

論文 / 著書情報
Article / Book Information

題目(和文)	合成音声評価のための音声の主観的特性に関する研究
Title(English)	
著者(和文)	稗田一郎
Author(English)	
出典(和文)	学位:博士(工学), 学位授与機関:東京工業大学, 報告番号:乙第3297号, 授与年月日:1999年3月31日, 学位の種別:論文博士, 審査員:
Citation(English)	Degree:Doctor (Engineering), Conferring organization: Tokyo Institute of Technology, Report number:乙第3297号, Conferred date:1999/3/31, Degree Type:Thesis doctor, Examiner:
学位種別(和文)	博士論文
Type(English)	Doctoral Thesis

合成音声評価のための音声の 主観的特性に関する研究

稗 田 一 郎

目次

第1章 序論	1
1.1 はじめに	1
1.2 音声合成の研究開発の現状と将来	2
1.2.1 音声合成技術の概要	2
1.2.2 将来展望	2
1.2.3 福祉応用の現状と将来	4
1.3 合成音声の評価の必要性和問題点	5
1.4 本研究の目的	5
1.5 論文の構成	6
第2章 合成音声評価の現状と問題点	11
2.1 はじめに	11
2.2 評価の工学的な意義	11
2.2.1 一般的な評価の特徴	11
(1) 社会生活における評価	11
(2) 評価を行う人	12
2.2.2 工学領域における評価	12
(1) 要素の評価と数値化	12
(2) 総合評価	13
2.3 合成音声の評価の現状と問題点	13
2.3.1 評価方法の分類	14
(1) 音声の構成単位に基づいた分類	14
(2) 情報の種類による分類	15
(3) 利用場面による分類	15
(4) 注意の分散の有無	16
2.3.2 合成音性評価の枠組み	17
(1) 音声の構成と情報の種類に着目した枠組み	17
(2) 利用する側の視点からの利用場面の分類	19
2.4 本研究の目的と位置づけ	22
2.5 章のまとめ	25
第3章 短期記憶による明瞭度評価法	27
3.1 はじめに	27

目次

3. 2	短期記憶による明瞭度評価法.....	28
3. 3	実験方法.....	30
3. 3. 1	実験装置.....	30
3. 3. 2	実験用音声の採取および作成.....	31
3. 3. 3	刺激提示手順.....	32
3. 4	実験.....	33
3. 4. 1	実験 1	33
(1)	実験方法.....	33
(2)	結果および検討.....	35
3. 4. 2	実験 2	35
(1)	実験方法.....	35
(2)	結果.....	41
3. 5	考察.....	41
3. 6	章のまとめ.....	44
第 4 章	単音節音声の音韻長の主観的等価性	47
4. 1	はじめに.....	47
4. 2	実験方法.....	48
4. 2. 1	実験の基本的な手続き.....	49
4. 2. 2	刺激音声の音韻継続時間の変化方法.....	49
(1)	連続変化.....	51
(2)	ランダム変化.....	52
4. 2. 3	刺激音声の作成.....	52
4. 2. 4	被験者.....	55
4. 3	単音節の主観的に等価な長さ	55
4. 3. 1	被験者ごとの弁別閾.....	58
4. 3. 2	単音節の主観的に等価な長さ	64
4. 4	比較刺激の刻み幅の最適値.....	65
4. 5	章のまとめ.....	68
第 5 章	合成音声の音質評価のための 主観評価項目の抽出.....	69
5. 1	はじめに.....	69
5. 2	実験方法.....	71
5. 2. 1	評価項目の選定.....	71
5. 2. 2	刺激系列の作成.....	71
5. 2. 3	刺激の作成法.....	75
5. 2. 4	評価実験の方法.....	76

5. 2. 5 主成分分析.....	83
5. 3 実験結果	83
5. 4 考察	85
5. 5 章のまとめ	88
第6章 音声の物理特性と心理的評価量の 関係解明の試み	89
6. 1 はじめに	89
6. 2 予備的分析	90
6. 2. 1 物理量.....	90
(1) ピッチおよびフォルマント.....	90
(2) スペクトルの積算値.....	93
(3) 音節パワー.....	93
(4) 立ち上がり, 立ち下がり of 急峻さと音韻継続時間.....	97
(5) 発話のテンポのばらつき.....	97
6. 2. 2 予備的分析の結果および検討.....	98
(1) ピッチおよびフォルマント.....	101
(2) スペクトル積算値.....	101
(3) 音節パワー, 立ち上がり, 立ち下がり of 急峻さ及び音韻継続時間.....	102
(4) 発話のテンポのばらつき.....	102
6. 2. 3 予備的分析のまとめ.....	102
6. 3 ピッチと発話テンポのばらつきの心理面への影響	102
6. 3. 1 男声, 女声サンプル間での差の検定.....	102
6. 3. 2 “高いー低い” のスコアの算出.....	103
6. 3. 3 発話テンポのばらつきの相互関係.....	109
6. 3. 4 相関係数の検定.....	109
6. 4 結果	109
6. 5 考察	109
6. 6 章のまとめ	112
第7章 結論	115
謝辞	119
参考文献	121
発表論文など	125
学会誌上	125

目次

学会口頭発表.....	125
その他の発表.....	125
学会誌上.....	125
学会口頭発表.....	126
その他.....	127

第1章

序論

1. 1 はじめに

手もとの国語辞典によれば，今日用いられる“評価”という言葉には大きく分けて3通りの意味がある¹⁾．科学技術分野において“評価”と言った場合，そのうちの“価値や成果を論じ決めること”の意味で用いられることがほとんどである．本研究の主題である，合成音声の評価方法の“評価”もこの意味で用いている．

社会生活の中で，評価の対象や，組織，社会全体にとって，公正で，有益な評価を行うことは難しい．科学技術，工学の領域に限定しても，適正で有意義な評価を行うことは容易ではない．近年では特に人類全体や地球環境にとって好ましいものかどうかといった地球規模の視点からの評価が問われるようになってきている．

本論文は，電子工学，情報工学の発展に歩調を合わせて着実に進歩してきた音声合成技術に対する評価技術について，生命工学工業技術研究所（およびその前身の製品科学研究所）にて行ってきたマンーマシンインターフェースの研究開発プロジェクトの成果の中からまとめたものである．

合成音声の評価方法の現状，問題点をふまえて，現在の評価方法全般の枠組みを明確にした上で，重要でありながら見落とされてきた領域，あるいは将来必要性の高まる領域に位置づけられる評価方法をいくつか提案し，実験を行って，その結果から提案する評価方法の有効性や問題点について考察を行っていく．検討を進めていく中で，評価の工学的な役割，重要性についても考えていきたい．

第1章

1. 2 音声合成の研究開発の現状と将来

1. 2. 1 音声合成技術の概要

音声合成は、今日ではマン・マシンインターフェースの重要な技術のひとつとなっている²⁻⁸⁾。音声合成の研究開発は、早くから音声応答装置の技術として始まり、コンピュータ、通信関連の技術の発展に伴いながら着実に進歩を続けており、今日では広く実用化されている。

音声合成の方法は、単音節、単語、文節あるいは短文の単位で録音した（今日では主にデジタル化した）自然音声（人の肉声）をつないで出力する編集合成方式、自然音声を何らかの分析手法によって経時的に必要な最低限のパラメータを抽出し、このパラメータから音声进行再合成する分析・合成方式、純粋に音声を合成する規則合成方式に大別できる。

編集合成方式は、録音（デジタル化）した自然音声をそのまま出力するので音質の面では一番良好であり、現在最も多く実用化されている。録音する単位を単音節よりは単語や文節、単語や文節よりはいくつかの連続した文節等と大きくとっていくほど音質（自然さ）も向上するが、同時に録音すべき組み合わせが増大し、必要なデータの記憶容量も増大して、これが従来の最大の制約であった。

しかし、近年、半導体メモリとハードディスク記憶装置は、激しい勢いで大容量化、高速化、高信頼性化、低価格化しており、単位容量あたりの価格の下落には著しいものがある。編集合成方式の一番の制約となっていた記憶容量とコストの関係が、大幅に改善されてきており、今後も実用的な装置には一番多く採用される方式と考えられる。

分析・合成方式は、もともと通信における音声情報圧縮技術として研究開発されてきた。1970年代前半にわが国で開発されたPARCOR分析・合成法⁹⁾は、得られる音質も良好な上、分析によって得られる係数から声道の形状を推定することも可能なため、今日でも広く利用されている。

規則合成方式は、人が成長する過程で自然に身につけたり、あるいは学習して得る、音声、言語の発生の規則をモデル化して、その規則に基づいて音声を純粋に合成する方式である。現段階では音質は他の方式に比較して劣るが、規則に従って合成するので汎用性が高い。印刷図書や文字だけの情報のデータベースを音読する、いわゆる人工知能の出力音声用として用いるのにふさわしい夢のある技術である¹⁰⁻¹²⁾。

1. 2. 2 将来展望

多くの国々の研究者たちの地道な努力により、着実な発展を遂げてきた音声合成技術であるが、研究開発を取り巻く技術的、経済的な情勢は、その発展の勢いをさらに加速するような状況になっている。

ひとつはコンピュータ関連技術の飛躍的な進歩である。近年の中央演算ユ

ニット（CPU または MPU）や周辺の集積回路の高性能化、半導体メモリ、ハードディスク装置の大容量化、低価格化は著しい。中でも半導体メモリ、ハードディスク装置の単位容量あたりの価格の下落は猛烈な勢いで進んでおり、10 年前はおろか、5 年前でさえ今日の状況は全く予測できなかったといえるほどである。

これまでは、音声合成装置の仕様づくりの際、記憶容量がコストに結びつき、大きな制約になっていたが、実用をめざした装置の開発ではこの制約が大幅に緩和されている。

一方、現在の世界的な経済の情勢も、今後音声合成技術を応用した装置のニーズを大きく後押しする要因になると考えられる。

ソビエト連邦の崩壊、冷戦の終結により、資本主義のもとでの市場競争が事実上唯一の経済原理となった。同時に、先進国が中心となって長年にわたり進めてきた経済の自由化がいよいよ最終段階に入り、金融等のサービス業界は国際間の厳しい競争に巻き込まれようとしている。

このような厳しい状況の中、各企業にとって、顧客の確保と新たな獲得のためにサービスの向上が至上命題となる。わが国の人件費の水準は、中進工業国に対してのみでなく、他の先進諸国と比較しても高水準であるが、雇用形態、物価の水準や医療、教育、社会福祉制度等の現状から人件費の水準の引き下げはあまり現実的ではない。高い人件費の水準を維持したまま経営の効率化、スリム化を行っていかねばならない。サービスの向上と経営の効率化という二律背反の課題を解決するために、顧客サービスを含む、業務内容のあらゆる面で大胆な自動化を進めていく必要がある。

金融業におけるサービスの改善で有力なのは、遠隔の顧客へのオンラインサービスの充実である。現在、オンラインサービスには、インターネットを利用したいわゆる“電子取引”と電話を利用した従来型のサービスがある。

インターネットを利用したサービスには、セキュリティ上の問題、高速通信回線の整備など多少の課題が残っているものの、動画を含む画像や音声、多種の字体、大きさの文字を用いるなど、いわゆる“マルチメディア”の将来性のある機能をそのまま利用できるため、今後の発展、普及が大いに期待されている。すでに一部の国内金融機関でも運用が開始されている。

一方の電話によるサービスは、電話のオペレーターや顧客係が行ってきた対応等の従来型のサービス形態であり、人件費削減、サービス拡大のためにこのサービスを可能な限り自動化しようというのが現在の風潮である。電話を通じた自動応答サービスの場合、ユーザーが利用できる入力装置は電話機の数字と二つの記号のボタンだけなので複雑なサービスには向かないが、国内外の出先からも、あるいは移動中でさえも利用できる、電子取引同様 24 時間年中無休体制でサービスが行えるという大きなメリットがある。将来、音声認識技術

第1章

が電話回線を通じた音声に対しても実用レベルに達し、入出力とも音声の利用を取り入れた、高度な自動応答システムが実用化すれば、インターネットを利用した電子取り引きが普及した後でも、存続する意義は十分にある。

わが国の人件費の水準の高さは、金融サービス業界のみならず、製造業でも大きな問題である。賃金の水準を下げることなく、わが国のこれまでの製造業の主力製品である、自動車、家電製品等の国際市場での競争力を維持していくためには、わが国が誇る製品の高い品質を維持すること、高付加価値化を今まで以上に進めることなどが特に重要である。合成音声を情報の出力メディアとして利用し、これらの製品のインテリジェント化を進めることは、高付加価値化の有効な手段の一つである。

このように、世界的規模の経済構造の変革により、電話による自動音声応答システム、情報の出力メディアなど、音声合成の応用技術には大きな需要が見込まれ、さらなる技術革新に期待が寄せられている¹³⁾。

1. 2. 3 福祉応用の現状と将来

音声合成技術の応用について、技術的、経済的な情勢から、大きな需要が期待できる状況にあるが、こうした変革の中では、社会的な弱者への配慮も重要である。

視覚障害者がパーソナルコンピュータなどの情報機器を利用する際、音声が重要なメディアとなることは言うまでもない。実際にこれまで合成音声装置を出力装置として備えた視覚障害者用のための読書器が開発され^{14,15)}、またワードプロセッサが開発されて製品として販売されてきた¹⁶⁾。こうした機器は、学習や就業のために、あるいはインターネットなどにより友人や社会とコミュニケーションをはかったり、情報収集を行ったりするために利用されている。

パーソナルコンピュータ、ワークステーションのマンーマシンインターフェースはここ数年でこれまでの文字単位の表示(CUI)からグラフィックスによる表示(GUI)に一気に変わったが、このことは視覚障害者や福祉機器の開発に携わる人たち等を不安にしている¹⁷⁾。CUIでは、ビデオ表示端末(VDT)は、たかだか数十の行、桁単位であり、通常、同じ字体、同じサイズの文字が表示されるだけなので、位置情報を認識し必要な情報をとりだすことはそれほど困難ではなかった。しかし、GUIでは様々な字体、サイズの文字が、図形と混在してドット単位で配置され、しかも外枠さえも浮動であるため、表示されている情報の構造や階層を理解したり、必要な情報を探し出すことが非常に困難である。健常者にとって、便利になる技術革新も、視覚障害者に対してはむしろ困難さを増加させている。

ところで、視覚障害者が利用できるようなパーソナルコンピュータ、ワークステーションやその他の情報機器を開発する方向としては、健常者が利用する最新の技術にわずかな付加装置、付加ソフトを組み込むことによって、利用可

能にするのが経済的であり、しかも視覚障害者のやる気も喚起できて、理想的といわれている。現在、先進諸国はほとんど例外なく財政難にみまわれ、福祉分野への補助金などの支出も今後大幅に増加することは期待できないことも背景にある。

このような見地から、音声合成装置や、位置情報を入出力するためのデバイスを付加して、視覚障害者が GUI のパーソナルコンピュータ、ワークステーションを利用可能にする技術の開発も行われている¹⁸⁾。また、パーソナルコンピュータのハードウェアや OS などのファームウェア等、基本的な部分についても設計段階で、障害者が利用しやすい用にするためのインターフェースを組み込んだり、組み込みやすくするなどの配慮がされては始めている。

技術を通して社会に貢献することが工学者の使命であるから、直接福祉機器の開発に直接関与しなくても、社会的弱者に対して何ができるか常日頃から念頭に置きたい。

1. 3 合成音声の評価の必要性和問題点

このように、着実に発展していて、しかも将来的な応用、展開が大いに期待されている音声合成技術であるが、社会生活、日常生活の中に定着させていくためには、使用する条件を考慮した評価の指針、およびそれに基づいた評価方法の策定が不可欠である。

合成音声に求められるもっとも基本的な条件は、音節、単語、短文などの文字的情報が正確に伝達されること、つまり明瞭性、了解性が高いことである。開発側が実験室、研究室で音声合成装置単体によって行う明瞭性、了解性の試験の条件と、実際の利用現場で音声合成装置がシステムに組み込まれた条件とでは明瞭性、了解性そのものに大きく影響するような重要な条件の違いがあっても、見過ごされることがある。実際の利用場面における合成音声の果たすべき役割、重要度などを十分に分析、把握して、実験条件を整える必要がある。

また、合成音声の感性、心理面への総合的な影響、疲労との関係等といった、明瞭性、了解性以外の特性については、研究が緒についたばかりである。今後基礎的な研究を積み重ねていく必要がある。

これらの合成音声評価の必要性和問題点については次章で詳しく述べる。

1. 4 本研究の目的

本研究の目的は、合成音声の品質向上に資するために、その評価方法の枠組みを明確にし、その中で、現状では欠けていたり、改善が必要な部分についていくつかの評価方法を提案し、実験的にその有効性を確認していくことである。

そのため、まず評価全般の工学的な意義について明らかにしてから、音声の

第1章

持つ構造的な特質と、音声伝達情報の種類による合成音声評価の枠組みを構成し、その中に現在行われている合成音声の評価方法について位置づけを行う。その枠組みによって、現在評価方法がない領域、あるいは評価方法の改善の必要のある領域の明確化をはかる。この他、利用者側からの視点、注意の分散の有無などによる評価方法の分類・位置づけの必要性について述べる。

枠組みの中で、必要性が明らかになった評価のうち、明瞭性の評価について新たな評価方法を提案して、実験的にその有効性を確認する。また、心理面への影響の評価として、単音節音声の音韻長の心理的に等価な長さ、合成音声の感性全般に及ぼす影響について実験的に検討を行う。

1. 5 論文の構成

本論文は、図1. 1に示すように、全7章で構成した。

第1章「序論」では、合成音声の研究開発の現状と将来について述べ、その中での合成音声の評価の必要性と問題点について触れている。また、本論文の構成を説明している。

第2章「合成音声評価の現状と問題点」では、音声合成研究の中での合成音声評価の必要性、現状における問題点について、日常生活で一般に行われている評価、科学技術分野で行われる評価との関連から詳しく論じる。こうした議論をふまえ、合成音声の評価の枠組みを明確にし、本研究の位置づけおよび必要性を明らかにする。

第3章「短期記憶による明瞭度評価法」では、新たな評価方法を提案し、実験によりその有効性を確認する。

明瞭性の評価法は、単音節明瞭度試験に代表されるように、文字情報の伝達の正確さを測る評価方法として、音響機器、設備、通信機器の評価に広く用いられ、合成音声の評価用としても定着している。これらの評価方法では、被験者は合成音声を聞き取ること集中できるので、多少品質の悪い音節についても、音響的ないろいろな手がかりから正答を推測することもできるうえ、慣れや学習の効果も現れやすい。

しかし、実際に合成音声を利用する場合、利用者は、音声を聞きながら、同時に判断や、操作・運転を行っているのが一般的であり、現在多く行われている明瞭性や了解性の試験のように、出力音声の文字的な情報だけに集中できるという条件は、むしろ特殊である。

そこで、本研究では、明瞭性の試験で注意の分散を生じさせるために、6～8の単音節を連続的に提示して再生する明瞭度評価法を提案する。6～8個の単音節が連続して提示された後、被験者は回答用紙に聞き取った単音節を筆記する。被験者は聞き取った単音節を忘れないように頭の中で繰り返しなが、続く単音節を聞き取りながら筆記するため、音声の聴取だけに集中することは

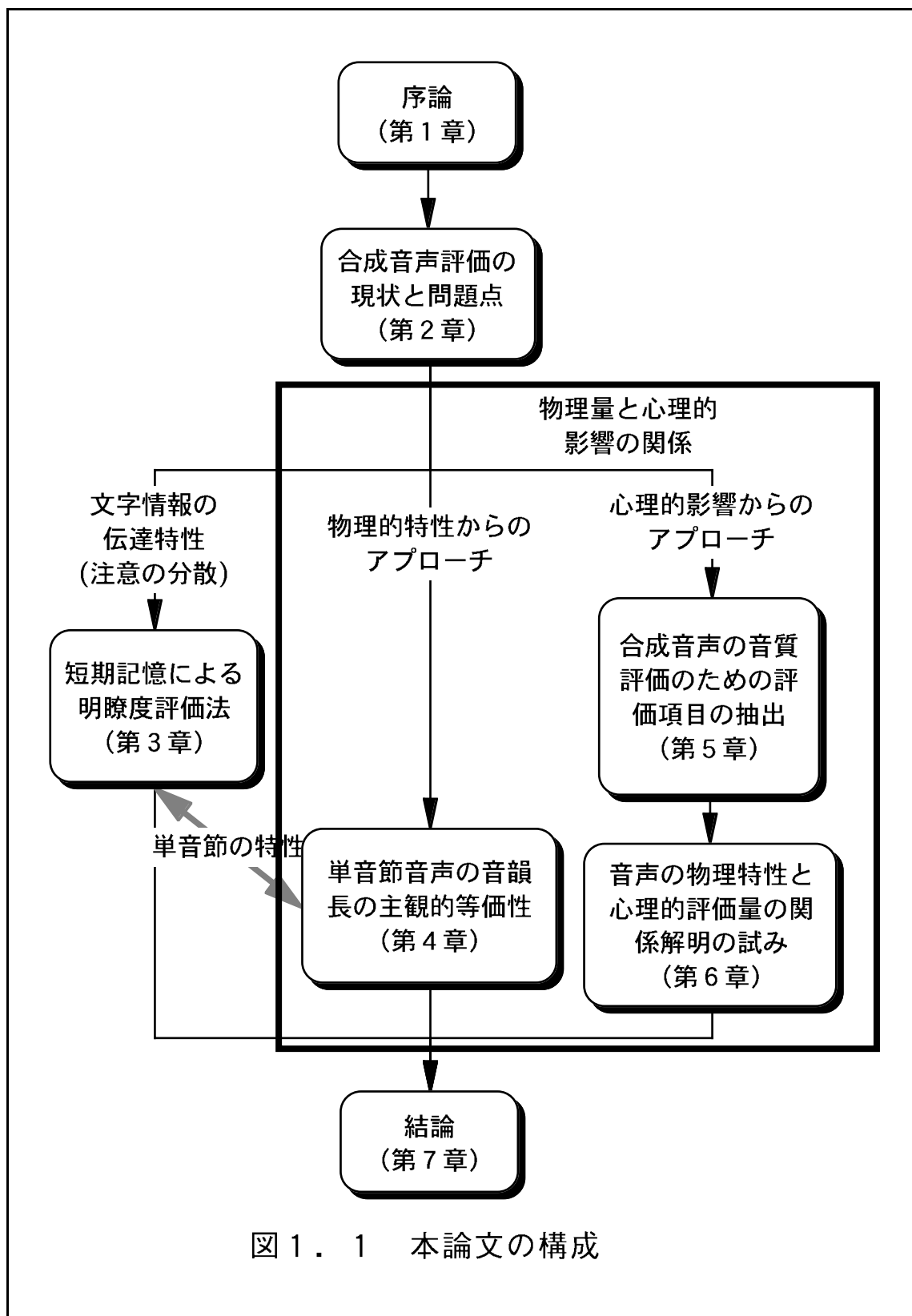


図 1. 1 本論文の構成

第1章

できない。

被験者を使った測定実験を行い、その結果について検討を加える。

第4章「単音節音声の音韻長の主観的等価性」では、広い意味で合成音声の評価のひとつといえる、音声の物理的な特性と心理的な特性の関係のうち、これまで調べられていない単音節の音韻継続時間の物理的な長さと心理的な長さの関係を調べる実験について説明し、その結果について検討を行う。

音声の物理的な特性とその心理的な影響、例えば音声の強弱、高低、長短等と、心理的に受ける強弱、高低、長短の印象との間にどのような関係があるかということは、マンーマシンインターフェースとして音声を用いる上で基本的な情報である。これらの関係の中で日本語の単音節の音韻継続時間の心理的影響についてはこれまで調べられていない。

そこで被験者に、母音の/a/を基準音声として最初に聞かせ、続いて比較する単音節を聴取させて長短を答えさせる実験を行い、単音節の長さの知覚の閾値、異種の単音節の心理的に等価な長さ等について検討する。

この検討結果は、単音節単位の編集合成方式における音韻継続時間の指標や、第3章で提案した短期記憶による明瞭度評価法を含む明瞭性試験に用いる標準音声の作成の指標にもなると期待される。

第5章「合成音声の音質評価のための主観評価項目の抽出」では、これまであまり行われてこなかった、合成音声の感性への総合的な影響の評価を試みる。

音声合成は実用化され広く普及しているが、社会生活の中に受け入れられていくためには、了解性、明瞭性が高いだけでは不十分で、感性への影響も総合的に考慮する必要がある。また、産業界においても、製品の付加価値を高めたり、サービスの質の向上をはかる上で感性への影響を測ることは非常に重要な課題である。

そこで、無意味つづり音声を用いて、意味内容や、感情などの付加価値によらない、音声の質（声質）そのものが人に心理的に及ぼす影響についての実験を行う。

実験では、男女の話者が読み上げた無意味つづり音声を、一人分ずつ被験者に聞かせて主観的に評価させ、その結果を主成分分析にかけて検討する。分析の結果、合成音声の感性への影響を評価する指標が抽出されると期待される。

第6章「音声の物理的特性と心理的評価量の関係解明の試み」では、さらに音声のどういう物理的な特性が心理面に影響を与えるか分析を試みる。

第5章の実験で用いた無意味つづり音声について、ピッチ、音節の立ち上がりのタイミングのばらつき等について分析を行い、その結果と第5章で得られた主観評価実験の結果を比較することで、主観評価と物理量の関係について調べる。

分析、検討結果から自然音声を分析して主観評価の結果との関係を調べるこ

とにより主観評価と物理量の間係を調べるという手法の有効性自体についても吟味する。

第7章「結論」では、本研究で得られた成果を総括する。

第2章

合成音声評価の現状と問題点

2.1 はじめに

本章では、はじめに評価全般の工学的分野での必要性や意義について述べる。

続いて、合成音声の評価の現状と問題点を明確にするために、音声の持つ構造的な特質に着目した分類、音声が伝達する情報の種類に基づく分類、および利用者の視点からの分類によって合成音声評価の枠組の構成を試みる。その枠組みの中で、現在行われている合成音声の評価方法を位置づけ、従来の評価方法でカバーしきれない領域、改良の必要な箇所、将来的に重要となると予想される領域等を明らかにしていく。

本論文で提案する評価方法を、構成した合成音声評価の枠組みの中に位置づけ、本研究の意義を明確にする。

2.2 評価の工学的な意義

2.2.1 一般的な評価の特徴

(1) 社会生活における評価

“評価”は人が考えて行うことなので本質的に主観的である。しかし、学業の評価、勤務評定などのように、評価は本人の進路や人生に重大な影響を与えることが多く、公正な評価が求められる。

公正な評価、言い換えれば誰が行っても同一の項目について同一の評価結果が得られるような評価を行うためには、評価基準、手順、その他の方法を明確（必要に応じて明文化）にする必要がある。また、評価をする人の教育、訓練が必要な場合も多い。

例としてスポーツ競技における判定、審判等があげられる。競技の多くは、

第2章

明文化されたルールに基づいて時間や得点を競い、誰の目からも明確に勝敗、順位が定まる。それでも微妙な判定は審判員に委ねられることが多い。時間や、得点を競う競技でも、審判員の判定が勝敗、成績に影響する場面は避けられない。

近年では、政治、経済、科学技術など社会生活の中で行われている評価全般について、これまでの自国の利益優先、経済成長優先といった視点の評価から、地球環境、人権を重視した視点の評価が問われるようになってきている。

(2) 評価を行う人

より公正で再現性の高い評価を行う上で、最も重要なポイントはいうまでもなく評価を行う“人”である。評価を行うためには経験、知識、即断力、より客観的な見方等が重要であり、教師、スポーツ競技の審判員等のように専門家に評価が委ねられることが多い。

公正で再現性の高い評価を行うのに必要な条件は、教育、訓練によって身につけられるものも多いが、先天的な能力（あるいは資質）が必要になる場合もある。例として、香料の調合、酒類の評価など味嗅覚による評価があげられる。

人間工学的研究により、味嗅覚の閾値（感度）は個人差が非常に大きく、年齢による違いがあることも明らかになっている¹⁹⁻²¹⁾。また、少なくない割合の人たちが、先天的にある種の味や臭いを感じられないことも明らかになっていて、香りや味の評価に先天的な能力が必要な場合のあることが裏付けられる。

このように評価を行う場合、専門的な教育、訓練が必要となり、場合によっては資質が必要とされることもある。その一方で、市場経済では購買者である“一般の人”を基準にした評価も重要である。趣味嗜好、生活の様式がますます多様化していく中、限られた調査対象（サンプル）から、市場戦略を打ち立てるのに有効な調査結果を得るのは困難な課題である。

2.2.2 工学領域における評価

社会生活全般における評価の特質について述べてきたが、科学技術、工学の領域でも評価の重要性は全く変わらない。また、研究・開発に対しても、人類全体への貢献、地球環境にとって有益であるか否かといったグローバルな見地からの評価が行われるようになってきている。

(1) 要素の評価と数値化

科学技術の分野では究極的には、“人類へ貢献”、“地球環境に優しい”という視点からの総合的な評価が行えればよい。しかし、たいていの場合、評価の対象となる事がらは多面的な要素を持っており、その要素の全てを的確にとらえて、総合的な評価を下すことは容易ではない。

そこで、総合評価を行う前の段階として、要素別の評価がよく行われる。細かい要素について着目すれば、再現性の高い評価を行いやすく、評価結果を物理量として数値で表すことができることもある。物理量を数値化して評価する

手法は“客観評価”と呼ばれ、実用化されているものも多い。

例として騒音の評価が挙げられる。一般的な騒音については、測定点の音圧レベルと、人がうるさく感じる度合いとの間の単調な関係が知られている。そこで、機器で測定した数値を騒音の指標そのものとして用いることができる²²⁾。しかし、騒音にも、可聴周波数よりも低い周波数成分が主となる超低周波騒音のようなものもある。この超低周波騒音は、人体へ悪影響を及ぼすことは明らかでありながら、数値化の方法が確立していない^{23,24)}。

臭いも、空気中に含まれる特定の物質の濃度に関係があることは明白であり、特定の臭いや香りについては測定装置が実用化されている。しかし、人の嗅覚は現在の化学（ガス）センサよりも一般に高感度で選択性も強く、しかもそれぞれの臭いの原因となる物質に対する閾値の個人差も大きいため、例えば“悪臭”といういろいろな要素を総合した評価を数値で表すのは容易ではない²⁵⁻²⁸⁾。

評価が完全に客観評価に置き換えられれば、規定されている通りに測定条件を整えることで、再現性の高い測定ができるので、科学技術の分野の評価としては好ましいが、客観評価には問題点や限界もあるので、制限等を十分理解した上で、数値として得られた結果を利用する必要がある。

物理量として測定できず、評価する人の主観で評価を行う場合でも、要素ごとに評価を行うほうが容易であり、再現性の高い結果を得やすい。

（２）総合評価

このようにして要素ごとに得た評価を総合して総合評価を得るのが一般によく用いられる方法である。

この手法の問題点として、要素ごとの総合評価に寄与する重みはそれぞれ異なるはずであるが、その重みを定量化することが難しいこと、評価しやすい（数値化しやすい）要素だけが選ばれ、総合評価に対して本来は重要な意味を持つ要素が見落とされる危険があること、などが挙げられる。

要素ごとに得られた評価から総合評価を直接求めるのが困難な場合は、いくつかの類似性のある要素をグループ化して、グループごとに評価を集約し、さらにそれらを総合して、最終的な総合評価とする方法も用いられる。この中間レベルの評価が複数の段階となることもある。

2.3 合成音声の評価の現状と問題点

合成音声に対して行われる音声の品質の評価は、これまでは、研究開発の目標への到達の度合いを測る自己評価的な性質が強かった。しかし、音声合成の研究開発が進み、実用的な装置が多く普及してきている今日、合成音声の特質をふまえ、再現性の高い評価方法の策定が重要な課題となってきた。

現在、合成音声に対して、文字情報の伝達の正確さを評価する明瞭性、了解性の試験がもっとも一般的に行われており、その他、自然性の評価、利用場面

第2章

に応じた評価などが試みられている^{29, 30)}。これらの現在行われている評価方法について、合成音声（または音声）のもつ特徴に基づいた視点から分類を行い、それぞれの評価方法の特徴、問題点等について検討を進めていく。

2.3.1 評価方法の分類

(1) 音声の構成単位に基づいた分類

これまで、合成音声の評価の分類は、主に音声の構成単位に対応させる形で行われてきた。音声の構成は、波形から文章へと、ミクロからマクロへの区分が容易であり、理解しやすいためである。

音声の言語学的な最小単位は音素とされるが、言語の単位に満たない部分的な波形、あるいは言語的な概念とは無関係な音声の物理的性質も心理的な影響を及ぼしていることは容易に想像される。そこで本論文では便宜上、こうした性質を“音声波形”または“波形レベル”の性質と呼び、音素よりもミクロな音声の単位とする。音素の次に大きい単位の考え方は、言語により異なるが、日本語では一般に単音節と考えられている。さらに、連音節、単語、文節、短文、文章の順で大きくなる。

既に述べてきたように、評価とは人への影響について、価値観の伴う判定を行うことであるが、意味を広げて、人への何らかの影響が予想される特性の分析ととらえると、既に確立して、広く用いられている、周波数スペクトル、フォルマント、ピッチ、ケプストラムなど²⁻⁶⁾の分析法も、音声波形レベルに対する一種の評価と考えることができる。

話者や、言語の特徴的な特性、自然性や感性へ影響を及ぼす要因の多くも、この音声波形のレベルに含まれていると考えられる。この音声波形の分析方法は、音声情報の圧縮技術、話者・音声認識技術の基礎的な研究開発として、数多くの手法が提案され、定着しているもの多い。

音素、単音節、連音節に対して行われる評価は明瞭性の評価が主である³¹⁻³³⁾。最も代表的なものが、単音節明瞭度試験である³⁴⁾。日本語の100個の単音節をランダムに音響装置で再生（提示）して被験者に聴取させ、書き取らせて正答率をみる。ホールなどの音響設備や通信回線等の試験用に開発されたが、合成音声の明瞭性の評価のためにも多く用いられている。

この他、残響の影響に配慮した連音節の明瞭度試験、母音・子音間の移行区間の特性を重視した、VCV（子音 C を同じ母音 V で挟んだ）などの連音節の明瞭度などが考案されている。

単語（または文節）や短文に対して行われる評価としては、主として了解性の評価が挙げられる。了解性の評価は、あらかじめ用意された、試験用の単語、短文リストの中からランダムに単語や短文が出力されて、これを聞き取る形式で行われる。リストの内容は限定的であり、単語や短文は単音節よりも情報の冗長性³⁵⁾が大きいので、音質が同じ場合、一般に単音節の明瞭性試験の結果と

比較して、高い正答率が得られる。しかし、後述するように、すべての利用者が、出力される単語、短文を予測可能という状況は特殊であり、研究開発上の指標とはなり得るが、実用的な評価としては十分でない。

文章のレベルには、感情や読み方の巧拙など、話者の特徴的な特性が現れる。

(2) 情報の種類による分類

人類が文字を発明して以来、音声の言語と文字の言語は、互いに強く影響しながら進化、発展を遂げている。この観点から、合成音声の評価についても、文字情報に関するものと、文字情報に関係しないものに分類することが可能である。文字情報の伝達特性の評価として、単音節、連音節に対する明瞭性の評価、単語や短文の了解性の評価があげられる。工学的には合成音声を情報出力メディアのひとつとして位置づけるので、文字情報の伝達の正確さが、もっとも重要視されている。

文字情報以外の情報としては、感性に影響する付加的な情報、話者の特徴、このほか、現時点では明確な関係が明らかでない物理的な性質なども含まれる。文字情報の占めるウェートが非常に大きく、その他の付加的な情報の中で、文字情報と並べられるような項目はないので、本論文では、これらを文字情報に対する、“その他の付加的情報”としてひとまとめにして扱うことにする。

これらの付加的な情報は、自然音声については“声質”と呼ばれるもので³⁶⁾、実体は物理的な諸々の特性である。これらの特性が原因となって、聞き取る人へ、感情やその他の感性の情報を伝えたり、心理面へいろいろな影響を及ぼしたり、また、聞き取る人の疲労度へ影響を及ぼすこともある。

具体的な例として、アクセント、イントネーションなどは、主にピッチ、振幅のパターンと発声のタイミングの変化といった物理的特性であることが知られている。アクセントやイントネーションは、文字情報伝達を助けるための冗長さとしての役割もあるが、話者の特徴や、そのときの感情などといった付加的な情報としての役割が大きい。

その物理的な特性と、感性や心理への影響の関係を解明していくことにより、サービスの質を高め、製品により高度な付加価値をつけていくという、工学的な波及効果も期待できる。

(3) 利用場面による分類

実用的な装置の開発を目指すうえで、利用者側の視点に立ち、利用場面によって合成音声の用途、必要な仕様を分類して、これらに対応した評価方法を整備していくことも重要である。こうした視点からの評価方法の研究は緒についたばかりである。これまでに利用する条件に配慮した実用的な評価としては、新聞の校正作業等にわずかに例がある程度である³⁷⁾。

利用する側の立場から見て、合成音声の評価するための重要な要素として、利用者のその装置への慣れ、熟練の度合い、および音声の情報メディアとして

第2章

の重要度のふたつが挙げられる。

合成音声は、聞き慣れてくると、もともとの品質があまり高くなくても、聞き取れるようになることが知られている。この性質は、利用する人が限られた作業や、出力される単語や文章が限定的で、利用者がある程度出力を予想できる場合などには有効であり、合成音声の質は多少低くても実用上支障ない。しかし、公共の施設、設備では、出力される単語、文章が限定的であっても、はじめて利用する人、聴覚の機能が低下しつつある高齢者、日本語（合成音声装置が出力する言語）を母国語としない人、文化的背景の異なる人等の利用にも配慮する必要があると、高い音声の品質が要求される。

