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
SP2014-91 (2014-11)

対話音声合成を目的とした発話中の笑い声の変動要因の検討

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　音声合成においては，自然な音声を合成するために前後の合成単位との繋がりやアクセント型といった音声の変動要因を利用している．笑い声に対しても変動要因を明らかにすることで，笑い声の合成に役立ち，対話音声合成の表現力向上に有用であると考えられる．本研究では，笑い声の韻律的な特徴である基本周波数，強度，継続時間と簡易的な分節的特徴である有声確率に関する統計量を利用した笑い声のクラスタリングを行い，どの特徴量がクラスタリングに寄与しているかを確かめることで発話中の笑い声の変動要因を検討する．各クラスタに属する笑い声の音響特徴量の分布の検定を行った結果，強度および継続時間，有声確率についての統計量が分類に大きく寄与していることを示した．更に，本研究ではクラスタリング結果をもとに言語音と笑い声を接続した刺激を作成し，自然性に関する主観評価実験を行った．主観評価実験の結果，言語音に元々後続する笑い声と同じカテゴリに属する笑い声を接続した場合の音声の自然性が元々後続する笑い声と異なるカテゴリに属する笑い声を接続した音声の自然性よりも有意に高くなることを示した．また，被験者による自然性評価実験の平均評価値の分布から，言語音と笑い声の境界付近において強度の変化が大きくなる笑い声を接続した場合に自然性が低下することと，言語音の最終モーラにおける母音と異なる母音のように聞こえる笑い声を接続した場合に自然性が低下することを示した．
キーワード　対話音声，笑い声，クラスタリング

Variable factor of laughter in an utterance for dialogue speech synthesis

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　In speech synthesis, variation factors of speech such as accent type and phonetic unit are used to synthesize natural speech. It is likely that the factors are also effective for the synthesizing laughter. In this study, we perform clustering of laughter using statistics about voicing probability, fundamental frequency, intensity, and duration. Effective factors for classifying laughter are identified by seeking features that contribute to laughter clustering. As the result of the clustering, statistics of intensity, duration and voicing probability were found to contribut greatly to the clustering. Furthermore, subjective evaluation was performed for sounds that were created by connecting speech sound and laughter. Result of naturalness test indicated that sounds composed of speech sound and connected laughter belonging to same cluster was more natural than the ones composed of speech sound and laughter belonging to different clusters. It also indicated that the naturalness was reduced when the connected laughter differed much in intensity, and when the connected laughter had a different vowel quality from the last mora of speech sound.
Key words　dialogue speech, laughter, clustering

1. はじめに

人間同士の音声コミュニケーションにおいては，バーバル情報だけではなくノンバーバル情報を活用している．日常的な対話場面では，特にノンバーバル情報が重要な役割を果たしており，円滑なコミュニケーションに必要不可欠である．近年では人間と機械とのコミュニケーションにおいても音声が用いられるようになってきており，そのような場面でもノンバーバル情

報によるやりとりが必要となることが予想される．

笑いはノンバーバル情報を伝達する代表的な行動であり，音声においてはその行動は笑い声として現れるため，笑い声は人間同士のインタラクションに留まらず，人間と機械のインタラクションにおいても重要な役割を果たすと考えられる．しかし，人間と機械とのインタラクションにおいて，笑い声を考慮するという研究は近年になるまでほとんど行われてこなかった．特に音声合成の分野においては，笑い声の合成は一部の予備的な検討を除いてほとんど行われていない．

笑い声の合成に関する研究には [1] や [2] がある．[1] は笑い声を周期的に振動するバネ-マス系による 2 次系システムによってモデル化し，線形予測に基づいて音声を合成する分析合成方式で笑い声母音を合成しており，[2] では Diphone 合成方式により，柔らかい笑い (soft)・中間的な笑い (modal)・激しい笑い (loud) を合成している．これらの研究では笑い声の合成を実現しているものの，状況や言語音に対して適切な笑い声を合成するという課題については触れていない．笑い声は単体で出現することもあれば，言語音に付随して出現することもある．そのような音声を合成する場合，付随する言語音や笑い声の出現する位置などを考慮し，適切な笑い声を選択する必要がある．

