修 士 論 文

連続発話母音の基本周波数変動に含まれる個人性に関する研究

指導教官　赤木 正人 助教授

北陸先端科学技術大学院大学
情報科学研究科情報処理学専攻

皆川知也

1998 年 2 月 13 日

目 次

図 目 次

表 目 次

第 1 章

序論

1.1 研究の目的

現在、音声合成、あるいは音声認識を人間と機械との情報交換の手段として用いることには今だに困難がある。機械の作り出す音声は、自然な音声というには程遠い品質である。また、音声認識は依然として限定された語彙、話者であるような場合でのみ可能である。音声をインターフェイスとして用いることの難しさの原因の１つに、音声の個人性がある。音声が生理的な発話器官で生成され、発話器官に個体差がある以上、音声に個人差が生じることは避けられない [15]。しかし、音声に含まれる個人性を抽出しパラメータ化することで、音声における個人性を表現できるようになる。つまり、パラメータを変化させることでより自然な音声を少ない情報から生成することができ、この個人性を表すパラメータを音韻性などを表すパラメータと分離することで、音声認識における認識率の向上につながる。

音声の個人性は声道特性と声帯特性とに大別できる [2][3][4]。声道特性としては、ホルマント周波数、スペクトル包絡の時間変化パターン、スペクトル包絡の形と傾斜、平均スペクトル包絡特性等が挙げられる。一方、声帯特性としては、平均基本周波数、基本周波数の時間変化パターン [1]、基本周波数の揺れ等がある。過去に、平均基本周波数や基本周波数の時間変化パターンについては分析が行なわれている。そこで、本研究では今までに議論されることの少なかった声帯特性の 1 つである基本周波数の揺れ [6][7]、とりわけ母音を連続発話しているときの基本周波数の揺れに着目し、個人性を検証することとした。

1.2　研究の背景・特色

音声には言語の意味を表す、という特性の他にいくつかの特徴が含まれている。その中でも重要で、かつ明確である特徴が話者の個人性である。この情報は、人の身体的構造には個人差があり、音声が人間の発話器官で生成される以上、不可避的に音声に含まれる。したがって、音声の個人性の情報は、より品質の良い音声合成、音声認識を達成するための重要な情報である。また、同じ音の物理的特徴が話し手によって変化することを利用して、音声による話者の識別に利用し、個人情報に基づくサービス等に用いることができる。つまり、音声情報から個人性を表すパラメータを抽出することは有用なことである。

現在までに、この個人性に関しては声道特性、声帯特性の両面から個人性を表す特徴を抽出する試みが行なわれてきた。声帯特性については、例えば藤崎モデルを用いた基本周波数の時間変化パターンに現れる個人性については分析が進んでいる。しかし、基本周波数の揺れに含まれると考えられる個人性についてはあまり分析がされていない。基本周波数の揺れは、一定の高さに保ち発話している母音から単語や文章発話時まで、あらゆる状況で存在する。そこで本研究では連続発話した母音の基本周波数に現れる基本周波数の揺れに着目し、その揺れに含まれると考えられる個人性について分析を行なった。

第 2 章

Laryngograph 出力信号からの音声の基本周波数の推定

2.1 Laryngograph について

声帯の開閉運動を電気信号として記録した波形を Electro-Glotto-Graph(EGG) という。本研究で用いる Laryngograph という装置は、EGG が得られる装置の 1 種であり、米国 Kay 社の製品である。EGG の原理は、甲状軟骨上の皮膚に電極板を置き、弱い高周波電流を両極間に流し、音声発生時の両極間のインピーダンス変化が高周波電流の振幅の変化として検出される、というものである。したがって、生体を侵食することなく、また声道の共振や環境雑音等に影響されずに声帯の振動を計測することができる。図 2.1は Laryngograph の写真である。

図 2.1: Laryngograph の写真

2.1.1 Laryngograph 出力波形の観測

大きな傑出した喉頭をもつ (一般には男性) 発声者に対しては、甲状軟骨の位置の決定が容易であり、喉頭が鋭い角度であるので声帯の振動、及び周辺の電界を容易に集中できる。逆に喉頭が傑出していない (一般には女性) とき、特に比較的厚い皮下組織の層におおわれているとき、喉頭の位置の決定が困難であり、波形の詳細な形を十分に明確にできない。そこで波形の信頼性は、波形の観測時にオシロスコープなどで波形を点検することで確かめねばならない。

2.1.2 波形の解釈

正常な Laryngograph の出力信号の波形 L_xは、声帯の振動と関係があり、以下のような特徴がある。

- 閉鎖／開放の一連の場面は規則的である
- 各周期における閉鎖／開放の一連の場面は類似している
- L_xの立ち上がりは声帯の閉鎖に対応し、立ち下がりは開放に対応している
- 声帯は開放する時よりも素早く閉鎖するため、L_xの立ち上がりの縁は、立ち下がりの縁よりも急峻である

図 2.2に 6 種類の声帯の接触状況と、各状態の L_x上での位置を示す。1〜3 が声帯が閉鎖する期間であり、3〜5 が開放する期間である。先に声帯の性質として、開放よりもく閉鎖の方が素早いと述べた。このために、接触が始まってから (1 の状態) 完全に接触する (3 の状態) までに要する時間の方が、完全に接触している状態 (3 の状態) から完全に離れる直前の状態 (5 の状態) に至る時間よりも短い。したがって、立ち上がり (1〜3) が立ち下がり (3〜5) よりも急峻になる。5 から 6 の間は声帯が完全に開放されている期間である。

また、L_xは病理学的な発声条件の物理的な解釈の基礎を提供する。検査できる特徴は、

- 振動の規則性
- 閉鎖の時期の限定
- 開放の時期の限定

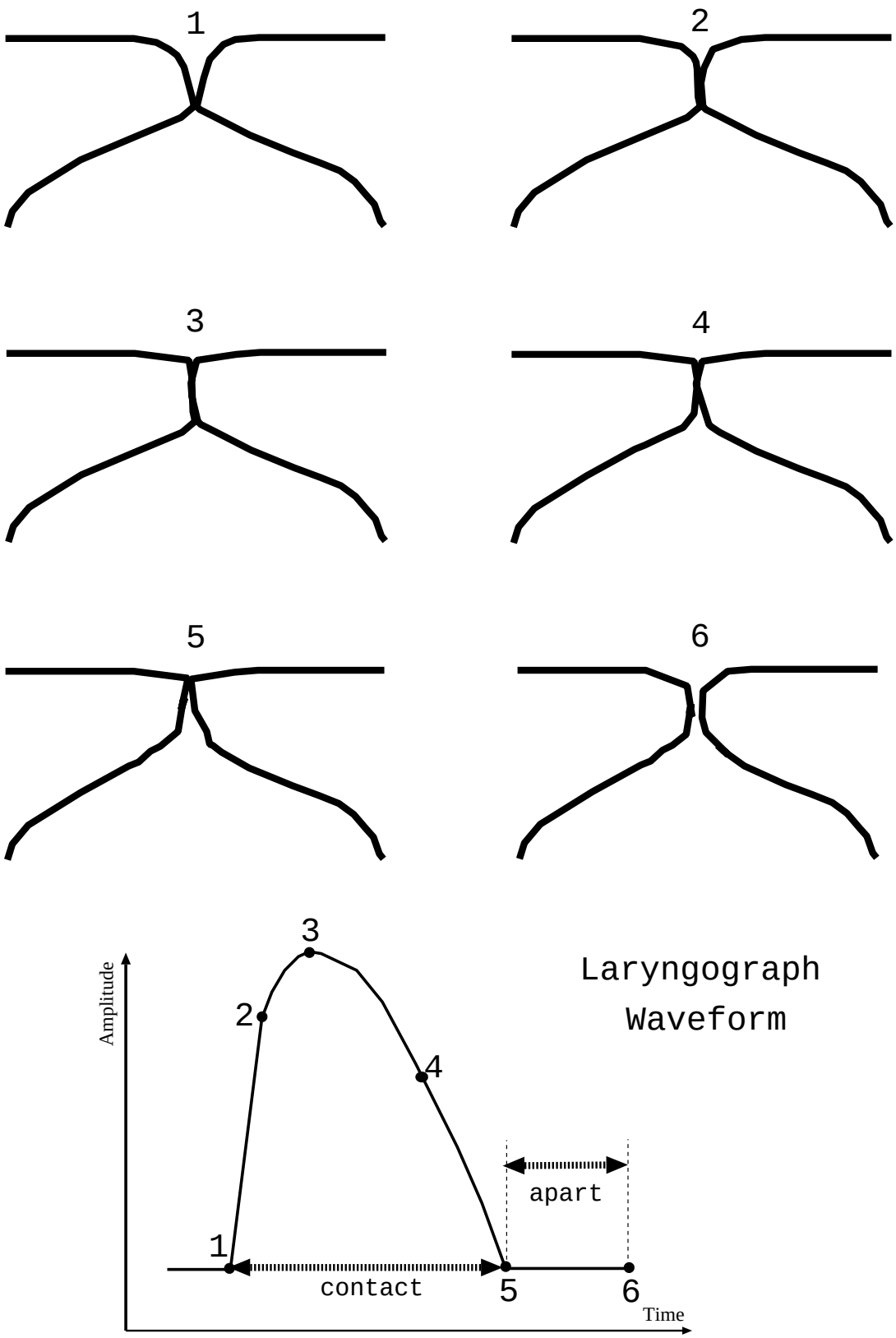

図 2.2: 声帯の開閉と L_x の関係図

- 開放時間／閉鎖時間の比
- 閉鎖／開放の一連の場面における形状

である。さらにこれから声帯に関して、質量の不均整、硬さの不均整、小さな節やポリープの位置、治療／療法が正常な状態への回復に作用しているか、ということがわかる。

2.1.3 Laryngograph を用いる利点

本研究では音声の基本周波数の細かい変動に個人性に関与するような特徴が含まれているかどうかを調べることが目的である。この細かい変動とは、人が一定の高さで母音を発生し続けている場合でも、声帯が音声が発生している間中、まったく同じ間隔で閉じる／開くということを繰り返すわけではないことに起因する、基本周波数の微妙な揺れのことを指す。従来の様々な手法、例えば自己相関関数やケプストラムを用いた方法、は分析フレームごとで基本周波数の推定を行なうために、推定できた値は最低 1 ピッチ周期以上の長さのフレーム (大抵 1 フレームは 20〜30ms 程度) 内の平均値である。つまり、ある種の平滑化を施された値であり、またフレーム周期ごとでしか値を得ることができないので、着目した細かな変動を含んだ基本周波数の推定は不可能である。

しかし、Laryngograph を用いると声道の共振や雑音に影響されずに声帯の振動を捉えることができる。また、その出力信号の周期の逆数が音声の基本周波数と推定できるので、フレームごとに区切るという処理を必要としない。このため、細かな変動を含んだ基本周波数の推定が可能になる。これが Laryngograph を使うことの利点である。

