修士論文

絵本読み聞かせ調音声の合成を目的とした
深層学習に基づく韻律制御に関する検討

2018 年 2 月 1 日

指導教員　峯松 信明 教授

電気系工学専攻
37-166875　尤　秀

内容梗概

近年、スマートフォンの音声ナビゲーションシステムやコミュニケーションロボットなど、テキストを読み上げる音声合成技術は活躍する場が広がっている。そうした中で、テキストを単調に読み上げるだけではなく、多様な読み方・話し方を実現する音声合成に向けた研究がなされている。このような研究の代表的なフィールドとして、喜びや怒りなどの感情を込められた感情音声合成があげられる。しかし、我々が実際に発話するときには、感情のみならず、相手との関係によって、発話態度や話し方を変えることがある。特に、日本では、目上の人には敬語で、ていねいに話す等、相手によって話し方がかなり異なる。このような相手との関係性を反映できる音声は、social な音声とも呼ばれる。

social な音声の中で、敬語の他、対乳児音声（Infant-directed speech; IDS）と呼ばれる例もある。これは、特に幼稚園や保育園の幼児に対して、母親が発話するときに見られる現象である。その特徴については、多くの研究が行われている。IDS 風の音声合成ができれば、より親しみ話し方で子供に語りかけるコミュニケーションロボットの運用が期待される。そこで、本研究では、IDS 音声の一例として、幼児に向かって絵本を読み聞かせるような音声を取り上げ、このような音声を合成するにあたって、以下のことを検討した。

1）読みあげ調と読み聞かせ調の両スタイル音声を収録されるパラレルコーパスをを用いて、両スタイル音声に対し、韻律特徴量を抽出し、読み聞かせ調音声が持つ独特な韻律特徴を統計的な手法により分析した。

2）分析より確認された韻律特徴を合成音声に再現させるため、それらの特徴を韻律ラベルとして、近年主流である DNN 音声合成に用いられて、音響モデルにより韻律を制御する形で読み聞かせ調音声の合成を試みた。

3）音響モデルに基づく韻律制御手法における問題点を踏まえ、音響モデルより直接に読み聞かせ調音声を合成する代わりに、まず読みあげ調音声を合成し、次に合成された読みあげ調音声の韻律を読み聞かせ調音声の韻律に変換することで、後者の音声合成を実現する手法、すなわち、韻律変換を用いた読み聞かせ調音声合成を提案した。また、韻律ラベルを変えることで、制御可能な発話スタイルへの変換可能性を実験した。

評価実験の結果、韻律変換を用いた音声合成では、より効果的に韻律を制御する可能性を示した。特に表現力の高く、変化が多様な音声に対して有効であることを確認された。

目次

図目次

表目次

第1章

序論

1.1　はじめに

近年、会話システムやコミュニケーションロボットなど、音声合成や音声対話機能が様々なアプリケーションに組み込まれるようになってきている。しかし、現行の音声合成応用例の多くは発話意図を表す言語情報を伝達することを主にし、テキストを単調に読み上げるものが多い。これは人間の発声と比べて、平坦な音調であったり、平均にポーズを取ったりするため、自然性に欠けてしまう。より人間らしく、expressive な音声合成技術が求められている。

expressive な音声を合成するためには韻律制御が重要である。これに対し、音声の言語情報の他、話者の個人性（つまり、話者性）、感情・気分，更には，相手との関係性などを反映できる韻律情報を対象とする研究が行われている。[1, 8, 14, 12] これらの研究から得られた知見を踏まえ、expressive な音声合成に関する研究が活躍の場に広がれている。

こうした研究では、感情音声を対象とした研究が盛んに行われているが、まだ研究の余地がある音声スタイルもある。例えば、Blizzard Challenge でも採択されたように，子供向け audio books を訓練データとした TTS システム構築の研究例があるが [15]、これは、話し相手を指定された話し方、すなわち、相手との関係性を反映できる発話スタイル。このよう発話スタイルの一例として、より年少の子供，幼稚園や保育園の園児に向かって主に母親が発話するときに見られる対乳児音声（Infant-directed speech; IDS）が存在する [9]。

IDS 風の音声合成が出来れば、より親しみ話し方で子供に語り掛けるコミュニケーションロボットの運用が期待できる。そこで、本研究は IDS 風音声を取り上げ、特に絵本を読み聞かせるような音声を対象とし、このような音声を合成することを目指す。

読み聞かせ調音声の合成を実現するには、読み聞かせ調音声における読み方の工夫を表現する韻律に関する制御が課題となる。そのため、本研究では、読み聞かせ調音声を合成する際の韻律制御に着目し、韻律素性から目標発話スタイルの韻律生成を、深層学習により制御する手法を検討する。

1.2　研究の目的

聞き手との関係性を反映できる IDS 風の音声スタイルを取り上げ、絵本を読み聞かせるような音声の合成を目標とする。

これを実現するために、発話スタイルを指定できる韻律素性を用いて、目標発話スタイルの韻律制御を深層学習による実現の可能性を検討する。

はじめに、韻律素性と物理データの韻律特徴量との相関を統計的手法を用いて分析し、素性より韻律の制御の可能性を検証する。次に、分析の結果を踏まえて、有効と見られる素性を韻律ラベルに変換し、深層学習の入力として用いられ、実際に読み聞かせ調音声を合成する実験を行い、韻律制御の性能を評価する。最終的に、韻律ラベルで示された読み聞かせ調音声の韻律特徴が適切に再現できる音声を合成するシステムの開発を目指す。

1.3　本論文の構成

本論文は全 6 章から構成される。まず第 1 章では、本論文の背景と目的について述べる。第 2 章では、本研究で用いられる音声コーパスの詳細と分析について述べる。第 3 章では、本研究で用いられる統計的音声合成手法の概要と流れについて述べる。また、BLSTM 音響モデルに基づ

く読み聞かせ調音声の韻律制御の性能を、実際に音声を合成し、検討する。第 4 章では、音響モデルにより韻律制御の問題点を踏まえて、韻律変換を用いた読み聞かせ調音声の合成システムを提案し、システムにおける主な処理について述べる。第 5 章では、提案システムの各モジュールの性能を評価する。また、実際に提案システムより合成された読み聞かせ調音声と、第 3 章の実験より合成された音声を、主観評価実験及び客観評価実験により比較し、評価する。最後に第 6 章で本論文をまとめ、今後の課題について述べる。

第2章

音声コーパスとその分析

2.1　はじめに

本研究は、聞き手との関係性を反映できる IDS(Infant directed speech) 風音声を取り上げ、具合的な例である絵本を読み聞かせ調音声を対象とする。このような音声を合成するには、読み聞かせ調音声を収録されるコーパスと、読み聞かせ調音声における読み方の工夫を表現できる特徴を定義する必要となる。

[2] では、読み聞かせ調音声の合成を実現するために、コーパスの構築、および読み聞かせ用の特徴ラベルの設計等の検討を行われた。本研究は、その研究成果を基にして、読み聞かせ調音声合成の実現を目指すこととする。

本章では、研究に用いられる音声コーパスとその特徴について述べる。第 2.2 節ではコーパスの仕様を説明する。第 2.3 節ではコーパスに収録される読み聞かせ調音声において、確認された韻律特徴について紹介する。第 2.4 節では、読み聞かせ調音声の韻律特徴量を統計的に分析する。

2.2　コーパスの仕様

本研究で用いられる音声コーパスは、[2] より構築されたパラレルコーパスである。このコーパスは、女性保育士 1 名[1]に、七冊の絵本（917 文）を
1) アナウンサーのような読み上げ調
2) 園児に語るような読み聞かせ調
の 2 通りの読み方で読ませたパラレルコーパスである。すなわち、一文に 2 スタイルの音声（読みあげ調・読み聞かせ調）が収録されている。

コーパスの文章としては、音素バランス性を考慮し、キャラクターの個性が比較的明確となっているシリーズものの絵本を中心に選定されたものである。収録された絵本は以下の 7 冊である。

- 『お月さんはきつねがすき？』, 神沢利子：作
- 『ババールおうさま』, ジャン・ド・ブリュノフ：原作, せなあいこ：訳
- 『まめうしとまめばあ』, あきやまただし：作
- 『ぼくひこうきにのったんだ』, わたなべしげお：作
- 『ひとまねこざるびょういんへいく』, マーガレット・レイ：作, 光吉夏弥：訳
- 『あいうえおん』, あきびんご：作
- 『ききみみずきん』, 木下順二：文, 前半部のみ

コーパスの詳細を表 2.1 に示す。

表 2.1: コーパスの仕様

話者	女性保育士 1 名
文数	917 文（絵本 7 冊分）
収録スタイル	読み上げ調スタイル
	読み聞かせ調スタイル
時間	各スタイルにつき約 1 時間

[1] 10 名の女性保育士から、お金を払いても雇いたいという条件で、オーディションによって選ばれた保育士である。

2.3　読み聞かせ調音声における韻律特徴

本研究は、IDS 風音声（読み聞かせ調）の合成を目指す。IDS は幼児の言語獲得に関係あると考えられているため、その特徴については様々な分析研究が行われている。こうした研究では、IDS に見られる特徴には、言語や話者によらず共通的な特徴（例えば、ピッチが全体的に高い、発話が短い、ポーズが長いなど）もあれば、言語や話者に依存して見られるものもあると様々な知見が得られている [9]。

[2] では、収録された音声コーパスを読み聞かせ調音声合成に用いられるため、コーパスベースの読み聞かせ調音声の特徴を見出すことを狙い、ラベリングを行われた。音声学を専攻する学生 3 名をラベラーとして、コーパスにある両スタイル音声を聞き比べることにより、以下の韻律特徴が確認され、ラベルを付与された。

1）長音化

長母音を極端に延ばして園児の注意をひくことが観測されている。この長音化は非長母音に対しても行われることがある。

2）抑揚の変化

アクセント句の韻律制御において、（読み上げ調と比較して）全体的により高い（揚）／低い（抑）音調で読むことがある。揚の場合も、抑の場合も、その度合いを強／中／弱でラベル化する。

3）上昇調 BPM[9]

読み聞かせ調音声で観測される現象であるが、アクセント句末で局所的にピッチが上がることがある (Boundary Pitch Movement)。その例を図 2.1 に示す。

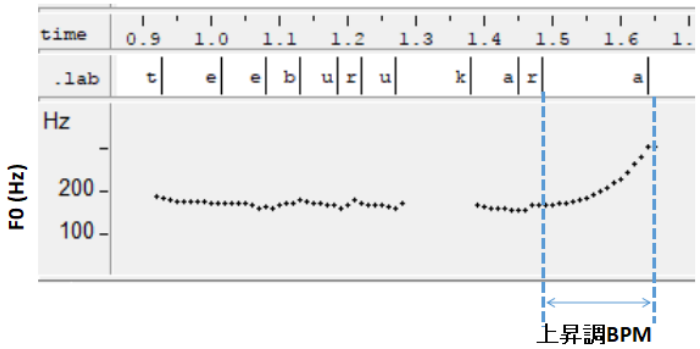


図 2.1: 上昇調 BPM の F0 曲線

4）キャラクター属性

登場人物により、大人（男・女）、子供（男の子・女の子）のラベルが付与されている。

特徴とラベルの対応付は表 2.2 に示す。

表 2.2: 韻律特徴とラベル標記の対応付け

韻律特徴	ラベル
上昇調 BPM	?
抑（弱・中・強）	[] ・ [[]] ・ [[[]]]
揚（弱・中・強）	{} ・ {{}} ・ {{{}}}
長音化	@

例として、
「まどぎわのテーブルから、ひろいひこうじょうが、とてもよくみえます。」
ラベル文：「まどぎわのテーブルから ?、[ひろ @ いひこうじょうが]、{ と @ て @ もよ く } みえます」
図 2.2 は該当文の読みあげ調音声・読み聞かせ調音声の F0 パターンである。

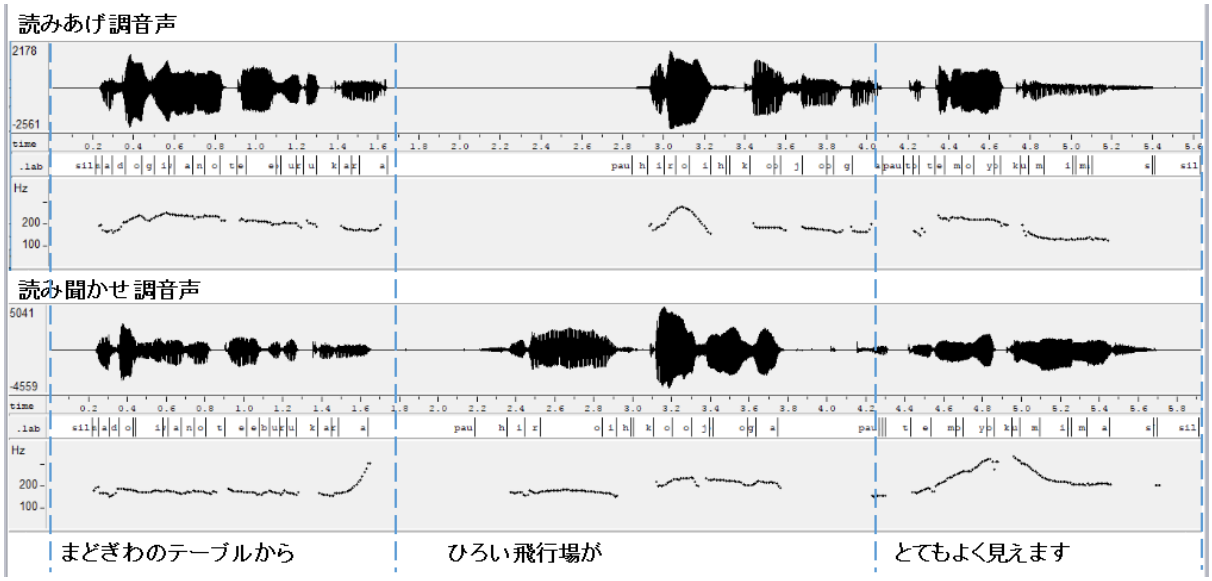


図 2.2: 「まどぎわのテーブルから、ひろいひこうじょうが、とてもよくみえます。」を発話するとき、読みあげ調音声（上）・読み聞かせ調音声（下）の F0 パターンの例

2.4　コーパス音声の韻律分析

第 2.3 では、コーパスベースの読み聞かせ調音声に特有な韻律特徴を述べた。それらを大きく分けて、F0、話速に分類することができる。例えば、「抑揚」、「上昇調 BPM」、「キャラクター属性」が特に F0 に関係あり、「長音化」は主に話速と関連する。もちろん、各韻律素性の間も相関があるが（例えば、「上昇調 BPM」に属するモーラがよく延ばして読まれて、ある程度「長音化」と関連する）、今回は各韻律特徴の概ねのトレンドを確認する狙いとし、それぞれ独立と考えて、分析を行った。

2.4.1　F0 分析

F0 分析は、読み聞かせ調音声の韻律特徴である（抑、揚）×（強、中、弱）、キャラクターラベル（大人、子供）×（男、女）を対象とし、それぞれの F0 を定量的に分析し、読みあげ調音声との差異や各特徴間の関係を確認することを目的とする。

コーパスにある 917 文の音声に対し、各韻律特徴において、アクセント句毎に F0 を抽出した。F0 の抽出は STRAIGHT [2] という音声分析ツールを用いた。結果を表 2.3 に示す。ここで、句数はアクセント句の数を表し、Δ は標準偏差を表す。他の項目はそれぞれアクセント句内の F0 の平均値 $F0_{mean}$、最大値 $F0_{top}$、最小値 $F0_{bottom}$ の平均を表す。

