TR-IT-0209

文音声におけるプロミネンスの抽出とその合成法について

星野 慎一　　Nick Campbell

1997.3.19

概要

音声対話システム性能向上のための重要な要因として、発話アクトや発話意図を入出力双方の分野で制御できるようにすることが挙げられる。そこでこれらの情報を扱う研究の一環として、入力分野では文音声においてのプロミネンス位置の推定、出力分野では出力イントネーションの補正方法についての研究を行った。

もくじ

1 はじめに 4

1.1 背景 4

1.2 目的 5

2 プロミネンスの抽出 6

2.1 抽出の基本方針 6

2.2 手順 6

2.3 アクセント型のプロミネンスに与える影響 8

2.4 プロミネンス抽出手段 9

2.4.1 手段1 9

2.4.2 手段2 10

2.4.3 手段3 11

2.4.4 手段4 12

2.5 認識結果 13

2.6 考察 14

2.6.1 手段間の違い 14

2.6.2 手段間の相関 16

2.6.3 プロミネンスの表現方法 16

2.7 プロミネンスの存在度数表現 17

2.8 存在度数表現を用いた実験 18

2.8.1 実験の基本方針 18
2.8.2 手順 18
2.9 実験結果 19
2.9.1 テキストのみを入力とした時の合成音声 20
2.9.2 プロミネンスを含まない文音声を用いた実行結果 20
2.9.3 第一文節にプロミネンスを含む文音声を用いた実行結果 21
2.9.4 第三文節にプロミネンスを含む文音声を用いた実行結果 22
2.10 実験の考察 23

3 イントネーション・エディタ **24**
3.1 基本方針 24
3.2 手順 24
3.2.1 話者性変換 25
3.2.2 イントネーション変換 25
3.3 システム環境による制約 26
3.4 考察 27

4 おわりに **28**
4.1 まとめ 28
4.2 今後の展望 28

第 1 章

はじめに

1.1 背景

音声には単語のアクセントや音声全体に渡るイントネーション機構など、文の内容によってある程度決定されてしまう要素と、プロミネンスやフォーカスといった、実際に発声されるまで決定されない要素がある。音声における後者の要素は発話アクトや発話意図と呼ばれる [1]。音声合成の分野において、発話アクト・意図を考慮した音声を合成することは、より良い合成システムを構築する上で必要な要因であると言える。

例えば chatr のような音声対話システムでは、これらの情報を扱えることにより、翻訳において「プロミネンスを含む音声を他の言語においてもプロミネンスを含めた形で出力する」(図 1.1) などといった操作が可能であると考えられる。

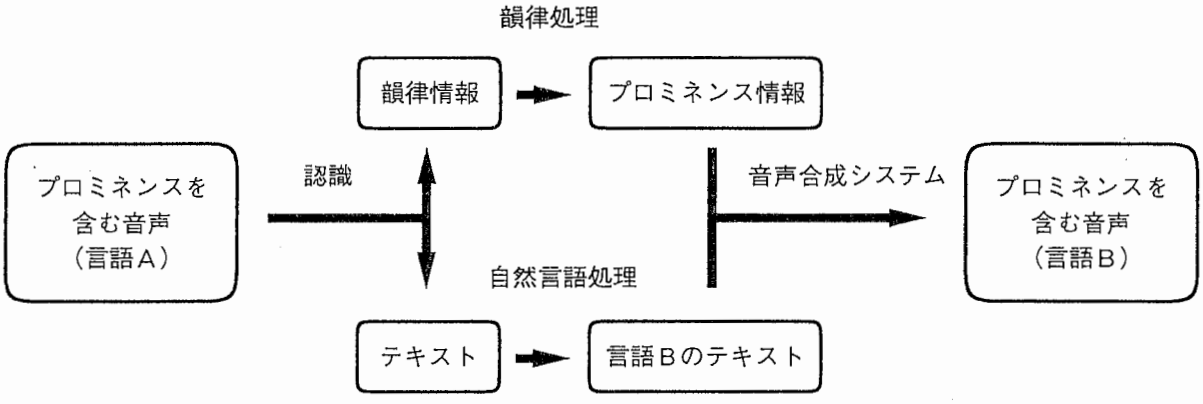


図 1.1: プロミネンスを含む音声の翻訳

そこで発話アクトなどの情報を扱う一環として、文音声におけるプロミネンスを音声対話システ

ムでの入出力双方で制御する方法についての研究を行おうと試みた。

1.2 目的

今回の研究の目的は大まかな分野として、音声対話システムの入力についての部分と出力についての部分に分けられる。それぞれ以下に示すような内容を行った。

- システムへの入力の制御

  文音声におけるプロミネンスの位置の推定

- システムからの出力の制御

  音声対話システムからの出力音声のイントネーション補正

前者は文音声中のプロミネンスの位置を韻律処理によって推定し、なおかつその情報を対話システム用にコード化することまでを目標としている。この分野については「プロミネンスの抽出」の章で詳しく述べていく。

後者は音声対話システムの合成音声に対し、イントネーションが気に入らない場合などにモデル音声を用意することにより、イントネーションの補正を行うことなどを目標としている。この分野については「イントネーション・エディタ」の章で詳しく述べていく。

第 2 章

プロミネンスの抽出

ここでは、プロミネンスを含む文音声を入力として、どこにプロミネンスが含まれているかを認識する作業を行い、最終的に情報の音声対話システムへの入力を前提としたコード化を目標とする。

2.1 抽出の基本方針

文音声からプロミネンスを抽出しようとする際、プロミネンスが含まれている部分をどのように区分化するかという問題がある。実際、一般的にプロミネンスが含まれているとみなされる区分は、文節単位であったり、助詞や名詞のみであったり、2 箇所に分断されていたりと様々である。

このうち今回は区分を文節単位という形に限定し、抽出の際も文中のどの文節にプロミネンスが含まれるかという判断の仕方を行うことにする。

2.2 手順

プロミネンス抽出の手順を簡単に示すと以下のようになる (図 2.1)。

1. 入力としてプロミネンスを含む文音声を用意する
2. 1 の音声に信号処理を行い、音声波形の基本周波数値などの韻律情報を得る
3. 韻律情報を分析することにより、プロミネンスに関する情報を得る
4. プロミネンスに関する情報をコード化し、音声対話システムに入力として渡せる形に変形させる

