相槌音声合成におけるパラ言語情報の感情次元による制御*

○森　大毅 (宇都宮大)

1　はじめに

相槌の機能のひとつは、対話を円滑にすることである。発話交替に関わる調整行動の中では、相槌のフィードバックは聴き手が発話権を取る意志がないことを話し手が知る手がかりとなっている。テレフォンショッピングタスク [1] のような、様式化された固い対話における相槌は、主としてこの種のメタメッセージとして機能しているものと考えられる。

一方、家族や親しい友人との会話における相槌はもっと多様であり、上記の働きに加え、聴き手の興味・驚き・同情・称賛などの反映であるパラ言語情報を伝達している [2, 4]。人間と機械の円滑な対話のため、音声対話システムにおいて相槌発話を生成する試みがある [1–3, 5, 6] が、これらはいずれも相槌生成タイミングを重視したもので、相槌によってパラ言語情報を積極的に表現しようというものではなかった。

相槌音声の合成において、パラ言語情報に関わる制御要因が取り入れられて来なかった 1 つの原因としては、制御要因となるべき一般的属性を定義することが困難であったことが考えられる。一方、宇都宮大学では自然で表情豊かな対話音声に含まれるパラ言語情報に関する研究への利用に適した音声対話データベース (以下、UUDB)[7] を開発し、広く公開している。UUDB においては、相槌を含む全ての発話に、音声から知覚される話者の感情状態を記述したパラ言語情報ラベルが付与されている。本研究は、UUDB を基にしたコーパスベース音声合成の枠組を利用し、機械から人間へのフィードバックとして発する相槌音声の音響的特徴を、任意に指定したパラ言語情報が伝達されるよう制御することを目的とするものである。

2　パラ言語情報の制御要因

現在公開されている版の UUDB には、各発話から知覚される話者の感情状態を 3 名のラベラが評価した結果が、「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」の 6 抽象次元により記録されている。今回は、話者自身の感情状態の最も一般的な指標であるとされている「快-不快」「覚醒-睡眠」の 2 次元に注目した。各ラベラの評定値は、4 を中立、1 を「非常に不快」または「非常に睡眠」、7 を「非常に快」または「非常に覚醒」とする 7 段階である。モデルの学習に用いるパラ言語情報の属性値としては、ラベラ 3 名の平均評定値を用いた。よって、各次元の属性が取り得る値は $1.00, 1.33, 1.67, \ldots, 6.67, 7.00$ であり、2 次元だと $19^2 = 361$ 通りの可能な組み合わせがある。今回は、相槌音声合成における制御要因もこれと同様の離散値で与えるものとする。

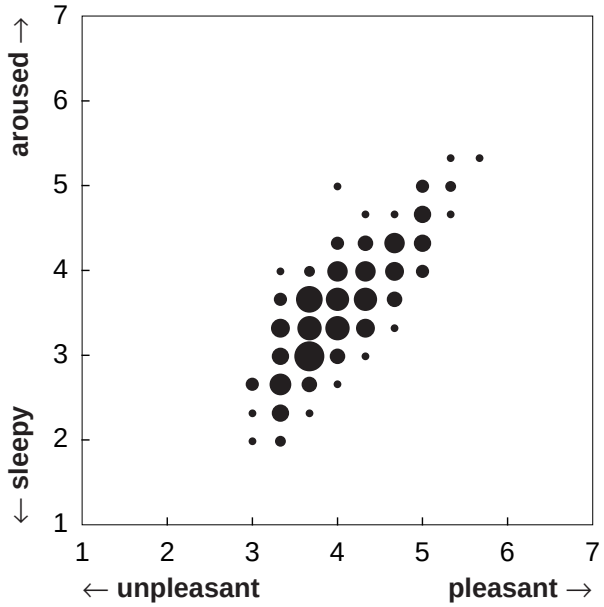

Fig. 1　発話「うん」の感情状態の平均評定値分布 (話者 FTS)

Fig. 1 に、UUDB に収録されている女性話者 FTS の相槌および相槌様発話「うん」のうち、孤立した 182 発話の平均評定値の分布を示す。図中、円の面積は発話数に比例している。全発話の分布 [7] と比較すると分布の範囲は狭くなっているが、それでも評定値にして 2 以上の広がりを持つ分布となっており、同じ相槌でも伝達されるパラ言語情報に多様性があることがわかる。