また、電話による音声応答システムのように、合成音声による情報が唯一の情報なのか、あるいは、画像表示装置があって、画像情報と合成音声が併用されているのか、あるいは画像情報が主で合成音声はあくまでガイド的な機能にとどまるのかによっても、合成音声に要求させる品質の高さは異なり、評価の基準も変わってくる。

（４）注意の分散の有無

音声の構成、性質からの評価の枠組み、利用場面による分類からも明確になりにくい、もうひとつの重要な要素が、“注意の分散”である³⁸⁾。

合成音声は、情報出力メディアの一形態である。音声には情報メディアとして、一過性で蓄積が効かないという特性があるので、実用場面で利用者は、受け取った情報をその場で、あるいは受け取りながら利用するということが最終的な目的となる。言い換えれば、利用者は音声を聞きながら、全体の意味内容を考えたり、次の操作のための判断等を行う。合成音声を聴取しながら、同時に別のタスクを（時には複数）行っている。

自動車の運転中に携帯電話を使用したために起こる交通事故がひとつの社会問題となっているように、同時にいくつかの操作や思考を行うと注意の分散が生じ、それぞれの操作や思考（タスク）の処理のパフォーマンスが低下する。このことは日常生活でしばしば経験する。

このように実用的な場面では、合成音声を聴取している際に他のタスクを行っているのがむしろ一般的である。しかし、現在行われている合成音声の品質の評価で、注意の分散の影響について配慮したものはほとんどない。多くの評価法では、聞こえてくる音声だけに集中することができる。したがって、品質が悪い場合でも、いろいろな手がかり[†]から正答を推測することもできる上、集中できる分、学習による効果も大きくなる。

このように、注意の分散は、操作、作業の安全性、効率を考える上で見落とすことのできない重要な問題である。実用的な場面における条件を想定した評

[†]音響特徴(acoustic cue)と呼ばれる²⁹⁾。

表 2. 1 合成音声の評価において今後検討が必要な課題

注意の分散	作業，操作，運転，判断，思考等を行いながら合成音声を聞く場合の明瞭性，了解性やその他の特性
感性への影響	製品やサービスの付加価値を高め，利用者が定着，拡大するために必要な感性への影響
疲労	合成音声を長時間聞き続けて作業を行うときの，疲労とその合成音声の特性の関係
高齢者・障害者への配慮	高齢者，障害者の聴覚特性，視覚特性等へ配慮した合成音声や，合成音声を利用した装置のガイドラインづくり
ソフト面	初心者も熟練者も利用する可能性のある場合，熟練度や利用者の要求に応じて冗長性を変えたり，繰り返しを行うなどが可能な柔軟なソフトウェア

価方法を設定するうえで考慮していく必要がある。

注意の分散，感性への影響の他にも，表 2. 1 に示すように，長時間の作業における疲労への影響，初心者，熟練者に対応したソフトウェア，高齢者や障害者への対応などについて，合成音声の評価のガイドラインや方法を策定する上で検討が必要である³⁹⁾。

2. 3. 2 合成音性評価の枠組み

(1) 音声の構成と情報の種類に着目した枠組み

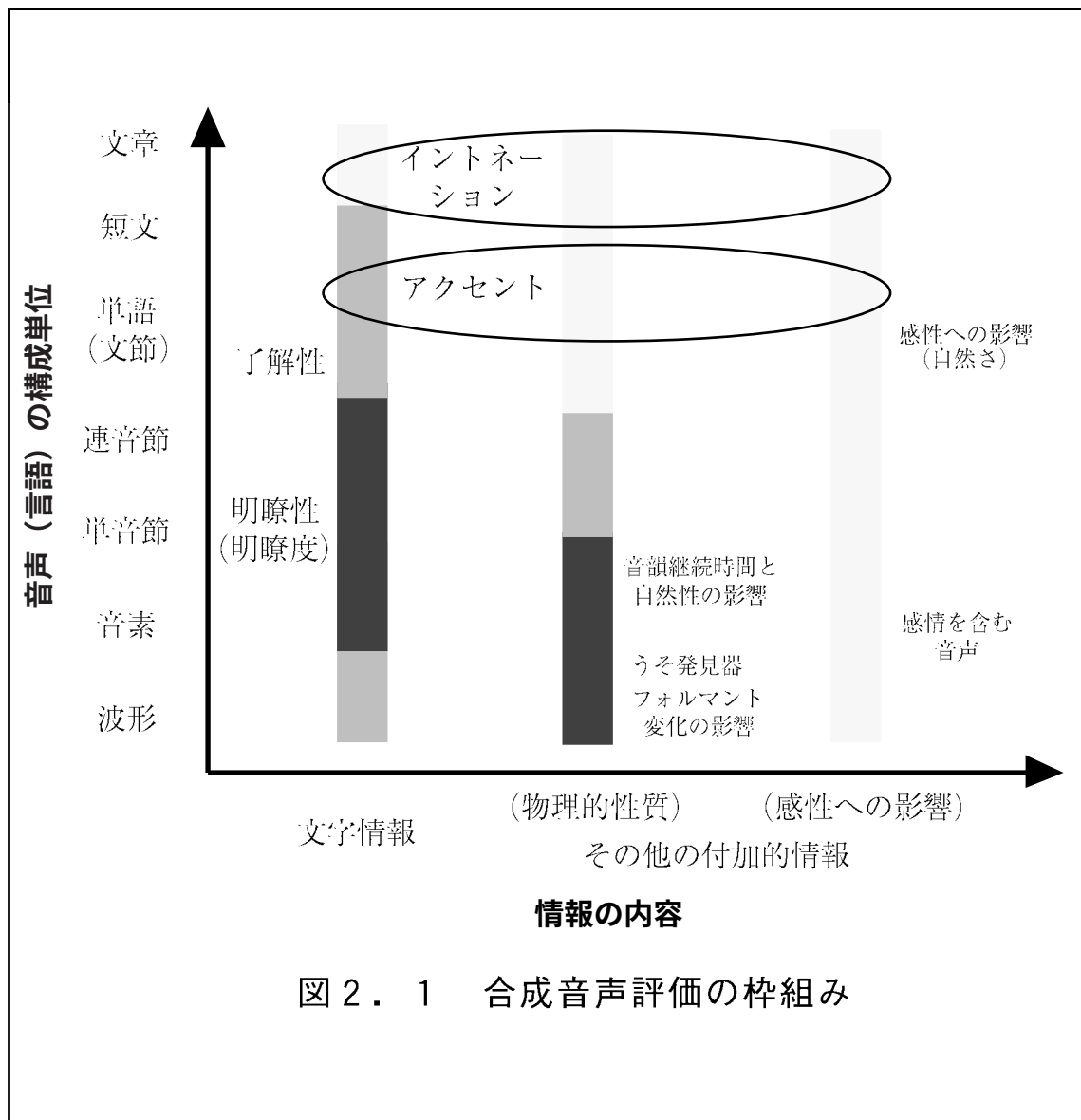
以上，検討を行ってきた合成音声の評価の分類をもとに，合成音声の評価の全体の枠組みを考え，その中で，従来の評価方法の位置づけを試みる。

合成音声の評価の枠組みとしては，音声のもつ特性を基本として，音声（言語）の構成単位と，伝達される情報の内容によって構成される空間を考える。音声の構成単位は，波形から文章へと，ミクロからマクロに並んでいる。

音声の伝達する情報の種類としては，文字情報とその他の付加的情報に分類した。その他の付加的情報の特性の評価に関する研究はこれまであまり行われてきていない。付加的情報へのアプローチには，音声（または合成音声）の物理的な特性を分析・制御して，その音声の感性への影響を調べるという手法と，感性への影響を調べてその原因である物理的な特性，性質を究明する，という手法が考えられる。そこで，物理的性質からのアプローチと感性への影響からのアプローチに分けて，便宜上，文字情報と並べてひとつの軸を構成することにした。

この二つの軸によって構成される空間が図 2. 1 である。この空間内で，例としてアクセント，イントネーションの分布について考えてみることにする。アクセント，イントネーションは，それぞれ，単語（文節）と文について，ピッ

第2章



チ、振幅、音韻長といった物理的な特性が制御されて生じることが知られている。どちらも、文字の表記上にはほとんど現れないが、単語や文の内容の伝達に重要であることは明らかである。一過性で揮発性という音声の情報メディアとしての短所を補う、冗長性ということもできる。

このようにアクセントとイントネーションは文字情報の伝達を手助けすると同時に、話者の特徴やそのときの感情など、文字以外の情報を伝達することも経験的に知られている。

合成音声の品質の評価として一般的に行われている了解性、明瞭性の評価は、いずれも合成音声のもつ文字情報そのものの評価である³¹⁻³³⁾。評価の対象が音素、単音節、連音節のものは“明瞭性の評価”，単語、短文については“了解性”の評価と呼ぶのが一般的である。

明瞭性では、一般に実験条件が明確に定義され再現性の高い試験が行われ、単音節明瞭度試験など事実上の標準試験となっているものもあるので、図2.1の中で濃い色で示している。了解性は、実験条件は明確にしやすいが、テストに用いられる単語、単文などが限定的で条件が実用の条件と一致しにくく、研究開発の中間評価的な性質が強いので薄い色とした。

合成音声の文字情報以外の付加的な情報の評価は、これまであまり行われてきていない。このうち、物理的な特性を分析・制御してその音声の感性への影響を調べる研究としては、音韻時間長を変化させたときの自然性の評価や、フォルマント周波数を変化させたときの聴感を調べる研究が行われている⁴⁰⁻⁴⁴⁾。

感性への影響からのアプローチの例としては、感情を含む音声の基礎的な研究などがある⁴⁵⁾。また、最近になって、テロ対策の技術として、音声进行分析して得られる話者の緊張度などの情報を利用した、うそ発見器も実用化されている。

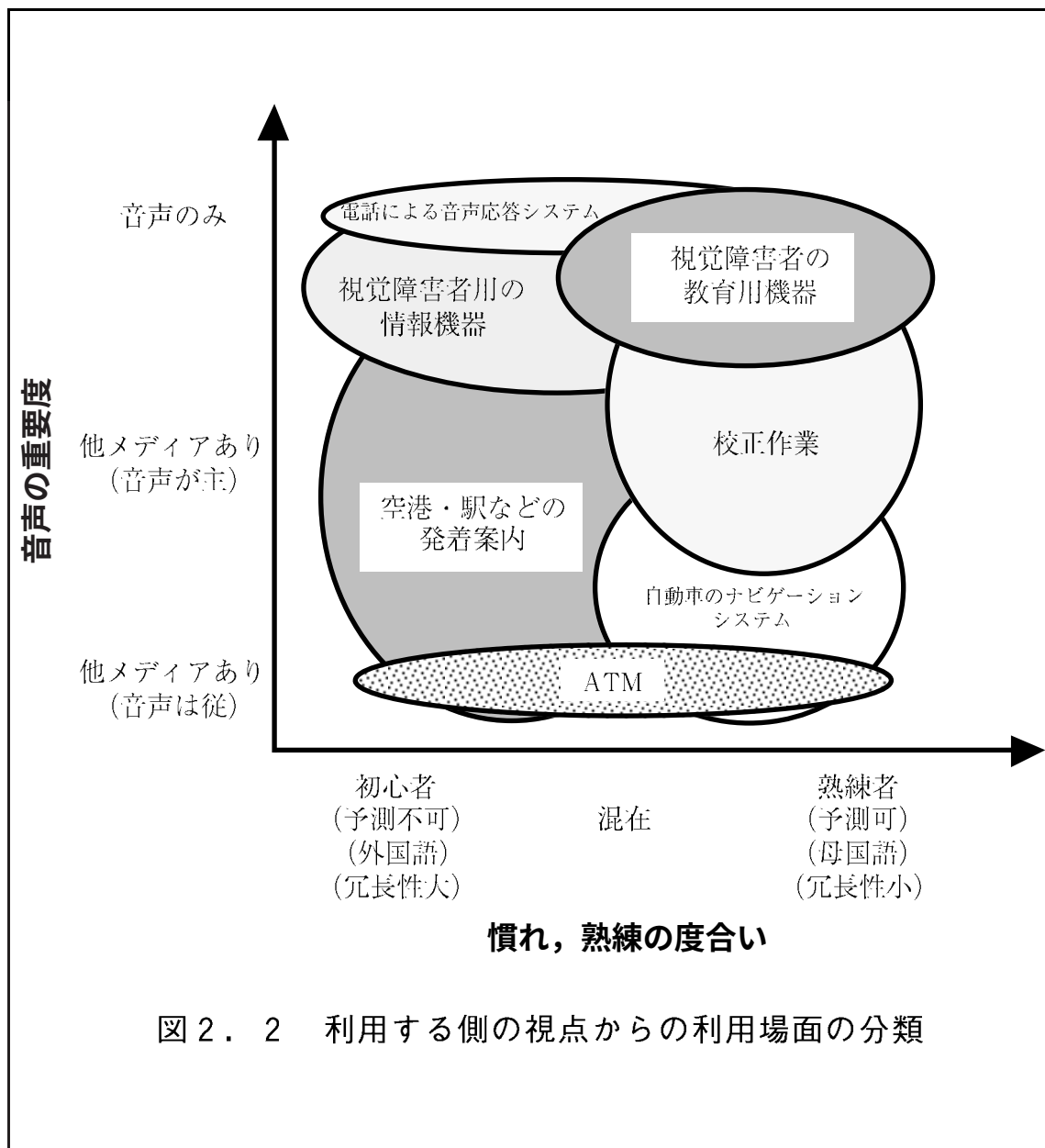
合成音声に対する文字情報以外の付加的な情報については、評価方法そのものの提案、研究の段階には達しておらず、自然音声や合成音声を用いて、物理量と感性への影響を調べる基礎的な実験が多く、将来的にその結果を活かして合成音声の品質改善や、評価方法の開発に発展させることを目指している。そこで、本論文では、“評価”を直接合成音声に対して行う狭い意味に限定せず、将来的に合成音声の評価に結びつけるための基礎研究や、合成音声の品質を改善やその他の工学的応用を目指した評価等も含めて検討していく。

(2) 利用する側の視点からの利用場面の分類

以上、検討してきた合成音声評価の枠組みは、音声の構成と特性に基づいたものであった。この他に、合成音声の日常生活、社会生活への普及、浸透をはかるうえで、利用者の視点に立った、利用場面の条件などを配慮した評価方法が重要になってくる。

図2.2は、利用する側の視点から、合成音声を利用する場面を分類したも

第2章



のである。利用する立場から重要になるポイントは2点ある。

ひとつは、利用する音声出力が、唯一のメディアであるのか、あるいは他のメディアと併用される場合には、合成音声の重要度はどの程度かどうかという点である。電話での音声応答システムでは、いうまでもなく音声が唯一の情報メディアとなる。

視覚障害者のための情報機器や、教育用の機器では、合成音声が唯一の情報メディアである場合が多いが、触覚を併用する場合もある。また、視覚障害者の中でも、全盲ではなく、明暗を感じることができたり、大きな文字、図形を読みとれる弱視者の場合は、そうした残存の視覚情報も利用できる。

ATM（金融機関の自動現金預け入れ・引き出し装置）では、ディスプレイ装置が主な情報メディアであり、音声はガイドや補助的な機能にとどまることが多い。

空港や駅など、公共交通機関における発着等の案内は中間的な性質を持っている。今日では、ディスプレイ装置と音声出力が併用されていることが多いが、利用者は通常、音声出力に注意し、必要に応じてディスプレイ装置で確認する。事故等で、ダイヤ、スケジュールに変更があるような場合、ディスプレイを注視するようになる。ディスプレイ装置と音声の案内という二つのメディアが本質的に同じ内容の情報を伝達することで冗長性を高め、受け取る側の都合で集中するメディアを選択できる。

これに対して、原稿の校正作業の場合には、音声とディスプレイの情報を照合しながら、誤字、脱字を確認するので、作業者は二つのメディアから同時に情報を求めている、音声もディスプレイ装置もどちらも主体であると言える。

また、自動車のナビゲーションシステムでは、走行時には音声の案内が主体になるが、駐停車時にディスプレイ装置で地図（案内図）を確認することのほうがより有効である。このディスプレイ装置によって直接確認している際には、音声出力は補助的な機能であり重要性は低い。

この他、航空機では、緊急時に警告音声が出力されるが、あくまで注意を喚起するアラーム音的な機能であり、情報メディアとしての性質は弱い。居眠り防止用の警告音声を自動車の運転等へ適用する研究も行われているが、同様に注意の喚起が主な役割となる。

第二のポイントは、利用者が初心者なのか、熟練者なのか、あるいは双方が混在するののかということである。利用者が熟練者のみであれば、慣れ、学習効果があるので、多少の音質の悪さは問題にならない。一方、初心者のみであれば、高い品質が要求され、冗長性の大きさ、要求に応じた繰り返しなど、ソフトウェアによる対応も重要である。

両者が混在するのは一番難しいケースである。音質が悪く、必要最小限の情報のみ出力されるだけでは初心者はとまどうが、反対に、あまりに冗長では熟

第2章

練者に敬遠される。利用者のレベルや要求に応じて柔軟に対応できる高度なシステムづくりが要求される。

初心者対熟練者の指標は、必要とされる情報メディアの冗長性の大小の指標と言い換えられる。この冗長性の大小に依存する利用者の属性としては、この他に、出力情報が限定的で簡単に予測できるか、可能性が広範囲で予測できないか等が当てはまる。空港、駅の発着案内など、公共の設備、施設では、その利用に慣れていない人のほか、日本語を母国語としない人の利用も十分考えられる。言葉の問題だけでなく、文化、習慣の違いから、発声される単語、文章も予測しにくいことが多い。

電話による応答システム、視覚障害者の情報機器では、空港・駅などの発着案内と同様に、初めて利用する人が主な領域から、比較的頻繁に利用する人が混在している領域にまたがっている。視覚障害者の教育用の機器、校正作業あるいは自動車のカーナビゲーションでは、習熟した利用者が利用するが、地名などの未知の単語や誤った単語などが出力されることが予想されるので、熟練者で予測可能な領域から、中間の領域に位置づけている。

ATM 装置では、同じ場所の同じ装置を頻繁に使用する熟練者、初めての場所で、初めての装置を利用する初心者、そして、両方の可能性を持つ人など、この指標について全ての可能性が考えられる。

以上、合成音声の情報メディアとして利用する側の視点からの指標により、合成音声を利用した機器の特性の分類を試みた。今後、マルチメディア機器など、合成音声を利用した情報機器、装置が日常生活で広く利用されるとともに、高度化、複雑化することが予想される。利用者側の視点に立ったシステムの開発がますます重要になっていく。

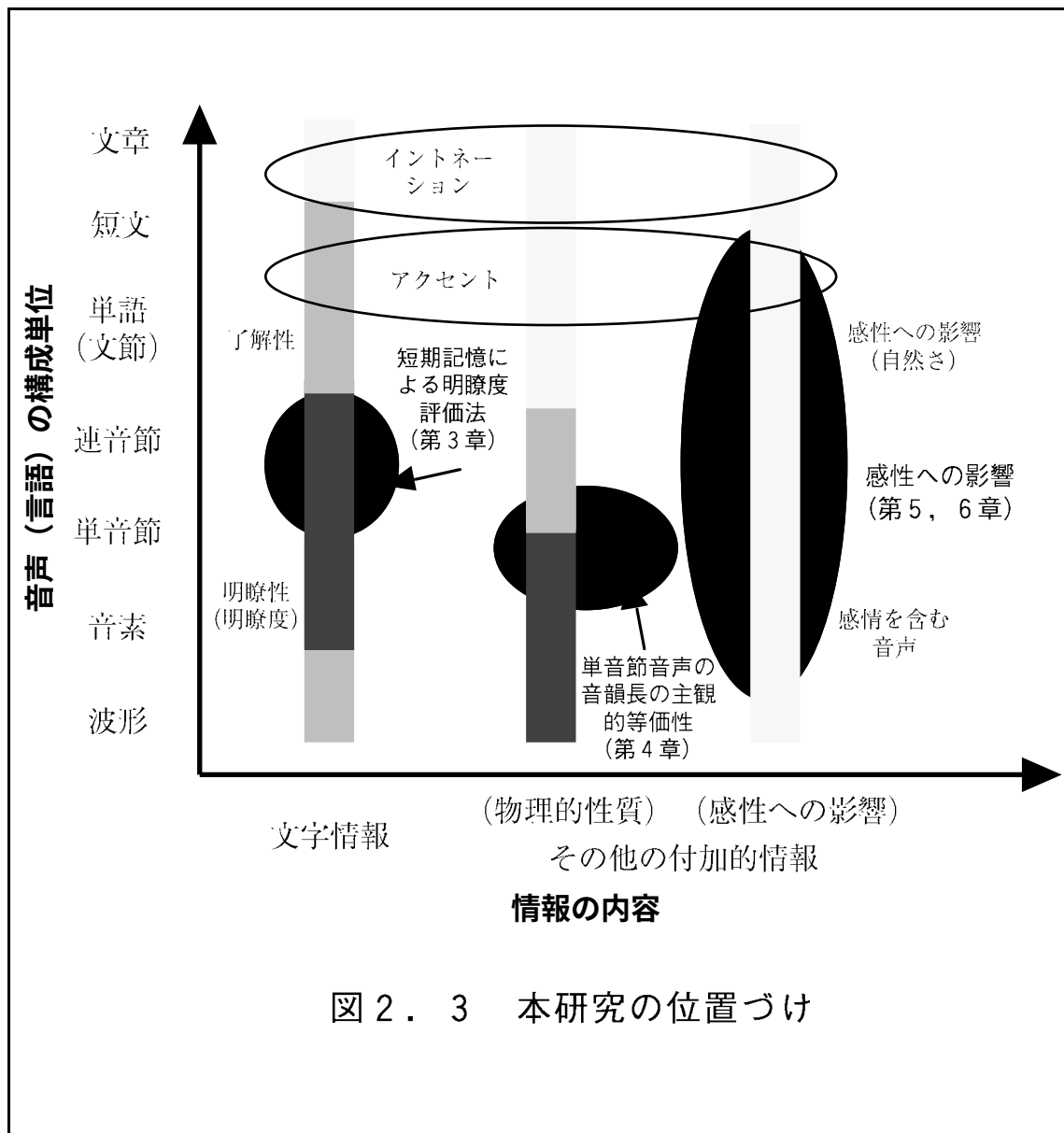
2. 4 本研究の目的と位置づけ

本研究の目的は、合成音声の品質向上に資するために、その評価方法の枠組みを明確にし、その中で現在欠けていたり、改善が必要な部分についていくつかの評価方法を提案し、実験的にその有効性を確認していくことである。

既に示したように、音声の持つ構造的な特質と、音声が伝達する情報の種類によって枠組みを構成し、その中に現在行われている合成音声の評価方法をそれぞれ位置づけた。この他、利用者側からの視点、注意の分散の有無などによる評価方法の分類・位置づけの必要性も述べた。

図2. 3は本研究で提案する評価方法の位置づけを、先の図2. 1の中に試みたものである。

図2. 1及び図2. 3が示しているように、文字情報の伝達の正確さの評価については、単音節、連音節、単語、短文等を用いた明瞭性、了解性の評価方法は多く提案されていて、そのうちのいくつかは合成音声の評価方法として標



第2章

準的な評価方法として用いられている。

これらの評価方法では、回答者（被験者）が、音声の聴取だけに集中できるため、音声の品質がある程度高くなると、正答率が100%近くとなり（天井効果）、その評価により差異が明らかにできる音質の領域も限定される。

こうした明瞭性、了解性の試験は、音声の基本的な品質の評価としての意味は十分ある。しかし、実際に利用する場合、利用者は合成音声を聞き取りながら、単語や文章全体の意味を考え、次の操作について考え、何らかの判断する、などというタスクを同時進行で行っていて、聞き取っている音声だけに集中することはできず、注意の分散が生じている。

そこで、6から8個のランダムな単音節音声を連続して聞かせて、聞き取った音節を忘れないように頭の中で繰り返させることにより[†]、音声の聴取のみに集中できないようにした、短期記憶による明瞭度評価法を提案して、実験的にその有効性を調べる（第3章）。

次に、文字情報以外の付加的な情報の伝達の領域の評価について、実験的に検討を進めていく。物理的な特性を制御してその感性への影響を測る例として、単音節音声の音韻長の心理的等価性について検討する。図2. 3では、単音節の物理的特性からの付加情報へのアプローチのひとつ（広義の評価）に相当する。

この領域では、これまで、音声の持つ特性のうち、強さや音の高さ（ピッチ）について、その心理的な効果が詳しく調べられてきた。しかし、時間的な要素についてはあまり調べられていない。そこで、日本語の単音節について音韻継続時間を伸縮した合成音声を作成して、被験者による比較実験を行う（第4章）。

最後に、図2. 3の感性への影響を出発点とした、合成音声の感性全般に及ぼす影響について調べていく。この領域の研究はこれまであまり行われてこなかったが、合成音声を利用した装置が社会や生活に広くゆきわたるようになるために重要な課題である。

その第一段階として、人が音声を聞くときに、どのような評価項目に着目しているか、広くアンケートを行って集めた表現をもとに評価項目を作成し、この評価項目によって自然音声を被験者に評価させる心理実験を行う。この心理実験の結果を主成分分析によって分析して、主に着目する評価項目を抽出する（第5章）。

第二段階では、物理量の例として、ピッチと音節の発声のタイミングのばらつき等を取り上げ、先の主観評価実験に用いた自然音声について分析する。こ

[†]短期記憶は記憶の分類のひとつで、感覚器に付随した感覚記憶から送られる情報が蓄積される記憶である。短期記憶に蓄えられた情報を保持するためには、頭の中で繰り返す必要があり、このことをリハーサルという^{38, 46, 47}。

の分析結果と、先の主成分分析の結果の間の相関を調べる主成分回帰の手法によって、主観評価の結果とこれらの物理量の関係を調べる（第6章）。

2.5 章のまとめ

本章では、合成音声評価の必要性、現状における問題点を明確にするために、合成音声の評価方法の分類方法、枠組みを明確にし、その中で現状での評価方法を分類し、現在の方法では補えきれなかったり、改善が必要な領域について検討した。

その上で、本論文で提案する評価方法の概要と評価の枠組みにおける位置づけを説明して、本研究の意義を明確にした。

第3章

短期記憶による明瞭度評価法

3.1 はじめに

第2章で述べたように、音声によって伝達される情報のうち、文字によって書き表すことのできる情報は大きなウェートを占めていて、その重要性についてはあらためて述べるまでもない。この文字情報の伝達の正確さ（明瞭性、了解性）を評価するための方法として、電話回線やホール等の音響設備の試験用に明瞭性、了解性の試験法が多く考案され、これらは、音声合成装置の出力音声の評価にも頻繁に用いられている²⁹⁻³³⁾。

合成音声の果たす重要な役割は、情報機器の出力メディアである。具体的には、電話による自動応答サービス、金融機関のATM（自動現金預け入れ・引き出し機）、自動車の付属装置、インテリジェントな家電製品などへの応用が挙げられる。これらの機器、装置などについて利用者側の立場からは、合成音声を聞き取って情報を受け取るまでが最終的な目的ではなく、希望する金額の現金を引き出す、目的地まで安全に自動車を運転するなど、最終的な目的が別に存在するのが普通である。これらの目的のために、合成音声から情報を受け取りながら、同時に思考・判断を行い、何らかの操作や機器等の運転を行う。

心理学では、人が思考、判断、操作、運転、その他の精神的な活動が伴う行為（以下“課題”という）を行う場合、全体として向けられる注意の合計には上限があるというのが定説になっている³⁸⁾。このため同時に複数の課題を行うと一つあたりの課題に向けられる注意は低下する。最近社会問題化している、自動車運転時に携帯電話を操作、あるいは通話して起こす事故は、その具体的な事例の一つである。

注意の考え方を当てはめると、合成音声の実用的な応用場面では、利用者は

第3章

合成音声を聞きながら、同時に別の課題を行っていて、注意はそれぞれに分散する。このとき、合成音声を聞くことではないほうの課題のほうが、その利用者にとって重要であることが多く、したがって、合成音声の聴取に割かれる注意は小さいことのほうが多くなる。

これに対して、合成音声に評価に用いられる多くの明瞭性（または了解性）の試験では、こうした注意の分散についての配慮はなく、被験者には唯一、聴取した音声を書き取るという課題が課せられるだけである。このため被験者は試験音声の聴取に注意を集中させることができる。集中できるので、いろいろな音響的な手がかり[†]から正答を推測することができる上、その音質に対する慣れ（学習効果）の効果も大きく出る。注意を集中できる結果、試験音声の正答率は高くなり、飽和してしまうので、実際には存在すると考えられる明瞭性の差が結果には現れなくなってしまう。

合成音声に対して従来行われてきた、音声の聴取のみに集中することができる明瞭性の試験を行うことは、研究開発の過程で基礎的な評価としての意味があるが、実用的な場面とは条件が異なることを理解する必要がある。

こうした観点から、本章では複数の単音節を連続的に聞かせ、その音節を筆記させる、より実際の条件に近い明瞭性の試験方法を提案する。提案する方法で実験を行い、その評価法の有効性を検証する。

3.2 短期記憶による明瞭度評価法

注意の分散を生じさせるひとつの方法として、音声の聴取とは関係のない課題（妨害課題）を音声の聴取と同時に課すことが考えられる。例えば、音声の聴取と同時に簡単な計算等を行わせる。この方法は実際の使用条件に近いが、主な課題と妨害課題とに関連性がないため、被験者ごとに、それぞれの課題へ向ける注意の重みが異なる可能性が考えられ、注意を均等化する工夫や、得られた結果から注意の度合いを評価する必要性が生じ、実験の手順や結果の分析・評価が複雑化する。

注意の分散を生じさせる別の方法として、音声を聞き取ってから回答するまでの必然的な手順を複雑化する手法が考えられる。

本章では、後者の一つである、短期記憶による明瞭度評価法を提案する。この方法は心理学の自由再生実験とほぼ同様の手順による^{46,47)}。すなわち、1.5～2.5 秒の間に6～8個のランダムな日本語の単音節音声を1系列の聴覚刺激として被験者に提示し、被験者は一度に提示された系列の最後の音節を聞き終えてから、回答用紙に聞き取った順に音節を書き取る。これによって、被験者の

[†]音響特徴(acoustic cue)と呼ばれる²⁹⁾。

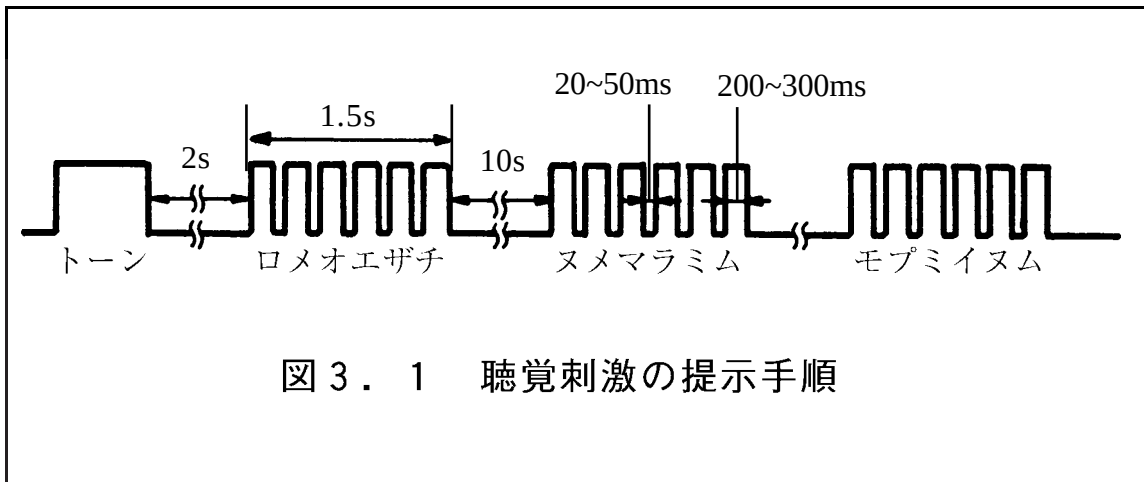


図 3. 1 聴覚刺激の提示手順

短期記憶にある音声情報を再生させる。

音声は聞き取った際、はじめに音声波形の情報のまま感覚記憶として保持されると考えられる³⁸⁾。次に、感覚記憶から、適合する言語情報に置き換え、短期記憶に記憶される。

多くの単音節を短い時間に連続して聞き取るにより、音声波形情報として感覚記憶にある音声情報を、文字情報に変換して、それを忘れないように繰り返し（“リハーサル”と呼ばれる[†]）、最後に書き取るという一連の課題が複雑化する。

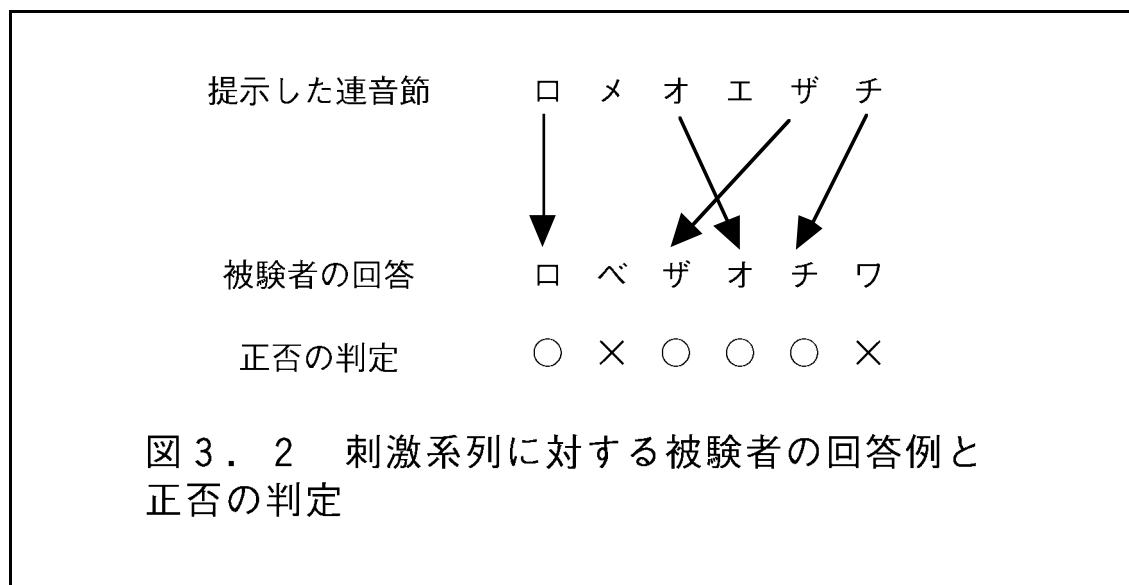
特に、連続して提示される連音節の前半部分を聞き取り、感覚記憶から該当する文字情報に変換した時点から、その文字情報を短期記憶内の情報を保持するためにはリハーサルが必須になる。このため後続の音節の聞き取りの際には、感覚記憶から文字情報への変換、リハーサルとに注意が分散することが予想される。

図 3. 1 は 6 音節を 1 系列とする場合の聴覚刺激の例である。実験開始の合図であるトーンの後、2 秒間おいて最初の系列が提示される。系列の提示から次の系列の提示まで 10 秒間の余裕があり、聞き終えた直後からこの 10 秒の間に回答を記入する。回答の際にリハーサルを行うような具体的な指示は与えないが、できる限り正確に回答するように教示した。また、回答を記入している間にリハーサルが妨げられて短期記憶が失われないように、回答はひらがな、カタカナのいずれでも、被験者が書きやすいものを選択するように教示した。

被験者には、聞き取った順番に書き取るように教示するが、回答の採点にあたっては、自由再生実験と同様に、系列内の順序の入れ替わりは配慮せず、脱落した音節と、完全に回答された音節の度数のみを計数した。例えば、図 3.

