一般社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report
SP2017-65 (2017-12)

自然対話における発話の文脈を考慮した笑い声合成の検討

永田　智洋† 　森　 大毅†

† 宇都宮大学大学院工学研究科　〒321–8585 栃木県宇都宮市陽東 7–1–2
E-mail: †{ken1,hiroki}@speech-lab.org

あらまし　従来の笑い声合成研究のほとんどは対話場面における笑い声を対象としておらず，対話場面に頻出する言語音を伴う笑い声を合成した際に発話全体として自然性が低下することが問題となっていた．そこで本研究では，自然対話における笑い声を対象とした．言語音を伴う笑い声を適切に合成するために，発話における笑い声位置といった笑い声に対するコンテキストが定義された．定義されたコンテキストと，隠れマルコフモデルに基づく音声合成方式によって文脈を考慮した笑い声が合成された．定義されたコンテキストの有効性を確かめるために，合成された笑い声の自然性評価が行われた．笑い声は言語音と共に呈示され，笑い声単独ではなく発話全体としての自然性が評価された．自然性評価では文脈を全く考慮しない，音韻環境を考慮，発話の文脈を考慮した笑い声の自然性が比較された．自然性評価の結果から，発話の文脈を考慮することによって発話全体としての自然性が改善したことが示された．
キーワード　笑い声, 自然対話音声コーパス, 自発音声, HMM 音声合成, ノンバーバル

Laughter synthesis considering the context of the utterance in natural conversation

Tomohiro NAGATA† and Hiroki MORI†

† Graduate school of engineering, Utsunomiya university　Yoto 7–1–2, Utsunomiya, Tochigi, 321–8585 Japan
E-mail: †{ken1,hiroki}@speech-lab.org

Abstract　Most of conventional studies on laughter synthesis do not deal with laughter in actual dialogue scene, and it became a problem that the overall naturalness of the utterance is reduce when synthesizing laughter accompaneied speech sounds. Therefore, this study focuses on laughter in spontaneous dialogue. In order to synthesize laughter accompanied by speech sounds, the context for laughter (i.e. the laughter position in the utterance) were defined. Laughter considering the context was synthesized by the defined context and HMM-based speech synthesis framework. To confirm the effectiveness of the defined context, we checked a naturalness of the synthesized laughter. The synthesized laughs were presented with speech sounds, and the subjects evaluated the overall naturalness of the utterance. The result of the naturalness evaluation showed that the overall naturalness of the utterance was improved by considering the context.
Key words　Laughter, Spontaneous speech corpus, spontaneous speech, HMM-based speech synthesis, non-verbal

1. はじめに

近年では，人間同士だけではなく，人間と機械のインタラクションにも関心が高まっている．ドコモの「しゃべってコンシェル」や Apple の「Siri」などに加えて，更に最近ではソフトバンクの「Pepper」といった，より人間同士のインタラクションに近い利用形態が広まりつつある．そのような利用形態においては，音声合成技術には言語情報を明瞭に表現することだけではなく，話者の意図や態度，感情といったパラ言語情報を表現することや話し言葉の合成，更には咳払い，笑い，ため息などの非言語音をも合成することが求められる．ここでは，そのような音声合成のことを対話音声合成と呼ぶ．

笑い声は代表的な非言語音の 1 つである．笑いはコミュニケーションや人間関係の潤滑油とも呼ばれる重要なノンバーバル行動であり，様々な形態および機能がある [1–3]．笑い声の機能と形態の関係を解き明かし，機能に沿った笑い声の形態の合成を実現することによって，人間と機械のインタラクションがよりヒューマンライクなものになると期待できる．そこで，本研究では種々の形態の笑い声を合成することに注目する．

笑い声の合成は非常に挑戦的な課題であり，現在でもあまり研究が進んでいない．数少ない笑い声合成に関連する研究には，分析合成に基づく手法 [4,5]，波形接続型方式に基づく手法 [6,7]，統計モデルに基づく手法 [8,9] がある．文献 [4] では，言語音母音から抽出された音源パラメータを調整することによって笑い声合成を実現している．文献 [5] では，笑い声波形の振幅の振動的な動きをバネ-マス系モデルでモデル化することにより，周期的な有声笑い声の合成を実現している．文献 [6,7] ではダイフォンに基づく笑い声合成を行っている．実際の笑い声素片を利用しているため，高品質な笑い声を合成可能である．文献 [8,9] では，隠れマルコフモデル (HMM) 音声合成方式による手法に基づいて笑い声を合成している．この手法では，笑い声を音素のような単位で記述し，それを音素と見立てて言語音における HMM 音声合成の枠組みに当てはめることにより笑い声を合成する．