2.2 基本周波数の推定

音声の基本周波数を推定するために、L_xからパルス列の抽出、異常値修正、内挿といった処理を行なう。この節では L_xに行なう信号処理を具体的に説明する。

2.2.1 L_xの採取について

分析を行なうための音声と Laryngograph のデータの採取方法について述べる。図 2.3 にデータ採取のためのハードウェア構成を示す。

データは以下に示す条件のもとで採取した。

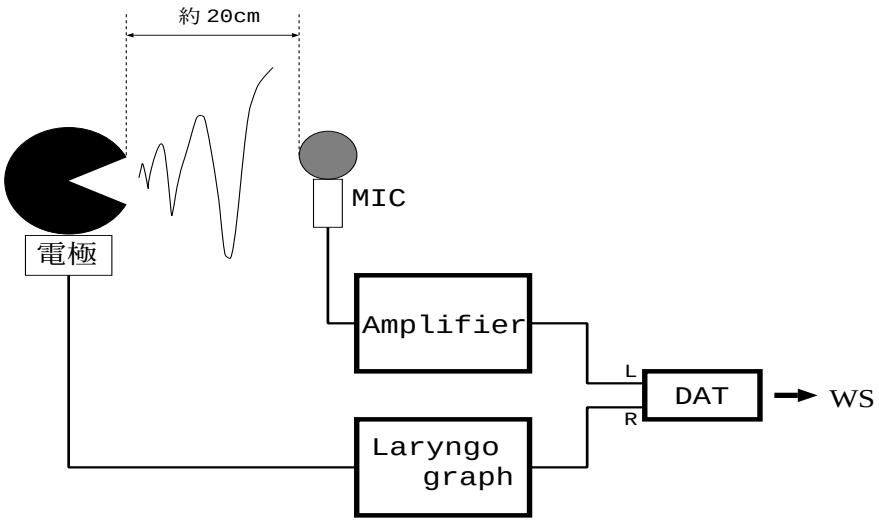

図 2.3: データ採取の為のハードウェア構成図

1. サンプリング周波数は 48kHz
2. 量子化数は 16bit
3. 録音は防音室内で行なう
4. 音声のみアンプを通す
5. 総話者数は 9 人 (全て男性話者)
6. 採取した音声は各母音と 4 種類の単語、5 種類の文章の 1 話者につき計 14 種類

一方、データ採取に当たって話者に課した条件を以下に示す。

- マイクロフォンと話者は約 20cm の距離を置く
- Laryngograph の電極はなるべく話者の甲状軟骨の真上になるように配置する
- 母音発声時にはヘッドフォンから 130Hz の純音を出力し、その高さに合わせて発声する

最初の条件についてであるが、これはできるだけ音声と Laryngograph の出力信号とを同期させるための条件である。話者が発声した時、声は声道を通り空気中を伝播しマイクロフォンまで到達する。一方 Laryngograph の出力信号は電気信号であるので、発声を行なった直後に出力が DAT まで到達する。したがって、両者の DAT までの到達時間には

差がある。この時間差を正確に測定することはできないが、ある程度まで軽減することはできる。音速を 350m/s、声道の長さが 15cm とすると、この条件下では声帯とマイクロフォンまでは 35cm の距離があるので、双方の信号の時間差は約 1ms になる。そこで、分析に用いるデータは全て 1ms の時間差があるものとし、その影響を取り除いて分析に用いた。

次の条件であるが、理想としては各話者の甲状軟骨の正確な位置を調べた上でデータ採取を行なえば良いのであるが、話者の甲状軟骨付近を視察するだけでは正確な位置を知ることは難しい。そこでなるべく甲状軟骨の上に電極を配置するようにした。

最後の母音発声における条件は、音声の基本周波数が高くなるにつれて基本周波数の細かなゆらぎの幅が大きくなるという報告がある [6]。故に基本周波数の高さがほぼ同じであるデータでないと、基本周波数の細かな変動に着目して分析を行なう時に、その基本周波数の高さの違いを考慮しなければならなくなるために加えた。

図 2.4に上述した条件で採取した音声波形と Laryngograph 出力信号波形 L_xの例を示す。これは男性話者が「めがね」と発音したときの波形である。上段が音声波形、下段が Laryngograph の出力信号波形である。

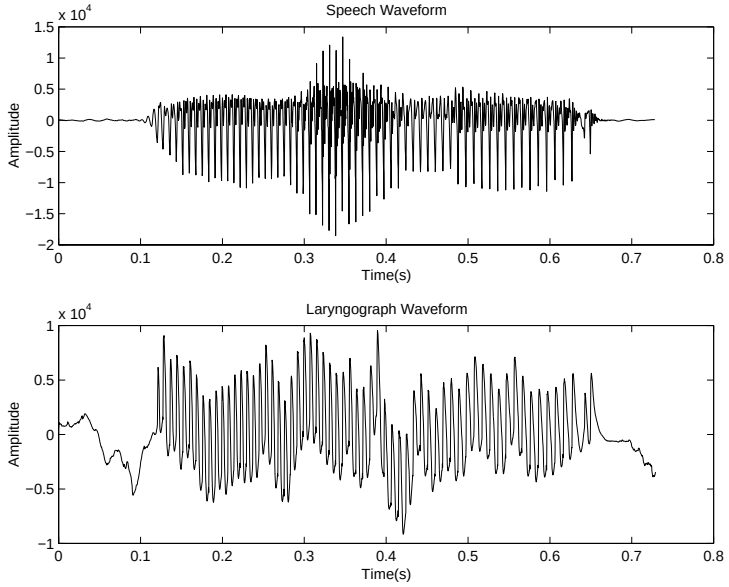

図 2.4: 採取したデータの例

2.2.2 瞬時基本周波数の抽出

声帯の閉塞／開放の一連の動きは、有声音生成時の音源生成によるものである。基本周波数とは声帯の振動周波数のことであるから、声帯が完全に閉塞し、再び完全に閉塞するまでの間隔の逆数が瞬時基本周波数となる。したがって、本研究では声帯の閉塞の瞬間を抽出し、この閉塞間隔の逆数を音声の瞬時基本周波数の推定値とした。この瞬時基本周波数を得るために、以下のような処理を行なう。

1. 移動平均を用いて Laryngograph 出力信号を平滑化
2. 微分処理により表れるパルス列を抽出 (=声帯の閉塞する瞬間の抽出)
3. いき値処理を行ないパルス間隔を抽出 (=基本周波数瞬時値)

まず平滑化処理であるが、これは Laryngograph 出力信号に含まれる雑音をある程度除去するためであり、実際の処理には移動平均を用いている。測定データを $x(n)$ $(n = 0, 1, \cdots, N-1)$、サンプリング周波数を f_s、カットオフ周波数を f_cとすると、$x(n)$ に M 点移動平均 (M は奇数) を行なった後の結果 $y(n)$ は次式で定義される。この計算式を用いると処理後に位相のずれが生じない [7]。

$$y(n) = \frac{1}{M} \sum_{m=-L}^{L} x(n+m) \tag{2.1}$$

$$L = \frac{M-1}{2} \tag{2.2}$$

$$M = \frac{0.443 \cdot f_s}{f_c} \tag{2.3}$$

カットオフ周波数は経験的に 2kHz としている。図 2.5の上段に Laryngograph の出力信号波形を、下段に上段の信号に移動平均を施した後の波形を示す。

次に微分処理について説明する。一般に声帯は素早く閉じ、ゆっくり開く、という性質がある。この性質は信号の立ち上がり部が他の部分より急峻になる、という特徴として Laryngograph 出力信号に現れることは既に述べた。そこで、この急峻な部分 (高周波成分＝声帯が閉じる瞬間) を取り出すために微分処理を行なう。図 2.6に微分処理を行なった後の波形を示す。微分処理により、パルスがほぼ規則的に並ぶ波形が得られることがわかる。

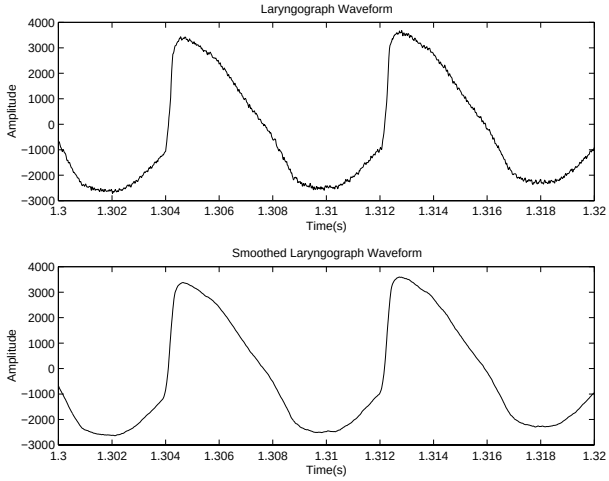

図 2.5: Laryngograph の出力と移動平均を施した後の波形の例

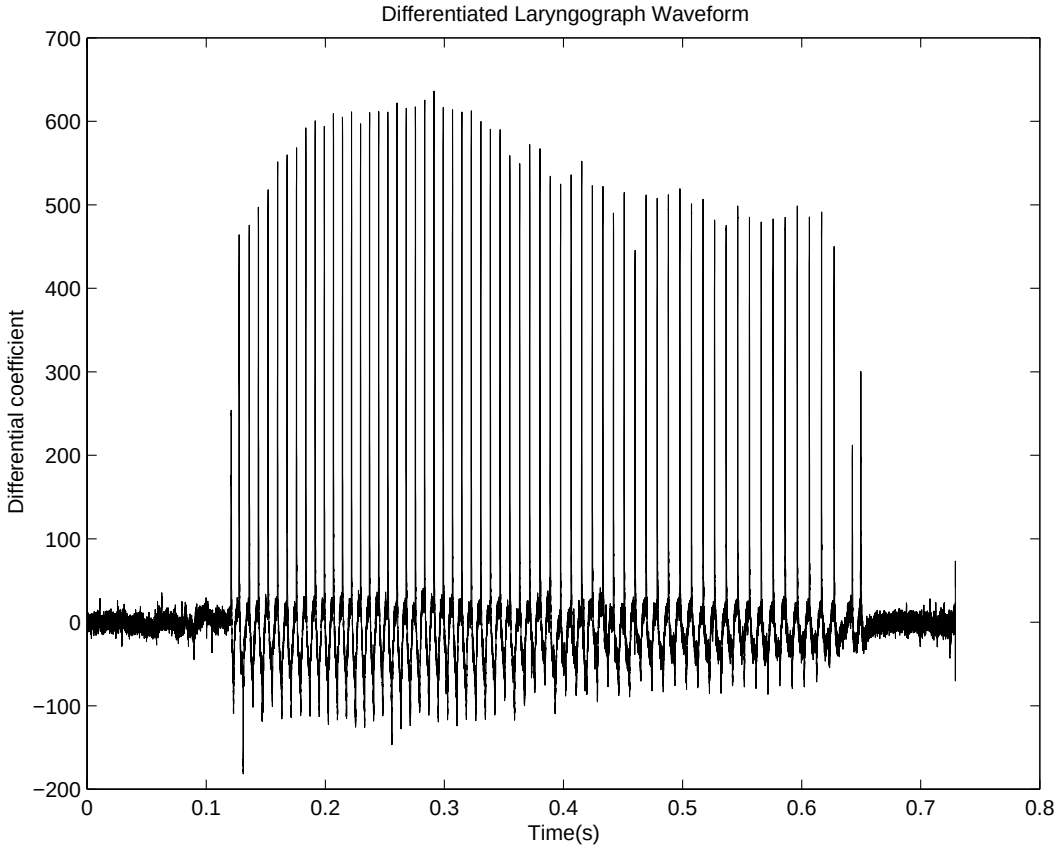

図 2.6: Laryngograph 出力信号を微分処理することで得られるパルス列

次にいき値処理によるパルス間隔の抽出を行なう。これを図2.7を用いて説明する。まずいき値処理を行ないパルス列を抽出する。次に時刻t_{n-1}、t_n、t_{n+1}にパルスが存在し、$t_n - t_{n-1} = P_n$、$t_{n+1} - t_n = P_{n+1}$であると仮定する。このとき、時刻t_nにおける基本周波数を$1/P_n$、時刻t_{n+1}における基本周波数を$1/P_{n+1}$とすることによって、瞬時基本周波数の抽出が可能となる。

