UU データベースを用いた対話音声合成におけるコンテキスト情報の効果*

○森　大毅，△佐藤 航，△倉持 直貴 (宇都宮大)

1　はじめに

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] を用いて実現することを検討している。これまでに、提案する枠組において、快-不快、覚醒-睡眠の次元で表現されたパラ言語情報をコンテキスト情報として与えることで合成音声にパラ言語情報を反映させることができ、合成音声からパラ言語情報を知覚させる実験の結果、主観評価値との間に相関係数 0.8 程度の高い相関が得られることがわかっている [2]。

ところで、パラ言語情報だけでなく、言語情報も話者の感情状態などの情報を伝達する。したがって、上記検討は、言語情報とパラ言語情報の同時伝達による感情状態の知覚を評価したものだと考えられる。対話音声合成を広く応用可能とするためには、言語情報とは独立にパラ言語情報を制御可能であることがより望ましい。そこで本研究では、言語情報を固定し、パラ言語情報のコンテキストだけを変化させて合成した音声を刺激とした主観評価実験について述べる。

2　パラ言語情報の主観評価実験

HMM 音声合成に用いるコンテキスト情報として、UU データベースに収録されている快-不快 (1:非常に不快, 7:非常に快) および覚醒-睡眠 (1:非常に睡眠, 7:非常に覚醒) の各次元に対する 3 人のラベラの平均評価値を与えた。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。

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、訓練用にセッション C002–C007 から話者 FTS の 589 発話、テスト用にセッション C001 から話者 FTS の 5 発話を用いた。テスト用の発話は、C001 の話者 FTS の 95 発話の発話内容から合成した音声を試聴し、その中から比較的明瞭性が高いものを選んだ。テストに用いた発話を Table 1 に示す。

Table 1　呈示音声セットの発話内容

発話 ID	発話内容
006	(F うんとね)
015	うん
077	うん聞こえんのかなって、うんそんな感じ
177	あ、そうだね
186	うん、ビーシーエーディーだね

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。

被験者は 8 名の大学生・大学院生である。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、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。

被験者は実験用 PC の前に座り、自分自身の操作によりヘッドホン (SONY MDR-Z600) から呈示される音声を聴取する。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。各刺激は 3 回まで聴取することを許した。また、以前の刺激を聞き直すことは禁止した。

呈示刺激は、上述した発話内容のそれぞれに対し、快-不快が 3, 4, 5, 6 の 4 通り、覚醒-睡眠が 3, 4, 5, 6 の 4 通りの計 16 通りのパラ言語情報を与えて合成したものである。呈示順序は 006, 015, 077, 177, 186, ... を 16 回繰り返したもの

*Effects of contextual factors in synthesizing dialogue speech using the UU Database.
by MORI, Hiroki, SATO, Wataru, KURAMOCHI, Naoki (Utsunomiya University)