このように，笑い声合成に関する研究は数少ないものの一定の成果を挙げており，品質の高い笑い声の合成が実現されている．しかしながら，これまでの笑い声合成に関する研究は，お笑いやジョークビデオによって誘発された笑い声を用いているものがほとんどであり，実際の対話場面における笑い声を対象としていない．実際の対話場面では笑ってから話す，あるいは話してから笑うというように，言語音を伴う場合が多い．一方で，映像刺激などによって誘発された笑い声が言語音を伴うことは稀であり，そのような笑い声を用いて言語音を伴う笑い声を合成すると，笑い声自体は高品質であっても発話全体としての自然性が低下することが指摘されている [6]．

そこで，本研究では映像刺激によって誘発された笑い声ではなく，実際の対話場面に現れる笑い声を対象とした笑い声の合成を検討する．自然対話音声コーパスに含まれる笑い声を用い，笑い声の発話における文脈を考慮することによって，合成される発話の発話全体としての自然性を改善することを目的とする．

2. 自然対話音声コーパス

本研究では，実際の対話場面に現れる笑い声が要求される．そこで，自然対話音声コーパス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) [10] および感情評定値付きオンラインゲーム音声チャットコーパス (OGVC) [11] の 2 つを対象とする．

2.1 UUDB

UUDB は自然で表情豊かな音声対話に見られる多様な音声学現象および言語学的現象の研究への用途を開発目的とした音声コーパスであり，親近性の高い大学生 7 ペア (女性 12 名，男性 2 名) による自然な対話音声が収録されている．

UUDB には収録されている音声に対してパラ言語情報をはじめとする様々なノンバーバル情報が与えられている．UUDB では非言語音に関する記述も与えられており，笑い声や吸気，咳払いといった非言語音の位置と時間情報が与えられている．UUDB に含まれる笑い声の総数は 280 個であり，各話者による内訳は 1 の通りである．

表 1　UUDB の各話者における笑い声の数

Speaker	Num	Speaker	Num	Speaker	Num
FJK	8	FKC	22	FMS	16
FMT	34	FNN	14	FSA	26
FSH	17	FTH	5	FTS	40
FTY	19	FUE	14	FYH	23
MKK	22	MKO	20		

表 2　OGVC の各話者における笑い声の数

Speaker	Num	Speaker	Num	Speaker	Num
01_MMK	26	03_FMA	74	05_MYH	52
01_MAD	118	03_FTY	41	05_MKK	87
02_MFM	142	04_MNN	154	06_FTY	251
02_MEM	145	04_MSY	246	06_FWA	175

2.2 OGVC

OGVC は友人同士がボイスチャットをしながらゲームをプレイしている時の対話を収録した自然対話コーパスである．また，自然対話を再朗読した音声を収録されており，自然対話音声と演技音声の比較を行うこともできる．本研究では，OGVC の自然対話音声のみを使用する．OGVC には 13 名 (女性 4 名，男性 9 名) の話者による 9114 発話の自然対話音声が収録されている．

OGVC では笑い声の位置が転記として与えられている．OGVC の自然対話音声に含まれる笑い声の総数は 1593 個であり，各話者における内訳を表 2 に示す．表から，OGVC には 1 人あたりの笑い声の数が比較的多いことがわかる．

3. 笑い声のセグメンテーション

笑い声の構造は階層的に理解されており，その構造は大きく分節レベル，音節レベル，フレーズレベルに分けられる [12]．図 1 はその構造を示す．分節レベルでは，笑い声は「笑い声の子音」や「笑い声の母音」によって分割される．笑い声子音および笑い声母音は音声学における子音および母音とは厳密には異なる．音節レベルでは笑い声子音と笑い声母音の組，あるいは笑い声母音単独で構成される「call」と呼ばれる単位で笑い声が分割される．フレーズレベルでは，笑い声は 1 つ以上の call で構成される「bout」と呼ばれる単位で笑い声が分割される．一般に，bout には吸気は含まれない．また，複数のフレーズの連続は文レベルとも呼ばれ，「笑い声 episode」と呼ばれる．

