宇都宮大学博士論文

対話音声が伝達するノンバーバル情報を表現可能な音声合成に関する研究

宇都宮大学大学院工学研究科
システム創成工学専攻

永田智洋

2018年3月

目 次

図 目 次

表 目 次

第1章　序論

1.1　研究の背景

1.1.1　人間同士のコミュニケーション

人間は他者とのコミュニケーションを行なうことで社会を形成し、今日まで発展してきた。コミュニケーションとは、人間同士が互いに意思・思考・感情を伝達し合うことであり、様々なメディアを通して実現されている。

人間が意思や思考を伝達し合う上でよく用いられるのが言語によるコミュニケーション (言語コミュニケーション、あるいはバーバルコミュニケーション) である。情報を送る側は自らの意思や思考を言語化し、文字に書き起こす、あるいは音声として発する。情報を受け取る側は言語化された内容を読み取り、送信者の意思や思考を理解する。

また、人間は言語のみを用いてコミュニケーションを行っているわけではない。ボクシングの試合のラウンド間にボクサーとセコンドが対話している場面を思い浮かべてほしい。セコンドが試合を継続させるかの判断をするためにボクサーに対して「行けそうか？」と尋ねる。それに対してボクサーが「大丈夫です」と答える。このやりとりを言語的な情報だけから判断すれば、セコンドは試合の継続を決めるであろう。しかし、ボクサーが肩で息をし、うつろな目をしており、弱々しい声で答えていたとしたらどうだろう。セコンドはボクサーの大丈夫というメッセージを強がりだと判断し、リタイアを決断するのではないだろうか。

このように、人間は表情、身体動作、視線といった言語以外の様々なものから意識的、あるいは無意識的に情報を発信・受信することでコミュニケーションを行っている。このようなメディアによって行われるコミュニケーションを非言語コミュニケーション、あるいはノンバーバルコミュニケーションと呼ぶ。

さて、コミュニケーションを行なうための手段には様々なものがあると述べたが、中でも音声は言語コミュニケーションおよびノンバーバルコミュニケーションの両方において重要な役割を果たしている。これは音声が言語情報以外にも様々な情報を伝達することが可能だからである。藤崎は音声によって伝達される情報を言語情報、非言語情報、パラ言語情報の 3 つに大別している [8]。ここで、非言語情報とパラ言語情報は文字に書き起こすことが不可能な情報を表す。また、非言語情報とパラ言語情報は話し手が制御可能であるかで分けられ、話し手自身は制御することができない性別、年齢、身体状態、感情といった情報は非言語情報、話し手自身が制御可能な発話意図や態度といった情報はパラ言語情報に分類されるとしている。言語情報が言語コミュニケーションに貢献していることは言うに及ばず、非言語情報とパラ言語情報がノンバーバルコミュニケーションに貢献することも想像に容易いであろう。

話者の制御下にあるという点で、音声によって伝達されるパラ言語情報は重要視される。先に述べた話者の発話意図や態度は特に注目される対象であり、伝達されるこれらの情報と声の高低や大小、発話速度、声質などの関係が調査されるなどしている [9]。

藤崎による分類は現在では非常に強い影響力を持っており、多くの研究がこの分類を採用している。しかしながら、感情の取り扱いに関してはしばしば解釈が分かれることがあり、感情をパラ言語情報に含める研究もある [10, 11]。特に、コミュニケーションの場面においては、あえて「怒ってみせ」たり「悲しんでみせ」たりするといった感情の模倣をすることも少なくない [12]。場合によっては、悲しみを押し殺して喜んでみせるといったように、本当の感情と別の感情を伝えるといった場面もある。このように感情については無意識的なものだけでなく、意識的なものをも含有している。これに対し、森は藤崎の分類を拡張した情報伝達のモデルを提案している [1]。このモデルを図 1.1 に示す。森による情報伝達モデルでは、話し手と聞き手が明確に区別されている。話し手が伝えようとするものをメッセージとし、パラ言語的メッセージには話者の意志による感情や態度の表出が含まれるとしている。

パラ言語情報の「パラ」という用語は「... のそば」や「... の近くに」という接頭辞に由来する。このことからパラ言語情報は言語に寄り添って伝達される

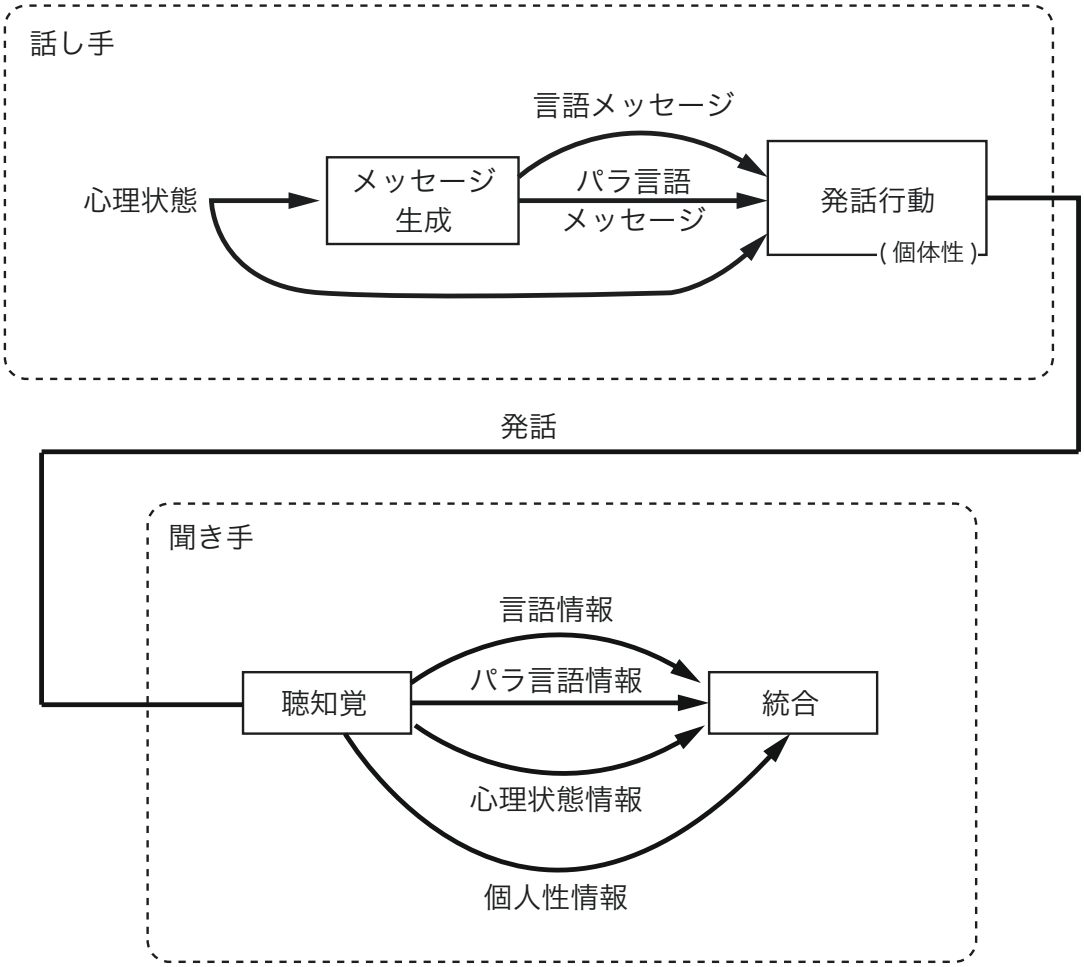


図 1.1: 音声による情報伝達モデル [1]

表 1.1: 笑いの種類 [6]

種類	タイプ	状況
快の笑い	本能充足の笑い	睡眠欲や食欲の満足
	期待充足の笑い	入試合格や試合勝利
	優越の笑い	他人と比較したときの優越感
	不調和の笑い	期待はずれや意味の取りちがえ
	価値逆転・低下の笑い	普段上位の者が下位の者に見下される
社交上の笑い	協調の笑い	挨拶
	防御の笑い	自分の本心を知られたくない時
	攻撃の笑い	他人の失敗や欠点、不道徳の非難
	無価値化の笑い	不都合な出来事をなかったことにする
緊張緩和の笑い	強い緊張の緩和	強いストレスからの解放
	弱い緊張の緩和	ダジャレやくすぐり

情報を意味しており、あくまで音声に付随して伝達される言語情報以外の情報として扱われることが多い。そのため、音声以外に発せられる音に対してはパラ言語として取り扱われないことがある。一方で、Trager は言語以外の音全てをパラ言語として扱っており [9]、鼻を鳴らす音や舌打ちによる音、笑い声や叫び声、うめきまでもパラ言語に含めている。ここではパラ言語が意味する範囲についてまで議論することはしないが、このような音声以外の音 (ここでは便宜的に非語彙音と呼ぶ) もノンバーバルコミュニケーションにおける情報伝達を担うメディアになっていることを否定する者はいないであろう。

中でも笑い声 (あるいは笑いそのもの) はよく注目される対象である。笑いは日常生活のいたるところで、ごくありふれたものとして観察され、時にはコミュニケーションにおける潤滑油とも呼ばれる。笑顔・爆笑・失笑・冷笑・微笑・嘲笑・苦笑・照れ笑い・愛想笑いというように、笑いに関する表現の多さもこのことを支持しているだろう。このようなことから、笑いの人間社会や対人コミュニケーションにおける役割などが精力的に研究されており、例えば Owren による研究では、笑いは社会的集団における建設的かつ協力的な関係を形成・維持するために進化したことが示唆されていると述べている [13]。また、志水らは、

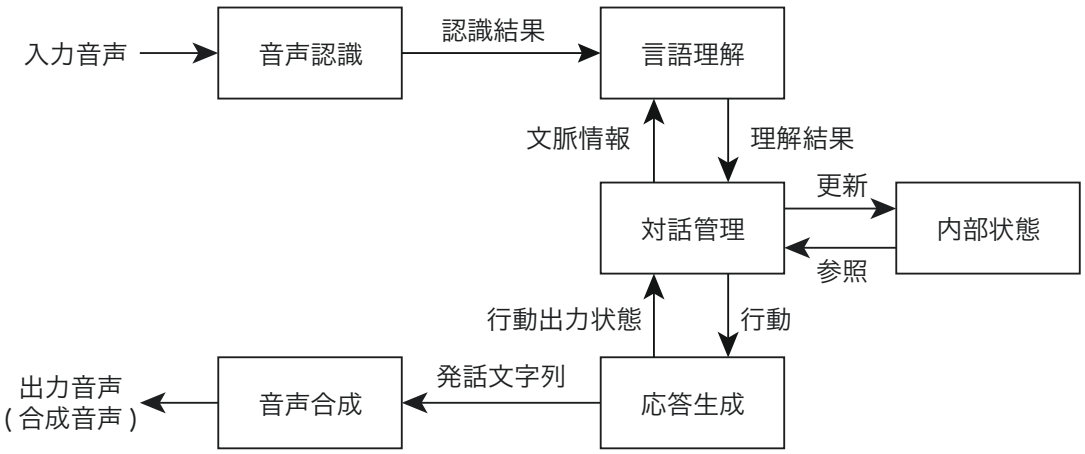


図 1.2: 一般的な音声対話システムの構成

コミュニケーションにおける笑いの役割の観点から快の笑い、社交上の笑い、緊張緩和の笑いの 3 つに分類している (表 1.1 参照)。このことからもわかるように、非語彙音の笑い声だけを取ってみても、様々な情報が伝達されていることがわかる。

1.1.2 人間と機械のコミュニケーション

近年では、人間同士のコミュニケーションだけではなく、人間と機械のコミュニケーションについても関心が高まっている。人間と機械の間のコミュニケーションに自然言語を使用するシステムのことを対話システムと呼び、特に自然言語を媒介するメディアとして音声を使用するシステムのことを音声対話システムと呼ぶ。

音声対話システムは音声認識、音声合成、自然言語処理、会話分析といった様々な研究分野の技術・成果が統合されたものである。音声対話システムの基本的な構成を図 1.2 に示す。音声対話システムは大きく 5 つのモジュールで構成される。入力音声は音声認識部により文字列に変換される。そして、変換された文字列は言語理解部により構文解析・意味解析が行われ、その結果が対話管理部に渡される。対話管理部では内部状態を参照・更新しながら入力に対して適切な行動を選択する。応答生成部では、その行動を反映する応答を (発話文字列) を生成する。発話文字列は音声合成部に渡され、その内容に沿った合成音声

が出力される。このような一連のプロセスを繰り返すことで人間と機械の対話を実現している。

音声対話システムは様々な場面で目にすることができる。コールセンターにおける電話応対システム [14,15] や車のカーナビゲーション [16] などは音声対話システムの代表的な実用例である。より身近な実用に、Apple の Siri や docomo のしゃべってコンシェルといったスマートフォン向けの音声対話エージェントなども登場している。現在、音声対話システムの実用例のほとんどは機械にある命令を実行させるためのコマンドとして音声を使用するタスク指向型の音声対話システムである。また、最近ではソフトバンクの Pepper のように、基本的にはタスク指向型の音声対話システムであるが、雑談を楽しむような非タスク指向的な用途で使用する需要も高まっている。

しかし、現在実用されている音声対話システムは、雑談を楽しむような用途として利用するのに不十分である。これは、これまでの音声対話システムは朗読音声や演技音声などを用いて構築されており、人間同士の対話で用いられている対話音声を想定したものではないからである。対話音声は朗読音声の対立概念である自発音声であり、前川は自発音声と朗読音声の大きな違いとして

- 言い間違い、言い淀み、言い直しといった非流暢性の存在
- 音声の特徴における聞き手の特性の有無

があると述べている [17]。また、日常に溢れる音声の多くは自発音声であり、自発音声を処理することができない音声処理システムおよび音声対話システムの実用性は極めて限られているとも指摘しており、朗読音声を対象とした音声対話システムの限定性を支持している。

また、朗読音声は主に言語的な情報のみを対象とすれば良いのに対し、対話音声ではノンバーバル情報をも取り扱わなければならない。なぜなら、対話音声は人間同士の対話、つまりコミュニケーション場面で用いられる音声であり、これを取り扱うということは人間と機械のノンバーバルコミュニケーションを実現しなければならないことに他ならないためである。

音声対話システムは前述したように様々な要素技術の集大成であるため、全ての要素技術において対話音声を取り扱うことを実現した音声対話システムは

存在しない。しかしながら、それぞれの要素技術においては対話音声を対象とするために必要となる技術の研究が行われている。

1.1.3 音声対話システムに関する研究

音声認識部に関する研究

対話音声は基本的に話す内容が事前に与えられておらず、発声するまでに発話内容のプランニングが行われる。このプランニングによる時間を埋めるなどのために「あのー」や「えー」のようなフィラーが挿入されることがある。Goto らは、フィラーがコミュニケーションにおいて重要な役割を果たしていると考え、自発音声認識においてフィラーを自動検出する手法について検討している [18]。Goto らは F0 の変化とスペクトル傾斜を用いてフィラーを検出する手法を提案しており、適合率 91.4%、再現率 84.9%を達成している。

言語理解部に関する研究

対話音声を扱う上では、何を話しているかという言語的な意味に加えてどのような意図や感情なのかといったパラ言語情報をも理解しなければならない。

藤江らは、発話に含まれる韻律情報から、forward selection 法により主要な特徴量を抽出し、それを用いて発話が肯定的であるか否定的であるかを識別する手法を提案し、その有効性を確認している [19]。対話例を以下に示す。

U: お昼ごはんなんだけど、どこかいいところないかな

R: カレーなんかどう

U: カレーか (否定的)

R: それじゃあ、弁当なんかどうかな

U: ああ、弁当ね (肯定的)

R: 弁当なら近くにホカ弁があるよ

ここで、U:はユーザーの発話、R:はシステムの発話を表す。ユーザーの応答が否定的であれば代案を、肯定的であれば具体的な提案を行うシステムとなっている。

対話管理部に関する研究

音声対話システムの構築において最も重要な要素技術が対話管理であり、様々な側面から研究が進められている。対話管理の根幹となる部分は言語処理であるが、音声対話システムにおいては言語処理は音声認識 (あるいは感情認識) や音声合成と統合して実現されるものであるため、言語処理単独というよりはそれらの技術と一体となって進められていることが多い。

NTT コミュニケーション基礎科学研究所は対話管理に着目した音声対話システムの研究として、ユーザーの音声入力に対して相槌を打つシステムである「飛遊夢」を開発した [20]。このシステムでは音声認識部に不特定話者による連続音声認識器 VoiceRex [21] を用いており、音声認識結果を逐次出力する。逐次的に出力される認識結果に対して、システムが相槌を打つか、ユーザーの理解結果を確認するための確認発話を行なうか、ユーザーに情報を要求する要求発話なのかを判断し、決定する。

応答生成部に関する研究

実際の対話場面では、聞き手が話し手の発話を最後まで聴く場合もあれば、途中で遮って割り込む場合がある。堂坂らは、ユーザーに伝達済みの情報を逐次管理しながら、発話を生成する手法を提案し [20]、システムの発話中にユーザーが割り込むと、その時点で伝達済みの情報と照合してユーザーの意図を理解するシステムを構築した。ユーザーが話の進め方を変更する意図を持っていた場合、対話管理部は発話を中断することを命じ、ユーザーの意図に合致するようにシステム応答文を変更したうえで、発話を再開する。

音声合成部に関する研究

従来、音声合成研究においては音声によるテキストの流暢な朗読を目標とした研究が精力的に進められてきており、現在では、非常に高品質なテキスト音声合成システムが市販されるまでに実用化されている。

しかしながら、それらはあくまで朗読調の音声を合成するためのものであり、対話音声に要求される言語的メッセージ以外のメッセージを伝達するに至っていない。対話音声の合成に関する研究は、これまで述べた各分野の研究と比較するとまだまだ限定的であるのが現状である。対話音声を対象とした音声合成に関する研究はほとんど行われておらず、感情音声合成に代表される演技音声を用いた擬似的な検討に留まっている。

感情音声合成は、代表的なパラ言語情報である感情を合成音声に反映させるための技術である。感情音声合成において、感情は「怒り」や「悲しみ」といった基本感情が取り扱われることが多く、それらの感情が反映された音声を合成する。飯田らは「怒り」、「喜び」、「悲しみ」、「驚き」、「平静」に相当する演技音声コーパスを収録し、波形接続方式によって感情音声を合成する手法を提案している [22]。また、都築らは「喜び」、「怒り」、「悲しみ」、「平静」の感情に対応する感情音声コーパスを収録し、各感情ごとに統計モデルで音響的特徴をモデル化し、そのモデルに基づいて感情音声を合成する手法を提案している [23]。

1.1.4 対話音声を対象とした音声合成

先にも述べたように、音声対話システムに関連する研究において、対話音声を対象とした音声合成の研究は他と比較しても少ない。音声合成研究、特にコーパスに基づくアプローチを取る手法では、大規模な音声データが要求される。対話音声を始めとする自発音声は、条件の統制が取れないことによる収録の困難性や収録した音声に対して適切な情報のアノテーションを行なう膨大なコストの問題から、大規模なコーパスを確保することが難しかったことが 1 つの原因に挙げられる。少なくとも、国内では「日本語話し言葉コーパス」(CSJ) が公開されるまで、大規模な自発音声コーパスは存在せず、一般的な分析および検討を行なうことができなかった。また、CSJ であっても話者の感情や意図といっ

たノンバーバルな情報はアノテーションされておらず、それらを反映するような自発音声の合成の検討は行えないでいた。しかし、現在では宇都宮大学パラ言語情報研究向け音声対話データベース (UUDB) [24] や感情評定値付きオンラインゲームチャットコーパス (OGVC) [25] といったノンバーバルな情報のアノテーションが施された自然対話音声コーパスが登場しており、対話音声合成を研究する土台が揃いつつあると言える。

対話音声を合成し、人間と機械のノンバーバルコミュニケーションを実現するためには、少なくとも 1.1.1 節で述べた音声から知覚されるパラ言語情報を表現することや笑い声に代表される非語彙音の合成を達成しなければならない。そしてそれは現在行われている感情音声合成の枠組みをそのまま使用しているだけでは達成することはできないと考えられる。

まず、大きな問題として自然対話音声から知覚されるパラ言語情報をどのように記述するかという問題がある。現在、感情音声合成のほとんどは基本感情を取り扱っているが、対話音声から知覚される感情を基本感情のカテゴリだけで表現可能であるかは不明である。OGVC では自然対話音声に対してカテゴリで表現される感情記述が与えられているが、基本感情の他に「その他」というカテゴリが用意されており、評価が難しいものはそこに属する。UUDB ではカテゴリよりもより抽象的な次元を用いて感情を始めとするパラ言語情報の記述を行っている。

更に、記述されたパラ言語情報をどのように合成音声に反映させるかという問題もある。カテゴリで表現された感情であれば、感情音声合成の枠組みをそのまま使用することができる。しかし、UUDB のように次元によって表現されたパラ言語情報は感情音声合成の枠組みにそのまま乗せることはできない。

合成された音声を評価する手法が確立されていないという問題もある。従来の音声合成の評価では主に明瞭性や原音声との一致性が評価の対象とされてきた。しかし、対話音声の評価ではそのような観点だけではなく、意図したパラ言語情報が聞き手に知覚されているかという観点からの評価も必要であり、そしてそれは時としてこれまでの評価項目で良いとされていた基準とは異なる場合がある。例えば、眠い状態を伝える音声を合成する場合を想像してほしい。話者の眠気を伝える音声として、明瞭性の高い音声が適切であると言えるだろう

か。このように、これまでの明瞭性の評価とパラ言語情報の評価をどのように統合するかという問題を孕んでいる。

また、非語彙音の合成についてはこれまでほとんど行われていないというのが現状である。非語彙の中でも笑い声は比較的メジャーな研究対象であるが、それでも限られた研究しか行われていない。笑い声の合成に関する研究において、自然対話中の笑い声を対象としたものは今のところ存在せず、映像刺激などによって誘発された笑い声を対象としたものしか存在しない。これはすなわち、コミュニケーション場面で用いられている笑い声を対象としていないことを意味し、コミュニケーションでどのような笑い声が必要か、必要である場合にどのような方法で合成しわけるのかといったことが検討されていないことを意味する。

1.2 研究の目的

人間同士のコミュニケーションでは、言語的なチャネルを用いた言語コミュニケーションと言語的ではないチャネルを用いたノンバーバルコミュニケーションが行われている。近年では、人間同士のコミュニケーションだけではなく、人間と機械のコミュニケーションについても関心が高まっている。これまで、人間と機械のコミュニケーションは音声を介して特定のタスクを達成するという目的で使用されることが多かったが、擬人化エージェントや対話ロボットなどの登場によって、音声を単なる指令コマンドとして使うという用途だけではなく、機械とのコミュニケーション自体を楽しむといった用途の需要も高まってきている。そして、そのためには人間同士のコミュニケーションのように、機械とのノンバーバルコミュニケーションの実現が必要である。

本研究では、人間と機械のノンバーバルコミュニケーションを視野に入れた、ノンバーバル情報を表現可能な音声合成の実現を目的としている。人間同士のコミュニケーションにおいては音声に付随して伝達されるパラ言語情報や非語彙音を通してノンバーバルコミュニケーションが実現されている。そこで、本研究では、

- 感情を中心としたパラ言語情報を反映可能な音声合成技術の確立
- 代表的な非語彙音である笑い声の合成

について論じる。実際の自然対話で使用されている音声、非語彙音を対象としたこれらの検討はこれまでほとんど行われていない。本研究では実際の自然対話における音声、非語彙音を対象とすることにより、これらを扱う困難性および要求される対処なども明らかにする。

1.3 本論文の構成

本論文は「序論」から「結論」までの 6 章で構成される。

第 2 章では、対話場面で用いられているノンバーバル情報について述べる。対話場面で要求されるノンバーバル情報表現を述べ、音声合成において、どのようなノンバーバル情報を表現することが必要なのかを論じる。

第 3 章では、ノンバーバル情報を表現可能な音声合成手法について述べる。本章では、代表的なノンバーバル情報であるパラ言語情報を対象とし、パラ言語情報の違いを合成音声に反映することができる音声合成方式を提案する。この章では統計的パラメトリック音声合成方式におけるモデルパラメータをパラ言語情報を説明変数とする重回帰モデルに基づいて変換することにより、パラ言語情報の反映された音声を合成する方式について論じる。また、本研究では実際の対話場面において伝達されるパラ言語情報に焦点を当て、自然対話音声コーパスを使用する。これまでのパラ言語情報を表現するための音声合成研究の多くは自然対話音声コーパスではなく、演技音声コーパスを採用してきた。そのため、自然対話音声を使用するうえでの問題点などが不明瞭であった。本章では、自然対話音声コーパスを使用する際に生じる問題点として、統計的パラメトリック音声合成における統計モデルパラメータの過推定問題について論じる。そして、統計モデルパラメータの過推定問題を解決するために、統計モデルパラメータのロバストな推定方法を提案する。合成された音声に対して客観評価および主観評価を行い、パラ言語情報の表現能力および自然性を評価し、パラ言語情報の反映手法およびロバスト推定手法の有効性について論じる。

第 4 章では、自然対話音声コーパスにおける笑い声を対象とした分析について述べる。対話場面では、ノンバーバル情報は言語音からだけではなく、非言語からも伝達される。笑い声はノンバーバル情報を伝達する代表的なキャリアで

ある。従来、特に笑い声の合成に関する研究はお笑い映像やライブによって誘発された笑い声を用いて行われている。これは多様な笑い声の中でも、限られた条件で発せられた笑い声のみを対象としていることを意味しており、対話場面に要求される笑い声を考慮していない。また、そもそも対話場面でどのような笑い声が要求されるのかということすらも検討されていないのが現状である。そこで、本研究では対話場面における笑い声に注目する。そのために、自然対話音声コーパスに含まれる笑い声を対象とした分析が行われる。しかし、自然対話音声コーパスには笑い声に関する詳細な情報が与えられていない。そこで本章では、まず自然対話音声コーパスに対して行われた笑い声のアノテーションについて述べる。その後、アノテーション結果をもとに行われた笑い声の分析について述べ、笑い声合成に要求される情報について論じる。

第 5 章では、自然対話コーパスに含まれる笑い声を用いた笑い声合成手法について述べる。自然対話音声コーパスに含まれる笑い声の合成では、お笑い映像などによって誘発された笑い声を合成する従来の研究と比較して十分な量の笑い声を用意することができない。そのため、少ないデータでも効率的に笑い声の音響的特徴をモデル化できるポテンシャルを持つ統計的パラメトリック音声合成手法を用いて笑い声合成を行なう。また、自然対話音声に含まれる笑い声には単独の笑い声だけではなく、話してから笑う、または笑ってから話すといった言語音に付随する笑い声が存在する。従来のお笑い映像などによって誘発された笑い声には、言語音に付随する笑い声は少なく、そのような発話を合成する際にはたとえ笑い声自体の品質が良くても発話全体としての自然性が低下する問題が報告されている。本研究では、自然対話音声コーパスを使用することにより、笑い声が置かれている文脈や状況を考慮した笑い声を合成することを提案する。統計的パラメトリック音声合成方式では、合成単位である音素の音響的特徴の変動要因 (コンテキスト) に依存するコンテキスト依存モデルを学習することによって、音素の置かれている環境および文脈を考慮している。そこで、本研究ではその方法を応用し、笑い声に対して適切なコンテキストを定義し、笑い声コンテキストに依存したモデルを構築することによって、状況に合った笑い声を合成することを提案する。提案手法の有効性は主観評価実験により論じられる。

最後に第 6 章では、本論文の結論を述べる。

第2章　対話におけるノンバーバル情報

2.1　はじめに

我々はコミュニケーションを行なう際、言語のみによるコミュニケーション以外にも、ジェスチャやボディランゲージ、表情、視線といった言語以外の様々なものを用いている。場合によっては、その言語以外によるものによる比重が多い場合も存在する。このように言語以外によるものを用いたコミュニケーションをノンバーバルコミュニケーションと呼ぶ。また、これまでの例には発声によるものを含んでいなかったが、音声の言語的メッセージ以外の情報もノンバーバルに含めるという認識が一般的である。

