修士論文

絵本読み聞かせ風音声合成のための コーパス構築と コンテキストラベルの設計

2016 年 2 月 4 日

指導教員　峯松 信明 教授

電気系工学専攻

37-146477　百武 恭汰

内容梗概

近年，文字情報を音声情報に変換する音声合成技術は，スマートフォンやコミュニケーションロボットの普及により活躍の場を広げている．そうした中で，アナウンサーがニュース原稿を読むようにテキストを単調に読み上げるだけでなく，多様で表現豊かな読み方・話し方を実現する音声合成が求められている．表現豊かな音声合成として，これまでに喜びや怒りなどの感情が込められた感情音声の合成に関する研究が多数なされている．

ところで，我々が実際に話す時には，感情のみならず，話す相手との関係性によっても話し方を変えている．前段で挙げた感情音声の合成はを expressive な音声合成と呼ばれるが，それに対して相手との関係性を反映した音声合成は social な音声合成と言うことができ，これを多様な音声合成を実現するための軸の一つと捉えることができる．

相手によって話し方が変わる例の一つとして，子ども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．その特徴についてはこれまでに様々な分析研究が行われており，数多くの知見が得られている．

IDS 風の音声を合成することができれば，コミュニケーションロボットが子どもにとってより親しみやすくなるなどの効果が期待される．また，IDS 風の合成音声が実現した場合に活躍が期待される場面として，絵本の読み聞かせが考えられる．そこで，本研究では，子供に向けた話し方の中でも特に読み聞かせに焦点を当てた検討を行った．

絵本読み聞かせ風の音声合成を実現するにあたって，近年広く用いられている HMM 音声合成で実現するためには，読み聞かせ音声のコーパスが必要となる．既存のコーパスで，条件を満たし，かつ利用可能なものはこれまでに存在しないため，本研究では新たに読み聞かせ音声コーパスの構築を行った．

構築したコーパスは女性保育士による絵本の文章の単調な読み上げ音声と読み聞かせ音声とを収録したパラレルコーパスである．話者の選定にあたっては，保育士から立候補を募り，各立候補者の読み聞かせ音声を用いたアンケートで上位に選ばれた者を選定した．また，収録用の文章には，音素バランス性や登場するキャラクターのバリエーションなどを考慮して選定した絵本 7 冊の文を用いた．

収録した音声を HMM 音声合成に用いるため，ポーズやアクセント情報などのラベリングを行った．ここで，読み聞かせ音声特有の読み方の工夫を合成音声において制御するため，通常の HMM 音声合成において必要となるアクセント情報のみではなく，読み聞かせ音声に特徴的に見られる現象に関しても着目しつつラベリングを進め，ラベリング項目の追加を行った．この結果，アクセント句末のピッチ上昇や長音化，抑揚の程度，緩急の変化といった現象が確認され，これらの現象に対してもラベルの付与を行った．

このようにして構築したコーパスを利用して，絵本読み聞かせ風音声の合成を目指し，HMM 音声合成に関する検討を行った．その結果，ポーズ情報のラベリングの不備が疑われ，機械によ

るポーズの自動検出によってこの問題が改善すること，および初期モデル学習のための時間情報の不備が疑われ，フラットスタート法の採用によりこの問題が改善することが確認された．

さらに，読み聞かせ音声の特徴を合成音声において再現，制御するために．コンテキストラベルの設計を行った．句末音調や長音化，抑揚の程度，緩急，セリフか地の文か，セリフのキャラクター属性といったコンテキストをコンテキストラベルに導入して音声合成を行った結果，読み聞かせ音声に特徴的な現象を合成音声において制御できることが確認された．この合成音声による主観評価実験を行ったところ，読み聞かせ用コンテキストを導入することで合成音声がより「読み聞かせらしい」と感じられる傾向が見られた．一部のコンテキストに関しては読み聞かせ音声の特徴を良く表現しきれていないおそれがある様子ではあったが，今回設計した読み聞かせ用コンテキストラベルの有効性が確認された．

目次

図目次

表目次

第1章

序論

1.1　本研究の背景

近年，文字情報を音声情報に変換する音声合成技術は，スマートフォンやコミュニケーションロボットの普及により活躍の場を広げている．そうした中で，アナウンサーがニュース原稿を読むようにテキストを単調に読み上げるだけでなく，多様で表現豊かな読み方・話し方を実現する音声合成が求められている．表現豊かな音声合成の代表的な検討例としては，喜びや怒りなどの感情が込められた感情音声の合成が挙げられる [1] [2].

ところで，我々が実際に話す場面を振り返ると，感情によって話し方を変えるだけでなく，目上の人には敬語を使って丁寧な口調で話すなど，相手によっても話し方を変えていることが伺える．前段で挙げた感情音声の合成は expressive な音声合成と呼ばれるが，それに対して相手との関係性を反映した音声合成は social な音声合成と言うことができ，これを多様な音声合成を実現するための軸の一つと捉えることができる．

相手によって話し方が変わる例としては目上/目下による変化の他にも，子ども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．IDS は幼児の言語獲得と関係があるのではないかと考えられているため，その特徴についてこれまでに様々な分析研究が行われている．こうした分析研究の結果，IDS には一般にピッチが高くなりそのレンジが広がる，ポーズが多く，長くなるなど，様々な特徴が見られるという知見が得られている [3].

IDS 風の音声を合成することができれば，コミュニケーションロボットがより子どもの興味を引きつけ，より親しみやすくなるなどの効果が期待される．また，IDS 風音声合成が実現した場合に需要の想定される場面として，絵本の読み聞かせが考えられる．そこで，本研究では，子供に向けた話し方の中でも特に読み聞かせに焦点を当てて検討を進めることとする．

読み聞かせ風音声の合成を実現するには，読み聞かせ音声を収録したコーパスと，読み聞かせ音声における読み方の工夫を表現するコンテキストラベルが必要となる．そのため，本研究ではコーパスの構築，および読み聞かせ用コンテキストラベルの設計を行うことを通して読み聞かせ風音声合成の実現を目指すこととする．

1.2　本研究の目的

聞き手との関係性を反映した話し方を合成音声により実現するため，特に分析研究による知見の豊富な IDS 風の音声合成を目指す．その中でも特に需要が期待される，絵本を読み聞かせるような音声の合成を目標とする．そのために必要となる読み聞かせ音声のコーパスを設計・構築し，読み聞かせ音声の特徴を合成音声に反映し，制御するために必要となるコンテキストラベルに関しても設計を行う．このコーパスとラベルを用いることで，読み聞かせ特有のラベルが付与されたテキストから，ラベルで示された特徴が適切に再現された音声を合成するモジュールを開発することを目指す.

1.3　本論文の構成

本論文は全 5 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，本研究との関連事項として，日本語のアクセントに関する知見，IDS に関する知見，お

よび HMM 音声合成システムについて述べる．第 3 章では，絵本読み聞かせ風音声合成のための絵本読み聞かせ音声コーパスの構築，および収録音声に対するラベリングについて述べる．第 4 章では，絵本読み聞かせ風音声の合成に向けた HMM の学習条件に関する検討やコンテキストラベルの設計，および設計したコンテキストラベルを用いた主観評価実験について述べる．最後に第 5 章で本論文をまとめ，今後の展望について述べる．

第2章

本研究との関連事項

2.1　はじめに

本章では，本研究と関連の深い事項として，IDS に関する知見，および HMM 音声合成システムに関して述べる．また，IDS の特徴として，特徴的な韻律制御がなされていることを述べるため，IDS に関する知見に先立って，アクセントなどの一般的な日本語の韻律制御についてまとめることとする．

2.2　日本語発声における韻律制御

2.2.1　アクセントとピッチ

語句の一部を他の部分と比べて音韻的に際立たせることをアクセントと呼ぶ．アクセントはその「際立たせ方」によって分類される．例えば英語では音の強弱が用いられており，このようなアクセントは強弱アクセント，あるいはストレスアクセントと呼ばれる．一方，日本語では音の高低が用いられており，これは高低アクセント，あるいはピッチアクセントと呼ばれる．

ピッチとは声の高さを表すパラメータであり，声が高くなるほどピッチは上がる．ピッチそのものは心理量であるが，音声の基本周波数（F_0）とほぼ対応しており，習慣的に基本周波数がピッチと呼ばれることもある．

2.2.2　単語アクセント

日本語のアクセントではリズムの基本単位であるモーラ（拍）ごとにピッチの高（High; H）と低（Low; L）を切り替えている．モーラは俳句や短歌の「5・7・5」や「5・7・5・7・7」を数える際の単位にあたるもので，一般にひらがな 1 文字が 1 モーラに対応している．ただし,「きゃ」「きゅ」「きょ」のような拗音など，2 文字で 1 モーラを形成するケースもある．

東京方言では，単語を発声する際は，原則として単語の先頭でピッチが L から H へと上がり，どこか 1 か所で H から L に下がり，その後は末尾まで L のままとなる．この高さが H から L へと下がる直前のモーラをアクセント核と呼ぶ．Fig. 2.1 は単語中のモーラごとのピッチの L,H を示したものである．ここに示されている「ひこうき」という単語では一般に「こ」がアクセント核になる．ただし,「おはなみ」のようにアクセント核がなく H のまま終わる単語も存在する．また，ここに示した例は東京方言のものであるが，アクセント核の位置は方言によって変化することがある．

2.2.3　アクセント句

前節では単語アクセントについて触れたが，実際の発話では単語単体ではなく複数の単語が連なった文として発声される．この時，単語アクセントで見られた「L から H に上がり，再び L に下がる」というピッチ変動の流れはしばしば複数の単語をまたいで現れる．したがって，アクセントという観点からは，この現象が見られる連続した複数の単語を一つのまとまりとして捉えることができる．このまとまりをアクセント句と呼ぶ．定義上，一つのアクセント句には 2 つ以上のアクセント核は存在しないことになる．

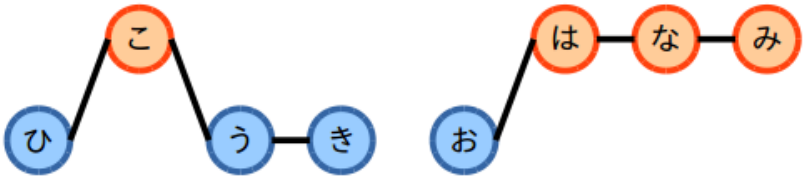


図 2.1: 日本語単語アクセントの例

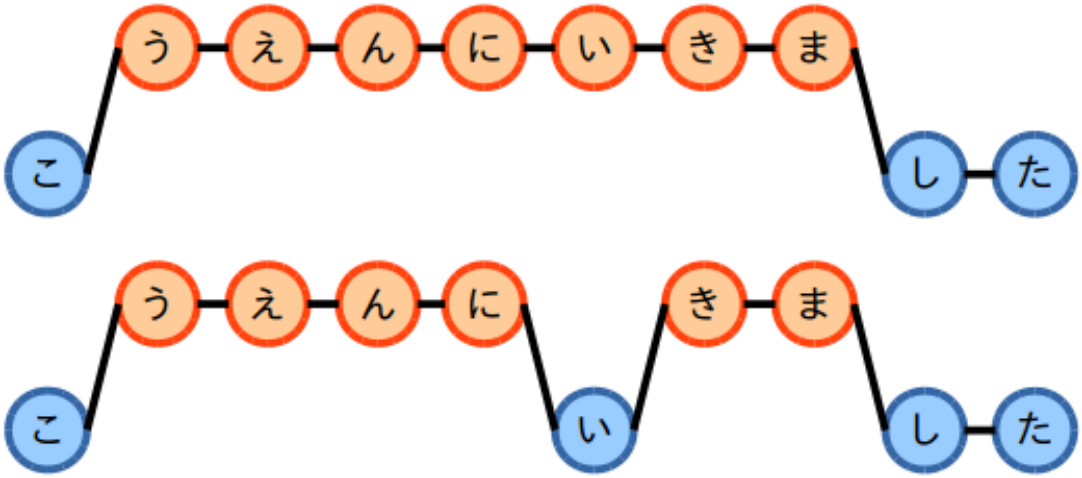


図 2.2: 日本語の文中のアクセントの例

ここで,「公園に行きました」という文を発話する場合について考えると，話し方，特に話す速度を変えることで Fig. 2.2 のように 2 通りのピッチ変動が考えられる．このことは，アクセント句の区切りが，文節のような単位とは異なり，テキストに対して一意ではないことを示唆している．図中の上の例では文全体で 1 つのアクセント句を形成しており，下の例では 1 文が「こうえんに」と「いきました」の 2 つのアクセント句に分かれていることになる．

2.2.4　イントネーション句（フレーズ）

発話中のピッチの変動には，これまで述べた以外にも長時間を通して緩やかに下降していく成分が含まれている．この緩やかな下降によるまとまりをイントネーション句（フレーズ）と呼ぶ．イントネーション句は 1 つあるいは複数のアクセント句からなっている．実際の発話にはアクセントによるピッチ変動も加わるため，イントネーション句の境界は判別が難しいケースもあるが，読点が置かれていてポーズ（休止）が入る部分などがフレーズ境界となることが多い．

アクセントとイントネーションを統合的に扱い，音声の F_0 パターンを表現するモデルとして、基本周波数パターン生成過程モデル（藤崎モデル）がある [4]．藤崎モデルでは、対数軸上での F_0 パターンをフレーズ成分とアクセント成分、および基底周波数成分の足し合わせによって近似的に表現して

