2004年度 卒業論文

「笑い声」合成のための音響特徴の分析

Analysis of acoustic features for laughing voice synthesis

提出日:2005年2月2日

指導教授

白井克彦 教授

早稲田大学 理工学部 情報学科

1G01P076-1

芳 賀 寿 昭

Toshiaki HAGA

目 次

図 目 次

表 目 次

第1章　序論

1.1　はじめに

人と人とのコミュニケーションにおいて感情音声は欠くことのできないものであるため、対話システムやCGキャラクターに人間らしさを表現させる際、感情音声の合成は当然の要求として求められる。また、CGアニメーションの作成は、専門家だけではなく一般にも普及し始めてきている一方で、その複雑さから感情音声との併用は未だ課題となっている。従って、今後はより容易な感情音声の合成技術が要求されると考える。

現在、音声合成の研究は盛んに行われており、音声に含まれる感情に関する研究もなされてきた[1][2]。従来では、「喜び」や「怒り」、「疑い」といった感情を含む対話音声を合成することが研究され、ある程度実現されてきた[3]。本研究では、生理的な発声による感情表現に着目することで人間らしさを実現するという観点から、「笑い」に注目した。

1.2　研究目的

本研究では、規則合成方式を用いて「笑い声」及び笑いながら喋る「喋り笑い」のいろいろな音声を自動的に合成することを目標としている。そのために、「笑い声」を分析して音声の合成規則や制御パラメータを明らかにすることを目下の研究目的とする。なお、本合成システムを構築する際に、ユーザインタフェースの操作項目として直感的に分かりやすい操作項目を用意することも検討する。

1.3　本論文の構成

本論文は全5章で構成される。本章では、序論として研究の背景と目的について述べた。以降の章では、以下のような内容について述べる。

第2章では、感情音声の分析・合成について本研究でのアプローチと、音声合成技術について述べる。

第3章では、音声収録実験の結果と、収録した笑い声の音声から「笑い声」のパラメータの時間変化パターンによる分類の検討、そして「通常発話」との比較により笑い声に特有な音響特徴を定量的に分析した結果を述べる。

第4章では、主観評価実験によって音響パラメータの笑い声らしさに対する貢献度を分析した結果を述べる。

第5章では、本研究のまとめ及び今後の展開についてを述べる。

第2章　感情音声の分析・合成

2.1　本研究でのアプローチ

本研究では、一般的で汎用性の高い「笑い声」の合成を目指すため、日常生活で出現頻度が高いと考えられる/ha/を基本とした笑い声を分析対象とする。分析する特徴パラメータとしては、音声パワー、ピッチ周波数、有声音・無声音の程度、スペクトル (音色)、音素タイミングを用いた。

まず、ユーザの要求に合わせた合成を行う為には、「笑い声」を単純に分類する必要がある。本研究では、大原ら [4] の行った「大爆笑」や「冷笑」といった感情での分類ではなく、その前段階として音の大小変化や高低変化といった聴覚的に分かり易い特徴からアプローチし、「笑い声」の特徴を確認していくだけでなく、ユーザインタフェースの操作項目になり得る要素を検討していく。さらに、その分類ごとに音響特徴を分析することで、分類を構成する要因を特定していく。

また、「笑い声」特有の音響特徴を知るためには、「通常発話」との比較が有効と考えられる。ここでは、ピッチ周波数、スペクトル、音素タイミングといった音響特徴を数値等によって比較することで定量的な分析を行った。

最後に、音響パラメータを必要に応じて操作して作成した合成音に対して主観評価実験を行い、「笑い声」の笑い声らしさに対する各音響パラメータの貢献度を調べた。

2.2　音声合成

音声合成方式には、主に以下の三つがある [5]。

- 録音編集方式
- パラメータ編集方式

- 規則合成方式

まず録音編集方式とは、あらかじめ録音しておいた音声の波形を蓄えておき、必要に応じてつなぎ合わせて再生する手法である。録音した音声をそのまま再生するので自然性は高いが、決まった音声しか再生できない。次にパラメータ編集方式とは、音声波形を分析して得られる物理的パラメータを蓄えておいて、そのパラメータによって音声合成器を制御する手法である。高圧縮の符号化をした音声を用いるので、録音編集方式よりも蓄積するデータ量は減るが、自然性はやや劣る。そして規則合成方式とは、単語より小さい音素・音節などの基本単位を用いて、それらの結合則や韻律情報（アクセント、イントネーションなど）に関する規則により音声合成器の制御パラメータを生成し、音声を出力する手法である。音素や音節といった要素を自由に組み合わせて合成するので、自由度が高い方式であるが、自然性は低くなる。

本研究では、任意の合成音を生成することが目的であるため、規則合成方式を用いて「笑い声」、さらには「喋り笑い」の音声合成を目指す。従来の規則合成方式では、音素をはじめとする言語的シンボルが有限であることから、あらかじめ蓄えられた有限個の音声パターンから任意の語彙を生成することは可能である。しかし、「笑い声」では背後に離散的なシンボルが存在せず、様々な笑い声やその程度を生成する必要がある。また、さらには「喋り笑い」は、音声の中に笑いの要素と言語的要素が混在しており、従来の言語的単位に応じた音声パターンを蓄積する手法では、様々な喋り笑いの音声を生成することは極めて非効率的となる。

規則合成方式による合成音生成の流れは次のようになる。まず、入力データとして言語的記号列を与える。これに対し、蓄積しておいた音声素片と、その合成規則・韻律制御規則から音声単位の伸縮・結合を経てパラメータ列を生成する。これを音声合成器にかけることで合成音を生成する。合成器の概要を図 2.1 に示す。