HMM 音声合成方式を代表とする統計モデルに基づく音声合成においては，音素などの分節的特徴やアクセント型といった韻律的特徴による音響特徴量の変動要因を記したコンテキストを用いた決定木クラスタリングを行い，様々な変動要因を考慮したコンテキスト依存モデルを学習することで，任意のテキストの音声の合成を実現している．この手法は笑い声に対しても有効であると考えられる．笑い声の音響的特徴の変動要因を明らかにし，笑い声における適切なコンテキストを定義することができれば，発話中の笑い声の分類および適切な笑い声の合成に有用であると考えられる．

そこで本研究では，発話中の笑い声の音響特徴量の変動要因について検討する．変動要因を明らかにすることで，統計モデルに基づくパラメータ生成による音声合成方式や，素片接続に基づく音声合成方式における適切な笑い声を選択，合成に役立てることを期待する．

2. 笑い声の変動要因

2.1 音響特徴量に基づく笑い声の分類

発話中に出現する笑い声は多様であるため，笑い声をいかに分類するかが肝要となる．笑いの分類には様々な視点があり，例えば [3] では笑いを生成様式の違いなどによって分類して分析している．また，[4] では笑いをコミュニケーションのための笑い，快の笑い，緊張が緩んだ際の笑いの 3 つに分類している．

本研究では，音響特徴量にもとづいた笑い声の分類を行う．音響特徴量にもとづく笑い声の分類を行っている研究には [5] がある．[5] では，笑い声に対して丁寧 (polite)，陽気 (mirthful)，冷笑 (derisive)，その他 (others) というアノテーションを行い，それを教師データとして決定木およびサポートベクターマシンを用いて笑い声の分類を行っている．

本研究では，正解となるラベルを用意せず，音響特徴量を用いた凝集型クラスタリングにより，笑い声のクラスタを作成する．作成したクラスタに属する笑い声の音響特徴量の分布の差を確認し，クラスタリングに有効な音響特徴量を検討する．クラスタリングの結果，クラスタリングにおいて有効であった音響特徴量が発話中の笑い声の変動要因として利用可能かを確かめるために，言語音に対して笑い声を接続して，自然性の評価を行う．クラスタリングにおいてクラスタ間に差のある音響特徴量が発話中の笑いの変動要因として有用であれば，同一クラスタに属する笑いを接続した音声の自然性は異なるクラスタに属する笑いを接続した場合の自然性よりも高くなると考えられる．

表 1　UUDB に収録されている笑い声の数
Table 1　The number of laughter in the UUDB

話者	笑いの数	話者	笑いの数	話者	笑いの数
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		

2.2 対象とするコーパス

本研究では，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) [6] を使用する．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の研究への用途を開発目的とした音声コーパスである．UUDB は，親近性の高い大学生 7 ペア (女性 12 名，男性 2 名) による自然な対話における音声から構成される．音声は「4 コマまんが並べ替え課題」というタスクを与えて収録されている．総発話数は合計で 4840 発話であり，収録されている音声は表情豊かなものとなっている．

UUDB には言語情報やパラ言語情報に関するアノテーションに加え，笑い声，吸気音，呼気音，咳といった非言語音に関するアノテーションも施されている．これらの非言語情報は開始時間と終了時間が記されている．UUDB に収録されている笑い声の総数は 280 個であり，話者ごとの内訳を表 1 に示す．

3. 笑い声のクラスタリング

3.1 クラスタリング条件

本研究では，UUDB の話者 FTS の発話のうち，発話の途中あるいは最後に笑いを含んでいる発話 21 発話を対象とした．また，途中に複数の笑い声を含んでいる発話については，最初の笑いの部分までを対象とした．

