表情豊かな対話音声合成におけるコンテキストラベルの作成*

☆人見貴嗣, 森大毅 (宇都宮大院・工学研)

1 はじめに

音声合成技術の発達とともに，読み上げ調の音声だけではなく話し言葉音声の合成にも関心が向けられている [1–2] ．しかし，話し言葉にはパラ言語情報の多様性が含まれ，高い自然性を保ちながら合成することは容易でない．

パラ言語情報を考慮した音声合成の研究には．丁寧，ぞんざいなどの異なる発話スタイルを制御したもの [3] がある．しかし，日常的な会話に見られる表情豊かな対話音声は多様である．したがって，音声を合成するためにはコーパスの設計を含む新しい枠組みが必要である．

我々は対話音声に含まれるパラ言語情報に関する研究向けの音声対話データベース（以下，UUDB）[4] を開発している．本研究では，UUDB を音声コーパスとして利用した HMM 音声合成を行うことで，感情状態を制御できるような TTS システムの実現を目指す．

HMM 音声合成には，韻律などの変動要因を考慮したコンテキストラベルが必要である．しかし，UUDB には音声合成のためのコンテキストラベルや形態素情報などが付与されていない．よって本稿では音声合成の第一歩として，UUDB の形態素解析を行い，韻律の変動要因を考慮したコンテキストラベルの作成を検討した．

2 音声コーパス

UUDB には，合計 7 ペア，14 名の大学生が「4 コマまんが並べ替え課題」というタスクを行った対話音声が収録されており，日常会話のような，感情表現に富み生き生きとした対話が多く含まれている．全発話に対し，知覚される話者の感情状態が 6 次元の抽象次元（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）によってラベリングされている．

3 コンテキストラベルの作成

3.1 使用するコンテキスト

韻律の変動要因の中でも，アクセント情報は音響特徴量に最も影響することが示されている [5]．このようなことから，本研究では Table 1 に示すコンテキストを使用する．発話の中の言いよどみ，言い直しの部分には韻律ラベルは付与しなかった．

Table 1 コンテキストの種類

・{ 先行, 当該, 後続 } 音素
・当該音素のアクセント句内でのモーラ位置，およびモーラ位置とアクセント型の差
・{ 先行, 当該, 後続 } アクセント句の長さ，アクセント型
・発話の長さ（モーラ数）

3.2 形態素解析とアクセント句の同定

韻律ラベルは通常手作業で付与するものであるが，本研究では自動で作成することを試みた．UUDB の書き起こしテキストを形態素解析し，アクセント句の同定を行うことで，Table 1 に示したコンテキストを自動で作成した．UUDB の形態素情報付与には形態素解析器として茶筌 (ChaSen)[6] を，形態素解析辞書として UniDic[7] を使用した．また，茶筌の出力データからアクセント句の情報を推定するため，アクセント結合規則による処理を実装したアクセント型決定モジュール ChaOne[8] を使用した．

表情豊かな対話の書き起こしを自動で高精度に形態素解析することは難しい．UUDB の対話音声には単語同士の融合や省略などを伴うくだけた発話が多く含まれている．辞書に新しく単語登録などをして機械的に修正を行い，それでも修正できなかったものに対しては手作業で修正した．辞書の修正前後における形態素解析精度を Table 2 に示す．辞書の修正後になお誤解析した例を Table 3 に示す.「と言った」が融合して変化した「つった」の部分が「吊った」という誤った語に解析されている．

UUDB には発音形転記が与えられているが，正書法転記で形態素解析すると発音形転記と食い違った結果を出力する場合がある．Table 4 は形態素解析結果の読みを誤解析していた例である．このような例が生じ，形態素解析精度はあまり高くなかった．

4 HMM 音声合成

作成したコンテキストラベルを用いて，音声合成を行った．音声データは UUDB に収録されている女性話者 FTS の発話を用いた．これは，FTS の全発話数が比較的多かったためである．FTS の全発話を学習データとし，その発話の中の一部分を合成音声と

*Prosodic label generation for HMM-based conversational speech synthesis. by HITOMI, Takatsugu, MORI, Hiroki (Utsunomiya University)