ノンバーバルコミュニケーションに関する研究の起源は Birdwhistell による動作学 (Kinesics) まで遡ることができる [26]。また、Hall は非言語によるコミュニケーション手段を沈黙のことば (silent language) と称し、人間と文化の関係の調査している [27]。現在では、ノンバーバルコミュニケーションの重要性が広まりつつあり、表情合成や音声合成におけるノンバーバル情報表現といった分野の研究も行われ始めている。

本章ではノンバーバルコミュニケーションの構成要素について述べる。その後、構成要素の 1 つである paralanguage(周辺言語、言語周辺、パラ言語とも) について説明する。そして、パラ言語によって伝達される情報について論じる。最後に、現状の音声合成研究におけるパラ言語表現の現状について述べ、今後要求される音声合成におけるパラ言語表現について論じる。

2.2　ノンバーバル情報の範囲

「ノンバーバル」という言葉は「言語的でない」という意味しか持っておらず、非常に対象の広い用語となっている。そのため、単にノンバーバル情報と言う場合にも研究者や文献によっては対象とする範囲が異なる場合があり、しばしば混乱を招く事態が発生する。

例えば Winner らは、ノンバーバルを非常に限定的なものとして扱っており、情報の送信者および受信者の積極性 (あるいは意識的であるか言った方が適切かもしれない) の及ぶ範囲をノンバーバルとしてしている。例えば、場所を伝える際に目的地の方向を指で指すといった身体動作はノンバーバルに含めるが、場所を思い出している間に無意識的に視線が泳いだり、顔を様々な方向に動かすといった動作はノンバーバルに含めないとしている。

一方、Bull は Winner らの扱うノンバーバルの範囲を少し広げ、無意識的であっても情報の受信側に送信者のメッセージが正しく受信されるのであれば、それもノンバーバルの対象とするとしている。また、Ekman と Friesen はノンバーバルの範囲を更に広げるべきであると述べており、広げた上でそれらを更に細分化する必要があると主張している。このように、ノンバーバル情報が対象とする範囲だけでも様々な説が存在しており、どこまでを対象とするかという単純な部分についてもまだまだ自明とは言えないのが現状である。

2.3　ノンバーバル情報を伝達するメディア

どこまでをノンバーバル情報とするかを厳密に定義することが難しいということは先に述べたが、現在では、ノンバーバルの範囲を比較的広く捉えることが一般的になりつつある。そのため、ノンバーバル情報を伝達するメディアも非常に多様に取り扱われている。例えば、Vargas はノンバーバルコミュニケーションの構成要素として、視線や動作、パラ言語、対人的空間といったメディアを挙げている [28, 29]。また、石黒らは図 2.1 に示すようにコミュニケーションにおけるメディアを言語的か否か、更に音声的であるか否かの観点から分類している [2]。このようにノンバーバル情報を伝達するメディアの分類について

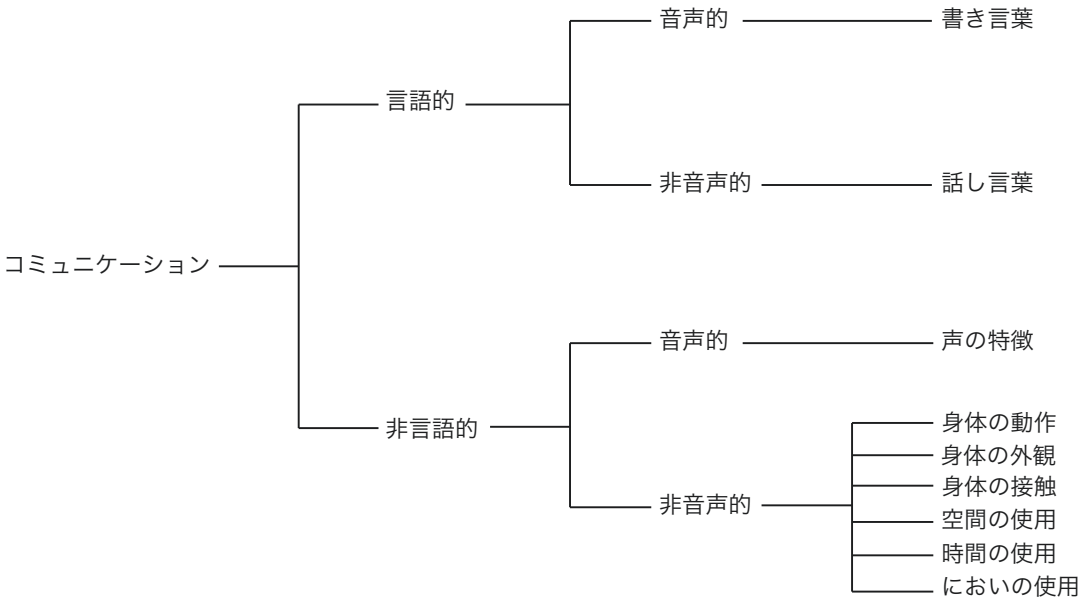


図 2.1: 石黒らによるコミュニケーションの分類 [2]

も様々な流儀がある。ここでは多くの分類で一般的に採用されているメディアについて述べる。

2.3.1 身体動作

身体動作は身体の運動によってメッセージを表現するメディアであり、身振り、表情、視線などがこれに属する。

身振りは身体の一部、あるいは複数の部位を動かす、または一定に静止するなどする動作である。身振りは発話などと同時に行われることもあり、メッセージの伝達性を助けることもある [30,31]。道を説明する時に、言葉だけではなく身振りも使用するといった行動はその 1 つの例であろう。更に、話し手が意図した情報の伝達を助けるだけではなく、聞き手が受ける話し手の印象などを決定する情報になることもある。藤原は、発話だけでなく手の動きという身振りを伴うことにより、より知的かつ自信があるように知覚される傾向があることを示している [32]。また、大神は「わかりやすい説明」をする話し手は身振りを多用している傾向があったことを報告している [33]。

表情は感情表出や感情知覚に大きく寄与するメディアである。機嫌を確かめる

ことを「顔色をうかがう」と言ったりすることからもわかるように、話し手あるいは聞き手の感情や態度を判断する重要な手がかりである。Ekman と Friesen は表情による感情表出・理解の様式は文化を超えたある程度までの共通性があるということを述べている [34]。普遍性を超えるものについては「表示規則 (display rules)」という概念を導入することで説明している。また、Ekman は表情の解剖学的変化に基づいた分類手法として、顔面動作符号化法 (FACS) [35] なども提案しており、表情研究の多くで採用されている。

視線はコミュニケーションにおいて様々な役割を果たす重要なメディアである。「目は口ほどにものを言う」は、コミュニケーションにおける視線の重要性を象徴する慣用句と言える。例えば視線の交差 (俗に言う目が合う) は対話の開始になることもあれば、対話中の相手への発話権を移譲などの役割を果たすこともある。また、聞き手の視線は対話や対話内容についての関心や意欲に関係するといったことが指摘されている [36]。

2.3.2 対人距離と対人接触

対人距離と対人接触は対話者間の空間的関係に関するメディアである。対人距離は対話の内容や対話参加者の人数などにも関係し、心地よい距離を得るように調節される。「パーソナルスペース」として知られるように、人には他人が入り込むと不快に感じる空間があり、それは相手との関係性などによって伸縮する。Hall は対人距離を親密さのレベルおよびインタラクションの種類に応じて、密接距離、個体距離、社会距離、公衆距離の 4 つに分類している [37]。

対人接触はパーソナルスペースに深く入り込んで相手の身体に触れる行為であり、対話者間の関係が親密である場合に行われる動作であったり、逆に険悪な間柄であったときに一方が他方に加えるいやがらせや暴力的な行動であったりする。

2.3.3 パラ言語

パラ言語はノンバーバルメディアの中で唯一の音声要素である。既に述べたように、パラ言語は音声に寄り添って言語以外のメッセージを伝達するメディ

アであり、声の抑揚などの韻律や声質といったものによって表現される。また、パラ言語は音声に限ったメディアであるとされることもあるが、Trager のように、音声だけではなく非語彙音をもパラ言語に含める場合もある [9]。

音声によって伝達される情報は、藤崎によって大別された言語情報、非言語情報、パラ言語情報という分類が広く受け入れられていることと、藤崎の分類では非言語情報に属していた感情が、近年ではパラ言語情報に属する傾向にあることは序論で述べた。本研究においても、感情もパラ言語情報に含めている。

パラ言語情報はどのような音響的特徴によって表現されるか、またどのように知覚されるかという観点から精力的に研究されている。前川らは話し手の発話意図と音響的特徴の関係について実験的に調査しており、発話の持続時間やピッチ (特に句末における動き) などによって話し手が意図を表現しようとしていることや、それらの意図は音声の韻律的特徴および分節特徴よってほぼ正確に伝達されうることを示している [38]。河津らは音声から知覚される感情の程度とピッチの関係を基本周波数生成パターンに基づいて分析している [39]。また、平賀らはピッチや振幅の変化パターンを用いて音声から感情情報を抽出する検討を行っている [40]。

パラ言語情報の認識に関する研究も盛んに行われている。特に音声による感情認識は最もポピュラーであり、代表的なものとしては音声の音響的特徴から基本感情を認識するという研究がある [41, 42]。また、Lee らや Ang らは主に韻律に関する情報を用いて、話者のネガティブな感情の認識について検討している [43, 44]。

パラ言語情報を反映した音声を合成する研究には感情音声合成がある。感情音声合成については既に簡単に述べており、序論では感情音声コーパスと統計モデルに基づく音声合成方式によって感情音声の音響的特徴が自然に反映される音声合成の研究について紹介した [22, 23]。また、規則に基づいて音響的特徴を制御し、感情を反映させる検討もされている [45, 46]。

2.4 人間と機械におけるノンバーバルコミュニケーション

人間は相手が機械であっても、コミュニケーションによって何らかの人間性を感じ、特に、対象が擬人化されている場合に顕著で、対話対象の身体動作、容貌や衣服などの外観や、音声に含まれる周辺言語を通して伝達される性別、年齢などの影響も受けることが明らかになっている [47]。このようなことから、現在、人間と機械のインタラクションにおいてもノンバーバル情報を活用することが重要視されており、様々な検討が行われている。ここでは、主にパラ言語情報に注目した人間と機械のノンバーバルコミュニケーションに焦点を置き、これまでにどのような研究行われているかについて述べる。また、そのうえで今後どのようなノンバーバル研究が要求されるかについても述べる。

音声対話システムにおいてパラ言語情報を活用する試みとして、ユーザの態度を認識する研究が藤江らによって検討されている [19]。藤江らによる研究では、ユーザの音声入力から抽出される韻律特徴と頭部動作からユーザの態度が肯定的であるか、あるいは否定的であるかを認識する手法を提案している。更に、実際に音声対話システムにその手法を組み込み、人と同等の認識能力を持つことと、ユーザがシステムの提案に肯定的であれば和台を継続し、否定的であれば代案を返すといったような、従来にない効率的な対話が実現されていることを報告している。

また、山口らは対話の文脈に応じて適切な形態の相槌を打つことが可能な傾聴対話システムの実現を目標として、相槌の形態の分析および予測を検討している [48]。この検討において、山口らは先行発話の韻律的／分節的特徴から相槌の生起位置および形態予測の有効性を示している。

パラ言語情報を取り扱う音声対話システムはいくつか存在する。しかしながら、そのほとんどは認識と、認識結果からを受けての対話管理の部分に限られており、出力にまで及んでいない。今後はパラ言語情報を反映した出力が可能となるように、音声合成部および合成に入力するための対話管理 (合成時にどのようにパラ言語情報を入力するかなど) の研究が必要とされる。

また、非語彙音を出力可能な音声対話システムは実現されていない。最も研

究されていると思われる笑い声ですら、合成に関しては現在まで予備的な検討しか行われていない。今後人間と機械のコミュニケーションを人間同士のように豊かにするためには、こういった非語彙音についても取り扱っていく必要がある。

2.5　おわりに

本章では、対話におけるノンバーバル情報について述べた。まず、ノンバーバル情報の多様性について述べ、その後、各メディアの中でも音声あるいは音声ではないが口などから発せられる音に注目して現在どのようなことが研究されているのかについて述べた。また、人間同士の対話だけではなく、人間と機械の対話についても注目し、人間と機械のノンバーバルコミュニケーションの現状と問題点について述べ、人間と機械の円滑なコミュニケーションを実現するために、今後どのようなことが要求されるかについて論じた。

第3章　パラ言語情報を反映可能な音声合成

3.1　はじめに

近年、音声合成の主なゴールは与えられたテキストに対して自然な音声を合成することであった。そのような研究の多くの貢献によって、現在ではかなり自然な音声の合成が実現されており、音声合成研究の焦点は言語音の明瞭に合成することから、話者の感情や意図、態度といったような対話音声のパラ言語的な側面に移りつつある。つまり、現在の音声合成研究の目標は言語情報だけではなくパラ言語情報をも表現することになっている。

本章の目的は、自然対話音声に含まれる豊富なパラ言語情報を表現可能な音声合成技術の確立である。現在、商用の音声合成器のほとんどはパラ言語情報の表現を考慮しておらず、アナウンサーやナレーターが原稿を読み上げる時の音声のような朗読音声を合成している。しかしながら、現在では会話エージェントやロボットなどが登場してきており、そのような応用では朗読音声ではなく豊かなパラ言語情報を表現可能な音声合成の需要が高まっている。

そのような音声合成を実現するために、パラ言語情報を表現可能な対話音声合成が望まれる。対話音声合成を実現する方法には以下の 3 つの手法が考えられる。

- フォルマント合成
- 波形接続型音声合成
- 統計的パラメトリック音声合成

フォルマント合成は音声生成モデルに基づいた音声合成方式であり、実際の音声データを必要としない音声合成方式である。そのため、大規模な音声コー

パスを用意する必要ない。フォルマント合成によってパラ言語情報を表現する研究には文献 [49] があり、基本 6 感情である「怒り」、「喜び」、「悲しみ」、「恐れ」、「驚き」、「嫌悪」の反映された音声を合成している。

波形接続型音声合成は、音声データベースに含まれる音声素片を接続することによって音声を合成する手法である。この手法は大規模な音声データベースが要求されるものの、実際の音声波形を使用するため、高品質な音声を合成することが可能である。Bulut らは基本感情ごとに収録された感情音声データベースを用いて感情音声合成システムを構築することを提案している [50]。また、Iida らも波形接続型合成方式によって感情音声合成を行っている [51]。更に、ダイフォン単位での波形接続によって感情音声を合成する研究も行われている [52, 53]。

統計的パラメトリック音声合成は、音声の音響的特徴を統計モデルによってモデル化し、そのモデルに基づいて音声パラメータを合成する手法である。一般に、そのモデルには隠れマルコフモデル (Hidden Marukov Model: HMM) が用いられる [54]。Yamagishi らは統計的パラメトリック音声合成方式を用いて感情音声合成を実現している [55]。この研究では、基本感情ごとに統計モデルを構築することで、各感情に対応した音声を合成することができる。

統計的パラメトリック音声合成によるアプローチは、与えられたデータのパラ言語的な特徴が反映されたモデルパラメータを自然に反映することができる。また、モデルパラメータを様々な適応技術によって柔軟に変換することでもパラ言語情報を反映することができることから、パラ言語情報を表現する音声合成はとして有望視されている。Tachibana らは HMM のモデルパラメータを線形補完することで様々な感情や発話スタイルの音声の合成を実現している [56]。Nose らによる研究では、HMM モデルパラメータをアフィン変換することによって、発話スタイルの制御を行っている [57]。更にはクラスタ適応学習手法を取り入れ、各感情に対応するクラスタの重み付け和によって様々な感情に対応する音声合成技術も提案されている [58]。

これまで、統計的パラメトリック音声合成においてパラ言語情報を表現する研究のほとんどは演技音声を用いて行われてきた。この演技音声とは、発話内容と感情や発話スタイルといったパラ言語情報が指定された音声であり、多くの場合はプロのナレーターや声優によって演技される。しかしながら、演技音

声は自然音声とは異なった音声であるため、音声対話などの場面で使用される音声としてはあまり適していないと考えられる。これは、いわゆる芝居がかった音声が対話で用いられていないことからも明らかであろう。

また、演技による感情音声が対話音声合成に適していない理由に、多くの場合で基本感情が用いられていることが挙げられる。演技音声を収録する場合、多くの研究で「怒り」や「悲しみ」といったような典型的な感情を指示して感情音声を収録している。しかしながら、実際の対話場面ではそのような典型的な感情が表出されることは稀であり、微妙なニュアンスの感情であったりと非常に多様な感情が存在する。演技音声によるコーパスではそのような多様な感情が含まれておらず、対話音声合成に使用するのは難しい。

一方で、演技音声ではなく自然対話音声コーパスを用いた対話音声合成の研究は驚くほどに少なく、HMM を用いたもの [59,60] やディープニューラルネットワークを用いたもの [61] などの一部の研究しかない。一般に、自然対話コーパスを用いた音声合成システムの構築は朗読音声と比較して難しいと言われている。その大きな要因が自然対話音声の多様性である。自然対話音声は韻律的にも多様であり、韻律パラメータを扱うだけでも非常に困難を伴う [62,63]。また、自然対話音声は演技音声とは異なり発話内容やパラ言語情報の統制が取られていない。したがって、音韻やパラ言語情報のバランスがとれたデータを収集することができない。すなわち、自然対話音声コーパスを用いた対話音声合成を行う場合には、実際に話された音声に対して言語情報やパラ言語情報のアノテーションといった作業が要求される。しかし、上述したように、現在主流であるカテゴリによる表現では自然対話音声から知覚されるパラ言語情報を記述することは難しい。このように、自然対話コーパスを使用した対話音声合成では乗り越えなければならない課題が多く存在する。

自然対話音声から知覚されるパラ言語情報を記述する手法として、カテゴリではなく次元による記述を行うアプローチが広く受け入れられるようになってきた。近年では、感情を Valence-Acitivation-Dominance (VAD) モデルといった 3 つの次元で表現する手法が広まっている [64,65]。次元による手法では、カテゴリよりも抽象的な次元を用いてパラ言語情報を記述するため、微妙なニュアンスを含む多様なパラ言語情報を表現できる可能性があり、自然対話音声のパ

ラ言語情報を表現可能な対話音声合成に有効であると考えられる。

自然対話音声コーパスと次元に基づくパラ言語情報記述を用いてパラ言語情報を表現可能な音声合成を実現する研究に Mori らによる研究がある [66]。Mori らによる研究では、HMM 音声合成手法を採用しており、音声の音響特徴量の変動要因に次元で記述されたパラ言語情報を加えることで、様々な感情を反映可能な音声合成を実現した。しかしながら、この研究では、HMM を学習する学習データに存在する範囲の感情しか表現することができないという問題があった。

そこで、本研究では自然対話音声コーパスと次元によるパラ言語情報記述を用いた新しい音声合成手法として、重回帰 HSMM [57, 67] に基づくパラ言語情報の表現方法を提案する。重回帰 HSMM では、音声の音響的特徴をモデル化するためのモデルパラメータを少数の次元で構成される重回帰モデルによって変換する手法である。その少数の次元にパラ言語情報を表現するための抽象次元を当てはめることでパラ言語情報を表現可能な音声合成を実現する。このような重回帰モデルを用いることで、学習データにない範囲のパラ言語情報をも合成音声に反映することができる可能性がある。

自然対話音声コーパスでは言語的・パラ言語的な統制が取られていないことは先に述べた。したがって、コーパスではパラ言語情報に偏ったデータになっていることが多い。また、自然対話音声コーパスは一般に演技音声のコーパスと比べて小規模であることが多い。そのようなコーパスから重回帰モデルを推定すると過推定が生じてしまう可能性がある。本章では、自然対話音声コーパスを用いた場合でも安定した重回帰モデルを推定するためのロバストな推定手法も提案する。

3.2 パラ言語情報の記述

3.2.1 記述手法

本研究では、パラ言語情報の定義を「音声に付随して伝達される話者の意図、心的態度あるいは感情を含む状態」とする。感情だけに焦点を当ててみても、自然対話においては「怒り」や「悲しみ」といったような典型的な感情が現れる

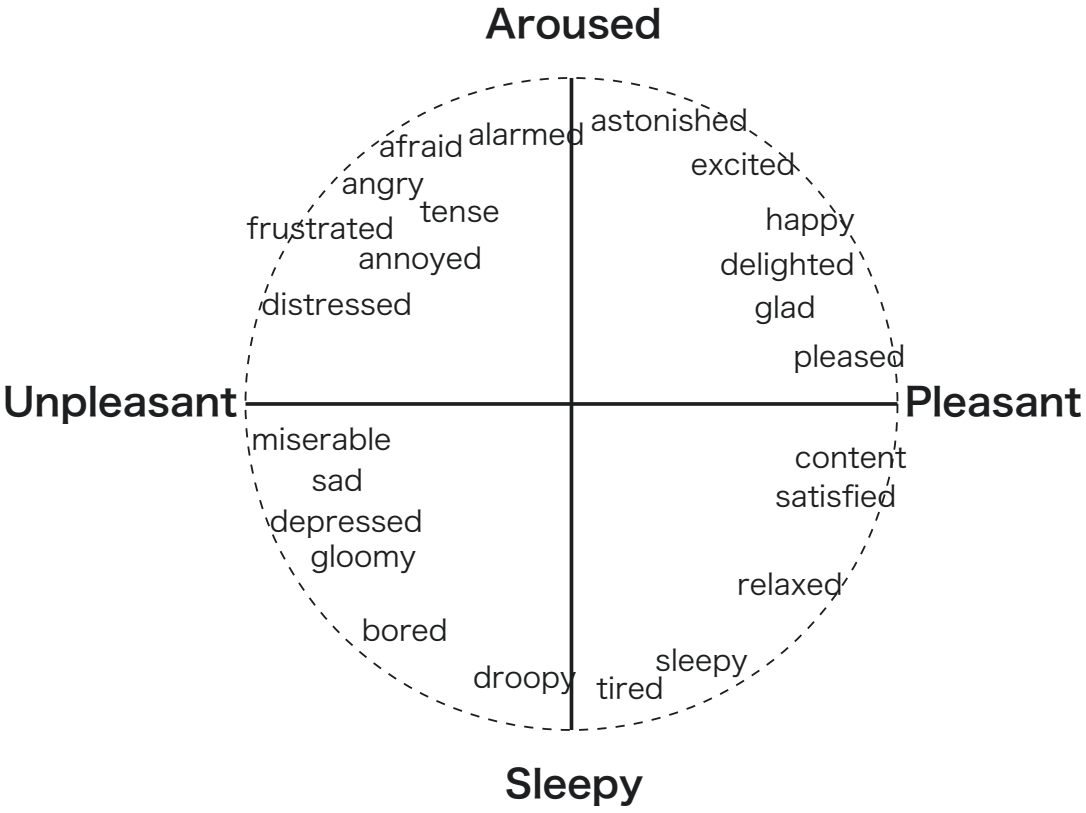


図 3.1: 感情の円環モデル [3]

ことはむしろ稀である。そのため、自然対話における音声が伝達するパラ言語情報を基本感情説のようなカテゴリで表現することは困難である。

そこで、本研究ではパラ言語情報の記述法として、感情の次元説に基づく手法を用いる。感情の次元説では、感情は離散的に特定できるものではなく、抽象的な次元の 1 つのベクトルとして表されると考える。図 3.1 は次元説の立場から Russell によって提唱された感情の円環モデルである。この円環モデルでは「快-不快」および「覚醒-睡眠」という 2 つの抽象的な次元上のベクトルとして感情を記述している。このようにカテゴリよりも抽象的な次元による記述を用いることによって、自然対話に頻繁に出現する典型的ではない感情の表現が可能になることを期待している。

3.2.2 パラ言語情報の記述を持つ自然対話コーパス

本研究では次元説に基づくパラ言語情報の記述を用いる。次元説に基づくパラ言語情報の記述を持つ自然対話コーパスに、宇都宮大学パラ言語情報研究向け音声対話データベース (UUDB) [24] がある。UUDB は自然で表情豊かな音声対話に見られる多様な音声学現象および言語学的現象の研究への用途を開発目的とした音声コーパスであり、親近性の高い大学生 7 ペア (女性 12 名、男性 2 名) による自然な対話音声が収録されている。

UUDB に収録されている対話音声は「4 コマまんが並べ替え課題」[68] というタスクを与えて収録されている。「4 コマまんが並べ替え課題」は，対話者がばらばらにされた 4 コマまんがの 2 コマ分をそれぞれ持ち，相手が持つ残り 2 コマの内容を見ることができない状態で、対話により元の順番を推定するというタスクである。このタスクにより収録された総発話数は合計で 4840 発話であり、収録されている音声は表情豊かなものとなっている。

UUDB の大きな特徴は、収録されている全ての発話に対して感情の次元説に基づいて記述されたパラ言語情報が与えられていることである。UUDB では、以下に示す 6 軸の評価項目を用いた抽象次元によりパラ言語情報を記述している。

- 快-不快
  - 発話者の気持ちや気分の良し悪し
- 覚醒-睡眠
  - 発話者の心理活動の活発さ
- 支配-服従
  - 発話者が相手とのコミュニケーションをリードしているか
- 信頼-不信
  - 発話者が相手のことをどの程度信じているか
- 関心-無関心

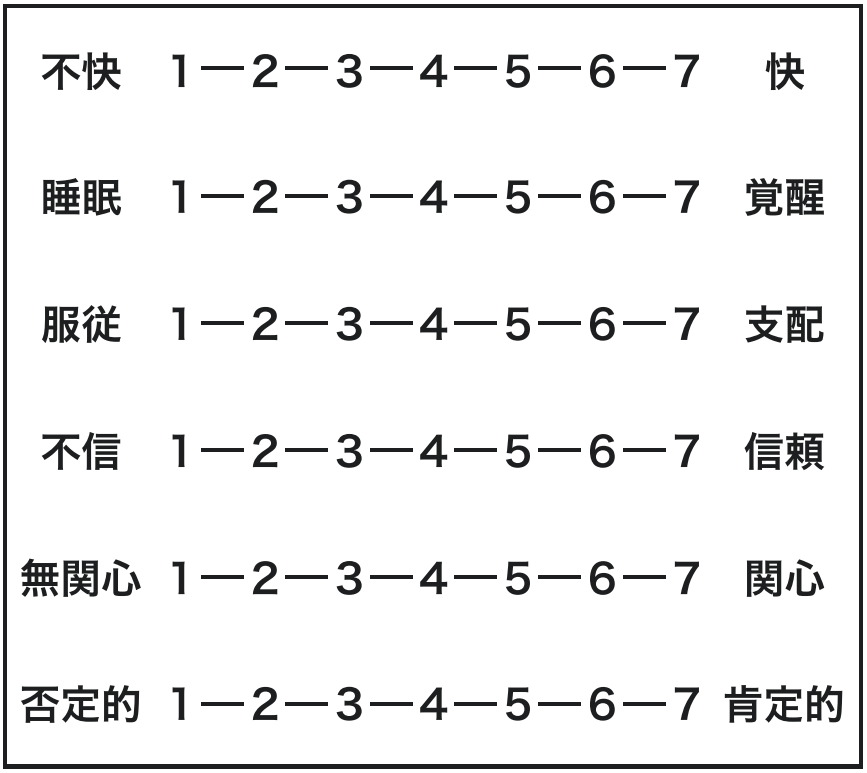


図 3.2: UUDB のパラ言語情報に関する記述

  - 発話者が相手や相手の発話に対してどの程度関心や興味があるか

- 肯定的-否定的
  - 発話者が相手の発話をどの程度肯定的に評価しているか

「快-不快」、「覚醒-睡眠」は話者自身の感情状態に関係する評価項目である。「支配-服従」、「信頼-不信」は話者間の対人関係に関する評価項目であり、「関心-無関心」、「肯定的-否定的」は相手に対する態度に関する評価項目である。UUDB では、これらの 6 軸の項目に対して図 3.2 に示すように 7 段階の評価を施している。このラベリングの安定性については文献 [69] によって確かめられている。ラベリングは 3 名によって行われており、評価の一貫性、平均評定値との類似性および設定項目の独立性が保たれている。

