WaveNet を用いた言語情報なし感情音声合成における感情の強さ制御の検討

Study of controlling the Strength of Emotions in Speech-like Emotional Sound Generated by WaveNet

松本 剣斗
Kento Matsumoto

岡山大学 阿部研究室
Abe Laboratory, Okayama University

概要 音声は言語情報と非言語情報を通じて感情情報を伝える．音声合成において，同じテキストから異なる感情音声を合成するためには言語情報と非言語情報を独立して扱えることが望ましい．我々はこれまでに WaveNet を用いた言語情報なし感情音声合成方式について検討した．人と機械の対話での応用のためには感情の強さの制御も必要である．本稿では，提案方式を拡張して感情ラベル操作によって感情の強さを制御する方式を検討する．

1 はじめに

近年，合成音声の明瞭度と自然性の向上は著しく，合成音声はスマートスピーカーや音声アシストアプリケーションなどの商用製品に幅広く用いられている．合成音声はこれらの製品やアプリケーションで重要な役割を持つが，その表現の多様性は，必ずしも十分であるとは言い難い．理由の 1 つは感情表現や話者性といった表現の不足である．

これまで感情表現可能な音声合成の研究がおこなわれてきた．波形接続に基づく音声合成は高品質な音声を合成可能である [1]．しかし，大量の感情音声データ収集のためのコストの課題がある．また，Hidden Markov Model に基づく感情音声合成は，補完や適応を用いることで柔軟なパラメータ制御が可能ではあるが，必ずしも十分な感情音声が合成できているとは言い難い [2] [3] [4] [5].

我々はこれまでに WaveNet [6] を用いて言語情報を含まないが感情情報を伝えることができる音 (Speech-like Emotional Sound： SES) を生成する方式を提案し，提案方式が 75% 程度正しく感情を伝えられることを示した [7]．Human computer interaction (HCI) への応用のためには，それぞれの感情を表現できるだけでなく，弱い怒りや強い怒りのように感情の強さを制御する必要があると考えられる．

本稿では，SES の感情の強さを制御する方式を検討する．従来方式 [7] をベースとして，合成時に入力する感情ラベルを操作することで感情の強さの制御をおこなう．

本稿は以下の通りの構成である．第 2 章では WaveNet の基礎について述べる．第 3 節では提案方式について述べる．第 4 節では評価結果および考察について述べる．第 5 節では結論と今後の課題を述べる．

2 WaveNet の概要

2.1 WaveNet

WaveNet[6] は，過去の波形から直接未来の波形を予測する Convolutional Neural Network である．WaveNet は過去の有限長のデータ点から未来の波形を予測する．そのとき，波形 $\mathbf{x} = \{x_1\ x_2\ \cdots\ x_T\}$ の結合確率 $p(\mathbf{x})$ は次の条件付き確率の積で表現される．

$$p(\mathbf{x}) = \prod_{t=1}^{T} p(x_t \mid x_1, ..., x_{t-1}) \tag{1}$$

WaveNet は，複数の Residual block から構成されており，各 Residual block 中で dilated causal convolution を一回おこなう．dilated causal convolution は畳込みをおこなう際，dilation という数だけ飛び越えて畳み込みをおこなう．その後，次の式で表される gated activation をおこなう．

$$z = \tanh(W_{f,k} * \mathbf{x}) \odot \sigma(W_{g,k} * \mathbf{x}) \tag{2}$$

ただし，$*$ は畳み込み演算，$\odot$ は要素積，$\sigma(\cdot)$ はシグモイド関数を表す．また，W_f と W_g は学習可能な畳込みフィルタであり，k はレイヤーインデックス，f と g はそれぞれ filter と gate を表す．そして，出力層では，μ-law アルゴリズムによって 8 bit に量子化された波形を，$2^8 = 256$ クラスの分類問題として予測する．

2.2 Conditional WaveNet

WaveNet は追加特徴量 $\mathbf{h}$ を補助特徴量として与えることで WaveNet の出力を制御することができる．補助特徴量を追加した際の gated activation は次の計算をおこなう．

$$z = \tanh(W_{f,k}*\mathbf{x}+V_{f,k}*\mathbf{y}) \odot \sigma(W_{g,k}*\mathbf{x}+V_{g,k}*\mathbf{y}) \tag{3}$$

