ヒューマンインタフェース 41−10
（１９９２．３．３）

音声合成における音声強調インタフェースの設計法

浜田 洋　千葉 仁一
ＮＴＴヒューマンインタフェース研究所

音声合成における強調の程度を変化可能な音声強調インタフェースの設計を例に、人間の認知特性に応じて複数の物理パラメータを制御するインタフェースの設計法を提案する。設計に先立ち、予め作成した文の対比較聴取実験に基づいて物理パラメータの強調表現能力分析を行った。この結果、複数の物理パラメータを制御することにより強調の程度を制御可能であることを確認した。次に、ＧＵＩを用いた音声コントローラにより強調音声の編集実験を行い、感覚と物理パラメータとの対応関係を抽出した。ＧＵＩによる音声コントローラによれば、音声に関する専門的知識を持たなくとも合成音声を制御することができる。編集実験の結果、本方法によれば短時間で装置の性能に適合したインタフェースの設計が可能であるとの見通しを得た。

Designing a Controller for Keyword Emphasis in a Text-to-Speech Synthesizer

Hiroshi HAMADA and Jin'ichi CHIBA
NTT Human Interface Laboratories
1-2356 Take, Yokosuka, Kanagawa 238-03 Japan

This paper proposes a design methodology that manipulates the physical parameters of a system by expressing system parameter controls in terms suitable for the operators cognitive ability. The example of controlling the emphasis level of a text-to-speech synthesizer is given. Preference experiments extract the characteristics of various keyword emphasis methods, and the results confirm that the emphasis level can be controlled by varying several physical speech parameters. Speech editing experiments performed with an interactive speech editor are used to extract the relation between emphasis level and physical speech parameters. The experiments confirm that the proposed design method can develop effective interface for the target system.

1．まえがき

ヒューマンインタフェースの設計において、人間の心理的な変数と装置やソフトウェアの物理的な変数との対応関係を求め、この結果をインタフェースに反映する必要がある。特に、音や音声の出力における感情、音質や、図形、映像などの出力における色、雰囲気など、視聴覚情報に対する人間の心理・感覚量と物理的な変数との対応を明確にすることが、マルチメディアのシステムやサービスのヒューマンインタフェースを向上する上では重要になってくる。

一般に人間の感覚は複数の物理量に対応している。そのため、専門的な知識を持たないユーザが自分の感覚に合わせて装置を制御するためには、複数の物理パラメータを調整しなければならない。例えば、オーディオアンプにおけるグラフィックイコライザなどはこれにあたる。もし、インタフェースが人間の感覚に対応したものとして提供されていれば、ユーザにとって操作性が格段に向上することは明らかである。本報告では、規則音声合成における音声強調を例に、感覚に対応して複数の物理パラメータを相互に連動させながら制御するインタフェースの設計方法を提案する。

規則音声合成は、電話を用いた情報照会・提供サービスや電話予約等の自動受付、ワードプロセッサにおける文章の読み合わせなどに利用されている。規則音声合成装置では、入力された日本語文を解析し読みに変換したのち、あらかじめ蓄えられた音声の素片を結合することにより音声を出力する。しかし、処理は自動で行われるため出力音声は平均的なもので、意図や感情などを反映したものではない。また、発話速度も一定で単調な音声であることが多い。一部の装置では、発話速度や抑揚などの制御機能が提供されているが、これらの機能を利用して所望の音声を作成することは、専門的な知識を持たないユーザにとっては困難である。しかし実際の利用時にはメッセージを伝える側の意図や強調を相手に伝えたい場合が多く、人間が発話する場合にはキーワードなどを強調して発声することを自然に行っている。

本報告では、グラフィカル・ユーザ・インタフェース(GUI)による音声コントローラを用いた音声編集実験を行い、強調の程度を可変制御することが可能な音声強調インタフェースを設計した結果について述べる。

2．音声強調インタフェースの設計

人間が強調して発話する場合の音響的特徴については、音声合成や言語学の立場から検討が行われている。人間がある部分を強調して伝える時、言語的に「～こそ」などの強調を意味する助詞を付与する、音響的に際だたせて発声する、などを行っている。音響的に際だたせて発声する場合、(1)音調を際だたせる、(2)強く発音する、(3)ゆっくり発音する、(4)その語の前に”間”を置く、(5)子音母音の調音を念入りに行う、などの手法によることが知られている[1]。