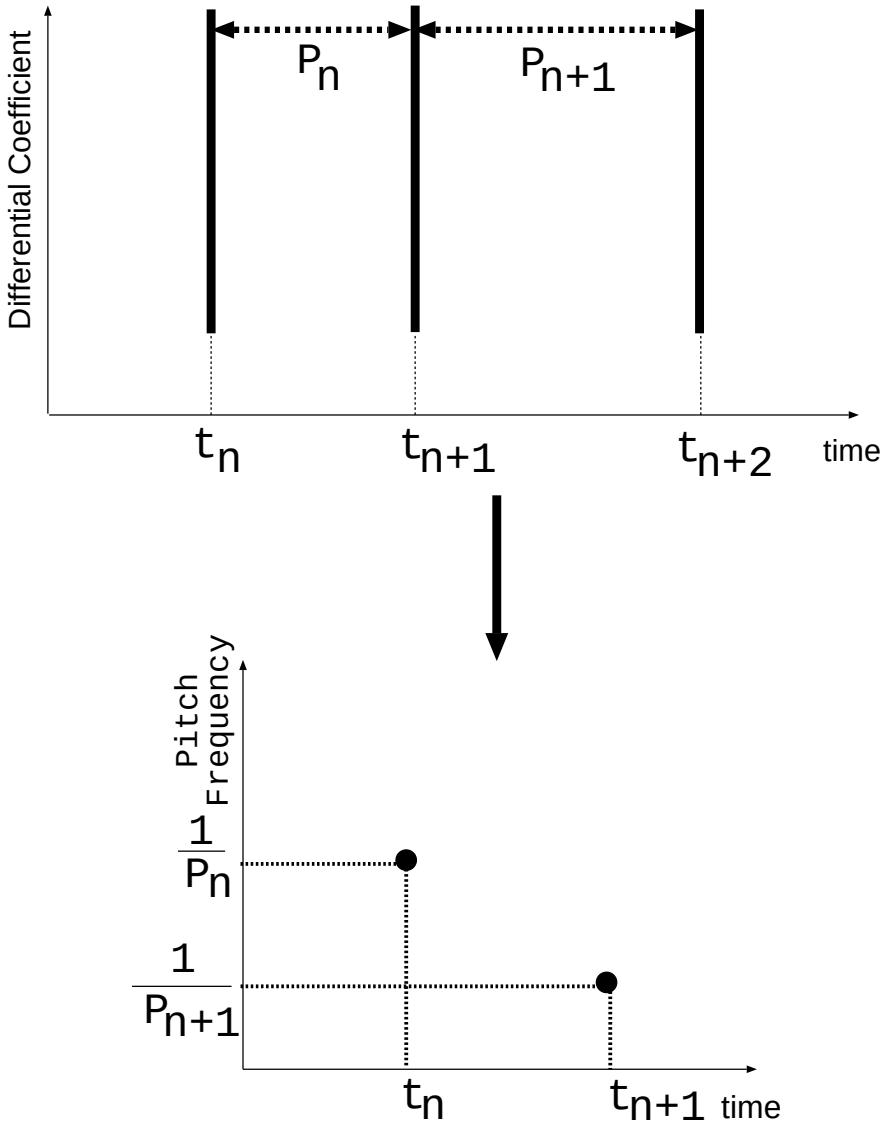

図 2.7: 瞬時基本周波数の抽出

2.2.3 瞬時基本周波数の異常値修正

図 2.8に Laryngograph の出力の一部に異常が見られる例を示す。

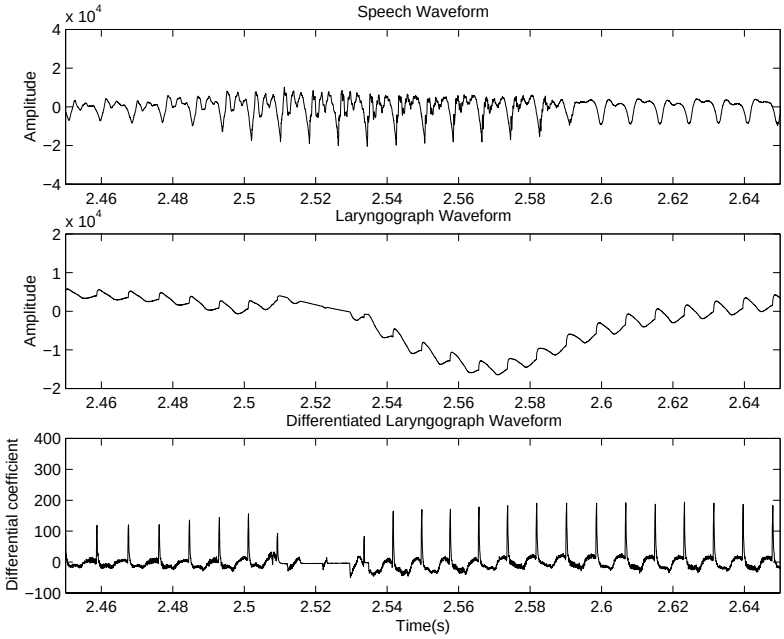

図 2.8: Laryngograph の出力異常の例

図 2.8の時刻 2.52 秒近辺で、音声は出力されているが、この時間帯だけ Laryngograph の出力が山と谷を繰り返す様になっておらず、ほぼ一直線に下降している。したがって、微分処理によってパルスの抽出を行なうことができない。このような出力異常等の理由により、瞬時基本周波数の抽出に支障を生じる場合がある。そこで、ある範囲に瞬時基本周波数の値が存在しない時に、メディアンフィルタを用いて異常値の修正を行なう。具体的な手法については図 2.9を用いて説明する。

まず時刻 t_{n-1}、t_n、t_{n+1}、t_{n+2}にパルスが存在し、それぞれのパルス間隔を P_{n-1}、P_n、P_{n+1}と仮定する。次に全てのパルス間隔が 50Hz〜800Hz の範囲 (音声の基本周波数と考えられる範囲) にあるかどうかを調べる。仮にこの範囲外の値をとるような瞬時基本周波数がある場合 (図中では時刻 t_{n-1}における瞬時基本周波数 P_n)、時刻 t_{n-2}、t_{n-1}、t_n、t_{n+1}、t_{n+2}の 5 点メディアンを求め、この値 (これが図中の$\hat{P_n}$) を時刻 t_nにおける瞬時基本周波数とする。また、音声の始まりや終り近辺で異常値が見受けられる場合は、最初、あるいは最後の 5 つの瞬時基本周波数のメディアンを求めて、同様の処理を行ない修正する。

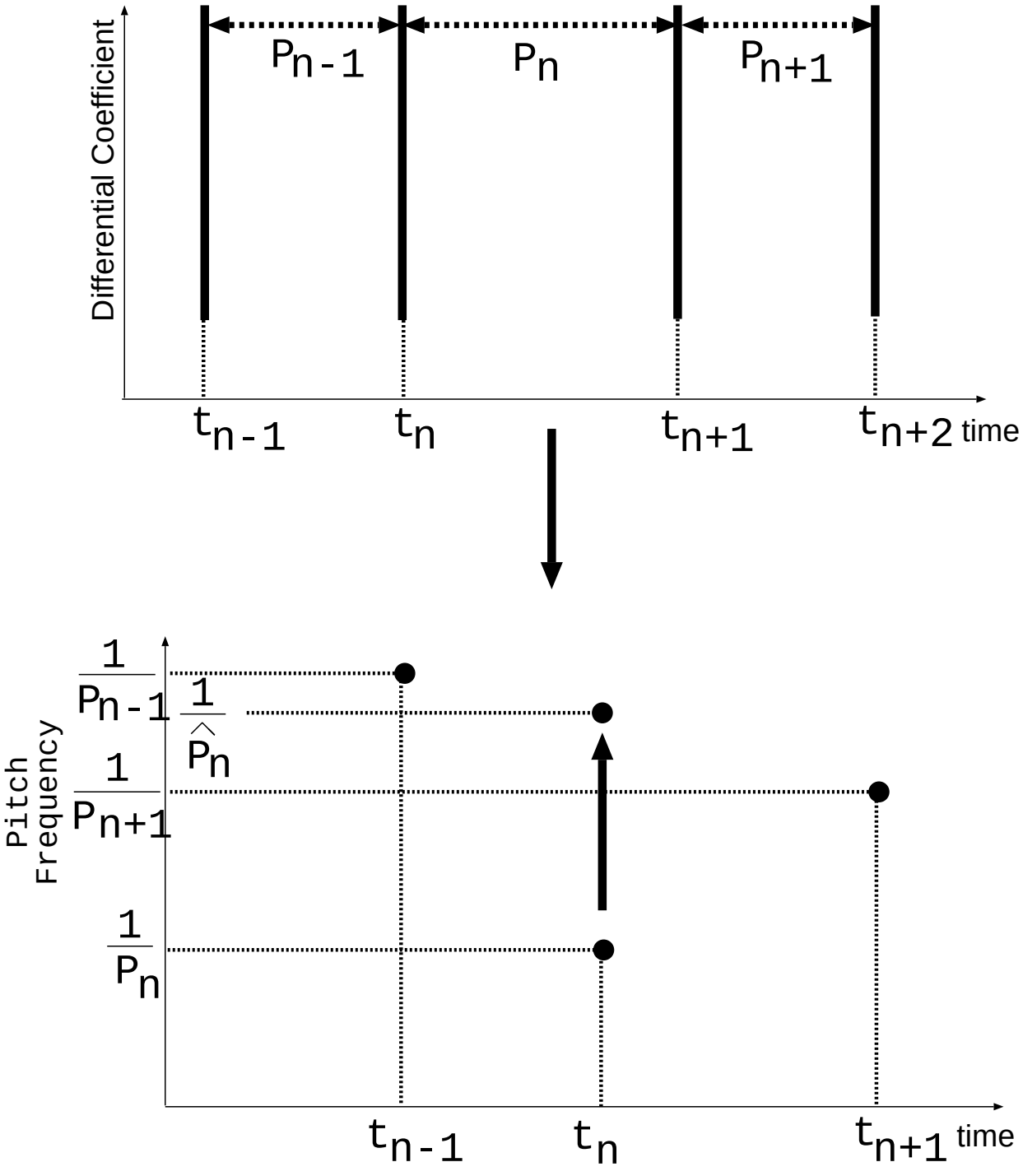

図 2.9: 基本周波数の異常値修正

2.2.4 内挿

前節までの処理により、有声音全体において瞬時基本周波数の抽出が可能となった。その系列の集合に線形補間の処理を行ない、音声の各時刻における基本周波数推定値を求める。図2.10に内挿の処理を示す。異常値修正までの処理が終ると、図中での P_{n-1}、$\hat{P}_n$、P_{n+1} が得られる。ここで時刻 t_{n-1}、t_n、t_{n+1}間を線形補間することで、音声の任意の時刻における基本周波数の推定値が求められる。

