対話音声合成のための
音声表現の多様化に関する研究

Studies on Diversification of Speech Expressions
for Conversational Speech Synthesis

2023年　2月

岩　田　　和　彦

Kazuhiko　IWATA

対話音声合成のための
音声表現の多様化に関する研究

Studies on Diversification of Speech Expressions
for Conversational Speech Synthesis

2023年　2月

早稲田大学大学院　基幹理工学研究科

岩　田　　和　彦

Kazuhiko　IWATA

目 次

図 目 次

表 目 次

第1章　序論

1.1　本研究の背景

人の声を機械を使って作り出そうという音声合成の試みの歴史は古く、その始まりは、18 世紀に制作された人の発声機構を機械仕掛けで模した装置にまでさかのぼる [1]。以来、現在に至るまでの間に、その実現の形態はアナログ部品で構成された電気回路 [2], [3] から、ディジタル・シグナル・プロセッサを搭載したハードウェア 9)、そしてコンピュータ・ソフトウェア 39) へと大きく変貌を遂げた。これに伴い、様々な方式や手法が誕生し、合成される音声の品質も大幅に向上した。人が通常の読み書きに用いるのと同じ形式のテキストを音声に変換するテキスト音声合成システム 8), 71) も開発され、音声合成技術の利用範囲は拡大した。現在、ロボットなどとの音声対話 5) をはじめとして、自動通訳 36) や福祉機器 43) のほか、様々な用途 42), 44) で利用されている。

こうした音声合成技術全般の開発の歴史の詳細は、数多（あまた）ある文献 78), [4]~[10] に譲るとして、ここではその中でも特に、本研究が取り組む対話音声の合成を巡る近年の動向について概観する。

1.1.1　モデル設計から学習データ設計へ

1990 年代になると、計算機の演算速度の高速化、メモリや記憶装置の大容量化が加速度的に進んだ。これにより、大規模な音声データを扱うことが可能となり、音声合成の研究にも大きな変化がもたらされた。

〔1〕 コーパスベース音声合成の登場

計算機の処理能力の劇的な向上も追い風となって、多種多様な音声データベースが次々と構築され、研究者向けに公開されるようになった [11]~[15]。この状況を

背景として登場したのが、コーパスベースの音声合成方式[16]~[21]である。大規模な音声データが利用できるようになると、音声認識で一足先に導入されていた時系列の統計モデルである隠れマルコフモデル（Hidden Markov Model; HMM）に基づく音声合成方式が開発された[22]。

それまでは、必ずしも大量とはいえない実例の注意深い観察によって獲得した知見に基づいて、音源－フィルタ理論[23]に立脚した音声波形の生成[24]~[27]と、統語構造などに着目したイントネーション（以下、「音調」と記す）やリズム、ポーズなどの韻律の制御[10), 11), [28]~[35]のそれぞれで、音声の生成過程に基づくモデル化や、モデルパラメータの制御規則の構築が進められていた。これに対してHMM音声合成方式は、波形生成と韻律制御を同じ枠組みを用いて同時にモデル化することを可能にした[36]。波形や音調、リズムなどの生成モデルを個別に設計し、制御パラメータを少数の実例に基づいて決定する従来のやり方から、モデルには汎用的な統計モデルを利用し、そのモデルパラメータを大量のデータから学習する方式へと、音声合成の研究の手法そのものを大きく転換させた。話者や発話スタイルなどの多様化も、容易に実現できるようになった[37]~[39]。

後に登場した深層学習（Deep Neural Network; DNN）の導入[40]によって、モデルはより精緻化され、合成される音声の明瞭性、自然性は飛躍的に向上した。今やDNNは、音声合成方式の主流となっている[41]~[48]。

〔２〕 大規模音声データ収集の重要性の増大

HMMやDNNに基づく音声合成方式は、モデルの学習に大量の音声データを必要とする。このため、合成音声のもととなる音声データ（以下、「音声セット」と呼ぶ）として、どのような音声を、どのようにして収集するかは、重要な課題である。

かつて、音声を収録するための発話文の一式（以下、「文セット」と呼ぶ）には、ATR音素バランス503文[11]（以下、「ATR503文」と記す）が、広く利用されてきた。新聞や雑誌の記事などの書き言葉による文章を集めた大量のテキストデータの中から、音素連鎖の出現頻度を考慮して抽出された文セットであり、ニュース原稿の読み上げなどの用途向けの音声合成の開発には有用である。しかし、話

し手の意図や態度を伝えるために対話の中では頻繁に用いられる文末詞[49]を伴った文は僅かしかなく、疑問文もないため、対話音声の合成に適しているとはいい難い。そのため、話し言葉による音声コーパスを構築するに当たり、一部の文の文末を対話調や対話調の疑問形に書き換えてから音声の収録に用いた、という事例が報告されている[50]。様々な感情や発話スタイルによる発話の収集にも用いられてはいるが[51],[52]、文の内容がスタイルにそぐわないものがあり、収録した音声の中には指示したスタイルに聞こえないものもあったと報告されている[51]。自然な表現を得る上では、所望の表現を表出するのにふさわしい発話内容の文セットを用意することも重要である。

明瞭で滑らかな音声を合成するには、調音結合によって隣接音素の影響を受けた音響特徴量の時間変化を忠実に再現する必要がある。コーパスベースの音声合成においても、音声セット中に音素連鎖の多彩なバリエーションがバランスよく含まれているかどうかは、合成音声の品質を左右する重要な鍵となる。更に、音素連鎖だけでなく、自然な韻律の生成に関わる要素のバリエーションも考慮した文セットの設計方法も提案されている[53],[54]。一方で、最近のDNN音声合成におけるモデル学習などでは、数十時間規模の音声データが利用されるようになっている[42],[45]。100時間を超える音声データを用いたコーパスベースの音声合成システム[55]も開発されている。ただ、含まれる音素連鎖などのバリエーションをあえて意識する必要がないほどの大規模な音声データは、誰にでも簡単に用意できるというものではない。そこで、オーディオブックなどの、既存の音声資源を活用しようという取り組みも進められている[56]~[59]。

統計モデルによる音声合成手法が主流となったことで、合成音声の品質向上の決め手、換言すれば、研究の課題の中心は、“どのような音声生成モデルを構築するか”から、“どのような学習データを用意するか”へと、移り変わってきている。

1.1.2　朗読口調から対話口調へ

音声対話は、人にとっては最も自然なコミュニケーションの手段である。そのため、古くから究極のヒューマン・マシン・インタフェースとして待望され、機械との間で声による情報の入出力を可能にするための技術[72),73)]として発展して

きた。近年、音声対話技術は、基礎研究から実用化の段階へと急速な進展を遂げた[60]。これにより、音声インタフェースを有するヒューマノイドロボット[61]や擬人化エージェント[62]をはじめ、実用的なサービスにも利用されるようになってきた[63],[64]。そして今や、ロボットを話し相手にして、対話を楽しむことができるようにまでなっている[65]~[67]。このとき、ロボットが、話の内容やその場の状況などに応じて表情豊かで多彩な口調[68]を使い分けて話し掛けてきたら、ロボットに対する親しみが増し、対話がより楽しいものになると期待される。音声対話が身近な技術となる中で、音声合成技術に求められる性能や機能も、大きく変化してきている。

〔１〕 音声表現がコミュニケーションにおいて果たす役割

人同士のコミュニケーションでは、言語情報だけでなく、話し手の意図や態度、感情などのパラ言語・非言語情報[69]~[71]も含めた情報交換が行われる。声の表情は、感情の表出以外にも、話し手と聞き手との関係[72]や、話題、場の雰囲気など、発話の背景にある様々な要因に応じて変化する。

本論文では、パラ言語・非言語情報が音声を媒体として伝達される際に韻律や声質†に現れる特徴を、「音声表現」と呼ぶことにする。人同士では、全てを言葉にしなくても、音声表現によって気持ちを伝えることができる。「そうですか」と同じ言語表現を使って返事をしたとしても、感心しているときと、落胆しているときとでは、音声表現は明らかに違ったものになる[70],[73]。その場合、聞き手は、音声表現から話し手の心的状態を知ることになる。「よかったね」と皮肉を込めた口調でいうなど、あえて言語表現と矛盾する音声表現を用いることで、より効果的な伝え方をすることもできる[74]。明日の予定を尋ねられて、行き先を伝えたいときは「朝から学校へ行くよ」、いつ家を出るかを伝えたいときは「朝から学校へ行くよ」と、重要な情報を音声表現を使って際立たせることもできる。どれが重要な語句かが聞き手に正しく伝わる基本周波数（以下、「F0」と記す）のパターンを合成音声の聴取実験によって明らかにした研究[1),7)]や、このときの音調をF0生成過程モデルに基づいてモデル化した研究[75],[76]なども報告されている。

†ここでは、発声の仕方の違いに起因する聞こえの違いを指し、音声器官の生理的・物理的違いによる話者の個人差に起因する聞こえの違いは含めない。

かつての合成音声の主な用途は、ニュースや天気予報、交通情報、株価情報などの原稿の読み上げであった[77],[78]。そこでは、情報が正しく伝わるよう正確に（テキスト解析技術）、聞き取りやすいよう明瞭に（音声波形生成技術）、そして、内容を理解しやすいよう自然な音調とリズムで（韻律制御技術）、読み上げられるようにすることが、技術開発の中心であった。しかし、対話音声には朗読音声とは異なる様々な特徴があり[79]~[82]、与えられたテキストを正確に、明瞭に読み上げるだけでは、自然な対話音声とはならない。ロボットの声がどれだけ聞き取りやすくても、ニュース原稿を読み上げるような口調で話し掛けられたのでは、ロボットへの親近感は湧かないであろう。

一方、音声対話の研究分野では、高度な対話制御の実現に向けて、人同士の豊かで効率的なコミュニケーションを支えている音声表現が積極的に利用されている。言葉では表されていない話し手の意図や態度を、音声表現に見られる特徴から把握する意図理解の研究[83]や、句末の音調を発話の継続と終了の判断に利用する研究[84],[85]などが進められている。

音声対話システムの応答音声を生成する音声合成システムとしては、より高度な音声対話の実現のために、コミュニケーションにおいて重要な役割を果たす多彩な音声表現を表出できる機能を備えておく必要がある。

〔２〕 対話特有の多様な音声表現の収集

音声対話システムとの対話がより親しみやすいものとなるようにすることを目指して、様々な視点から、対話らしい表情をもった音声を収集する取り組みが進められている。

従来の韻律制御は、単独で発話された文を単位としてモデルが構築されていた。しかし、ひとまとまりの文章を全体を通して読むのと、文章中の各文をばらばらにして単独で読むのとでは、韻律に大きな違いが生じる[40)]。対話システムが、ニュース記事などの情報を分かりやすく伝えるには、記事全体を通しての韻律制御が必要である。そこで、核となる重要な情報を印象付けられるように設計した原稿を用いて、ニュース記事の単位で発話させた音声が収録された[20)]。この音声を用いて、ニュースや話題を一方的に伝えるのではなく、利用者と対話しながら、メリ

ハリのある話し方で伝えることができる音声合成システムが開発されている[6)]。

通常、音声セットを構築するに当たっては、一人の話者が発話した音声が収録される。これに対して、“モノローグ”では得られない対話らしい表現を収集することを狙って、2名の声優に、対話の台本を目の前にいる相手と実際に掛け合いをしながら演じさせた音声を収録する、“非モノローグ”の音声収集手法が提案されている[86],[87]。また、複数の実験参加者に課題を与え、参加者が協力してそれに取り組んでいる間に発した音声を収録した、自発音声による対話音声データベースが構築されている[14],[88]。対話音声では頻繁に現れる笑い声やフィラーなどを含む音声も、積極的に収集されている[89]~[91]。

感情音声の合成は、表現の多様化に向けた取り組みの一つとして、多くの研究者によって長年研究されてきている[92]~[98]。感情認識などの目的も含め、様々な言語において、基本感情[99]を中心とした感情音声コーパスの整備が進行している[100]~[108]。

合成音声の表現力の多様化に向けて、多種多様な音声表現が収集されている。音声合成で感情や意図を表現する技術を開発する際の指針となる、“話し方”を分類したガイドライン[109]も策定されている。その一方で、基本感情は必ずしも日常のやり取りの中で実際に頻繁に表出されるものではない、との指摘もある[107]。どのような音声表現を提供することが、音声対話システムの応答を表情豊かなものにする上で効果的かについては、検討の余地がまだある。

1.1.3　情報伝達から意図表出へ

対話においては、単に情報を伝えるだけではなく、その情報に対する話し手の態度や判断なども併せて伝達される[110]~[112]。話し手の意思や態度を聞き手に正確に伝えることは極めて重要であり、それらが正しく伝わらなければコミュニケーションに齟齬(そご)を来すことになる。音声対話システムと利用者との間で高度なコミュニケーションが行われるようになってくれば、システムからの応答音声には、システムの意図を利用者に正しく伝えられる表現力が求められるようになるはずである。しかし、合成音声による意図表現に関する研究としては、間投詞「ん」の音調の形状と高さの組合せに着目した印象表現の研究[113]や、句境界の末尾にお

けるF0の特徴的な動きがもつ機能を明らかにしようとした研究[114]などが、幾つか見られる程度である。

〔1〕 文末音調がコミュニケーションにおいて果たす役割

日本語の話し言葉では、文末に現れる終助詞や助動詞などのいわゆる文末詞[49]が、話し手の態度や判断などのモダリティを表す役割を果たす。言語学の分野では、個々の文末詞の意味や用法に関する数多くの研究があり[115]~[119]、文末詞の音調と意図の表現との関係についても活発な議論がなされている[120]~[133]。

文末を上昇音調で発話した「どうして遅かったの↗」は通常の〈質問〉であるが、「どうして遅かったの↘」と下降音調で発話すると、遅かったことを〈非難〉したり、〈詰問〉したりする意図が伝わる[120]。つまり、同じ言語表現でも、文末の音調表現の違いに応じて、異なる意図を表現することができるわけである。更に、文末詞の音調が発話全体の意味を支配することを示す実験結果も、報告されている[123]。これらのことは、裏を返せば、合成音声にも適切な文末音調を付与しないと、誤ったメッセージが伝わってしまう危険性があることを意味している。

実際、文末詞の音調の違いで意味が変わる表現は、日本語を外国語として学ぶ学習者にとっては、コミュニケーション上の誤解を生む要因となっている[120]。しかし、母語の音調の干渉が習得を難しくしているため[134],[135]、日本語教育における音調教育の必要性が提起されている[136]。

これらのことからも、文末音調がコミュニケーションにおいて果たす役割の重要性をうかがい知ることができる。

〔2〕 言語学における文末音調の分類

言語学の分野で議論されている文末音調の分類には諸説あるが、定説とされているものはないといわれている。表1.1は、そのうちの幾つかの説を、文末音調の分類数ごとに整理したものである。

「上昇調」と「下降調」の2種類とする説には、前者が聞き手の反応を伺う意味を表し、後者が特に伺わない意味を表すという原則によって文末音調の諸現象は説明されるとするものがある[124]。また、前者は対話の構造の継続性を、後者

表 1.1　言語学における文末音調の分類例

分類数	文末音調の分類
2	・上昇調，下降調 [124], [125]
4	・昇調，降調，平調，降昇調 [126] ・降昇調，昇調，降調，昇降調 [127]
5	・平調，昇調 1，昇調 2，降調，@型類 [128] ・普通の上昇調，浮き上がり調，反問の上昇調，強めの上昇調，つり上げ調 [129] †1 ・基本音調，のぼり音調，くだりのぼり音調，くだり音調，ひくめ音調 [130] †2 ・疑問型上昇調，強調型上昇調，顕著な下降調，上昇下降調，平調 [131] ・平坦調，疑問上昇調，アクセント上昇調，下降調，上昇下降調 [132] †3
6	・疑問型上昇調，強調型上昇調，平坦調，急下降調，無音調，上昇下降調 [133]

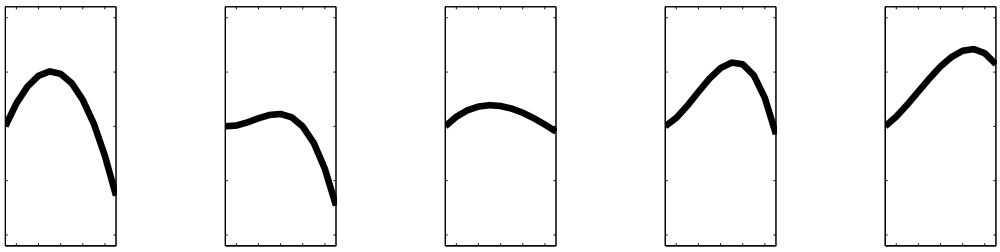
†1　上昇調を対象とした分類。

†2　声の強さの変化が最大の特徴であるとする「つよめ音調」も挙げられているが、本研究は F0 の形状による表現を対象としているため除いた。

†3　それぞれで「順接」、「低接」を想定すると、理論上 10 種類が可能としている。

は区切りを示すものと定義することで、終助詞“よ”と“ね”の機能と文末音調の相互作用から生じる様々な現象に説明が与えられるとする説もある [125]。更に細かく、4 種類 [126], [127]、あるいは 5 種類 [128]~[132] に分類した上での議論も行われている。

音調には、表 1.1 に示したように、F0 の動きに由来した、あるいは使われ方や機能に由来した名称が与えられている。しかし、それぞれの説の間には呼び方や機能の捉え方に違いも見られる。特殊な気分を表すための平坦な F0 形状を「平調」、平叙文の文末に見られる F0 の降下を「降調」と呼ぶ説 [126] がある。一方で、



(a)　(b)　(c)　(d)　(e)

図 1.1　「上昇下降調」とは？

機能的には中立で平叙文の文末に見られる F0 が緩やかに下降する音調のことを「平調」[128],[131]、あるいは「基本音調」[130] と呼び、この音調との対立として、特別な意図を伝えるために文末における F0 の自然降下と比べて際立って下降する音調を「降調」[128]、「顕著な下降調」[131]、あるいは「くだり音調」[130] と呼ぶ説もある。そこで、各説を整理した上で音調を 6 種類に分類し、名称も見直して、前者の説 [126] の平調を「平坦調」、後者の説 [128],[131] の平調を「無音調」とする提案がなされている [133]。また、「くだりのぼり音調」には、聞き手の行動などに対する話し手のあからさまな〈不信〉、あるいは〈不満〉の表明になる [130] とする説があるが、上昇する前に特に下降することが常に必要なわけではないとする説 [129]、「急下降調」と「疑問型上昇調」の臨時的複合と見なす説 [133] もある。ただ、いずれの説も、文末音調が話し手の意図を表現する上で重要な役割を担っていることを指摘している点では、一致している。

文末での F0 の動き（以下、「文末 F0 形状」と呼ぶ）の考察は、かつては研究者の内省に基づいて行われていた。ところが、音声の聴取実験によって、音声言語などの研究者の間でも“上昇”、“下降”、“平坦”などの聴覚的判断がばらつくことが指摘されていた [123]。現在は、ディジタル音声処理技術の普及の後押しもあって [137]、実音声から抽出した F0 の観測に基づいた研究が進められている。これにより、提案する分類のそれぞれの音調の典型的な F0 形状の実例が示されるようにはなってきた。

ただ、「上昇下降調」ひとつを取ってみても、図 1.1 に示すような様々な上昇の仕方、下降の仕方がありそうなことが想像できる。ところが、表 1.1 のそれぞれ

の「上昇下降調」が、どのようなF0形状を指しているのか、あるいは、想定しているのかは、必ずしも明確ではない。そもそも、文末音調としてどのようなF0形状のバリエーションが存在するのかが、大規模な音声データを対象とした観察に基づいて具体的に示されたことはない。図 1.1 に例を示したような形状の違いが意図表現における機能の違いを生むかどうかも、詳しく調べられてはいない。文末音調と発話意図とを統合したアノテーションを施した、音声コーパスの整備の必要性も提言されている [138]。文末音調の研究は、言語学の分野においても、まだ途上にあるといえる。

1.2　本研究の目的

これまで見てきたように、対話音声合成の技術は目覚ましい進歩を遂げている。しかし、音声対話システムの表情豊かな応答を実現する技術に高めるまでには、取り組むべき課題はまだ多い。

合成音声は、その明瞭性が格段に向上したことで、音声対話技術を利用した実用的なサービスやロボットの応答音声にも利用されるようになり、日常で耳にする機会が増えている。しかし、いくら発音が明瞭であっても、ニュース原稿を読み上げるような淡々とした口調や、話の内容に関わらずいつも同じ口調で話し掛けられたのでは、システムと対話している感覚にはなれないはずである。対話音声らしい表情豊かな合成音声を提供することは、喫緊の課題である。

また、音声対話を利用したシステムは、一方的に情報提供などのサービスを行うわけではなく、利用者と対話をしながらサービスを提供するものである。ロボットとのおしゃべりを楽しむことを提供するサービスでは、対話すること自体が目的となっている。人同士の対話では、言葉による表現だけではなく、音声表現も駆使した高度なコミュニケーションが行われている。このような、コミュニケーションを円滑に進める上で不可欠な表現力も、対話音声合成の機能の一つとして備えておく必要がある。

以上のような背景を踏まえ、音声対話システムの応答音声としての実用に供する合成音声を提供することで音声情報処理技術の実用化の推進に貢献することを目指して、合成音声における音声表現の多様化の研究に取り組んだ。

その一つとして、まず、発話の中で特に重要な語句を強調して伝えるための音声表現に着目した。強調したい語句が正確に聞き手に伝わるように際立たせるためには、発話文の中の語句のアクセント型の並びに応じた適切な音調のパターンを付与する必要があることを、合成音声の聴取実験を通して明らかにした。この聴取実験の結果に基づいて、重要な語句を強調して表現するための音調の制御規則を構築した[1), 7)]。

そして、次に、更なる音声表現の多様化に向け、新たに二つの具体的な目標を設定した。

【目標 1】　対話状況に応じた音声表現の多様化

対話の状況に応じて異なる音声表現を使い分けることができる機能を提供する。

【目標 2】　文末音調による意図表現の多様化

話し手の意図を確実に伝えることができる文末音調による表現手法を確立する。

本論文では、これら二つの目標の実現に向けた取り組みと、その研究成果について詳述する。

1.2.1　対話状況に応じた音声表現の多様化

生き生きとした対話を実現するために必要な表情豊かな合成音声の提供を目指して、次の二つの点に焦点を当て、異なる音声表現の使い分けを可能にする音声合成の枠組みの構築に取り組む。

〔1〕 どのような音声表現を提供するか

1.1.2 で述べたように、音声対話技術は、単なる声によるインタフェース技術というだけではなく、ロボットなどを相手に対話を楽しむことを可能にする技術にまでなってきている。ところが、ロボットがどのような場面でも常に同じ口調でしゃべっていたのでは、話は一向に弾まないだろう。

まずは、音声対話システムの応答を表情豊かなものにするためには、どのような音声表現を提供することが効果的かを検討する。人同士の対話を観察すれば、対話を取り巻く状況に応じて多彩な音声表現が用いられていることが分かる。そこで、対話の状況に応じた音声表現の違いに着目する。とはいえ、あらゆる対話の状況を網羅して、音声表現を収集することは現実的ではない。様々な対話の中で利用できて、少数でも効率よく広範囲をカバーできるような幾つかの対話状況に絞り込むことを検討し、提供すべき音声表現を選定する[12), 46)]。

〔2〕 どのような収集方法によって異なる表現の調和を保つか

音声対話システムなどのアプリケーションでは、合成音声は対話のコンテキストの中で使用される。感情などを表現した合成音声の品質の評価としては、一文ごとにどの表現に聞こえるかを判定させるだけでは十分ではなく、それらがコンテキストの中で表出されたときにも自然であるかを確認する必要があることが、指摘されている[139]。本論文では、複数の異なる音声表現を対話のコンテキストの中で使い分けたときの表現全体を通した自然性を、個々の表現の自然性と区別するために、「調和性」と呼ぶことにする。

対話の状況に応じて異なる音声表現を使い分けることができる機能を作り込む際には、それぞれの表現をどのような対話のコンテキストの中で表出させても違和感が生じないようにすることにも、留意しなければならない。そこで、一つひとつの発話の明瞭性や自然性だけでなく、これまでは注意が払われることがなかった、複数の音声表現全体の調和性を確保することにも力点を置く。どのようにすれば全体の調和を保ちながら異なる表現表現を収集できるかを検討し、音声セットの新たな収集手法を構築する[4)]。

1.2.2　文末音調による意図表現の多様化

コミュニケーションを円滑に進めるために必要な表現力の提供を目指して、次の二つの点を明らかにし、文末音調によって発話意図を伝える表現手法の確立に取り組む。

〔１〕 どのような文末音調のバリエーションがあるか

1.1.3 で述べたように、話し手が意図を表現する際には、文末音調が重要な役割を果たす。長年、言語学の分野で議論されてきているが、上昇や下降といった音調の違いは、かつては研究者の内省によって判断されていた。文末音調としてどのようなF0形状のバリエーションが存在するのかが、大量の音声データに基づいて具体的に示されたこともなかった。

これまで「上昇調」や「上昇下降調」などのような名称を与えられてきた文末音調を合成音声で実現するためには、まずは、それらの具体的なF0形状を知る必要がある。そこで、大規模な対話調の音声データを対象として、文末F0形状を収集することからはじめる。それらを分類し、整理することで、対話音声の文末F0形状にどのようなバリエーションがあるかを明らかにする。更に、得られたF0形状を、合成音声に様々な文末音調を付与するためのテンプレートとして利用する方法を検討する。

〔２〕 どのような文末音調が意図表現に有効か

音声合成システムで使用する記号を定めた規格 [140] では、文末音調を指定する韻律記号として、「通常」と「疑問」の２種類が規定されている。しかし、同じ言語表現であっても、聞き手は、文末の音調に応じて違う意図を感じ取るものである [13), 45)]。文末音調がコミュニケーションにおいて重要な役割を果たすことを考えれば、音声対話システムの応答音声として利用する合成音声においては、文末音調の緻密な制御は必須である。

そこで、音声データから取得した様々な文末F0形状のテンプレートを利用して文末音調を付与した合成音声を作成し、それらの聴取実験を通して、文末音調の違いに応じてどのような発話意図が聞き手に伝わるかを明らかにする。文末詞にも意図を伝える機能があることから、両者の関係に着目し、文末詞と文末音調の組合せによる意図表現の手法を確立することを目指す [2), 14), 47), 48)]。更に、これを発展させて、主たる意図に付加される言外の意図ともいえる微妙なニュアンスの違いも、文末音調の違いによって表現可能かを探る [3), 15), 49)]。

1.3　本論文の構成

本論文では、次章以降で以下のような具体的な議論を進める。

第 2 章〜第 3 章では、【目標 1】の実現に向けて、提供すべき音声表現の選定と、それらの収集手法について論じる。

第 2 章では、どのような対話の状況を想定した音声表現を提供するかを、コミュニケーションロボット [65] が行う対話タスクを手掛かりに選定する。それぞれの対話の状況にふさわしい音声表現ができるだけ自然な形で得られるよう、対話タスクにおけるロボットの発話内容を参考にして、音声収録のための文セットを設計する。この文セットを用いて収録した音声セットに基づいて状況ごとの音声合成モデルを作成し、複数の音声表現を使い分けることができる枠組みを構築する。選定した複数の対話状況ごとの音声表現が、それらを対話の流れの中で使い分けることで対話らしさを向上させる効果をもたらすかを、模擬対話を用いた合成音声の主観評価実験によって検証する。

第 3 章では、第 2 章で収集した音声セットにおいて二つの課題が浮かび上がったことを踏まえ、多様な音声表現の収集手法の改善に取り組む。一点めは、状況ごとの合成音声の品質の、改善と均質化である。第 2 章での文セットの設計では、異なる音声表現を表出するのにふさわしい内容の発話文となるようにすることを重視し、合成音声の明瞭性や自然性に影響を及ぼす音素連鎖などの要素のバリエーションを確保することを考慮しなかった。このことは更に、状況ごとの合成音声の品質のばらつきも招いた。所望の音声表現を得るのにふさわしい発話内容とすることは元より、明瞭性や自然性に影響を及ぼす要素の多様なバリエーションを確保した上で、状況ごとのばらつきをできるだけ抑えた文セットとなるような設計方法を検討する。二点めは、異なる音声表現全体の調和性の確保である。第 2 章の音声セットでは、状況ごとの音声に、別人の声のようにも聞こえるほどの極端な声質の違いが生じてしまっていた。その原因は、複数の異なる音声表現を収集する際に、それぞれが適切な表現となるようにすることに重点を置き、異なる表現間の調和を保つことを考慮しなかったことにあると考えた。そこで、音声セットの収集手法の根本的な見直しに取り組む。複数の異なる音声表現を互いの調和

を保った状態で収集できるようにするための方策として、それぞれの状況が順に満遍なく現れるような対話のシナリオを導入する。その上で、新たな設計による文セットと音声収集手法とを用いて実際に音声セットを収集し、第 2 章で収集した音声セットと新たな音声セットのそれぞれを学習した音声合成モデルを作成する。文セットの構成、音声セットの音響特徴量、及び合成音声の品質の、三つの側面から両者を比較して、多様な音声表現の新たな収集手法の有効性を検証する。

第 4 章～第 7 章では、【目標 2】の実現に向けて、文末詞とその音調の組合せと、それによって聞き手に伝わる話し手の意図との関係を明らかにする。

そのための準備として、第 4 章～第 5 章では、大規模な対話調の音声データに基づいて、どのような文末音調のバリエーションが存在するかを明らかにする。第 4 章では、大規模な音声データから文末 F0 形状を抽出し、対話音声で用いられている文末音調の実例を収集する。それらを合成音声に様々な文末音調を付与するための F0 形状のテンプレートとして利用できるようにするために、抽出した F0 形状の時間軸と周波数軸を正規化する方法を検討する。更に、収集したそれらの実例をクラスタリング手法を用いて階層的に分類し、文末 F0 形状にどのようなバリエーションが存在するかを整理する。

第 5 章では、階層的クラスタリングによって得られた多数のクラスタのセントロイドの中から、文末音調を生成するためのテンプレートとして利用する F0 形状を選定するための絞り込みの方法を検討する。音調としての差異が顕著なテンプレートを選び出すために、聴感上の判定を基準として導入する。すなわち、クラスタのセントロイドを音調として付与した音声を聴き比べて異なる音調と感じられるかどうかを確認する聴取実験を行い、その結果を、テンプレートを絞り込む際の判断基準として利用する方法を検討する。

続いて、第 6 章～第 7 章では、文末音調によってどのような意図を伝えることができるかを、テンプレートによる様々な F0 形状の文末音調を付与した合成音声の聴取実験を通して明らかにする。第 6 章では、特定の意図を聞き手に確実に伝えることができる、文末詞とその音調の組合せを明らかにする。文末での言語表現として用いられる文末詞も、話し手の意図を表現する役割を担っている。その

ため、文末詞とその音調の組合せごとの機能を明らかにする必要がある。聴感上で音調としての差異が顕著なテンプレートを幾つか選び出し、それらによる文末音調を付与した合成音声を作成して、発話意図の聴取実験を行う。その結果から、文末詞とその音調の組合せと聞き手に伝わる意図との関係を詳細に考察し、音声合成で利用可能な知見となるよう整理する。得られた知見に基づき、文末音調によって話し手の意図を伝える表現手法を構築する。

第 7 章では、第 6 章で検討した文末音調による意図表現を更に発展させ、主たる意図に付加される言外の意図ともいえるニュアンスの違いも、文末音調の違いによって伝わるかを確認する。第 6 章の実験には使用しなかった様々なテンプレートで文末音調を付与した合成音声を聴取してみると、話し手の心情や話し手と聞き手の関係などが反映されているような、微妙なニュアンスの違いが感じられるものが幾つもあった。そこで、使用するテンプレートの種類を増やし、文末音調の違いに応じて聞き手は異なるニュアンスを感じ取るかを、合成音声の聴取実験によって詳しく調べる。得られた結果から、ニュアンスの違いをもたらす文末 F0 形状の特徴を整理し、文末音調がもつ表現力の更なる可能性を探る。

最後に、第 8 章で、本研究の成果のまとめと今後の課題について述べる。

第2章　対話状況に応じた音声表現の違いに着目した表現の多様化

2.1　まえがき

音声対話技術の進展[60]により、音声対話を利用したサービスや製品が一般にも広く浸透してきている。しかし、システムからの応答がどのような場面でも常に同じ口調では、“システムと対話している”感覚は湧かないはずである。

では、音声対話システムの応答を表情豊かにするためには、どのような音声表現を提供することが有効だろうか。感情音声の合成[92]~[98]は、その問いに対する答えの一つである。しかし、基本感情[99]は日常では頻繁に表出されるものではない、との指摘もある[107]。音声対話システムで利用される実用的な技術の開発を目指すのであれば、“怒り”をあらわにする声や、“悲しみ”に打ちひしがれる声が、音声対話を使ったサービスの中でどの程度必要とされているかは、考慮に入れるべきであろう。

人同士の対話を観察すれば、対話を取り巻く様々な状況に応じて多彩な音声表現が用いられている様子が見て取れる。話し手と聞き手の関係やそれぞれの人物像、その発話がなされる背景などの状況設定を詳細に指示して演技させることで、多彩な音声表現を収集する手法も提案されている[141]。そこで、このような対話の状況に応じた音声表現の違いに着目する。とはいえ、音声セットを構築する上では、あらゆる状況を網羅的に想定してそれぞれにふさわしい音声表現を収集するということは、現実的ではない。様々な対話の中で利用できて、少ない種類でも効果的に対話らしさを向上させられると期待できる、幾つかの実現可能な数の状況に絞り込む必要がある。どのような対話状況における音声表現を提供するかを具体的に選定し、実際に複数の対話状況における音声表現を収集して、状況ごとの音声合成モデルを作成する。その上で、選定した音声表現を対話の流れの中

で使い分けることが対話らしさを向上させる効果をもたらすかを、模擬対話を用いた合成音声の主観評価実験によって検証する。

2.2　対話状況に応じた音声表現の収集

2.2.1　基本方針

異なる音声表現を収集するための複数の対話状況を選定するに当たり、音声対話の具体的な活用例の一つである、高齢者施設でのレクリエーションに参加して人同士のコミュニケーションを活性化させる役割を担うロボット [65] の対話タスクを考えることにした。

少数でも効率よく広範囲をカバーできるような状況を選定する手掛かりとして、感情の研究における円環モデル [142] を参考にした。円環モデルでは、図 2.1 に示す「快－不快」と「覚醒－睡眠」の 2 軸で表された平面上に、様々な感情が配置される。そこで、原点と各象限に対応付けられる状況 0〜状況 IV の 5 種類を、代表的な対話の状況として用意することを考えた。すなわち、状況 I には、話し手の心的状態なども含めた対話を取り巻く状況が「快」で「覚醒」の状態にあるような場面を、状況 III には「不快」で「睡眠」の心的状態にあると考えられる場面を選定することにした。状況 0 は、心的状態としては「平静」に相当する中立的な状況である。

2.2.2　文セットの設計と音声収録

前述のロボットが行う対話タスクには、高齢者施設において難読漢字の読み方を当てるゲームに参加して場を盛り上げる「難読タスク」と、施設を訪れた高齢者を出迎えて健康状態などを尋ねる「挨拶タスク」がある。各タスクでのロボットが果たす役割から考えて、難読タスクから状況 I と状況 II、挨拶タスクから状況 III と状況 IV に当てはまる対話の状況を選び出した。その上で、それぞれの状況でのロボットのセリフとして想定される発話文を、500 文程度ずつ創作した。状況 0 の音声の収録には、音素バランスが考慮されている ATR503 文 [11] を使用した。各状況の設定や発話文の例などを、表 2.1 に示す。難読タスクから選定した状

