自然対話コーパスを用いた音声コミュニケーション場面における笑い声合成の検討 *

☆永田智洋, 森大毅 (宇都宮大)

1 はじめに

近年では人間同士だけではなく、人間対機械によるインタラクションにも関心が高まっており、人間同士の対話で用いられるような音声を合成する対話音声合成の需要が高まっている。現在、対話音声合成は話し言葉や話者の感情の制御が検討されているが、言語音を対象とした検討がほとんどであり、更なる表現力向上のため、笑い声を含むノンバーバル情報を伝達可能な非語彙的な音声現象の合成を期待している。

笑いは相手に自身の好意的な態度を伝えるために意識的に発するものや、話者の感情の高まりによって突発的かつ無意識的に発せられるものなど非常に多様であるため、録音した笑い声を再生するだけでは不十分である。[1] による研究では、Diphone 音声合成により笑い声を合成する手法が提案されているが、笑い声を含む発話は全体としての自然性が低くなることが報告されている。このことから、笑い声を合成する場合には状況や文脈を考慮する必要があると考えられる。

一方、HMM 音声合成方式では様々な要因によって変動する音素の韻律的／文節的特徴をコンテキストとして表現する。そこで、本研究では、次節で述べる Call と呼ばれる笑い声を構成する断片に対してコンテキストを定義し、HMM 音声合成と同様の方法でコンテキストクラスタリングを行うことにより笑い声のモデル化を行う。

2 笑い声に対するコンテキスト

笑い声の構造は階層的であり、1 回の呼気に対応する Bout(句に相当) は 1 個以上の Call(音節に相当) によって構成される [2](Fig. 1)。

本研究で定義した笑い声を構成する Call のコンテキストを Table. 1 に示す。声道形状および有声性により特徴付けられる音の違い [3] を表現するため、各 Call の音声学的転記 ([ha] など) をコンテキストに加えている。また、HMM 音声合成の場合と同様、前後の音の違いによる影響に対処するために先行、当該、後続の音素／ Call の音声学的転記を考慮した。

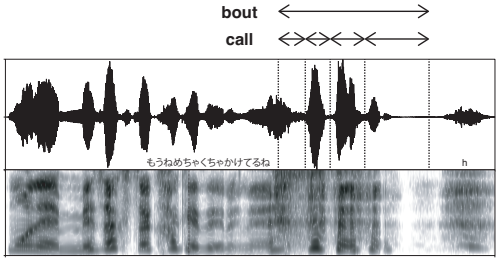

Fig. 1 笑い声の構造

Table 1 笑い声を構成する Call のコンテキスト

先行音素／ Call の音声学的転記
当該音素／ Call の音声学的転記
後続音素／ Call の音声学的転記
当該 Call が口音／鼻音
当該 Bout における Call 位置
当該 Bout の Call 数
当該発話における Bout 位置

Call は Bout 中における位置によって継続長に差がある (e.g. 第 1 Call は長い [2])。この傾向をモデル化するため Call 位置に関するコンテキストを追加している。

3 HMM 音声合成に基づく笑い声合成

モデル構築用および検証用の笑い声データとして、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)[4] に収録されている女性話者 FTS の笑い声 40 Bouts を用いた。初期モデルの学習のために必要な Call の開始／終了時刻は人手で与えた。各条件を Table. 2 にまとめる。

各 Bout の合成は、leave-one-out で学習された

Table 2 モデルの学習条件

モデル	5 状態 left-to-right HSMM
特徴ベクトル	0-34 次のメルケプストラム係数, 対数基本周波数, Δ, ΔΔ
分析条件	サンプリング周波数 16 kHZ の音声に対し窓長 25 ms, 分析周期 5 ms のハミング窓

* Consideration of laughter synthesis in speech communication with spontaneous speech corpus . by NAGATA, Tomohiro, MORI, Hiroki (Utsunomiya University)

モデルに合成対象 Bout のコンテキストラベル列を与えて行った。笑い声の合成結果の例として、「フフフ」を合成した場合の波形とスペクトログラムを Fig. 2 に示す。図より、Call の位置によって継続長および波形の振幅が変化していることが確認でき、コンテキストが反映されていることがわかる。また、第 1Call の継続長が以降の Call と比較して長くなっており、これは [2] による笑い声の傾向と一致していることがわかる。

