Vol.2018-MUS-119 No.30
Vol.2018-SLP-122 No.30
2018/6/16

LSTMによる人物の印象を考慮した音声合成手法の検討

森 優斗[†1,a)]　井上 勝文[†1,b)]　吉岡 理文[†1,c)]

概要：音声合成では，機械音声の不自然さを除去した，人間の声に近い音声が求められている．特に，娯楽作品では登場人物の魅力や特徴を出すため，「明るくて元気な少女」といった人物の印象が表現できる音声合成が求められている．本研究では，ニューラルネットワーク，特に Long Short Term Memory(LSTM) を用いた複数人話者の音声の学習に，人物の印象を表す印象ベクトルを追加入力し，人物の印象を付与した音声を生成することを目的とする．

Speech synthesis using LSTM with impression

Mori Yuto[†1,a)]　Inoue Katsufumi[†1,b)]　Yoshioka Michifumi[†1,c)]

1. はじめに

任意の文章から対応した音声を自動的に作り出す技術をテキスト音声合成，または単に音声合成という．音声合成は，人の声を録音するのに比べて，新しい音声を作成するのが非常に容易である．例えば，声を録音する際には，防音室のような限られた収録環境の準備や，話者がその場にいる必要がある．しかし，音声合成では機材さえあれば，時，場所を問わず音声を作成できる．音声合成の手軽さから，視覚障がい者のための文字読み上げやカーナビゲーションなど，人手で大量の音声を作るのが難しいような場面で多く用いられている．ところが，音声合成はゲームや動画といった娯楽作品においては用いられることが少ない．理由として，娯楽作品においてはキャラクタの多様な個性を表現する必要があることが挙げられる．キャラクタの個性とは，髪型，性格などに特徴を加え，作品内で他のキャラクタと差別化を図るものであり，声から受ける人物の印象も個性を表現する．従来の音声合成技術による音声の多くは，情報を伝達するということのみを目的としているため，抑揚のない平坦なものであり，このような娯楽作品の需要に堪えるものではない．そのため，多様な個性を表現できる新たな音声合成技術が求められている．

近年，文章と音声の結びつきを，ニューラルネットワーク (NN) を用いた統計モデルで表現する音声合成が盛んに行なわれている [1]．NN による音声合成は，従来の隠れマルコフモデルを用いた音声より品質が改善されたことが示されている [2,3]．さらに多様な音声合成を実現させるために，複数人の音声を用いて様々な音声を生成する試みが行なわれている [4–10]．

NN を用いた音声合成では，NN の入力層に文章から得られる言語特徴量を入力するが，さらに，様々な種類の音声を生成するために，話者ごとに異なる情報を追加する手法がいくつか提案されている．話者コードと呼ばれる話者を one-hot ベクトルで識別したもの [4,5] や，i-vector と呼ばれる話者の類似度を混合ガウスモデルで表現したもの [5,6] を入力に用いることで，多様な音声の生成を行なっている．追加入力をするだけではなく，出力層を話者ごとに分岐させる手法が提案されている [7,8]．また，年齢，性別 [9] や感情 [10] を付与することで様々な表現を持つ音声が生成できることが示唆されている．しかし，話者から受ける印象について考慮された音声合成はあまり行なわれていない．

そこで本研究では NN を用いた統計モデルの入力層に，人物の印象を表す印象ベクトルを追加入力することで，指定した印象を持つような音声を生成する．人物の印象とは，「明るい」，「激しい」，「大人らしい」といった，音声

†1 現在，大阪府立大学 大学院工学研究科
Presently with Graduate School of Engineering, Osaka Prefecture University
a) mori@sig.cs.osakafu-u.ac.jp
b) inoue@cs.osakafu-u.ac.jp
c) yoshioka@cs.osakafu-u.ac.jp

