} = [1, s_1, s_2, \cdots, s_L]^\top = [1, \boldsymbol{s}^\top]^\top \tag{2}$$

$\boldsymbol{H}_{bi}$，$\boldsymbol{H}_{pi}$ はそれぞれ出力確率分布の平均ベクトル，状態継続長分布の平均ベクトルに対する回帰行列であり，重回帰モデルにおける独立変数で構成されるベクトル $\boldsymbol{s}$ の次元数を L としたとき，$M \times (L+1)$, $1 \times (L+1)$ の行列となる。ここで，M は特徴ベクトルの次元数である。また，$\boldsymbol{\xi}$ は制御ベクトルである。

本研究では重回帰モデルの独立変数に，UUDB に記述されているパラ言語情報ラベルを用いることで合成音声のパラ言語情報の制御を行う。

$$\boldsymbol{\xi} = [1, v_{\mathrm{pl}}, v_{\mathrm{ar}}, \ldots]^\top \tag{3}$$

ここで，$v_{\mathrm{pl}}, v_{\mathrm{ar}}, \cdots$ はそれぞれ「快-不快（pleasantness)」,「覚醒-睡眠（arousal)」，$\cdots$ の項目に対応する変数である。したがって，式 (3) で求められる出力確率分布と状態継続長分布の平均ベクトルを用いて音声合成を行う。

3 実験

3.1 実験条件

学習には UUDB の対話セッション C002 から C007 に含まれる話者 FTS の 550 発話を用いた。ここで，550 発話の総時間は 15 分 50 秒である。スペクトルパラメータには，サンプリング周波数 16 kHz の音声信号から，分析周期 5 ms，分析窓長 25 ms のハミング窓を用いて求めた 0 次から 24 次のメルケプストラム係数を用いた。F0 パラメータは対数基本周波数とし，特徴ベクトルはこれらのパラメータにそれぞれの Δ, $\Delta\Delta$ パラメータを加えた 78 次元のベクトルとした。回帰行列の再推定に必要となる初期回帰行列は文献 [3] による初期化手法を用いた。制御ベクトルは，感情状態を表す一般的な指標とされている「快-不快」,「覚醒-睡眠」を用いた式 (4) とした。

$$\boldsymbol{\xi} = [1, v_{\mathrm{pl}}, v_{\mathrm{ar}}]^\top \tag{4}$$

また，各次元の値には UUDB に記述されているラベラ 3 名による平均評価値を用いた。テストセットには UUDB の対話セッション C001 の話者 FTS の発話を用いた。

評価実験はヘッドホンによる両耳聴取により行った。被験者は全員男性であり，音声の研究室に所属する大学院生 3 名及び大学生 5 名の合計 8 名とした。

3.2 パラ言語情報の再現性評価実験

パラ言語情報の可制御性を示すためには，合成時に付与したパラ言語情報が聞き手側に伝達されているかを確かめる必要がある。そこで，合成音声から

*Conversational speech synthesis with controllability of paralinguistic information using MRHSMM. by NAGATA, Tomohiro, MORI, Hiroki (Utsunomiya Univercity), NOSE Takashi (Tokyo Tech)

Table 1　パラ言語情報の再現性評価実験結果

	快-不快	覚醒-睡眠
相関係数	0.618	0.831

知覚されるパラ言語情報の評価実験を行った。合成する発話内容はテストセット中の 70 発話とし，合成時に付与したパラ言語情報は UUDB に記述されている 3 名のラベラによる平均評価値とした。評価方法は合成音声から受ける印象を,「快-不快」,「覚醒-睡眠」の項目について 7 段階で評価する形式で行った。

Table 1 に，合成時に付与したパラ言語情報と被験者による平均評価値の相関係数を示す。表より，どちらの次元に対しても正の相関が得られていることから，意図したパラ言語情報が伝達されていたことがわかる。また，合成時に付与した「快-不快」及び「覚醒-睡眠」の評価値の平均はそれぞれ 4.90，5.17 であるのに対し，被験者による平均評価値の平均はそれぞれ 4.19，4.74 となり，どちらの次元に対しても付与した評価値より低く知覚される傾向があった。