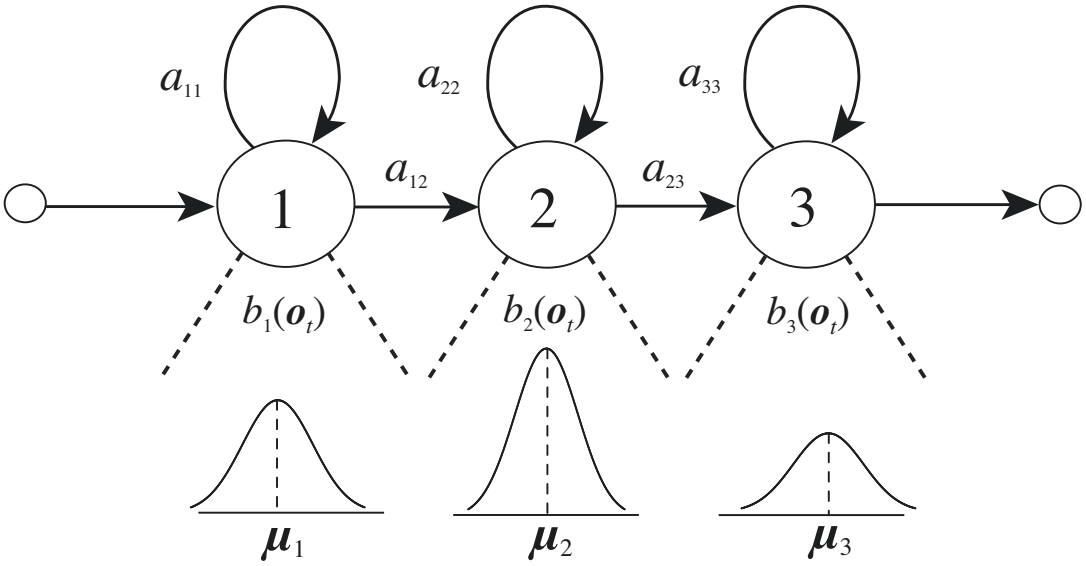


図 3.3: 隠れマルコフモデルの例

3.3 パラ言語情報を反映する音声合成手法

3.3.1 HMM 音声合成

隠れマルコフモデル (Hidden Markov Model: HMM) は、観測シンボルの出力確率分布を持つ信号源 (状態) と状態遷移確率によって定義される確率モデルである。図 3.3 は、音声関連の応用でよく用いられる left-to-right 型の HMM である。ここで、a_{ij} は状態 i から状態 j への状態遷移確率であり、$b_i(\cdot)$ は状態 i における観測シンボルの出力確率分布を表す。一般の HMM では、任意の状態間での遷移が許されているが、音声のモデル化においては因果性を表現するために、状態を横 1 列に並べたときに左方向への遷移がない left-to-right 型の HMM が用いられる。音声関連の応用において、観測シンボルはメルケプストラムや音声の基本周波数などの音声の音響特徴量によって構成される音響特徴量ベクトル $\boldsymbol{o}_t$ となる。また、音響特徴量ベクトルは連続量であり、出力確率分布 $b_i(\cdot)$ には一般にガウス分布が用いられる。

また、HMM は時々刻々と変化する音声の音響的な特徴のモデル化に用いられるが、音声の重要な特徴の 1 つである音韻継続長に相当するパラメータが存在しない。そこで、音韻継続長に相当するパラメータとして、状態継続長分布 HMM に

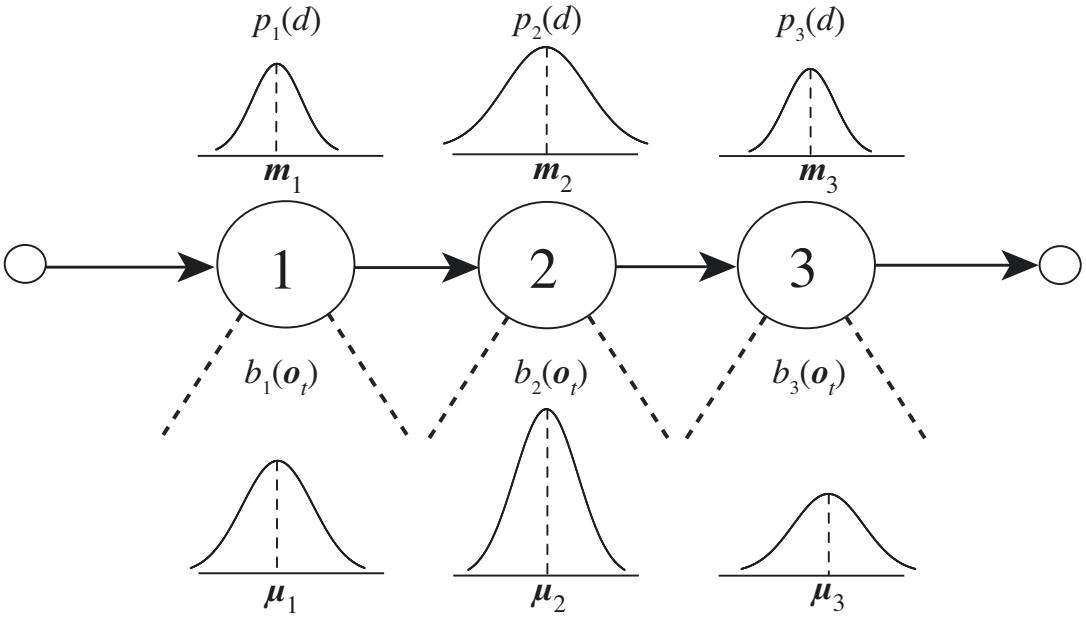


図 3.4: 隠れセミマルコフモデルの例

組み込んだ隠れセミマルコフモデル (Hidden Semi-Markov Model: HSMM) [70] が用いられることがある。図 3.4 は left-to-right 型の HSMM の例である。各状態は自己遷移を行わず、各状態に滞在する時間を状態継続長分布として含有している。

HMM 音声合成の基本構成を図 3.5 に示す。学習部では、与えられた音声データから、各分析フレームごとに静的特徴量としてスペクトルパラメータと基本周波数 (F0) パラメータが抽出される。そして、静的特徴量から動的特徴量を計算し、それらを合わせた音響特徴量ベクトルを用いて HMM が学習される。

HMM の学習は通常、音素単位で行われる。音素は同じ音素であっても、音韻環境により音響的特徴が大きく異なることが知られている。また、音韻環境だけでなく、アクセント型や文長といった音素文脈 (コンテキスト) によっても大きく異なることが知られている [71]。そのため、コンテキストを考慮したコンテキスト依存 HMM が学習される。コンテキストの組み合わせは膨大であり，全ての組み合わせを網羅的に学習することは現実的ではないため、決定木を用いたコンテキストクラスタリングが行われる [72]。

合成部では、与えられた入力テキストをコンテキスト依存ラベル列に変換し、

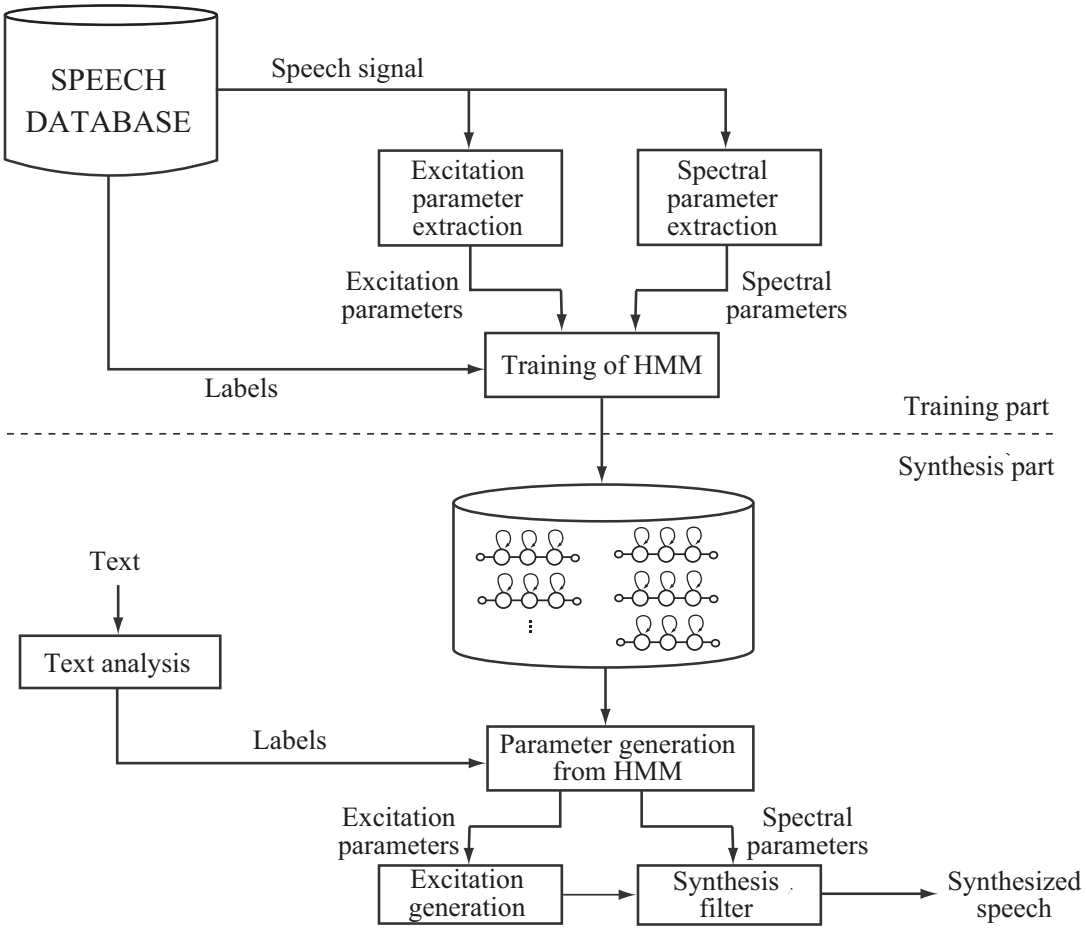


図 3.5: HMM 音声合成の構成 [4]

各ラベルに対応するコンテキスト依存 HMM を連結した文 HMM を作る。そして、文 HMM から動的特徴量を考慮したパラメータ生成アルゴリズムを用いて F0 パラメータとスペクトルパラメータ系列を生成し、ボコーダ方式により合成音声を生成する。

3.3.2 重回帰 HSMM

HMM 音声合成方式では、出力確率分布のモデルパラメータを適切に変換することで、音声の音響的特徴を変化させることができる。本節では、その変換手法の 1 つである重回帰 HSMM [57, 67] について述べる。

HSMM の各状態の出力確率分布および継続長分布に、単一ガウス分布を仮定した場合を考える。状態 i における出力確率分布 $b_i(\cdot)$ と継続長分布 $p_i(\cdot)$ は、次式で定義される。

$$b_i(\boldsymbol{o}) = \mathcal{N}(\boldsymbol{o}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \tag{3.1}$$

$$p_i(d) = \mathcal{N}(d|m_i, \sigma_i^2) \tag{3.2}$$

ここで、$\boldsymbol{o}$ および d は、それぞれ観測ベクトルと状態 i に滞在する継続時間である。$\boldsymbol{\mu}_i$ と $\boldsymbol{\Sigma}_i$ は出力確率分布の平均ベクトルと共分散行列であり、m_i と σ_i^2 は継続長分布の平均と分散である。重回帰 HSMM では、出力確率分布の平均ベクトルと継続長分布の平均が次式で表すような線形モデルで表現可能であると仮定する。

$$\boldsymbol{\mu}_i = \boldsymbol{H}_{b_i}\boldsymbol{\xi}, \tag{3.3}$$

$$m_i = \boldsymbol{H}_{p_i}\boldsymbol{\xi} \tag{3.4}$$

$$\boldsymbol{\xi} = [1, v_1, v_2, \cdots, v_L]^\top = [1, \boldsymbol{v}^\top]^\top \tag{3.5}$$

ここで、$\boldsymbol{H}_{b_i}$ および $\boldsymbol{H}_{p_i}$ は、それぞれ $M \times (L+1)$ および $1 \times (L+1)$ の行列である。また、M は観測ベクトルの次元数であり、L はベクトル $\boldsymbol{v}$ の次元数である。$\boldsymbol{v}$ は重回帰モデルにおける説明変数によって構成されるベクトルであり、制御ベクトルと呼ばれる。すなわち、出力確率分布および継続長分布は次式で表される。

$$b_i(\boldsymbol{o}|\boldsymbol{H}_{b_i}, \boldsymbol{v}, \boldsymbol{\Sigma}_i) = \mathcal{N}(\boldsymbol{o}|\boldsymbol{H}_{b_i}, \boldsymbol{v}, \boldsymbol{\Sigma}_i) \tag{3.6}$$

$$p_i(d|\boldsymbol{H}_{p_i}, \boldsymbol{v}, \sigma_i^2) = \mathcal{N}(d|\boldsymbol{H}_{p_i}, \boldsymbol{v}, \sigma_i^2) \tag{3.7}$$

重回帰 HSMM における回帰行列を推定する一般的な手法は、最尤基準を用いることである。観測系列 $\{\boldsymbol{O}^{(1)}, \cdots, \boldsymbol{O}^{(K)}\}$ と、それに対応する制御ベクトル系列 $\{\boldsymbol{v}^{(1)}, \cdots, \boldsymbol{v}^{(K)}\}$ が与えられた時、出力確率分布に対する回帰行列は、以下の補助関数を最大化することによって推定される。

$$Q_{b_i}(\lambda, b_i) = \sum_{k=1}^{K}\sum_{t=1}^{T_k}\sum_{d=1}^{t}\gamma_t^d(i)\sum_{s=t-d+1}^{t}\log b_i(\boldsymbol{o}_s^{(k)}|\boldsymbol{H}_{b_i}\boldsymbol{v}^{(k)}, \boldsymbol{\Sigma}_i) \tag{3.8}$$

ここで、T_k は k 番目の観測系列 $\boldsymbol{O}^{(k)}$ のフレーム数、$\boldsymbol{o}_s^{(k)}$ は $\boldsymbol{O}^{(k)}$ の時刻 s における観測ベクトルである。また、$\gamma_t^d(i)$ は状態占有確率と呼ばれるものであり、時刻 $t-d-1$ から時刻 t までに状態 i に滞在する確率である。補助関数を最大にする回帰行列は、上式を $\boldsymbol{H}_{b_i}$ で微分して 0 とおくことで求めることができ、

$$\sum_{k=1}^{K}\sum_{t=1}^{T_k}\sum_{d=1}^{t}\gamma_t^d(i)\sum_{s=t-d+1}^{t}\boldsymbol{\Sigma}_i^{-1}\boldsymbol{o}_s^{(k)}\boldsymbol{\xi}^{(k)\top} = \sum_{k=1}^{K}\sum_{t=1}^{T_k}\sum_{d=1}^{t}\gamma_t^d(i)\cdot d\cdot\boldsymbol{\Sigma}_i^{-1}\boldsymbol{H}_{b_i}\boldsymbol{\xi}^{(k)}\boldsymbol{\xi}^{(k)\top}$$

を得る。この式を整理することで、再推定式

$$\overline{\boldsymbol{H}}_{b_i} = \left\{\sum_{k=1}^{K}\sum_{t=1}^{T}\sum_{d=1}^{t}\gamma_t^d(i)\sum_{s=t-d+1}^{t}\boldsymbol{o}_s^{(k)}\boldsymbol{\xi}^{(k)\top}\right\}\cdot\left\{\sum_{k=1}^{K}\sum_{t=1}^{T}\sum_{d=1}^{t}\gamma_t^d(i)\cdot d\cdot\boldsymbol{\xi}^{(k)}\boldsymbol{\xi}^{(k)\top}\right\}^{-1}$$

を得る。また、同様の手順により、継続長分布についての回帰行列 $\boldsymbol{H}_{p_i}$ の再推定式は次式となる。

$$\overline{\boldsymbol{H}}_{p_i} = \left\{\sum_{k=1}^{K}\sum_{t=1}^{T}\sum_{d=1}^{t}\gamma_t^d(i)\cdot d\cdot\boldsymbol{\xi}^{(k)\top}\right\}\cdot\left\{\sum_{k=1}^{K}\sum_{t=1}^{T}\sum_{d=1}^{t}\gamma_t^d(i)\boldsymbol{\xi}^{(k)}\boldsymbol{\xi}^{(k)\top}\right\}^{-1}$$

3.3.3 重回帰 HSMM に基づくパラ言語情報の反映

本研究では、重回帰 HSMM における制御ベクトルにパラ言語情報についての抽象次元を用いることで合成音声にパラ言語情報を反映させる。すなわち、

$$\boldsymbol{v} = [v_{\text{pleasantness}}, v_{\text{arousal}}, \cdots] \tag{3.9}$$

と定義することで、パラ言語情報についての抽象次元によって音声の音響的特徴を線形変換する。

図 3.6 に、重回帰 HSMM に基づく音声合成方式を用いたパラ言語情報の反映手法のダイアグラムを示す。学習部では、まず音声データベースの音声波形からスペクトルおよび音源パラメータが抽出される。その後、音声波形に対応するコンテキストラベルとパラ言語情報ラベルから、重回帰 HSMM モデルパラメータである回帰行列が推定される。

合成部では、コンテキストラベルとパラ言語情報ラベルを与える。その後、パラ言語情報ラベルと推定された回帰行列から計算される平均パラメータを持つ HSMM から音響特徴量が生成され、ボコーダにより音声が合成される。

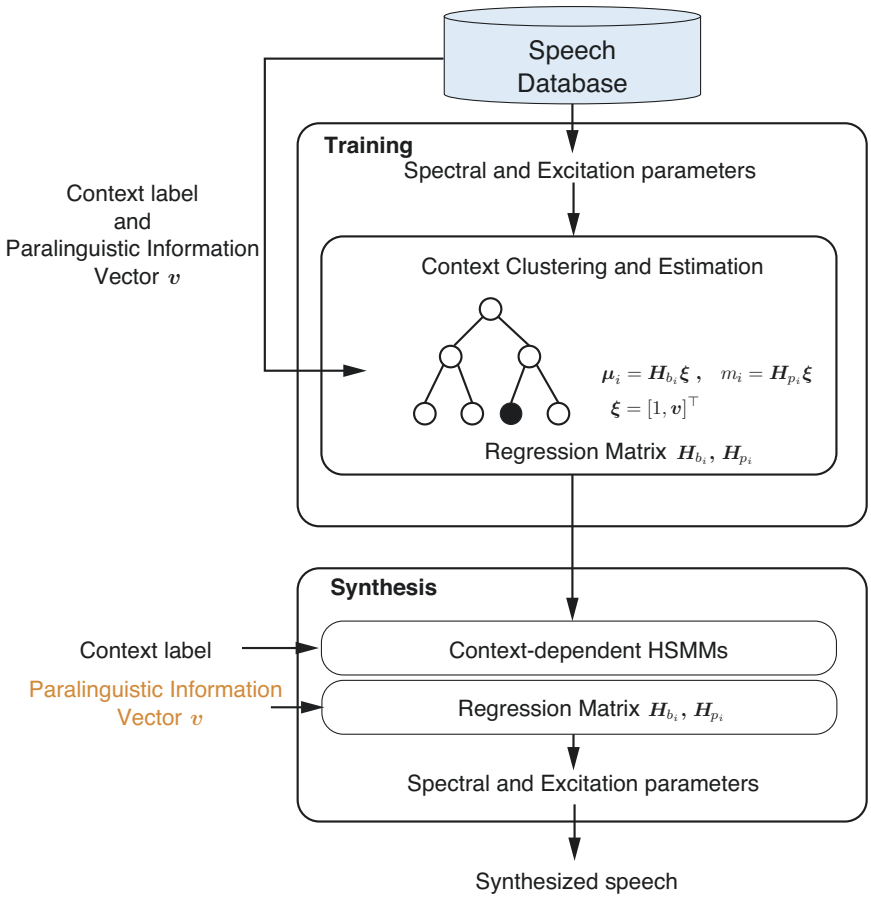


図 3.6: 重回帰 HSMM に基づく音声合成によるパラ言語情報の反映

3.3.4 自然対話コーパスにおける過推定問題

先に述べたように、重回帰 HSMM における回帰行列は一般に最尤基準によって推定される。この最尤基準は、データが十分に集められる場合には適切なパラメータを推定できる。重回帰 HSMM に基づく従来の音声合成に関する研究 [57, 67] では、原稿を用意し、音韻やパラ言語情報を統制した比較的多くのデータを収集することで、十分なデータを確保していた。

しかしながら、自然対話コーパスを用いる場合では原稿を用意するといったような統制をとることができない。また、自然対話コーパスの規模は一般に演技音声のコーパスに比べて小さい。このように規模が小さく、かつ音韻・パラ言語情報の偏ったコーパスから最尤基準で回帰行列を推定すると、過推定問題

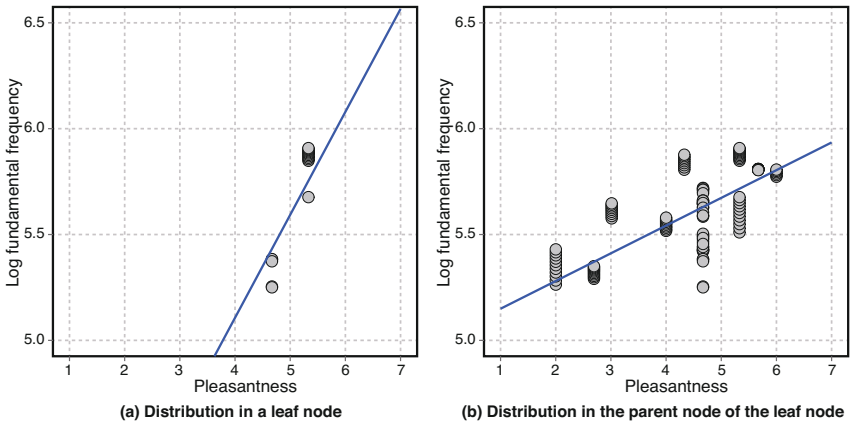


図 3.7: クラスタリングによるパラ言語情報の偏り

が生じる。

重回帰 HSMM に基づく音声合成方式では、決定木クラスタリングによって得られる木のリーフノードで回帰行列が推定される。決定木クラスタリングにおいてノードを分割する質問には、韻律的・分節的な質問が用いられる。決定木の構築の際にはパラ言語情報は使用されないが、適用される質問によっては、パラ言語情報の分布をより偏らせることがある。

その例を図 3.7 に示す。図 3.7 (a) は、あるリーフノードにおける「快-不快」の評価値に対する対数基本周波数の分布である。「快-不快」の評価値の分布が非常に狭い。一方、図 3.7 (b) は、そのリーフノードの親ノードにおける分布である。親ノードでは、パラ言語情報が広く分布していることがわかる。このように、適用される質問によってパラ言語情報の分布が偏ることがあり、そのような分布から推定される回帰行列は非常に極端な値を取ることがある。そのため、重回帰 HSMM に基づく音声合成方式において自然対話コーパスを用いる場合には、偏った分布でも回帰行列をロバストに推定する手法が必要となる。

3.3.5 重回帰 HSMM パラメータのロバストな推定

本研究では、重回帰 HSMM に基づく音声合成における回帰行列のロバストな推定手法として、最大事後確率 (Maximum a posteriori: MAP) 基準による回帰

行列の推定を提案し、推定式を導出する。MAP 基準では、以下の事後確率についての補助関数を最大化する。

$$Q_{b_i}(\lambda, b_i) = \sum_{k=1}^{K}\sum_{t=1}^{T_k}\sum_{d=1}^{t}\gamma_t^d(i)\sum_{s=t-d+1}^{t}\log b_i(\boldsymbol{o}_s^{(k)}|\boldsymbol{H}_{b_i}, \boldsymbol{v}^{(k)}) + \log P(\boldsymbol{H}_{b_i})$$

ここで、$P(\boldsymbol{H}_{b_i})$ は回帰行列 $\boldsymbol{H}_{b_i})$ の事前分布であり、以下で定義される行列正規分布 [73] である。

$$P(\boldsymbol{H}_{b_i}) = (2\pi)^{-\frac{M(L+1)}{2}}|\boldsymbol{\Omega}_{b_i}|^{-\frac{M}{2}}|\boldsymbol{\Phi}_{b_i}|^{-\frac{L+1}{2}}\cdot \exp\left\{-\frac{1}{2}\mathrm{tr}\left(\boldsymbol{H}_{\mathrm{b_i}} - \boldsymbol{W}_{\mathrm{b_i}}\right)^{\top}\boldsymbol{\Omega}_{\mathrm{b_i}}^{-1}\left(\boldsymbol{H}_{\mathrm{b_i}} - \boldsymbol{W}_{\mathrm{b_i}}\right)\boldsymbol{\Phi}_{\mathrm{b_i}}^{-1}\right\} \quad (3.10)$$

ここで、$\boldsymbol{W}_{b_i}, \boldsymbol{\Omega}_{b_i}$ および $\boldsymbol{\Phi}_{b_i}$ はそれぞれ $M\times(L+1), M\times M$ および $(L+1)\times(L+1)$ の行列である。$\boldsymbol{W}_{b_i}$ は行列正規分布における平均パラメータであり、$\boldsymbol{\Omega}_{b_i}$ と $\boldsymbol{\Phi}_{b_i}$ は分散パラメータに相当する。$\boldsymbol{\Phi}$ を単位行列 $\boldsymbol{I}_{(L+1)}$ であると仮定し、式 (3.3.5) を $\boldsymbol{H}_{b_i}$ で微分して 0 とおくと、

$$\sum_{k=1}^{K}\sum_{t=1}^{T_k}\sum_{d=1}^{t}\gamma_t^d(i)\sum_{s=t-d+1}^{t}\boldsymbol{\Sigma}_i^{-1}\boldsymbol{o}_s^{(k)}\boldsymbol{\xi}^{(k)\top} + \boldsymbol{\Omega}_{b_i}^{-1}\boldsymbol{W}_{b_i} = \sum_{k=1}^{K}\sum_{t=1}^{T_k}\sum_{d=1}^{t}\gamma_t^d(i)\cdot d\cdot\boldsymbol{\Sigma}_i^{-1}\boldsymbol{H}_{b_i}\boldsymbol{\xi}^{(k)}\boldsymbol{\xi}^{(k)\top} + \boldsymbol{\Omega}_{b_i}^{-1}\boldsymbol{H}_{b_i} \quad (3.11)$$

を得る。更に、$\boldsymbol{\Omega}_{b_i} = \tau_{\mathrm{out}}^{-1}\boldsymbol{\Sigma}_i$ として $\boldsymbol{H}_{b_i}$ について整理すると

$$\overline{\boldsymbol{H}}_{b_i} = \left\{\sum_{k=1}^{K}\sum_{t=1}^{T_k}\sum_{d=1}^{t}\gamma_t^d(i)\sum_{s=t-d+1}^{t}\boldsymbol{o}_s^{(k)}\boldsymbol{\xi}^{(k)\top} + \tau_{\mathrm{out}}\boldsymbol{W}_{b_i}\right\}\cdot \left\{\sum_{k=1}^{K}\sum_{t=1}^{T_k}\sum_{d=1}^{t}\gamma_t^d(i)\cdot d\cdot\boldsymbol{\xi}^{(k)}\boldsymbol{\xi}^{(k)\top} + \tau_{\mathrm{out}}\boldsymbol{\Phi}_{b_i}\right\}^{-1} \quad (3.12)$$

を得る。ここで、τ_{out} は出力確率分布についてのハイパーパラメータである。同様の手順により、継続長についての MAP 推定式は以下となる。

$$\overline{\boldsymbol{H}}_{p_i} = \left\{\sum_{k=1}^{K}\sum_{t=1}^{T_k}\sum_{d=1}^{t}\gamma_t^d(i)\cdot d\cdot\boldsymbol{\xi}^{(k)\top} + \tau_{\mathrm{dur}}\boldsymbol{W}_{p_i}\right\}\cdot \left\{\sum_{k=1}^{K}\sum_{t=1}^{T_k}\sum_{d=1}^{t}\gamma_t^d(i)\boldsymbol{\xi}^{(k)}\boldsymbol{\xi}^{(k)\top} + \tau_{\mathrm{dur}}\boldsymbol{\Phi}_{p_i}\right\}^{-1} \quad (3.13)$$