工学的な観点では、主に規則合成のための韻律生成規則の作成の研究が行われている。藤崎・廣瀬らは、談話条件を変化させた場合の基本周波数パタンを統語構造との関連で分析し、規則音声合成のための韻律生成規則として表現した[2]。また、白井らは２つの韻律語（局所的な基本周波数パタンの起伏に対応する語連鎖）からなる文章において、一方の韻律語を強調した場合の基本周波数パタンの変化を韻律語のアクセント型との関連で分析し、規則化した[3]。一方、武田らは、強調表現するために音声パワーや発話速度など複数の韻律特徴量の制御の規則化を行った結果を報告している[4]。以上の例はいずれも自然音声の分析と作成した合成音の評価により行われている。また、これらの例では強調をする／しないの何れかが実現可能である。しかしながら人間が発話する場合には、強めの強調、弱めの強調、などレベルを変えながら行っている。そこで、強調のレベルを変化することが可能な、音声合成における音声強調インタフェースの設計を行った。

図１に目標とする音声強調インタフェースの例

を示す。ユーザが自分の意図を表現するために、ある特定部分を強調する場合、まず、強調スイッチをONにし、次に自分の望む強さとなるように強調のレベルをボリュームにより調整する。

このようなインタフェースを設計するためには、強調という感覚的な尺度と音声合成における種々の物理的パラメータとの対応を求める必要がある。従来この種のインタフェースの設計には分析的なアプローチが中心に行われてきた。しかしながら分析的な手法は、(1)用いる装置の性能や機能を十分反映したものとなり難い、(2)被験者への負荷が大きく、また、被験者にとって受動的な実験となりがちなため積極的な姿勢をひきだし難い、(3)解析に専門的な知識が必要である、ために設計に多くの時間を要する。

今回提案する設計法では、音声に関する専門的な知識を必要としないＧＵＩによる音声コントローラを用いて被験者に望ましいと考える音声を作成して貰うことにより人間の感覚と物理パラメータとの対応を求める。従って、対象とする装置の機能・性能を最大限活かして感覚とパラメータの対応を求めることが可能である。実験ではまず、パラメータを変化させた音声の比較聴取実験により強調レベルを変えた合成音の作成が可能であることを確認し、次に音声編集実験により複数のパラメータと感覚との対応を求める。

3．規則音声合成装置の概要

実験では市販の規則音声合成装置を用いた。実験に使用した規則音声合成装置「しゃべりん坊ＨＧ」(NTTｲﾝﾃﾘｼﾞｪﾝﾄﾃｸﾉﾛｼﾞ社)は、ＣＯＣ音声合成方式を１ボードで実現したもので市販のパーソナルコンピュータに組み込んで使用する[5]。この音声合成装置では、自然音声中に表れる音声のバリエーションを表現するための音声素片を統計的手法により自動生成しており、滑らかで自然な音声が合成可能である。

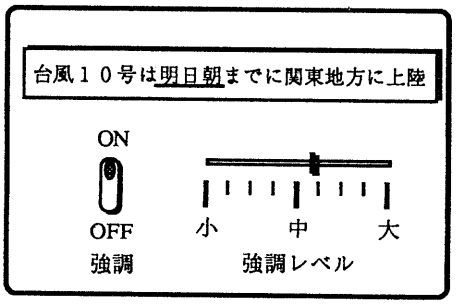

図１　規則合成における音声強調インタフェースの例

本装置は、パソコンから出力された日本文を読みの系列に変換し、アクセント、ポーズ等を付与した後、最適な音声素片を選択、結合し出力する。また、ユーザが制御コードを出力日本文中に挿入することにより、(1)音量16段階、(2)発話速度8段階、(3)基本周波数の高低（話調成分の高さ）5段階、(4)抑揚（基本周波数のダイナミックレンジ）5段階、(5)信号音出力5種類、(6)男／女声切替、などの制御を行うことができる[6]。今回はこれらの制御機能を用いて音声強調インタフェースの設計を行う。

4．聴取実験による物理パラメータの特徴分析[7]