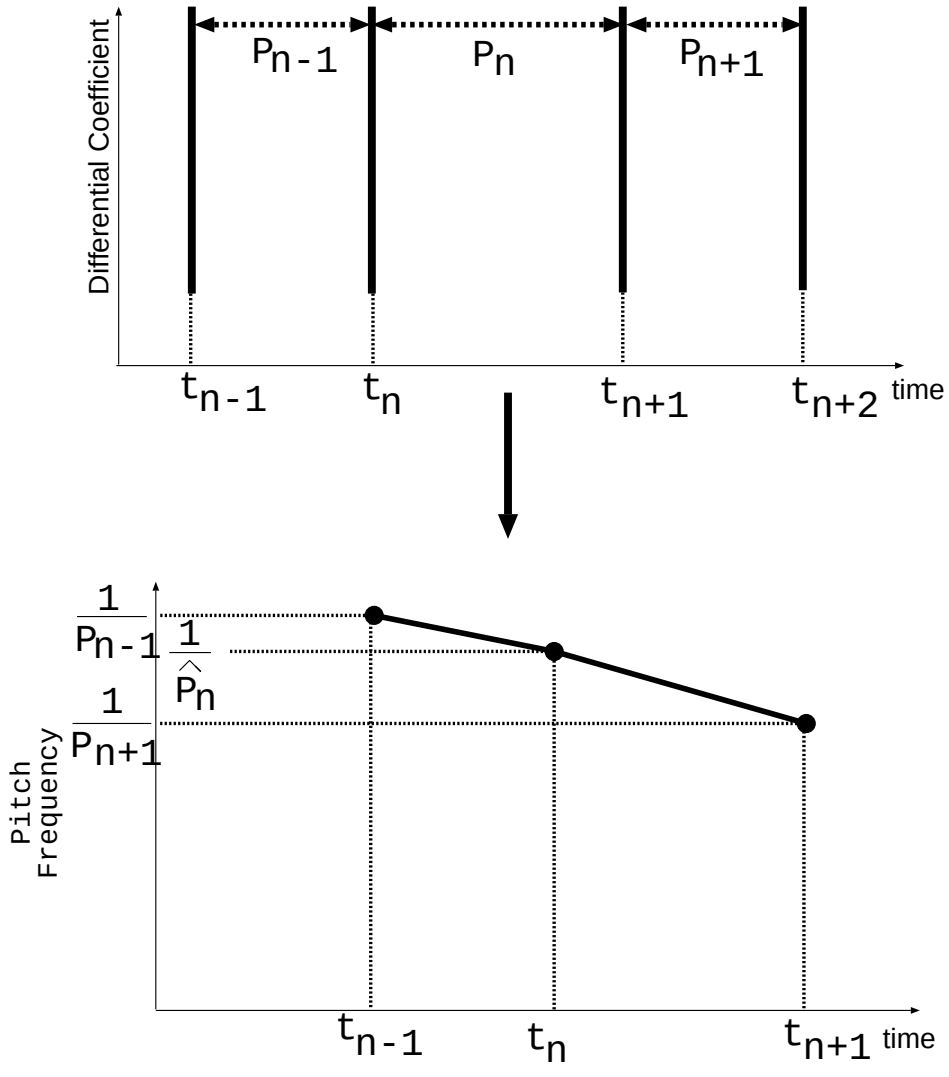

図 2.10: 内挿

2.3 まとめ

Laryngograph の出力を採取した後、瞬時基本周波数の抽出、異常値修正、内挿と経て任意の時刻における音声の基本周波数が得られた。図 2.11に、先に説明した方法で求めた基本周波数の例を示す。この図は男性話者が「あ」を連続発話したときの基本周波数を推定した例である。図 2.11上段が自己相関関数より基本周波数を推定した場合、下段が Laryngograph を用いて音声の基本周波数の推定を行なった結果である。この図から Laryngograph を用いた方が基本周波数の細かい変動を抽出できていることがわかる。

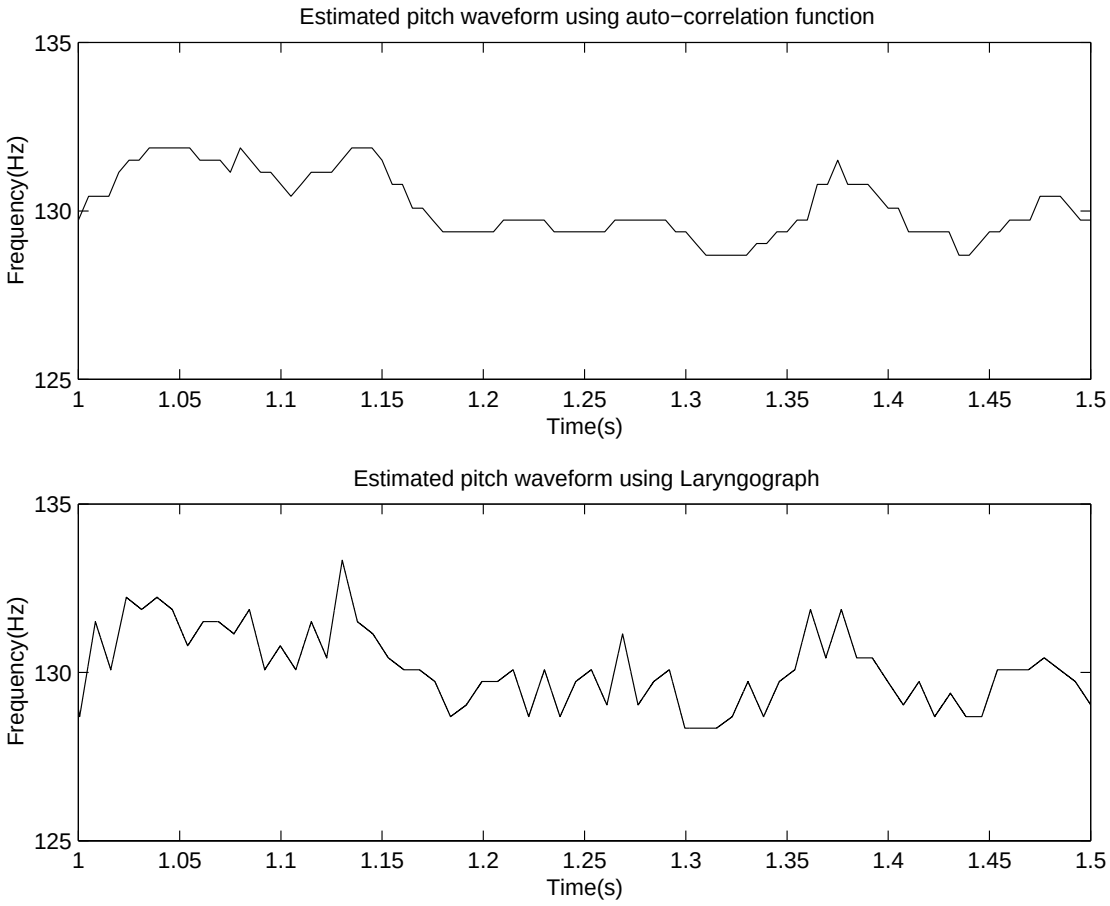

図 2.11: 推定した基本周波数の例

第 3 章

連続発話母音における基本周波数の細かい変動の分析

3.1　分析に用いるデータ

分析に用いるデータは、実際に採取した Laryngograph 出力信号から 2 章で述べた方法を用いて推定した基本周波数の推定値 2 秒間のデータである。この 2 秒間のデータは音声波形が定常である時間帯を視察により確認し、その時間帯の基本周波数推定値を切り出したものである。この切り出した 2 秒間のデータを話者 9 人分、1 人につき 5 母音、計 45 種類用意し、このデータを分析した。

3.2　基本周波数変動の分布

まず、ヒストグラムを用いて、基本周波数変動の分布に個人差が存在するか、ということについて分析を行なった。

3.2.1　ヒストグラム上での差異

図 3.1〜図 3.5は基本周波数のヒストグラムである。このヒストグラムは分析対象である 9 人の話者の各母音の基本周波数推定値から構成した。どの話者の場合も、基本周波数の推定値の最大値と最小値の間を 0.5Hz きざみで分割しプロットしてある。これら 5 つの

