パラ言語情報正規化学習による表情豊かな対話音声合成の検討*

☆高橋俊介, 森大毅 (宇都宮大)

1　はじめに

筆者らは，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] をコーパスとして用いた，感情状態を制御可能な対話音声合成の研究を行ってきた。これらの研究ではいずれも決定木コンテキストクラスタリングによる音素単位のモデル学習を行なっているが，モデルに感情状態を反映させるアプローチが異なる。[2] は感情評価値を数値属性としてコンテキストに加え決定木の構築を行うことで，感情評価値の大小をパラメータ分布の違いに反映させるアプローチである。この手法では，データ量が不十分なコンテキストでは感情状態が反映されにくい傾向があった。一方，[3] は，感情評価値を説明変数とした重回帰 HMM を学習するアプローチである。この手法では学習データに含まれない感情の表現も可能であるが，感情評価値に対するパラメータの変化が線形であると仮定しているため，パラ言語の表現力に限界があった。

本研究では，第 3 のアプローチとして，話者正規化学習 (SAT)[4] の手法を応用し，平均的な感情状態のモデルから，目標の感情状態への適応化を行うことにより，自然な感情状態の表現を目指す。

発話スタイル適応を目標とした過去の研究 [5] との違いは，本研究では自発音声のコーパスを元にしている点，および感情の次元説に基づく表現によるパラ言語情報の制御が可能な点である。

2　パラ言語情報正規化学習による感情制御

SAT を用いた音声合成では複数話者間の特徴の違いを正規化し，平均的な特徴を持ったモデルを学習する。そして平均声モデルから目標話者の音声データを用いて適応化を行うことで特定の話者性を表現する。

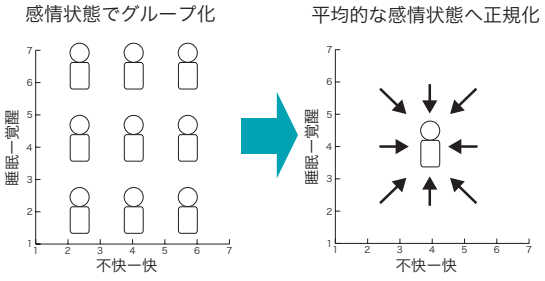

Fig. 1　感情状態の正規化

本研究では，感情状態の異なる同一話者の音声を，SAT における異なった話者の音声とみなす (Fig. 1)。学習データとして用いる発話には，その発話から知覚されるパラ言語情報が付与されていると仮定する。各発話は，その特徴がパラ言語的に平均的な発話に近づくよう線形変換される。この変換後の発話からパラ言語的に平均的なモデルが学習される。合成時には，パラ言語的平均モデルから目標とするパラ言語情報へ逆向きの線形変換を行う。この手法をパラ言語情報正規化学習と呼ぶ。パラ言語情報正規化学習を行うことにより，学習に使用可能なデータ量が少ない対話音声合成においても自然性の高いパラ言語情報の表現が可能となる事が期待される。

UUDB では収録されている全ての発話に対し，音声から知覚されるパラ言語情報の評価値が記述されている。パラ言語情報は,「快-不快」「覚醒-睡眠」「支配-服従」「信頼-不信」「関心-無関心」「肯定的-否定的」の 6 次元に対し，中立的な状態を「4」とした「1」から「7」の 7 段階評価が行われており，3 名の評価者による平均評価値がパラ言語情報ラベルとして付与されている。

各発話に対する平均評価値をそのまま適応目標とすると，適応データが少ない箇所において不自然に偏った適応化が行われる可能性がある。そこで，平均評価値が近い音声は知覚される感情状態が似ているとしてグループ化を行う。感情状態を「楽しい」「悲しい」等のカテゴリで表し，そのカテゴリを意識して発話された音声を用いて適応化を行う場合とは違い，聞き手がどのように感じたかをモデルに直接反映させ，感情の度合いや多様なニュアンスを表現することが可能であると考えられる。

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 551 発話を用いた。ここで，学習に用いた 551 発話の総時間は 16 分 12 秒である。学習データの作成条件は [2] と同様である。ただし，メルケプストラム係数を 35 次元とし，108 次元の特徴ベクトルを用いた。

パラ言語情報の制御には話者個人の感情を表す「快-不快」「覚醒-睡眠」の 2 次元を用いた。それぞれの感

*Expressive conversational speech synthesis using paralinguistic information-adaptive training. by TAKAHASHI, Shunsuke, MORI, Hiroki (Utsunomiya University)