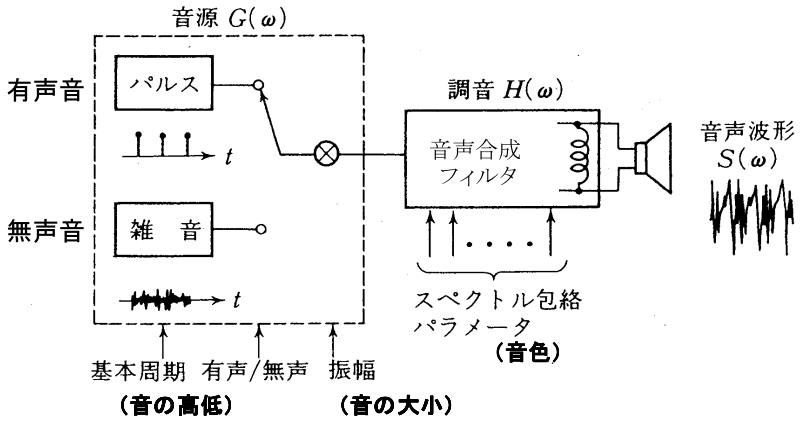

図 2.1: 音声合成器の構成

図 2.1 のように、駆動音源には有声音なら周期を持つパルス列、無声音なら白色雑音を用い、音声波形を作り出す。ここで、音の高低や大小が決められる。そして、調音部の音声合成フィルタにおいて、スペクトル包絡パラメータが与えられて音色が決定される。調音部は声道の共振特性を模擬する部分である。

本研究では入力として必要となる合成規則・韻律制御規則を特定していく。

第3章　音響特徴の抽出

3.1　音声収録

分析に用いる笑い声のデータ収集と、対話においてどのような笑い声が発声されるかを調べるため、男子大学生7人を対象に大学内のスタジオにて音声収録を行った。収録内容は、自然な「笑い声」及び「喋り笑い」収録のための二名または三名の被験者による30分〜1時間程度の自然対話である。収録した音声はDATに保存される。これにより収録した対話音声から、「笑い声」及び「喋り笑い」を抽出した。サンプリング周波数は16KHzである。

なお、収録ではヘッドセットマイクを用いた。これは、話者の可動性・自由度を向上させることでより自然な対話を促すのと同時に、マイクと発話者の口との距離を常に一定にすることで収録音声のパワー変化を純粋に発話の抑揚のみによって決めるという狙いがある。

収録データからの笑い声の抽出数を表3.1に示す。下段には/ha/以外を基本とした笑い声の総数を示した。笑い声には、主に/ha/、/hi/、/fu/、/he/、/ho/を基本とした笑い声が存在すると考えられるが、表3.1より、どの音素を基本とした笑い声を多く発声するかには個人差があることがわかる。話者Eの場合、/fu/を基本とした笑い声の方が多く、この傾向が顕著に現れる。しかし、全体でみればほぼ半数が/ha/を基本とした笑い声であることがわかり、/ha/を基本とした笑い声が一般的な笑い声であることを示している。一方で、収録音声には/ha/とも/fu/とも判別できない発音の笑い声も多数含まれており、今後はこれらを分類することも必要になると考える。また、/ha/を基本とした笑い声のうち、7割は「は」を2回または3回発声するものであった。

表 3.1: 話者別の笑い声の抽出数

	話者							
	A	B	C	D	E	F	G	合計
/ha/からなる笑い	26	31	38	29	9	13	11	157
その他の笑い	17	26	38	9	43	25	9	167

3.2 音素ラベリング

収録音声は音素ラベリングにより、無音区間である “Pause”、“音素/h/”、“音素/a/” に分ける。この際、Pause と/h/の境界はパワーの立ち上がり開始点、/h/と/a/の境界は有声・無声の変移開始点、/a/と Pause の境界はパワーの減衰終了点としている（図 3.1）。有声部分の分析は音素/a/の部分だけでよい。

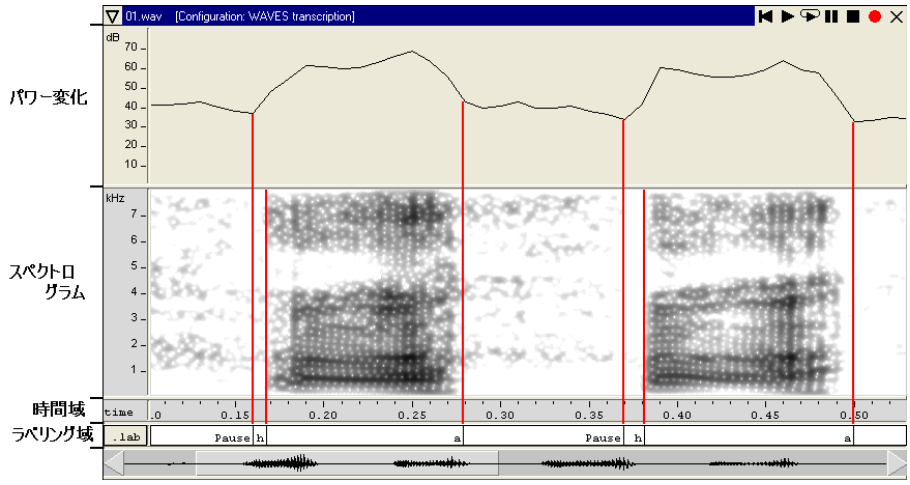

図 3.1: 音素ラベリング

3.3 音声パワー・ピッチ周波数の時間変化

「笑い声」に特有な音声パワー及びピッチ周波数の時間変化パターンを統計的に調べた。また、音声パワーやピッチ周波数といったパラメータは聴覚的にわかりやすい特徴であるので、パターン分けが「笑い声」の合成をシステム化する際のユーザインタフェース