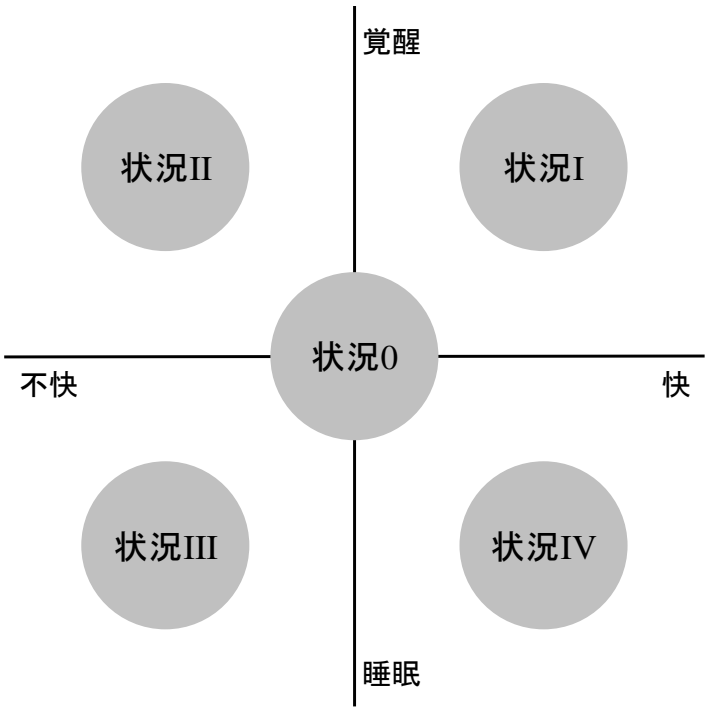


図 2.1　感情の円環モデルと 5 種類の対話状況との関係

況 I と状況 II では、音素連鎖のバリエーションを増やす目的で、発話文中の難読ゲームの解答に相当する単語（表中の『　』内）の部分に ATR 音素バランス 216 語 [11] を埋め込んだ。この点以外では音素バランスに関する考慮はしておらず、状況にふさわしい発話内容とすることに注力した。

女性声優 1 名に、状況ごとの文セットを、それぞれの状況の違いが明確になるように演じ分けさせた。一方、状況ごとの表現は一定に保つことができるように、一つの状況の音声を全て収録し終えてから、次の状況の音声を収録するようにした。収録したそれぞれの音声について、STRAIGHT[143] を用いて F0 とスペクトルを抽出し、波形とスペクトログラムの目視によって音素境界を決定して、モデルの学習に用いる音声データを作成した。

2.2.3　対話状況ごとの音声合成モデルの作成

状況ごとの音声合成モデルの作成には、HMM 音声合成ツールキット [144] を利用した。当初、状況 0〜状況 IV のそれぞれの音声データのみを用いて状況ごとの

表 2.1　5 種類の対話状況の設定と発話文

状況	タスク	対話状況の設定	発話文の例	文数
0	(ATR503 文)	平静	あらゆる現実を、すべて自分のほうへねじ曲げたのだ。 一週間ばかり、ニューヨークを取材した。 テレビゲームやパソコンで、ゲームをして遊ぶ。	503
I	難読タスク	場を盛り上げようとする	わかった、『勢い』だ！ これで、『具合』って読むなんて、不思議だね！ これって、『台所』っていう意味なんだって！	558
II	難読タスク	少し不機嫌に振る舞う	わかんないよ、『勢い』だなんて。 これで、『具合』って読むなんて、変だよ！ これって、ホントに『台所』っていう意味なのかなぁ。	525
III	挨拶タスク	相手の不調などを心配する	鈴木さん、頭痛いの？ 田中さん、鼻痛いんだって？ 伊藤さんは虫歯痛いみたいだよ。	520
IV	挨拶タスク	相手の好調などに安堵する	鈴木さん、頭痛いの治ってよかったね。 田中さん、鼻痛くなくなったんだね。 伊藤さんは虫歯痛いの治ったみたいだよ。	508

モデルを学習させたが、状況I〜状況IVでは、十分な品質の合成音声が得られなかった。文セットを創作するに当たって、音素連鎖などのバリエーションを十分に確保することを考慮しなかったことが原因と考えられる。そこで、状況I〜状況IVの状況のモデルの学習には、音素バランスが考慮されている状況0の音声データを併用することで、品質の向上を図ることにした。その結果、それぞれのモデルから生成される合成音声の品質には、改善が見られた。学習に用いた状況I〜状況IVの原音声がもつ話し方や声の特徴が、それぞれでよく再現されていることも確認できた。

2.3　対話状況に応じた音声表現の有効性の検証

感情などの表現の合成における主観評価実験では、文単位の合成音声を用いて、表現ごとにその再現性や自然性が評価されている[92]~[98]。しかし、合成音声を音声対話システムなどのアプリケーションで利用するのであれば、対話のコンテキストの中で表出されたときの自然さも評価すべきである[139]。

ロボットが行う対話タスクを参考にして選定した少数の代表的な対話状況の音声表現が、様々な種類の対話の中で幅広く利用できるものとなっているかを、模擬対話を用いた合成音声の主観評価実験で確認する。対話の流れの中で状況ごとのモデルを使い分けたときの自然さや対話らしさを総合的に評価し、選定した音声表現の有効性を検証する。

2.3.1　実験方法

ロボットが人と対話する具体的な場面を想定して、3種類の模擬対話を創作した。

(D1) 難読ゲームに参加している人とロボットとの対話

(D2) 初対面の人とロボットとの対話

(D3) 道順を尋ねる人と案内するロボットとの対話

(D1)は難読タスク、(D2)は挨拶タスクの一場面を想定したものだが、音声の収録に用いた文セットには含まれていない発話文で構成した。(D3)は、難読タスク、

表 2.2　道順を尋ねる人と案内するロボットとの対話

状況	発話文
	〔 すみません。〕
I	はい、こんにちは、僕に何かご用ですか？
	〔 早稲田大学への行き方を教えていただきたいんですが。〕
II	えー、困ったな。
II	僕も知らないんですが。
	〔 小林研究室に行きたいんです。〕
I	ああ何だ、小林研究室か。
IV	それなら知ってますよ。
III	ただ、ちょっと分かりにくいんだけど大丈夫かな。
	〔 そうなんですか。〕
III	うーんと、どの行き方が良いのかな。
I	そうだ、あの道だったら分かりやすいかもしれないな。
IV	えーと、まずこの道をまっすぐ行って…。

〔　〕は道順を尋ねる人の発話（録音音声を使用）

挨拶タスクとは異なるタスクである。それぞれについて、以下の二通りの音声刺激を作成した。

従来方式　発話の内容に関わらず常に同じ話し方をする従来の音声合成システムを想定して、ロボットの全ての発話を状況 0 のモデルで合成。

提案方式　発話文ごとに、発話の内容に応じて状況 I〜状況 IV の中からふさわしいと思われる状況のモデルを選んで合成。

表 2.2 に、(D3) の対話の内容と、提案方式においてロボットの各発話にどの状況のモデルを用いたかを示す。人の発話には、録音した音声を使用した。(D1)、(D2) についても同様に、提案方式において、状況 I〜状況 IV によるロボットの発話がほぼ同数ずつ現れるような構成にした。

評価者 16 名に、二つの音声刺激におけるロボットの発話を比較して、どちらがより対話らしいかを 5 段階で判定させた。このとき、提示順序の影響を考慮して、

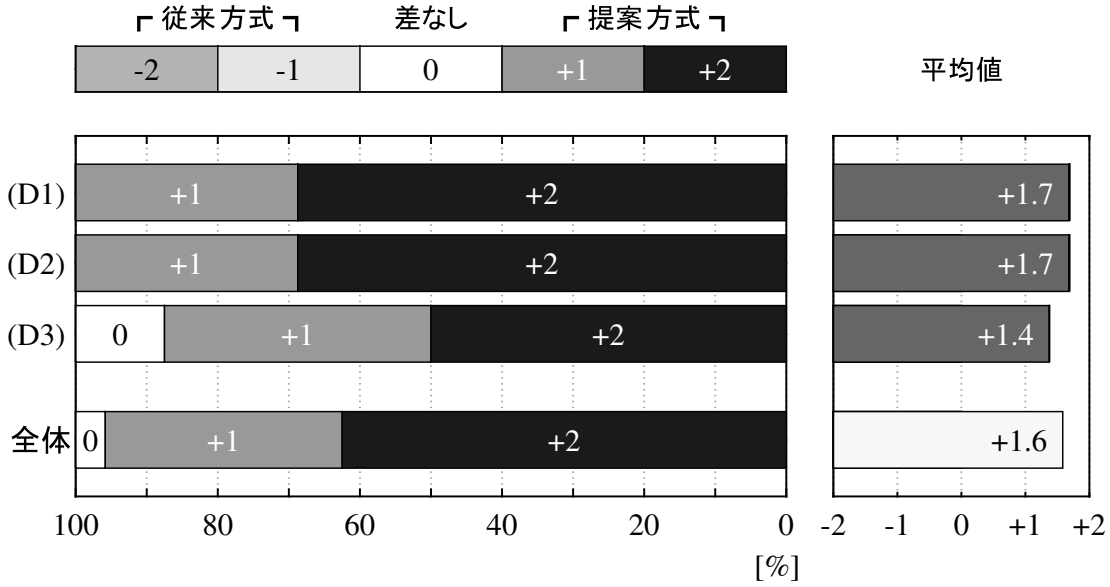


図 2.2　模擬対話の対比較評価結果

評価者を 8 名ずつの 2 グループに分け、一方のグループには従来方式、提案方式の順で、もう一方のグループにはその逆順で、それぞれの音声刺激を提示した。評価者には、二つの音声刺激がどのように異なるかや、提示順序に関する説明はしていない。

2.3.2　結果

各評価者の判定に対して、より対話らしいとされた音声刺激が、提案方式であった判定には +1、+2 を、従来方式であった判定には −1、−2 を、差はないとされた判定には 0 を、評点として付与した。従来方式と提案方式の提示順序を変えた二つのグループの各模擬対話の評点の平均値の間に、有意な差は認められなかった（両側 t 検定、有意水準 0.1%）。いずれの模擬対話においても、提示順序の影響はなかったといえる。

そこで、二つのグループの結果を統合し、それぞれの評点が付与された判定の割合と評点の平均値を求めた。結果を、図 2.2 に示す。(D3) のロボットのそれぞれの発話内容は、表 2.1 に示した音声収録時の状況設定に必ずしも合致しているとはいえない。難読タスクを想定した (D1) では、状況 III と状況 IV の音声表現を

割り当てた発話については、状況 III と状況 IV の音声収録時の状況設定と合致しているとはいえない。挨拶タスクを想定した対話の (D2) での、状況 I と状況 II の音声表現を割り当てた発話についても、同様のことがいえる。しかし、いずれの模擬対話においても、従来方式の方が対話らしいとされた判定（−1、−2）は一つもなく、提案方式が高い評価を得た。

以上の結果によって、選定した 4 種類の対話の状況に応じた音声表現は、収録用の文セットの設計時に設定した対話状況以外の場面での発話にも広く適用でき、これらを対話の流れの中で使い分けることで、対話らしさを向上させることができることが示された。

2.4　対話状況に応じた音声表現の収集手法の課題

模擬対話を用いた主観評価実験で、選定した対話状況ごとの音声表現を学習させたモデルを使い分けることの有効性を確認することができた。その一方、合成音声の品質上の課題も指摘された。

〈課題 1〉 発音の明瞭性や韻律の自然性が低くなることがあり、品質が安定していない。

〈課題 2〉 状況ごとに品質に差がある上に、それぞれの音声表現の違いが極端で別人の声のようにも聞こえるため、違和感がある。

〈課題 1〉からは、調音結合や韻律を再現するために必要な、音素連鎖や韻律構造のバリエーションが不足していることが考えられる。また〈課題 2〉は、状況ごとの文セットに含まれているそれらのバリエーションに、ばらつきがあることをうかがわせる。いずれも、発話文を創作する際に、合成音声の品質に関わる要素のバリエーションを確保することを考慮をしなかったことが原因であり、文セットの設計方法を見直す必要がある。

〈課題 2〉からは更に、複数の音声表現の収集方法の問題点が浮かび上がった。すなわち、個々の表現が適切になるようにすることに注力し、異なる表現同士の関係を考慮せずに収集したことで、それぞれの表現の乖離（かい）を招いたと考えられる。

このことは、多様な音声表現を収集する際には、個々の表現が自然であるかだけではなく、複数の表現が全体としての調和を保てているかにも注意を払う必要があることを示唆している。これは、対話の流れの中でそれぞれの表現を使い分ける評価を通して見つかった、新たな課題である。これまで見過ごされてきた重要な視点であり、音声の収録方法についても、根本的な見直しが必要であることが明らかとなった。

2.5　むすび

音声対話システムの応答を表情豊かにするためにはどのような音声表現を提供することが効果的かを検討し、対話の状況などに応じて異なる音声表現を使い分けることができる機能を提供する音声合成の枠組みを設計した。

2 次元平面で表される感情のモデルを手掛かりとして、4 種類の対話状況を選定し、それぞれの音声表現を表出するのに適した発話内容の文セットを創作した。2 次元平面の原点に相当する状況の文セットには、音素バランスが考慮されている ATR503 文を使用した。

5 種類の音声表現を収集して対話状況ごとの音声合成モデルを作成し、模擬対話を用いた合成音声の主観評価実験によって、対話の流れに応じてこれらのモデルを使い分けることで対話らしさが向上することを確認した。更に、それぞれの音声表現を収集するために設定した対話の状況とは必ずしも一致していない対話の場面であっても、これらの表現を使い分けることには対話らしさを向上させる効果があることも分かった。選定した音声表現は、様々な対話の場面で汎用的に利用できることが示された。

一方で、合成音声の明瞭性、自然性に加え、複数の音声表現全体の調和性に関する課題があることも浮き彫りとなった。このことによって、異なる表現を対話のコンテキストの中で使い分けたときの自然性を確認することの重要性 [139] が、改めて示されたといえる。

第 3 章では、これらの課題の改善に向けて、多様な音声表現の新たな収集手法の設計に取り組む。

第3章　多様な表現全体の調和を保つ音声収集手法の設計

3.1　まえがき

感情などの音声合成の研究では、DNN などの機械学習の性能の検証を目的として表現の再現性や可制御性を確認するために、収録した音声やそれらを用いて合成した音声が“意図したとおりの表現に聞こえるか”が評価されている[51],[92]~[98]。しかし、それらの表現を音声対話システムなどのアプリケーションで利用するのであれば、対話のコンテキストの中で表出されたときの自然さの確認は不可欠である[139]。音声対話システムの研究者の間では、対話の中で異なる音声表現を使い分けることの効果が議論されている[145]~[147]。ただ、合成音声の生成には既存の音声合成システムが使用されているため、合成音声そのものの自然性や表現の適不適は、評価の対象にはなっていない。

第 2 章では、複数の異なる音声表現の合成音声を使い分けることができる枠組みを構築し、音声対話システムが対話の流れに応じてその場にふさわしい音声表現を選択して使い分けることで、対話らしさを向上させることができることを示した。一方で、収集したそれぞれの音声表現の違いが極端で、別人の声のようにも聞こえるため、使い分けた際に違和感が生じるという課題が浮かび上がった。そこで、明瞭性や自然性の確保は元より、これまで注意が払われてこなかった、複数の音声表現全体の調和を保つことにも力点を置いた音声収集手法を設計し、新たな音声セットを収集する。

以下では、第 2 章で構築した文セットと音声セットを総称して「初期セット」、本章で構築する文セットと音声セットを「改良セット」と呼ぶ。様々な角度から両者を比較することで、新たな音声収集手法の有効性を検証する。

3.2　音声収集手法の設計

3.2.1　基本方針

多様な音声表現を互いに他の表現とは無関係に表出させると、表現の乖離を招き兼ねない（**2.4**〈課題 2〉）。複数の音声表現を、全体の調和を保ちながら収集するための方策として、話し手の心的状態が次々と変化して、収集したい音声表現が満遍なく出現するように進行する対話シナリオ（以下、「スキット」と呼ぶ）を導入することを考えた。これまでのように、それぞれの音声表現を個別に表出させるのではなく、対話の流れの中で表出させることで、全体としての調和が保たれた表現が得られるものと期待される。音声を収録するための文セットは、既存の大規模なテキストコーパスから抽出されることが多い [11], [53], [54], [148]。しかし、上述した要件を満たすようなスキットを既存のコーパスから探し出すことは難しいと考え、発話文は全て創作することにした。

高品質な合成音声を得るには、明瞭性や自然性に影響を及ぼす要素のバリエーションを確保することが不可欠である（**2.4**〈課題 1〉）。2 音素連鎖や 3 音素連鎖における音素の組合せは明瞭で滑らかな発声の再現に、文の長さや係り受け構造、対話特有の文末表現などは自然な韻律の再現に、それぞれ重要な役割を果たす。そこで、これらの要素のバリエーションを増やすことにも注力する。しかし、スキットの発話文だけで、それぞれの要素のバリエーションを十分に確保することは容易ではないと判断し、スキット以外の発話文も併用することにした。大規模コーパスから文セットを抽出する場合には、エントロピーを利用して音素連鎖の種類などの要素の出現頻度のバランスを考慮する手法がしばしば用いられる [11], [54], [148]。これに対して、発話文を創作する場合は、それとは異なるアプローチによって、それぞれの要素のバリエーションを確保することを考えなければならない。そこで、必要とする音素連鎖を含む単語を選び出し、それらを使った発話文を創作していくことで、音素連鎖や文の長さ、文末表現などのバリエーションを確保する。

3.2.2　文セットの作成方法の設計

基本方針を踏まえ、文セットを作成する際に考慮する要件を定めた。

（1）明瞭で滑らかな合成音声を得るために、音素連鎖をより細かく分類することを考え、音素の定義を以下のように拡張する。これらは、それ自体が音素連鎖であるものや、同じ音素の異音であるものではあるが、説明の便宜上、拡張したものも含めて「音素」と称する。下記の説明文中で、《　》は注目している音声区間を、〔　〕は母音が無声化した音節を表す。

(a) 長母音（例：「ホイール／ほ《いー》る」）は、同一母音の連続（例：「退院／た《いい》ん」）と区別する[149]。そのために、短母音とは別の音素として扱う。

(b) 母音が無声化した音節（例：「宿題／《〔しゅ〕》くだい」）や促音とそれに後続する子音（例：「一瞬／い《っしゅ》ん」の促音＋子音部分）、促音とそれに後続する母音が無声化した音節（例：「合宿／が《っ〔しゅ〕》く」）のように、音響的に不可分な音声区間は、複数のセグメントに分割せず一つの音素として扱う。

(c) 撥音は、後続する音素の調音位置に応じて異なる音で実現される[149]。よって、後続音素の調音位置が異なる撥音（例：「心配／し《ん》ぱい」、「心頭／し《ん》とー」、「進化／し《ん》か」、「できませ《ん》」）は、それぞれ別の音素として扱う。

（2）対話状況の枠組みとしては、図 2.1 に示した 5 種類を踏襲する。ただし、それぞれの状況の具体的な設定はロボットの対話タスク（表 2.1）には限定せず、様々な対話の場面を想定する。初期セットでは状況 0 の収集に ATR503 文を利用したが、ここでは状況 0 についても文セットを創作する。

（3）対話の中では、文末表現は、話し手の意図などを伝える重要な役割を担っている。「友達口調」（例：「～さ」、「～だよね」、「～だよ」）や、女性専用とされる終助詞[115]を伴った「女性口調」（例：「～わね」、「～のよ」、「～わ」）などの多様な口調を含める。また、文末表現のバリエーションとして、疑問文も含める。

（4）対話の流れの中で話し手の心的状態が次々と変化して、状況 0～状況 IV の

各状況が満遍なく表出されるように進行するスキットを創作する。

（5）発話内容からそれぞれの状況を容易に想像でき、状況にふさわしい音声表現が自然に表出されるような文（以下、「典型表現」と呼ぶ）を創作する。

（6）スキットや典型表現の他に、友人同士の雑談（旅行の相談や趣味の話題など)、カーナビゲーションシステムと利用者との対話、医師と患者とのやり取りなどの様々な対話の場面を想定し、それぞれで状況0～状況IVにふさわしい状況設定をした上で、状況ごとの発話文を創作する。

（7）状況ごとの文セット間で音素連鎖などのバリエーションに極端な差異が生じないようにするために、全ての状況で共通に用いる発話文も創作する。

3.2.3　音声の収録手順の設計

異なる音声表現全体の調和を保てるよう、文セットの構成と合わせて、以下のような手順で音声を収録することを考えた。

（1）収録時に用いる原稿には、設計した通りの音素連鎖が得られるよう、読み仮名を付しておく（例:「日本(にっぽん)」、「日本(にほん)」)。長母音や母音が無声化した音節なども一つの音素として扱い、それらを含めて音素連鎖のバリエーションを考慮しているため、読み仮名中に記号を用いて明示しておく（例:「映画(えーが)」／長母音、「季節(き)(せつ)」／無声化)。

（2）発話者には、各収録日の収録開始前に、典型表現を用いてそれぞれの表現が極端にならないように注意しながら、望ましい表現になるまで練習させる。2日目以降は、前日までに収録した音声を聴取させてから、練習を行う。

（3）各収録日の最初は、5種類の状況の音声表現を自然に発話し分ける感覚がつかめるよう、スキットや典型表現の収録から始める。

（4）スキットの収録では、収録対象外の登場人物の発話文は黙読させ、対話の流れに従って変化する話し手の心的状態を追いながら発話させる。

（5）スキットや典型表現以外の発話文の収録では、全体としての調和が保たれた表現が維持されるよう、文セットを 30 文程度を目安に適度に分割して、5 種類の状況について少しずつ順に収録していく。

3.3　文セットの作成と音声表現の収集

3.2 で述べた設計に基づき、実際に文セットを創作し、5 種類の対話の状況ごとの音声表現を収集した†。

まず、状況ごとの文セットの規模は、それぞれ 700 文とすることにした。その上で、10 種類のスキットと、状況ごとに 25 文ずつの典型表現を含めた、計 3,500 文からなる文セットを創作した。スキットの一例を表 3.1 に、典型表現の幾つかの例を表 3.2 に、それぞれ示す。スキットの創作では、対話の流れの中で、5 種類の対話の状況が次々と満遍なく、ほぼ同じ回数ずつ出現するよう工夫した。典型表現を含むスキット以外の発話文についても、状況ごとに話し手の心的状態を想定して、それぞれの状況にふさわしい表現が自然に表出されるような内容となるよう工夫した。

ある程度の数の文ができたところで各文を解析し、含まれている音素、2 音素連鎖、及び 3 音素連鎖、更に、文の長さのバリエーションの出現数をそれぞれ数え上げた。文の長さは、モーラ数とアクセント句数の 2 通りで算出した。まだ出現していない、あるいは出現数が少ない音素連鎖を洗い出し、それらを含む適当数の単語を選び出して、単語候補リストを作成した。音素連鎖のバリエーションを効率的にバランスよく増やしていくために、不足している音素連鎖を複数含む単語が単語候補リストにあれば優先的に選択し、更にそのような単語を複数用いて、発話文を創作した。長い文が多くなった場合は複数の短い文に分割し、短い文が多くなった場合は複数の文を結合して長い文に作り変えるなどして、文の長さの分布が偏らないように調整した。構文構造についても、右枝分かれ、左枝分かれなどの様々なバリエーションを増やすことを意識して、文を構成した。この作業を、状況ごとの発話文がそれぞれ 700 文になるまで繰り返した。

†「改良セット」構築をご支援いただいた（株）ATR–Trek の関係各位に深謝します。

表 3.1　スキットの例

発話者	状況	発話文
ト書き	0	老人ホームのレクリエーションで、難読ゲームに参加した鈴木さん。
	0	女の子のロボットと一緒に、難問に挑戦しますが、大丈夫でしょうか。
	0	司会者は、ホワイトボードに問題を書きました。
〔　司会者		じゃぁ、次の問題はこれです。＜手風琴＞　〕
ロボット	I	ねぇ、鈴木さんなら、この字読めるんじゃない？
〔　鈴木さん		えー、これは見たことないから、分からないわ。〕
ロボット	III	そう、何でも知ってる鈴木さんでも、分からないことあるんだぁ。
	II	鈴木さんにも分からないんじゃぁ、尚更、私に分かるはずないじゃないよ。
	II	なんか、ヒントないんですかぁ？
ト書き	0	司会者はヒントを１つ出すことにしました。
〔　司会者		そうね、ヒントは、楽器です。〕
ロボット	IV	鈴木さんはまだ分かってないみたいだけど、私、何だか分かった気がするよ。
〔　司会者		そう？　じゃぁ、答え言ってみて。〕
ロボット	IV	もしかしてぇ、アコーディオンって読むんじゃない？
〔　司会者		あっ、凄い！　正解 !!　〕
ロボット	I	やったー、やっぱ私って天才かも。

〔　　〕内の“ 司会者 ”と“ 鈴木さん ”の発話文は収録対象外

表 3.2　典型表現の例

状況	口調	発話文
0	丁寧口調	ファックスの送付先の番号を申し上げますので、メモをご用意ください。
	友達口調	来週の月曜日って、祝日だよね？
	女性口調	会議室にコーヒーを５つ、持ってきてくださるかしら。
I	丁寧口調	皆さん、今日は最後まで楽しんで行ってくださいね。
	友達口調	すごーい、あなたって天才じゃないの？
	女性口調	抜けるような青空で、気分爽快だわ。
II	丁寧口調	折角ですが、そういったご要望にはお応え致しかねます。
	友達口調	いつまでそんなこと言ってるつもりなの？
	女性口調	ここの料理って、最低だわね。
III	丁寧口調	大変残念なお知らせを、しなければなりません。
	友達口調	この調子じゃ、今回もうまく行きそうもないんじゃない？
	女性口調	最近、めまいがひどくて心配だわ。
IV	丁寧口調	大したことがなくて、ホットしました。
	友達口調	もうすぐ、退院できそうなんだって？
	女性口調	来週からは少しのんびりできそうで、良かったわ。

続いて、これらの文セットを用いて、女性声優1名（初期セットとは別の話者）による音声を収録した。収録は一日あたり約3時間、600文程度ずつとして、2週間の期間のうちの6日間で行った。スキットや典型表現を適宜利用することで、状況ごとの音声表現が極端にならないよう注意しながら、収録を進めた。

3.4　音声収集手法の有効性の検証

本手法の有効性を検証するために、以下の三つの視点から、初期セットと改良セットを比較する。

（1）文セットの設計面からの検証

（2）音声セットの音響特徴量面からの検証

（3）合成音声の品質面からの検証

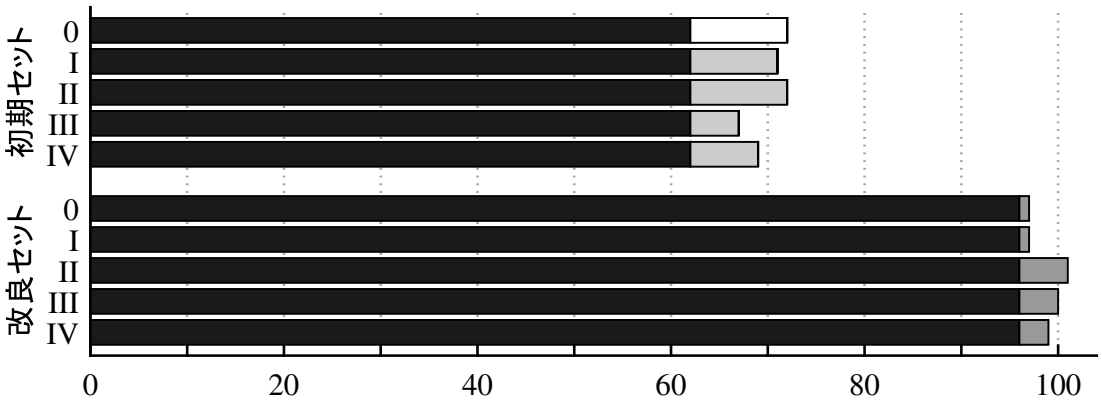
3.4.1　文セットの設計面からの検証

状況ごとの合成音声の品質にばらつきがあると、それらを対話の中で使い分けたときに違和感を生む原因となり兼ねない。状況ごとの文セット間で、含まれている明瞭性や自然性に影響を及ぼす要素の差異は極力小さいことが望ましい。そこで、各文セットに含まれている明瞭性や自然性に影響を及ぼす各種の要素についてそれぞれの種類数を数え上げ、状況ごとの文セット間の差異を確認した。

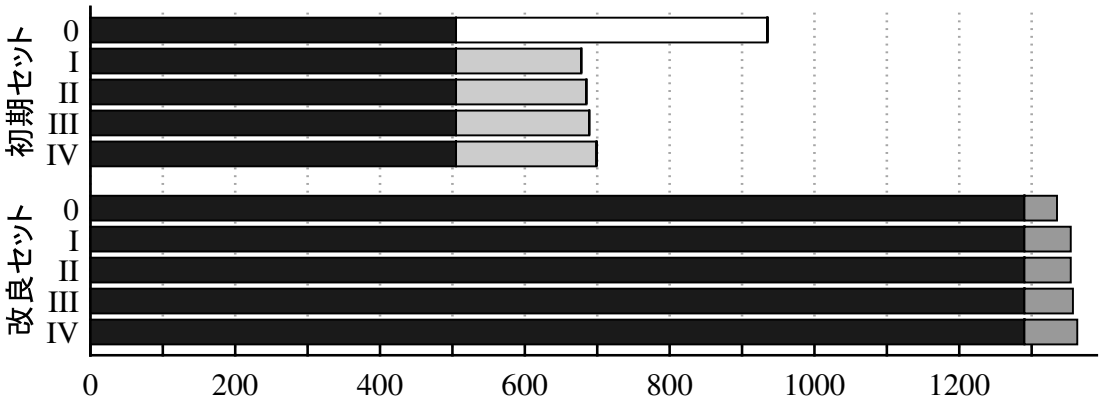
〔1〕 明瞭性に影響を及ぼす要素

各セットに含まれている音素、2音素連鎖、及び3音素連鎖の種類数をそれぞれ数え上げた結果を、図 3.1 に示す。棒グラフの黒塗り部分は、初期セット、改良セットのそれぞれで、5種類の状況の全てに共通して含まれていた音素連鎖の種類数を表している。

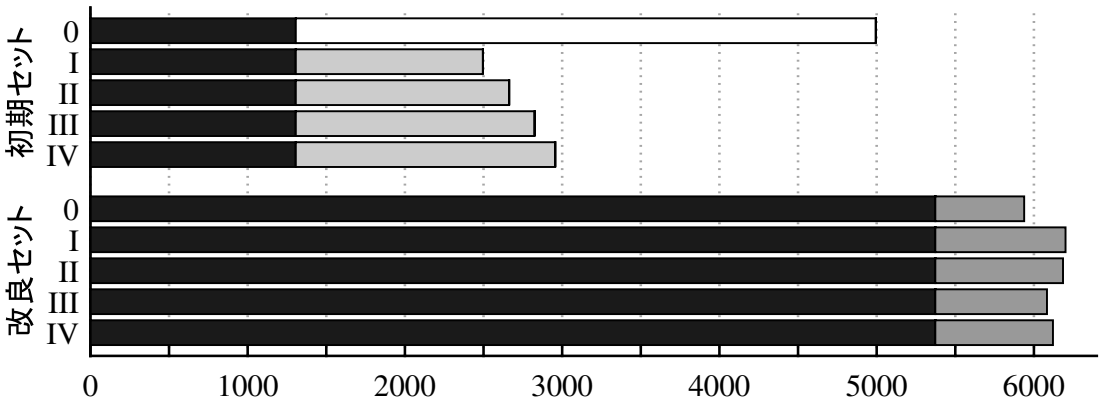
初期セットの状況I～状況IVは、表 2.1 に示したように発話文の数ではそれぞれ、状況0の文セットとして用いたATR503文を若干上回っているが、音素連鎖の種類数は少なく、3音素連鎖では40～50%下回っていた。



(a) 音素の種類数



(b) 2 音素連鎖の種類数



(c) 3 音素連鎖の種類数

図 3.1　明瞭性に影響を及ぼす音素連鎖の種類数

表 3.3　韻律の自然性に影響を及ぼす要素の具体例

発話文	今日も	天気が	抜群に	良くて	、	気分	最高ですよ	。
係り受け関係		無	無	有	(ポーズ)	有		
モーラ数	3	4	5	3		3	7	
アクセント型	1	1	0	1		1	5	
アクセント型種類	頭高	頭高	平板	頭高		頭高	中高	
(a) モーラ数 – アクセント型	3 – 1	4 – 1	5 – 0	3 – 1		3 – 1	7 – 5	
(b) 係り受け関係とアクセント型種類の連鎖	【2 アクセント句連鎖】							
	〔 頭高	無	頭高 〕〔	平板	有	頭高 〕	〔 頭高	有 中高 〕
		〔 頭高	無	平板 〕				
	【3 アクセント句連鎖】							
	〔 頭高	無	頭高	無	平板 〕			
		〔 頭高	無	平板	有	頭高 〕		

これに対して改良セットでは、文数を700に増やしたこともあるが、各音素連鎖とも大幅に種類数を増やすことができている。5状況の全てに共通して含まれていた音素連鎖の種類数（棒グラフの黒塗り部分）だけでも、ATR503文全体を上回っている。音素の定義を拡張した上でバリエーションを考慮したことや、ATR503文に含まれていない音素（「シェフ／《しぇ》ふ」、「略語／《りゃ》くご」、「ウェーブ／《うぇー》ぶ」など）を含めるようにしたことで音素自体の種類数（図 3.1(a)）が大幅に増えた。その音素に基づいて2音素連鎖、3音素連鎖のバリエーションを考慮したことが、それぞれの種類数（図 3.1(b)，図 3.1(c)）を増やすことにつながったと考えられる。また、初期セットに比べ、状況ごとの文セット間での種類数の差が小さい上に、5状況に共通の音素連鎖が多数を占めており、それぞれの文セットに含まれている音素連鎖のバリエーションのばらつきも小さく抑えられていることが分かる。

〔2〕 自然性に影響を及ぼす要素

アクセント句のモーラ数とアクセント型の組合せをエントロピー計算に加えることで、明瞭性だけでなく韻律の自然性に影響を及ぼす要素の出現頻度のバランスも考慮して、大規模コーパスから文セットを抽出する手法が提案されている[54]。アクセント句のモーラ数とアクセント型の組合せのバランスを考慮することは、アクセント句ごとの音調形状を再現する上で有効と考えられる。一方、文レベルの音調形状は、呼気段落内の連続する二つのアクセント句間の係り受け関係の有無とアクセント型の種類の組合せの影響を受けることが知られており、文の音調形状の決定には不可欠な情報である[28]。また、文の長さは、音調形状だけでなく発話速度にも影響を及ぼすと考えられる。そこで、収録した音声に基づいて決定した呼気段落の区切りとアクセント型を用いて、各セットに含まれている韻律の自然性に影響を及ぼすと考えられる、次の (a)〜(c) の要素の種類数を数え上げた。(a)、(b) については、表 3.3 に具体例を示す。

(a) アクセント句のモーラ数とアクセント型の組合せ

(b) 呼気段落内のアクセント句の連鎖（2〜3連鎖）における隣接アクセント句

間の係り受け関係（2 種類：有／無）と、それぞれのアクセント型の種類（3 種類：平板型または尾高型／頭高型／中高型）の組合せ

(c) 文の長さ（モーラ数）

(a)〜(c) の種類数をそれぞれ数え上げた結果を、図 3.2 に示す。棒グラフの黒塗り部分は、初期セット、改良セットのそれぞれで、5 種類の状況の全てに共通して含まれていた要素の種類数を表している。改良セットでは、初期セットに比べ、状況ごとの種類数の差が小さく、かつ、共通の種類が占める割合が高くなっていることから、状況ごとのばらつきを抑えることができていることが分かる。また、改良セットの全ての状況で、ATR503 文を上回る豊富な種類が含まれていることも確認できた。

3.4.2　音声セットの音響特徴量面からの検証

収集した音声セットについて、状況ごとに特徴の異なる音声表現が得られているか、また、違和感につながるほどの表現の乖離を回避できているかを、それぞれの音声の音響特徴量の分布を調べることにより確認した。収録した音声は 16kHz でサンプリングし、STRAIGHT を用いて F0 とスペクトルを 5ms ごとに求めた。音素境界は、音声波形とスペクトログラムの目視によって決定した。その上で、発話単位での音声表現の違いが現れると考えられる、次の (a)〜(d) の音響特徴量を発話ごとに抽出した。