[†]短期記憶内の情報を忘れないように意識的もしくは無意識的に頭の中で繰り返すことをリハーサルという。

第3章



2のように“ロメオエザチ”という連音節が提示され、これに対してある被験者の回答が“ロベザオチワ”である場合について考える。最初の“ロ”は位置も単音節も合致しているので問題なく正答である。以下の5音節は位置まで一致を求めると正答はない。しかし、3番目から5番目までの“ザオチ”の3音節は、どれも提示された“ロメオエザチ”に含まれているので正答とする。これら3個の音節は、図の矢印が示すように、感覚記憶から短期記憶に移行する際、リハーサル中、あるいは書き取りに際して順番が前後したものと考えることにする。

誤りについては、例えば、図3. 2の回答の2番目の音節“ベ”は、提示された連音節中の2番目の“メ”を“ベ”と誤ったのか、あるいは4番目の“エ”を“ベ”と誤ったの断定できない。また、回答の“ワ”のように、提示された連音節との関連を見いだすのが困難な回答が含まれたり、空欄があることも予想される。そこで、誤った回答の集計は行わなかった。

以上のように、提案する短期記憶による明瞭度評価法は、自由再生実験の手法を基本とするが、回答を被験者が自分で書き取ることと、聞き取った順番に再生する（書き取る）ように教示することの2点が自由再生実験と異なっている。

3. 3 実験方法

3. 3. 1 実験装置

図3. 3は本章の実験に用いた装置の概略である。A/D、D/A変換器を備えたマイクロコンピュータ、テープレコーダ、低周波アンプおよびローパスフィルタから構成されている。

3.3.2 実験用音声の採取および作成

実験では音質を評価する対象となる聴覚刺激として、日本語単音節音声进行 PARCOR 法によって分析・合成した単音節音声を使用した⁹⁾。

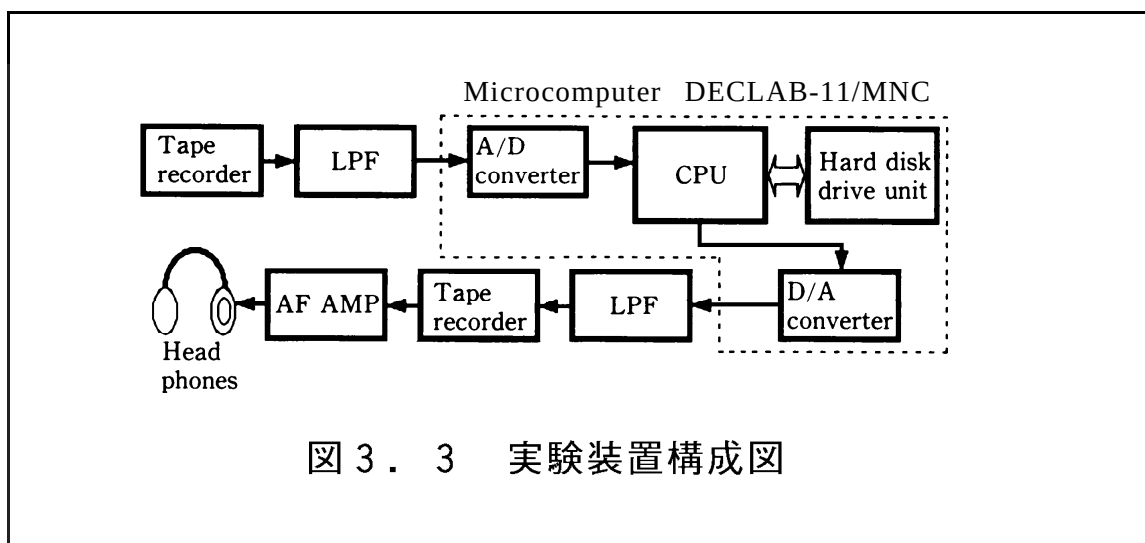
テープレコーダで再生された明瞭度試験用テープ³⁴⁾の音声信号はローパスフィルタを経て、マイクロコンピュータのA/D変換器に加えられ、サンプル周波数10kHz、12bitで量子化し、一音節ずつハードディスクに記録した。

ハードディスクに記録した各音節について、立ち上がり、立ち下りの閾値を設定して、音声部分を切り出した。これらの閾値は、実験者が実際に聞き取りながら、不自然な立ち上がり、立ち下りにならない最大の値になるように音節ごとに微調整した。また、同様に、聞き取りながら聴感的強度を一定にするという前処理を行った。この前処理を加えた音節を実験の結果を比較する参照用の刺激として使用した(以下、原音声という)。

使用した計算機の処理能力上、実時間で音声合成をすることができないので、

表3.1 PARCOR分析・合成の仕様

サンプル周波数	10 kHz
量子化ビット数	12 bit
窓長	25.6 ms
フレーム周期	12.8 ms
予測次数	15 次
窓関数	ハミング窓



第3章

前処理を加えた音節について、表3. 1の仕様に従い、ソフトウェアによってあらかじめPARCOR分析・合成を行って実験用の合成音声を作成し、ハードディスクに記録した（以下、合成音声という）。

以上のようにして作成した原音声と合成音声の単音節音声データを実験のスケジュールに従ってD/A変換器で再生し、テープレコーダで録音した。各音節の再生は、ハードディスク装置からそれぞれの音節の波形データが半導体メモリー上に転送され、このデータがサンプル周波数ごとにD/A変換器に転送される。このため、前の音節と次の音節の間に、ハードディスクのアクセス時間と半導体メモリーへのデータの転送時間の分、不必要な無音の時間が生じ、ハードディスクに関する条件によってその無音時間の長さがばらついてしまう。この不必要な無音の時間を最小限にするために、サンプル周波数を1/4の2.5kHzとして、ローパスフィルタのカットオフ周波数を下げ、テープレコーダのテープ速度を落として録音した。再生の際には、録音時の4倍の速度で再生することで、分析時と同じ周波数情報、音韻継続時間をもつ音声を得られる。同時に、D/A変換器のクロックの不正確さによる歪みも低減することができる。実験では、このあらかじめ録音した刺激を被験者に提示した。

3. 3. 3 刺激提示手順

聴覚刺激の提示手順は、まず原音声の系列を、次に合成音声の系列を提示し、以下それぞれの系列を交互に提示する。系列内の単音節の間隔は、提示プログラムでは特に設定しなかったが、先に述べたように、ハードディスクのアクセス時間などによる遅れによって無音時間が生じ、この無音時間を減らす工夫をした。しかし、無音時間をなくすることはできず、その長さは単音節の種類や条件によって差がある。この状態で1系列を6～8音節とすると、系列全体の長さは1.5～2.5秒となった。

4組ないし5組の系列（原音声と合成音声の刺激を合わせて8または10系列）を1シリーズとし、シリーズの間は30秒から3分間の小休止を規則的にとった。

順序による影響を少なくするために、原音声の刺激から提示したシリーズの次のシリーズでは合成音声の刺激から提示するようにした。被験者1名に対し、このシリーズを合計で20～24連続的に提示した。

このようにして、原音声と合成音声について同条件で実験を行い、その結果を比較することによってPARCOR分析・合成によるそれぞれの音節の明瞭性の低下を調べた。

3. 4 実験

3. 4. 1 実験1

(1) 実験方法

実験者が被験者となって簡単な予備実験を行ったところ、1系列の音節数が8以下であれば、音質が良好な場合には比較的容易に全音節を回答できることがわかったので、1系列を8音節として実験を行うことにした。提案する評価法は通常の明瞭度試験よりも厳しい評価結果が得られると予想されるので、用意した合成音声について、あらかじめ単音節ごとの明瞭度試験を行い、正答率の低いものを除外して表3. 2に示す、32の単音節を使って実験用聴覚刺激を作成した。使用する単音節の数を減らすと実験の一般性は低くなるが、それぞれの単音節の出現回数を多くできるので、提示した順番による正答率の変化の傾向などを調べるのに適している。

疑似乱数によって系列内の単音節を決定し、原音声の刺激4系列と合成音声の刺激4系列が交互に現れるように配慮し、これを1シリーズとして1名の被験者に全部で20シリーズ提示した。両耳とも正常な聴力を持つ6名の成人（うち女性2名）を被験者として実験を行った。1回の実験には約50分を要した。

表3. 2 実験に使用した単音節

	ザ	ガ	ラ	マ	ハ	ナ	カ*	ア
				ミ		ニ	チ	イ
プ	ズ		ル	ユ	ム	ヌ	ツ	ク
ペ	ベ	ゼ			メ	ネ*		ケ
ポ	ゾ		ロ		モ			オ

* 印の単音節は6音節による実験では使用しない

表3. 3 1系列8音節のときの誤り率 (%)

	短期記憶	単音節明瞭度試験
原音声	27.1	1.6
合成音声	29.5	7.4
$p<0.01$		○

表中の○印は、両側のt検定の結果、有意水準0.01で原音声と合成音声の誤り率に差が認められたことを表す。

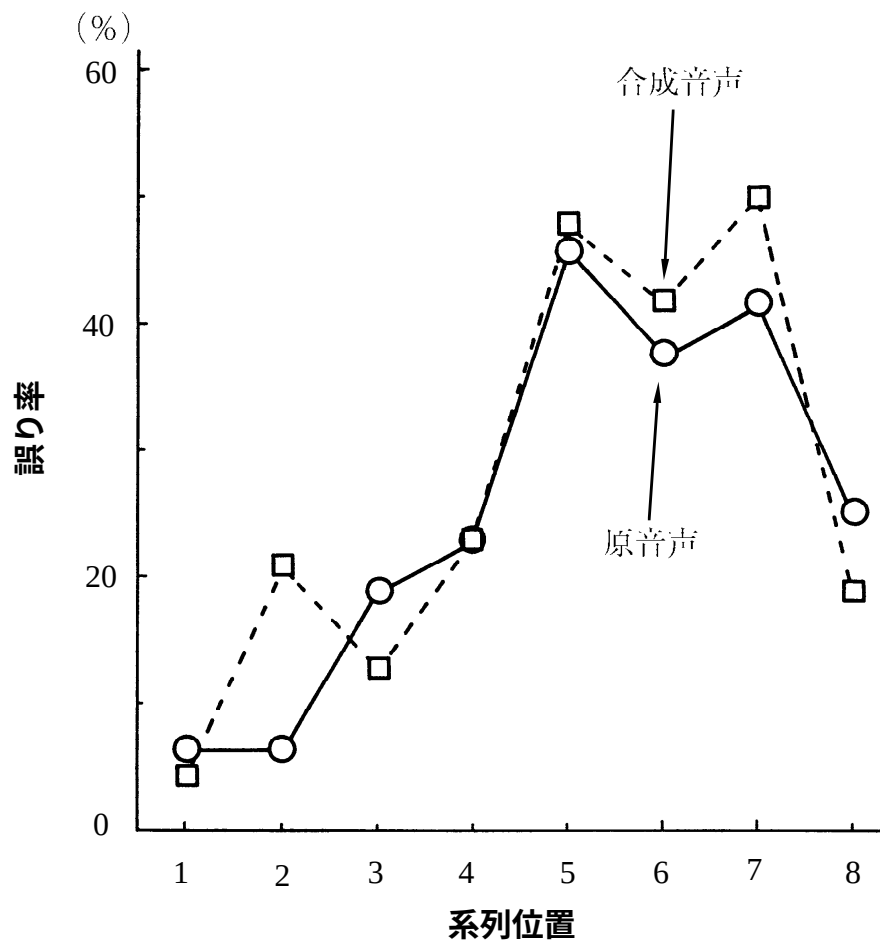


図3. 4 1系列を8音節としたときの系列位置と誤り率の関係

(2) 結果および検討

表3. 3は実験の結果得られた原音声と合成音声の誤り率と、同じ単音節について通常の明瞭度試験によって求めた誤り率である。被験者ごとの誤り率から求めた平均値と標準偏差によりtテスト^{48,49)}による検定を行うと、1%の有意水準で通常の明瞭度試験で合成音声の誤り率が、原音声よりも大きいことが認められる。しかし、短期記憶による明瞭度試験では、原音声と合成音声の誤り率に有意な差は認められなかった。

図3. 4は聴覚刺激の提示順（系列位置）と誤り率の関係を示している。誤り率は、系列位置が1番目のものが低く、後になるに従って徐々に上昇し、5～7番目で原音声、合成音声ともに50%にも達し、最後の8番目は再び減少している。そして、原音声、合成音声とも系列位置ごとの誤り率はほとんど等しい。tテストでは、全部の系列位置ごと及び全平均のいずれについても原音声と合成音声の間に有意な差は認められなかった。

この誤り率の増減は、心理学の自由再生実験における、初頭効果と新近性効果³⁸⁾と同様の効果の現れと考えられる[†]。系列位置が5～7番目の誤り率がほぼ一定となっているのは、1系列の音節数が多く、短期記憶の容量の限界に達しているためと考えられる。

また、原音声と合成音声の誤り率に差がなかったのは、1系列の音節数が多く短期記憶の限界に達して、短期記憶の容量を超えた分が切り捨てられるため、音質の優劣の差が結果に現れなかったものと考えられる。

3. 4. 2 実験2

(1) 実験方法

実験1の結果では全体に誤り率が高く、合成音声と原音声の誤り率の差は小さく、有意な差は認められなかった。その原因は1系列中の音節数が多く、短期記憶の容量を超えたためと考えられるので、1系列の音節数を6に減らして再度実験を行った。今回の被験者は、実験1とは別の11名の成人（うち女性6名）で全員正常な聴力をもつ。

使用した音節は先の8音節の実験で使った32音節のうちの30である（表3. 2参照）。

この実験では、系列位置効果を音節別にみるため、原音声と合成音声の刺激とも、全体を通して各音節がそれぞれ4回ずつ、系列中の1番目から6番目に均等に出現するようにした上で、可能な限りランダムになるような工夫をした。各シリーズは原音声と合成音声それぞれ5系列ずつからなり、実験は全体で

[†] 一般に、系列的に刺激を提示する学習や記憶の実験では、最初のうちに提示された刺激の誤り率は低く（初頭効果）、徐々に誤り率は上昇し、系列の最後のほうでは再び誤り率が低下する（新近性効果）。こうした系列位置と誤り率の関係を系列位置効果³⁸⁾と呼ぶ。

表3. 4 各音節の誤り率 (%)

音節	短期記憶			単音節明瞭度試験		
	原音声	合成音声	p<0.01	原音声	合成音声	p<0.01
ア	7.2	7.6		0.0	0.0	
イ	6.1	5.7		4.6	0.0	
エ	13.6	4.2		0.0	4.6	
オ	9.5	4.9		0.0	0.0	
ク	2.3	4.6		0.0	0.0	
ケ	13.3	28.4		0.0	0.0	
チ	100.0	96.2	○	86.4	36.4	○
ツ	70.5	72.4		9.1	9.1	
ナ	18.9	56.8	○	0.0	68.2	○
ニ	10.2	11.7		0.0	0.0	
ヌ	23.5	24.2		0.0	0.0	
ハ	4.2	3.8		4.6	0.0	
マ	17.4	14.8		0.0	0.0	
ミ	24.2	16.3		0.0	0.0	
ム	21.6	18.6		0.0	0.0	
メ	11.7	17.4		0.0	13.6	
モ	13.6	41.3	○	0.0	0.0	
ユ	3.8	2.3		0.0	4.6	
ラ	12.9	13.3		0.0	4.6	
ル	18.9	18.2		0.0	0.0	
ロ	19.7	13.6		0.0	0.0	
ガ	6.4	8.0		8.0	0.0	
ザ	13.3	33.3	○	0.0	0.0	
ズ	13.6	11.0		0.0	9.1	
ゼ	22.4	29.2		0.0	4.6	
ゾ	17.4	16.3		0.0	0.0	
ベ	29.9	45.5	○	0.0	0.0	
プ	23.5	17.4		0.0	0.0	
ペ	18.6	81.8	○	0.0	86.4	○
ポ	20.8	73.1	○	0.0	50.0	○
平均	19.6	26.4	○	3.5	9.7	○

○印は、両側の t 検定の結果、有意水準 0.01 で原音声と合成音声の誤り率に差が認められた音節。

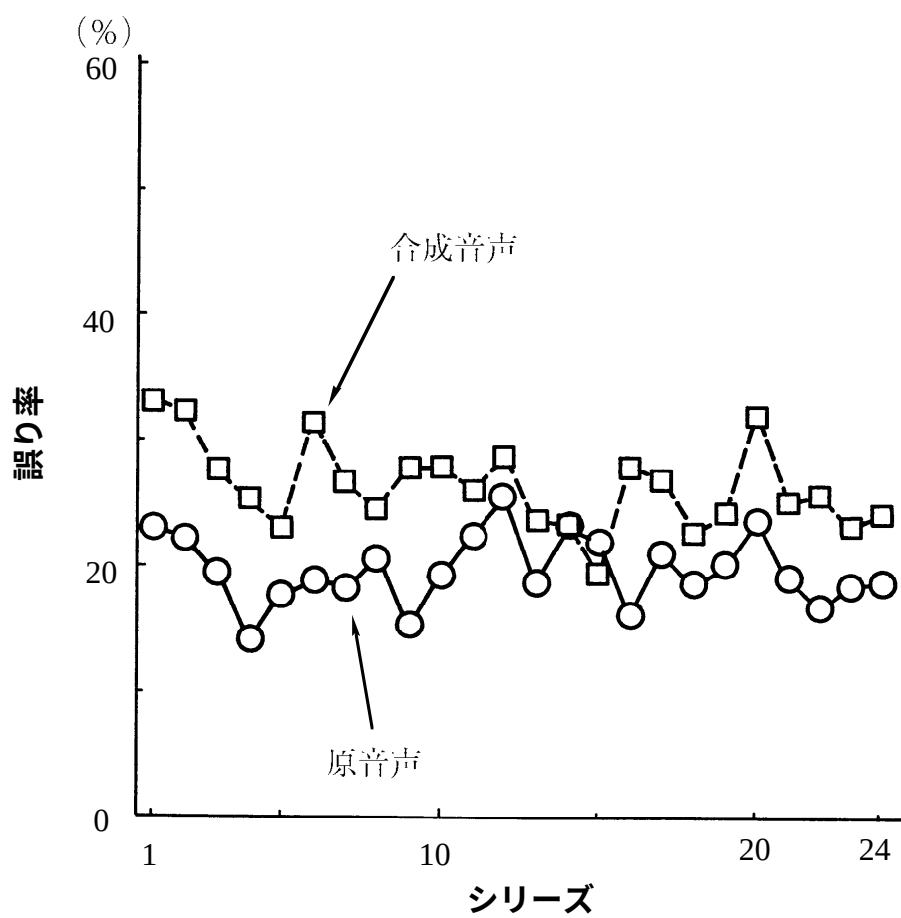


図 3. 5 刺激のシリーズと誤り率の関係

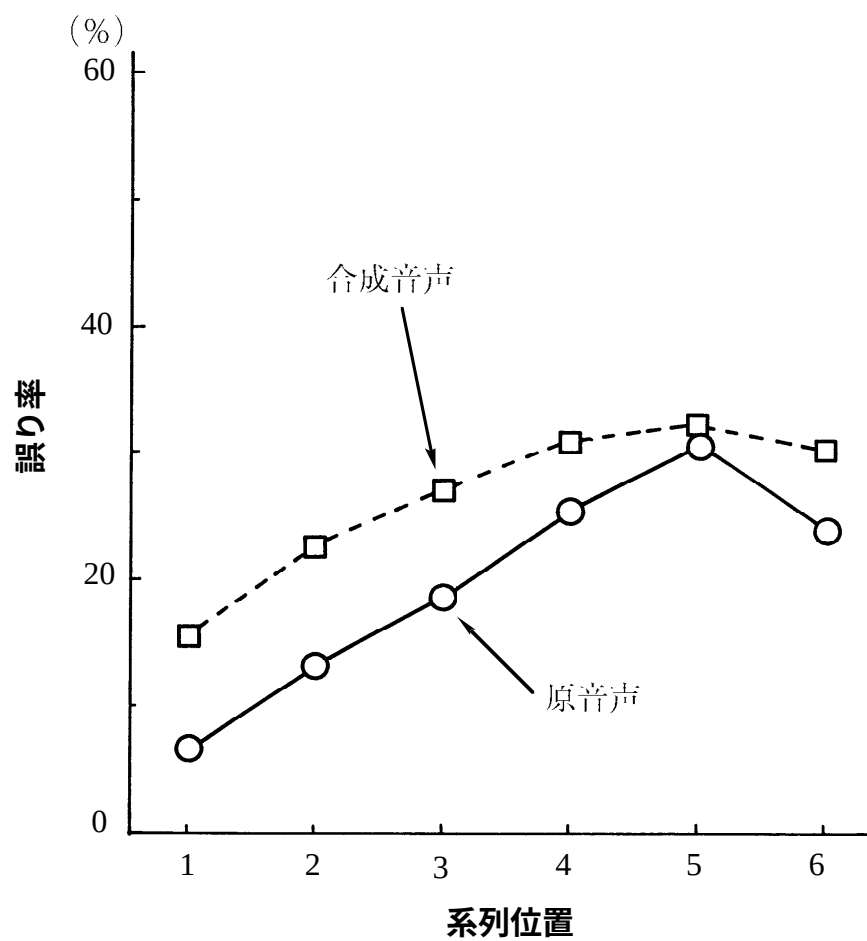
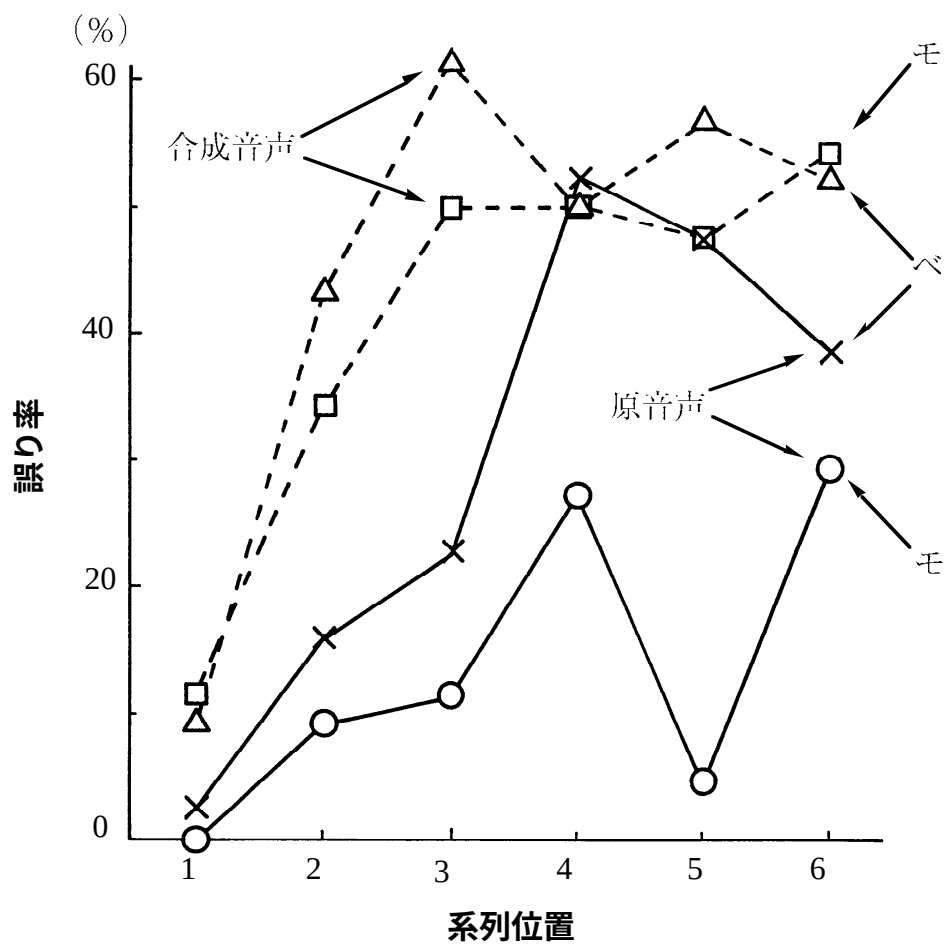


図3.6 1系列を6音節としたときの系列位置と誤り率の関係



(a) “モ”と“ベ”

図 3. 7 系列位置と誤り率の関係

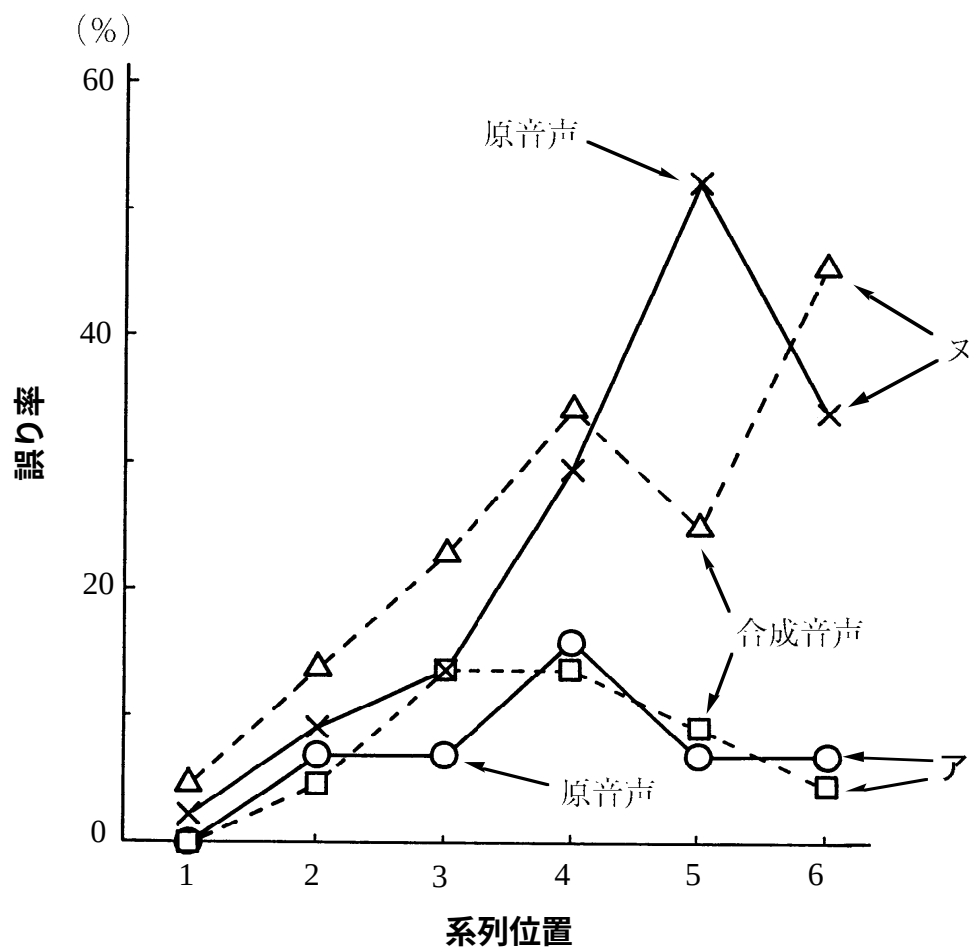


図 3. 7 (b) “ア”と“ヌ”

24シリーズで構成され、1時間を要した。

(2) 結果

表3.4は、この実験で刺激として提示したそれぞれの音節の誤り率と、通常の明瞭度試験による誤り率を示す。○印を付した音節は、被験者毎の誤り率から求めた標準偏差を用いてtテストの結果1%の有意水準で合成音声の誤り率が原音声よりも高いと認められたものである。

図3.5にシリーズ毎の合成音声、原音声それぞれの誤り率の推移を、図3.6に系列位置と誤り率の関係を合成音声、原音声それぞれについて示している。また、図3.7には、合成音声と原音声で誤り率に大きな差のあった“モ”，“ベ”と、誤り率に差のなかった“ア”，“ヌ”について、系列位置と誤り率の関係について示した。

3.5 考察

表3.4で、音節全体の誤り率の平均値を見ると、8音節を1系列とした実験1の結果とは異なり、いずれの評価方法においても合成音声の誤り率が原音声よりも高くなっている。

原音声の全体の誤り率は、明瞭度試験で3.5%と低いので、実験に使用した原音声の音節の品質は十分に高いといえる。この原音声について短期記憶では誤り率が19.6%になる。したがって、実験2で再生された音節数の平均個数は、

$$6 \times (1 - 0.196) = 4.824$$

およそ、4.8個である。

一方、8音節を1系列とした実験1（表3.3）では、明瞭度試験における原音声の誤り率は1.6%、短期記憶の実験の誤り率は27.1%となっている。短期記憶から再生された単音節数の平均値は、

$$8 \times (1 - 0.271) = 5.832$$

およそ5.8個である。

短期記憶の容量は、心理学ではチャンクと呼ばれる。頭の中で一つの事象としてとらえることのできる、文字、数字、その他身近なものの名前などは全て1チャンクである。心理学の実験結果では、短期記憶の平均的な容量は7チャンク前後とされている^{46, 47)}。本章の実験では、単音節をランダムに提示しているので、音質が十分によくて、聴取している単音節音声の感覚記憶から文字情報への移行ができれば、1音節が記憶の単位の1チャンクに相当すると考えられるので、こうした心理学の実験結果を適用すれば正答は最大で7前後になる。6音節の実験で4.8個、8音節の実験で5.8個の正答が得られた結果は、正答数が最大で7個前後になるという従来の知見と大きく矛盾しない。

実験2の結果である表3.4について個々の音節に着目すると、“ケ”，“モ”，

第3章

“ザ”、“ベ”などは、いずれも明瞭度試験では原音声、合成音声ともに誤りが少ないが、短期記憶実験では原音声に比べ、合成音声は非常に高い誤り率を示している。このうち“モ”、“ザ”、“ベ”については、明瞭度試験では有意な差が認められていないが、短期記憶において1%の有意水準で合成音声の誤り率が原音声よりも高いと認められている。

これに対して、“ア”、“イ”、“ハ”、“ユ”、“ガ”などは、原音声、合成音声とも誤り率が低く、両者の値の差も小さい。

これらの結果から、“モ”、“ザ”、“ベ”などの一部の音節について、通常の明瞭度試験では明らかにできない程度の（PARCOR 分析・合成による）音質の劣化を、提案する評価法によって検出できた、言い換えると、提案する短期記憶による明瞭度評価法の単音節の明瞭性に対する感度が通常の明瞭度試験より高いと言える。

表3. 4の単音節の中には、“オ”、“チ”、“プ”、“ミ”、“ガ”など、短期記憶実験または明瞭度試験において、原音声のほうが誤り率が高くなっている単音節がある。誤り率の差については“チ”を除き有意な差は認められなかった。

“チ”は短期記憶実験、明瞭度試験とも原音声の誤り率が高いことが有意に認められる。回答を調べると、短期記憶実験、明瞭度試験ともほとんどの場合“シ”と聞き取られてしまう“異聴”を起こしていて、聴覚刺激を提示するために用いた音響装置に何らかの問題があったと考えられる。しかし、短期記憶、明瞭度試験とも原音声の誤り率が有意に高いこと、この他の原音声の誤り率が高い単音節については誤り率の差は有意ではないことから、単に、当該の刺激が実験には不適切な音質であり、提案する試験法の問題ではないと考えられる。

ところで、合成音声の音質について、ある程度音質が悪くても、学習効果（あるいは慣れ、順応）^{29, 50)}によって、次第に正しく認識できるようになると言われている。本章の実験は1時間に及び続けられるので、被験者の疲労、集中力の低下、それに伴う信頼性の低下も懸念される。図3. 5は第1～第24シリーズについて、原音声、合成音声それぞれのシリーズごとの誤り率を提示の順番によって示したものである。実験開始から第4シリーズまではどちらの誤り率も減少しており、学習効果が現れていると考えられる。しかし、その後はどちらも変動を伴うものの、全体としての増加、減少の傾向なく推移している。学習効果による誤り率の低下も全体から見れば大きくなく、また、後半に疲労等からくる誤り率の上昇も見られない。したがって結果に大きな影響を及ぼすような顕著な学習効果や疲労等の影響は認められないといえる。

図3. 6には実験に使用した単音節全体の系列位置と誤り率の関係を示している。この結果にも、1系列を8音節とした実験1の結果（図3. 4）と同様、

初頭効果と新近性効果が現れている[†]。また、どの系列位置においても合成音声の誤り率のほうが大きくなっている。被験者ごとの誤り率から求めた標準偏差によって行った t 検定では、系列位置の 1, 2, 3 番目と、全体の平均値において、1%の有意水準で、合成音声の誤り率が原音声よりも高いと認められた。

図 3. 6 の、図 3. 4 との大きな違いは、誤り率の折れ線に平坦な部分がないことである。誤り率は原音声、合成音声のどちらについても系列位置の 5 番目まで増加してピークとなり、6 番目では減少している。平坦部がないことから、提示する 1 系列の情報量が短期記憶の容量を大きく越えていないと推測される。

図 3. 7(a) は、提案する評価法において原音声と合成音声とで誤り率に大きな差が現れた音節（表 3. 4 参照）のうちの“モ”と“ベ”の系列位置と誤り率を示したものである。いずれの音節についても、合成音声は 1 番目に提示したときの誤り率は低いが、2 番目から急激に上昇している。原音声との比較では、2～3 番目において特に大きな差が生じている。

このような誤り率のパターンを示す原因について、次のように推測される。連続的に音節が提示された場合、ひとつひとつの音節に対して時間的に余裕がないので、明瞭性の低い音節を感覚記憶から短期記憶に移すために、聞き取った音声に対応する文字情報に正確に変換することは非常に難しい。さらに、短期記憶内にある音節の情報のリハーサルを行いながらでは、注意が分散するため、感覚情報から音声情報に変換することはより困難となる。このため、リハーサルを開始するとみられる系列位置 2～3 番目において、明瞭度の劣る合成音声である“モ”、“ベ”などの誤り率が急に上昇するのであろう。

図 3. 7(b) は、短期記憶実験において原音声と合成音声の誤り率に差がない“ア”と“ヌ”の系列位置と誤り率の関係を示している。誤り率の低い“ア”においては、初頭効果、新近性効果ともよく現れており、原音声と合成音声のパターンは類似している。一方、誤り率の比較的高い“ヌ”は、原音声、合成音声とも系列位置が後になるほど誤り率が高まる傾向にあり、“ア”と、“モ”、“ベ”の合成音声との中間的な特性を示している。これらから、原音声と合成音声の“ヌ”の明瞭性についても、“ア”と合成音声の“モ”、“ベ”の中間的な特性と考えられる。

以上のように、提案する試験方法では、各音節を感覚記憶にある音声情報から短期記憶で保持する文字情報に変換する時間に制約があり、しかも短期記憶を失わないためにリハーサルを繰り返すことによって、注意の分散が生じる。この試験方法を用いると、1 音節ずつ聞く明瞭度試験では十分に高い正答率が得られる音節の中に、誤り率が著しく上昇する音節があることが認められた。

[†] p.35 の脚注参照。

第3章

特にこうした音節は、系列位置が2～3番目に提示される時に、誤り率に大きな差異が生じる。

合成音声が実用的に用いられる場面では、合成音声を聞きながら、何らかの思考、判断、操作、運転などを同時に行い、合成音声の聴取に向けられる注意は分散して低下しているにもかかわらず、従来、合成音声の文字情報の伝達の正確さを評価する明瞭性、了解性の試験では、注意の分散について考慮されたものはほとんどない。そこで今回、連音節を被験者に提示し、リハーサルを行うことにより、聴取する音声に対する注意の分散を生じさせる明瞭度評価法を提案し、同じ音声に対して従来の明瞭度評価法との間で明瞭性に差が生じることが明らかにできた。この評価方法は、短い単語を用いた了解性の試験等にも応用可能である。

短期記憶のリハーサルにより、聴取する音声に対する集中を妨げることはできたが、リハーサルを行うことと、実際の合成音声の利用条件とでは、聴取者（被験者）のタスクの性質や注意の配分が異なるので、これらの差異が及ぼす明瞭性への影響についてさらに検討を行う必要がある。より実用的な指標を得るためには、各種方式による合成音声を用い、また年齢等の差を加味して実験を繰り返す必要がある。そうした実験をもとに実用的といえる明瞭度の水準の標準化等についての検討を行っていく必要である。

3. 6 章のまとめ

合成音声は、情報機器の出力情報メディアのひとつとして用いられる。実用的な場面では、利用者はその出力情報を聴取しながら何らかの思考や判断、あるいは機器や車両、航空機などの操作、運転を行うため、音声の聴取にのみ集中することができないのがむしろ一般的である。しかし、従来行われてきた合成音声の明瞭性や了解性の評価方法では、被験者は提示される音声だけに集中できるので実用場面とは条件が大きく異なるといえる。

そこで、音声の聴取にのみ集中できないように、連続して出力される単音節音声を取り、短期記憶にある音声情報を再生する明瞭度評価法を提案し、実際に合成音声の評価を行った。

その結果、通常の明瞭度試験では正答率が十分高くても、連続的に発声されると正答率が低下する音節があることが明らかになった。これらの音節は連続系列中で、出現する順序によっても正答率が著しく変化することも認められた。

このことは、従来の明瞭性評価だけでは、実用上問題が生じる可能性を示しており、提案する評価方法など、注意の分散を加味した評価方法の必要性を裏付けている。

さらに検討の結果、短期記憶による明瞭度評価法が、全般的に通常の明瞭度試験よりも、音節の音質に対してより高い感度をもつことが確認された。

以上から、提案する明瞭度の評価方法が、より実用的な明瞭性の評価試験といえる。この試験方法は、短い単語を用いた了解性の試験にもそのまま応用が可能である。

第4章

単音節音声の音韻長の主観的等価性

4.1 はじめに

現在、実用化されている多くの情報機器の出力音声には文節（単語）や短文単位の編集合成や規則合成が用いられている。これらの合成音声では、単音節の明瞭性が多少低くても、アクセントやイントネーションといった付加情報の助けがあるので、高い了解性が得られる。しかし、高度な情報機器においては、人名、地名やその他の固有名詞、あるいは外来語、時には全く無意味と思われる文字列など、規則合成装置が持つアクセントをはじめとする情報を登録した辞書（以下特にことわりなく辞書といえ、この規則合成装置の辞書を指す）に登録されていない単語を出力する必要がしばしば生じる。そうした単語は利用者にとって重要性が高いことが多いと考えられる。むしろ、定形の文や辞書に登録された語句以上に明瞭性が高いことが要求される。

ところで、生命工学工業技術研究所の前身である製品科学研究所では、視覚障害を持つ児童生徒が授業等で利用することを目的とした盲人用読書機の研究開発を行った^{14, 15)}。この読書機は、家庭内、教室内で個人で利用することを想定したので、価格、大きさ、重量等はパーソナルコンピュータの規模を目指した。

この装置の出力音声には、単音節編集合成方式を採用した。開発当時の技術水準では、出力音声に規則合成を用いることは不可能であったことが主な理由である。現在の技術水準であれば、規則合成の音声合成装置を個人向けの装置に用いることは、技術的にも価格的にも可能である。しかし、規則合成では辞書の容量は有限であり、教育での利用を考えた場合、文字を一つずつ拾って読む必要もあるので、規則合成を主として用いるにしても、単音節単位の編集合