ここで，$\mathbf{y}$ は $\mathbf{y} = f(\mathbf{h})$ によって音声波形と同じ長さになるように変換された特徴量を表す．また，$V * \mathbf{y}$ は 1×1 の畳み込みを表す．

3 提案方式

提案方式は Conditional WaveNet を使用しており，2 つの学習ステップ (Step 1 と Step 2) と音声合成ステップによって構成される．提案方式は従来方式 [7] と比べて 2 つの学習ステップは同様であり，音声合成ステップのみ異なる．提案方式の概要を図 1 に示す．

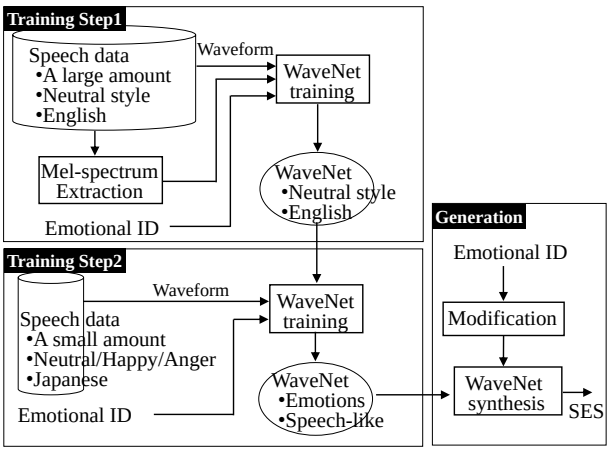

図 1: Outline of the proposed method

3.1 学習 Step 1

Step 1 は，大規模な Neutral 発話データを用いて音声の基礎部分を学習するためのステップである．補助特徴量として，メルスペクトログラムと感情 ID を用いて学習をおこなう．感情 ID は one-of-K 表現を用いている．

3.2 学習 Step 2

Step 2 は，感情音声を用いて再学習をおこない，感情表現を学習するためのステップである．再学習には，感情音声とそれに対応する感情 ID を用いる．また，Step 2 では言語情報を取り除くためメルスペクトログラムは使用しない．

3.3 音声合成

WaveNet に感情ラベルと最初のデータ点を与えることで WaveNet は連続的に音声を生成する．感情の強さを制御する際は，感情 c を示す one-hot ベクトル $\mathbf{e} = \{e_1\ e_2\ e_3\}$ に対して定数 α を用いて次のように修正をおこなう．

$$e'_i = \begin{cases} e_i - \alpha = 1 - \alpha & (i = c) \\ e_i + \frac{\alpha}{2} = 0 + \frac{\alpha}{2} & (i \neq c) \end{cases} \quad (4)$$

ただし，e'_i は修正後の感情 c を示すベクトルの値，α は $0 \leqq \alpha \leqq 1$ とする．本方式では，one-hot ベクトル中の目標となる感情の値と他の感情の値を定数 α によって修正することで感情の強さ制御をおこなう．

4 評価実験

提案方式による有効性を評価するために，感情の強さに関して主観評価実験をおこなった．

4.1 実験条件

実験条件を表 1 に示す．Step 1 では，学習データとして LJ Speech Dataset [9] を使用した．LJ Speech Dataset は，13,100 個の音声ファイル (約 24 時間) から構成されており，女性話者 1 名が英語の文章を “Neutral” 感情で読み上げた音声である．Step 2 では，学習データとして声優統計コーパス [10] を使用した．

表 1: Experimental conditions

Training data	
Corpus (Step 1)	The LJ Speech Dataset 13,100 utterances (24 hours)
(Step 2)	声優統計コーパス 300 utterances (51 minutes)
Sampling freq.	16 kHz
Speech analysis	
Window length	64 msec
Frame shift	16 msec
WaveNet configuration	
Iterations	Step 1: 770,000 iterations Step 2: 40,000 iterations
Mini batch size	4
Optimization	Adam[8]
Residual blocks	30 blocks
Dilations	$[2^0, 2^1, 2^2, ..., 2^9]$ was repeated three times
Input(Step 1)	Waveform: 256 classes × 7680 samples Mel-spectrum: 80 band × 30 frames Emotion ID: 3 types × 1 samples
Input(Step 2)	Waveform: 256 classes × 7680 samples Emotion ID: 3 types × 1 samples
Output	256 classes × 1 samples