図の横軸が周波数で、縦軸が度数である。各話者の母音ごとの基本周波数推定値の平均値と標準偏差は表 3.1〜表 3.5に示す。

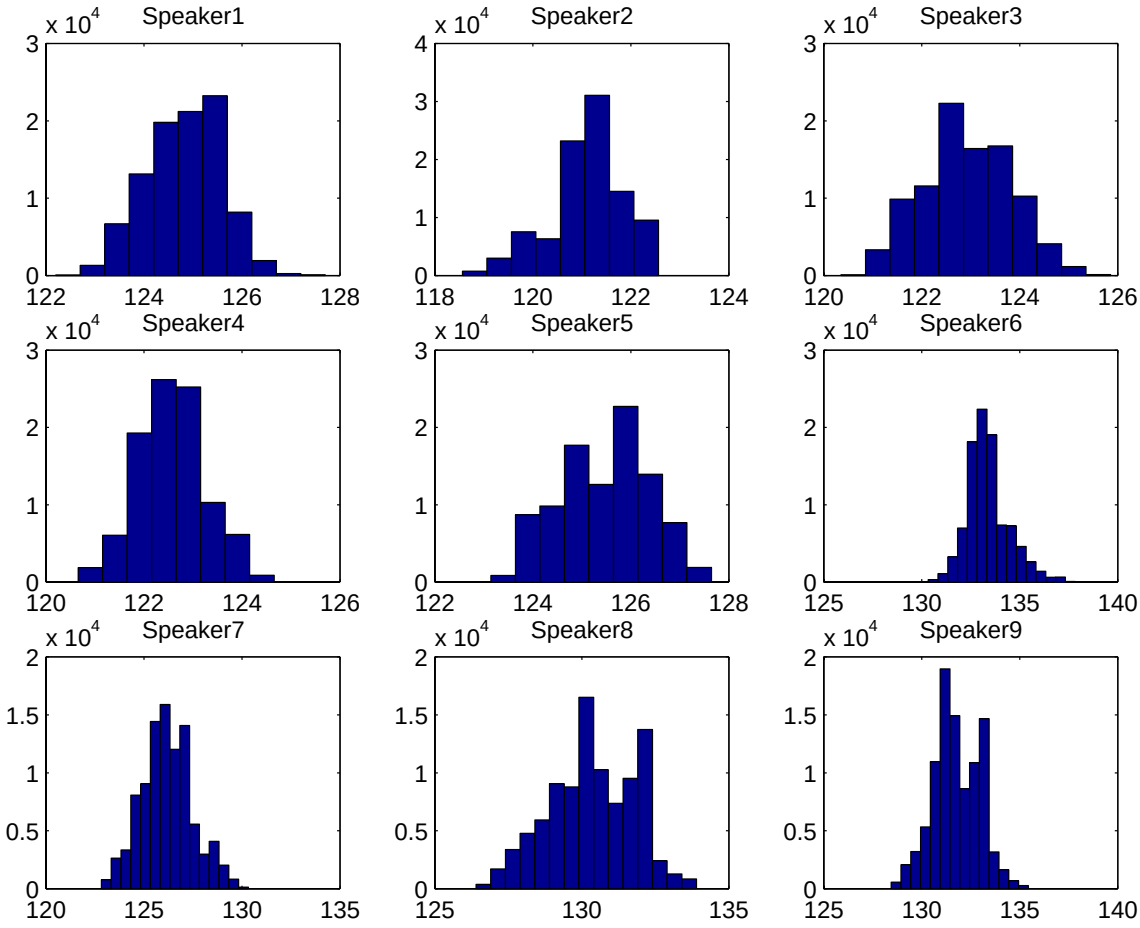

図 3.1: 各話者「あ」の基本周波数推定値 2 秒間のデータから構成したヒストグラム

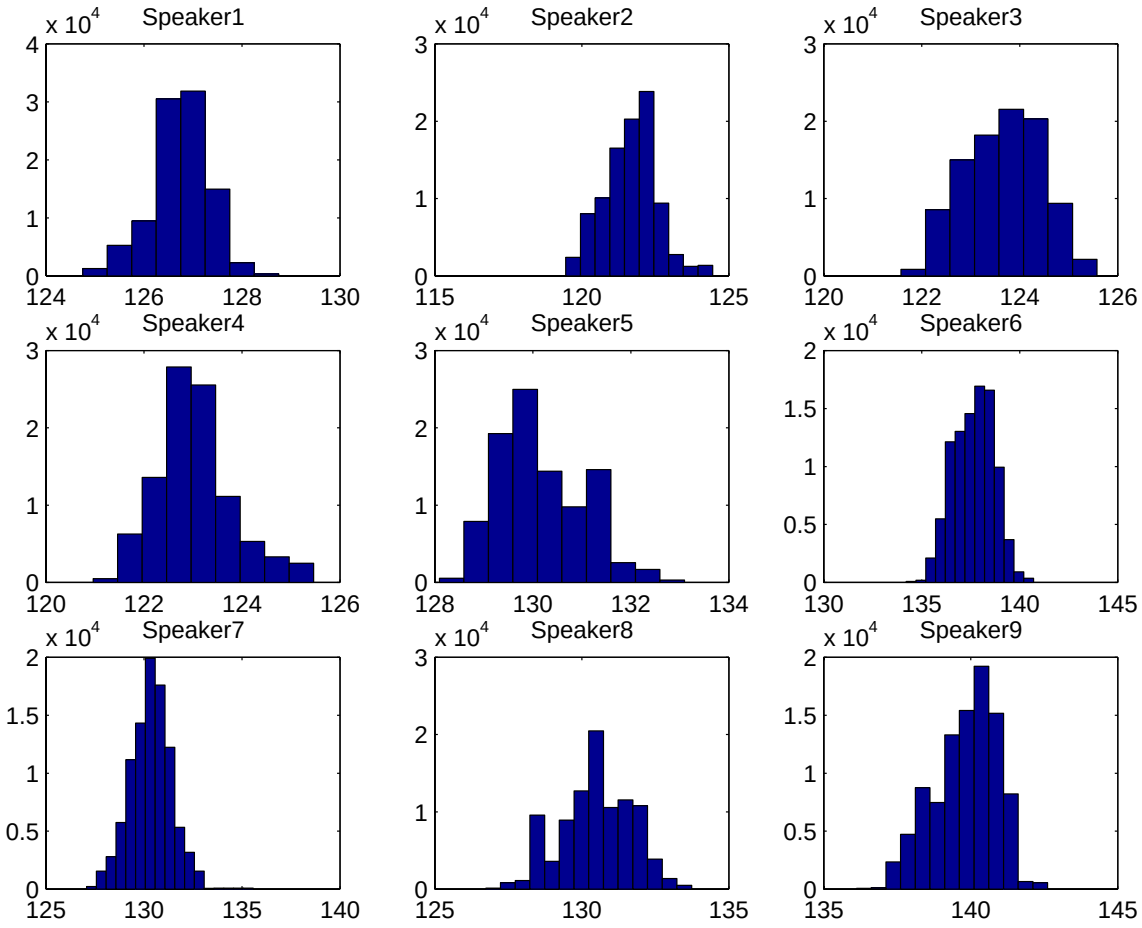

図 3.2: 各話者「い」の基本周波数推定値 2 秒間のデータから構成したヒストグラム

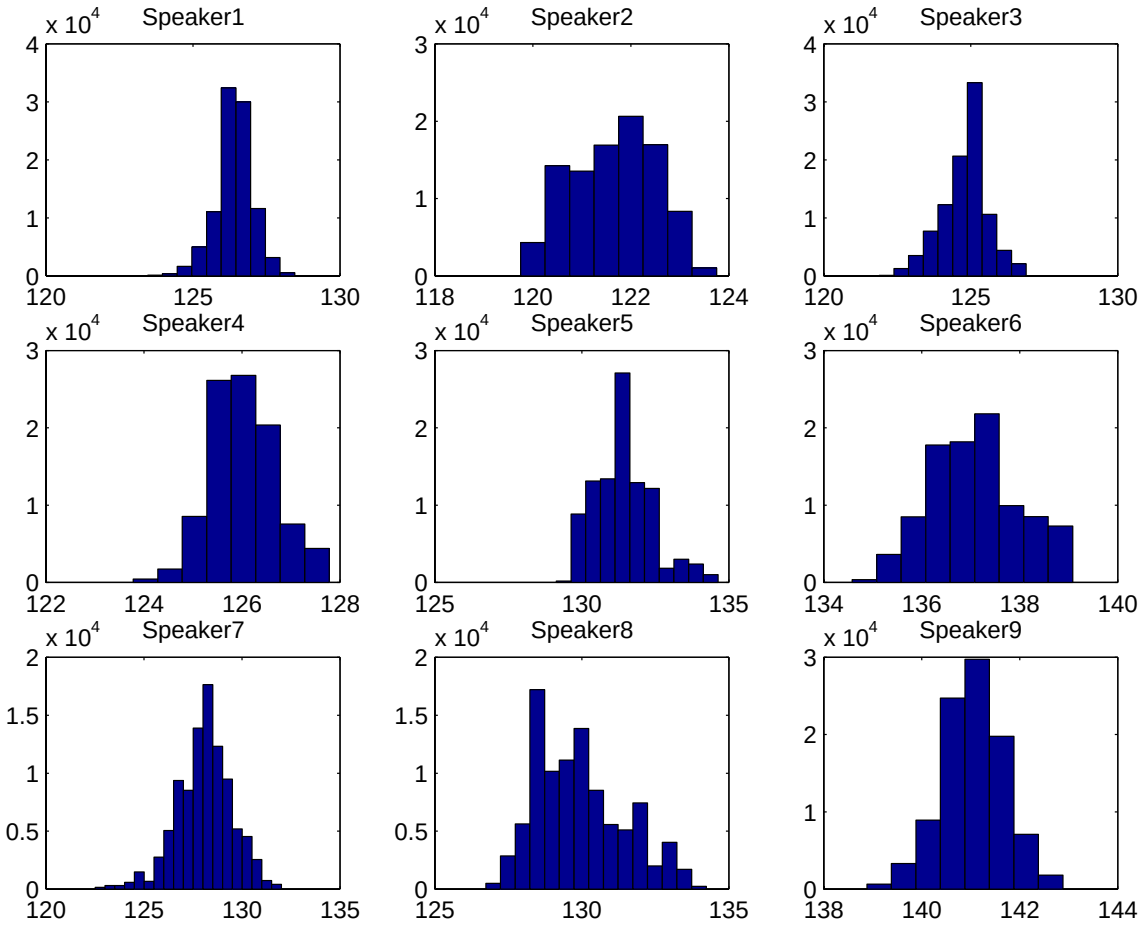

図 3.3: 各話者「う」の基本周波数推定値 2 秒間のデータから構成したヒストグラム

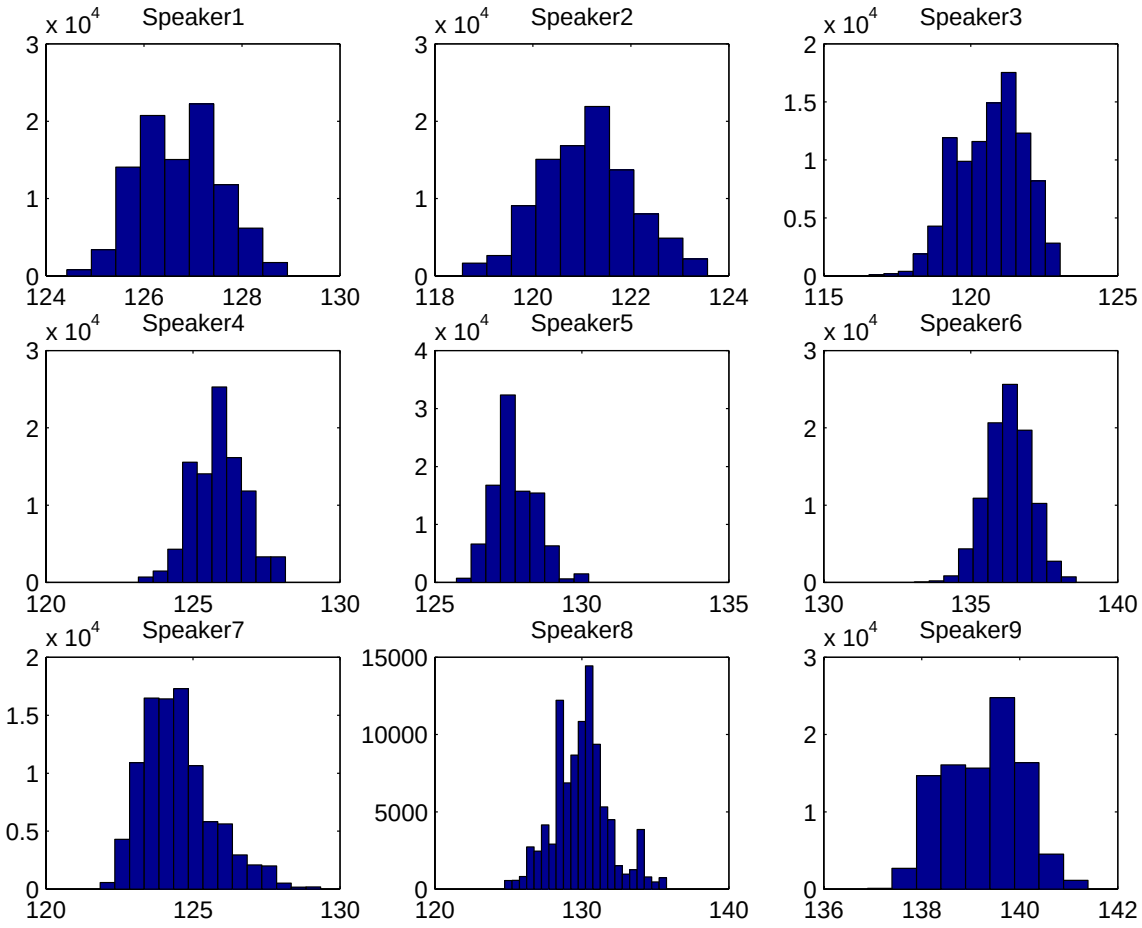

図 3.4: 各話者「え」の基本周波数推定値 2 秒間のデータから構成したヒストグラム

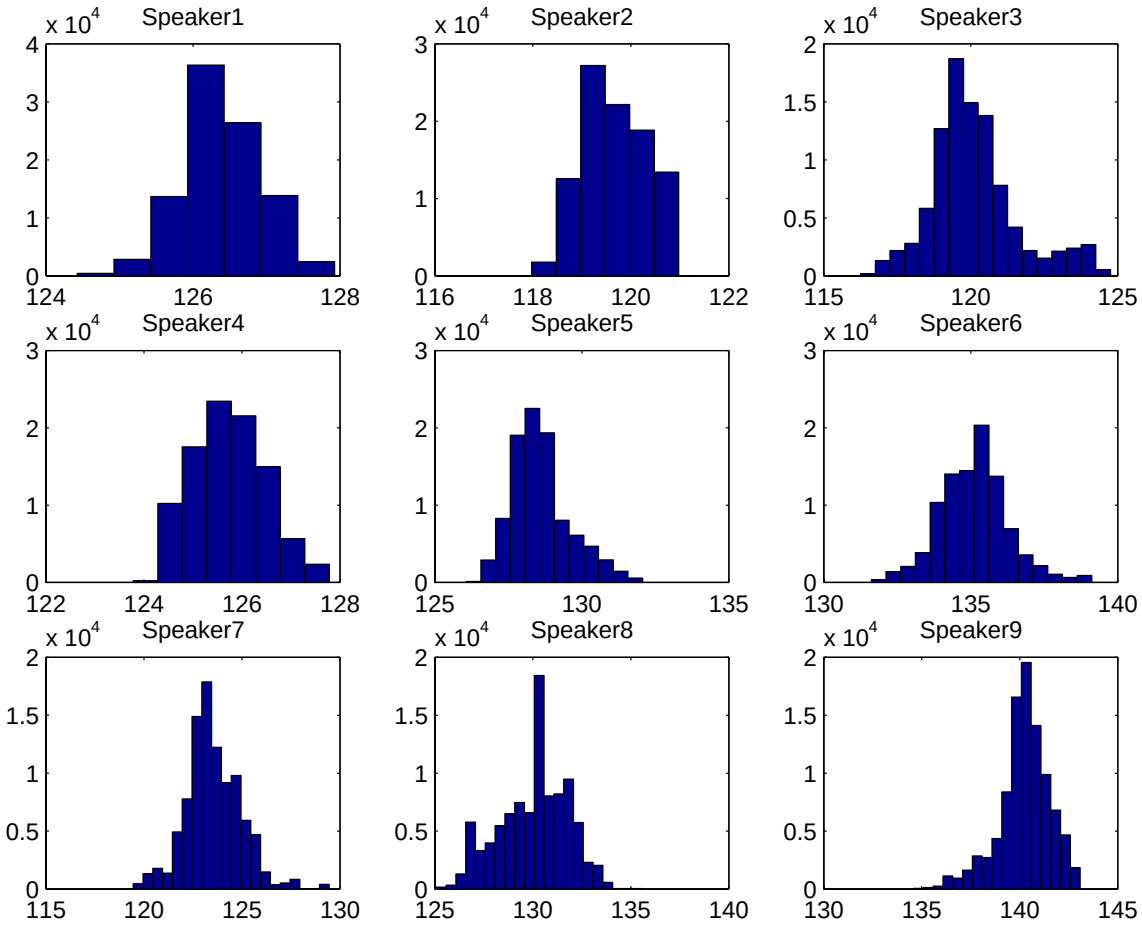

図 3.5: 各話者「お」の基本周波数推定値 2 秒間のデータから構成したヒストグラム

表 3.1: 各話者「あ」の基本周波数推定値の平均値と標準偏差

	平均値 (Hz)	標準偏差 (Hz)
話者 A	124.6887	1.6782
話者 B	121.1034	0.7562
話者 C	122.9154	0.9026
話者 D	122.5690	0.6832
話者 E	125.4538	0.9034
話者 F	133.3756	1.0844
話者 G	126.1885	1.3025
話者 H	130.3683	1.4567
話者 I	131.7784	1.1802

表 3.2: 各話者「い」の基本周波数推定値の平均値と標準偏差

	平均値 (Hz)	標準偏差 (Hz)
話者 A	126.7374	0.5799
話者 B	121.6866	0.8927
話者 C	123.6824	0.7627
話者 D	123.0437	0.7605
話者 E	130.1726	0.8545
話者 F	137.6483	1.0148
話者 G	130.3533	1.0520
話者 H	130.5041	1.2036
話者 I	139.8301	1.0634

表 3.3: 各話者「う」の基本周波数推定値の平均値と標準偏差

	平均値 (Hz)	標準偏差 (Hz)
話者 A	126.3872	0.6268
話者 B	121.6449	0.8361
話者 C	124.8052	0.7786
話者 D	126.0104	0.6530
話者 E	131.4086	0.9651
話者 F	137.1001	0.9120