第4章

成の手法は欠かせない。

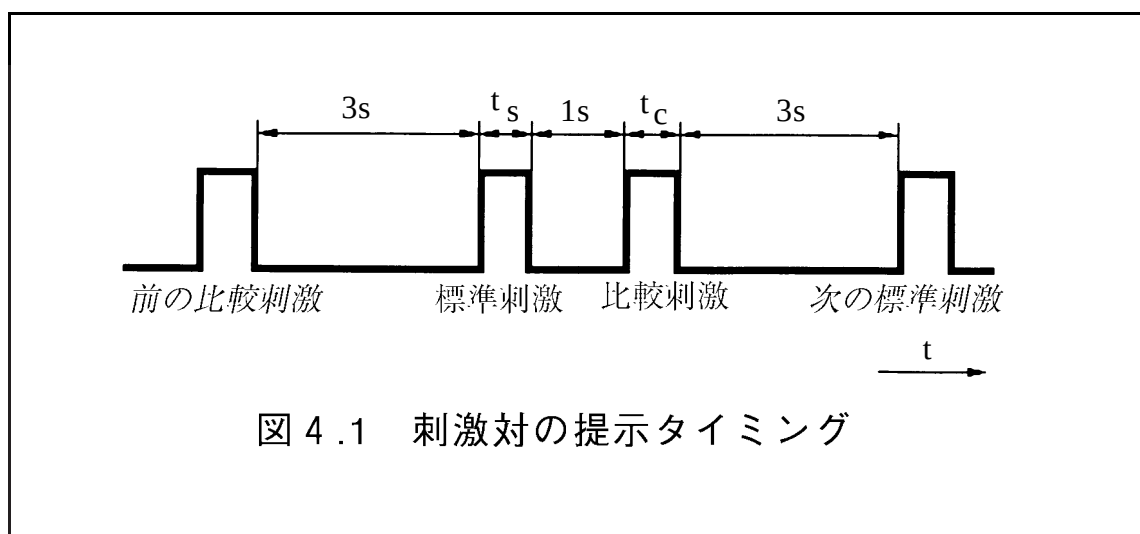
単音節単位の編集合成で、より聞きやすい出力音声を得るためには、単音節個々の物理的な特性と心理的な影響の関係を解明して、特性の不均一に起因する不自然さをなくしていくことが重要である。振幅、ピッチ等とその心理的な影響については、これまで行われてきた内外の研究成果を適用することができるものもある^{2,3)}。しかし、音韻長など別に検討しなければならない単音節固有の特性もある。

これら単音節固有の物理量と心理的な影響のうちの、音韻継続時間については、音声学的見地から、自然に近い条件で発声された単語や文章を分析し、発声された音節の音韻継続時間は音節の種類、出現順序、前後の音素などに依存していることが報告されている⁵¹⁻⁵⁶⁾。また、単音節の音韻継続時間の伸縮が明瞭性に及ぼす影響についての報告もあるが⁵⁷⁾、音節の音韻継続時間の物理的な増減とその直接的な心理的影響については詳しく調べられていない。この特性は、単音節単位の編集合成装置における音韻継続時間の指標として必要であり、また前章で提案したような単音節による明瞭性の試験に用いる標準音声の指標としても重要である。

そこで、本章では、この単音節音声の音韻継続時間の直接的な心理的影響の評価、すなわち2つの音韻継続時間や種類の異なる音節が連続して提示されたとき、聴取した人はそれぞれの単音節の長さについてどう感じるかについて実験を行い、結果について統計的に検討する。

4.2 実験方法

この章では、大きく分けて3種類の実験について説明している。以下、実験の手順、聴覚刺激である音節の音韻継続時間の変化方法、刺激音声の合成法、



および被験者の詳細について述べる。

4. 2. 1 実験の基本的な手続き

それぞれの実験では、標準刺激と比較刺激の対で構成される聴覚刺激を使用する。このうち標準刺激には、本章の全ての実験で常に音韻継続時間が 200ms の単音節/a/を用いる。比較刺激は比較の対象となる音節で、音韻継続時間は提示のつど実験の手順に従って増減される。

図4. 1に刺激提示のタイムチャートを示す。通常は図のように標準刺激を先に提示し、1秒間において比較刺激を提示する。刺激対と前後の刺激対との間には3秒の無音状態をおいている。標準刺激と比較刺激の時間間隔については、これまで同種の実験は行われていないので、1秒間に設定して簡単な予備実験を行ったところ、不都合が認められなかったので、本実験でもこの値をとることとした。前後の単音節の音韻時間長を比較するので、この刺激間隔の長さが影響する可能性も十分予想される。本章の実験では、刺激の間隔との関係について検討は行わず、間隔の影響が出ないように、刺激の間隔はどの実験条件でも1秒間に固定した。無音状態の3秒間についても、予備実験の結果妥当な値と認められた。この無音状態の時間は、短すぎると刺激対ごとの区別がつきにくくなるうえ、回答を筆記する余裕がなくなる。反対に長くすると、実験全体の所要時間が増えて、被験者の負担増となり、実験結果の信頼性の低下をまねく危険がある。

被験者は一对の刺激を聞き終えた直後に、標準刺激と比較刺激の“長さ”について主観的に判定し、次の対が提示されるまでの間に回答するように教示される。回答は回答用紙の所定の枠内に筆記する。標準刺激よりも比較刺激が長いと判断したら＋を、短いと判断したら－を、同じと判断したら○を記すように教示される。

実験は一度に1名ずつ行い、聴覚刺激はヘッドフォンを介して提示される。毎回の実験に先立ち、実験に用いる音節と同じ音節を、実験とは関係のない系列をそれぞれの被験者に聞かせて、聞きやすい音量に調節させた。

4. 2. 2 刺激音声の音韻継続時間の変化方法

本章の実験では、いずれも対になった聴覚刺激のうち、比較刺激の音韻継続時間を提示するごとに変化させて被験者に提示する。

刺激の物理的な特性を変化させて、感覚における等価な点（主観的等価点）などを求める測定実験を精神物理学的測定法と呼ぶ³⁸⁾。精神物理学的測定法では、主として“調整法”，“極限法”，“恒常法”の3種類の基本的な刺激の変化方法と、これらの基本的な方法から派生した各種の手法が用いられる。

調整法は、明らかに大きいか小さいかが誰にも明確に解る点から出発し、被験者自身が調整つまみ等の操作によって刺激を連続変化させて、等しいと感じる点に導く実験方法である。行きすぎたら何度戻ってもよい。明らかに大きい

第4章

点、小さい点からの実験を同数行い、各回の平均値を主観的等価点とする。また、確率誤差 PE（標準偏差 SD の 0.6745 倍）を弁別閾（区別できる最小の差）とする。光や連続音のように、刺激が連続的に提示されていて、刺激の物理的な性質の変化とつまみの対応をつけやすいものの実験に適している。

極限法は、調整法のように刺激を被験者が操作して連続変化させるのではなく、あらかじめ設定された刻み幅で、刺激が段階的に変化し、各刺激に対して大きいか、小さいか、等しいかを答えさせる。回答の選択肢が“等しい”を含めて3種類のものを3件法、“大きい”、“小さい”に限定されるものを2件法と呼ぶ³⁸⁾。本章の実験では全て3件法を用いている。明らかに大きい点から初めて等しい点を通り明らかに小さい点に達する下降系列と、その逆の上昇系列とを同数行う。下降系列で答えが大きいから等しいに転じる値と、上昇系列で小さいから等しいに転じる値を平均して主観的等価点とする。また、両者の差の半分を弁別閾とすることもある。

恒常法は、あらかじめ決められた刺激値をランダムな順番で被験者に提示する。3件法では、大きい、小さいの回答が 50%となる刺激値を求め、その平均値を主観的等価点、その差の半分を弁別閾とする。

本章の実験では、図4. 1のように標準刺激と比較刺激の対を提示するのが基本となる。刺激の長さの合計は約 0.4~1 秒になるので、次の刺激対との間隔 3 秒を加えた 1 サイクルは 4.4~5 秒となる。このため、調整法を採用した場合、被験者の操作が、比較刺激の時間長に反映されるまで最大 5 秒ほどかかってしまう。次の対との時間間隔、標準刺激と比較刺激の間隔を狭めていくと、刺激対の間の区別が付きにくくなるので、狭めたとしても、1 サイクル 3.5 秒程度までが限界であろう。調整法は、連続的に提示されている刺激の心理的等価点を求める実験にはむくが、本章の実験のように刺激が離散的に発生される場合、被験者の調整が刺激の長さの変化に反映され、被験者に刺激として提示されるまでの時間遅れが大きく、等しい点を求める操作が非常に難しい。また、後述するように、使用する実験装置では性能的に実時間の合成が不可能なので、調整法が前提とする連続変化は厳密には実現できない。このため、あらかじめ実験に使用する音韻長の単音節を合成しておく必要があるが、擬似的な連続を実現するためには、音韻長の刻みを小さくとる必要があるため、各音節について用意すべき音韻長の種類は非常に多くなってしまう。

このように、実験での被験者の操作の困難さや、実験装置の制約があるため、本章の実験では極限法を採用することにした。極限法であれば、あらかじめ、設定した音韻継続時間の刻み幅の単音節を合成しておけばよく、被験者の操作が入らないので実験手順も単純である。この他弁別閾の検証のために、恒常法に基づいた方法も一部用いた。

以下、実際に実験で用いた刺激の変化方法について、具体的に説明する。

表 4.1 実験に使用した単音節（*印を付したもの）

a*	i*	u*	e*	o*	ba*	bi*	bu*	be*	bo*
ka*	ki*	ku*	ke	ko	pa	pi	pu	pe	po
sa*	fi*	su*	se*	so*	kja		kju		kjo
ta*	t/i*	tsu*	te*	to*	fa*		fu*		fo*
na*	ni*	nu	ne	no	tfa*		tfu*		tfo*
ha*	hi*	hu*	he*	ho*	nja		nju		njo
ma*	mi*	mu*	me*	mo*	hja*		hju*		hjo*
ja*		ju		jo	mja*		mju*		mjo*
ra*	ri*	ru	re	ro	rja		rju		rjo
wa*					gja		gju		gjo
ga*	gi	gu	ge	go	d3a		d3u		d3o
za*	d3i	zu	ze	zo	bja*		bju*		bjo*
da*			de	do	pja		pju		pjo

（1）連続変化

用いた方法は極限法を基本としている．比較刺激の音韻継続時間は変化させる範囲内の最も小さい（大きい）値から開始し，提示するごとにあらかじめ設定した刻み幅だけ増加（減少）させ，変化させる範囲の上限（下限）の値まで提示を繰り返す．この増加（上昇系列），減少（下降系列）を交互に数回繰り返す．

後で詳しく説明する，4. 3. 1の被験者ごとの弁別閾について検討するための実験では，この提示手順に基づき，標準刺激として継続時間が200msの/a/を，比較刺激にも/a/を用い，継続時間は150～270msを10ms刻みで使用した．

同実験では刺激系列を2組用意した．一方は上昇系列から始まり，上限に達した時点でとぎれなくそのまま下降系列に移行し，以降，上昇・下降を繰り返す．刺激の系列は単純なので，規則性を判断の手がかりにされる危険もある．そこで，系列が折り返す上限の値は250, 260, 270ms，下限の値は150, 160, 170msのそれぞれの中から折り返すごとに違う値を用いた．被験者に提示する刺激は練習のための上昇系列1と，続く下降・上昇2組の合計5系列で，練習をのぞく4系列について集計した．

もう1種類の系列は練習のための系列が下降系列から始まり，その後上昇・下降系列が続く他は，同じである．

第4章

4. 3. 2のそれぞれの単音節音声の主観的に等価な長さを求める実験では標準刺激として音韻継続時間が200msの/a/を用いて先に提示した。比較刺激は主観的に等価な長さを求める対象の音節を用い、標準刺激の1秒後に提示した。実験に先立ち、実験に使用する標準刺激と比較刺激を実験者が聴取して簡単な予備実験を行い、それぞれの音節の音韻継続時間の範囲を決定した。

同実験では、各音節の最初の刺激系列は常に上昇系列で、比較刺激は音韻継続時間の範囲内の最小値から出発し、20msずつ増加して、範囲の最大値まで提示を繰り返す。最大値で5秒間の休止を設け、下降系列に転じる。同じく20msずつ減少して提示を繰り返す。上昇系列と下降系列は交互に合計4系列提示される。

この多数の音節を用いた実験では、実験に要する時間が非常に長くなる。実験が長時間に及ぶと被験者の集中力が低下して、実験結果の信頼性の低下をまねくことが懸念される。そこで、少しでも被験者の負担を軽減するために、他の実験方法よりも回答方法を簡略化した。上下それぞれの系列で、最初に標準刺激と比較刺激の長さが同じに聞こえた刺激に○をつけるように教示した。上昇、下降系列で○のつけられた比較刺激の音韻継続時間の平均値を、標準刺激と等しく聞こえた音韻継続時間とした。

1つの音節の上昇、下降4系列の提示のあとは30秒または1分間の小休止をとり、小休止後に次の音節の実験に移る。比較刺激の音節は表4. 1に示す日本語の単音節のうち＊印を付したものをを用いた。実験は一度に1名ずつ、13～16種の音節を用いて行った。

(2) ランダム変化

連続変化の実験は、目的とするそれぞれの音節の主観的な長さを求めることに主眼をおいているが、各被験者の弁別閾をより詳しく調べ、適切な刺激の変化の刻み幅について検討するために、恒常法（比較刺激をランダムに変化）に基づく実験を行った。4. 4の実験がこの方法による。

比較刺激は160～260msの範囲内で、10ms刻みとして、それぞれの音韻継続時間の音節の出現回数が同じになるようにして、可能な限りランダムな系列を作成した。

また、同じ4. 4では、音韻継続時間を変化させる比較刺激を先に提示して、1秒間の無音状態の後200msの標準刺激を提示する、他の実験とは逆の提示手順による実験も行った。この実験では先に提示した比較刺激よりも後の標準刺激が長いと判断したら＋、短いと判断したら－、同じと判断したら○を記すように教示した。その他の条件は同一である。

4. 2. 3 刺激音声の作成

本章の実験では、比較刺激として広い範囲の音韻継続時間をもつ音節を提示する必要がある。しかも、音韻継続時間に関わらず、振幅、ピッチ周波数、フォ

ルマント周波数などの特性を等しくする必要がある。そこで、もともになる音声を用意して、この音声に時間的な伸縮を施すことにした。

音声のピッチやフォルマントなどの特性を変化させずに伸縮を行う方法として、適当な間隔で、同じ部分を繰り返したり、適当な長さを間引くことによる方法と、分析・合成による方法などが古くから考案されている^{2,3)}。

繰り返したり間引く方法は、古くは特殊な回転ヘッドを用いたテープレコーダによって実現されていた。これは円筒形の回転ヘッドに2、3個等、複数個の再生ヘッドを装備した、特殊なテープレコーダである。テープは回転ヘッドの外周に、ヘッドが2個の場合は180度（ヘッドが3個の場合は120度）で接するようにできていて、回転ヘッドがどの角度であっても常時一つのヘッドがテープに接するようになっている。

テープを通常で再生して、回転ヘッドを回転させなければ、固定式ヘッドの通常のテープレコーダと全く同じで、録音したときと同じ速度で再生される。テープを通常よりもゆっくり再生して、遅くなった分を補うように、テープの走行方向とは逆方向にヘッドを回転させる（テープと回転ヘッドの相対速度は録音時と一致させる）と、ピッチなど周波数特性に変化はないが再生に要する時間が長くなる。回転ヘッドによって同じ部分が繰り返して再生されるためである。同様にテープの速度を通常よりも速くして、速くした分だけヘッドをテープと同じ方向に回して、テープと回転ヘッドの相対速度は録音時と同じに保つと、テープに録音された音声は間引かれて再生され、ピッチなどの周波数特性に変化はなく、再生時間が短縮される。原理的には単純で、50%～200%の伸縮に向いているとされる。

最近の半導体メモリーの大容量化、高速化、低価格下化により、原理的に回転ヘッドのテープレコーダと同じ処理を、メモリーに蓄えた音声データに対して行うことが可能になった。メモリー上にある音声データを数十～数百ミリ秒単位のフレームに分割して、伸ばす場合には各フレームの最後の部分を、伸ばす割合に応じた長さの分繰り返して再生する。縮める場合は、各フレームの最後の部分を縮める割合に応じて読み飛ばす。

以上の方法では、原理は簡単で、特別な信号処理も行わずに、音声を伸縮することができる。しかし、音声波形の位相等の条件を全く考慮せずに機械的に切ってつなぐため、つなぎ目ごとに雑音が発生してしまう。ヘッドの切り替わり（繰り返しや、間引きの切り替わり）の際、切る側の振幅を徐々に絞る、つなぐ側の振幅を徐々にあげるような工夫を加えることで雑音の発生は改善できて、長い文章を聞き取る場合には十分な了解度を得られるようになるが、それでも位相を合わせずにつなぐので、本章の実験のように単音節だけを聞く場合には、明瞭性の低下は重大である。

この繰り返したり間引く方法でも、波形の周期性を見極めて、波形の位相を

第4章

合わせながら切り貼りすれば、音声の品質をあまり損なわずに単音節の伸縮が可能である。しかし、このためには一般に目視等で周期性を見極めて切り貼りする手作業が必要となる。

最近では、周期の識別を自動的に行い、最適な間引き、繰り返しを決定するアルゴリズムを適用して、自動的に波形を間引いたり、繰り返すことで、準リアルタイムに音声を伸縮できる装置が、話速変換装置として実用化を目指して開発されている^{58, 59)}。

もう一つの伸縮法は、音声の分析・合成法による方法である。もともと分析・合成法は、音声の圧縮伝送を目的に研究が行われていた。分析・合成法では、音声波形の中から、時間的にあまり大きく変化しない情報を分析により抽出して、もとの音声波形自体を伝送するよりも少ない情報量として伝達することで、音質の劣化を最小にとどめて音声波形を合成することを目指している。

代表的なものがPARCOR分析・合成である⁹⁾。PARCOR分析では、音声を数十ミリ秒に区切ったフレームそれぞれについて、PARCOR（偏自己相関）係数と呼ばれる数値を求める。PARCOR係数の他、ピッチ、振幅および有声度[†]などが求められる。

合成の際にはPARCOR係数、ピッチ、振幅、有声度から音声を合成するが、合成時の1フレームの長さを分析時の1フレームの長さよりも長くしたり、短くしたりすることで、音声を伸縮することができる。

本研究では、回転式ヘッドを利用したテープレコーダの原理による方法は、位相に関係なく音声波形を切り、つなぐため、単音節としては音質に及ぼす影響が無視できないので採用しなかった。また、波形の周期性を見きわめて、間引き、繰り返しによって実験のための音声を作成するには、膨大な手作業が必要となり、実験を行った当時は効率よくこの作業を行うための装置も利用できなかったもので、この方法の採用も見送った。

音韻長の伸縮には、比較的音質の劣化の少ないPARCOR分析・合成による方法を採用ことにした。以下、PARCOR分析・合成法による音声の伸縮方法について具体的に説明する。

分析・合成の音源には、市販されている明瞭度試験用テープを用いた³⁴⁾。このテープから男性アナウンサーの発声した単音節音声を、図4.2の装置によって、サンプル周波数20KHz, 12ビットでA/D変換し、数値として計算機に取り込み、表4.2に示す仕様によってPARCOR分析を行った。音節の音声波形をそれぞれ12.8msの時間窓によって分割して窓関数をかけ、PARCOR係数、ピッチ周期、振幅値、有声度を求めた。このうち振幅値については、それぞれ

[†]分析しているフレームにおいて、音声がある音声としての性質が優勢であるか、無声音としての性質が優勢であるか、あるいはその中間の性質であるかを示す係数。

の音節の最大値が等しくなるように音節全体の振幅を調整した。このほかの係数には加工処理は行わなかった。合成の際に各時間窓の時間長を調整することにより、全体の継続時間が目的とする時間長と等しく、かつ原音声と同じピッチ、フォルマントの変化パターンをもつ任意の時間長の音節を合成することができる。

実験に用いた計算機ではその性能上、実時間の音声合成が不可能なので、実験に用いる音韻継続時間の音節を事前に合成して、ハードディスク装置に保存した。実験の際には、実験用プログラムで、タイミングをとりながら、ハードディスク装置から合成音声のデータを D/A 変換器へ送り、オーディオアンプ、ヘッドフォンを介して被験者に提示する。

4. 2. 4 被験者

被験者は年齢 26～48 歳の 8 名の成人男性で、全員正常な聴力を有する。

4. 3 単音節の主観的に等価な長さ

音韻継続時間を比較する際の弁別閾などについては、過去の文献にはほとん

表 4. 2 A/D変換およびPARCOR分析の仕様

サンプル周波数	20 kHz
量子化ビット数	12 ビット
分析窓長	25.6 ms (512 サンプル)
窓の移動周期	12.8 ms (256 サンプル)
予測次数	15 次

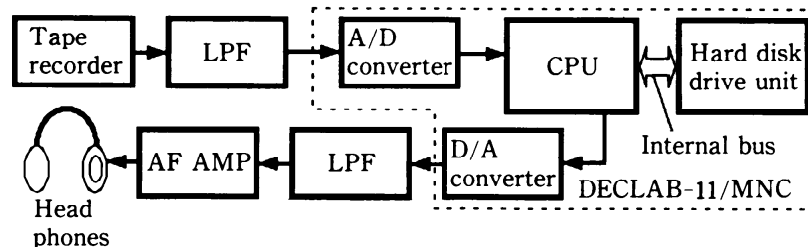


図 4.2 実験装置構成図

第4章

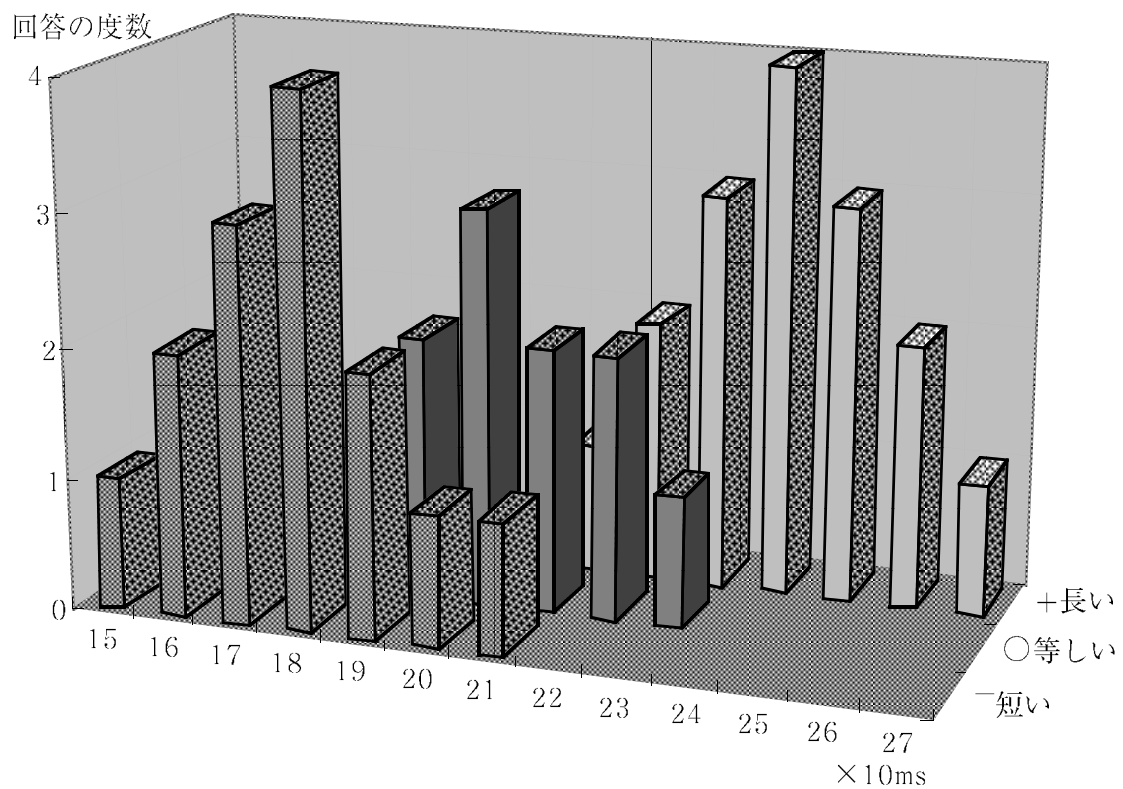
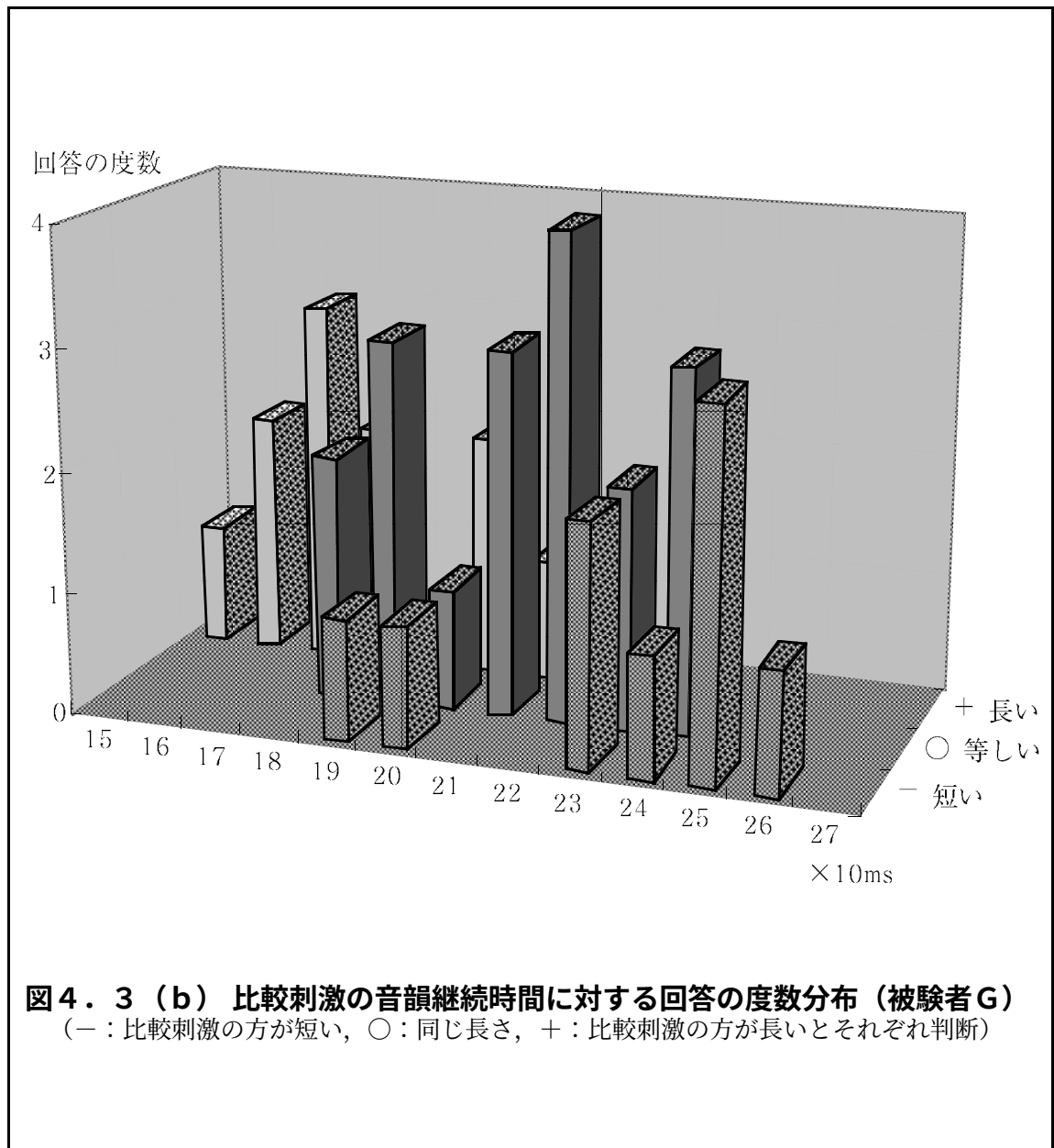


図4.3 (a) 比較刺激の音韻継続時間に対する回答の度数分布 (被験者D)
 (- : 比較刺激の方が短い, ○ : 同じ長さ, + : 比較刺激の方が長いとそれぞれ判断)



第4章

ど情報がない．そこで、最初に予備的な実験を行って、各被験者の音韻継続時間に対する反応の傾向を調べ、その結果をもとに被験者の選別を行い、引き続き音節/a/と他の音節との主観的な長さを等価にする実験を行った．

4.3.1 被験者ごとの弁別閾

被験者は合計8名で、4.2.2(1)に記述した連続的に音韻継続時間を変化させる方法を用い、それぞれの系列について4名ずつの被験者に提示する実験を行った．回答は、それぞれの刺激対について、比較刺激が長いと感じたら“+”を、短いと感じたら“-”を、等しいと感じたら“○”を筆記させた．

図4.3は、被験者DとGの回答の分布である．図中の+、○、-で示されているパターンの棒グラフは、それぞれ被験者が+、○、-を回答した度数を表している．被験者DとGの回答の分布は非常に異なったパターンになっている．図4.3(a)の被験者Dの回答では、○は横軸の中心付近に集中して分布しており、+はそれよりも音韻継続時間が長いほうに、-は短いほうに分布している．そして、+、-が○と重複している領域は小さい．

これに対して図4.3(b)の被験者Gの回答では、+と○、-と○が広い範囲で重複している．この図は被験者DとGとでは音韻継続時間の長短を判断する能力、あるいは判別するために被験者が得た何らかのスキルに差があることを示している．

表4.3に、被験者が○を付した比較刺激の音韻継続時間の平均値、各系列の全刺激の平均値、および系列別の被験者の集計を示す．この表に基づき、各被験者の回答の分布について統計的な検討を試みることにする．

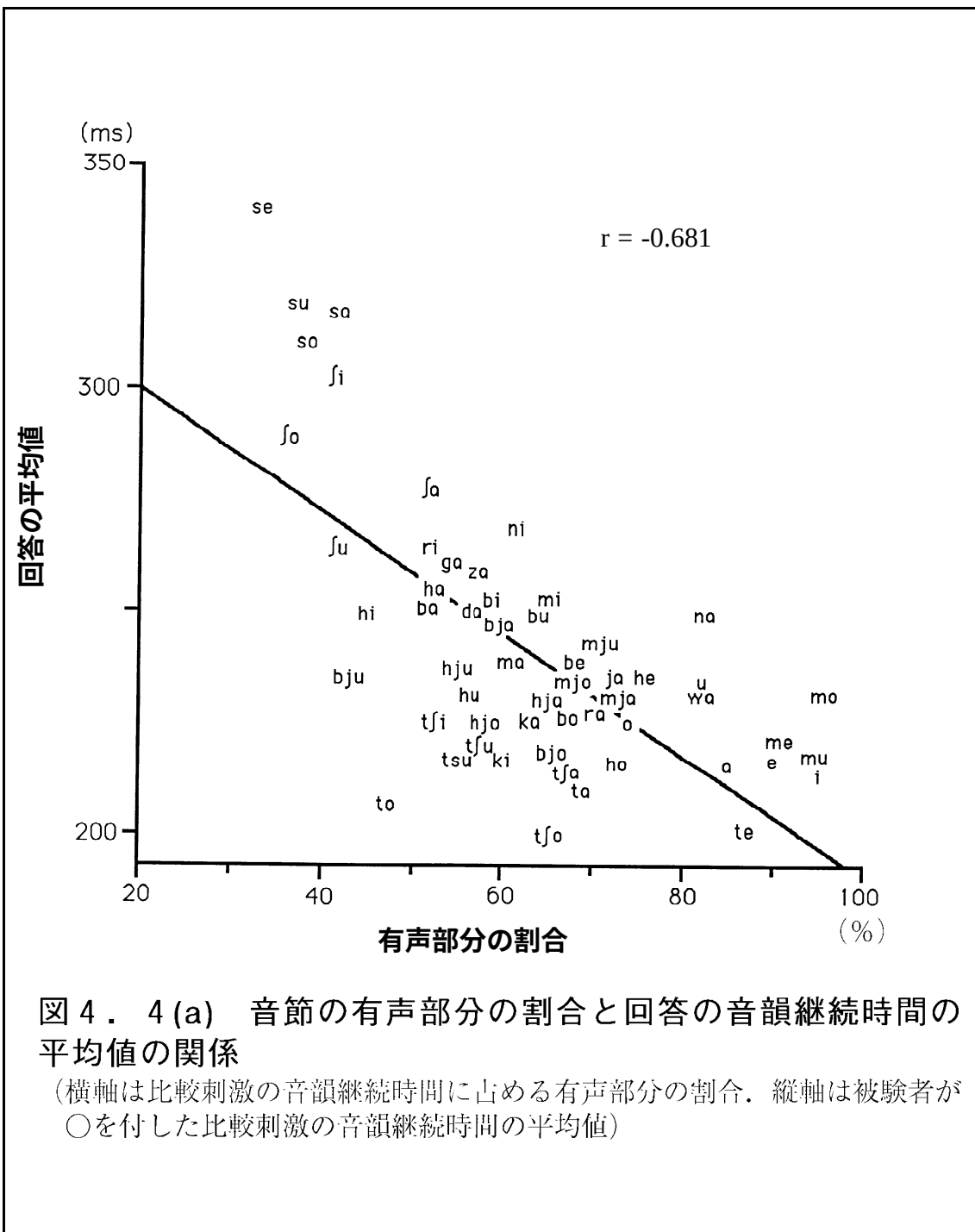
8名の被験者の各平均値は201.3～215.7msと集中して分布している．仮に被験者が音韻継続の時間の差異を判別できず、ランダムに回答しているとする、回答の平均値は、表4.3に示す比較刺激全体の平均値である205.9msまたは211.0msに近づくため、平均値から各被験者の回答の傾向を探ることは困難である．

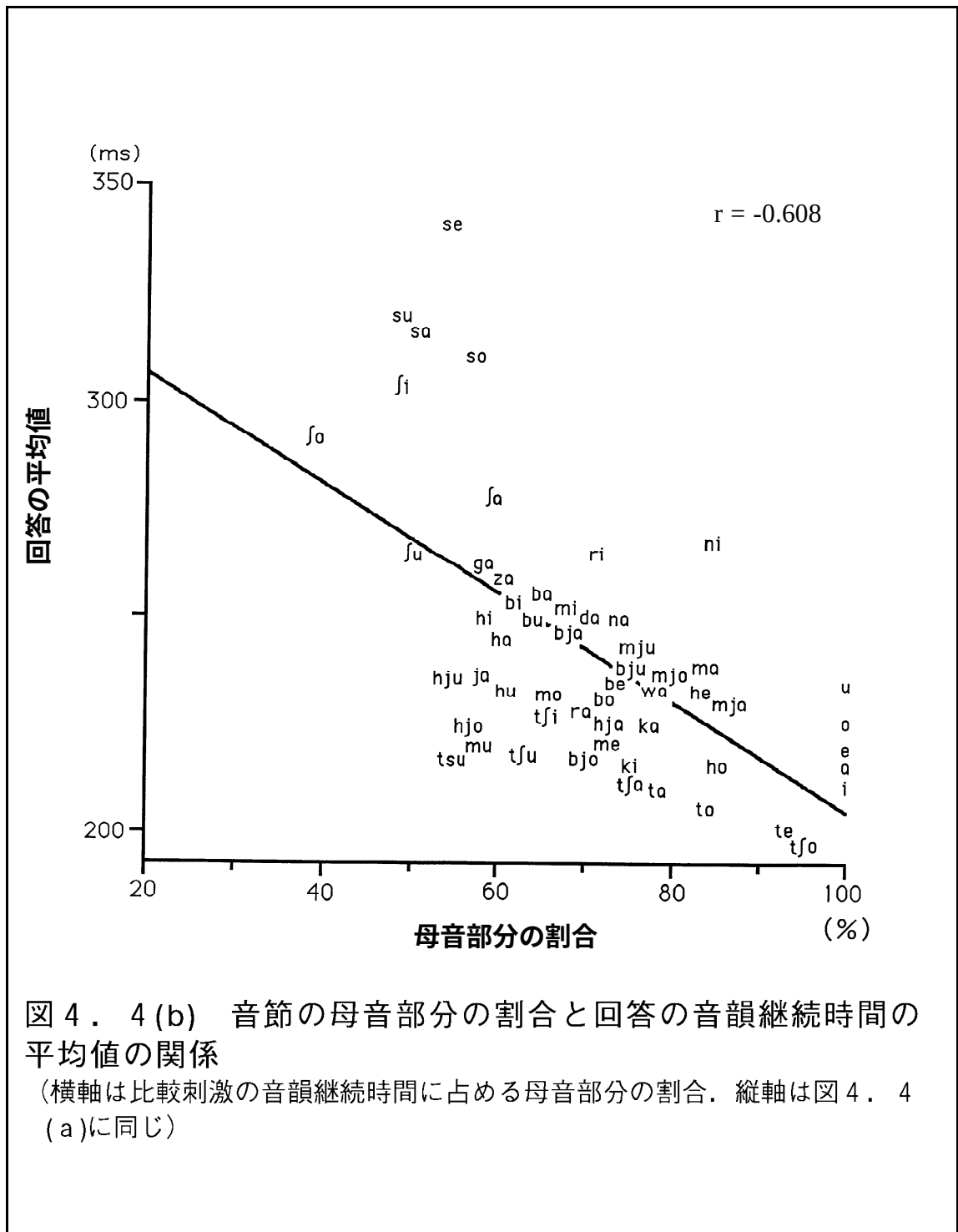
一方、標準偏差は8.7～20.7msと個人差が大きい．被験者が音韻継続時間の差異を判別できず、ランダムに回答していると仮定すると、その被験者の○を付した回答は、平均値だけでなく、標準偏差も、刺激系列の全刺激の標準偏差とほぼ等しくなる．反対に、被験者が正確に音韻継続時間を判別（弁別閾が小さい）すれば、○の回答分布はある特定の値に集中し、標準偏差は小さくなる．このことから、各被験者の○の標準偏差と、刺激系列全刺激の標準偏差の差異が、その被験者の音韻継続時間の差異を判別する能力・スキルを反映していると考えることができる．

そこで、標準偏差の代わりに分散を用いて、全刺激の分散とそれぞれの被験者の回答の分散との差について、Fテストによる検定を行った^{48, 49)}．この検定における帰無仮説は“被験者が○を付した刺激の音韻継続時間の分散は、全刺