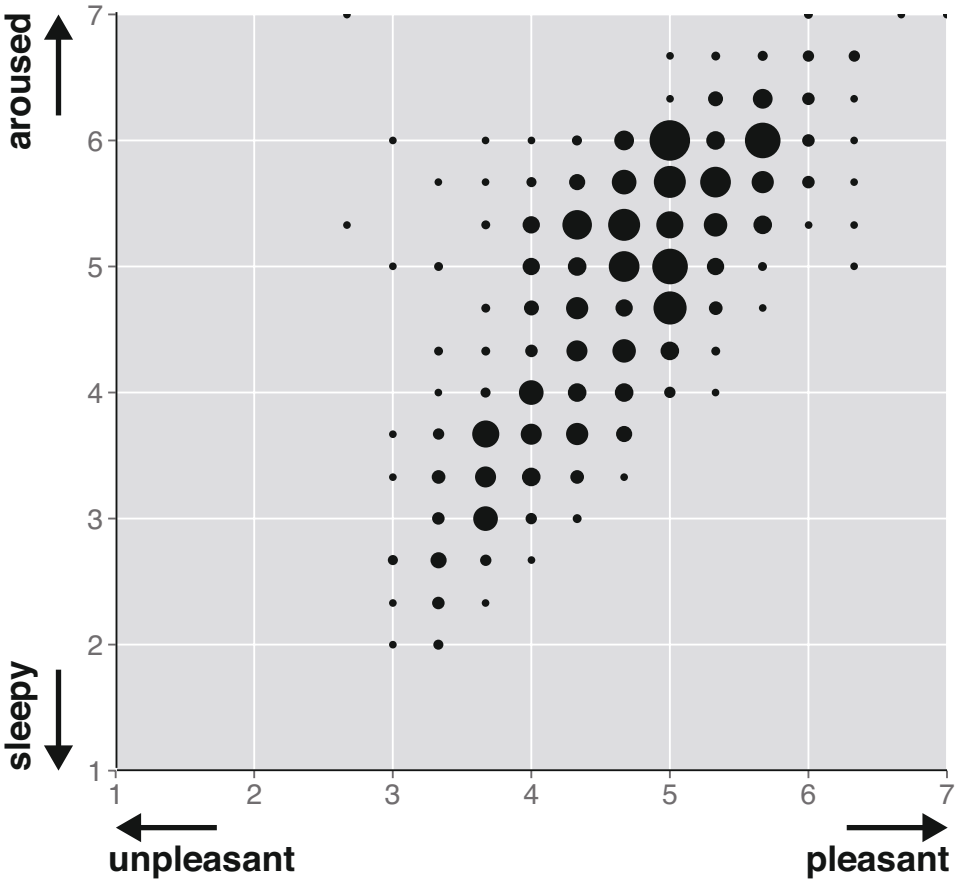


図 3.8: 話者 FTS の発話に与えられたパラ言語情報の分布

これらの式はデータが十分にある場合には、最尤推定における推定式に漸近し、データが少ない場合には事前分布の平均パラメータが支配的となる。この事前分布の平均パラメータを適切に設定することで、データが少ない場合でもロバストな回帰行列が推定できると期待される。

3.4 自然対話コーパスを用いた重回帰HSMMによる音声合成

3.4.1 合成条件

本研究では自然対話コーパスにUUDBを使用する。モデル学習および合成は、UUDBの話者FTSを対象とする。学習に使用するデータは589発話であり、合成には95発話内容を使用した。話者FTSはUUDBにおいて、最も感情表現の豊かな話者の1人であり、モデルの構築に適した話者であると考えられている。図3.8に、話者FTSの発話の「快-不快」、「覚醒-睡眠」の評価値の分布を示す。図より、FTSの発話は感情空間を広くカバーしていることがわかる。

音声分析は分析周期を5 msとしたSTARIGHT分析 [74] によって行われた。STRAIGHT分析によって得られたSTRAIGHTスペクトルから39次のメルケプストラム係数を抽出し、スペクトルパラメータに用いた。音源パラメータには基本周波数パターン生成モデル [75] によってスムージングされた対数基本周波数が用いられた。このスムージングは基本周波数の抽出誤りや図3.9に示すようなマイクロプロソディを取り除くために用いられ、これを行うことにより合成音声の品質が改善されることが報告されている [76]。また、混合励振源として、0–1 kHz, 1–2 kHz, 2–4 kHz, 4–6 kHzおよび6–8 kHz周波数帯域の非周期性指標の平均が用いられた。特徴ベクトルはこれらのパラメータと、それぞれのデルタおよびデルタデルタパラメータを加えた138次元のベクトルとした。

モデルは5状態のleft-to-right multiple-regression HSMM (MRHSMM) とした。重回帰モデルにおける説明変数にはUUDBの「快-不快」と「覚醒-睡眠」の2次元を使用した。すなわち、制御ベクトルは次式である。

$$\boldsymbol{v} = [v_{\text{pleasantness}}, v_{\text{arousal}}]^\top \tag{3.14}$$

ここで、$v_{\text{pleasantness}}$ と v_{arousal} はそれぞれ「快- 不快」および「覚醒-睡眠」に表す説明変数である。

前節で提案したMRHSMMにおける回帰行列のMAP推定の有効性を確認するために、以下に示す3つのモデルを使用して音声を合成した。

1. ML基準で推定したMRHSMM (ML)

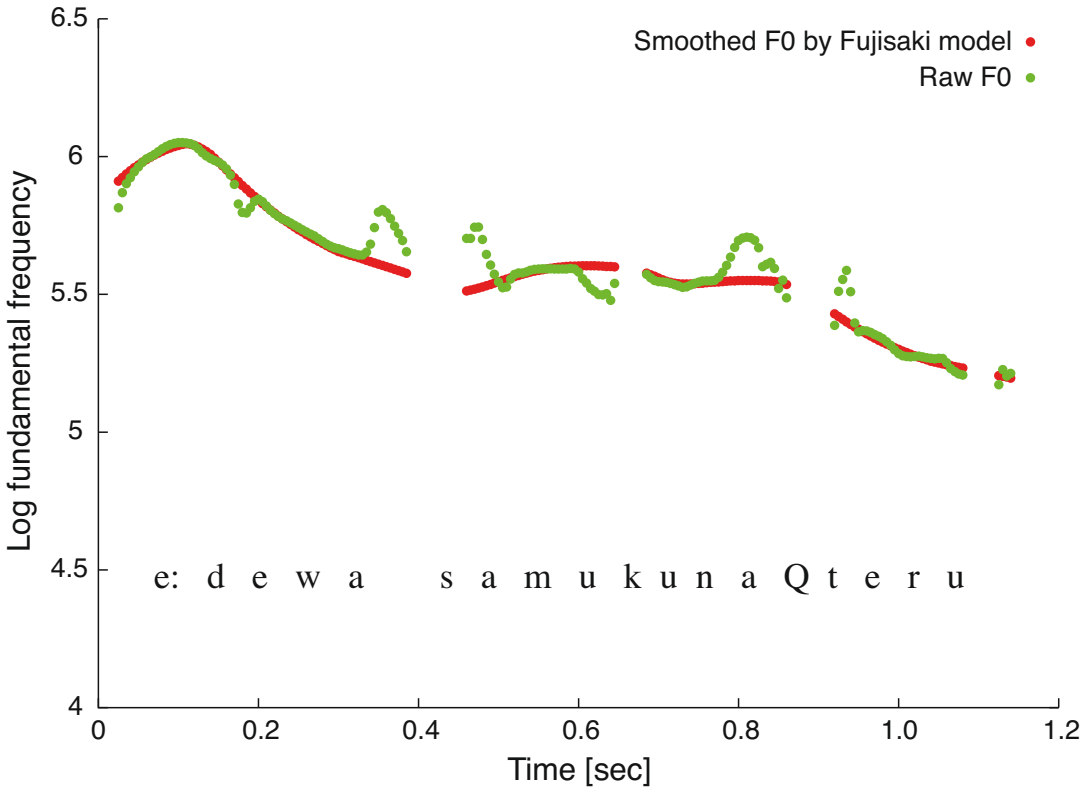


図 3.9: 基本周波数パターン生成モデルに基づく対数基本周波数のスムージング

2. MAP の近似手法 [67] で推定した MRHSMM (MAP-like)

3. MAP 基準で推定した MRHSMM (MAP-MRHSMM)

MAP-like では、次式で表す回帰行列の線形和によって回帰行列の推定を行った。

$$\overline{\boldsymbol{H}}_{b_i} = \frac{\tau_{\text{out}}\hat{\boldsymbol{H}}_{b_i} + \sum_{k=1}^{K}\sum_{t=1}^{T_k}\sum_{d=1}^{t}\gamma_t^d(i)\cdot d\cdot \boldsymbol{H}_{b_i}^{\text{ML}}}{\tau_{\text{out}} + \sum_{k=1}^{K}\sum_{t=1}^{T_k}\sum_{d=1}^{t}\gamma_t^d(i)\cdot d} \tag{3.15}$$

$$\overline{\boldsymbol{H}}_{p_i} = \frac{\tau_{\text{dur}}\hat{\boldsymbol{H}}_{p_i} + \sum_{k=1}^{K}\sum_{t=1}^{T_k}\sum_{d=1}^{t}\gamma_t^d(i)\cdot \boldsymbol{H}_{p_i}^{\text{ML}}}{\tau_{\text{dur}} + \sum_{k=1}^{K}\sum_{t=1}^{T_k}\sum_{d=1}^{t}\gamma_t^d(i)} \tag{3.16}$$

ここで、$\boldsymbol{H}_{b_i}^{\text{ML}}$ と $\boldsymbol{H}_{p_i}^{\text{ML}}$ は最尤基準で推定された回帰行列である。また、$\hat{\boldsymbol{H}}_{b_i}$ および $\hat{\boldsymbol{H}}_{p_i}$ はバックオフパラメータであり、今回の検討では MAP-MRSHMM における事前分布の平均パラメータに等しい。本研究では、回帰行列を推定するリーフノードの直近の親ノードで最尤推定された回帰行列が使用された。また、ハイパーパラメータ τ_{out} および τ_{dur} は、0.1 から 10000 の範囲で試行錯誤的に探索し、1000 に設定した。

MAP-MRHSMM では、事前分布の平均パラメータが必要となる。本研究では、MAP-like の場合と同様に回帰行列を推定する直近の親ノードで最尤推定された回帰行列を事前分布の平均パラメータとした。また、ハイパーパラメータ τ_{out} および τ_{dur} は、0.1 から 10000 の範囲で試行錯誤的に探索し、1 に設定した。

合成時には図 3.10 に示すパラ言語情報を与えた。ML 法において非常に低品質な音声が合成されることから (例: 20 kHz を超える基本周波数、通常音声の 200 倍の振幅)、与えられるパラ言語情報は 3 から 5 に制限された。この制限は ML 法に有利な条件だと思われるが、後の節で述べる自然性評価実験で使用する合成音声との条件を揃えるために定められた。

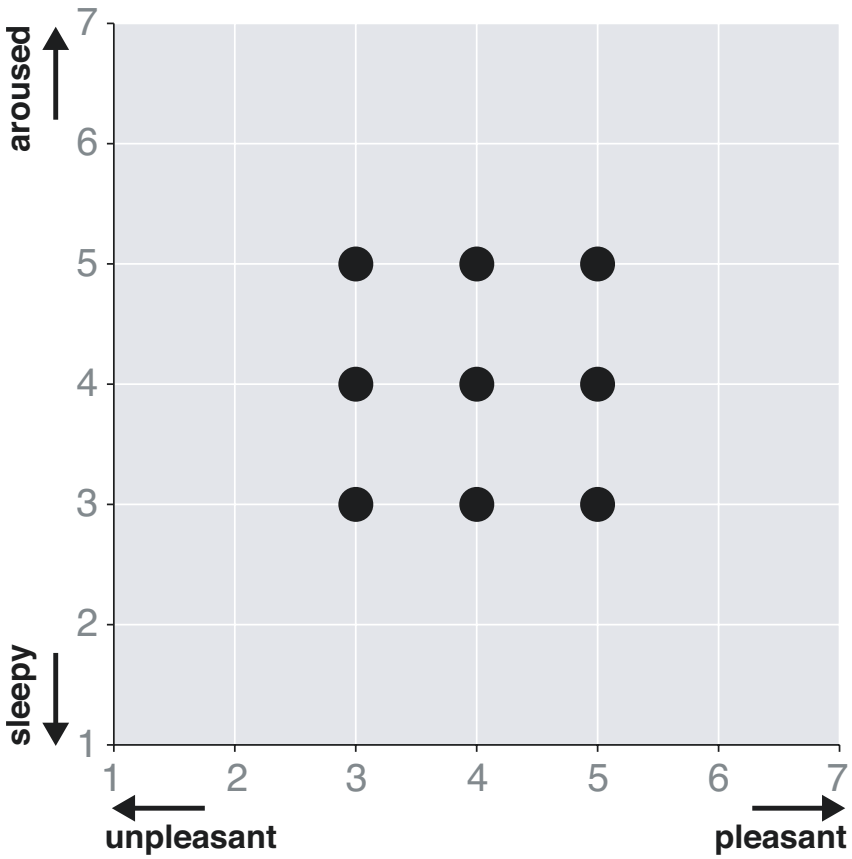


図 3.10: 合成時に与えたパラ言語情報

3.4.2 合成結果

合成された音声の例として、MAP-MRHSMM で合成された音声「そうだね」の対数基本周波数軌跡を図 3.11 に示す。この例では、「覚醒-睡眠」の次元を固定し、「快-不快」の次元を変化させた時の対数基本周波数を示している。図より、助詞である「ね」の部分において顕著な違いが現れていることがわかる。助詞「ね」は発話の末尾に存在し、パラ言語情報を伝達する重要なキャリアであると考えられるため、この部分に顕著な違いが現れているのは妥当な結果であると考えられる。

更に、音声「うん」を合成した時のランニングスペクトルを図 3.12 に示す。図 3.12 (a) および図 3.12 (b) ではスペクトルピークのゲインが大きくなっており、非常に低品質な音声が合成された。一方、提案法によって合成された「う

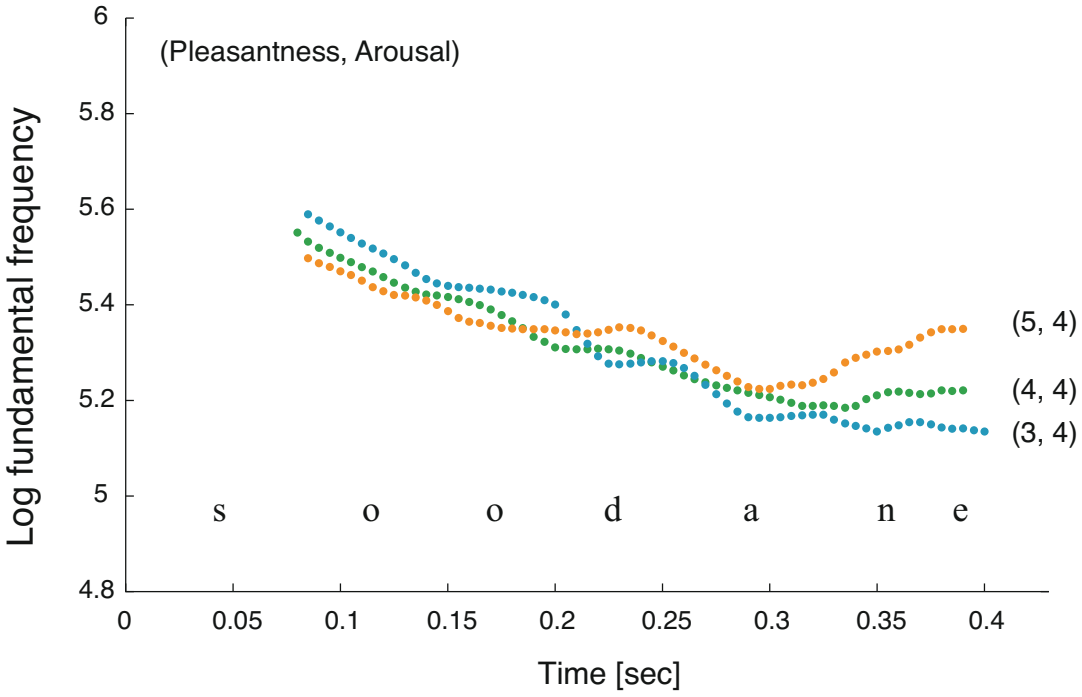
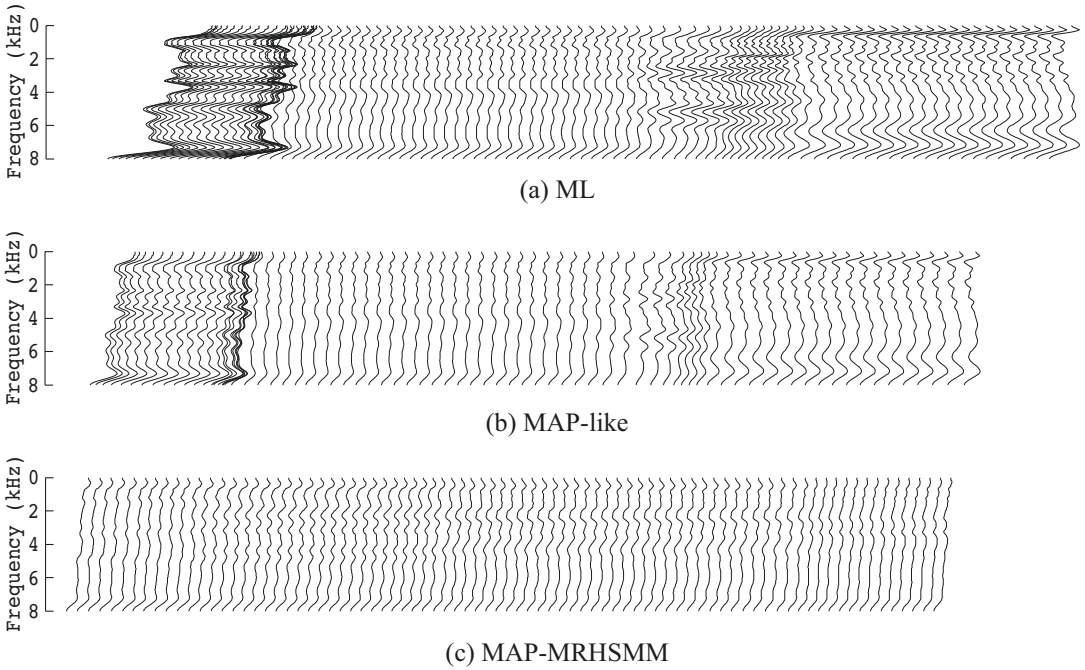


図 3.11: MAP-MRHSMM で合成された「そうだね」の対数基本周波数軌跡

ん」のランニングスペクトルではその現象が抑制されており、自然性が改善されていることがわかる。

提案法の有効性を更に詳しく確かめるために、合成された音声の音響特徴量の分布を調査した。図 3.13 に、合成された音声の 0 次メルケプストラム係数の最大値の分布を示す。0 次メルケプストラム係数はゲインに相当するパラメータであり、ML および MAP-like では非常に大きな値を持つ音声が存在しており、音声として破綻した音となっている。一方、MAP-MRHSMM では、大きな値が抑制されており、妥当な範囲に収まっていることがわかる。

同様に、対数基本周波数についての分布を図 3.14 に示す。図では、ML 法、MAP-like 法、MAP-MRHSMM によって合成された音声の対数基本周波数の平均値、最大値、最小値を示している。最も顕著な違いが見られたのは、図 3.14 (a) の平均値である。「覚醒-睡眠」が高くなるにつれて、全ての手法で基本周波数平均値が高くなっていることがわかる。また、同様の傾向が最大値 (図 3.14 (b)) および最小値 (図 3.14 (c)) でも見られる。



(a) ML

(b) MAP-like

(c) MAP-MRHSMM

図 3.12: 合成音声「うん」のランニングスペクトル

また、対数基本周波数の統計量についても 0 次メルケプストラム係数の場合と同様に、提案法である MAP-MRHSMM を用いることによって極端な値が抑制されている。特に、基本周波数の最小値では、MAP-MRHSMM を用いることで極端に低い基本周波数が抑制されていることがわかる。以上のことから、提案法の有効性を客観的に確認することができる。

3.5 主観評価実験

本節では、提案されたパラ言語情報の反映手法の有効性を確認するために実施された主観評価実験について述べる。ここで、主観評価実験は以下に示す 3 つの観点を検討する実験が行われた。

1. パラ言語情報伝達性
2. 自然性
3. 自発性

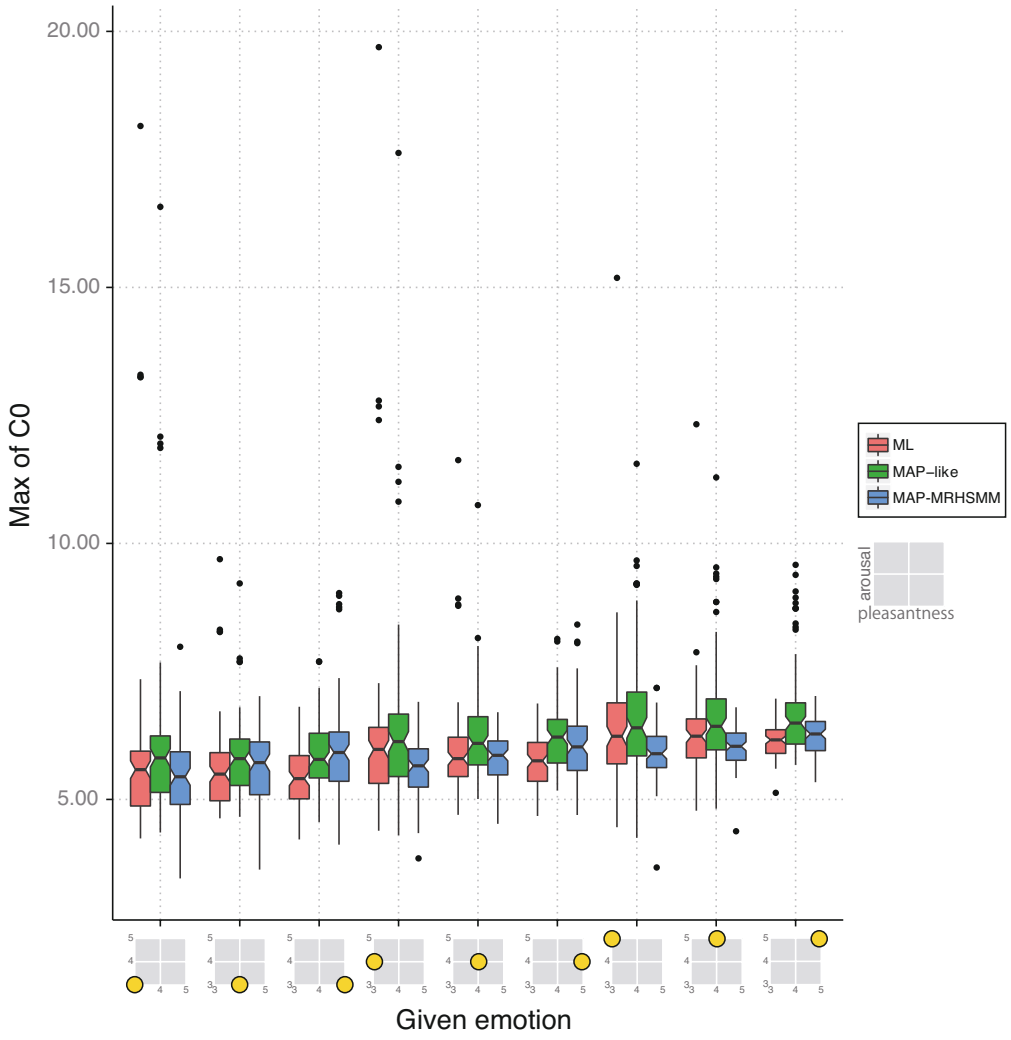
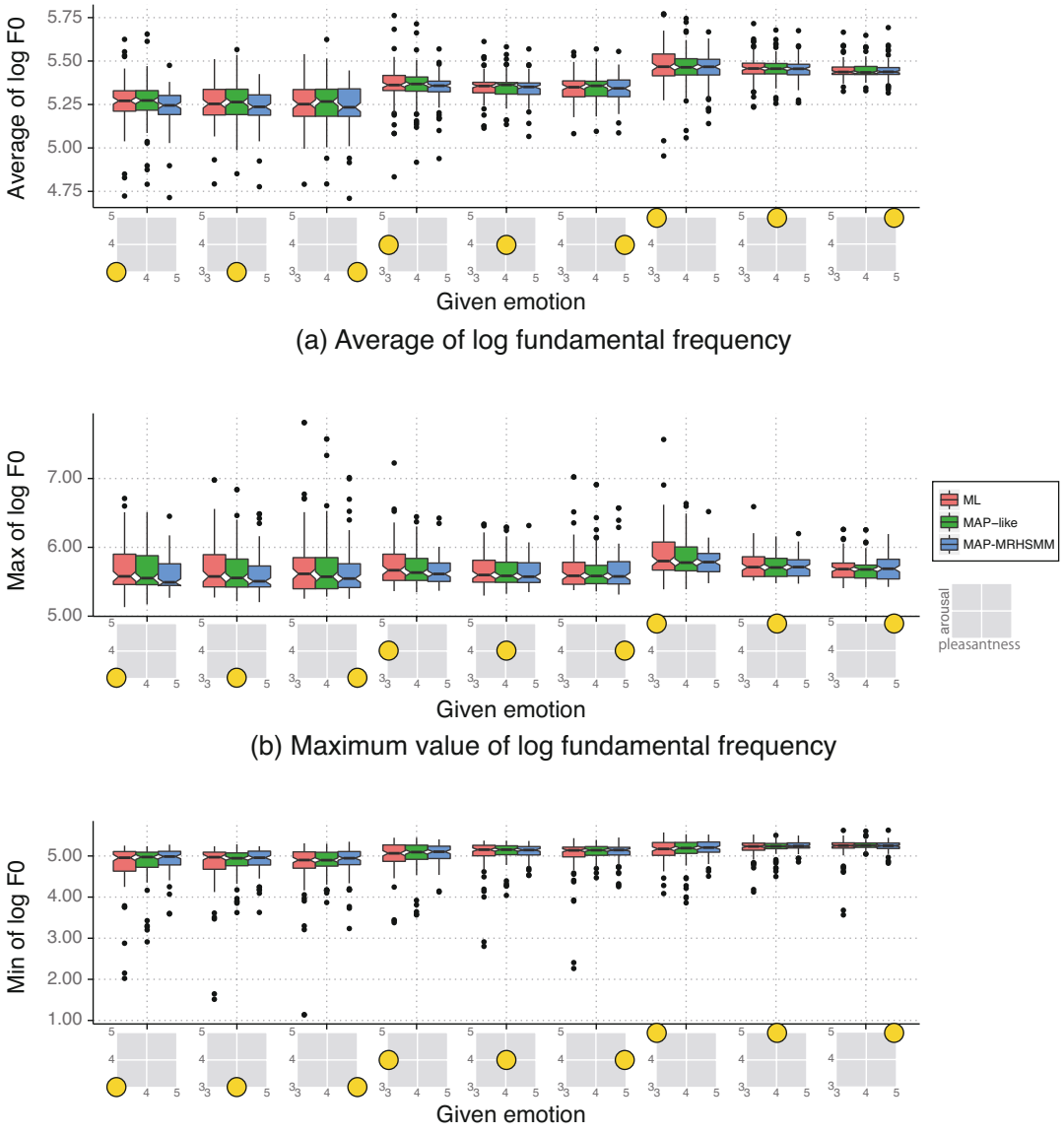


図 3.13: 0 次メルケプストラム係数最大値の分布



(a) Average of log fundamental frequency

(b) Maximum value of log fundamental frequency

(c) Minimum value of log fundamental frequency

図 3.14: 対数基本周波数統計量の分布

1つ目の実験はパラ言語情報の伝達性を確認するための実験であり、合成音声から知覚されるパラ言語情報の度合いを評価する。知覚されるパラ言語情報と、合成時のパラ言語情報を比較することによって、意図したパラ言語情報が伝達されているかを確認する。

2つ目の実験は自然性を確認するための実験である。自然対話音声コーパスを用いたMRHSMMでは、重回帰モデルパラメータの過推定問題により極端な値の音響特徴量を持つ音声が合成されることがある。ここでは、提案法であるロバストな推定手法を用いることで、自然性が改善されるかどうかを評価する。