して作成した（closed 実験）．

作成した合成音声の f_0 軌跡とサウンドスペクトログラムを Fig. 1 と 2 に示す．それぞれの合成音声は，明瞭性が高く，対話音声らしいリズムが感じ取れるものであった．

対話音声は，多様なパラ言語情報を伝達している．しかしながら，今回の HMM 音声合成ではパラ言語情報を表すようなコンテキストを使用していないため，書き起こしが同一の発話は全て同じ合成音声となってしまった．このことは「うん」に代表される表情豊かな対話音声を特徴付ける発話において不自然な印象を与える．合成音声における感情状態の制御について今後検討を行う必要がある．

5　おわりに

本報告では，UUDB で HMM 音声合成を行うために，その第一歩としてコンテキストラベルの生成方法を示した．UUDB の形態素解析には茶筌を，アクセント句の生成には ChaOne を用いてコンテキストラベルを自動で生成し，合成音声を作成した．今後は，作成した合成音声を用いて客観に基づく聴取実験を行う．

UUDB には話者の感情状態を記述したラベルが付与されている．今後の課題として，合成音声の音響的特徴を，感情状態ラベルによって制御しようと考えている．

参考文献

[1] 伊藤 他，“話し言葉音声合成の韻律制御に関する検討，” 情処研報，2009-NL-191，2009-SLP-76，1–8，2009.

[2] 郡山 他，“平均声に基づく対話音声合成に関する検討，” 信学技報，SP2009-101，33–38，2010.

[3] 大西 他，“HMM 音声合成における異なる発話スタイル生成の検討，” 信学技報，SP2002-172，17–22，2003.

[4] Mori et al.，Speech Communication (in press).

[5] 横溝 他，音講論 (春)，1-P-10，2010.

[6] http://chasen-legacy.sourceforge.jp/

[7] http://www.tokuteicorpus.jp/dist/

[8] 嵯峨山 他，“擬人化音声対話エージェントツールキット Galatea，” 情処研報，SPL96(55)，7–12，2003.

Table 2　辞書の修正前後における形態素解析精度

	修正前	修正後
誤解析形態素数	3239	3119
正解析率 [%]	87.49	87.95

Table 3　形態素「つっ」の修正例

修正前				
こいつ	何	つっ	た	っけ
代名詞	代名詞	動詞	助動詞	助詞
修正後				
こいつ	何	つっ	た	っけ
代名詞	代名詞	助動詞	助動詞	助詞

Table 4　誤解析した例

UUDB の 書き起こしテキスト	UUDB の 形態素解析後テキスト
カ イテ	エガ イテ
イッテ	オコナッテ
イ レル	ハイ レル
ホカ ノヒト	タ ノヒト

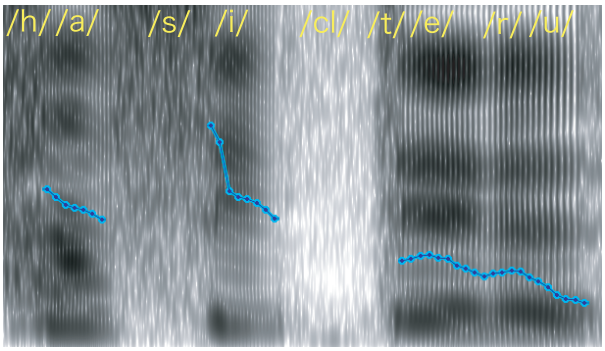

Fig. 1　合成音声「走ってる」

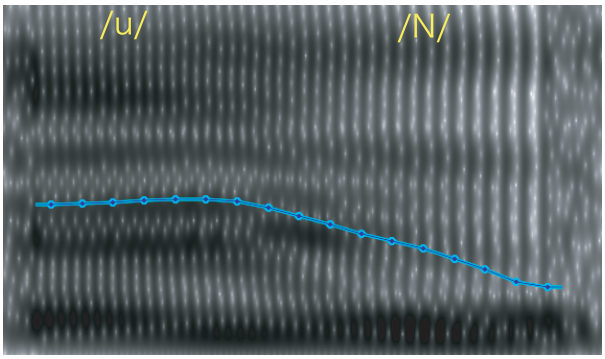

Fig. 2　合成音声「うん」