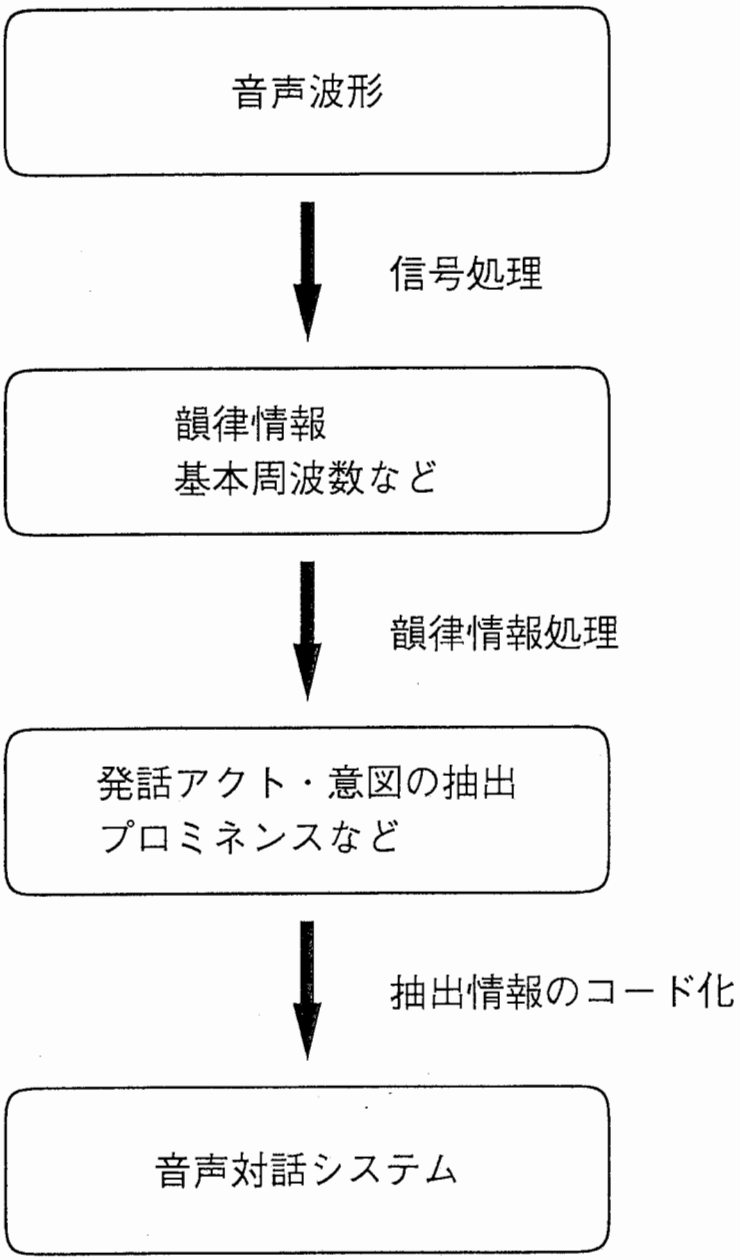


図 2.1: プロミネンス抽出の手順

2.3 アクセント型のプロミネンスに与える影響

抽出手段の解説に入る前に日本語の文節音声のアクセント型とプロミネンスの関係について述べる。アクセント型とは音声発声時の声の高さ、すなわちアクセントの高低の推移規則を表したもので、種類は文節のモーラ数に等しいだけ存在すると言われている。例えば文節モーラ数が5である文節はアクセント型の種類も5種類存在する (図2.2)[2]。日本語の文節音声にはアクセント型というものが必ず存在しており、形は文節中に含まれる単語に固有のものである。

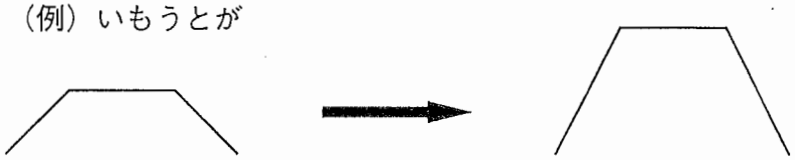
お花見が	妹が	湖が	飲み物が	三月が
おはなみが	いもうとが	みずうみが	のみものが	さんがつが

図 2.2: 文節モーラ数5のアクセント型一覧

実際にアクセントを含む音声を何種類か聞いてみた結果、アクセント型が異なる場合、プロミネンスのパターンもそれにつれて異なっているように感じられた。その傾向は大きく2種類のパターンに分けられる。一つは発声単位全体に渡ってアクセント推移幅が大きくなるもの、もう一つは発話単位の後半部にアクセント推移の上昇が見られるものである (図2.3)。

type 1　発声単位全体にわたりアクセント推移幅が大きくなる

（例）いもうとが



type 2　発声単位の後半部にアクセント推移の上昇が見られる

（例）さんがつが

図 2.3: プロミネンスにおけるアクセント推移の例

ここでは2種類のパターンを挙げたが、その他にも個人差、地域差などからさらに多くの種類が存在するだろう。よってプロミネンスの抽出手段についても、一つの手段で全てのパターンをカバーし得る方法というものは存在しないように思われる。そこで、次節では何通りかのプロミネンス抽出手段の例を挙げ、それぞれの認識率を調べてみたい。

2.4 プロミネンス抽出手段

プロミネンス抽出の手段として主に用いた韻律情報は、声の高さを表現するといわれている基本周波数値である。ここで言う抽出とは、基本周波数値の推移を何らかの基準を設けて処理することにより、プロミネンスが最も含まれているであろう確率が高い箇所を特定化するものである。ちなみに、2.1節でも述べた通り、特定化する箇所の単位としては文節を用いた。

以下では4通りのプロミネンス抽出のための手段を挙げる。

2.4.1 手段1

- 基本周波数推移の最大値をプロミネンスの位置とするもの (図2.4)

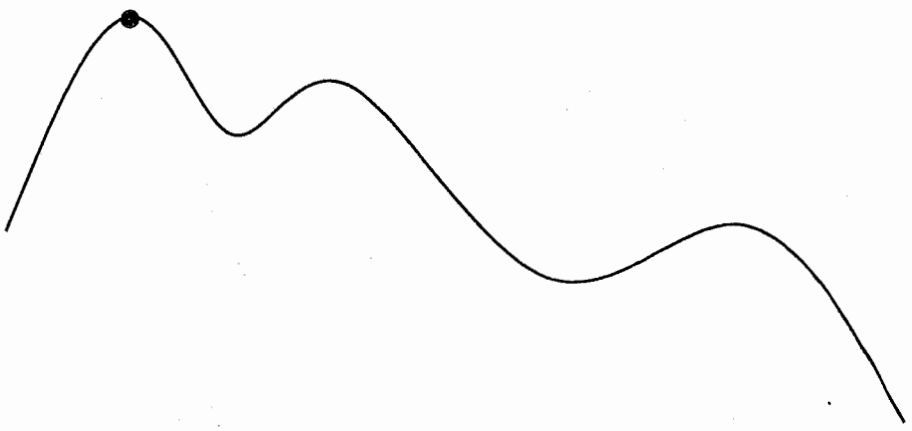


図2.4: 手段1による抽出の方法

プロミネンスを含む音声を聞くにつけまず初めに気づくことが、プロミネンスの部分では基本周波数値が高くなっている傾向がある、ということである。プロミネンスのほとんどが、前節でも述べた通りの発話区間全体でアクセント推移が大きくなっているものであるので、その傾向が強いと言えよう。

2.4.2 手段 2

- 基本周波数推移の直線近似を行い、直線からの距離の正の向きでの最大値をプロミネンスの位置とするもの (図 2.5)