3つ目の実験は自発性を確認するための実験である。対話音声は朗読音声と異なり、非常に自発的な音声である。合成された音声がどれだけ自発的かを評価にすることにより、対話音声らしい音声が合成できているかを確認する。

3.5.1 パラ言語情報知覚実験

本実験では、合成音声から知覚されるパラ言語情報を評価する。実験には、UUDBの話者FTSの15発話内容に対して、3.4.1節と同様に、図3.10に示されるパラ言語情報を与えて合成された音声が使用された。したがって、被験者に呈示される合成音声は3(手法) × 15(発話内容) × 9(パラ言語情報) = 405個である。ここで、合成時に与えるパラ言語情報は3から5に制限された。3.4.1節でも述べたように、この範囲を超えたパラ言語情報を与えた場合に、ML法による合成音声の品質が著しく低下するためであり、そのような合成音声を被験者に呈示することは非常に苦痛を強いることになるためである。

実験には、7人の男子大学生および9人の男子大学院生の計16名の被験者が参加した。被験者には、評価を行う前に感情の次元説に関する基本的な理論と各次元が意味することについての説明を行った。その後、被験者には合成音声から知覚されるパラ言語情報を7段階で評価するように指示した。また、知覚されるパラ言語情報は発話内容そのものではなく、話し方といったような印象から評価するように指示した。実験は静かな研究室でヘッドホン (AKG K271 MKII) による両耳聴取によって行われた。

実験結果として、合成時に与えたパラ言語情報と被験者によって知覚されたパラ言語情報の分布を図3.15および図3.16に示す。図の横軸は合成時に与えた

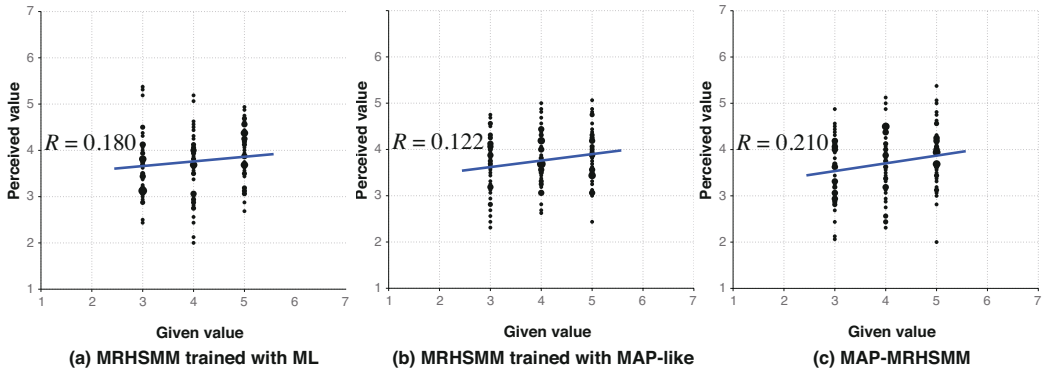


図 3.15: 「快-不快」の平均評価値の分布

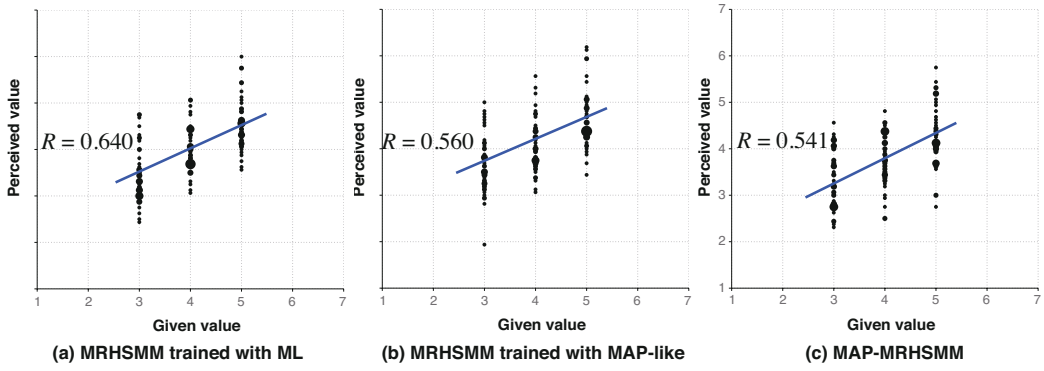


図 3.16: 「覚醒-睡眠」の平均評価値の分布

パラ言語情報であり、縦軸は被験者によって知覚されたパラ言語情報の平均値である。

図 3.15 は「快-不快」に関する結果である。図より、全ての手法において弱い正の相関があることがわかる。また、3 つの手法間で相関係数に差は確認されなかった ($\chi^2(2) = 0.57, p > .05$)。

図 3.16 は「覚醒-睡眠」に関する結果である。図より、どの手法においても強い正の相関があることが確認できる。すなわち、覚醒よりに合成した音声は被験者に対しても覚醒よりに知覚されていることを意味している。また、相関係数は ML 法が高く見えるが、3 つの手法間で相関係数の差は有意ではなかった ($\chi^2(2) = 1.74, p > .05$)

以上の結果から、意図したパラ言語情報が被験者にある程度伝達されていることがわかる。また、3 つの手法間でパラ言語情報の反映能力には差がないことも示された。

3.5.2 自然性評価実験

本実験では、合成音声の自然性を評価する。パラ言語情報知覚実験と同様に、UUDB の話者 FTS の 15 発話内容が実験に使用された。合成時には図 3.17 に示す 5 通りのパラ言語情報が与えられた。したがって、被験者に呈示される刺激の総数は 3(手法) × 15(発話内容) × 5(パラ言語情報) = 225 個である。

実験には、7 人の男子大学生および 9 人の男子大学院生の計 16 名の被験者が参加した。被験者には、合成音声の自然性を 5 段階 (1:不自然, 2:やや不自然, 3:どちらでもない, 4:やや自然, 5:自然) で評価するよう指示した。実験は静かな研究室でヘッドホン (AKG K271 MKII) による両耳聴取によって行われた。

自然性評価実験の結果として、図 3.18 に被験者による平均評価値 (Mean Opinion Score: MOS) を示す。ML 法および MAP-like 法では、MOS が 1.0 以上 1.5 未満と評価された合成音声の数から、不自然だと評価した合成音声が多く存在していることがわかる。一方、提案法である MAP-MRHSMM では、MOS が 1.5 未満の合成音声は存在せず、不自然な合成音声が減少していることが確認できる。この結果は、3.4.2 節で述べたような極端な音響特徴量が抑制されたことを反映していると考えられる。

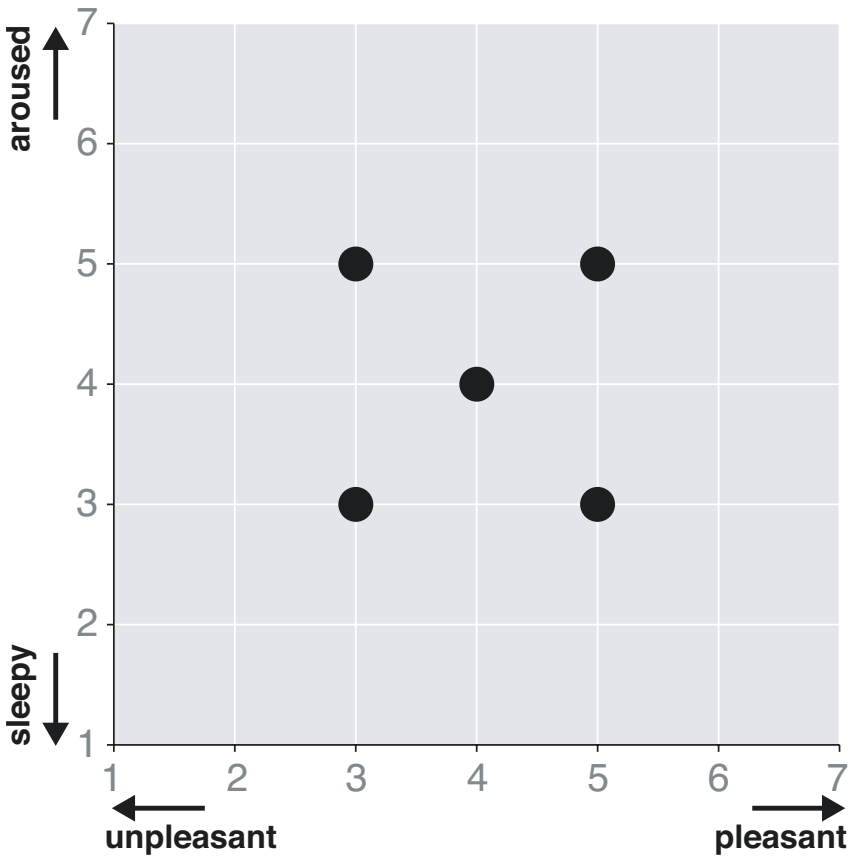


図 3.17: 自然性評価実験において与えられたパラ言語情報

ML 法、MAP-like 法、MAP-MRHSMM における MOS の平均値はそれぞれ 2.91, 2.97, 3.24 である。手法を要因とした分散分析を行った結果、主効果が有意であった ($F(2, 222) = 3.21, p < .05$)。そこで、Tukey の HSD 法によって多重比較を行った結果、ML 法と MAP-MRHSMM の間の差が有意であった ($p < .05$)。以上の結果から、MRHSMM パラメータのロバストな推定を行うことによって、合成音声の自然性が改善されることを確認した。

3.5.3 自発性評価実験

この実験では、合成音声の自発性を評価する。実験には、UUDB の話者 FTS の 80 発話内容を用い、合成時には発話に元々与えられている UUDB のパラ言語情報の平均評定値を与えた。

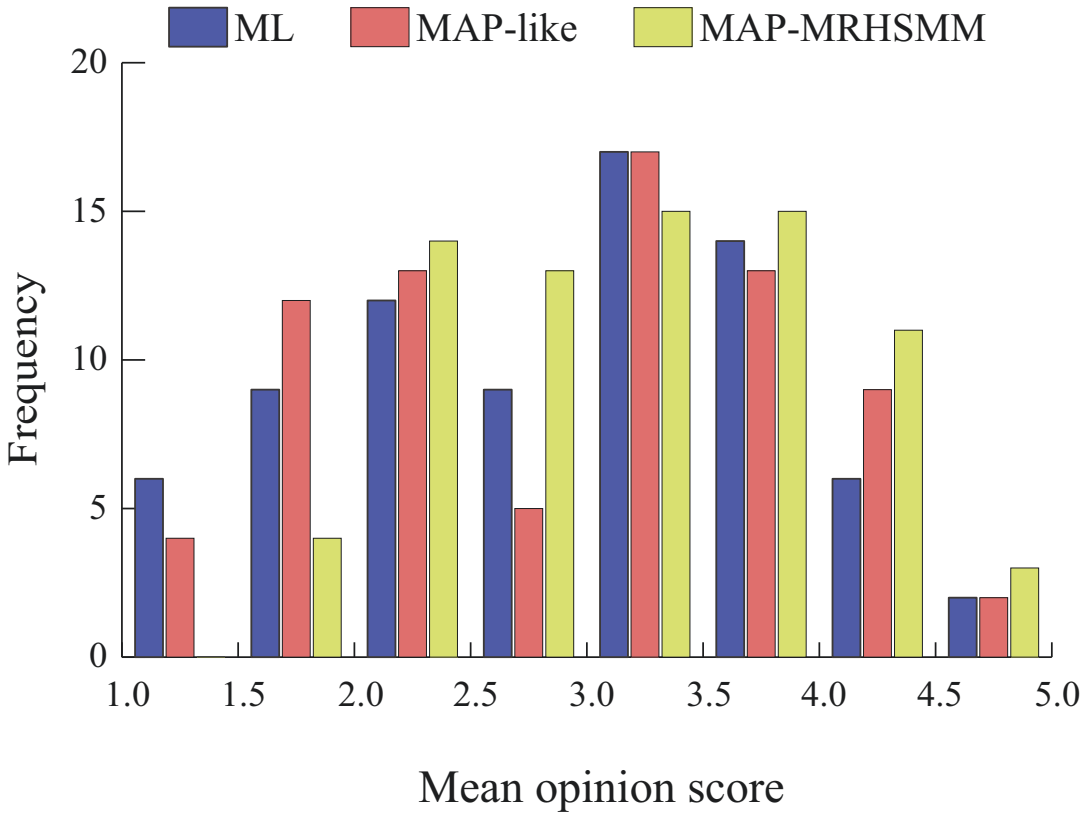


図 3.18: 自然性に関する平均評価値

本実験では、3.4.1 節で述べた 3 手法に、従来の HSMM(ここでは READ と呼ぶ) によって合成された音声がベースラインとして加えられた。READ モデルは Nitech データベースに含まれる 1 名の話者による音素バランス音声 503 文を用いて学習された [4]。音響特徴量の抽出条件および特徴ベクトルの構成は話者 FTS のものと同様である。HSMM は 5 状態の left-to-right HSMM とした。以上のことから、被験者に呈示される刺激の総数は 4(手法) $\times$ 80 (発話内容) $=$ 320 である。

実験には、5 人の男子大学生および 6 人の男子大学院生の計 11 名が参加した。被験者には、合成音声の自発性を 5 段階 (1:自発でない, 2:やや自発でない, 3:どちらでもない, 4:やや自発, 5:自発) で評価するよう指示した。実験は静かな研究室でヘッドホン (AKG K271 MKII) による両耳聴取によって行われた。

READ, ML 法, MAP-like 法および MAP-MRHSMM の自発性評価の平均値

はそれぞれ 2.22, 2.87, 2.87 および 2.91 であった。手法を要因とした分散分析を行った結果、主効果が有意であった ($F(3, 316) = 23.84, p < .01$)。そこで、Tukey の HSD 法による多重比較を行った結果、READ と UUDB によって学習された 3 つのモデルの間の差が有意であった ($p < .05$)。また、UUDB によって学習された 3 つのモデルの間では、自発性に有意な差は確認されなかった ($p > .05$)。

以上の結果から、自然対話コーパスを利用した MRSHMM に基づく音声合成方式によって、従来の HSMM を用いた場合よりも自発性の高い対話音声らしい音声が合成されていたことを確認できた。

3.5.4 考察

この節では、合成音声の客観的評価および主観的評価の結果について考察する。

まず、客観的評価について考察する。図 3.13 に示したように、合成された音声の 0 次メルケプストラム係数は「覚醒-睡眠」の変化によって組織的に影響を受ける。この結果は、ポジティブな感情においては発話の強度最大値が上昇するというこれまでの知見 [24] と一致しており、妥当な結果であると言える。同様に、図 3.14 に示した対数基本周波数の結果でも、「覚醒-睡眠」の違いにより組織的な影響を受けていた。「快-不快」については組織的な変化が見られなかったものの、図 3.11 に示したように、終助詞の基本周波数などに影響を与えており、「快-不快」の知覚に役立つ韻律パターンを反映している可能性がある。しかし、その仮説を証明するためには更なる証拠が必要である。

合成音声の客観的評価において最も重要な結果は、MAP-MRHSMM を用いることによって、極端な音響的特徴が出力されることが抑制されていることである。対数基本周波数では、提案法である MAP-MRHSMM と提案法の近似手法である MAP-like 法を用いた場合で極端な音響特徴量の出力が抑制されていた。すなわち、対数基本周波数に関しては両手法の性能は同程度であると言える。しかし、0 次メルケプストラム係数では、MAP-MRHSMM のみが極端な音響特徴量の出力を抑制している。このことから、MAP-MRHSMM を導入することの有効性が確認された。

次に、主観評価実験の結果について述べる。パラ言語情報知覚実験の結果では、「快-不快」および「覚醒-睡眠」の両次元において、合成時に与えた値と被験

者に知覚された値の間に正の相関があることが確認された。しかしながら、「快-不快」の相関は「覚醒-睡眠」の相関に比べて弱い相関であった。この理由としては、「快-不快」の次元は音響特徴量の単純な変化によって知覚されるものではないということが考えられる [77–79]。また、他の理由として、合成された発話内容の選択方法が考えられる。3.4.2 節で示したように、「快-不快」の違いによる音響特徴量の変化は組織的ではなく、終助詞といったような限定的な場面に現れる。すなわち、「快-不快」を表現するためのキャリアが含まれていない発話内容が選択された場合、「快-不快」があまり反映されない音声が呈示されることになる。したがって、「快-不快」の次元の評価については、合成する発話内容の選択基準を再設計した実験を行う必要があると考えられる。しかしながら、そういった考慮を全くしなかった場合でも合成時の「快-不快」の値と知覚される「快-不快」の間に正の相関が確認されているため、ある程度は意図した「快-不快」が伝達できていると考えられる。

自然性評価実験の結果では、提案法である MAP-MRHSMM によって自然性が改善されていることを確認した。特に、自然性が著しく低い合成音声 ($1.0 \leq \mathrm{MOS} < 1.5$) が減少し、やや自然な合成音声 ($4.0 \leq \mathrm{MOS} < 4.5$) が増加しており、極端な音響特徴量の出力が抑制されたことが反映された結果であるといえる。MAP-MRHSMM では回帰行列のロバストな推定を行うために、推定される回帰行列が保守的になり自然性の高い合成音声が減少する懸念があったが、今回の結果よりその懸念は払拭された。

3.6 おわりに

本章では、合成音声にパラ言語情報を反映させる手法として、次元で記述されたパラ言語情報を持つ自然対話音声コーパスと重回帰 HSMM に基づく音声合成方式について述べた。更に、自然対話音声コーパスを用いるうえでの問題点として、重回帰モデルパラメータの過推定問題について述べた。そして、その問題を解決する手法であるロバストな推定手法として、重回帰モデルパラメータの MAP 推定を提案し、その推定式を導出した。

提案法により合成された音声に対して客観的評価および主観的評価を行った。

客観的評価では、「覚醒-睡眠」を変化させることで合成音声の音響特徴量を組織的に変化させることができることと、「快-不快」を変化させることで終助詞の音響特徴量を変化させていることを確認した。

また、主観的評価では、合成時に与えたパラ言語情報と被験者に知覚されたパラ言語情報との間に正の相関があることを確認し、意図したパラ言語情報が合成音声に反映されていることを示した。更に、提案法である重回帰モデルパラメータのロバストな推定手法を用いることで、自然対話コーパスを使用する際の重回帰モデルパラメータの過推定問題を改善できることも示した。

以上のことから、自然対話音声におけるパラ言語情報を反映可能な音声合成手法を実現できたと考えられる。

第4章　対話音声における笑い声の記述と分析

4.1　はじめに

これまで、笑い声に関連する多くの研究では、お笑いビデオやライブといった映像刺激によって誘発された笑い声を対象としてきた。これは、笑い声研究のために大量の笑い声を収集する必要があることが1つの要因である。映像刺激によって笑い声を誘発することは多くの笑い声を収集するうえでは効率的あるといえるが、それによって生じる笑い声のほとんどは感情喚起による笑い声になると予想される。すなわち、笑い声の中でも非常に限定されたシチュエーションに生じる笑い声しか収集することができず、コミュニケーション場面で用いられる多様な笑い声をカバーしていないと考えられる。

そこで本研究では、実際の対話場面で生じた笑い声を対象とする。つまり、自然対話音声コーパスに含まれる笑い声を対象とし、最終的にはそれらを用いて笑い声の合成を行うことを目標としている。

4.2　対話音声コーパスに含まれる笑い声

本研究では、自然対話音声コーパスとして3章でも用いたUUDBを使用する。また、UUDBに加えて、オンラインゲーム音声チャットコーパス(OGVC) [25]に含まれる笑い声も対象とする。

4.2.1　UUDB

UUDBの詳細については3.2.2で述べたとおりである。UUDBには収録されている音声に対してパラ言語情報をはじめとする様々なノンバーバル情報が与

表 4.1: UUDB の各話者における笑い声の数

Speaker	Num	Speaker	Num	Speaker	Num
FJK	8	FKC	22	FMS	16
FMT	34	FNN	14	FSA	26
FSH	17	FTH	5	FTS	40
FTY	19	FUE	14	FYH	23
MKK	22	MKO	20		

表 4.2: OGVC の各話者における笑い声の数

Speaker	Num	Speaker	Num	Speaker	Num
01_MMK	26	03_FMA	74	05_MYH	52
01_MAD	118	03_FTY	41	05_MKK	87
02_MFM	142	04_MNN	154	06_FTY	251
02_MEM	145	04_MSY	246	06_FWA	175

えられている。UUDB では非言語音に関する記述も与えられており、笑い声や吸気、咳払いといった非言語音の位置と時間情報が与えられている。UUDB に含まれる笑い声の総数は 280 個であり、各話者による内訳は 4.1 の通りである。

4.2.2 OGVC

OGVC は友人同士がボイスチャットをしながらゲームをプレイしている時の対話を収録した自然対話コーパスである。また、自然対話を再朗読した音声を収録されており、自然対話音声と演技音声の比較を行うこともできる。本研究では、OGVC の自然対話音声のみを使用する。OGVC には 13 名 (女性 4 名、男性 9 名) の話者による 9114 発話の自然対話音声が収録されている。

OGVC では笑い声の位置が転記として与えられている。OGVC の自然対話音声に含まれる笑い声の総数は 1593 個であり、各話者における内訳を表 4.2 に示す。表からわかるように、OGVC には比較的多くの笑い声が含まれる。特に、1 人あたりの笑い声の数が多く、笑い声の合成などで特定話者モデルを構築する際に有利であると考えられる。

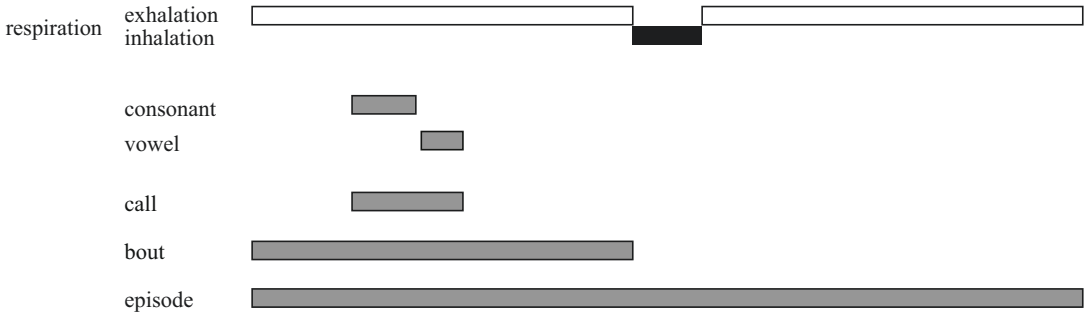


図 4.1: 笑い声の階層構造 ([5] 一部改変).

4.3 笑い声のセグメンテーション

笑い声の構造は階層的に理解されており、その構造は大きく分節レベル、音節レベル、フレーズレベルに分けられる [5]。図 4.1 はその構造を示す。分節レベルでは、笑い声は「笑い声の子音」や「笑い声の母音」によって分割される。笑い声子音および笑い声母音は音声学における子音および母音とは厳密に異なる。音節レベルでは笑い声子音と笑い声母音の組、あるいは笑い声母音単独で構成される「call」と呼ばれる単位で笑い声が分割される。フレーズレベルでは、笑い声は 1 つ以上の call で構成される「bout」と呼ばれる単位で笑い声が分割される。一般に、bout には吸気は含まれない。また、複数のフレーズの連続は文レベルとも呼ばれ、「笑い声 episode」と呼ばれる。

本研究では、それぞれの笑い声は bout 単位および call 単位で分割される。

4.3.1 笑い声の記述

笑い声 episode は 1 つ以上の bout と吸気によって構成される。本研究では、笑い声 episode は「{laugh}」という記号で表現される。まず、発話の中から笑っている部分を笑い声 episode として分割する。次に、笑い声 episode を bout と吸気に分割する。ここで、bout は「b」、吸気は「h」という記号で表現する。

最後に bout 部分を call 単位に分割する。また、分割する際には call の音韻性についての記述を与える。笑い声の HMM 音声合成に関する研究 [80, 81] で使用されていた笑い声コーパスである AVLaughterCycle データベース [82] では、国際音声字母 (IPA) に基づく厳密な記述を与えている。笑い声の厳密な音韻転記

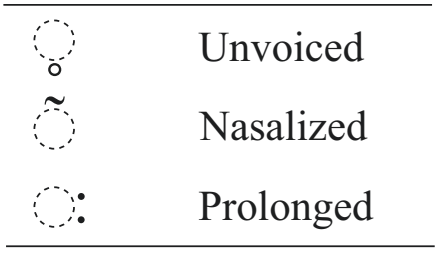


図 4.2: call 転記に使用される補助記号

を与えるのは音声合成の品質の観点から見れば有用であると考えられるが、そのような厳密な記述を与えることは非常にコストがかかる。そこで本研究では簡易的な記述として、call を仮名表記を用いる。笑い声を仮名表記で記した後、音素系列に変換して call の音韻を表現する。

また、音素系列だけでは無声化、鼻音化、長音化といった現象を表現することができない。そこで、図 4.2 に示す補助記号を定義し、異なる音韻として記述することにした。

4.3.2 アノテーション

前節で定義された記述方式を用いて笑い声のアノテーションが行われた。アノテーションは筆者であり、アノテーションは Praat [83] を用いて行われた。アノテーションは各コーパスの女性話者による笑い声を対象に行われた。これは、女性話者の方が男性話者よりも頻繁に笑いが発生していたためである。ここでは、UUDB の話者 FTS、OGVC の話者 03_FTY、06_FTY および 06_FWA の笑い声を対象にアノテーションが実施された。また、このアノテーションは後の章で説明する対話場面における笑い声の合成に用いることを考慮しているため、単独の笑い声ではなく、言語音に伴う笑い声が対象とされた。

4.3.3 アノテーション結果

実際にアノテーションされた笑い声の例を図 4.3 に示す。笑い声のアノテーションは図の①から③に示される 3 つのレイヤーで行われる。まず、①では笑い声 episode と言語音部分の転記が記述される。先に述べたように、{laugh} で

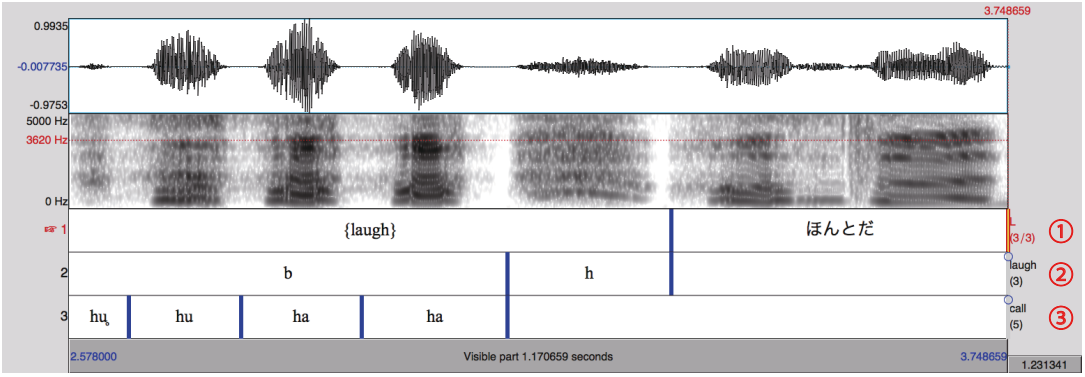


図 4.3: 笑い声に対するアノテーション例

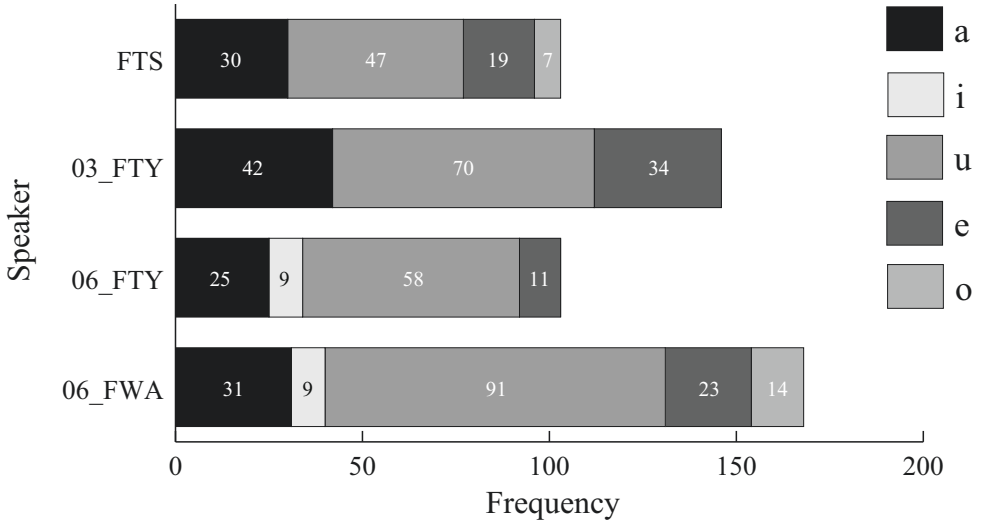


図 4.4: 笑い声母音の数

示される部分が笑い声であり、それ以外の部分が言語音である。②では笑い声 episode の構造が bout と吸気によって記述される。b が bout を表しており、h が吸気を表している。この例では、発話の先頭の笑い声 episode が 1 つの bout と 1 つの吸気によって構成されていることがわかる。③では、bout を構成する call の音韻が記述されている。この音韻記述には補助記号によって拡張された音素表記が使用されている。この例では、[hu̥huhaha] という 4 つの call で構成されている bout であることがわかる。