4　主観評価実験

前節で合成した笑い声を用いて自然性評価実験を行った。本研究では文脈や状況に適した笑い声の合成を目的としているため、笑い声のみを合成するだけではなく、言語音を含む発話における笑い声の自然性を評価する。言語音の部分は自然音声の分析合成音で置き換えた。言語音を含む発話の刺激は 30 個、笑い声のみの刺激は 7 個であり、刺激の総数は 37 個である。

被験者は男性の大学生 3 名と大学院生 5 名である。評価は笑い声部分の自然性を 5 段階 (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) で評価するよう指示した。

被験者による評価値の平均 (MOS) に対し母音性 (a, u, e, o)、有声／無声を要因とする 2 要因分散分析を行ったところ、有声／無声の主効果のみが有意 ($p < .05$) であった。有声／無声別の MOS の分布を Fig. 3 に示す。ここで、Voiced は有声 Call のみで構成される Bout であり、Unvoiced/mixed は無声 Call のみで構成される Bout と無声 Call および有声 Call が混在している Bout である。図より、無声 Call を含む Bout の方が自然性が高い傾向があり、MOS の分布が広いことがわかる。

MOS の低い有声のみの Bout には各 Call の音響的特徴に変化が現れない傾向がみられた。こ

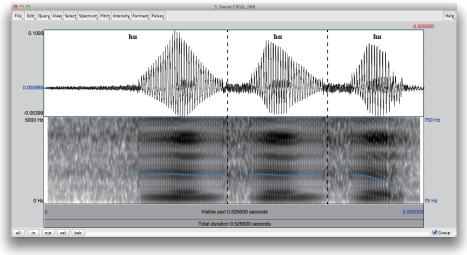

Fig. 2　合成された笑い声の波形とスペクトログラム

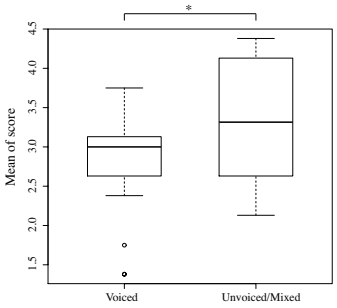

Fig. 3　合成された笑い声の自然性に関する MOS の分布

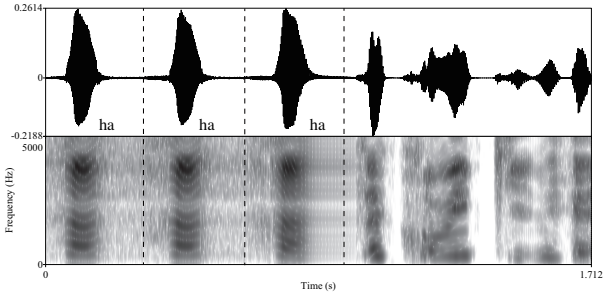

Fig. 4　各 Call の音響的特徴が変化しない例

の例を Fig. 4 に示す。このことから Call 間の音響的特徴の変化は自然性に強く影響すると考えられ、Call 間の関係を陽にモデル化する必要があると考えられる。

5　おわりに

本研究では、UUDB の笑い声に対して Call のラベリングを行い、ラベリング結果と HMM 音声合成の枠組みを利用して笑い声の合成を行った。合成された笑い声に対して自然性に関する主観評価実験を行い、比較的自然な笑い声が合成可能であることを示した。また、今後は Call 位置による音響的特徴の違いを陽にモデル化などを検討する必要がある。

謝辞 本研究は JSPS 科研費 26280100, 26284062 の助成を受けている。

参考文献

[1] Trouvain et al., Proc. Workshop on Affective Dialogue Systems, 229–232, 2004.
[2] Bachorowski et al., J. Acoust. Soc. Am. **110**, 1581–1595, 2001.
[3] 森, 音講論 (秋), 2-1-10, 2015.
[4] Mori *et al.*, Speech Communication, 53, 36-50, 2011.