[2] $http://www.wakayama-u.ac.jp/\ kawahara/STRAIGHTadv/$

表 2.3: 各種韻律ラベルにおける F0 分布

ラベル		句数	$F0_{mean}$	$\Delta F0_{mean}$	$F0_{top}$	$\Delta F0_{top}$	$F0_{bottom}$	$\Delta F0_{bottom}$
読みあげ		1502	192.5	12.7	285.9	39.2	139.6	22.6
抑	弱	261	188.3	23.1	260.9	60.7	136.3	26.2
	中	36	178.4	24.7	238.1	54.8	139.6	19.5
	強	13	182.3	22.3	275.4	42.5	143.4	9.4
揚	弱	348	258.1	42.3	367.8	51.2	154.3	35.1
	中	34	249.5	36.7	348.9	60.9	150.6	34.6
	強	44	264.8	36.1	373.7	42.7	163.3	35.0
男		260	215.5	32.8	289.9	60.2	130.5	30.6
女		167	242.5	41.7	394.1	42.2	136.6	41.4
男の子		236	254.2	53.3	399.9	40.8	143.6	41.6
女の子		103	267.9	45.3	414.3	45.2	145.3	57.7

読み上げ調音声と読み聞かせ調音声を比較すると、後者が全体的に高い音調で発話する傾向が観察されている。また、読み聞かせ調音声において、全体的にピッチ変化の幅が大きい（F0 の最小値 $F0_{bottom}$ と最大値 $F0_{top}$ の差が大きい）ことが分かる。それに、どの F0 項目においても、読み聞かせ調音声の標準偏差が極めて大きいことが分かる。これは、読み聞かせ調音声は、抑揚の制御や、キャラクター属性を表現するため、イントネーションを幅広く変化しながら発話するのであると考えられる。

次に読み聞かせ調音声の各種特徴ラベルの F0 分布に着眼する。ラベルの定義から予想された通りに、どの項目においても、「揚」＞「抑」、「女の子」＞「男の子」＞「女」＞「男」といった順序が観測されている。抑・揚の強・中・弱ラベルであるが、平均値的にはラベルの意図通りの分布にはなっていないことも分かる。これは、抑・揚のラベルを付与されたとき、ラベル同士の比較ではなく、読みあげ調音声とを比較し、ラベラーの感覚により弱・中・強を判断するため、統計的な物理データと誤差が生じると思われる。

偏差が少しあるが、この分析では、（抑、揚）×（強、中、弱）、キャラクターラベル（大人、子供）×（男、女）の韻律ラベルが、F0 分布を区分化する可能性を示していると考えられる。つまり、これらの韻律特徴ラベルを使って、F0 を制御するのが期待できる。

2.4.2　話速分析

読み聞かせ調音声において、読みあげ調音声と比べて、ゆっくり発話する現象が観測された。また、幼児の注意をひくために、個別のモーラにて母音を極端に伸ばして発話すること（長音化）も確認された。話速を定量的に分析するために、両スタイル音声において、音素毎に話速（音素継続長）を抽出した。各音素における話速の平均値、標準偏差を図 2.3 に示す。青線は読みあげ調音声 (ad)、赤線は読み聞かせ調音声 (id) を表す。

読み上げ調音声と読み聞かせ調音声を比較すると、後者において、話速の標準偏差が大きいことが分かる。特に母音においては、読み聞かせ調音声が全体的に伸ばして読む場合が多い。これに対し、子音においては、両スタイル音声の差のバリアンスが極めて大きい。例えば、「b」、「d」、「r」などの子音において、両スタイル音声の平均値の差も小さいし、標準偏差も小さいが、「gy」、「hy」のような拗音の子音の場合、両音声の差が極めて大きい上に、標準偏差も大きい。この原因

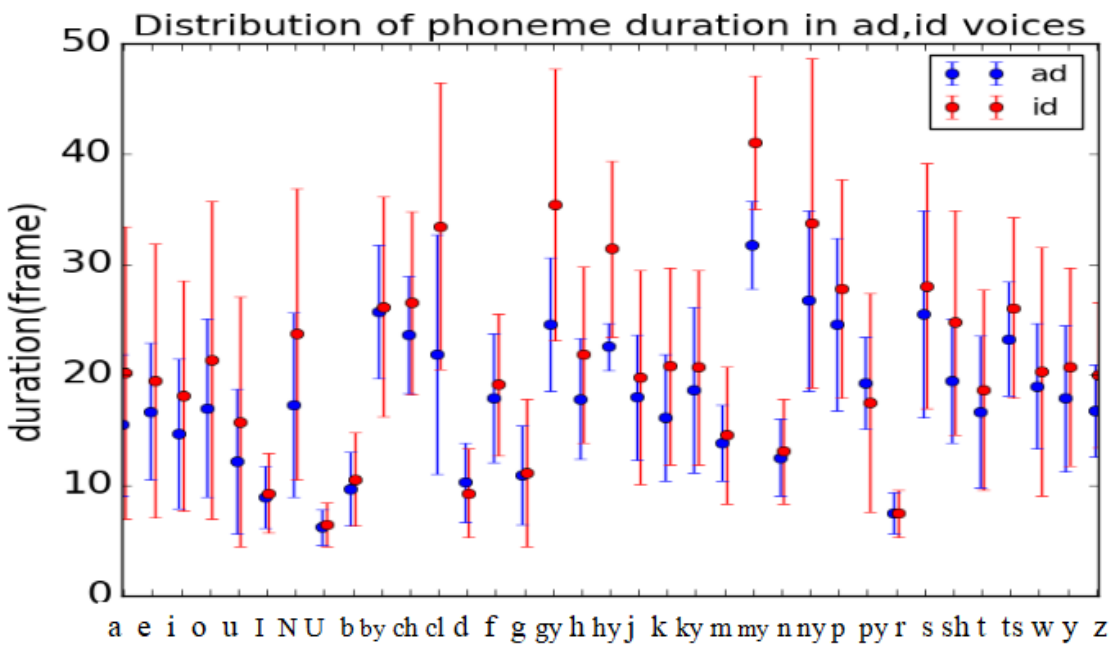


図 2.3: 音素継続長分布

を調べたところ、拗音の場合、例えば、「きゅうり」や「キューピー」など、極端に延ばして発話している例が存在することが確認されている。

2.5　むすび

本章では、絵本を読み聞かせ調音声合成を実現するために必要となる読み聞かせ調音声コーパスと、その特徴について述べた。次章では、音声合成技術について紹介し、本章で説明した韻律特徴を入力とし、実際に読み聞かせ調音声合成の実験を検討する。

第3章

BLSTMに基づく
読み聞かせ調音声の合成

3.1　はじめに

本研究は、絵本の文章を入力とし、読み聞かせ調で読ませる音声合成を目標とする。こうしたテキストからそれに対応する音声を生成することを、テキスト音声合成（Text-To-Speech;TTS）とも呼ばれる。音声合成技術は、コンピューター技術の発展と共に、ルールベースのアプローチからデータベースに基づく手法へと発展してきた。データベースに基づく音声合成システムには、主に波形接続型音声合成と統計的音声合成と二つに大別される。

中には、波形接続型音声合成 [3] は、名の通り、大量の音声波形素片から適切なものを選んで接続する手法である。音声のデータ量が十分であれば、高品質かつ多様な声質の音声合成が可能である。一方で、うまく接続できる音声単位が音声波形素片にない場合、合成された音声の中に不連続な単位が挿まれていたため、品質が低下する。

これに対し、統計的音声合成は、音声波形ではなく、音声から抽出する音響パラメータを、統計モデルを用いて学習し、入力のテキストと出力の音声の音響パラメータの間のマッピングを統計モデルにより捉える手法である。合成するときは、入力テキストに対応する音声の音響パラメータを統計モデルから予測し、この音響パラメータを用いて、音声を生成される。統計的音声合成手法は、統計モデルを用いることで、少量のデータにも対応でき、また、音響パラメータから音声波形を生成するので、声質を変化させたりすることも可能であるなどの利点があるため、この数十年注目を集めってきた。

本研究で検討するような読み聞かせ調音声合成は、多様な変化が含まれるため、波形接続型のほうがこれらの変化を忠実に再現し、自然な音声を生成できるはずだが、それなりのデータ量を用意するのは、容易ではない。似たような作業で、2005 年から毎年開催されている Blizzard Challenge[15]（子供向け audio books を訓練データとし、TTS システムを構築する研究）では、統計的音声合成は、条件によって、波形接続型と肩を並べる音声品質を達成可能であることを示した。

そこで、コーパスのデータ量を考慮し、本研究は統計的音声合成手法を用いて、絵本を読み聞かせ調音声合成を検討することにする。第 3.2 節では、統計的音声合成の概要を簡潔に説明してから、統計的音声合成に用いられる特徴量と、核心である音響モデルについて紹介する。第 3.3 節では、BLSTM-RNN（Bidirectional LSTM Recurrent Neural Network）を用いた音声合成の詳細を述べる。第 3.4 節では、実際にコーパス音声を用いて、BLSTM-RNN に基づく読み聞かせ調音声合成の実験を述べる。

3.2　統計的音声合成

3.2.1　統計的音声合成の概要

統計的音声合成は、入力テキストから統計モデルにより、音声波形を生成するための音響パラメータを予測する手法であり、概ねに、テキスト解析部と波形生成部に分けて理解できる。テキスト解析部では、テキストから音素、単語、品詞、アクセント型など言語特徴量系列を解析する。一方、波形生成部では、言語特徴量系列から音声の音響特徴量を推定し、音響特徴量系列から音声波形を合成する。また、言語特徴量系列 w から音響特徴量系列 o を生成する確率 $P(o|w,\lambda)$ を音響モデルと呼ばれる統計モデルによって学習される。

統計的音声合成システムは学習部分と合成部分から構成される。図 3.1 に統計的音声合成システムの概要を示す。学習部では、学習用の音声から抽出された音響パラメータ O と、テキストか

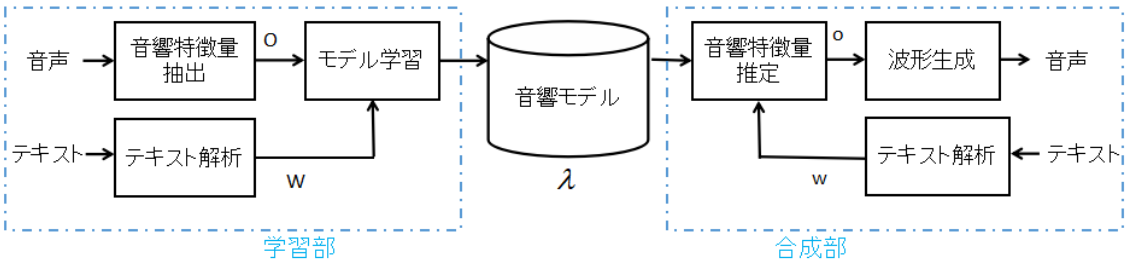


図 3.1: 統計的音声合成システムの概要

ら解析された言語情報 W とを学習し、音響モデル λ を構築する。合成部では、音響モデルを用いて、新たに入力されたテキスト w に対し、それに対応する音響パラメータ o を生成し、音声波形を合成する。確率的な表現を用いて、学習部と合成部は下式により表すことができる。

$$\hat{\lambda} = \arg\max_{\lambda} P(O \mid W, \lambda) \tag{3.1}$$

$$\hat{o} = \arg\max_{o} P(o \mid w, \hat{\lambda}) \tag{3.2}$$

ここで、λ が音響モデルである。λ の学習は、与えられた学習用の音響パラメータ O とテキスト W に対し、$P(O \mid W, \lambda)$ を最大にするような $\hat{\lambda}$ を求めることである。合成する際に、新たに与えられたテキスト w と学習された $\hat{\lambda}$ を用いて、$P(o \mid w, \hat{\lambda})$ を最大となるような $\hat{o}$ を統計的に推定する。最後に音響パラメータ $\hat{o}$ を用いて、音声波形を合成する。

このように、統計的音声合成では、音響モデルによって予測される音響特徴量系列が合成音声の品質に強く影響を与えるため、精度の高い音響モデルを用意することが統計的音声合成の一つの課題と言える。

3.2.2　音声合成に用いられる特徴量

i)　音響特徴量

音響特徴量は、音声波形を再合成するために、音響モデルから予測されるものである。一般的にスペクトルパラメータと励振源パラメータを採択される。中には、スペクトルパラメータはメルケプストラム、励振源パラメータは基本周波数（F0）が広く用いられる。

学習する際に、音響特徴量はフレームごとに音声から抽出し、連続値系列として用いられる。合成する際に、音響モデルを用いて音響特徴量を推定するわけであるが、出力されるパラメータは各フレームの平均値であるため、フレームの連続部が階段状となることがある。これにより、不連続な所が生じてしまい、合成する音声の品質が低下する原因の一つと考えられる。

この問題を解決するために、音声合成では、各フレームに、静的な音響特徴量の他、動的特徴量 [4] と呼ばれる各静的な特徴量の時間方向の 1 次微分、2 次微分に対応するパラメータを加えることとする。これらを連結して、ベクトルの列の形で用いられる。

$$o_t = \left[c_t^T, \Delta c_t^T, \Delta^2 c_t^T\right]^T \tag{3.3}$$

$$\Delta c_t = \frac{1}{2}\left(c_{t+1} - c_{t-1}\right) \tag{3.4}$$

$$\Delta^2 c_t = c_{t-1} - 2c_t + c_{t+1} \tag{3.5}$$

最終的に、音声合成に採択される音響特徴量を図 3.2 に示す。

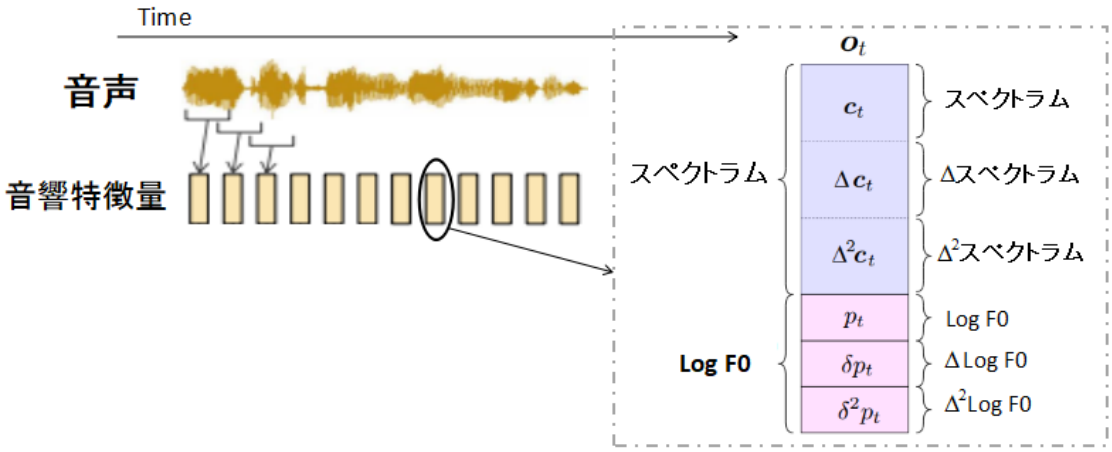


図 3.2: 音声から音響特徴量の抽出及び音響特徴量の構成

ii)　言語特徴量（コンテキストラベル）

言語特徴量は音響モデルの入力として、音素毎にテキストから解析して得られ、音響特徴量と対の形で学習に用いられる。

音声合成に用いられる言語特徴量は、今の時刻で発声される音素の他に、単語、品詞、アクセント型、アクセント句の長さなど様々な情報が含まれている。これは、今発声する音素の音響特徴量は、その前後の音素の違いや発話状況などによって、異なってくるためである。例えば、同じ /a/ という音素は、「か」と「は」を発音した際に現れる /a/ が、特に音素の前半においては異なる特徴を持っている。このような音響特徴量に影響を与える様々な要因はコンテキストと呼ばれる。異なるコンテキストに対し、音素がすべて同一なものとして対応すると、ほかの音素との接続が不連続なことが発生してしまい、合成した音声の品質が低下する原因となる。