(a) 対数 F0 平均値

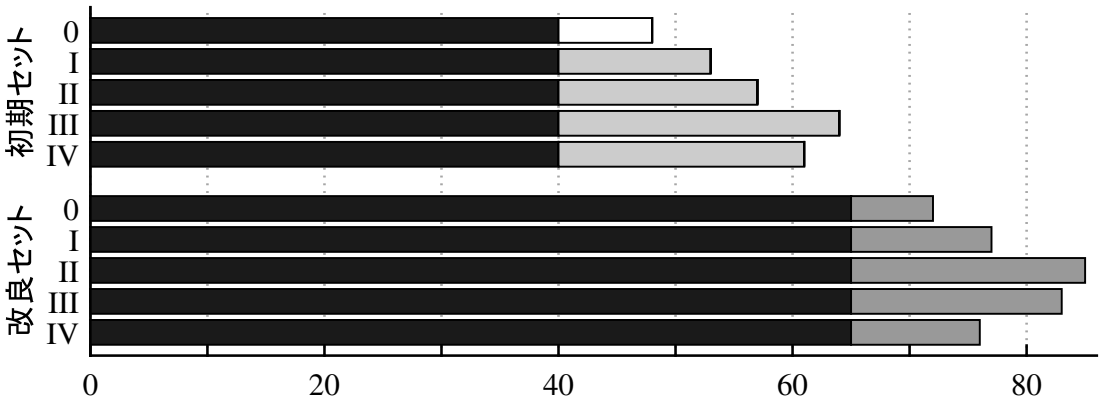
母音及び撥音区間フレームにおける対数 F0 値の発話内平均値

(b) 対数 F0 変動幅

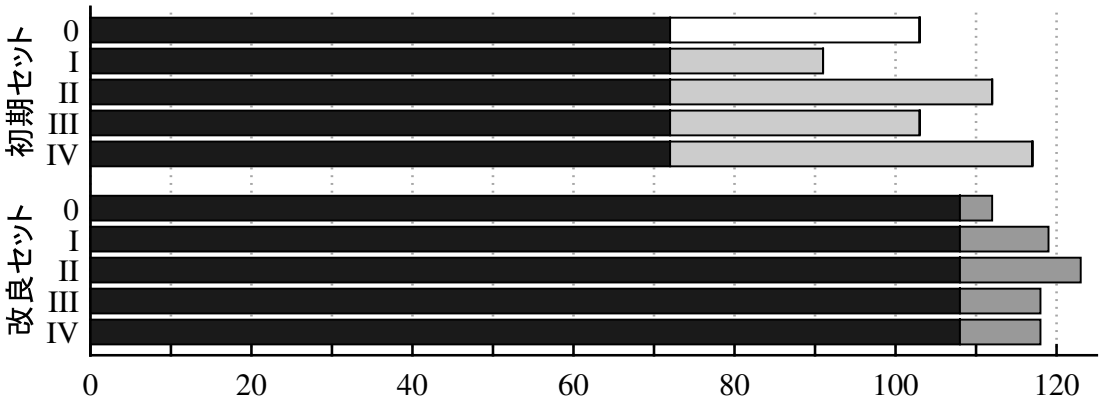
母音及び撥音区間フレームにおける対数 F0 値の発話内最大値と発話内最小値の差

(c) 発話速度

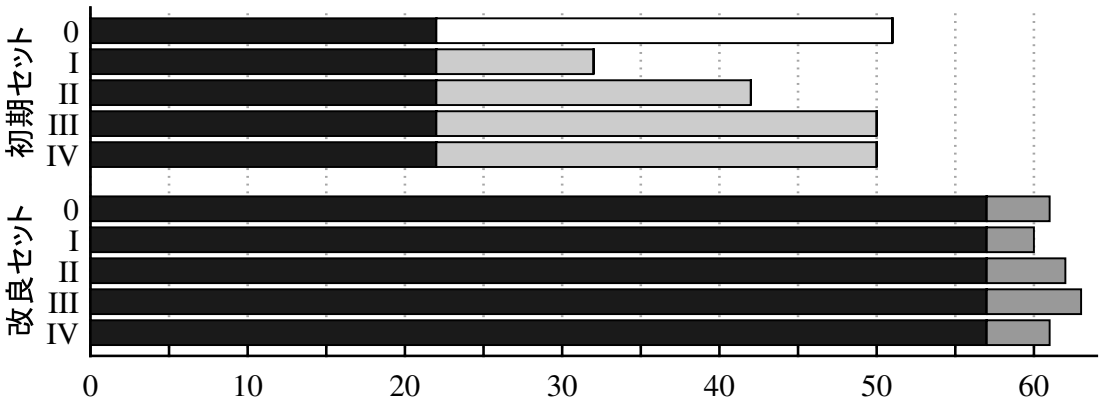
文のモーラ数を発話の始端から終端までの時間長で除した値



(a) モーラ数 – アクセント型の組合せの種類数



(b) 係り受け関係とアクセント型の組合せの種類数



(c) 文の長さ（モーラ数）の種類数

図 3.2　韻律の自然性に影響を及ぼす要素の種類数

(d) スペクトル傾斜

母音区間フレームにおける対数パワースペクトルの平均値（以下、「平均スペクトル」と記す）の傾斜

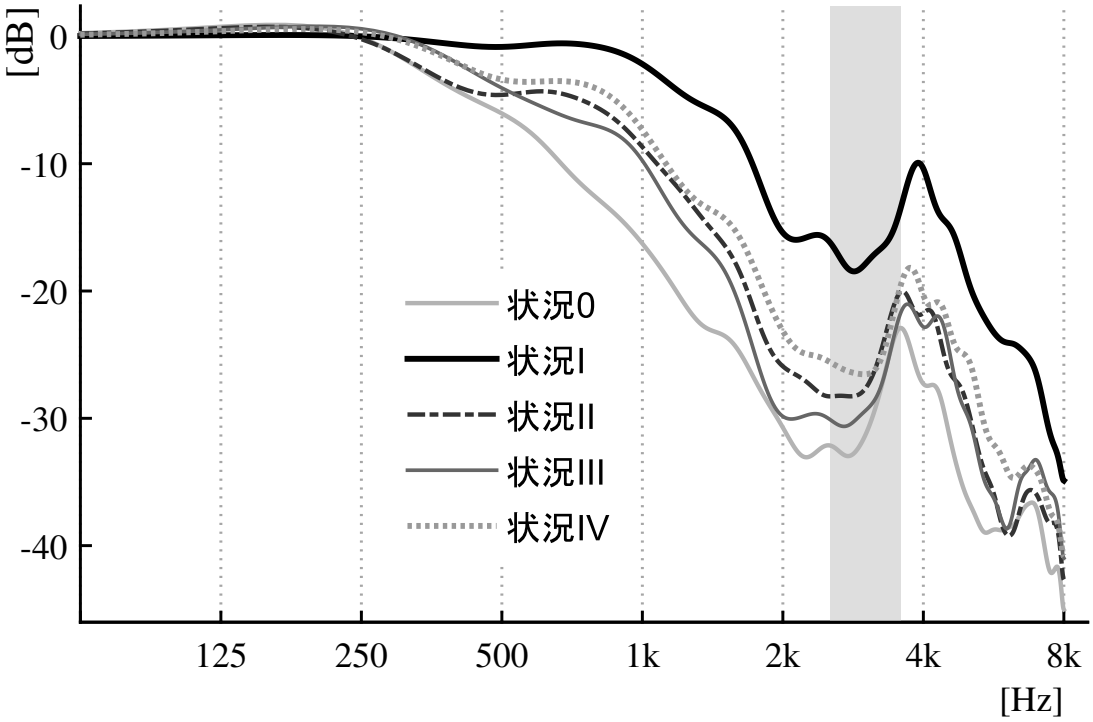
スペクトルの傾斜には、発声の仕方の違いが反映される [150]。状況ごとの全発話の母音区間フレームの平均スペクトルを求めた結果を、図 3.3 に示す。それぞれの平均スペクトルの 0Hz における値を、0dB の基準にして表示してある。スペクトル傾斜の算出には、対数パワースペクトルを 1 次ケプストラム係数で近似したときの 0Hz と 3kHz における強度の差で表す方法 [151] を参考にした。ここでは、STRAIGHT 分析によって既にスペクトルの包絡が得られているためケプストラムによる近似は行わず、代わりに、3kHz を中心周波数とする 1/2 オクターブ幅の帯域（図 3.3 の灰色で示した範囲）における強度の平均値を用いて、スペクトル傾斜を表すことにした。

〔１〕 状況ごとの音声表現の音響的な違い

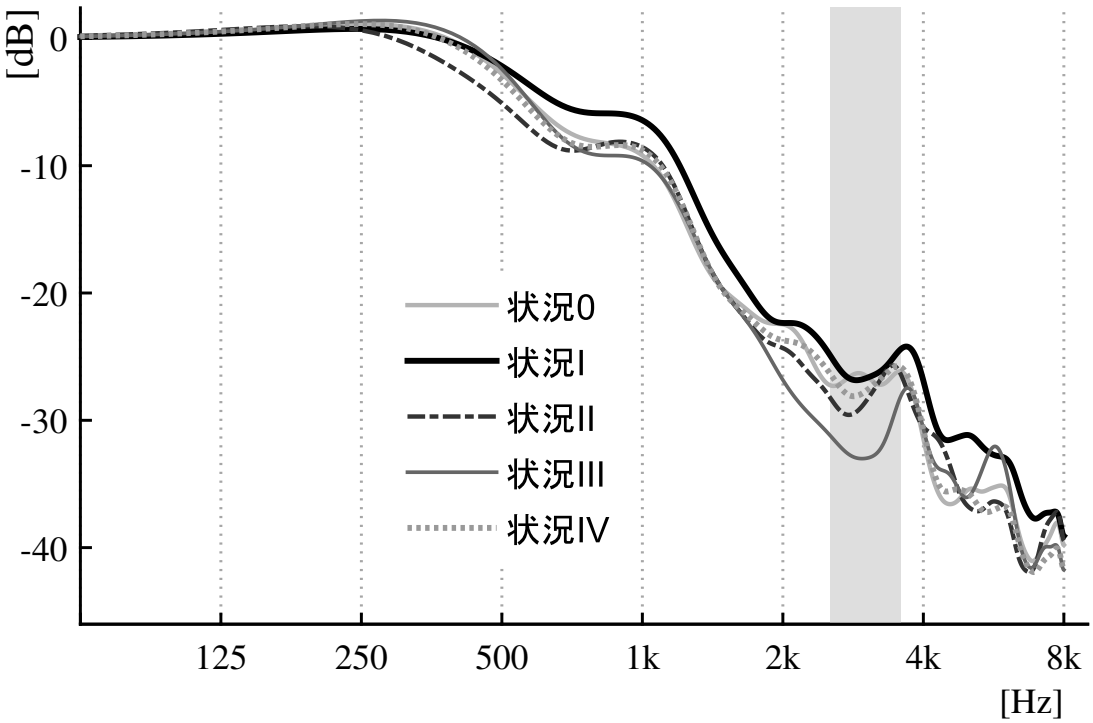
音響特徴量の分布のヒストグラムを、図 3.4 に示す。バーの高さは、そのビンに含まれる発話数の、総発話数に対する割合を表している。平均と標準偏差も、併せて表示してある。

状況ごとに音響的な特徴の異なる音声表現が得られているかを確認するために、音響特徴量の分布について、初期セット、改良セットのそれぞれの 5 状況間の多重比較検定を行った。その結果、初期セットでは、いずれの音響特徴量のいずれの状況の平均値の間にも、有意差が認められた（スペクトル傾斜の状況 II と状況 IV の間のみ $p < 0.05$、それ以外は全て $p < 0.001$）。一方、改良セットでは、スペクトル傾斜の状況 0 と状況 IV の平均値の間に有意差は認められなかったものの、その他ではいずれも有意差が認められた（発話速度の状況 0 と状況 I の間、及び状況 II と状況 III の間は $p < 0.01$、それ以外は全て $p < 0.001$）。

初期セット、改良セットともに、状況ごとに、音響的に異なった特徴を有する音声表現が収集できているものと考えられる。

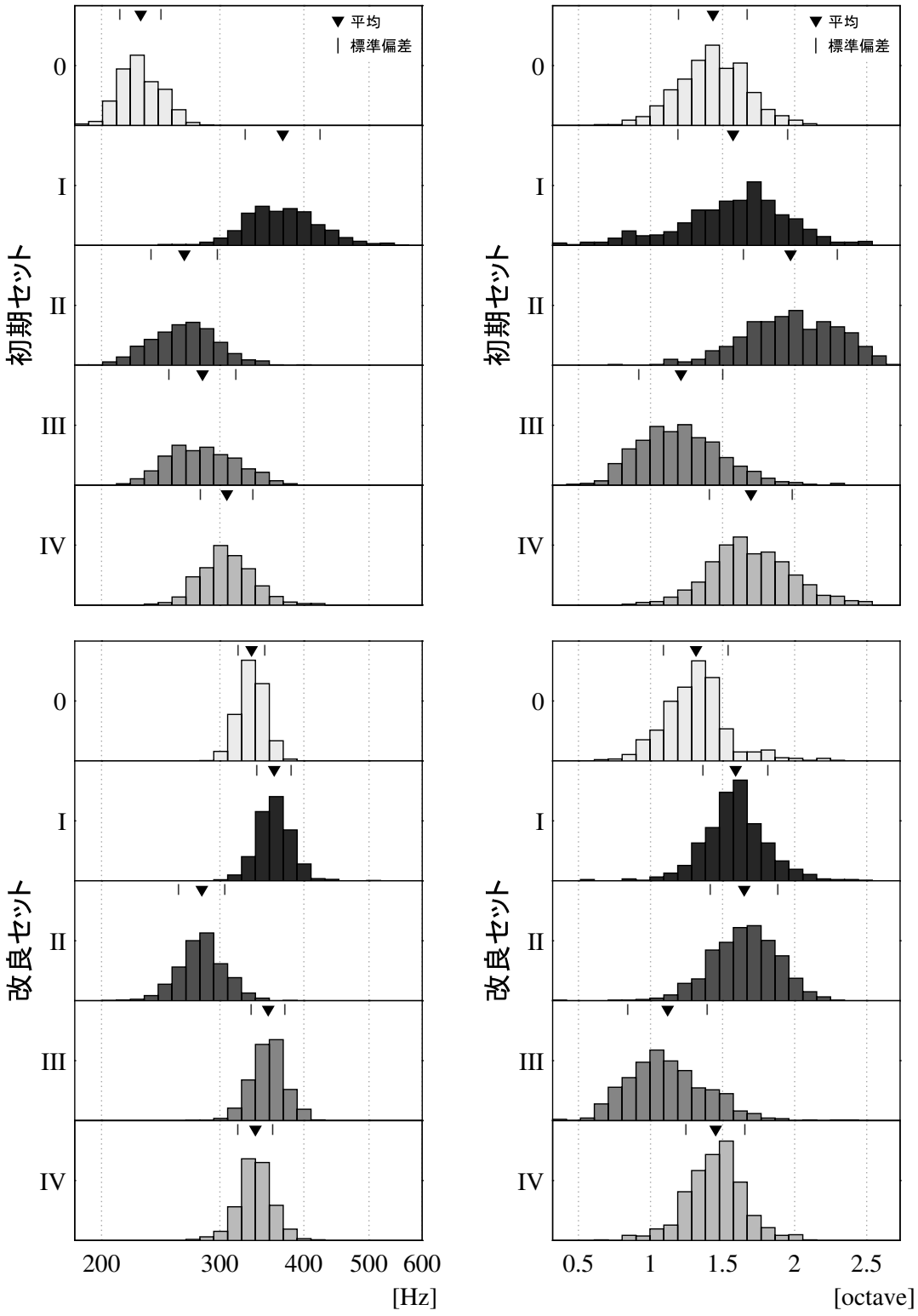


(a) 初期セット



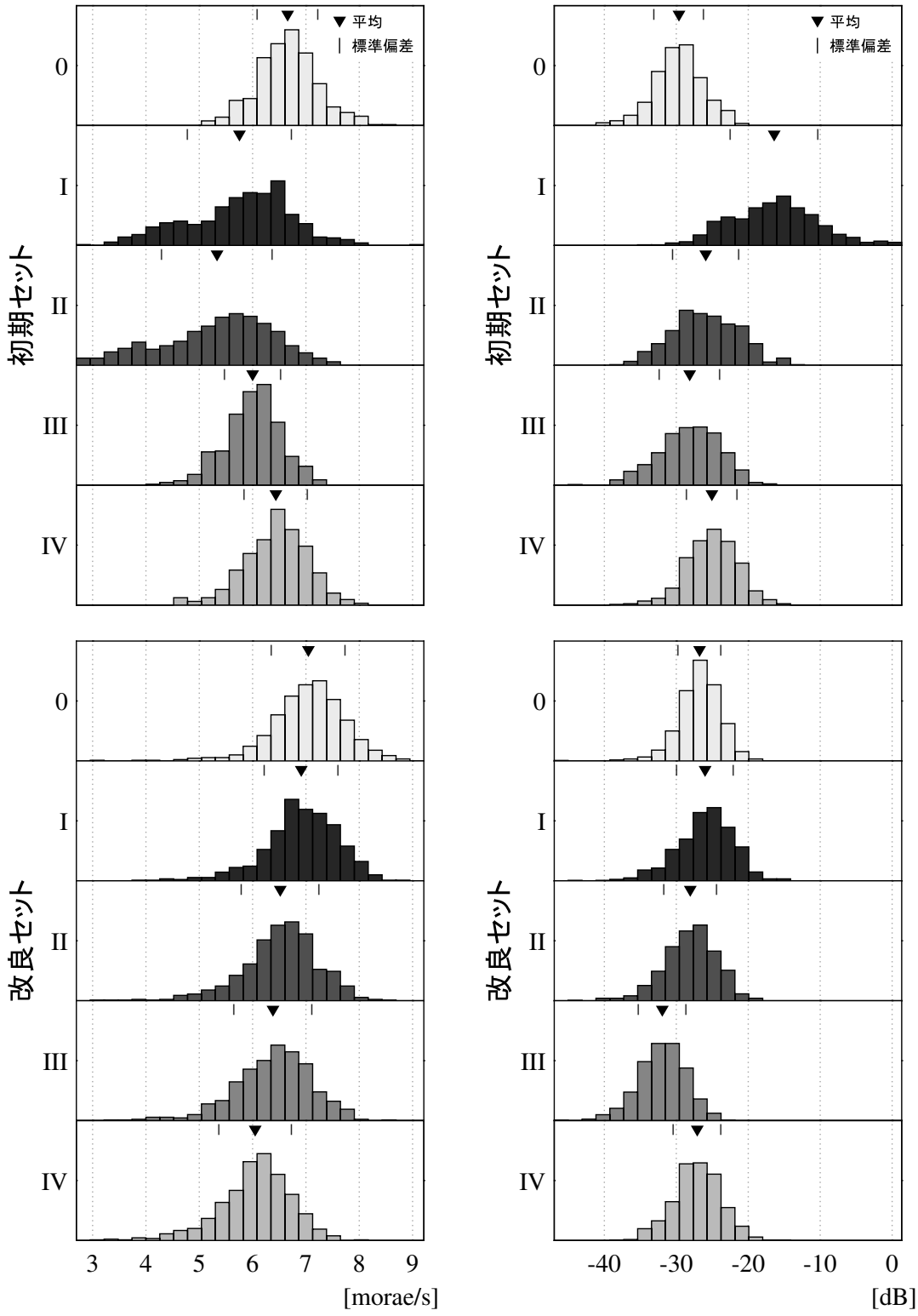
(b) 改良セット

図 3.3　状況ごとの音声セットの平均スペクトル



(a) 対数 F0 平均値　　(b) 対数 F0 変動幅

図 3.4　音響特徴量の分布



(c) 発話速度　　(d) スペクトル傾斜

図 3.4　音響特徴量の分布（続き）

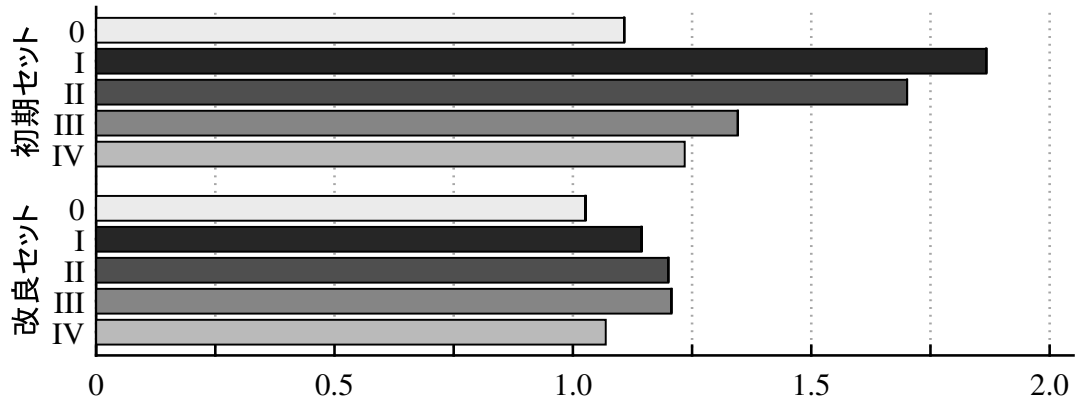


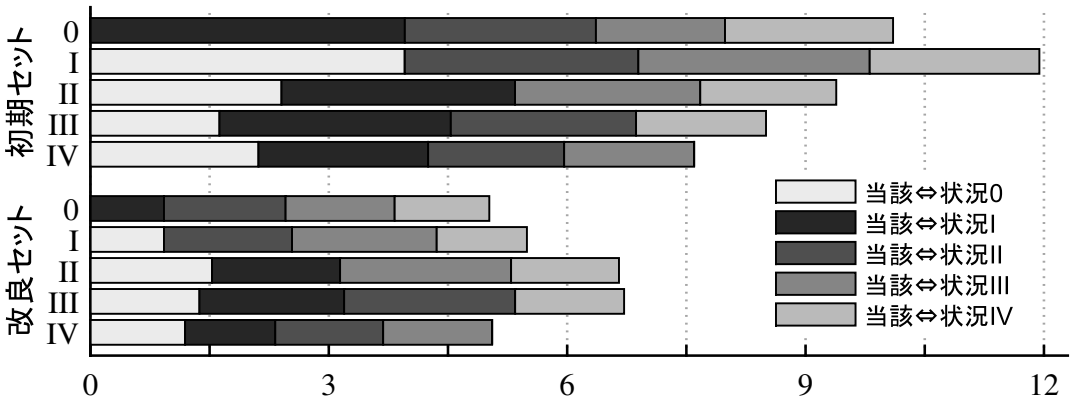
図 3.5　音響特徴量に基づく状況内距離

〔２〕 状況ごとの音声表現の音響的な隔たり

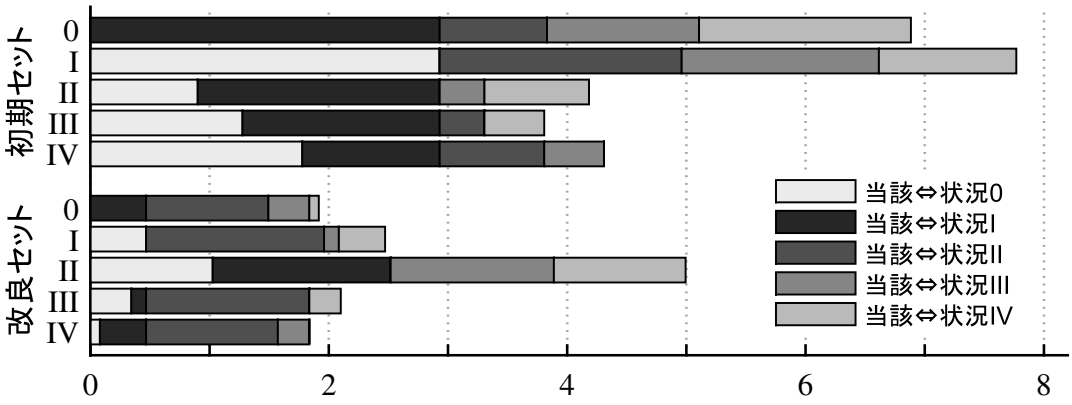
状況ごとに特徴が異なっていることは必要であるが、その違いが極端であると違和感を生じ兼ねない。状況ごとの表現の隔たりの程度を確認するために、音響特徴量空間における距離を調べた。初期セットと改良セットの全発話を対象として、4 種類の音響特徴量をそれぞれ平均 0、分散 1 となるように正規化した。

まず、状況内での表現のばらつきの程度を確認するために、状況ごとに重心と各発話との間のユークリッド距離の平均値を、状況内距離として算出した。結果を、図 3.5 に示す。初期セットの状況 I や状況 II の音響特徴量が、大きくばらついている様子がうかがえる。これらに比べると、改良セットの状況内距離の値はいずれも小さく、状況ごとの差異も小さく抑えられている。改良セットでは、それぞれの状況で、ばらつきの少ない安定した音声表現が収集できているものと考えられる。

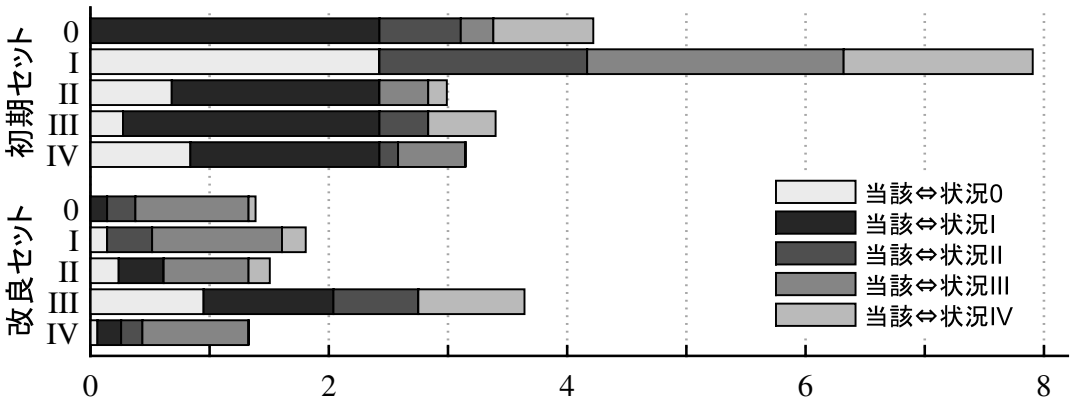
次に、状況間の隔たりを確認するために、状況ごとの重心間のユークリッド距離を、状況間距離として算出した。4 種類の音響特徴量全体の空間における状況間距離を、図 3.6(a) に示す。当該状況と、他の 4 状況との間のそれぞれの距離を積み上げて表示してある。初期セットに比べ、改良セットのいずれの状況も、他の状況との距離は小さく抑えられている。また、音響特徴量ごとの距離のうち、対数 F0 平均値での距離を図 3.6(b) に、スペクトル傾斜での距離を図 3.6(c) に示す。



(a) 全音響特徴量による状況間距離



(b) 対数 F0 平均値の状況間距離



(c) スペクトル傾斜の状況間距離

図 3.6　音響特徴量に基づく状況間距離

これらを見ると、初期セットの状況 I の他との隔たりが突出していることが分かる。このことが、初期セットの状況ごとの音声表現の間に別人の声に聞こえるほどの印象の違いをもたらし、対話の中で使い分けたときの違和感につながっていると考えられる。一方、改良セットでは、状況 II が対数 F0 平均値で、状況 III がスペクトル傾斜で、他との隔たりが大きくなってはいるものの、初期セットほどの大きな距離の差とはなっていない。

音響特徴量空間における距離がどの程度になると違和感を生じるかが明らかではないため初期セットとの比較による議論とはなるが、改良セットには違和感を生むような表現の隔たりがある様子は見られない。

3.4.3　合成音声の品質面からの検証

音声セットの音響特徴量の分析結果からは、改良セットでは、表現全体の調和が保たれた合成音声が得られることが期待される。そこで、明瞭性や自然性の改善の確認として発話文の単位で、調和性の改善の確認としてスキットの単位で、実際に初期セットと改良セットによる合成音声を作成して、それらの対比較による主観評価を行った。合成音声の生成には、DNN 音声合成ツールキット [152] を利用した。初期セット、改良セットの状況ごとに、ツールの諸設定は全て同一にして、DNN の学習と音声の合成を行った。**2.2.3** で述べた HMM の学習では、状況 I〜状況 IV のモデルの学習には状況 0 の音声データを併用したが、ここでは、当該状況の音声データのみを用いてそれぞれのモデルを学習させた。

〔１〕 発話文を単位とした明瞭性・自然性の評価

基本的な品質を確認するために、発話文を単位とした評価を行った。評価用の発話文として、複数の作家の複数の小説作品の中の登場人物のセリフを用いた。5 種類の状況ごとに、ふさわしい発話内容の 100 文を選び出し、できるだけ多くの種類の 3 音素連鎖を含むように、その中から 40 文を抽出した。40 文に含まれる 3 音素連鎖の種類数、総数ともにほぼ同規模の、状況ごとの評価用の文セットを構築した。その上で、初期セット、改良セットのそれぞれで、状況ごとの DNN を用いて合成音声を作成した。

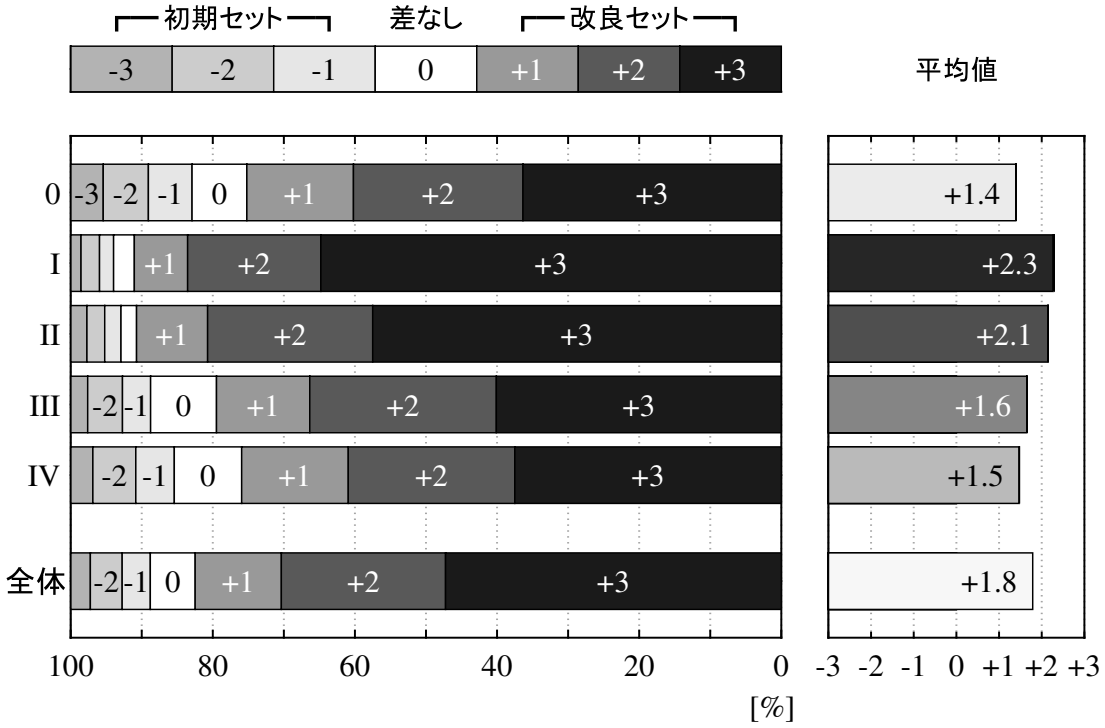


図 3.7　発話文単位の対比較評価結果

評価者 24 名に、200 文（40 文×5 状況）の合成音声の対を、初期セット、改良セットの順序も含めてランダムに提示し、発音の明瞭さ、滑らかさ、リズムやイントネーションの自然さ、対話らしさを総合して、どちらの方が高品質かを 7 段階で判定させた。高品質とされた合成音声が改良セットによるものであった判定には +1、+2、+3 を、初期セットによるものであった判定には −1、−2、−3 を、差はないとされた判定には 0 を、評点として付与した（±1 :「どちらかといえば」、±3 :「顕著に」）。各評点の判定数の割合と平均値を、図 3.7 に示す。改良セットの方が高品質であるとされた判定（+1〜+3）は、状況 0〜状況 IV の状況ごとでは 75.2〜91.0%、5 状況全体では 82.5%を占めた。有意差検定（両側 t 検定）の結果、状況 0〜状況 IV、及び全体のいずれについても帰無仮説「評点の平均値は 0」は棄却（$p < 0.001$）され、両者の品質に有意な差があることが確認できた。

改良セットの設計に当たっては、音素連鎖の種類や文の長さ、文末表現などのバリエーションを増やすことにも注力した。このことが、合成音声の基本的な品質の改善につながったと考えられる。

表 3.4　観光案内をするロボットと利用者の対話

状況	発話文
I	いらっしゃいませ、私は東京都のオフィシャル観光ガイドです。
	〔 夏休みにお薦めの観光スポットを教えてください。〕
IV	それでしたら、ぜひ小笠原で、雄大な自然を満喫して来てください。
	〔 小笠原には行ってみたいけど、メチャクチャ遠いじゃないですか。〕
II	うーん、世界遺産なんですけど、残念だなー。
III	じゃあどこが良いかなあ。
I	そーだ、浅草で下町情緒を楽しむってゆうのは、いかがですか。
	〔 浅草には、もう何度も行ったことあるんですよ。〕
III	えーえ、既に浅草はツウだったんですか。
IV	それなら、谷中なんかどうですか。
	〔 えー、お墓ですか。〕
II	いえいえ、お墓だけってわけじゃあないんです。
	〔 えっそうなの。〕
I	最近じゃあ、海外からの観光客にも大変人気で、とっても賑わってるんですよ。

〔　〕は利用者の発話

（初期セット、改良セットのいずれとも異なる女性声優の音声セット[6]に基づく DNN 合成音声）

〔２〕 スキットを単位とした調和性の評価

異なる音声表現を対話の流れの中で使い分けても、違和感なく自然と感じられる表現となっているかを確認するために、スキットを単位とした評価を行った。評価用として、新たに 3 種類のスキットを作成した。

(S1) 観光案内をするロボットと利用者の対話

(S2) ニュースや話題を提供するロボットと利用者の対話

(S3) 小説の読み聞かせ

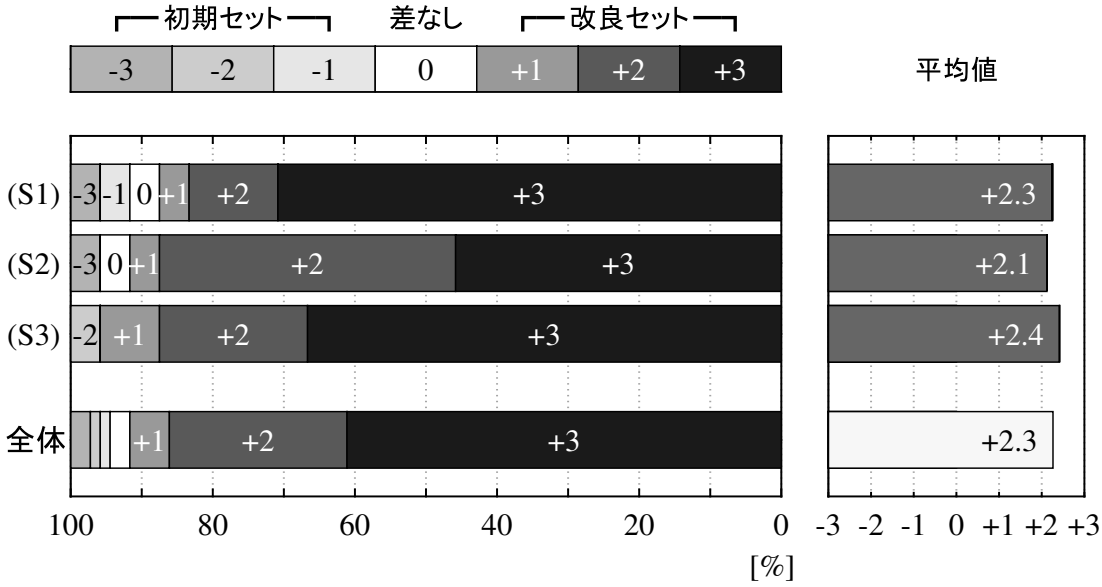


図 3.8　スキット単位の対比較評価結果

(S1) の対話の内容を、表 3.4 に示す。異なる表現全体の調和性を評価させるために、対話の流れの中で状況 I〜状況 IV の音声表現が次々と現れる構成とした。(S2) でも同様に、ロボットの発話に状況 I〜状況 IV の合成音声をそれぞれ 2〜3 発話ずつ用いた。(S3) では、小説の中から複数の登場人物による対話の一場面を抜き出し、状況 I〜状況 IV の合成音声をセリフ部分のそれぞれ 2〜3 発話ずつに用い、ト書き部分の 5 文には状況 0 の合成音声を用いた。