*Paralinguistic manipulation for backchannel synthesis based on abstract dimensions of emotion.
by MORI, Hiroki (Utsunomiya University)

3 HMM 音声合成

HMM 音声合成 [8] の枠組を用いて、相槌発話のパラ言語情報を制御することを試みた。音声データには、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 685 発話と比較的多く、また FTS は相槌や感情表出系感動詞がパラ言語的に多様なためである。

発話内容は、ある程度のデータ量が確保できる「うん」に限定した。孤立した「うん」182 発話に対し、19 次のメルケプストラム分析と f_0 抽出を行った。相槌や相槌様の「うん」には比較的不明瞭な発話が多く、また発話先頭および末尾周辺で非 modal の声質を呈することがある。そこで、視察ならびに分析再合成音の聴取により f_0 軌跡の連続性をチェックし、f_0 推定値の修正または無声区間への変更を行った。

HMM の構造は「うん」全体で 5 状態の単一ガウス分布とし、学習は **2** で述べたパラ言語情報の属性値をコンテキスト情報として行った。ガウス分布のクラスタリングは $\log f_0$ とメルケプストラムのそれぞれに対してボトムアップ的に行った。パラメータ調整の結果、得られたモデルの分布数は $\log f_0$ が 207, メルケプストラムが 17, 物理 HMM 数は 42, 論理 HMM 数は 44 となった。なお継続時間モデルのクラスタリングは行わなかった。

4 評価実験

ここでは、**3** で得られた HMM から合成した相槌音声が、任意に指定したパラ言語情報を伝達するかを検証するための聴覚実験について述べる。被験者は音声の研究室に所属する大学院生 3 名および卒研生 4 名 (全員男性) で、感情状態の抽象次元については十分に理解している。

評価対象の発話集合は、HMM の学習に使用したものと同一のセットから、パラ言語情報の偏りが最小となるように選んだ 97 発話である。呈示する刺激は、原音声の分析再合成音 97 発話+パラ言語情報ラベルから HMM 合成した 97 発話の計 194 発話である。被験者には、各発話から知覚されるパラ言語情報を、「快-不快」「覚醒-睡眠」それぞれ 7 段階で評価させた。

UUDB で与えられたパラ言語情報ラベルと、7 名の被験者の平均評価値との相関係数を表 1 に示す。表より、HMM 音声合成による相槌は、原音声に比べれば劣るものの、与えた抽象次元の値を反映したパラ言語情報を伝達できていることが、聴覚的に確かめられた。

Table 1　実験結果

	原音声	HMM 合成音声
快-不快	0.67	0.41
覚醒-睡眠	0.88	0.58

原音声と HMM 合成音声の聴覚印象を直接比較すると、HMM 合成音声の方がパラ言語情報ラベルの違いによる差が小さい傾向があった。また、快-不快の次元に関しては、全体に原音声 (平均 3.98) に比べて HMM 合成音声 (平均 3.51) の方が不快寄りに知覚される傾向があった。相関係数がやや低下したのは、これらの影響を反映した結果だと考えられる。

5 おわりに

HMM 音声合成の枠組により、パラ言語情報をコンテキスト情報として様々な相槌「うん」の音声を合成し、ユーザに与える印象をある程度制御することが可能であることを示した。今後は、これ以外の発話を含めさらに検討を進めたい。

今回はコーパスに出現しないパラ言語情報の値の組み合わせについては考慮しなかったが、将来任意の文を発話できる音声フィードバック合成器を実現するためには、抽象次元上で何らかの一般化を行い、未出現のコンテキストに対しても HMM を用意する必要がある。また、対話システム [9] における評価も今後の課題である。

参考文献

[1] 岡登 他, 情処論, 40(2), 469–478, 1999.
[2] Ward, N., 情処研報, SLP 96(55), 7–12, 1996.
[3] Hirasawa et al., Proc. Eurospeech '99, 1391–1394, 1999.
[4] 梅野 他, 音講論 (春), 313–314, 2003.
[5] 藤江 他, 人工知能学会研資, SIG-SLUD-A401, 15–20, 2004.
[6] Kitaoka et al., J. JSAI, 20(3), 220–228, 2005.
[7] 森 他, 音講論 (秋), 311–312, 2007.
[8] http://hts.sp.nitech.ac.jp/
[9] 長塚, 森, 音講論 (春), 485–486, 2009.