UU データベースを用いた対話音声合成におけるパラ言語情報制御の効果*

○森　大毅，人見 貴嗣 (宇都宮大)

1　はじめに

これからの音声翻訳機・音声対話システム・知的エージェントなどでは、音声合成により話者の意図・態度・感情状態などのパラ言語情報を表現することが求められる。

コーパスベース音声合成における感情音声合成の典型的なものは、喜び・怒り・悲しみなどの複数の感情 (または発話スタイル) で読み上げられたコーパスを用いている [2]。しかしながら、そのような基本感情語は自発的で表情豊かな音声の記述にはしばしば向いておらず、結果として得られる感情音声はプロトタイプ的になりすぎるおそれがある。

本論文では、表情豊かな会話音声の合成を目的として、HMM 音声合成の枠組に基づく音声合成器を、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 (UU データベース)[1] を用いて構築した。UU データベース中の全発話には、知覚された感情状態ラベルが付与されている。次元説に基づいた記述を採用しているため、典型的な感情だけではなく、自発音声から知覚される微妙なニュアンスも表現されている。この情報を HMM 学習時のコンテキストとして利用することで、合成音声のパラ言語情報を、感情次元により任意に制御できることが期待される。

2　コンテキストラベル

表情豊かな会話 TTS システムを構築するために準備したコンテキスト情報として、トライフォン・アクセント・モーラ数情報など従来のものに加え、我々は発話のパラ言語情報を追加することを検討した。今回は、快-不快および覚醒-睡眠の各次元に対する 3 人のラベラの平均評価値を用いた。他のコンテキスト情報と異なり、パラ言語情報は名義属性でなく数値属性となる。図 1 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。

```
n-a+N/A:1_0/C:1_0_x_x+3_1_x_x_x+x_x_x_x
triphone  mora position  preceding AP  current AP  succeeding AP

/E:4/PLEASANTNESS:300/AROUSAL:500
utt. len.  paralinguistic information
```

Fig. 1　コンテキストラベルの例

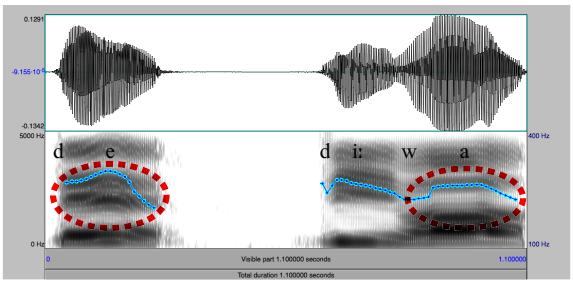

Fig. 2　Synthesized utterance “de diiwa.”

3　モデル学習と合成

UU データベース中の 7 セッションを訓練／テストデータとして用いた。これらは 2 名の女性話者 FTS および FTH による対話であり、訓練用にそれぞれ 589 発話および 654 発話、テスト用にそれぞれ 95 発話および 94 発話を用いた。話者 FTS は UU データベース中で最も感情表現の豊かな話者であるため、最初の試みとしては話者 FTS の声が最適だと考えている。

訓練用音声データには 24 次メルケプストラム分析 (フレーム長 25 ms, フレームシフト 5 ms, Hamming 窓) を施した。f_0 抽出は Praat で行い、creaky 声などが原因の異常な f_0 値を持つフレームは無声とみなした。

HMM は話者別に学習した。JNAS から学習した話者独立モノフォン HMM をモノフォンモデル再推定の初期モデルとして用いた。次に、フルコンテキストに展開した HMM を、そのコンテキストラベルに従い、MDL 基準を利用した決定木アルゴリズムによりクラスタリングした。

テストセット中の合成音声の例「で、ディーは」を図 2 に示す。合成音の f_0 軌跡からは、イントネーション句「で」と「ディーは」の両方の

*Effectiveness of paralinguistic information control in synthesizing dialogue speech using the UU Database.
by MORI, Hiroki, HITOMI, Takatsugu (Utsunomiya University)

