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
SP2014-90 (2014-11)

HMM に基づく対話音声合成におけるパラ言語情報制御手法の比較

森　　大毅† 　高橋　俊介† 　永田　智洋†

† 宇都宮大学大学院工学研究科
〒 321-8585 宇都宮市陽東 7 丁目 1-2
E-mail: †hiroki@speech-lab.org

あらまし　パラ言語情報の制御が可能な対話音声合成の実現を目指し，快-不快や覚醒-睡眠などの抽象次元に基づくパラ言語情報ラベルを有する対話コーパスに基づく HMM 音声合成を提案している．パラ言語情報制御を実現する方法として，これまでコンテキスト情報に基づく方法とパラ言語情報正規化学習・変換に基づく方法を提案してきた．本報告では，合成音声の自然性およびパラ言語情報の可制御性の観点からこれらの手法の有効性を比較検討する．16 名の被験者による主観評価実験の結果，変換に基づく方法では，コンテキスト情報に基づく方法と比較して，多くの場合で自然性に影響を与えることなくパラ言語情報の可制御性を向上できることがわかった．
キーワード　感情，ノンバーバル，話し言葉，対話コーパス，HMM 音声合成，コンテキスト，適応，正規化学習

A comparative study of paralinguistic information control methods for HMM-based dialogue speech synthesis

Hiroki MORI†, Shunsuke TAKAHASHI†, and Tomohiro NAGATA†

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

Abstract　Toward the realization of dialogue speech synthesis with capability to control paralinguistic information, we have proposed the HMM-based speech synthesis based on dialogue corpus that accompanies with paralinguistic information labels of abstract dimensions such as pleasantness or arousal. As methods for controlling paralinguistic information, context information-based method and adaptive training/conversion-based method have been proposed so far. In this report, effectiveness of these methods are compared from the viewpoints of naturalness and controllability of paralinguistic information. The results of subjective evaluation tests by 16 subjects revealed that the conversion-based method can better control paralinguistic information than the context information-based method, without sacrificing naturalness.
Key words　Emotion, nonverbal, spoken language, dialogue corpus, HMM-based speech synthesis, context, adaptation, adaptive training

1. はじめに

話し言葉は書き言葉と大きく異なる．話し言葉は，単に語を伝えるだけでなく，話者の意図，態度，感情状態などのパラ言語情報も運ぶ．音声言語が持つこれらのノンバーバルな側面は，音声によるコミュニケーションの本質と言ってもよい．音声対話システムや知的エージェントとの対話においても，状況に応じて話し方が適切に制御されることが，自然なコミュニケーションの実現のためには重要だと考えられる．対話音声合成とは，様相の上でも機能の上でも朗読音声 (=書き言葉の音読) とは異なる，話し言葉としての音声を合成する技術を指すために我々が用いている概念である．

このような対話音声合成の実現には，いくつもの課題がある．ここでは，とりわけ重要な 3 つの課題を挙げたい．

第 1 の課題は，そのような話し方の制御のため，音声合成器にどのような情報を入力すればよいか，という問題である．音声合成器への入力は当然に言語メッセージを含むが，言語メッセージを表現する方法が問題となることはほとんどない．これに対し，合成音声によって伝達されるべきパラ言語メッセー

ジ [1] を表現する方法は自明ではない．いわゆる感情音声合成の研究では，例えば対象を 6 大感情 [2] に限定し，伝達される話者の感情は 6 種類+中立のうちのどれかであるといった閉世界仮説に立脚することで，入力の表現の問題を自明化することが多い．しかし，6 大感情は感情の全てではない．さらに，感情そのものではないが感情に関連した話者の態度や気分などの状態もまた音声コミュニケーションの重要な構成要素である．パラ言語情報の表現の問題は，つきつめれば「音声は何を伝えているか」[1] を科学的に (再現可能な形で) 解明するという究極の問題となってしまう．