表 4. 3 実験結果およびF検定の結果

被験者別集計	被験者	系列	n	平均音韻継続		分散	F値	p
				時間 (ms)	標準偏差			
被験者別集計	A	2	8	205.0	8.7	75.0	10.4	<0.01
	B	1	9	211.1	9.9	98.8	8.9	<0.01
	C	2	8	201.3	11.7	135.9	5.7	<0.05
	D	1	10	207.0	12.7	161.0	5.5	<0.01
	E	2	9	204.4	14.2	202.5	3.9	<0.05
	F	1	14	215.7	18.8	353.1	2.6	
	G	2	18	212.2	19.6	384.0	2.2	
	H	1	17	211.8	20.7	426.3	2.2	
系列別全刺激の集計		1	41	211.0	31.1	964.9		
系列別被験者の合計		2	39	205.9	29.4	865.2		
(B, D, F, H)		1	50	211.8	17.4	302.7		
(A, C, E, G)		2	43	207.2	16.2	261.8		





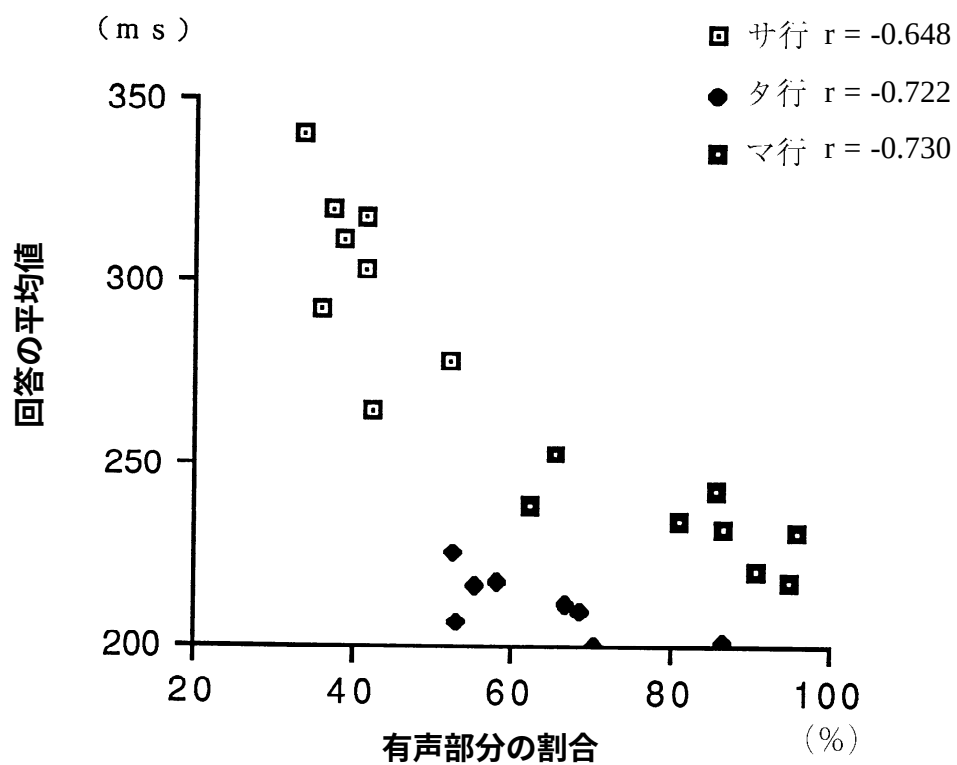
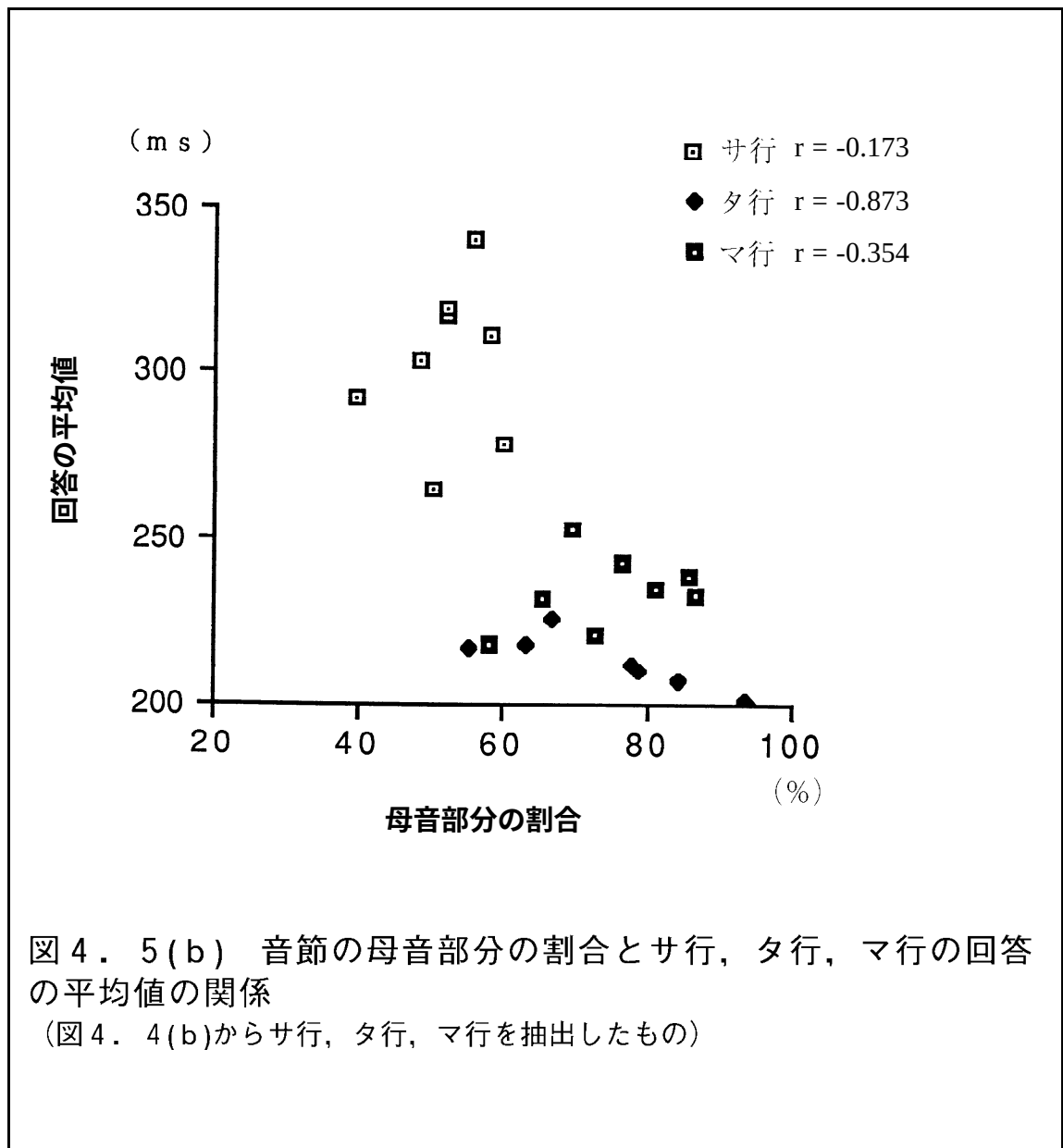


図4.5(a) 音節の有声部分の割合とサ行、タ行、マ行の回答の平均値の関係

(図4.4(a)からサ行、タ行、マ行を抽出したもの)



第4章

激の分散に等しい”であり、両側検定である。帰無仮説が棄却されて両者の分散が等しくないと言えれば、その被験者の○を付した音韻継続時間の分散は全刺激の分散とは異質であるといえ、その被験者は音韻継続時間を判別する何らかの能力をもつとみなすことができる。被験者Aの場合についてみると、Aには系列2が提示されたので、表4. 3から系列2の刺激数 n_1 は39、分散 s_1^2 は865.2である。一方、○を付した刺激の数 n_2 は8で、その分散 s_2^2 は75.0である。以上から、AのF値は以下ようになる。

$$F = \frac{s_1^2 n_1 (n_2 - 1)}{s_2^2 n_2 (n_1 - 1)} \quad (4. 1)$$
$$= 10.4$$

その他の被験者のF値も同様に計算し、表4. 3に示した。

求めたF値と自由度から、F検定の表を参照して各被験者の結果について検定を行った。表4. 3で○印を付したものは、それぞれの有意水準が1%と5%で、回答と刺激全部の分散の差が有意であることを示している。

表4. 3の系列別の被験者の集計では、平均値、標準偏差とも差がないので、系列の違いによる実験結果への影響はほとんどないといえる。

以上のように、簡単な判定の実験にも関わらず、結果の分散（標準偏差）に大きな差が現れた。今後の実験では刺激対に異種の音節を用いたり、音韻継続時間をランダムに変化させるので、判定の困難さが大きく増すと考えられる。そこで、F検定の結果と回答の分布を考慮して、有意水準5%で有意であったAからEの5名の被験者のみ、実験に参加させることにした。

4. 3. 2 単音節の主観的に等価な長さ

4. 2. 2 (1) に記述した方法により、標準刺激/a/と各音節とを比較することにより、それらの音節の主観的に等価な長さを求めた。

図4. 4(a), (b)はいずれも、それぞれの単音節が200msの/a/と心理的に等価な長さとなったときの物理的な音韻継続時間の平均値をまとめたものである。縦軸にはいずれも物理的な音韻継続時間の平均値をとり、(a)では、横軸にそれぞれの単音節の音韻時間中に占める有声部分の割合を示している。この有声部分の割合とは、PARCOR分析・合成により単音節を伸縮する際、各フレームについて、有声音としての性質が強い、無声音としての性質が強いを示す変数として求められる有声度が0.5以上になるフレーム数の、その単音節を構成する全フレーム数に対する割合である。PARCOR分析によって得られる有声度をそのまま用いているので、母音/a/, /i/, /u/, /e/, /o/についても、声のかすれ等により、無声的な性質が優勢となったフレームも存在するので、有声部分の割合が

100%にならなかった。図中の直線は、それぞれの関係について最小2乗法で求めた直線（回帰直線⁶⁰⁻⁶²）という）である。

一方、図4. 4(b)では、横軸に各単音節の音韻継続時間に占める母音部分の割合をとった。子音を含む単音節は、各フレームの有声音度を参考にしながら、PARCOR係数にFFTをかけて求めた、平滑化された周波数スペクトルのパターンを観察して、フォルマントに変化が出始めたフレーム以降を母音のフレームとした。母音だけの単音節は、母音の占める割合を100%とした。

図4. 4(a), (b)は、いずれも横軸と、得られた音韻継続時間の平均値とが負の相関を示しているが、横軸に有声音の割合をとった図4. 4(a)のほうは-0.681と、より大きい相関を示している。

図4. 5(a), (b)は図4. 4(a), (b)の部分集合で、サ行、タ行およびマ行の音節と、有声音部分、母音部分との関係を示している。特に図4. 5(a)では、サ行、タ行、マ行などの同じ行で、主観的に等価な長さをもつ音節の音韻継続時間と、音節の有声音の割合、および音節の母音部分の割合は限定された範囲に分布しており、相関係数もきわめて大きい値となっている。

図4. 4, 5から、各単音節の有声音部分の割合と主観的に同じ長さとなる物理的な音韻継続時間とは、全体的に負の相関、もしくは逆比例の関係があり、さらに、子音の種類を特定すれば、より厳密に主観的な長さを等しくすることが可能であることが示唆される。

4. 4 比較刺激の刻み幅の最適値

前節では、音節の音韻継続時間の差異を認識する何らかの能力もしくはスキルを有していると解釈された5名の被験者のみ異種の音節を比較する実験に参加させた。そこで、同じ5名の被験者について、4. 2. 2(2)で説明したランダムな系列を用いる方法（恒常法）によって実験を行い、それぞれの被験者の音韻継続時間の差異に対する回答の傾向をより詳しく調べた。

この節の実験では標準刺激、比較刺激ともに/a/を用い、比較刺激の刻み幅は10msとした。最初に基準刺激を提示する実験では、後に提示する比較刺激が前の基準刺激よりも長いと感じたら“+”，短いと感じたら“-”，等しいと感じたら“○”を所定の回答用紙の欄に記すように教示した。また、刺激対の順番を逆にした実験では、後に提示する標準刺激の長さが長いと感じたら“+”，短いと感じたら“-”，等しいと感じたら“○”を回答欄に記させた。

表4. 4は被験者が○を付した比較刺激の音韻継続時間の平均値と標準偏差である。表4. 5は比較刺激を先に提示した実験の平均値と標準偏差である。

表4. 4では、被験者E以外は、音韻継続時間を連続的に変化させた実験結果（表4. 3）よりも標準偏差が大きくなっている。これは同種の実験を繰り返したことによる練習効果よりも、疑似乱数による提示により予測が不可能な

表4.4 疑似乱数による刺激系列を提示したときの結果

被験者	n	平均音韻継続 時間 (ms)	標準偏差	(4.3)式の値 (ms)
A	17	200.6	12.1	30.3
B	12	213.3	11.8	29.5
C	11	212.7	21.4	53.5
D	9	211.1	15.2	38.0
E	13	207.7	8.9	22.3
合計	62			
平均		208.2	13.9	34.7

表4.5 疑似乱数による系列で，かつ刺激対の標準刺激と比較刺激の提示順を逆にしたときの結果

被験者	n	平均音韻継続時間 (ms)	標準偏差
A	19	190.5	16.4
B	13	186.9	12.6
C	17	198.2	23.6
D	12	194.2	13.8
E	15	202.7	22.4
合計	76		
平均		194.6	19.4

ことによる困難さのほうがこの実験の結果に支配的に現れることを示している。

表4. 5に示した刺激順が逆の場合の実験結果では、すべての被験者の標準偏差はそれぞれ表4. 4よりも大きくなっている。一見すると刺激順が逆の場合のほうで、被験者にとって音韻継続時間の差異の判別が困難なようである。

これは、次のような理由によると考えられる。被験者は頭の中に長さの基準を置いて、この基準を毎回基準刺激で校正して、次にくる比較刺激の長さと比較する。このため、刺激対の順番を逆にすると、頭の中の長さの基準が校正できず、正確な判定ができなくなるのであろう。

もうひとつ、標準偏差が大きくなった原因として考えられるのは、2番目の刺激の立ち上がりのタイミングがとりにくくなることである。標準刺激が先に提示される実験では、200msの/a/が提示され、無音の1秒間を経過した後に比較刺激の提示が開始される。従って、比較刺激が開始されるのは、標準刺激の/a/の提示が開始されてから1.2s後で常に一定である。被験者は標準刺激を聞いて基準を構成しながら、次に提示される比較刺激についても、提示される立ち上がりを予想した上で十分に注意を集中した上で、聞き取ることができる。

これに対して比較刺激が先に提示されると、2番目に提示される標準刺激の立ち上がりは、比較刺激の開始からは比較刺激の音韻継続時間の変動に伴って変動する。このために、2番目に提示される標準刺激に対して十分に注意が集中できないと考えられる。

表4. 4には、すべての被験者の平均値が200ms以上となる順序効果が表れている。表4. 5でも同様の傾向がみられる。これらの結果は、被験者にとっては最初に提示した刺激のほうで後から提示した刺激よりも長く感じられるということを示していて興味深い。

日本語を母国語とする話者に2つのモーラからなる単語を自然に発話させ、その音声进行分析した報告によると、その前後のモーラ[†]の音韻継続時間に差はないとされている^{54,55)}。したがって、知覚される時間の長さと発話される音韻長との関係は特にないといえる。

ここで、標準偏差に着目すると、表4. 4では標準偏差は8.9～21.4msの値をとっている。それぞれの音韻継続時間の音節を4回ずつ提示するとして、0でない最小の標準偏差は、ある音韻継続時間の度数が4、それに隣接する音韻継続時間の度数が1の場合である。比較刺激の音韻継続時間の刻み幅を a とすると、この場合の標準偏差 σ は、

$$\sigma^2 = \frac{\sum (x_i - \bar{x})^2}{n} \quad (4. 2)$$

[†]音節の長さを測る単位。通常は短母音を含む単音節。

第4章

$$\sigma = \frac{2}{5}a$$

である。反対に、0でない最小の標準偏差の値を x としたいとき、刻み幅 a は、

$$a = \frac{5}{2}x \quad (4.3)$$

となる。被験者の標準偏差について4.3式の値を計算したものが表4.4に示してある。

4.3節の実験条件のうち、音韻継続時間の刻み幅を大きくしていったとき、その値が表4.4の(4.3)式の値に近づくと、その刻み幅は被験者にとって十分小さいとはいえず、その結果の精度は粗いということになる。

異種の音節を比較した4.3.2の実験では、音韻継続時間の変化の刻みを20msとしたが、その実験結果では、最も標準偏差の小さい被験者Eの場合に(4.3)式の値が22.3msと、刻み幅とほぼ同じ値となる程度で、他の被験者の値は20msよりもかなり大きな値になっている。したがって、刻み幅が20ms程度の場合は、刻みの粗さの実験結果の精度への影響は大きくないということが裏づけられる。

加えて、この実験では比較刺激は/a/以外の音節を使用しているので、異種の音節を比較する困難さも加わる。このことから、刻み幅は20ms程度で、結果の精度に及ぼす影響は少ないといえる。

このように、表4.4の結果から、音節の音韻継続時間の刻み幅が結果の精度に及ぼす影響について検討することはできた。

4.5 章のまとめ

音声の伝達する情報のうち、文字情報以外の付加情報の物理量と心理量の関係の一つである単音節音声の音韻継続時間について、主観的に等価に聞こえるときの物理的な長さについて、実験的に検討を行った。その結果、音韻継続時間の差異を認識する何らかの能力またはスキルについて個人差が大きいこと、主観的に等価に聞こえる音節の物理的な音韻継続時間は、その音節の有声音の割合に概ね逆比例することなどが判明した。

以上から本章の実験方法によって、異なる音節の主観的な長さを等しくすることが可能であることが示された。この方法は、明瞭性等の試験用の標準音声を作成する場合に適用可能である。

また、単音節単位の編集合成法は今日でも、実用装置における地名、人名などの固有名詞の出力や、福祉目的の装置に十分適用可能であり、本章の実験方法や結果を活用することができる。

第5章

合成音声の音質評価のための 主観評価項目の抽出

5.1 はじめに

第4章では、音声のもつ物理的特性のうち、心理面への直接的影響が推測できる関係のうち、単音節音声の音韻継続長の心理評価実験を行い、異種の単音節間の心理的に等価な長さを求めた。

しかし、音声の質を表現する評価項目（形容詞等）のうち、音声の物理的性質との結びつきが明らかなものはむしろ限られている。音声の質を表す評価項目のうち、音声の複数の物理的特性と複雑に関係すると推測されるものもあり、また、物理量との結びつきが、推測できないものも多い。

音声のもつ物理特性が心理面へどのように影響するかを調べる試みは、声質と物理量の関係³⁶⁾、言い換えれば、音声が伝達する、多重で複雑に関係し合った付加的な情報が感性へ及ぼす影響を解明していくひとつのステップである。

音声の伝達する付加的情報の評価において最終目的である、音声の物理的な諸特性と感性への総合的な影響を求めていくための方法論として、次の二つの手法が考えられる。

ひとつは、分析・合成等の手法によって合成した音声について主観評価を行う方法である⁴⁰⁻⁴⁴⁾。第4章の試みもこの手法のひとつに分類できる。この方法では多くの場合、目的とする物理的特性を精密に制御した合成音声が作成できて、しかも目的としない他の特性を一定に保つことができる。また、心理的な影響もある程度予想できるので、心理実験における評価項目を絞り込むことができる。したがって目的の物理特性と主観評価の関係が精密に求められる。この手法によって、種々の物理的特性を制御して主観評価実験を行い、関係を積

第5章

み重ねていく。

この方法の問題として、それぞれ別々に求められた関係の、感性への影響全体に対する重みづけが難しい点が挙げられる。そのほか、特定の特性のみ変化させるので、結果として実験に用いる音声は多彩ではなく、均質であり、実験条件は一般的とはいえない。また、音声の物理的な特性を変化させたときの感性への影響が実験前からある程度予想できるので、反対に、予想されている以外の影響が見だしにくい。さらに、合成音声のため不自然さが残るので、主観評価において、例えばそう快感や自然さ等に関する関係を調べることも難しい。

もう一つの方法は、自然音声[†]の主観評価を行い、その音声の種々の物理特性を分析して求めて主観評価の結果との関係を調べる方法である。

この方法では、主観評価に自然音声を用いるので、評価結果は自然音声そのものに対する評価であり、刺激音声の選び方に注意すれば、より一般的な結果が得られると期待される。反面、自然音声から安定に、信頼できる物理特性を分析して得ることは非常に難しく、評価項目をあらかじめ絞り込むことも難しい。また、音声の伝える意味、内容、読み方の巧拙などと、声質そのもののうちのどれが感性へ影響しているかを分離していくことも難しい。

ある話者の音声を聞いたときの印象は、ある程度の時間連続して音声を聞いて総合的に形成されると考えられるのに対して、音声を分析して得られる特性の多くは瞬時あるいは短い時間窓内の特性を表していることも留意すべき問題である。比較する物理量が、その話者の音声の全体的な傾向を表す特性となるように、分析の対象となる時間の長さについても考慮する必要がある。

本章と次の第6章では、後者の、自然音声を用いて主観評価を行い、刺激音声の物理特性との関係を求める手法の有効性について確認するために、評価が、話し方、読み方の巧拙、意味や内容によらない、声質に対して行われるように工夫した主観評価実験を行い、評価に用いた刺激音声からいくつかの物理的特性を分析し、主観評価との関係を調べていく。

そのうち、本章では、次のような手順で、人が音声を聞いたとき、声質に対して主に着目する主観的評価項目を抽出する。

まず、アンケートなどによって一般の人から、人の声の質を評価する言葉を集める。集めた言葉を整理して、一般的に使用される頻度の高い言葉を選択して、主観評価実験のための評価項目とする。次に、主観評価実験のための音声刺激を作成する。音声刺激には、一般の人の発声した肉声（自然音声）を録音する。発声された音声に、意味内容や、読み方の巧拙、感情などが加わらないように工夫する。集めた自然音声を刺激として提示しながら、先に選定した評

[†]合成音声に対して、人の肉声を自然音声と呼ぶ。

価項目による主観評価実験を行い、結果を主成分分析によって分析する。その結果として、人が音声を聞くとときに主に着目する評価項目が抽出される。

5.2 実験方法

5.2.1 評価項目の選定

はじめにアンケート調査によって、音声の特徴を主観的に評価する表現を収集した⁶³⁾。次のような回答者への指示を記した用紙を実験者が直接20名の成人に手渡し、回答を依頼した。

“人が話し声の質を表す言葉について、例にならって20以上お答えください。
例) ダミ声、キンキン声、鼻づまりの声、怒った(ような)声”

依頼者全員から回答を回収することができた。

回答者の年齢は22～55歳、20名のうち7名は女性であった。回答数は1名あたり17～96で、のべ939個、種類にして516の表現が集められた。

次に、筆者を含む3人の実験者によって、寄せられた回答から一般性に欠ける表現、および意味に差のない表現を取り除く作業を行った。最終段階では、3人の実験者による合議により選択を行った。これまで、声質そのものの主観評価に用いる評価項目自体、あるいは、そうした主観評価実験の指針も示されていなかったもので、取捨選択の最終的な判断は実験者の主観によらざるを得なかったが、一般的な表現を選ぶように心がけた。この作業によって26対の表現を、音声を評価するための項目として選択した。

表5.1に選択された表現を示す。選択された表現の中には、“暗い”―“明るい”、“澄んだ”―“濁った”などのように、表現が対象をなすものと、“よくとおる”、“つやのある”などのように明確に対立した表現のないものがある。これら対立した表現のないものについては、“～ない”という否定の表現を対にした。実験では-3から+3までの7段階の評定をするので、集計の便宜上、それぞれの対立する表現をプラスの表現とマイナスの表現とに分類した。

回答用紙は図5.1に示すように、1ページを1刺激分として、26対の評価項目を並べた。評価項目の順序はページごとにランダムに並べ替えた。左右の項目はプラスの表現とマイナスの表現を偏りなく配置した。被験者の回答の傾向に左右に偏りがあることが考えられるので、それぞれの評価項目について、出現する順番はランダムに入れ替わるが、プラスとマイナスの表現の左右の関係は常に一定にした。このようにすれば、評価項目ごとに平均値を引くことにより、左右への偏りの影響を排除することができる。

5.2.2 刺激系列の作成

本研究では、合成音声が伝達する文字情報以外の付加的な情報が感性へ及ぼ

表5. 1 実験に使用した評価項目

	マイナスの表現	プラスの表現
1	かすれた	なめらかな
2	低い	高い
3	暗い	明るい
4	弛緩した	緊張した
5	冷たい	暖かい
6	とおらない	よくとおる
7	濁った	澄んだ
8	女性的な	男性的な
9	おだやかな	激しい
10	平凡な	個性の強い
11	やさしい	きびしい
12	やわらかい	かたい
13	うっとおしい	さわやかな
14	つやのない	つやのある
15	にくらしい	かわいい
16	いろっぽくない	いろっぽい
17	鈍い	鋭い
18	甘い	渋い
19	軽い	重い
20	細い	太い
21	かわいた	湿った
22	ききにくい	ききやすい
23	悪い	良い
24	粗野な	繊細な
25	とげとげしい	まろやかな
26	不自然な	自然な

1:	うっとおしい	-- -- -- -- -- --	さわやかな
2:	太い	-- -- -- -- -- --	細い
3:	自然な	-- -- -- -- -- --	不自然な
4:	とげとげしい	-- -- -- -- -- --	まろやかな
5:	男性的な	-- -- -- -- -- --	女性的な
6:	かわいた	-- -- -- -- -- --	湿った
7:	高い	-- -- -- -- -- --	低い
8:	かたい	-- -- -- -- -- --	やわらかい
9:	冷たい	-- -- -- -- -- --	暖かい
10:	個性の強い	-- -- -- -- -- --	平凡な
11:	やさしい	-- -- -- -- -- --	きびしい
12:	つやのある	-- -- -- -- -- --	つやのない
13:	よくとおる	-- -- -- -- -- --	とおらない
14:	暗い	-- -- -- -- -- --	明るい
15:	かすれた	-- -- -- -- -- --	なめらかな
16:	緊張した	-- -- -- -- -- --	弛緩した
17:	軽い	-- -- -- -- -- --	重い
18:	良い	-- -- -- -- -- --	悪い
19:	にくらしい	-- -- -- -- -- --	かわいい
20:	ききやすい	-- -- -- -- -- --	ききにくい
21:	渋い	-- -- -- -- -- --	甘い
22:	濁った	-- -- -- -- -- --	澄んだ
23:	いろっばい	-- -- -- -- -- --	いろっばくない
24:	おだやかな	-- -- -- -- -- --	激しい
25:	鈍い	-- -- -- -- -- --	鋭い
26:	粗野な	-- -- -- -- -- --	繊細な

図 5. 1 回答用紙の例

表5. 2 録音に用いた無意味つづり

系列 #	無意味つづり			
1	けそつんに	つりでびめ	なみさなお	いにしらぬ
2	のりうのう	せふこしは	おでてこい	なのがりさ
3	てくおいそ	ていくまと	みりしがか	うめいふい
4	ちぼてへの	れかうせで	くたしわり	てたぎなど
5	にそでたや	りたじおと	のあでなふ	きまうよし
6	かほいわん	くしてくそ	るらとあに	おきんのま
7	じのはかり	つまいふが	ねあおよう	おいぜねい
8	のでわりし	ぐがさよず	さしんでさ	えしとおか
9	いぎかあの	うもぶにね	うまなるお	まはとのお
10	いききこり	わいせずん	でてたてみ	めきいとい
11	えこうにる	いまさおや	につていで	かさにりへ
12	とらこるあ	とおきじに	あぼきがみ	んあまので
13	りらちみい	のかおぎょく	らおういな	てしおかた
14	じのがてく	うにがまも	ごみわろら	でかおりけ
15	しゅもちいの	そひとるど	んたしこか	がわいただ
16	いまたくん	てりぶなち	とれかにの	てわんまみ
17	にきまよに	れつすこど	たらだのん	はたわうか
18	のにみはう	がいづちび	にんわてい	なはよらや
19	もだらばも	らべしょるよ	りどけきづ	おうぶにて
20	そたがわし	りぼみよな	やなたみな	いまかえに

す影響の究明を目指しているが、音声の主観評価を行う場合、聴覚刺激として単語や文章を朗読したものをを用いると、読み手の抑揚の付け方、巧拙、単語や文章の意味などが結果に及ぼす影響を無視することができない。

声の音質そのものの主観評価項目を抽出するために、意味や読み方の影響が少ない無意味つづりを刺激として用いた。

しかし、全くの乱数によって作成した無意味つづりでは、それぞれの単音節の出現頻度が日本語の日常的な口語文と大きく異なり、不自然になることが懸念される。日本語の合成音声の評価に用いるための評価項目の抽出を目的としているので、この無意味つづりに含まれる単音節の構成比率は、日本語の日常的な口語文と大きく異ならないことが望ましい。

そこで、小学4年生の国語の教科書を用い、乱数表を用いて、ランダムなページと行を選び出し、本文の文章の途中からかなにして20文字分、縦に書きだした。促音は省略し、拗音は1つの単位とした。この作業を繰り返し、20文字×20行の行列を5組作成した。

作成した行列を横に読むことによって、単音節の構成比が日常的な口語文と大きく異なることのない無意味つづりを作成できる。

こうして用意した、全部で100の無意味つづりの中から、筆者を含む実験者3名の合議によって、日本語として意味のあるもの、不自然なものを除外して、最終的に表5. 2に示す20組の無意味つづりを選択した。

5. 2. 3 刺激の作成法

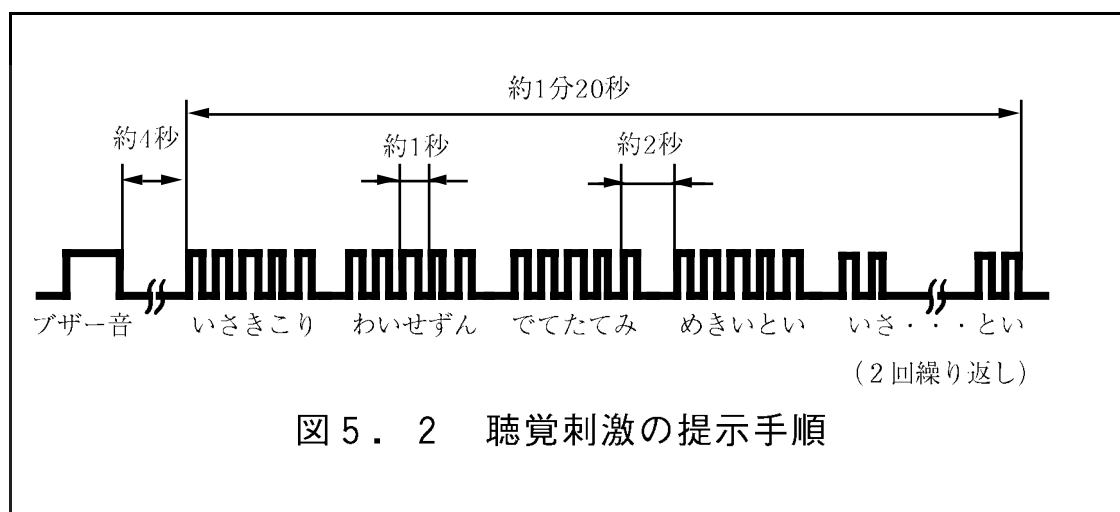
作成した無意味つづりを21名の成人(25~49歳、うち女性が11名)に読ませ、テープレコーダで録音した。

声質の主観評価を行いたいのので、読み手による、速度やイントネーションの違いが最小であることが望ましい。そこで、あらかじめ録音したモデルリーディングをヘッドフォンで各人に聞かせ、その後について読ませた。モデルリーディングでは、1系列の20音節を1音節ごとに約1秒間隔、5音節ごとに約2秒休止をおいて読んでおり、1系列あたり約25秒の継続時間となっている。それぞれの人に20系列の無意味つづりを全部読ませ、テープレコーダで録音した。

録音した21人分、20系列の中から、21人全員が特に不自然な読み方にならなかった系列を2種類選択した(表5. 2の系列#10と#18)。そのうちの1種類(系列#18)を使って、男声刺激が連続10人分現れた後に女声刺激が11人分現れる系列と、男女の刺激が交互に現れる刺激の2種類を作成し、これらの刺激によって被験者3~4名による予備実験を行い、結果について相関行列の計算から始める主成分分析を施した^{60, 61, 64-67}。

予備実験の結果から、実験の手順としては刺激提示の継続時間が1分20秒ほどあれば、刺激提示中に26項目の評定を完了できることがわかった。

第5章



主成分分析の結果からは男声を連続に提示した後に女声を提示する場合と、交互に提示する場合とでは、各主成分に寄与率が高い評価項目の種類やそれぞれの主成分の重み係数、刺激の主成分スコアに著しい変化が現れないことが確認できた。

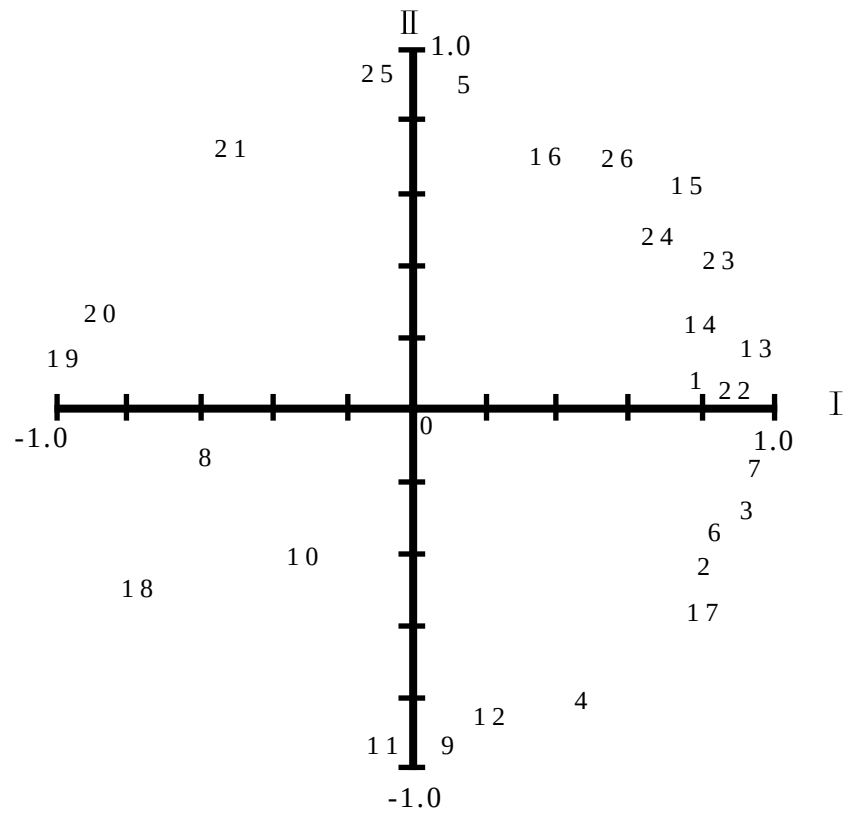
本実験用の刺激として、図 5. 2 に示すように、予備実験で使用しなかった系列#10 の無意味つづりを用いて、継続時間が 1 分 20 秒になるように一人分の音声を 3 回繰り返し、男声 10 人分と女声 11 人分を別々のテープに編集録音した。

5. 2. 4 評価実験の方法

本実験は一度に 4～6 名の被験者で行った。実験には特別な音響対策のとられていない会議室を使用した。実験室内には、被験者の机、その前方 4 m に刺激提示用スピーカー、さらにその前方に実験者の机を配置した。実験者は 1 名

表 5. 3 各成分の全分散に対する寄与率，
累積寄与率および固有値

成分	寄与率 (%)	累積寄与率 (%)	固有値
1	48.5	48.5	12.6
2	33.8	82.3	8.8
3	6.6	88.9	1.7
4	3.1	92.0	0.8
5	2.3	94.3	0.6
6	1.2	95.5	0.3
7	1.0	96.5	0.3



