Vol.2019-MUS-123 No.61
Vol.2019-SLP-127 No.61
2019/6/23

WaveNetによる言語情報を含まない感情音声合成方式の検討

松本 剣斗[1,a)]　原 直[1,b)]　阿部 匡伸[1,c)]

概要：本稿では，人間の発声した音声のように聞こえるが，言語情報より感情情報の伝達に主眼をおいた音声合成方式を検討する．感情情報は人間のコミュニケーションにおいて重要な役割をもち，音声は感情を伝える便利な方法の一つである．通常，音声は感情情報とともに言語情報も伝える．我々は言語情報を含まず，感情情報を伝える音を生成することができれば，人間と機械の対話をより円滑にするための一助となる場合があると考えている．高品質な音声を合成するために WaveNet を使用する．提案方式は WaveNet の学習に必要な感情音声データの量を減らすために 2 つの学習ステップからなる．第一のステップでは，WaveNet は音韻情報を大規模な音声コーパスを用いて学習する．第二のステップでは，小規模な感情音声コーパスを用いて感情音声のような音を合成できるように WaveNet の再学習をおこなう．主観評価実験では，提案方式による合成音声は人に感情を伝えることができ，人間の声のような音と判断されることが示された．

Study of emotional speech synthesis using WaveNet without linguistic information

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 (Amazon Echo や Google Home など) や音声アシストアプリケーション (Apple Siri) 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，その品質は，必ずしも十分であるとは言い難い．理由の 1 つは感情表現の不足である．人間と機械との対話において，自然な応答を生成するためには，感情表現は重要な役割を果たすと考えられる．様々な研究が過去 20 年間にわたり行われているものの，感情音声の合成は道半ばといえる [1] [2]．

感情合成音声の課題の一つは感情音声データ収録の難しさにある．高品質な音声を合成するためには大量の音声データが必要である．しかし，プロのナレーターや声優以外の人にとって長時間一つの感情を維持して発話することは難しいため大量の感情音声を集めることは難しい．さらに，感情表現は状況や文脈，フレーズに依存しているので適切な文章を作ることも困難である．これらの理由のため，波形接続型の音声合成 [3][4] に基づく感情音声合成の実現は困難であると考えられる [5]．一方，Hidden Markov Model (HMM) に基づく音声合成 [6] は，補完または適応 [7] [8] [9] 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 [10]．

音声は 2 つのチャネルを通して感情情報を伝える．1 つ目は，単語やフレーズといった言語チャネルである．2 つ目は，声質やイントネーションといった音響的特徴の非言語チャネルである．例えば，我々は知らない言語で話しているのを聞いても感情を感じることがある．また，会話において感情を表現するとき，非言語的な感情の発声が重要な場合があることも報告されており，言語情報を必要とする Text-To-Speech (TTS) によって感情音声を合成するだけでは，必ずしも十分とは言えないと考えられる [11] [12]．これは，感情音声を合成するためには言語チャネルと非言語チャネルを独立して扱えることの必要性を示唆している．

本稿では，言語情報を含まないが感情情報は伝える，人間の発声した音声のような音を生成する方式を提案する．

[1] 岡山大学 大学院ヘルスシステム統合科学研究科
[a)] k_matsu@a.cs.okayama-u.ac.jp
[b)] hara@okayama-u.ac.jp
[c)] abe-m@okayama-u.ac.jp

Vol.2019-MUS-123 No.61
Vol.2019-SLP-127 No.61
2019/6/23

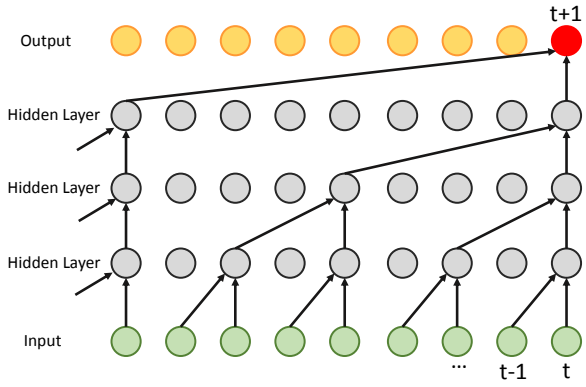

図 1　Dilated causal convolution

提案方式では，言語チャネルとは無関係に，非言語チャネルによって感情情報を伝えることが可能になる。提案方式は高品質な音声を合成するために WaveNet [13] を採用している．そして，学習に必要な感情音声データの量を減らすために，2 つのステップからなる学習方式としている．