音声コミュニケーションのあらゆる側面をカバーするようなノンバーバル情報の普遍的なコード体系は目下存在しない．当面の解決策として，著者らは対話音声が伝達するパラ言語情報を，感情の次元説に基づいて定義された抽象次元により表現することを提案している．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) [3] はこの考えを体現するものである．対話音声の合成器への入力として UUDB と同様の抽象次元に基づくパラ言語情報の表現を用いることにより，極端な感情だけでなく，感情や態度などに関連した話し方の微妙なニュアンスを合成音声により表現できることが期待できる．

第 2 の課題は，合成音声のパラ言語情報制御手法である．合成音声のパラ言語情報制御とは，言語情報を一定とした上で，韻律・声質等の音響的特徴を変化させて多様なパラ言語情報が伝達されるようにすることを指す．技術的には，声質変換や合成音声の声質制御と密接な関係がある．近年発達したコーパスベース音声合成は，このような音声の多様性を合成音声で表現するための枠組として有効である [4]．上にも述べたような，6 大感情などを対象とした感情音声合成の場合には，単に必要な種類のコーパスを収録しさえすれば原理的には単純に実現できる．しかし，6 大感情のような極端な感情よりも広い範囲のパラ言語情報の制御を目標とした場合には，どのようなコーパスを用いるか，どのようなアノテーションが必要か，そしてそれをコーパスベース音声合成で実現するにはどうするか，といった問題を全て解決しなければならない．我々の一連の対話音声合成の研究は，UUDB をコーパスとした HMM 音声合成において，抽象次元により記述されたパラ言語情報が入力として与えられたとき，合成音声の韻律・声質をいかに制御すべきかを探究する試みである．

第 3 の課題は，対話音声合成をいかにして評価すべきかという問題である．通常のテキスト音声合成に求められる明瞭性や自然性 (肉声らしさ) は対話音声合成にも求められるが，実際の対話では明瞭でない発話も現れる．不明瞭性には，非流暢性 [5] と同様に話者の認知的・感情的状態のモニターとしての役割があり，音声コミュニケーションの無視できない構成要素である．したがって，一概に不明瞭な合成音声はよくない，とは言えない．また，一般にはパラ言語情報は言語情報と独立に自由自在に制御可能であるべきだと考えられることが多いが，感情などの話者の心理状態は言語メッセージとパラ言語メッセージの両方に同時に影響を与える [1] ため，これらは決して独立ではない．パラ言語情報の知覚実験によく見られる，発話内容を固定し韻律や声質だけを変化させる実験手続きは，実際にはない独立性の仮定の上に成り立っている点で危険をはらんでいる．対話音声合成の研究があまり進んでいない原因のひとつには，このように妥当な評価法がほとんど確立していない点も挙げられよう．

本報告は，主として第 2 の課題に関連した研究に関するものである．具体的には，HMM 音声合成による対話音声の合成におけるパラ言語情報制御を実現する方法として，これまでに提案した 2 手法の有効性を比較検討する．第 1 の手法は，UUDB に付与されたパラ言語情報の評価値を，HMM 学習時のコンテキストクラスタリングにおけるコンテキスト情報に追加する方法である [6]．第 2 の手法は，適応化に基づくパラメータの変換を利用した方法である [7]．

上述のように，対話音声合成の確立した評価法は存在しない．本報告では，自然性およびパラ言語情報の可制御性の観点からこれら 2 手法の有効性を比較検討する．

2. コーパスとアノテーション

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) [3] は，自然で表情豊かな対話音声に含まれるパラ言語情報の運用・構造および効果を明らかにする目的で設計・構築された，話し言葉のコーパスである．UUDB は，同学年の親しい友人同士である大学生による自然な対話音声から成る．対話参加者には，「4 コマまんが並べ換え課題」と呼ぶ，ばらばらにされた 4 コマまんがの 2 コマ分をそれぞれ持ち，対話により元の順番を推定する課題を遂行させた．この課題では，納得の行くストーリーを想像するプロセスや，女子学生が好むキャラクターを取り入れるなどの課題に没入させるための工夫により，極めて自発性が高く，言語的にも音声学的にも多様な対話音声が収録できる．