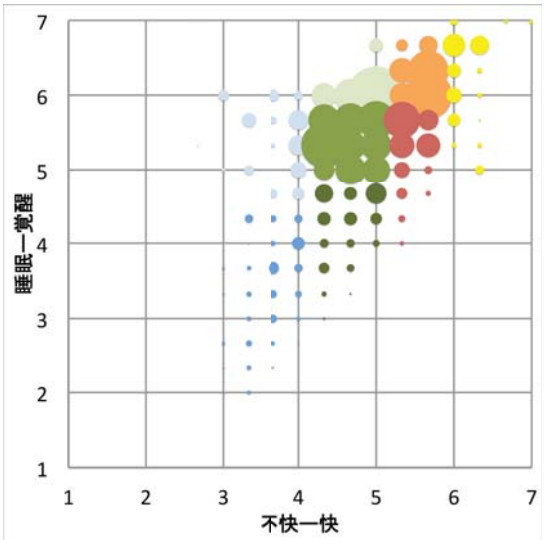

Fig. 2　話者 FTS の感情評価値の分布

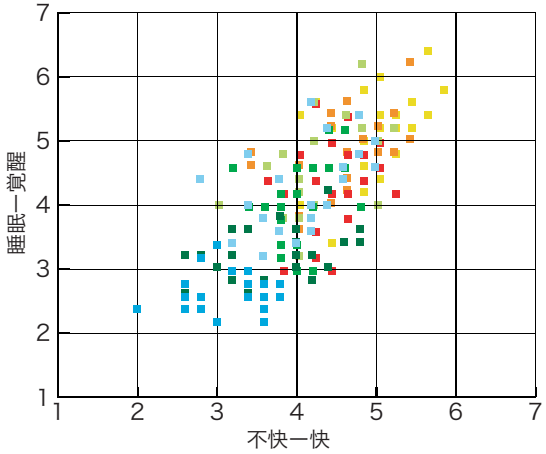

Fig. 3　各評価音声の平均評価値

情状態であると評価された発話に含まれる音素数を求め，各グループ内の音素数が少なくなり過ぎないように，感情評価値を基準に Fig. 2 のような 8 つのグループに分割を行った。このグループを用いてパラ言語情報正規化学習を行い，平均的な感情状態のモデルを学習する。その後，各感情のグループへ適応化することにより目標の感情状態の音声を合成する。

3.2　合成音声の感情表出評価実験

合成された音声が目標とした感情状態のとおりに知覚されるかを確かめるため，合成音声から知覚される感情の評価実験を行った。

評価には UUDB の対話セッション C001 に含まれる話者 FTS の発話内容から選択した 22 発話を用いた。適応化を行った 8 パターンの感情の異なるモデルから，同一内容の 22 発話を合成した。感情状態をランダムに入れ替えた 22 発話を 1 セットとし，計 8 セット全 176 発話に対して音声から知覚される感情を評価した。評価者は各音声を 1 度ずつ聴取し，どのような感情で発話されたと感じたかを判断する。評価方法は UUDB の評価基準に基づく 7 段階の評価とし，音声からどのような印象を受けたかを「快-不快」「覚醒-睡眠」の 2 次元に対して評価する。被験者は音声の研究室に所属する大学院生 5 名とした。

Fig. 3 に各音声の平均評価値の分布を示す。Fig. 3 を Fig. 2 と比較することにより，合成音声から知覚される感情状態は，目標とする感情状態，すなわち Fig. 2 の各グループの平均的な感情状態と類似した傾向を示していることがわかる。

Table 1 に，感情状態の目標値と知覚された感情状態の評価値との相関係数を示す。Table 1 に示すように，平均評価値と適応目標値との間にはどちらの次元に対しても相関が見られ，特に「覚醒-睡眠」の次元に関して高い相関が見られる。この結果から感情の次元に基づく評価値を基準に適応化を行うことで，異なる感情状態を表現した対話音声合成が可能であるといえる。

Table 1　各評価者の相関

	評価者					
	#1	#2	#3	#4	#5	平均
pleasant	0.35	0.48	0.45	0.51	0.36	0.58
arousal	0.71	0.56	0.70	0.65	0.51	0.78

4　おわりに

HSMM に基づく対話音声合成において，感情の次元に基づいたパラ言語情報正規化学習を行い，目標とする感情状態への適応化を行った。主観評価実験により，高い精度で意図した感情状態が知覚されていることがわかった。

本稿では適応データとして 8 グループに分割を行ったが，グループ化する感情評価値の範囲を変化させた際の影響等も検討したい。

参考文献

[1] Mori et al., Speech Communication, **53**, 36–50, 2011.
[2] Mori and Hitomi, Proc. Oriental COCOSDA 2012, 135–140, 2012.
[3] Nagata et al., Proc. Interspeech 2013(to appear).
[4] Yamagishi and Kobayashi, IEICE Trans. Inf. & Syst. **E90-D**, 533–543, 2007.
[5] 橘，小林，信学技報，SP2007-87，7–12，2007.