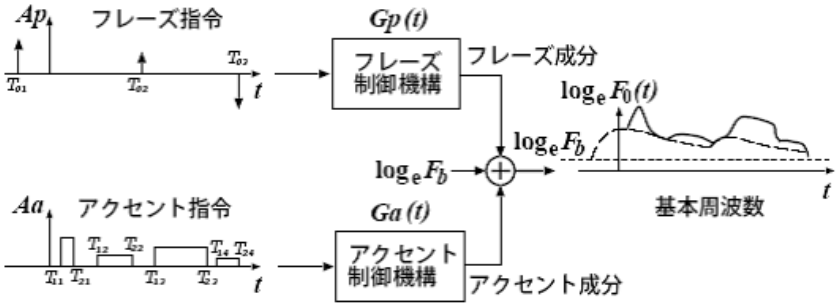


図 2.3: 基本周波数パターン生成過程モデル（藤崎モデル）

いる。その概要はFig. 2.3のようになっており、フレーズ成分（PHRASE COMPONENTS）とアクセント成分（ACCENT COMPONENTS）はそれぞれフレーズ指令（PHRASE COMMANDS）、アクセント指令（ACCENT COMMANDS）に対する応答として表現されている。

フレーズ成分は句頭から句末に向けてゆるやかに下降していく成分であり、インパルス応答として表現される。アクセント成分は個々の単語や連続した単語に付随して現れる F_0 の局所的な変動に対応する成分であり、ステップ応答として表現される。基底周波数成分は話者に固有な成分であり、定数値となっている。

2.2.5　句末境界音調

日常会話などの発話中では，アクセント句末で局所的なピッチ変動が観測されることがある．こうした音調は句末境界音調と呼ばれている [5]．句末境界音調は，単純にピッチが下降してアクセント句が終わる単純境界音調と，句末で上昇を伴う複雑なピッチの変動が見られる複合境界音調（boundary pitch movement; BPM）とに大別される．

BPM を伴った F_0 パターンの例を Fig. 2.4 に示す．BPM は主に上昇調 1（H%），上昇調 2（LH%），上昇下降調（HL%），上昇下降上昇調（HLH%）によって表現することができる．これらの音調は質問，強調などの意味や発話意図の伝達に重要な役割を果たしている．BPM 以外の箇所のピッチパターンは語彙情報による制約を受けるため韻律的強調による変化は限定的になり，一方でそうした制約のない BPM は意図などが表現されやすいと考えられている．

2.3　対乳児発話（IDS）

2.3.1　対乳児発話とは

主に母親が子供に話しかける際に見られる特徴的な発話をマザリーズ，あるいは対乳児発話（IDS）と呼ぶ．これに対し，大人に向かって話しかける際の，いわば普通の発話は ADS（Adult-Directed Speech）と呼ばれる．IDS に見られる特徴には言語によらず普遍的に見られるものと言語に依存して見られるものとがある．こうした特徴は乳幼児の言語の獲得と深い関係があるのではないかと考えられており，IDS の分析研究によりこれまでに様々な知見が得られている．

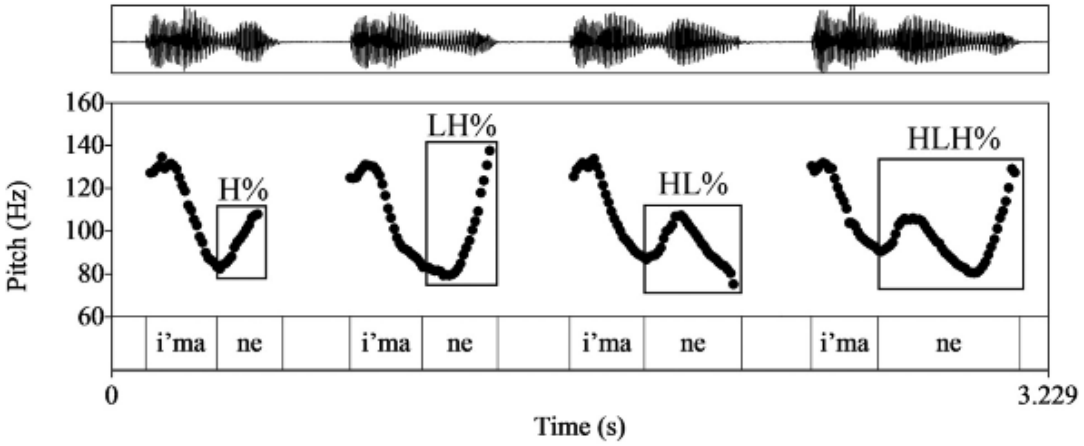


図 2.4: BPM を伴う音声の波形と F_0 パターン（[6] より）

2.3.2　多言語で見られる特徴

IDS において，日本語や英語，フランス語，イタリア語，ドイツ語で共通の特徴が存在することが確認されている [7]．これらの言語で共通して見られる特徴としては，F_0 の値が全体的に高いこと，発話が短いこと，ポーズ（休止）が長いことなどが挙げられる．また，F_0 については，全体的な上昇に加え，レンジも拡大することが確認されている．

2.3.3　日本語 IDS 固有の特徴

日本語 IDS の特徴については，理化学研究所の言語発達研究チームが『理研母子会話コーパス（R-JMICC）』の構築を進めており，このコーパスに対し様々な分析を行っている [8]．その結果，日本語 IDS にはこれまで他の言語で報告されてきたものとは違った特徴が見られることが確認されている．

まず，発話速度に関しては，過去に IDS では通常より遅いとする研究があったが [9]，実はそうではないとする研究結果が出ている [10]．一般に，発話速度は単位時間あたりのモーラの数によって定義されており，モーラの継続長が長いほど発話速度は遅いということになっている．この研究では，IDS，ADS を問わず，イントネーション句の最後ではモーラの継続長が長くなる final lengthing が観測されることに着目している．イントネーション句の最終モーラとそれ以外のモーラに分けて継続長の平均をとると，IDS，ADS の両方で最終モーラが非常に長くなったが，IDS と ADS の間では有意差が見られないという結果が出ている．IDS では発話が短いため，final lengthing の起きているモーラの数が相対的に多くなり，そのことによって IDS でモーラ継続長の平均が長くなっているだけ，とされている．

また，日本語 IDS の韻律的特徴として，「おそさがり」現象が挙げられる [11].「おそさがり」とは，アクセントがあると知覚される音節よりも後にピッチの最高点が存在し，ピッチの「さがり」が「おそ」くなる現象である.「ママ」という単語について，「おそさがり」を伴う音声と伴わない音声の F_0 パターンを図 2.5 に示す.「おそさがり」を伴わない右の場合では F_0 は 1 モーラ目で最高値をとっているが，「おそさがり」を伴う左の場合では 2 モーラ目で最高値をとっている.「おそさ

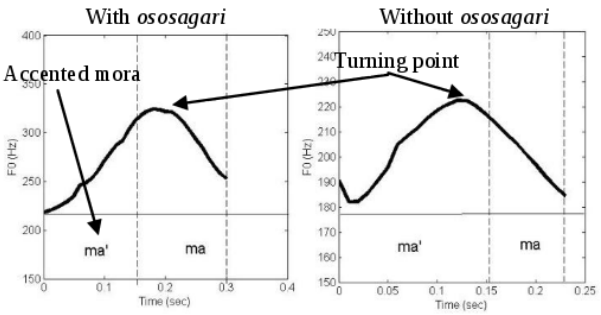


図 2.5: 「おそさがり」を伴う音声と伴わない音声の F_0 パターン（[11] より）

がり」の現れる合成音声はより「女性らしい」聴覚印象を与えるという報告もなされている [12]. IDS においては，ピッチの上限が大きく上昇することでピッチレンジが拡大しており,「おそさがり」はピッチ立ち下がりの遅れのみならず最高点の上昇と相関を持って現れるという仮説が提示されている [13]．なお，ピッチレンジの拡大に関しては，BPM の部分で顕著に見られ，それ以外の部分ではほとんど見られないという分析結果が出ている [8].

さらに，分節的特徴として，母音のバリエーションが増えている可能性が指摘されている [14]. 模擬講演音声では極端な調音（hyper articulation）による母音の弁別性や明瞭性を高める特徴が見られることが知られているが，IDS ではこれとは真逆の特徴が見られており，決して明瞭な音声にはなっていないことが示唆されている.

2.4　HMM 音声合成

2.4.1　音声合成システム

入力されたテキストから音声波形を生成し出力するシステムは音声合成システム，あるいは TTS（Text-to-speech）システムと呼ばれる．音声合成システムには様々な方式があるが，近年用いられているものは波形接続合成と統計的音声合成の 2 つに大別される.

波形接続合成とは，音声波形そのものを分割した断片を保持しておき，それらの断片をつなぎ合わせることで合成音声とする方式である．波形接続合成は限られたテキストに対しては非常に高品質の合成音声を出力するが，発話スタイルを変えて合成することなどが困難となるほど柔軟性に欠けるという短所を持ち合わせる.

統計的音声合成とは音声から抽出されたパラメータを統計モデルを用いてモデル化し，そのモデルから生成されたパラメータを用いて音声波形を合成する方式である．統計的音声合成による合成音声の品質は波形接続合成による最高品質には劣るが，各種のモデル適応技術が容易に使えるため，話者性や発話スタイルを変えることが比較的容易であり，柔軟な音声合成が可能である.

2.4.2　HMM 音声合成の概要

HMM 音声合成 [15] は統計的音声合成の中でも特に，隠れマルコフモデル（Hidden Markov Model; HMM）により音響パラメータをモデル化する手法であり，統計的音声合成の中では最も

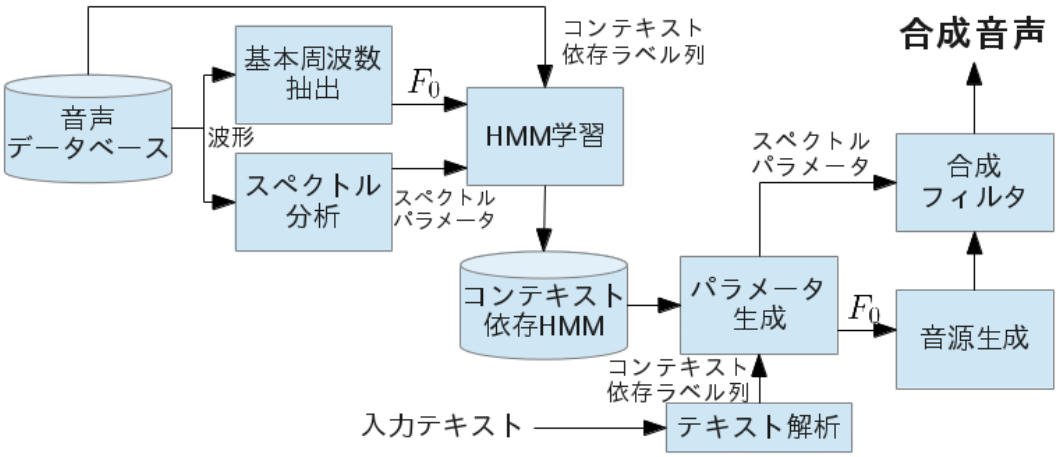


図 2.6: HMM 音声合成の概要

幅広く用いられている手法である．

その概要を図 2.6 に示す．まず学習用音声から F_0，スペクトルパラメータなどの特徴量を抽出する．各特徴量を連結したベクトルと，次節で述べるコンテキストラベルを用いて HMM を学習する．学習時にはコンテキストラベルを入力とし，HMM から尤度最大化基準で生成された特徴量を用いて音声波形を合成する．

HMM 音声合成は柔軟性に優れることから，スタイル制御や感情音声合成などの検討がなされている．

2.4.3　隠れマルコフモデル (HMM)

隠れマルコフモデル（HMM）は，不確定な時系列データのモデル化に有効な統計モデルである．音声は発声によって時間的に長さ，速さが変化し，また，現在どの音素のどの部分であるかなどの情報を音響特徴量から直接観測することができないという特性を持つため，音声のモデル化にはしばしば HMM が利用される．そのため，音声合成のみならず，音声認識でも既に広く用いられている．図 2.7 に示すのは典型的な left-to-right 型の HMM である．a_{ij} は状態 S_i から状態 S_j への遷移確率を表し，$b_i(x)$ は状態 S_i における観測シンボル x の生起確率分布を表す．HMM 音声合成では，観測特徴量 o_i が各フレームにおける音声の特徴量の連結ベクトルとなる．生起確率分布 $b_i(x)$ には多くの場合，ガウス分布に基づくものが用いられる．

ここで，入力 λ が与えられたときに，出力シンボル系列 $\boldsymbol{o} = \{\boldsymbol{o}_1, \boldsymbol{o}_2, \cdots, \boldsymbol{o}_T\}$ が観測される確率は，forward-backward アルゴリズムによって効率的に求めることができる．前向き確率を $\alpha_t(\cdot)$ とし，後向き確率を $\beta_t(\cdot)$ とし，次式のように定義する．