音響特徴量には基本周波数，強度，有声確率，継続長を使用する．基本周波数，強度，有声確率については

（1） 平均値
（2） 最大値
（3） 最小値
（4） 最大値位置
（5） 最小値位置

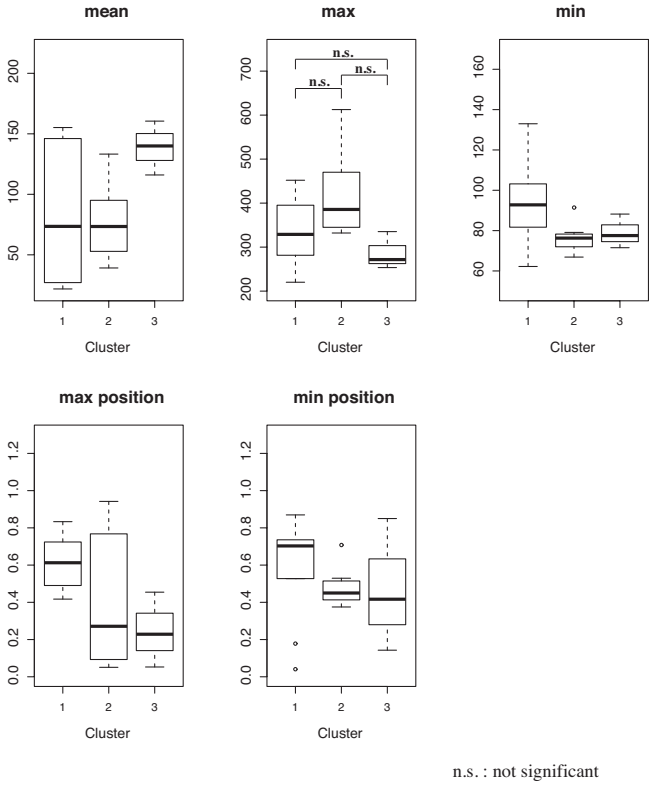

図 1　各クラスタにおける基本周波数に関する特徴量の分布

Fig. 1　The distribution of statistics of fundamental frequency for each cluster

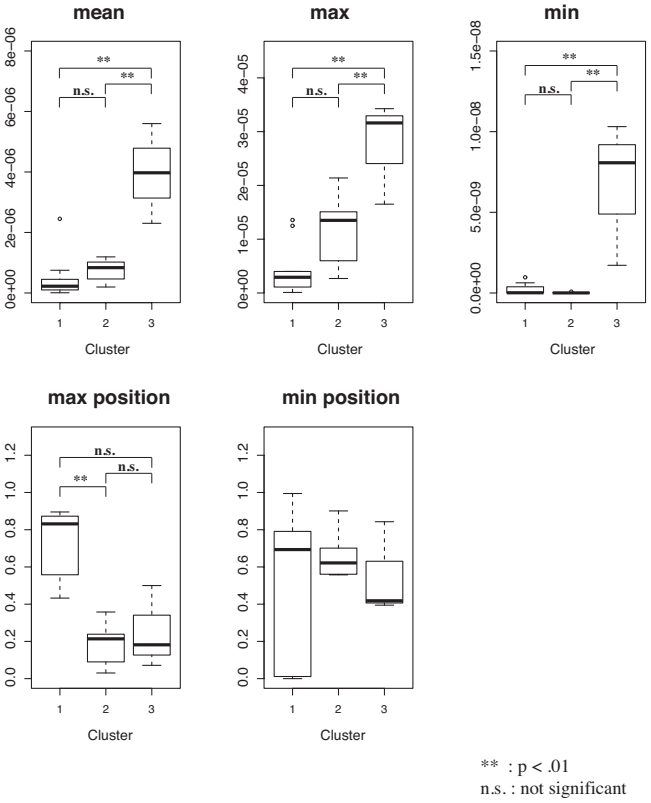

図 2　各クラスタにおける強度に関する特徴量の分布

Fig. 2　The distribution of statistics of intensity for each cluster

を求め，それを音響特徴量とした．特徴量の抽出は openSMILE [7] を用いて行った．このときの分析はサンプリング周波数 16 kHz の音声に対して，フレーム長 25 ms，フレーム周期 5 ms のハミング窓を使用して行った．また，強度最大値位置および強度最小値位置については笑い声の継続長全体の割合としている．