UUDB の各発話には，音声から知覚されるパラ言語情報のラベルが 6 次元ベクトルの形で付与されている．それらは

（1） 快-不快
（2） 覚醒-睡眠
（3） 支配-服従
（4） 信頼-不信
（5） 関心-無関心
（6） 肯定的-否定的

である．このうち，快-不快と覚醒-睡眠は，次元説 [8] に基づく感情研究において広く受け入れられている，話者個人の基本的な感情状態の評価項目である．

パラ言語情報ラベルの付与作業は，一貫性・代表性・評定項目の独立性の観点から選出された信頼できる評定者 3 名によって行われた [3]．評定者は，発話順に 1 発話ずつ聴取し，各次元に対し 1 から 7 までの 7 水準で評価した．各水準は，例えば，快-不快の次元では，1: 非常に不快, 2: かなり不快, 3: やや不快, 4: どちらでもない, 5: やや快, 6: かなり快, 7: 非常に快，である．

```
<Utterance UtteranceID="136" Channel="L" ...
  <EmotionalState>
    <Rating AnnotatorID="#23" Pleasantness="3" Arousal="6" ...
    <Rating AnnotatorID="#26" Pleasantness="3" Arousal="6" ...
    <Rating AnnotatorID="#27" Pleasantness="3" Arousal="6" ...
  </EmotionalState>
  <Chunk ChunkID="1" OrthographicTranscription="ディーで" ...
    :
    :
<Utterance UtteranceID="137" Channel="R" ...
  <EmotionalState>
    <Rating AnnotatorID="#23" Pleasantness="3" Arousal="2" ...
    <Rating AnnotatorID="#26" Pleasantness="4" Arousal="5" ...
    <Rating AnnotatorID="#27" Pleasantness="4" Arousal="3" ...
  </EmotionalState>
  <Chunk ChunkID="1" OrthographicTranscription="うん" ...
    :
    :
<Utterance UtteranceID="138" Channel="R" ...
  <EmotionalState>
    <Rating AnnotatorID="#23" Pleasantness="4" Arousal="3" ...
    <Rating AnnotatorID="#26" Pleasantness="4" Arousal="5" ...
    <Rating AnnotatorID="#27" Pleasantness="4" Arousal="4" ...
  </EmotionalState>
  <Chunk ChunkID="1" OrthographicTranscription="ああ" ...
    :
    :
<Utterance UtteranceID="139" Channel="L" ...
  <EmotionalState>
    <Rating AnnotatorID="#23" Pleasantness="6" Arousal="7" ...
    <Rating AnnotatorID="#26" Pleasantness="6" Arousal="7" ...
    <Rating AnnotatorID="#27" Pleasantness="6" Arousal="7" ...
  </EmotionalState>
  <Chunk ChunkID="1" OrthographicTranscription="あ" ...
    :
    :
<Utterance UtteranceID="140" Channel="R" ...
  <EmotionalState>
    <Rating AnnotatorID="#23" Pleasantness="4" Arousal="4" ...
    <Rating AnnotatorID="#26" Pleasantness="5" Arousal="5" ...
    <Rating AnnotatorID="#27" Pleasantness="4" Arousal="3" ...
  </EmotionalState>
  <Chunk ChunkID="1" OrthographicTranscription="はい" ...
    :
    :
```

図 1　UUDB XML 文書に記述された各発話のパラ言語情報ラベルの例

図 1 に，UUDB XML 文書の一部を示す．パラ言語情報ラベルは，各発話の EmotionalState 要素として，評定者ごとに記述されている．

3.　パラ言語情報の制御手法

コーパスベース対話音声合成の方式として，本研究では継続時間長分布をモデルパラメータとして持つ隠れセミマルコフモデル (HSMM) に基づく HMM 音声合成 [9] を利用している．本報告では，UUDB 中でも感情表現の豊かな話者である女性話者 FTS および FTH のペアによる 7 セッション分の発話をコーパスとして用いる．