話者 G	128.0861	1.3873
話者 H	129.9406	1.5181
話者 I	141.0583	0.6289

表 3.4: 各話者「え」の基本周波数推定値の平均値と標準偏差

	平均値 (Hz)	標準偏差 (Hz)
話者 A	126.7213	0.8213
話者 B	121.1022	0.9360
話者 C	120.6749	1.1165
話者 D	125.8519	0.8824
話者 E	127.7229	0.7471
話者 F	136.2824	0.7461
話者 G	124.4818	1.2143
話者 H	129.9933	1.9198
話者 I	139.2496	0.8009

表 3.5: 各話者「お」の基本周波数推定値の平均値と標準偏差

	平均値 (Hz)	標準偏差 (Hz)
話者 A	126.3858	0.5335
話者 B	119.6764	0.6500
話者 C	120.1218	1.4701
話者 D	125.7372	0.7278
話者 E	128.6030	0.9886
話者 F	135.1181	1.1524
話者 G	123.5341	1.4109
話者 H	130.0479	1.7469
話者 I	140.2322	1.2963

3.2.2 考察

音声の基本周波数からヒストグラムを構成すると、その分布に差が生じることがわかった。分布の中心となる値が異なるのはもちろんであるが、中心の周りに分布する様子が各データごとに違う。ただし、これは 1 人の話者についてはどの母音でも同じ特徴が見られるわけではなく、同じ話者でも母音が異なると、分布の形状も異なってくる。同一の話者でも母音によって分布が異なることより、ヒストグラムの分布と度数から話者を特定することはできない。しかし、この分布に差が現れていることは明らかである。

3.3 基本周波数変動の違い

人間が一定の高さで母音を発生し続けた場合、その母音の基本周波数は一定ではなく平均値を中心に揺れている。この揺れの様相を手がかりに基本周波数が、基本周波数の平均値の周りにどのように分布しているか、ということを分析する。

3.3.1 基本周波数の変動

図 3.6に男性話者が「あ」と連続発音した時の基本周波数推定値の例を示す。この図から基本周波数推定値には、細かく山あるいは谷を繰り返す動きと、それとは別に全体として周波数が高く、あるいは低くなる、という動きがあることがわかる。この両者の動きは基本的には各データごとに違う。しかし、グラフを視察することで基本周波数の変動の様相が似ている基本周波数をまとめると、いくつかの集団に基本周波数を分類することが可能であると考えられる。