話し方や声質が対話の流れや発話内容にふさわしいものとなっているかを評価の視点に加え、評価者 24 名（発話文を単位とした評価と同じ評価者）に、同様にして 7 段階で判定させた。結果を、図 3.8 に示す。改良セットの方が高品質であるとされた判定は、(S1)〜(S3) のスキットごとでは 87.5〜95.8%、全体では 91.7%を占めた。有意差検定（両側 t 検定）で、(S1)〜(S3)、及び全体のいずれでも「評点の平均値は 0」は棄却（$p < 0.001$）され、両者の品質に有意な差があることが確認できた。

3.4.4　検証結果のまとめと考察

文セットの設計面からの検証によって、改良セットの状況ごとの文セットでは、初期セットよりも音素連鎖や韻律に影響を及ぼす要素のバリエーションが増え、状況ごとのセット間のばらつきも小さく抑えられていることが示され（図 3.1、図 3.2)、初期セットからの改善が図れていることが確認できた。

音声セットの音響特徴量面からの検証では、改良セットは、状況ごとに異なる音響的特徴をもち、かつ、全体の調和が保たれた音声表現となっていることが期待できる音声セットとなっていることが示された（図 3.4〜図 3.6)。

合成音声の品質面からの検証では、明瞭性や自然性の評価で、いずれの状況においても、改良セットによる合成音声の方が高品質であるとの判定が多数を占めた（図 3.7)。状況ごとの評点の平均値を見ると、両者の品質の差は、状況 I が最も大きく、以下、状況 II、状況 III、状況 IV、状況 0 の順となっている。これを例えば、状況ごとの文セットに含まれる 3 音素連鎖の種類数の結果（図 3.1(c)）と比較してみると、同じ順で、初期セットでの種類数が少なく、両者の差が大きくなっている。すなわち、音素連鎖の種類数の差が大きいほど、品質の差も大きくなっていることが分かる。含まれる音素連鎖のバリエーションの多寡が、合成音声の品質に影響を及ぼしている様子が見て取れる。

表現全体の調和性の点においても、改良セットは、初期セットに比べて高品質であるとの評価を得た（図 3.8)。発話文単位とスキット単位の全ての評価結果を対象として、評点の平均値の差の検定（両側 t 検定）を行ったところ、「両者の評点の平均値は等しい」は棄却（$p < 0.01$）された。この結果は、スキット単位での改良セットと初期セットとの品質の差は、発話文単位での品質の差以上に大きいと判定されたことを意味している。表現全体の調和が取れていなかった初期セットに対して、改良セットでは、明瞭性や自然性だけでなく、調和性の点でも改善が図れていることの現れと考えられる。

3.5　むすび

複数の異なる表現全体の調和を保つことを考慮に入れた、多様な音声表現の収集手法について論じた。それぞれの音声表現が適切になるようにすることだけでなく、それらを対話の流れの中で使い分けることを念頭に置き、異なる表現全体の調和を保つことにも注力する必要があることを指摘した。その上で、全体の調和を保つために、異なる音声表現を対話の流れの中で次々と表出させるスキットなどを導入した音声収集手法を設計した。

明瞭性や自然性に影響を及ぼす要素のバリエーションを考慮した文セットの設計面からの検証、表現全体の調和を保つことに力点を置いて収集した音声セットの音響特徴量面からの検証、及び、対話の状況に応じた音声表現を再現した合成音声の品質面からの検証を通して、本手法の有効性を確認することができた。

感情音声などの収録においても、それらを音声対話システムなどのアプリケーションで利用することを考えるのであれば、異なる表現全体の調和を保つことを考慮に入れることは必要であろう。

第4章　文末音調による意図表現の確立に向けた文末F0形状の分類

4.1　まえがき

言語学の分野では、文末詞の機能が、文末音調との関係とともに議論されている[120]~[133]。その中で、文末音調は「普通の上昇調」、「浮き上がり調」、「反問の上昇調」などの名称を用いて分類されている。しかし、これらの分類は研究者の内省に基づくものであり、合成音声でそれらの音調を実現するためには不可欠な、具体的なF0形状が明らかになっているわけではない。また、大規模な音声データを調査対象として、文末F0形状にどのようなバリエーションが存在するかが調べられたこともない。加えて、従来では同じ音調に分類される形状の中に機能が異なるバリエーションはないのかや、従来は取り上げられてこなかった形状で特別な機能をもつものはないのかなど、文末音調がもつ意図表現の機能に関する興味は尽きない。

文末音調によって発話意図を表現する音声合成技術を確立するための準備として、対話調の音声データから文末F0形状を抽出し、文末音調の具体的な形状を収集することからはじめる。収集した文末F0形状を正規化し、それらをクラスタリングによって分類することを試みる。クラスタリングの手法には、大別すると階層的手法と非階層的手法の2種類がある。非階層的手法では、あらかじめ定めた分割数に対する最適な分割を求める。しかし、文末音調の分類数については、表 1.1に示したように議論が分かれている。そこで、階層的手法による分類を試みる。階層的クラスタリングでは分類の階層構造が樹形図によって示されるため、どのような形状が存在しているかを見渡しやすくなることも期待できる。各クラスタのセントロイドは、合成音声に文末音調を付与するためのF0形状のテンプレートとしての利用が可能と考えられる。

4.2　音声試料

文末 F0 形状の収集には、**2.2.2** で作成した、状況 I〜状況 IV の音声データを使用した。これらは、自分の意思を表明したり、相手の反応を伺ったりする話し言葉による対話調の文を、女性声優 1 名が発話した音声である。収録に当たっては、特定の意図を表現させる指示はせず、文の内容に応じて声優自身の判断で自由に表現させた。ただ、各文の文末では、様々な文末詞を伴って、対話音声らしい表情豊かな表現がなされていることから、ここから文末音調の多彩なバリエーションを収集できるものと考えた。

これらの音声のうち、文末の母音が無声化しているものを除いた 2,092 発話を対象として、文末 F0 形状を収集した。

4.3　文末 F0 形状の抽出

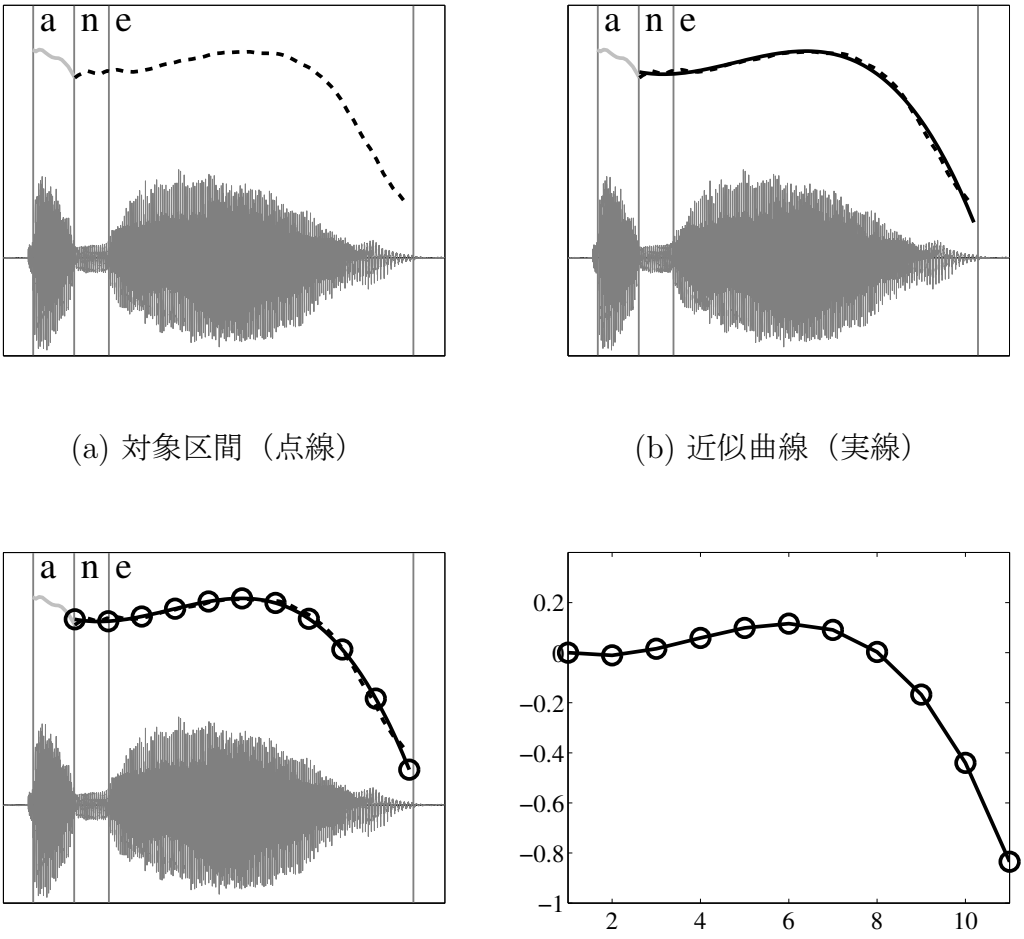
使用した音声データでは、「〜でしょう」や「〜ください」などで終わる発話の「しょう」や「さい」の区間全体で文末音調としての F0 形状が形作られている様子が見て取れたことから、文の最終音節の母音区間を文末 F0 形状の抽出対象区間とすることにした。また、子音が有声子音 /n/、/m/、/y/、/r/、/w/ のときは、子音区間の F0 の動きが後続する母音のそれと連続しており、音声の振幅も比較的大きい。このため、合成時にはこの区間も含めて連続した滑らかな F0 形状を生成する必要があると考え、これらの子音区間も対象に含めることとした。

図 4.1 に、終助詞 “ね” を文末にもつ発話を例として、文末 F0 形状の抽出手順を示す。クラスタリングによる分類を可能にする目的で、時間軸と周波数軸の正規化も併せて施す。

（1）対象区間の F0 を抽出し、対数 F0 値を得る（図 4.1(a) 点線）。

（2）揺らぎなどの影響による局所的な F0 の変動を取り除くために、対数 F0 値を 3 次の最小 2 乗曲線で近似する（図 4.1(b) 実線）。

（3）対象区間の継続時間長を 10 等分し、始端・終端を含めた 11 点で近似曲線の値をサンプリングする（図 4.1(c) 丸印）。



(a) 対象区間（点線）　(b) 近似曲線（実線）

(c) サンプリング点（丸印）　(d) 正規化した文末 F0 形状

図 4.1　文末 F0 形状の抽出手順

（4）発話ごとに異なっている声の高さを始端を基準としてそろえるために、サンプリングした 11 点を、全体の形状は保ったまま始端の高さが 0 となるよう周波数軸方向に平行移動する（図 4.1(d)）。

以上の手順によって、時間軸と周波数軸を正規化した文末 F0 形状を得る。

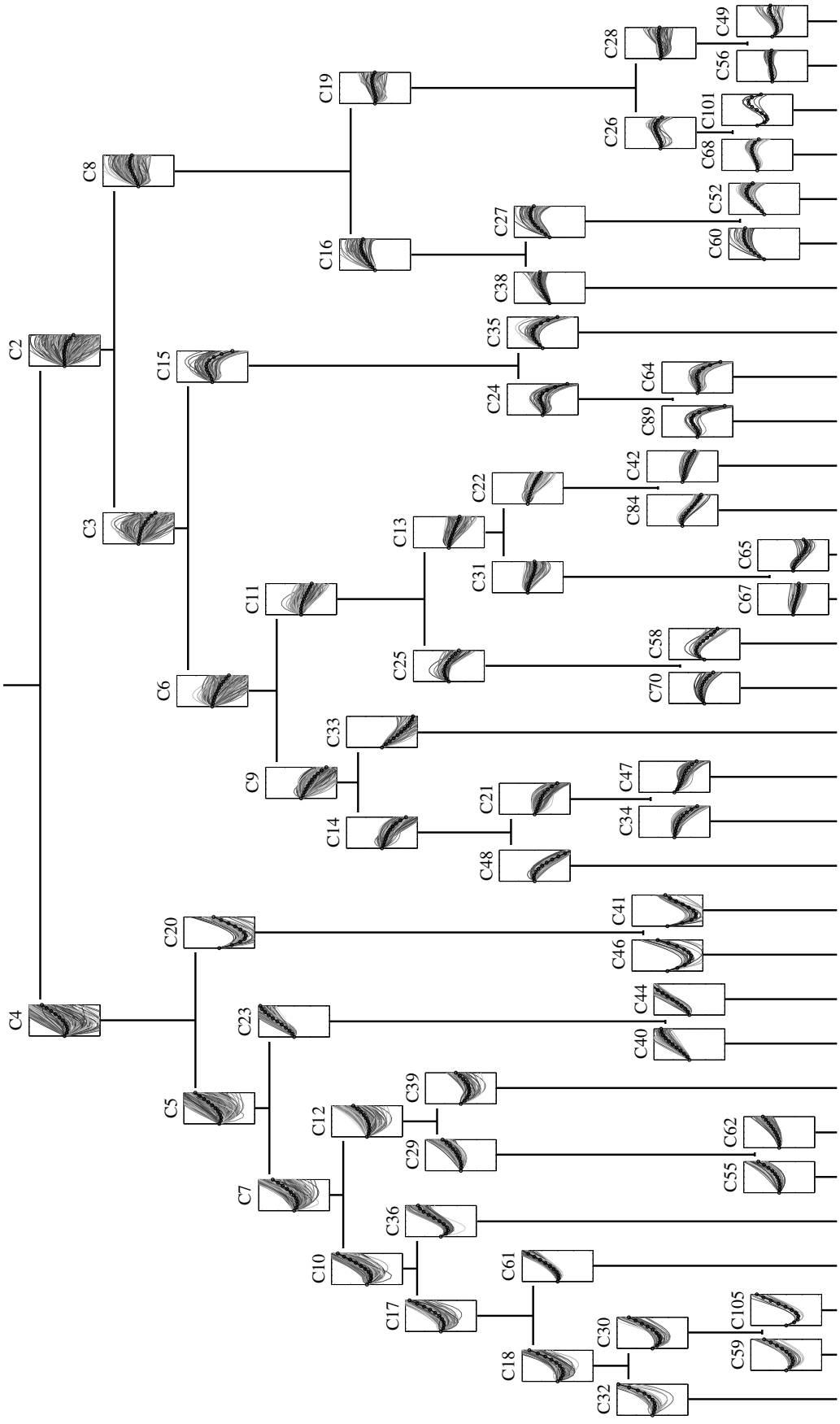
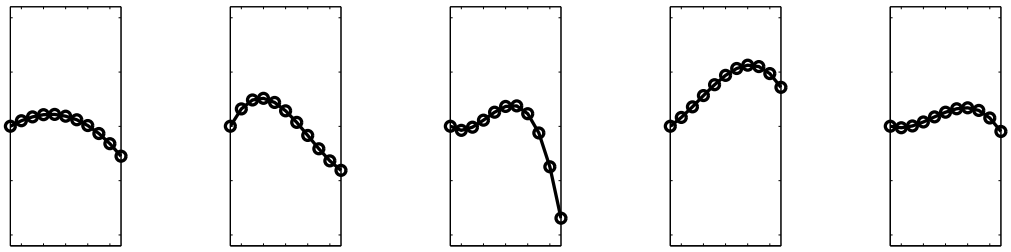


図 4.2　文末 F0 形状のクラスタリング結果（32 分割まで）



(a) C70　(b) C58　(c) C89　(d) C52　(e) C68

図 4.3　「上昇下降調」のバリエーション

4.4　クラスタリングによる文末F0形状の分類

文末音調のバリエーションを知るために、収集した文末F0形状をクラスタリングにより階層的に分類した。分類は、正規化した文末F0形状の時間変化を表す10次元のベクトル（図 4.1(d) のF0値の時間差分）を特徴量として、Ward法[153]により行った。図 4.2 に、全データを32クラスタに分割するまでの、各クラスタに分類された文末F0形状の全て（細線）とそのセントロイド（太線）を示す。紙面の都合で縦方向の枝の長さはクラスタ間の距離を反映していないが、分割の順序は保持してある。C2などの番号は便宜的に付与したクラスタの識別番号であり、当該クラスタが分割される順番を表している。図示はしていないが、根ノードに位置する全データによるクラスタがC1である。Ward法は凝集型のクラスタリング手法であり、図 4.2 の樹形図は、実際には葉ノード側からクラスタを順次結合していくことで構成されたものである。しかし、この後の処理での図の参照方向に合わせて、ここでは便宜上、根ノードからクラスタを順次分割していくものとして説明を行う。

各クラスタのセントロイドを見ると、単純な上昇調（C5など）、下降調（C6など）だけでなく、下降上昇調（C20など）や上昇下降調（C15など）などと呼べる、多彩な形状があることが分かる。更に、上昇調と呼べる音調には、上昇の仕方や上昇の幅の異なる様々なバリエーション（C59、C62、C44、C60など）も存在している。上昇下降調には、図 4.3 に示すような、上昇の仕方、下降の仕方の異なる様々なバリエーションが存在している。

なお、文末のアクセントのレベルが「高」（文末のアクセント句が平板型または尾高型）の場合と「低」（頭高型または中高型）の場合とで別々に文末 F0 形状を分類することも試みたが、音声試料中に前者のデータが 1 割強程度しかなかったことや、それぞれの分類で得られたクラスタのセントロイドに顕著な差異が認められなかったことから、ここでは両者を区別せずに扱うことにした。

4.5　むすび

対話調の音声データから文末 F0 形状を抽出し、文末音調としてどのような具体的な F0 形状が存在するかを明らかにした。抽出した F0 形状の時間軸と周波数軸を正規化した上で、階層的クラスタリングの手法を用いて分類を行った。

クラスタリングによって得られた樹形図の観測から、単純な上昇調、下降調だけでなく、下降上昇調や上昇下降調などの多彩な形状があることが分かった。更に、例えば上昇下降調には、上昇の仕方、下降の仕方の異なる様々なバリエーションが存在していることが確認できた。樹形図上の各クラスタのセントロイドは、似通った形状を幾つか集めて算出した平均値であり、いずれも、合成音声に文末音調を付与する際の具体的な F0 形状のテンプレートとして利用することが可能である。

第 5 章では、これらの中から発話意図を伝える文末音調の生成に利用するテンプレートを絞り込む手法を検討する。

第5章　意図表現に用いる文末音調のF0形状テンプレートの獲得

5.1　まえがき

第 4 章では、対話調の音声データから文末の F0 形状を抽出し、クラスタリングの手法を用いてそれらを階層的に分類することで、文末音調の具体的な形状のバリエーションを明らかにした。図 4.2 の樹形図上に示した各クラスタのセントロイドは、どれも合成音声に文末音調を付与するためのテンプレートとして利用することは可能である。ただ、実用上は、樹形図上の多数のクラスタのセントロイドの中から幾つかを選んで、文末音調を付与するためのテンプレートとして利用することになる。

クラスタリングによる分類は、あくまでも F0 形状を表す特徴量に基づいて、計算よって求めたものである。このため、見た目の形状に違いはあっても、音調として聴取したときに差異が感じられないものが存在することは十分あり得る。また、一つのクラスタを分割した二つのクラスタのセントロイドによる音調に聴感上の差異がなければ、分割してもテンプレートのバリエーションを増やすことにはならない。しかし、ある程度まで分割数を増やしていけば、聴感上の差異がない F0 形状が出てくる可能性は高くなる。

文末音調の付与に使用するテンプレートとして用意しておく F0 形状は、実際に音調として聴取したときにそれぞれの差異が感じられるものである必要がある。そこで、テンプレートの絞り込みに、聴感上の判定を基準として利用する方法を検討する。

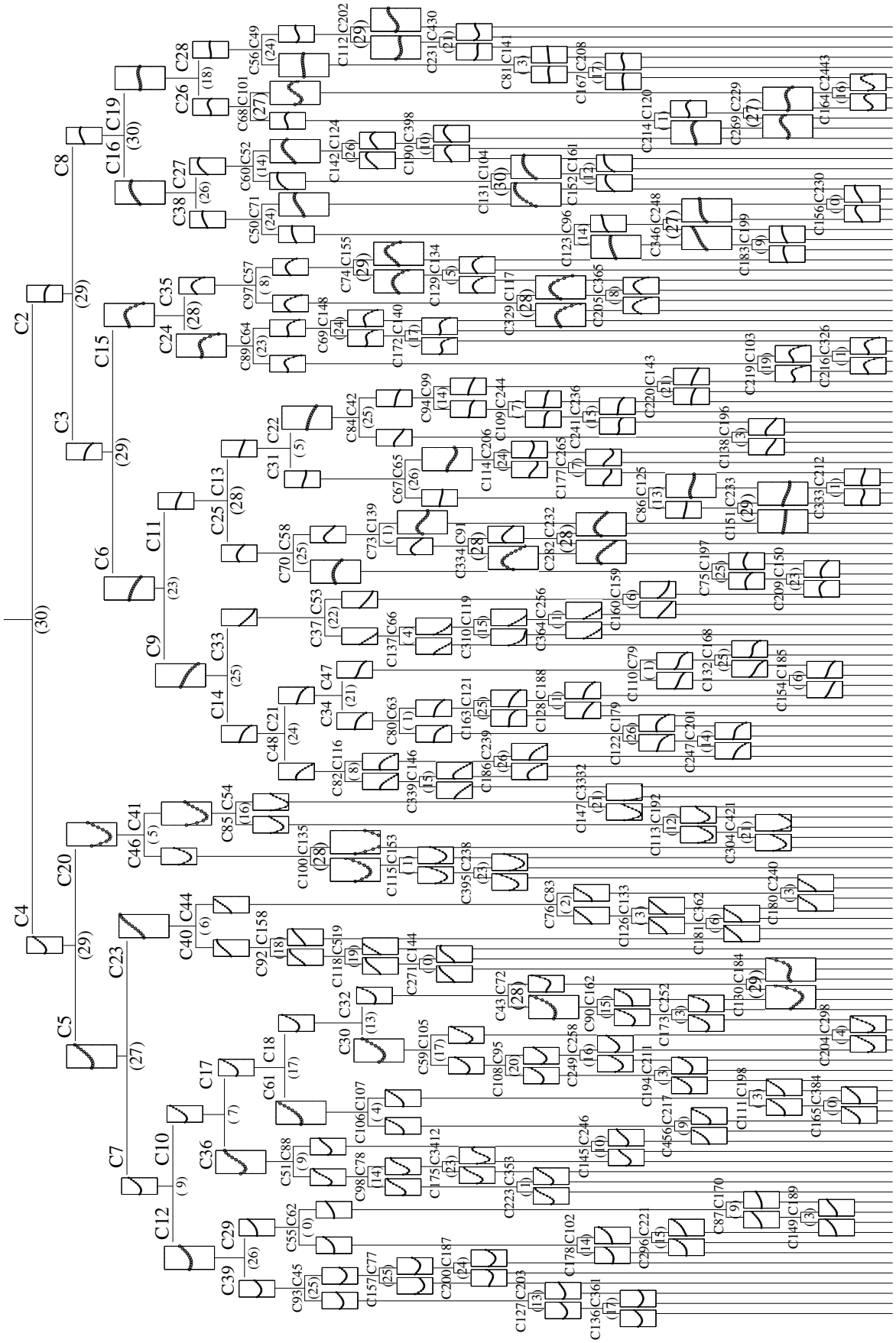


図 5.1　クラスタセントロイドによる音調の対比較実験結果

5.2　セントロイドによる音調の対比較実験

聴感上の判定の基準を得るために、樹形図の各分割点につながる二つのクラスタのセントロイドをF0形状として付与した音声を聴き比べて、音調としての差異を感じられるかどうかを判定させる対比較実験を行う。その結果から、二つのクラスタのセントロイドによる音調に差異がなければその分割はしないものとすることで、聴感上の判定を利用した絞り込みを実現することを考えた。

5.2.1　実験方法

音調を付与する音声には、相槌(づち)の音声「はぁ」を使用した。相槌「はぁ」には、

- 言語的には意味をもたない。
- 発話時間が短い。
- 発話全体にテンプレートによる音調を適用できる。

という特徴があることから、

- 評価者が音調の違いの比較に集中できる。
- 評価の所要時間を短くできる。

などの効果が期待された。更に、

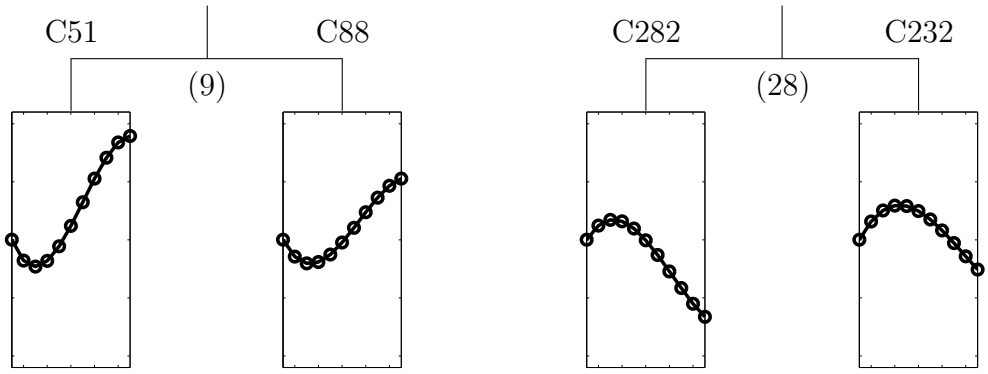
- 音調によって、様々な意図やニュアンスを表現する相槌となり得る。

と考えられるため、

- 音調としての差異があるかどうかを判断しやすくなる。

ことも期待された。

あらかじめ収録してあった相槌の音声「はぁ」をSTRAIGHTで分析し、再合成する際に、母音区間のF0形状をテンプレートによる形状に置き換えた。今後、テンプレートとして利用する種類数などを考慮して、文末F0形状の全データを128クラスタに分割するまでの127の分割点において生成されたクラスタ対の各セントロイドを対象とすることにした。すなわち、127対、254種類の音調を付与した相槌の「はぁ」を、STRAIGHTにより合成した。



(a) 見掛け上の違い：大　　(b) 見掛け上の違い：小

図 5.2　音調としての違い

127 対の順序の入れ替えも含め、全 254 対の「はぁ」を、評価者 15 名にランダムに提示した。二つの音調を「同じ」と感じるか、「違う」と感じるかの二者択一で強制判定させて、セントロイドの対ごとに 30 回答を得た。

5.2.2　結果と考察

図 5.1 に、結果を示す。括弧内の数値が、その上方に示されているクラスタが分割されてできた左右二つのクラスタのセントロイドによる音調が「違う」と判定された回答数を表している。例えば、C4 が分割されてできた C5 と C20 のセントロイドによる音調が「違う」と判定された回答は、30 回答中 29 あった。C5 が分割された C7 と C23 では、30 回答中 27 が「違う」の判定であった。

ここで、二組のクラスタ対の例を、図 5.2 に示す。見掛け上は、図 5.2(a) に示した二つのセントロイドの形状の差は大きく、これと比較すると、図 5.2(b) の二つの形状の差は小さいように思われる。ところが、聴感上は、図 5.2(a) の二つのセントロイドは音調として「同じ」であるとの判定が多数を占めているが、図 5.2(b) は「違う」と判定されており、見掛け上とは逆の傾向を示している。この結果は、テンプレートの絞り込みに聴感上の判定を導入することの必要性を示唆している。

5.3　聴感上の判定に基づくテンプレートの絞り込み

相槌の音声「はぁ」に付与した音調の対比較実験結果を、テンプレートの絞り込みに利用する。具体的には、

（1）「違う」の回答数にしきい値を設定

（2）それを基準として、根ノードから順に以降の分割を進めるかどうかを判断

（3）最終的に葉ノードとして残ったセントロイドを採用

という手順によって絞り込むことを考えた。

そこで、音調としての差異が顕著なものを選び出すことを想定した絞り込みを試みた。しきい値を 29（30 回答中 29 以上が「違う」であれば分割を進める）と設定した場合の絞り込みの手順を、図 5.3(a) を参照しながら説明する。

（1）全データのクラスタ C1（図には示していない）を分割してできる子クラスタの C4 と C2 による音調が「違う」と判定された回答は 30 あるので、C1 は C4 と C2 に分割する。

（2）C4 の子クラスタ C5 と C20 では「違う」の判定が 29 あるので、分割する。

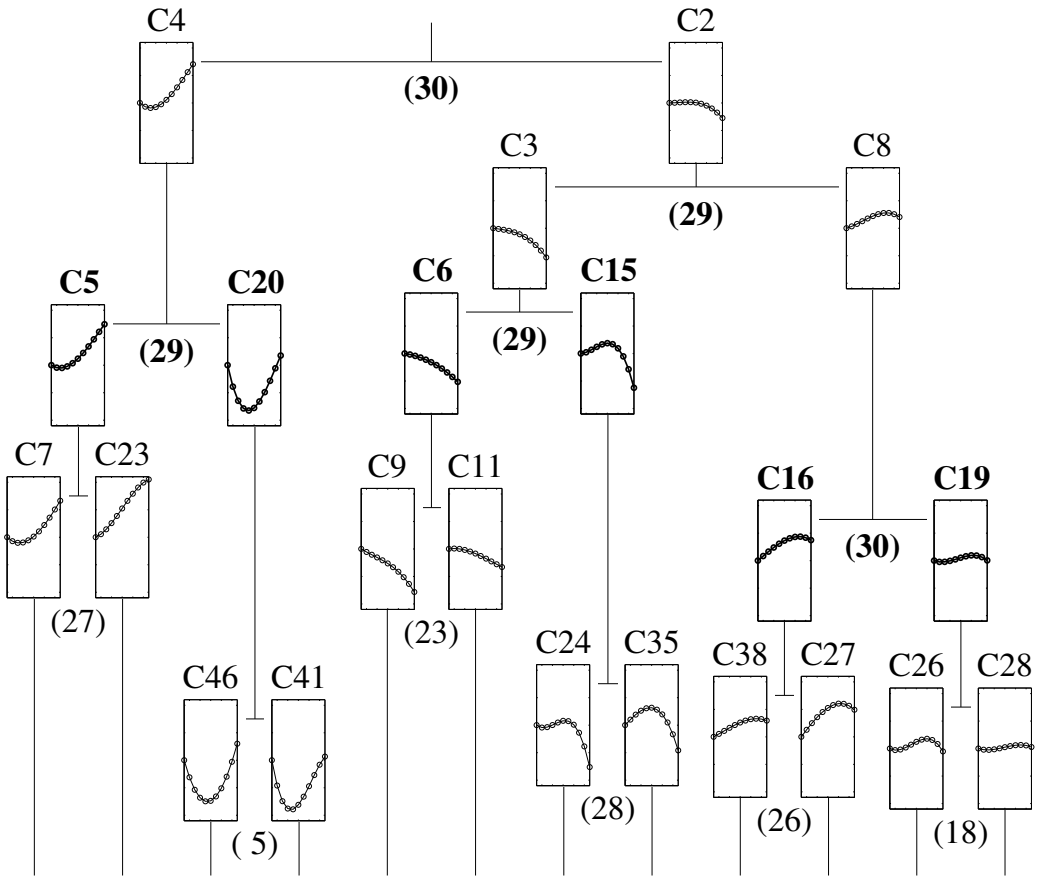
（3）C5 の子クラスタ C7 と C23 の「違う」の回答数 27 は、しきい値 29 に満たないため分割はせず、C5 をテンプレートとして選定する。

（4）C20 の子クラスタ C46 と C41 の「違う」も 29 未満であるため分割せず、C20 をテンプレートとして選定する。

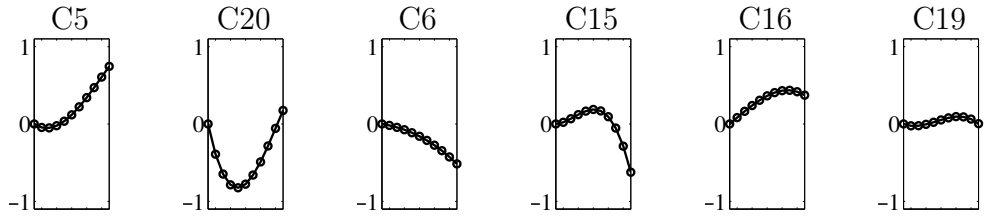
以下同様にして、

（5）C2 は、C3 と C8 に分割する。

（6）C3 は C6 と C15 に、C8 は C16 と C19 に、それぞれ分割する。



(a) クラスタセントロイドによる音調の対比較実験結果（一部）



のぼり音調	くだりのぼり音調	基本音調／くだり音調				[130]
疑問型上昇調		無音調／急下降調	上昇下降調	強調型上昇調	平坦調	[133]

先行研究 [130], [133] との対応

(b) 絞り込まれたテンプレート

図 5.3　聴感上の判定に基づくテンプレートの絞り込み

（7）C9 と C11、C24 と C35、C38 と C27、及び C26 と C28 の対は、「違う」と判定された回答数が 29 未満であるため、それぞれの親クラスタである C6、C15、C16、及び C19 を、テンプレートとして選定する。

以上のようにして絞り込んだ結果、全データを六つに分割するクラスタが選ばれた。図 5.3(b) に、選定したクラスタのセントロイドを示す。これらの形状を、表 1.1 に挙げた分類 [130],[133] と照らし合わせると、「のぼり音調」、「くだりのぼり音調」、「基本音調」、「上昇下降調」、「強調型上昇調」、及び「平坦調」などとの対応が見られる。音調としての差異が顕著なものを選ぶ絞り込みによって、言語学の分野で議論されてきた音調の分類に相当する、具体的な F0 形状を得ることができた。

5.4　むすび

文末 F0 形状の階層的クラスタリングによって得られた多数のクラスタのセントロイドの中から幾つかを、合成音声の文末音調の付与に利用するテンプレートとして選び出すために、聴感上も音調としての違いが感じ取れるかどうかを判断基準に取り入れた絞り込みの手法を考案した。

一つのクラスタの分割により新たに生成された二つのクラスタのセントロイドを F0 形状として付与した音声の対を聴き比べ、聴感上も異なる音調と感じられるかを確認する対比較実験を行った。似通った形状でも聴感上は違う音調であると感じられるものや、見掛け上の差は大きくても音調としての差異がほとんど感じられないものがあり、テンプレートの絞り込みに聴感上の判定を基準として導入することの必要性が示された。聴取実験の結果に基づいて、音調としての差異が顕著なものを選ぶような絞り込みを試したところ、言語学の分野で議論されてきた音調の分類との対応も見られる F0 形状が選定された。