(a) I - II 平面

図 5. 3 4 つの主成分についての各評価項目の
重み係数による空間分布

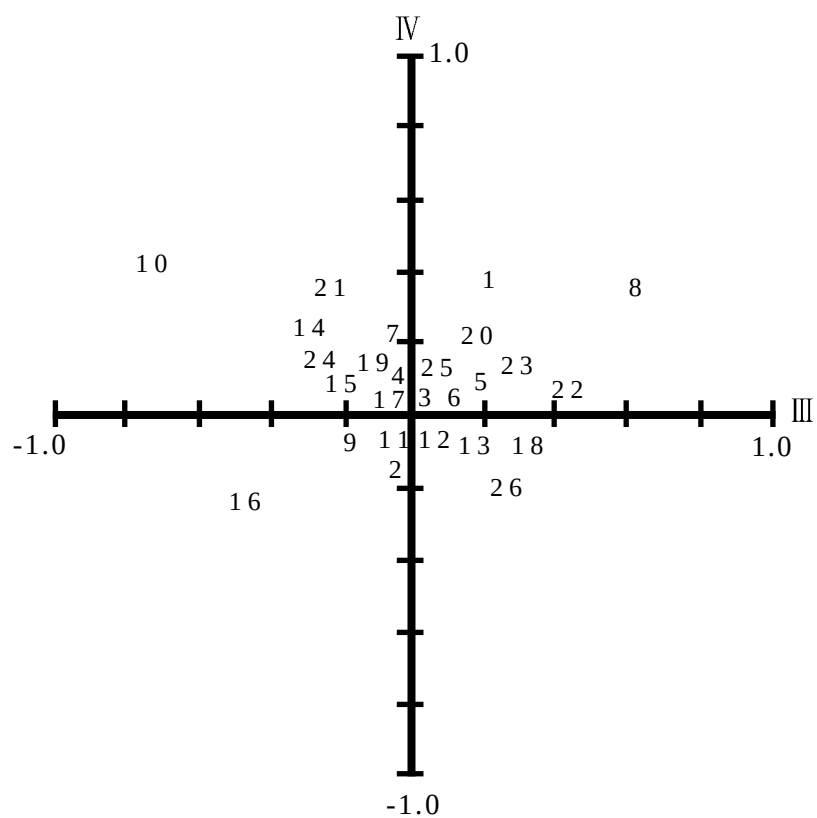
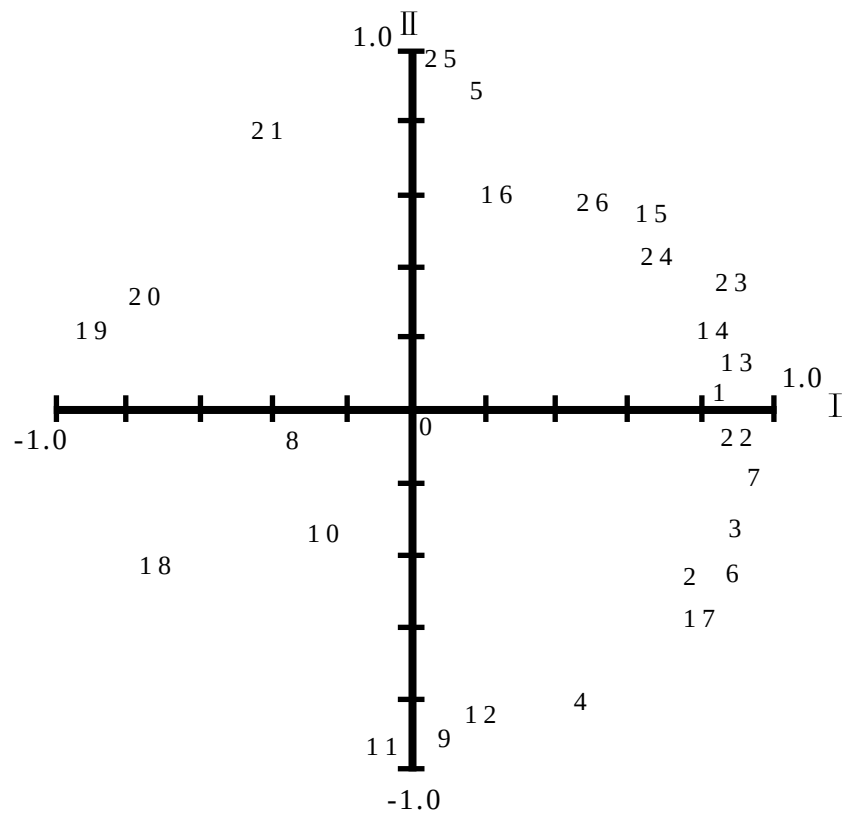


图 5. 3 (b) III—IV平面



(a) I - II 平面

図 5. 4 バリマックス回転後の評価項目の
重み係数による空間分布

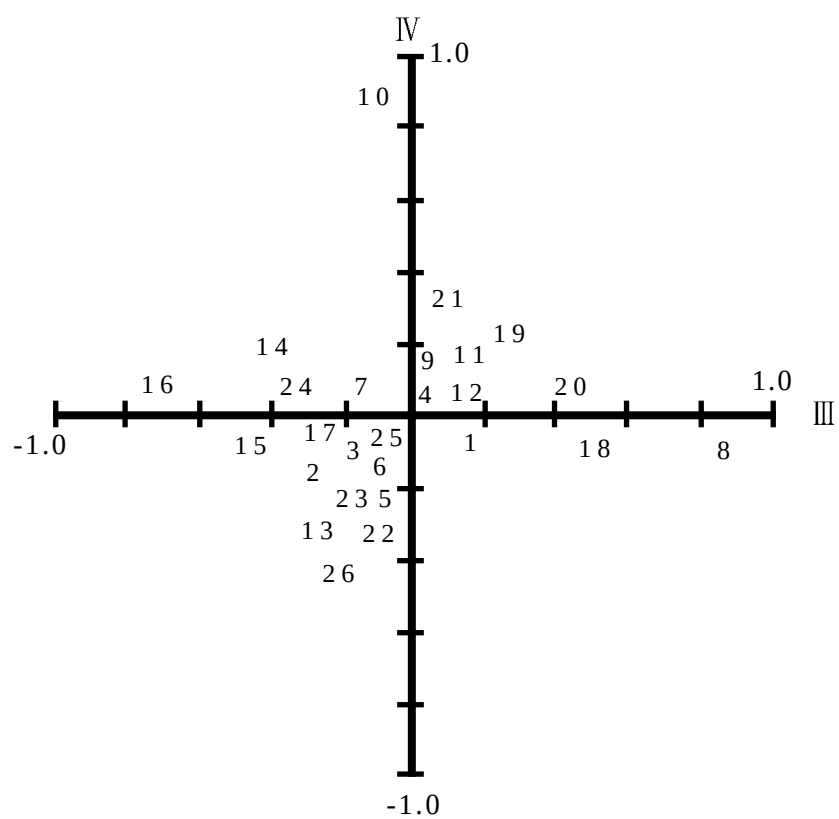
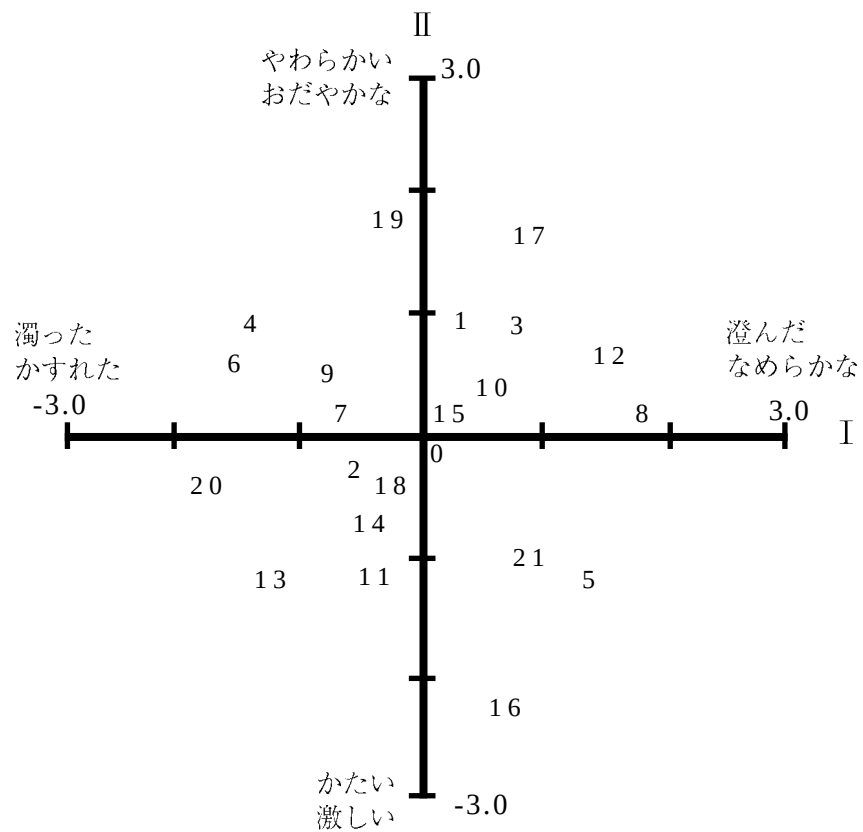


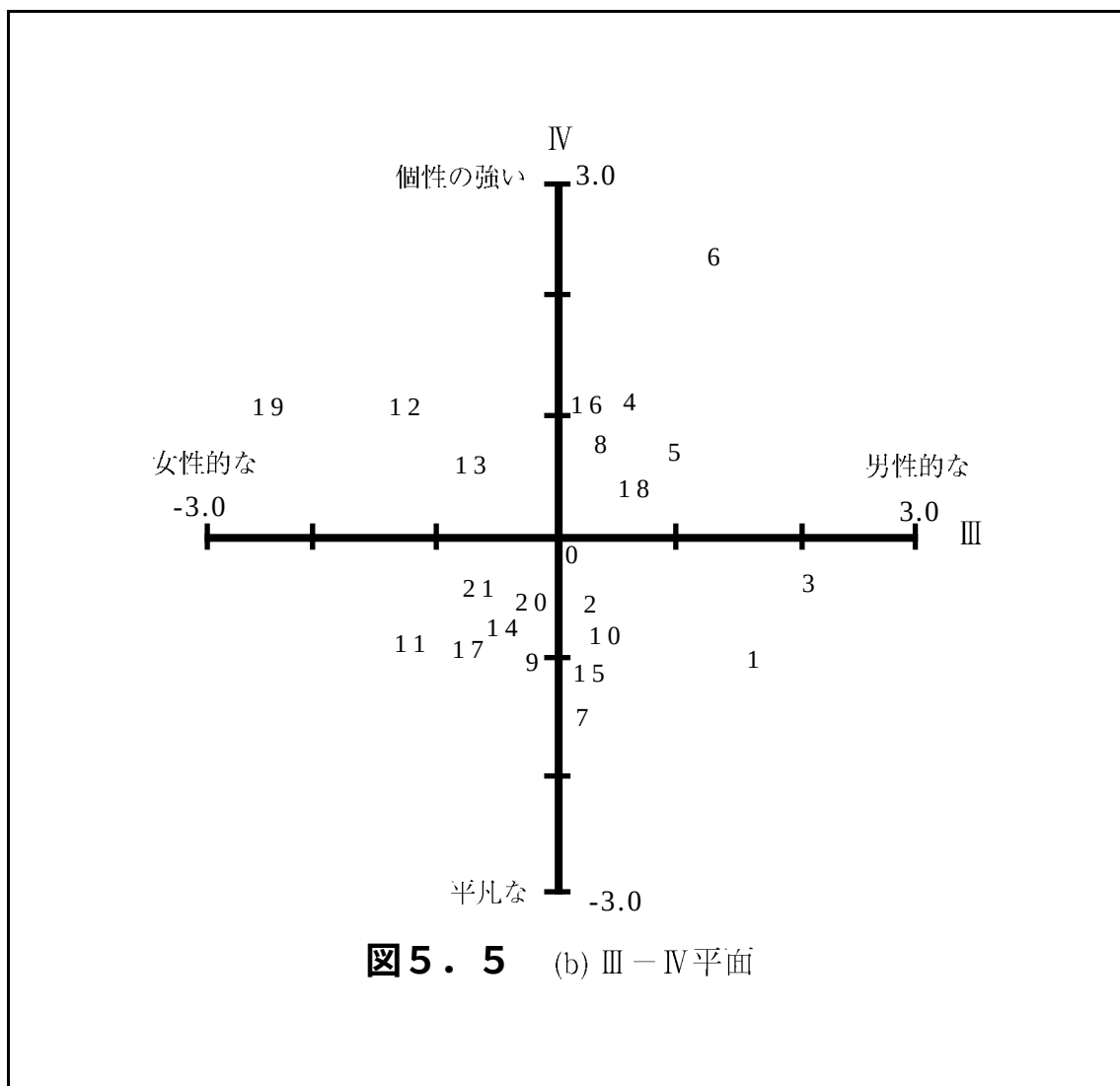
图 5. 4 (b) III—IV平面



(a) I - II 平面

図 5. 5 バリマックス回転後の音声刺激の主成分値の空間分布

(1～10：男性, 11～21：女性)



で、刺激を提示する機器の操作と被験者の回答の終了の確認を行った。

スピーカーから無意味つづり音声を1人分ずつ提示し、その印象を26項目全部について7段階評定させ、回答用紙に筆記させた。

評定は刺激系列の提示の開始とともに開始させ、実験者は目視によって被験者全員の回答（筆記）の終了を確認してから次の刺激の提示に移行した。総数18名（20～48歳，うち女性9名）の被験者について，半数の9名（うち女性7名）は男声刺激が先，残りの9名（うち女性2名）は女声刺激が先という条件とした。

5.2.5 主成分分析

被験者の回答は各話者の評価項目ごとに平均値と標準偏差を求める。この平均値について，主成分分析で解析を行った。

主成分分析は，ある結果と関係すると考えられる多数の変数を，情報の損失をなるべくなくして，少数の変数に集約する手法である。

まず，変数相互間の相関係数を全て求めて相関係数の行列を作る[†]。この行列について，繰り返し法によって固有値，固有ベクトルを求める。得られた最大の固有値に対する固有ベクトルの各要素（重み係数）が第1主成分であり，以下第2，第3，第4主成分となる。

固有値は，得られた軸に対する各サンプルの分散と一致するので，固有値がその主成分によって説明できる大きさを示しているといえる。また，相関係数行列から求めた固有値の総和は変数の数 n と一致し，各固有値を総和で割った値を寄与率と呼ぶ。第1主成分から第 m 主成分までの寄与率の和を累積寄与率と呼ぶ。

5.3 実験結果

求められたそれぞれの主成分の全分散に対する寄与率，累積寄与率および固有値は表5.3に示すとおりである。第1，第2主成分の寄与率がそれぞれ48.5%，33.8%と大きい。固有値は第3主成分までが1より大きくなっている。

図5.3に，表5.1に示す評価項目の主成分の重み係数による空間分布を示す。図に示しているのは評価項目のうちのプラスの表現で，マイナスの表現はプラスの表現と原点対象の点に分布する。

図5.3(a)から第1主成分の軸（第I軸と呼ぶ[‡]，他の主成分についても同様）のプラス側で重み係数が0.8以上と，非常に大きくなっている項目を大きい

[†] 最初に変数間の共分散からなる共分散行列を作る分析する方法もある。得られた固有ベクトルの各要素を各要素の2乗の和で割れば，相関係数行列から始める結果と同じになる。

[‡] 主に図中で刺激の番号と軸の番号の混同を避けるために，本論文では空間の軸の番号にはローマ数字を用いることとした。

第5章

順に拾うと，“澄んだ”，“さわやかな”，“明るい”，“よくとおる”，“ききやすい”，“つやのある”，“高い”，“良い”，“鋭い”，“なめらかな”の10項目である．マイナス側では“重い”，“太い”の2項目の絶対値が大きくなっている．

第Ⅱ軸では，プラス側で“まろやかな”，“暖かい”が，マイナス側では“きびしい”，“激しい”，“かたい”，“緊張した”の4項目の重み係数の絶対値が大きい．

図5．3(b)のⅢ－Ⅳ平面では，第Ⅲ軸では“個性の強い”，“男性的な”，“いろっぽい”の重み係数がそれぞれ-0.72, 0.63, -0.47となっている．第4主成分の寄与率は3.1%，固有値は0.81である．

何番目までの主成分を分析結果として採択するかについて，通常次のような点が判断の基準とされる．

1. 固有値が1よりも大きい
2. 累積寄与率（上位からその主成分までの寄与率の総和）が60%以上

表5．3では，累積寄与率は第2主成分までで82.3%と60%を越えているが，第3主成分の固有値が1.7となっているので，第3主成分までの採択については問題ない．

第4主成分については，第3主成分までの累積寄与率は88.9%となっており，しかも第4主成分の固有値は0.8と1よりも小さくなっているので主成分として採択しないことも十分妥当である．

しかし，図5．3（b）のⅢ－Ⅳ平面における変数（評価項目）の分布を見ると，8“男性的な－女性的な”，10“個性の強い－平凡な”，という二つの異質な評価項目が他の評価項目よりも目立って大きな重み係数となっている．さらに，8と関係が深いと想像される16“いろっぽい－いろっぽくない”は，8とは原点をはさんで反対側にやや小さい値で布置している．以上から，座標空間を回転させることで，8と16が一つの軸をなし，10がその軸と直交する別の軸を構成する可能性が読みとれる．そこで，第4主成分までの分析結果をバリマックス回転[†]にかけた^{60, 61)}．

表5．4にバリマックス回転後の各主成分の寄与率を示す．

図5．4は，回転後の評価項目の各主成分の重み係数による空間分布である．図5．4(a)の第Ⅰ軸のプラス側では，“澄んだ”，“なめらかな”，“ききやすい”，“さわやかな”，“よくとおる”，“明るい”，“良い”，“つやのある”などの項目の重み係数が0.8以上となっている．マイナス側では，回転前と同様，“重い”，“太い”の2項目の重み係数の絶対値が大きい．

[†] 主成分分析と同じ多変量解析の一つである因子分析で主に用いられる手法で，各主成分の重み係数の2乗の分散の総和が最大となるように軸を回転させる．主成分分析への適用例もある．

表 5. 4 回転後の各主成分の全分散に対する寄与率および累積寄与率

成分	寄与率 (%)	累積寄与率 (%)
1	43.5	43.5
2	32.7	76.2
3	10.0	86.2
4	5.8	92.0

第Ⅱ軸では、プラス側で“ まろやかな ”，“ 暖かい ”の，マイナス側では“ きびしい ”，“ 激しい ”，“ かたい ”，“ 緊張した ”の4項目の重み係数の絶対値が大きい。

第Ⅰ軸，第Ⅱ軸とも，回転前と重み係数の大きな項目の分布に変化はほとんどない。以上から第Ⅰ軸は“ 透明感 ”，“ なめらかさ ”の評価軸，第Ⅱ軸は声の“ かたさ ”，“ 激しさ ”の評価軸と解釈できる。

図 5. 4 (b)のⅢ－Ⅳ平面では，第Ⅲ軸にはプラス側で“ 男性的な ”，マイナス側で“ いろっぽい ”の重み係数の絶対値が大きく，第Ⅳ軸ではプラス側で“ 個性の強い ”の重み係数が大きい。この図と図 5. 3 (b)を比較すると，バリマックス回転によりⅢ－Ⅳ平面が負の方向に 45° 回転したことがわかる。図から第Ⅲ軸は“ 男性的な－女性的な ”の軸，第Ⅳ軸は“ 個性の強さ ”の軸と解釈される。

図 5. 5 はバリマックス回転後の 21 人の音声刺激の主成分スコアによる分布である。図中の数字は刺激の番号で，1～10 は男声刺激，11～21 は女性刺激である。この図から分布が特徴的とみられる刺激 6,8,16 および 19 の7段階評定の平均値と標準偏差を表 5. 5 に示す。

5. 4 考察

表 5. 4 では，図 5. 5 (a)の第Ⅰ軸で正側に偏っている刺激 8 は，“ 澄んだ ”，“ ききやすい ”，“ よくとおる ”，“ 明るい ”などの，そう快さや明るさを表す項目で，評定値の平均が 1 以上となっている。第Ⅱ軸で大きく負側に偏っている刺激 16 は，評定値に絶対値の大きなものが目立つが，なかでも“ 激しい ”，“ かたい ”がそれぞれ 1.37 と，比較的大きな値をとっている。

図 5. 5 (b)の第Ⅲ軸で負に大きく偏っている刺激 19 は，“ 男性的な ”で，7 段階評定で-2.16 と極端な値をとっており，“ いろっぽい ”も大きな評定値となっ

第5章

表5.5 特徴的な音声刺激の7段階評定（－3～＋3）の平均値と標準偏差

評価項目	平均値（標準偏差）			
	刺激番号6	刺激番号8	刺激番号16	刺激番号19
1 なめらかな	-0.68 (1.17)	0.74 (0.71)	0.16 (0.59)	-0.32 (1.62)
2 高い	-1.89 (1.02)	1.16 (1.39)	1.63 (1.09)	-0.68 (1.56)
3 明るい	-2.37 (0.93)	1.05 (1.00)	0.84 (0.93)	-0.95 (1.32)
4 緊張した	-0.53 (1.57)	0.79 (1.00)	1.37 (0.98)	-0.79 (1.24)
5 暖かい	-0.32 (1.13)	0.00 (1.12)	-1.05 (1.00)	0.89 (1.29)
6 よくとおる	-2.05 (1.05)	1.37 (0.93)	1.53 (1.04)	-0.68 (1.52)
7 澄んだ	-1.58 (0.99)	1.16 (0.87)	0.68 (1.22)	-0.05 (1.39)
8 男性的な	1.11 (1.17)	-0.32 (1.22)	-0.47 (0.94)	-2.16 (0.81)
9 激しい	-0.84 (1.18)	-0.37 (1.04)	1.26 (0.96)	-1.32 (0.92)
10 個性の強い	1.68 (0.92)	0.68 (1.03)	1.42 (1.14)	0.84 (1.04)
11 きびしい	-0.26 (1.07)	-0.21 (1.06)	1.37 (0.93)	-1.58 (0.82)
12 かたい	-0.53 (1.39)	0.53 (1.39)	1.37 (1.04)	-1.68 (0.80)
13 さわやかな	-2.11 (1.07)	0.16 (1.39)	-0.53 (1.60)	-0.53 (1.14)
14 つやのある	-1.00 (1.52)	0.47 (1.04)	-0.16 (1.60)	0.42 (1.66)
15 かわいい	-1.16 (1.04)	0.58 (1.18)	-1.00 (1.08)	1.05 (0.69)
16 いろっぼい	-1.37 (1.38)	-0.79 (1.40)	-1.63 (1.04)	1.42 (1.18)
17 鋭い	-1.79 (1.24)	0.89 (0.91)	1.21 (1.40)	-0.89 (1.25)
18 渋い	0.89 (1.17)	-0.47 (0.94)	0.37 (0.58)	-1.00 (1.30)
19 重い	2.16 (0.87)	-1.05 (0.89)	-0.26 (1.16)	0.63 (1.35)
20 太い	1.47 (1.27)	-1.00 (1.12)	-0.74 (1.37)	0.00 (1.21)
21 湿った	1.05 (1.82)	-0.05 (1.15)	-0.84 (0.99)	1.00 (1.45)
22 ききやすい	-2.05 (1.00)	1.05 (1.10)	0.37 (1.31)	-0.32 (1.22)
23 良い	-1.26 (1.25)	0.58 (0.94)	-0.53 (1.43)	0.21 (1.32)
24 繊細な	-0.84 (0.87)	1.16 (0.93)	-0.79 (1.10)	1.00 (0.92)
25 まろやかな	0.74 (1.02)	0.05 (1.19)	-1.42 (0.88)	1.26 (0.78)
26 自然な	-1.32 (1.26)	0.16 (1.09)	-1.32 (1.08)	0.42 (0.88)

ている。

第IV軸で正の大きな値をとっている刺激 6 は、表 5. 4 で多数の項目において評定値の絶対値が大きくなっているなかで、“個性の強い”の評定値も 1.68 と大きな値をとっている。

以上のように、図 5 で特徴的な分布をしている刺激の評定結果は、4 つの評定軸がそれぞれ“透明感”や“なめらかさ”，“かたさ”や“激しさ”，“男性的な”および“個性の強さ”を示す評定軸であることを裏づけている。

図 5. 5 (b) で、男声の刺激と女声の刺激の第III軸上の主成分スコアの平均値について、t-検定を行った^{48,49)}。男声、女声それぞれの主成分スコアの平均値と標準偏差は表 5. 6 に示すとおりであり、 $t=3.72$ が得られる。男女のサンプル数の和から 2 を引いた自由度 19 について t 分布表から 1% のときの両側検定の t の値は 2.86 が得られるので、有意水準 1% で刺激音声の第 3 主成分の主成分スコアは男声と女声の平均値に差があると認められた。このことも第III軸が“男性的な—女性的な”の軸であることを裏づけている。

問題点としては、第 1 主成分に対する寄与率が 0. 5 以上、つまり主成分への重み係数が 0. 7 以上のものをあげると“澄んだ”，“なめらかな”，“ききやすい”，“さわやかな”，“よくとおる”，“明るい”，“良い”，“軽い”，“つやのある”，“鋭い”，“細い”，“高い”の 12 項目にも及び、全項目の半数近くなる。

刺激の数が 21 と評価項目の数 26 よりも少なく、刺激の採取にあたって特別な配慮をしなかったので、刺激の音質があまり多彩とはいえず、本来、相関の低い、いくつかの項目を分離できなかった可能性もある。

刺激の種類や数と同様に、分析結果に影響を及ぼす条件として評価項目の種類、数が挙げられる。評価項目収集のためのアンケートは、対象が比較的広い年齢の範囲に及び、寄せられた回答も多いので、特に問題はない。

反面、得られた評価項目の選択は実験者が合議制で行ったが、最終的な判断

表 5. 6 第 3 主成分の主成分スコアの男女別平均値，標準偏差と t 検定の結果（両側検定）

	男声		女声		t 検定の結果	
	平均	標準偏差	平均	標準偏差	t	p
第 3 主成分“男性的な—女性的な”の主成分スコア	0.70	0.72	-0.64	0.84	3.71	p<0.01

第5章

は主観に頼らざるを得なかったことから、その内容が実験結果に及ぼす影響は少なくないと考えられる。

従来行われてきた汎用的な音声合成システムの研究・開発では、明瞭性、了解度の次の問題として、人間に近い、より自然な音声の合成を目標としている。“人の声に近い”というのではあまりに漠然としていて、適切で一般的な評価の方法を定めにくい。そうしたなかで、今回の実験結果で“透明感”や“やわらかさ”などといった具体的な評価指標を明らかにできたことは有意義と考えられる。

しかし、実際に合成音声の評価するために信頼性、再現性の高い標準的な評価指標を得るためには、実験に用いる評価項目の種類、数を考慮し、被験者数を増やして実験を繰り返す必要がある。

5.5 章のまとめ

合成音声の感性への総合的な影響を調べるために、人が音声を聞いたときに主に注目する、声質の評価項目の抽出を試みた。音声の持つ、意味内容や話者の感情、読み方の巧拙などを排除した、声質そのものの評価になるように配慮しながら、刺激音声、評価項目を作成、選定して、これらによって主観的评价実験を行なった。その結果について主成分分析を行い、透明感、かたさ、男性的一女性的、個性の強さを表す合計4つの評価軸を主成分として抽出した。

抽出された評価項目は、各種合成音声の感性への影響を評価する際の指標として用いることができる⁶⁸⁾。

次章では、本章のような手順で主観評価実験によって得られる評価（心理量）と、その評価に用いた刺激音声の物理量との関係を解明する試みについて述べる。

第6章

音声の物理特性と心理的評価量の関係解明の試み

6.1 はじめに

前章では、音声の持つ物理的な特性と心理的な評価量の関係を求める前段階として、自然音声を用いて主観評価実験を行い、人が自然音声を聞いたときに、主に着目する心理的な評価項目を抽出した。本章では、次の段階として音声のどの物理量（物理パラメータ、統計量）が、これらの心理評価にどの程度影響を及ぼすか、あるいは逆にどの心理評価はどのような物理量によって形成されるのかといった、音声の物理量と心理評価との関係について調べていく。

音声の物理量としては、周波数スペクトル、フォルマント、基本周波数などの周波数領域の特性、音圧およびその最大値と最小値の比など強度に関する特性などが知られている^{2,3)}。

これらの物理量と、前章で得られた主成分スコアのような心理量を比較する場合問題となるのは、物理量のほとんどは、数十～数百 ms という狭い時間窓内の波形を分析して得られる、人の感覚からすると瞬時的な特性と言えるのに対して、音声から人が感じとる印象は、瞬時的な波形の特性から文章までのそれぞれの特徴までを総合して形成されるという点である。

そこで、心理量と比較する際には、瞬時的である物理量に何らかの統計処理を施して、その話者の音声全体の特性として得る工夫が必要となる。

本章では、前章の心理実験に用いた音声について、物理量として、ピッチ、発話のテンポのばらつき等の統計量を求め、主成分分析の結果と比較、検討を行い、音声の主観評価と物理量との関係を調べる。

音声を周波数領域、振幅、時間領域に関するいくつかの特性について分析し

第6章

て、主成分分析の結果と単純な相関係数を調べ、その中で特に検討が必要と認められたいくつかの物理量について心理量との関係をさらに詳しく検討する。

分析、検討結果から、自然音声进行分析して主観評価の結果との関係を調べることにより主観評価と物理量の関係を解明していくという方法論の有効性についても合わせて吟味する。

6.2 予備的分析

6.2.1 物理量

前章の主観評価実験で被験者に提示した音声刺激について、物理量を網羅的に分析して求め、それぞれについて同じ前章の主成分分析の結果得られた主成分スコアとの相関を調べる。

文献等によれば、音声波形の物理量として周波数領域、振幅、時間領域に関する各種特性がある^{2,3)}。そのうち、周波数領域では基本周波数、フォルマント、周波数スペクトルが挙げられる。振幅特性については、振幅の他、音声短区間平均パワー、音節パワー[†]などの指標が用いられることもある。

時間的な特性については、言語により特徴が異なることもあって、一般的な指標はない。日本語の場合は、単音節単位の時間長の制御が重要と考えられているので⁵¹⁻⁵⁵⁾、単音節間の時間間隔、単音節の音韻継続時間、各単音節の立ち上がりの急峻さなどに着目して、物理量の抽出を試みた。

得られる物理量の多くは、数十～数百 ms という狭い時間窓内の波形を分析して得られる瞬時的な特性なので、総合的な特性とするためには、分析結果について全体の特性を反映するような統計処理を施す必要がある。

以下、前章で得られた主成分スコアとの相関を調べるために網羅的に求めた物理量について述べる。

(1) ピッチおよびフォルマント

主観評価実験で提示した音声を、カットオフ周波数 4kHz のローパスフィルタにかけた後、サンプル周波数 10kHz, 12bit で A/D 変換して、各話者ごとに計算機に取り込んだ。この音声データにハミング窓⁶⁹⁾をかけたのち、予測次数 21 次として PARCOR 分析を施した⁹⁾。この PARCOR 分析から得られる残差信号からピッチを求めた。

ここで、カットオフ周波数 4kHz, サンプル周波数 10kHz としたのは、低次のフォルマントと、残差信号からのピッチの抽出を目的としたからである。

[†] 10ms の間に放出される音声のパワーの平均値は、音声短区間平均パワーと呼ばれることがあり³⁾、本論文では便宜上この用語を使用する。同様に、音節内での音声短区間平均パワーの最大値は音節パワーと呼ばれることがあり、本論文ではこの用語も用いる。

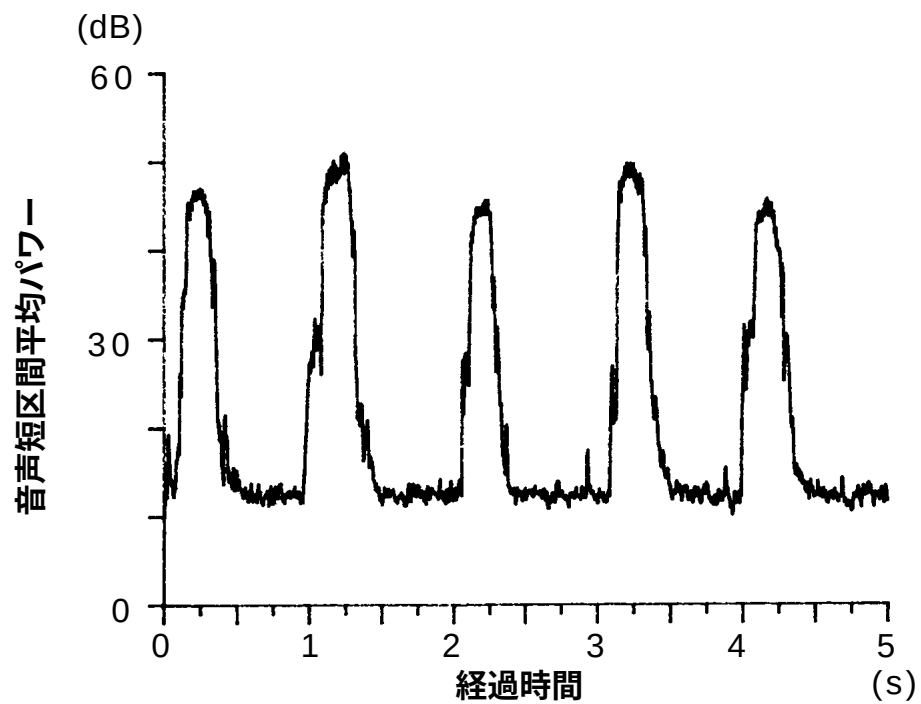


図 6. 1 話者 1 の最初の 5 音節の音声短区間平均パワー（10ms内に放出されるパワーの平均値）の時間変化

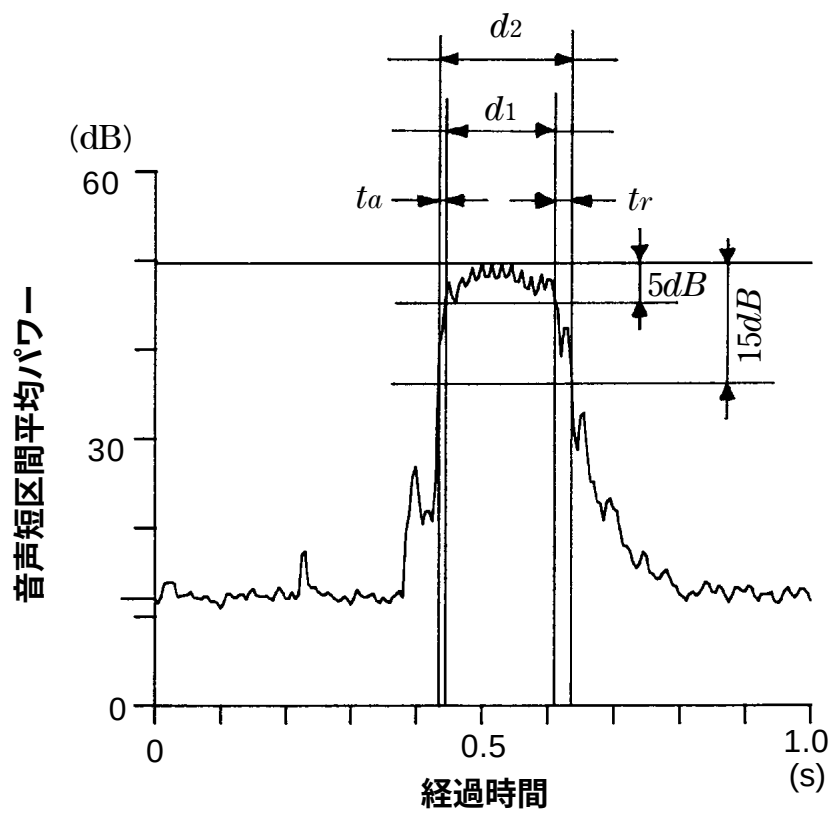


図 6. 2 単音節の立ち上がりと立ち下りの急峻さの計測

各話者毎に発声した 20 音節のピッチの平均値を求め、このときの標準偏差をピッチのばらつきの指標とした。

フォルマントは、PARCOR 分析により推定されたスペクトルから求めた³⁻⁶⁾。全 20 音節の中から、5 母音と“ン”を含むように、“コ”、“セ”、“ズ”、“ン”、“タ”、“ミ”の 6 音節を選んで、PARCOR 分析により推定されたスペクトルの時間変化を作図して、その図から第 1、第 2 フォルマント周波数をそれぞれ読み取り、話者ごとの平均値を求めた。

(2) スペクトルの積算値

スペクトルは音声波形の重要な物理的特性の 1 つであるが、時間窓内の特性であり、心理的特性との関係を考えるうえでその時間窓は狭く、瞬時的といえる。そこで、発声している時間全体の特性を反映されるため、時間窓のスペクトルを、各話者が発声している全時間について周波数毎に積算し、広義の長時間スペクトルとして求めることにした。

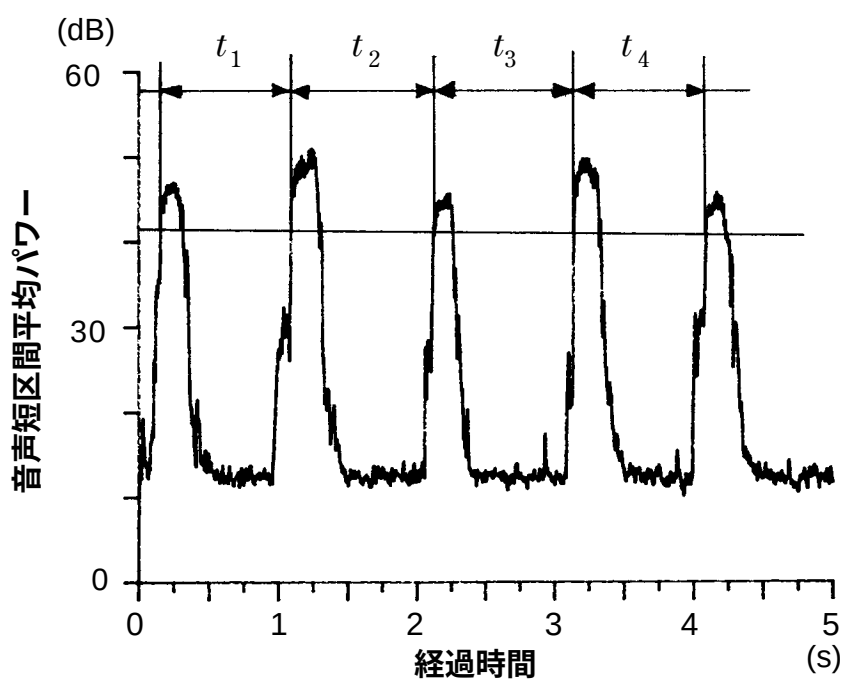
PARCOR 分析ではスペクトルからの低次フォルマントの推定と、残差からのピッチの抽出を目的としていたので帯域を 4kHz としたが、この分析では可聴周波数の影響を広く探ることを目的として、ローパスフィルタのカットオフ周波数を 8kHz とし、サンプル周波数 20kHz 12bit で A/D 変換した音声データを用いた。それぞれの話者について、最初の音節の立ち上がりの直前から最後の音節の立ち下がりの直後までを、25.6ms (512 サンプル) の時間窓にかけて F F T を施した。直流成分～10kHz の 257 点の周波数それぞれについてパワーを求めるため、周波数毎に絶対値の 2 乗をとって積算した（ただし、F F T の前処理で直流成分を除去するので、F F T の結果に直流成分は現れない）。窓の移動周期は窓長と同じ 25.6ms (512 サンプル) とし、発声している全区間について積算した。録音状態等により、積算されるパワーの絶対値にはばらつきが出るので、話者ごとにこのスペクトル積算値の全周波数の積算値の和で各周波数成分を割って規格化した。