このため、同じ音素であっても、コンテキスト（発話状況）の違いを明示する必要がある。音声合成では、コンテキストの違いを表示するには、コンテキストラベルと呼ばれるものを用いられる。このコンテキストラベルは、音声認識においても採択されるが、音声合成の場合、コンテキストラベルを入力として音響特徴量予測するため、より大量のコンテキストラベルが必要となる。表 3.1 には、一般的に、音声合成で用いられるコンテキストラベルを示す。

iii)　音素継続長

音素継続長は、話速のことであり、音声の重要なパラメータとして、特に話者性や、発話スタイルなどを反映する。音声を合成するにあたって、話速を予測し、明示した上で、音響パラメータを生成する必要となる。音声から話速を抽出することは、所詮、音素と音響特徴量とのアライメントを取ることである。つまり、音声を音素毎に区切り、各音素に対応する音響特徴量のフレーム数を求めることである。その詳細は図 3.3 に示す。

表 3.1: 通常、音声合成に用いられるコンテキストラベルに含まれる情報

2 つ前の音素の種類
先行音素の種類
当該音素の種類
後続音素の種類
2 つ後の音素の種類
アクセント型とモーラ位置との差
当該モーラのアクセント句内の位置（先頭から）
当該モーラのアクセント句内の位置（末尾から）
先行アクセント句の長さ
先行アクセント句のアクセント型
当該アクセント句の長さ
当該アクセント句のアクセント型
当該呼気段落中のアクセント句の位置（アクセント句単位，先頭から）
当該呼気段落中のアクセント句の位置（アクセント句単位，末尾から）
当該呼気段落中のアクセント句の位置（モーラ単位，先頭から）
当該呼気段落中のアクセント句の位置（モーラ単位，末尾から）
後続アクセント句の長さ
後続アクセント句のアクセント型
先行呼気段落の長さ（アクセント句単位）
先行呼気段落の長さ（モーラ単位）
当該呼気段落の長さ（アクセント句単位）
当該呼気段落の長さ（モーラ単位）
文中での当該呼気段落の位置（呼気段落単位，先頭から）
文中での当該呼気段落の位置（呼気段落単位，末尾から）
文中での当該呼気段落の位置（アクセント句単位，先頭から）
文中での当該呼気段落の位置（アクセント句単位，末尾から）
文中での当該呼気段落の位置（モーラ単位，先頭から）
文中での当該呼気段落の位置（モーラ単位，末尾から）
後続呼気段落の長さ（アクセント句単位）
後続呼気段落の長さ（モーラ単位）
文の長さ（呼気段落単位）
文の長さ（アクセント句単位）
文の長さ（モーラ単位）

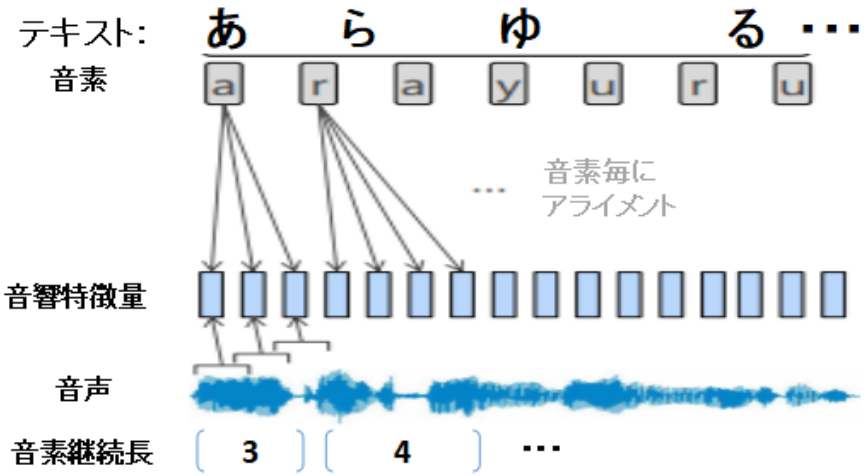


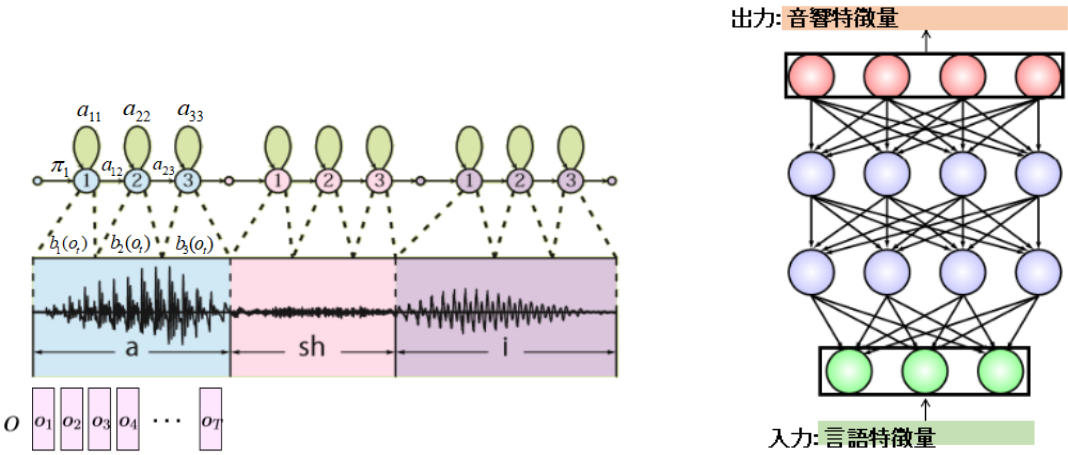
図 3.3: 音声から音素継続長の抽出

3.2.3　音響モデル

音響モデルは、入力される言語特徴量から、音響特徴量を予測することを回帰問題として捉え、統計的音声合成システムの核心として、合成音声の品質に強く影響を与える。

これまで、統計的音声合成の音響モデルとして、隠れマルコフモデル (Hidden Markov Model;HMM)[5, 6] は、音声パラメータ系列を有効にモデリングできる手法として広く利用されている。近年、Deep Neural Network (DNN) を用いた手法が音声認識の分野で高い性能を示したため、HMM の代わりに、DNN を音響モデルとした手法が盛んに研究されてきた。従来の HMM 音声合成に比べ、DNN を用いた音声合成は高品質な音声が合成可能であることが示されている [7]。以下、この 2 つの音響モデルについて説明し、比較する。

HMM 音響モデルは、各音素において、一つの HMM により入力の言語特徴量と出力の音響特徴量をモデリングする。図 3.4a に 3 状態からなる HMM モデルを示す。



(a) HMM 音響モデル　　(b) DNN 音響モデル

図 3.4: 音響モデル

HMM の各状態が、生成確率分布 $b_i(o_t)$ を持つ信号源とみなすことができ、音響特徴量 o_t を出力する。また、各状態の間に、遷移確率 a_{ij}（状態 i から j 状態への遷移確率）をもって接続される。図 3.4a に示しているのは「a」「sh」「i」といった音素の HMM モデルである。しかし、第 (3.2.2) 節でも述べたように、一つの音素であっても、コンテキストによって、対応する音響特徴量が異なるため、対応する HMM モデルも異なってくる。つまり、コンテキストの違いを考慮した HMM モデルの数が膨大なものであり、すべてのコンテキストに対応できる学習データを用意するのは不可能である。このため、HMM 音声合成では、類似したコンテキストを持った音素を同一と見なし、決定ツリーに基づくクラスタリングが行われる [5]。各ノードにおいて、コンテキストに関する質問（「直前の音素は /a/ であるか？」）によりクラスタを二分していき、類似したコンテキストのモデルパラメータを共有することで、学習データに存在しないコンテキストに対応する。図 3.5 に決定ツリーの様子を示す。木をルートノードからリーフノードに辿ることにより、すべてのコンテキストは必ずいずれかのリーフノードに属することとなる。適切な決定木構造を選択することで、高精度なコンテキスト依存モデルを学習する。

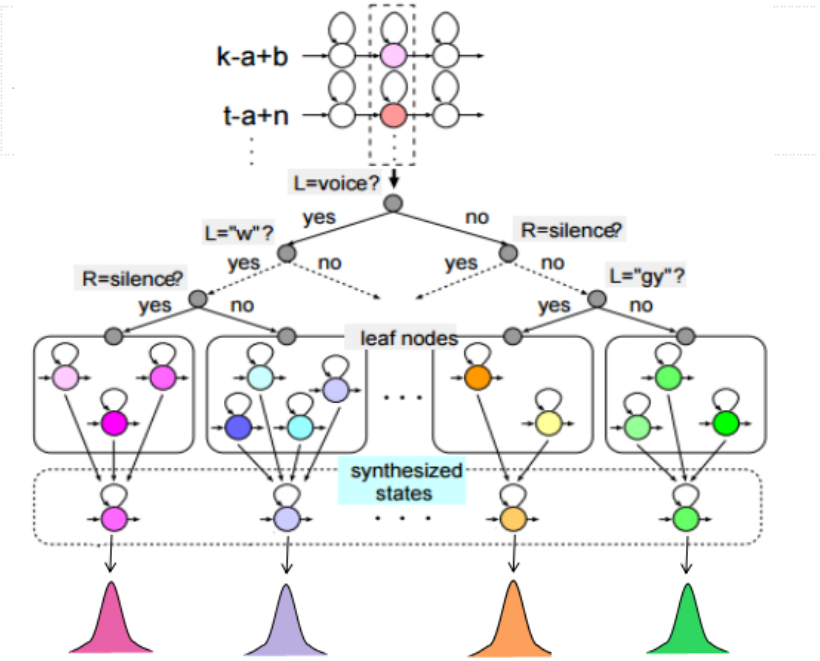


図 3.5: 決定ツリーによるクラスタリング

結局、HMM 音声合成は、決定ツリーに基づくコンテキスト依存 HMM により、言語特徴量と音響特徴量の関係を表現するのである。

これに対し、DNN に基づく音響モデルは、もっとシンプルで、言語特徴量と音響特徴量をそれぞれ DNN の入力と出力とし、入出力の特徴量の関係を一つの DNN より捉える。図 3.4b に示す。

HMM モデルと比較し、

1) 決定ツリーにに基づくコンテキスト依存 HMM は、コンテキストに関する質問によって二分していくため、言語特徴量と音響特徴量の関係が人間によって理解しやすいと言えるが、非線形変換関数等によって表現される複雑な関係を捉えることが難しい。これに対し、DNN モデルは、言語特徴量と音響特徴量の関係を読み取って理解することが難しいが、より複雑なマッピングを表現することができる。

2)HMM モデルは、決定ツリーによって、クラスタリングした結果、学習データが分割されてしまい、他のクラスターを学習するに用いられるデータ量が少なくなる。これに対し、DNN モデルは、すべての学習データを単一のモデルを学習するため、学習データを効率に利用することができる。
このため、DNN に基づく音声合成は、従来の HMM 音声合成に比べ、高精度に音響特徴量を予測することが可能となり、高品質な音声が合成可能である [7]。

3.3　BLSTM-RNN に基づく音声合成

3.3.1　Bidirectional LSTM Recurrent Neural Network(BLSTM-RNN)

第 3.2.3 節では、DNN を用いた音声合成が HMM により高品質な音声を合成可能であると述べた。しかし、DNN は一つの欠点がある。それは、DNN はフレーム単位で独立に学習し、音声における時間的な連続性が無視されてしまうと思われる。これに対し、RNN は、再帰的な構造を持ち、DNN のように単一フレームを用いるのではなく、過去入力されたフレームの隠れ層の情報を用いて、現在の隠れ層を学習し、時間的な連続性を表現することができるため、音声のような時系列的なデータを扱うのに有効と認識される。言語特徴量を $x = (x_1, ..., x_T)$ とすると、RNN の以下のように隠れ層 $h = (h_1, ..., h_T)$ と出力層 $y = (y_1, ..., y_T)$ を計算する。

$$h_t = H(\mathbf{W}_{xh}x_t + \mathbf{W}_{hh}h_{t-1} + b_h) \tag{3.6}$$
$$y_t = \mathbf{W}_{hy}h_t + b_y \tag{3.7}$$

ここで、$\mathbf{W}$ は重み行列で（例えば、$\mathbf{W}_{xh}$ は入力と隠れ層の間の重み行列）、H は活性化関数と呼ばれる非線形関数で、シグモイド関数・tanh 関数などが用いられる。b は偏差ベクトルを表す。RNN の隠れ層では、h_{t-1} によって、これまで入力された音素の情報も暗に持つことになる。また、音素をベクトルとして連続的に表現しているため、音素間の関係や意味の類似度を考慮した上での効果的な学習が行われると期待される。上記に理由から音響モデルの性能としては RNN のほうが優れているという結果が報告されている [10]。

しかし、RNN では長い系列を処理する場合、更新値が指数的に減衰してしまい、勾配消失と呼ばれる問題が発生する。そこで、RNN の隠れ層に memory cell という特別な構造を加えた Long Short-Term Memory(LSTM) と呼ばれる Network を利用した LSTM-RNN 音響モデルが提案されてきた。LSTM は図 3.6 のような構造をしている。

一つの memory cell に三つの gate (forget gate,input gate,output gate) を設けている。これらのゲートにより、中間層の情報 h_t を次の式によって制約される。

$$i_t = \sigma\left(\mathbf{W}_{xi}x_t + \mathbf{W}_{hi}h_{t-1} + \mathbf{W}_{ci}c_{t-1} + b_i\right) \tag{3.8}$$
$$f_t = \sigma\left(\mathbf{W}_{xf}x_t + \mathbf{W}_{hf}h_{t-1} + \mathbf{W}_{cf}c_{t-1} + b_f\right) \tag{3.9}$$
$$c_t = f_t c_{t-1} + i_t \tanh\left(\mathbf{W}_{xc}x_t + \mathbf{W}_{hc}h_{t-1} + b_c\right) \tag{3.10}$$
$$o_t = \sigma\left(\mathbf{W}_{xo}x_t + \mathbf{W}_{ho}h_{t-1} + \mathbf{W}_{co}c_t + b_o\right) \tag{3.11}$$
$$h_t = o_t \tanh\left(c_t\right) \tag{3.12}$$

LSTM-RNN が、memory cell を加えたため、長期間の依存性に対応でき、基本的な RNN より、長時間系列の学習に適用される [11]。

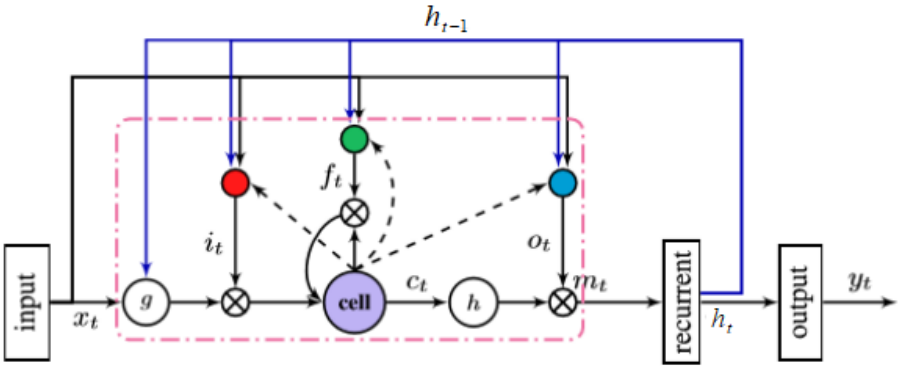


図 3.6: Long Short-Term Memory cell

さらに、BLSTM-RNN は LSTM-RNN の一種類の拡張であり、 二つ逆方向の LSTM 層を重ねることによって、過去の入力情報だけではなく、未来の情報も利用できる。BLSTM-RNN モデルに基づく音声合成システムの全体像を図 3.7 に示す。