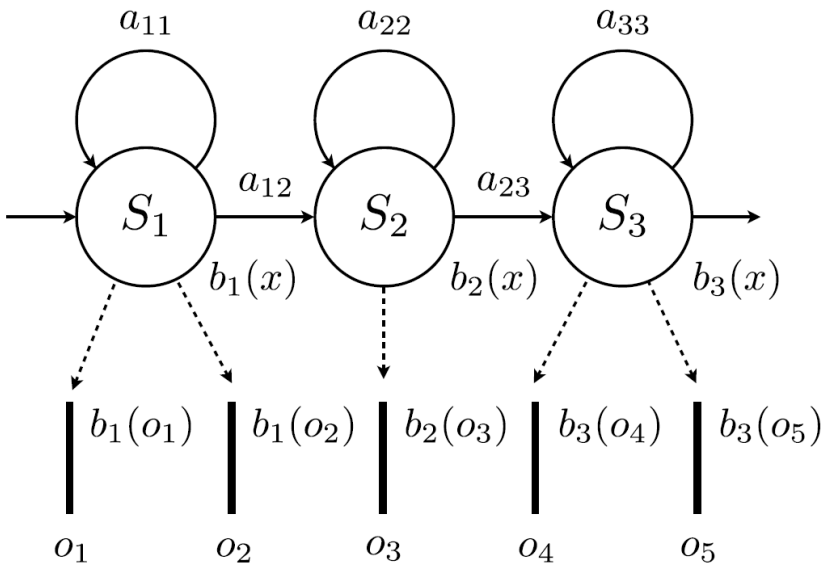


図 2.7: 隠れマルコフモデル (HMM)

$$\alpha_0(j) = \begin{cases} 1 & j = 1 \\ 0 & \text{otherwise} \end{cases} \tag{2.1}$$

$$\begin{aligned} \alpha_t(j) &= P(\boldsymbol{o}_1, \cdots, \boldsymbol{o}_t, q_t = j|\lambda) \\ &= \sum_{i=1}^{N} \alpha_{t-1}(i) a_{ij} b_j(\boldsymbol{o}_t) \quad \left(\begin{array}{l} t = 1, 2, \cdots, T \\ 1 \le j \le N \end{array} \right) \end{aligned} \tag{2.2}$$

$$\beta_{T+1}(i) = \begin{cases} 1 & i = N \\ 0 & \text{otherwise} \end{cases} \tag{2.3}$$

$$\begin{aligned} \beta_t(i) &= P(\boldsymbol{o}_{t+1}, \cdots, \boldsymbol{o}_T, |q_t = i, \lambda) \\ &= \sum_{j=1}^{N} a_{ij} b_j(\boldsymbol{o}_{t+1}) \beta_{t+1}(j) \quad \left(\begin{array}{l} t = T-1, \cdots, 1 \\ 1 \le i \le N \end{array} \right) \end{aligned} \tag{2.4}$$

ここで，$q_t = j$ は時刻 t に j 番目のステートにいる状態のことを指す．すると，入力 λ が与えられたときに，出力シンボル系列 $\boldsymbol{o}$ が観測される確率は次式によって求まる．

$$\begin{aligned} P(\boldsymbol{o}|\lambda) &= \sum_{i=1}^{N} P(\boldsymbol{o}, q_t = i|\lambda) \\ &= \sum_{i=1}^{N} \alpha_t(i) \beta_t(i) \end{aligned} \tag{2.5}$$

また，Baum-Welch アルゴリズムによって局所解を求めることで，学習データからモデルパラメータ $\theta = \{a_{ij}, b_i(\boldsymbol{x})\}$ を求めることができる．

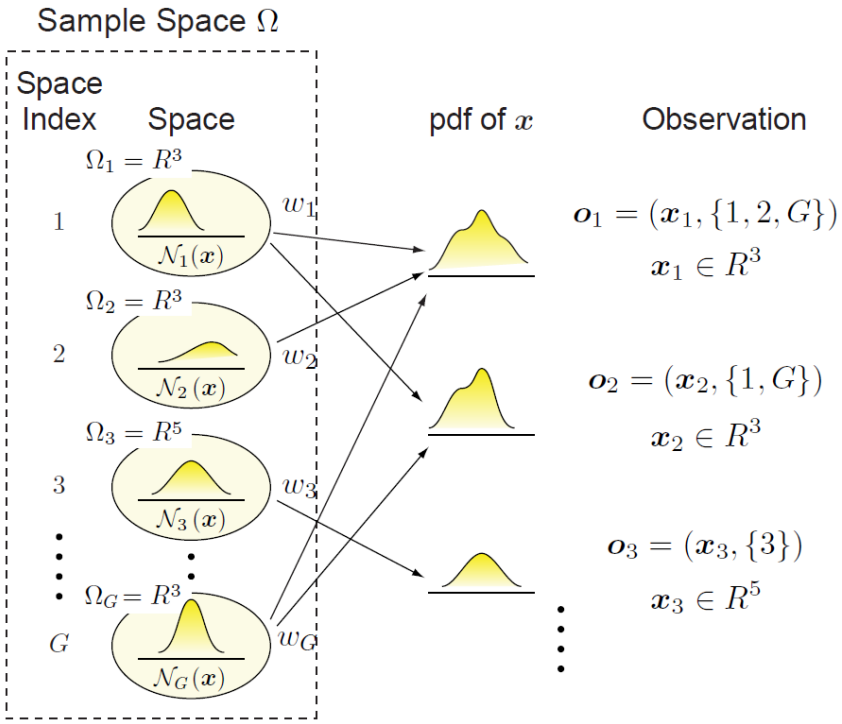


図 2.8: 多空間確率分布（[16] より）

2.4.4　多空間確率分布 HMM

F_0 は無声区間では定義されないため，通常の HMM では F_0 を取り扱うことができない．HMM 音声合成では，F_0 を取り扱うため，多空間確率分布という概念を導入している．

多空間確率分布について，図 2.8 を例に説明する．G 個の空間 $\Omega_1, \Omega_2, \cdots, \Omega_G$ からなる標本空間 Ω を考える．

$$\Omega = \bigcup_{g=1}^{G} \Omega_g \tag{2.6}$$

各空間 Ω_g は n_g 次元の実空間 R^{n_g} とする．G 個の空間の次元 n_g は，互いに異なる値でも一部が同じ値でも良い．各空間 Ω_g が採択される確率を w_g とし，各空間のもつ確率密度関数を $p^{(g)}(\boldsymbol{x})$, $\boldsymbol{x} \in R^{n_g}$ とする．ただし，

$$\sum_{g=1}^{G} w_g = 1 \tag{2.7}$$

かつ，

$$\int p^{(g)}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = 1 \tag{2.8}$$

であるとする．また，$n_g = 0$ のときは，$p^{(g)}(\cdot) = 1$ と定義する．そして，空間インデックスの集合 X とベクトル $\boldsymbol{x}$ から，$\boldsymbol{o}$ が観測されると考える．つまり，

$$\boldsymbol{o} = (X, \boldsymbol{x}) \tag{2.9}$$

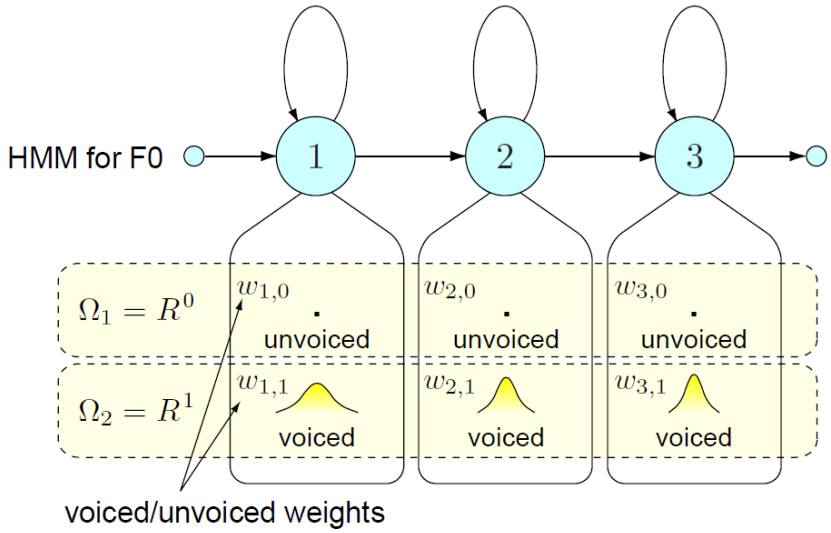


図 2.9: 多空間確率分布 HMM（[16] より）

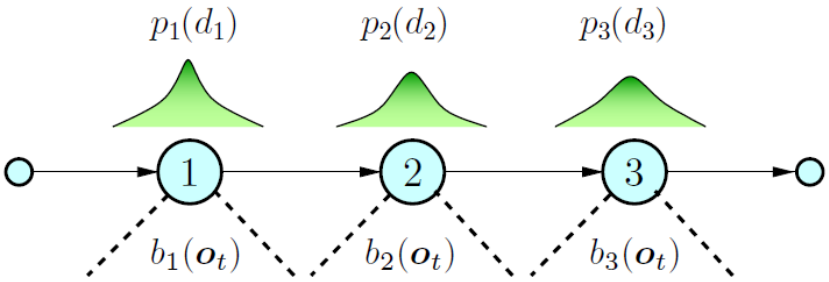


図 2.10: 隠れセミマルコフモデル (HSMM)　（[16] より）

である．ただし，X に含まれる空間インデックスが表す空間は，全て同じ次元でなければならない．このとき，$\boldsymbol{o}$ の観測確率は次式で表される．

$$b(\boldsymbol{o}) = \sum_{g \in S(\boldsymbol{o})} w_g p^{(g)}(V(\boldsymbol{o})) \tag{2.10}$$

ただし，

$$S(\boldsymbol{o}) = X, \quad V(\boldsymbol{o}) = \boldsymbol{x} \tag{2.11}$$

である．

この多空間確率分布を HMM に導入したモデルが多空間確率分布 HMM(MSD-HMM) である [17]．これにより，図 2.9 のように，有声部の値を出力する分布と，無声部の値を出力しない（すなわち 0 次元の）分布とを内包したモデルで F_0 を取り扱うことができる．

2.4.5　状態継続長モデル

HMM は音声認識で幅広く用いられてきたモデルであるが，HMM 音声合成で用いるにあたって，継続長のモデル化に問題が生じてしまう．なぜなら，遷移確率は必ず 1 以下であるため，継続

長が長くなるほど尤度が低くなってしまい，尤度最大化基準でそのままパラメータを推定すると各ステートを 1 回しか通らないパスが選択されてしまうためである．そのため，初期の HMM 音声合成では，学習時に各ステートに留まる回数をカウントしておき，別途継続長をモデル化するという事が行われていた．その後，継続長分布を明示的に含んだ隠れセミマルコフモデル (Hidden Semi-Markov Model; HSMM) を利用する手法が，全らによって提案され，広く用いられている [18]．

HSMM において，入力 λ' が与えられたときに，出力シンボル系列 $\boldsymbol{o} = \{\boldsymbol{o}_1, \boldsymbol{o}_2, \cdots, \boldsymbol{o}_T\}$ が観測される確率は，従来の HMM と同様に forward-backward アルゴリズムによって効率的に解くことができる．前向き確率を $\alpha'_t(\cdot)$ とし，後向き確率を $\beta'_t(\cdot)$ とし，次式のように定義する．

$$\alpha'_0(j) = \begin{cases} 1 & j = 1 \\ 0 & \text{otherwise} \end{cases} \tag{2.12}$$

$$\begin{aligned} \alpha'_t(j) &= P(\boldsymbol{o}_1, \cdots, \boldsymbol{o}_t, q_t = j | q_{t+1} \neq j, \lambda') \\ &= \sum_{d=1}^{t} \sum_{\substack{i=1 \\ i \neq j}}^{N} \alpha'_{t-1}(i) a_{ij} p_j(d) \prod_{s=t-d+1}^{t} b_j(\boldsymbol{o}_s) \quad \left(\begin{array}{l} t = 1, 2, \cdots, T \\ 1 \leq j \leq N \end{array} \right) \end{aligned} \tag{2.13}$$

$$\beta'_{T+1}(i) = \begin{cases} 1 & i = N \\ 0 & \text{otherwise} \end{cases} \tag{2.14}$$

$$\begin{aligned} \beta'_t(i) &= P(\boldsymbol{o}_{t+1}, \cdots, \boldsymbol{o}_T, | q_t = i, q_{t+1} \neq i, \lambda') \\ &= \sum_{d=1}^{T-t} \sum_{\substack{j=1 \\ j \neq i}}^{N} a_{ij} p_j(d) \prod_{s=t+1}^{t+d} b_j(\boldsymbol{o}_s) \beta'_{t+d}(j) \quad \left(\begin{array}{l} t = T-1, \cdots, 1 \\ 1 \leq i \leq N \end{array} \right) \end{aligned} \tag{2.15}$$

ここで，$q_t = j$ は時刻 t に j 番目のステートにいる状態のことを指す．すると，入力 λ' が与えられたときに，出力シンボル系列 $\boldsymbol{o}$ が観測される確率は次式によって求められる．

$$P(\boldsymbol{o}|\lambda') = \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{d=1}^{t} \alpha'_{t-d}(i) a_{ij} p_j(d) \prod_{s=t-d+1}^{t} b_j(\boldsymbol{o}_s) \beta'_t(j) \tag{2.16}$$

また，Baum-Welch アルゴリズムによって局所解を求めることで，学習データからモデルパラメータ $\theta = \{a_{ij}, b_i(\boldsymbol{x})\}$ を求めることも従来の HMM と同様に実現できる．

HSMM は状態継続長用に別途モデルを構築する必要がないことに加え，従来の HMM における話者適応等の技術をそのまま利用することができ，なおかつ継続長も含めた適応も容易に実現できるという利点を持つ．