本研究では，自然対話音声コーパスの笑い声は bout 単位および call 単位でセグメンテーションされる．また，call 単位のセグメントは，その後音韻についての転記情報が与えられる．

3.1 セグメンテーションおよびアノテーション

セグメンテーションは，まず発話の中から笑っている部分を笑い声 episode として分割する．笑い声 episode は 1 つ以上の bout と吸気によって構成される．次に，笑い声 episode は bout と吸気に分割され，最後に bout 部分は call 単位に分割される．

call レベルで分割されたセグメントは音韻についての転記が

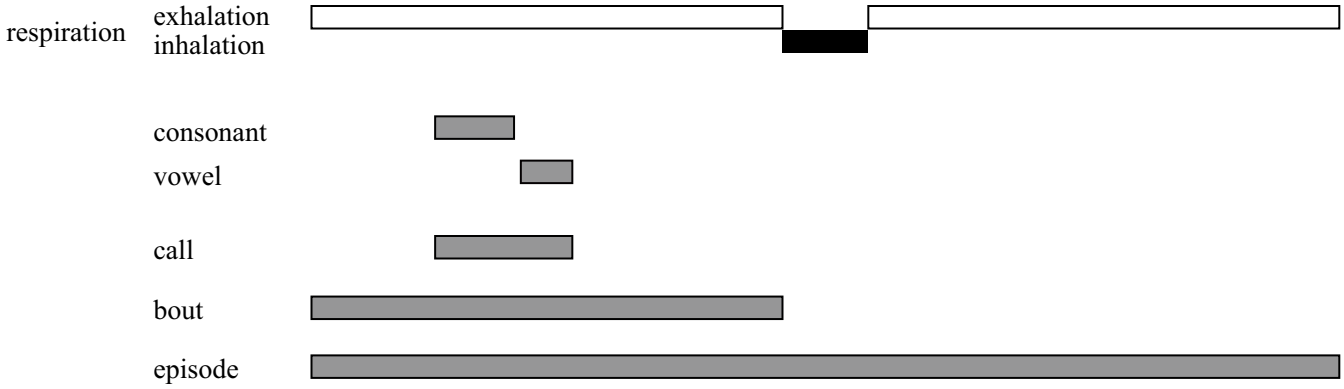


図 1　笑い声の階層構造 ([12] 一部改変).

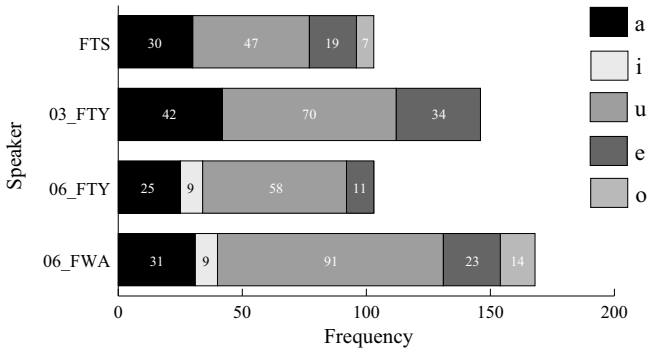


図 3　笑い声母音の数

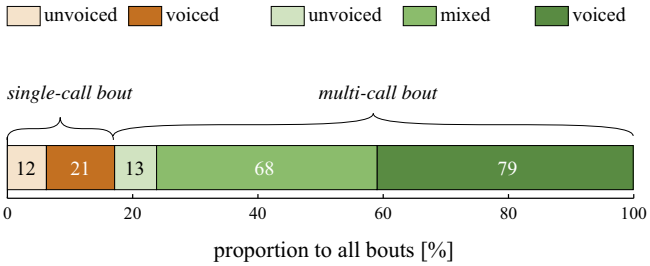


図 4　bout の有声/無声構造の内訳

与えられる．ここで，転記は音節を音素で表記したものを用いる．また，この際，音響的特徴が大きく変化する無声化，鼻音化，長音化といった現象を伴うセグメントは異なる音韻として記述した．