アノテーション結果として、call の笑い声母音のヒストグラムを図 4.4 に示す。対話場面における call の笑い声母音としては、/a/系、/u/系、/e/系が支配

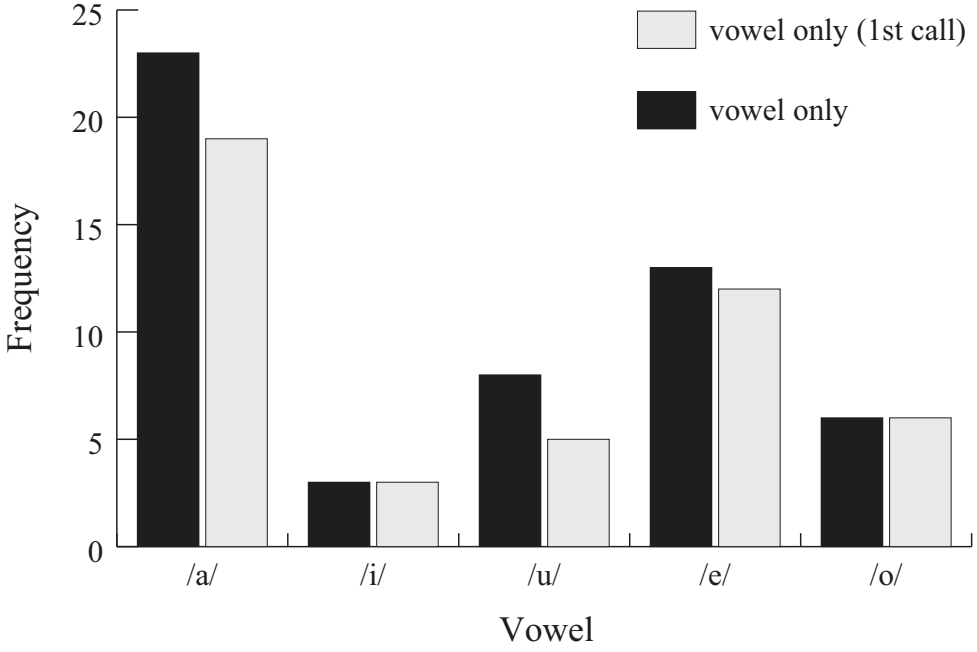


図 4.5: 母音のみで構成される call の数

的ある。/i/系や/o/系の笑い声母音は稀であり、話者によっては全く出現していない。

最も頻繁に出現した笑い声母音は/u/であり、そのほとんどは笑い声子音を伴う call であった。このことを図 4.5 に示す。図は笑い声母音のみで構成される call の数を示している。図中の黒い棒は笑い声母音のみで構成される call の数を示しており、グレーの棒はその中でも先頭の call であるものの数を示している。図より、笑い声母音のみで発せられた/u/系の call が他の母音性と比べても少ないことがわかる。対象的に、/a/系と/e/系の笑い声は笑い声母音のみで構成される call が比較的多いことがわかる。また、全体として笑い声母音のみで構成される call は先頭 call であることがわかる。

自然対話音声コーパス中の笑い声の構造について調査した研究 [84] と同様に、bout 構造の内訳について調査した。この結果を図 4.6 に示す。図より、1 つの call のみから成る bout(single-call bout) が全体の 17%であり、そのうちの 35%程度が無声であることがわかる。一方で、2 つ以上の call から成る bout(multi-call bout) では、全体が無声で構成されることはあまりないことがわかる。この結果

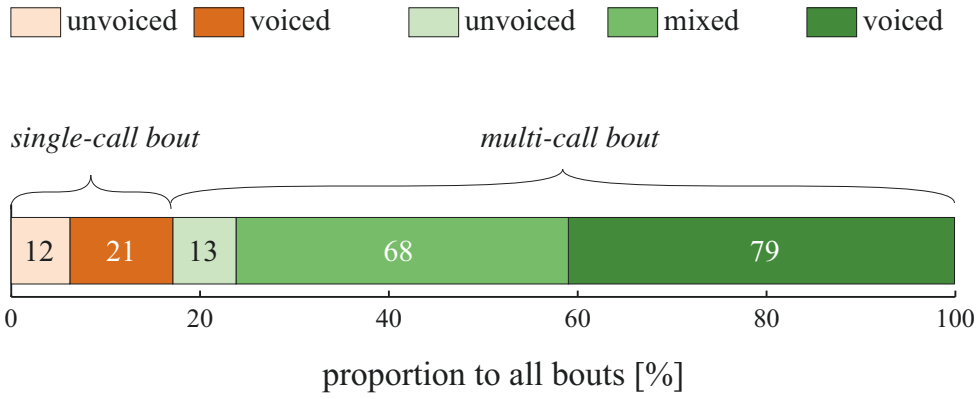


図 4.6: bout の有声/無声構造の内訳

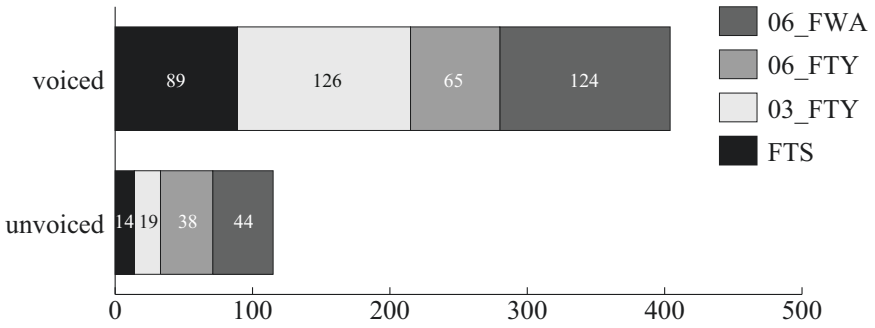
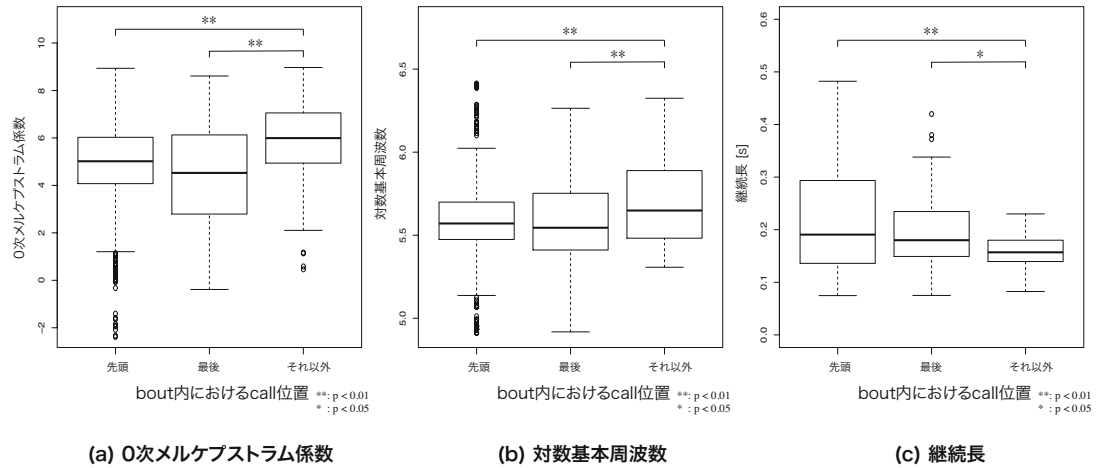


図 4.7: call の有声性分布



(a) 0次メルケプストラム係数　(b) 対数基本周波数　(c) 継続長

図 4.8: 各 call 位置における音響特徴量の分布

は UUDB の笑い声のみを対象とした結果 [84] と概ね一致する。

また、call の有声性の分布を図 4.7 に示す。出現した call の多くは有声であり、無声 call は有声 call の半分に満たないことがわかる。

4.4 笑い声の音響的特徴の分析

ラベリングされた笑い声から抽出された音響特徴量による分析を行った。ここでは、call 単位および bout 単位での音響特徴量の比較を行った。call 単位では、1 bout における call の位置に比較を行った。また、bout 単位の比較では 1 発話における bout の位置ごとに比較を行った。

call 位置に関する音響特徴量の差

bout 内の call の位置における音響特徴量に関する分析を行った。ここでは call 位置を先頭、最後、それ以外の 3 つに分類して各位置について 0 次メルケプストラム係数、対数基本周波数、call 継続長の分布を調べた。各位置における音響特徴量の分布を図 4.8 に示す。

図より、どの特徴量においても先頭 call とそれ以外の call、最後 call とそれ以外の call の間に差が確認された。ここで、どの特徴量においても等分散性を仮定できなかったため多重比較は Games-Howell 法を用いて行った。

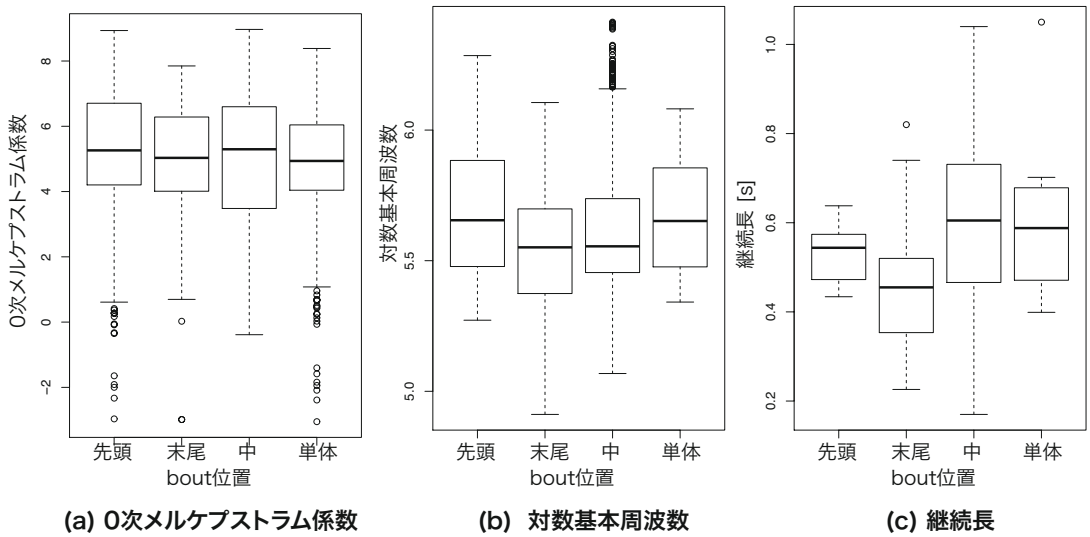


図 4.9: 各 bout 位置における音響特徴量の分布

0 次メルケプストラムについては、先頭および最後の call がそれ以外よりも小さい傾向があった。これは最初の call が無声化しやすいことや肺圧の低下といった現象を反映していると考えられる。

対数基本周波数についても、0 次メルケプストラム係数と同様の傾向が得られた。

継続長は他の特徴と異なり、先頭と最後がそれ以外のものよりも長くなるという傾向が現れた。また、先頭でも最後でもない call の継続長の分布が狭くなっていることがわかった。

bout 位置に関する音響特徴量の差

発話内の bout 位置における音響特徴量の分布を調べた。ここでは bout 位置を先頭、末尾、中、単体の 4 つに分類して、各位置について call の場合と同様の 3 種類の音響特徴量の分布を調べた。音響特徴量の分布を図 4.9 に示す。

0 次メルケプストラム係数は bout 位置によって差が現れていない。このことから、この特徴量は bout 位置によって影響を受けない特徴量であることがわかる。

次に対数基本周波数に注目する。対数基本周波数の分布は、先頭および単体

が末尾および中よりも高い位置に分布していることがわかる。

最後に継続長分布について述べる。先頭および末尾の bout に含まれる call の継続長が中および単体の call の継続長よりも短くなる傾向があった。

4.5 おわりに

本章では、ノンバーバル情報を伝達する代表的な媒介である笑い声について述べた。自然対話における笑い声を対象とするために、自然対話音声コーパスである UUDB および OGVC の 2 つのコーパスにおける笑い声を含む発話に対して、笑い声の時間情報および転記情報を与えた。

与えられた転記情報をもとに、笑い声の構造や母音性・有声性についての分析が行われた。分析の結果、笑い声の構造によって母音性や有声性などに差があることが確認された。

更に、笑い声の音響特徴量についての分析が行われた。対象となる音響特徴量は統計的パラメトリック音声合成において広く用いられているメルケプストラム係数、対数基本周波数、継続長であり、call 単位および bout 単位で比較された。比較の結果、call 位置や bout 位置によって笑い声の音響特徴量に差があることが明らかとなった。この結果は次章で行なう笑い声の合成における音響特徴量の変動要因として有効に働くと考えられる。

第5章　自然対話コーパスを用いた笑い声合成

5.1　はじめに

近年では、人間同士だけではなく、人間と機械のインタラクションにも関心が高まっている。そのような場面で音声合成技術を用いるために、対話音声合成が望まれる。対話音声合成は言語情報だけではなく、話者の意図や態度、感情といった情報を表現することが要求される。3章では、そのような情報を表現するための技術について論じた。そこで対象としていたのは言語音の合成音声にパラ言語情報を反映させる方法である。しかしながら、対話場面では言語音だけではなく、言語音ではないものの様々な音が存在する。そして、それは実際のコミュニケーションにおいて重要な役割を果たしていると考えられる。

その1つに笑い声がある。笑いはノンバーバル情報を伝達する典型的な行為であり、最近では研究者たちの関心を集めている [85–88]。特に、笑い声の検出や認識に関する分野が現在精力的に研究されている。Truong ら [85] や Neuberger ら [87] は、混合ガウス分布やサポートベクトルマシンを用いた笑い声の検出について研究している。Knox らはニューラルネットワークを用いた笑い声の認識手法について提案しており、90%以上の精度を達成している [86]。また、Petridis らは視覚情報をも利用した笑い声と言語音の分類手法についても提案している [88]。

一方で、笑い声の合成に関しては、現在でもあまり研究されていない。数少ない笑い声合成に関連する研究には、分析合成に基づく手法 [89,90]、波形接続型方式に基づく手法 [91,92] がある。Sathya らによる研究では、言語音母音から抽出された音源パラメータを調整することによって笑い声合成を実現している [89]。Sundaram らは、笑い声波形の振幅の振動的な動きをバネ-マス系モデルでモデル化することにより、周期的な有声笑い声の合成を実現している [90]。Trouvan

らや Lasarcyk らはダイフォンに基づく笑い声合成を行っている [91, 92]。この手法では実際の笑い声素片を利用しているため、高品質な笑い声を合成可能である。また、有声の笑い声だけでなく、無声の笑い声をも合成可能である。しかしながら、これらの研究では笑い声を含む発話を合成する場合、笑い声そのものの品質が高かったとしても、発話全体としての自然性が低下してしまう問題について指摘している。

笑い声はどれも似ているわけではない [93]。笑い声は話者の感情の高まりによって発せられる無意識的な笑いや、話者の肯定的な態度を相手に伝えるために意識的に発せられる笑いといったように、非常に多様である。笑い声の形態も多様である。Bachorowski らは笑い声の形態的な特徴を調査しており、「有声の歌うような笑い声」、「鼻息のような笑い声」、「無声の語断片のような笑い声」の 3 つのタイプに分類している [93]。また、Campbell は「有声笑い声」、「静かな笑い」、「気息音のような笑い」、「鼻笑い」に分類している [94]。更に、Tanaka らは「丁寧な笑い」、「陽気な笑い」、「冷笑的な笑い」といった 3 つのタイプに分類している [95]。このように笑い声は様々な種類があり、文脈や状況に応じて適した笑い声が要求される。このことには音韻的な文脈だけではなく、パラ言語的な文脈をも含む。波形接続型合成方式において、このように様々な文脈や状況を考慮した笑い声合成を実現するためには、それだけ膨大な笑い声を収集しなければならないという問題がある。

そこで本研究では、波形接続型合成方式ではなく、統計的パラメトリック合成手法を用いて笑い声合成を実現する。3 章でも述べたが、代表的な統計的パラメトリック合成方式である HMM 音声合成では、音声の音響的特徴の変動要因をコンテキスト依存モデルで表現することにより、様々な文脈を考慮した音声を合成する。笑い声に対しても適切なコンテキストを定義し、コンテキスト依存モデルに基づいた合成を行うことで、文脈を考慮した笑い声合成を実現できると期待される。

関連した研究に、HMM 音声合成方式による笑い声合成の研究がある [80, 81]。この研究は本研究と非常に近い目標を持った研究である。本研究との大きな違いは「映像刺激によって誘発された笑い声」を対象としている点である。すなわち、対話場面における笑い声を対象にしていない部分が異なる。映像刺激に

よって誘発された笑い声は基本的にコミュニケーションを目的として発せらた笑い声ではないため、「話してから笑う」や「笑ってから話す」といったものが基本的には存在しない。すなわち、笑い声が置かれている文脈や状況を考慮していないといえる。

本研究では実際の対話場面で発せられる笑い声を用いた笑い声合成を行う。そのためには、笑い声が置かれている文脈や状況を考慮するためのコンテキストが要求される。そこで、笑い声合成のための笑い声コンテキストを定義する。定義されたコンテキストに基づいて笑い声のコンテキスト依存モデルを学習し、そのモデルに基づいて笑い声合成を行う。また、コンテキストに基づく笑い声合成を行なう前に、笑い声の形態的分類に基づいた合成を行なうことで、笑い声の合成が可能かどうかについても確認する。

実際に合成された笑い声が文脈や状況に適しているかどうかを評価するために、笑い声に対する主観評価実験を行う。ここでは、笑い声だけではなく、言語音に付随する笑い声を被験者に呈示し、発話全体の自然性を評価することで、文脈に適しているかを評価する。

5.2 笑い声の形態的分類に基づいた合成

形態的分類に基づく笑い声の合成では笑い声の形態ごとに HSMM を学習し、そのモデルに基づいて笑い声を合成する。

宇都宮大学パラ言語情報研究向け音声対話データベース (UUDB) に収録されている笑い声の形態的分類は [7,96] で行われている。[7] における笑い声の分類基準を表 5.1 に示す。この分類では、以下の 4 つの属性に基づいて笑い声を分類している。

- call 数
- 有声/無声
- 鼻音/口音
- 吸気/呼気

本研究では、この分類に基づいて笑い声の合成を行う。

表 5.1: 笑い声の分類基準 ([7] 一部改変)

	Property			
Form ID	call の数	voiced	nasal	ingressive
110	single	-	+	-
120	single	-	-	-
130	single	+	-	-
150	single	-	-	+
160	single	+	-	+
210	multi	-	+	-
220	multi	-	-	-
230	multi	+	-	-
240	multi	+/-	-	-
250	multi	-	-	+
260	multi	+	-	+
270	multi	+/-	-	+

表 5.2: 形態毎の bout の数

Form ID	110	120	130	150	160	210	220	230	240	250	260	270
Num	0	1	2	0	0	5	0	18	13	0	0	0

5.2.1 合成条件

本発表では特定話者の笑い声を合成する。対象話者は UUDB の話者 FTS とする。対象とする笑い声は 40 bout であり、形態ごとの bout 数を表 5.2 に示す。表に示すように、属する bout が存在しない形態も存在するので、今回の検討では bout の属する 120, 130, 210, 230, 240 を合成対象の形態とする。

学習および合成に使用するパラメータはスペクトルパラメータと音源パラメータである。スペクトルパラメータは STRAIGHT 分析によって得られた STRAIGHT スペクトルから求められた 0 次から 34 次のメルケプストラム係数とした。音源パラメータは対数基本周波数を使用した。分析条件はサンプリング周波数 16 kHz、分析窓長 25 ms、フレームシフト 5 ms とし、分析窓はハミング窓とした。特徴ベクトルは得られたスペクトルパラメータおよび音源パラ

メータとそれらの Δ 及び ΔΔ パラメータを含めた 108 次元のベクトルとした。

学習するモデルは 10 状態の left-to-right HSMM とした。

5.2.2 合成結果

合成結果について述べる。合成した 5 つの笑い声のうち、120(単発無声笑い) と 130(単発有声笑い) については自然な笑い声が合成されていた。また、210(連発鼻笑い) についても比較的自然な笑い声が合成されていた。単発笑いに関しては、単純な音素 HMM による音声合成の延長であると考えられるのでこの結果は妥当である。連発鼻笑いが比較的自然に合成されていたのは予想と反していたが、この要因には、この形態に属する笑い声の call の種類が少なかったことが考えられる。この形態に属する笑い声は「ンフフフ」や「フフン」がほとんどであり、call の数にもほとんどばらつきがなかった。そのため、比較的安定したモデルが学習できたと推測される。

一方、230(連発有声笑い) および 240(連発有声無声混在笑い) の合成結果は自然性が低かった。この要因としては、これらの形態に属する笑い声には call の数や call の種類に様々なバリエーションがあることが考えられる。

以上のことから、単発笑いに関してはうまく合成できており、連発笑いに関しては鼻笑い以外はうまく合成できていないという結果が得られた。しかしながら、単発笑いの合成は実現できていることから、call 単位での笑い声の合成が有効であることが考えられる。したがって、次節以降で call 単位における笑い声合成について検討する。

5.3 笑い声コンテキストの定義

本研究で定義されたコンテキストを表 5.3 に示す。声道形状と音韻によって特徴付けられる音の違いを表現するために、当該 call の音韻転記が定義された。ここで、call の転記には前章で定義された仮名表記を音素系列に変換したものが用いられた。また、音韻の種類だけではなく補助記号によって表される有声/無声、口音/鼻音、長音の有無も明確に区別された。

表 5.3: 本研究で定義された笑い声コンテキスト

c_c：当該 call の音韻転記	
c_l：先行セグメントの音韻転記	A
c_r：後続セグメントの音韻転記	A
p_l：発話における bout 位置	B
p_c：bout における call 位置	B
n_c：bout を構成する call 数	B

更に、2 つのグループがコンテキストに追加された。グループ A は狭い意味での文脈を考慮する要素の集合であり、先行および後続の「セグメント」の音韻転記である。ここで、「セグメント」とは前の音が笑い声である場合には call に相当し、言語音である場合には音素に相当する。

グループ B では、より広い意味での文脈を考慮するための要素の集合である。このグループには、笑い声の発話における位置が定義されている。この位置は発話の先頭、発話内、末尾かどうかを区別するために使用される。また、call の継続長は call 位置によって異なるということが報告されている [93]。このことを考慮するために、bout における call 位置が定義された。更に、HMM 音声合成方式による言語音の合成では、発話のモーラ数が変動要因として広く用いられている。これを参考に、笑い声の 1 bout における call 数が定義された。

5.4 笑い声合成

5.4.1 合成条件

本研究では、UUDB の話者 FTS および OGVC の話者 06_FTY、06_FWA の計 3 名の笑い声を対象とする。ラベリングの段階では 03_FTY の笑い声も対象としていたが、収録品質の低さから笑い声合成の検討では除かれた。モデルの学習に使用する笑い声の総数は 109 bouts である。笑い声の音響特徴量の分析条件およびモデル学習に使用する特徴量ベクトルとモデルの構成を表 5.4 に示す。また、モデル学習には共有決定木を用いた話者適応学習技術 [97, 98] が使用された。

表 5.4: 笑い声モデルの学習条件

モデル	5 状態の left-to-right HSMM
特徴量ベクトル	0 から 39 次のメルケプストラム係数, 対数基本周波数, それぞれの Δ および ΔΔ を 含めた 123 次元ベクトル
分析	サンプリング周波数 16 kHz の笑い声に対して 窓長 25 ms、フレームシフト 5 ms とした ハミング窓による分析

テスト用の笑い声は、学習に使用する笑い声の中から選択し、その笑い声を除いて学習されたモデルから合成された (Leave-one-out 法)。各合成笑い声は元々の話者のモデルに話者適応された。

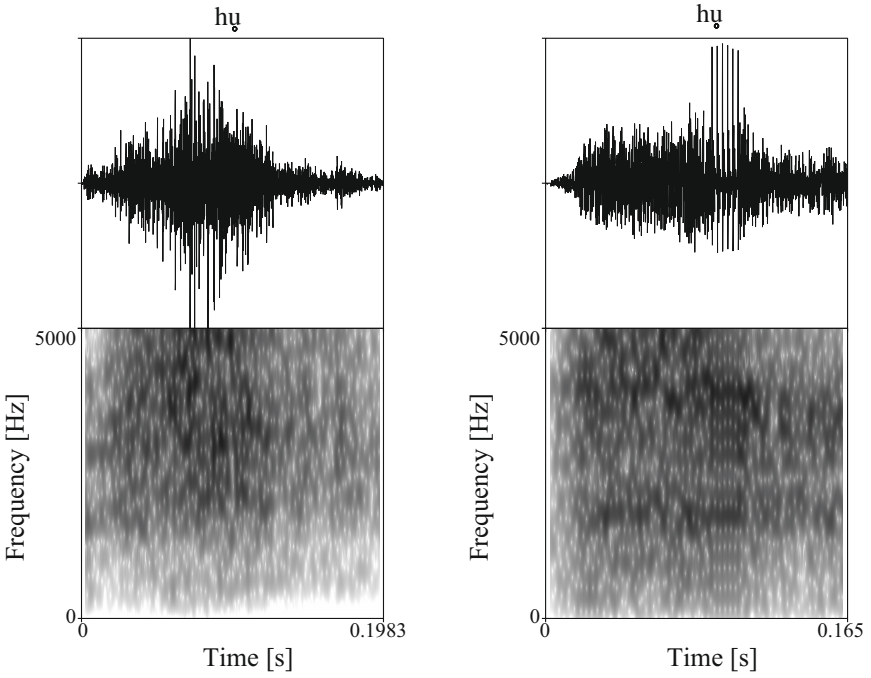
5.4.2 合成結果

合成された笑い声の例として、single-call bout の笑い声 [hu̥] の波形およびサウンドスペクトログラムを図 5.1 に示す。笑い声 [hu̥] は図 5.1 (a) に示すように声帯の振動を伴わない摩擦音であり、実際の対話場面では自嘲的に笑う場面や忍び笑いといった場面に現れる典型的な笑い声である。図 5.1 (b) が合成された [hu̥] の波形とサウンドスペクトログラムであり、自然笑い声の乱流によく似た笑い声が合成されていることを確認することができる。

同様に、multi-call bout の笑い声 [hu̥huhu] の波形およびサウンドスペクトログラムを図 5.2 に示す。図 5.2 (a) は自然笑い声のものであり、図 5.2 (b) は合成された笑い声の結果である。図 5.2 (b) から、自然笑い声の波形に見られる波形の振幅が徐々に減少するという傾向が反映されていることがわかる。このことから、比較的に自然笑い声に近い笑い声が合成されていることがわかる。

