一般社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report
SP2012-121 (2013-02)

［ポスター講演］対話音声合成の自然性とパラ言語情報の可制御性に対する話者正規化学習の効果

高橋　俊介†　　森　　大毅†

† 宇都宮大学大学院工学研究科　〒 321–8585 栃木県宇都宮市陽東 7–1–2
E-mail: †{shun,hiroki}@speech-lab.org

あらまし　本稿では，HMM を用いた対話音声合成において，複数話者のコーパスを用いた学習による学習データ量の増加および話者正規化学習法が合成音声の自然性とパラ言語情報の可制御性にもたらす効果について検討する．本研究では，パラ言語情報を少数の次元から構成される空間上の座標として表現し，この座標を用いて決定木コンテキストクラスタリングを行うことで，多様なパラ言語情報を表現する．モデルの学習時に発話毎に与えられた評価値を用いて学習し，任意の評価値を与えた音声を合成し，それらの自然性とパラ言語情報の可制御性の評価を行った．
キーワード　HMM 音声合成，話者正規化学習法，話者適応

[Poster Presentation] The effect of Speaker Adaptive Training on the naturalness of conversational speech synthesis and the controllability of paralinguistic information

Shunsuke TAKAHASHI† and Hiroki MORI†

† Graduate School of Engineering, Utsunomiya University　Yoto 7–1–2, Utsunomiya-shi, Tochigi, 321–8585 Japan
E-mail: †{shun,hiroki}@speech-lab.org

Abstract　The Speaker Adaptive Training (SAT) was applied to the HMM-based dialogue speech synthesis based on the UU Database. Its effect on the naturalness and controllability of paralinguistic information was investigated. The naturalness for the SAT model was 3.2 in the 5-grade MOS, which was comparable to that for the speaker independent (SI) model, and slightly better than that for the speaker dependent (SD) model. The controllability test showed that the correlation coefficient between given and perceived Arousal for the SAT model was 0.64, which was significantly higher than that for the SD model, and slightly worse than that for the SI model.
Key words　HMM-based speech synthesis, SAT, speaker adaptation

1. はじめに

我々はこれまでに，決定木に基づくコンテキストクラスタリングにおいて，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) に記述されているパラ言語情報の評価値をコンテキストとして用いる事で，対話音声合成におけるパラ言語情報の制御を行ってきた [1]．しかし，特定話者では学習データが少ないため，合成音声の自然性が不十分であるだけでなく，合成時に与えたパラ言語情報が合成音声に十分に反映されない場合もあった．

そこで本研究では，平均声モデルの学習法として有効性が示されている話者正規化学習法 [2], [3] を用いた不特定話者学習により学習データ量を増やす事により，自然性が高くパラ言語情報を自在に表現可能な対話音声合成の実現を目指す．

2. 決定木コンテキストクラスタリングによる感情制御

本研究では，コンテキストラベルに「音素」,「モーラ位置」,「アクセント句」,「文長」の情報に加え，UUDB に付与されているパラ言語情報の評価値の平均値を与えている．このラベルにより決定木に基づくコンテキストクラスタリングを行い，感情状態の違いを考慮したモデル学習を行っている．

UUDB に収録されている音声において，話者毎に感情表出の偏り方が異なるため，複数の話者のデータを学習に用いること

表 1　話者毎の発話数と発話時間
Table 1　Time and the number of utterances of each speaker.

話者	発話数	発話時間	話者	発話数	発話時間
FTS	551	16 m 12 s	FTH	720	13 m 39 s
FNN	169	3 m 42 s	FSH	141	3 m 35 s
FTY	258	4 m 48 s	FUE	203	4 m 14 s
MKK	577	11 m 00 s			

により，多様なパラ言語情報の表現が可能になると考えられる．

3. 話者正規化学習の効果の検証

3.1 実 験 条 件

学習には UUDB に含まれる 14 名の話者の内，女性話者 6 名，男性話者 1 名の計 2616 発話 (57 分 13 秒) を用いた．話者毎の発話数および発話時間を表 1 に示す．学習データの作成条件は文献 [1] と同様である．ただし，メルケプストラム係数を 0 次から 34 次とし，108 次元の特徴ベクトルを用いた．コンテキストラベルには UUDB に付与されている 6 次元のパラ言語情報から，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」の 2 次元を用いた．特定話者学習 (SD) では，話者 FTS の 551 発話を学習データとした．不特定話者学習 (SI) では 7 名の話者の 2616 発話を学習データとし，話者 FTS の 551 発話を用いて話者適応を行った．話者正規化学習 (SAT) では SI で学習したモデルを初期モデルとし，SI と同様のデータを用いて SAT を行い，話者 FTS の 551 発話を用いて話者適応を行った．被験者は 4 名で，ヘッドホンによる両耳聴取により評価を行った．