4.1　実験に用いた文章

実験に用いた文章は、ニュース、道路交通状況案内等実際のテレホンサービスから抽出選択した単文７種類であり、各文の特定の文節を強調対象とした。強調対象としては、特定の名詞（キーワード）と形容詞とがあるが、今回の実験ではキーワードを強調の対象とした（例：アメリカのブッシュ大統領は、以前から続いている南アフリカへの経済制裁解除を発表しました）。

4.2　強調方法

強調の方法として規則合成装置の音声出力制御機能のうち以下の４通りを採用した。

(1)当該文節をゆっくり発話
(2)当該文節の前後にポーズを挿入
(3)当該文節の音量を増大
(4)当該文節の基本周波数を上昇

図２に強調処理を施さない音声、および、それぞ

れの処理を施した音声の波形、対数パワー、基本周波数パタンの例を示す。

各制御パラメータの変化幅は、処理を施すことによる差を知覚できる最小値を予備実験により求め、ポーズ:前後に約800msecのポーズ挿入、基本周波数:最も高くなる部分で約15%上昇、音量:6dB増加、発話速度:0.6倍とした。合成ボードで設定できる値は離散的であるため、これらの値は人間が弁別できる閾値とは異なる。なお、強調部分を除く全体の発話速度は4.6ﾓｰﾗ/秒である。

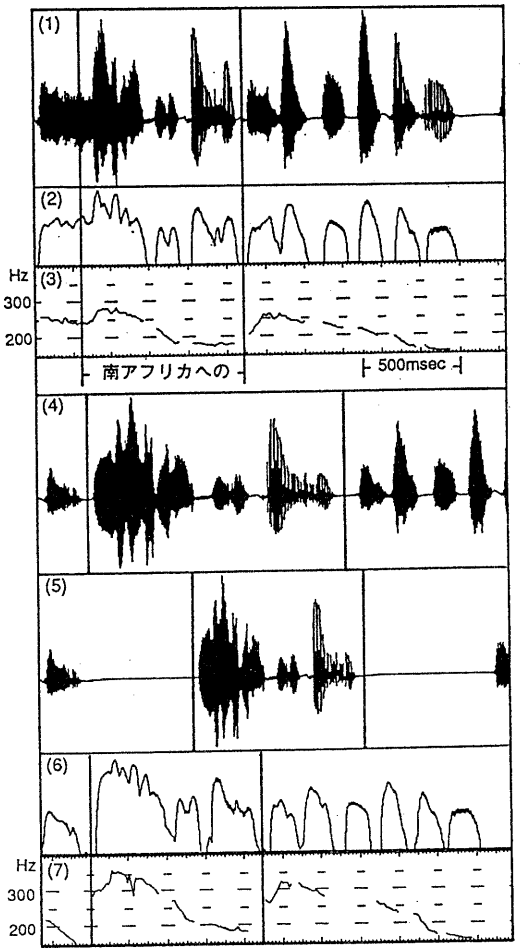

図2　強調処理の例
「…ている南アフリカへの経済制裁解除を…」
(1)～(3)強調処理を施さない場合の音声波形、対数パワー、基本周波数パタン
(4)ゆっくり発話した場合の音声波形
(5)前後にポーズを挿入した場合の音声波形
(6)音量を大きくした場合の対数パワー
(7)基本周波数を高くした場合の周波数パタン

4.3　実験方法

対象とした各文章に対して4.2で示した処理の何れか一つの方法により強調した試験音声2種を被験者に呈示し「強く強調されていると感じた」方を選択させる対比較試験を行った。被験者は規則合成音を聞き慣れていない成人男女20名で、実験に先立ち実験の趣旨について説明を行い、「自然性の評価でなく、強調の度合について評価する」ことを指示した。

さらに、単独の処理で効果の大きかった発話速度、音量、基本周波数のうち2種を組み合わせた場合について対比較試験を行った。被験者、文章等は単独の処理による場合と同じで、7文章に対して、(1)速度＋音量、(2)速度＋基本周波数、(3)音量＋基本周波数、の3種の組み合せに、(4)速度、(5)音量、を加えた計5種の強調方法により処理を施した音声を用いて対比較試験を行った。