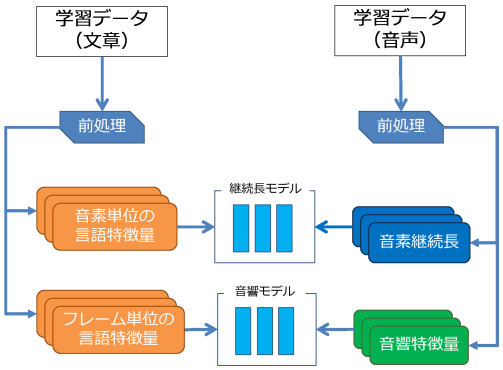


図 1　学習部の流れ

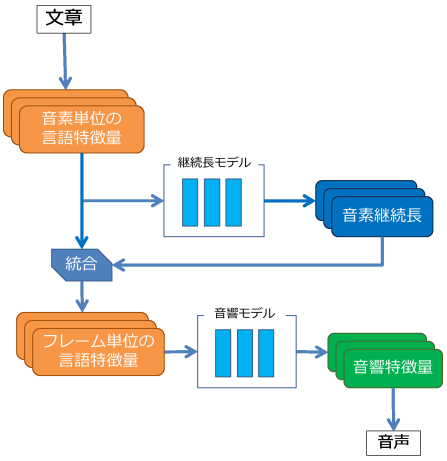


図 2　合成部の流れ

を聴いたときに感じられるもので，印象ベクトルはそれらを数値化したものである．話者コードや i-vector を用いた手法で，話者情報を NN のモデル構造に組み込むことができたのと同様に，印象情報をモデル構造に組み込むことができると考える．また，モデル構造に Long Short-Term Memory(LSTM) [11] を用いる．LSTM は音声のような時系列データの長期依存関係を学習できる．音声から受ける印象は，音声の断片的な情報ではなく，全体を通したひとまとまりで判断されるものと考え，印象ベクトルを用いた学習に，時系列情報が有効であると考える．

2. 従来手法

本節では，提案手法の基礎となる従来手法 [1,12]．を説明する．従来手法は前処理部と学習部および合成部で構成される．前処理部で音声と文章の特徴量抽出を行なった後，学習部で統計モデルの学習を行ない，合成部では学習したモデルを用いて音声を生成する．学習部における統計モデルは音素を推定するための継続長モデルと，音響特徴量を予測するための音響モデルの 2 つである．前処理部と学習部の概要を図 1，合成部の概要を図 2 に示す．以下でそれぞれの処理の詳細を示す．

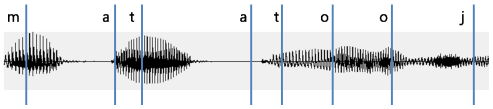


図 3　音素分割の例

2.1 前処理部

図 1 のように，学習部で統計モデルの学習を行なうために，学習データに前処理を行なう．まず，文章から言語特徴量を抽出する．また，文章に対応する音声から，各音素の長さを表す音素継続長と，音声波形を表す音響特徴量を抽出する．以下，各特徴量について説明する．

2.1.1 言語特徴量

言語特徴量とは，音声の最小単位である音素の種類，アクセントの位置や長さといった情報を含むベクトルである．言語特徴量の抽出において，まず，文章を形態素解析して音素を推定する．その後，一つの音素に対して，「直前の音素は/ a /であるか？」,「現在の単語に含まれる音素の数はいくつか？」といった質問に対する二値または連続値での回答を繋ぎ合わせ，その数値データを言語特徴量とする．また，2.2 節で述べる継続長モデルと音響モデルでは言語特徴量の形式が異なる．継続長モデルでは，各音素ごとに言語特徴量を抽出するが，音響モデルで用いる言語特徴量では，各音素が何フレームの長さを持つかの情報も加え，フレームごとに言語特徴量を抽出する．以下，継続長モデルに用いる言語特徴量を「音素単位の言語特徴量」，音響モデルに用いる言語特徴量を「フレーム単位の言語特徴量」とする．

