4S-06

多様な笑い声生成のための有声音・無声音間隔の制御

木全　亮太朗†　　上乃　聖†　　李　晃伸†
†名古屋工業大学

1　はじめに

音声の持つ情報には，言語情報・パラ言語情報・非言語情報がある．そして，音声合成における言語情報の表現の品質は近年のニューラルネットワークによるEnd-to-End 型音声合成技術の発展により向上し，対話システムへの適用が期待される．一方で，対話においては言語情報だけではなく笑い声のようなパラ言語情報も重要な役割を果し，状況にあった使い分けが求められる．会話の状況にあった多様な笑い声を生成することで，より自然な会話を実現できる．

また，小山ら [1] は笑い声を「高圧的な笑い」「嘲笑い」「悲観的な笑い」「苦笑い」「愛想笑い」「照れ笑い」「陽気な笑い」の 7 つに分類し，それぞれの笑い声に含まれる呼気，無声吸気，有声吸気の割合が異なることを示した．この結果から本研究では，有声音・無声音 (Voiced / Unvoiced; VU) の間隔を制御することで多様な笑い声の生成できる笑い声音声合成モデルの実現を目指す．

2　笑い声音声合成モデル

笑い声はパラ言語情報であるため，笑い声合成の入力に用いる記号列は標準化されていない．したがって，テキスト表現をする際は bout (呼気音) や吸気音，有声音・無声音などを人手でつけられることが多い．そうした中で Hsu らは，パラ言語情報の一種である NSVs (non-speech vocalizations) を離散表現である疑似音素にするために HuBERT を用いている [2]．HuBERT は，一つの音声を 20 ミリ秒ごとに分割しそれぞれの音声に対して言語情報（例えば音素）や非言語情報（例えば話者性）をもつ特徴を抽出することができる．得られた特徴から，K-means 法を用いてクラスタ割り当てを行うことで離散化し，疑似音素の生成を行う．しかし，笑い声を記号化した疑似音素は直接制御することができないため，多様な笑い声を生成することは難しい．

本研究ではこの手法を拡張し，有声音・無声音の間隔を変化させる機構を新たに導入することで複数の笑い声の生成を目指す．

Controlling voiced and unvoiced intervals for generating various laughter
†Ryotaro KIMATA　†Sei UENO　†Akinobu LEE
†Nagoya Institute of Technology

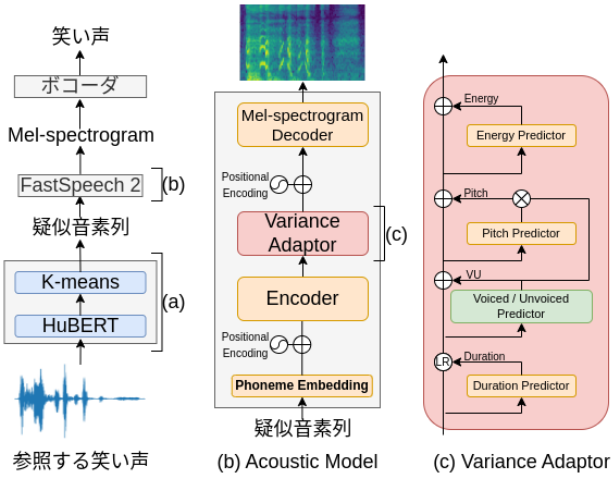

図 1: 提案手法のモデル図

3　有声音・無声音を用いた笑い声音声合成モデル

本研究では， FastSpeech 2 で有声音・無声音の間隔を制御できるような機構を導入し，笑い声から得られた疑似音素を用いて複数の笑い声を出力できる笑い声音声合成モデルを提案する．モデル図を図 1 に示す．

3.1　ベースラインモデル

本研究のモデルは，Hsu らのモデル [2] を参考にしており，3 つの構造からなる．図 1 中の (a) では笑い声を入力として HuBERT を用いて疑似音素に変換する．モジュールとして Libri-Light の 60,000 時間を事前学習した HuBERT と クラスタ数が 200 の K-means 法を用いている．次に図 1 中の (b) の Acoustic model は疑似音素を入力として FastSpeech 2 を用いてメルスペクトログラムを予測する．FastSpeech 2 はエンコーダ，Variance adaptor，デコーダから構成されている．図 1 中の (c) の Variance adaptor には出力音声の音量，ピッチ，音素継続長を予測する Energy predictor， Pitch predictor, Duration predictor が構築されている．それぞれの予測値は Variance Adaptor 内で用いるため，各予測値を変更することで合成音声を制御することができる．本研究では，ピッチの予測値に対して定数をかけることで，制御を行う．最後にボコーダで，メルスペクトログラムを Griffin-Lim アルゴリズムを用いて音声に変換する．