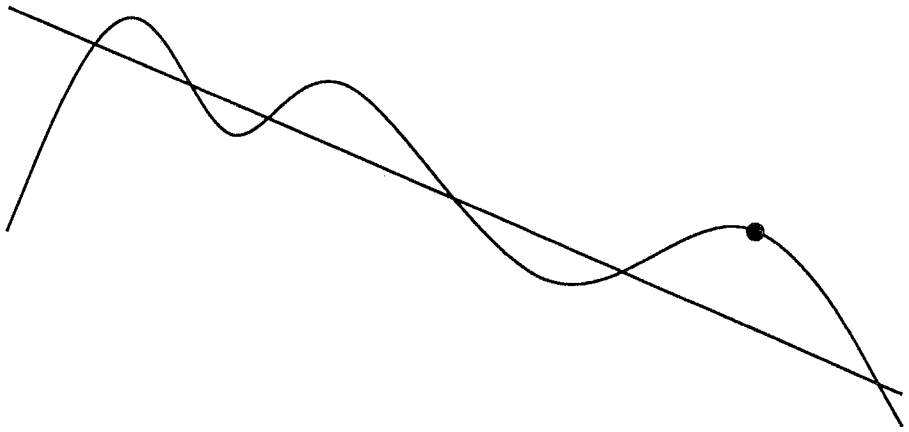


図 2.5: 手段 2 による抽出の方法

文音声における基本周波数推移は大局的に見ると、時間とともに下降していくという動きをすると言われている [3]。従って文頭の文節と文末の文節では、プロミネンスを含まない状態でもかなりの基本周波数値の差が見られる。

よってその差を解消するために基本周波数推移曲線を直線近似し、直線の上部にある点のうち直線からの距離が最も長い位置の点にプロミネンスが存在するとした。これによりどの位置の文節においても、プロミネンスによる影響が出やすい状況をつくり出すことが出来ていると思われる。

2.4.3 手段3

- 基本周波数推移の音声単位毎の直線近似を行い、直線と基本周波数推移の曲線とがなす距離の平均値の最大値をプロミネンスの位置とするもの (図 2.6)

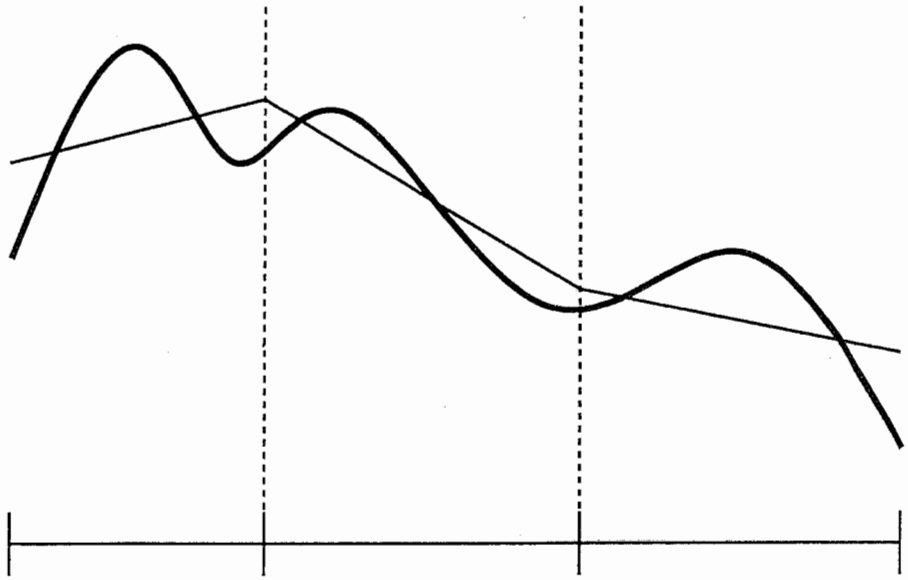


図 2.6: 手段3による抽出の方法

これ以後2つの手段については、韻律情報として文節の開始位置と終了位置がわかっているものとした。その上で本手段は以下のような操作を行った。

1. 基本周波数推移曲線を文節毎に直線近似
2. 基本周波数推移曲線と近似直線がなす距離の総和を面積として求める
3. 求めた面積を文節区間長で割る
4. 3の数値の文節ごとの数値を求め、数値が最大である文節中にプロミネンスが含まれているものとする

既に述べているように、プロミネンスを含む文節のほとんどがアクセントによる推移幅が音声区間全体に渡って大きくなるパターンをとっている。文節の区間がわかっている状態なら、区間内における曲線と直線の距離の平均値を求めることにより、文全体を通しての傾向ではなく文節毎の傾向がわかるのではと思われる。

2.4.4 手段 4

- 基本周波数推移の音声単位毎の直線近似を行い、直線と基本周波数推移の曲線とがなす距離に対し、単位の後半部分に重み付けをしたものの平均値の最大値をプロミネンスの位置とするもの (図 2.7)

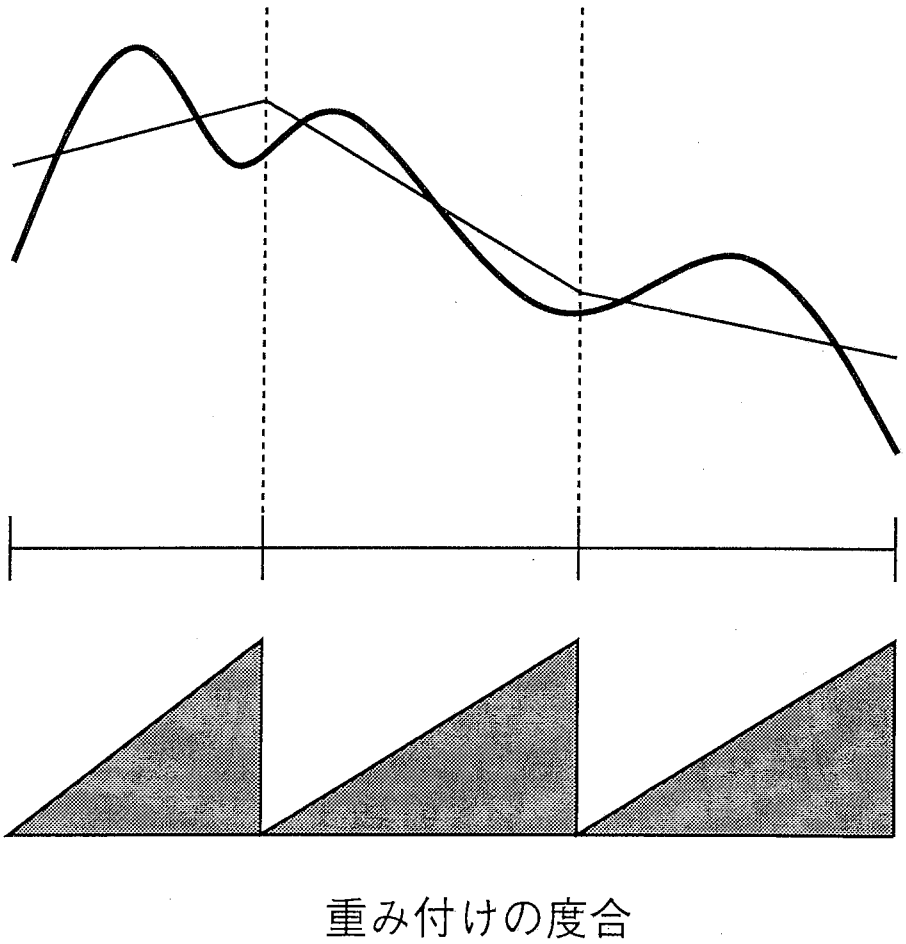


図 2.7: 手段 4 による抽出の方法

本手段は以下のような操作を行った。

1. 基本周波数推移曲線を文節毎に直線近似
2. 基本周波数推移曲線と近似直線がなす距離に、図のように区間末により大きな重み付けをし、距離の総和を面積として求める