の操作項目になり得るものとして検討している。

各パラメータの時間変化パターンの出現割合を、/ha/の連続回数別に分けた表 3.2、3.3 を以下に示す。変化パターンの種類は以下の通りである。ただし、音声パワーなら 3[dB]、ピッチ周波数なら 15[Hz] の差異が直前の音節との間にあったときに “変化した” と判断するものとする。

a 下り調子…全体的に減少変化する

b 上り調子…全体的に増加変化する

c 上って下る…増加変化の後、減少変化する

d 下って上る…減少変化の後、増加変化する

e 上下過変化…増加と減少を繰り返す

f ほとんど不変…大きな増減が無い

表 3.2: /ha/の連続回数と音声パワーの変化パターンの出現割合

		変化パターン					
		a	b	c	d	e	f
/ha/の 連続回数	2 回	62.7 %	3.4 %	0.0 %	0.0 %	0.0 %	33.9 %
	3 回	67.3 %	6.1 %	4.1 %	2.0 %	0.0 %	20.4 %
	4 回	76.0 %	4.0 %	4.0 %	0.0 %	0.0 %	16.0 %
	5 回〜	73.9 %	8.7 %	0.0 %	0.0 %	0.0 %	17.4 %

表 3.3: /ha/の連続回数とピッチ周波数の変化パターンの出現割合

		変化パターン					
		a	b	c	d	e	f
/ha/の 連続回数	2 回	44.1 %	35.6 %	0.0 %	0.0 %	0.0 %	20.3 %
	3 回	34.7 %	30.6 %	16.3 %	12.2 %	0.0 %	6.1 %
	4 回	32.0 %	12.0 %	12.0 %	20.0 %	8.0 %	16.0 %
	5 回〜	56.5 %	0.0 %	13.0 %	21.7 %	8.7 %	0.0 %

表 3.2 より、音声パワーは全体的に減少変化をするか、またはあまり変化がないことがわかる。また、発声音素数が多くなると、より下り調子のパターンが増える。下り調子の

生理的要因として、「笑い声」は呼気の流出が「通常発話」より多く、また、/ha/の発声回数が増えるほど肺の呼気圧が減少することが考えられる。数値でみても音声パワーは全体的な単調減少変化をしており、これは、話し声における語調成分の下降と一致している。

また、表3.3より、ピッチ周波数は発生音素数が多くなると下り調子になるパターンの割合が多くなる傾向がみられたが、全体としては変化パターンにはバラつきが見られ、様々なピッチ周波数の変化が現れるので、ピッチ周波数の変化パターンでの分類は有効であると考えられる。ユーザが合成音を作成する際、ピッチ周波数の変化パターンを操作項目として笑い声を作り分けることができると考えられる。

3.4 「笑い声」と「通常発話」の比較

「通常発話」の音声と比較することで、「笑い声」に特有な音響特徴を調べた。

3.4.1 ピッチ周波数の比較

「笑い声」と「通常発話」それぞれの発話におけるピッチ周波数の違いを有声部分である後続母音の音素/a/について調べた。以下に、発話者二名分のデータにおけるピッチ周波数の度数分布を図3.2に示す。

図3.2より、「笑い声」の方が全体的に50〜100Hzほどピッチ周波数が高いという傾向があることがわかる。また、「笑い声」は、この全体的にピッチ周波数が高いという傾向だけでなく、さらに周波数の高い400Hz付近にもピッチ周波数の分布が現れることがわかる。これは、笑いの発声過程で裏声が現れることに起因するものと考えられる。今回対象とした二話者に関して、この高い周波数帯における分布割合には明らかな違いが見られたが、これは笑い方に個人差があることによるものと考えられる。今後は、通常の笑い声と裏声による笑い声を区別していくことも考えている。

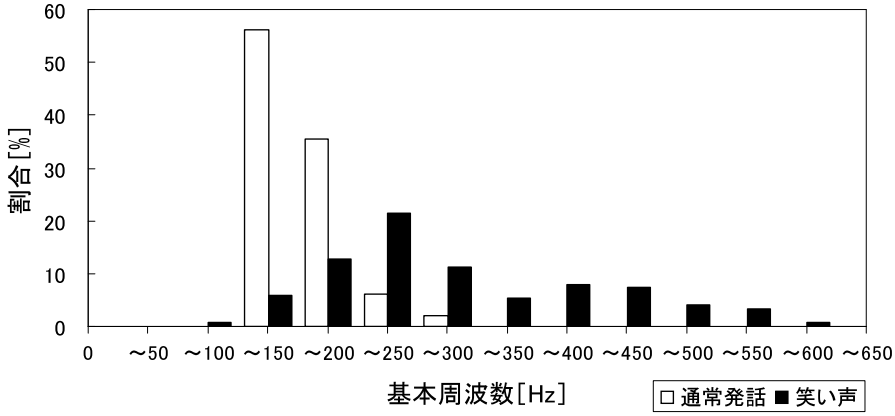

図 3.2: /a/のピッチ周波数分布

3.4.2 スペクトル包絡の比較

有声部分である後続母音の音素/a/について、「笑い声」と「通常発話」のスペクトル包絡の概形を視覚的に比較し、音色の違いを検討した。それぞれのスペクトル包絡を図3.3、3.4に示す。なお、「笑い声」と「通常発話」それぞれに関して、ピッチ周波数の違いによるスペクトル包絡の概形の差異が少なくなるように、ピッチ周波数が近いもの同士を比較した。

図3.4より、通常発話のスペクトルでは有声音であるために高調波構造がはっきり現れているが、図3.3の「笑い声」のスペクトルでは高調波構造が曖昧である。このように、「笑い声」のスペクトルは有声音ではなく無声音や雑音の特徴を示しているが、これは声道からの過剰な呼気の流出による影響が大きいためと考えられる。また、このことから「笑い声」は子音による影響が強く、「通常発話」と「笑い声」では音色が異なるということが考えられる。