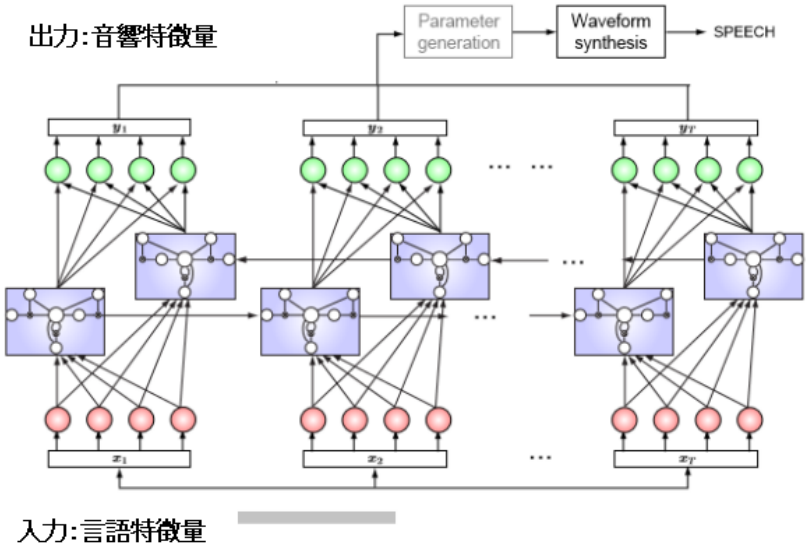


図 3.7: BLSTM-RNN に基づく音声合成システムの全体像

3.3.2　学習部

学習部では、言語特徴量と音響特徴量との間の非線形変換関数を BLSTM より学習する。前節（ 3.2.2）でも述べたように、通常、言語特徴量は音素単位で、音響特徴量はフレーム単位で抽出されるため、入出力の時間単位が異なる。しかし、BLSTM は一対一で入出力の対応関係をトレーニングするため、時間単位の異なる言語特徴量と音響特徴量を学習できない。このため、学習する前に、データの準備として、入出特徴調の時間長を揃える必要がなる。具合的には、各音素において、その音素の言語特徴量を該当音素の音素継続長と同じ長さになるように、時間軸に沿ってコピーする。時間順位を明示するために、個々のコピーデータの最後に、「第 N フレーム」と

いう数値データを加える。また BLSTM の入力とする言語特徴量を、数値データとして表現する必要もある。このため、コンテキストに関する質問に対して、二値 (0/1) または連続値での回答に変換し、数値データとして連結する言語特徴量を用意する。

学習するときに、言語特徴量と音響特徴量をフレーム単位で、二乗誤差最小化によって、BLSTM モデル $\hat{\lambda}$ を学習する。

$$\hat{\lambda} = \arg\min_{\lambda} \frac{1}{2} \sum_{t=1}^{T} \left\| o_t - g_\lambda \left(w_t \right) \right\|^2 \tag{3.13}$$

ただし、w_t と o_t はそれぞれフレームにおける言語特徴量と音響特徴量、λ は BLSTM モデルパラメータ、$g_\lambda(.)$ は BLSTM により表現される非線形変換関数である。

確率的な表現である式 3.1 との関係を考えると、BLSTM の学習は次式のように書き換えることができる。

$$\hat{\lambda} = \arg\max_{\lambda} P(o \mid w, \lambda) \tag{3.14}$$

$$= \arg\max_{\lambda} \prod_{t=1}^{T} \mathcal{N} \left(o_t \mid \bar{\mu}_t, \bar{\Sigma}_t \right) \tag{3.15}$$

ここで、$\bar{\mu}_t$ は非線形変換関数 $g_\lambda(w_t)$ のことであり、BLSTM の出力である音響特徴量は正規分布の平均ベクトルとして表す。

3.3.3　合成部

合成部では、入力された言語特徴量から BLSTM を用いてフレームごとに音響特徴量を推定する。まず、音素単位の言語特徴量から音素継続長を推定し、フレーム単位の言語特徴量に変換する。次に、フレーム単位の言語特徴量を BLSTM へと入力し、BLSTM からの出力を連結することで、音響特徴量系列を得る。また、音響特徴量を生成する際に、動的特徴量を用いてパラメータ生成アルゴリズムを採択する。

静的特徴量 c_t^T と動的特徴量 $\Delta c_t^T, \Delta^2 c_t^T$ から音響特徴量 $o = [o_1^T, ..., o_T^T]^T$ が構成されるとき、音響特徴量は以下のように求められる。

$$o = Wc \tag{3.16}$$

ここで、W は静的特徴量系列 $c = [c_1^T, ..., c_T^T]^T$ から動的特徴量を含む音響特徴量系列 o を求める行列である。動的特徴量を用いて制約として、パラメータ生成アルゴリズムは次式を表す。

$$\hat{c} = \arg\max_{c} \mathcal{N}(Wc \mid \bar{\mu}, \bar{\Sigma}) \tag{3.17}$$

ここで、$\bar{\mu}$ は $[\bar{\mu}_1^T, ..., \bar{\mu}_T^T]^T$ であり、$\bar{\mu}_t$ は言語特徴量 w_t を入力としたとき BLSTM の出力 $g_\lambda(w_t)$ であり、$\bar{\Sigma}$ は $diag[\bar{\Sigma}_1, ..., \bar{\Sigma}_T]$ である。BLSTM の出力を連結した平均ベクトル $\bar{\mu}$ を用い、式 3.17 を解くことで、滑らかな音響特徴量が生成され、BLSTM に基づく音声合成が実現される。

3.4　BLSTM に基づく読み聞かせ調音声を合成するための実験

以上、BLSTM-RNN に基づく音声合成の流れについて詳述した。本節では、BLSTM 音声合成の枠組みにおいて、読み聞かせ調音声コーパスを用いて、音声合成の実験について述べる。

3.4.1　読み聞かせ調音声の韻律特徴をコンテキストラベルに追加

合成音声に読み聞かせ調音声の特徴を反映するために、BLSTM 音響モデルの入力に読み聞かせ調音声の特徴を明示する必要となる。そこで、第 2.3 節で述べた韻律特徴（「長音化」、「上昇調 BPM」、「抑揚」、「キャラクター属性」）を BLSTM の入力であるコンテキストラベルに追加し、出力の韻律特徴量を制御する。追加したコンテキストラベルの種類は表 4.2 に示す。

表 3.2: 読み聞かせ調音声の韻律を明示するコンテキストラベル

当該モーラに上昇調 BPM があるか
地の文/セリフ、キャラクター属性（大人（男・女)、子供（男・女))
抑揚ラベルの属性（なし、抑（弱・中・強)、揚（弱・中・強))
先行モーラが長音化しているか
当該モーラが長音化しているか
後続モーラが長音化しているか

3.4.2　実験設定

第 2.2 節で紹介したパラレルコーパスを用いて、読みあげ調音声と読み聞かせ調音声合計 1834 文の内、1634 文を学習データとし、100 文をテストデータとし、音声の合成実験を行った。

出力の音響特徴量として、STRAIGHT 分析により得られたメルケプストラム、Log F0、およびそれらの Δ、Δ^2 を採用する。中には、メルケプストラムは各フレームにおいて 26 次元であり、Log F0 に関しては、各時刻の前後 4 フレーム、合計 9 フレームの Log F0 である。

入力のコンテキストラベルとして、表 3.1 に示している一般的に音声合成に用いられるコンテキストラベルの他、第 3.4.1 節で述べた読み聞かせ調音声の韻律特徴を反映する追加ラベルも採用する。これらのコンテキストラベルを数値データへと変換し、連続的なベクトルとして用いられる。

実験条件の詳細を表 3.3 に示す。

表 3.3: BLSTM-RNN に基づく読み聞かせ調音声合成の実験設定

データセット	パラレルコーパス
	読みあげ調・読み聞かせ調
学習/テストデータ	1634 文 /100 文
コンテキストラベル（入力）	(36×5) 次元カテゴリー特徴量
	38 次元数的特徴量
音響特徴量（出力）	26 次元 mgc、9 次元 Log F0 、Δ, Δ^2
BLSTM-RNN	2 BLSTM 層　256-256-128-64 ノード/層
	sigmoid

3.4.3　客観評価実験

100 文のテストデータを用いて、実際に音声を合成し、客観評価実験を行った。

本実験の目的は、第 3.4.1 節で述べた韻律ラベルを用いて、BLSTM 音響モデルより読み聞かせ調音声の韻律制御の性能を評価することである。つまり、韻律ラベルを追加したことで、合成された音声が読み聞かせらしくなるか否かを評価する。このため、実験では、同じ文章に対し、2 種類のテストデータ (S_{id}・S_{ad}) を用いた。

S_{id} は、読み聞かせ調音声の韻律を明示する追加ラベルを有効にしたコンテキストラベルを用いたテストデータであり、S_{ad} は、追加韻律ラベルを off にしたコンテキストラベルを用いたテストデータである。理想的には、S_{id} の合成音声が読みあげ調音声の韻律を表現するため、S_{ad} の合成音声より、読み聞かせ調に似ることが望ましい。

これを評価するために、韻律に強く関係する F0 パターンを用いて、コーパスにある読み聞かせ調音声を評価データとする。評価方法は、図 3.8 で示すように、S_{id} と S_{ad} の合成音声の F0 パターンと評価データ（読み聞かせ調）の F0 パターン間の相関を相互相関係数によって比較し、どれが読み聞かせ調音声に近いかを相関係数の大小で判別する。

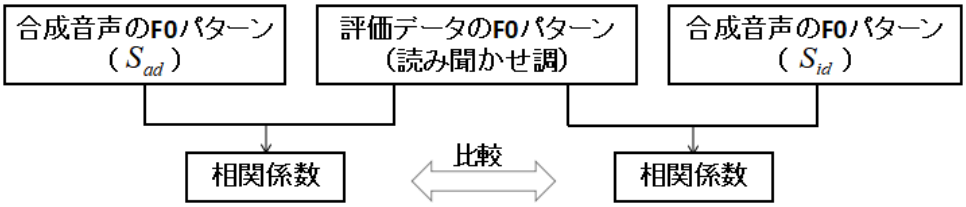


図 3.8: 客観評価実験の方法

相互相関係数の算出は式 3.18 を用いる。

$$r(x,y)\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2\sum(y_i-\bar{y})^2}} \tag{3.18}$$

比較の結果が、ファイル総数で表す。表 3.4 に示す。

表 3.4: BLSTM 音響モデルを用いた韻律制御の客観評価の結果

	S_{id}	S_{ad}
ファイル数	82	18

100 文の内、S_{id} の合成音声が目標音声の韻律に近いと評価された文数が 82 である結果が得られた。これにより、韻律ラベルを用いて、BLSTM 音響モデルを用いた読み聞かせ調音声合成の韻律制御が有効であると考えられる。しかし、S_{id} の合成音声において、各相関値におけるファイル数の分布を調べると（図 3.9 に示す）、相関値の低いファイル数が 5 分の 1 程度あることもわかる。これは、BLSTM 音響モデルにより韻律の制御は可能であるが、まだ学習が不十分な所もあると考えられる。実際に S_{id} の合成音声を聞くと、意図せぬ箇所でピッチが変動したり、不自然な音声となることも確認されている。これは、読み聞かせ調音声における多様な発話スタイルが原因であると思われる。コーパスデータが十分であれば、制御の性能を上げることが可能であるが、1634 文の学習データで、音響モデルによる韻律制御が容易ではないといえる。

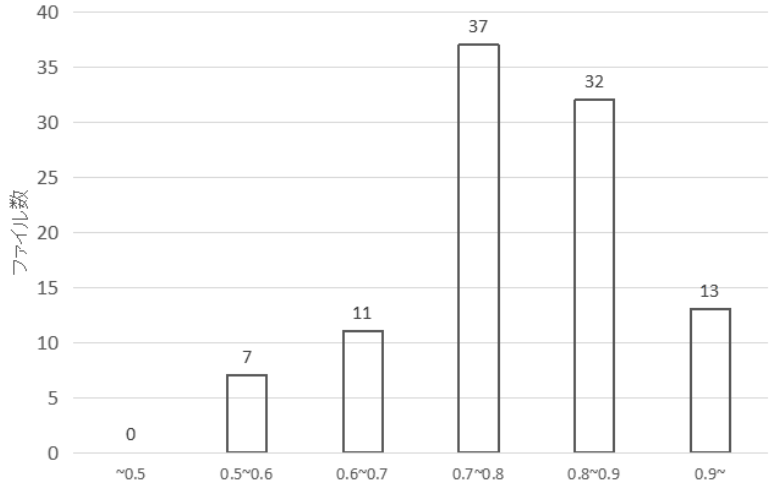


図 3.9: S_{id} の合成音声における各相関値のファイル数の分布

3.5　むすび

本章では、統計的音声合成の概要と BLSTM 音響モデルに基づく読み聞かせ調音声合成の実験を述べた。実際に韻律制御を加えて、音声を合成した結果、意図せぬ箇所で F0 が変動するなど、不自然な音声となることが分かる。これは、音響モデルが、韻律情報に比べ、言語情報の伝達に重きを置いており、学習データが少ない場合、韻律に関する制御を上手く捉えられないと考えられる。次章では、この結果を踏まえ、テキストから読み聞かせ調音声を合成する代わりに、最初に読みあげ調音声を合成し、次に合成された読みあげ調音声の韻律パラメータを韻律変換と呼ばれる手法を用いて、読み聞かせ調音声の韻律へと変換することで、読み聞かせ調音声の合成を実現する手法を紹介する。

第4章

韻律変換を用いた
読み聞かせ調音声合成の提案

4.1　はじめに

第 3 章では、BLSTM を用いた音声合成の枠組みにおいて、読み聞かせ調音声の合成を実行した。これは、読み聞かせ調音声が持つ韻律的情報と言語的情報を同時に音響モデルを用いて学習させるという考えを踏まえた手法と考えられる。

しかし、音響モデルは、所詮言語情報を捉えることに重きを置いており、学習データが少ない場合、韻律に関する制御を上手く捉えられないことがある。特に多様な変化を含まれる読み聞かせ調音声において、ストーリーやキャラクター属性によって、ピッチや話速のバリアンスが非常に大きいので、合成された音声において、意図せぬ箇所でピッチが変動するなど、不自然な音声となることが第 3 章の実験評価により分かる。これに対し、学習データが十分であれば、このような問題はないと思われるが、今のコーパスのデータ量では、音響モデルを用いた読み聞かせ調音声の韻律制御の精度がよくないと言える。しかし、データが多ければ、多様な韻律変化は、音響モデルより平均化されてしまう恐れもあると考えられる。これらの原因で、韻律特徴量と言語特徴量を同時に音響モデルに学習させ、テキストから読み聞かせ調音声を合成するのは容易ではない。

expressive な音声合成の分野でも、韻律に関する制御が特に困難と思われ、音声の感情を規則化したり、制御する研究が行われている。こうした研究の中で、テキストではなく音声を入力とし、声質や話者性を変えずに、目標の発話スタイルへと韻律だけを変換する方法が研究されている。これは、複雑な変化が含まれる韻律情報に着目し、処理するモジュールを導入することで、合成音声の韻律だけを変化させたり、制御したりすることを目的とする。

韻律変換に関する研究では、主に読みあげ調音声から感情音声への変換 [14, 8] に主題を置いており、入力音声と出力音声の韻律パラメータを連続的な変換関数により捉えることが多い。これは、どのような音韻素性、韻律素性がどのように変化されたのか、について記述することなく変換を実装することに起因すると考えられる。しかし、我々が発話するとき、発話の状況や内容等によって、感情や、発話スタイルも変わっていく。すなわち、言語そのものが韻律の変化と暗に関係あると考えられる。

この考えを踏まえて、本研究では、テキストから読み聞かせ調音声の韻律を予測してから韻律変換する手法を提案する。つまり、読みあげ調音声だけではなく、テキスト情報も利用し、テキストから生成された韻律を、読みあげ調音声に付与することで、読み聞かせ調音声へと変換する手法である。韻律の内、F0 と話速を重要であると考え、特にこの 2 つの韻律特徴量の変換を検討する。

本章では、提案手法の重要な処理をモジュールに分けて説明する。

4.2　韻律変換に基づく読み聞かせ調音声合成システム

4.2.1　韻律変換に基づく読み聞かせ調音声合成システムの概要

本節では、提案する韻律変換に基づく読み聞かせ調音声合成システムの概要を紹介する。システムの全体像を図 4.1 に示す。本システムは読みあげ調音声合成器、韻律生成モジュール、韻律変換モジュールから構成される。