2.1.2 音素継続長

音素継続長は一つの音素が何フレームの長さを持つかを表す．学習時には学習データから抽出した音素継続長を用いる．音声に対応する文章が既知であるため，任意の文章解析手法と音声認識手法を用いて，対応する音素の波形を検出することができる [12]．そして，音素ごとに波形を区切り，各音素のフレーム数を音素継続長とする．図 3 に音素の分割例を示す．合成時には音素単位の言語特徴量から予測した音素継続長を用いる．詳細は 2.3 節で述べる．

2.1.3 音響特徴量

音響特徴量は以下の 4 つを用いる [12]．学習時には学習データから抽出した音響特徴量を用い，合成時にはフレーム単位の言語特徴量から予測した音響特徴量を用いる．詳細は 2.3 節で述べる．

(1) メルケプストラム (Mel-frequency Cepstral coefficients:MC)
人間の声道の特性であるスペクトル包絡を少量のパラメータで表現したものがケプストラムである．MC は低周波数領域では感度が高く，高周波数領域では感度が乏しいという，人間の聴覚特性を考慮したケプストラムである．また，特徴量として MC の変化量である

動的特徴量（δ パラメータ）を加える．さらに，変化量の変化量である動的特徴量（δ' パラメータ）も加える．

(2) 非周期成分 (Band APeriodicity:BAP)
BAP は音声の非周期的な成分であり，声のかすれ等を表現する．Morise の手法 [13] を用いて，音声信号と信号の非周期成分との周波数ごとの強さの比を BAP とする．また，MC と同様に，δ パラメータと δ' パラメータを加える．

(3) 基本周波数 (Fundamental frequency:F_0)
F_0 は音声の周期性を表現し，声の高さを表す．森勢らの手法 [14] を用いて，フレームごとに音声波形をフーリエ変換し，最も低い周波数成分の周波数を F_0 とする．また，MC，BAP と同様に，δ パラメータと δ' パラメータを加える．

(4) 有声無声フラグ (Voiced/UnVoiced:VUV)
VUV は有声音か無声音かを表す 0,1 のビット列である．有声音とは声帯の振動を伴う音で，無声音とは声帯の振動を伴わない音である．具体的には，F_0 が 0 より大きければ VUV を 1（有声音），F_0 が 0 であれば VUV を 0（無声音）とする．

2.2 学習部

継続長モデルと音響モデルという 2 つの NN の学習を行う．継続長モデルは，図 1 に示すように音素単位の言語特徴量から音素継続長を予測するモデルである．学習データから抽出した音素継続長と予測した音素継続長の二乗誤差が最小になるような重みを学習する．

$$\hat{\boldsymbol{\lambda}} = \arg\min_{\boldsymbol{\lambda}} \sum_{n=1}^{N}\sum_{k=1}^{K} ||\mathbf{d}_{nk} - f_{\boldsymbol{\lambda}}(\mathbf{l}_{nk})||^2 \qquad (1)$$

$\mathbf{d}_{nk}$ と $\mathbf{l}_{nk}$ はそれぞれ，K 個の音素 (K は 1 つの発話文章から得られる音素の個数) のうち，k 番目の音素における音素継続長と音素単位の言語特徴量，$\boldsymbol{\lambda}$ は NN の重み，N は総データ数，$f_{\boldsymbol{\lambda}}(\cdot)$ は NN によって表される言語特徴量から音素継続長への非線形変換関数である．

音響モデルは，フレーム単位の言語特徴量から音響特徴量を予測するモデルである．学習データから抽出した音響特徴量と予測した音響特徴量の二乗誤差が最小になるような重みを学習する．

$$\hat{\boldsymbol{\Lambda}} = \arg\min_{\boldsymbol{\Lambda}} \sum_{n=1}^{N}\sum_{t=1}^{T} ||\mathbf{o}_{nt} - g_{\boldsymbol{\Lambda}}(\mathbf{l}_{nt})||^2 \qquad (2)$$