2.4.6　動的特徴量とマルチストリーム化

HMM 音声合成では尤度最大化基準でパラメータを生成するため，各 HMM から出力されるパラメータは，基本的に各 HMM の生起確率分布であるガウス分布の平均値となる．しかし，このままでは出力されるパラメータは同じ HMM に所属する限り，各ステートのガウス分布の平均値となってしまう．これによりパラメータが階段状となり，不連続となる箇所が生じるため，合成音声の品質も低下してしまう．

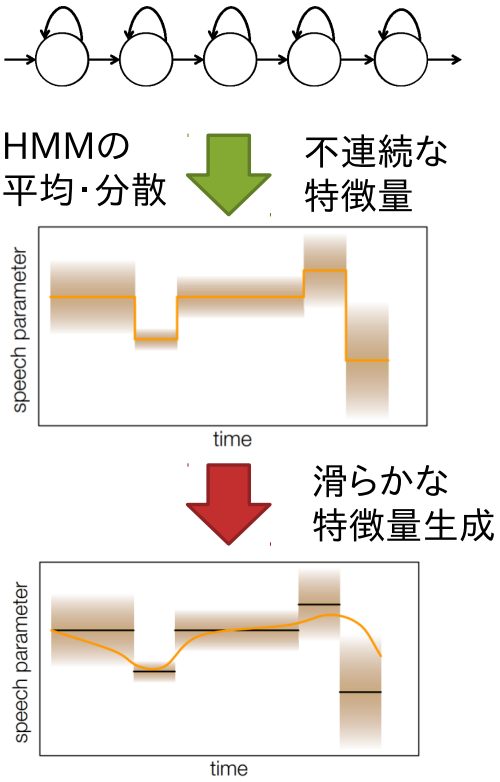


図 2.11: HMM 音声合成における特徴量生成の様子

この問題を解決するために，動的特徴量が用いられる．次式で示されるように，HMM の観測シンボルには，静的特徴量だけでなく，それらの 1 次微分，2 次微分も特徴量として加えられている．

$$\boldsymbol{o}_t = [\boldsymbol{c}_t^\top, \Delta\boldsymbol{c}_t^\top, \Delta^2\boldsymbol{c}_t^\top]^\top \tag{2.17}$$

$$\Delta\boldsymbol{c}_t = \frac{1}{2}(\boldsymbol{c}_{t+1} - \boldsymbol{c}_{t-1}) \tag{2.18}$$

$$\Delta^2\boldsymbol{c}_t = \boldsymbol{c}_{t-1} - 2\boldsymbol{c}_t + \boldsymbol{c}_{t+1} \tag{2.19}$$

これにより，パラメータ生成時に動的特徴量を制約として加えることで，滑らかなパラメータ生成を可能にしている．この様子を図 2.11 に示す．

ここで，音韻的特徴と韻律的特徴の時間的対応が取れるように，スペクトルパラメータと F_0 は同一のベクトルにまとめられているが，これらの独立性は比較的高いことが知られている．さらに，F_0 における有声/無声の判定は，空間選択の重みだけによって決定されるため，F_0 の動的特徴量はロバスト性が低くなっている．そのため，各特徴量はマルチストリーム化によってそれぞれ重みをつけた独立の分布として取り扱われている．

以上より，HMM 音声合成で用いられる特徴量ベクトルの構成は図 2.12 のようになる．ここで，$\boldsymbol{c}_t$ がスペクトルパラメータであり，p_t が F_0 である．また，観測ベクトル $\boldsymbol{o}_t$ は次式によって求め

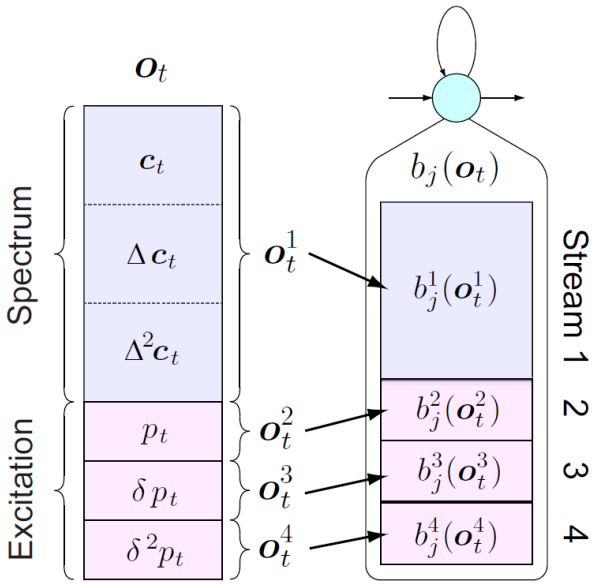


図 2.12: 特徴量ベクトルの構成（[16] より）

ることができる．

$$b_j(\boldsymbol{o}_t) = \prod_{s=1}^{S} (b_j^{(s)}(\boldsymbol{o}_t^{(s)}))^{w_s} \tag{2.20}$$

2.4.7　コンテキストラベルとクラスタリング

発話内においては，同一の音素であっても発音される状況によって異なった特徴を持つ音声となることがしばしばある．

例えば，同じ/a/という音素でも,「か」という文字を読んだ際に現れる/a/と「さ」という文字を読んだ際に現れる/a/とでは，特に音素の前半部において異なる特徴を持ちうる．こうした異なる状況下の音素が区別されず，/a/という音素が全て同一の波形が出力されると仮定すると，他の音素との接続部で不連続が発生し，合成音声の品質低下の一因となる．

そこで，現在の HMM 音声合成においては，こうした違いを区別するためのラベルが用いられている．このような状況の違いをコンテキストと呼び，コンテキストを表現するラベルをコンテキストラベルと呼ぶ．

現在広く用いられている HMM 音声合成ツールキットである HTS-2.2[1] のデモスクリプトにて用いられているコンテキストラベルに含まれる情報を表 2.1 に示す．この表に示されたコンテキストにはテキストから自動的に生成されるものも存在するが，そうでないものも存在し，韻律に関するコンテキストがこれにあたる．自動で生成できないコンテキストの情報を得るためには，2.2 節で示したアクセント核やアクセント句境界，イントネーション句境界といった情報を，実際に音声を聞いた上でラベリングする必要がある．一方で，合成用のテキストについてもコンテキストラベルに変換する必要があり，そのために韻律情報を指定，あるいは予測する必要がある．テキストからアクセント核位置やアクセント句境界などの韻律情報を自動的に予測する手法も研究

[1] HTS-2.2, http://hts.sp.nitech.ac.jp

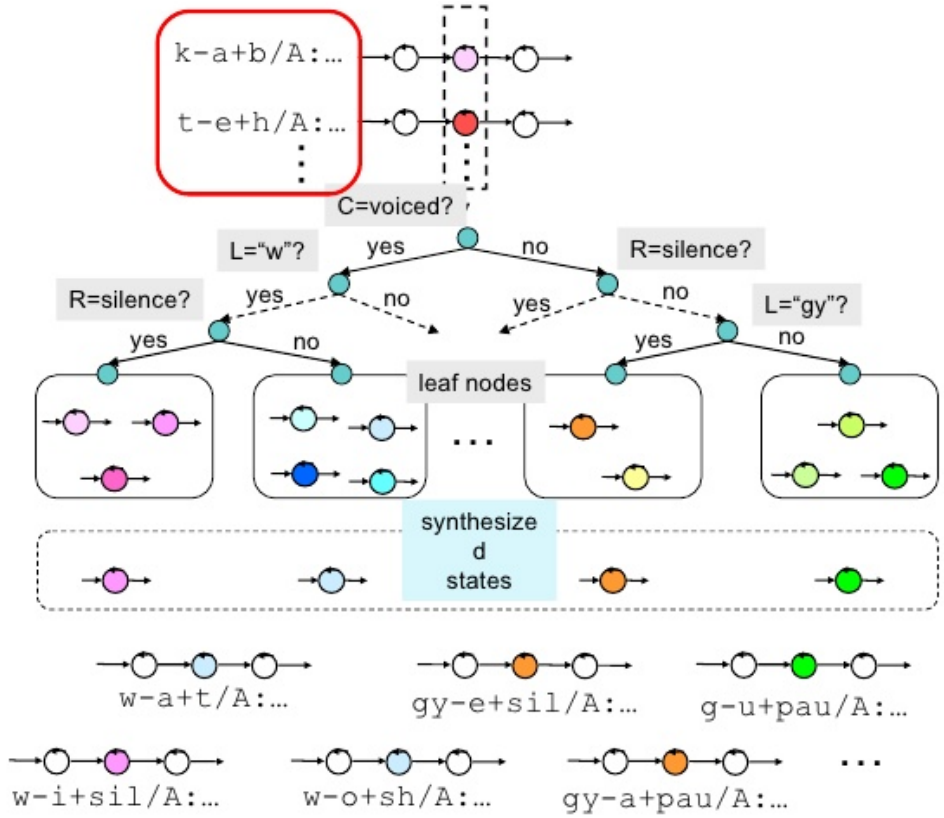


図 2.13: コンテキストクラスタリングのイメージ図

されている [20]

また，表 2.1 からは，1 種類の音素であっても，コンテキストラベルにより区別すると非常に大きな数に区別されうることが読み取れる．したがって，全てのコンテキストに対応した音声を含む学習データを準備することは不可能である．そこで，決定木によるクラスタリングを行うことで，学習データに存在しないコンテキストが入力された場合にも対応している．決定木によるコンテキストクラスタリングの様子のイメージを図 2.4.7 に示す．

HMM のクラスタ S が質問 q によってクラスタ S_{q+} とクラスタ S_{q-} に分割される際に，その分割を適応するか否かを決定する基準となるのが次式で示される最小記述長 (MDL) 基準である [21]．

$$\Delta q = \frac{1}{2}\left\{\Gamma(S_{q+})\log|\Sigma_{S_{q+}}| + \Gamma(S_{q-})\log|\Sigma_{S_{q-}}| - \Gamma(S)\log|\Sigma_S|\right\} + K\log|\Gamma(S_0)| \qquad (2.21)$$

ここで，$\Gamma(S)$ はクラスタ S に含まれる学習データの量，Σ は各クラスタの共分散行列，K は特徴量ベクトルの次元数，S_0 は決定木のルートクラスタである．

2.4.8　HTS-2.2 による HMM 音声合成の手順

HTS-2.2 のデモスクリプトにおいて，HMM の学習および音声の合成は以下の手順で行われる．事前に用意すべきデータは学習用音声ファイルと，各音声に対するモノフォンラベル（音素書き

表 2.1: HTS-2.2 デモスクリプトで用いられてい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
先行アクセント句と当該アクセント句間のポーズの有無
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
当該アクセント句と後続アクセント句間のポーズの有無
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）

起こし）およびコンテキストラベル，そして合成用のテキストと韻律情報を表現するコンテキストラベルである．まず，学習用音声とモノフォンラベルを利用し，Julius によって強制アライメントを行い，時間情報つきのモノフォンラベルを得る．次に，音声を分析して F_0 やスペクトルパラメータなどの特徴量を抽出し，学習用データセットを構成する．

時間情報つきのモノフォンラベルを利用し，Viterbi 学習によりモノフォンの初期モデルを学習し，モデルパラメータの再推定を行う．その後，時間情報を用いず，Baum-Welch アルゴリズム連結学習を繰り返し，モノフォン HMM を学習する．学習されたモノフォン HMM から，各モノフォンのモデルをコンテキストごとのモデルに変換し，再び Baum-Welch アルゴリズムによる連結学習を行う．この学習により得られたコンテキスト依存 HMM に対し，コンテキストクラスタリングを行う．なお，この時，決定木は各特徴量ごと，HMM の各ステートごとに構築される．クラスタリングがなされた後は再度連結学習を行い，クラスタリングされたモデルのパラメータを再推定する．

このようにして学習された HMM に対し，合成用のテキストおよび韻律情報から生成したコンテキストラベルを入力することで，決定木により適切なモデルが選択され，尤度最大化基準によってパラメータ系列が生成される．このパラメータ系列を利用して音声波形が生成され，合成音声として出力される．

以上の手順はあくまでデモスクリプトで用いられているものではあるが，HTS による HMM 音声合成では概ねこの手順に沿って HMM の学習と音声の合成を行うことが多く，本研究でも概ねこの手順に基づいた検討を行っていく．

第3章

絵本読み聞かせ音声コーパスの構築

3.1　はじめに

本章では，絵本読み聞かせ風音声合成を実現するために必要となる音声コーパスの構築について述べる．3.2 節では話者や文章などといったコーパスの仕様に関して述べる．3.3 節では収録した読み聞かせ音声に対するラベリングに関して述べる．その後，3.4 節では，本章でのコーパス構築を踏まえ，次章において目指す音声合成について示す．

3.2　コーパスの仕様

3.2.1　話者選定

音声の収録対象とする話者について，絵本の読み聞かせに熟達した保育士から候補を募った結果，6 名の立候補者が集まった．うち 1 名についてはアナウンサーのような読み方であり，今回の趣旨に合致しないため候補から除外した．残る立候補者 5 人について，それぞれ 3 通りの読み聞かせ音声サンプルを 43 名の日本人に web 上で聴取させ,「自分の子どもに読み聞かせを依頼する場合，誰を選ぶか」という観点から優先順位を問うアンケートを行った．