1) 読みあげ調音声合成器では、入力されるコンテキストラベル w に対し、音声波形を生成するための音響パラメータ o（スペクトラム・F0）を予測する。ここの音響パラメータのスタイルは、読みあげ調音声に対応する。つまり、ここの読みあげ調音声合成器は、一般の合成器と同様である。

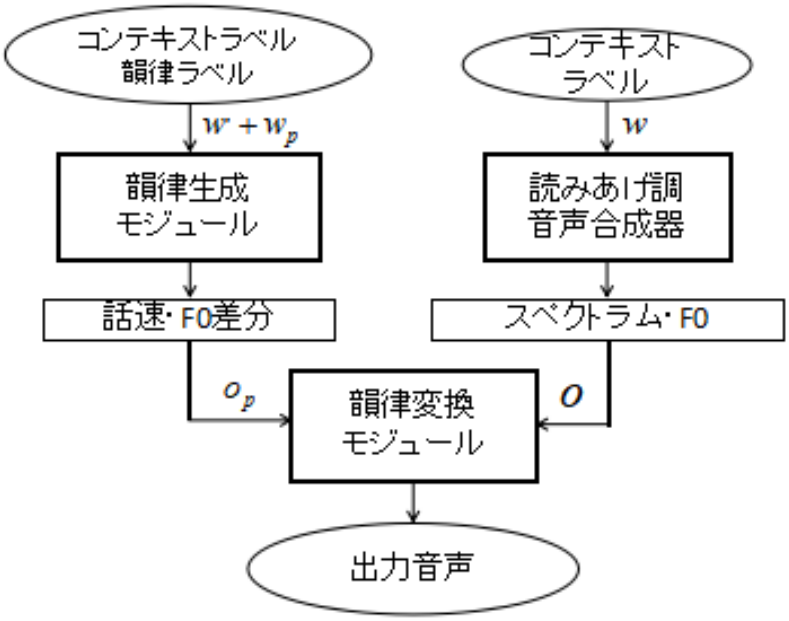


図 4.1: 韻律変換に基づく読み聞かせ調音声合成システム

2) 韻律生成モジュールでは、読み聞かせ調音声の韻律 o_p（話速・読み聞かせ調音声と読みあげ調音声の F0 差分）を生成する。生成される韻律が、入力のコンテキストラベル w'(読みあげ調音声合成器に用いられる w の一部である）と読み聞かせ調音声の韻律ラベル w_p により制御できる。
3) 韻律変換モジュールでは、読みあげ調音声合成器から生成された音響パラメータ o と、韻律生成モジュールから予測された韻律特徴量 o_p を入力とし、o に o_p を付与することで、読みあげ調音声から読み聞かせ調音声への変換を実現する。

4.2.2　読みあげ調音声合成器

読みあげ調音声合成器の目標は、入力のコンテキストラベルに対し、読みあげ調音声を生成することである。すなわち、入力テキストの言語情報を上手く捉えることである。これは提案システムの最初の処理であり、他の処理の基礎でもあるため、システム全体の性能に大きく影響するといえる。

合成器は、 3.3 節で述べた BLSTM-RNN に基づく音声合成の枠組みを用いる。コーパス内の読みあげ調音声（917 文）の内 817 文を学習データとし、100 文をテストデータとする。

音響パラメータのスペクトラムと F0 は STRAIGHT 分析により求める。各時刻に対して 26 次のメルケプストラムを求め、F0 に関しては、各時刻の前後 4 フレーム、合計 9 フレームの Log F0 を採用する。入力のコンテキストラベルとして、第 3 章の表 3.1 に示した一般的に音声合成に用いられるコンテキストラベル（読み聞かせ調音声の韻律ラベルを含まず）を用いる。

4.2.3　韻律生成モジュール

従来の韻律変換の研究では、入出力音声の韻律パラメータを対象とし、それらを連続的な変換関数により捉えることが多い。これは、どのような音韻素性、韻律素性がどのように変化された

のか、について記述することなく変換すると思われる。これに対し、本システムは、言語の裏に持つ韻律特徴を考慮し、言語情報から韻律を生成する形で目標音声の韻律を捉えることを提案する。

提案システムに用いられる韻律生成モジュールでは第 2.3 節で述べたラベルを韻律素性として捉え、合成器の入力テキスト情報から得られる一部のコンテキストラベルと目標先（読み聞かせ調音声）の韻律ラベルに基づき、目標音声の韻律を予測する。表現力の高いニューラルネットワークを用いて、韻律（話速・両スタイル音声の F0 差分）を、目標先の韻律ラベルを用いて、制御可能な形で生成するモデルを用いる。

生成モデルは 2 つあり、DNN を用いた話速生成モデルと、BLSTM を用いた F0 差分生成モデルである。入力特徴量として，言語的情報を表すコンテキストラベル（表 4.1 に示す）の他，目標音声の韻律特徴を反映する韻律ラベル（表 4.2 に示す）も用いられる。

表 4.1: 韻律生成モデルに用いられる言語的コンテキストラベル

先行音素の種類
当該音素の種類
後続音素の種類
当該モーラのアクセント句内の位置（先頭から）
当該モーラのアクセント句内の位置（末尾から）
当該アクセント句の長さ（モーラ単位）
当該フレーズの長さ（音素単位）

表 4.2: 韻律生成モデルに用いられる韻律ラベル

当該モーラに上昇調 BPM があるか
地の文セリフ、キャラクター属性（大人（男・女)、子供（男・女))
抑揚ラベルの属性（なし、抑（弱・中・強)、揚（弱・中・強）
先行モーラが長音化しているか
当該モーラが長音化しているか
後続モーラが長音化しているか
当該フレーズの長さ（モーラ単位）

i)　話速生成モデル

話速生成モデルは、入力のコンテキストに対し、目標音声の音素継続長を生成する。つまり、音素単位で、話速を予測することである。4 層の DNN を用いた。各隠れ層のノード数は 256 – 256 – 128 – 64 である。1830 文の内 1730 文を学習データとし、100 文をテストデータとした。

ii)　F0 差分生成モデル

第 2.4 節において、読み聞かせ調音声は、キャラクターの属性や、抑揚などを表現するため、F0 の標準偏差が極めて大きいと報告した。要するに、読みあげ調音声から読み聞かせ調音声へと

変換する際に、F0 に関する制御が非常に重要であるといえる。このため、もし両スタイル音声の F0 パターンの差異を上手く捉えられれば、目標音声により近い韻律を模倣することが可能であると考えられる。

そこで、本システムは、各時刻における読みあげ調音声と読み聞かせ調音声の F0 パターンの差分を対象とし、この F0 差分を予測できる生成モデルを導入する。ここで、パラレルデータ（同じ内容を、読みあげ調と読み聞かせ調より読まれた音声）を活用し、読みあげ調音声の F0 パターンと、読み聞かせ調の F0 パターンの差異をモデル化する。具合的には、対となる発話（読み上げ・読み聞かせ）の F0 パターンを比較し、両者の差を時系列としてモデル化する。比較のためには両者の時間長を揃える必要がある。ここで、DTW(Dynamic Time Warping) 用いる。

図 4.2 に示している DTW とは、非線形マッピングで二つの時系列の各点の距離を総当たりで比較した上で、系列同士の距離が最短となるパスを見つける技術である。

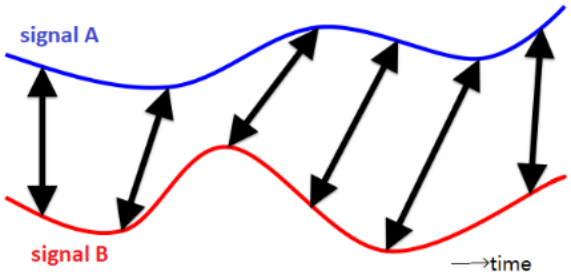


図 4.2: DTW(Dynamic Time Warping)

音声の場合、対となる発話に対し、周波数領域でのメルケプストラムを DTW に用いられることで、両信号のマッピングパスが得られる。両信号の時間伸縮は得られるパスに従って行われる。距離を計算する際に、式 4.1 で定義されるメルテプストラム歪み (MCD) を用いる。

$$MCD(c^{id}, c^{ad})[dB] = \frac{10\sqrt{2}}{\ln 10}\sqrt{\sum_{d=1}^{D}\left(c_d^{id} - c_d^{ad}\right)^2} \tag{4.1}$$

DTW の精度が F0 差分の計算に強く関係あるため、ここで、以下の 2 つのマッピング方法を試みた。

1)utterance を単位としたマッピング方法

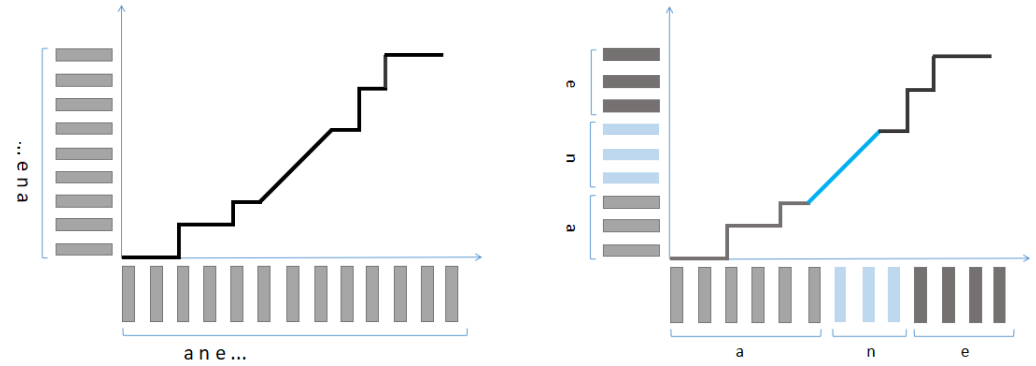
これは、両音声に対し、ポーズや、句点などに関わらず、一文と見なして、全体に対して DTW を計算し、マッピングを行う方法である。その詳細を図 4.3a に示す。

2) 音素ラベル情報を用いた音素単位でのマッピング方法

これは、両音声において信頼できる音素ラベル情報が存在すると仮定し、音素ラベルに従って、音素を単位として区間対を作成し、各区間に対して DTW を計算し、マッピングを行う方法である。その詳細を図 4.3b に示す。

両方法におけるマッピングの性能を比較するために、予備実験として、各マッピング方法の DTW 距離とパスを計算した。距離の結果を表 4.3 に示す。

どの方法においても、DTW の距離が大きい（6.8 以上）ことが分かる。これは、読み聞かせ調音声が持つ極めて多様な発話スタイルの影響だと思われる。例えば、距離の最も大きい文（utterance 単位の距離が 8.9、音素単位の距離が 8.3）では、話者が物まねをするような形でセリフ「いじわ



(a) utterance を単位とするマッピング方法　(b) 音素ラベル情報を用いた音素単位でのマッピング方法

図 4.3: DTW を用いたマッピング方法

表 4.3: DTW 距離（utterance 単位・音素単位）

方法	utterance 単位	音素単位
距離	7.6	6.8

る、いじわる」を読むため、内容が同じであっても、発話スタイルがかなり異なるため、両音声の距離が大きい。

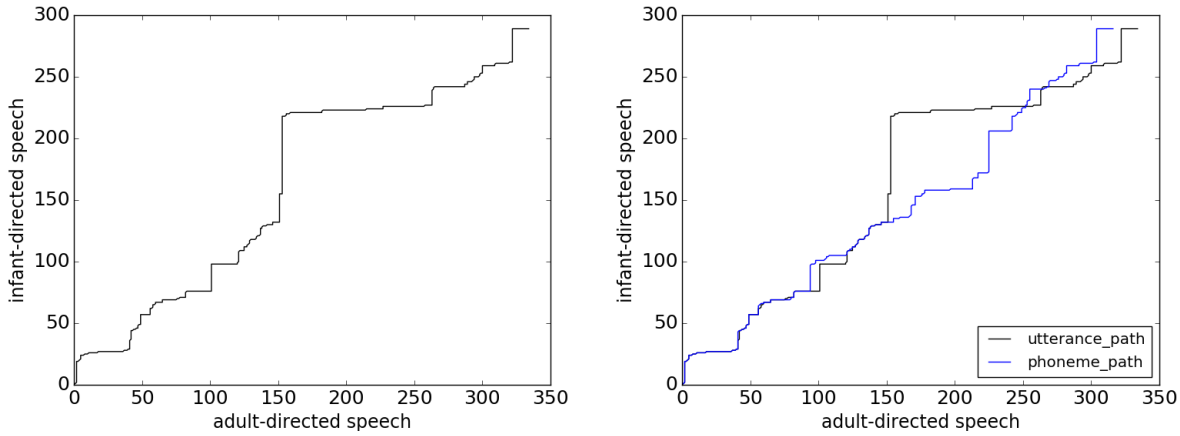
また、両方法を用いたマッピングのパスを調べたところ、utterance 単位方法においては、パスがうまく取れなかった場合が観測された。これは、読み聞かせ調音声において、一般的にポーズを長く延ばしたり、あるいは、読みあげ調にポーズのないところにポーズを入れ入ったりする傾向があるので、両音声の対応パスが斜線の代わりに、垂直線や水平線になってしまうのであると考えられる。また、パラレルコーパスとして構築されたが、両スタイル音声の発話内容が必ずしも一致とは言えない。例えば、話者が収録の時、不注意によって「は」を「が」に読み間違い等のことがあると、音素ラベルを調べることより観測されている。

これらの問題に対して、音素単位でマッピングする方法では、音素ラベルを参考とし、音素区間を分割したうえで、各音素の区間内に DTW を計算するため、ポーズを削除したり、同一音素であるかどうかを確認したりすることができ、ポーズのミスマッチや読み間違いの問題を解決できると考えられる。

また、「いじわる、いじわる」の例でいうと、utterance 単位での方法では、図 4.4a に示しているようなパスが得られた。音素単位での方法を用いて、DTW のパスが図 4.4b に示しているようになり、utterance 単位での方法よりある程度改善されていることが分かる。

そこで、最終的に、音素ラベル情報を用いた音素単位でのマッピング方法を用いる。F0 差分の計算は図 4.5 に示しているように、対となる発話（読み上げ・読み聞かせ）に対して、まず、音素を単位とし区分化する。次に各音素区間において、得られた音響パラメータのスペクトラムを用いて DTW を行い、得られたパスに従って F0 差分を計算する。

生成モデルは、BLSTM を用いて、計算された F0 差分を時系列としてモデル化する。ネットワークが二層の NN と二層の BLSTM で、ノード数は 256 – 128 – 64 – 64 に設定した。学習データとテストデータはそれぞれ 810 文と 100 文である。



(a) utterance 単位でのマッピングパス　(b) 音素単位でのマッピングパス

図 4.4: 「いじわる、いじわる」を発話する例で、utterance 単位 (a) と音素単位 (b) でマッピング方法より得られるパス

4.2.4　韻律変換モジュール

韻律変換モジュールでは、読みあげ調音声合成器から生成された読みあげ調音声の音響パラメータを、韻律生成モジュールから予測された韻律特徴量を用いて、所望の値に変換させ、その後音声を出力する。ここで、話速と F0 を変換対象とする。以下、それぞれの変換方法を紹介する。

i)　話速変換

変換モジュールでは入力の音響パラメータに対して、最初に行う変換は話速変換である。即ち、入力の読みあげ調音声の時間長制御を、生成モデルより予測された読み聞かせ調音声のそれと置換する。

話速生成モデルでは、音素毎に子音・母音の継続長を予測するが、変換にあたってはこれを直接採用せず、モーラ単位で時間伸縮する。

話速生成モデルの性能評価（第 5.2.2 節）により、DNN が極めて長い音素長を予測する傾向がある。これは、読み聞かせ調音声が持つ極めて多様な発話スタイルの影響だと考えられる。また、母音と比較して、子音の予測精度が低い。これは、母音においては、長音化等を表現するために、音素長を極端に伸ばす場合があるが、子音においては、音素長をきわめて長く予測されると、発声の自然性を壊すことがある。