$\mathbf{o}_{nt}$ と $\mathbf{l}_{nt}$ はそれぞれ，T 個のフレーム (T は 1 つの発話音声から得られるフレーム数) のうち，t 番目のフレームにおける音響特徴量とフレーム単位の言語特徴量，$\boldsymbol{\Lambda}$ は NN の重み，$g_{\boldsymbol{\Lambda}}(\cdot)$ は NN によって表される言語特徴量から音響特徴量への非線形変換関数である．

2.3 合成部

図 2 に示すように，まず，入力された任意の文章から得られる音素単位の言語特徴量を継続長モデルに入力することでそれぞれの音素継続長が何フレームか予測する．その後，予測した音素継続長を用いて，音素単位の言語特徴量からフレーム単位の言語特徴量を生成する．それを音響モデルに入力することで音響特徴量を予測し，任意のボコーダを通して音声を生成する．

3. 提案手法

本節では，従来手法との差分について述べる．提案手法では，指定した印象を聴き手に与える音声を生成できるように，従来手法の学習部と合成部において話者の印象を表す印象ベクトルを統計モデルに追加して入力に加える．提案手法の学習の流れを図 4 に示す．話者の j 種類の印象を表す特徴量として主観評価で定めた j 次元の印象ベクトルを統計モデルに追加して入力する．これによって，入力した印象ごとに音素継続長や音響特徴量に差異が表れ，聴き手に印象の違いを感じさせることができる．

3.1 印象ベクトル

印象ベクトルとは音声を聴いたときの印象を主観評価したものである．本研究では印象ベクトルとして，「明るさ」「激しさ」「大人らしさ」の各評価値を持つベクトルを用いる．具体的には，「明るさ」の評価値は，1 から 5 の連続値を持ち，「明るい」という印象を持つ音声ほど 5，「暗い」ほど 1 に近づく．

学習に用いる印象ベクトルの計算はシェッフェの方法 [15] を用いて以下のように行う．まず，評価者が評価軸に則って全話者の音声を 1 対 1 で比較する．音声 A と音声 B を連続で聴き，音声 B は音声 A に比べてどの程度異なるかを評価する．例えば，「明るさ」の項目では「明るい:+2，やや明るい:+1，変わらない:0，やや暗い:-1，暗い:-2」の 5 段階である．評価値を回帰分析して算出した偏回帰係数を 1 から 5 の連続値で正規化し，印象ベクトルの要素とする．同様に「激しさ」と「大人らしさ」においても，「激しい:5—穏やか:1」「大人らしい:5—子供らしい:1」として 1 から 5 の連続値とする．

テスト時は，印象ベクトルの要素にそれぞれ任意の実数値を入力することによって，指定した印象を持つ音声を生成する．例えば，「明るい」音声を生成したいときは，「明るさ」の要素を 5 にする．

3.2 モデル構造

本研究では，NN のモデル構造として，全結合 (Full Connect:FC) 層と LSTM(Long short-term memory) 層 [11] を用いる．FC 層のみで構成されたモデル構造では，音声の連続した関係性を学習できない．しかし，LSTM 層は要素

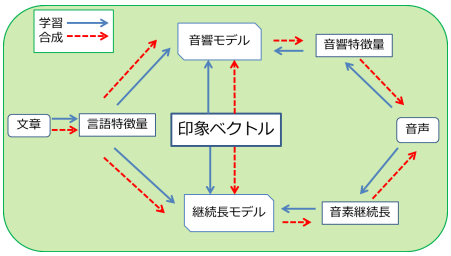


図 4　提案手法の学習及び音声合成処理の流れ

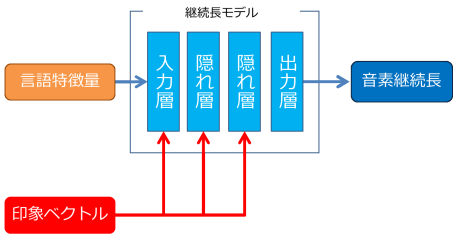


図 5　印象ベクトルの入力方法の概略図

が順番に並んでいる時系列データを扱えるモデル構造であり，音素の前後の依存関係を学習することが期待できる．本研究では，この依存関係を学習するかどうかで，生成した音声がどのような影響を受けるのか検証するために，FC 層のみのモデルと，FC 層と LSTM 層を組み合わせたモデルを用いる．