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 章では提案方式について述べる．第 4 章では評価実験および考察について述べる．最後に第 5 章では結論と今後の課題を述べる．

2. WaveNet の概要

2.1 WaveNet

WaveNet[13] 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 R 個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．

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

式 (1) は，波形の各データ点 x_t は過去の R 個のデータ点によって条件付けされていることを示している．予測の際に使用する有限長の過去のデータ点数 R は受容野 (receptive field) の大きさを表す．波形を予測する場合，大きな受容野を確保する必要がある．そこで，図 1 のような dilated causal convolution という畳込み手法を利用する．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．

図 2 は WaveNet のネットワーク構造を示す．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．図 2 中の 1×1 は，1×1 のフィルタによる畳み込みを表している．また，Gated は gated activation を表しており，次の計算をおこなっている．

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

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．

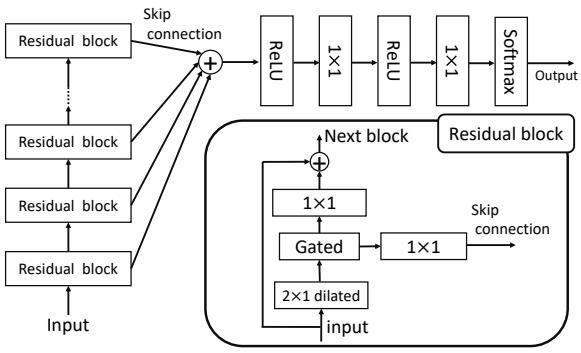

図 2　WaveNet

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条件付き分布 $p(\mathbf{x} \mid \mathbf{h})$ をモデリングすることができる．この場合，式 (1) は次の式に書き換えられる．

$$p(\mathbf{x} \mid \mathbf{h}) = \prod_{t=1}^{T} p(x_t \mid x_1, ..., x_{t-1}, \mathbf{h}) \qquad (3)$$

追加特徴量により条件付けをおこなうことで，WaveNet の出力を制御することができる．また，補助特徴量を追加した際の Gated では，次の計算をおこなう．

$$z = \tanh(W_{f,k} * \mathbf{x} + V_{f,k} * \mathbf{y}) \odot \sigma(W_{g,k} * \mathbf{x} + V_{g,k} * \mathbf{y}) \quad (4)$$

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，$V * \mathbf{y}$ はここでは 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 つのステップ (Step 1 と Step 2) によって構成されている．2 つのステップにおいて WaveNet の構造は同じであるが，使用する補助特徴量と学習データが異なる．提案方式の概要を図 3 に示す．

3.1 学習 Step 1

Step 1 は，言語情報を含む音声を生成できるように学習をおこない，音声の基礎部分を学習するためのステップである．大規模な通常の発話 (以下 “normal” と表記) データを用いて学習をおこなう．補助特徴量としては，メルスペクトログラムと感情ラベルを用いる．感情ラベルは one-of-K 表現を用いている．Step 1 では感情はすべて “normal” であるので感情ラベルは常に “normal” に対応する部分にの

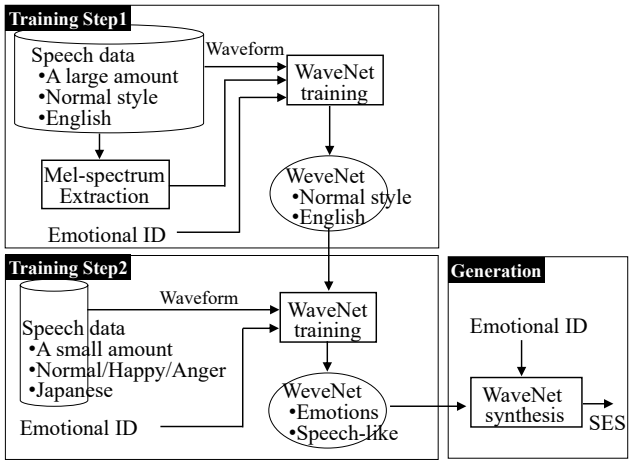

図 3 Outline of the proposed algorithm

み 1 を設定する．Step 1 の場合，式 (4) は次のように書き換えられる．

$$z = \tanh(W_{f,k} * \mathbf{x} + V_{f,k} * \mathbf{y_{EID}} + U_{f,k} * \mathbf{y_{mel-spectrum}}) \odot \sigma(W_{g,k} * \mathbf{x} + V_{g,k} * \mathbf{y_{EID}} + U_{f,k} * \mathbf{y_{mel-spectrum}}) \quad (5)$$