コーパスが十分にあれば、自然性を壊す音素長を予測することはないと思われるが、本実験では上記の結果が得られたので、各子音に対して、子音長の最大値、最小値を設定した。DNN の予測がこれを超えた場合は、最大値／最小値を採択することとした。

モーラ単位での時間伸縮であるが、例えば CV に対して、DNN による C の予測長が最大値を超えた場合、最大値を継続長として採択し、その分、V の継続長を予測長より増分する。CV 単位では、予測長通りに時間伸縮する。

次に、入力音声の音響パラメータ（スペクトラム・F0・パワー）を、音素毎に目標音声に合うように時間軸方向に伸縮する。伸縮方法は、Catmull-Rom スプライン補間法を用いる。

スプライン補間法は、与えられる複数の離散的な点を制御点とし、すべての制御点を通過する

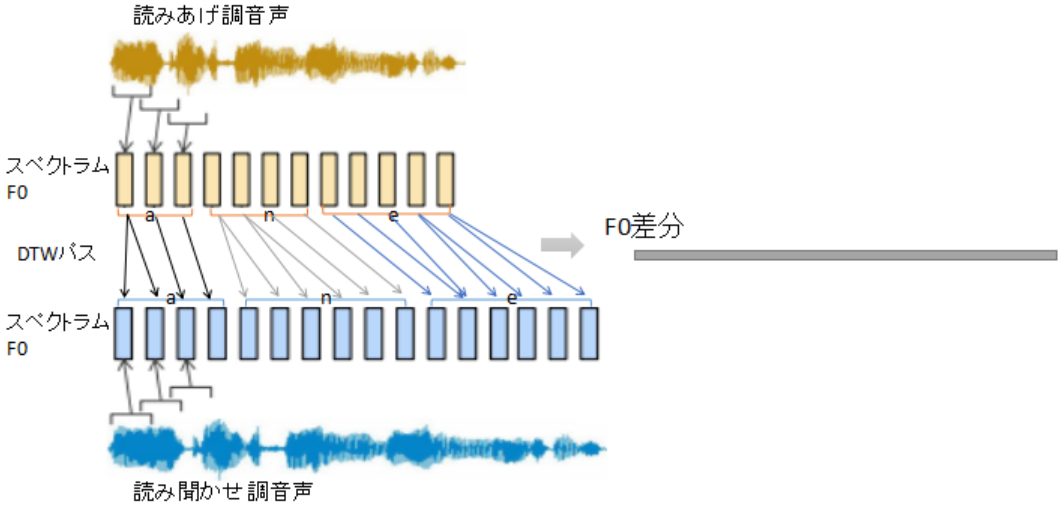


図 4.5: 音素単位で DTW を用いた F0 差分の計算

関数を求め、制御点の間の値を推測する方法である。これにより、伸縮する音響パラメータ間を Catmull-Rom スプラインを用いて補間し、補間された曲線から所望の長さとなるように等間隔にサンプリングするで、目標音声の時間長に合うような音響パラメータが得られる。図 4.6 に示す。

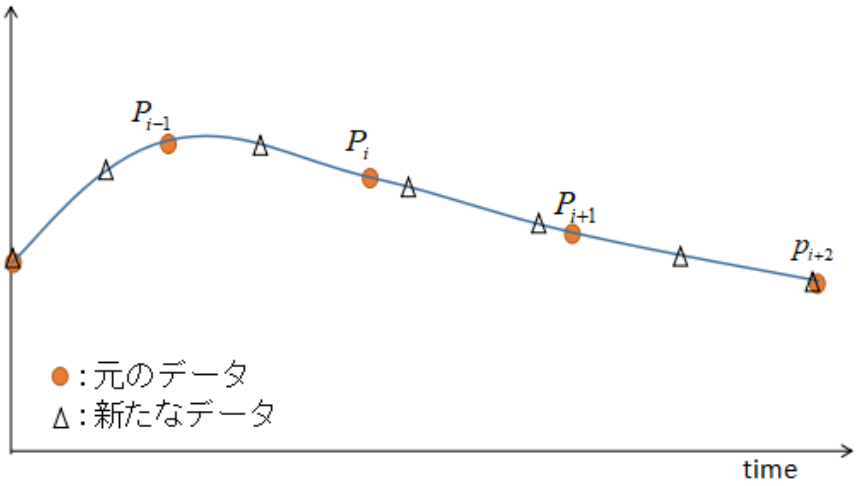


図 4.6: Catmull-Rom スプライン補間法

Catmull-Rom スプラインは四つの制御点 $P_{i-1}, P_i, P_{i+1}, P_{i+2}$ を用いて計算され、P_i, P_{i+1} の両点を通過し、その間を滑らかにつなぐ特徴がある。P_i, P_{i+1} の間の補間値 $x_i(t)$ を式 4.2 により求

める。

$$x_i(t) = \frac{1}{2}\left[t^3\ t^2\ t\ 1\right]\begin{bmatrix} 2 & -2 & 1 & 1 \\ -3 & 3 & -2 & -1 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}\begin{bmatrix} P_i \\ P_{i+1} \\ v_i \\ v_{i+1} \end{bmatrix} \tag{4.2}$$

$$v_i = \frac{1}{2}\left(P_{i+1} - P_{i-1}\right) \tag{4.3}$$

$$v_{i+1} = \frac{1}{2}\left(P_{i+2} - P_i\right) \tag{4.4}$$

図に 4.7 モーラを例「re」とした伸縮前後のスペクトラムを示す。

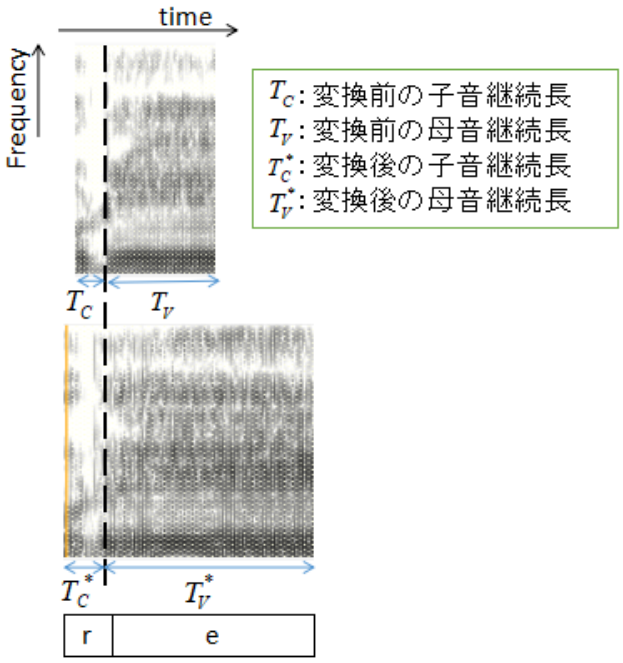


図 4.7: モーラ「re」の時間伸縮例

ii)　F0 の変換

F0 の変換は、話速変換の後に行われる。これは、F0 差分生成モデルは読み聞かせ調の時間軸に合わせて設計されており、話速変換が行われた後であれば、F0 差分モデルの出力をそのまま適用することが可能であるためである。これは、話速変換された入力音声の各時刻の F0 に対して、この F0 差分モデルの出力を適応することを意味するが、そのまま実行すると、局所的に大きな変動を示すことがあった。これも読み聞かせ調コーパスの多様性の大きさ、及び用意できたコーパスサイズに起因すると思われるが、継続長制御同様、適切な制約を導入することでこれに対応した。

ここでは、ピッチターゲットモデル [14] を採用し、モーラ単位で F0 パターンをモデリングすることにする。ピッチターゲットモデルでは、実際の F0 パターンが以下の二つの要素から制御されると仮定する。

1) 発話するときの音調意図を表すピッチターゲットと呼ばれるもの。

2) 調音的制約を表すもの。

図 4.8 に示しているように、ピッチターゲット（破線）は、F0 の変化のトレンドを表し、実際の F0 パターン（実線）は調音的制約を受けた後、漸近的にピッチターゲットに近似する。3 つの垂直線はモーラの境界を表す。一つのモーラのピッチターゲットに到着すると、次のモーラのピッチターゲットに近似し始める。

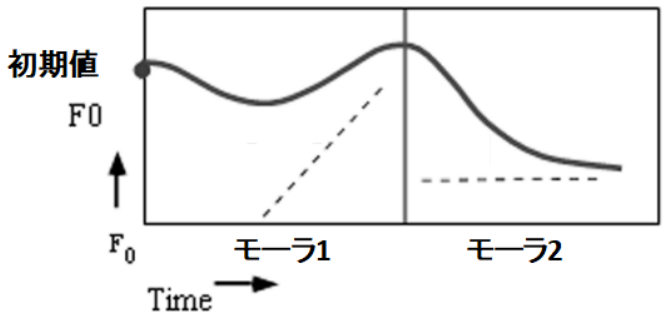


図 4.8: ピッチターゲット

モーラの時間長を [0,D] とすると、実際の F0 パターンが以下の $y(t)$ で近似される。

$$T(t) = at + b \tag{4.5}$$

$$y(t) = \beta \exp(-\omega t) + T(t) \qquad (\omega > 0) \tag{4.6}$$

ここで、$T(t)$ がピッチターゲットを表す。これに対し、$\beta \exp(-\omega t)$ は調音的制約を表す。$y(t)$ が調音的制約を受けた後、最終的の F0 パターンである。β は、$t = 0$ の時の実際の F0 値とピッチターゲットとの距離を、ω は、ターゲットに到達するまでの速さを表す。各モーラの F0 パターンが、4 つのパラメータ a, b, β, ω により近似できる。

モデルパラメータは、非線形回帰により推測される。式 (4.6) を簡略化するために、パラメータを以下のように F0 の観測値に置き換える。

1)(t_0, y_0) を F0 パターンの最初の点と仮定し、式 (4.6) に代入すると、式 (4.7) が得られる。

$$y(t) = (y_0 - b) \exp(-\omega t) + at + b \tag{4.7}$$

つまり、$t_0 = 0$ である。

2)t_1 の時、ピッチターゲットが y_1 に到着すると仮定する。つまり、$\exp(-\omega t)$ が 0 となる（ここで、0 に近似すると意味する）。式 (4.6) が式 (4.8) となる。

$$y_1 = at_1 + b \tag{4.8}$$

また、式 (4.8) を用いて、b を置き換えると、y_t が (4.9) と書き換える。

$$y_t = (y_0 - y_1 + at_1) \exp(-\omega t) + at + y_1 - at_1 \tag{4.9}$$

ここで、Levenbery-Marquardt[21] アルゴリズムを用いて、パラメータを推定する。もし、非線形回帰手法が失敗すると、線形回帰手法を採択する。つまり、β, ω が 0 に設定される。

ピッチターゲットにより F0 パターンのモデリング精度を調べるために、コーパスデータにおいて、上記の方法よりモデルパラメータを推定し、y_t（式 4.6）より近似される F0 パターンと実際の F0 パターン間の誤差を RMSE(式 4.10) を用いて計算した。

$$RMSE(f^{syn}, f^{ref}) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(f_i^{syn} - f_i^{ref}\right)^2} \tag{4.10}$$

1700 文のデータに対し、平均 RMSE は 1.45Hz である。ピッチターゲットより F0 パターンを近似することが可能であると考えられる。図 4.9 が上記の方法より実際の F0 パターンを近似する例である。

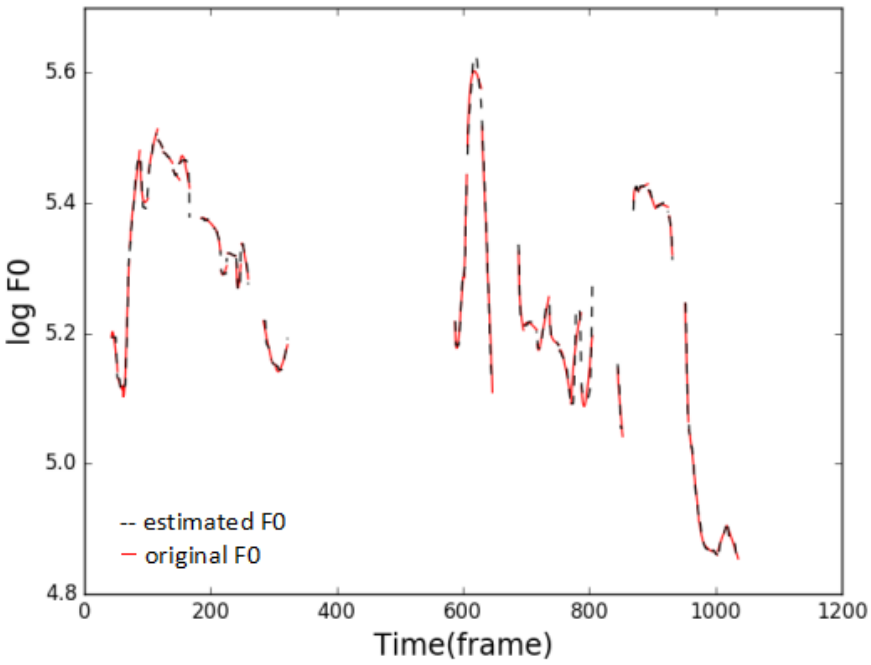


図 4.9: ピッチターゲットモデリングより近似された F0 パターンの例（赤線：近似の結果、黒線：実際の F0 パターン）

本システムでは、読み聞かせ調音声への変換が、発話時の音調意図 $T(t)$ の変化であると考え、F0 差分モデルが出力する F0 差分の時系列から、読み聞かせ用の発話意図 $T_i(x)$ を導出することを考えた。

具体的には、読みあげ調音声の F0 パターンの各モーラに対して推定されたパラメータ a, b, β, ω に対して、a と b のみを調整する。即ち、F0 差分生成モデルより予測された F0 差分の時系列に対してピッチターゲットモデルを適用し、そのパラメータ $\Delta a, \Delta b, \Delta\beta, \Delta\omega$ を得る。対応する読みあげ調音声中の当該モーラのパラメータも参照し、Δa、Δb を利用し、最終的な F0 パターンを下記で得る。

$$y(t) = \beta \exp(-\omega t) + T_i(t) \tag{4.11}$$

$$T_i(t) = (a + \Delta a)t + (b + \Delta b) \tag{4.12}$$

4.3　むすび

本章では、韻律変換を用いた読み聞かせ調音声合成システムを提案し、構築した。システムの各モジュールにおける処理を詳しく説明した。次章では、このシステムを用いて、実際に読み聞かせ調音声を合成する評価実験を述べる。最初に、各モジュールの処理性能について評価し、次に、実際の合成音声の品質を評価する。

第5章

韻律変換を用いた
読み聞かせ調音声合成の評価実験

5.1　はじめに

4 章にて、読み聞かせ調音声を合成する際の韻律制御において、深層学習により生成された韻律特徴量を用いて、目標音声の韻律へと I 韻律変換手法を提案した。本章では、100 文のテストデータを用いて、提案システムにより実際に音声を合成し、評価実験を行った。以下に評価の結果を報告する。まず、システムの各モジュールの精度を報告する。次に、合成音声を評価する。

5.2　各モジュールの性能評価

提案システムは、3 つのモジュール（読みあげ調音声合成器・韻律生成モジュール・韻律変換モジュール）から構成される。各モジュールにて、韻律変換ためのパラメータを予測し、処理を行う。これらの処理は密接に関係し、最終的の合成音声の品質に強く影響を与える。そこで、各段階の処理の性能を調べるために、各モジュールにおいて性能評価実験を行った。本節では、それぞれの精度を報告する。