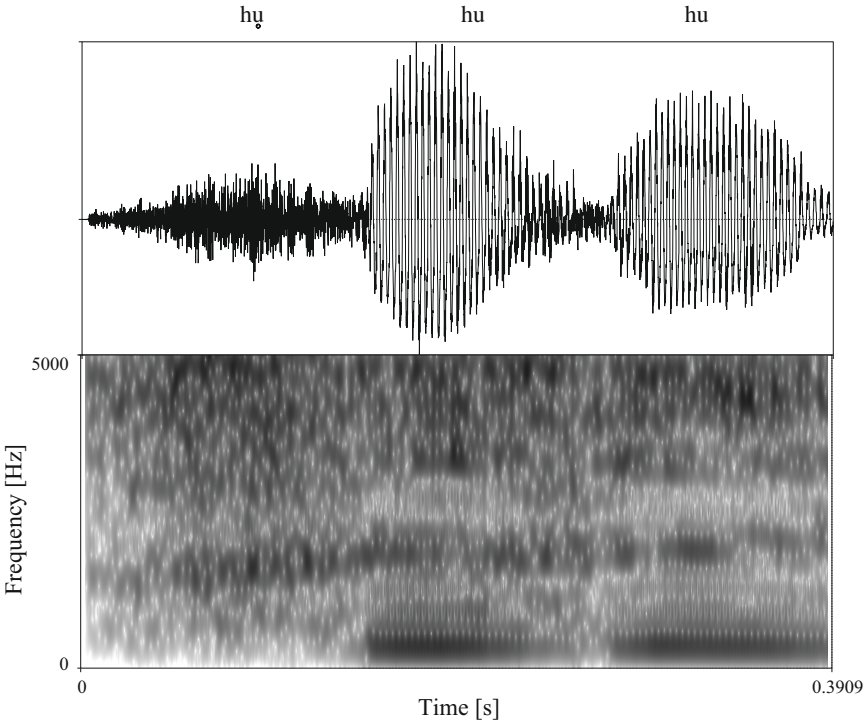
5.5 自然性評価実験

この節では、前節で合成された文脈を考慮した笑い声が、発話全体としての自然性を改善しているかを確かめるために、自然性評価を行う。

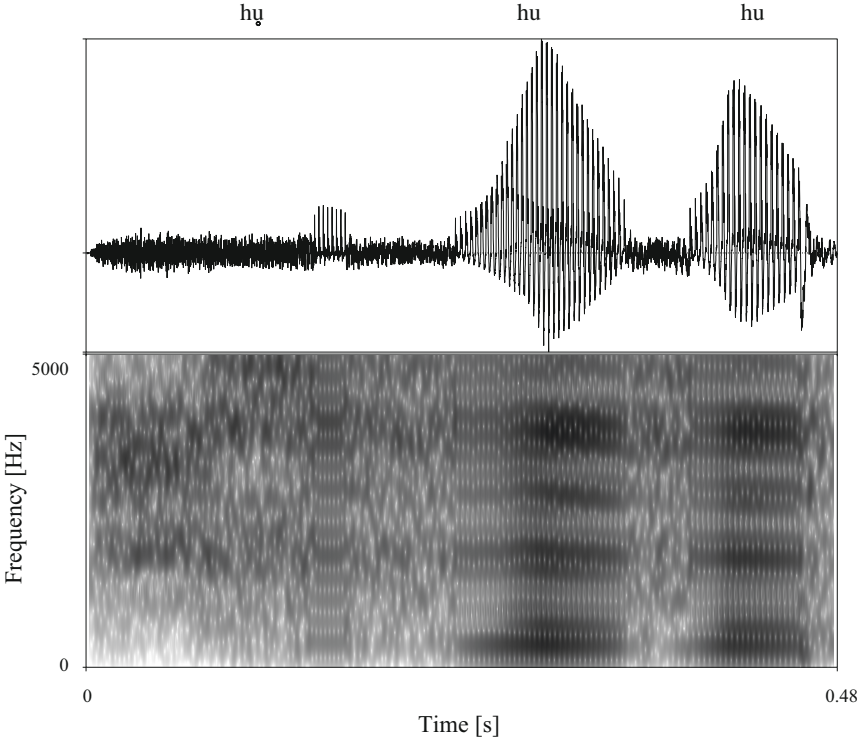


(a) natural *single-call bout*　　(b) synthesized *single-call bout*

図 5.1: single-call bout の例



(a) natural *multi-call bout*



(b) synthesized *multi-call bout*

図 5.2: multi-call bout の例

発話全体としての自然性を評価するために、合成された笑い声だけではなく、言語音に伴う笑い声が実験に使用される。更に、笑い声コンテキストの有効性も確認するために、適用するコンテキストを段階的に増やした時の自然性を比較する。

5.5.1 実験条件

実験に使用する笑い声は以下の 3 つの条件で合成された。

- ベースライン (BL)
  当該 call のみを考慮したコンテキストを用いて合成された笑い声
- ベースライン+グループ A (BL+A)
  BL におけるコンテキストに加えて、先行・後続セグメントの音韻を考慮したコンテキストを用いて合成された笑い声
- ベースライン+グループ A+グループ B (BL+AB)
  BL+AB におけるコンテキストに加えて、発話における bout 位置、bout における call 位置、bout を構成する call 数を考慮したコンテキストを用いて合成された笑い声

上記の条件で合成された笑い声が言語音に接続された。言語音部分は自然音声の分析合成によって合成され、各条件の笑い声に対して接続された。刺激の数は各条件で 46 個ずつであり、総数は 138 個である。

実験には 5 人の男子大学生および 5 人の男子大学院生が参加した。また、参加者は全員日本語話者である。被験者には、各刺激の自然性を 5 段階 (1:不自然, 2:やや不自然, 3:どちらともいえない, 4:やや自然, 5:自然) で評価させた。自然性は「言語音部分と笑い声部分の分節的・韻律的整合性の度合い」と定義した。これは、言語音部分と笑い声部分のパラ言語的な整合性の度合いについても暗に評価している。

実験はヘッドホン (AKG K271 MKII) による両耳聴取によって行われた。刺激は静かな研究室にて呈示され、各刺激は被験者に対し 1 回だけ呈示された。

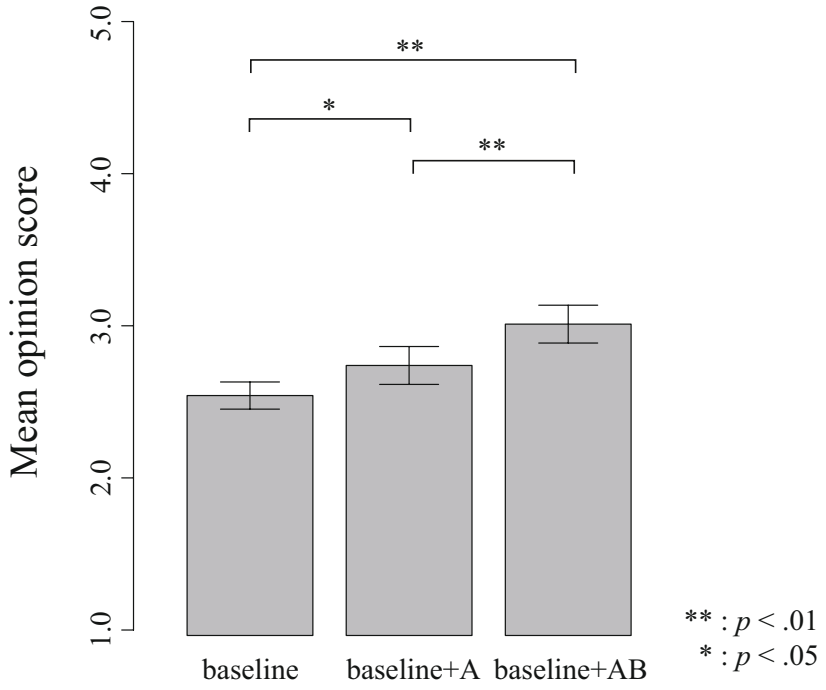
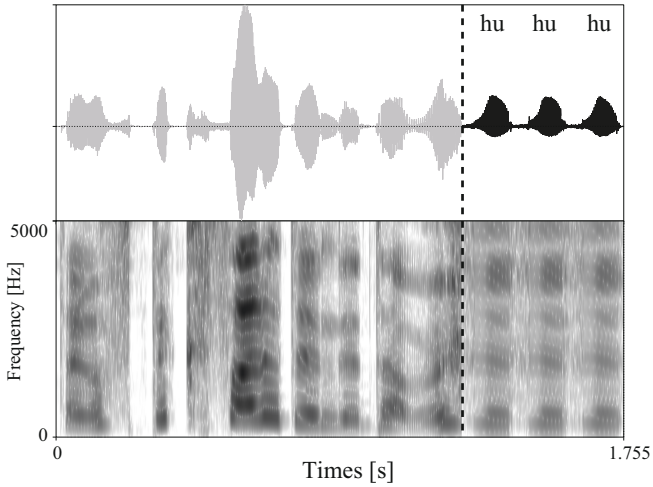


図 5.3: 自然性評価の平均評価値の分布

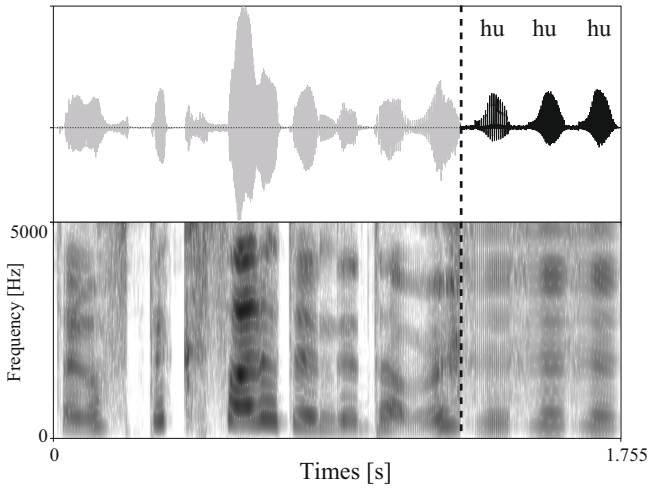
5.5.2 実験結果

自然性評価実験の結果として、被験者による平均評価値 (MOS) の分布を図 5.3 に示す。図中のバーは 95%信頼区間を表す。BL および BL+A、BL+AB の MOS の平均はそれぞれ 2.54, 2.74, 3.01 である。これらの分布に対して合成条件を要因とする分散分析を行った結果、主効果が有意であった ($F(2, 135) = 17.34, p < .01$)。そのため、Tukey HSD 法による多重比較を行った。その結果、BL と BL+A ($p < .05$)、BL+A と BL+AB ($p < .01$)、BL と BL+AB ($p < .01$) の間の差が有意であった。これらの結果から、文脈に関するコンテキストを増やすことによって自然性が上昇することを確認した。最も自然であった条件は BL+AB であり、前後の音韻だけではなく、笑い声が置かれている文脈についての情報が自然性改善に寄与していることがわかる。

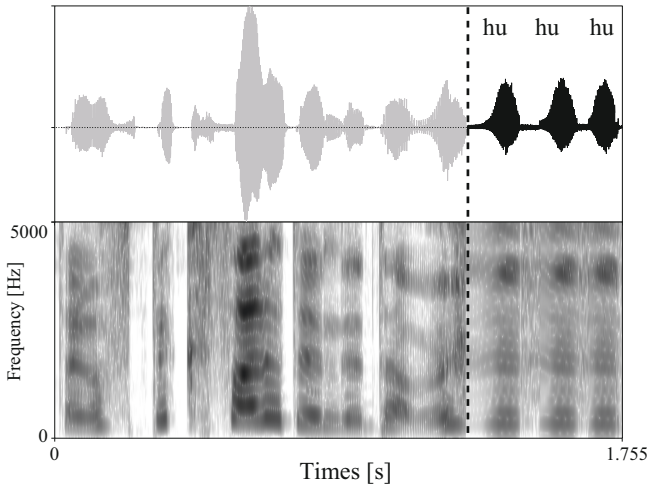
自然性が改善された例を図 5.4 に示す。図は笑い声を含む発話の波形およびサ



(a) Synthesized with baseline



(b) Synthesized with baseline+A



(c) Synthesized with baseline+AB

図 5.4: 全体の自然性が向上した例

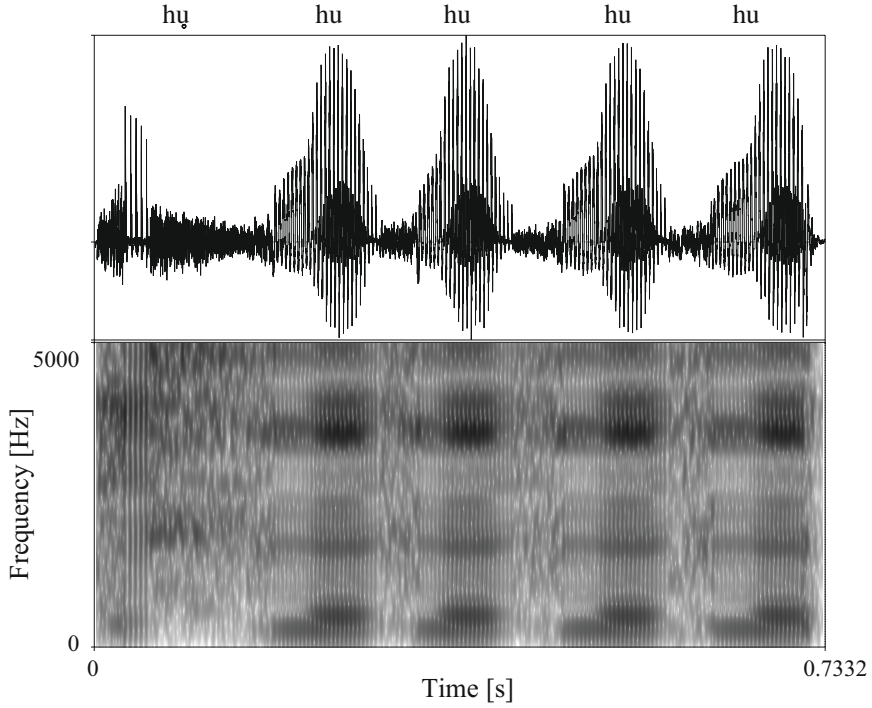


図 5.5: 全体の自然性が低い刺激の例

ウンドスペクトログラムを示しており、薄くなっている部分は言語音部分を表す。これは multi-call bout [huhuhu] を合成した例であり、発話の末尾に笑い声が位置している。図 5.4 (a) は BL によって合成された笑い声が接続されている。BL によって合成された笑い声は全ての call が同じ音響的特徴を持っている。これは、同じ音韻であれば全て同じモデルから合成されるからであり、このように音響特徴量の変化のない単調な音は不自然に知覚される。図 5.4 (b) は先行・後続のセグメントを考慮して合成された笑い声が接続された例である。前後の音を考慮したことにより、全ての call が同じ音響特徴量で出力されることはなく、多少自然性が改善されている。図 5.4 (c) はより広義の文脈を考慮して合成された笑い声が接続された例である。単に各 call の音響特徴量が異なっているだけではなく、call の振幅が言語音の部分の振幅に近い値になっていることがわかる。このように、文脈を考慮することによって言語音部分と笑い声部分の韻律的・分節的な特徴が整合されたことが自然性改善に寄与していると考えられる。

しかしながら、いくつかの刺激については自然性が改善されない例があった。

それは笑い声自体の自然性が低い合成笑い声である。図 5.5 に、笑い声自体の自然性が低い合成笑い声の波形およびサウンドスペクトログラムを示す。これは、[hu̥huhuhuhu] を合成した例である。文脈を考慮して合成したにも関わらず、[hu] の部分が同じ音響特徴量となっており、非常に単調で不自然な笑い声となっている。コンテキストに基づく手法では、このように必ずしも音響特徴量の変化を反映するように学習されるとは限らないため、音響特徴量の時間的な変化を明示的に反映するような手法が必要になると考えられる。

5.6 パラ言語情報知覚実験

本節では、笑い声が対話においてどのような機能を果たしているかを確かめるために、笑い声を含む刺激と笑い声を含まない刺激から知覚されるパラ言語情報を比較した。また、笑い声の形態的な違いと機能の関係についても論じる。

5.6.1 実験条件

本実験では、重回帰 HSMM(MRHSMM) に基づいて合成された言語音と、HMM 音声合成方式によって合成された笑い声が使用される。

言語音部分のモデルには、5 状態の left-to-right MRHSMM が用いられた。重回帰モデルの説明変数には「快-不快」、「覚醒-睡眠」の 2 次元が用いられた。スペクトルパラメータには STRAIGHT 分析によって得られた STRAIGHT スペクトルから抽出された 39 次のメルケプストラム係数が使用された。音源パラメータには基本周波数パターン生成モデルによってスムージングされた対数基本周波数が用いられた。特徴量ベクトルはこれらのパラメータとそれぞれのデルタ、デルタパラメータを含めた 123 次元のベクトルとした。言語音のモデル学習には UUDB の話者 FTS の 651 発話が使用された。

笑い声部分のモデル学習条件については 5.4.1 節と同様である。また、笑い声コンテキストには表 5.3 で定義された全ての要素を用いている。

言語音部分の合成内容は UUDB の笑い声を含む発話から選択され、/a/系, /u/系, /e/系の笑い声を含む発話をそれぞれ 10 発話ずつ (笑いが先頭の 5 発話と笑

いが末尾の 5 発話) 選択した。また、合成時には UUDB に与えられている評価値の平均を与えて合成した。

笑い声部分については以下の条件を系統的に変化させて合成された。

- 有声/無声
- single-call/multi-call

したがって、刺激の総数は 10(発話内容) $\times$ 3(母音性) $\times$ 2(有声性) $\times$ 2(*bout* 構造) = 120 に笑い声を含まない刺激 30 個を加えた計 150 個である。

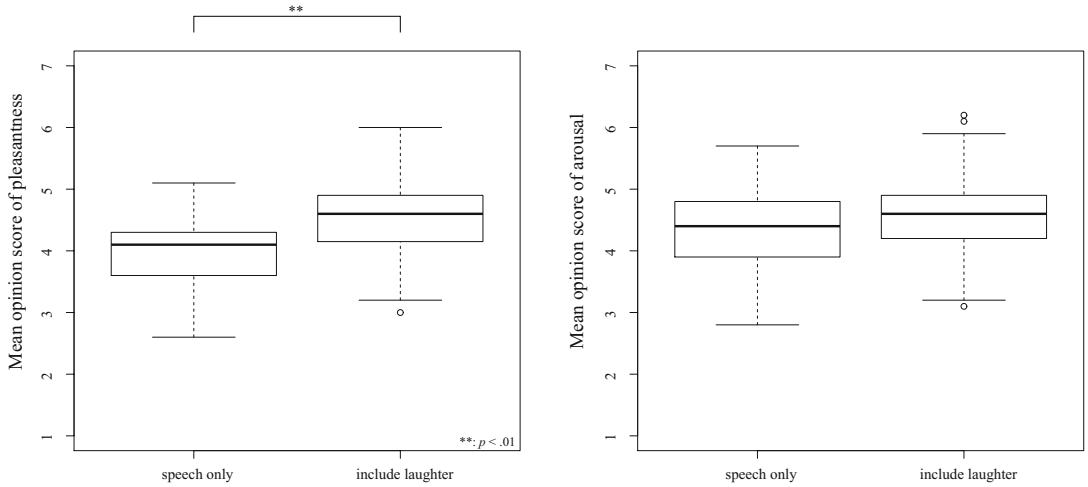
実験には、男子大学生 5 名および男子大学院生 5 名が被験者として参加した。被験者には呈示された刺激から知覚される「快-不快」、「覚醒-睡眠」の度合いを 7 段階で評価するように指示した。刺激の呈示回数は 1 回である。刺激は静かな実験室内による両耳聴取によって行われた。刺激の呈示にはヘッドホン (AKG K271 MKII) が使用された。

5.6.2 実験結果

実験結果として、笑い声を含む発話と含まない発話から知覚される「快-不快」、「覚醒-睡眠」の分布を図 5.6 に示す。図より、笑い声を含む発話は含まない発話よりも快よりに知覚されていることがわかる (対応のない t 検定: $t(148) = -4.49, p < .01$)。一方、「覚醒-睡眠」については差がなかった (対応のない t 検定: $t(148) = -1.49, p > .05$)。

この「快-不快」の分布の違いがどの要因によるものなのかを調べるために、有声/無声、笑い声の位置、笑い声の形態を要因とした 3 要因の分散分析を行った。各要因の水準は有声/無声、先頭/末尾、単一 call/複数 call の 2 水準である。分散分析の結果、有声/無声および笑い声の形態による主効果が有意であった (有声/無声: $F(1, 112) = 15.1, p < .01$, 形態: $F(1, 112) = 9.20, p < .01$)。また、1 次および 2 次の交互作用は見られなかった。

有声/無声および形態ごとの「快-不快」の評価値の分布を図 5.7 に示す。図 5.7(a) より、有声の笑い声の方がより快に知覚される傾向があることがわかる。また、図 5.7(b) より、複数 call による笑い声の方が快よりに知覚される傾向があることもわかる。



(a) Pleasantness (b) Arousal

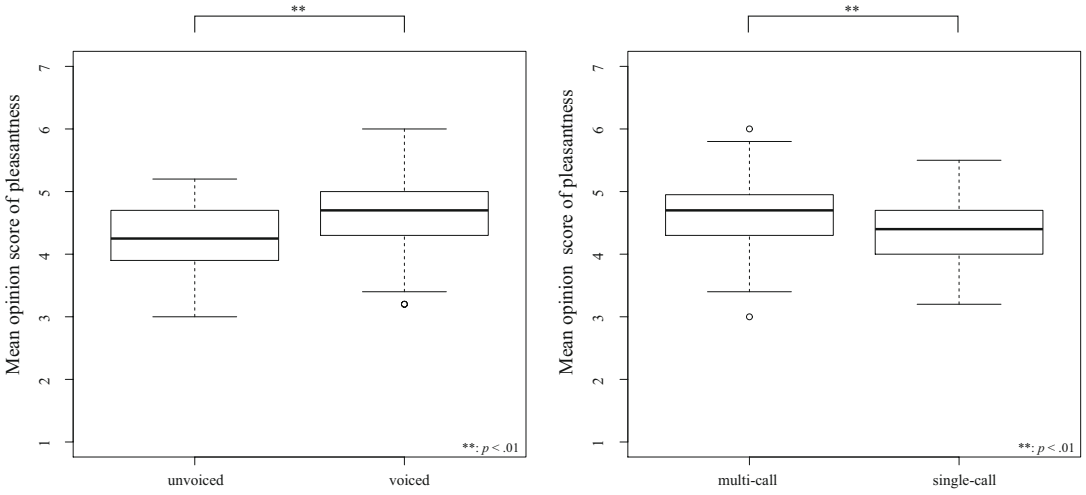
図 5.6: 笑い声を含む発話と含まない発話のパラ言語情報の分布

「覚醒-睡眠」についても同様に 3 要因の分散分析を行った。分散分析の結果、有声/無声による主効果が有意であった ($F(1, 112) = 28.0, p < .01$)。1 次および 2 次の交互作用は確認されなかった。図 5.8 に「覚醒-睡眠」の MOS の分布を示す。「快-不快」と同様に、有声の方が覚醒よりに知覚される傾向があることがわかる。

5.7 おわりに

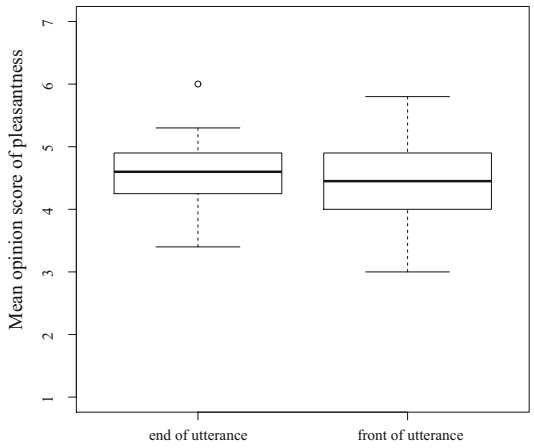
本章では、自然対話音声コーパスに含まれる笑い声を対象とし、統計的パラメトリック合成方式を用いた笑い声の合成を行った。統計的パラメトリック合成手法には HMM 音声合成方式が用いられた。HMM 音声合成の枠組みで様々な文脈を考慮した笑い声を合成するために、笑い声に対するコンテキストが定義された。基本的なコンテキストとして、前後の音韻転記が定義された。また、更に広義の文脈を考慮するために、笑い声の位置や長さに関する要因が定義された。合成された笑い声と自然笑い声を比較し、自然笑い声の特徴に近い笑い声が合成されていることを確認した。

文脈を考慮したことの有効性を確認するために、合成された笑い声に対する主観評価実験が行われた。主観評価実験では、文脈を考慮した笑い声が考慮し



(a) voicing

(b) form

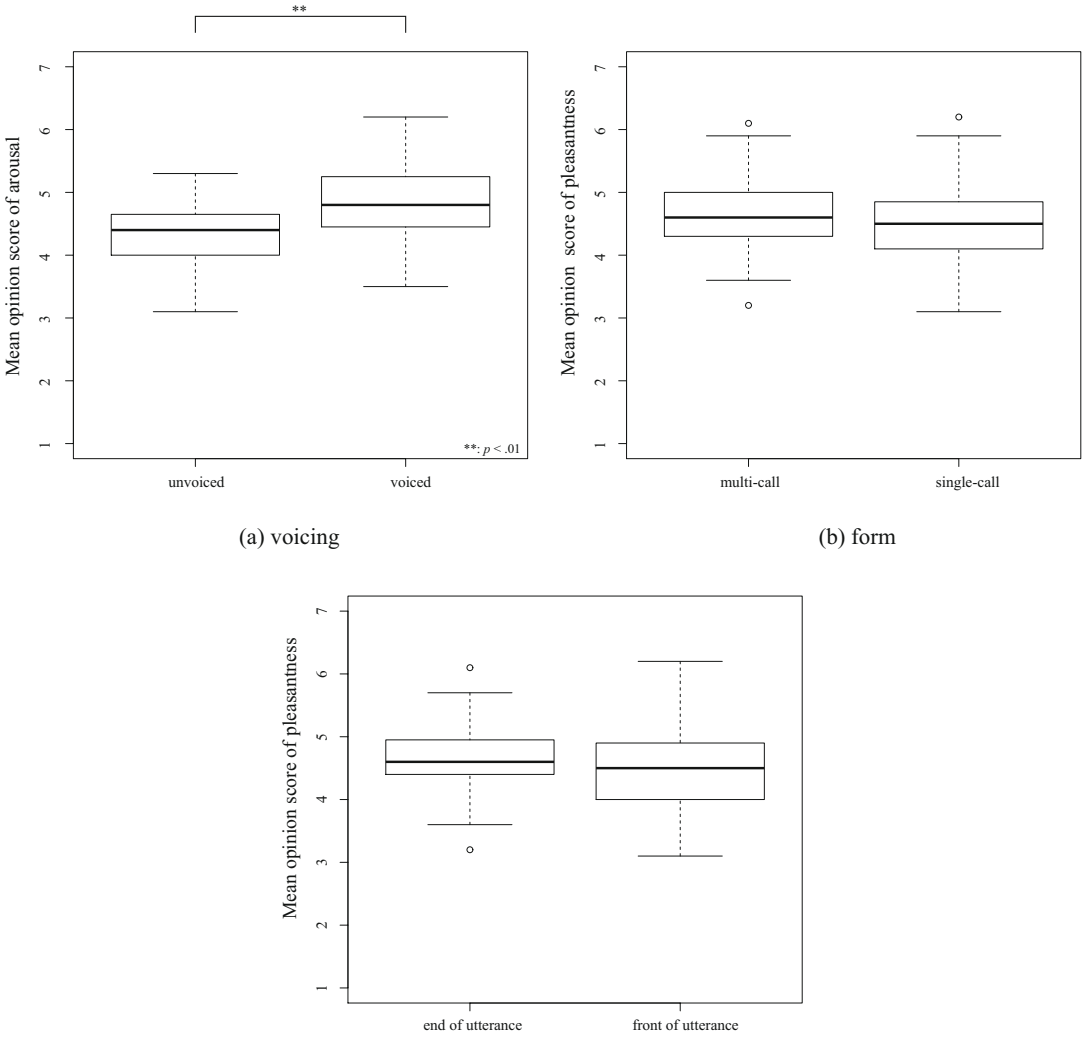


(c) position

図 5.7: 「快-不快」の MOS の分布

ない笑い声よりも発話全体としての自然性を改善させることを確認するために、笑い声単体ではなく、言語音に伴う笑い声に対して自然性評価実験が行われた。自然性評価実験では言語音部分と笑い声部分の分節的・韻律的な整合性の度合いの観点から評価が行われた。自然性評価実験の結果から、文脈を考慮することで発話全体としての自然性が改善されることを明らかにした。