3.3　パラ言語情報の表出制御実験

3.2 では，言語情報がパラ言語情報の知覚に影響を及ぼしている可能性があった。そのため，ここでは言語情報を固定し，付与するパラ言語情報を変化させて評価実験を行った。評価に用いる発話内容はテストセット中の 25 発話とした。この発話内容には,「はい」や「うん」といった短い発話や,「そのよこにじいちゃんがすわってて」といった比較的長い発話が含まれている。付与したパラ言語情報の値は Fig. 1 に示されている 7 通りの値とし，各合成音声について付与した評価値の重複がないようランダムに入れ替えた 7 セットの計 175 発話の合成音声を呈示した。評価方法は合成音声から受ける印象を,「快-不快」,「覚醒-睡眠」の次元について 7 段階で評価する形式で行った。

Table 2 に，合成時に付与したパラ言語情報と被験者による平均評価値の相関係数を示す。こちらの結果についても,「快-不快」,「覚醒-睡眠」の両項目において正の相関が得られたことから，パラ言語情報が言語情報に依らずに伝達されていることが示された。

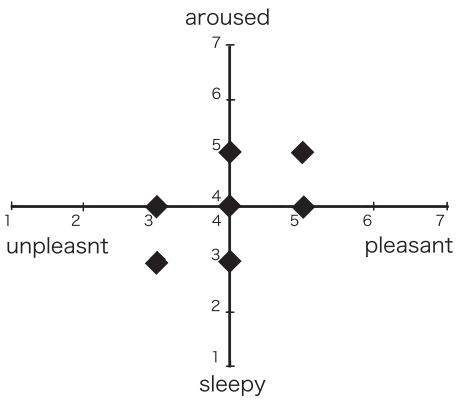

Fig. 1　合成時に付与したパラ言語情報

Table 2　パラ言語情報の表出制御実験結果

	快-不快	覚醒-睡眠
相関係数	0.488	0.646

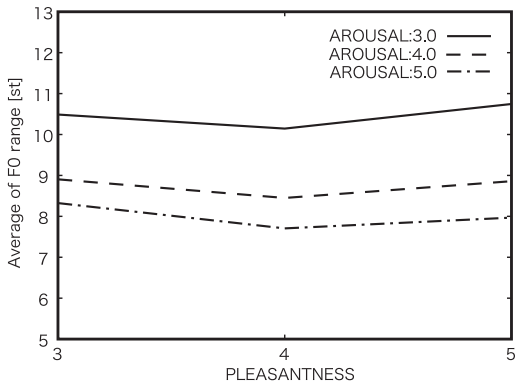

Fig. 2　各評価値における発話全体の F0 レンジの平均

合成時に付与したパラ言語情報の違いによる，合成音声の音響特徴量の変化について示すために，Fig. 2 に「快-不快」,「覚醒-睡眠」のそれぞれの項目に対し，3.0 から 5.0 まで 1.0 刻みで変化させた値を与えた場における合成音声の発話全体の F0 レンジの平均を示す。図より，F0 レンジは「快-不快」よりも「覚醒-睡眠」の値による影響を大きく受けている。「覚醒-睡眠」の相関係数が「快-不快」と比較して高くなっているのは，この影響を反映した結果であると考えられる。

4　おわりに

本研究では，重回帰 HSMM に基づく音声合成を用いて，対話音声合成におけるパラ言語情報の制御を行った。合成された音声に対して主観評価実験を行い，パラ言語情報が制御可能であることを示した。

今回は 2 次元によるパラ言語情報制御を行ったが，今後は他の次元を用いた制御についても検討する。

参考文献

[1] Mori et al., Speech Communication, 53, 36-50, 2011.
[2] Nose et al., IEICE Trans. Inf. & Syst., E90-D(9), 1406-1413, 2007.
[3] 能勢, 小林, 音講論 (秋), 329–330, 2011.