以下では，著者らがこれまでに提案したパラ言語情報制御手法の中で，本報告で検討対象とした 2 手法について説明する．

3.1　コンテキスト情報による方法

HMM 音声合成における HMM の学習では，決定木によって似た分布を持つコンテキストを同一視するコンテキストクラスタリングが行われる．使用されるコンテキスト情報には，トライフォン・アクセント・モーラ数などの情報がある．ここに，各発話から知覚されるパラ言語情報を追加することで，パラ言語情報の相違による特徴の分布のバリエーションを説明するモデルが構築できる [6]．

図 2 に，感情状態が 3(やや不快) と 5(やや覚醒) と知覚された発話に対するコンテキストラベルを例示する．今回は，快-不快および覚醒-睡眠の各次元に対する 3 名の平均評価値をパラ言語情報コンテキストとした．この属性は順序尺度であるため，決定木の質問には各次元に対し考えられる全ての境界値に関する「より大きい」を追加した．

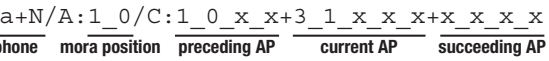

図 2　コンテキストラベルにパラ言語情報を追加した例

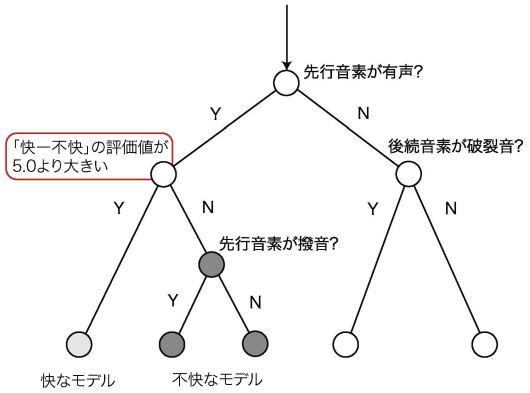

図 3　パラ言語情報を質問に加えた決定木の例

構築される決定木の模式図を図 3 に示す．この例では，快が 5.0 より大きい発話とそれ以外の発話の特徴の分布の違いが表現されている．

決定木を用いた手法は，既存のシステムに単純にコンテキストラベルと質問を追加するだけで実装できる容易さの点で優れている．しかし，自然な対話コーパスは読み上げコーパスに比べ多様であるためデータスパースネスが問題となる．データ量が十分ではない場合には，パラ言語情報によるデータ分割のためパラメータ推定が不安定となるおそれがあり，またこれを回避するために分割を抑制すると，パラ言語情報に関する質問が 1 度も適用されないリーフノードが多数となるため，異なるパラ言語情報を入力しても合成音声が変化しない場合が増加するという問題がある．

3.2　パラ言語情報正規化学習・変換による方法

HMM 音声合成における話者正規化学習・変換 [10] 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した上で，目標話者の音声データへ適応化を行うことで特定の話者性を表現する方法である．

パラ言語情報正規化学習 [7] の基本的な考え方は，パラ言語情報が異なる同一話者の音声を，話者正規化学習における複数話者の音声とみなすというものである (図 4)．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平均声のモデルが学習される．合成時には，パラ言語的平均声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．