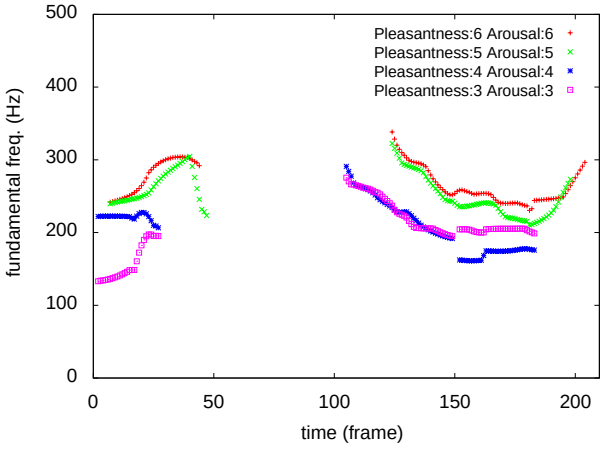

Fig. 1　合成された音声パラメータ (f_0) に対するパラ言語情報コンテキストの影響

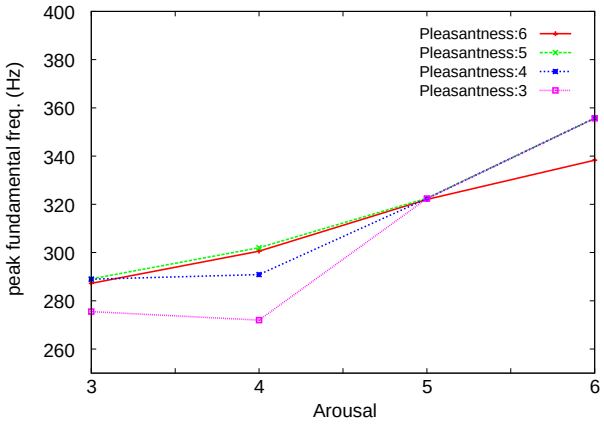

Fig. 2　パラ言語情報コンテキストと合成された f_0 最大値との関係

であり、同一の発話内容に対し与えるパラ言語情報コンテキスト (16 通り) の順序はランダム化した。

呈示した合成音声 f_0 軌跡の例 (発話 ID 177) を Fig. 1 に示す。この発話では、コンテキスト情報として与えた快-不快および覚醒-睡眠の値が大きいほど f_0 軌跡が高い声側にシフトしていることがわかる。

Fig. 2 に、快-不快 4 通り×覚醒-睡眠 4 通りの全てのパラ言語情報コンテキストの組み合わせに対し、発話 ID 177 から合成した音声パラメータの f_0 最大値を示す。f_0 最大値の変化は覚醒-睡眠に対して比較的敏感であり、覚醒寄りの音声の f_0 最大値は大きくなる傾向がある。

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表 2 に示す。“平均” は 8 人の平均評価値に対して計算された相関係数である。全体として、覚醒-睡眠の方が

Table 2　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 (*: $p < 0.05$, **: $p < 0.01$)

被験者	#1	#2	#3	#4
快-不快	0.20	0.18	0.06	0.10
覚醒-睡眠	0.26*	0.56**	0.61**	0.36**

#5	#6	#7	#8	平均
0.08	0.06	0.13	0.19	0.31**
0.54**	0.47**	0.38**	0.48**	0.66**

快-不快に比べ高い相関係数を示した。また、全ての被験者に対して快-不快のコンテキストと主観評価値との間の相関係数は有意ではなかった。この原因の 1 つは、Fig. 2 にも示したように、合成される音声パラメータが快-不快コンテキストに対して鈍感な傾向があるためである。

各発話の合成時に与えるパラ言語情報コンテキストを、UU データベースにより与えられている平均評価値 (1 通り) とした過去の検討 [2] と比較すると、パラ言語情報コンテキストを 16 通りに振った今回の主観評価実験で得られた相関係数は、快-不快、覚醒-睡眠ともに全体として小さくなった。このことより、対話合成音声から知覚される話者の感情状態は、パラ言語情報だけでなく言語情報によっても伝達されていることが裏付けられた。

3　おわりに

コンテキスト情報の操作により、対話音声合成においてパラ言語情報を聴き手に伝達できるかを調べ、覚醒-睡眠の次元については、指定したパラ言語情報が全ての被験者において意図した方向に知覚されることがわかった。

現在、パラ言語情報の制御手法として、コンテキスト情報の他に重回帰 HMM を利用する方法を検討している [3]。この方法でも本研究と同様に言語情報を固定した実験を行い、快-不快および覚醒-睡眠の各次元に対しそれぞれ 0.49, 0.65 の相関係数が得られている。今後、さらなる分析により両手法の得失を明らかにしたい。

参考文献

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

[2] 森, 人見, 音講論 (秋), 331–332, 2011.

[3] 永田, 森, 能勢, 音講論 (春), 2012 (発表予定).