ここで，$\mathbf{y_{EID}}$ は感情ラベルを入力信号の長さに合わせて複製した行列を表し，$\mathbf{y_{mel-spectrum}}$ は，メルスペクトログラムを入力信号の長さに合わせて transposed convolution によってアップサンプリングをおこなった行列を表す．また，ここでは，$U * \mathbf{y}$ は 1×1 のフィルタによる畳込みを表す．

予備実験によって，音声データの量は 20 時間程度とした．音声データの量が少ない場合，WaveNet は人間の声とは異なる，周期的に同じような波形を繰り返す音を生成する傾向がみられた．大量の学習データによって，様々なスペクトルを持つ音声を生成可能になる．すなわち，大量の感情音声データは収録が難しいので，感情音声データのみでは WaveNet の学習をおこなうことは困難であると考えられる．

3.2 学習 Step 2

Step 2 は，Step 1 で学習した重み係数を初期値として利用することで，人間の発声した音声らしさを保存しつつ，感情情報の制御を学習するためのステップである．感情音声データとそれに対応する感情ラベルを用いて Step 1 で学習された WaveNet モデルの再学習をおこなう．Step 2 では，メルスペクトログラムは学習に用いない．Step 1 の場合，式 (5) からメルスペクトログラムの項を除いた次のように書き換えられる．

$$z = \tanh(W_{f,k} * \mathbf{x} + V_{f,k} * \mathbf{y_{EID}}) \odot \sigma(W_{g,k} * \mathbf{x} + V_{g,k} * \mathbf{y_{EID}}) \quad (6)$$

表 1 Experimental conditions

Training data	
Corpus	Step 1: The LJ Speech Dataset (24 hours) Step 2: 声優統計コーパス (137 minutes)
Sampling freq.	16 kHz
Training data	Step 1: 13,100 utterances (24 hours) Step 2: 285 utterances (51 minutes)

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[14]
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 感情ラベル: 3 types × 1 samples
Input(Step 2)	Waveform: 256 classes × 7680 samples 感情ラベル: 3 types × 1 samples
Output	256 classes × 1 samples

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．

4. 評価実験

提案方式による有効性を評価するために主観評価実験をおこなった．主観評価実験では，合成音声の自然性と感情認識率について調べる．また，音響特徴量を用いて主観評価実験の考察および合成音声の分析をおこなう．

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 the LJ Speech Dataset [15] を使用した．The 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 “normal” 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 [16] を使用する．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．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 評価実験用音声

主観評価実験のためにイタリア語とドイツ語で発話された感情音声データを使用した．イタリア語としては the EMOVO Corpus [17] の中から，“normal” と “angry”，“happy” とラベル付けされた音声データを使用した．ドイツ語としては Berlin Emotional Speech (EMO-DB) [18] の中から，“normal” と “angry”，“happy” とラベル付けされた音声データを使用した．なお，使用する音声は女性話者の音声である．また，実験には各感情から 5 発話を選び使用し，合計 30 発話 (2 言語 × 5 発話 × 3 感情) となる．選んだ 30 発話に対して，WORLD ボコーダ [19] と MLSA ボコーダ [20] によって分析合成をおこなう．以下，合成された分析合成音をそれぞれ，WORLD と MLSA と呼ぶ．MSLA ボコーダでは WORLD により分析されたスペクトル包絡から近似した 0–39 次メルケプストラムを用いて合成をおこなった．このとき，フレームシフト長は 5 ms とした．MSLA ボコーダの音声合成フィルタには MLSA フィルタを用いた．

主観評価実験に用いる音声データは，分析合成をおこなっていない元音声 (以下，ORIGINAL) と WORLD，MLSA，提案方式による合成音声 (以下，WAVENET) の 4 種類である．WAVENET は 3 種類の感情ラベルにより合成された 30 発話 (10 発話 × 3 感情) を用いる．なお，WAVENET は，学習に用いた声優統計コーパスの各音声の長さと同じ長さで生成をおこない，さらに，ランダムに 3 sec を抽出した音声である．以上より，主観評価実験に使用する音声は 120 発話となる．