また，図 5 のように，隠れ層において話者の特徴を失うことなく学習を行うため [16]，入力層だけではなく隠れ層にも補助的な入力として印象ベクトルを加える．図 5 では継続長モデルにおける補助入力を表しているが，音響モデルにおいても同様の処理を行う．

4. 実験

4.1 実験条件

本実験では，データセットに声優統計コーパス [17] と JSUT(Japanese speech corpus of Saruwatari Lab, University of Tokyo) [18] を用いた．これらのデータセットは音声データと文章データのセットである．声優統計コーパスは女性 3 人が感情表現として「通常」,「喜び」,「怒り」の 3 パターンで日本語の 100 文章を読んだ音声データである．本実験では話者の人数を増やすため，同じ話者でも，異なる感情の音声は異なる話者として扱い，合計 9 人分の音声として扱った．JSUT は女性 1 人が日本語のおよそ 7500 文章を読んだ音声データである．2 つのデータセットから合わせて女性話者 10 人が共通の 100 文章を読んだ音声，合計 1000 個を実験に用いた．学習に 900 個，検証に 50 個，テストに 50 個を用いた．また，言語特徴量抽出に OpenJTalk [19]，音素継続長推定に Julius [20]，音響特徴量抽出に WORLD [21] と SPTK [22]，ボコーダに WORLD [21] を用いた．

前述のとおり，継続長モデルと音響モデルには NN を用いた．[5, 12] をベースに，隠れ層の構成を図 6 の 4 通りを用意し，各モデル構造ごとに実験を行なった．継続長モデルと音響モデルは同じ構造を用いた．継続長モデルの入力は言語特徴量に印象ベクトルを加えた計 579 次元，出力は音素継続長 1 次元とした．また，音響モデルの入力は言語特徴量に印象ベクトルを加えた計 580 次元，出力は音響特徴量 199 次元（MC:60*3，BAP : 5*3，F_0 : 1*3，VUV : 1）とした．各モデルの詳細は以下の通りである．入力と出力は全て共通である．

(1) FeedForward-model(FF)
　各層のユニット数 512 の全結合層 4 つ
(2) LSTM-model(LSTM)
　全結合層 3 つ，LSTM 層 3 つ．各ユニット数は 50，200，400，300，200，100.
(3) auxiliary-FeedForward-model(auxFF)
　各層のユニット数 512 の全結合層 4 つに補助的な入力を加えたもの
(4) auxiliary-LSTM-model(auxLSTM)
　全結合層 3 つ，LSTM 層 3 つに補助的な入力を加えたもの．各ユニット数は 50，200，400，300，200，100.

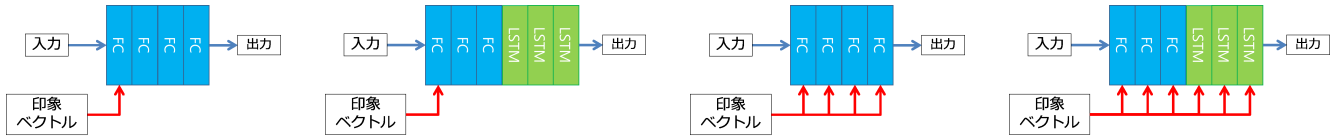
表 1　各話者の印象ベクトル

話者_感情	明るさ	激しさ	大人らしさ
fujitou_angry	1.99	4.43	4.41
fujitou_happy	5.00	4.52	3.32
fujitou_normal	2.29	1.32	5.00
tsuchiya_angry	1.00	1.00	3.94
tsuchiya_happy	4.95	3.83	1.00
tsuchiya_normal	2.26	1.23	3.79
uemura_angry	2.92	5.00	3.29
uemura_happy	4.91	3.63	1.25
uemura_normal	2.70	1.31	4.10
JSUT	2.53	1.48	4.07