笑い声のセグメンテーションおよびアノテーションは UUDB の女性話者 FTS および OGVC の女性話者 03_FTY, 06_FTY および 06_FWA の笑い声を含む発話を対象に行われた．セグメンテーションおよびアノテーションは 1 名によって行われた．

3.2　セグメンテーションおよびアノテーション結果

実際にアノテーションされた笑い声の例を図 2 に示す．①では笑い声 episode と言語音部分の転記が記述される．{laugh} で示される部分が笑い声 episode であり，それ以外の部分が言語音である．②では笑い声 episode の構造が bout と吸気によって記述される．b が bout を表しており，h が吸気を表している．③では，bout を構成する call の音韻が記述されている．この例では，[huhuhaha] という 4 つの call で構成されている bout であることがわかる．

アノテーションの結果として，call の笑い声母音の出現頻度分布を図 3 に示す．対話場面における call の笑い声母音としては，/a/系，/u/系，/e/系が支配的である．/i/系や/o/系の笑い声母音は稀であり，話者によっては全く出現していない．

bout 構造の内訳を図 4 に示す．UUDB の bout 構造の内訳については文献 [13] にて既に調査されているが，本研究では OGVC の笑い声も含む．図より，1 つの call のみから成る bout が全体の 17%であり，そのうちの 35%程度が無声であることがわかる．一方で，2 つ以上の call から成る bout では，全体が無声で構成されることはあまりないことがわかる．この結果は UUDB の笑い声のみを対象とした結果 [13] と概ね一致する．

4.　笑い声合成

4.1　笑い声コンテキストの定義

本研究で定義されたコンテキストを表 3 に示す．声道形状と音韻によって特徴付けられる音の違いを表現するために，当該 call の音韻転記が定義された．

更に，2 つのグループがコンテキストに追加された．グループ A は狭い意味での文脈を考慮する要素の集合であり，先行および後続の「セグメント」の音韻転記である．ここで，「セグメント」とは前の音が笑い声である場合には call に相当し，言語音である場合には音素に相当する．

グループ B では，より広い意味での文脈を考慮するための要素の集合である．このグループには，笑い声の発話における位置が定義されている．この位置は bout が発話の先頭，末尾，それ以外のどこに位置するかを示す．また，call の継続長は call 位置によって異なるということが報告されている [1]．このことを考慮するために，bout における call 位置が定義された．更に，HMM 音声合成方式による言語音の合成では，発話のモーラ数が変動要因として広く用いられている．これを参考に，笑い声の 1 bout における call 数が定義された．

4.2　合 成 条 件

本研究では，UUDB の話者 FTS および OGVC の話者 06_FTY，06_FWA の計 3 名の笑い声を対象とする．モデルの学習に使用する笑い声の総数は 109 bouts である．笑い声

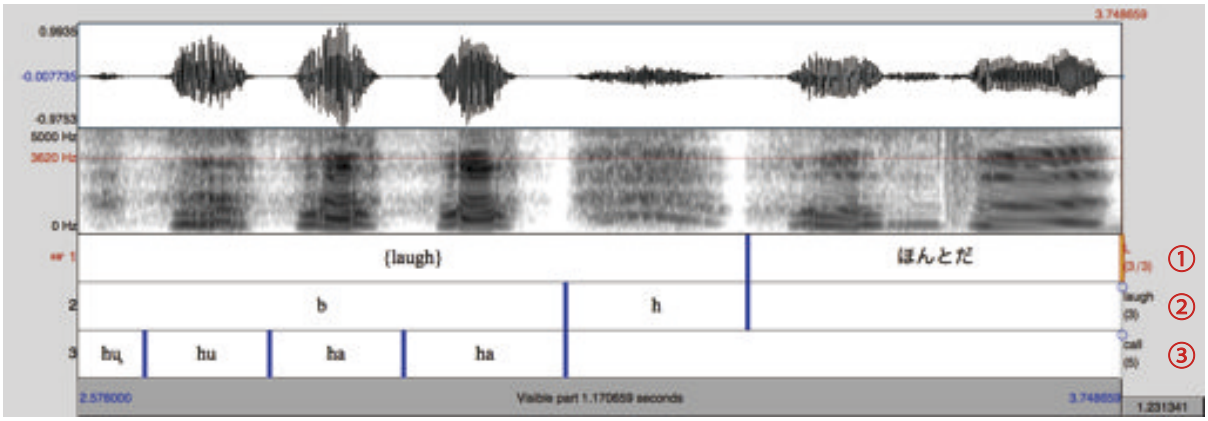


図 2　笑い声に対するアノテーション例

表 3　本研究で定義された笑い声コンテキスト

c_c：当該 call の音韻転記	
c_l：先行セグメントの音韻転記	A
c_r：後続セグメントの音韻転記	A
p_l：発話における bout 位置	B
p_c：bout における call 位置	B
n_c：bout を構成する call 数	B