4.3 感情認識に関する実験

4.3.1 実験方法

合成音声の感情表出度を評価するために感情認識テストをおこなった．実験に使用する発話は 4.2 節で述べた 120 発話である．各発話は 2 回出現するようにして，合計発話数は 240 である．音声を流す順番はランダムとする．言語チャネルの影響をなくし，非言語チャネルの感情情報に集中するために，実験参加者は日本語を母語とするイタリア語とドイツ語を知らない参加者とした．また，実験参加者には，音声データはどこかの国の言葉であると説明をおこなった．実験参加者は 11 人であり，実験参加者は音声を聞いて 3 つの選択肢 (“normal”，“angry”，“happy”) の中から感情を選択した．

4.3.2 実験結果

表 2 は ORIGINAL と MLSA，WORLD，WAVENET の実験結果の混同行列を示している．ORIGINAL と MLSA，WORLD では，“angry” は 75%以上正しく認識されている．しかし，“happy” は 40%程度である．ORIGINAL の結果より，“happy” は “normal” と似ていると考えられ，この理由により認識率が低下していると考えられる．一方，

表 2 Confusion matrixes of Experimental results

Confusion matrix of ORIGINAL			
	Subject-perceived emotions		
Correct emotions	Normal	Angry	Happy
Normal	**0.900**	0.073	0.027
Angry	0.232	**0.750**	0.027
Happy	0.523	0.023	**0.455**

Confusion matrix of MLSA			
	Subject-perceived emotions		
Correct emotions	Normal	Angry	Happy
Normal	**0.909**	0.082	0.009
Angry	0.195	**0.782**	0.023
Happy	0.582	0.045	**0.368**

Confusion matrix of WORLD			
	Subject-perceived emotions		
Correct emotions	Normal	Angry	Happy
Normal	**0.900**	0.082	0.018
Angry	0.150	**0.827**	0.023
Happy	0.564	0.023	**0.414**

Confusion matrix of WAVENET			
	Subject-perceived emotions		
Correct emotions	Normal	Angry	Happy
Normal	**0.927**	0.050	0.023
Angry	0.300	**0.600**	0.100
Happy	0.200	0.077	**0.723**

表 3 Confusion matrix of training data

	Subject-perceived emotions		
Correct emotions	Normal	Angry	Happy
Normal	**1.000**	0.000	0.000
Angry	0.000	**0.994**	0.006
Happy	0.000	0.011	**0.989**

WAVENET に関して他の方式と比較すると，“happy” ははるかに良い結果，“angry” は少し悪い結果になっている．このような結果となった理由を調べるために WAVENET の学習データに対して同様の感情認識テストをおこなった．結果を表 3 に示す．学習データに対する感情認識テストの結果より，学習データのおかげで WAVENET は “happy” に関して認識率が向上したと考える．これは，WAVENET が Step 2 で感情表現の特徴を効果的に取得したことを示している．また，学習データセットとしては，感情表現ができるだけハッキリと分かれていることが重要であると言える．

4.3.3 基本周波数による分析

感情により違いが現れる音響特徴量の 1 つに基本周波数 (Fundamental frequency: F0) がある．ここでは F0 と F0 の動的特徴量 ΔF0 によって分析をおこなう．分析には，Step 2 で用いた学習データ 300 発話 (各感情 100 発話) と

Vol.2019-MUS-123 No.61
Vol.2019-SLP-127 No.61
2019/6/23

表 4　コーパス音声の感情毎対数 F0 および ΔF0 の統計量

	F0		ΔF0	
	Avg	SD	Avg	SD
normal	2.366 (217 Hz)	0.120	−5.016 Hz/5ms	33.892
angry	2.421 (264 Hz)	0.114	−5.921 Hz/5ms	36.786
happy	2.622 (419 Hz)	0.097	−11.977 Hz/5ms	69.621

表 5　合成音声の感情毎対数 F0 および ΔF0 の統計量

	F0		ΔF0	
	Avg	SD	Avg	SD
normal	2.427 (267 Hz)	0.063	−0.717 Hz/5ms	14.469
angry	2.463 (290 Hz)	0.074	−2.854 Hz/5ms	28.900
happy	2.606 (404 Hz)	0.082	−6.559 Hz/5ms	54.739