4.2 客観評価

まず 11 人の評価者に学習に用いるデータセットの評価をしてもらい，印象ベクトルの算出をした．印象ベクトルの一覧が表 1 である．

次に，印象ベクトルを追加入力する上記のモデル構造に対して，客観評価を用いて比較した．学習に用いた印象ベクトルをそのまま追加入力とし，生成した音声とテスト音声，それぞれの音響特徴量の誤差を求め，その誤差を客観評価指標として用いた．誤差が小さいほどテスト音声に近い音声が生成でき，上手く学習できていると言える．継続長モデル及び音響モデルの入力として，テスト文章とテスト音声から得られるフレーム単位の言語特徴量を用いた．評価指標として，各音響特徴量の誤差を次のように定めた．メルケプストラムの誤差として MCD（MC Distortion）を式 (3) のように定義した．



[1]FF [2]LSTM [3]auxFF [4]auxLSTM

図 6 実験に用いたモデル構造

表 2 音響モデルの客観評価

モデル構造	MCD(dB)	BAPD(dB)	F_0-RMSE(Hz)	VUV-MAPE(%)
FF	7.984±0.081	0.7068±0.0195	54.43±0.73	20.02±0.18
auxFF	7.981±0.123	0.6956±0.0091	53.95±0.76	20.14±0.55
LSTM	7.579±0.025	0.6658±0.0035	52.25±1.71	18.88±0.25
auxLSTM	7.480±0.079	0.6626±0.0139	49.63±1.06	18.37±0.53

$$\mathrm{MCD[dB]} = \frac{10}{\ln 10}\sqrt{2\sum_{i=1}^{60}\left(m_i^{(y)} - m_i^{(\hat{y})}\right)^2} \qquad (3)$$

$m_i^{(y)}$ はテスト音声の 1 フレームあたりのメルケプストラム係数，$m_i^{(\hat{y})}$ は生成した音声の 1 フレームあたりのメルケプストラム係数である．また，非周期成分の誤差として BAPD（BAP Distortion）を式 (4) のように定義した．

$$\mathrm{BAPD[dB]} = \frac{1}{10}\sqrt{\sum_{i=1}^{5}\left(b_i^{(y)} - b_i^{(\hat{y})}\right)^2} \qquad (4)$$

$b_i^{(y)}$ はテスト音声の非周期成分，$b_i^{(\hat{y})}$ は生成した音声の非周期成分である．また，F_0 の誤差として平均二乗誤差平方根 (Root Mean Square Error:RMSE)，VUV の誤差として平均絶対誤差率 (Mean Absolute Percentage Error:MAPE) を用いた．

生成した音声とテスト音声で求めた音響特徴量の誤差を表 2 に示す．統計モデルの学習，音声の生成および誤差の測定は 5 回ずつ行い，その平均を取った．表 2 より，FC 層のみのモデルより，LSTM を用いたモデルの方がテスト音声に近い音声が生成されたことが分かる．これは時系列情報を加えることで，音素間の繋がりといった，より多くの特徴量を捉えられたからだと考えられる．また，auxLSTM で生成した音声がテスト音声に最も近かった．隠れ層における話者の特徴を失わないために，補助入力が有効であると言える．