5.2.1　読みあげ調音声合成器の評価

読みあげ調音声合成器は、読みあげ調音声の合成を目標とするため、読みあげ調音声を評価データとする。その性能は、合成器より生成された音響特徴量（メルケプストラム・F0）が評価データに近い結果になるか否かによって評価する。つまり、評価データとの誤差によって評価する。ここで、音声合成によく用いられるメルケプストラム歪み（式 5.1）と、F0 の二乗平均平方根誤差 RMSE（式 5.2）を用いて評価する。

$$MCD(c^{syn}, c^{ref})[dB] = \frac{10\sqrt{2}}{\ln 10}\sqrt{\sum_{i=1}^{N}\left(c_i^{syn} - c_i^{ref}\right)^2} \tag{5.1}$$

$$RMSE(f^{syn}, f^{ref}) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(f_i^{syn} - f_i^{ref}\right)^2} \tag{5.2}$$

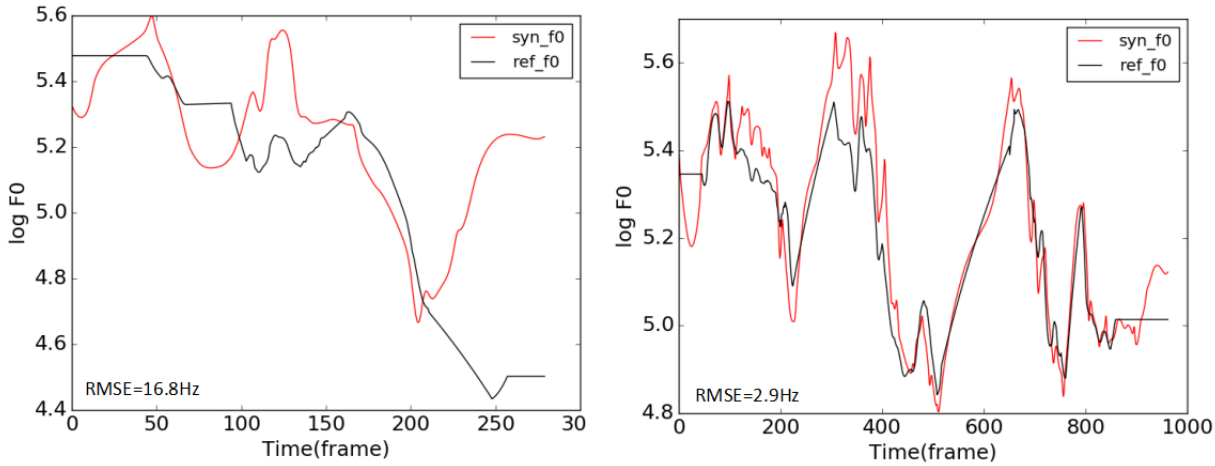
その結果を表 5.1 に示す。

表 5.1: 読みあげ調音声合成器の評価結果

MCD(dB)	RMSE(Hz)
5.1	16.7

理想的には、誤差が小さいほど、性能がいいである。ここで、韻律に強く関係する F0 に着目する。16.7 といった結果が得られたが、F0 の予測精度のバリアンスが大きいことが、個々のテストデータの精度を調べたところで分かった。図 5.1 に示しているのは、100 文のテストデータの内、F0 誤差が一番高い (RMSE=25.8Hz) と一番低い (RMSE=3.9Hz)F0 パターンと、それに対応する評価データの F0 パターンである。

図 5.1a に示しているように、個別のファイルにて F0 の予測がうまくいかなかった例があるが、読みあげ調音声におけるピッチの変化は、比較的に平らかであるため、合成器は、ほぼ期待通り



(a) F0 パターンの予測誤差が一番高い例　　(b) F0 パターンの予測誤差が一番小さい例

図 5.1:　読みあげ調音声合成器から生成された F0 パターンの例（左：誤差が一番高い例（RMSE=25.8Hz）、右：誤差が一番小さい例（RMSE=3.9Hz）、赤線：生成された F0 パターン、黒線：評価データの F0 パターン）

の性能を示した。しかし、学習データのサイズの制限がある以上、実験結果で誤差の高い F0 パターンの予測精度を引き上げることは難しい。これを解決のためには、コーパスの拡張や、複数話者の音声コーパスにより平均的な音響モデルを学習するなどが考えられるが、今回は、システムの構築を一人の話者に限定し、コーパス話者により構築されたこの読みあげ調音声合成器のもとで、検討する。

5.2.2　韻律生成モジュールの評価

i)　話速生成モデルの評価

話速生成モデルでは、読み聞かせ調音声の音素継続長を予測する。そのため、生成された音素継続長と評価データ（読み聞かせ調音声）の音素継続長との誤差によって評価する。ここで、式 5.3 で定義された継続長偏差を用い、子音と母音別々に評価した。

$$D(d^{syn}, d^{ref}) = \sqrt{\frac{1}{N}\sum_{n=1}^{N}\left(\frac{d_n^{syn} - d_n^{ref}}{d_n^{ref}}\right)^2} \tag{5.3}$$

結果を表 5.2 に示す。母音と比較し、子音の予測精度が明らかに低くなった。

表 5.2: 音素長予測精度

評価データ	子音	母音	全音素
$D(\%)$	32.2	21.6	27.1

個々の子音の精度を調べるために、音素毎に話速の予測精度を評価した。その結果を図 5.2 に示す。

子音における予測の性能が不安定であることが分かる。特に「ky」といった音素の予測率が低く、D が 50 % に至ることが確認されている。その原因を調べると「きゅうり」や「キューピー」

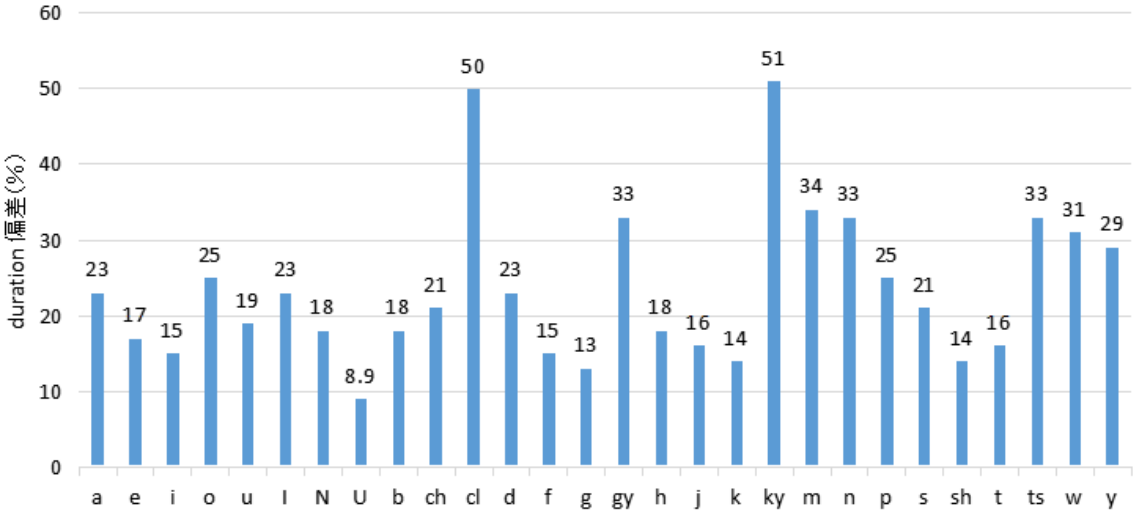


図 5.2: 各音素における継続長の予測結果

など、母音を極端に延ばして発話している例が評価データに存在しており、子音の部分も伸ばして読まれていることが原因である。各予測音素長の平均値を調べると、全体的に評価データより長い音素長を予測される傾向があることが確認されている。これも、コーパス音声の多様性の原因であると考えられる。このような多様な変化を含ませる音声に対し、話速生成モデルの性能の向上が期待される。

ii)　F0 差分生成モデルの評価

F0 差分の評価は、式 5.2 で定義される二乗平均平方根誤差 RMSE を用いた。

また、各韻律ラベルにおいて、評価を行った。その結果を表 5.3 に示す。「抑」ラベルの予測率が比較的に低くなっていた。これは、コーパスにある読み聞かせ調音声の F0 が、読みあげ調より全体的に高いため、「抑」ラベルの学習データ数が少ないことが原因であると考えられる。

表 5.3: 各韻律ラベルにおける F0 差分の予測結果

ラベル	抑	揚	BPM	キャラクター
RMSE(Hz)	4.3	3.1	2.3	3.5

5.2.3　韻律変換モジュールの評価

韻律変換モジュールでは、生成された韻律特徴量を読みあげ調音声合成器より合成された音響パラメータに適用し、読み聞かせ調音声の音響パラメータへの変換を実現する。つまり、変換後の音響パラメータは読み聞かせ調音声に対応する。このため、読み聞かせ調音声を評価データとして、変換後の話速と F0 を評価する。

i)　話速変換の評価

話速生成モデルでは、特に発声の自然性を壊しやすい子音における予測精度が低いため、話速変換では、モーラ単位で子音に制約条件を与えた上で実施すると第 4.2.4 節で述べた。そこで、話速変換の評価は、モーラ単位で時間伸縮した後の音素長と評価データとの誤差を用いて評価する。話速生成モデルの精度より高くなるか否かを確認する。ここで、話速生成モデルの評価と同様で、式 5.3 で定義された継続長偏差を用い、子音と母音別々に評価する。その結果を表 5.4 に示す。

表 5.4: 話速変換後の音素長偏差

評価データ	子音	母音	全音素
$D(\%)$	26.9	22.4	25.1

音素毎に話速変換の精度を図 5.3 に示す。

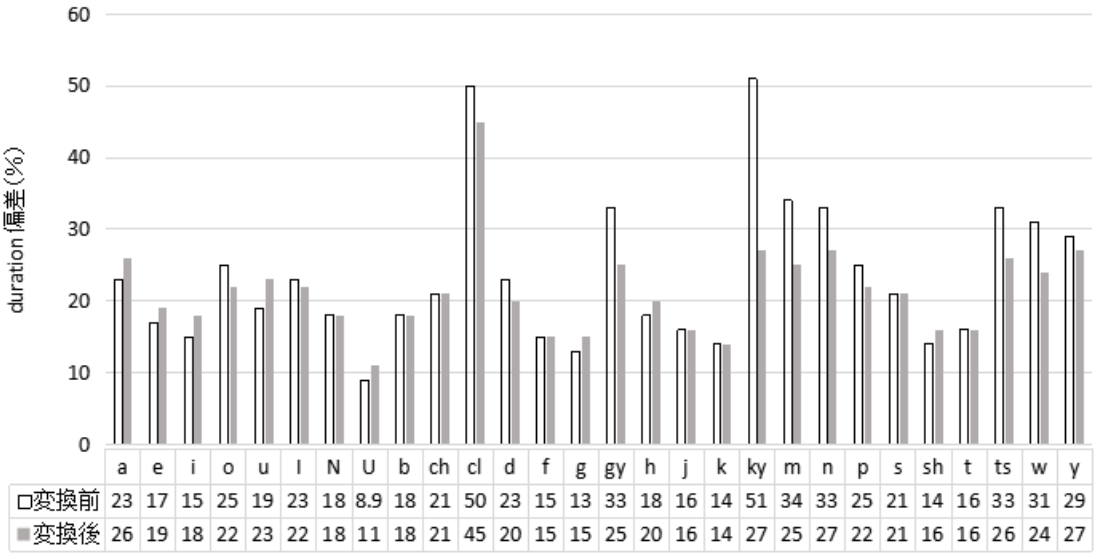


図 5.3: 音素毎に話速変換前と変換後の音素長精度の比較の図

話速生成モデルの精度（子音：32.2、母音：21.6、全音素：27.1）と比べ、母音においては、話速変換後の誤差が高くなっているが、子音および全音素においては、変換後の誤差がかなり低くなっていることが分かる。これにより、モーラ単位で制約条件を加えて話速変換の結果、全体的に音素長の精度が改善されていると考えられる。

ii)　F0 変換の評価

F0 変換の評価では、読みあげ調音声の F0 パターンから読み聞かせ調音声の F0 パターンへの変換性能を評価する。つまり、変換後の F0 パターンが変換前と目標のいずれに近いかを評価する。そのため、図 5.4 で示すように、変換前の F0 パターン（読みあげ調）と変換後の F0 パターン、そして目標となる F0 パターン（読み聞かせ調）間の相関を相互相関係数によって比較する。

どちらに近いかを相関係数の大小で判別し、それぞれのファイル総数を統計する。理想的には変換後の F0 パターンは、変換前より目標に似ることが望ましい。

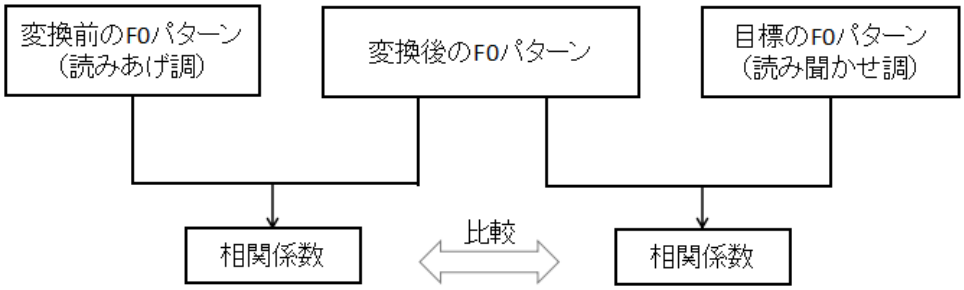


図 5.4: F0 変換の評価方法

その結果を表 5.5 に示す。

表 5.5: F0 変換の評価

	変換前に近い	目標に近い
ファイル数	27	73

結果として、評価データは変換前より目標音声に近いと評価されたファイル数が多いという結果が得られた。一方で、変換前に近いと判別されたファイルもある。そのファイルの詳細を調べると、地の文等発話スタイルの変化の少ない文が評価データに存在することが原因であると思われる。全体的に、F0 変換では、期待通りに変換していると考えられる。

なお、図 5.5 に実際の F0 の変換例を示す。上の図が変換前の F0 パターン（つまり読みあげ調音声の F0 パターン)、下の図が変換後の F0 パターン（赤）と目標音声（読み聞かせ調）の F0 パターン（青）である。

5.3　合成音声の評価

本節では、提案システムより合成された音声と第 3 章で述べた BLSTM 音声合成システムより合成された音声を比較し、評価する。評価項目は韻律制御の精度であり、つまり読み聞かせ調らしさであり、これを主観評価と客観評価によって評価した。主観評価では、韻律制御の精度を知覚的に、客観評価では F0 パターンを相関係数により評価した。以下それぞれの実施内容と結果を示す。

5.3.1　主観評価

主観評価実験は提案システムより合成された音声と BLSTM 音声合成システムより合成された音声を実際に耳で聞き比べ、評価音声の「読み聞かせらしさ」を知覚的に評価する。実験は Web ページで実施し、10 名の被験者に対して評価を行わせた。

テストデータは、両システムで統一しているように、第 2 章で述べたコーパスから学習データに含まない 100 文を選択した。聴取実験を実施するにあたって、100 文のテストデータの内、ラ

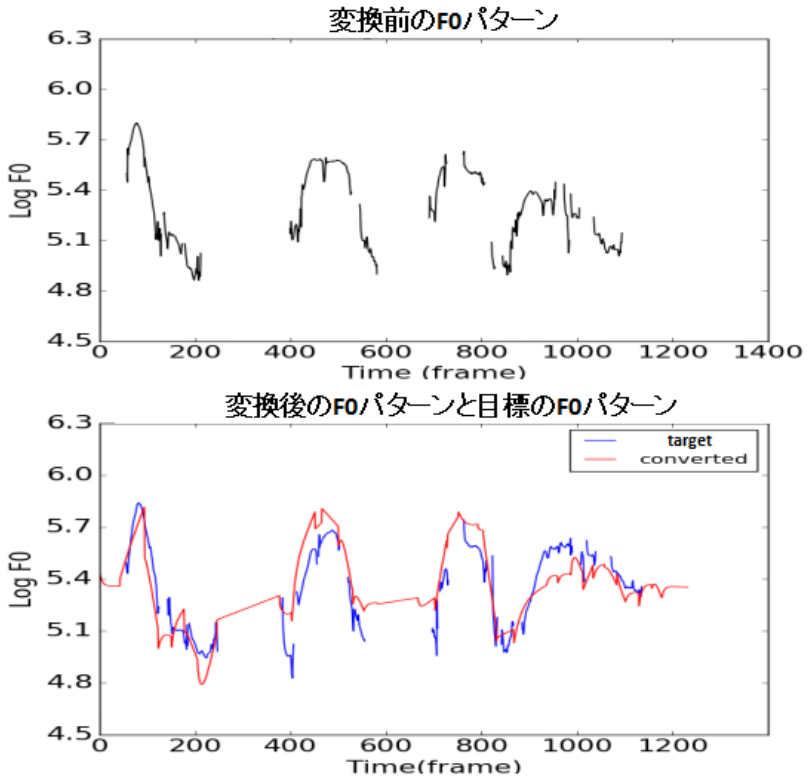


図 5.5: 「ほうら、これがセレストビルの街ですよ」を発話したときの F0 の変換例（上：変換前の F0 パターン、下：変換後（赤）と目標（青）の F0 パターン）