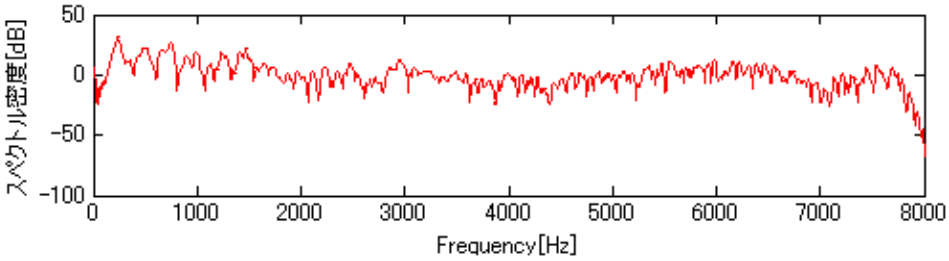

図 3.3: 笑い声の/a/のスペクトル

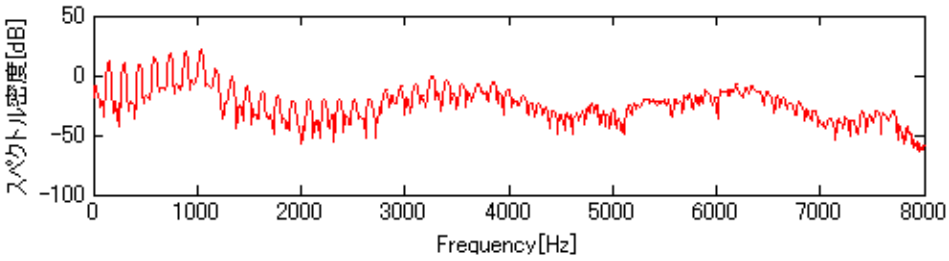

図 3.4: 通常発話の/a/のスペクトル

3.4.3 音素タイミングの比較

ラベリング情報を元にして音素タイミングを比較した。以下に、収集データから算出した Pause 長、音素/h/の音素持続時間、音素/a/の音素持続時間の度数分布を図 3.5、3.6、3.7 に示す。

図 3.5 より、Pause 長は、30[msec]〜110[msec] 程度である。「笑い声」でも「通常発話」でもほとんど同じ程度の長さがあるが、「笑い声」の方が「通常発話」よりも Pause 長が長いといえる。平均値を算出すると、「笑い声」における Pause 長は 77.4[msec]、「通常発話」における Pause 長は 68.4[msec] 程度であった。前述のとおり「笑い声」は呼気の流出が「通常発話」より多く、その分だけ呼気時間が長くなるため、この 9[msec] のわずかな差が生じていると考えられる。

図 3.6 より、音素/h/の音素持続時間は「笑い声」も「通常発話」も大きな差はないといえる。過剰な呼気の流出により無声化が強まり、子音である/h/の音素持続時間は長くなると考えていたが、前述のスペクトル包絡に対する検討を踏まえれば無声化の強まりは無声部分の音素持続時間に対してではなく、有声部分の音色に対して影響を与えていると

考えられる。

図 3.7 より、音素/a/の音素持続時間には「笑い声」と「通常発話」の間に大きな差があることがわかった。「笑い声」の方が「通常発話」よりも音素/a/の音素持続時間がかなり短いといえる。平均値を算出すると、「笑い声」における音素/a/の音素持続時間は 66.7[msec]、「通常発話」における音素/a/の音素持続時間は 112.6[msec] 程度であった。これは、「通常発話」は一つ一つの音素を明確に発声しているために後続母音の持続時間が長くなると考えられる。また、この差は有声部分のものであるため、聴覚的にも違いが感じられる要因になると考えられる。