3. 求めた面積を文節区間長で割る

4. 3の数値の文節ごとの数値を求め、数値が最大である文節中にプロミネンスが含まれているものとする

手段3との違いは面積を求める際に求めた距離をそのまま総和していくのではなく、区間頭の値は軽く区間末の値は重くなるように重み付けした値を総和していく点にある。これは2.3節で述べたプロミネンスによるアクセント推移の種類のうち、区間末のみアクセント推移が上昇するパターンに対応させるための手段と考えた。

2.5 認識結果

まず、それぞれの手段による認識に成功した文の個数(表2.1)を示す。手段1,2については特定化した点がプロミネンス文節に含まれていれば認識成功、手段3,4については特定化した文節がプロミネンス文節に等しければ認識成功とみなすこととした。ただし、認識に用いた文は76文音声で、必ずプロミネンスを含むものとした。その上で手段1と2、手段3と4の違いを見るために、共通して認識成功した文および失敗した文の個数(表2.2)を示す。

表2.1: 4手段の認識結果

手段	認識成功文数	第一文節	第二文節	第三文節	第四文節以降
手段1	51	13	16	13	9
手段2	52	6	13	14	19
手段3	54	11	15	14	14
手段4	52	10	15	13	14
全候補数	76	15	20	19	22

表2.2: 手段間での共通認識の結果

共通事項	認識成功文数	第一文節	第二文節	第三文節	第四文節以降
手段1,2共に成功	38	6	11	12	9
手段3,4共に成功	44	9	14	11	10
手段1,2共に失敗	11	2	2	4	3
手段3,4共に失敗	14	3	4	3	4
全候補数	76	15	20	19	22

2.6 考察

2.6.1 手段間の違い

まず全体を通して見てみると、手段1は予想通り認識率が最も悪かった。しかしそれでも6割以上の確率で認識出来たことは予想外であった。それに対し手段2以降ではかなりの認識率向上が見られると予想していたが、手段1に比べほとんど差のない結果となってしまっている。これにはプロミネンスという概念自体の曖昧さから、このあたりが自動認識の限界であるのかも知れないと思われる。

手段1の認識結果の傾向としては、第一文節にプロミネンスがある文の認識率が最も良かった。ただしこれは手段1の認識方法が第一文節を選びやすい構造になっているためでもある。その代償として第四文節以降の認識率は半分以下にまで落ち込んでいる。

手段1とは対象的に、手段2は第一文節での認識率は最も低いが、第四文節以降の認識率はかなり良いものになっている。しかしこれは手段2の認識方法の構造上の点から考えると、期待とは少し違った結果であると言える。なぜなら、直線近似を行った以上どの文節においても同様な認識率が得られるはずだからである。その理由としていくつか考えられることがある。

- 直線近似がうまく行えていなかった
- 基本周波数推移の大局的な動きは正確には直線ではなく放物線に近いため、ずれが生じてしまっている
- 文が長い場合、文の途中で基本周波数の持ち直しが見られることがあり、一つの大局的な流れでは表現できない

これらの点から一つの直線および曲線による近似での認識率向上はある程度までしか望めないのが正直なところであろう。

以上2手段が双方とも偏った結果になったのに対し、手段3,4はどの文節位置の認識率も6割を越え、平均的に良い手段であると言えるだろう。強いて言えば双方とも、第四文節以降の認識率が第三文節までと比べ少し劣っている。これは四文節以上ある文の場合、候補となる文節数が増えるため、自然に認識率も低下しているのではないかと思われる。

手段3においては、距離の総和を求める際に、そのまま総和を行うのではなく距離を何乗かしてから総和を求める方法も試してみた。これは距離をそのまま足すより何乗かしてから足し合わせた方が、プロミネンスによる距離の増大の影響をより大きく表現することが出来るのではないかと思われたからである。その結果、1.1から1.4乗付近で最も良い結果が得られ、それ以降は単

調減少となる (表 2.3) ことがわかった。予想では 2.0 乗以上で最も良い結果が得られるものと思っていたが、実際に実験してみた結果、大部少ない値で最大値が得られてしまった。

表 2.3: 手段 3 における距離の冪乗数と認識成功文数との関係

距離の冪乗数	1.0	1.1	1.2	1.3	1.4	1.5	2.0	2.5	3.0	4.0	5.0
認識成功文数	54	55	55	55	55	54	54	53	52	49	47

手順 4 における重み付けの係数についてであるが、今回の実験に用いた方法は以下の 3 つの不連続な規則に従った。

区間内の距離に対し、

- 先頭から全体の $\frac{1}{2}$ の区間までは 0.1 倍
- 全体の $\frac{1}{2}$ から全体の $\frac{3}{4}$ の区間までは 2.0 倍
- 全体の $\frac{3}{4}$ から末尾までの区間までは 5.0 倍

という重み付けを行った後総和を求める

この手順の目的はプロミネンスによるアクセント推移が、音声単位の後半部以降にのみ上昇が見られる場合の対処のつもりであった。しかし、このパターンに当てはまる文における認識文数 (表 2.4) を見ると、手順 4 でなくてもそれなりにいい結果をが残せている状態だ。全候補数が少ないこともあり正確な割合はわからないが、他の方法でもそれなりの認識率は出ているので、この手順における対処は必要ないと言えるかも知れない。

表 2.4: 後半部にアクセント推移上昇が見られる文の認識成功文数

手順	手順 1	手順 2	手順 3	手順 4	全候補数
認識成功文数	9	8	8	8	10

2.6.2 手段間の相関

今回の認識実験のもう一つの注目点として、手段1と2、および手段3と4を対比して行ったことが挙げられる。それぞれの対比項目は以下の通りである。

- 手段1,2間の対比点

  直線近似を行うことにより、認識結果がどのように変わるか

- 手段3,4間の対比点

  重み付けを行うことにより、音声単位の後半部にアクセント推移上昇が見られる文に対応することができるか

まず手段1,2間の相関であるが、共通に認識が成功した文数の少なさから見てもわかるように、手段1と手段2では全く違う傾向が現れてしまった。手段1では文頭から文末に進むにつれて、逆に手段2では文末から文頭に進むにつれ認識率が下がっている。手段1の結果は予想通りであったが、手段2の結果は予想と反するものとなった。

また、手段1でのみ認識がうまくいった文に見られる傾向は、もちろんプロミネンス位置が第一もしくは第二文節にあること。それと対応するように手段2でのみ認識がうまくいった文に見られる傾向は、プロミネンス位置が第三文節以降にあることであった。ただし前節の考察より、これらの結果は直線近似を行ったことによる変化とは言い切れない。

次に手段3,4間の相関であるが、こちらはほぼ似通った結果が現れた。そのため両者の間にこれといった相関は見い出せなかった。また、対比点として挙げた対応についても、前節の考察より差が現れなかった。言い替えれば、手段4には手段3に対し有効な差が見い出せないというのが認識実験による考察であろう。

2.6.3 プロミネンスの表現方法

以上の考察をまとめると以下のようになる。

- 手段3が最も認識率が良く、文のどの位置の文節にも平均的に認識が行えている
- プロミネンスにおける単語のアクセント型による変化については、そのほとんどが発話単位全体に渡りアクセント推移幅が大きくなるパターンであるので、それほど注意が必要なさそうである
- 自動認識による認識率は70%台が限界ではないだろうか