この対比較実験の結果は、文末音調による発話意図の表現に用いるテンプレートを絞り込む際の判断基準として、この後の第 6 章、第 7 章で実際に利用する。

第6章　文末詞とその音調の組合せによる発話意図の伝達

6.1　まえがき

文末音調は、話し手の意図を伝える上で重要な役割を担っている[120]~[133]。合成音声においても、付与する文末音調が適切でないと、聞き手に誤ったメッセージを伝えてしまうことになり兼ねない。ところが、合成音声で意図や態度を表現することを目的とした文末音調の制御に着目した研究は、余り多くは見られない上に、文末詞がもつ機能との関係は論点となっていない[13),114]。音声合成システムで使用する記号を定めた規格[140]では、文末音調を指定する韻律記号として、「通常」と「疑問」の2種類が規定されているだけである。大まかには、「通常」は平叙文の文末での“下降音調”を、「疑問」は疑問文の文末での“上昇音調”を想定したものと推察されるが、それぞれの具体的な実現方法は個々の音声合成システムの設計に委ねられている。ただ、平叙文であれば必ず下降音調になる、というものではない[13),45)]。今後、合成音声が音声対話システムの応答音声として利用される機会が増えてくれば、システムの意図を正しく伝えられる表現力を備えた合成音声が求められることになるはずである。

文末詞の音調が発話全体の意味を支配することを示す実験結果[123]が報告されているが、実際には、文末だけでなく文全体の音調や、更には継続時間長や声質などの音調以外の特徴量も変化している[70]。このことから、どのような意図でも文末音調の制御だけで伝えられるわけではないことは、想像に難くない。とはいえ、文末音調の違いだけでも十分伝えられる意図の表現方法が見いだせれば、それらは音声合成における表現力の向上に有用な知見となるはずである。そこで、文末詞とその音調の組合せに応じて聞き手にどのような意図が伝わるかを明らかにすることで、合成音声によって話し手の意図を確実に伝えることができる表現

手法を確立することを目指す。

言語学における文末音調と意図表現との関係の研究は、「意図を伝えるために話し手はどのような表現を用いているか」を明らかにしようとする立場から、設定した意図に即して発話された音声や、実際に何らかの文脈の中で発話された音声を資料とした分析が中心のアプローチにより進められている[122]。これに対して、本研究の目的は音声合成技術の開発にあり、「伝えようとした意図が聞き手に正しく伝わったか」が重要となる。そのため、様々な文末音調を付与した合成音声の聴取実験に依拠したアプローチを採る。

6.2　文末音調によって伝わる発話意図の聴取実験

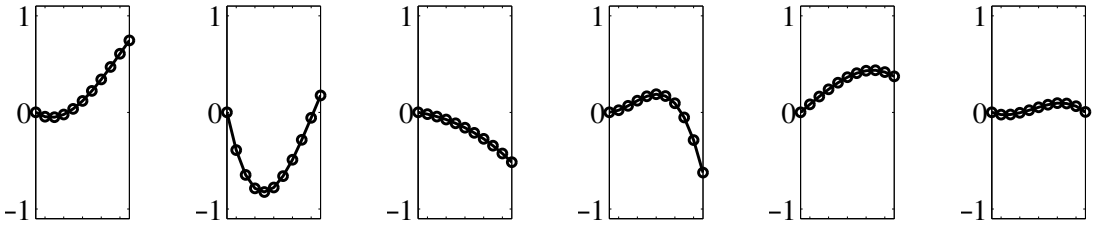
意図を聞き手に確実に伝えるためにはどのような文末詞と音調の組合せが適しているかを、テンプレートによる様々な文末音調を付与した合成音声の聴取実験を通して明らかにする。

6.2.1　発話意図の聴取実験に用いるテンプレートの選定

まずは、どのような文末F0形状を対象として、伝わる発話意図との関係を調査するかを決める必要がある。文末音調の分類は、音声合成システムで使用する記号を定めた規格[140]では2種類、言語学の分野では2〜6種類[124]~[133]が挙げられている。そこで、対象とするテンプレートの数を、言語学における分類と同数程度に絞り込むことにした。

5.3で、音調としての差異が顕著なものを選び出す絞り込みを試みた。その結果、言語学における分類の「疑問型上昇調」や「上昇下降調」などとの対応が見られる、代表的な6種類の形状が選ばれている。そこで、先行研究との整合性も高いこれら6種類を用いて、文末詞と音調の組合せに応じて伝わる発話意図の聴取実験を実施することにした。

選定した6種類のテンプレートを、改めて図 6.1に示す。本論文では以降、これらのセントロイドを参照する際には、それぞれの形状の把握を容易にするため



(a) C5[↗]　(b) C20[∨]　(c) C6[↘]　(d) C15[∧]　(e) C16[→]　(f) C19[−]

図 6.1　発話意図の聴取実験に用いるテンプレート

に、形状から連想される記号をクラスタの識別番号に添えてC5[↗]、C20[∨]、C6[↘]、C15[∧]、C16[→]、C19[−] と表記することにする。

6.2.2　実験方法

意図を伝える発話文には、評価者が音調による意図表現の聴き取りに集中できるよう、「動詞＋文末詞」の形の短い文を用いることにした。文末詞の定義には諸説あるが、終助詞と助動詞 [115] のうち、複数の意味をもち、それらが音調によって表現し分けられる可能性があるものを対象とすることにした。このような文末詞では、伝えたい意図に適した音調を付与しないと違う意図が伝わって、コミュニケーションに齟齬を来すことにもなり兼ねない [120]。そこで、先行する語の品詞などによって意味が決まるものや「男性専用」[115] などのように使い方が限定的なものは除き、終助詞と助動詞、及びそれらの連鎖の中から以下の 13 種類を、調査の対象として選んだ。

（1）“か”

（2）“だって”

（3）“だってね”

（4）“だってよ”

（5）“だろう”

（6）“だろうね”

（7）“な”

（8）“ね”

（9）“の”

（10）“のか”

（11）“よ”

（12）“よな”

（13）“よね”

伝えたい意図は、これらの文末詞によって表現される意図のうち、複数の文末詞に共通する 11 種類とした。言語学の分野でも特に多くの議論がなされている“ね”、“よ”、“よね”[116]~[118] など、実用上も重要と考えられる文末詞と、それらが表す意図を含めるように選定した。

（1）〈依頼〉

（2）〈命令〉

（3）〈非難〉

（4）〈問い返し〉（以下、〈問返〉と表記）

（5）〈伝聞〉

（6）〈質問〉

（7）〈推量〉

（8）〈勧誘〉

（9）〈主張〉

（10）〈確認〉

（11）〈提案〉

有核の動詞「食べる」に上記の文末詞を後続させた 31 種類の発話文の音声を、**2.2.3** で作成した状況 IV の音声合成モデルを用いて合成した。このとき、文末音節については、母音の継続時間長を **4.2** で文末 F0 形状の抽出に用いた音声データにおける平均値（313 ms）に固定した。その上で、F0 形状を図 6.1 の各形状に置き換えて、発話文ごとに文末音調だけが異なる 6 種類の合成音声を用意した。

伝えたい意図ごとに、それぞれが表現される対話の状況を設定し、上述の発話文の中からその状況において発話されることが想定できる文を全て選び出した。各文の 6 種類の合成音声を集めて全体をランダムに並べ、意図ごとの音声刺激とした。評価者 20 名に、意図ごとに対話の状況設定を説明した上で対応する音声刺激を聞かせ、それぞれの発話の文末詞と音調による表現が当該の意図を伝えるのに適しているかどうかを −2（全く別の意図に取れ、その意図は伝わってこない）〜2（その意図が確実に伝わってくる）の 5 段階で評価させた。

6.3　結果と考察

伝えたい意図ごとに、それを聞き手に確実に伝えるにはどのような文末詞と音調の組合せが適しているかを詳しく考察し、音声合成で利用可能な有用な知見としてまとめる。図 6.2〜図 6.6 に、音調の違いに応じて異なる意図が伝わった発話文についての、5 段階評価の平均値を示す。

6.3.1 〈依頼〉,〈命令〉,〈非難〉

〔 1 〕“ね”と音調の組合せ

図 6.2 に、「食べて“ね”」、及び「食べないで“ね”」の結果を示す。

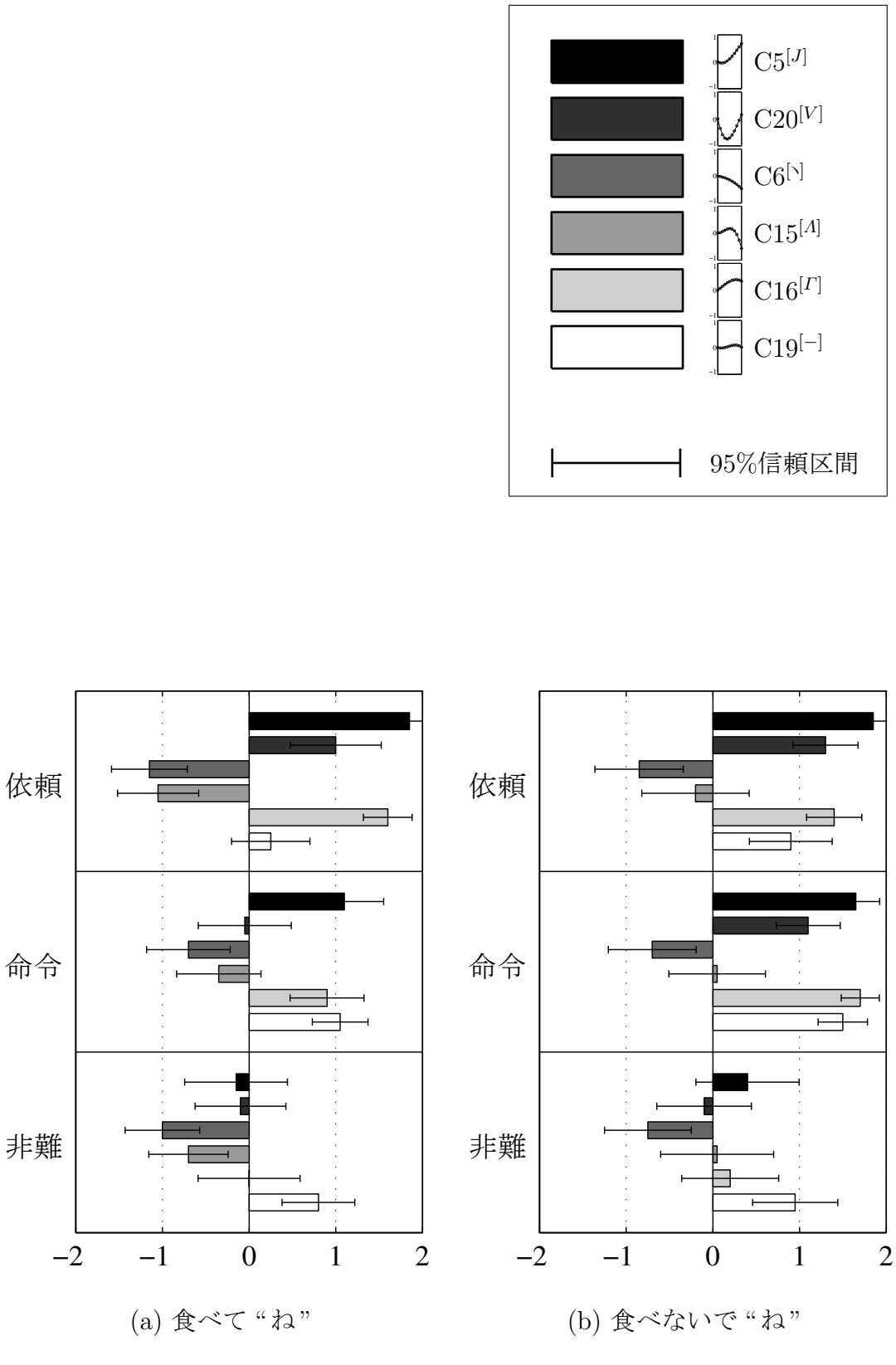
- 相手の同意を求める気持ちを表す“ね”によって、〈依頼〉の意図が伝わることが予想された。「食べて“ね”」(図 6.2(a)) では、上昇調の $C5^{[J]}$ でそれが最も良く伝わっている。

  これと、〈命令〉及び〈非難〉の伝わり方との間で有意差を検定（両側 t 検定）したところ、それぞれに有意差（$p < 0.001$）が認められた。以下、この三つの意図の間の比較においては、有意水準 5 %未満で他の二つの意図のいずれとの間にも有意差が認められた場合に、大きかった方の p 値で有意水準を示す。

- 〈依頼〉は、上昇を伴う $C20^{[V]}$（$p < 0.01$）、$C16^{[\Gamma]}$（$p < 0.05$）でも良く伝わっている。

- 平坦調の $C19^{[-]}$ では〈依頼〉よりも要求の度合いが強いと捉えることができる〈命令〉が伝わっている。

- 否定文「食べないで“ね”」(図 6.2(b)) では、〈命令〉の印象が $C16^{[\Gamma]}$（$p < 0.05$）、$C19^{[-]}$（$p < 0.05$）で強くなっている。



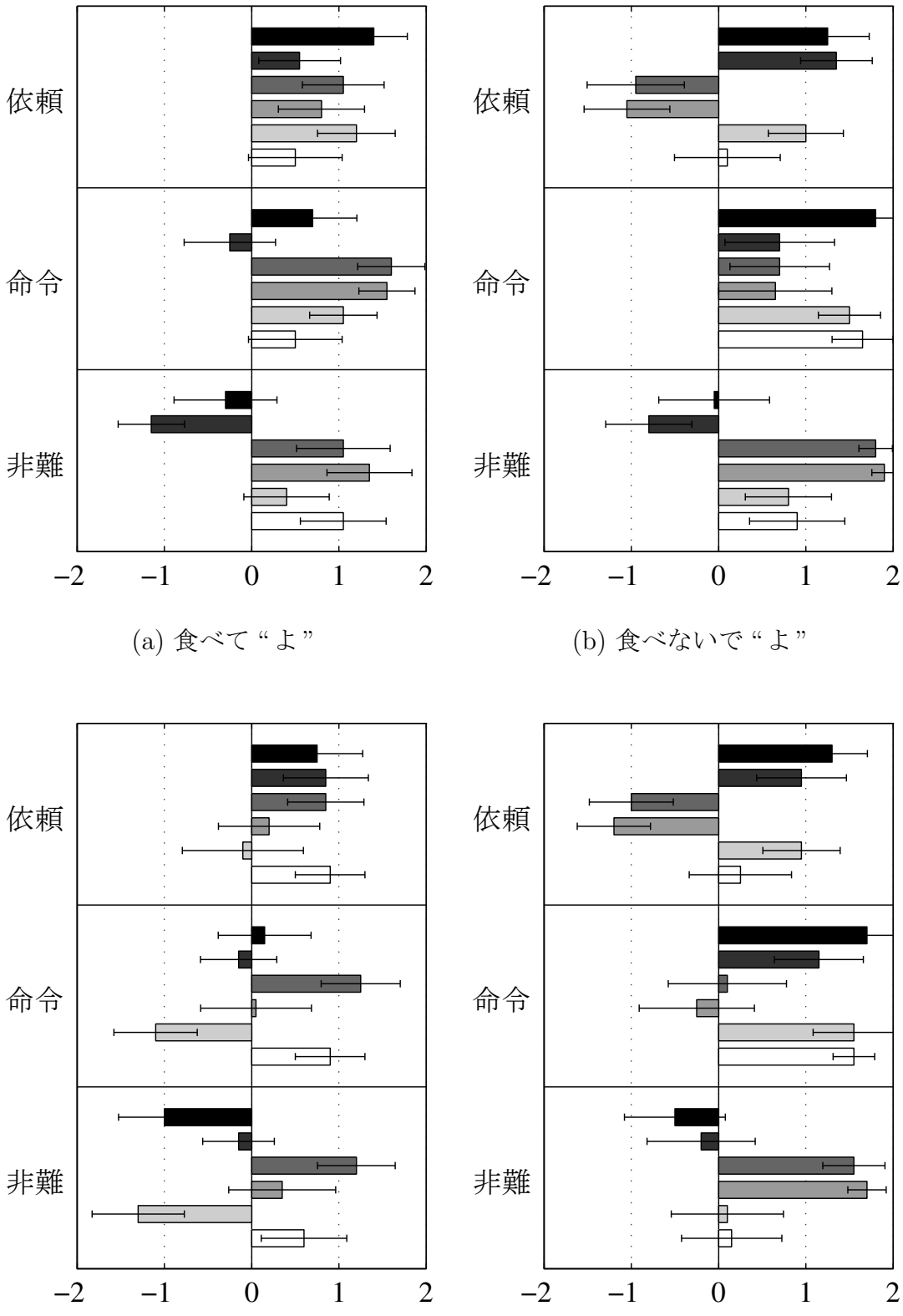
(a) 食べて“ね”　　(b) 食べないで“ね”

図 6.2　文末詞と音調の組合せに応じて伝わる発話意図の聴取実験結果 (1)

〔 2 〕 “よ” と音調の組合せ

図 6.3 に、「食べて “よ”」、「食べないで “よ”」、「食べろ “よ”」、及び「食べるな “よ”」の結果を示す。

- 「食べて “よ”」（図 6.3(a)）では、「食べて “ね”」では現れなかった、下降を伴う C6[ㇾ]、C15[ㅅ] での〈命令〉や〈非難〉の印象が出ている。
- 禁止を表す否定文「食べないで “よ”」（図 6.3(b)）では、「食べないで “ね”」と同じく〈命令〉が C16[ㅣ]（$p < 0.05$）、C19[–]（$p < 0.05$）で伝わる一方で、〈非難〉が C6[ㇾ]（$p < 0.01$）、C15[ㅅ]（$p < 0.001$）ではっきりと伝わっている。この結果は、命令文で “よ” が下降音調を伴うと異議の申し立てを表し、否定命令ではより明確な形で現れるとする考察 [121] を支持している。
- 「食べろ “よ”」、「食べるな “よ”」は命令文に “よ” を付加した形であるが、「食べろ “よ”」（図 6.3(c)）は、上昇を伴う C5[ㅈ]、C20[ㅣ] によって〈命令〉よりも要求の程度が軽い〈依頼〉の印象となっている。
- 否定命令の「食べるな “よ”」（図 6.3(d)）では、〈命令〉が平坦調の C19[–]（$p < 0.001$）で伝わる一方で、〈非難〉がやはり下降を伴う C6[ㇾ]（$p < 0.001$）、C15[ㅅ]（$p < 0.001$）で明確に伝わっている。



(a) 食べて“よ”　(b) 食べないで“よ”

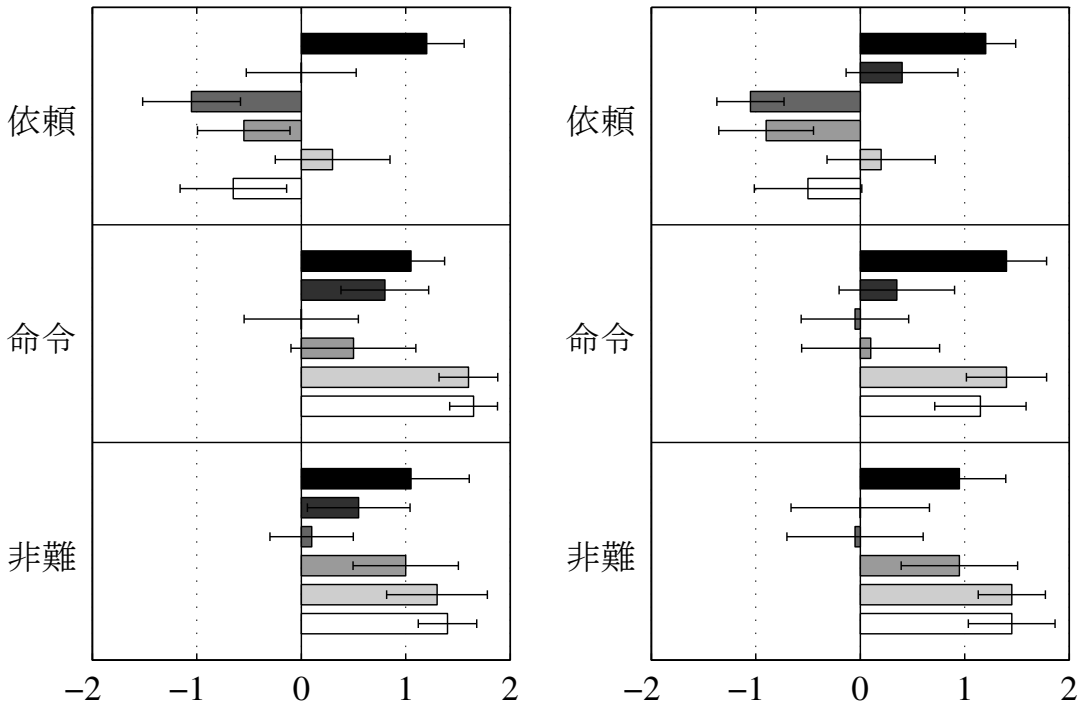
(c) 食べろ“よ”　(d) 食べるな“よ”

図 6.3　文末詞と音調の組合せに応じて伝わる発話意図の聴取実験結果 (2)

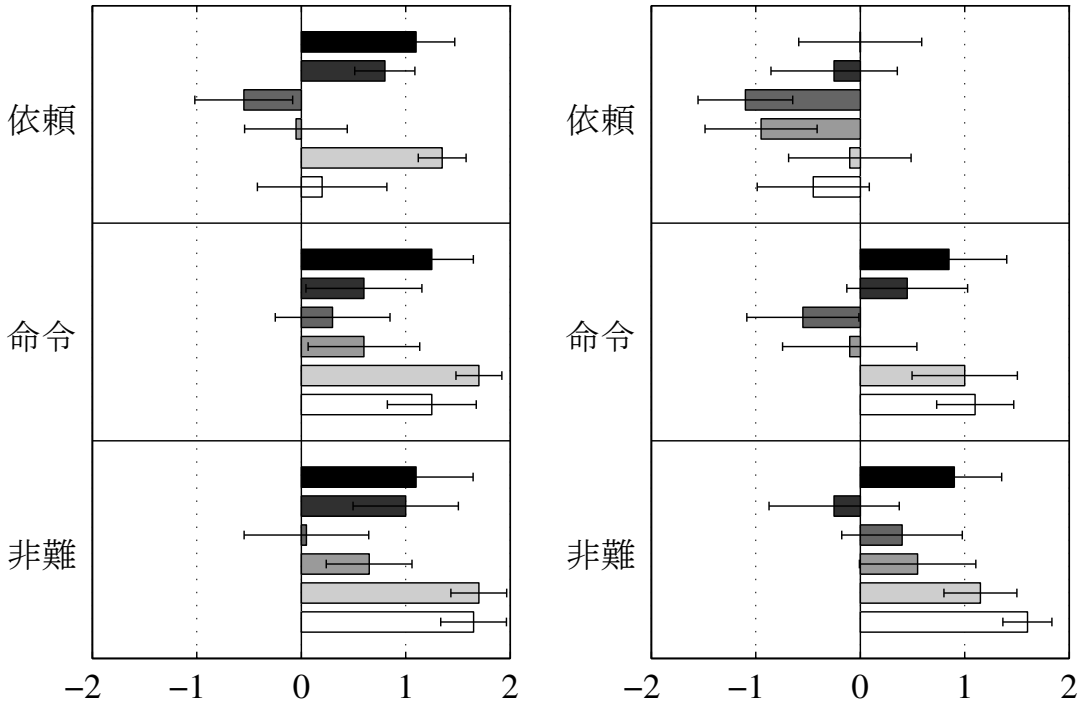
〔3〕“よね”，“よな”と音調の組合せ

図 6.4 に、「食べて“よね”」、「食べないで“よね”」、「食べろ“よな”」、及び「食べるな“よな”」の結果を示す。

- “よね”（図 6.4(a)、図 6.4(b)）では、C16[↗]、C19[–] で〈命令〉や〈非難〉が伝わりやすくなっている。この傾向は、“ね”（図 6.2(a)、図 6.2(b)）よりも強く出ているだけでなく、〈命令〉や〈非難〉が主に下降を伴う C6[↘]、C15[↗] で伝わる“よ”（図 6.3(a)、図 6.3(b)）とも異なっている。
- “よな”（図 6.4(c)、図 6.4(d)）の C16[↗]、C19[–] で〈命令〉や〈非難〉が伝わりやすくなっている点は、“よね”と共通している。
- 「食べるな“よな”」（図 6.4(d)）では、どの音調でも〈依頼〉は伝わっていない。これは、否定命令文に更に念を押す意を表す文末詞を付加することで、強い禁止の意図が言語で表現されていることに起因しているものと思われる。



(a) 食べて“よね”　　(b) 食べないで“よね”



(c) 食べろ“よな”　　(d) 食べるな“よな”

図 6.4　文末詞と音調の組合せに応じて伝わる発話意図の聴取実験結果 (3)

6.3.2 〈問返〉,〈伝聞〉

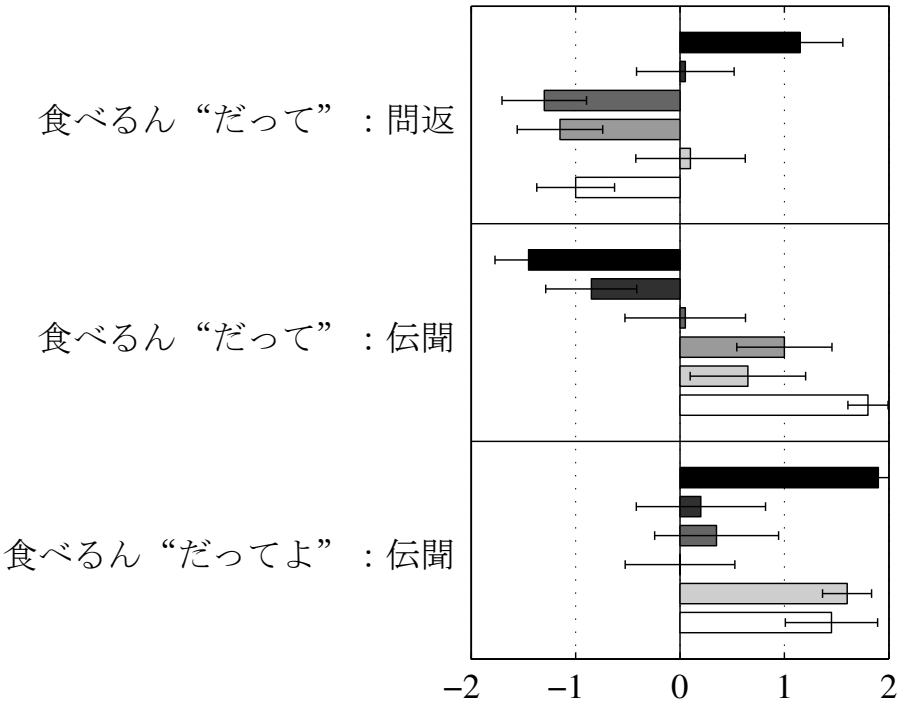
図 6.5(a) に、「食べるん “だって”」、及び「食べるん “だってよ”」の結果を示す。

- 「食べるん “だって”」では、〈問返〉が上昇調の $C5^{[J]}$（$p < 0.001$）で、〈伝聞〉が下降を伴う $C15^{[A]}$（$p < 0.001$）、平坦調の $C19^{[-]}$（$p < 0.001$）で、それぞれ伝わっている。
- 「食べるん “だってよ”」では、$C19^{[-]}$ に加え、上昇調の $C5^{[J]}$ や $C16^{[I]}$ も〈伝聞〉を表現するのにふさわしいとされた。

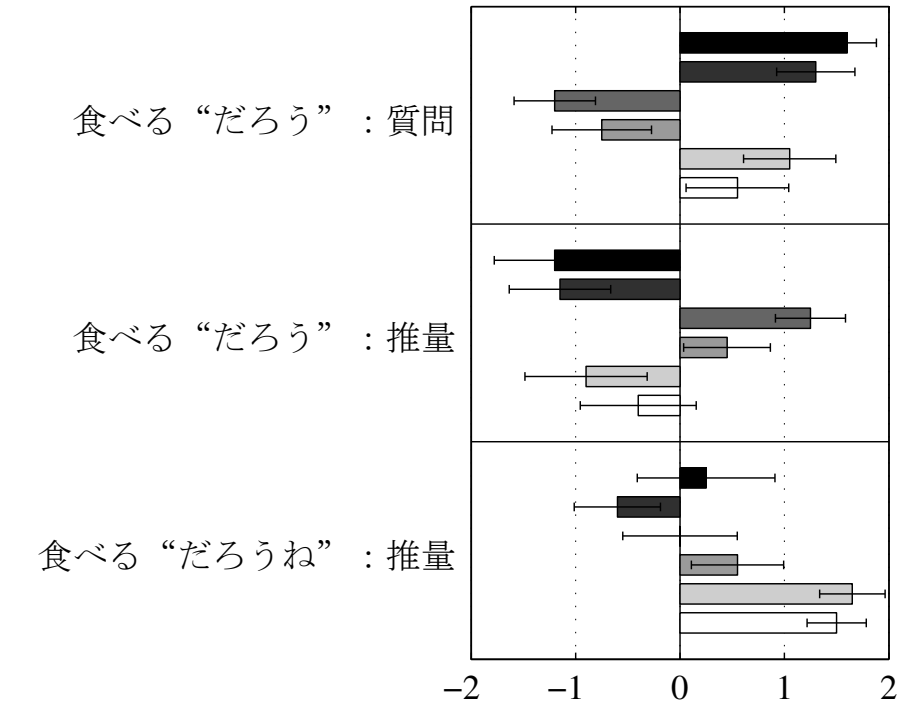
6.3.3 〈質問〉,〈推量〉

図 6.5(b) に、「食べる “だろう”」、及び「食べる “だろうね”」の結果を示す。

- 「食べる “だろう”」では、〈質問〉には上昇を伴う $C5^{[J]}$、$C20^{[V]}$、$C16^{[I]}$（いずれも $p < 0.001$）が、〈推量〉には下降調の $C6^{[\backslash]}$（$p < 0.001$）が適している。
- “ね” を付加した「食べる “だろうね”」で〈推量〉を伝えるには、上昇調の $C16^{[I]}$、平坦調の $C19^{[-]}$ が適しているとされた。



(a) 食べるん{“だって”，“だってよ”}



(b) 食べる{“だろう”，“だろうね”}

図 6.5　文末詞と音調の組合せに応じて伝わる発話意図の聴取実験結果 (4)

6.3.4 〈勧誘〉

図 6.6(a) に、〈勧誘〉を表現する 3 種類の発話文「食べよう“よ”」、「食べよう“か”」、及び「食べない“か”」の結果を示す。

- 「食べよう“よ”」では下降を伴う C6[ヽ]、C15[ㇾ] が適しているとされたが、「食べよう“か”」では C6[ヽ] のみであった。
- 「食べない“か”」では、上昇を伴う C5[ノ]、C20[V] が適しているとされた。上昇調により相手の意向を伺う意図が表現され、〈勧誘〉が伝わったものと考えられる。

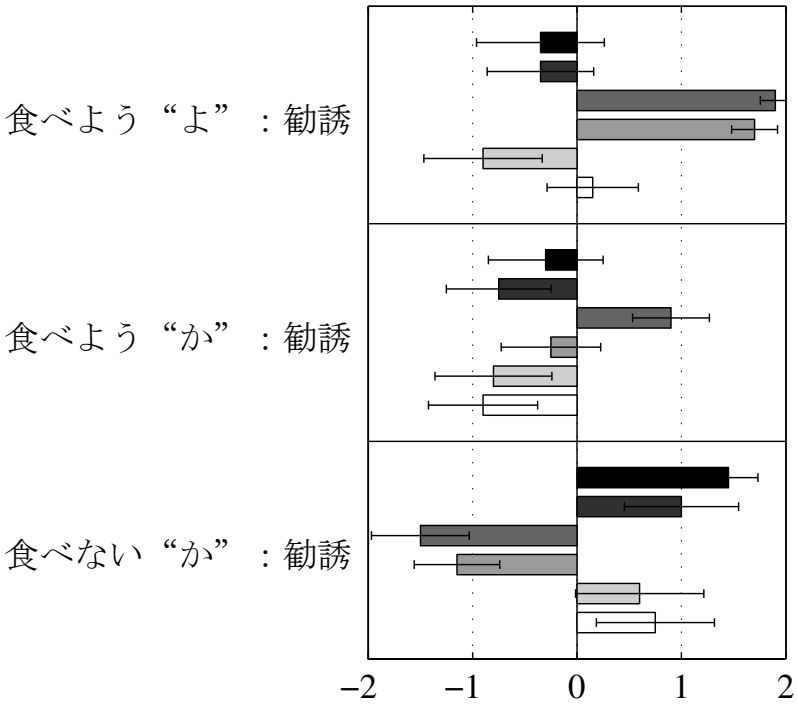
6.3.5 〈問返〉,〈主張〉

図 6.6(b) に、「食べる“の”」の〈問返〉と〈主張〉、及び「食べる“よ”」の〈主張〉の結果を示す。

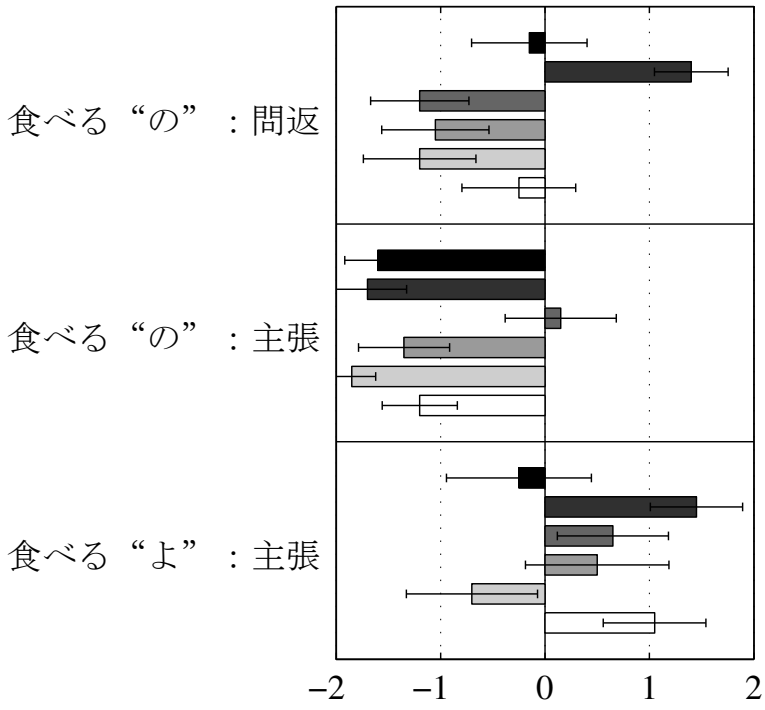
- 「食べる“の”」では、〈問返〉が C20[V]（$p < 0.001$）で伝わっており、くだりのぼり音調が聞き手の言明に対する不信の表現になるとする考察 [130] と整合する。
- 〈主張〉については、「食べる“の”」ではどの音調でも伝わらなかった。「食べる“よ”」では C20[V] や C19[−] で伝わっているが、〈確認〉や〈提案〉との間に有意差は認められなかった。

  合成音声を聴取したところ、「食べる“の”」では、“の”の母音の継続時間長が〈主張〉を表現するには長過ぎると感じられた。試みに、これを短く（100 ms）した合成音声を作成して聴取したところ、〈主張〉らしさが感じられるようになった。「食べる“よ”」においても、下降調の C6[ヽ] などで〈主張〉らしさが増す印象を受けた。

文末音調の違いだけでは伝えられなかった意図については、他の特徴量の制御と併せた表現方法の検討が必要である。



(a) 食べよう｛“よ”，“か”｝，食べない“か”



(b) 食べる｛“の”，“よ”｝

図 6.6　文末詞と音調の組合せに応じて伝わる発話意図の聴取実験結果 (5)