表 4　笑い声モデルの学習条件

モデル	5 状態の left-to-right HSMM
特徴量ベクトル	0 から 39 次のメルケプストラム係数, 対数基本周波数, それぞれの Δ および ΔΔ を含めた 123 次元ベクトル
分析	サンプリング周波数 16 kHz の笑い声に対して 窓長 25 ms，フレームシフト 5 ms とした ハミング窓による分析

の音響特徴量の分析条件およびモデル学習に使用する特徴量ベクトルとモデルの構成を表 4 に示す．また，モデル学習には共有決定木を用いた話者適応学習技術 [14] が使用された．

テスト用の笑い声は，学習に使用する笑い声の中から選択し，その笑い声を除いて学習されたモデルから合成された (Leave-one-out 法)．各合成笑い声は元々の話者のモデルに話者適応された．

4.3　合 成 結 果

合成された笑い声の例として，笑い声 [hu̥huhu] の波形およびサウンドスペクトログラムを図 5 に示す．図 5 (a) は自然笑い声のものであり，図 5 (b) は合成された笑い声の結果である．図 5 (b) から，自然笑い声の波形に見られる波形の振幅が徐々に減少するという傾向が反映されていることがわかる．このことから，比較的に自然笑い声に近い笑い声が合成されていることがわかる．

5.　自然性評価実験

この節では，前節で合成された文脈を考慮した笑い声が，発話全体としての自然性を改善しているかを確かめるために，自然性評価を行う．

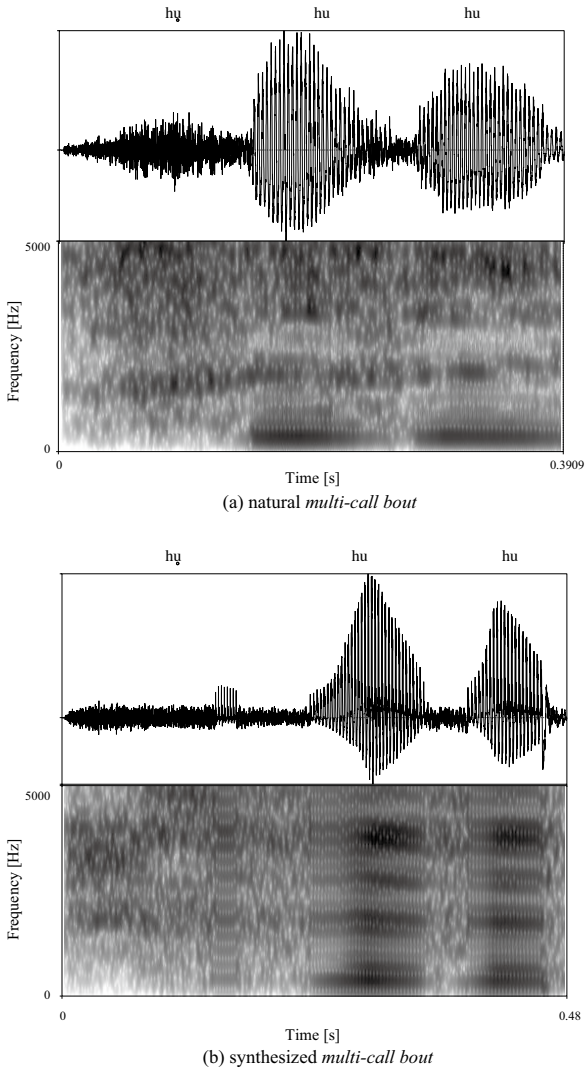


図 5　実際の笑い声と合成された笑い声の比較

発話全体としての自然性を評価するために，合成された笑い声だけではなく，言語音に伴う笑い声が実験に使用される．更に，笑い声コンテキストの有効性も確認するために，適用するコンテキストを段階的に増やした時の自然性を比較する．