この結果，それぞれの順位に選ばれた頻度，および平均順位は表 3.1 のようになった．これを踏まえ，平均順位上位 2 名である話者 A，話者 C の 2 名を今回の収録話者として選定した．このうち話者 C のサンプル音声では地の文とセリフの文との読み方の変化が特徴的であった．また，話者 A のサンプル音声では句末に BPM が頻繁に見られた．

表 3.1: 話者選定アンケートの結果

話者	1 位	2 位	3 位	4 位	5 位	平均順位
A	30	32	41	26	3	2.55
B	19	39	31	34	9	2.81
C	**70**	28	19	10	5	**1.88**
D	13	24	19	27	49	3.57
E	0	9	22	35	66	4.20

3.2.2　文章選定

コーパスの文章としては，絵本の文章をそのまま用いることとした．用いる絵本に関しては，図書館の職員の意見を参考に，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した．その上で，様々な属性・関係性を持ったキャラクターが登場するよう注意した．また，音素バランス性も考慮し，日本人原作のものと外国人原作のものが共に存在するよう選定した．登場頻度の低いモーラをカバーするため，かるた遊びのように各々のひらがなについて例文を提示している絵本も 1 冊採用した．この結果選ばれた絵本は以下の 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

文数の合計は 917 文となった．ただし，文数については,「『おはよう』と，おかあさんが言いました」のような引用文は基本的に 2 文としてカウントしている．

これらの文章に関して各音素の出現頻度をカウントし，音声合成の研究においてしばしば学習データとして用いられる ATR 音素バランス文 [19] と比較したものを Fig. 3.1, 3.2 に示す．音素数，モーラ数共に合計数は ATR 音素バランス文を上回っており，なおかつ登場回数が ATR 音素バランス文に比べて極度に少ない音素も存在しないため，音素バランス性は損なわれていないと言える．なお，モーラごとに見ると絵本の文章には「し」「た」「ま」の 3 つが特に多く現れていた．これは，絵本において「〜ました.」「〜しました.」という文が多いことに起因していると考えられる．

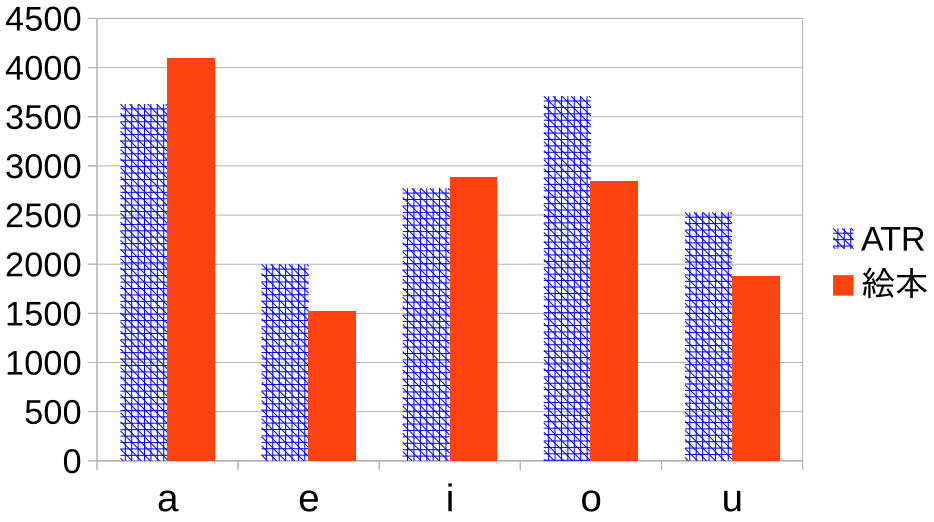


図 3.1: 母音の出現頻度

3.2.3　収録条件

以上により選定された文章について，選定された 2 話者に

- アナウンサーがニュース原稿を読むような単調な読み方（読み上げスタイル）
- 子どもに絵本を読み聞かせるような読み方（読み聞かせスタイル）

の 2 通りのスタイルによる発声を依頼し，それぞれについて音声を収録した．音声の収録は防音室で行ったが，各スタイルの特徴がより出やすくなるよう，読み上げスタイルについては絵本の文章を漢字仮名混じりに書き直した文章のみを呈示し，読み聞かせスタイルについては挿絵など

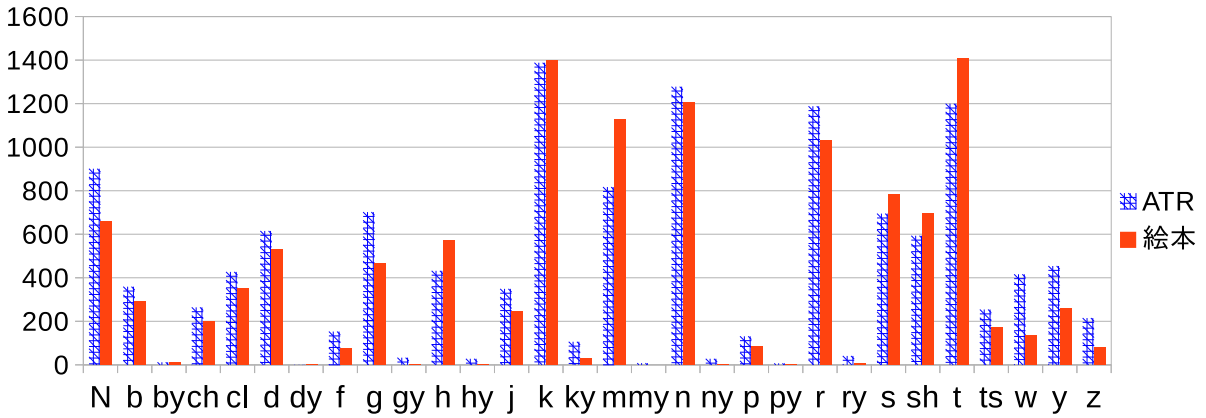


図 3.2: 子音の出現頻度

表 3.2: 絵本読み聞かせコーパスの仕様

話者	女性保育士 2 名（話者 A，C）
文数	917 文（絵本 7 冊分）
収録スタイル	読み上げスタイル
	読み聞かせスタイル
時間	各話者・スタイルにつき約 1 時間

が入った絵本そのものを呈示した．なお，絵本の文章は漢字表記のあるものもあったが，全てについてふりがなが記されていた．

以上のようにして収録されたコーパスの仕様を表 3.2.3 に示す．

3.3　収録音声へのラベリング

3.3.1　ラベリングの枠組み

前節で収録した音声を音声合成システムで利用するため，アクセント情報などのラベリングを行った．ラベリング作業に関しては，アクセント核の位置などに対する感覚が優れた者が行う必要がある．そのため，音声学を専攻する学生 3 名をラベラーとして起用した．

作業をより行いやすくするため，web 上にラベリングインターフェースを構築した．その様子を図 3.3 に示す．このインターフェース上では同一の文に対して読み上げ，読み聞かせ両スタイルの音声を聞き比べながらそれぞれに対するラベリングを並行して進めることができる．アクセント情報に関してはスタイルを問わず一致する点も多いため，読み上げスタイルのラベリング結果を読み聞かせスタイルにコピーし，初期値として用いることができるような機能を備えた．その上で，聞き比べることにより，新たに着目すべき読み聞かせ音声の特徴を見出すことも狙いとした．なお，情報のない状態から全てのアクセント核の位置などをラベリングすることは負担が

大きい作業となるため，あらかじめ鈴木らの手法 [20] によりテキストから自動推定されたアクセント境界およびアクセント核の位置を示し，これを実際の音声に対応させて修正する，という方策をとった．

まず，全ラベラーによる同一の 50 文程度に対するラベリングを行った．ラベリング項目は暫定的に

- ポーズ
- アクセント核
- アクセント句境界
- フレーズ境界
- 無声化
- 句末のピッチ上昇（BPM）

の 6 項目とした．これらのうち，BPM 以外については通常の「読み上げ」調べる音声合成用コーパス構築でも行われるラベリングである．

この作業によって，異なるラベラー間での不一致が最小限となるよう，ラベリング基準・戦略の統一を図った．それと共に，新たに指摘された

- 長音化（長音が更に長くなる場合を含む）
- 抑揚の大きさ（抑・中・揚の 3 通り）
- 緩急（緩・急の 2 通り）

の 3 項目を読み聞かせ音声のラベリング項目に加えることとした．ラベリング基準・戦略の同意を得た後は，各ラベラーで異なる文章についてラベリングを進行した．

3.3.2　ラベリングを通して確認された読み聞かせ音声の特徴

前節で触れたように，ラベリングを通して読み聞かせ音声の特徴が数点挙げられ，新たなラベリング項目として追加されている．ここで，それぞれの特徴について例と共に示すこととする．なお，各項目の「音声」アイコンをクリックすることで，当該音声のサンプルを聴取することができる．

i)　上昇調 BPM

アクセント句末でピッチを上昇させる上昇調 BPM が多くの文章で確認された．特に多く見られたのは

じょーじは，かけらを ひとつ，とりだしてみました。 音声

という文の「じょーじは」のように，主語を示す句の末尾のピッチを上げるものであった．この他にも，

そして、その わらいごえが、だんだん うつくしい うたごえに かわっていきます。
音声

という文の「そして」のような接続詞の末尾のピッチを上げる場合が多く見られた．また，上昇調 BPM が生じたアクセント句末の直後にはポーズが挿入される場合が非常に多かった．

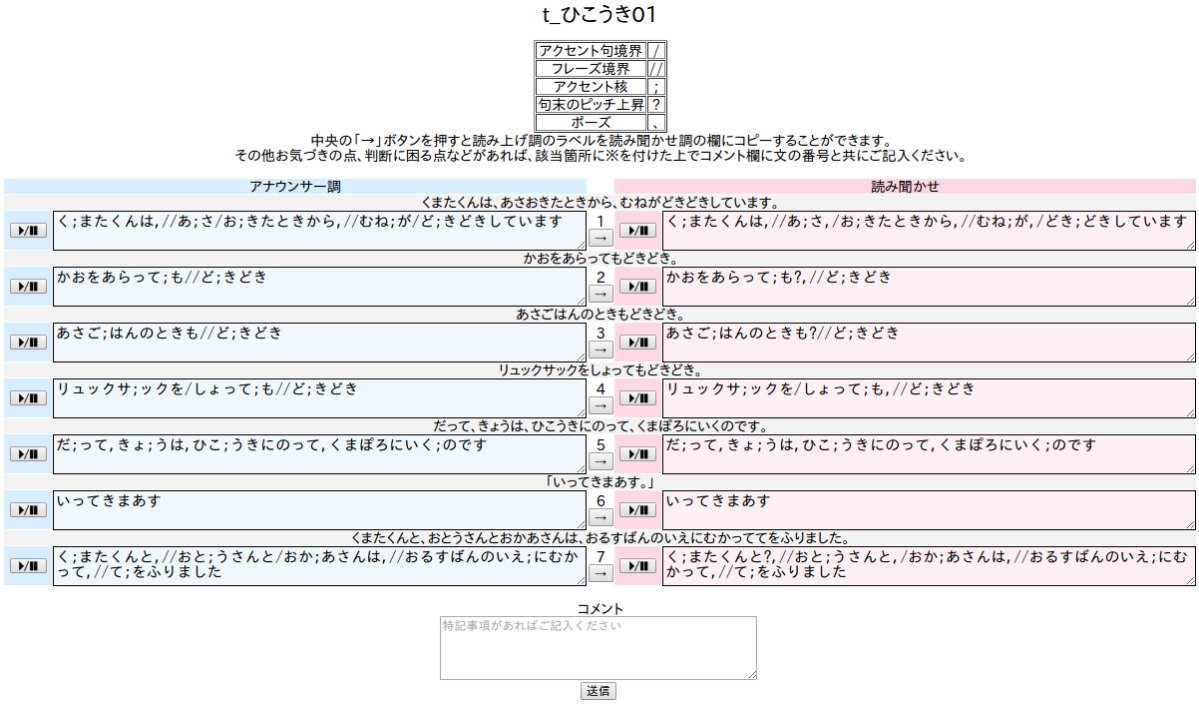


図 3.3: ラベリングインターフェースの様子

ii)　長音化

テキスト上では長音でない部分を伸ばして読む長音化，あるいは長音を極端に長く伸ばして読む現象が確認された．例として，

なんだ，もう，あけちゃったのか。 音声

というセリフの文において,「だ」が長音化すると共に「もう」という部分が極端に長く発音されていた．

iii)　抑揚の程度の変化

第 2 章で述べた通り，日本語ではピッチの上下，すなわち抑揚によりアクセントが表現される．読み聞かせ音声においては，アクセント句内での抑揚の程度を変えながら発話するケースが多く見られた．例えば，

ぼうしは ぺっしゃんこ。 音声

という文について,「ぼうしは」という句における各モーラのピッチは読み上げ音声では LHHH となっているが，読み聞かせ音声では 4 モーラとも低く読まれており，その後の「ぺっしゃんこ」という句では抑揚の幅が大きくなっている．この他にも，

ながいろうかを とおって，かんごふさんが しんさつしつへ つれていってくれました。
音声

という文では,「かんごふさんが」が低く平らに読まれ，その後の「しんさつしつへ」「つれていって」で段階的に抑揚の幅が大きくなっていく様子が見られた．これらの例のように，2 段階，あるいは 3 段階で抑揚の程度が大きくなっていく様子が多くの文で確認された．

iv)　話速の緩急

句ごとに話速を変化させながら発話している箇所が多く確認された．例えば，

> はじめは ゆっくり，それから，だんだん はやく，そして，しまいには，くるくる，くるくる，くるくる，くるくる——音声

という文では，文の意味に合わせるように，文の最初の部分を遅く発話し，その後次第に発話を速くしていく様子が確認された．このように，速さや長さを表現する単語があると，発話内容に合わせるように緩急を変化させる傾向が見られた．