表 6.1　文末詞と音調の組合せによって伝えることができる発話意図

意図	発話文	文末詞の音調（文末 F0 形状） C5[J]	C20[V]	C6[ヽ]	C15[Λ]	C16[Γ]	C19[−]
依頼/命令/非難	食べて“ね”	**依頼*****	**依頼****	——	——	**依頼***	**命令**
	食べないで“ね”	**依頼**	**依頼**	——	命令/非難	**命令***	**命令***
	食べて“よ”	**依頼****	依頼**	**命令**	**命令**	**依頼**	**非難**
	食べないで“よ”	**命令**	**依頼***	**非難****	**非難*****	**命令***	**命令***
	食べろ“よ”	依頼	依頼**	**命令**	非難	——	依頼/命令
	食べるな“よ”	**命令**	**命令**	**非難*****	**非難*****	**命令**	**命令*****
	食べて“よね”	**依頼**	命令	非難	**非難**	**命令**	**命令**
	食べないで“よね”	**命令**	依頼	——	非難**	**非難**	**非難**
	食べろ“よな”	**命令**	**非難**	命令	非難	**命令/非難**	**非難**
	食べるな“よな”	非難	命令*	非難**	非難	**非難**	**非難***
問返/伝聞	食べるん“だって”	**問返*****	問返*	伝聞***	**伝聞*****	伝聞	**伝聞*****
質問/推量	食べる“だろう”	**質問*****	**質問*****	**推量*****	推量***	**質問*****	質問**
問返/主張	食べる“の”	——	**問返*****	主張***	——	——	——

各発話文と音調の組合せにおいて 5 段階評価の平均値が最も高い正の値となった意図。このうち、太字は平均値が 1.0（その意図がほぼ伝わってくる）以上だったもの。***、**、* は、他の意図（〈依頼〉、〈命令〉及び〈非難〉の比較においては p 値が大きかった方の意図）との間に、それぞれ有意水準 0.1%、1%、5% で有意差が認められたもの。—— は平均値が正の値となった意図がなかったことを表す。

6.4　文末詞と音調の組合せによる発話意図の表現手法

以上の結果に基づいて、「文末詞を伴った発話文によって所望の意図を伝えるには、どのような文末音調が適しているか」を整理した。文末音調に応じて異なる意図が伝わった発話文について、5 段階評価の平均値が最も高くなった意図を、表 6.1 に示す。

音声を合成するに当たって、文末詞を伴った発話文とともに、それによって伝えたい意図が指定された場合は、表 6.1 を参照し、その文末詞と伝えたい意図との組合せに適した文末 F0 形状のテンプレートを選択する。その形状を文末詞の音調として付与することによって、話し手の意図を確実に伝えることができる音声の合成が可能となる。幾つかの具体的な例を挙げて、説明する。

- 「食べて“ね”」といって〈依頼〉の意図を伝えたいときは、“ね”の音調として C5$^{[J]}$、C20$^{[V]}$、または C16$^{[\Gamma]}$ を付与することが可能である。いずれも 5 段階評価の値が高く、〈命令〉及び〈非難〉の伝わりやすさとの間に有意差が認められるので、確実に伝わることが期待できる。どれを使うかは任意であるので、適宜使い分ければ、常に同じ話し方になってしまうことも避けられる。

- 「食べないで“よ”」で〈依頼〉を伝えるには、C20$^{[V]}$ を用いる。C16$^{[\Gamma]}$ や C19$^{[-]}$ では〈命令〉の印象が、C6$^{[\backslash]}$ や C15$^{[\Lambda]}$ では〈非難〉の印象が伝わるため、最適な音調を選択することが必要となる。

- 「食べる“だろう”」と〈推量〉するには、C6$^{[\backslash]}$ か C15$^{[\Lambda]}$ を付与すればよい。ただし、C6$^{[\backslash]}$ の方が 5 段階評価の値が高いのでより確実である。

6.5　むすび

文末詞を伴った発話文と伝えたい意図の組合せに適した文末音調を、テンプレートの中から選択して付与する発話意図の表現手法を提案した。

第 5 章での結果を利用して、音調としての差異が顕著であることを基準として、代表的な 6 種類のテンプレートを選んだ。それらを種々の文末詞の音調として付

与した合成音声の聴取実験の結果から、文末詞と伝えたい意図の組合せに応じて、文末に付与する音調としての適不適があることを例証した。

この結果は、適切な文末音調を付与しないと、聞き手に誤ったメッセージを伝えてしまうことになり兼ねないことも意味している。音声対話システムの応答音声として利用する上では、文末音調の緻密な制御が必要であることが示された。

図 5.1 に示した多彩な文末音調の中には、主たる意図とは別に、何らかの微妙なニュアンスの違いが感じられるものが幾つもあった。第 7 章では、使用するテンプレートの種類を増やして、文末音調によって微妙なニュアンスの違いまでも伝えることができるかを探る。

第7章　文末詞とその音調の組合せによる付加的ニュアンスの伝達

7.1　まえがき

人同士の間では、言葉で表現された以上に深いコミュニケーションが図られている。食べようと思っていた好物を食べられてしまって「勝手に食べるなよ」と相手を〈非難〉するときの音声表現が、腹を立てているのか、がっかりしているのかによって違ったものになるであろうことは、容易に想像できる。話し手は、言語では表現していない心情を音声表現によって伝える。音声対話システムが今後更に普及すれば、このような言外の意図ともいうべき微妙なニュアンスの違いをも伝えられる表現力が、合成音声にも求められることになるだろう。

実際の発話では、話し手は文末だけでなく文全体の音調のほか、音素の継続時間長や振幅などの様々な音響特徴量を変化させて意図を表現している[70]。文末音調の形状は無限に変容するもので、細々としたニュアンスを詳述することは不可能である、ともいわれてきた[124]。文末音調だけを変化させた発話から、聞き手が異なるニュアンスを感じ取ることができるかは、必ずしも自明ではない。意図に付加されるニュアンスにはどのようなものが考えられるか、それらはどのように表現されるか、といった踏み込んだ議論もほとんどされていない。それだけに、文末音調の違いだけで異なるニュアンスが伝わる表現が幾つかでも見つかれば、それらは対話音声の合成に有用な新たな知見となり得る。また、文末音調が違うだけでも伝わるニュアンスが変わることがあるとなると、それは裏を返せば、適切な文末音調を付与しないと聞き手に誤解を与えてしまう危険性があることを意味する。

第 4 章でのクラスタリングで得られた様々なテンプレートによって文末音調を付与した合成音声を聴取してみると、話し手の心情や話し手と聞き手の関係など

が反映されているような、微妙なニュアンスの違いが感じられるものが幾つもあった。そこで、聞き手は、様々なテンプレートによる文末音調を付与した合成音声から微妙なニュアンスを感じ取るかを調査し、文末音調による意図表現の更なる可能性を探る。

7.2　文末音調によって伝わるニュアンスの聴取実験

第 6 章で実施した発話意図の聴取実験の結果を起点として、文末音調を付与するテンプレートの種類を増やし、主たる意図に付加されるニュアンスの聴取実験を行う。

7.2.1　主たる意図を伝える表現の選定

まず、第 6 章で扱った意図の中から、何らかのニュアンスが付加される可能性があると考えられる〈依頼〉、〈命令〉、及び〈非難〉の三つの意図を、主たる意図として選んだ。パラ言語情報の伝達では、韻律的特徴の調節によって意図や態度の離散的な区分の違いが表現されるだけでなく、同じ区分内での連続的な程度の違いも表現される [69]。話し手の要求度合いが〈依頼〉から〈命令〉、〈非難〉へと連続的に強くなっていくと捉えられることや、要求を発する状況にも様々あると考えられることから、これらの表現の中で何らかのニュアンスの違いが生まれることを期待した。

次に、〈依頼〉、〈命令〉、及び〈非難〉を伝えるのに適した発話文と文末音調の組合せを、表 6.1 の中から選んだ。当該意図の伝わりやすさと他の二つの意図の伝わりやすさとの間に有意差が認められたものを中心に、表 7.1 に示す組合せを選び出した。

7.2.2　ニュアンスの聴取実験に用いるテンプレートの選定

聴取実験に使用する文末音調のバリエーションを増やす方法として、6 種類の文末音調のセントロイドを与えるクラスタを親クラスタとして分割を更に進めて、複数の子クラスタに展開することを考えた。その上で、表 7.1 の組合せにおいて、

表 7.1　〈依頼〉・〈命令〉・〈非難〉を伝える発話文と文末音調の組合せ

発話文	文末音調					
	C5[J] 	C20[V] 	C6[ヽ] 	C15[ㅅ] 	C16[ㄱ] 	C19[ㅡ] 
食べて“ね”	依頼	依頼			依頼	
食べて“よ”	依頼	依頼			依頼	
食べないで“ね”					命令	命令
食べないで“よ”			非難	非難	命令	命令
食べるな“よ”			非難	非難	命令	命令
食べるな“よな”					非難	非難

展開した子クラスタのセントロイドを元のテンプレートの代わりに使用することで、より多くの種類の文末音調を付与した合成音声を用意することにした。

具体的には、**5.3** でテンプレートを 6 種類に絞り込んだ際の条件を緩和し、「違う」の回答数が 30 回答中 27 以上であれば分割を進めることにした。図 5.1 に示した対比較実験結果において、上記の条件を満たしている回答数を太字で表示したものを、図 7.1 に示す。ここでの目的は多彩なバリエーションを選び出すことにあるので、絞り込みの条件を更に緩和して、30 回答中 27 未満のクラスタ対に到達しても、それよりも葉ノードに近いクラスタで条件を満たすクラスタ対が見つかれば、そこまで分割を進めることにした。例えば、C16[ㄱ] を分割した C38 と C27 の対（「違う」の回答が 30 回答中 26）は、分割の条件を満たしていない。しかし、それぞれを更に葉ノードに向かってたどっていくと、C96 を分割した C346 と C248 の対（同 27）、及び C60 を分割した C131 と C104 の対（同 30）が、それぞれ条件を満たしている。これらよりも葉ノードに近い側に条件を満たす対はないため、最終的に C16[ㄱ] は、C123(ㄱ)、C346(ㄱ)、C248(ㄱ)、C71(ㄱ)、C131(ㄱ)、C104(ㄱ)、C52(ㄱ) の 7 種類の子クラスタに分割されることになる。これらを、表 7.1 の組合せにおいて、C16[ㄱ] に代わるテンプレートとして使用する。

以上のようにして、新たなテンプレートとして選ばれた子クラスタを、表 7.2 に整理する。子クラスタの識別番号には、親クラスタのセントロイドの概形を表

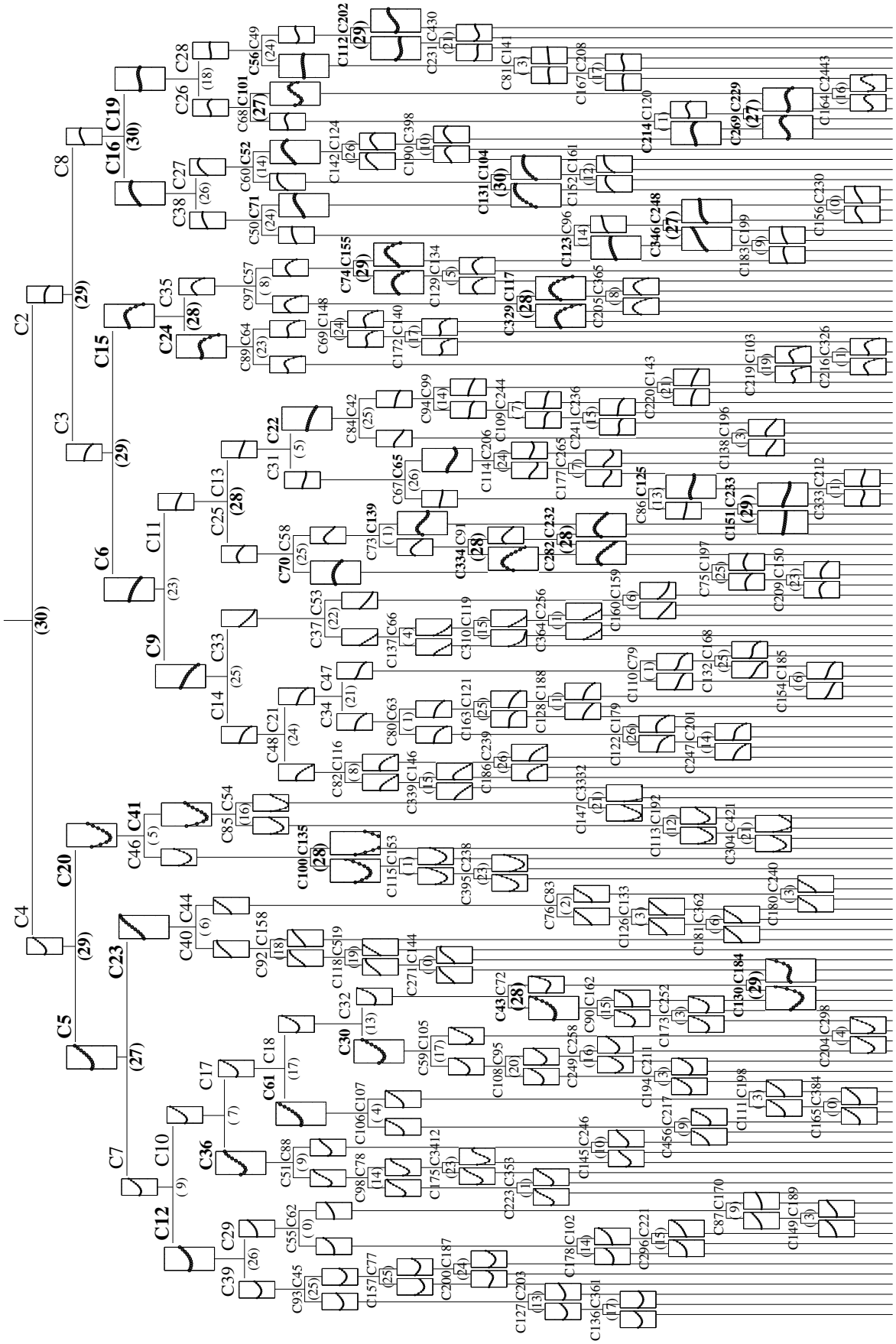


図 7.1　ニュアンスの聴取実験に用いるテンプレートの絞り込み

表 7.2　ニュアンスの聴取実験に用いるテンプレートに選定した子クラスタ

親クラスタ	子クラスタ (1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
C5$^{[J]}$	C12$^{(J)}$	C36$^{(J)}$	C61$^{(J)}$	C30$^{(J)}$	C43$^{(J)}$	C130$^{(J)}$	C184$^{(J)}$	C23$^{(J)}$			
											
C20$^{[V]}$	C100$^{(V)}$	C135$^{(V)}$	C41$^{(V)}$								
											
C6$^{[\backslash]}$	C9$^{(\backslash)}$	C70$^{(\backslash)}$	C334$^{(\backslash)}$	C282$^{(\backslash)}$	C232$^{(\backslash)}$	C139$^{(\backslash)}$	C151$^{(\backslash)}$	C233$^{(\backslash)}$	C125$^{(\backslash)}$	C65$^{(\backslash)}$	C22$^{(\backslash)}$
											
C15$^{[\Lambda]}$	C24$^{(\Lambda)}$	C329$^{(\Lambda)}$	C117$^{(\Lambda)}$	C74$^{(\Lambda)}$	C155$^{(\Lambda)}$						
											
C16$^{[\Gamma]}$	C123$^{(\Gamma)}$	C346$^{(\Gamma)}$	C248$^{(\Gamma)}$	C71$^{(\Gamma)}$	C131$^{(\Gamma)}$	C104$^{(\Gamma)}$	C52$^{(\Gamma)}$				
											
C19$^{[-]}$	C214$^{(-)}$	C269$^{(-)}$	C229$^{(-)}$	C101$^{(-)}$	C56$^{(-)}$	C112$^{(-)}$	C202$^{(-)}$				
											

す記号を添えてある。図 7.1 では、テンプレートとして選定された子クラスタの識別番号を太字で表示するなどしてある。

7.2.3　実験方法

まず、表 7.1 に示した発話文の音声を、**6.2.2** と同様にして合成した。文末音節の母音の継続時間長も同じく、**4.2** で用いた音声データにおける平均値（313 ms）に固定した。その上で、各音声の文末 F0 形状を、表 7.1 の組合せにおける 6 種類の文末音調のそれぞれに対応する子クラスタのセントロイド（表 7.2）で置き換えた。例えば、〈依頼〉の意図を伝える「食べて“ね”」としては、C5[J]、C20[V]、及び C16[I] のそれぞれの子クラスタのセントロイド（8 種類、3 種類、及び 7 種類）を文末音調として付与した計 18 種類の音声刺激を作成した。〈非難〉の意図の「食べるな“よ”」には、C6[ɔ] と C15[ʌ] の子クラスタのセントロイド（11 種類と 5 種類）による文末音調を付与した計 16 種類を用意した。

次に、予備検討として、作成した全ての音声刺激を筆者が聴取し、主たる意図ごとに共通して感じられたニュアンスを 3 種類ずつ挙げて、表 7.3 に示す評価項目を作成した。評価項目は、系統的に網羅された、汎用性の高いものであることが望ましい。しかし、現状では付加的なニュアンスとしてどのようなものがあり得るのかも明らかになっているわけではないため、予備検討を実施して、聞き手に伝わる可能性のあるニュアンスの候補を用意することにした。結果的に、一部は主たる意図に強く依存したニュアンスとなり、汎用性の高いニュアンスを定義することには至らなかった。なお、(r1) と (r2)、(o1) と (o2) は、それぞれ相反する心情を表すものとなっている。ただ、文末音調が異なるだけでも違ったニュアンスが伝わるかを確認する、という目的は、これらの評価項目でも十分達成できると判断した。

評価者 20 名に、各発話文により主たる意図が表現される対話場面の設定と、表 7.3 の評価項目とを記した回答用紙を配付した。更に、評価者がこれらとは別のニュアンスを感じた場合は、随時それを評価項目に追加するよう指示した。音声刺激は自由に繰り返し聞くことを許し、評価者自身で追加した項目も含めて、それぞれの評価項目のニュアンスを感じるか否かを、項目ごとに独立に、4 段階

表 7.3　主たる意図ごとの付加的なニュアンスの評価項目

主たる意図	付加的なニュアンス 「話し手は…」
依頼	(r1) 熱心に勧めてくれているらしい
	(r2) 心から言っているとは思えない
	(r3) 不機嫌そうだ
命令	(o1) 本当に食べてしまうかもしれないと思っているらしい
	(o2) 本当に食べてしまうとまでは思っていないらしい
	(o3) 少し怒っているらしい
非難	(b1) 慌ててやめさせようとしているらしい
	(b2) 食べられてしまってひどく怒っているらしい
	(b3) 食べられてしまってガッカリしているらしい

(0：感じない、1：やや感じる、2：感じる、3：強く感じる）で評価させた。

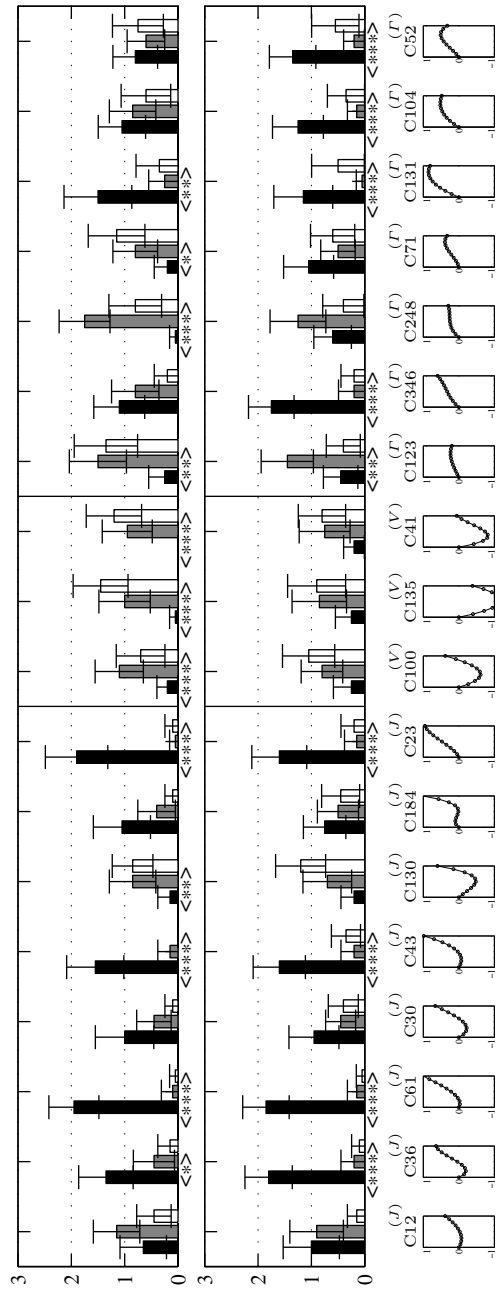
なお、(r1) と (r2)、(o1) と (o2) はそれぞれに相反する心情と捉えることができるので、対を成すニュアンスを一つの軸の両端に配した評価尺度とするのが一般的とも考えられる。しかし、異なる評価尺度の混在は評価者を混乱させることにつながるため、今回はそれを避けることを優先し、全ての項目を同一の尺度で評価させることにした。

7.3　結果と考察

聴取実験の結果から、聞き手は文末音調の違いに応じて異なるニュアンスを感じ取ったかを確認する。特定のニュアンスを伝える文末音調としての利用が期待できるテンプレートを中心に、ニュアンスの違いをもたらす形状の特徴に注目し、先行研究で指摘されている文末音調がもつ機能とも照らし合わせながら考察する。

7.3.1　〈依頼〉に付加されるニュアンス

$C5^{[J]}$、$C20^{[V]}$、及び$C16^{[I]}$の子クラスタのセントロイドによる文末音調を付与した「食べて“ね”」と「食べて“よ”」の合成音声の聴取実験の結果として、20



(a) 食べて“ね”（上段）／(b) 食べて“よ”（下段）

図 7.2　主たる意図〈依頼〉に付加されるニュアンスの聴取実験結果

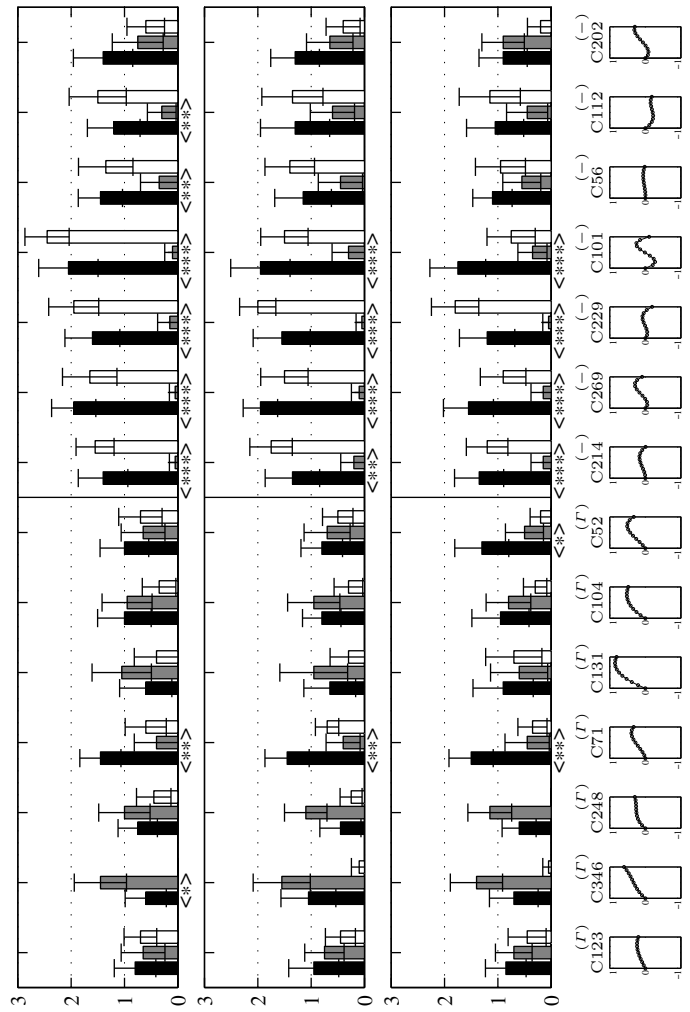
名の評価者による評価の平均値（以下、「評価値」と記す）を図 7.2 に示す。(r1) と (r2) は、相反する心情を表しており、排他的な関係にある。このため、たとえどちらかのニュアンスの評価値が十分に高いテンプレートであっても、相反する心情の伝わりやすさとの間に有意な差が認められなければ、そのニュアンスを表現する文末音調として利用することはできない。そこで、テンプレートごとに、(r1) と (r2) の伝わりやすさの間の有意差の有無の確認も行った（両側 t 検定）。有意差が認められたものについては、有意水準を併せて表示してある。

- (r1) は、どちらの発話文でも、終端の高い値に向かって大幅に上昇する $\mathrm{C61}^{(J)}$ や $\mathrm{C43}^{(J)}$、$\mathrm{C23}^{(J)}$ などで伝わっており、それぞれの音調での (r2) の伝わりやすさとの間に有意差が認められた。更に、「食べて“ね”」では $\mathrm{C131}^{(\Gamma)}$、「食べて“よ”」では $\mathrm{C36}^{(J)}$ や $\mathrm{C346}^{(\Gamma)}$ などの、やはり大きく上昇する音調で伝わっている。

- これに対して (r2) は、「食べて“ね”」において、わずかな上昇を伴う $\mathrm{C123}^{(\Gamma)}$、及び $\mathrm{C248}^{(\Gamma)}$ で伝わっており、(r1) の伝わりやすさとの間に有意差が認められた。他に、「食べて“よ”」の $\mathrm{C123}^{(\Gamma)}$ でも伝わっている。(r1) と (r2) の伝わり方の比較から、聞き手は、上昇の程度の違いに《熱心さ》の程度の違い、あるいはその有無を感じ取っていると推察される。

- (r3) は余りはっきりとは伝わっていないが、「食べて“ね”」の $\mathrm{C135}^{(V)}$ などや、「食べて“よ”」の $\mathrm{C130}^{(J)}$ などでは、やや感じ取られている。言明に対する不信の表現となるくだりのぼり音調 [130] によって、《不機嫌》そうな印象が伝わったものと推察される。

- この他、自由記述により追加された評価項目の中に、表現はそれぞれで異なるものの《親近感》と解釈できるものが「食べて“ね”」の複数の評価者の回答で見られ、$\mathrm{C23}^{(J)}$ や $\mathrm{C248}^{(\Gamma)}$、$\mathrm{C131}^{(\Gamma)}$ などで感じるとされていた。これらは、のぼり音調の機能は聞き手に対する好意的な態度の積極的な表明にある [130] とする指摘に通じるものと考えられる。



< • >　(o1), (o2) 間の有意差　(*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$)

(a) 食べないで“ね”（上段）／(b) 食べないで“よ”（中段）／(c) 食べるな“よ”（下段)

図 7.3　主たる意図〈命令〉に付加されるニュアンスの聴取実験結果

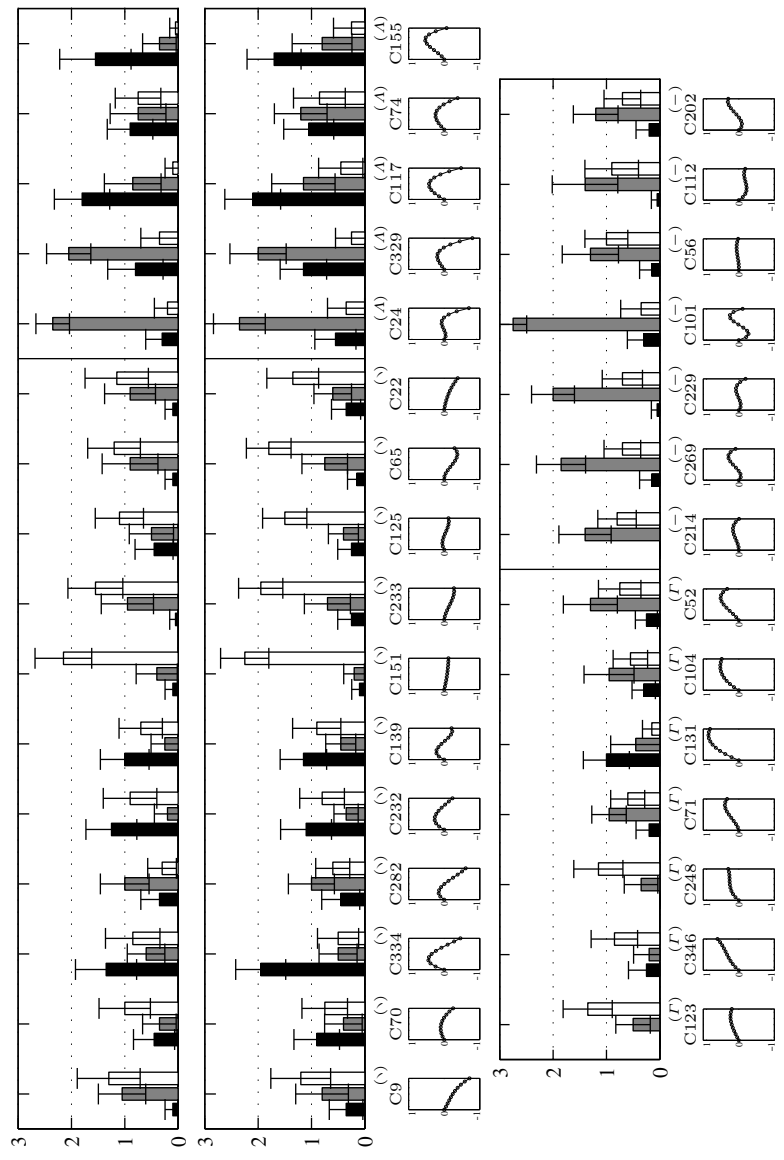
7.3.2 〈命令〉に付加されるニュアンス

C16$^{[\Gamma]}$ と C19$^{[-]}$ の子クラスタのセントロイドによる文末音調を付与した「食べないで“ね”」、「食べないで“よ”」、及び「食べるな“よ”」の、それぞれの評価値を図 7.3 に示す。(o1) と (o2) は、相反する心情を表しており、排他的な関係にある。両者の伝わりやすさの間に有意差が認められたものについては、有意水準を表示してある。

- (o1) は、上下に大きくうねる特徴的な動きのある C269$^{(-)}$ や C229$^{(-)}$、C101$^{(-)}$ で伝わっており、いずれも (o2) の伝わりやすさとの間に有意差が認められた。C269$^{(-)}$ と比べるとうねりは小さいものの似た形状の C71$^{(\Gamma)}$ や C214$^{(-)}$ でも、(o1) が感じ取られている。

- 一方、(o2) が感じ取られ、(o1) の伝わりやすさとの間に有意差が認められたものは、「食べないで“ね”」における C346$^{(\Gamma)}$ のみであった。C346$^{(\Gamma)}$ と C71$^{(\Gamma)}$ は、共に C16$^{[\Gamma]}$ の子クラスタであるが、相反するニュアンスが伝わっている。全体的に緩やかに上昇する似通った形状をしているが、終端付近でそのまま上昇するか下降するかのわずかな違いが見られる。

- C229$^{(-)}$ では、(o1) に加えて (o3) のニュアンスも伝わっている。(o3) は、「食べないで“ね”」と「食べないで“よ”」では、C214$^{(-)}$ や C269$^{(-)}$、C101$^{(-)}$ でも伝わっている。一方で、C269$^{(-)}$ に似ている C71$^{(\Gamma)}$ では (o3) の印象は弱く、全体的に上昇する傾向にある C346$^{(\Gamma)}$ や C248$^{(\Gamma)}$ などでもほぼ伝わっていない。



(a) 食べないで“よ”（上段）／(b) 食べるな“よ”（中段）／(c) 食べるな“よな”（下段）

図 7.4　主たる意図〈非難〉に付加されるニュアンスの聴取実験結果

7.3.3 〈非難〉に付加されるニュアンス

C6[ヽ]、C15[Λ] の子クラスタのセントロイドによる文末音調を付与した「食べないで“よ”」と「食べるな“よ”」、及びC16[Γ]、C19[−] の子クラスタのセントロイドによる文末音調を付与した「食べるな“よな”」の、それぞれの評価値を図 7.4 に示す。

- 「食べないで“よ”」、「食べるな“よ”」では、C334(ヽ) やC117(Λ)、C155(Λ) で、(b1) が伝わっている。上昇から下降へと転じる急激な動きが、《慌てて》いる印象を与えているものと推察される。

- 同じ上昇下降調でも、前半の上昇は小幅で、後半に終端の非常に低い値にまで大きく下降するC24(Λ) やC329(Λ) では、(b2) が良く伝わっている。命令文で下降音調を伴うと異議の申し立ての表現になる“よ”[121] が、大幅な下降を伴うことによって〈非難〉の度合いが強まり、《怒って》いる印象が伝わったものと考えられる。

- 一方、C151(ヽ) やC233(ヽ)、C65(ヽ) などの緩やかに下降する音調では、(b3) が伝わっている。

- 「食べるな“よな”」では、(b1) はほとんど伝わっていないが、(b2) はC269(−) やC229(−)、C101(−) のような上下にうねる音調ではっきりと伝わっている。〈命令〉の意図において、これらの音調によって《怒って》いる印象が伝わっている傾向（図 7.3）と共通している。(b3) は、わずかな上昇を伴うC123(Γ) でやや感じ取られている。

表 7.4　付加的なニュアンスの違いをもたらす文末音調形状

文末音調				主な	主たる	付加的な
音調の分類	動きの特徴	テンプレートの例		文末詞	意図	ニュアンス
上昇調	上昇幅・大	C61$^{(J)}$ 	C131$^{(\Gamma)}$ 	“ね”	〈依頼〉	《熱心さ》あり
	上昇幅・小	C123$^{(\Gamma)}$ 	C248$^{(\Gamma)}$ 	“ね”	〈依頼〉	《熱心さ》なし
		C23$^{(J)}$ 	C248$^{(\Gamma)}$ 	“ね”	〈依頼〉	《親近感》
〔	上昇幅・大				〈質問〉	《驚き》　〕
平坦調	うねり・大	C269$^{(-)}$ 	C101$^{(-)}$ 	“ね”／“よ” “よな”	〈命令〉 〈非難〉	《怒って》
平坦調／下降調	うねり・小	C151$^{(\backslash)}$ 	C233$^{(\backslash)}$ 	“よ”	〈非難〉	《ガッカリして》
〔	うねり・無					《無関心》　〕
上昇下降調	上昇幅・大	C117$^{(\Lambda)}$ 	C155$^{(\Lambda)}$ 	“よ”	〈非難〉	《慌てて》
	下降幅・大	C24$^{(\Lambda)}$ 	C329$^{(\Lambda)}$ 	“よ”	〈非難〉	《怒って》

〔　〕は先行研究[13]による

7.4　ニュアンスの違いをもたらす文末音調形状の特徴

以上の分析から明らかになった、ニュアンスの違いをもたらす文末音調形状の特徴を、表 7.4 に整理する。

文末詞“よ”で〈非難〉の意を表明する際の上昇下降調には、伝わるニュアンスに違いがあるバリエーションが存在することが分かった。それらの F0 形状には、上昇幅や下降幅に微妙な違いが見られる。このことは文末音調がもつ表現力の高さを裏付けており、このようなバリエーションを積極的に利用することで、表現の幅を広げられるものと期待できる。しかし、文末音調を 5〜6 種類 [132],[133] に分類する従来の議論では、このような F0 形状の微妙な違いまでは区別されていない。それぞれで伝わるニュアンスの違いを考慮せずに音調として付与していては、聞き手に誤ったメッセージを伝えてしまうことにもなり兼ねない。そうした事態を避けるためには、文末音調を、形状の微妙な違いにも注目して従来よりも細分化し、それぞれがどのような意図やニュアンスを表現する機能をもつかを明らかにしていくことが必要である。