3.2 合成音声の自然性評価実験

多様なパラ言語情報を表現するにあたり，様々な感情状態を与えて合成した音声の自然性が十分に確保できている必要がある．そこで，任意の感情状態のラベルを与えた合成音声を用いて，音声の自然性の主観評価実験を行った．合成する発話内容は，話者 FTS の発話から，学習に用いた 551 発話に含まれない 20 発話を用いる．感情状態のラベルには図 1 に示した 5 組の感情状態値を与えて合成した．実験には，20 発話を学習方法 (3 条件) および感情状態値 (5 条件) を変えて合成した 15 セット，計 300 発話を用い，呈示順序はランダム化した．評価は「肉声らしさ」と「対話音声らしさ」の観点を総合的に評価し，「自然：5」,「やや自然：4」,「どちらでもない：3」,「やや不自然：2」,「不自然：1」の 5 段階から選択する形式で行った．

合成時に与えた感情状態値毎の全被験者の平均評価値を表 2 に示す．表 2 より，SI を行った際に話者性の違いにより自然性が低下しているが，SAT を行うことにより自然性が向上していることが分かる．「覚醒-睡眠」に高い値を与えた際に，SI や SAT において自然性が高くなっている．これは，学習データにおける「覚醒-睡眠」の分布が広がり，決定木のノード分割に反映されたことによりモデルの精度が向上したものと考えられる．

3.3 感情状態の表出制御実験

3.2 節と同一の音声について，与えた感情状態値が合成音声から受ける印象に反映されているか評価を行った．評価方法は合成音声から受ける印象を 7 段階で評価する形式とした．

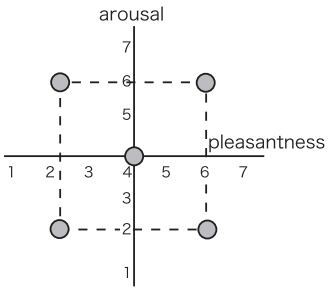

図 1　付与した感情状態値
Fig. 1　The emotional score that I give.

表 2　各学習方法における MOS 値
Table 2　MOS in each learning method.

感情状態値		学習方法		
Pleasantness	Arousal	SD	SI	SAT
2.00	2.00	3.15	2.73	3.10
2.00	6.00	3.00	3.53	3.58
4.00	4.00	3.38	2.86	2.90
6.00	2.00	3.29	2.89	3.19
6.00	6.00	3.33	3.50	3.56
	average	3.23	3.10	3.27

表 3　正解評価値と平均評価値の相関係数
Table 3　Correlation coefficient of correct score and average score.

感情状態	SD	SI	SAT
pleasantness	0.02	0.14	0.18
arousal	0.31	0.72	0.64

与えた感情状態値と評価値との相関係数の平均値を表 3 に示す．SD においてはどちらの軸に対しても相関が見られないが，SI および SAT において「覚醒-睡眠」について高い相関が得られた．しかし，SAT を行ったことにより「覚醒-睡眠」に関する相関が低下していることから，本手法においては SAT を行うことにより感情状態による音響特徴量の違いも緩和されてしまった可能性が考えられる．

4. お わ り に

本研究では，SAT を用いた対話音声合成を行い，次元説に基づくパラ言語情報の制御を行った．主観評価実験の結果から，SAT により自然性が向上する見込みが得られたが，パラ言語情報の表出において表現の幅を狭めている可能性も見られた．今後の課題として，パラ言語情報がよりクラスタリングに反映される学習条件の検討が挙げられる．

文　　献

[1] 人見貴嗣，“表情豊かな対話音声の合成に関する研究，” 宇都宮大学修士論文，2011.
[2] Anastasakos et al., “A Compact Model for Speaker-Adaptive Training,” ICSLP. 1137–1140,1996.
[3] 郡山知樹，能勢隆，小林隆夫，“平均声に基づく対話音声合成に関する検討，” 信学技報，SP2009–101, pp.33–38, 2010.