抽象次元によるパラ言語情報表現は本来は連続値であるが，

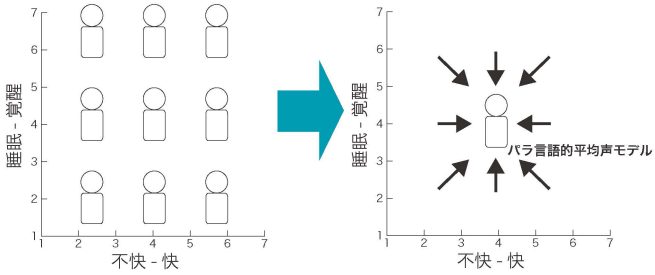

図 4　パラ言語情報正規化学習の概念図

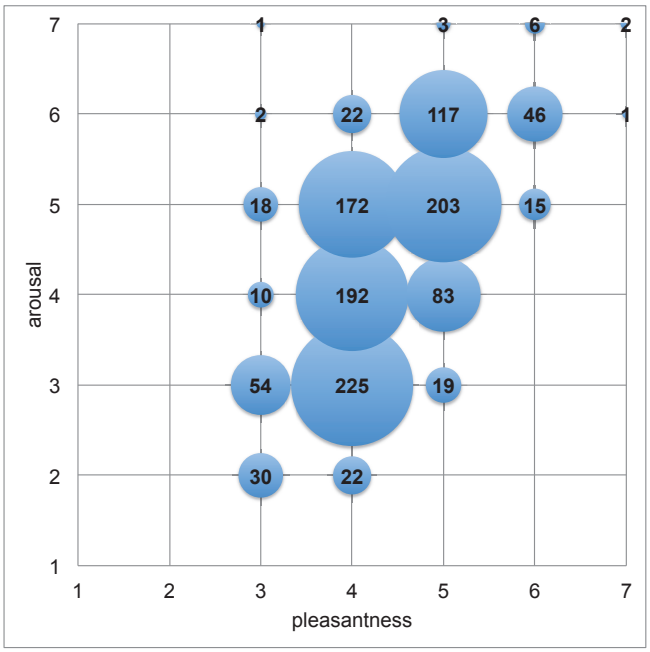

図 5　学習データのパラ言語情報の分布

パラ言語情報正規化学習・変換では原理上これを離散化する必要がある．今回は，3 名の評定者による評価値の中央値を求め，快-不快および覚醒-睡眠の両方の中央値が一致する発話を同一のパラ言語情報を持つものとみなした．実験のセクションで述べる学習データのパラ言語情報の分布を図 5 に示す．図中の数値は発話数を示している．

モデル学習の手順は以下の通りである．はじめに，初期モデルとして全学習データを用いた話者非依存・パラ言語情報非依存モデルを作成する．次に，初期モデルから各話者の各パラ言語情報のデータへの変換写像を CMLLR [11] により求め，逆写像により話者性およびパラ言語情報を捨象したデータへと変換し，このデータを用いて話者非依存・パラ言語情報非依存モデルのパラメータを再推定する．

合成時には，目標の話者およびパラ言語情報を表現するようにモデルを変換する．まず CMLLR によりモデルパラメータを変換する．次に，変換後のパラメータを事前分布とし，目標話者・目標パラ言語情報のデータから変換先のモデルパラメータを MAP 推定により求めることで，与えられた話者・パラ言語情報を表現するモデルが得られる．

パラ言語情報正規化学習では，パラ言語情報に関わらず，全ての発話からいったんパラ言語的平均声のモデルが学習される．コンテキスト情報による方法の場合には，パラ言語情報によりデータが分割されるため，データスパースネスの問題が助長されるおそれがある．パラ言語情報正規化学習ではこの問題が回避されるため，UUDB をコーパスとして用いる場合のように大量の学習データが確保できない状況でも安定したモデル学習が行えると期待される．さらに，複数話者のデータから話者・パラ言語情報同時正規化学習を行うことにより，学習データの不足に対応することも可能だと考えられる．

4. 実　　験

対話音声合成におけるパラ言語情報制御手法として，以下の比較検討を行う．

- コンテキスト情報による方法 (CON)
- パラ言語情報変換による方法 (PI-I)
- パラ言語情報正規化学習・変換による方法 (PI-AT)

CON は **3.1** で説明された方法であり，話者正規化学習により複数話者の学習データを併用している．PI-AT は **3.2** で説明された方法であり，話者正規化・変換も同時に行う．PI-I は PI-AT との比較のためのもので，話者非依存・パラ言語情報非依存モデルを単純に学習してから目標の話者およびパラ言語情報へ変換する方法である．