残された課題として、更なる自然性の改善が挙げられる。これには、学習データの増加や更なる変動要因の検討などが挙げられる。特に、変動要因については、発話者に関する情報だけではなく、対話相手とのインタラクションについても考慮する必要があると考えられる。Truong らは対話相手の笑い声に重複するように発せられた笑い声は、重複しない笑い声と異なる音響的特徴を持つということを報告している [99]。これはすなわち、対話場面における笑い声を合成するためには、発話者だけでなくインタクラションについても注目しなければならないことを意味している。



(a) voicing

(b) form

(c) position

図 5.8: 「覚醒-睡眠」の MOS の分布

第6章　結論

人間同士のコミュニケーションでは言語以外の手段を用いたノンバーバルコミュニケーションが行われている。人間と機械のコミュニケーションを目的としたソフトウェアやアプリケーションに応用可能であると考えられるノンバーバル情報を表現可能な音声合成技術として、パラ言語情報を表現可能な音声合成技術および代表的なパラ言語である笑い声の合成について論じた。

パラ言語情報を表現可能な音声合成では、統計的パラメトリック音声合成方式におけるモデルパラメータをパラ言語情報を説明変数とする重回帰モデルに基づいて変換することにより、パラ言語情報の反映された音声を合成する方式について論じた。また、本研究では実際の対話場面において伝達されるパラ言語情報に焦点を当て、自然対話音声コーパスを使用する。これまで、パラ言語情報を表現するための音声合成研究の多くは自然対話音声コーパスではなく、演技音声コーパスを採用しており、自然対話音声を使用した研究はほとんどなかった。そのため、自然対話音声を用いる場合の困難さや問題点などが不明瞭であった。本研究では、自然対話音声コーパスを使用する際に生じる問題点として、統計的パラメトリック音声合成における統計モデルパラメータの過推定問題があることを明らかにし、その過推定問題を解決するための、統計モデルパラメータのロバストな推定方法として、最尤基準ではなく事後確率最大基準を用いることを提案し、その推定式を導出した。

提案したパラ言語情報の表現手法は客観評価および主観評価により有効性が検討された。客観評価では、提案したロバスト推定手法を用いることにより、極端な音響的特徴の出力が抑制されることを確認した。また、主観評価はパラ言語情報の伝達性および自然性により評価された。これらの結果より、提案法は比較的高いパラ言語情報の伝達性があることと、ロバスト推定手法により自然性が改善されることを示した。

笑い声の合成では、自然対話音声コーパスに含まれる笑い声を対象として合成が行われた。一般に、自然対話コーパスに笑い声に関する情報はほとんど与えられていないため、笑い声のアノテーションが実施された。自然対話音声コーパスに含まれる笑い声の合成では、お笑い映像などによって誘発された笑い声を合成する従来の研究と比較して十分な量の笑い声を用意することができない。そのため、少ないデータでも効率的に笑い声の音響的特徴をモデル化できるポテンシャルを持つ統計的パラメトリック音声合成手法を用いて笑い声合成が行われた。

また、自然対話音声に含まれる笑い声には単独の笑い声だけではなく、話してから笑う、または笑ってから話すといった言語音に付随する笑い声が存在する。従来のお笑い映像などによって誘発された笑い声には、言語音に付随する笑い声は少なく、そのような発話を合成する際にはたとえ笑い声自体の品質が良くても発話全体としての自然性が低下することが問題視されていた。本研究では、自然対話音声コーパスを使用することにより、笑い声が置かれている文脈や状況を考慮した笑い声を合成することを提案した。統計的パラメトリック音声合成方式では、合成単位である音素の音響的特徴の変動要因 (コンテキスト) に依存するコンテキスト依存モデルを学習することによって、音素の置かれている環境および文脈を考慮している。そこで、笑い声に対して適切なコンテキストを定義し、笑い声コンテキストに依存したモデルを構築することによって、文脈を考慮した笑い声を合成することを提案した。提案手法は主観評価により有効性が検討された。主観評価では、文脈を考慮するためのコンテキスト要素を増やすことにより、発話全体としての自然性が改善されることを示し、提案法の有効性が示された。

これらの結果から、本研究は人間と機械のコミュニケーションへの応用が期待される音声合成におけるノンバーバル情報の表現を実現できたと言える。本研究の成果が音声合成を始めとするノンバーバル研究の足掛かり、発展に寄与することを期待している。

本研究ではパラ言語情報を表現可能な音声合成技術と、パラ言語情報の伝達に関係の深い代表的な非言語音である笑い声の合成を行い、有効性を確認した。しかし、解決しなければならない問題は依然多く残されている。

特に課題が残されているのは、笑い声の合成についてである。本研究で明らかにできたのは、笑い声に対しても言語音の場合と同様に適切なコンテキストを定義することによって、様々な笑い声を合成しわけることができたという点、文脈を考慮した笑い声を合成可能であるという点、そして文脈を考慮するための簡易的な要素を確認したという点のみである。

今後解決しなければならない問題として、形態的に異なって合成された笑い声を用いて、笑い声が対話においてどのような役割を果たしているのかという機能を調査することが挙げられる。これは言い換えれば、笑い声の形態と機能との対応関係を明らかにすることである。すなわち、意図したノンバーバル情報を伝えるためにはどのような形態を合成すればよいかという、笑い声合成を行うための入力を用意することに等しい。現在、笑い声を合成する際には、合成するためのコンテキストは既知であるとして与えている。しかしながら、実際に対話システムなどで利用することを考えた場合には笑い声のコンテキストを用意する必要があり、その入力をどのように決定するのかという問題を解決しなければならない。

謝辞

ご多忙であるにもかかわらず、主任指導教員として熱心に御指導くださいました宇都宮大学大学院工学研究科 森大毅准教授に心より感謝を申し上げます。研究方針や研究内容への御指摘・御助言はもちろんのこと、生活面、精神面においても細やかなご配慮を頂き、深く感謝しています。国内だけではなく、国外でも研究発表の機会に恵まれたことは、森先生のお力添えがあってのものだと感じています。

本研究を進めるにあたり、丁寧な御指導を頂きました副指導教員の平田光男教授、東剛人准教授に深く感謝致します。研究に関する客観的な御指摘・御意見は本論文をまとめる助けとなりました。

副専門研修において丁寧に御指導頂いた横田隆史教授、白石和男氏、依田秀彦准教授に深く感謝致します。他分野に関する研修はとても新鮮であり、自分の視野を広げる有益な機会を得ることができました。

研究を行うにあたり、研究に関する貴重な御意見・御助言を頂き、さらには技術的にも御支援頂いた東北大学大学院工学研究科 能勢隆准教授に深く感謝致します。

研究に関する鋭い御指摘・御意見を頂いた帝京大学理工学部 有本泰子氏に深く感謝致します。研究に関することだけではなく、勉強会などにもお付き合い頂き、感謝しています。

研究面だけではなく、生活面においても精神的援助を頂いた森研究室の皆様に感謝を申し上げます。時には苦しいこともありましたが、乗り越えてこれたのは森先生と、皆様の支えがあってこそだと思っています。同期であり、卒業後も時間を作って研究のディスカッションにお付き合い頂いた田澤祥亨氏、忌憚のない意見や質問から研究に関する様々なヒントを頂いた高橋俊介氏、物忘れの激しい私をサポートして頂いた鈴木圭氏。多くのご迷惑をおかけしました

が、これまで支えてくださったことに深く感謝します。
　最後に、これまでの生活を支えてくださった父に深く感謝します。

参考文献

[1] 森大毅，“音声が伝えるものとは？，” 日本音響学会秋季研究発表会講演論文集，pp.239–242，2012.

[2] 石黒昭博，山内信幸，赤楚治之，北林利治，菊田千春，伊藤徳文，須川精致，川本裕未，“現代の言語学，” 東京：金星堂，1996.

[3] J.A. Russell, “A circumplex model of affect,” Journal of Personality and Social Psychology, vol.39, no.6, pp.1161–1178, 1980.

[4] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A.W. Black, and K. Tokuda, “The HMM-based speech synthesis system (HTS) version 2.0,” Proceedings of workshop on Speech Synthesis, pp.294–299, 2007.

[5] J. Trouvain, “Segmenting phonetic units in laughter,” Proceedings of ICPhS, pp.2793–2796, 2003.

[6] 志水彰，角辻豊，中村真，“人はなぜ笑うのか—笑いの精神生理学，” 講談社，1994.

[7] 大塚祥平，“笑い声の音響モデルに関する基礎的研究，” 宇都宮大学卒業論文，2015.

[8] H. Fujisaki, “Prosody, models, and spontaneous speech,” Computing Prosody, eds. by Y. Sagisaka, N. Campbell, and N. Higuchi, pp.27–42, Studies in Linguistics, 1996.

[9] G.L. Trager, “Paralanguage : A first approximation,” Studies in Linguistics, vol.13, pp.1–12, 1958.

[10] D. Crystal, “Palalinguistics,” The Body as a Medium of Expression, edited by Jonathan Benthall and Ted Polhemus, pp.162–174, 1975.

[11] D.R. Ladd, “Intonational phonology,” Cambridge: Cambrige Uviversity Press, 1996.

[12] R. Bense and K.R. Scherer, “Acoustic profiles in vocal emotion expression,” Journal of Personality and Social Psychology, vol.70, no.3, pp.614–636, 1996.

[13] M.J. Owren and J.-A. Bachrowski, “The evolution of emotional expression: a selfish-gene account of smiling and laughter in early hominids and humans,” Mayne T., Bonanno G.A. (Eds.), Emotions: Current issues and future development, pp.152–191, 2001.

[14] S. Seneff, E. Hurley, R. Lau, C. Pao, P. Schmid, and V. Zue, “Galaxy-II: A reference architecture for conversational system development,” Proceedings of ICSLP, pp.931–934, 1998.

[15] V. Zue, S. Senef, J. Glass, J. Polifroni, C. Pao, T.J. Hazen, and L. Hetherington, “JUPITER: A telephone-based conversational interface for weather information,” IEEE Transactions on Speech and Audio Processing, vol.8, no.1, pp.100–112, 2000.

[16] 有田正剛, 島津秀雄, “カーナビゲーションシステム用音声対話インタフェース,” 人工知能学会研究会資料, 1995.

[17] 前川喜久雄, “コーパスを利用した自発音声の研究,” 東京工業大学大学院情報理工学研究科博士学位論文, 2011.

[18] M. Goto, K. Itou, and S. Hayamizu, “A real-time filled pause detection system for spontaneous speech recognition,” Proceedings of Six European Conference on Speech Communication and Technology, pp.227–230, 1999.

[19] 藤江真也，江尻康，菊池英明，小林哲則，“肯定的／否定的発話態度の認識とその音声対話システムへの応用，” 電子情報通信学会論文誌，vol.88-D2，no.3，pp.489–498，2005.

[20] 堂坂浩二，安田宣仁，宮崎昇，中野幹生，相川清明，“音声対話システム「飛遊夢 (ひゅうむ)」，” 電子情報通信学会講演論文集，pp.506–507，2001.

[21] 野田喜昭，山口義和，大附克年，小川厚徳，中川聡，今村明弘，“音声認識エンジン VoiceRex の開発，” 音響学会秋季講演論文集，pp.93–94，1999.

[22] 飯田朱美，ニック・キャンベル，安村通晃，“感情表現が可能な合成音声の作成と評価，” 情報処学会論文誌，vol.40，no.2，pp.479–486，1999.

[23] 都築亮介，全炳河，徳田恵一，北村正，“HMM 音声合成における感情表現のモデル化，” 電子情報通信学会技術報告，vol.78，pp.25–30，2003.

[24] H. Mori, T. Satake, M. Nakamura, and H. Kasuya, “Constructing a spoken dialogue corpus for studying paralinguistic information in expressive conversation and analyzing its statistical/acoustic characteristics,” Speech Communication, vol.53, pp.36–50, 2011.

[25] Y. Arimoto, H. Kawatsu, S. Ohno, and H. Iida, “Naturalistic emotional speech collection paradigm with online game and its psychological and acoustical assessment,” Acoustical Science and Technology, vol.33, pp.359–369, 2012.

[26] R.L. Birdwhistell, “Introduction to kinesics: An annotation system for analysis of body motion and gesture,” University Press of Kentucky, 1952.

[27] E.T. Hall, “The silent language,” New York: Doubleday, 1959.

[28] M.F. Vargas, “LOUDER THAN WORDS – an introduction to nonverbal communication –,” Iowa State University Press, 1986.

[29] マジョリー・F・ヴァーガス (著)，石丸正 (訳)，“非言語コミュニケーション，” 新潮社，1986.

[30] R.M. Krauss, P. Morrel-Samupels, and C. Colasante, “Do conversational hand gestures communicate?,” Journal of Personality and Social Psychology, vol.61, pp.743–754, 1991.

[31] S.D. Kelly, D.J. Barr, R.B. Church, and K. Lynch, “Offering a hand to progmatic understanding: the role of speech and gesture in comprehension and memory,” Journal of Memory and Language, vol.40, pp.577–592, 1999.

[32] 藤原武弘, “態度変容と印象形成に及ぼすスピーチ速度とハンドジェスチャーの効果,” 心理学研究，vol.57，pp.200–206，1986.

[33] 大神優子，“「わかりやすい説明」の特徴—発話の身振りの分析から—,” 日本教育心理学第 41 回発表論文集，p.395，1999.

[34] P. Ekman and W.V. Friesen, “Unmasking the face: A guide to recognizing emotions from facial cues,” Engle wood Cliffs, New Jersey: Prentice-Hall, 1975.

[35] P. Ekman and W.V. Friesen, “Facial action coding system,” Consulting Psychologists Press, 1975.

[36] M. Argyle and M. Cook, “Gaze and mutual gaze,” Chambrige: Chambrige University Press, 1976.

[37] E.T. Hall, “The hidden dimension,” Garden City, NJ: Doubleday Anchor, 1966.

[38] 前川喜久雄，北川智利，“音声はパラ言語情報をいかに伝えるか,” 認知科学，vol.9，no.1，pp.46–66，2002.

[39] 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 f0 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，vol.89，no.8，pp.1811–1819，2006.

[40] 平賀裕，斎藤善行，森島繁生，原島博，“音声に含まれる感情情報抽出の一検討,” ヒューマンコミュニケーション，vol.93，no.439，pp.1–8，1994.

[41] F. Dellaert, T. Polzin, and A. Waibel, “Recognizing emotion in speech,” Proceedings of 4th International Conference on Spoken Language Processing, vol.3, pp.1970–1973, 1996.

[42] S. McGilloway, R. Cowie, E. Cowie, S. Gielen, M. Westerdijk, and S. Stroeve, “Approaching automatic recognition of emotion from voice: A rough benchmark,” Proceedings of ISCA Workshop on Speech and Emotion, pp.1367–1370, 2001.

[43] C. Lee, S. Narayanan, and R. Pieraccini, “Recognition of negative emotions from the speech signal,” Proceedings of 7th IEEE Workshop on Automatic Speech Recognition and Understanding, pp.240–243, 2001.

[44] J. Ang, R. Dhillon, A. Krupski, E. Shriberg, and A. Stolcke, “Prosody-based automatic detection of annoyance and frustration in human-computer dialog,” Proceedings of 7th International Conference on Spoken Language Processing, pp.2037–2040, 2002.

[45] Y. Kitahara and Y. Tohkura, “Prosodic control to express emotions for man-machin speech interaction,” IEICE Transactions on Fundamentals, vol.75, no.2, pp.155–163, 1992.

[46] M. Schröder, “Emotional speech synthesis: A review,” Proceedings of EUROSPEECH, pp.561–564, 2001.

[47] 黒川隆夫，渡辺富夫，“ノンバーバルコミュニケーションとインタフェース，” ヒューマンインタフェース学会誌，vol.3，no.2，pp.91–98，2001.

[48] 山口貴史，井上昂治，吉野幸一郎，高梨克也，N.G. Ward，河原達也，“傾聴対話システムのための言語情報と韻律情報に基づく多様な形態の相槌の生成，” 人工知能学会論文誌，vol.31，no.4，pp.C–G31_1–10，2016.

[49] J.E. Cahn, “Generation of affect in synthesized speech,” Proceedings of American voice I/O society, pp.251–256, 1989.

[50] M. Bulut, S.S Narayanan, and A. Syrdal, "Expressive speech synthesis using a concatenative synthesizer," Proceedings of Interspeech, pp.1265–1268, 2002.

[51] A. Iida, N. Campbell, F. Higuchi, and M. Yasumura, "A corpus-based speech synthesis system with emotion," Speech Communication, vol.40, pp.161–187, 2003.

[52] J. Vroomen, R. Collier, and S. Mozziconacci, "Duration and intonation in emotional speech," Proceedings of Eurospeech, pp.577–580, 1993.

[53] E. Rank and H. Pirker, "Generating emotiona speech with a concatenative synthesizer," Proceedings of ICSLP, vol.3, pp.671–674, 1998.

[54] T. yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum pitch and duration in HMM-based speech synthesis," IEEE Transactions on Information and Systems, vol.83, no.11, pp.2347–2350, 1999.

[55] J. Yamagishi, K. Onishi, T. Masuko, and T. Kobayashi, "Modeling of various speaking styles and emotions for HMM-based speech synthesis," Proceedings of Eurospeech, pp.2361–2464, 2003.

[56] M. Tachibana, J. Yamagishi, T. Masuko, and T. Kobayashi, "Speech synthesis with various emotional expressions and speaking styles by style interpolation and morphing," IEICE Transactions on Information and Systems E, vol.88-D(3), pp.1092–1099, 2005.

[57] T. Nose, J. Yamagishi, and T. Kobayashi, "A style control technique for HMM-based expressive speech synthesis," IEICE Transactions on Information and Systems, vol.90, no.9, pp.1406–1413, 2007.

[58] J. Latorre, V. Wan, M.J.F. Gales, K.K. Chin, K. Knill, and M. Akamine, "Speech factorization for HMM-TTS based on cluster adaptive training," Proceedings of Interspeech, pp.971–974, 2012.

[59] A. Andersson, J. yamagishi, and R. Clark, "Prediction and realisation of conversational characteristics by utilising spontaneous speech for unit selection," Speech Communication, vol.54, no.2, pp.175–188, 2012.

[60] R. Dall, M. Wester, and M. Carley, "The effect of filled pauses and speaking rate on speech comprehension in natural, vocoded and synthetic speech," Proceedings of Interspeech, pp.56–60, 2014.

[61] M. Wester, O. Watts, and G.E. Henter, "Evaluating comprehension of natural and synthetic conversational speech," Proceedings of Speech Prosody, pp.766–770, 2016.

[62] T. Koriyama, T. Nose, and T. Kobayashi, "On the use of extended context for HMM-based spontaneous conversational speech synthesis," Proceedings of Interspeech, pp.2657–2660, 2011.

[63] T. Koriyama, T. Nose, and T. Kobayashi, "An f0 modeling technique based on prosodic events for spontaneous speech synthesis," Proceedings of ICASSP, pp.4589–4593, 2012.

[64] M. Schröder, "Speech and emotion research: an overview of research frameworks and a dimensional approach to emotional speech synthesis," PhD thesis, Saarland University, 2011.

[65] D. Wu, T. Parsons, E. Mower, and S.S. Narayanan, "Speech emotion estimation in 3d space," Proceedings of ICME, pp.737–742, 2010.

[66] H. Mori and T. Hitomi, "Annotating conversational speech for corpus-based dialogue speech synthesizer - a first step," Proceedings of Oriental COCOSDA, pp.135–140, 2012.

[67] T. Nose, M. Tachibana, and T. Kobayashi, "HMM-based style control for expressive speech synthesis with arbitrary speaker's voice using model adaptation," IEICE Transactions on Information and Systems, vol.92, no.3, pp.489–497, 2009.

[68] H. Mori, H. Kasuya, M. Nakamura, and M. Amanuma, “Some considerations for designing spoken dialogue database from the view point of paralinguistic information,” Acoustical Science and Technology, vol.24, no.6, pp.376–378, 2003.

[69] 森大毅，相澤宏，粕谷英樹，“対話音声のパラ言語情報ラベリングの安定性，” 日本音響学会誌，vol.61，no.12，pp.690–697，2009.

[70] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, “Hidden semi-marukov model based speech synthesis,” Proceedings of ICSLP, pp.1397–1400, 2004.

[71] 吉岡元貴，田村正統，益子貴史，小林隆夫，徳田恵一，“HMM 音声合成における韻律の変動要因の検討，” 電子情報通信学会技術報告，vol.80，no.34，pp.51–56，2001.

[72] 全炳河，徳田恵一，北村正，“決定木に基づく音素コンテキスト・次元・状態位置の同時クラス タリング，” 日本音響学会秋季研究発表会講演論文集，vol.1，pp.39–40，2002.

[73] A.K. Gupta and T. Varga, “Elliptically contoured models in statistics,” Kluwer Academic Publisers, 1993.

[74] H. Kawahara, I. Masuda-Katsune, and A. 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, pp.187–207, 1999.

[75] H. Fujisaki and K. Hirose, “Analysis of voice fundamental frequency contours for declarative sentences of japanese,” Journal of acoustical society of Japan (E), vol.5, no.4, pp.233–242, 1984.

[76] H. Hashimoto, K. Hirose, and N. Minematsu, “Improved automatic extraction of generation process model commnads and its use for generating

fundamental frequency contours for training HMM-based speech synthesis," Proceedings of Interspeech, pp.458–461, 2012.

[77] R. Tato, R. Santos, R. Kompe, and J.M. Pardo, "Emotional space improves emotion recognition," Proceedings of ICSLP, pp.2029–2032, 2002.

[78] M. Grimm and K. Kroschel, "Emotional estimation in speech using a 3d emotion space concept," Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop, pp.381–385, 2005.

[79] M. Grimm, K. Kroschel, and S. Narayanan, "Support vector regression for automatic recognition of spontaneous emotions in speech," Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), vol.4, pp.1085–1088, 2007.

[80] J. Urbain, H. Cakmak, and T. Dutoit, "Development of HMM-based acoutic laughter synthesis," Proceedings of Interdisciplinary Workshop Laughter and Other Non-Verbal Vocalisations in Speech, pp.26–27, 2012.

[81] J. Urbain, H. Cakmak, and T. Dutoit, "Evaluation of HMM-based laughter synthesis," Proceedings of ICASSP, pp.7835–7839, 2013.

[82] J. Urbain, T. Dutoit, A. Moinet, R. Niewiadomski, C. Pelachaud, B. Picart, J. Tilmanne, and J. Wagner, "The AVLaughterCycle database," Proceedings of LREC, pp.2996–3001, 2010.

[83] P. Boersma and D. Weenink, "Praat: doing phonetics by computer [computer program]". http://www.praat.org/.

[84] 森大毅，"Affect burstの音声学的分析 ー感情表出系感動詞の言語的・パラ言語的特徴ー，" 日本音響学会秋季研究発表会講演論文集，pp.293–296，2015.

[85] K.P. Truong and D.V. Leeuwen, "Automatic detection of laughter," Proceedings of Interspeech, pp.485–488, 2005.

[86] M.T. Knox and N. Mirghafori, "Automatic laughter detection using neural networks," Proceedings of Interspeech, pp.2973–2976, 2007.

[87] T. Neuberger, A. Beke, and M. Gósy, "Acoustic analysis and automatic detection of laughter in hungarian spontaneous speech," Proceedings of ISSP, pp.281–284, 2014.

[88] S. Petridis and M. Pantic, "Is this joke really funny? judging the mirth by audiovisual laughter analysis," Proceedings of ICME, pp.1444–1447, 2009.

[89] A.T. Sathya, K.K. Sudheer, and B. Yegnanarayana, "Synthesis of laughter by modifying excitation characteristics," Journal of Acoustical society of America, vol.133, pp.3072–3082, 2013.

[90] S. Sundaram and S. Narayanan, "Automatic acoustic synthesis of human-like laughter," Journal of Acoustical society of America, vol.121, pp.527–535, 2007.

[91] J. Trouvain and M. Schroder, "How (not) to add laughter to synthetic speech," Proceedings of Workshop on Affective Dialogue Systems, pp.229–232, 2004.

[92] E. Lasarcyk and J. Trouvain, "Imitating conversational laughter with an articulatory speech synthesis," Proceedings of Interdisciplinary Workshop Phonetics of Laughter, pp.43–48, 2007.

[93] J.A. Bachorowski, M.J. Smoski, and M.J. Owren, "The acoustic features of human laughter," Journal of Acoustical society of America, vol.110, pp.1581–1597, 2001.

[94] N. Campbell, H. Kashioka, and R. Ohara, "No laughing matter," Proceedings of Interspeech, pp.465–468, 2005.

[95] H. Tanaka and N. Campbell, "Classification of social laughterin natural conversational speech," Computer Speech and Language, vol.28, pp.314–325, 2014.

[96] 森大毅，“Affect burst の形態論的分類 — uudb を対象とした検討，” 日本音響学会 2015 年春季研究発表会講演論文集，pp.403–404，2015.

[97] 山岸順一，益子貴史，徳田恵一，小林隆夫，“HMM 音声合成におけるコンテキストクラスタリング決定木を用いた話者適応の検討，” 電子情報通信学会技術研究報告，vol.SP2003-79，pp.31–36，2003.

[98] J. Yamagishi, M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi, “A context clustering technique for average voice models,” IEICE Transactions on Information and Systems, vol.E86-D, pp.534–542, 2003.

[99] K.P. Truong and J. Truvain, “On the acoustics of overlapping laughter in conversational speech,” Proceedings of Interspeech, pp.851–854, 2012.

発表論文

学協会誌論文

1. T. Nagata, H. Mori, and T. Nose, “Dimensional paralinguistic information control based on multiple-regression HSMM for spontaneous dialogue speech synthesis with robust parameter estimation,” Speech Communication, vol. 88, pp. 137–148, 2017.

国際会議論文

1. T. Nagata, H. Mori, and T. Nose, “Robust estimation of multiple-regression HMM parameters for dimension-based expressive dialogue speech synthesis,” Proc. Interspeech 2013, pp.1549–1553, 2013.

国内口頭発表

1. 永田 智洋, 森 大毅, 能勢 隆, “パラ言語情報を表現可能な対話音声合成のための重回帰 HSMM の検討,” 電子情報通信学会技術研究報告, SP2011-97, pp. 179–184, 2011.

2. 永田 智洋, 森 大毅, 能勢 隆, “重回帰 HSMM を用いたパラ言語情報を制御可能な対話音声合成の検討,” 日本音響学会 2012 年春季研究発表会講演論文集, pp. 435–436, 2012.

3. 永田 智洋, 森 大毅, 能勢 隆, “重回帰 HSMM に基づく音声合成における回帰行列の MAP 推定,” 日本音響学会 2013 年春季研究発表会講演論文集, pp. 493–494, 2013.

4. 永田 智洋, 森 大毅 “対話音声合成を目的とした発話中の笑い声の変動要因の検討,” 電子情報通信学会技術研究報告, SP2014-91, pp. 7–12, 2014.

5. 森 大毅, 高橋 俊介, 永田 智洋, “HMM に基づく対話音声合成におけるパラ言語情報制御手法の比較,” 電子情報通信学会技術研究報告, SP2014-90, pp. 1–6, 2014.

6. 永田 智洋, 森 大毅, “クラスタリングに基づく発話中における笑い声の変動要因の検討,” 日本音響学会 2015 年春季研究発表会講演論文集, pp. 403–404, 2015.

7. 永田 智洋, 森 大毅, “自然対話コーパスを用いた音声コミュニケーション場面における笑い声合成の検討,” 日本音響学会 2016 年春季研究発表会講演論文集, pp. 309–310, 2016.

8. 森 大毅, 有本 泰子, 永田 智洋, “複数の会話コーパスを対象とした笑い声イベントのアノテーション,” 日本音響学会 2017 年秋季研究発表会講演論文集, pp. 217–218.

9. 永田 智洋, 森 大毅, “自然対話における発話の文脈を考慮した笑い声合成の検討,” 電子情報通信学会技術研究報告, vol. 117, no. 368, SP2017-65, pp. 93–98, 2017.