(3) 音節パワー

第 5 章の実験では、単音節単位で刺激音声を発声していることから、音声の振幅、強さ、パワーの指標として用いられる単位、測度等のうち、音節ごとの特性を代表する指標である音節パワー[†]を求めることにした。

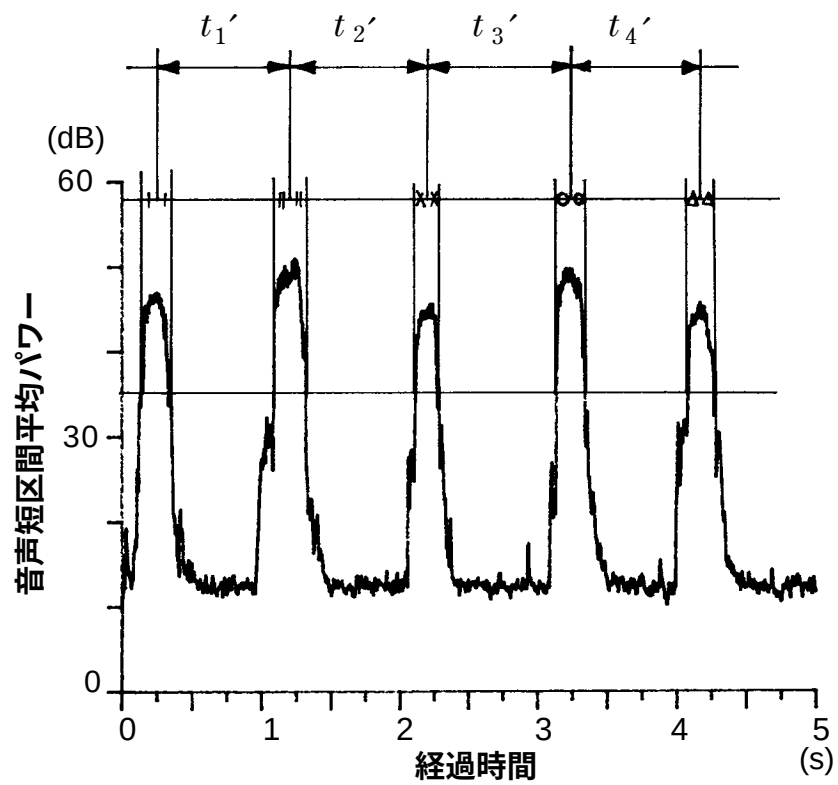
具体的には、次のような手順で話者ごとに音節パワーを求めた。(2) のスペクトルの積算値を求めたときと同様、20kHz でサンプルした各話者の音声データについて、10ms (200 サンプル) の時間窓（方形窓）をかけ、振幅の 2 乗の値の時間窓内における総和を計算し、窓を 10ms ずつ移動することによって、図 6. 1 に示すように、音声短区間平均パワーを時間変化として求めた。

[†] p.90 脚注参照。



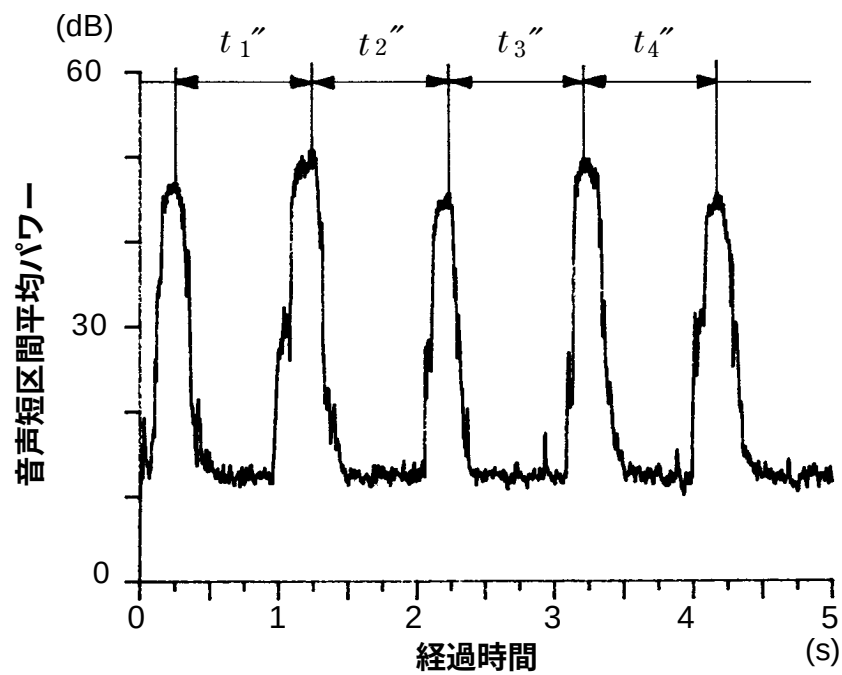
(a) 各単音節の立ち上がりの間隔

図 6. 3 発話テンポのばらつきの測定方法



(b) 各単音節の中間点の間隔

図 6. 3 発話テンポのばらつきの測定方法



(c) 各単音節の音声短区間平均パワー最大の点の間隔

図6. 3 発話テンポのばらつきの測定方法

次に音節ごとに音声短区間平均パワーの最大値を抽出し、音節パワーを求め、話者ごとに音節パワーの平均値と標準偏差を求めた。音節パワーの平均値は録音条件によって話者間に差があるので、標準偏差を音節パワーで割って規格化した値を音節パワーのばらつきの指標とした。この指標は、それぞれの話者の音節ごとの強弱の付け方の大小を表しているといえる。

(4) 立ち上がり、立ち下がり急峻さと音韻継続時間

単音節の振幅に関する特性として、それぞれの単音節の立ち上がり、立ち下がり急峻さ（アタック、リリースの強弱）についても心理面への影響が予想される。そこで、各話者の全音節について、立ち上がり、立ち下がり急峻さを求めた。急峻さの指標としては、図6. 2に示すように、音節ごとに、音声短区間平均パワーが下の基準値から上の基準値まで立ち上がるのに要する時間 t_a を立上りの急峻さとし、反対に上の基準値から下の基準値まで立ち下がるのに要する時間 t_r を立ち下がり急峻さとし、それぞれ、話者ごとに平均値を求めた。

試行錯誤の結果、上下の基準値をそれぞれ、音節パワーの -15 dB （約 $1/30$ ）と -5 dB （約 $1/3$ ）にすると、主観評価実験に用いた全ての話者の全音節について立ち上がりと立ち下がりを求めることができた。音節ごとの特性を計算するため、音節ごとの音節パワーを基準にしている。

また、 -5 dB 以上の継続時間 d_1 と -15 dB 以上の継続時間 d_2 についても単音節の音韻継続時間として、話者ごとに平均値を求めた。

(5) 発話のテンポのばらつき

発話のタイミングの測定には、3種類の方法を用いた。いずれも、音節パワー、立ち上がり立ち下がり急峻さを求めたときと同じ、音声短区間平均パワーから求めている。発話のテンポのばらつき（規則性）は、話者の発声した全部の音節によって形成されるので、立ち上がりの基準となる値は、立ち上がり、立ち下がり急峻さを求めたときとは異なり、各話者の全音節の中の最大の音節パワーを用いた。

テンポのばらつきを求める第1の方法は、図6. 3(a)に示すように、各音節間の立上り同士の間隔で、最初の音節から続く音節の間隔 $t_1, t_2, t_3, \dots$ を計測した。立ち上がりの基準となる閾値は、最大の音節パワーから -10 dB （ $1/10$ ）とした。

2番目の方法では、図6. 3(b)に示すように、音節の立ち上がり立ち下がりの中央の間隔 $t'_1, t'_2, t'_3, \dots$ を計測した。立ち上がり立ち下がりの閾値は、最大の音節パワーから -15 dB （約 $1/30$ ）とした。立ち上がり間隔の計測と同じ -10 dB とすると、音節の途中で音声短区間平均パワーが低減して、正しく立ち下がりが検出できない音節が多く出るという理由による。

3番目の計測方法では、図6. 3(c)に示すように、各音節の音声短区間平均

第6章

パワーが最大となる点（音節パワーが検出される点）同士の間隔 $t_1'', t_2'', t_3'' \dots$ を計測した。

刺激音声は5音節ずつ区切って発声されているので、第5音節と第6音節、第10音節と第11音節、第15音節と第16音節の間隔は除いて、それぞれの間隔の標準偏差を求めて、発話のテンポのばらつきの指標とした。

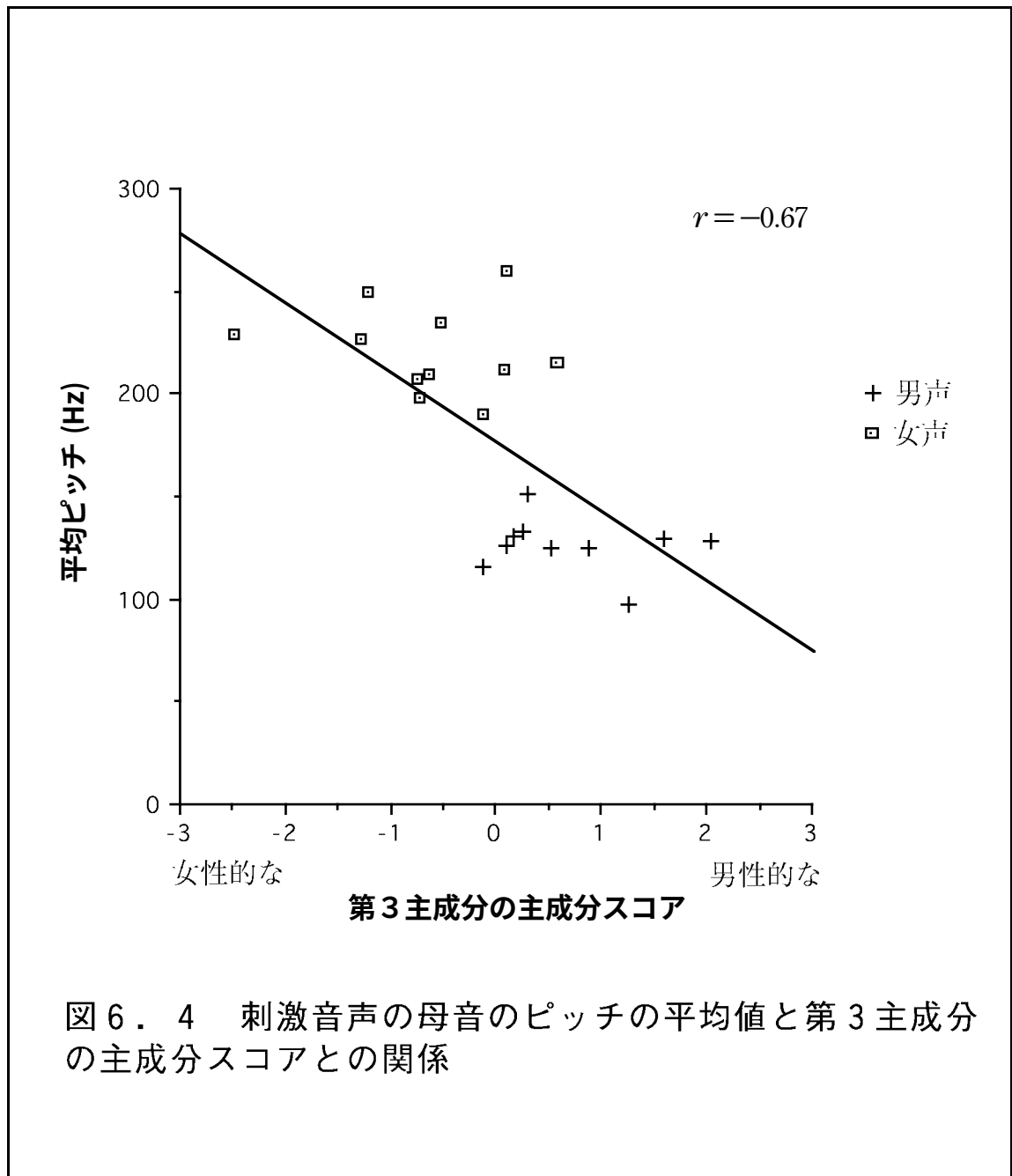
6.2.2 予備的分析の結果および検討

求めた物理量について、前章の主成分スコアとの相関係数を求めた。物理量と主成分との相関係数の絶対値が、原則として0.5以上となった関係について、物理量と主成分に何らかの関係がある可能性が示唆されたとして、表6.1にまとめた。スペクトルの積算値については、相関係数の絶対値が0.6以上の関係が複数認められたので、絶対値が0.6以上のものを掲げた（周波数を表示）。また、主成分スコアと絶対値で0.5以上の相関係数が得られなかった関係については、0.4以上の関係も掲げた。

表6.1 予備的分析の結果，主成分と関係のある可能性が示された物理量

†印の用語は p.90 の脚注参照

主観評価結果	物理量	相関係数
第1主成分 (澄んだ) (なめらかな)	スペクトルの積算値(3867Hz)	0.62
	音節短区間平均パワー†最大時の	-0.53
	間隔の標準偏差	
第2主成分 (やわらかい) (おだやかな)	第1フォルマントの平均値	-0.59
	スペクトルの積算値	0.67
	(820Hz, 977Hz, 1563Hz, 2422Hz)	
	音韻継続時間の平均値(-5dB)	-0.46
	音節の立上りの間隔の標準偏差	-0.53
	音節の中央の間隔の標準偏差	-0.51
第3主成分 (男性的な)	スペクトルの積算値	0.66
	(234Hz, 9101Hz)	
	ピッチの平均値	-0.67
	第2フォルマントの平均値	-0.57
	音節パワー†の標準偏差	0.54
第4主成分 (平凡な)	なし	



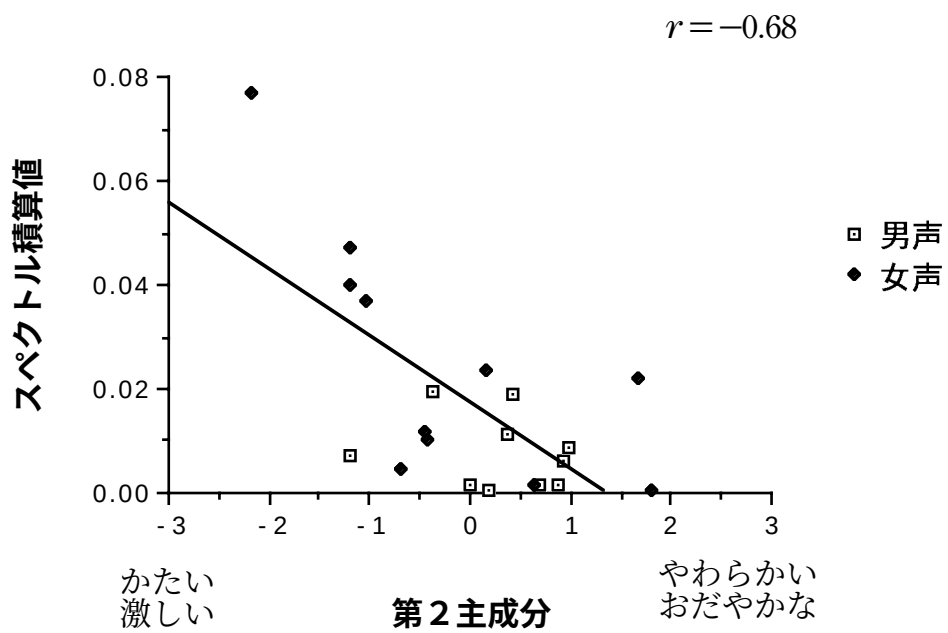


図6.5 820Hzのスペクトル積算値と第2主成分の関係

(1) ピッチおよびフォルマント

ピッチと第2フォルマントの平均値は第3主成分の主成分スコアとそれぞれ -0.67, -0.57 という負の相関が認められた。この結果をそのまま受け入れると、ピッチと第2フォルマントの周波数が低ければ、より男性的、高ければより女性的に感じられるということになる。図6. 4に、第3主成分の主成分スコアとピッチの平均値の関係を示す。図中の直線は最小2乗法により求めた回帰直線である。

周波数が低ければ、より男性的、高ければより女性的にという解釈自体に大きな矛盾はないが、図6. 4に示すピッチと第3主成分の主成分スコアを見ると、男声と女声の分布が分離しており、しかも男女それぞれの分布は、回帰直線の方ではなく、左右の方向に広がって分布している。

第2フォルマントについても同様に男声、女声が分離して分布をしており、これら分布からピッチ、第2フォルマントと第3主成分“男性的な—女性的な”の関係を結論づけることできない。

一方、同じ周波数領域の変数でありながら第1フォルマントについては、第2主成分“やわらかい（おだやかな）—かたい（激しい）”との間に、相関係数が -0.59 という負の相関が認められた。

ピッチ、フォルマントは音声の知覚において特に重要とされており、分析や合成においても重要なパラメータである。そこで、第3主成分の主成分スコアと負相関の認められたうちのピッチ、および異なる主成分である第2主成分と負の相関の認められた第1フォルマントの2者についてさらに検討を行う。

(2) スペクトル積算値

スペクトル積算値については、複数の周波数のパワーと第1から第3主成分の主成分スコアとの間で相関係数の絶対値が0.6以上となっている。図6. 5にもっとも相関係数の絶対値が大きくなった820Hzのスペクトル積算値と、第2主成分との関係を示している。左上に女声の特異なデータがあり、大きな相関係数の絶対値は、このデータによるものと考えられる。他の刺激の分布から相関関係は読みとりにくい。

他の周波数についても同様に1つ2つの特異なデータがあり、これらにより大きな絶対値の相関係数が得られたと見られる。これらの特異データは、ピッチによって生じる周波数スペクトル上のリップルに起因すると考えらる。

人が音声を知覚している際、周波数領域については、ピッチやフォルマントを認識しているという考え方が一般的であり、ある特定の絶対的な周波数の振幅値と心理的な影響を結びつける説明は難しい。個別にみても例えば820Hzという特定のパワーだけが第2主成分の主成分スコアと強い相関があるのは不自然であり、また、第1から第3主成分の主成分スコアとそれぞれ強い相関が認められたいくつかの周波数は、離散していて関連性は見いだせない。

第6章

以上の理由からそれぞれの周波数成分と主観評価の関係を認めるのは適当でない⁶²⁾。

(3) 音節パワー，立ち上がり，立ち下がり急峻さ及び音韻継続時間

これらのパラメータについては，いずれの主成分スコアとも顕著な相関は認められなかった。

(4) 発話のテンポのばらつき

発話のテンポについては，音声短区間平均パワーが最大となる点の間隔の標準偏差は，第1主成分の主成分スコアとの間に-0.53という負の相関が認められた。また，音節の立ち上がりの間隔の標準偏差，音節の中央の間隔の標準偏差はともに第2主成分の主成分スコアと負の相関が認められ，相関係数はそれぞれ-0.53, -0.51であった。同じテンポのばらつきについて，その基準の取り方を変えることで違う主成分スコアと相関が認められたのは興味深い。

6.2.3 予備的分析のまとめ

以上のように，分析を試みた物理量のうちいくつかのものについて主成分スコアとの間に比較的大きい相関が認められた。

主成分スコアと相関が認められた物理量のうち，ピッチ，第2フォルマントは，第3主成分の主成分スコアとの分布で男女のデータが分離して分布していて，このために相関係数が大きくなったもので，主成分スコアとの関係を結論づけることはできない。

ピッチはその心理的な効果も明確なので，ピッチと主成分スコアとの関係をさらに詳しく調べることにする。第1フォルマントは同じ周波数領域のパラメータであるのにピッチとは違う第2主成分，“やわらかい（おだやかな）—かたい（激しい）”と相関が認められたので，ピッチと比較しながら検討を行う。

発話のテンポについては，テンポの基準を変えることで違う主成分スコアと相関が認められた。さらに詳しく検討を行う。

6.3 ピッチと発話テンポのばらつきの心理面への影響

予備的な分析についての検討の結果，主成分スコアと大きな相関係数の認められた，ピッチ，第1フォルマントおよび発話テンポのばらつきについて，主成分スコアとの関係をさらに詳しく調べていく。

6.3.1 男声，女声サンプル間での差の検定

ピッチと第1フォルマント，第3主成分“男性的な—女性的な”の主成分スコアについて，それぞれ男声と女声のサンプルの間で分布に差があるか，t検定によって調べた（両側検定[†]）^{48,49)}。

[†]以降，特に断りのない場合は両側検定。

6.3.2 “高いー低い”のスコアの算出

ピッチの高低は、知覚される音声の高低と直接結びついていることが知られている。主成分分析の結果（図5.4）では、“高いー低い”の固有ベクトルの重み係数は、第1主成分に対して0.78と大きな値をもつものの、第2主成分に対しても-0.52と、無視のできない大きさを持ち、第1主成分を構成する要素の1つと見るのは難しい。

そこで、第1主成分から第4主成分までの固有ベクトルの重み係数を使って、各刺激の主成分スコアから、“高いー低い”の心理スコアを計算することにした。主成分分析により求められる固有ベクトルの重み係数は、ある評価項目の、主成分に対する重みを表しているのので、各刺激の“高いー低い”のスコア x_h は次式で表すことができる。

$$x_h = 0.78x_1 - 0.52x_2 - 0.20x_3 - 0.01x_4 \quad (6.1)$$

ただし x_1, x_2, x_3 および x_4 は、音声刺激の第1～4主成分それぞれの主成分スコア、0.78, -0.52, -0.20 および -0.01 は第1～4主成分それぞれの固有ベクトルの重み係数である。

ピッチ、第1フォルマントについて、男女のデータの規格化を行った。男声、女声のサンプルについて別々に平均値と標準偏差を求め、男声、女声のサンプルから、対応する平均値を引き、標準偏差で割って、規格化した数値を得た。

この規格化したピッチと第1フォルマントについて、第1～第4主成分の主成分スコアと、6.1式によって得られる“高いー低い”のスコアとの間の相関を調べた。

表6.2 ピッチ、第1フォルマント、第3主成分の主成分スコアのそれぞれの男女別平均値、標準偏差とt検定の結果（両側検定）

第3主成分の主成分スコアの検定結果については表5.6を再掲した。

	男声		女声		t検定の結果	
	平均	標準偏差	平均	標準偏差	t	p
ピッチ (Hz)	125.90	13.55	221.09	21.61	-11.94	p<0.01
第1フォルマント	452.12	36.68	554.77	60.26	-4.65	p<0.01
第3主成分“男性的なー女性的な”の主成分スコア	0.70	0.72	-0.64	0.84	3.71	p<0.01

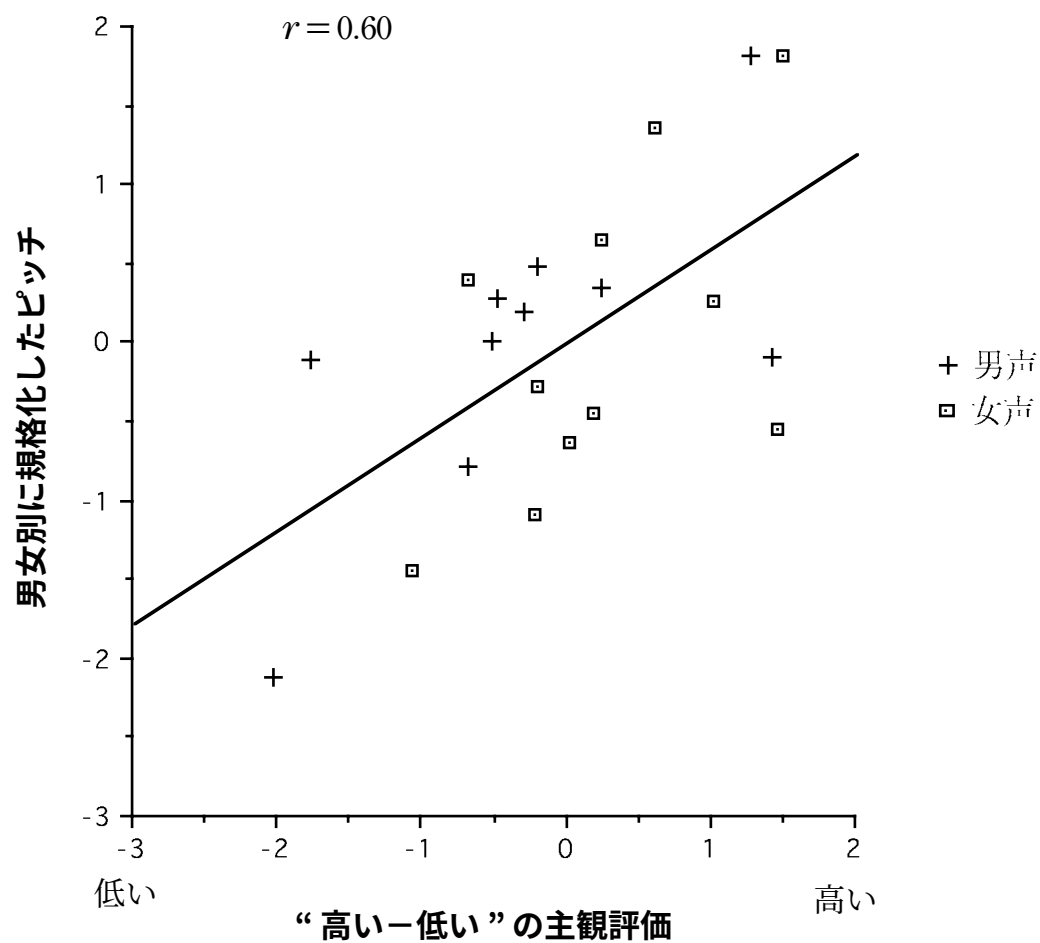
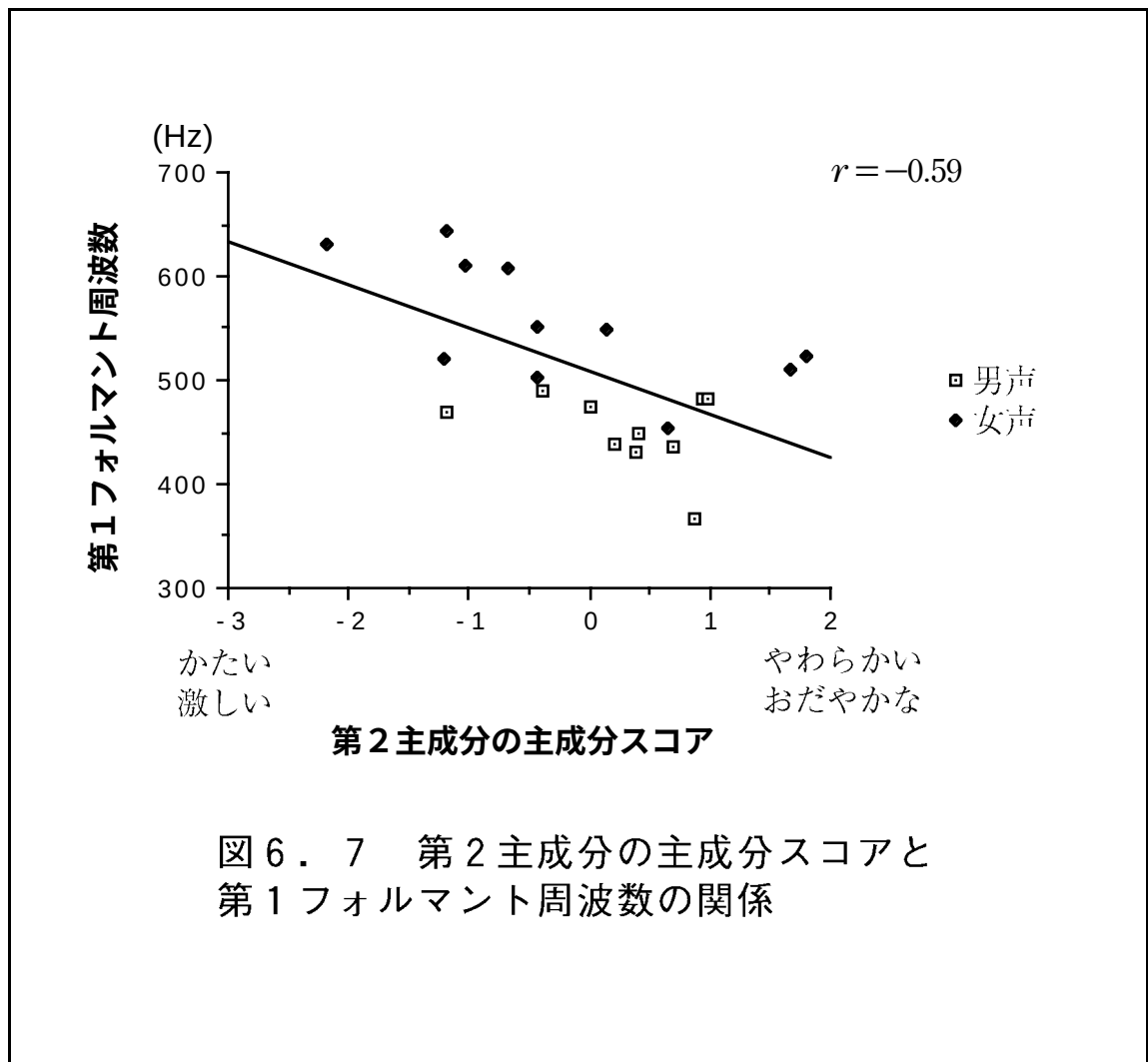


図 6. 6 男女別に規格化したピッチ周波数と
“高い—低い”の評価との関係



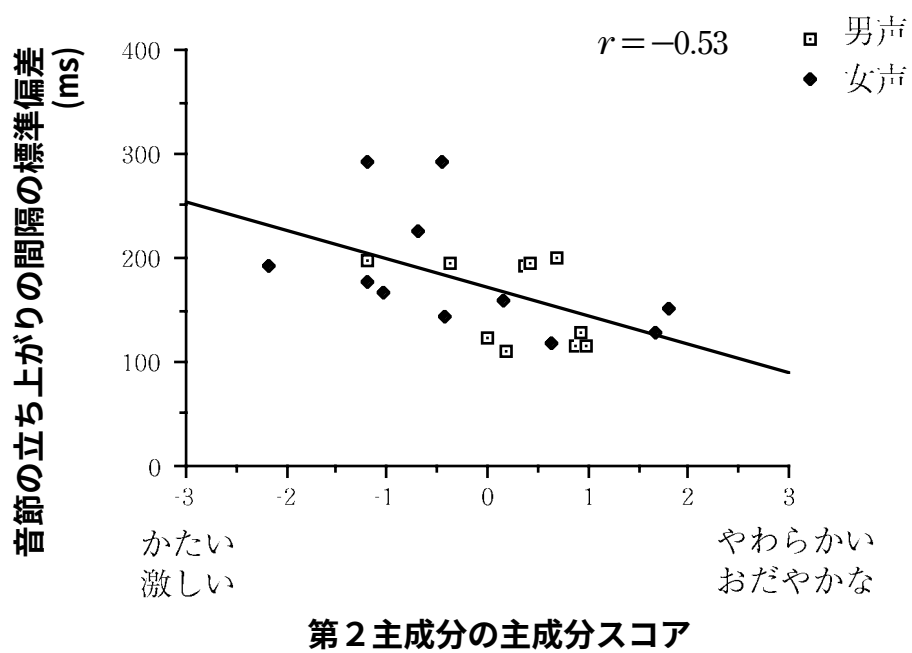


図6.8 音節の立ち上がりの間隔の標準偏差と第2主成分の主成分スコアの関係

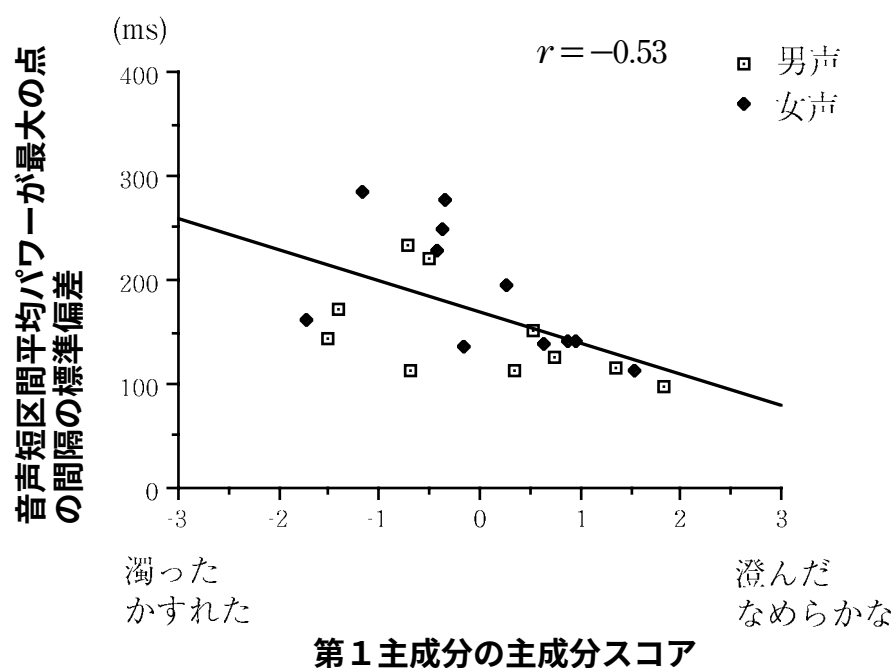


図6.9 音声短区間平均パワーが最大の点の間隔の標準偏差と第1主成分の主成分スコアとの関係

表 6. 3 分析して得た物理量と有意な相関が認められた心理量

心理量として、第1～第4主成分の主成分スコアおよび、6. 1式によって求めた“高いー低い”のスコアを用いた。表中の左側の数値は相関係数を表し、右側のかっこ内は有意と認められた相関の有意水準（両側検定）を表す。
+印の用語についてはp.90の脚注参照。

心理量	ピッチの平均値	男女別に規格化したピッチ	第1 フォルマント	音節の立ち上がり間隔の標準偏差	音声短区間平均パワー+最大の点の間隔の標準偏差
第1主成分 澄んだー濁った なめらかなーかすれた					-0.53 (p<0.05)
第2主成分 やわらかいーかたい おだやかなー激しい			-0.59 (p<0.01)	-0.53 (p<0.05)	
第3主成分 男性的なー女性的な	-0.67 (p<0.01)				
第4主成分 個性の強いー平凡な 高いー低い		0.60 (p<0.01)			

6.3.3 発話テンポのばらつきの相互関係

音声短区間平均パワーが最大となる点の間隔の標準偏差は第1主成分と相関

があり、音節の立ち上がり間隔の標準偏差と中央の間隔の標準偏差は第2主成分と相関が認められたが、これら3種類のパラメータは、測定する基準だけを変えて得られたので、完全に独立しているとは考えられない。そこで、3種のパラメータ間の相関を調べ、相互の依存性と主成分スコアとの関係について検討を行うことにした。

6.3.4 相関係数の検定

ピッチ、第1フォルマント、発話テンポのばらつきについて、相関が認められる主成分スコアとの間で相関係数の検定を行う。いずれの場合も両側検定とした。

6.4 結果

表6.2に、ピッチ、第1フォルマントそれぞれの男女別平均値と標準偏差、およびこれらに基づいて行ったt検定の結果を示している。第3主成分“男性的な—女性的な”の主成分スコアについては、第5章の表5.6を再掲した。男声、女声のサンプル数は、いずれの場合も第5章の実験の話者の数であるので、それぞれ、10と11である。いずれについても両側検定で、有意水準1%で、男声、女声の分布に差があることが認められた。

図6.6は男女別に規格化したピッチ周波数と“高い—低い”のスコアとの関係を表している。図6.7に第2主成分の主成分スコアと第1フォルマント周波数の関係を示している。図6.8に音声の立ち上がり間隔の標準偏差と第2主成分の主成分スコアの関係を示している。それぞれの図の中の直線は回帰直線である。

表6.3に有意な相関が認められた関係をまとめた。

6.5 考察

ピッチ、フォルマントについては、男女の発声のメカニズムの寸法に起因する差があることが知られている。主観評価実験に用いた音声について男女の差についてt検定を行った表6.2の結果でも、実験に用いた音声は、ピッチ、第1フォルマントともに男女間で有意な差が認められた。

第3主成分“男性的な—女性的な”の主成分スコアについても、前章の“5.4 考察”で示したように同様に男女のスコアの間で有意水準1%で差が認められている。

男女の分布に差がある“男性的な—女性的な”の主成分スコアに対する、同

第6章

様に男女の分布に差があるピッチおよび第1フォルマントの分布を求めると、図6. 4で明らかなように、男声のサンプルと女声のサンプルは、別々のグループを形成する分布になってしまう。このような場合、男声のサンプル、女声のサンプルそれぞれについては相関がなくても相関係数としては大きな値になる場合がある⁶²⁾。

こうしたケースでは、男女のサンプルについて別々に主成分スコアとの関係を調べる方法と、男女のサンプルについて同時に扱えるような規格化の処理を行う方法とが考えられる。前者のメリットとしては、データに処理を施さない、男女のサンプルが別々の心理量との関係を明らかにできる可能性がある点が挙げられるが、主成分スコアと比較するサンプル数が少なくなるので求められる関係の信頼性が低くなるというデメリットがある。後者の方法は、ちょうど前者の特徴とは反対で、データを規格化処理するのでその処理の妥当性の検討も必要になる、サンプル数は減少しないなどの得失を持つ。