16 種類の音響特徴量を 16 次元の特徴量ベクトルとし，特徴量ベクトルのユークリッド距離をもとに構造的クラスタリングを行った．ここで，特徴量ベクトルの各次元はスケールが異なるため，各次元のデータを標準化してからユークリッド距離を求めた．

クラスタリング手法には Ward 法を用い，クラスタの数は 3 とした．

3.2　クラスタリング結果

基本周波数に関する音響特徴量のクラスタ間における分布を図 1 に示す．各特徴量についてクラスタを要因とする一元配置分散分析を行った結果，基本周波数の最大値のみクラスタの主効果がだった ($F(2,20) = 3.82, p < .05$)．そこで基本周波数最大値について Tukey の HSD 法により有意水準 5%の多重比較を行った．しかし，基本周波数最大値についてはクラスタ間の平均に差は見られなかった．このことから，今回のクラスタリングにおいては基本周波数に関する特徴量は有効に働いていないと考えられる．

各クラスタにおける強度についての特徴量の分布を図 2 に示す．各特徴量についてクラスタを要因とする一元配置分散分析を行った結果，強度最小値位置以外でクラスタの主効果が有意だった (強度平均値: $F(2,20) = 23.77, p < .01$，強度最大値: $F(2,20) = 16.72, p < .01$，強度最小値: $F(2,20) = 24.55, p < .01$，強度最大値位置: $F(2,20) = 29.77, p < .05$)．差の確認できた特徴量に対して，Tukey の HSD 法による多重比較を行い，どのクラスタ間に差があるのかを確かめた．2 中にこの結果を示す．強度平均値，強度最大値，強度最小値についてはクラスタ 1 とクラスタ 3 およびクラスタ 2 とクラスタ 3 において差があることがわかった．また，強度最大値位置についてはクラスタ 1 とクラスタ 2 およびクラスタ 1 とクラスタ 3 において差が確認された．以上のことから，クラスタ 1 には強度のピークが後半に現れる笑いが属しており，クラスタ 3 には強度の強い笑いが属していることがわかる．

有声確率および継続長のクラスタ間における分布を図 3 に示す．他の特徴量と同様に一元配置分散分析を行った結果，有声確率に関する特徴量については有声確率最小値 ($F(2,20) = 5.44, p < .05$)，有声確率最大値位置 ($F(2,20) = 7.85, p < .01$) および有声確率最小値位置 ($F(2,20) = 6.79, p < .01$) においてクラスタの主効果が有意であった．差の確認された特徴量について Tukey の HSD 検定を行い，どのクラスタ間に差があるのかを確かめた．この結果を図 3 中に示す．有声確率最小値についてはクラスタ 1 とクラスタ 3 およびクラスタ 2 とクラスタ 3 の間に差が確認された．有声確率最大値位置についてはクラスタ 1 とクラスタ 2 およびクラスタ 1 とクラスタ 3 の間に差が確認された．クラスタ 3 は有声確率の最小値が比較的高いことか