4.1　合成音声の作成

話者 FTS のセッション C002 から C007 までの 6 セッション (589 発話，17.4 分) と話者 FTH のセッション C001 から C007 までの 7 セッション (748 発話，14.7 分) を訓練データ，話者 FTS のセッション C001(96 発話) をテストデータとして共通に用いる．特徴ベクトルは，36 次元のメルケプストラム係数 (0 次を含む)，対数 F0，およびそれらの Δ と ΔΔ からなる 108 次元である．モデル学習のためのメルケプストラムは，分析周期 5 ms の STRAIGHT スペクトルを経由して推定する [12]．モデル学習のための F0 は，UUDB に対して付与された藤崎モデルパラメータ [13] に基づいてスムージングされたものを用いる [14]．モデル構造は 5 状態単一ガウス分布 left-to-right HSMM である．テストデータからのパラメータ生成においては GV は使用しない．音声合成のための音源には STRAIGHT の混合励振源を用いる [12]．

主観評価実験で呈示する発話は，セッション C001 の中から，多様なパラ言語情報を伝達し得る内容のものを，発話長のバランスを考慮して選んだ 15 文とした．

パラ言語情報評価実験において目標として合成音声に与えるパラ言語情報は，訓練データの分布を考慮して，4-3, 4-4, 4-5, 4-6, 5-4, 5-5, 5-6, 6-5, 6-6, 6-7 の 10 パターンとした (表記: P-A の P が快-不快，A が覚醒-睡眠)．呈示音声セットは，パラ言語情報制御手法 (3) × 目標 (10) × 文 (15) = 450 発話からなる．セット全体は，15 発話のサブセット#1 から#30 より構成されている．文の配列は全サブセットで同一であるが，パラ言語情報制御手法と適応目標の順序はランダム化している．

自然性評価実験において目標として合成音声に与えるパラ言語情報は，被験者の負担を考慮して 4-3, 4-6, 5-5, 6-7 の 4 パター

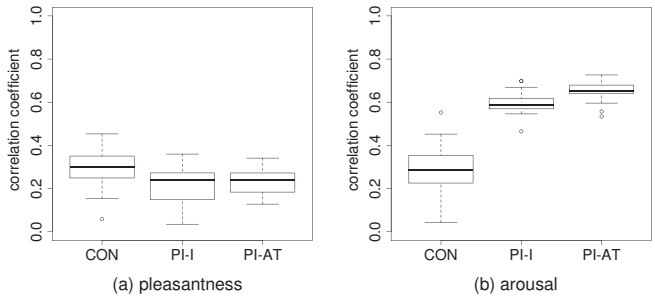

図 6　パラ言語情報制御の目標と各被験者の評価との相関係数: (a) 快-不快, (b) 覚醒-睡眠

表 1　パラ言語情報制御の目標と平均評価値との相関係数

	CON	PI-I	PI-AT
快-不快	0.443	0.328	0.326
覚醒-睡眠	0.438	0.760	0.811

ンに減らした．呈示音声セットは，パラ言語情報制御手法 (3) × 目標 (4) × 文 (15) = 180 発話からなる．セット全体は，15 発話のサブセット#1 から#12 より構成されている．呈示順序はパラ言語情報評価実験と同様にランダム化した．

4.2　主観評価実験

被験者は音声科学に関する特段の知識を持たない大学生 16 名である．実験に先立ち，被験者は著者のうち 1 名から感情心理学の基礎 (感情次元に関する内容を含む) についての講習を受け，ついで実験に使用しなかった合成音声 (全 50 発話) に対する評価の練習を行った．

パラ言語情報評価実験では，被験者は呈示音声サブセット#1 から#30 までの合成音声を順にヘッドホンにより 1 度ずつ両耳聴取し，合成音声から話者がどのような感情状態で発話しているように知覚したかを評価する．評価方法は，快-不快および覚醒-睡眠の各次元に対し，**2.** で述べた UUDB のパラ言語情報の評価と同様，1 から 7 までの 7 水準で行った．