なお、実験は女声の合成音で行い、試験音声の呈示には全てヘッドホンを用いた。

4.4　実験結果と考察

単独の処理により強調した場合の対比較実験結果から全文章、および、文章毎に心理尺度値を求めた結果を図3に示す。図3で値が大きいほど強い強調である。平均的に見ると、「ゆっくり発話する（発話速度）」場合に最も強い強調効果が得られ、続いて「音量を大きくする（音量）」、「基本周波数を高くする（F0）」、「前後にポーズを挿入する（ポーズ）」の順に強調効果が減少している。この結果を被験者毎、文章毎に見ても、心理尺度上での間隔には差があるものの、順序は一定している。

今回の実験では7種の文章に対して4種の方法によりキーワードの強調を行った。被験者が強く強調されていると判断した順序は、文章6で基本周波数と音量の順序が入れ替わっている以外一定であったが、各処理間の効果は文章毎に異なる。この理由としては、各パラメータの制御は離散的にしか行えないため、アクセントの位置、イントネーションとの関連からピッチや音量制御の結果

出力された音声の知覚的な効果に差があることが上げられる。

ポーズ挿入による強調の効果は、ポーズ挿入位置の構文的な結合の強さと関連があると考えられる。言い替えれば、読点を挿入しても不自然でない位置ほどポーズ挿入の効果は小さい。今回の実験では、キーワードを含む文節を強調対象としたため、ポーズ挿入の効果が小さかったとも考えられる。また、今回は該当する文節の前後に同じ長さのポーズを挿入したが、前後のポーズ長は通常異なっている。

武田らの結果[8]とは、ポーズ挿入、発話速度低減の強調表現能力に関して傾向が異なっている。これは、構文や文の長さ、強調対象となる部分の長さの差によるものと考えられる。

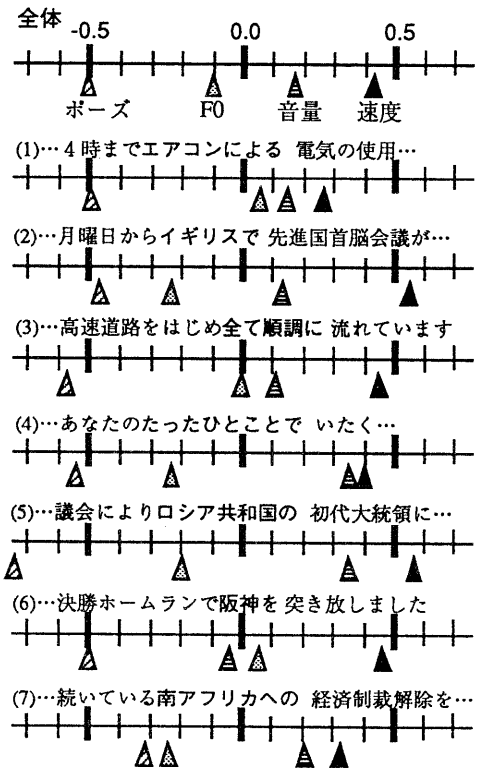

図3　単独の強調処理による対比較実験結果

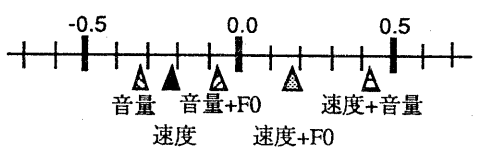

図4　処理の組み合せによる対比較実験結果

次に、処理を組み合わせた場合の対比較試験結果から求めた心理尺度値（全文章の平均）を図4に示す。単独処理の場合と比較して心理尺度上の差は小さい。また、被験者によるばらつきは単独処理の試験結果と比較して大きかった。しかし、文章毎に結果を見ると、実験1と同様文章6でピッチと音量の順序が入れ替わっている以外強い強調と感じられる順序は一定であった。

組み合せによる強調処理を行った場合の強調効果の順序は単独強調処理の場合の結果と対応している。すなわち、最も強調効果が大きかった”速度”と次に強調効果が大きかった”音量”との組み合せで、強調効果が最も大きくなる。また、処理を組み合わせることにより、単独の処理を行った場合よりも強調効果が大きくなる。