WAVENET を含む合成音声 300 発話 (各感情 100 発話) を用いる．以下，Step 2 で用いた学習データをコーパス音声，WAVENET を含む合成音声を合成音声と呼ぶ．コーパス音声と合成音声の対数 F0 の平均と標準偏差および ΔF0 の平均と標準偏差をそれぞれ表 4，表 5 に示す．表 4 と表 5 中の Avg は平均値，SD は標準偏差を示す．ただし，F0 の平均値と標準偏差は対数に変換してから算出した．表 4 と表 5 から，感情ごとの F0 の平均値においてはコーパス音声と合成音声は近い値となっている．しかし，F0 の標準偏差においては，合成音声の方がコーパス音声よりも小さい値となっている．一方，ΔF0 の平均値と標準偏差に関しては，コーパス音声の方が合成音声よりも各値の絶対値が大きい傾向がある．しかし，合成音声の各感情ごとの ΔF0 の平均値と標準偏差の大小関係はコーパス音声と同じようになっている．ΔF0 の平均値においては，大きい順に normal，angry，happy となっており，ΔF0 の標準偏差においては，大きい順に happy，angry，normal となっている．ΔF0 は F0 の傾きを表しているので ΔF0 の標準偏差が大きい場合，音声の中に F0 の急な変化があることを示している．コーパス音声の “happy” は表 4 より F0 の急な変化が多いことがわかり，同様に合成音声の “happy” も表 5 より F0 の急な変化が多いことがわかる．

合成音声はコーパス音声の特徴を完全に再現できているわけではないが，感情間の大小関係が維持されていることや対数 F0 の平均値が近いことなど部分的に再現できていると考えられる．特にコーパス音声の “happy” は他の感情と比べて F0 の違いが大きいため，合成音声においても “happy” は他の感情と比べて F0 の違いが現れている．そのため，感情認識実験では WAVENET の “happy” の正解率が高くなっていると考えられる．

4.4　自然性に関する実験

4.4.1　実験方法

合成音声の自然性を評価するために Mean Opinion Score (MOS) テストをおこなった．実験参加者と使用した発話データは 4.3 と同様とし，それぞれ 11 人と 240 発話である．実験参加者は各発話に対して 5 段階 (5: 非常に人間の声のように聞こえる，4:人間の声のように聞こえる，3:どちらでもない，2:機械音のように聞こえる，1:非常に機械音のように聞こえる) で評価をおこなった．また，5 段階の判定は感情の種類ごとにおこなった．すなわち，ある感情に関して，ORIGINAL，WORLD，MLSA，WAVENET の中からランダムに流し，実験参加者が 5 段階で判定し，その後，他の感情に関しても同様におこなった．

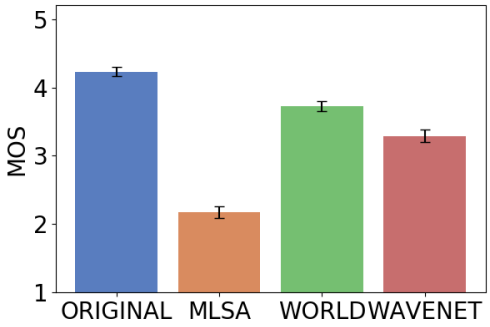

図 4　Average MOS score for all emotions

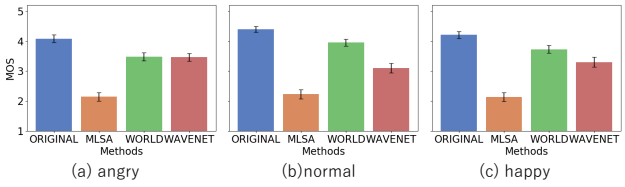

図 5　MOS score for each emotion

4.4.2　実験結果

図 4 に自然性に関する実験の結果を示す．ただし，図 4 は各方式におけるすべての感情の平均値を示している．また，エラーバーは 95%信頼区間を示す．結果を見ると，ORIGINAL の MOS 値が必ずしも 5 であるとは限らないことがわかる．WORLD は ORIGINAL を分析して得たパラメータから合成をおこなった分析合成音である．一方，WAVENET は感情ラベルのみから生成された音声である．しかし，WORLD と WAVENET の MOS 値の差はわずかである．これは提案方式は感情音声について上手く学習していたことを示している．

図 5 は，方式毎かつ感情毎の実験結果を示している．図 5 より，“angry” と “happy” に関して WORLD と WAVENET の間に大きな差はない．しかし，“normal” に関しては比較的差が大きい．これは主に “angry” と “happy” のスペクトルや基本周波数において急な変化があるため，WORLD の分析合成が困難であったためと考えられる．これに対して，“normal” は，急激な変化が少ないため，その困難さが少なく，品質のよい合成音声が生成されていたと考えられる．