自然性評価実験では，被験者は呈示音声サブセット#1 から#12 までの合成音声を順にヘッドホンにより 1 度ずつ両耳聴取し，合成音声の自然性を評価する．自然性は，「肉声らしさ」および「対話音声らしさ」の 2 つの観点から総合的に評価させた．評価方法は，1:不自然, 2:やや不自然, 3:どちらともいえない, 4:やや自然, 5:自然，の 5 水準で行った．

4.3　実験結果および考察

まずパラ言語情報評価実験の結果を示す．被験者一致性の尺度である級内相関係数は，快-不快で $\mathrm{ICC}(3,1) = 0.41$, 覚醒-睡眠で $\mathrm{ICC}(3,1) = 0.54$ であり，中程度の一致であった．図 6(a) に快-不快の目標と各被験者の評価との相関係数の分布を，図 6(b) に覚醒-睡眠の目標と各被験者の評価との相関係数の分布をそれぞれ示す．また，表 1 に目標と平均評価値との相関係数を示す．

各手法を比較すると，快-不快ではコンテキスト情報による方法がパラ言語情報変換による方法よりもやや高い相関係数となっている．しかし，いずれの方法も各被験者の評価との相関係数は 0.2 から 0.3 付近に多く分布している．したがって，快-不快の制御はある程度はできているが十分とは言えない．一方，覚醒-睡眠ではコンテキスト情報に基づく方法に比べパラ言語情報変換による方法が明らかに相関係数が大きい．このことから，覚醒-睡眠の制御に関してはパラ言語情報変換による方法が有効だと結論できる．また，PI-I と PI-AT を比較すると，快-不快では同程度，覚醒-睡眠ではやや PI-AT の方が相関係数が高くなっており，パラ言語情報正規化学習の効果が見られた．

表 2　自然性の平均オピニオンスコア

CON	PI-I	PI-AT
3.123	2.888	2.999

次に，自然性評価実験の結果を示す．表 2 に自然性の MOS(平均オピニオンスコア) を示す．3 手法の自然性はいずれも 3(どちらともいえない) 程度であった．パラ言語情報制御手法 (3) および目標 (4) を要因とする分散分析の結果，目標の主効果が有意 ($p < .01$) であったが制御手法の主効果は有意ではなかった ($p > .05$)．また，交互作用が有意 ($p < .01$) であったため下位検定を実施したところ，目標が 6-7 の場合に限って制御手法の単純主効果が有意 ($p < .01$) であり，CON の MOS(3.171) が PI-I および PI-AT の MOS(2.409, 2.630) に比べて有意 ($p < .01$) に高かった．

目標が 6-7 の場合に限ってパラ言語情報変換による方法である PI-I および PI-AT で自然性が低下した原因としては，この目標のデータが不足していたことが挙げられる．図 5 に示したように，学習データの中で 3 名の評価値の中央値が 6(かなり快)-7(非常に覚醒) であるものは全部で 6 発話に過ぎない．中央値を用いた離散化には，このように上下限の値を取るデータが極端に少なくなる傾向がある．変換の自然性を保つためには，データ数の偏りをある程度是正するようなグルーピングが必要と考えられる．

これらの結果をまとめると，大半の条件では自然性の点でパラ言語情報制御手法の違いはないが，データ量が少ない目標へ変化させようとした場合，パラ言語情報変換による方法である PI-I および PI-AT では自然性が低下する傾向があった，と言える．

以上の結果を総合してパラ言語情報制御手法の得失を述べると，パラ言語情報変換による方法では，コンテキスト情報による方法と比較して，多くの場合で自然性に影響を与えることなく覚醒-睡眠の次元の可制御性が大きく向上している．ただし，パラ言語情報の大きな変化に対しては，コンテキスト情報による方法は保守的ではあるが自然性の低下の少ない安全な方法だと考えられる．