3番目の項目についてであるが、プロミネンスを含む文の人間による聴取実験においてもその程度の認識率しか期待できない。それはやはりプロミネンスのもつ曖昧性という問題から来るものである。また今回の認識実験では、全ての文にプロミネンスが含まれているように選択したが、プロミネンスを含まない文音声に対しては対応できないのが現状である。すなわち、プロミネンスを含まない文音声に対しても、どこかにプロミネンスが存在するものとして認識を行ってしまう状態にある。

この点は一見欠点のように思えるが、逆に考えるとプロミネンスを全く含まない音声というものは存在しないと思われる。プロミネンスを含まないつもりで発声していても、微小ではあるにしても文中のどこかにプロミネンスに相当するものが存在すると考えてもおかしくはないだろう。

このように考えると、これまでプロミネンスを存在する存在しないというように認識する方法について述べてきたが、余程意識して発声された音声でもない限り完全に分類を行うことは不可能であるだろう。そこで、プロミネンスの存在非存在ではなく、どのくらい存在しそうかという、存在度数として表す方法について考えてみたいと思う。

2.7 プロミネンスの存在度数表現

存在度数表現の基本方針も、基本周波数制御であることは変わらない。前出のプロミネンス認識などと同様の方法で基本周波数値の推移から、プロミネンスの存在非存在のかわりにどのくらい存在しそうかという数値を求めるものである。

プロミネンスの存在度数の定義は以下のように行う。

- 音声単位毎に、何らかの形でプロミネンスがどのくらい存在しそうかという指数を数値化する (プロミネンス指数)
- プロミネンス指数を全ての音声単位 (本研究では文節) について求める
- それぞれの音声単位同士でプロミネンス指数を比較した上で、相対的な度数を数値化する (プロミネンス存在度数)

このように存在非存在ではなく存在度数を求めることにより、プロミネンスを曖昧に含む文や、プロミネンスをほとんど含まない文などにも、それぞれの文に合わせた値が得られると期待できる。

次節ではプロミネンス存在度数を用いた実験を示す。

2.8 存在度数表現を用いた実験

2.8.1 実験の基本方針

chatr 音声合成システムに、日本語のローマ字によるテキストのみを入力として用いた場合、ある程度のイントネーションをもって合成音声出力が行われる。その出力にはもちろんプロミネンスは含まれていない。

ここでプロミネンス存在度数とは、プロミネンスを含む一つの文における基本周波数推移の音声単位毎の特徴を表したものであると言える。よって、プロミネンス存在度数が適正に作用するものならば、この合成音声に対し存在度数を適応することにより、「プロミネンスを含むように聞こえる合成音声」が作り出せるのではないかと想像できる。

本実験はプロミネンス存在度数というものを用いて、基本周波数推移の細かい制御なしに音節単位毎のプロミネンスを合成音声に組み込もうとするものである。これがもし可能になれば、異なる話者間、および異なる言語間でもプロミネンスを維持したままの翻訳がスムーズに行えるようになると言えるだろう。

今回はその準備として合成音声用のデータベースのある話者 (MHN) が発声したプロミネンスを含む音声を用い、出力として得る合成音声にも同一話者を用いた。

2.8.2 手順

実験では人間の話す「プロミネンスを含む」文音声と、chatr 音声合成システムの作り出す「プロミネンスを含むように聞こえる」合成音声の比較を行い、存在度数表現がうまく行えているかを見ていく。その手順 (図 2.8) は以下のようになる。

1. 「プロミネンスを含む」文音声の「テキストのみ」を chatr の入力として用いる。

   ここで音声を出力として選べば、chatr から発声されるのは「プロミネンスを含まない」合成音声である

2. 1 の出力として「プロミネンスを含まない」合成音声のセグメントを得る

3. 文音声からプロミネンス存在度数を求める

4. 存在度数と 2 で得られたセグメントを処理し、chatr に入力として用いる

5. 4 の出力として「プロミネンスを含むように聞こえる」合成音声を得る

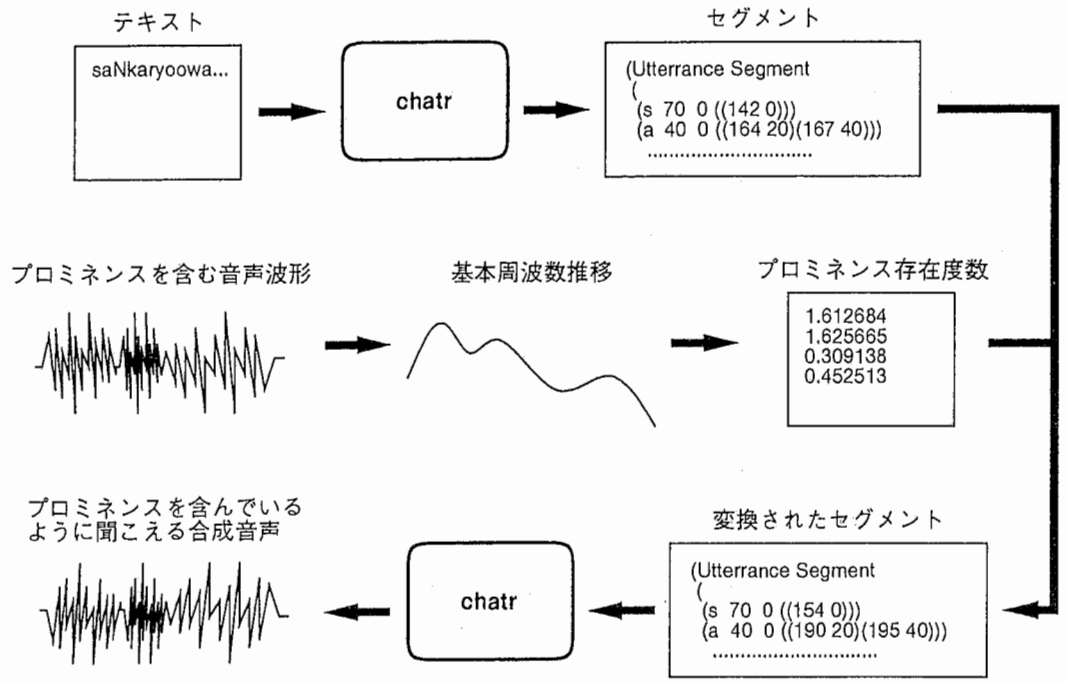


図 2.8: 実験の流れ

2.9 実験結果

実験に使用した音声と chatr への入力テキストとして用いたものは以下の通りである。

- 音声 「参加料には、予稿集代と歓迎会費が含まれています」
- ローマ字テキスト 「saNkaryooniwa yokoosyuudaito kaNgeekaihiga hukumareteimas#」

実験結果として次に挙げる音声波形の基本周波数推移のグラフを載せる。

- テキストのみを入力とした時の合成音声 (図 2.9)
- 文音声と、文音声におけるプロミネンス存在度数を適応した合成音声の組
  - プロミネンスを含まない文音声 (図 2.10),(図 2.11)
  - 第一文節にプロミネンスを含む文音声 (図 2.12),(図 2.13)
  - 第三文節にプロミネンスを含む文音声 (図 2.14),(図 2.15)