以上の結果から、種々の物理パラメータを変化させることにより、強調のレベルを変化させた合成音声を実現可能であるとの見通しを得た。

5．GUIによる音声コントローラを用いた音声強調インタフェースの設計

5.1　GUIによる音声コントローラ

実際に人間の発話において強調する場合、複数の物理パラメータの組み合せにより強調を行っている。また、予め作成した音声の聴取実験では被験者が真に望んでいる合成音を設計することは困難であり、対象としている音声合成装置の機能、性能に適合しているいるとは言い難い。そこで次に、被験者に自然性などを考慮して自分の考えた好ましい音を作成させることにより、音声強調インタフェースの設計を行った。この方法によれば、装置の性能・制約に合った合成音を作成でき、また、聴取実験では無視していた自然性も考慮することができる。

被験者は音声処理の専門家ではないため、専門的な知識を意識させずに種々のパラメータを簡便

に制御するための制御ソフトウェアが重要となる。さらに、最終的には音として評価されるべきであるため常に音を聞きながら音声合成装置の制御パラメータを決定する必要がある。そこで、ＧＵＩにより規則合成音の制御を簡単に行うことができる対話型ソフトウェアを作成した。

感覚の変化は、物理パラメータの相対的な変化により与えられるものである。すなわち、ある特定部分を強調することは、当該部分のパラメータを強調する方向により変化することのみでなく、当該部分以外のパラメータを強調する方向とは反対に変化させることによっても実現可能である。これらを考慮して音声コントローラ（ソフトウェア）を作成した。この音声コントローラは以下の機能を持つ。

(1)強調部分の発話速度、音量の大小、基本周波数の高低、強調部分前後のポーズ長の制御

(2)強調部分を除いた文章全体の発話速度、音量、基本周波数の制御

(3)作成した音声、および、比較のための強調を施していない音声の合成出力

図５にグラフィックインタフェースによる音声コントローラの画面を示す。画面全体は、３段階の強調レベル（強め、普通、弱め）用に３分割されており、それぞれのレベルに対して前述したパラメータをマウス操作により変更することができる。速度、音量、基本周波数については、強調部分とそれ以外の部分に対するパラメータを２次元のグラフ上で変化させることができる（横軸方向：強調部分以外の文章全体のパラメータ値、たて軸方向：強調部分のパラメータ値）。また、編集中にマウスのボタンをクリックすることによりいつでも作成した音声を聞くことができる。さらに、画面下の「標準音声出力」部をクリックすることにより強調処理を施していない音声を比較のために聞くことができる。

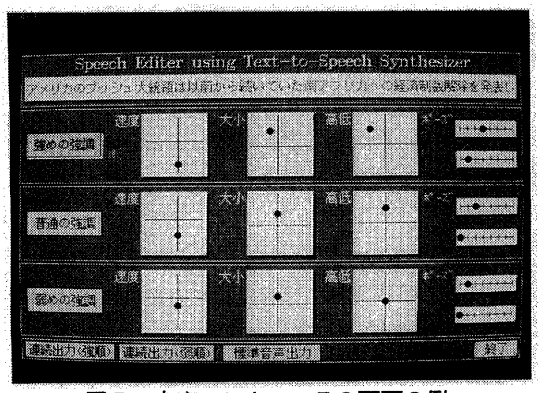

図５　音声コントローラの画面の例

5.2 実験方法

被験者は、音声コントローラを用いて強め、普通、弱めの３種類の強調音声を作成する。実験手順は以下の通りである。

(1)操作方法を修得するため、簡単な文を用いて１種類の強調音声を作成する。この時は画面には１種類の音声の編集部分（図５の「普通」に相当する部分）のみが表示される。

(2)次に、図５の画面を用いてレベルを変えた強調音声を作成する。このとき、初めに普通の強さの強調を作成し、次に普通の強調よりも強めと弱めの強調音声を作成するよう指示した。なお、途中で普通の強さの強調のパラメータを変更することも許した。

(3)半日以上の間隔をおいて、別の文章で(2)を繰り返し、さらに、半日以上の間隔をおいて別の文章で(2)を繰り返す。ひとりの被験者は合計３回の編集を行う。

通常編集を行う際には左側のパラメータから制御するため、画面の中でのパラメータの並び（図５では左から速度-音量-基本周波数-ポーズ）が結果に影響を与えることも考えられる。そこで、画面の中でのパラメータの順序を、聴取実験により強調効果が大きいとされた順（図５）と、その反対との２種類を作成し実験毎に変えて行った。