5.　おわりに

HMM 音声合成による対話音声の合成におけるパラ言語情報制御を実現する方法として，対話コーパスに付与されている抽象次元で表現されたパラ言語情報を HMM 学習時のコンテキストクラスタリングにおけるコンテキスト情報に追加する方法と，適応化に基づくパラメータの変換を利用した方法について検討し，主観評価実験により自然性およびパラ言語情報の可制

御性を評価した．その結果，変換による方法では，コンテキスト情報による方法と比較して，多くの場合で自然性に影響を与えることなくパラ言語情報の可制御性を向上できることがわかった．

本報告では，パラ言語情報の中でも，合成音声から知覚される感情状態の制御だけを対象とした．しかし，感情状態はパラ言語情報の一部に過ぎない．話者のメッセージや状態を反映する別の側面を同時に伝達する合成音声の実現が今後の課題であるが，これは単に音声合成の技術的問題 (**1.** で述べた第 2 の課題) なのではなく，パラ言語情報の表現 (**1.** の第 1 の課題) および評価 (**1.** の第 3 の課題) と一体で解決を目指すべき問題であることを重ねて強調したい．

謝辞　本研究の一部は JSPS 科研費 26280100 の助成を受けている．

文　献

[1] 森　大毅，前川喜久雄，粕谷英樹，音声は何を伝えているか — 感情・パラ言語情報・個人性の音声科学—，コロナ社，2014.
[2] P. Ekman, “An argument for basic emotions,” Cognition & Emotion, vol.6, pp.169–200, 1992.
[3] H. Mori, T. Satake, M. Nakamura, and H. Kasuya, “Constructing a spoken dialogue corpus for studying paralinguistic information in expressive conversation and analyzing its statistical/acoustic characteristics,” Speech Commun., vol.53, pp.36–50, 2011.
[4] 能勢　隆，“統計モデルに基づく音声合成における話者・スタイルの多様化，” 信学技報，SP2012–109，2013.
[5] 伝　康晴，渡辺美知子，“音声コミュニケーションにおける非流暢性の機能，” 音声研究，vol.13，no.1，pp.53–64，2009.
[6] H. Mori and T. Hitomi, “Annotating conversational speech for corpus-based dialogue speech synthesizer — A first step,” Proc. Oriental COCOSDA 2012, pp.135–140, 2012.
[7] 高橋俊介，森　大毅，“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，” 日本音響学会研究発表会講演論文集 (秋)，pp.355–356，2013.
[8] J.A. Russell, “How shall an emotion be called?,” Circumplex Models of Personality and Emotions, eds. by R. Plutchik and H.R. Conte, pp.205–220, American Psychological Association, Washington, DC, 1997.
[9] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, “A hidden semi-Markov model-based speech synthesis system,” IEICE Transactions on Information and Systems, vol.E90-D, no.5, pp.825–834, 2007.
[10] J. Yamagishi and T. Kobayashi, “Average-voice-based speech synthesis using HSMM-based speaker adaptation and adaptive training,” IEICE Transactions on Information and Systems, vol.E90-D, no.2, pp.533–543, 2007.
[11] M.J.F. Gales, “Maximum likelihood linear transformations for HMM-based speech recognition,” Computer Speech and Language, vol.12, pp.75–98, 1998.
[12] H. Zen and T. Toda, “An overview of Nitech HMM-based speech synthesis system for Blizzard challenge 2005,” Proc. Interspeech 2005, pp.93–96, 2005.
[13] 渡邊諒馬，森　大毅，“表情豊かな対話音声の感情状態に関連する F0 モデルパラメータの検討，” 日本音響学会研究発表会講演論文集 (春)，pp.511–512，2013.
[14] H. Hashimoto, K. Hirose, and N. Minematsu, “Improved automatic extraction of generation process model commands and its use for generating fundamental frequency contours for training HMM-based speech synthesis,” Proc. Interspeech 2012, pp.458–461, 2012.