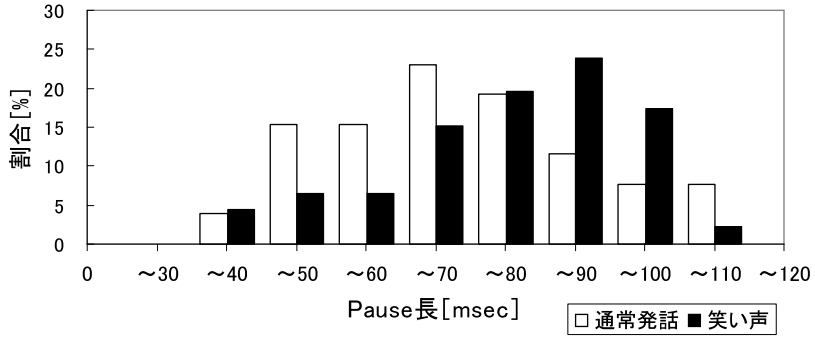

図 3.5: Pause 長の分布

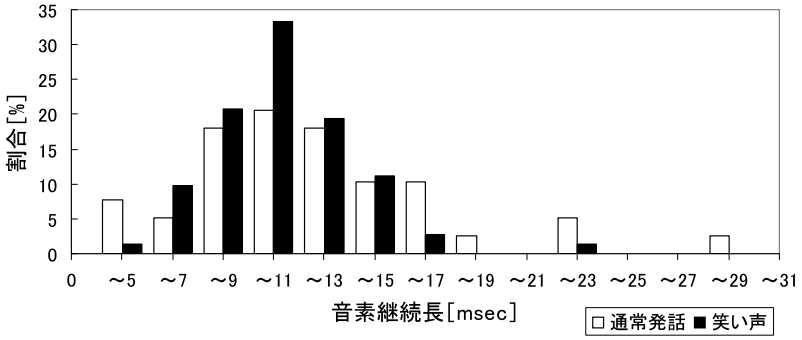

図 3.6: 音素/h/の音素持続時間の分布

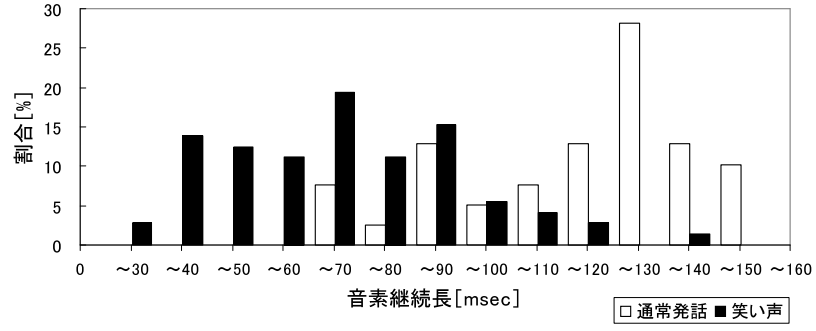

図 3.7: 音素/a/の音素持続時間の分布

第4章　笑い声らしさに対する貢献度の評価

4.1　概要

一部の音響パラメータを、単に “は” を3回発声した「通常発話」の音声の音響パラメータで差し替えて合成した「笑い声」の音声と、差し替えを行わない「笑い声」の音声との相違を主観的に比較することで、その音響パラメータの笑い声らしさに対する貢献度を調べた。分析・合成のパラメータとして、音源情報として音声パワー、ピッチ周波数、有声・無声判定情報、さらに声道フィルタ（調音）情報として16次のPARCOR係数によるスペクトル包絡情報を用いた。なお、分析の際のウィンドウサイズは30[msec]、シフトサイズは5[msec]である。

この主観評価実験の際に分析対象となるパラメータは、前述の音源情報3つと声道フィルタ情報1つ、そしてPauseの長さや音素/h/,/a/の音素持続時間に関する音素タイミングの情報を加え、以下の5つとなる。

- 音声パワー
- ピッチ周波数
- 有声・無声判定情報
- スペクトル包絡情報
- 音素タイミング

これらのパラメータについて、様々な差し替え方をした音声を用意する。具体的には、「笑い声」の各パラメータを「通常発話」のパラメータで順次差し替えて音声を合成する。全てのパラメータを差し替えた場合、「笑い声」は「通常発話」に変換されることになる。また、差し替えるのではなくデータを加工することによって対象の笑い声のデータに近づけたパターンも用意した。ここでいう加工とは、対象となる笑い声のデータのピッチ周波

数、音素持続時間の平均を算出し、それに合うように通常発話のデータのピッチ周波数を一様に上昇させる、または音素タイミングを一様に短くすることをさす。この加工データは、パラメータを一様に変化させたものであるので、ピッチ周波数の時間変化や音素出現の時間変化について笑い声のデータを踏襲していないことが差し替えデータとの相違点となる。また、「笑い声」と「通常発話」のパラメータを差し替える際の時間軸の対応付けは、図 4.1 に示すように、第 3.2 節のラベリングに基づいて音素区間を一致させて対応づけし、各音素区間内では時間軸を線形伸縮させた。

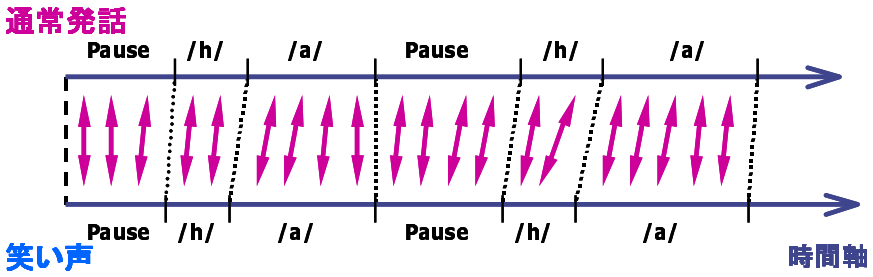

図 4.1: パラメータ差し替えのイメージ

4.2 主観評価実験

「笑い声」の笑い声らしさに対する各音響パラメータの貢献度を調べる評価実験を男子大学生 6 人を対象に行った。まず最も「笑い声」に近い音声と「通常発話」に近い音声を提示した後、作成した合成音について、「笑い声」と「通常発話」のどちらに近いかを 1〜5 点の 5 段階で評価させる。評点は、最も「笑い声」と判断される場合に 5 点、最も「通常発話」と判断される場合に 1 点とし、笑い声らしい音声ほど高い評点となる（図 4.2)。評価は、ピッチ周波数と音素タイミングの加工データを含めた 5 種類のパラメータ全ての組み合わせ計 72 種類に対して、基準となる笑い声を、笑い声 A、B2 種類用いたので合計 144 個の音声を対象としている。また、対象となる音声は「笑い声」のデータ、「通常発話」のデータともに “はははは” という 3 モーラのデータのみを用いた。

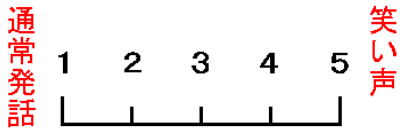

図 4.2: 評価スケール

4.3 結果と考察

評価実験の結果に対し、選ばれた音声の評点とパラメータの関係、選ばれた音声の評点とパラメータの組み合わせの関係の 2 点について以下に示す。

4.3.1 評点平均からみた結果

各音響パラメータにおいて、使用したデータと評点の平均について表にしたものが表 4.1 である。表中の各行は、それぞれの音響パラメータにおいて、左側が通常発話のデータを用いた合成音声の評点平均、中央が笑い声のデータを用いた合成音声の評点平均、そして右側が加工を施した合成音声の評点平均を示す。