5. おわりに

本稿では，WaveNet を用いた言語情報を含まないが感情情報は伝える音を生成する方式を提案した．提案方式は 2 つのステップからなる．Step 1 では巨大な音声データによって音声の基礎部分を学習し，Step 2 では少量の感情音声データによって感情表現を学習する．実験結果より，提案方式は言語情報を含まず感情情報含む音声を生成できていると考えられる．さらに，音声の品質に関しては人間が発話した音声と同等である．

今後の課題として，様々な感情への対応や学習データ量と合成音声の品質との関係を明らかにすることが挙げられる．

参考文献

[1] Schröder, M.: Emotional speech synthesis: A review, *Proc. of EUROSPEECH*, pp. 561–564 (2001).
[2] Schröder, M.: Expressive speech synthesis: past, present, and possible futures, *Affective Information Processing*, pp. 111–126 (2009).
[3] Black, A. and Campbell, N.: Optimising selection of units from speech databases for concatenative synthesis, *Proc. of EUROSPEECH*, International Speech Communication Association, pp. 581–584 (1995).
[4] Mizuno, H., Asano, H., Isogai, M., Hasebe, M. and Abe, M.: Text-to-speech synthesis technology using corpus-based approach, *NTT Technical Review*, pp. 70–75 (2004).
[5] Iida, A. and Campbell, N.: Speech database design for a concatenative text-to-speech synthesis system for individuals with communication disorders, *International Journal of Speech Technology*, Vol. 6, No. 4, pp. 379–392 (2003).
[6] Zen, H., Tokuda, K. and Black, A. W.: Statistical Parametric Speech Synthesis, *Speech Communication*, Vol. 51, No. 11, pp. 1039–1064 (2009).
[7] Yamagishi, J., Onishi, K., Masuko, T. and Kobayashi, T.: Modeling of various speaking styles and emotions for HMM-based speech synthesis, *Eighth European Conference on Speech Communication and Technology*, pp. 2461–2464 (2003).
[8] Masuko, T., Kobayashi, T. and Miyanaga, K.: A style control technique for HMM-based speech synthesis, *Proceedings of the 8th International Conference of Spoken Language Processing* (2004).
[9] Yamagishi, J., Kobayashi, T., Tachibana, M., Ogata, K. and Nakano, Y.: Model adaptation approach to speech synthesis with diverse voices and styles, *Proc. ICASSP*, pp. 1233–1236 (2007).
[10] Barra-Chicote, R., Yamagishi, J., King, S., Montero, J. M. and Macias-Guarasa, J.: Analysis of statistical parametric and unit selection speech synthesis systems applied to emotional speech, *Speech communication*, Vol. 52, No. 5, pp. 394–404 (2010).
[11] Trouvain, J. and Schröder, M.: How (Not) to Add Laughter to Synthetic Speech, *Proc. Workshop on Affective Dialogue Systems*, pp. 229–232 (2004).
[12] Schröder, M., Heylen, D. K. and Poggi, I.: Perception of non-verbal emotional listener feedback, *Proc. of Speech Prosody*, pp. 43–46 (2006).
[13] van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A. and Kavukcuoglu, K.: WaveNet: A Generative Model for Raw Audio, the Computing Research Repository(CoRR) abs/1609.03499 (2016).
[14] Kingma, D. P. and Ba, J.: Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980 (2014).
[15] Ito, K.: The LJ Speech Dataset, `https://keithito.com/LJ-Speech-Dataset/` (2017). accessed Nov. 2018.
[16] y_benjo and MagnesiumRibbon: Voice-Actress Corpus, `http://voice-statistics.github.io/`. accessed Nov. 2018.
[17] Costantini, G., Iadarola, I., Paoloni, A. and Todisco, M.: EMOVO Corpus: an Italian Emotional Speech Database, `https://core.ac.uk/download/pdf/53857389.pdf`. accessed March. 2019.
[18] Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. and Weiss, B.: A Database of German Emotional Speech, *INTERSPEECH*, pp. 1517–1520 (2005).
[19] Morise, M., Yokomori, F. and Ozawa, K.: WORLD: a vocoder-based high-quality speech synthesis system for real-time applications, *IEICE transactions on information and systems*, Vol. E99-D, No. 7, pp. 1877–1884 (2016).
[20] Imai, S., Sumita, K. and Furuichi, C.: Mel log spectrum approximation (mlsa) filter for speech synthesis, *Electronics and Communications in Japan (Part I: Communications)*, Vol. 66, No. 2, pp. 10–18 (1983).