声優統計コーパスは日本人話者が日本語の文章を 3 種類の感情 (“Normal” と “Angry”，“Happy”) で読み上げた音声データである．各感情の音声データの長さは約 17 分であり，3 感情合わせて 51 分である．なお，本実験では感情 “Normal” は “Neutral” として扱った．補助特徴量のメルスペクトログラムは短時間フーリエ変換により算出した．

4.2 感情の強さ制御に関する実験

4.2.1 実験方法

感情ラベル操作による感情の強さの制御性能を評価するために Mean Opinion Score (MOS) テストをおこなった．3.3 節で述べたように感情を表すベクトルを定数 α により修正し，修正後のベクトルを用いて音声を合成した．本実験では定数 α として $\{0.5, 0.4, 0.3, 0.2, 0.1, 0.0\}$ の 6 種類を使用した．実験に使用した音声は，感情毎に 5 発話 ×6 種類 = 30 発話であり，3 回繰り返した．また，実験に使用するすべての合成音声の長さは 4 秒間とし，音声の最初と最後にフェード処理をおこなった．実験参加者は 20 代の男女 12 名であり，各発話に対して 5 段階 (5:とても感情が強い – 1:とても感情が弱い) で評価をおこなった．また，実験は “Neutral” を除いた “Angry” と “Happy” に関しておこない，感情ごとに分けて実験をおこなっ

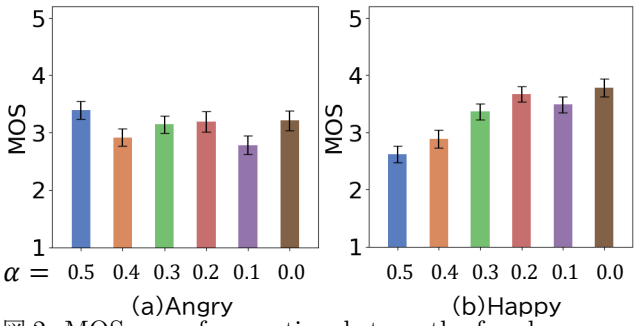

図 2: MOS score for emotional strength of each emotion

表 2: Confusion matrix of conventional method [7]

	Subject-perceived emotions		
Correct emotions	Neutral	Angry	Happy
Neutral	**0.927**	0.050	0.023
Angry	0.300	**0.600**	0.100
Happy	0.200	0.077	**0.723**

た．本実験に使用した合成音声の一部は Web ページ上で試聴できる[*1]．

4.2.2 実験結果

図 2 は感情毎かつ 6 種類の α による実験結果を示す．なお，エラーバーは 95% 信頼区間を示す．また，表 2 は従来方式 [7] の感情認識に関する実験結果の混同行列の再掲である．詳細は [7] に記しているが，実験参加者は 11 名，実験に用いた音声は 30 発話 (10 発話 × 3 感情) である．図 2(a) から，“Angry” では α の違いによって知覚される感情の強さの変化が小さい．また，α の大きさに応じて感情の強さも上下するような相関がなく，感情ラベル操作によって感情の強さの制御が可能とは言い難い．表 2 から “Angry” は “Neutral” との混同が多く，差が小さいため感情ラベルによる制御が難しくなったと考えられる．

一方，図 2(b) から，“Happy” では $\alpha = 0.0$ の one-hot ベクトルの際に最も強い感情として知覚されている．さらに，α を小さくし one-hot ベクトルに近づくにつれて，概ね知覚される感情が強くなっており感情ラベル操作によって感情の強さを制御できているといえる．

“Angry” と “Happy” における知覚される感情の強さの違いを調べるために基本周波数 (fundamental frequency: F0) を分析した．図 3 は実験に使用した “Angry” と “Happy” の 6 種類の音声の対数 F0 と ΔF0 の分布を示す．また，図 4 に “Neutral” の $\alpha = 0.0$ の音声の対数 F0 と ΔF0 の分布を示す．

図 3(a) より “Angry” に関して，MOS スコアの低い $\alpha = 0.1$ や $\alpha = 0.4$ は，図 4 の “Neutral” の分布と似た分布になっており，“Angry” としての感情の強さは弱くなったと考えられる．