2.9.1 テキストのみを入力とした時の合成音声

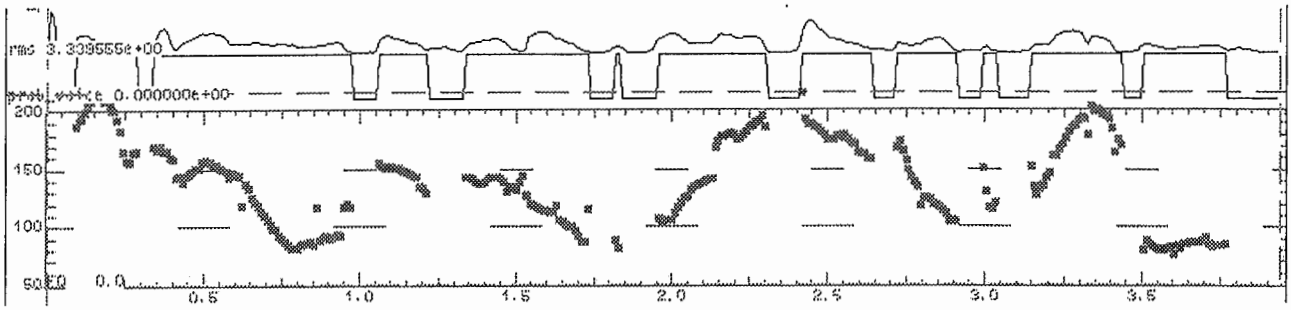


図 2.9: テキストのみを入力とした合成音声

2.9.2 プロミネンスを含まない文音声を用いた実行結果

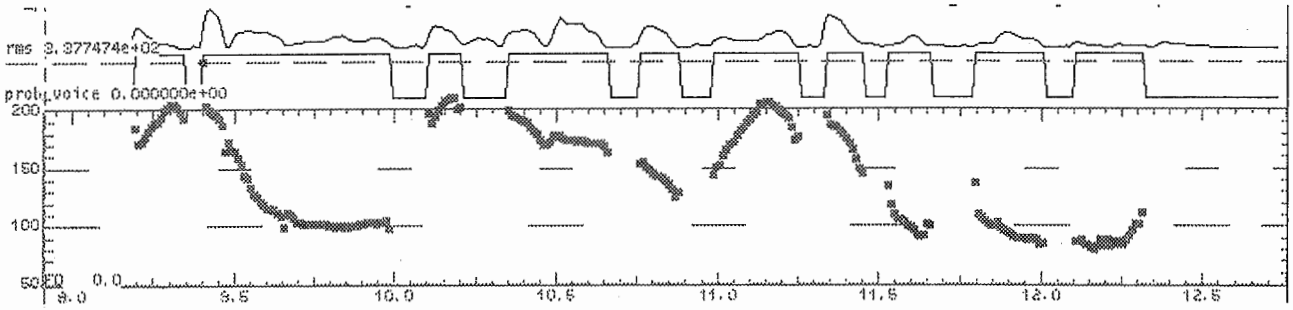


図 2.10: プロミネンスを含まない文音声

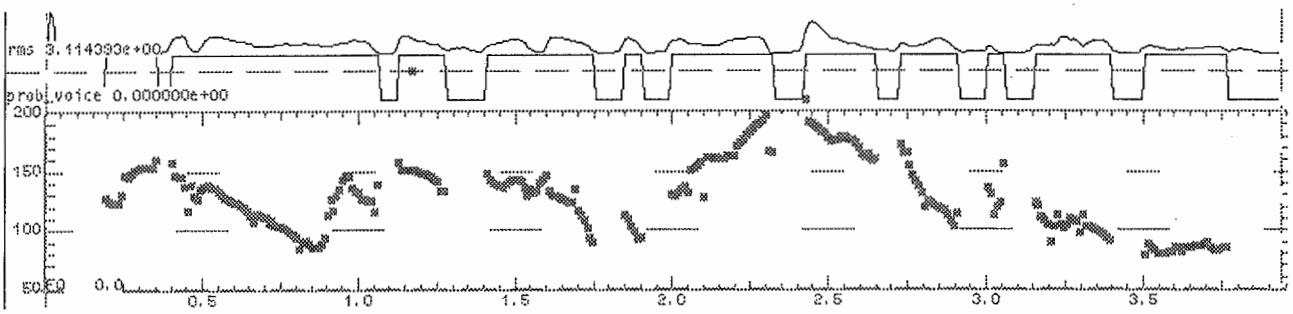


図 2.11: 文音声におけるプロミネンス存在度数を適応した合成音声

2.9.3 第一文節にプロミネンスを含む文音声を用いた実行結果

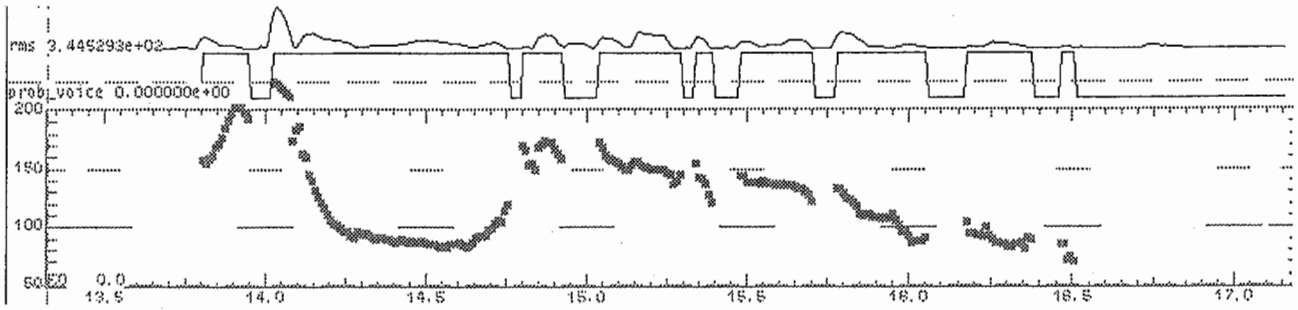


図 2.12: 第一文節「参加料には」にプロミネンスを含む文音声

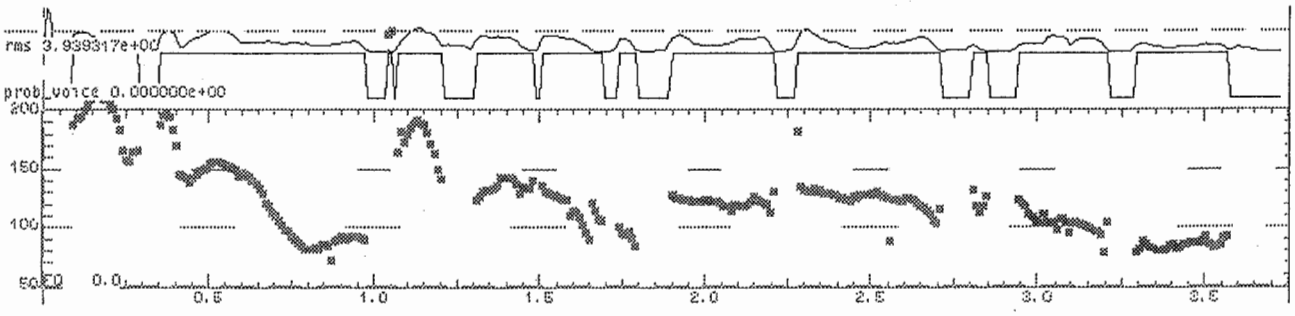


図 2.13: 文音声におけるプロミネンス存在度数を適応した合成音声

2.9.4 第三文節にプロミネンスを含む文音声を用いた実行結果

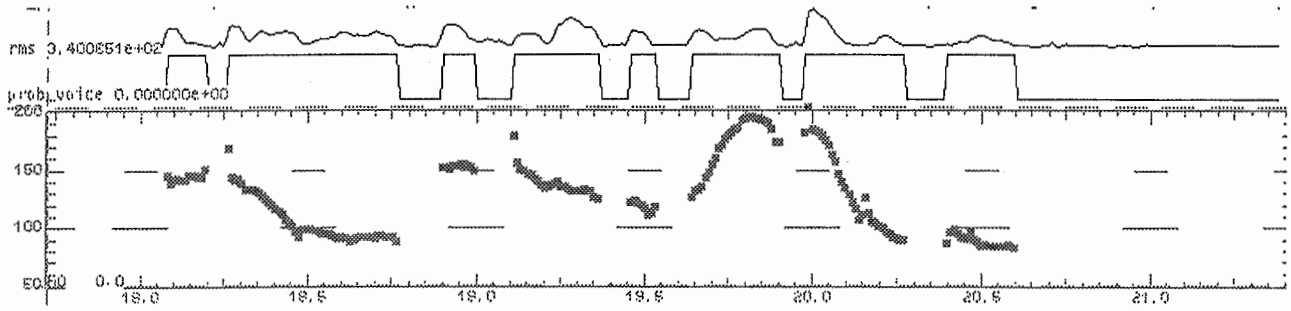


図 2.14: 第三文節「歓迎会費が」にプロミネンスを含む文音声

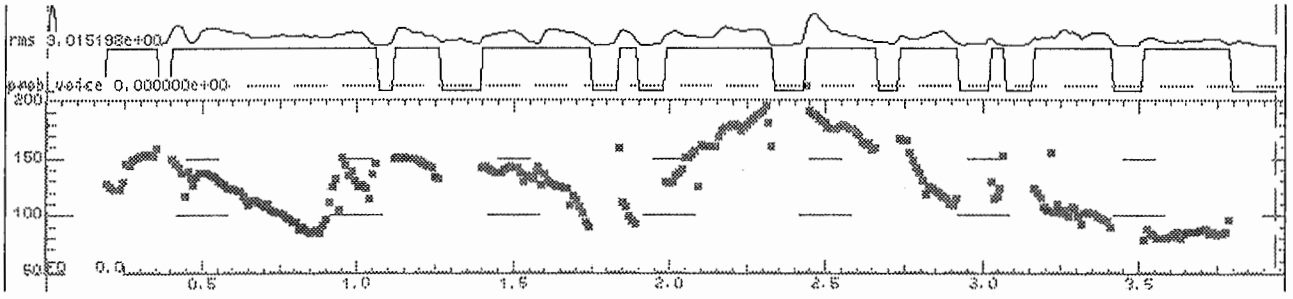


図 2.15: 文音声におけるプロミネンス存在度数を適応した合成音声

2.10 実験の考察

文節毎の基本周波数推移という点だけからすると、主観評価ではある程度の結果が得られたのではないかと思われる。

ただ、基本周波数推移のみを変換しただけであるので、音声全体の長さがあっていなかったり、音素の長さがおかしかったりする実験結果が見られた。これはプロミネンス存在度数を適応するべき、「テキストのみを chatr の入力として用いた合成音声」がそのような構造をしていたためである。

今回プロミネンス存在度数として基本周波数推移の変化率しか用いなかったが、上記の問題の対策として音素長などの情報を用いることも音質の向上に繋がると思われる。

第 3 章

イントネーション・エディタ

音声対話システムにおいては合成音声の韻律情報を何らかの形で変換したい場合がある。例えば合成音声のイントネーションが気に入らないので直したいといった時や、自分の喋った音声を他の話者の声で出力したい時などがそれにあたるだろう。

ここでは、音声対話システムの出力としての合成音声におけるイントネーションに対して、そういった話者性の変換やモデル音声による補正などを行う際の方針を述べる。

3.1 基本方針

イントネーション・エディタの方針として今回考えた操作は次の 2 つである。

- 話者 A による音声のイントネーションを話者 B にあったイントネーションに変換する (話者性変換)
- ある音声に対し、モデル音声を用意することによりそのイントネーションを補正する (イントネーション変換)

イントネーションに関連する変換はこの他にも多数ありそうだが、とりあえずここでは、以上の 2 つの操作について述べていく。

3.2 手順

前節で挙げた 2 つの方針における手順を順次説明していく。