ンダムに 30 文を選択し、評価用データとして Web ページにて提示する。

提示する音声は、目標音声（コーパスにある人に読まれた読み聞かせ調音声）と両システムより合成された音声との三つである（どの音声も複数聴取することができる）。「どちらが読み聞かせ調らしいか」という基準で、可能な限り二つの合成音声から選択し、どちらとも判断が付かない場合は、「どちらにも似ていない」を選択するよう被験者に指示した。聴取実験の回答数は 300 であり、回答の分布を図 5.6 に示す。

主観評価では、68% の回答が韻律変換システムより合成された音声が読み聞かせらしいと判別する結果が得られた。26% の回答数である BLSTM システムと比較して、韻律変換システムの性能が上回っているといえる。

実際に両システムより合成された音声を聞くと、どの音声もある程度読み聞かせらしい読み方で合成されていることが分かる。しかし、極めて多様な変化が含まれる音声において、韻律変換を用いた手法より合成された音声は特に高精度でそれらの韻律変化を再現することができるが、BLSTM システムより合成された音声はそれらの変化を平均化されてしまう傾向がある。これにより、韻律変換を用いた手法が特に多様な韻律変化の制御においては有効であると考えられる。

一方で、韻律変換システムで失敗した例も存在する。それを確認すると、多くの場合、読みあげ調音声合成器より合成された音声の品質が良くない原因であることが分かった。また、被験者からも、一部の音声において雑音が聞こえたという意見があった。これは、音声合成器の学習は不十分であると思われる。つまり、現階段では読みあげ調音声合成器の予測結果によって読み聞かせ調音声への韻律変換の品質に差が生じると言える。従って、学習データを増やしたり、音響

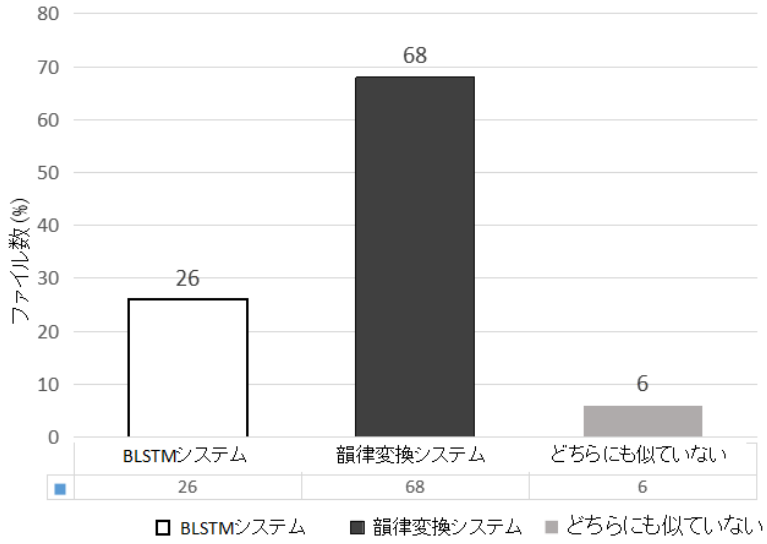


図 5.6: 主観評価の結果

モデルの構造を改善するなど音声合成器の構築における工夫が今後の課題と言える。

5.3.2　客観評価

客観評価実験は合成音声ではなく、韻律制御に関係する F0 パターンのみを用いる。テストデータは主観評価と同様である。100 文のテストデータを用いて評価した。評価方法は F0 変換の評価方法と同じで、相互相関係数を用いた。両システムより合成された音声の F0 パターンと目標音声の F0 パターン間の相関係数を計算し、どの合成音声が目標音声の韻律に近いかを判断する。その結果を図 5.7 に示す。

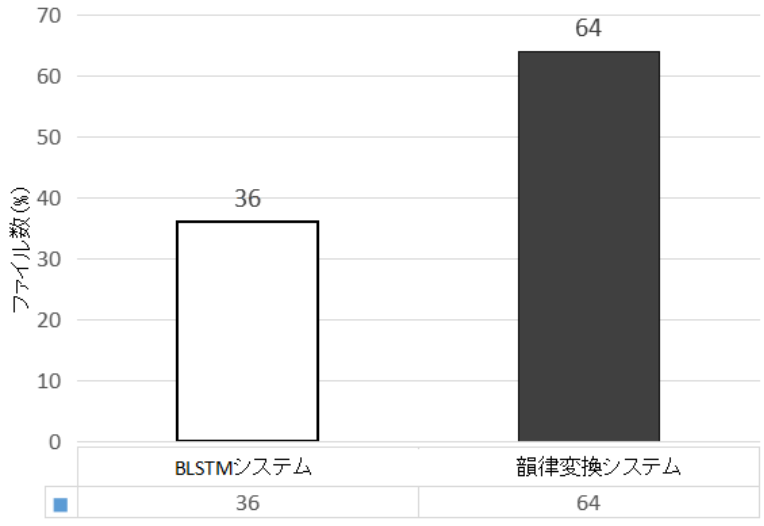


図 5.7: 客観評価の結果

また、各システムの相関係数におけるファイル数の分布を図 5.8 に示す。横軸に相関値を表し、

縦軸に各相関値のファイル数（合計 100 個）を表す。

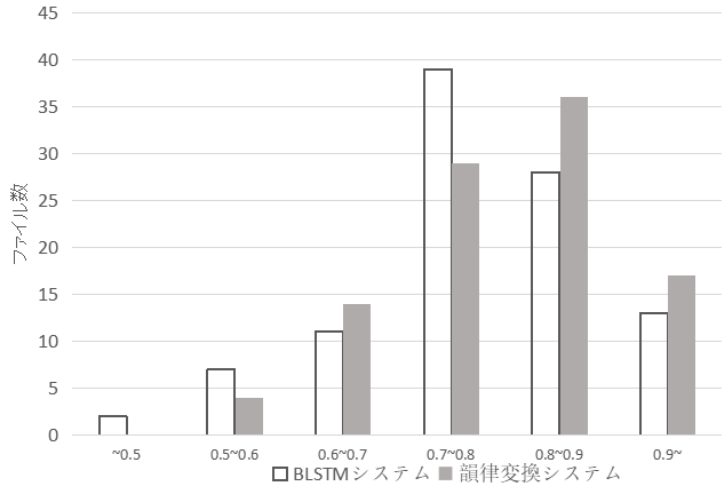


図 5.8: 各相関値におけるファイル数の分布

韻律変換システムでは半分以上の F0 パターンが、BLSTM システムでも半分程度の F0 パターンが 0.8 以上の相関を示した。相関係数の値が大きいほど予測の精度が高いであるから、この結果は両システムともある程度目標音声の韻律を制御する可能性を示したと考えられる。特に韻律変換システムは相関が 0.9 以上の F0 パターンが全体数の 5 分の 1 に達し、良好な性能に達成したと言える。

5.4　むすび

本章では、第 3 章で提案した韻律変換を用いた読み聞かせ調音声合成システムの評価実験を行った。まず、システムの各モジュールの性能を評価し、最後に、BLSTM システムと比較し、合成音声の評価を行った。主観評価実験と客観評価実験を実施した結果、韻律変換システムが、読み聞かせ調音声を合成する際の韻律制御にあたって、特に多様な韻律変化を含まれる音声において、BLSTM 音響モデルより有効であることを示した。

第6章

結論

6.1 まとめ

本論文では、聞き手との関係性を反映できるIDS風の音声を取り上げ、特に絵本を読み聞かせるような音声の合成を目指した。そのため、発話スタイルに強く関係する韻律特徴量に着目し、読み聞かせ調音声を合成する際の韻律制御に関する検討を行った。

検討のため、読みあげ調音声と読み聞かせ調音声の両スタイル音声を収録されているパラレルコーパスを用いた。読みあげ調音声と比較し、読み聞かせ調音声に特有な韻律特徴（長音化、抑揚の変化、上昇調BPM、キャラクター属性など）を分析し、これらの韻律を制御可能であることを確認した。

韻律制御にあたって、最初に、BLSTM音響モデルに基づく方法を実験した。読み聞かせ調音声の韻律特徴をコンテキストラベルに追加し、言語的特徴量と一緒にBLSTM音響モデルの入力として用いられることで、合成音声の韻律を制御する実験を行った。その結果、意図せぬ箇所でピッチが変動するなど、不自然な音声となることがあった。これは読み聞かせ調音声にける多様な韻律変化は、音響モデルより平均化されてしまうおそれがあると考えられる。

そこで、言語情報の伝達に重きを置く音響モデルを用いてテキストから読み聞かせ調音声を合成するではなく、読みあげ調音声に対し、声質や話者性を変えずに、目標発話スタイルへと韻律（話速、F0）だけを変換する手法を提案した。

提案システムは、読みあげ調音声合成器、韻律生成モジュール、韻律変換モジュールから構成される。読み聞かせ調音声の実現は、読みあげ調音声合成器から合成された音声を、韻律生成モジュールから生成された韻律（話速・F0）を用いて、目標発話スタイルへと変換することである。なお、韻律変換にあたって、話速の変換において、モーラ単位で子音に制約条件を与えた上での実施の有効性が確認された。また、F0の変換において、ピッチターゲットモデルを導入することで、局所的にF0が大きく変動することの改善が見られた。

さらに、提案システムとBLSTMシステムの韻律制御の性能を、主観評価実験と客観評価実験によって示した。その結果、BLSTMシステムと比べ、提案システムは高い性能で読み聞かせ調音声の韻律を制御する可能性を示した。特に、多様な変化を含まれる音声において有効であることが分かった。

6.2　今後の課題

今回の実験では特に長音化、抑揚の変化、上昇調 BPM、キャラクター属性等の韻律特徴に注目し、これらの特徴における韻律制御について検討した。読み聞かせ調音声の独特な韻律特徴は他にも考えられる。特にパワーと言った韻律特徴量は今回の検討に含まれていなかったが、発話内容によってパワーの変動に関する分析・制御が期待できる。また、今回は韻律制御に用いられた韻律ラベルがラベラーにより手動で付与されたものである。このような読み聞かせ用の韻律ラベルを自動的にテキストから予測することは今後の課題となっている。さらに、読み聞かせ調音声の評価基準においては、今回はコーパスにある人に読まれる音声を参考音声として評価実験を行われたが、発話スタイルには正解がないため、今後は発話スタイル評価に特化した指標が期待される。

謝辞

まず本研究を進めるにあたり、指導教員である峯松信明教授には多大なご指導ご鞭撻を頂きました。また、論文の執筆において熱心にご指導をいただいたこと、心より深く感謝いたします。

また、不自由のない研究生活のために日頃の研究活動を支えてくださった高橋登技術専門員、秘書の池上恵さん、折茂結実子さんにも、感謝の意を表します。

さらに、齋藤大輔講師、博士課程の趙イ氏には、研究方針や実装方法など何か困った時にいつでも相談させていただけたお陰で、この研究がここまでくることができました。様々な御助言をいただき、心から感謝いたします。

峯松研究室の皆様には、修士 2 年間の研究生活において大変お世話になりました。研究はもちろんのこと、様々な面でお世話になった研究室の方々に感謝いたします。快く充実した大学院生活を送ることができたのは皆様のお陰です。本当にありがとうございました。

最後に、学生生活を支えてくださった家族に心より感謝の意を表します。

2018 年 1 月 31 日

尤秀

参考文献

[1] 能勢，“統計モデルに基づく音声合成における話者･スタイルの多様化”，信学技報SP112(422)，67–72, 2013.

[2] 百武他, “絵本読み聞かせ風音声合成ためのコンテキストラベル設計に関する実験の検討”, 信学技報 SP115(521), 255–260, 2013.

[3] J.Hunt et al., “Unit selection in a concatenative speech synthesis system using a large speech database,” Proc. IEEE International Conference on Acoustics, Speech, and Signal Processiong (ICASSP), 1520–6149, 1996.

[4] H. Zen et al., “Reformulating the HMM as a trajectory model by imposing explicit relationships between static and dynamic feature vector sequences,” Computer Speech & Language, 153–173,2006

[5] T. Yoshimura et al., “Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synsthesis,” Proc. Eurospeech, 2347–2350,1999.

[6] K. Tokuda et al., “Speech synthesis based on hidden Markov models,” Proc. IEEE 1234–1252, 2013.

[7] O. Karaali et al., “Speech synsthesis with neural networks,” World Congress on Neural Networks, 45–50,1998.

[8] 大野他，” 韻律変換実現ための一試行：高橋みなみ風の音声を小嶋陽菜風に変えてみた ” エンタテインメントコンピューティングシンポジウム 2015 論文集，483–486，2015.

[9] 西海枝他，“『理研母子会話コーパス (R-JMICC)』構築の試みと研究成果 –対乳児自発音声における日本語特有の韻律的・分節的特徴の解明を目指して–,” 第 3 回コーパス日本語学ワークショップ論文集, 383–392, 2013.

[10] Z.Chen et al., “An investigation of implementation adn performance analysis of DNN based speech synthesis system,” in Proc. ICSP, 577–582, 2014.

[11] Y.Fan et al., “TTS synthesis with bidirectional LSTM based recurrent neural networks,” in Proc. Interspeech, 1964–1968,2014.

[12] T. Saito et al., ”Speech-to-singing synthesis: converting speaking voices to singing voices by controlling acoustic features unique to singing voices,” Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 215–218, 2007.

[13] K. Hashimoto et al., “The effect of neural networks in statistical parametric speech synthesis,” in Proc. ICASSP, 4455-4459, 2015

[14] J. Tao et al., “Prosody conversion from neutral speech to emotional speech,” Proc. IEEE Transactions on Audio, Speech, and Language Processing, 1145–1154, 2006.

[15] http://www.synsig.org/index.php/Blizzard_Challenge_2017

[16] H. Zen et al., “Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis,” Proc. IEEE International Conference on Acoustics, Speech, and Signal Processiong (ICASSP), 4470–4474， 2015.

[17] Y.Xu et al., “Pitch targets and their realization: Evidence from mandarin Chinese,” Speech Communication, 33, 31-337, 2001.

[18] J. Tao et al., “Emotional speech generation by using statistic prosody conversion methods,” Affective Information Processing, 127-141,2009.

[19] A. Ryo et al., “Gmm-based emotional voice conversion using spectrum and prosody features,” American Journal of Signal Processing, 134-138, 2012.

[20] Y. Kang et al.,’ “Applying pitch target model to convert F0 contour for exressive mandarin speech synthesis,” Proc. IEEE International Conference on Acoustics, Speech, and Signal Processiong (ICASSP), 1520-6149, 2006.

[21] X.Sun, “The determination, analysis and synthesis of fundamental frequency,” Ph.D. dissertation, Univ.Reading, Reading, U.K., 2001.

発表文献

国内研究会・全国大会

[1] 尤秀, 峯松信明, 齋藤大輔, 趙イ “深層学習に基づく韻律生成を用いた読みあげ調から読み聞かせ調への韻律変換に関する検討” 日本音響学会春季講演論文集, 2018. (発表予定)