5.1　実 験 条 件

実験に使用する笑い声は以下の 3 つの条件で合成された．

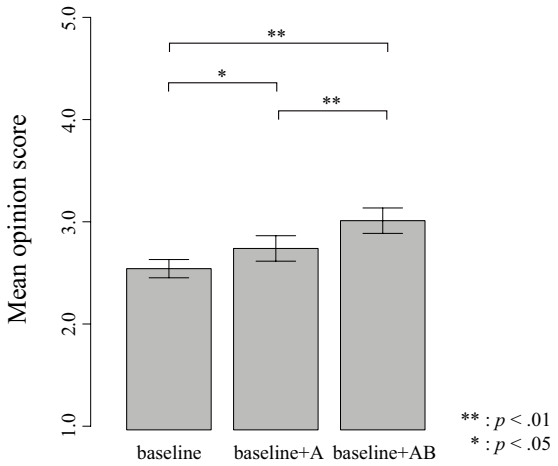


図 6　自然性評価の平均評価値の分布

- ベースライン (BL)
- ベースライン+グループ A (BL+A)
- ベースライン+グループ A+グループ B (BL+AB)

ここで，グループ A およびグループ B は表 3 の分類を指す．

上記の条件で合成された笑い声が言語音に接続された．言語音部分は自然音声の分析合成によって合成され，各条件の笑い声に対して接続された．刺激の数は各条件で 46 個ずつであり，総数は 138 個である．

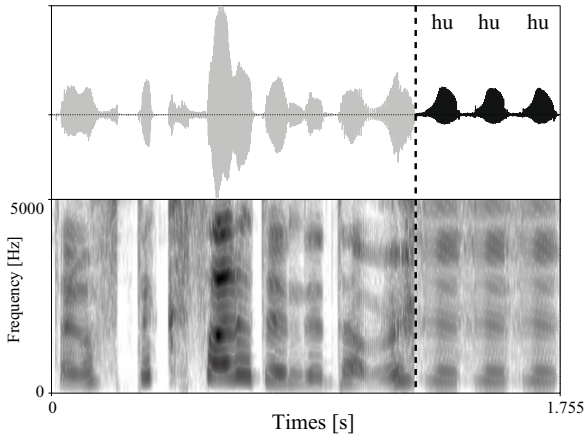
実験には 5 人の男子大学生および 5 人の男子大学院生が参加した．また，参加者は全員日本語話者である．被験者には，各刺激の自然性を 5 段階 (1:不自然, 2:やや不自然, 3:どちらともいえない, 4:やや自然, 5:自然) で評価させた．自然性は「言語音部分と笑い声部分の分節的・韻律的整合性の度合い」と定義した．これは，言語音部分と笑い声部分のパラ言語的な整合性の度合いについても暗に評価している．

実験はヘッドホン (AKG K271 MKII) による両耳聴取によって行われた．刺激は静かな研究室にて呈示され，各刺激は被験者に対し 1 回だけ呈示された．

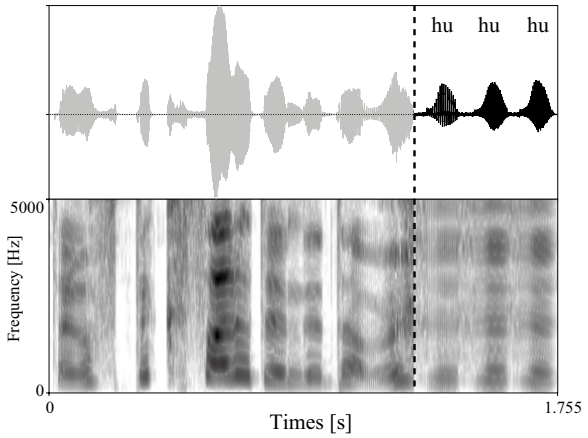
5.2 実験結果

自然性評価実験の結果として，被験者による平均評価値 (MOS) の分布を図 6 に示す．図の区間は 95%信頼区間を表す．BL および BL+A，BL+AB の MOS の平均はそれぞれ 2.54, 2.74, 3.01 である．これらの分布に対して合成条件を要因とする分散分析を行った結果，主効果が有意であった ($F(2, 135) = 17.34, p < .01$)．そのため，Tukey HSD 法による多重比較を行った．その結果，BL と BL+A ($p < .05$)，BL+A と BL+AB ($p < .01$)，BL と BL+AB ($p < .01$) の間の差が有意であった．これらの結果から，文脈に関するコンテキストを増やすことによって自然性が改善されることが確認された．最も自然であったのは BL+AB であり，前後の音韻だけではなく，笑い声が置かれている文脈についての情報が自然性改善に寄与していることがわかる．