3.2.1 話者性変換

話者性変換が必要な場面として考えられるのが、章の導入部でも述べたように自分の音声を内容を変えずに他の話者の音声に変換して出力したいといった時などがある。また、次節で述べるイントネーション変換のための一過程としても応用できる。chatr システムにおいては、多様な話者データベースが存在するため、このような操作を行う環境として問題はないと思われる。

この際、話者の韻律情報を処理する必要がある。その手順を話者 A の音声を話者 B の音声として出力したい場合を例として以下に示す。

1. 話者 A の音声から韻律情報を得る
2. データベース中にある話者 B の音声についての韻律情報としての特徴を得る
3. 話者 A の韻律情報を話者 B の韻律情報特徴に合うように変換する
4. 3 の変換された韻律情報を元に話者 B のデータベースを用い音声を合成する

3.2.2 イントネーション変換

イントネーション変換は様々な場面で必要になると思われる。合成音声の出力のイントネーションが意図しているものと違ってくるということは、音声合成を行う上で避けられない問題である。リアルタイムに翻訳を行いたい場合などにおいては対処のしようがないが、場内放送などリアルタイムで発声する必要のない場合に有効であると考えられる。

その手順を合成音声 A をモデル音声 B を用いて補正する場合を例として以下に示す。

1. モデル音声 B から韻律情報を得る
2. モデル音声 B の韻律情報に対し話者性変換を行い合成音声 A の話者に近付ける
3. 2 の変換された韻律情報を用い合成音声 A の話者のデータベースを用い音声を再合成する

3.3 システム環境による制約

イントネーション・エディタという分野は音声対話システム全般に適応し得るものであると考えられる。そこでこれらの操作を chatr システムに繰り込む際の問題点となりそうな部分について、chatr というシステムの環境を考慮しながら述べていきたい。

環境の問題として以下の 2 点がある。

- 有限個の音声データベースから音声素片を繋ぎ合わせる合成方法であるため、指定した通りのイントネーションが結果として出力できない場合がある
- 音声の認識率に限界があるため、音素情報などを完全に信頼することができない

前者の問題はデータベースの絶対量に関連してくる。よって絶対量を多くすることにより対処は十分可能である。

chatr では合成音声に対し信号処理を行わず音声の品質を保ったまま合成を行うことを目的としている。よってイントネーション・エディタで指定されたイントネーションに対応するデータベースがない場合、それに近い音声素片を繋ぎ合わせることで対処している。従って、折角の補正が活かされない可能性が大いにありそうだと言える。

後者の問題は音声認識技術の完全性が必要であるため、実質的な対処は非常に難しいと言わざるを得ない。

韻律情報の内、基本周波数推移や音声パワー、有声無声判定などは使用に耐え得る結果が得られるが、音素種類、音素長など言語に固有の情報は現状ではまだ問題がある。ただしイントネーション変換において、音素長などの情報を用いなかった場合には次のような問題点が考えられる。