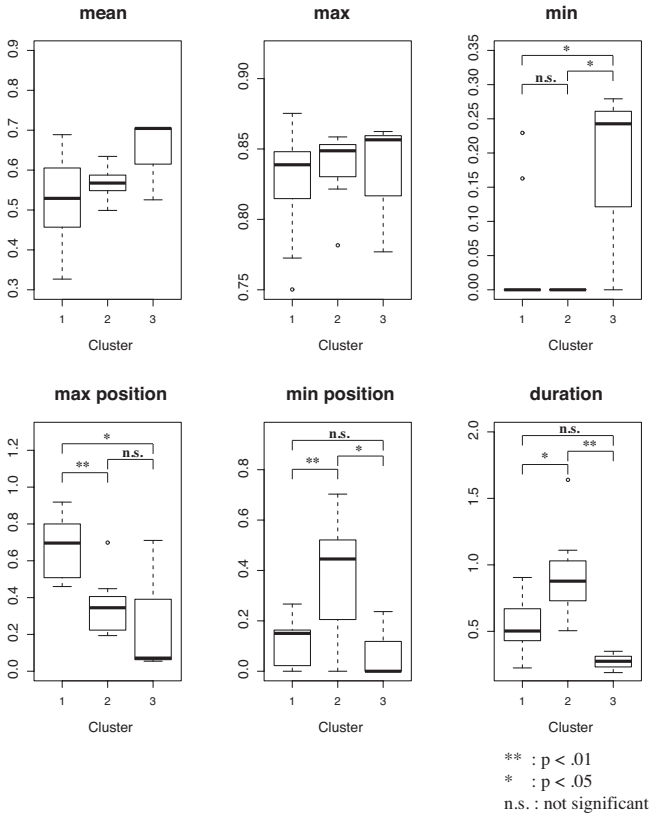

図 3　各クラスタにおける有声確率に関する特徴量と継続長の分布

Fig. 3　The distribution of statistics of voicing probability and duration for each cluster

ら，無声区間の少ない笑いが属していると考えられる．

同様の検討を笑いの継続長についても行った．一元配置分散分析の結果，クラスタの主効果が有意であった ($F(2, 20) = 8.65, p < .01$)．したがって，Tukey の HSD 法により多重比較を行い，どのクラスタ間で差があるのかを確認した．この結果を図 3 中に示す．クラスタ 1 とクラスタ 2 およびクラスタ 2 とクラスタ 3 の間に差が確認された．したがって，クラスタ 2 には継続長の長い笑いが属していることが確認された．

以上より，発話中の笑い声のクラスタリングについては強度および有声確率，継続長に関する特徴量が用いられていることがわかった．

4.　発話中の笑い声の自然性評価実験

3. における笑い声のクラスタリングの結果が発話中の笑い声の変動要因として有用であるかを確認するために，言語音に対して笑い声を接続した音声を用いた自然性評価実験を行った．言語音に元々後続する笑い声と同一のクラスタに属する笑い声を接続した音声と，異なるクラスタに属する笑い声を接続した音声の自然性を比較する．

4.1　実 験 条 件

3.2 における笑い声のクラスタリング結果より，クラスタ 1 およびクラスタ 2 に属する笑い声を含む発話をそれぞれ 5 個ずつ無作為に選択し，その言語音部分を切り出す．笑い声についても同様に，クラスタ 1 およびクラスタ 2 に属する笑い声を，

表 2　全被験者の平均評価値

Table 2　MOS and standard diviation at all examinee.

同一クラスタ	異なるクラスタ
3.297	2.747

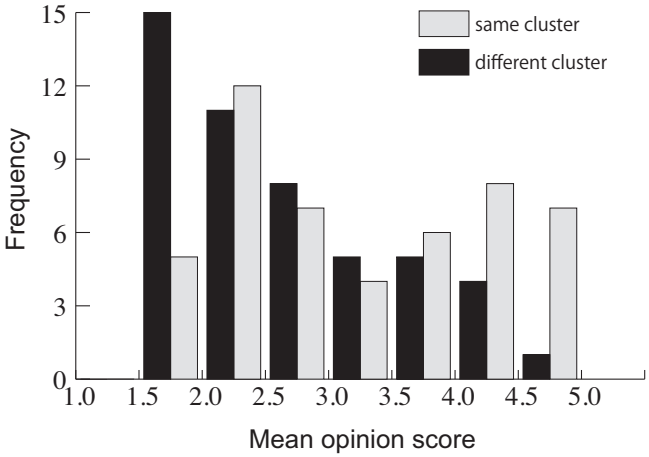

図 4　同一クラスタの笑いを接続した場合の平均評価値のヒストグラムと異なるクラスタの笑いを接続した場合の平均評価値のヒストグラム

Fig. 4　The histgrams for connecting laughter in different cluster and same cluster