表 4.1 より、ピッチ周波数を保存した音声は、それぞれ「笑い声」と「通常発話」に近いと判断されており、両者の違いを表すパラメータといえる。すなわち「笑い声」の笑い声らしさに対する貢献度が高いことがわかる。しかし、ピッチ周波数を加工したデータは実際の笑い声のデータよりもやや笑い声らしくないと判断されている。つまり、笑い声らしさにはピッチ周波数の時間的変化の影響があると考えられる。一方で、音素タイミングについては、笑い声のデータと加工したデータは近い評価値を得ており、時間的な変化の影響はないといえる。つまり、第 3.4 節でみられた Pause 長のわずかな差異は、聴覚的に大きな影響がでない程度の差異であるとわかる。また、表 4.1 から有声・無声判定情報は笑い声らしさに対する貢献度は非常に低いこともわかる。つまり、「笑い声」は過剰な呼気の流出により無声化が強まるが、そのために生成される無声音を聴取者が感じることは無いといえる。

表 4.1: 各パラメータの評価平均点

	使用したデータ		
	通常発話	笑い声	加工
音声パワー	2.94	3.28	—
ピッチ周波数	2.08	3.98	3.26
有声無声判定情報	3.08	3.13	—
スペクトル包絡情報	2.91	3.30	—
音素タイミング	2.38	3.48	3.46

4.3.2 組み合わせからみた結果

選ばれた音声における具体的なパラメータの組み合わせに着目する。4 以上の高い評価を受けた組み合わせについて表 4.2〜4.5 に示す。なお、表 4.2〜4.5 中の “●” は笑い声のデータ、“—” は通常発話のデータ、“▲” は加工したデータを用いていることを表す。表中のパラメータ名称は以下の通りである。

power 音声パワー
pitch ピッチ周波数
有声/無声 有声・無声判定情報
spectrum スペクトル包絡情報
timing 音素タイミング

笑い声らしさに対する基礎的な要素の検討

5 段階評価のうち 4 以上の評価を得られた音声は「笑い声」として認識されたとし、笑い声らしさに対する基礎的な要素を含んでいると考えられる。そこで、評価者 6 人全員が 4 以上の評価をした差し替えパターンに関する表 4.2、4.3 について検討する。

表 4.2 には、ピッチ周波数が低い通常発話のデータは選ばれていないことがわかる。このことから、ピッチ周波数が高いことが、「笑い声」として評価されるための最低限の要因となっていると考えられる。すなわち、ピッチ周波数は「笑い声」の笑い声らしさに対する貢献度が最も高いパラメータであり、笑い声らしさに対する基礎的なパラメータであ

るといえる。

また、表 4.2 のみにおいて笑い声 A、B に対する結果を比較すると、単純に音素タイミングが笑い声らしさに対する貢献に不可欠なパラメータであるとはいえないが、表 4.3 によれば、どのような笑い方であっても音素タイミングが短いほうが「笑い声」として人が認識しやすくなるといえる。つまり、音素タイミングも「笑い声」の笑い声らしさに対する貢献度が高いパラメータであるといえる。ここで、組み合わせに注目すると、表 4.3 の 1 行目と 2 行目、3 行目と 4 行目のそれぞれの組み合わせは、音素タイミングが笑い声のデータであるか加工データであるかの違いだけである。5 行目についても、表 4.2 の笑い声 B においては同様のペアをつくることができ、表 4.3 における 6 つの選出パターンのうち 5 つがこのようなペアをつくれるということになる。このことから、音素タイミングは笑い声に特有な時間変化は無いと考えられ、特別な時間変化を重視せず、単純に短くするだけで笑い声らしく聞こえることがわかる。

次に、音声パワーとスペクトル包絡情報は、表 4.2 において共に全体の 75%の組み合わせが笑い声のデータを使用している組み合わせである。表 4.3 においては全体に占める割合がさらに増え、83.3%になる。ピッチ周波数や音素タイミングに比べると「笑い声」の笑い声らしさに対する貢献度は低いようであるが、どのような笑い声であってもこのパラメータが笑い声のデータであることが少なからず笑い声らしさに影響を与えると考えられる。この点については、次項で検討していく。なお、使用した「通常発話」の音声は、最後の/ha/のみ音声パワーが減少する音声であった。これに対し、笑い声 A は同様の変化、笑い声 B は音素ごとに音声パワーが一様に減少していく単調減少変化をするものであった。表 4.2 において、笑い声 B よりも笑い声 A の方が音声パワーを通常発話のデータで差し替えている組み合わせが多く選出されているが、このことと変化パターンが関係していると考えられる。つまり、笑い方、ピッチ周波数の時間変化と音声パワーの時間変化に相関があるということが考えられる。

笑い声らしさを高める要素の検討

より本物の「笑い声」として人が認識するための要因を調べるため、5 段階評価で最高の評価を得た差し替えパターンに関する表 4.4、4.5 について検討する。

まず、表 4.4 からもピッチ周波数が高いことと音素タイミングが短いことは笑い声らしさにおいて重要な要素であることがわかる。表 4.5 によれば、この二つのパラメータは笑い声によらず重要なパラメータであることもわかる。前述のとおり、「笑い声」の笑い声らしさを作り出すための最も基礎的なパラメータがこの二つであると考えられるので、当然の結論といえる。但し、ピッチ周波数に関して、加工したデータは 5 人以上から最高評価を得られていないことがわかる。つまり、笑い声らしさを強めるためには高い周波数であることが基礎的な要因であるが、単に周波数が高いだけでは十分とはいえない。前項で検討した通り、「笑い声」には特有の時間変化パターンがあり、主観的により本物の「笑い声」として認識するにはこの時間変化パターンを考慮する必要があるといえる。一方で、前項で検討した結果に加え、表 4.5 からもタイミングの時間変化パターンを考慮する必要はないと考えられる。