> 合成音声とモデル音声の全体の長さが異なる場合、どちらかを伸縮させてもう片方に合わせなければならない。そういった場合、音素長情報を用いずに行うと、合成音声とモデル音声の音素の位置がずれてしまう (図 3.1)。

こういった理由から音素情報を無視してイントネーション変換という作業を行うことはできない。音声認識が不完全な状況下では何らかの対応策が必要であると思われる。

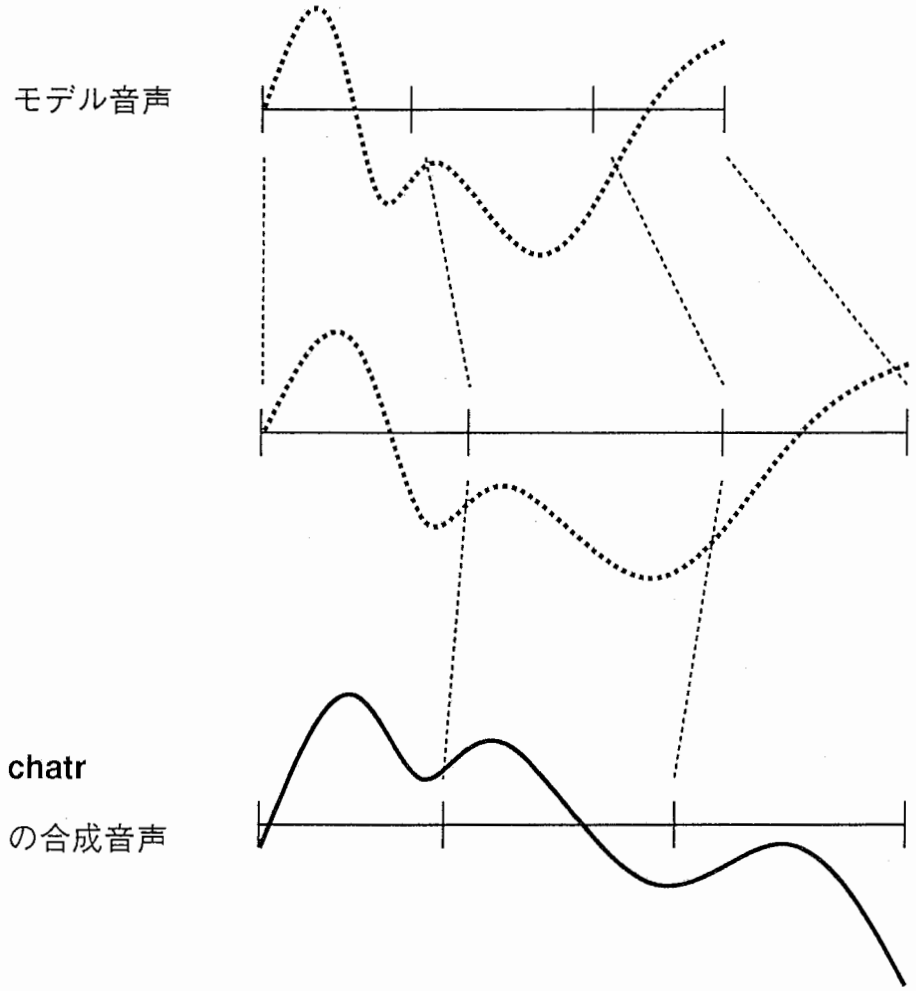


図 3.1: 音素の位置がずれてしまった例の模式図

3.4 考察

これらの方針は chatr システムにとって必要なものと思われるのだが、前節で述べた通り現システムにそのまま対応できるものではなさそうだ。今回実際に予備的実験を行ってみたが、予想通り目的のイントネーションとは少し異なる出力しか得られなかった。システムの向上によっても問題は解決されるが、現システムにも対応できるような新たな対策も考えていきたい。

第 4 章

おわりに

4.1 まとめ

本論文では音声対話システムの入力部と出力部の制御について、実験およびその考察を行った。

入力部についてはプロミネンスの抽出と題し、文音声におけるプロミネンス情報を数値化することを目的として作業を行った。まずプロミネンス推定のための手段を 4 つ述べた。次に実際にプロミネンスを含んでいる音声データを用い、プロミネンス認識実験を行いその結果を考察した。その上でプロミネンス存在度数という形でプロミネンスがどのくらい存在しそうかという概念を数値化した。その数値を使った音声変換実験も行ってみた。

出力部についてはイントネーション・エディタと題し、文音声におけるイントネーションを変換する方法についての方針を立てそれについての考察を行った。方針として話者性変換とイントネーション変換の 2 種類を立て、それぞれの実現手段を予想した。その上で chatr システム上での制約事項を考慮して問題点を考えていった。

4.2 今後の展望

今後に改良の余地の残る点として、プロミネンスの抽出については、プロミネンス存在度数表現のアルゴリズムが確立していない点が挙げられる。今回の実験ではプロミネンス推定に用いた手段 3 を元に存在度数を求めていったが、この方法はあくまでも予備実験のための措置であり、アルゴリズムとして最適とは言えない。また、基本周波数値のみを用いているので、音声パワーや音素長などその他の影響も考慮した新しいアルゴリズムを考える必要があると思われる。

イントネーション・エディタについては、今回は方針を述べるにとどめたが、現システムでも適応できる対策を考え、実験を行うことが必要であろう。

今回の研究では、音声対話システムにおける認識と合成の統合にむけての方針付けを行ってみたつもりである。入力分野においてはプロミネンス以外の発話アクト、意図についても順次適応を行っていき、最終的には入力音声におけるそれらの情報を出力音声にもそのまま適応できるようにすることが目的である。出力分野においてはシステムの向上と共にイントネーション・エディタを完成させ、自由に出力音声の補正を行えるようにすることが目的である。その上で、chatr翻訳システムに適応することにより認識と合成の統合という問題が解決すると思われる。

参考文献

[1] 藤尾茂,Nick Campbell, 樋口宜男," 韻律を用いたテキスト非限定型発話アクト識別方法", 音講論集,1-4-14(1996,3)

[2] 日本放送協会編," 日本語発音アクセント辞典" (1966)

[3] 匂坂芳典," 韻律制御研究の現状と課題 - より自然な音声を求めて -", 音響学会誌,49,854-859(1993)

謝辞

本研究を進めるにあたって、暖かく見守りつつ、多くの御指導を頂いた匂坂芳典第一研究室室長、樋口宜男第二研究室室長に心から感謝致します。

さらに、本研究に対する適切な助言を頂き、様々な相談にのって下さったATR音声翻訳研究所第二研究室の皆様に深く感謝致します。特に、Nick Campbell氏には、chatrシステムやツールの使いかたといった初歩的なところから研究の方針まで全分野に渡って丁寧に指導、助言して下さったことに心から感謝の意を表します。

またお忙しい中、音声認識プログラムの使用法を教えて下さった山本博史氏、様々なデータベースについての説明をして下さった太田洋子、下田京子両氏にも感謝致します。

最後に、本研修の機会を与えて下さった早稲田大学の白井克彦教授、ATR音声翻訳通信研究所の山崎泰弘社長に心から感謝致します。

1997年3月

星野 慎一