4.3 主観評価

生成した音声が入力した印象ベクトル通りになっているか主観評価で検証を行なった．この検証では学習に用いていない文章 16 文を，客観評価で最も良い評価を得た auxLSTM に入力して音声を生成した．印象ベクトルは「明るい:5，暗い:1」「激しい:5，穏やか:1」「大人らしい:5，子供らしい:1」それぞれ 5 または 1 にした組み合わせ 8 通りである．評価者 17 人が，各音声について各印象ベクトルの要素が 5 または 1 のどちらであるかを回答した．その正答率を表 3 に示す．

表 3 より，「明るさ」は正答率が高かった．「明るい」を入力すると声の高い音声が生成されやすく，「暗い」を入力すると声の低い音声が生成されやすいため，声の高低で「明るさ」を評価したのだと考えられる．同様に「激しさ」も正答率が高かった．「激しい」を入力すると早口な音声が生成されやすく，「穏やか」を入力するとゆっくりした音声が生成されやすいため，声の速度で「激しさ」を評価したのだと考えられる．しかし，「大人らしさ」は音声によって正答率に差が出た．「暗い」「大人らしい」音声を正しく「大人らしい」と回答した人が多い一方，「暗い」「子供らしい」を「大人らしい」と多く誤回答していた．また，「激しい」音声より「穏やか」な音声の方が「大人らしい」と判断されていることが分かる．したがって「大人らしい」音声は「暗い」「穏やか」という印象に相関があると考えられる．さらに，「明るい」「大人らしい」を「子供らしい」と多く誤回答し，「激しい」方が「子供らしい」と正答率が高いことから，「子供らしい」音声は「明るい」「激しい」に相関があると考えられる．

5. おわりに

本研究では NN を用いた音声合成時に，人物の印象を表す印象ベクトルを入力することで，人物の印象を加味できるかどうかを検証した．音声を生成するための統計モデルの学習において，時系列データを扱えるモデル構造がテスト音声により近い音声を生成できたことから，時系列情報は音声生成において重要な要因であると考えられる．印象の聞き分けを行う主観評価において，「大人らしさ」のような複数の要素で成り立つような印象は音声に反映できなかったが，「明るさ」は声の高さ，「激しさ」は喋る速さ，といった一つの要素である程度判別可能な印象は音声に反映できることが分かった．

今後の課題としてより多くの印象を表現できる特徴量抽出やモデル化を行うことが挙げられる．

表 3 生成した音声の主観評価

	印象ベクトル			正答率 (%)		
文章番号	明るさ	激しさ	大人らしさ	明るさ	激しさ	大人らしさ
1	暗い	穏やか	子供らしい	100	100	5.9
2	暗い	穏やか	子供らしい	94.1	100	0.0
3	暗い	穏やか	大人らしい	100	100	100
4	暗い	穏やか	大人らしい	100	100	100
5	暗い	激しい	子供らしい	88.2	35.3	29.4
6	暗い	激しい	子供らしい	82.4	58.8	5.9
7	暗い	激しい	大人らしい	76.5	76.5	94.1
8	暗い	激しい	大人らしい	94.1	64.7	100
9	明るい	穏やか	子供らしい	52.9	88.2	82.4
10	明るい	穏やか	子供らしい	82.4	47.1	94.1
11	明るい	穏やか	大人らしい	47.1	82.4	35.3
12	明るい	穏やか	大人らしい	47.1	88.2	52.9
13	明るい	激しい	子供らしい	88.2	76.5	100
14	明るい	激しい	子供らしい	94.1	76.5	100
15	明るい	激しい	大人らしい	82.4	100	58.8
16	明るい	激しい	大人らしい	58.8	94.1	41.2
平均				80.5	80.5	62.5

参考文献

[1] 橋本佳, 高木信二. 深層学習に基づく統計的音声合成. 日本音響学会誌, 73(1):55–62, 2017.

[2] H. Ze, A. Senior, and M. Schuster. Statistical parametric speech synthesis using deep neural networks. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pp. 7962–7966. IEEE, 2013.