図 3(b) より “Happy” に関して，$\alpha = 0.0$ にして one-hot ベクトルに近づくにつれて，高域に分布が広がっていることがわかる．また，ΔF0 に関しては，$\alpha = 0.0$ や $\alpha = 0.2$ の分布が広くなっており，急な F0 の変化が多くある．そのため，$\alpha = 0.0$ や $\alpha = 0.2$ は感情表現が強いと判断されたと考えられる．図 5 は本実験の “Happy” において最もスコアの高い発話 ($\alpha = 0.0$) と低い発話 ($\alpha = 0.5$) の F0 を示す．最もスコアの低い発話と比べて最も高い発話は F0 の平均が高い．また，600 Hz 付近から 300 Hz 付近への F0 の急な変化も存在している．“Happy” においては，感情ラベルを α によって修正することで合成される音声の F0 を変化させることができ，感情の強さの制御が可能であると考えられる．

5 まとめ

本稿では，WaveNet を用いた言語情報なし感情音声合成方式において，感情の強さを制御するために感情ラベルの操作を検討した．評価実験から感情ラベルの操作によって，“Happy” に関しては感情の強さを制御可能であることがが示された．しかし，“Angry” に関しては感情の強さが上手く制御できないことがわかった．

今後の課題として，感情の強さを制御するために，人間の知覚情報の利用や制御するための新たな特徴量を検討したい．

参考文献

[1] R. Barra-Chicote, J. Yamagishi, S. King, J.M. Montero, and J. Macias-Guarasa, “Analysis of statistical parametric and unit selection speech synthesis systems applied to emotional speech,” Speech communication, vol. 52, no. 5, pp. 394–404, 2010.

[2] H. Zen, K. Tokuda, and A.W. Black, “Statistical Parametric Speech Synthesis,” Speech Communication, vol. 51, no. 11, pp. 1039–1064, 2009.

[3] J. Yamagishi, K. Onishi, T. Masuko, and T. Kobayashi, “Modeling of various speaking styles and emotions for HMM-based speech synthesis,” Eighth European Conference on Speech Communication and Technology, pp. 2461–2464, 2003.

[4] T. Masuko, T. Kobayashi, and K. Miyanaga, “A style control technique for HMM-based speech synthesis,” Proceedings of the 8th International Conference of Spoken Language Processing, 2004.

[5] J. Yamagishi, T. Kobayashi, M. Tachibana, K. Ogata, and Y. Nakano, “Model adaptation approach to speech synthesis with diverse