先ほど選択した発話以外の発話からそれぞれ 5 個ずつ無作為に選択する．次に，これらの次に，これらの言語音と笑い声を接続して刺激音声を作成する．作成した刺激音声は 10 個の言語音と 10 個の笑い声のすべての組み合わせ $10 \times 10 = 100$ 通りである．

被験者は男子大学生 2 名，男子大学院生 4 名の計 6 名であり，ヘッドフォンによる両耳聴取によって評価を行った．10 発話を 1 セットとするセットリストを 10 セットの用意し，各セット間で言語音と笑い声の組み合わせが重複しないようにランダムに入れ替えて呈示した．

評価は笑い声が言語音部分に対して自然であるかを 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．

4.2　実 験 結 果

言語音に対して同一のクラスタに属する笑い声を接続した刺激を呈示した場合と異なるクラスタに属する笑い声を接続した刺激を呈示した場合の全被験者における平均評価値 (MOS: Mean Opinion Score) を表 2 に示す．表 2 より，言語音に元々後続する笑い声と同一クラスタの笑い声を接続した場合の自然性の方が，異なるクラスタの笑い声を接続した場合の自然性よりも有意に高かった ($t(49) = -2.94, p < .01$)．

図 4 に全被験者の MOS のヒストグラムを示す．言語音に元々後続する笑い声と異なるクラスタの笑い声を接続した場合の MOS は 1.5 より大きく 2.0 以下の部分に多く分布している．この部分に属している刺激は笑い声の直前の言語音と笑い声の

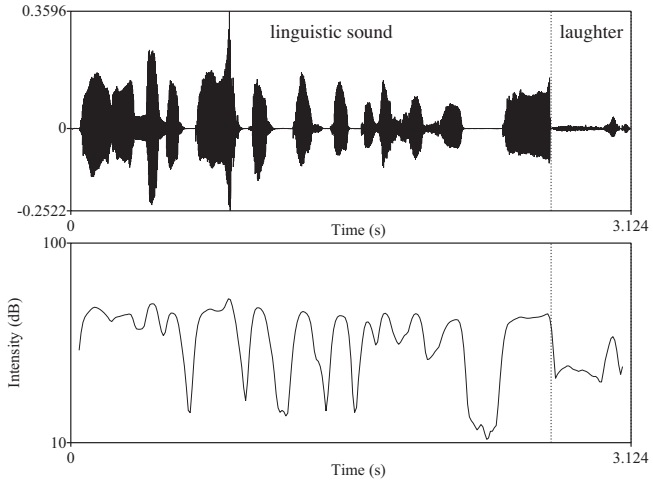

図 5　笑い声の接続部分で強度に急激な変化のある音声

Fig. 5　The speech with a rapid change in intensity at the connection part of laughter.

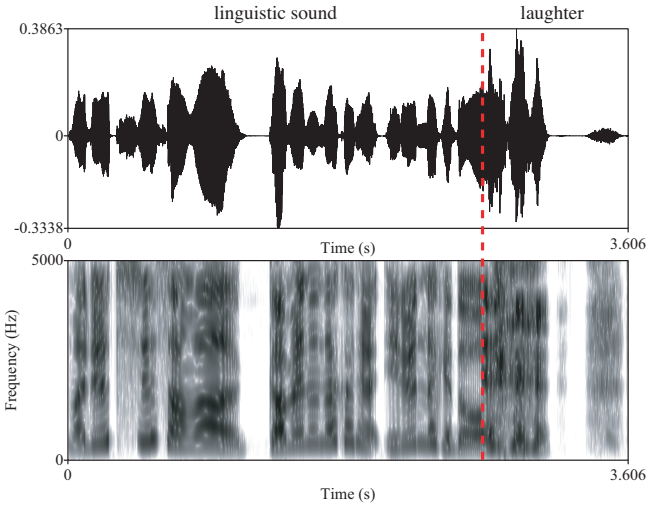

図 6　言語音の最終モーラと異なる音韻の笑いが接続された音声

Fig. 6　The speech with laughter that is different from last mora of speech sound.