被験者は聴取実験の被験者20名中の12名、対象文は聴取実験に用いた7文中の3文である。

5.3 実験結果と考察

実験によって得られたパラメータの値を文章毎に平均し図６に示す。発話速度は強調部分と強調部分を除く文章全体との平均発話速度の比で、基本周波数は最大周波数の比、音量は差(dB)で示している。図６から以下のことが言える。

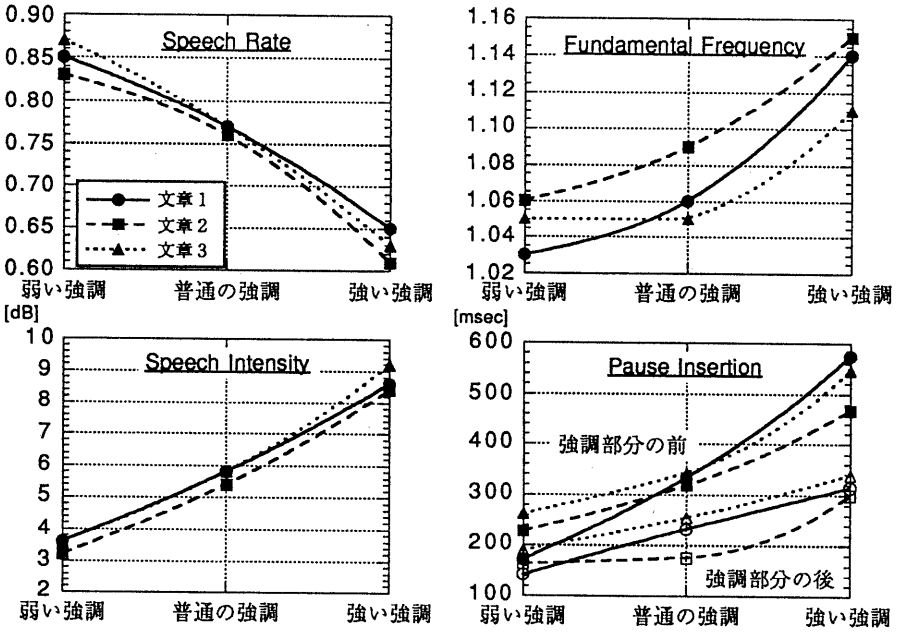

図6　音声コントローラを用いた音声編集実験結果

(1)音量(dB)は弱い-普通-強いの順に約2.5dBずつ等間隔で増加している。この傾向に文章の種類による差は見られない。

(2)発話速度の低減(係数)も文章による差が無く、「弱い」と「普通」の間の差よりも「普通」と「強い」の間の差が大きい。これは、対比較実験で発話速度の強調能力が大きかったこととも対応しており、強い強調を表現する場合に発話速度を低減することの効果が確認された。

(3)基本周波数上昇（係数）とポーズ時間長は、文章により値が異なる。この理由として、基本周波数上昇は強調対象や前後の語のアクセントが、ポーズ長は強調部分や文全体の長さが影響していることがあげられる。また、実験に用いた音声合成装置のパラメータの制御ステップが大きいことも要因と考えられる。

(4)強調部分の前のポーズ長は、後のポーズ長の1.5～2倍が適当である。

図6に示した値は、文章毎の平均であるが、パラメータ値は被験者によって異なっており、分散は大きい。しかし、編集画面におけるパラメータの並びの差による結果への影響はなかった。

先に述べたように、強調効果は強調対象部分とそれ以外の部分とのパラメータの相対的関係により得られる。強調部分とそれ以外の部分の物理パラメータの関係を基本周波数の場合を例に図7に示す。図7では、標準値（強調処理をしない場合）の基本周波数を1とした場合の値で示してある。この図から、強調のために必要な差を確保しつつ自然性への悪影響を避けるために、強い強調の場合には強調部分の基本周波数の値を高くする

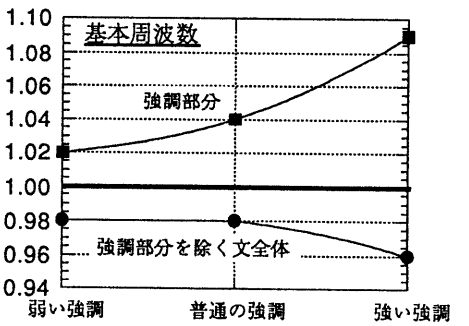

図7　強調部分と文全体との関連
(強調していない音声の値を1とした場合)