この他にも，

> おなかの れんとげんしゃしんを とるからね 音声

という文では,「おなかの」の部分が低く平らに，かつ遅く発話されており,「れんとげんしゃしんを」の部分は抑揚の幅が大きく，かつ速く発話されている．この例のように抑揚と緩急を組み合わせた読み方は複数箇所見られ，このような読み方の工夫が文中の強調箇所をより際立たせる効果を担っていると考えられる．

3.4　本研究で目指す音声合成

次章では，本章で構築したコーパスとそのラベリング結果を踏まえて音声合成に関する検討を行っていく．本研究では，合成用のテキストに対し，本章で確認された読み聞かせ音声の特徴を示すラベルを付与した状態のものを入力とし，ラベルによって指定された特徴の現れた音声を合成することを目標とする．つまり，例えば，3.3.2 節の iii) において示した音声サンプルのような合成音声を，

> な@が@いろうかをとおって，[かんごふさんが]，[[しんさつしつへ]][[[つれていって]]]くれました．

のようなラベル付きテキストを入力することで出力するモジュールを開発する．なお，この例において，@は直前のモーラが長音化することを，[] は抑揚の程度が小さいことを，[[]] は抑揚の程度が中程度であることを，[[[]]] は抑揚の程度が大きいことを，それぞれ表す記号である．

こうしたラベルをテキストから自動的に推定することに関しても需要はあると考えられるが，これに関しては本研究の対象とせず，後続の研究に譲ることとする．

第4章

絵本読み聞かせ風音声の合成のための検討

4.1　はじめに

本章では，前章で構築した音声コーパスを用いた音声合成に関する諸検討について述べる．

4.2　共通の検討条件

本章では，HMM 音声合成の枠組みにおいて，学習データとして用いる音声の種類やコンテキストラベルの種類を様々に変えながら検討を進めていく．そうした複数種類の条件の間で共通して用いる条件を表 4.1 に示す．HMM の学習データには女性話者 C の音声を用いた．音声の分析は STRAIGHT を用いて [22]，F_0，スペクトル包絡特徴量，非周期性指標を抽出した．分析条件は，フレーム周期 1 [msec] であり，F_0 の探索範囲は最小値 120 [Hz]，最大値 800 [Hz] である．HMM に用いた特徴量は，0 から 39 次元までのメルケプストラムと 0-1，1-2，2-4，4-6，6-8 [kHz] の 5 帯域の平均非周期性指標，対数 F_0，およびそれらの Δ，Δ^2 を含めた 138 次元のベクトルとした．メルケプストラムと平均非周期性指標は，STRAIGHT を用いて抽出したスペクトル包絡特徴量と非周期性指標から，それぞれ SPTK [1] を用いて求めた．また，HMM の学習においては，分析により得られた特徴量を 5 フレームごとに取り出し，フレーム周期を 5 [msec] とした．この時，5 通りの取り出し方が考えられるが，5 通りの特徴量系列に全てに対しそれぞれ同一のラベルを付与したものを学習データとして用いた．すなわち，学習に用いるデータ量は，単純にフレーム周期 5 [msec] の条件下で分析を行ったものをそのまま用いる場合に比べて 5 倍となっている．特徴量の取り出し方のイメージを図 4.2 に示す．HMM は HTS-2.2 を用いて構築した．状態継続長分布を明示的に含んだ 5 ステート left-to-right HSMM を用い，各状態の出力は単一の対角共分散ガウス分布とし，決定木によるコンテキストクラスタリングを行い，木の停止基準には MDL 基準を用いた．

表 4.1: 共通の条件

HMM 学習に用いる特徴量	対数 F_0，0〜39 元のメル一般化ケプストラム，5 次元の非周期性指標とそれらの Δ，Δ^2
話者	女性話者 C
F_0 探索範囲	120〜800[Hz]
分析フレーム周期	1 [msec]
音声の分析ツール	STRAIGHT
波形合成ツール	STRAIGHT
HMM の構築ツール	HTS-2.2

4.3　読み上げスタイルの音声を用いた音声合成

アクセント情報やポーズに関するラベリングの妥当性を検証するため，読み上げスタイルの音声を学習データとして用いた音声合成に関する検討を行った．

[1] SPTK, http://sp-tk.sourceforge.net/

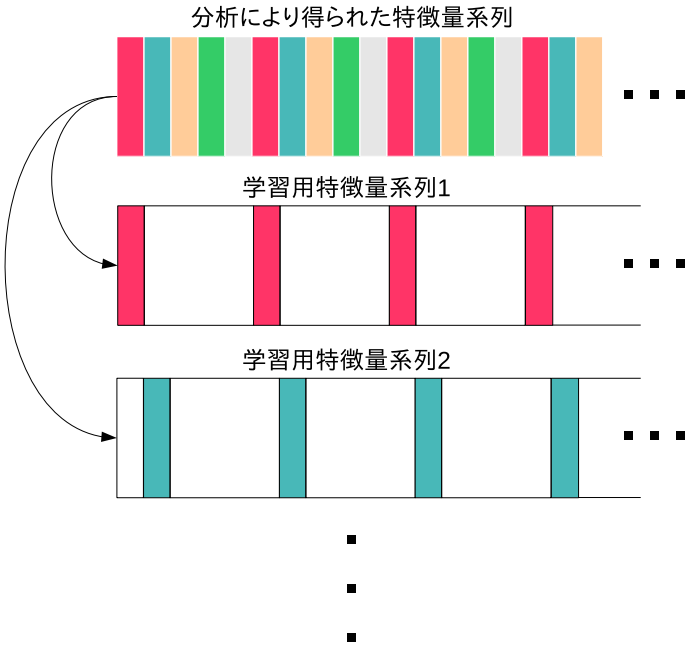
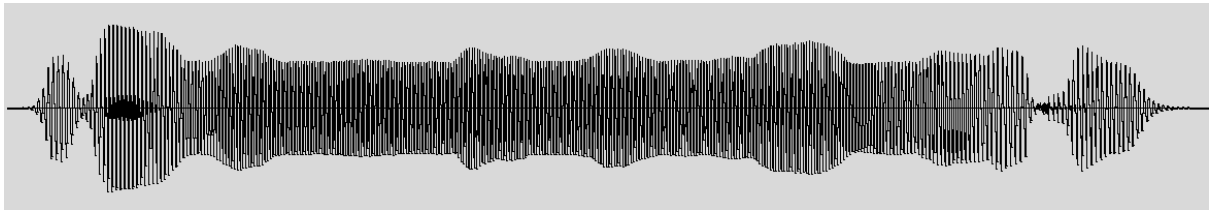


図 4.1: 学習用特徴量の取り出し方の様子

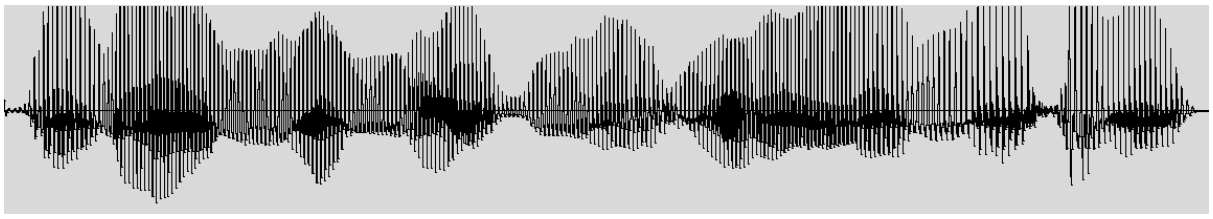
まず，テキストから鈴木らの手法 [20] により自動推定されたアクセント情報を学習用のラベルに用いて HMM 音声合成を実行したところ，ピッチの変動やポーズの挿入位置が明らかに不適切なものとなり，品質の低い，不自然な音声が合成された．音声 ここで用いたラベルでは，ポーズの位置が読点の位置と完全に一致していることを仮定しており，これは実際の音声とは必ずしも一致しない．また，アクセントの位置や句境界についても実際の発話に表れているものとは必ずしも一致しない．そのため，合成音声が不自然なものとなったことは概ね予想通りであると言える．

続いて，ラベラーによって手動で付与されたアクセント境界，アクセント核，ポーズの情報を学習用ラベルに反映して合成音声を作成した．これにより合成音声の品質が改善することが想定された，アクセント，イントネーションなどの F_0 制御には改善が見られた．一方で，モーラごとに切れ切れの音声となっており，品質の改善が十分でないことも確認された．図 4.3(b) にこの時の音声波形の一例を示す．図 4.3(a) に示した市販されている音声合成ソフトウェア N2[2] による合成音声 音声 の波形と比べて，振幅が大きくなったり小さくなったりしている様子が見て取れる．また，ポーズを指定しなかった箇所にもポーズが入った音声が合成されている箇所も確認された．この原因として，ポーズのラベルが HMM 音声合成に用いるには不適切な形で付与されていることが考えられた．具体的には，本来ポーズがある位置にポーズラベルが付与されておらず，ラベルのないポーズ部分が前後いずれかの音素の一部として学習されることで，ある音素の HMM の末尾あるいは先頭のステートが無音的な状態となっており，そのため，意図しないところで波形振幅が極端に小さくなっていることが疑われた．

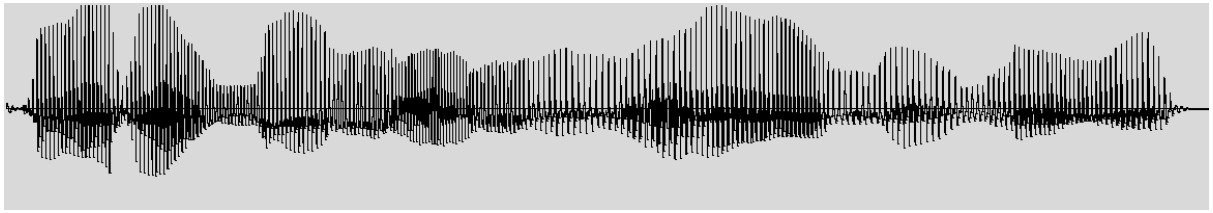
[2] http://www.kddilabs.jp/products/audio/n2tts/product.html



(a) N2 による合成音声の波形



(b) モーラごとに切れ切れとなる合成音声の波形



(c) ポーズの自動検出を適用した場合

図 4.2: 読み上げスタイル音声による合成音声の波形（テキストは「それは，泥沼のような逆境から抜け出したいという，切ないほどの願望だろうか」という文の「泥沼のような」の部分）

そこで，学習用音声に対し HTK[3] を用いたポーズの自動検出を行い，その情報を用いることにした．学習用音声の音素書き起こしに対し，母音と/N/の直後にポーズが存在しうるという条件で音声認識を施し，最終的に，ポーズラベルを含む音素アライメント結果を得た．この結果，1 モーラ分程度の短いポーズが見落とされ，ポーズラベルが付与されていない箇所が多数確認された．これらのポーズは波形表示ソフトを用いると明確に確認できたが，今回のラベリングでは作業の効率化のために波形表示ソフトを用いず web 上のインターフェースを用いたこと，読み聞かせ音声の特徴に注意が向いていたことから精査されず，ポーズとは認められなかったものと考えられる．

自動検出によるポーズ情報をラベルに反映して合成音声を作成した結果，波形は 4.3(c) のようになった．市販のソフトウェアによる合成音声ほど振幅は安定していないが，切れ切れの聴取印象はほとんどなくなっており，十分な品質の改善が確認された．音声 このため，次節以降の読み聞かせスタイルの音声を用いた音声合成でもポーズに関しては自動検出による情報を利用することとした．

[3] http://htk.eng.cam.ac.uk/

4.4　読み聞かせスタイルの音声を用いた音声合成

4.4.1　初期モデルに関する検討

HMM 音声合成において初期モデルを用意する方法には，2.4.8 節で示したような時間情報を用いた Viterbi 学習のみでなく，全ての学習データからモデルパラメータの平均を計算し，それを全音素，全ステートの平均ベクトルとして採用するフラットスタート法がある．つまり，フラットスタート法における初期モデルでは全ての音素・ステートが物理的に同一のパラメータ値を有する．フラットスタート法は音素の時間情報を必要としないという利点を持つ．Julius による強制アライメントは長音化が頻繁に発生している読み聞かせ音声において正しく機能しないおそれがある．仮に誤った時間情報を用いると初期モデルが不適切に学習されてしまい，合成音声の品質が低下の一因となりうる．

読み聞かせスタイルの音声を用いて合成音声を作成した際，後述する長音化ラベルを導入した場合でも，意図せぬところで長音化が発生する，声質が安定しないなど，合成音声の品質が非常に低い状態であることが確認された．音声 この原因が前述した初期モデルの不適切な学習である可能性を考え，フラットスタート法による HMM の学習を検討した．

この結果，継続長の不具合が改善されることを確認できた．そのため，以降の検討では，Julius による時間情報を利用した初期モデルの学習は行わず，フラットスタート法を採用することにした．