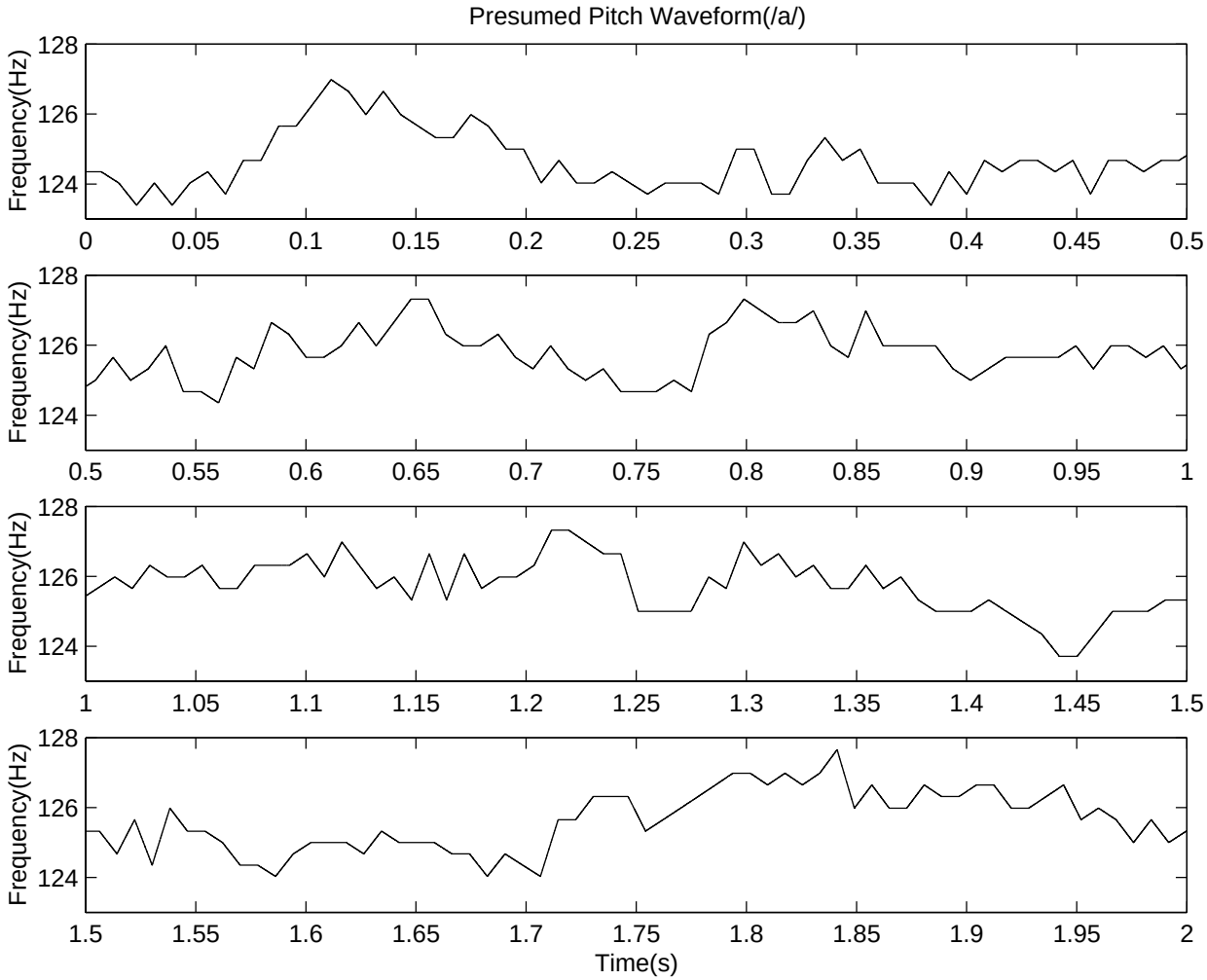

図 3.6: 男性話者による連続発音「あ」の基本周波数推定値

3.3.2 基本周波数のパワースペクトルにおける特徴

基本周波数のパワースペクトルの典型的な例が図 3.7である。

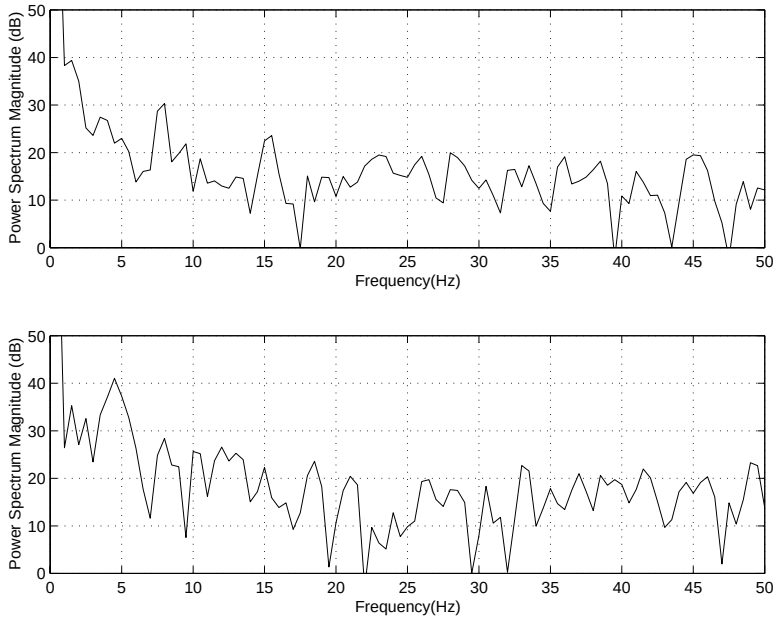

図 3.7: 基本周波数のパワースペクトルの例

この図より、基本周波数のパワースペクトルには 10Hz 近辺に山、あるいは谷が存在することがわかる。この特徴は 45 種類の基本周波数全てに表れる。したがって、パワースペクトルの 10Hz 近辺に山が存在するのか、あるいは谷が存在するのか、という特徴が基本周波数を分類する時の目安になると考えられる。

3.3.3 基本周波数変動に基づく分類

基本周波数の全体的な変動に対応するのは基本周波数変動の低い周波数成分であり、同様に細かい山 (谷) を繰り返す動きは基本周波数変動の比較的高い周波数成分に対応していると考えられる。この変動の様相から基本周波数をいくつかのグループに分類できそうであることは述べた。そこで、基本周波数推定値をパラメータによって分類することを試みる。

ここで用語について定義しておく。以後、基本周波数中の細かな山 (谷) の繰り返しである動きを基本周波数の「細かな変化」、全体的に値が高く、あるいは低くなるような動

きを基本周波数の「緩やかな変化」と呼ぶことにする。

「細かな変化」と「緩やかな変化」を基準として基本周波数を分類するために、基本周波数から両者の片方ずつの成分のみ含まれる波形を抽出する必要がある。そこで、両者を以下のように定義して求めることとした。

- 「緩やかな変化」:基本周波数推定値の変動において 10Hz 以下の周波数成分のみから構成される波形
- 「細かな変化」:基本周波数推定値の変動において 10Hz より大きい周波数成分から構成される波形

この処理は、基本周波数推定値を FFT した後、必要な周波数成分はそのまま残し、不必要な周波数領域の値を 0 とした後に IFFT して時間領域の波形に戻す、というものである。この処理で得られる波形を図 3.8、3.9、3.10に示す。この 3 つの図はそれぞれ話者 A、B、C のものであり、いずれの図も上段が基本周波数推定値の波形、中段が「緩やかな変化」の波形、下段が「細かな変化」の波形である。

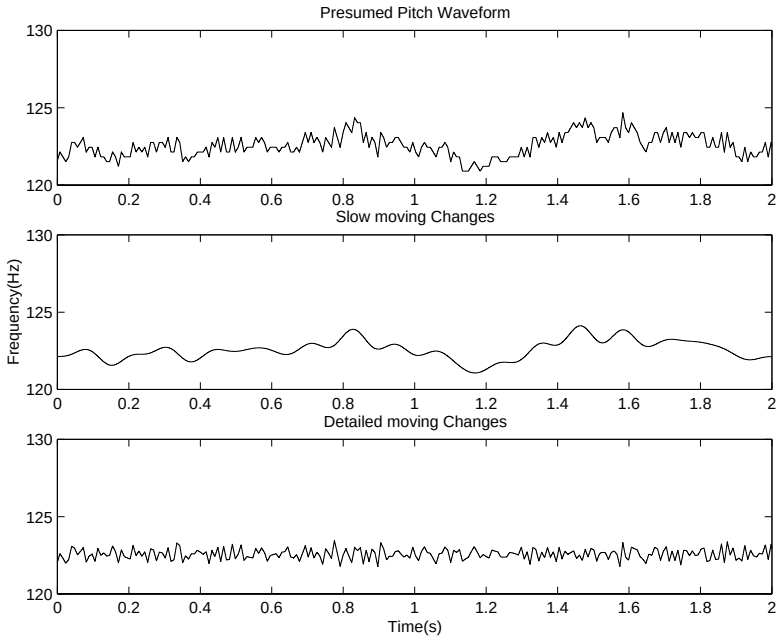

図 3.8: 話者 A(/a/) の基本周波数、「緩やかな変化」、「細かな変化」

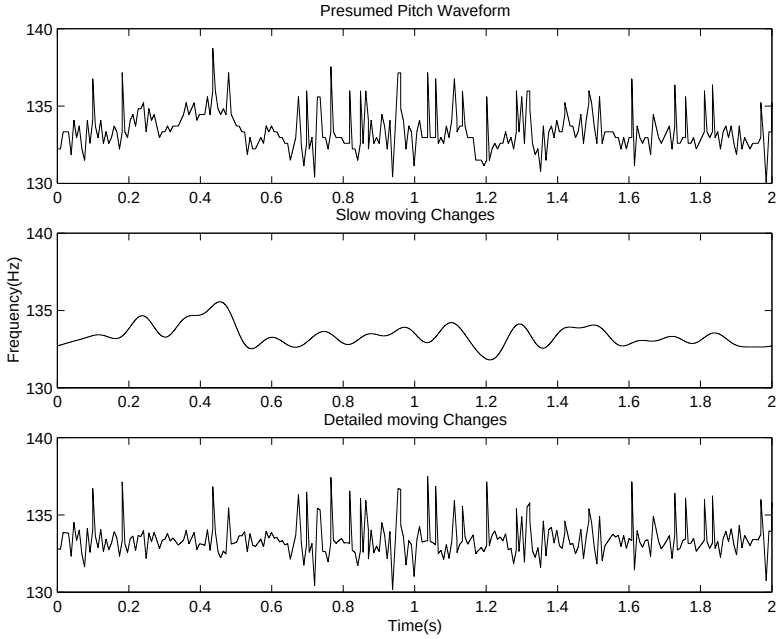

図 3.9: 話者 B(/a/) の基本周波数、「緩やかな変化」、「細かな変化」

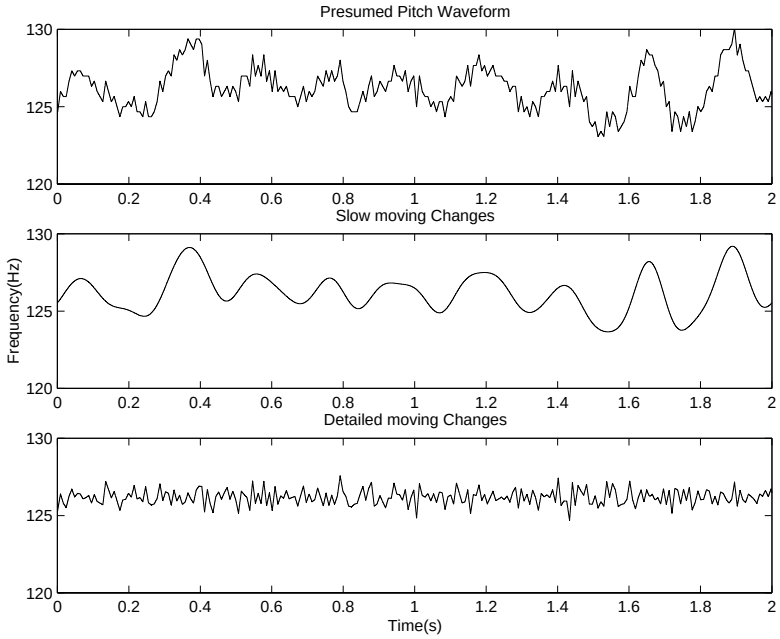

図 3.10: 話者 C(/a/) の基本周波数、「緩やかな変化」、「細かな変化」

基本周波数を分類するために、「細かな変化」、「緩やかな変化」双方の統計学での変動係数 [15] を用いる。変動係数は標準偏差を平均値で割った値である。変動係数が設定した閾値を越えるか、越えないかということで基本周波数を分類することにする。データ総数が少ないので、設定した閾値は暫定的なものであるが、「細かな変化」の閾値を 0.0045、「緩やかな変化」の閾値を 0.0075 として分類を行なう。

この分類により基本周波数は 4 種類の集団に分けることができる。

- 第 1 集団:「細かな変化」が閾値を越えず、「緩やかな変化」も閾値を越えないような基本周波数
- 第 2 集団:「細かな変化」が閾値を越え、「緩やかな変化」は閾値を越えないうな基本周波数
- 第 3 集団:「細かな変化」が閾値を越えず、「緩やかな変化」は閾値を越えるような基本周波数
- 第 4 集団:「細かな変化」が閾値を越え、「緩やかな変化」も閾値を越えるような基本周波数

この基準に沿って分析対象であるデータを分類した結果を表 3.6に示す。この表から、第 4 集団に属する基本周波数がないことがわかる。これはデータ総数の絶対数が少ないためであると考えられる。また、先述した話者 A、B、C「あ」の基本周波数はそれぞれ上記の 1、2、3 の集団に分類される。

表 3.6: 基本周波数の分類

	あ	い	う	え	お
話者 A	1	1	1	1	1
話者 B	2	1	1	1	2
話者 C	3	3	3	3	3
話者 D	1	1	1	1	1
話者 E	1	3	1	3	1
話者 F	1	1	1	3	2
話者 G	1	1	3	1	3
話者 H	3	3	3	3	3
話者 I	3	3	1	1	3

3.3.4 考察

基本周波数の変動の様相から、基本周波数を 4 種類の集団に分類することができた。前節で述べた 4 種類の集団は、「細かな変化」が高周波成分に、「緩やかな変化」が低周波成分に対応すると考えられるので、

- 第 1 集団:基本周波数の変動がほとんどない
- 第 2 集団:基本周波数の高周波成分の変動が大きい
- 第 3 集団:基本周波数の低周波成分の変動が大きい
- 第 4 集団:基本周波数の変動が大きい