と共にそれ以外の部分の基本周波数を低くしていることがわかる。音量の場合も基本周波数の場合ほど顕著ではないが、同様な傾向が見られた。

実験後に実施した被験者へのアンケートでは、強調設定に際して重視したパラメータは、速度、音量がほぼ同数で多く、続いて、基本周波数、前ポーズ、後ポーズの順であるとの回答が得られた。また、他に制御したい項目として、強調部分の中でのパラメータの制御（なめらかな変化を作りたい）、発話速度、基本周波数、ポーズ長のより精密な設定、などが挙げられた。また、聴取による対比較試験と比較して被験者は積極的であり、本方法がインタフェース設計の一方法として有効であるとの見通しを得た。

以上の実験で得られた結果を図１に示した音声強調インタフェースの形で適用すれば、人間の感覚に合致して複数の物理パラメータを制御することができ、結果として音声の知識を有していない利用者でも容易に規則音声合成の制御を実施することが可能となる。

６．まとめ

規則音声合成における強調レベルを変化させることが可能な音声強調インタフェースの設計を例に、感覚に対応して複数の物理パラメータを制御するインタフェースの設計法を提案した。

設計に先立ち、対比較聴取実験により物理パラメータの強調表現力を分析した。その結果、対象文節の発話速度を遅くする方法により最も強い強調が得られ、続いて、当該文節の音量を大きくする、当該文節の基本周波数を高くする、当該文節の前後にポーズを挿入する、の順に効果が減少することが判った。また、複数の処理を組み合わせることにより強調効果が増加する。この結果、複数パラメータを制御することにより強さを変えた強調音声の出力が可能であることを確認した。

つぎに、ＧＵＩによる音声コントローラを用いた強調音声の作成実験を行い、強調と複数の物理パラメータ間の対応を求め、強調インタフェースを作成した。グラフィックインタフェースを用いることにより専門的な知識を持たない被験者でも合成音声を制御することが可能である。さらに、本方法によれば被験者の積極性を引き出すことが可能であり、短時間で対象とする装置に適合したインタフェース設計が実現できる見通しを得た。

今後、他の品詞を強調の対象とした場合や、言語構造との関連の分析を行うと共に、作成した音声強調インタフェースの評価を行う予定である。

【謝辞】　被験者として協力して頂いた皆様に深謝します。また、実験方法に関して討論して頂いたマルチメディア処理研究部・徳永主幹研究員、小川主幹研究員、音声合成装置に関してご協力頂いた音声情報研究部・広川主幹研究員、NTTｲﾝﾃﾘｼﾞｪﾝﾄﾃｸﾉﾛｼﾞ(株)・箱田課長に感謝します。

【参考文献】

[1]和田實：アクセント　イントネーション　プロミネンス、徳川宗賢編、「アクセント」（論集日本語研究２），有精堂（昭和55年）

[2]廣瀬啓吉，藤崎博也：音声合成とアクセント・イントネーション、信学誌，Vol.70, No.4 (1987)

[3]白井克彦，岩田和彦：音声合成のための単語の強調表現の規則化、信学論(A), Vol.J70-A, pp.816-821 (1987)

[4]武田昌一，市川薫：日本語文音声のプロミネンス生成規則の作成と評価、音学誌，Vo.47, No.6, pp.397-404 (1991)

[5]箱田和雄，広川智久，水野秀之，中嶌信弥：COC法を用いたテキスト合成ボードの試作、信学会音声研究会資料 SP90-55 (1990)

[6]NTTインテリジェントテクノロジ(株)：しゃべりん坊ＨＧ操作マニュアル (1991)

[7]浜田洋，千葉仁一：音声メッセージ出力における強調方法とその効果、第7回HIｼﾝﾎﾟｼﾞｳﾑ論文集，pp.129-132 (1991)

[8]武田昌一，市川薫：日本語文規則合成音声におけるプロミネンス表現力の改善検討、音学会講論集，2-6-19 (1991.10)