区間において強度が急激に変化している刺激が多かった．特に笑い声の開始部分で強度が弱くなる刺激の自然性が低くなるという傾向が見られた．この例を図 5 に示す．図の上側は音声の波形であり，下側は音声の強度である．図より，笑い声の部分で強度が急激に減少していることがわかる．**3.2** より，クラスタ 1 に属する笑い声は強度最大値位置が笑い声の後半，クラスタ 2 に属する笑い声は強度最大値位置が前半に多く分布するようにクラスタリングされているので，強度最大値位置は笑い声の変動要因として有用であると考えられる．

これに対し，言語音に元々後続する笑い声と同一クラスタの笑い声を接続した場合の MOS は 1.5 より大きく 2.0 以下の部分が少なく，4.0 より大きく 4.5 以下および 4.5 より大きく 5.0 以下の部分に多く分布している．言語音に元々後続する笑い声と同一のクラスタの笑い声を接続した場合でも不自然に知覚される刺激については，言語音の最終モーラの母音とは異なる母音のように聞こえる有声音から始まる笑い声が多く含まれていた．図 6 にその例を示す．図の上側は音声波形，下側はスペクトログラムを表す．図より，笑い声の部分で音声のフォルマント周波数に大きな変化が生じていることがわかる．このことから，分節的特徴に関連する音響特徴量を用いたクラスタリングを行うことによって自然性が向上する可能性がある．

以上の結果から，発話中の笑い声の変動要因としては，音声の強度および分節的特徴に関する音響特徴量が有効であると考えられる．

5. おわりに

本研究では，発話中の笑い声の変動要因を明らかにする目的で，笑い声の音響特徴量に基づくクラスタリングを行い，クラスタ間で差のある音響特徴量を確認した．クラスタリングの結果，基本周波数に関する特徴量はクラスタリングにあまり寄与しておらず，音声の強度および有声確率，継続長についての特徴量が大きく寄与していた．クラスタリングの結果に基づき，言語音に元々後続する笑い声と同一のクラスタに属する笑い声を接続した音声と異なるクラスタに属する笑い声を接続した音声を用いて自然性の評価実験を行った結果，同一のクラスタに属する笑い声を接続した場合の方が自然性が高いという結果が得られた．また，自然性評価実験の平均評価値の分布より，発話中の笑い声の変動要因として強度最大値位置が有効であることを示した．更に，分節的特徴に関する音響特徴量が変動要因として有効となる可能性についても示した．

今後の課題としては，他の被験者についての傾向についての調査および分節的特徴に関する音響特徴量を用いた実験が挙げられる．また，笑い声だけではなく，言語音を含む発話全体に関する音響的特徴を用いたクラスタリングについても検討する必要があると考えられる．更に，UUDB 以外の自然対話音声コーパス [8] [9] に含まれる笑い声に関しても検討する必要がある．

文　献

[1] S. Sundaram and S. Narayanan, “Automatic acoustic synthesis of human-like laughter,” Journal of the Acoustical Society of America, vol. 121, no. 1, pp. 527–535, 2007.

[2] J. Trouvain and M. Schroder, “How not to add laughter to synthetic speech,” Proc. the Workshop on Affective Dialogue Systems, pp. 229–232, 2004.

[3] J.-A. Bachorowski, M.J. Smoski, and M.J. Owren, “The acoustic features of human laughter,” The Journal of the Acoustical Society of Americal, vol.110, no.3, pp.1581–1595, 2001.

[4] 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.

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

[6] 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.

[7] F. Eyben, F. Weninger, M. Wollmer, and B. Schuller, “openSMILE the munich open speech and musicc interpretation by latge space extraction toolkit,” TU Munchen,

MMK, 2013.
[8] 有本泰子, 河津宏美, “音声チャットを利用したオンラインゲーム感情音声コーパス,” 日本音響学会 2013 年秋季研究発表会講演論文集, 1-P-46a, pp. 385–388, 2013.
[9] Y. Den and M. Enomoto, “A scientic approach to conversational informatics: Description, analysis,and modeling of human conversation,” Conversational informatics: An engineering approach, pp. 307–330, 2007.