3.2 提案手法

本研究では図 1 中の (c) の Variance adaptor に新たに Voiced / Unvoiced predictor (VU predictor) を追加することで有声音・無声音の間隔の制御を可能にする．VU predictor は 1 つの疑似音素に対して 0 から 1 の連続値の予測を行い，しきい値 vu（デフォルトで vu=0.5）以上で値は 1（有声音）に，それより小さい値で 0（無声音）になる VU 列を生成する．このしきい値 vu を変更することで，有声音・無声音の間隔を制御することができる．また，Pitch predictor により予測されたピッチに VU predictor により予測された VU 列を反映させることで，有声音・無声音かの予測と有声音であるときのピッチの値の予測をすることができる．

また，笑い声コーパスは通常の音声合成に用いられるコーパスに比べてサイズが小さいため FastSpeech 2 の学習が難しい [3]．そのため，学習には読み上げ音声の LJSpeech で事前学習をした後に笑い声コーパスでファインチューニングを行う．

4 評価実験

本研究では， FastSpeech 2 の事前学習に一人の女性の読み上げ音声である LJSpeech コーパス (合計 13100 発話，総時間約 24 時間) を用い，ファインチューニングに笑い声コーパスの UNILAG Laughter Corpus と Vocalsound の女性の発話のみ（合計 6432 発話，総時間約 4.8 時間）を用いて学習を行った．また，笑い声コーパスは，学習・検証・テスト用に 7:2:1 に分割し，22050 Hz にダウンサンプリングした．本実験には 20 代の学生 17 人が参加し，10 発話の笑い声をランダムな順番に聞いて評価した．

評価実験は， 生成する笑い声の多様性について検証するため，1 発話の笑い声に対して従来手法と提案手法で音声を 3 つずつ生成し，多様性の高さを A/B テストで評価した．従来手法で生成する音声は，ピッチの推定値 × 定数 p (=1.2，1.0， 0.8) により変化させた．提案手法で生成する音声は，従来手法同様にピッチを変更し，さらにしきい値 vu を 1.0（すべて有声音），0.5（間隔の変更なし），0.0（すべて無声音）の中からランダムに選び，生成した．詳細は表 1 に示す．

表 2 に各発話に対する A/B テストの結果を示す．10 発話の笑い声の内，8 発話が多様性があると回答された．また，その 8 発話では大多数が 7 割を超える結果となった．これにより，従来手法より提案手法の方が多様性のある笑い声を生成できることが示された．

従来手法の方が多様性があると回答された発話，または従来手法と提案手法の結果の差があまりない発話である，F0213027_7 や f1777_0_laughter，f2482_0_laughter は，一部発話の生成に失敗したため従来手法と提案手法との差が出づらい結果となった．

表 1: 音声生成時に定数 p としきい値 vu を変化させたときの構成の詳細

発話 id	p	vu
SF0101005_L7	1.2 \1.0 \0.8	0.5 \1.0 \0.5
SF0101005_L16	1.2 \1.0 \0.8	0.0 \0.5 \1.0
SF0112018_L20	1.2 \1.0 \0.8	0.0 \0.5 \1.0
F0113021_17	1.2 \1.0 \0.8	0.5 \0.5 \0.5
F0114001_9	1.2 \1.0 \0.8	0.5 \1.0 \0.0
F0213027_7	1.2 \1.0 \0.8	1.0 \0.0 \0.5
f1777_0_laughter	1.2 \1.0 \0.8	1.0 \0.5 \0.5
f2355_0_laughter	1.2 \1.0 \0.8	0.0 \0.5 \1.0
f2482_0_laughter	1.2 \1.0 \0.8	0.5 \1.0 \1.0
f3125_0_laughter	1.2 \1.0 \0.8	0.0 \1.0 \0.5

表 2: A/B テストの結果

発話 id	従来手法 (%)	提案手法 (%)
SF0101005_L7	11.8	**88.2**
SF0101005_L16	11.8	**88.2**
SF0112018_L20	17.6	**82.4**
F0113021_17	17.6	**82.4**
F0114001_9	17.6	**82.4**
F0213027_7	**64.7**	35.3
f1777_0_laughter	35.3	**64.7**
f2355_0_laughter	23.5	**76.5**
f2482_0_laughter	**52.9**	47.1
f3125_0_laughter	17.6	**82.4**

5 むすび

本研究では FastSpeech 2 の Variance adaptor に VU predictor を追加して，多様な笑い声を出力できる笑い声音声合成モデルを提案した．被験者実験の結果では，VU predictor の導入が，多様な笑い声を生成することに貢献することが示された．今後の展望としては，感情ごとの有声音・無声音の間隔を指定できるようにし，感情表現のできる笑い声合成モデルを検討する．

参考文献

[1] 小山俊樹, 有本泰子. ゲームプレイ中に表出した社会的な笑い声のラベリングとその音響分析. 日本音響学会 2020 年春季研究発表会講演論文集, pages 827–828, 3 2020.

[2] Hsu, Chin-Cheng. Synthesizing Personalized Non-speech Vocalization from Discrete Speech Representations. 2022.

[3] Matsumoto Kento, Sunao Hara, and Masanobu Abe. ”Controlling the Strength of Emotions in Speech-Like Emotional Sound Generated by WaveNet.” INTERSPEECH. 2020.