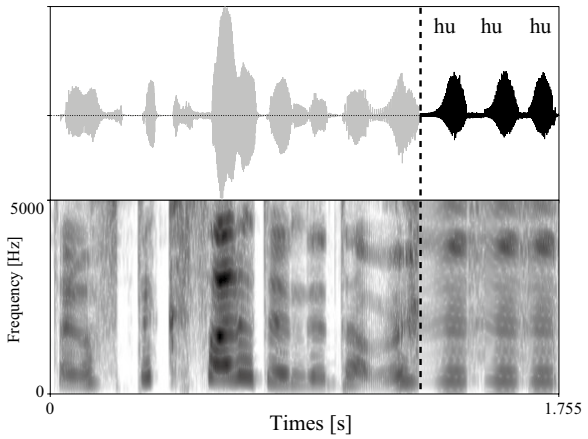
自然性が改善された例を図 7 に示す．図は笑い声を含む発話の波形およびサウンドスペクトログラムを示しており，薄くなっている部分は言語音部分を表す．これは multi-call bout [huhuhu] を合成した例であり，発話の末尾に笑い声が位置している．図 7 (a) は BL によって合成された笑い声が接続されている．BL によって合成された笑い声は全ての call が同じ音響的特徴を持っている．これは，同じ音韻であれば全て同じモデルから合成されるからであり，このように音響特徴量の変化のない単調な音は不自然に知覚される．図 7 (b) は先行・後続



(a) Synthesized with baseline



(b) Synthesized with baseline+A



(c) Synthesized with baseline+AB

図 7　自然性改善例

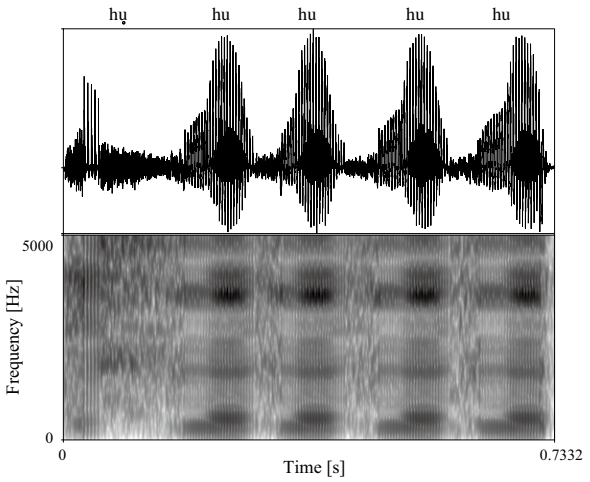


図 8　全体の自然性が低い刺激の例

のセグメントを考慮して合成された笑い声が接続された例である．前後の音を考慮したことにより，全ての call が同じ音響特徴量で出力されることはなく，多少自然性が改善されている．図 7 (c) はより広義の文脈を考慮して合成された笑い声が接続された例である．単に各 call の音響特徴量が異なっているだけではなく，call の振幅が言語音の部分の振幅に近い値になっていることがわかる．このように，文脈を考慮することによって言語音部分と笑い声部分の韻律的・分節的な特徴が整合されたことが自然性改善に寄与していると考えられる．

しかしながら，いくつかの刺激については自然性が改善されない例があった．それは笑い声自体の自然性が低い合成笑い声である．図 8 に，笑い声自体の自然性が低い合成笑い声の波形およびサウンドスペクトログラムを示す．これは，[hu̥huhuhuhu] を合成した例である．文脈を考慮して合成したにも関わらず，[hu] の部分が同じ音響特徴量となっており，非常に単調で不自然な笑い声となっている．コンテキストに基づく手法では，このように必ずしも音響特徴量の変化を反映するように学習されるとは限らない．この問題は笑い声の時間的な構造を異なるモデルでモデル化することにより改善される可能性がある．