[3] Z. Wu, C. Valentini-Botinhao, O. Watts, and S. King. Deep neural networks employing multi-task learning and stacked bottleneck features for speech synthesis. In *Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4460–4464. IEEE, 2015.

[4] N. Hojo, Y. Ijima, and H. Mizuno. An investigation of dnn-based speech synthesis using speaker codes. In *INTERSPEECH*, pp. 2278–2282, 2016.

[5] Y. Zhao, D. Saito, and N. Minematsu. Speaker representations for speaker adaptation in multiple speakers' blstm-rnn-based speech synthesis. In *INTERSPEECH*, pp. 2268–2272, 2016.

[6] Z. Wu, P. Swietojanski, C. Veaux, S. Renals, and S. King. A study of speaker adaptation for dnn-based speech synthesis. In *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[7] S. Pascual and A. Bonafonte Cávez. Multi-output rnn-lstm for multiple speaker speech synthesis with a-interpolation model. In *SSW9: September 13-15*, pp. 112–117. Institute of Electrical and Electronics Engineers (IEEE), 2016.

[8] B. Li and H. Zen. Multi-language multi-speaker acoustic modeling for lstm-rnn based statistical parametric speech synthesis. In *INTERSPEECH*, pp. 2468–2472, 2016.

[9] H. Luong, S. Takaki, G. E. Henter, and J. Yamagishi. Adapting and controlling dnn-based speech synthesis using input codes. In *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, pp. 4905–4909. IEEE, 2017.

[10] 井上勝喜, 原直, 阿部匡伸, 北条伸克, 井島勇祐. Dnn 音声合成における感情付与のためのモデル構造の検討. 電子情報通信学会技術研究報告 IEICE technical report: 信学技報, 117(106):23–28, 2017.

[11] Y. Fan, Y. Qian, F. Xie, and F. K. Soong. Tts synthesis with bidirectional lstm based recurrent neural networks. In *Interspeech*, 2014.

[12] Z. Wu, O. Watts, and S. King. Merlin: An open source neural network speech synthesis system. *Proc. SSW, Sunnyvale, USA*, 2016.

[13] M. Morise. D4c, a band-aperiodicity estimator for high-quality speech synthesis. *Speech Communication*, 84:57–65, 2016.

[14] 森勢将雅, 河原英紀, 西浦敬信. 基本波検出に基づく高 snr の音声を対象とした高速な f0 推定法. 電子情報通信学会論文誌 D, 93(2):109–117, 2010.

[15] H. Scheffé. *The Analysis of Variance*. LWW, 1960.

[16] S. O. Arık, G. Diamos, A. Gibiansky, J. Miller, K. Peng, W. Ping, J. Raiman, and Y. Zhou. Deep voice 2: Multi-speaker neural text-to-speech. *arXiv preprint arXiv:1705.08947*, 2017.

[17] y_benjo and MagnesiumRibbon. Voice-actress corpus. http://voice-statistics.github.io/, 2017.

[18] R. Sonobe, S. Takamichi, and H. Saruwatari. Jsut corpus: free large-scale japanese speech corpus for end-to-end speech synthesis. *arXiv preprint arXiv:1711.00354*, 2017.

[19] H. W. Group, et al. Open jtalk. http://open-jtalk.sourceforge.net/, 2016.

[20] A. Lee and T. Kawahara. Recent development of open-source speech recognition engine julius. In *Proceedings: APSIPA ASC 2009: Asia-Pacific Signal and Information Processing Association, 2009 Annual Summit and Conference*, pp. 131–137, 2009.

[21] M. Morise, F. Yokomori, and K. Ozawa. World: a vocoder-based high-quality speech synthesis system for real-time applications. *IEICE TRANSACTIONS on Information and Systems*, 99(7):1877–1884, 2016.

[22] S. W. Group, et al. Speech signal processing toolkit (sptk). *http://sp-tk.sourceforge.net*, 2009.