[*1] https://ktmatu.github.io/SES-emotional-strength/

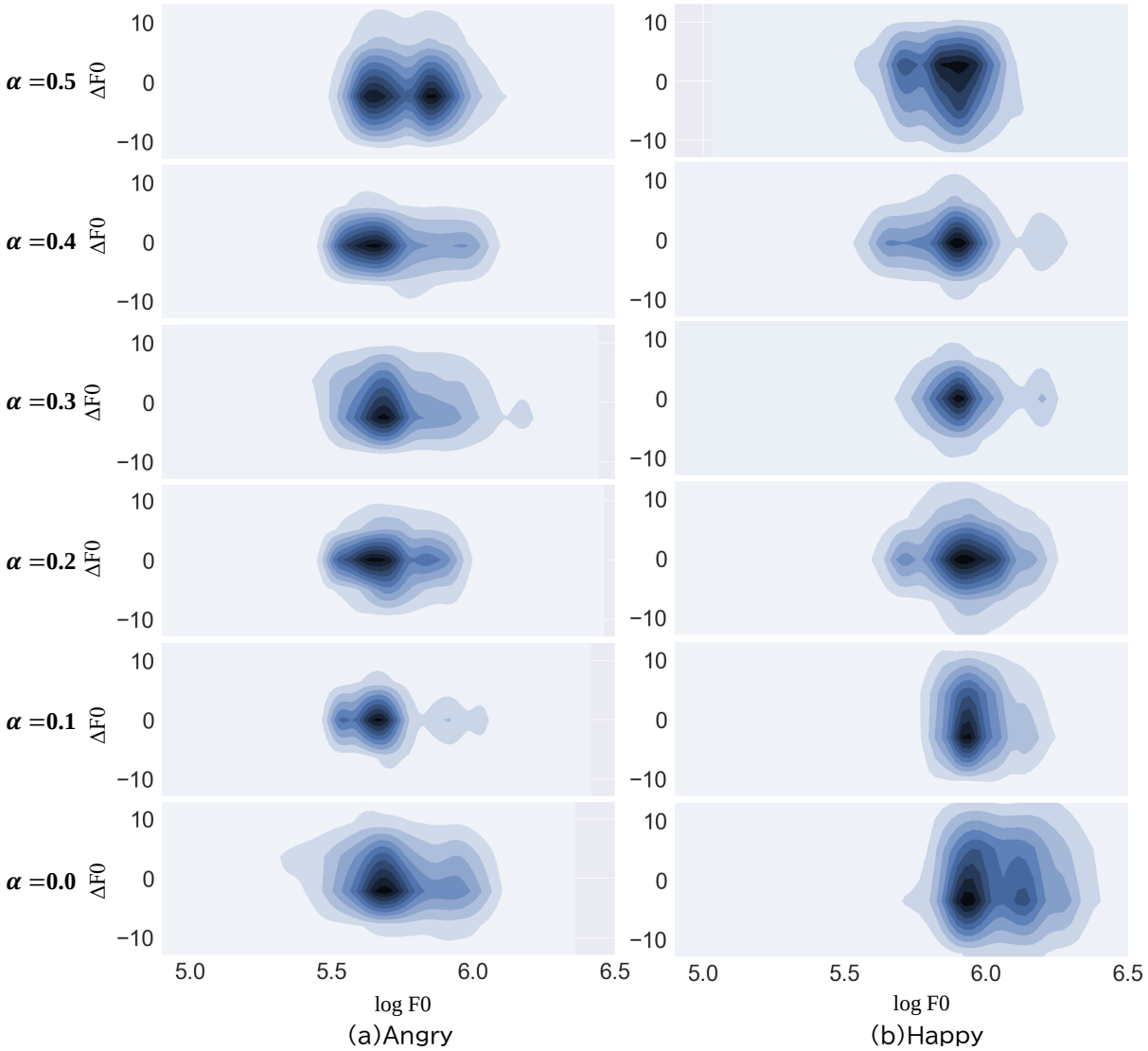

図 3: The distribution of F0 and differential F0

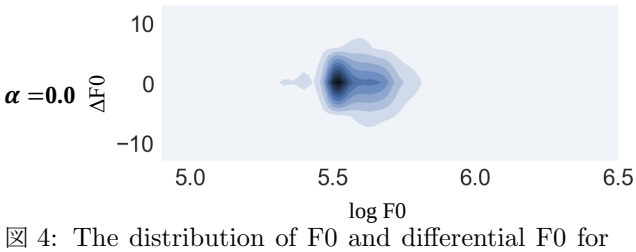

図 4: The distribution of F0 and differential F0 for "Neutral"

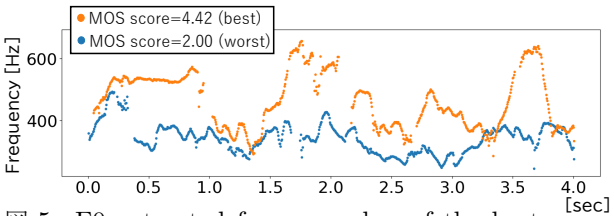

図 5: F0 extracted from speeches of the best score and the worst score for Happy

voices and styles," Proc. ICASSP, pp. 1233–1236, 2007.

[6] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "WaveNet: A Generative Model for Raw Audio," the Computing Research Repository(CoRR) abs/1609.03499, 2016.

[7] K. Matsumoto, S. Hara, and M. Abe, "Speech-like emotional sound generator by wavenet," APSIPA Annual Summit and Conference 2019, pp. 143–147, 2019.

[8] D.P. Kingma, and J. Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980, 2014.

[9] K. Ito, "The LJ speech dataset," `https://keithito.com/LJ-Speech-Dataset/`, 2017, accessed Nov. 2018.

[10] y_benjo, and MagnesiumRibbon, "Voice-Actress Corpus," `http://voice-statistics.github.io/`, accessed Nov. 2018.