次に、前項で、スペクトル包絡情報と音声パワーは笑い声らしさに対する貢献が少なからずあると考えたが、表 4.5 より、スペクトル包絡情報の方がその貢献度は高いといえる。表 4.4 によれば、笑い方によらず 5 人以上から最高評価を得られた音声はすべてスペクトル包絡情報に笑い声のデータを用いている音声であることがわかる。このことからも、より本物の「笑い声」として人が認識する合成音を生成するためにはスペクトル包絡情報を操作し、音色を笑い声特有のものにする事が有効であると考えられる。一方で、音声パワーは多少の貢献度はあるものの、スペクトル包絡情報ほど重要視する必要は無く、第 3.3 節で述べた語調成分の下降にのみ注意をすれば良いと考えられる。

表 4.2: 6 人全員が 4 以上の評価をつけた組み合わせ

	パラメータ				
	power	pitch	有声/無声	spectrum	timing
笑い声 A	●	●	●	●	●
	●	●	●	●	▲
	●	▲	●	●	▲
	●	▲	●	−	▲
	●	▲	●	−	●
	●	●	−	●	●
	●	●	−	●	▲
	●	▲	−	●	●
	−	●	●	●	▲
	−	●	−	●	●
	−	●	−	●	▲
	●	●	−	−	▲
	−	●	−	−	●
笑い声 B	●	●	●	●	●
	●	▲	●	●	●
	●	●	●	●	▲
	●	●	●	●	−
	●	●	●	−	●
	●	●	−	●	●
	●	●	−	●	▲
	−	●	●	●	●
	−	●	●	●	▲
	●	●	−	●	−
	●	●	−	−	▲

表 4.3: 6 人全員が笑い声A、B共通して 4 以上の評価をつけた組み合わせ

パラメータ				
power	pitch	有声/無声	spectrum	timing
●	●	●	●	●
●	●	●	●	▲
●	●	—	●	●
●	●	—	●	▲
—	●	●	●	▲
●	●	—	—	▲

表 4.4: 5 人以上が最高評価をつけた組み合わせ

	パラメータ				
	power	pitch	有声/無声	spectrum	timing
笑い声 A	●	●	●	●	▲
	●	●	—	●	●
	—	●	●	●	●
笑い声 B	●	●	●	●	▲
	●	●	—	●	●
	—	●	●	●	▲
	●	●	●	●	—
	●	●	—	●	—

表 4.5: 4 人以上が笑い声A、B共通して最高評価をつけた組み合わせ

パラメータ				
power	pitch	有声/無声	spectrum	timing
●	●	●	●	▲
●	●	—	●	●
—	●	●	●	●
—	●	●	●	▲

第5章　まとめと今後

5.1　本研究のまとめ

本研究では、任意の合成音を生成可能な規則合成方式による「笑い声」さらには「喋り笑い」の音声合成を目指している。そこで、「笑い声」の音声について、規則合成方式による音声合成に必要な音響特徴を調べていった。本研究では、「笑い声」に関して以下のことがわかった。

笑い声一般

はじめに、男子大学生7人分の自然対話の音声を収録し、分析対象データとして笑い声の音声を抽出した。抽出結果から、どの音素を基本とした笑い声を多く発声するかには個人差があることがわかったが、全体として/ha/を基本とする笑い声は出現頻度が高いことがわかった。また、その中で出現頻度の高い笑い声は、“はは” または “ははは” という2〜3モーラの笑い声であった。

音声パワー

笑い声の発声において、音声パワーの時間変化は下り調子の変化をするか、またはほとんど変化をしないかであり、多くは全体的な単調減少の時間変化をすることがわかった。これは、「笑い声」は呼気の流出が「通常発話」より多く、/ha/の発声回数が増えるほど肺の呼気圧が減少するためであり、発声音素数が多い場合はより下り調子の変化になりやすいことが原因と考えられる。

また、主観評価実験の結果によって、音声パワーのもつ「笑い声」の笑い声らしさに対する貢献度はそれほど高くないことがわかった。合成音を生成するときは、音声パワーの時間変化パターンが前述のように全体的な単調減少の変化となるように注意すれば十分

であるといえる。但し、ピッチ周波数の変化パターンとの相関にも着目する必要があると考えられる。

ピッチ周波数

「笑い声」と「通常発話」の/ha/の音についてピッチ周波数を比較したところ、「笑い声」は「通常発話」よりも全体的に50〜100Hzほどピッチ周波数が高く、話者によっては裏声がよく現れ、400Hz付近の音声になることがあるとわかった。話者による裏声の出現頻度の違いは、笑い方に個人差があるためと考えられる。

また、音声パワーの減少に伴い、発声音素数が多くなるとピッチ周波数も減少していくが、ピッチ周波数の時間変化には全体として様々な変化パターンが存在し、それらのパターンの出現頻度に大きな偏りは無かった。そこで、ユーザが合成音を作成する際、聴覚的にわかりやすいという点からもピッチ周波数の変化パターンを操作項目にして笑い声を作り分ける事が有効であると考えられる。しかし一方で、主観評価実験の結果から、変化パターンによって笑い声らしさに差が出るということがわかった。合成音を生成する際、より本物の「笑い声」として人に認識させるためには、笑い声特有の時間変化パターンを考慮する必要があると考えられる。

さらに、この主観評価実験の結果から、ピッチ周波数の低いものは「笑い声」として認識され難く、ピッチ周波数が高いことが笑い声らしさのための重大な要因であることがわかった。このことから、ピッチ周波数は「笑い声」の笑い声らしさに対する最も基礎的なパラメータであるといえる。