4.4.2　読み聞かせ用コンテキストラベルの設計

通常の HMM 音声合成に用いられるコンテキストラベルに含まれるコンテキストの種類を 2.4.7 節で示した．この「読み上げ」用のラベルのみを用いて読み聞かせスタイルの音声を用いた音声合成を行っただけでは，意図せぬ箇所で F_0 が変動するなど，非常に不自然な音声が合成されてしまうことが確認された．音声 このことから，読み聞かせらしい音声を合成するには，学習データとして読み聞かせ音声を用いるだけでなく，読み聞かせ音声における読み方の工夫を適切に表現するコンテキストラベルの導入が必要であると考えられる．

そこで，コーパスに対するラベリングを通して確認された読み聞かせの特徴をコンテキストラベルに追加した．追加したコンテキストの種類を表 4.2 に示す．ここで，F_0 の制御に関わりが深いと考えられるものを A，継続長の制御に関わりが深いと考えられるものを B，その他のものを C として分類した．A と B に分類されたコンテキストは共に読み聞かせ音声コーパスに対するラベリングにおいて確認された読み聞かせ音声の特徴をコンテキストラベルに組み込んだものである．C に分類されたコンテキストは，話者選定時に挙げられたこの話者の読み聞かせ音声の特徴であり，音響特徴量との関連が明確ではないが，読み聞かせ音声の特徴を表現するために必要と考えられるものである．

また，追加したコンテキストの影響が反映されるよう，コンテキストクラスタリングに用いる質問セットにも，これらのコンテキストに関係する質問を追加した．追加した質問の種類を表 4.3 に示す．ここでも，追加したコンテキストと同様の分類を行っている．

通常のコンテキストラベルおよび質問セットに対し，これらのコンテキストと質問を追加して，合成音声を作成した．学習用音声と同一のテキストに対し，異なる箇所において BPM や長音化，抑揚の程度の変化を生じさせるように指定したコンテキストラベルを用いて（text-closed,

表 4.2: 追加したコンテキストの種類

A	当該アクセント句に上昇調 BPM があるか
	抑揚ラベルの属性（なし，抑，中，揚）
B	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
	緩急ラベルの属性（なし，緩，急）
C	地の文/セリフ，キャラクター属性（男女，大人子供）

prosody-open）合成音声を作成した結果，意図した通りの制御が可能であることが確認された．
音声 音声

また，学習用音声とは異なるテキストに対しても，今回設計した読み聞かせ用コンテキストラベルを用いる（text-open, prosody-open）ことで，概ね意図した通りの読み方を実現できることが確認された．音声

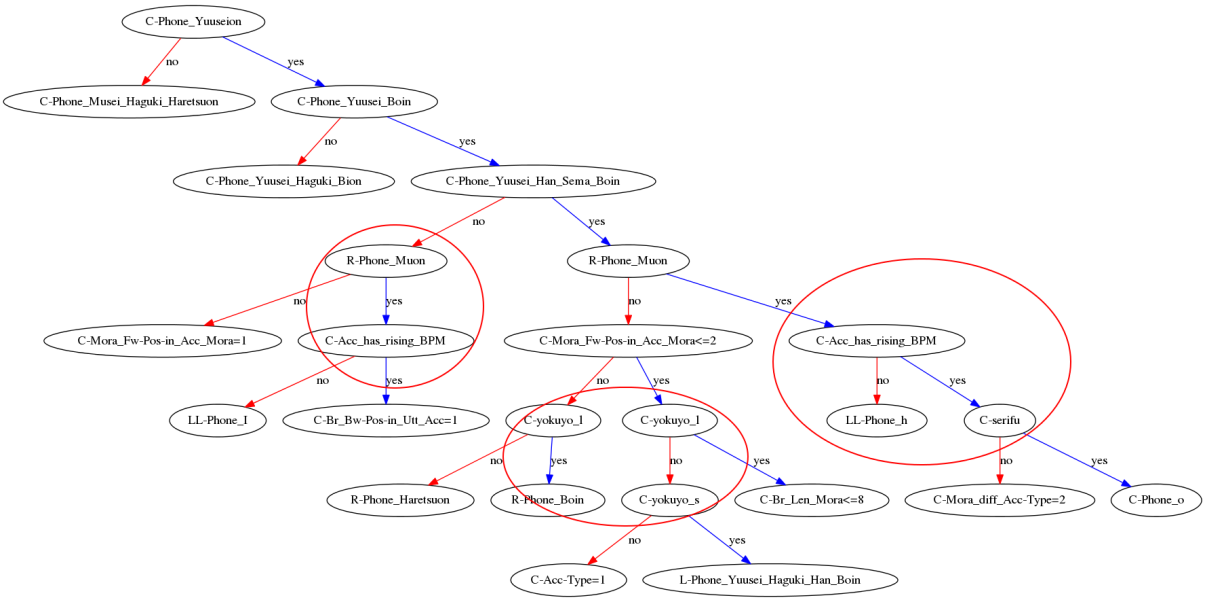
これらの制御が正しく行われていることは，コンテキストクラスタリングにおける決定木の様子から見て取ることができる．図 4.4.2 に，A,B,C 全種の読み聞かせ用コンテキストラベルを用いた際の決定木の様子を示す．それぞれ，読み聞かせ用コンテキストに関する質問が選択されている部分を赤い丸で囲んである．

図 4.4.2(a) は F_0 の 4 番目のステートに関する決定木を最上位から一部抜粋したものである．この決定木では，後続音素がポーズであるか否かを示す「R-Phone_Muon」が Yes であった場合，直下に当該アクセント句に上昇調 BPM があるか否かを示す「C-Acc_has_rising_BPM」の質問が選択されている．これは，ポーズの前に位置するアクセント句の末尾の音素の後半部の F_0 の決定について，上昇調 BPM の有無が重要であることを意味している．また，抑揚の程度が「揚」であるかを否か示す「C-yokuyo_l」や抑揚の程度が「抑」であるかを否か示す「C-yokuyo_s」といった質問も比較的上位に選択されており，これらの情報も F_0 の決定において重要度が高いことを示している．

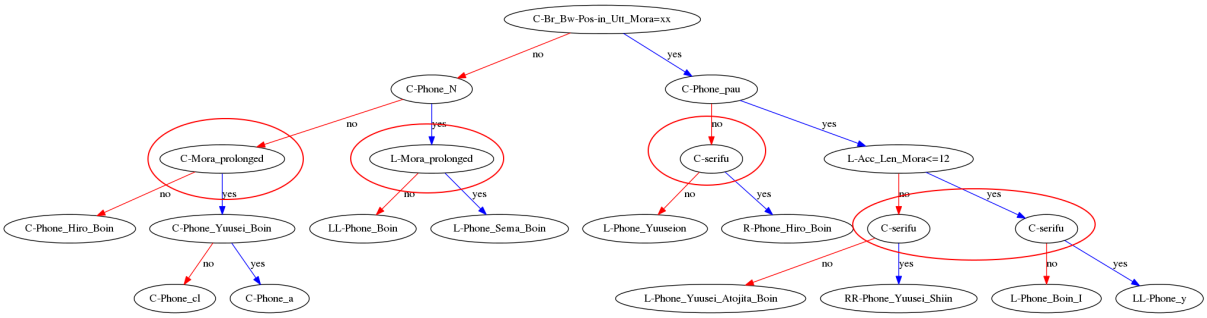
また，図 4.4.2(b) は音素の継続長に関する決定木を最上位から一部抜粋したものである．この決定木では，セリフ部分であるか否かを示す「C-serifu」の質問が積極的に選択されており，地の文であるかセリフであるかどうかは継続長にとって重要な情報であることが示唆されている．また，当該モーラが長音化しているか否かを示す「C-Mora_prolonged」や先行モーラが長音化しているか否かを示す「L-Mora_prolonged」といった質問も上位で選択されており，長音化を示すコンテキストの有効性を確認することができる．また，当該モーラが長音化している場合，直下の質問は当該音素が有声母音であるか否か，そしてその質問が No の場合さらに直下の質問は当該音素が促音であるか否かとなっている．このことから，長音化の影響を大きく受けているのは有声母音と促音であると言える．

表 4.3: 追加した質問の種類

A	当該アクセント句に上昇調 BPM があるか
	抑揚の程度が「抑」であるか
	抑揚の程度が「中」であるか
	抑揚の程度が「揚」であるか
B	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
	先行モーラが長音化しているか
	緩急のつけ方が「緩」であるか
	緩急のつけ方が「急」であるか
C	セリフ部分であるか
	男性キャラクターのセリフであるか
	女性キャラクターのセリフであるか
	子供キャラクターのセリフであるか
	大人キャラクターのセリフであるか
	子供男性キャラクターのセリフであるか
	子供女性キャラクターのセリフであるか
	大人男性キャラクターのセリフであるか
	大人女性キャラクターのセリフであるか



(a) F_0 の 4 ステート目にの決定木の様子



(b) 継続長の決定木の様子

図 4.3: 読み聞かせ用コンテキストラベルを導入した際の決定木の様子

表 4.4: 評価実験に用いたコンテキストラベルと質問セットの条件

条件名	用いたコンテキストと質問の種類
conventional	通常の HMM 音声合成に用いられるコンテキストと質問
yokuyo_off	conventional + B + C
kankyu_off	conventional + A + C
serifu_off	conventional + A + B
all_labels	conventional + A + B + C

4.4.3　読み聞かせ用コンテキストラベルの評価実験

前節において設計した読み聞かせ用コンテキストラベルの効果を確認するため，聴取実験を行った．読み聞かせスタイルの表 4.4 に示す 5 通りのコンテキストラベルと質問セットを用いて HMM 音声合成を行った．なお，表中の A，B，C は前節におけるラベルと質問の分類を示している．合成に用いたテキストはコーパスには含まれない絵本の一部であり，以下に示す 3 セットである．

- うずらちゃんが ひよこちゃんとかくれんぼを はじめました。
  「じゃんけん ぽん！」「あいこで しょ！」
  うずらちゃんの かち！
  さいしょは うずらちゃんが かくれます。
- 「さて、この いす、どこへ おこうかな。」
  ちょっと かんがえると たちまちいい かんがえが うかびました。
  うさぎさんは たてふだを ひとつつくりました。
- たまごが おちていました。
  「やあ、なんて おおきな たまごだろう。
  おつきさまぐらいの めだまやきが できるぞ」と、ぐりが いいました。

各セット内において，5 種類のコンテキストラベルによる合成音声から 2 音声の組み合わせ 10 通り全てを呈示し，「どちらの音声がより読み聞かせらしいか」という基準で 7 名の被験者に対して評価を行わせた．この結果から，サーストンの一対比較法に基づき尺度値を求めた [23].

一対比較による回答数の分布を表 4.5 に示す．なお，表中の回答数は縦軸の条件による合成音声と横軸の条件による合成音声を比較した際，縦軸の手法がより読み聞かせらしいと評価した回答の数である．この回答数をもとに算出された標準得点とサーストンの尺度値は表 4.6 の通りである．尺度値は yokuyo_off で最も高く 0.692，次いで高いのは all_labels で 0.307 であった．また，尺度値が最も低かったのは conventional の-0.919 であり，読み聞かせ用コンテキストを導入した全ての条件でこの値を大きく上回っていた．

4.4.4　考察

聴取実験の回答数は 21 であり，これは十分な数であるとは言えないが，読み聞かせ用コンテキストを導入した全ての条件において尺度値が conventional の尺度値を大きく上回っており，読み聞かせ用コンテキストを導入することで合成音声がより「読み聞かせらしい」と感じられる傾向

表 4.5: 聴取実験における回答数の分布

	conventional	yokuyo_off	kankyu_off	serifu_off	all_labels
conventional	–	1	4	6	6
yokuyo_off	20	–	16	15	9
kankyu_off	17	5	–	10	8
serifu_off	15	6	11	–	9
all_labels	15	12	13	12	–

表 4.6: 聴取実験の標準得点とサーストンの尺度値

	conventional	yokuyo_off	kankyu_off	serifu_off	all_labels	尺度値
conventional	–	-1.668	-0.876	-0.566	-0.566	-0.919
yokuyo_off	1.668	–	0.712	0.566	-0.180	**0.692**
kankyu_off	0.876	-0.712	–	-0.060	-0.303	-0.050
serifu_off	0.566	-0.566	0.060	–	-0.180	-0.030
all_labels	0.566	**0.180**	0.303	0.180	–	0.307

にあることが確認できる．

yokuyo_off，kankyu_off，serifu_off の 3 条件を all_labels と比較した場合では，serifu_off において尺度値が最も下がっており，セリフであるか否か，およびキャラクター属性を示すコンテキストが「読み聞かせらしさ」を担っていると考えられる．

一方で，yokuyo_off では all_labels よりも尺度値が高くなっていた．yokuyo_off の音声では BPM を示すコンテキストと質問が存在しないにも関わらず，それらが存在する場合と同一の箇所で BPM が発生している様子が確認された．これは BPM の有無を示すコンテキストがなくても，他のコンテキストで BPM の有無に関する情報がある程度表現されていることを意味している．そこで，yokuyo_off の条件における F_0 の決定木の様子を確認したところ，図 4.4.4 のようになっていた．これを図 4.4.2(a) と比較すると，「R-Phone_Muon」が Yes であったとき，図中の 2 つのケースでは共通して，数段下で「C-serifu」の質問が選択されている様子が確認できる．なお，図中では「C-serifu」の質問を赤い丸で囲んである．したがって，セリフであるか否かが BPM の有無に何らかの関係を持っていることが考えられる．