6. おわりに

本論文では，自然対話音声コーパスに含まれる笑い声に対して文脈を考慮するためのコンテキストを定義し，HMM 音声合成の枠組みで笑い声を合成した．文脈を考慮したことの有効性を確認するために，合成された笑い声に対する主観評価実験が行われた．主観評価実験では，文脈を考慮した笑い声が考慮しない笑い声よりも発話全体としての自然性を改善させることを確認するために，笑い声単体ではなく，言語音に伴う笑い声に対しての自然性評価が行われた．自然性評価実験の結果から，文脈を考慮することで発話全体としての自然性が改善されることを明らかにした．

残された課題として，自然性以外の観点による評価が挙げられる．今回，発話全体の自然性の観点から主観評価実験を行ったが，笑い声が発話全体から知覚されるパラ言語情報に与える影響などについては調べていない．今後，笑い声の形態と機能の関係を調べるためには，そのような観点からの評価が必要になると考えられる．

また，実際の対話場面におけるインタラクションについても注目する必要がある．例えば，対話相手の笑い声に重複するように発せられた笑い声は，重複しない笑い声と異なる音響的特徴を持つということが報告されている [15]．対話場面における笑い声を合成するためには，このようなインタクラションを考慮する必要があると考えられる．

謝辞 本研究は JSPS 科研費 26280100 の助成を受けた．

文　献

[1] J.A. Bachorowski, M.J. Smoski, and M.J. Owren, “The acoustic features of human laughter,” Journal of Acoustical society of America, vol.110, pp.1581–1597, 2001.

[2] H. Tanaka and N. Campbell, “Classification of social laughterin natural conversational speech,” Computer Speech and Language, vol.28, pp.314–325, 2014.

[3] 志水彰，角辻豊，中村真，人はなぜ笑うのか—笑いの精神生理学，講談社，1994.

[4] A.T. Sathya, K.K. Sudheer, and B. Yegnanarayana, “Synthesis of laughter by modifying excitation characteristics,” Journal of Acoustical society of America, vol.133, pp.3072–3082, 2013.

[5] S. Sundaram and S. Narayanan, “Automatic acoustic synthesis of human-like laughter,” Journal of Acoustical society of America, vol.121, pp.527–535, 2007.

[6] J. Trouvain and M. Schroder, “How (not) to add laughter to synthetic speech,” Proceedings of Workshop on Affective Dialogue Systems, pp.229–232, 2004.

[7] E. Lasarcyk and J. Trouvain, “Imitating conversational laughter with an articulatory speech synthesis,” Proceedings of Interdisciplinary Workshop Phonetics of Laughter, pp.43–48, 2007.

[8] J. Urbain, H. Cakmak, and T. Dutoit, “Development of hmm-based acoutic laughter synthesis,” Proceedings of Interdisciplinary Workshop Laughter and Other Non-Verbal Vocalisations in Speech, pp.26–27, 2012.

[9] J. Urbain, H. Cakmak, and T. Dutoit, “Evaluation of hmm-based laughter synthesis,” Proceedings of International Conference on Acoustics, Speech and Signal Processing, pp.7835–7839, 2013.

[10] H. Mori, T. Satake, M. Nakamura, and H. Kasuya, “Constructing a spoken dialogue corpus for studying paralinguistic information in expressive conversation and analyzing its statistical/acoustic characteristics,” Speech Communication, vol.53, pp.36–50, 2011.

[11] Y. Arimoto, H. Kawatsu, S. Ohno, and H. Iida, “Naturalistic emotional speech collection paradigm with online game and its psychological and acoustical assessment,” Acoustical Science and Technology, vol.33, pp.359–369, 2012.

[12] J. Trouvain, “Segmenting phonetic units in laughter,” Proceedings of International Congress of Phonetic Sciences, pp.2793–2796, 2003.

[13] 森大毅，“Affect burst の音声学的分析 －感情表出系感動詞の言語的・パラ言語的特徴－，” 日本音響学会秋季研究発表会講演論文集，pp.293–296，2015.

[14] J. Yamagishi, M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi, “A context clustering technique for average voice models,” IEICE Transactions on Information and Systems, vol.E86-D, pp.534–542, 2003.

[15] K.P. Truong and J. Truvain, “On the acoustics of overlapping laughter in conversational speech,” Proceedings of Interspeech, pp.851–854, 2012.