本実験では、もともとサンプルの数が男声10、女声11と多くない。そこで、結果の信頼性の低下を避けることを優先して、後者の規格化の手法を用いた。

規格化したピッチは、第1～第4主成分から求めた“高いー低い”のスコアと $r=0.60$ の正の相関が認められた。図6. 6にその関係を示している。回帰直線から広がりはあるものの、回帰直線方向に長く分布している。また、相関係数の検定の結果、有意水準1%で有意な相関といえる（両側検定）。

この結果は、男女のピッチの分布に大きな差があるにもかかわらず、被験者は無意識的に、規格化のような処理を行い、同じスケール内で男女のピッチを評価する傾向のあることを示している。

このように、男女の音声の性質に隔たりがあり分布が不連続であるにも関わらず、被験者は無意識的に規格化のような処理を行って、同じスケール内で評価を行うことがある。その一方、一連の検討で物理的な特性との関係は認められていないが、“男性的なー女性的な”のように男女別々の枠の中で評価を行うこともある。音声の物理特性のうち、特に周波数領域のパラメータのように男女間の数値に差がある場合には慎重に吟味する必要がある。

一方、同じ周波数領域のパラメータのうち、第1フォルマントは、表6. 2に示すように、男女の数値に有意な差が認められたにもかかわらず、予備的な分析で、男女の数値に差がある第3主成分“男性的なー女性的な”の主成分スコアではなく、第3主成分とは直交している[†]第2主成分“やわらかい（おだやかな）ーかたい（激しい）”の主成分スコアと、 -0.59 の負の相関が認められた。この関係を示したのが図6. 7である。t検定の結果から判明した男女の第

[†]主成分分析によって求められた主成分はそれぞれ直交している。

1 フォルマントの差が図にも現れている。しかし、図6. 3のピッチとは異なり、男女は別々に離れて分布はしておらず、一部男女のサンプルが重なって連続している。このため第3主成分の主成分スコアとの間に、ピッチのときのような見かけ上の相関が現れなかったと考えられる。

図6. 7では、男女のサンプルとも回帰直線から $\pm 100\text{Hz}$ くらいの幅を持って回帰直線の方に広がって分布していて、検定の結果でも、この相関は有意水準 1%で有意と認められた。

以上のように、第1フォルマントの周波数が、前章で第2主成分として抽出された、“やわらかい（おだやかな）—かたい（激しい）”の印象の形成に参与している可能性が示唆された。

音節の発話のテンポのばらつきと、主観評価との関係について、立ち上がりの時間間隔の標準偏差、音節の中央の間隔の標準偏差、音声短区間平均パワー最大の点の間隔の標準偏差の3者相互の相関を調べると、表6. 4のようになる。立ち上がりの間隔、音節の中央の間隔それぞれの標準偏差の間には相関係数が0.96という非常に強い相関関係があり、別のパラメータとして扱うのには無理がある。そこで、代表して立ち上がりの間隔の標準偏差のみについて検討していく。この他の関係についても相関係数が0.85, 0.73と、主観評価の結果と物理量の関係、あるいは主観評価の結果同士の関係等と比較すると強い相関であり、表6. 4の関係はすべて有意水準 1%で有意な関係である。しかし、予備的な分析で、立ち上がり（音節の中央の点）の間隔の標準偏差と音声短区間平均パワー最大の点の間隔の標準偏差は違う主成分の主成分スコアとそれぞれ相関が認められたのでそれぞれについて検討する。

図6. 8の立ち上がりの間隔の標準偏差と第2主成分の主成分スコアの関係では、女声のサンプルがふたつほど200msの値となって回帰直線から離れて分布しているが、その他のサンプルは回帰直線にそって、回帰直線の方に広がって分布している。相関係数-0.53も1%の有意水準で有意と認められる。音節の間隔のばらつきが大きくなると、“激しい”印象を与えることには矛盾がないので、音節の立ち上がりの間隔の標準偏差が第2主成分“やわらかい（おだやかな）—かたい（激しい）”の印象の形成に何らかの関係がある可能性が示唆された。

図6. 9の音声短区間平均パワー最大点の間隔の標準偏差と第1主成分の主成分スコアの関係では、主成分スコアが正の領域では、サンプルが回帰直線に沿うように分布している。負の領域では、回帰直線を挟むように、上下に分かれて分布している。負の領域での回帰直線からのかい離は小さいとはいえない。しかし、音節の発話間隔のばらつきが小さいことが、なめらかさの印象を与えるという解釈は無理がない、全体の傾向として、第1主成分の主成分スコアと音声短区間平均パワー最大点の間隔の標準偏差との相関関係は、回帰直線に沿

第6章

う傾向があり、 -0.53 の負の相関も 1%の有意水準で有意と認められる、などの理由から音声短区間平均パワー最大点の間隔の標準偏差が、第1主成分“澄んだ（なめらかな）—濁った（かすれた）”の主成分スコアと何らかの関係があることは否定できない。

同じ発話のテンポのばらつきを示す指標が、基準の取り方を変えることで、異なる主成分と関係を持っている可能性が認められたが、一連の実験結果、分析結果からはその根拠は明らかにできていない。今後、規則合成や単音節単位の編集合成などで、より自然な音声を合成していくためには、発声のテンポの基準をどのようにとるべきか、さらに主観評価実験を行ったり、自然音声を分析していく必要性がこれらの結果から明らかにされた。

今回、物理量と心理量の関係を求める試みを行い、いくつかの物理量と心理量との関係が明らかになった。得られた相関の強さはいずれも関係を結論づけるには十分とはいえない。また、今回の手法では、物理量から心理量への関係としては必要条件が得られたにすぎない。今後、信頼できる物理量と心理量との関係を求めるためには、今回のような手法により多くの関係を発見し、物理量から心理量への十分条件を認められるような実験方法を考案して、それらの関係について検討を行っていく必要がある。

6. 6 章のまとめ

主観評価と物理量の関係を求める方法の1つとして、自然音声の分析を行って得られた物理量と、主観評価実験の結果との関係を調べた。

心理実験に用いた刺激音声について網羅的行った予備的な分析の結果相関が認められた、ピッチ、第1フォルマント、および発声テンポのばらつきについて、主成分分析の結果と比較検討を行った。

その結果、ピッチのように男女で特性が大きく異なる場合、心理実験の結果と比較する際に、男女別々に規格化する必要性が認められた。規格化したピッチと“高い—低い”のスコアの間に相関関係が認められ、規格化の手法の妥当性が確認できた。

一方、第1フォルマントについては、ピッチと同様に男女のデータの間で分布に差が認められたにもかかわらず、規格化の手続きなしで、第2主成分“やわらかい（おだやかな）—かたい（激しい）”の印象の形成に関係する可能性が認められた。

発話のテンポのばらつきについては、2種の標準偏差がそれぞれ別の主成分との間で有意な相関を持つことが認められ、テンポの基準の取り方を変えることで、違う次元の特性として感性に影響を与える可能性が示唆された。

条件を再検討して、さらに詳しく関係調べるための実験を行う必要があるが、規則合成等において、立ち上がりと強勢とを別々に制御することで、聞き手に

音声の物理的特性と心理的評価量の関係解明の試み

与える印象を制御することが可能になると考えられる。

また、自然音声に対する主観評価結果と、その音声を分析して得た物理量とを比較することにより、主観評価と物理量の間関係を調べていくという方法論の有効性も確認された。

第7章

結論

本論文は、通商産業省工業技術院生命工学工業技術研究所（およびその前身の製品科学研究所）にて行ってきた、「盲人用読書器の出力音声の評価法の研究」、「高齢者におけるヒューマン・エラーの生理心理的特性の研究」等、マン・マシンインターフェース関連の研究プロジェクトにおける合成音声の評価方法に関する成果をまとめたものである。この章では、本研究で得られた主要な成果を要約し、意義を明らかにして、今後の展望などを述べて結論とする。

以下に、各章ごとの結果をまとめる。

第1章「序論」では、本研究の必要性、目的および位置づけ等について述べ、論文の構成について説明した。

第2章「合成音声評価の現状と問題点」では、音声合成研究の中での合成音声評価の必要性、現状における問題点について、社会生活で一般的に行われている評価、科学技術分野で行われる評価との関連から詳しく述べた。こうした議論をふまえ、合成音声の評価の枠組みを明確にし、本研究の位置づけおよび必要性を明らかにした。

第3章「短期記憶による明瞭度評価法」では、音声の伝達する文字情報の正確さを評価する明瞭性の評価方法について、従来の明瞭性の試験方法では配慮されていなかった注意の分散の影響を加味した、新たな明瞭度の評価方法を提案し、実験的にその有効性を確認した。

通常、合成音声の明瞭性、了解性の評価には、単音節や2～3連音節を提示する明瞭度の試験や、意味のある単語や文を提示する了解度試験が行われる。しかし、合成音声が使用される実用的な場面では、連続的に音声提示されるうえ、音声を聞きながら、何らかの思考、判断、運転、操作など（タスク）を同時に行うのが一般的である。この同時に行うタスクにより、音声の聴取に対

第7章

する注意の度合いが低下するので、従来の音声の聴取のみに集中できる明瞭性の評価方法では条件が不十分といえる。

そこで、6～8の単音節を連続的に提示することにより、聞き取った音声を忘れないように頭の中で繰り返す“リハーサル”が必須になるような明瞭度評価法を考案した。この方法によれば、リハーサルに注意が向く分、音声の聴取に対する注意が低下する。この評価方法の有効性を確認するため実験を行った。

一連の実験の結果、通常の明瞭度試験で正答率が十分高くても、連続的に発声されると正答率が著しく低下する単音節があることが明らかになった。これらは連続系列中に出現する順序によっても正答率が変化することも認められた。この結果から、提案する短期記憶による明瞭度評価法が、通常の明瞭度試験よりも単音節の音質に対してより高い感度をもつことが確認された。より実用的な明瞭性の評価方法であるとともに、単語を用いた了解性の試験にも応用可能である。

第4章「単音節音声の音韻長の主観的等価性」では、音声が伝達する文字以外の付加的な情報のうち、単音節レベルの音声の物理的特性とその心理的な影響について調べた。

音声の物理的な特性とその心理的な影響、すなわち音声の強弱、高低、長短等を人が聞いた場合どのような尺度で受けとめているかという関係は、基本的に重要な情報である。これらの関係の中で日本語の単音節の音韻継続時間に関する物理的な特性と心理的な特性との関係は、明瞭性の試験などの標準音声を作成したり、単音節単位で音声を出力する音声合成装置の設計に必要であるにもかかわらず、これまで詳しく調べられていなかった。

そこで被験者に母音の/a/を基準音声として最初に聞かせ、続いて比較する単音節を聴取させ、音韻継続時間を直接比較させる実験を行った。比較する単音節には、音韻継続時間の区別できる閾値を調べる実験などでは同じ/a/を用い、各単音節の心理的に等価な長さを調べる実験では、それぞれの単音節を用いた。

一連の実験の結果、音韻継続時間の差異を認識する何らかの能力またはスキルには個人差が大きいこと、心理的に等価に聞こえる単音節の物理的な音韻継続時間は、その単音節の有声音の割合に概ね逆比例することなどが判明した。

単音節音声の音韻継続時間の基本的な特性が得られたことで、単音節単位の編集合成方式における音韻継続時間の指標として用いることができるほか、明瞭性試験などで用いる試験用標準音声の作成の指標にもなる。

第5章「合成音声の音質評価のための主観評価項目の抽出」では、人が音声を聞く際に、どのような評価項目に注目しているか、無意味つづりを読み上げた自然音声を用いた心理実験を行い、結果を分析して調べた。

音声合成技術は既に広く普及しているが、日常生活に広く浸透していくためには、了解性、明瞭性が高いだけでは不十分で、感性への影響も考慮する必要

がある。しかし、合成音声の使用を想定した条件のもとで、感性への影響が調べられたことはこれまでほとんどなかった。そこで、無意味つづり音声を用いて、意味内容や、感情などの付加価値によらない、声質が人に心理的に及ぼす影響について実験を行った。

実験は、男女 21 名の読み上げた無意味つづり音声を、被験者に聞かせて、アンケートに基づいて作成した評価項目によって評価させるという方法で行った。得られた結果から相関行列を作成して主成分分析を行った。

その結果、透明感、かたさ、男性的-女性的、個性の強さを表す合計 4 つの評価軸が得られた。これらの評価軸は、合成音声の感性への影響を評価する際の指標として用いることができる。

第 6 章「音声の物理特性と心理的評価量の関係解明の試み」では、第 5 章の実験で用いた無意味つづり音声の物理的特性のいくつかについて分析を行い、その結果と第 5 章で得られた主観評価実験の結果を比較して、主観評価と物理量の関係について調べた。

はじめに音声波形の物理特性（統計量）について網羅的に分析を行い、それぞれの結果と第 5 章で抽出された各主成分の主成分スコアとの間の相関を調べた。その結果、比較的強い相関の認められたピッチ、第 1 フォルマント、および発話テンポのばらつきの指標についてさらに詳しく調べた。

その結果、ピッチのように男女の音声の特性が大きく異なる場合、男女別々に規格化する必要性が確認された。また、発声のタイミングについては、2 種の基準の取り方による標準偏差がそれぞれ別の主成分との間で有意な相関が認められた。タイミング測定の基準の取り方を変えることで、違う次元の特性として感性に影響を与える可能性が示唆された。さらに、第 1 フォルマントは、かたさを表す主成分スコアとの間で有意な相関が認められた。

いくつかの関係が求められたことから、自然音声进行分析して主観評価の結果との関係を調べて主観評価と物理量の関係を探るという手法の有効性が確認された。また、それぞれ相関関係が認められた物理量は、それぞれ関係の認められた主成分の印象の形成に関係しているという可能性が示唆された。今後、条件をしばって個々の関係についてより詳しい実験を行う必要があるが、これら、関係が認められた物理量を制御することにより、聞き手に与える印象を制御することができると思われる。

音声合成は、着実な進歩を続けており、既に広く実用化もされている。職場や家庭にパーソナルコンピュータが文字どおり個人単位に普及し、高速大容量のネットワークで結ばれるようになる近未来には、音声合成の社会生活における役割はさらに大きくなるであろう。

このように社会に広く普及していく上で、本研究で明らかにしたような、実験の条件による明瞭性の低下、合成音声の心理的に及ぼす影響についてさらに

第7章

広く研究を行い，使用目的，利用環境，作業内容等に応じた評価の指針を作成していく必要がある．

謝辞

本論文を結ぶにあたり，ご指導ご鞭撻をいただきました，東京工業大学 小林隆夫教授に，心から感謝の意を表させていただきます。また，貴重なご討論，ご意見をいただきました，同大学 内川恵二教授，長橋宏教授，河原田弘教授，佐藤誠教授に感謝いたします。

本論文のもととなる，マンーマシンインターフェースおよび人間工学関連の研究プロジェクトのリーダーであり，ご指導いただいた元生命工学工業技術研究所主席研究官 永村寧一氏，および同研究所 口ノ町康夫部長にも感謝いたします。

筆者が研究の道に入るきっかけとなる助言をいただき，今はともに故人となっている叔父 石原和喜，伯母 矢野寿恵子の両氏に感謝して冥福を祈ります。

また，東京工業大学の先輩として，貴重なご助言をいただいた名古屋大学 生田幸士教授，友人としてアドバイスしていただいた電子技術総合研究所 小野栄一主任研究官にも謝意を表します。

そして，卒業研究，修士課程において親身にご指導いただき，修了後も公私にわたりたいへんお世話になっている東京工業大学 深尾正教授に，深く感謝の意を表させていただきます。

参考文献

- 1) 精選国語辞典: 明治書院, 東京, 1994.
- 2) 電子通信学会編: 聴覚と音声, 電子通信学会, 東京, 1975.
- 3) 電子情報通信学会編: 新版 聴覚と音声, 電子通信学会, 東京, 1980.
- 4) 中田和男: 音声, コロナ社, 東京, 1977.
- 5) 安居院猛, 中嶋正之: コンピュータ音声処理, 産報出版社, 東京, 1980.
- 6) Rabiner. L. R., 鈴木久喜訳: 音声のディジタル信号処理, コロナ社, 東京, 1983.
- 7) 佐藤大和: 音声合成の特集にあたって, 日本音響学会誌 49 巻 12 号, 845-846, 1993.
- 8) 粕谷英樹, 渡辺隆夫, 杉山雅英, 小林隆夫, 阿部匡伸, 武田一哉: 音声研究の現状と課題, 日本音響学会誌 51 巻 8 号, 618-620, 1995.
- 9) Itakura, F., & Saito, S.: An audio response unit based on partial correlation, IEEE Transaction of Communications, COM-20(4), 792-797, 1972.
- 10) 広川智久: 規則合成における音声合成単位及び音声合成法 —より高品質を求めて—, 日本音響学会誌 49 巻 12 号, 847-853, 1993.
- 11) 匂坂芳典: 韻律制御研究の現状と将来 —より自然な音声を求めて—, 日本音響学会誌 49 巻 12 号, 854-859, 1993.
- 12) 山下洋一: 朗読から対話音声合成へ, 日本音響学会誌 49 巻 12 号, 860-865, 1993.
- 13) 三留幸夫: 音声合成の応用と今後の展望, 日本音響学会誌 49 巻 12 号, 875-880, 1993.
- 14) 北島宗雄, 篠原正美: 文字・音声変換システムの概要, 昭和 58 年度電子通信学会全国大会, 1454, 1983.
- 15) 篠原正美, 北島宗雄: 文字 - 音声変換システム, 情報処理学会, コンピューターにおけるヒューマン・インターフェースシンポジウム報告集, 7. 21-23, 1983.
- 16) 大田博志: 福祉から見た音声合成装置, 日本音響学会誌 49 巻 12 号, 871-874, 1993.
- 17) 相川哲弥: プログラム実行途中でも説明文ファイルを読める盲人用補助機能の設計, 電子情報通信学会技術研究報告ヒューマンコミュニケーション基礎, HCS97-7, 1-5, 1997.
- 18) 技術研究組合医療福祉機器研究所編: 高福祉社会をめざして～産業科学技術研究開発制度(医療福祉機器技術研究開発)～, 財団法人日本産業技術振

参考文献

- 興協会, 東京, 1997.
- 19) 大山正, 今井省吾, 和気典二編: 新版感覚知覚ハンドブック, 誠信書房, 東京, 1994.
 - 20) 斉藤幸子: ニオイと官能検査, 香料, 158, 63-70, 1988.
 - 21) 金田弘拳, 前島こず恵, 小早川達, 綾部早穂, 斉藤幸子: 味, 匂いの弁別, 同定における高齢者と若者の比較, 日本味と匂学会誌, 4(3) 507-510, 1997.
 - 22) 野呂影勇編: 図説エルゴノミクス, 日本規格協会, 東京, 1990.
 - 23) 柳井晴夫, 岩坪秀一, 石塚智一編: 人間行動の計量分析, 東京大学出版会, 東京, 1990.
 - 24) 栗山洋四, 江原勝夫, 山根木正人編: 感性計測先端技術技術集成, サイエンスフォーラム, 東京, 1991.
 - 25) 金村早穂, 斉藤幸子, 飯田健夫: ニオイの快・不快の直接的・間接的評価法の検討, 人間工学, 25(2), 77-85, 1989.
 - 26) Saito, S. and Iida T.: Psychophysics of Gustation and Olfaction, Sensors and Materials, 4 (3), 121-133, 1992.
 - 27) 関根道昭, 斉藤幸子, 飯尾心: 連続的に提示させた臭気に対する金属酸化物半導体ガスセンサの応答特性, 臭気の研究, 26(2), 46-127, 1995.
 - 28) 山口佳子, 斉藤幸子: 複数種のガスセンサとニオイの感覚特性を利用した食品の腐敗度判定, 人間工学, 33(2), 71-78, 1997.
 - 29) 粕谷英樹: 音声合成技術の評価, 日本音響学会誌, 48(1), 46-51, 1992.
 - 30) 粕谷英樹: 合成音声の品質評価法, 日本音響学会誌 49 巻 12 号, 866-870, 1993.
 - 31) 中田和男: 音声入出力装置の設計・評価の立場から, 日本音響学会第 1 回シンポジウム「試験用音声の標準化」, 73-75, 1985.
 - 32) 飯田茂隆: 建築空間における明瞭度研究について, 日本音響学会第 1 回シンポジウム「試験用音声の標準化」, 1-9, 1985.
 - 33) 小川有子, 西隆司, 古河宣一: 無意味三連単音節による明瞭度試験, 日本音響学会第 1 回シンポジウム「試験用音声の標準化」, 10-16, 1985.
 - 34) 単音節明瞭度試験テープ, 日本音響コンサルタント協会, 1979.
 - 35) アブラムソン著, 宮川洋訳: 情報理論入門, 好学社, 東京, 1978.
 - 36) 粕谷英樹, 楊長盛: 音源から見た音質, 日本音響学会誌, 51(11), 869-875, 1995.
 - 37) 粕谷英樹, 糟谷清基, 森田清紀: 校正支援用規則合成音声の発声速度の役割, 日本音響学会誌, 47(2), 96-98, 1991.
 - 38) 新版 心理学事典: 平凡社, 東京, 1989.
 - 39) 板橋秀一: テキスト音声合成装置性能評価法のガイドライン, 日本音響学会誌, 52(2), 144-149, 1996.

- 40) Greene, B. G., Logan, J. S., & Pisoni, D. B.: Perception of synthetic speech produced automatically by rule: intelligibility of eight text-to-speech systems, *Behavior Research, Methods, Instruments, & Computers*, 18, 100-107, 1986.
- 41) Klatt, D. H., & Klatt, L. C.: Analysis, synthesis, and perception of voice quality variations among female and male talkers, *The Journal of Acoustical Society of America*, 87, 820-857, 1990.
- 42) Pisoni, D. B., & Hunnicutt, S.: Perceptual evaluation of Mitalk: the MIT unrestricted text-to-speech system, *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, 80 No. II, 572-575, 1980.
- 43) Pisoni, D. B., Nusbaum, H. C., & Greene, B. G.: Perception of synthetic speech generated by rule, *Proceedings of IEEE*, 73, 1665-1676, 1985.
- 44) Spiegel, M. F., Altom, M. J., Macchi, M. J., & Wallace, K. L.: Comprehensive assessment of the telephone intelligibility of synthesized and natural speech, *Speech Communication (Netherlands)*, 9, 279-291, 1990.
- 45) 伊東久美子: 感情を含む音声に関する基礎研究(III), *人間工学*, 22(4), 211-217, 1986.
- 46) Kintsch, W.: *Memory and Cognition*, Wiley, New York, 1977.
- 47) Loftus, G. R. and Loftus, E. F. 著, 大村彰道訳: *人間の記憶*, 東京大学出版会, 東京, 1980.
- 48) 岩原信九郎: *教育と心理のための推計学*, 日本文化科学社, 東京, 1979.
- 49) 芝祐順: *統計的方法II 推測*, 新曜社, 東京, 1977.
- 50) 熊耳憲一, 粕谷英樹: 規則合成音声への使用者の順応過程の客観的評価, *日本音響学会誌*, 47(4), 243-249, 1991.
- 51) Beckman, M. : Segment Duration and the "Mora" in Japanese, *Phonetica*, 39, 113-135, 1982.
- 52) Fowler. C. A. : Converging Sources of Evidence on Spoken and Perceived Rhythms of Speech: Cyclic Production of Vowels in Monosyllabic Stress Feet, *J. of Experimental Psychology: General*, 112 (3), 386-412, 1983.
- 53) Hoequist, C. : Syllable Duration in Stress-, Syllable- and Mora-Timed Languages. *Phonetica*, 203-237, 1983.
- 54) Port, R. F. , Al-ani, S. and Maeda, S. : Temporal Compensation and Universal Phonetics, *Phonetica*, 37, 235-252, 1980.
- 55) Port, R. F. , Dalby, J. and O'Dell, M. : Evidence for mora timing in Japanese, *J. of Acoustical Society of America*, 81(5), 1574-1585, 1987.

参考文献

- 56) 匂坂芳典, 東倉洋一: 規則合成による音声合成のための音韻時間長制御, 電子通信学会論文誌, Vol. J67-A, No. 7, 629-636, 1984.
- 57) 片桐滋, 東倉洋一, 古井貞熙: 単音節知覚における時間情報の役割, 日本音響学会誌, 42(2), 97-105, 1986.
- 58) 中村章, 清山信正, 池沢龍, 都木徹, 宮坂栄一: リアルタイム話速変換型受聴システム, 日本音響学会誌, 50(7), 509-520, 1994.
- 59) 都木徹, 清山信正, 宮坂栄一: 複数の窓幅から得られた自己相関関数を用いる音声基本周波数抽出法, 電子情報通信学会論文誌, Vol. J80-A, No. 9, 1341-1350, 1997.
- 60) 圓川隆夫: 多変量のデータ解析, 朝倉書店, 東京, 1988.
- 61) 菅民郎: 多変量解析の実践 上, 現代数学社, 京都, 1993.
- 62) 市原清志: バイオサイエンスの統計学, 南江堂, 東京, 1990.
- 63) 竹内晴彦, 稗田一郎: 調査報告一音声特徴の記述に関する調査, 計量国語学, 16(2), 70-80, 1987.
- 64) Lawley, D. N., & Maxwell, A. E.: Factor analysis as a statistical method (2nd edition), Butterworth, London, 1971.
- 65) 奥野忠一他: 多変量解析法, 日科技連出版社, 東京, 1971.
- 66) 奥野忠一他: 続多変量解析法, 日科技連出版社, 東京, 1976.
- 67) 柳井晴夫, 高根芳雄: 多変量解析法, 朝倉書店, 東京, 1977.
- 68) 渡辺俊朗: 規則合成音の自然性評価法の検討, 電子情報通信学会論文誌 A, J74-A(4), 599-609, 1991.
- 69) 広中平祐他編: 現代数理科学事典, 大阪書籍, 大阪, 1991.

発表論文など

学会誌上

- 1) 稗田一郎, ロノ町康夫: 短期記憶による単音節音声の明瞭度評価法, 人間工学, 23(5), 325-331, 1987.
- 2) 稗田一郎: 合成音声の音質評価のための主観評価項目の抽出, 人間工学, 24(6), 387-393, 1988.
- 3) 稗田一郎: 単音節音声の心理的に等価な長さ, 人間工学, 28(1), 21-28, 1992.
- 4) 稗田一郎, ロノ町康夫: 無意味綴り音声における発声のテンポの乱れが感性へ及ぼす影響の検討, 人間工学, 31(6), 415-416, 1995.
- 5) 稗田一郎, ロノ町康夫: 無意味綴り音声のピッチ周波数と主観評価の関係, 人間工学, 32(3), 159-161, 1996.
- 6) Hieda, I. & Kuchinomachi, Y.: Preliminary Study of Relations between Physical Characteristics and Psychological Impressions of Natural Voices, Perceptual and Motor skills, 85, 1483-1491, 1997.

学会口頭発表

- 1) 稗田一郎, ロノ町康夫, 永村寧一: 単音節音声の心理的に等価な長さ, 日本音響学会研究会資料, S8447, 365-372, 1984.
- 2) 稗田一郎, ロノ町康夫, 永村寧一: 短期記憶による明瞭度評価法, 日本音響学会研究会資料, S8533, 247-254, 1985.
- 3) 稗田一郎: 合成音声の音質の評価指標の抽出, 日本人間工学第28回大会講演集, 144-145, 1987.
- 4) 稗田一郎: 音声のパワースペクトルと主観的評価の関係, 電子情報通信学会春期全国大会講演論文集 分冊 1, 1-15, 1989.
- 5) 稗田一郎: 音声の主観的評価と物理特性の関係, 日本人間工学会第30回大会講演集, 226-227, 1989.

その他の発表

学会誌上

- 1) 竹内晴彦, 稗田一郎: 音声特徴の記述に関する調査, 計量国語学, 16(2), 70-80, 1987.

発表論文など

- 2) 口ノ町康夫, 稗田一郎: 視覚刺激負荷時の Rheoencephalogram(REG)変化率におよぼす電極配置の効果, 生理心理, 6 (1), 41-47 1988.
- 3) Kuchinomachi, Y., Kodama, H. & Hieda, I.: The Posture Effect on The Volume of The Body Fluid at a Lower Limb, Designing for Everyone, Vol. 3, 9-10, 1991.
- 4) 口ノ町康夫, 児玉廣之, 稗田一郎: インピーダンス法による膀胱内尿量の長時間計測, 人間工学, 28(3), 125-131, 1992.
- 5) Hieda, I., Sato, S. & Kuchinomachi, Y.: Impedance Mapping System Using Multiple Electrodes For Blood Flow Measurement In The Brain, Japan Sensor Newsletter, 7(1), 21-22, 1993.
- 6) 口ノ町康夫, 熊田孝恒, 児玉廣之, 稗田一郎: 加齢に伴う脳動脈硬化の簡易計測, 人間工学, 30(4), 233-240, 1994.
- 7) Kodama, H., Hieda, I., Kuchinomachi, Y., Takeichi, H. & Matsumoto, H.: Non-invasive Ultrasonic Urination Sensor for Ambulatory Patient Support, Methods of Information in Medicine, 33, 97-100, 1994.
- 8) Kodama, H., Hieda, I., Kuchinomachi, Y., Takeichi, H. & Matsumoto, H.: Ultrasonic Urination Sensor and Monitoring of Urine Volume in the Bladder, Biobehavioral Self-Regulation (Eastern and Western Perspectives), 263-267, 1995.
- 9) Kodama, H., Kuchinomachi, Y., Yamauchi, Y., Hieda, I., Sato, S., Takeichi, H. & Matsumoto, H.: Ultrasonic Urination Sensor for Home Use, Proceedings of the 13th Triennial Congress of the International Ergonomics Association, 5, 582-584, 1997.

学会口頭発表

- 1) 稗田一郎: 生体インピーダンス計測における組織抵抗率変化の端子に現れる大きさ(感度)の分布, 昭和61年度電気関係学会東北支部連合大会, 1986.
- 2) 稗田一郎: 生体インピーダンス計測における電極配置と感度分布の関係, 昭和62年度電気学会全国大会, 1987.
- 3) 稗田一郎, 口ノ町康夫, 児玉廣之: 携帯型尿意センサーの試作, 日本人間工学会第31回大会, 1990.
- 4) 稗田一郎, 口ノ町康夫, 佐藤滋, 児玉廣之: インピーダンス法による脳血流分布解析装置, 電気関係学会東北支部連合大会, 1991.
- 5) 稗田一郎, 口ノ町康夫, 佐藤滋, 児玉廣之: 尿意センサのための電極形状の工夫, 日本人間工学会第34回大会, 1993.
- 6) 稗田一郎, 児玉廣之, Rob Gill: 遠隔医療に適用するための医用画像フィルタの試み, 第12回日本ME学会秋季大会, 1998.

その他

- 1) 稗田一郎, ロノ町康夫, 永村寧一: 合成音声の「自然さ」に関する基礎的研究
無意味つづり音声からの主観的評価項目の抽出, 製品科学研究所研究報告,
106, 55-60, 1987.
- 2) Hieda, I., Kuchinomachi, Y., Sato, S. & Kodama, H.: Impedance Mapping
Method and System for Blood Flow Measurement in Brain, and
Mechatronics, Journal of Robotics, 4(4), 314-320, 1992.
- 3) 稗田一郎, 佐藤滋, ロノ町康夫: 多点電極を用いた脳血流分布解析装置, セン
サ技術, 10(11), 1992.
- 4) 稗田一郎, ロノ町康夫, 永村寧一: 音声の主観評価と物理的特性の関係, 製品
科学研究所研究報告, 124, 15-26, 1992.

論文要旨

本論文は、音声合成の研究開発に資する目的で行った、音声の主観的評価の研究についてまとめている。合成音声の研究開発は進み、広く実用化され、今日では、重要なマン・マシンインターフェースのひとつとなっている。合成音声の研究開発において、他の科学技術の研究開発と同様、評価は重要な課題の一つである。評価の中でも、伝達しようとする情報が聞き手に正確に伝わるかの評価と、聞き手の心理面への影響の評価は特に重要である。

本論文は全7章で構成されている。以下に各章の概要を記す。

第1章「序論」では、合成音声の研究開発の現状と将来について述べ、その中での合成音声の評価の必要性和問題点について触れた。また、本論文の構成について述べた。

第2章「合成音声評価の現状と問題点」では、音声合成研究の中での合成音声評価の必要性、現状における問題点について、日常生活で一般に行われている評価、科学技術分野で行われる評価との関連から詳しく論じた。こうした議論をふまえ、合成音声の評価の枠組みを明確にし、本研究の位置づけおよび必要性を明らかにした。

第3章「短期記憶による明瞭度評価法」では、文字情報の評価方法に、注意の分散の影響を応用した新たな評価方法を提案した。

明瞭性の評価法は、文字情報の伝達の正確さを測る評価方法として定着している。しかし、実際に合成音声を利用する場合、利用者は、音声を聞きながら、同時に判断や、操作・運転を行っているのが一般的であり、従来の評価方法による試験だけでは、実用的な明瞭性の評価として不十分である。

そこで、本研究では6～8の単音節を連続的に提示する明瞭度評価法を提案する。6～8個の単音節が連続して提示された後、被験者は回答用紙に聞き取った単音節を筆記する。被験者は聞き取った単音節を忘れないように頭の中で繰り返しながらかつて単音節を聞き取りを続けて筆記するため、注意の分散が生じる。被験者による実験を行い、その結果について検討を加え、提案する評価方法の有効性を示した。

第4章「単音節音声の音韻長の主観的等価性」では、広い意味で合成音声の評価のひとつといえる、音声の物理的な特性と心理的な特性の関係のうち、これまで調べられていない単音節の音韻継続時間の物理的な長さと心理的な長さの関係を調べる実験について説明し、その結果について検討を行った。

音声の物理的な特性とその心理的な影響、例えば音声の物理的な強弱、高低、

長短等と、心理的に受ける強弱、高低、長短の印象との間にどのような関係があるかということは、マンーマシンインターフェースとして音声を用いる上で基本的な情報である。これらの関係の中で日本語の単音節の音韻継続時間については、物理的な特性と心理的な特性との関係はあまり調べられていない。

そこで被験者に、母音の椰鐔椰を基準音声として最初に聞かせ、続いて比較する単音節を聴取させて長短を答えさせる実験を行い、単音節の長さの知覚の閾値、異種の単音節の心理的に等価な長さ等を求めた。

第5章「合成音声の音質評価のための主観評価項目の抽出」では、これまであまり行われてこなかった、合成音声の感性への総合的な影響の評価を試みた。

音声合成が広く普及しているが、社会生活で受け入れられていくためには、了解性、明瞭性が高いだけでは不十分で、感性への総合的な影響も考慮する必要がある。そこで、無意味つづり音声を用いて、意味内容や、感情などの付加価値に寄らない、音声の質（声質）そのものが人に心理的に及ぼす影響についての実験を行った。

男女の話者が読み上げた無意味つづり音声を、一人分ずつ被験者に聞かせて、評価実験を行い、その結果を主成分分析にかけて検討した。分析の結果抽出された指標は、合成音声の感性への影響の評価に利用できると期待される。

第6章「音声の物理的特性と心理的評価量の関係解明の試み」では、第5章の実験で用いた無意味つづり音声について、ピッチ、音節の発話のタイミングのばらつきと等の物理的特性について分析を行い、第5章で得られた主観評価項目との関係について検討した。その結果、音節の立ち上がりタイミングのばらつき等、いくつかの物理量と主観評価の関係が示唆された。

あわせて自然音声を分析して主観評価の結果との関係を調べることにより主観評価と物理量の関係を調べるという手法の有効性についても吟味した。

第7章「結論」では、本研究で得られた成果を総括した。

音声合成技術が社会に広く普及していく上で、本研究で明らかにしたような、実験の条件による明瞭性の低下、合成音声が心理面に及ぼす影響についてさらに広く研究を行い、使用目的、利用環境、作業内容等に応じた評価の指針を作成していく必要がある。

Study of Subjective Characteristics of Voice for Assessment of Synthesized Speech

By Ichiro HIEDA

ABSTRACT

This thesis summarizes about study of subjective evaluation of voice, whose results would be applied to assessment of synthesized speech.

Synthesized speech becomes one of important man-machine interfaces. Evaluation is important to measure accuracy of information transmitted through the man-machine interface. Psychological effects of synthesized voice to human should be also evaluated. This thesis discussed framework of generic evaluation usable for synthesized speech today. New evaluation methods for synthesized speech were proposed to complete the framework.

Synthesized speech mainly transmits character information. Intelligibility tests of phonemes and syllables evaluate accuracy of the transmission. In these tests, subjects can concentrate on hearing of test voice. On the other hand, in most practical cases, a user is doing important tasks while listening to synthesized speech, for example, driving a vehicle, operating a device, or, thinking the next operations. Intelligibility test by short-term memory was proposed as evaluation reflecting fluctuation of attention, and effectiveness of the test was discussed from the result of the experiments.

Next, influence of synthesized speech to feeling of listeners was examined. Monosyllable is regarded as smallest unit of Japanese language. However, the relation between duration of monosyllable and its psychological effect has not been examined. Subjects compared the monosyllables of variable duration to standard monosyllable that had constant duration. The results of the experiment could be applied to monosyllable-based speech synthesis and intelligibility tests.

Total influence to feelings of listeners of synthesized voice is difficult subject left unsolved. As a preliminary study of the relations between physical characteristics and feelings of listeners, subjective evaluation was performed using natural voice. The results were analyzed by principal component analysis. Consequently, four pairs of evaluation items were extracted representing aspects to which most people were observed to pay attention when listening to voices.

Natural voices, which were used in the subjective evaluation, were analyzed for their physical parameters, for example, pitch, formants, utterance timings and correlations to principal component scores were calculated. As the result, relations between some physical parameters and psychological effects were found.