7.4.1　上昇調の上昇幅の異なり

第 6 章で、例えば「食べて“ね”」といって〈依頼〉の意図を伝えたいときは、“ね”の音調として $C5^{[J]}$、$C20^{[V]}$、または $C16^{[\Gamma]}$ を付与することが可能であることを示した。しかし今回、$C5^{[J]}$ の子クラスタの $C61^{(J)}$ や $C16^{[\Gamma]}$ の子クラスタの $C131^{(\Gamma)}$ などと、$C16^{[\Gamma]}$ の子クラスタの $C123^{(\Gamma)}$ や $C248^{(\Gamma)}$ とでは、相反するニュアンスが伝わることが分かった（図 7.2(a)）。いずれも大きく分ければ上昇調ではあるが、上昇幅の違いが伝わるニュアンスの違いを生んでいると考えられる。〈依頼〉の意図の表明においては、上昇幅の大小で聞き手への働きかけの《熱心さ》の程度の違いを表現できることが期待できる。

また、上昇調は聞き手の反応を伺う意味を表す [124] とされており、上昇幅の違いで反応を伺う気持ちの強さの違いが表現できるとも考えられる。〈質問〉文での F0 の大幅な上昇による《驚き》の表出 [13] は、相手から更に情報を引き出そうとする積極的な気持ちの現れと解釈できる。

7.4.2　上下にうねる動きの有無

〈命令〉の意図において、《怒って》いる印象が C269$^{(-)}$ や C101$^{(-)}$ などで伝わっている（図 7.3)。これらは、始端と終端の高さがほぼ同じであることから平坦調の一種ともいえるが、上下に大きくうねる動きが特徴的である。〈非難〉の意図を伝える「食べるな“よな”」でも、同様の傾向が見られている（図 7.4(c)）。聞き手は、F0 の大きな上下動から感情の起伏のようなものを感じ取るのではないかと推察される。

これに対して、〈非難〉の意図を伝える「食べないで“よ”」、「食べるな“よ”」での C151$^{(\backslash)}$ や C233$^{(\backslash)}$ などでは、《ガッカリして》いる印象が伝わっている（図 7.4(a), 図 7.4(b)）。これらは、上下のうねりは小さく、全体的には平坦か、あるいは緩やかに下降する動きを示している。平坦な音調で《無関心》さが伝わる傾向 [13] と合わせて考えると、上下にうねる動きのある音調とは対照的に、これらでは抑制された感情が伝わるものと推察される。

7.4.3　上昇下降調の変動幅の異なり

上昇下降調と呼べる音調を付与した〈非難〉の意図を伝える“よ”で、音調形状の微妙な差により異なるニュアンスが伝わっている（図 7.4(a)，図 7.4(b)）。C117$^{(\Lambda)}$ や C155$^{(\Lambda)}$ などでは《慌てて》いる印象が、C24$^{(\Lambda)}$ や C329$^{(\Lambda)}$ では《怒って》いる印象が、それぞれ伝わっている。前者の音調は、後者に比べて前半部分の上昇幅が大きい。後者は、後半の下降部分がより低い値に向かって急しゅんになっている点が特徴的である。このような上昇と下降のそれぞれの変動幅の違いが、ニュアンスの違いを生んでいると考えられる。

7.5　むすび

話し手の様々な心情が表出されることで生じる付加的なニュアンスの違いに着目し、文末音調の違いによって異なるニュアンスを伝える表現手法の実現可能性を示した。

従来の文末音調の分類よりも細かく、微妙に形状の異なる音調を文末に付与した合成音声の聴取実験を実施した。その結果、一般に話し手は文末音調以外にも様々な音響特徴量を変化させて意図を表現しているのに対して、文末音調が違うだけでも、聞き手には異なるニュアンスが伝わることを確認した。従来の分類では同じ音調として扱われると考えられる文末 F0 形状の中には、F0 の動きが微妙に異なり、伝わるニュアンスに違いが見られる様々なバリエーションがあることも具体的に明らかにした。

これらの結果は、文末音調がもつ表現力の高さを裏付けるものであり、主たる意図だけでなく言外の意図ともいうべきニュアンスについても、文末音調による表現が可能であることが分かった。文末音調の分類を従来よりも細分化し、それぞれによって伝わる意図やニュアンスを明らかにしていくことで、対話音声合成における表現の幅を更に広げていける可能性を示すことができた。

第8章　結論

8.1　本研究のまとめ

合成音声における音声表現の多様化に向けて二つの具体的な目標を掲げ、それぞれを実現するための研究に取り組んだ。

【目標 1】　対話状況に応じた音声表現の多様化

対話の状況に応じて異なる音声表現を使い分けることができる機能を提供する。　…… （第 2 章～第 3 章）

【目標 2】　文末音調による意図表現の多様化

話し手の意図を確実に伝えることができる文末音調による表現手法を確立する。　…… （第 4 章～第 7 章）

対話状況に応じた音声表現の多様化では、対話を取り巻く状況の違いに応じた複数の音声表現を使い分けることが、音声対話システムからの応答を表情豊かにすることに有効であることを示した。更に、対話の流れの中で異なる音声表現を使い分ける上では、個々の表現が適切であることだけでなく、複数の表現全体の調和が保たれていることも不可欠であることを指摘し、全体の調和を保ちながら複数の異なる音声表現を収集する手法を設計した。多様な表現全体の調和性の確保は、本研究で着目した対話状況に応じた音声表現の多様化に限った問題ではなく、感情や発話スタイルなどの観点での多様化にも共通する課題と考えられる。

文末音調による意図表現の多様化では、話し手の意図を伝える上で文末詞とその音調の組合せには適不適があること、更には、文末音調の形状の微妙な違いに応じて主たる意図とは別の付加的なニュアンスの違いさえも伝わることを明らかにした。発話の意図が聞き手に正しく伝わるようにするためには、適切な文末音調を付与する必要があることを指摘し、文末詞とその音調の組合せによる発話意

図の表現手法を確立した。合成音声の聴取実験に依拠したアプローチによって、文末音調が、長年、言語学の分野で議論されてきている以上に、高い表現力をもっていることを示す結果が得られたことは、大変意義深い。

8.1.1　対話状況に応じた音声表現の多様化に関する成果

対話の状況に応じた音声表現の違いに着目し、複数の対話状況における音声表現を収集した。それらの表現を対話の流れの中で使い分けることができる、音声合成の枠組みを構築した。

第 2 章では、音声対話システムの応答を表情豊かにするためには、どのような音声表現を用意すべきかを論じた。人同士の対話では、そのときの話題や場の雰囲気などの要因に応じて様々な話し方や声質などの音声表現が用いられることから、対話を取り巻く状況の違いに応じた音声表現に着目した。感情の円環モデルの 2 次元平面を手掛かりとして、四つの象限に対応するような状況 I〜状況 IV と原点に対応する状況 0 の、5 種類の対話の状況を選定した。ロボットと人との模擬対話を用いた主観評価実験で、ロボットの全ての発話を状況 0 のモデルで合成した従来型の応答に比べ、発話文ごとに状況 I〜状況 IV のモデルを適宜使い分けて合成した応答は、より自然で対話らしいとの評価を得た。選定した音声表現は様々な対話の場面で汎用的に利用可能であり、それらを対話の中で使い分けることで、音声対話システムの応答が表情豊かなものになることが確認できた。一方で、合成音声の品質に関する課題も見つかった。

第 3 章では、第 2 章の評価実験で浮かび上がった課題の改善に取り組んだ。特に、対話の流れに応じて異なる音声表現を使い分けたときの、表現全体の調和性を確保することに力点を置いた。これまでのように、異なる音声表現をそれぞれ別々に単独で表出させていると、表現同士の乖離を招き兼ねない。そこで、対話の流れの中で異なる音声表現を順に表出させることを考え、話し手の心的状態が次々と変化するように進行するスキットを利用した音声収集手法を提案した。合成音声の評価においても、対話の流れに沿って異なる音声表現が出現したときの全体の自然さや対話らしさを評価するために、各表現が次々と満遍なく出現する模擬対話を用いた評価を実施した。その結果、新たに収集した音声セットによる

合成音声では、音素連鎖のバリエーションなどを考慮した文セットにより明瞭性、自然性が向上し、異なる音声表現を対話の中で使い分けたときの調和性も保たれていることを確認することができ、提案した音声収集手法の有効性を示すことができた。

8.1.2　文末音調による意図表現の多様化に関する成果

文末音調の形状のバリエーションを分類し、そこから複数の形状を、合成音声に文末音調を付与するためのテンプレートとして選定した。それらを用いて、文末詞と文末音調の組合せによって話し手の意図を伝える表現手法を構築した。

第 4 章では、対話音声の文末で、実際にどのような F0 形状が用いられているかを具体的に明らかにした。対話調の音声データから抽出した文末 F0 形状を、合成音声の文末音調の付与に用いるテンプレートとして利用できるようにするために、時間軸と周波数軸を正規化する方法を検討した。その上で、階層的クラスタリングによって形状を分類し、どのような文末音調が存在するかを見渡せる樹形図に示した。これらの中には、上昇調、上昇下降調などの、言語学の分野での分類に相当する特徴をもった形状も見られた。

第 5 章では、クラスタリングによって分類した多数の文末 F0 形状の中から、合成音声の文末音調の付与に利用するテンプレートを選び出すために、聴感上の判定を基準として導入した絞り込みの手法を考案した。一つのクラスタを分割した後の二つのクラスタのセントロイドが音調として差異があるかどうかを、相槌の音声「はぁ」に付与した音調の対比較実験によって明らかにした。その結果、似通った形状でも聴感上ははっきり違うと感じられるものや、見掛け上は違いがあっても音調としての差異はほとんど感じられないものがあり、テンプレートの絞り込みに聴感上の判定を基準として導入することの必要性が示された。この結果を利用して、音調としての差異が顕著なものを選ぶような絞り込みを試みたところ、言語学の分野の先行研究における分類との対応が見られる形状が選ばれた。「上昇調」や「くだりのぼり音調」などのような名称で呼ばれている音調を合成音声に付与するための、具体的な F0 形状のテンプレートを獲得することができた。

第 6 章では、文末詞とその音調の組合せに着目し、それらと聞き手に伝わる話

し手の意図との関係を明らかにした。聴感上も音調としての差異が顕著な 6 種類の F0 形状を文末音調として付与した合成音声の聴取実験により、文末音調には、文末詞と伝えたい意図の組合せに応じて適不適があることを示した。その結果を整理して、文末詞と伝えたい意図の組合せに適した文末 F0 形状をテンプレートの中から選択する、発話意図の表現手法を提案した。音調に応じて異なる意図が伝わる可能性がある文末詞を伴う発話文では、伝えたい意図に適した文末音調を付与しないとコミュニケーションに齟齬を来すことになる。本研究により、対話音声合成における文末音調の適切な制御の必要性を、改めて示すことができた。

第 7 章では、第 6 章で扱った主たる意図に付加される微妙なニュアンスに着目し、文末音調が違うだけでも聞き手には異なるニュアンスが伝わることを明らかにした。これまでほとんど議論されてこなかった言外の意図ともいうべき付加的なニュアンスについても、文末音調による表現手法の実現可能性を示すことができた。従来の文末音調の分類では同じ音調として扱われると考えられる F0 形状の中には、F0 の動きが微妙に異なり、伝わるニュアンスに違いが見られる、様々なバリエーションがあることも明らかになった。これらの結果は、文末音調を従来よりも細分化した上で、それぞれによって伝わる意図やニュアンスを具体的に明らかにしていくことで、合成音声による表現の幅が更に広がる可能性があることを示している。

8.2　今後の課題

対話状況に応じた音声表現の多様化では、5 種類の対話の状況を想定した音声表現を用意したが、これで十分というわけではない。どのような音声表現を提供することが有用かについては、更なる検討の余地がある。また、本研究では、異なる音声表現を使い分けられる機能を提供することを目標としたが、対話システムが、これらの音声表現を対話の流れの中でどのようにして使い分けるかや、異なる音声表現を使い分けることが対話の活性化につながるかなどは、対話制御の研究における新たな課題となり得るもので、興味深い。

文末音調による意図表現の多様化では、文末音調がもつ表現力の高さを例証することができたが、研究はまだ緒に就いたばかりである。実験で確認した範囲は

限定的であり、更に多くの意図や文末詞を対象とした調査が必要である。文末音調だけでは十分に伝わらなかった意図でも、継続時間長を制御することで表現できる可能性も示された。他の特徴量の制御も併せた表現方法についても、探っていく必要がある。また、クラスタリングによって得られたテンプレートを使った音調を聞き比べてみると、何らかのニュアンスの違いを感じられるものが幾つもあった。これらによって様々な意図にどのようなニュアンスを付加する表現が可能かなど、表現の幅を広げるための研究課題は数多くあり、興味は尽きない。

現在、機械学習に基づくコーパスベースの音声合成が主流となっており、大量の音声データを集めさえすれば、比較的容易に合成音声を生成することができるツール類の整備も進んでいる。しかし、更なる品質向上を図ったり、新たな機能を導入したりするには、何に着目すべきか、何を学習させればよいかが重要な鍵となる。その手掛かりを見いだすには、音声や言葉、コミュニケーションに関する知識や知見が不可欠である。今後とも、こうした知見の獲得を目指した研究に取り組み、合成音声の表現の引き出しを増やす新たな提案につなげて、表情豊かで生き生きとした音声対話システムの実現に貢献したいと考えている。

謝辞

本研究を進めるに当たって、多大なるご指導、ご教授を賜りました早稲田大学理工学術院 小林哲則教授に、心より感謝申し上げます。本論文の柱の一つである「文末音調による発話意図の表現」は、筆者が早稲田大学・大学院在学中に、修士論文の研究テーマの候補として構想していたものの、長い間眠らせてしまっていたものです。四半世紀以上の時を経て、改めてその研究に取り組み、成果を博士論文としてまとめることができたことは、望外の喜びです。これもひとえに、本研究に着手する機会と、研究に必要な環境を与えていただけたことによるものであり、感謝の念に堪えません。

本論文の執筆に当たっては、名古屋工業大学 徳田恵一教授、早稲田大学 甲藤二郎教授、小川哲司教授から貴重なご意見、ご助言を賜りました。ここに深く感謝申し上げます。

早稲田大学・理工学部在学中の 1983 年に “WABOT–2” プロジェクトと出会ったことで、それまで漠然と興味を抱いていた「音声合成」が、生涯の研究テーマとなりました。音声研究の道へと導いていただいた白井克彦 早稲田大学名誉教授に、心より感謝申し上げます。

修士課程修了後は、日本電気株式会社にて、音声合成技術の研究開発、音声認識応答装置の製品開発、音声インタフェースを有する応用システムの実用化などに従事しました。日本電気株式会社を早期定年退職した後も、小林先生の下で、早稲田大学にて研究員の職務に従事しながら、音声合成の研究を継続することができました。40 年にわたり一貫して、音声合成をはじめとする音声情報処理技術

の研究と製品開発に携わり続けられたことが、本論文をまとめ上げる原動力となりました。早稲田大学、日本電気株式会社に在籍中のそれぞれの場で、ご指導いただいた諸先輩、本研究を含め、これまでの研究開発活動の推進をサポートしていただいた多くの関係者の皆様の、温かいご支援の賜物と感謝しております。

また、本研究の成果をまとめるに当たって不可欠であった、数々の評価実験にご協力いただいた小林研究室所属の学生の皆様に、この場を借りて厚く御礼申し上げます。

最後に、日本電気株式会社を早期退職して以来、今日まで、博士学位の取得を応援し続けてくれた妻と、大学院修士課程にまで進学させてくれた亡き両親に、心から感謝します。

参考文献

[1] H. Dudley and T.H. Tarnoczy, “The speaking machine of Wolfgang von Kempelen,” J. Acoust. Soc. Am., Vol.22, No.2, pp.151–166, March 1950.
DOI:10.1121/1.1906583

[2] J.Q. Stewart, “An electrical analogue of the vocal organs,” Nature, Vol.110, No.2757, pp.311–312, Sept. 1922.
DOI:10.1038/110311a0

[3] H. Dudley, R.R. Riesz, and S.S.A. Watkins, “A synthetic speaker,” J. Franklin Institute, Vol.227, No.6, pp.739–764, June 1939.
DOI:10.1016/S0016-0032(39)90816-1

[4] J.L. Flanagan, “Voices of men and machines,” J. Acoust. Soc. Am., Vol.51, No.5A, pp.1375–1387, May 1972.
DOI:10.1121/1.1912988

[5] D.H. Klatt, “Review of text-to-speech conversion for English,” J. Acoust. Soc. Am., Vol.82, No.3, pp.737–793, Sept. 1987.
DOI:10.1121/1.395275

[6] M.R. Schroeder, “A brief history of synthetic speech,” Speech Commun., Vol.13, No.1–2, pp.231–237, Oct. 1993.
DOI:10.1016/0167-6393(93)90074-U

[7] M.E. Beckman, “Speech models and speech synthesis,” in Progress in Speech Synthesis, eds. J.P.H. van Santen, J.P. Olive, R.W. Sproat, and J. Hirschberg, pp.185–209, Springer, New York, 1997.
DOI:10.1007/978-1-4612-1894-4_15

[8] 広瀬啓吉，“音声合成技術,” 情報処理，Vol.38，No.11，pp.984–991，Nov. 1997.

[9] M. Schröder, “Expressive speech synthesis: Past, present, and possible futures,” in Affective Information Processing, eds. J. Tao and T. Tan, pp.111–126, Springer, London, 2009.
DOI:10.1007/978-1-84800-306-4_7

[10] B.H. Story, “History of speech synthesis,” in The Routledge Handbook of Phonetics, eds. W.F. Katz and P.F. Assmann, pp.9–33, Taylor and Francis, London, 2019.
DOI:10.4324/9780429056253-2

[11] 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 Commun., Vol.9, No.4, pp.357–363, Aug. 1990.
DOI:10.1016/0167-6393(90)90011-W

[12] 小林哲則，板橋秀一，速水　悟，竹沢寿幸，“日本音響学会研究用連続音声データベース,” 音響誌，Vol.48，No.12，pp.888–893，Dec. 1992.
DOI:10.20697/jasj.48.12_888

[13] 前川喜久雄，籠宮隆之，小磯花絵，小椋秀樹，菊池英明，“日本語話し言葉コーパスの設計,” 音声研究，Vol.4，No.2，pp.51–61，Aug. 2000.
DOI:10.24467/onseikenkyu.4.2_51

[14] 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 Commun., Vol.53, No.1, pp.36–50, Jan. 2011.
DOI:10.1016/j.specom.2010.08.002

[15] S. Takamichi, R. Sonobe, K. Mitsui, Y. Saito, T. Koriyama, N. Tanji, and H. Saruwatari, “JSUT and JVS: Free Japanese voice corpora for accelerating speech synthesis research,” Acoust. Sci. & Tech., Vol.41, No.5, pp.761–768, Sept. 2020.
DOI:10.1250/ast.41.761

[16] 匂坂芳典，ニックキャンベル，“音声合成のための規則とデータの表現，獲得，評価,” 信学論（D-II），Vol.J83-D-II，No.11，pp.2068–2076，Nov. 2000.

[17] 匂坂芳典，“コーパスベース音声合成技術の動向［I］ —コーパスベース音声合成の過去・現在・将来,” 信学誌，Vol.87，No.1，pp.64–69，Jan. 2004.

[18] 阿部匡伸，“コーパスベース音声合成技術の動向［II］ —音声合成単位を例題に,” 信学誌，Vol.87，No.2，pp.129–134，Feb. 2004.

[19] 河井　恒，津崎　実，“コーパスベース音声合成技術の動向［III］ —コーパスの設計と評価尺度,” 信学誌，Vol.87，No.3，pp.227–231，March 2004.

[20] 小林隆夫，徳田恵一，“コーパスベース音声合成技術の動向［IV］ —HMM音声合成方式,” 信学誌，Vol.87，No.4，pp.322–327，April 2004.

[21] N. Campbell, “コーパスベース音声合成技術の動向［V・完］ —大規模音声コーパスによる音声合成,” 信学誌，Vol.87，No.6，pp.497–500，June 2004.

[22] 徳田恵一，“隠れマルコフモデルによる音声認識と音声合成,” 情報処理，Vol.45，No.10，pp.1005–1011，Oct. 2004.

[23] G. Fant, Acoustic theory of speech production, Mouton, The Hague, 1960.

[24] L. Rabiner, “Speech synthesis by rule: An acoustic domain approach,” The Bell System Technical J., Vol.47, No.1, pp.17–37, Jan. 1968.
DOI:10.1002/j.1538-7305.1968.tb00029.x

[25] A.E. Rosenberg, “Effect of glottal pulse shape on the quality of natural vowels,” J. Acoust. Soc. Am., Vol.49, No.2B, pp.583–590, Feb. 1971.
DOI:10.1121/1.1912389

[26] K. Ishizaka and J.L. Flanagan, “Synthesis of voiced sounds from a two-mass model of the vocal cords,” The Bell System Technical J., Vol.51, No.6, pp.1233–1268, July-Aug. 1972.
DOI:10.1002/j.1538-7305.1972.tb02651.x

[27] D.H. Klatt, “Software for cascade/parallel formant synthesizer,” J. Acoust. Soc. Am., Vol.67, No.3, pp.971–995, March 1980.
DOI:10.1121/1.383940

[28] 箱田和雄，佐藤大和，“文音声合成における音調規則,” 信学論（D），Vol.J63-D，No.9，pp.715–722，Sept. 1980.

[29] H. Fujisaki and K. Hirose, “Analysis of voice fundamental frequency contours for declarative sentences of Japanese,” J. Acoust. Soc. Jpn. (E), Vol.5, No.4, pp.233–242, Oct. 1984.
DOI:10.1250/ast.5.233

[30] 匂坂芳典，東倉洋一，“規則による音声合成のための音韻時間調制御,” 信学論（A），Vol.J67-A，No.7，pp.629–636，July 1984.

[31] M.E. Beckman and J.B. Pierrehumbert, “Japanese prosodic phrasing and intonation synthesis,” Proc. ACL, pp.173–180, New York, USA, July 1986. DOI:0.3115/981131.981156

[32] 酒寄哲也，佐々部昭一，北川博雄，“規則合成のための数量化 I 類を用いた韻律制御,” 秋季音響講論集，3–4–17，pp.245–246，Oct. 1986.

[33] J.B. Pierrehumbert and M.E. Beckman, Japanese Tone Structure, The MIT Press, Cambridge, 1988.

[34] 海木延佳，匂坂芳典，“局所的な句構造によるポーズ挿入規則化の検討,” 信学論（D-II），Vol.J79-D-II，No.9，pp.1455–1463，Sept. 1996.

[35] 武藤博子，井島勇祐，宮崎　昇，水野秀之，阪内澄宇，“合成音声への自然なポーズ挿入のための音声の自然性に影響を与えるポーズ位置に関する要因の分析と評価,” 情処学論，Vol.56，No.3，pp.993–1002，March 2015.

[36] 吉村貴克，徳田恵一，益子貴史，小林隆夫，北村　正，“HMM に基づく音声合成におけるスペクトル・ピッチ・継続長の同時モデル化,” 信学論（D-II），Vol.J83-D-II，No.11，pp.2099–2107，Nov. 2000.

[37] 田村正統，益子貴史，徳田恵一，小林隆夫，“HMM に基づく音声合成におけるピッチ・スペクトルの話者適応,” 信学論（D-II），Vol.J85-D-II，No.4，pp.545–553，April 2002.

[38] M. Tachibana, J. Yamagishi, T. Masuko, and T. Kobayashi, “A style adaptation technique for speech synthesis using HSMM and suprasegmental features,” IEICE Trans. Inf. & Syst., Vol.E89-D, No.3, pp.1092–1099, March 2006.
DOI:10.1093/ietisy/e89-d.3.1092

[39] 能勢　隆，“統計モデルに基づく多様な音声の合成技術,” 信学論（D），Vol.J100-D，No.4，pp.556–569，April 2017.
DOI:10.14923/transinfj.2016JDS0001

[40] H. Zen, A. Senior, and M. Schuster, “Statistical parametric speech synthesis using deep neural networks,” Proc. ICASSP 2013, SP–P17.3, pp.7962–7966, Vancouver, Canada, May 2013.
DOI:10.1109/ICASSP.2013.6639215

[41] Y. Qian, Y. Fan, W. Hu, and F.K. Soong, "On the training aspects of deep neural network (DNN) for parametric TTS synthesis," Proc. ICASSP 2014, pp.3829–3833, Florence, Italy, May 2014.
DOI:10.1109/ICASSP.2014.6854318

[42] X. Wang, S. Takaki, and J. Yamagishi, "A comparative study of the performance of HMM, DNN, and RNN based speech synthesis systems trained on very large speaker-dependent corpora," Proc. 9th ISCA Workshop on Speech Synthesis, pp.118–121, Sunnyvale, USA, Sept. 2016.
DOI:10.21437/SSW.2016-20

[43] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "WaveNet: A generative model for raw audio," Proc. 9th ISCA Workshop on Speech Synthesis, p.125, Sunnyvale, USA, Sept. 2016.

[44] J. Sotelo, S. Mehri, K. Kumar, J.F. Santos, K. Kastner, A. Courville, and Y. Bengio, "Char2Wav: End-to-end speech synthesis," Proc. ICLR 2017 – Workshop track, W5, pp.1–6, Toulon, France, April 2017.

[45] Y. Wang, R.J. Skerry-Ryan, D. Stanton, Y. Wu, R.J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, and R.A. Saurous, "Tacotron: Towards end-to-end speech synthesis," Proc. INTERSPEECH 2017, pp.4006–4010, Stockholm, Sweden, Aug. 2017.
DOI:10.21437/Interspeech.2017-1452

[46] Y. Taigman, L. Wolf, A. Polyak, and E. Nachmani, "VoiceLoop: Voice fitting and synthesis via a phonological loop," Proc. ICLR 2018, no. 31, pp.1–14, Vancouver, Canada, April 2018.

[47] W. Ping, K. Peng, A. Gibiansky, S. Arik, A. Kannan, S. Narang, J. Raiman, and J. Miller, "Deep Voice 3: Scaling text-to-speech with convolutional sequence learning," Proc. ICLR 2018, no. 41, pp.1–16, Vancouver, Canada, April 2018.

[48] N. Li, S. Liu, Y. Liu, S. Zhao, and M. Liu, "Neural speech synthesis with transformer network," Proc. AAAI Conference on Artificial Intelligence, pp.6706–6713, Honolulu, USA, Jan. 2019.
DOI:10.1609/aaai.v33i01.33016706

[49] 藤原与一，文末詞の言語学，三弥井書店，東京，1990.

[50] 山田修平，能勢　隆，伊藤彰則，“多様な対話音声合成のための話し言葉音声コーパスの構築と評価,” 情処研報，Vol.2015-MUS-107，No.72，pp.1–6，May 2015.

[51] J. Yamagishi, K. Onishi, T. Masuko, and T. Kobayashi, “Acoustic modeling of speaking styles and emotional expressions in HMM-based speech synthesis,” IEICE Trans. Inf. & Syst., Vol.E88-D, No.3, pp.502–509, March 2005. DOI:10.1093/ietisy/e88-d.3.502

[52] 清山信正，世木寛之，今井　篤，都木　徹，“感情制御用音声データの評価,” 映情学冬季予稿集，13-4，Dec. 2013.
DOI:10.11485/itewac.2013.0_13_4

[53] 河井　恒，樋口宜男，山本誠一，“基本周波数及び音素持続時間を考慮した音声合成用波形素片データセットの作成,” 信学論（D-II），Vol.J82-D-II，No.8，pp.1229–1238，Aug. 1999.

[54] T. Nose, Y. Arao, T. Kobayashi, K. Sugiura, and Y. Shiga, “Sentence selection based on extended entropy using phonetic and prosodic contexts for statistical parametric speech synthesis,” IEEE/ACM Trans. Audio, Speech, Lang. Process., Vol.25, No.5, pp.1107–1116, May 2017.
DOI:10.1109/TASLP.2017.2688585

[55] 河井　恒，戸田智基，山岸順一，平井俊男，倪　晋富，西澤信行，津崎　実，徳田恵一，“大規模コーパスを用いた音声合成システム XIMERA,” 信学論（D），Vol.J89-D，No.12，pp.2688–2698，Dec. 2006.

[56] O. Boeffard, L. Charonnat, S. Le Maguer, D. Lolive, and G. Vidal, “Towards fully automatic annotation of audiobooks for TTS,” Proc. LREC 2012, pp.975–980, Istanbul, Turkey, May 2012.

[57] M. Charfuelan and I. Steiner, “Expressive speech synthesis in MARY TTS using audiobook data and EmotionML,” Proc. INTERSPEECH 2013, Tue–O–22–6, pp.1564–1568, Lyon, France, Aug. 2013.
DOI:10.21437/Interspeech.2013-395

[58] H. Zen, V. Dang, R. Clark, Y. Zhang, R.J. Weiss, Y. Jia, Z. Chen, and Y. Wu, “LibriTTS: A corpus derived from LibriSpeech for text-to-speech,”

Proc. INTERSPEECH 2019, pp.1526–1530, Graz, Austria, Sept. 2019.
DOI:10.21437/Interspeech.2019-2441

[59] W. Nakata, T. Koriyama, S. Takamichi, N. Tanji, Y. Ijima, R. Masumura, and H. Saruwatari, “Audiobook speech synthesis conditioned by cross-sentence context-aware word embeddings,” Proc. 11th ISCA Speech Synthesis Workshop, pp.211–215, Budapest, Hungary, Aug. 2021.
DOI:10.21437/SSW.2021-37

[60] 河原達也，“音声対話システムの進化と淘汰 —歴史と最近の技術動向,” 人工知能誌，Vol.28，No.1，pp.45–51，Jan. 2013.

[61] 小林哲則，藤江真也，松坂要佐，白井克彦，“人間形会話ロボット —パラ言語の生成・理解機能を持つマルチモーダルインタフェース,” 音響誌，Vol.61，No.2，pp.85–90，Feb. 2005.
DOI:10.20697/jasj.61.2_85

[62] 嵯峨山茂樹，西本卓也，中沢正幸，“擬人化音声対話エージェント,” 情報処理，Vol.45，No.10，pp.1044–1049，Oct. 2004.

[63] 大浦圭一郎，山本大介，内匠　逸，李　晃伸，徳田恵一，“キャンパスの公共空間におけるユーザ参加型双方向音声案内デジタルサイネージシステム,” 人工知能誌，Vol.28，No.1，pp.60–67，Jan. 2013.

[64] 翠　輝久，水上悦雄，堀　智織，柏岡秀紀，“音声対話による観光案内システムの開発と多言語化 —音声対話システム AssisTra の研究開発から得られた知見と課題,” 人工知能誌，Vol.28，No.1，pp.68–74，Jan. 2013.

[65] 藤江真也，松山洋一，谷山　輝，小林哲則，“人同士のコミュニケーションに参加し活性化する会話ロボット,” 信学論（A），Vol.J95-A，No.1，pp.37–45，Jan. 2012.

[66] Y. Matsuyama, A. Saito, S. Fujie, and T. Kobayashi, “Automatic expressive opinion sentence generation for enjoyable conversational systems,” IEEE/ACM Trans. Audio Speech Lang. Process., Vol.23, No.2, pp.313–326, Feb. 2015.
DOI:10.1109/TASLP.2014.2363589

[67] 藤堂祐樹，西村良太，山本一公，中川聖一，“複数の対話エージェントを用いた雑談指向の音声対話システム,” 信学論（D），Vol.J99-D，No.2，pp.188–

200，Feb. 2016.
DOI:10.14923/transinfj.2015JDP7010

[68] 郡　史郎，“日本語の「口調」にはどんな種類があるか，” 音声研究，Vol.10，No.3，pp.52–68，Dec. 2006.
DOI:10.24467/onseikenkyu.10.3_52

[69] H. Fujisaki, “Prosody, models, and spontaneous speech,” in Computing Prosody: Computational Models for Processing Spontaneous Speech, eds. Y. Sagisaka, N. Campbell, and N. Higuchi, pp.27–42, Springer, New York, 1997.
DOI:10.1007/978-1-4612-2258-3_3

[70] 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，March 2002.

[71] 森　大毅，前川喜久雄，粕谷英樹，音声は何を伝えているか —感情・パラ言語情報・個人性の音声科学，コロナ社，東京，2014.

[72] 郡　史郎，“対人関係・対人態度を反映する韻律的特徴 —特に目上に対する話し方について，” 日本語の教育から研究へ，土岐哲先生還暦記念論文集編集委員会（編），pp.167–176，くろしお出版，東京，2006.

[73] 永野マドセン泰子，鮎澤孝子，“日本語における感情表現とイントネーション —男女差，および日本語学習者の母語背景の視点を加えて，” 音声文法，杉藤美代子（編），pp.61–83，くろしお出版，東京，2011.

[74] A. Mehrabian, Silent Messages, Wadsworth Publishing, Belmont, California, 1971.

[75] 広瀬啓吉，藤崎博也，河井　恒，山口幹雄，“基本周波数パターン生成過程モデルに基づく文章音声の合成，” 信学論（A），Vol.J72-A，No.1，pp.32–40，Jan. 1989.

[76] 越智景子，広瀬啓吉，峯松信明，“基本周波数パターン生成過程モデルの指令の差分に着目した発話の焦点制御，” 信学論（D），Vol.J98-D，No.3，pp.524–533，March 2015.
DOI:10.14923/transinfj.2014JDP7084

[77] 世木寛之，清山信正，田高礼子，都木　徹，大出訓史，今井　篤，西脇正通，小山隆二，“高品質な株価音声合成装置の開発とデジタルラジオ放送での試

験運用,” 映情学誌，Vol.62，No.1，pp.69–76，Jan. 2008.
DOI:10.3169/itej.62.69

[78] 世木寛之，田高礼子，清山信正，都木　徹，斎藤英雄，小澤愼治，“音声合成のためのテンプレートを用いた録音文セット生成システムとラジオ番組「気象通報」への適用について,” 映情学誌，Vol.65，No.1，pp.76–83，Jan. 2011.
DOI:10.3169/itej.65.76

[79] 宮武正典，匂坂芳典，“種々の発話様式に見られる韻律特徴とその制御,” 信学論（D-II），Vol.J73-D-II，No.12，pp.1929–1935，Dec. 1990.

[80] N. Kaiki and Y. Sagisaka, “Prosodic characteristics of Japanese conversational speech,” IEICE Trans. Fundamentals, Vol.E76-A, No.11, pp.1927–1933, Nov. 1993.

[81] 広瀬啓吉，阪田真弓，“対話音声と朗読音声の韻律的特徴の比較,” 信学論（D-II），Vol.J79-D-II，No.12，pp.2154–2162，Dec. 1996.

[82] グリーンバーグ陽子，加藤宏明，津崎　実，匂坂芳典，“語彙が与える印象に基づく対話韻律生成,” 音響誌，Vol.67，No.2，pp.65–74，Feb. 2011.
DOI:10.20697/jasj.67.2_65

[83] 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,” 信学論（D-II），Vol.J88-D-II，No.3，pp.489–498，March 2005.

[84] 堀内靖雄，小磯花絵，土屋　俊，市川　熹，“自発的音声対話における話者交替の制御に関わる発話末の統語的・韻律的特徴,” 情処研報，Vol.96，No.21，pp.45–50，Feb. 1996.

[85] C.T. Ishi, “The functions of phrase final tones in Japanese: Focus on turn-taking,” 音声研究, Vol.10, No.3, pp.18–28, Dec. 2006.
DOI:10.24467/onseikenkyu.10.3_18

[86] 翠　輝久，水上悦雄，志賀芳則，川本真一，河井　恒，中村　哲，“ユーザの相づち・うなずきを喚起する音声対話システム,” 信学論（A），Vol.J95-A，No.1，pp.16–26，Jan. 2012.