有声部・無声部の差異

「笑い声」は声道からの過剰な呼気の流出により無声化が強まるが、主観評価実験の結果によれば、そのために生じる無声音を聴取者が感じることは無いといえる。無声化の影響は、主に音色に影響していると考えられる。

スペクトル包絡・音色

「笑い声」と「通常発話」の/ha/の音について、有声部である後続母音の音素/a/のスペクトル包絡の概形を比較したところ、「笑い声」の場合は無声音や雑音のように高調波構造が曖昧になっていることがわかった。声道からの過剰な呼気の流出によって無声化が進み、子音の影響が強まることによって「通常発話」の音色とは異なる音色になると考えられる。

主観評価実験の結果から、このスペクトル包絡の構造の違いによる音色の違いが主観的にも感じられるものであるとわかった。合成音を生成する際、より本物の「笑い声」として人に認識させるためには、スペクトル包絡情報を操作することにより音色を「笑い声」特有の音色に近づける必要があると考えられる。

音素タイミング

定量的な分析により、「笑い声」の音声と「通常発話」の音声では、Pause の長さや音素/h/の音素持続時間に差はほとんどなく、音素/a/の音素持続時間に大きな差異があるとわかった。音素タイミングに関して、音素/a/の音素持続時間の差が「通常発話」との違いであるといえる。主観評価実験の結果により、この差が「笑い声」と「通常発話」の聞こえ方の違いに対する重大な要因になっていることがわかった。音素タイミングは「笑い声」の笑い声らしさに対する基礎的なパラメータの一つであるといえる。また、合成の際に音素タイミングの時間変化に注目する必要はなく、笑い声らしさを強調するためには単に音素タイミングを一様に短くすれば良いと考えられる。

5.2 今後について

研究全体としては、定量化した「笑い声」の特徴をもとに、いろいろな「笑い声」の自動生成システムを構築する方法を検討していく予定である。また、「喋り笑い」の分析も進めていく予定である。

目下、以下の点を解決していく。まず、スペクトル包絡や音色に関して、本稿での主観評価実験では/h/と/a/をまとめて扱っており、/h/と/a/ではどちらが笑い声らしさに対

する貢献度が高いパラメータであるかは不明な状態である。無声化の影響で子音部の/h/による貢献度が高いと考えられるが、音色が変わっている以上、後続母音の/a/を無視することはできないと考えられる。引き続き主観評価実験を行って、結論をだす予定である。さらに、どのような音色が「笑い声」の音色なのかも特定していく予定である。また、ピッチ周波数の時間変化パターンの違いによって笑い声らしさに差が出るということがわかったので、第 3.3 節で調べたピッチ周波数の時間変化パターンそれぞれについて評価実験を行い、有効な時間変化パターンを特定していく予定である。

参考文献

[1] 門谷 信愛希, 阿曽 弘具, 鈴木 基之, 牧野 正三: “音声に含まれる感情の判別に関する検討”, 信学技報,SP2000-82,2000.

[2] 武田 昌一, 西澤 良博, 大山 玄: “「怒り」の音声の特徴分析に関する 1 考察”, 信学技報,SP2000-164,2001.

[3] 飯田 朱美, ニック・キャンベル, 安村 通晃: “感情表現が可能な合成音声の作成と評価”, 情報処理学会論文誌,vol40,No2,1999.

[4] 大原 遼, 柏岡 秀紀, ニック・キャンベル: “対話音声の笑い声や笑い方についての分析”, 日本音響学会講演論文集,2004,9.

[5] 古井 貞熙: “電子情報通信工学シリーズ　音声情報処理”, 森北出版株式会社,1998.

[6] 古井 貞熙: “電子・情報工学入門シリーズ　音響・音声工学”, 近代科学社,1992.

[7] 鹿野 清宏, 中村 哲, 伊勢 史郎: “音声・音情報のディジタル信号処理”, 昭晃堂,1997.

[8] 勝木 保次 他: “聴覚と音声”, 電子通信学会,1974.

謝辞

本研究に当たり、白井克彦教授には、大学総長という非常に多忙な立場のなかで御時間を割き、ゼミ等でこまめに指導をして頂きました。白井教授から研究に対する考え方や鋭い指摘を伺うことで自分の研究を何度も深く考えさせられ、そのお蔭でここまで研究を進めることができました。心から感謝いたします。

また、音声班音響信号処理チームのゼミにおいて、右も左もわからない私に対して基礎技術の御指導や的確なアドバイス、研究の適切な方向性の示唆を与えて下さった誉田雅彰教授（早稲田大学スポーツ科学部)、樽松明教授（早稲田大学理工学総合研究センター)、大川茂樹助教授（千葉工業大学情報科学部)、金子格助教授（東京工芸大学工学部）には大変感謝しております。

そして、研究室配属当初から研究の進め方から研究室での身の振り方まで様々なことを教わり、随分とお世話になった諸先輩方の皆様に深く感謝いたします。特に博士後期過程2年の谷口さんを筆頭とする音声班の大久保さん、小林さん、椿さん、山本さん、塩崎さんにはプログラムや論文の書き方から研究の指針まで親身に教えていただき、深く感謝しております。また、画像班の皆様とは顔を合わせる機会は少なかったですが、全体ゼミでの貴重な意見や実験への協力など非常に感謝をしております。皆様、ありがとう御座いました。

さらに、自分達の学年は人数が少ないながらも共に卒業論文という大変な作業を頑張ってきたB4の彦坂君、松井君には感謝の気持ちでいっぱいです。来年以降も共に頑張っていきましょう。

最後に、一年間の浪人生活を経て、本大学への進学を理解し4年間の学業生活を支え、さらに大学院への進学にも理解を示して下さった両親に深く感謝いたします。

2005年2月

芳賀 寿昭