これ以外にも，A に分類されるコンテキストと質問を導入しない場合の方が尺度値が高くなった理由として，抑揚の程度に関するコンテキストとして前後のアクセント句の抑揚の程度を採用しておらず，複数の句をまたいで現れる抑揚の程度の変化を必ずしも良く表現できていないことから，conventional と比較した際に，all_labels では yokuyo_off に比べてはっきり良くなったと感じられなくなっていることが考えられる．なお，all_labels と yokuyo_off の直接の比較では all_labels の方が読み聞かせらしいとする回答が半数以上であり，また，BPM や抑揚の程度に関するコンテキストを除外するとこれらのを明示的に制御することはできなくなることも考慮すると，除外は必ずしも望ましいとは言えない．したがって，これらのコンテキストに関して，より良い表現となるよう検討していく必要があると考えられる．

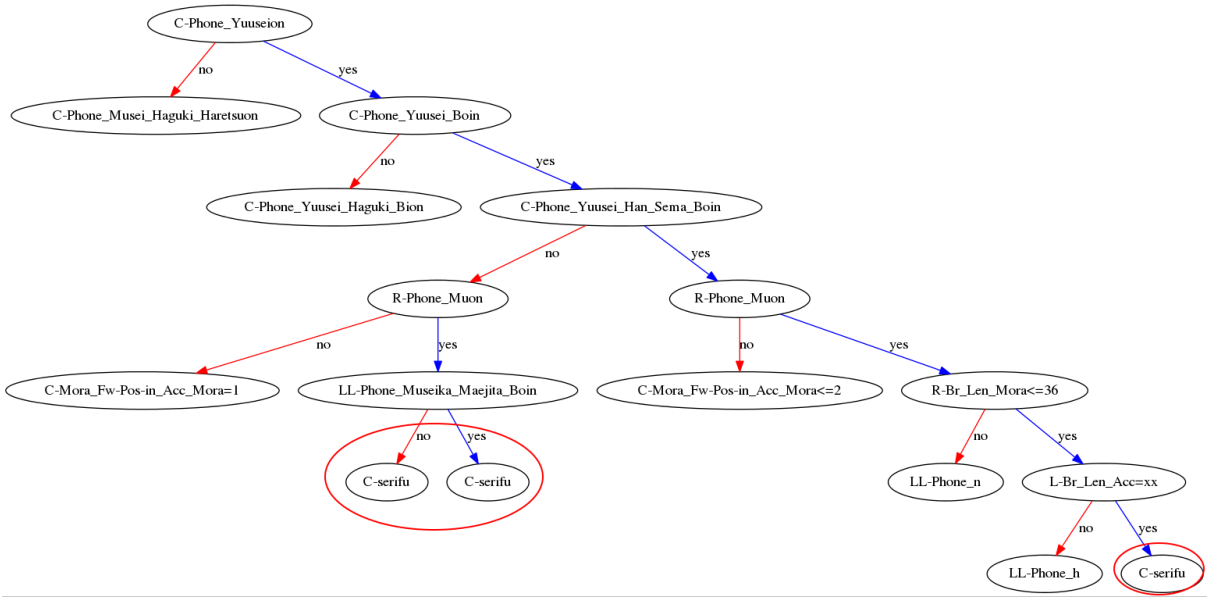


図 4.4: yokuyo_off の条件における F_0 の 4 ステート目の決定木の様子

第5章

結論

5.1　本論文のまとめ

本論文では，表現豊かな音声合成として，聞き手との関係を反映した social な音声合成という軸のもとで，子供に向けた話し方，特に絵本を読み聞かせるような音声の合成を目指した．そのためのアプローチとして，絵本読み聞かせ音声コーパスを設計・構築し，HMM 音声合成における学習手順とコンテキストラベルの拡張に関する検討を行った．

絵本読み聞かせ音声コーパスは女性保育士による絵本の単調な読み上げ音声と読み聞かせ音声を収録したもので，文章についても音素バランスが損なわれていないことが確認されている．収録した音声に対してラベリングを行い，音声に基づいたアクセントラベルを付与すると共に，読み聞かせ音声に特徴的な現象についても確認し，それらの現象に対するラベルの付与を行った．具体的には，アクセント句末の BPM，長音化，抑揚の程度の変化，緩急の変化といった現象が確認され，読み聞かせ音声に対するラベルとして採用された．

構築した読み聞かせ音声コーパスを用いた HMM 音声合成に関する検討では，ポーズの自動検出およびフラットスタート法の有効性が確認された．また，読み聞かせ用コンテキストとして，前段で用いた読み聞かせ音声に対するラベルの情報とセリフであるか否か，およびセリフである場合はキャラクター属性の情報を採用したコンテキストラベルを設計し，これを用いた HMM 音声合成を行った．この結果，コンテキストクラスタリングによる決定木の様子から，合成音声における BPM や長音化などの制御が可能であることが確認された．さらに，聴取実験による評価でも，サンプル数が十分ではない中ではあるが，設計したコンテキストラベルを用いることでより読み聞かせらしい合成音声となる傾向が確認された．尺度値としては BPM と抑揚の程度に関するコンテキストを導入しない場合が最も読み聞かせらしいと評価される結果となっていた一方で，本研究で提案した全てのコンテキストを導入した場合は他の全条件との比較で過半数の支持を得ていた．BPM や抑揚の程度の制御が妥当になされていることも確認されていることも考慮すると，尺度値をそのまま信用してコンテキストを除外することが必ずしも望ましいとは言えない．

5.2　今後の展望

本研究で設計した読み聞かせ用コンテキストラベルに関して，前後のアクセント句の抑揚の程度を示すコンテキストと質問の導入やキャラクター属性の更なる拡張など，様々な改善案が考えられる．特にキャラクター属性に関しては，さらに多くの絵本の読み聞かせ音声を収録することでも，多種多様なキャラクターが登場し，より豊かな表現が可能になっていくことが期待できる．

また，任意のテキストに対し視覚的でわかりやすい編集を通して今回用いたコンテキストラベルへの変換が可能となるエディタが作成されれば，任意の文章を，所望の読み方で「読み聞かせ」てくれるアプリケーションが実現されることも期待される．このような，合成音声を自在にコントロールするための制御・編集機能は統計的音声合成技術における課題となっている [24]．

さらに，本論文中でも触れた通り，テキストからの読み聞かせ用ラベルの予測も今後の研究課題となっている．

謝辞

まず本研究を進めるにあたり，2 年間指導教員として多大なご指導をして頂いた峯松信明教授に深く感謝いたします．また，ご退官までの 1 年間ではありましたが，広瀬啓吉名誉教授にも様々なご指導を賜り，たいへん感謝いたしております．また，日頃の研究活動を支えて下さった高橋登技術専門員，池上恵事務補佐員，折茂結実子事務補佐員にも，ここに感謝の意を表します．

さらに，齋藤大輔助教，博士課程の橋本浩弥氏にも，音声合成研究の先達として多くの貴重な意見を頂きましたことを，深く感謝いたします．両名の大きな支えなくして，この研究がここまで来ることはなかったと，痛切に感じております．

そして，コーパスの話者募集のお知らせにご協力いただいた電気系の事務補佐員の皆様，その募集に対し立候補していただいた保育士の皆様，ラベリングを引き受けていただいた國學院大學の坂本薫氏，竹内はるか氏，木野景子氏など，本研究は私が直接存じ上げている方からそうでない方まで，非常に多くの方々のご協力のもとに初めて成り立ちました．ここに，全ての方々への深い感謝の意を記します．

また，研究はもちろんのこと，時には他愛のない会話に興じるなど，様々な面でお世話になった研究室の方々にも感謝いたします．特に，同期の鈴木颯氏とは非常に多くの時間を共に過ごし，あらゆる面で支えて頂いたことを深く感謝しております．

そして最後に，これまで私を支えてくださったすべての方々に深く感謝いたします．本当にありがとうございました．

2016 年 2 月 4 日

百武 恭汰

参考文献

[1] H. Kanagawa, T. Nose, and T. Kobayashi ,“Speaker-independent style conversion for HMM-based expressive speech synthesis,” Proc. ICASSP, pp. 7864–7868, 2013.

[2] 河津宏美, 長島大介, 大野澄雄, “生成過程モデルに基づく感情表現における F_0 パターン制御規則の導出と合成音声による評価,” 電子情報通信学会論文誌 D, Vol. J89-D, No. 8, pp. 1811–1819, 2006.

[3] M. Soderstrom,“Beyond babytalk: Re-evaluating the nature and content of speech input to preverbal infants,” Developmental Review, 27, pp. 106–119, 2007.

[4] H. Fujisaki, and K. Hirose, “Analysis of voice fundamental frequency contours for declarative sentences of Japanese,” J. Acoust. Soc. Japan (E), vol.5, no.4, pp.233–242, 1984.

[5] 前川喜久雄, 五十嵐陽介, 菊池英明, 米山聖子, “『日本語話し言葉コーパス』のイントネーションラベリング Version 1.0,” 2004.

[6] Y. Igarashi, K. Nishikawa, K. Tanaka, and R. Mazuka, “Phonological theory informs the analysis of intonational exaggeration in Japanese infant-directed speech,” The Journal of Acoustical Society of America, 134(2), pp. 1283–1294, 2013.

[7] A. Fernald, T. Taeschner, J. Duunn, M. Papousek, B. Boysson-Bardies, and, I. Fukui, “A cross-language study of prosodic modifications in mothers’ and fathers’ speech to preverbal infants,” Journal of Child Language, Vol 16(3), pp. 477–501, 1989.

[8] Y. Saikachi, K. Watanabe, T. Konishi, N. Ito, A. Kanato, Y. Igarashi, K. Miyazawa, K. Nishikawa, and R. Mazuka, “『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, pp. 383–392, 2013.

[9] S. Amano, and T. Kondo, “対乳児音声の発声速度の長期的変化,” 電子情報通信学会技術報告, SP2008-38, pp. 105–108, 2008.

[10] Y. Igarashi, and R. Mazuka, “母親特有の話し方 (マザリーズ) は大人の日本語とどう違うか,” 電子情報通信学会技術研究報告 SP2006-90, pp.31–35, 2006.

[11] Y. Saikachi, M. Kitahara, K. Nishikawa, A. Kanato, and R. Mazuka, “The F0 fall delay of lexical pitch accent in Japanese Infant-directed speech,” Proc. INTERSPEECH, 2012.

[12] Y. Hasegawa, “The function of F0-peak delay in Japanese,” Proc. of the 21st Annual Meeting of the Berkeley Linguistics Society, 141-51, 1995.

[13] M.Kitahara, K. Nishikawa, Y. Igarashi, T. Shinya, and R. Mazuka, “対乳児発話におけるピッチアクセントの性質について,” 電子情報通信学会技術研究報告 SP, 108(338), pp. 133–136, 2008.

[14] K. Miyazawa, H. Kikuchi, T. Shinya, and R. Mazuka, “対乳児発話の母音の時間構造,” 電子情報通信学会技術研究報告 SP, 2009.

[15] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi, “Hidden Markov models based on multispace probability distribution for pitch pattern modeling,” Proc. ICASSP, pp.229–232, 1999.

[16] K. Tokuda, and H. Zen, “Fundamentals and recent advances in HMM-based speech synthesis,” Proc. INTERSPEECH tutorial, 2009.

[17] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi, “Hidden Markov Models Based on Multi-Space Probability Distribution for Pitch Pattern Modeling,” Proc. ICASSP, pp. 229–232, 1999.

[18] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, “A hidden semi-Markov model-based speech synthesis system,” IEICE Trans. Inf. & Syst., vol. E90-D, no. 4, pp. 825–834, 2007.

[19] A. Kurematsu, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara, and K. Shikano, “ATR Japanese speech database as a tool of speech recognition and synthesis,” Speech Communication, vol.9, pp.357–363, 1990.

[20] 鈴木雅之, 黒岩龍, 印南圭祐, 小林俊平, 清水信哉, 峯松信明, 広瀬啓吉, “条件付き確率場を用いた日本語東京方言のアクセント結合自動推定,” 電子情報通信学会論文誌 D, vol.J96-D, No.3, pp.644–654, 2013.

[21] K. Sinoda, and T. Watanabe, “MDL-based context-dependent subword modeling for speech recognition,” J. Acoust. Soc. Jpn. (E), vol. 21, no. 2, pp. 79–86, 2000.

[22] H. Kawahara, I. Matsuda-Katsuse, and A. de Cheveigné, “Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds,” Speech Communication, vol. 27, no. 3–4, pp. 187–207, 1999.

[23] L. Thurstone, “A law of comparative judgment,” Psychological review, vol. 34(4), pp.273, 1927.

[24] 徳田恵一, “統計的音声合成技術の現在・過去・未来,” 第 2 回自然言語処理シンポジウム & 第 17 回音声言語シンポジウム, 招待講演, http://www.sp.nitech.ac.jp/ tokuda/tokuda_SIG-SLP_2015_for_pdf.pdf, 2015.

発表文献

国内研究会・全国大会

[1] 百武 恭汰, 齋藤 大輔, 峯松 信明, “絵本読み聞かせ音声コーパスの構築とそのラベリングに関する検討,” 日本音響学会秋季講演論文集, 2015.

[2] 百武 恭汰, 齋藤 大輔, 峯松 信明, “絵本読み聞かせ風音声合成のためのコンテキストラベル設計に関する実験的検討,” 電子情報通信学会音声研究会, 2016.（発表予定）