末尾において、「説明調上昇下降句末音調」が見られる。これは、若者のくだけた発話スタイルの特徴を反映している。

合成した音声がどの程度表情豊かな対話音声らしく聴こえるかを評価するための実験を実施した。大学生 9 名を被験者とし、FTS セット (95 発話) に対する合成音声が、実際のセッションと同じ順序で呈示された。被験者は、各発話の対話音声としての自然性を 5 段階で評価するよう指示された (5:自然、4:やや自然、3:中立、2:やや不自然、1:不自然)。評価に先立って、被験者には UU データベースの一部の原音声 (セッション C003, 107 発話) を自発的な対話の例として聴かせた。

95 発話に対する MOS は 3.38 で、中立からやや自然の間であった。詳細に分析すると、自然性は発話によって開きがあった。このスコアの開きには発話長との関連が見られ、「うん」「で」「そうだよね」などの短い発話に対する合成音声は一般に自然であると評価されていた。

4 パラ言語情報の主観評価実験

本論文における表情豊かな音声合成器の目標は、言語情報だけでなくパラ言語情報をも伝達することである。このため、指定したパラ言語情報が、合成音声によって聴者にどの程度伝達されるかを調べるための実験を実施した。

被験者は 10 名の大学生である。評価に先立ち、被験者には感情次元の理論およびそれぞれの次元の説明を行った。その後、前節と同様に、自然な一連の会話を例として聴かせた。

被験者には、FTS セットおよび FTH セットの合成対話音声を、実際のセッションと同じ順序で呈示した。被験者は、各発話に対して知覚されたパラ言語情報を、UU データベースにおける自然発話に対する評価と同じ方法で 7 段階評価するよう指示された。パラ言語情報評価の項目は、快-不快と覚醒-睡眠の 2 次元である。例えば快-不快においては、1 が非常に不快、4 が中立、7 が非常に快に対応する。被験者は、話している内容そのものでなく、話し方から受ける印象を評価するよう教示された。

合成時にコンテキストとして指定したパラ言語情報と、主観評価結果との間の相関係数を評価指標とした。表 1 に結果を示す。“平均” は 10 人の平均評価値に対して計算された相関係数である。最も感受性が高い被験者は#3, #4, #10 であり、快-不快に対しては 0.6 以上の、覚醒-睡眠に対しては 0.7 以上の相関係数が得られた。これは、合成した発話によって、指定されたパラ言語情報が正しく聴者に伝わったことを意味する。覚醒-睡眠次元の方がより制御が容易であり、最も感受性の低い被験者に対してさえも 0.592 の相関係数が得られた。平均評価値は一般に個々の被験者の評価値に比べて音声パラメータとの相関が高いことがわかっているが、ここでの相関係数も 0.8 程度と高かった。

Table 1　指定したパラ言語情報と主観評価結果との間の相関係数

被験者	#1	#2	#3	#4	#5
快-不快	0.527	0.641	0.631	0.651	0.447
覚醒-睡眠	0.675	0.698	0.783	0.748	0.691

#6	#7	#8	#9	#10	平均
0.470	0.342	0.371	0.512	0.601	0.772
0.664	0.689	0.602	0.592	0.765	0.837

5 おわりに

HMM 音声合成に基づいてパラ言語情報の制御能力を持つ表情豊かな会話音声を合成する試みについて述べた。自然性テストでは 3.38 の MOS が得られた。また、指定した快-不快および覚醒-睡眠の値と主観評価の平均値との間に非常に高い相関 ($R \simeq 0.8$) が見られ、合成された発話が指定されたパラ言語情報を正しく伝達できたことがわかった。

本論文で用いられているコンテキスト情報は、自発的な対話音声をモデル化するには十分とは言えない。例えば、異なった句末音調を区別することはできない。韻律的句、談話構造、声質などの要素も自然性を改善するために必要であると考えられる。また、感情次元では表現できないパラ言語情報の記述法も今後の課題である。

参考文献

[1] Mori et al., Speech Communication **53**, 36–50, 2011.

[2] Tachibana et al., IEICE Trans. Inf. & Syst., **E88-D**, 2484–2491, 2005.