[87] K. Sugiura, Y. Shiga, H. Kawai, T. Misu, and C. Hori, “A cloud robotics approach towards dialogue-oriented robot speech,” Advanced Robotics, Vol.29,

No.7, pp.449–456, March 2015.
DOI:10.1080/01691864.2015.1009164

[88] Y. Arimoto, H. Kawatsu, S. Ohno, and H. Iida, “Naturalistic emotional speech collection paradigm with online game and its psychological and acoustical assessment,” Acoust. Sci. & Tech., Vol.33, No.6, pp.359–369, June 2012.
DOI:10.1250/ast.33.359

[89] K. El Haddad, I. Torre, E. Gilmartin, H. Çakmak, S. Dupont, T. Dutoit, and N. Campbell, “Introducing AmuS: The amused speech database,” in Statistical Language and Speech Processing, eds. N. Camelin, Y. Estève, and C. Martín-Vide, pp.229–240, Springer, Cham, 2017.
DOI:10.1007/978-3-319-68456-7_19

[90] É. Székely, G.E. Henter, J. Beskow, and J. Gustafson, “Spontaneous conversational speech synthesis from found data,” Proc. INTERSPEECH 2019, pp.4435–4439, Graz, Austria, Sept. 2019.
DOI:10.21437/Interspeech.2019-2836

[91] H. Guo, S. Zhang, F.K. Soong, L. He, and L. Xie, “Conversational end-to-end TTS for voice agents,” Proc. SLT 2021, pp.403–409, Shenzhen, China, Jan. 2021.
DOI:10.1109/SLT48900.2021.9383460

[92] I.R. Murray and J.L. Arnott, “Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion,” J. Acoust. Soc. Am., Vol.93, No.2, pp.1097–1108, Feb. 1993.
DOI:10.1121/1.405558

[93] 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,” 情処学論，Vol.40，No.2，pp.479–486，Feb. 1999.

[94] J. Lorenzo-Trueba, G.E. Henter, S. Takaki, J. Yamagishi, Y. Morino, and Y. Ochiai, “Investigating different representations for modeling and controlling multiple emotions in DNN-based speech synthesis,” Speech Commun., Vol.99, pp.135–143, May 2018.
DOI:10.1016/j.specom.2018.03.002

[95] M. Yokoyama, T. Nagata, and H. Mori, “Effects of dimensional input on paralinguistic information perceived from synthesized dialogue speech with

neural network," Proc. INTERSPEECH 2018, pp.3053–3056, Hyderabad, India, Sept. 2018.
DOI:10.21437/Interspeech.2018-2042

[96] P. Wu, Z. Ling, L. Liu, Y. Jiang, H. Wu, and L. Dai, "End-to-end emotional speech synthesis using style tokens and semi-supervised training," Proc. APSIPA ASC 2019, pp.623–627, Lanzhou, China, Nov. 2019.
DOI:10.1109/APSIPAASC47483.2019.9023186

[97] M. Tahon, G. Lecorvé, and D. Lolive, "Can we generate emotional pronunciations for expressive speech synthesis?," IEEE Trans. Affective Comput., Vol.11, No.4, pp.684–695, Oct.–Dec. 2020.
DOI:10.1109/TAFFC.2018.2828429

[98] K. Inoue, S. Hara, M. Abe, N. Hojo, and Y. Ijima, "Model architectures to extrapolate emotional expressions in DNN-based text-to-speech," Speech Commun., Vol.126, pp.35–43, Feb. 2021.
DOI:10.1016/j.specom.2020.11.004

[99] P. Ekman, "An argument for basic emotions," Cognition and Emotion, Vol.6, No.3–4, pp.169–200, 1992.
DOI:10.1080/ 02699939208411068

[100] J.M. Montero, J.M. Gutiérrez-Arriola, S. Palazuelos, E. Enríquez, S. Aguilera, and J.M. Pardo, "Emotional speech synthesis: from speech database to TTS," Proc. ICSLP 98, paper 1037, Sydney, Australia, Nov. 1998.

[101] F. Burkhardt, A. Paeschke, M. Rolfes, W.F. Sendlmeier, and B. Weiss, "A database of German emotional speech," Proc. INTERSPEECH 2005, pp.1517–1520, Lisbon, Portugal, Sept. 2005.
DOI:10.21437/Interspeech.2005-446

[102] R. Barra-Chicote, J.M. Montero, J. Macias-Guarasa, S.L. Lufti, J.M. Lucas, F. Fernandez-Martinez, L.F. D'haro, R. San-Segundo, J. Ferreiros, R. Cordoba, and J.M. Pardo, "Spanish expressive voices: Corpus for emotion research in Spanish," Proc. Workshop on Corpora for Research on Emotion and Affect, pp.66–70, Marrakech, Morocco, May 2008.

[103] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J.N. Chang, S. Lee, and S.S. Narayanan, "IEMOCAP: interactive emotional

dyadic motion capture database," Lang. Resources and Eval., Vol.42, No.4, pp.335–359, Dec. 2008.
DOI:10.1007/s10579-008-9076-6

[104] G. Costantini, I. Iaderola, A. Paoloni, and M. Todisco, "EMOVO Corpus: An Italian emotional speech database," Proc. LREC 2014, pp.3501–3504, Reykjavik, Iceland, May 2014.

[105] E. Takeishi, T. Nose, Y. Chiba, and A. Ito, "Construction and analysis of phonetically and prosodically balanced emotional speech database," Proc. O-COCOSDA 2016, pp.16–21, Bali, Indonesia, Oct. 2016.
DOI:10.1109/ICSDA.2016.7918977

[106] J. James, L. Tian, and C.I. Watson, "An open source emotional speech corpus for human robot interaction applications," Proc. INTERSPEECH 2018, pp.2768–2772, Hyderabad, India, Sept. 2018.
DOI:10.21437/Interspeech.2018-1349

[107] N. Tits, K. El Haddad, and T. Dutoit, "Emotional speech datasets for english speech synthesis purpose: A review," in Intelligent Systems and Applications, eds. Y. Bi, R. Bhatia, and S. Kapoor, pp.61–66, Springer, Cham, 2020.
DOI:10.1007/978-3-030-29516-5_6

[108] K. Zhou, B. Sisman, R. Liu, and H. Li, "Emotional voice conversion: Theory, databases and ESD," Speech Commun., Vol.137, pp.1–18, Feb. 2022.
DOI:10.1016/j.specom.2021.11.006

[109] 音声入出力方式標準化専門委員会，JEITA IT-4012 音声合成技術で感情や意図を表現するための話し方種別のガイドライン，電子情報技術産業協会，東京，2018.

[110] 仁田義雄，日本語のモダリティと人称，ひつじ書房，東京，1991.

[111] 日本語記述文法研究会（編），現代日本語文法 4 第 8 部 モダリティ，くろしお出版，東京，2003.

[112] 益岡隆志，日本語モダリティ探究，くろしお出版，東京，2007.

[113] グリーンバーグ陽子，津崎　実，加藤宏明，匂坂芳典，"一語発話「ん」の対話韻律とその印象表現の対応について，" 音声文法，杉藤美代子（編），pp.199–208，くろしお出版，東京，2011.

[114] J.J. Venditti, K. Maeda, and J.P.H. van Santen, “Modeling Japanese boundary pitch movements for speech synthesis,” Proc. 3rd ESCA/COCOSDA Workshop on Speech Synthesis, pp.317–322, Blue Mountains, Australia, Nov. 1998.

[115] 国立国語研究所，国立国語研究所報告 3 現代語の助詞・助動詞 —用法と実例，秀英出版，東京，1951.
DOI:10.15084/00000991

[116] 益岡隆志，“終助詞「ね」と「よ」の機能,” モダリティの文法，pp.92–107，くろしお出版，東京，1991.

[117] 伊豆原英子，“終助詞「よ」「よね」「ね」の総合的考察 —「よね」のコミュニケーション機能の考察を軸に,” 名古屋大学日本語・日本文化論集，Vol.1，pp.21–34，1993.

[118] 犬養　隆，“低く短く付く終助詞「ね」,” 文法と音声 III，音声文法研究会（編），pp.17–29，くろしお出版，東京，2001.

[119] 宮崎和人，安達太郎，野田春美，高梨信乃，“終助詞の機能,” 新日本語文法選書 4 モダリティ，pp.261–288，くろしお出版，東京，2002.

[120] 杉藤美代子，“イントネーションの記号論,” 文化言語学 —その提言と建設，文化言語学編集委員会（編），pp.1068–1055，三省堂，東京，1992.

[121] 井上　優，“発話における「タイミング考慮」と「矛盾考慮」 —命令文・依頼文を例に,” 国立国語研究所報告 105 研究報告集 14，pp.333–360，秀英出版，東京，1993.
DOI:10.15084/00001138

[122] 御園生保子，“文末に現れるジャナイの用法と韻律の分析をめぐる問題について,” 日本語　意味と文法の風景 —国広哲弥教授古稀記念論文集，山田　進，菊地康人，籾山洋介（編），pp.343–355，ひつじ書房，東京，2000.

[123] 杉藤美代子，“終助詞「ね」の意味・機能とイントネーション,” 文法と音声 III，音声文法研究会（編），pp.3–16，くろしお出版，東京，2001.

[124] 森山卓郎，“文の意味とイントネーション,” 講座日本語と日本語教育　第 1 巻　日本語学要説，宮地　裕（編），pp.172–196，明治書院，東京，1989.

[125] 片桐恭弘，“終助詞とイントネーション,” 文法と音声，音声文法研究会（編），pp.235–256，くろしお出版，東京，1997.

[126] 大石初太郎，“日本語の音声表現について,”日本語の発見　ことばの勉強 1，山本安英の会（編），pp.183–195，未来社，東京，1969.

[127] 小山哲春，“文末詞と文末イントネーション,”文法と音声，音声文法研究会（編），pp.97–119，くろしお出版，東京，1997.

[128] 吉沢典男，“イントネーション,”国立国語研究所報告 18　話しことばの文型 (1) —対話資料による研究，pp.249–288，秀英出版，東京，1960.
DOI:10.15084/00001230

[129] 川上　蓁，“文末などの上昇調について,”国語研究，Vol.16，pp.25–46，Aug. 1963.

[130] 上村幸雄，“日本語のイントネーション,”ことばの科学 3，言語学研究会（編），pp.193–220，むぎ書房，東京，1989.

[131] 郡　史郎，“イントネーション,”朝倉日本語講座 3　音声・音韻，北原保雄（監），上野善道（編），pp.109–131，朝倉書店，東京，2003.

[132] 轟木靖子，“東京語の終助詞の音調と機能の対応について —内省による考察,”音声言語 VI，pp.5–28，近畿音声言語研究会，2008.

[133] 郡　史郎，“日本語の文末イントネーションの種類と名称の再検討,”言語文化研究，Vol.41，pp.85–107，March 2015.
DOI:10.18910/51420

[134] 福岡昌子，“イントネーションから表現意図を識別する能力の習得研究 —中国 4 方言話者を対象に自然・合成音声を使って,”日本語教育，No.96，pp.37–48，March 1998.

[135] 湧田美穂，“「い形容詞＋ナイ」の表現意図と韻律的特徴 —北京語・上海語話者を対象とした録音実験から,”日本語教育と音声，戸田貴子（編著），pp.183–208，くろしお出版，東京，2008.

[136] 鮎澤孝子，“イントネーションと日本語教育,”日本語学，Vol.10，No.7，pp.98–113，July 1991.

[137] 前川喜久雄，“イントネーション研究発展の要因,”音声研究，Vol.10，No.3，pp.7–17，Dec. 2006.
DOI:10.24467/onseikenkyu.10.3_7

[138] 岡田祥平，江崎哲也，“「文末音調と発話意図とを統合したアノテーション」を施した音声コーパスを考える際に必要となる視点は何か? —「同意要求表現」を中心に,” 第 1 回コーパス日本語学ワークショップ予稿集，pp.329–338，March 2012.

[139] Y. Yamashita, “A review of paralinguistic information processing for natural speech communication,” Acoust. Sci. & Tech., Vol.34, No.2, pp.73–79, Feb. 2013.
DOI:10.1250/ast.34.73

[140] 音声入出力方式標準化専門委員会，JEITA IT-4006 日本語テキスト音声合成用記号，電子情報技術産業協会，東京，2010.

[141] 宮島崇浩，菊池英明，白井克彦，大川茂樹，“演技指示の工夫が与える音声表現への影響 —表現豊かな演技音声表現の獲得を目指して,” 音声研究，Vol.17，No.3，pp.10–23，Dec. 2013.
DOI:10.24467/onseikenkyu.17.3_10

[142] J.A. Russell, “A circumplex model of affect,” J. Personality and Social Psychology, Vol.39, No.6, pp.1161–1178, Dec. 1980.

[143] H. Kawahara, I. Masuda-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 Commun., Vol.27, No.3–4, pp.187–207, April 1999.
DOI:10.1016/S0167-6393(98)00085-5

[144] 全　炳河，大浦圭一郎，能勢　隆，山岸順一，酒向慎司，戸田智基，益子貴史，ブラックアラン，徳田恵一，“HMM 音声合成システム（HTS）の開発,” 信学技報，SP2007-147，Dec. 2007.

[145] J.C. Acosta and N.G. Ward, “Achieving rapport with turn-by-turn, user-responsive emotional coloring,” Speech Commun., Vol.53, No.9–10, pp.1137–1148, Nov.–Dec. 2011.
DOI:10.1016/j.specom.2010.11.006

[146] 加瀬嵩人，能勢　隆，千葉祐弥，伊藤彰則，“発話状態推定に基づく協調的感情音声合成による音声対話システムの評価,” 信学論（A），Vol.J99-A，No.1，pp.25–35，Jan. 2016.

[147] J. James, B.T. Balamurali, C.I. Watson, and B. MacDonald, “Empathetic speech synthesis and testing for healthcare robots,” Int. J. Social Robotics, Vol.13, No.8, pp.2119–2137, Dec. 2021.
DOI:10.1007/s12369-020-00691-4

[148] N. Hojo, Y. Ijima, H. Sugiyama, N. Miyazaki, T. Kawanishi, and K. Kashino, “DNN-based speech synthesis considering dialogue-act information and its evaluation with respect to illocutionary act naturalness,” Proc. Speech Prosody 2020, pp.975–979, Tokyo, Japan, May 2020.
DOI:10.21437/SpeechProsody.2020-199

[149] 金田一春彦，日本語音韻の研究，東京堂出版，東京，1967.

[150] 粕谷英樹，楊　長盛，“音源から見た声質,” 音響誌，Vol.51，No.11，pp.869–875，Nov. 1995.
DOI:10.20697/jasj.51.11_869

[151] K. Maekawa and H. Mori, “Voice-quality analysis of Japanese filled pauses: A preliminary report,” Proc. 7th Workshop on Disfluency in Spontaneous Speech, pp.61–64, Edinburgh, Scotland, Aug. 2015.

[152] Z. Wu, O. Watts, and S. King, “Merlin: An open source neural network speech synthesis system,” Proc. 9th ISCA Workshop on Speech Synthesis, pp.202–207, Sunnyvale, USA, Sept. 2016.
DOI:10.21437/SSW.2016-33

[153] J.H. Ward, Jr., “Hierarchical grouping to optimize an objective function,” J. Am. Statistical Assoc., Vol.58, No.301, pp.236–244, March 1963.

研究業績

【論文】

1) 白井克彦，岩田和彦，“音声合成のための単語の強調表現の規則化,” 電子情報通信学会論文誌 A，Vol.J70-A，No.5，pp.816–821，May 1987.

2) 岩田和彦，小林哲則，“対話音声合成の表現力向上に向けた文末詞と音調の組合せによる発話意図の表現に関する実験的検討,” 電子情報通信学会論文誌 D，Vol.J100-D，No.11，pp.938–948，Nov. 2017.
DOI:10.14923/transinfj.2017JDP7010

3) 岩田和彦，小林哲則，“対話音声合成の表現力向上に向けた文末音調の制御による付加的なニュアンスの表現に関する実験的検討,” 電子情報通信学会論文誌 D，Vol.J102-D，No.6，pp.442–453，June 2019.
DOI:10.14923/transinfj.2018JDP7055

4) 岩田和彦，小林哲則，“異なる音声表現の調和を保つことに力点を置いた対話音声合成のための音声収集手法の設計,” 電子情報通信学会論文誌 D，Vol.J106-D，No.1，pp.57–65，Jan. 2023.
DOI:10.14923/transinfj.2022JDP7026

5) 白井克彦，小林哲則，岩田和彦，深沢克夫，“ロボットとの柔軟な対話を目的とした音声入出力システム —WABOT-2 における会話系,” 日本ロボット学会誌，Vol.3，No.4，pp.104–114，Aug. 1985.
DOI:10.7210/jrsj.3.362

6) 高津弘明，福岡維新，藤江真也，岩田和彦，小林哲則，“会話によるニュース記事伝達のための音声合成,” 人工知能学会論文誌，Vol.34，No.2，pp.B–I65_1–15，March 2019.
DOI:10.1527/tjsai.B-I65

【国際会議】

7) K. Shirai, K. Iwata, and T. Ohno, “Pitch contour control in Japanese conversational speech,” Proc. 1986 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp.2043–2046, Tokyo, Japan, April 1986.
DOI:10.1109/ICASSP.1986.1168660

8) K. Iwata, K. Ozawa, Y. Mitome, and T. Watanabe, “A Japanese text-to-speech conversion system using pitch-controlled residual wave excitation,” Journal of the Acoustical Society of America, Vol.87, No.S1, p.S104, May 1990.
DOI:10.1121/1.2027795

9) K. Iwata, Y. Mitome, J. Kametani, M. Akamatsu, S. Tomotake, K. Ozawa, and T. Watanabe, “A rule-based speech synthesizer using pitch controlled residual wave excitation method,” Proc. 1st International Conference on Spoken Language Processing (ICSLP 90), pp.185–188, Kobe, Japan, Nov. 1990.
DOI:10.21437/ICSLP.1990-47

10) K. Iwata, Y. Mitome, and T. Watanabe, “Pause rule for Japanese text-to-speech conversion using pause insertion probability,” Proc. 1st International Conference on Spoken Language Processing (ICSLP 90), pp.837–840, Kobe, Japan, Nov. 1990.
DOI:10.21437/ICSLP.1990-107

11) K. Iwata and Y. Mitome, “Prosody generation models constructed by considering speech tempo influence on prosody,” Proc. 2nd International Conference on Spoken Language Processing (ICSLP 92), pp.1155–1158, Banff, Canada, Oct. 1992.
DOI:10.21437/ICSLP.1992-139

12) K. Iwata and T. Kobayashi, “Conversational speech synthesis system with communication situation dependent HMMs,” in Proceedings of the Paralinguistic Information and its Integration in Spoken Dialogue Systems Workshop, eds. R.L.-C. Delgado and T. Kobayashi, pp.113–123, Springer, New York, 2011.
DOI:10.1007/978-1-4614-1335-6_13

13) K. Iwata and T. Kobayashi, “Expressing speaker’s intentions through sentence-final intonations for Japanese conversational speech synthesis,” Proc. INTERSPEECH 2012, Mon.P2b.03, pp.442–445, Portland, USA, Sept. 2012. DOI:10.21437/Interspeech.2012-153

14) K. Iwata and T. Kobayashi, “Speaker’s intentions conveyed to listeners by sentence-final particles and their intonations in Japanese conversational speech,” Proc. 2013 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), SP–L9.5, pp.6895–6899, Vancouver, Canada, May 2013.
DOI:10.1109/ICASSP.2013.6638998

15) K. Iwata and T. Kobayashi, “Expression of speaker’s intentions through sentence-final particle/intonation combinations in Japanese conversational speech synthesis,” Proc. 8th ISCA Workshop on Speech Synthesis (SSW), OS6–4, pp.235–240, Barcelona, Spain, Aug. 2013.

16) T. Kobayashi, Y. Komori, N. Hashimoto, K. Iwata, Y. Fukazawa, J. Yazawa, and K. Shirai, “Speech conversation system of the musician robot,” Proc. ’85 International Conference on Advanced Robotics (ICAR), pp.483–488, Tokyo, Japan, Sept. 1985.

17) S. Doi, K. Iwata, K. Muraki, and Y. Mitome, “Pause control in Japanese text-to-speech conversion system with lexical discourse grammar,” Proc. 3rd International Conference on Spoken Language Processing (ICSLP 94), pp.743–746, Yokohama, Japan, Sept. 1994.
DOI:10.21437/ICSLP.1994-188

18) K. Takahashi, K. Iwata, Y. Mitome, and K. Nagano, “Japanese text-to-speech conversion software for personal computers,” Proc. 3rd International Conference on Spoken Language Processing (ICSLP 94), pp.1743–1746, Yokohama, Japan, Sept. 1994.
DOI:10.21437/ICSLP.1994-191

19) K. Takahashi, K. Iwata, Y. Mitome, and K. Nagano, “Japanese text-to-speech conversion software system,” Proc. 47th International Federation for Information and Documentation (FID) Conference and Congress, pp.194–197, Omiya, Japan, Oct. 1994.

20) I. Fukuoka, K. Iwata, and T. Kobayashi, “Prosody control of utterance sequence for information delivering,” Proc. INTERSPEECH 2017, pp.774–778, Stockholm, Sweden, Aug. 2017.
DOI:10.21437/Interspeech.2017-708

【国内講演】

21) 岩田和彦，阿部匡伸，小熊幸雄，白井克彦，“素片編集規則合成の実験用システム,” 昭和 59 年度電子通信学会総合全国大会講演論文集（分冊 6），1643，p.196，March 1984.

22) 岩田和彦，白井克彦，“規則合成システムにおける韻律パタンの生成,” 日本音響学会音声研究会資料，S84–45，pp.349–356，Oct. 1984.

23) 岩田和彦，大野　毅，白井克彦，“単語合成における音韻性を考慮したピッチパタン制御,” 日本音響学会昭和 60 年度春季研究発表会講演論文集，1–5–18，pp.137–138，March 1985.

24) 岩田和彦，大野　毅，白井克彦，“会話文章におけるピッチパタンについて,” 日本音響学会昭和 60 年度秋季研究発表会講演論文集，2–3–21，pp.195–196，Sept. 1985.

25) 岩田和彦，大野　毅，白井克彦，“会話文章における基本周波数パタンの制御規則について,” 日本音響学会音声研究会資料，S85–42，pp.317–324，Oct. 1985.

26) 岩田和彦，大野　毅，白井克彦，“対話文章合成のための韻律制御規則の検討,” 昭和 61 年度電子通信学会総合全国大会講演論文集（分冊 6），S26–2，pp.367–368，March 1986.

27) 岩田和彦，三留幸夫，伏木田勝信，スティーブクレイトン，“英語テキスト音声変換システムの試作,” 昭和 62 年電子情報通信学会情報・システム部門全国大会講演論文集（分冊 1），169，p.170，Nov. 1987.

28) 岩田和彦，市川昌子，小澤一範，渡辺隆夫，“残差制御による音声合成システムの検討,” 日本音響学会昭和 63 年度秋季研究発表会講演論文集，3–2–7，pp.183–184，Oct. 1988.

29) 岩田和彦，三留幸夫，小澤一範，“残差制御型規則音声合成システムにおける韻律制御規則の検討,” 日本音響学会平成元年度春季研究発表会講演論文集，2–7–17，pp.203–204，March 1989.

30) 岩田和彦，市川昌子，小澤一範，三留幸夫，渡辺隆夫，“残差制御型音声合成方式を用いた規則音声合成システム,”日本音響学会平成元年度春季研究発表会講演論文集，2–7–19，pp.207–208，March 1989.

31) 岩田和彦，小澤一範，三留幸夫，渡辺隆夫，“残差制御型合成方式を用いた日本語テキスト音声合成システム,”日本音響学会平成元年度秋季研究発表会講演論文集，3–P–18，pp.311–312，Oct. 1989.

32) 岩田和彦，小澤一範，三留幸夫，渡辺隆夫，“残差制御型合成方式を用いた規則音声合成システムの評価,”日本音響学会平成 2 年度春季研究発表会講演論文集，1–4–22，pp.221–222，March 1990.

33) 岩田和彦，三留幸夫，友竹世光，赤松　実，亀谷　潤，“残差制御型音声合成方式を用いた規則音声合成システムの試作,” 1990 年電子情報通信学会秋季全国大会講演論文集（分冊 1），A–127，p.129，Oct. 1990.

34) 岩田和彦，三留幸夫，友竹世光，赤松　実，亀谷　潤，小澤一範，渡辺隆夫，“残差制御型音声合成方式を用いた日本語テキスト音声変換システム,”信学技報，SP90–56，Vol.90，No.335，pp.15–22，Nov. 1990.

35) 岩田和彦，三留幸夫，“発声速度が韻律に及ぼす影響,”日本音響学会平成 4 年度春季研究発表会講演論文集，1–2–21，pp.247–248，March 1992.

36) 岩田和彦，三留幸夫，宮本　牧，渡辺隆夫，“自動通訳システム INTERTALKER における日本語音声合成,” 情報処理学会第 44 回全国大会講演論文集（分冊 3），6P–8，pp.225–226，March 1992.

37) 岩田和彦，三留幸夫，“発話テンポと種々の音声区間長との関係,”日本音響学会平成 4 年度秋季研究発表会講演論文集，1–5–12，pp.239–240，Oct. 1992.

38) 岩田和彦，三留幸夫，“発話テンポの影響を考慮した継続時間長制御モデル,”日本音響学会平成 5 年度春季研究発表会講演論文集，3–8–3，pp.241–242，March 1993.

39) 岩田和彦，高橋一裕，三留幸夫，永野敬子，“パソコン向けソフトウェア日本語テキスト音声合成,”日本音響学会平成 5 年度秋季研究発表会講演論文集，2–8–13，pp.245–246，Oct. 1993.

40) 岩田和彦，“文章朗読音声における韻律的特徴の分析,”日本音響学会平成 6 年度春季研究発表会講演論文集，2–Q–11，pp.285–286，March 1994.

41) 岩田和彦，三留幸夫，“発話テンポに依存しない韻律構造のモデル化,” 1994 年電子情報通信学会春季大会講演論文集（第 1 分冊），SA–5–2，pp.486–487，March 1994.

42) 岩田和彦，高橋一裕，三留幸夫，西浦　充，“マルチメディア・パソコンにおけるテキスト音声合成の利用,” 1995 年電子情報通信学会総合大会講演論文集（情報・システム 1），SD–9–8，pp.381–382，March 1995.

43) 岩田和彦，三留幸夫，“パソコン向け音声合成ソフトウェアを利用したコミュニケーション支援装置の試作,” 情報処理学会第 51 回全国大会講演論文集（分冊 1），4T–8，pp.373–374，Sept. 1995.

44) 岩田和彦，高野　優，磯　健一，“音声認識・合成ソフトウェアを利用した音声 I/F を持つ電子メールシステムの試作,” 1996 年電子情報通信学会総合大会講演論文集（情報・システム 1），D–656，p.215，March 1996.

45) 岩田和彦，小林哲則，“対話音声における文末音調と聞き手に伝わる発話意図との関係の分析,” 日本音響学会 2011 年秋季研究発表会講演論文集，2–8–5，pp.295–296，Sept. 2011.

46) 岩田和彦，小林哲則，“対話状況に応じた HMM を持つ音声合成システム,” 日本音響学会 2011 年秋季研究発表会講演論文集，3–Q–3，pp.355–356，Sept. 2011.

47) 岩田和彦，小林哲則，“対話音声における終助詞に応じた文末音調と聞き手に伝わる発話意図との関係,” 日本音響学会 2012 年秋季研究発表会講演論文集，3–Q–15，pp.393–394，Sept. 2012.

48) 岩田和彦，小林哲則，“終助詞とその音調とによって聞き手に伝わる発話意図の分析,” 信学技報，SP2012-77，Vol.112，No.281，pp.31–36，Nov. 2012.

49) 岩田和彦，小林哲則，“発話意図に様々なニュアンスを付加して伝える文末音調の分析,” 日本音響学会 2013 年秋季研究発表会講演論文集，3–P–24，pp.455–456，Sept. 2013.

50) 白井克彦，阿部匡伸，富田孝司，岩田和彦，“素片編集規則合成システムにおける素片の選択と補間法,” 日本音響学会音声研究会資料，S83–65，pp.509–516，Jan. 1984.

51) 市川昌子，岩田和彦，三留幸夫，伏木田勝信，“規則合成における単位音声セットの検討,” 信学技報，S87–6，pp.41–48，April 1987.

52) 市川昌子，岩田和彦，伏木田勝信，“波形編集型規則合成における単位音声セットの検討,”日本音響学会昭和62年度秋季研究発表会講演論文集，1–6–8，pp.171–172，Oct. 1987.

53) 市川昌子，岩田和彦，小澤一範，渡辺隆夫，“残差を用いた規則合成方式におけるピッチ制御の一検討,”昭和63年電子情報通信学会秋季全国大会講演論文集（基礎・境界グループ），A–19，pp.A–1–19，Sept. 1988.

54) 坂井信輔，村木一至，岩田和彦，“中間言語からの韻律情報の生成,”日本音響学会平成2年度春季研究発表会講演論文集，2–4–9，pp.243–244，March 1990.

55) 北風晴司，岩田和彦，三留幸夫，平野　哲，山崎淳二，阿部博之，“盲人用点字パソコンの研究開発,”第6回リハ工学カンファレンス講演論文集，pp.133–136，Aug. 1991.

56) 宮本　牧，三留幸夫，岩田和彦，“声帯音源波モデルを用いた規則音声合成,”日本音響学会平成3年度秋季研究発表会講演論文集，1–6–16，pp.221–222，Oct. 1991.

57) 土井伸一，亀井真一郎，村木一至，三留幸夫，岩田和彦，“機能語に着目した文内文脈構造抽出手法 (LDG) と音声出力の制御,”人工知能学会研究会資料，SIG–SLUD–9301–4，pp.27–34，May 1993.

58) 高橋一裕，岩田和彦，三留幸夫，永野敬子，森谷祐一，“パソコン向け音声合成ソフトウェア,”情報処理学会第47回全国大会講演論文集（分冊2），5V–6，pp.377–378，Sept. 1993.

59) 磯　健一，岩田和彦，野口　淳，畑崎香一郎，“音声認識／合成ソフトウェアのAPI開発,”1996年電子情報通信学会総合大会講演論文集（情報・システム1），D–654，p.213，March 1996.

60) 野口　淳，岩田和彦，畑崎香一郎，“音声認識・合成カスタムコントロールの開発,”1996年電子情報通信学会総合大会講演論文集（情報・システム1），D–655，p.214，March 1996.

61) 近藤玲史，岩田和彦，磯　健一，畑崎香一郎，三留幸夫，渡辺隆夫，水野正典，“パソコン向け音声認識合成プラットフォームの構築とアプリケーションの試作,”情報処理学会第53回全国大会講演論文集（分冊2），7N–9，pp.363–364，Sept. 1996.

62) 大町　基，岩田和彦，小林哲則，“距離感を与える音声の特徴分析と合成,”信学技報，SP2009-89，Vol.109，No.356，pp.159–163，Dec. 2009.

63) 大町　基，岩田和彦，小林哲則，“音声における距離感の変換方法の検討,”日本音響学会 2010 年春季研究発表会講演論文集，1–7–6，pp.295–296，March 2010.

64) 大町　基，岩田和彦，小林哲則，“音声における距離感の変換方法の評価,”日本音響学会 2010 年秋季研究発表会講演論文集，3–P–5，pp.339–340，Sept. 2010.

65) 大町　基，岩田和彦，小林哲則，“音声における距離感表現のための基本周波数パターン変換方法の検討,”日本音響学会 2011 年春季研究発表会講演論文集，1–Q–60(d)，pp.421–422，March 2011.

66) 大町　基，岩田和彦，小林哲則，“距離感を変換した合成音声の評価,”日本音響学会 2011 年秋季研究発表会講演論文集，3–Q–14，pp.385–388，Sept. 2011.

67) 大町　基，岩田和彦，小林哲則，“実環境における距離感変換音声の評価,”日本音響学会 2012 年春季研究発表会講演論文集，2–11–5，pp.343–344，March 2012.

68) 小林哲則，岩田和彦，“会話向け音声合成システム,”信学技報，SP2014-93，Vol.114，No.303，pp.19–24，Nov. 2014.

69) 福岡維新，岩田和彦，小林哲則，“発話系列を扱う会話音声合成,”信学技報，SP2016-74，Vol.116，No.414，pp.59–64，Jan. 2017.

70) 福岡維新，岩田和彦，小林哲則，“まとまりのある情報の伝達を目的とした会話音声合成システム,”日本音響学会 2017 年春季研究発表会講演論文集，1–6–6，pp.215–218，March 2017.

【その他著書】

71) 岩田和彦，高橋一裕，三留幸夫，“音声合成,” NEC 技報，Vol.47，No.8，pp.65–71，Sept. 1994.

72) 岩田和彦，桜井基宏，塚田　聡，吉田富士夫，小林平生，“音声認識の応用装置・システム,” NEC 技報，Vol.51，No.11，pp.99–102，Nov. 1998.

73) 岩田和彦，“音声認識技術とその応用,” NEC 技報，Vol.56，No.9，pp.19–22，Nov. 2003.

74) K. Shirai, T. Kobayashi, Y. Komori, N. Hashimoto, K. Iwata, Y. Fukazawa, and J. Yazawa, “Speech I/O system realizing flexible conversation for robot —The conversational system of WABOT-2,” Bulletin of Science and Engineering Research Laboratory, Waseda University, No.112, pp.53–79, Sept. 1985.

75) T. Watanabe, K. Iso, K. Iwata, K. Hatazaki, and Y. Mitome, “Speech interface,” NEC Research & Development, Vol.35, No.4, pp.426–432, Oct. 1994.

76) 塚田　聡，岩田和彦，桐山良雄，吉田富士夫，“大語彙音声認識技術と音声応答技術,” NEC 技報，Vol.51，No.11，pp.41–46，Nov. 1998.

77) 桐山良雄，岩田和彦，幅崎直行，友岡靖夫，高木啓三郎，辻　善丈，“電話音声認識応答装置 DS-X(T),” NEC 技報，Vol.53，No.6，pp.7–10，June 2000.

78) 白井克彦（編著），音声言語処理の潮流，コロナ社，東京，2010.
（3 章分担執筆）

【登録特許】

79) 岩田和彦，音声合成装置，日本電気株式会社，特許第 1945018 号，1995.06.23.

80) 岩田和彦，音声セグメンテーション装置，日本電気株式会社，特許第 2031088 号，1996.03.19.

81) 岩田和彦，音韻継続時間長決定装置，日本電気株式会社，特許第 2513266 号，1996.04.30.

82) 岩田和彦，音声合成方法および装置，日本電気株式会社，特許第 2062933 号，1996.06.24.

83) 岩田和彦，音声合成装置，日本電気株式会社，特許第 2551041 号，1996.08.22.

84) 岩田和彦，音韻継続時間長決定装置，日本電気株式会社，特許第 2581130 号，1996.11.21.

85) 岩田和彦，音声ピッチ抽出装置，日本電気株式会社，特許第 2638829 号，1997.04.25.

86) 岩田和彦，ピッチパタン生成装置，日本電気株式会社，特許第 2643408 号，1997.05.02.

87) 岩田和彦，テキスト音声合成装置，日本電気株式会社，特許第 2679623 号，1997.08.01.

88) 岩田和彦，音声収録装置，日本電気株式会社，特許第 2734028 号，1998.01.09.

89) 岩田和彦，ポーズ挿入位置決定方式，日本電気株式会社，特許第 2748445 号，1998.02.20.

90) 岩田和彦，ピッチパタン生成装置，日本電気株式会社，特許第 2785628 号，1998.05.29.

91) 岩田和彦，音声合成装置，日本電気株式会社，特許第 2910587 号，1999.04.09.

92) 岩田和彦，継続時間長決定方法，日本電気株式会社，特許第 2936773 号，1999.06.11.

93) 岩田和彦，ポーズ挿入位置決定装置，日本電気株式会社，特許第 3001210 号，1999.11.12.

94) 岩田和彦，ポーズ挿入位置決定装置，日本電気株式会社，特許第 3076047 号，2000.06.09.

95) 岩田和彦，テキスト音声合成装置，日本電気株式会社，特許第 3094622 号，2000.08.04.

96) 岩田和彦，音声認識型通信制御装置，日本電気株式会社，特許第 3327213 号，2002.07.12.

97) 岩田和彦，音声応答装置，日本電気株式会社，特許第 3601411 号，2004.10.01.

【表彰】

98) 日本電気株式会社（渡辺隆夫，三留幸夫，畑崎香一郎，坂井信輔，磯　健一，服部浩明，岩田和彦，篠田浩一，高木啓三郎，野口　淳，高橋一裕，山田栄子，永野敬子），“パソコン音声入出力ソフトウェアの開発,” 第 3 回技術開発賞，日本音響学会，May 1995.