と言い替えることができる。ただし、低周波成分、高周波成分が具体的にどのくらいの帯域であるのか、ということについてはさらに分析が必要である。

また、この分類だけで個人を特定することはできない。それは音声の個人性はスペクトル構造や基本周波数の平均値等にも現れるからである。したがって、着目した基本周波数の細かい変動のみで個人を区別する特徴を抽出することは難しい。しかし、基本周波数の変動に差がある以上、この変動の差が音質に影響を与えることは不可避であると考えられる。そして、音質に差が存在するならば、その差を個人識別の手がかりとしていることは十分に考えられる。

第 4 章

聴取実験

4.1　実験の目的

変動係数を手がかりに基本周波数を分類することはできた。そこで、実験 1 として各集団から取り出した基本周波数から合成音を作成し、被験者が基本周波数の変動を基に集団間を知覚できるかということを検討する。その後、実験 2 として各集団ごとの変動を聞き分ける時に、変動のどの帯域に着目しているかということを調べる。

4.2　実験 1

実験 1 について以下で述べる。

4.2.1　実験条件

実験条件を以下に列挙する。

基本周波数データ

第 1、第 2、第 3 の各集団より 1 つずつ取り出す (話者 A、B、C の「あ」の基本周波数)

刺激音

刺激音はKlatt ホルマント合成器により作成した。これは、基本周波数の細かい変動を合成音に反映させるためである。合成時に必要なホルマント周波数と帯域幅は用いる基本周波数に対応する音声のサウンドスペクトログラムより得た。刺激音作成時、基本周波数の変動以外の要素 (ホルマント周波数と帯域幅、基本周波数の平均値) は話者 A のものに固定した。刺激音の長さは2 秒である。また、刺激音の前後部 $50ms$ を sin 関数により重みづけした。

被験者

正常聴力を有する男性 5 人。これらの話者は基本周波数データに用いた 3 名の声に日常から接している。

実験方法

評価させる刺激音は話者 A-A、A-B、A-C、B-B、B-C、B-A、C-C、C-A、C-B の組合せ 9 種類である。この 9 種類を呈示することを 1 回の実験として、呈示する順番を変えて計 5 回の実験を行なった。被験者はヘッドフォンにより刺激音を受聴する。1 対の刺激音を約 1 秒の間隔を置いて呈示し、先に呈示された音声と後に呈示された音声が同じ音声か違う音声かということを回答させた。

4.2.2 実験結果

実験の結果を表 4.1に示す。この表は各被験者が 45 組の刺激音対の先に呈示された音声と後に呈示された音声が同じ音声であるか、異なる音声であるかということを正しく判断した回数である。

表 4.1: 実験 1 の結果

	正当した回数
被験者 1	45
被験者 2	45
被験者 3	44
被験者 4	45
被験者 5	45

4.2.3 考察

実験 1 の結果より、被験者全員が、全ての刺激音の組合せについて違う音、あるいは同じ音であることをほぼ正確に聞き分けているということが考えられる。したがって、分類間の基本周波数変動の差を人は知覚できるということになる。

4.3 実験 2

各集団の基本周波数の変動を人が知覚する上で、重要である帯域を調べるための聴取実験を行なった。

4.3.1 実験条件

実験に用いた基本周波数データ、刺激音、被験者、実験方法、実験システムについて説明する。

基本周波数データ

実験に用いる刺激音を合成する際に用いた基本周波数は、1、2、3 の各集団から 1 つずつ取り出した。いずれも男性話者「あ」の基本周波数推定値 2 秒間分である。

刺激音

刺激音は Klatt ホルマント合成器により作成した。これは、基本周波数の細かい変動を合成音に反映させるためである。合成時に必要なホルマント周波数の値と帯域幅は、用いる基本周波数に対応した音声波形のサウンドスペクトログラムより求めた。刺激音の長さは 2 秒である。また、刺激音の前後部 $50ms$ を sin 関数により重みづけした。以下で 1 つの集団から取り出した基本周波数から作成した合成音の種類について述べる。

- 「Laryngograph」:Laryngograph より推定した基本周波数を用いて合成した音声
- 「Mean」:基本周波数を基本周波数推定値 2 秒間の平均値で一定して合成した音声
- 「$LPF(HPF) : F_c = x$」:Laryngograph より推定した基本周波数の周波数成分を LPF、あるいは HPF を用いて操作した基本周波数を用いて合成した音声で、x はカットオフ周波数である (カットオフ周波数は 10Hz、30Hz、60Hz、100Hz)
- 「LPF セット」:「Laryngograph」、「Mean」、「$LPF : F_c = x$」の計 6 つの合成音

- 「HPF セット」:「Laryngograph」、「Mean」、「$HPF : F_c = x$」の計 6 つの合成音

フィルタを用いて基本周波数の周波数成分を操作する時は前章の緩やかな変化の波形を求める場合と同じ手法を用いている。

被験者

被験者は正常聴力を有する男性 5 人。これらの話者は基本周波数データに用いる 3 者の声に日常から接している。

実験方法

対比較により行なった。2 つの刺激音を約 1 秒の間隔を置いて呈示し、先に呈示された音声と後に呈示された音声が同じ音声に聞こえたかということを 5 段階で評価させる。5 段階の評価は以下のように設定した。

0 完全に違う音に聞こえた時

1 違う音に聞こえるが、全く違うというほどではない場合

2 判断不能の時

3 同じ音に聞こえるが、全く同じというほどではない場合

4 完全に同じ音に聞こえた時

1 回の実験に用いる刺激音は、1 人の話者の LPF セットか HPF セットのどちらか 1 つである。双方のセットとも 6 種類の音声があるので、1 回の実験で評価する刺激音の組は 36 組となる。また、実験結果の信頼性向上のため、1 つの音声セットに対して呈示する順番を変更し、2 度の評価を行なわせた。

被験者は防音室内でヘッドフォンにより受聴した。受聴は各被験者の聞きやすいレベルによる両耳受聴である。被験者には聞き直しを許し、Macintosh を用いて回答させた。

4.3.2 実験システムの概要

聴取実験に用いたシステムの全体図を図 4.1に示す。被験者は防音室内でヘッドフォンより刺激音を受聴し、Macintosh の画面上のボタンをクリックすることで回答する。刺激音が呈示され被験者が回答するまでの間は、Macintosh の HDD は停止するので、HDD によるノイズは発生しない。

このシステムを利用することで、DAT に刺激音を録音し、それを順番に呈示する場合と比べて、呈示音の聞き直しが可能となるため、DAT を用いる場合よりも被験者の精神的疲労の軽減と、聞き逃しによる回答精度の下落を防ぐことが期待できる。また、回答用紙を用いる場合と比べ、被験者自身の過去の回答による影響を取り除くことができる。

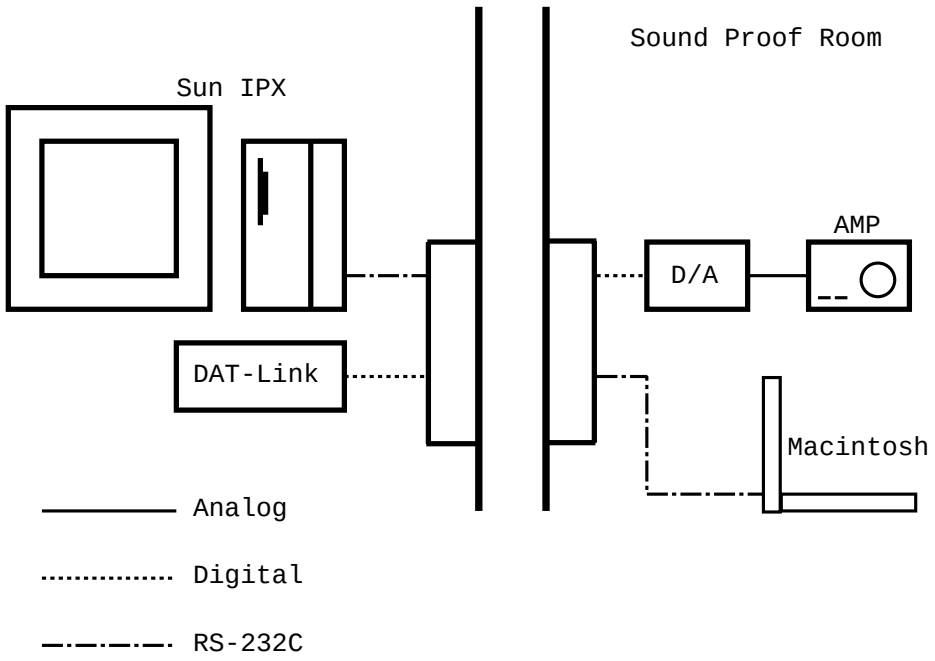

図 4.1: 実験システムの全体図

4.3.3 実験結果

本実験の結果から、各々の刺激音セットの場合について心理距離行列が得られる。それを表4.2、4.3、4.4に示す。この 3 つの表は被験者の回答の平均値となっている。また、左側がLPF セットについての結果、右側がHPF セットについての結果である。表中の「M」は「Mean」、「L」は「Laryngograph」を、数字はLPF セットの図であるなら「$LPF : F_c = x$」を、HPF セットの図であるなら「$HPF : F_c = x$」(x はカットオフ周波数) を示す。

表 4.2: 心理行列:話者 A(第 1 集団)

LPF	M	10	30	60	100	L	HPF	M	100	60	30	10	L
M	3.5	1.0	1.1	0.6	0.7	0.5	M	3.2	3.0	3.5	3.2	2.4	0.4
10	1.2	2.3	2.4	2.2	1.7	1.8	100	2.6	2.6	2.7	2.5	3.3	0.7
30	1.4	1.9	2.1	2.4	1.6	2.5	60	2.7	2.6	2.8	2.8	3.4	0.4
60	0.6	2.6	2.2	2.6	2.0	1.6	30	3.1	2.3	2.7	3.1	3.5	0.7
100	1.2	2.7	2.6	1.4	2.0	2.5	10	2.9	1.9	2.7	2.3	3.0	0.6
L	0.7	1.9	1.8	2.3	2.8	2.2	L	1.1	0.6	0.8	0.4	0.1	2.2

表 4.3: 心理行列:話者 B(第 2 集団)

LPF	M	10	30	60	100	L	HPF	M	100	60	30	10	L
M	4	1.1	0.5	0.4	0	0.4	M	4	3.6	3.6	0.1	0	0
10	1.2	3.2	1.2	1	0.5	0.5	100	3.9	3.9	3.4	0.3	0	0
30	0	0.1	2.6	2.4	2.3	2.4	60	1.4	2.5	2.8	0.1	0	0.4
60	0.3	0.3	1.4	2.9	2.8	2.2	30	0.1	0.2	0.6	2.3	2.7	1.1
100	0.3	0	0.1	2.9	2.6	2.6	10	0.4	0	0	2	3.9	1.8
L	0.4	0.8	0.7	1.8	2.5	2.7	L	0	0	0.3	0.8	2.1	3.3

表 4.4: 心理行列:話者 C(第 3 集団)

LPF	M	10	30	60	100	L	HPF	M	100	60	30	10	L
M	4	1.1	0.5	0.4	0	0.4	M	2.9	2.7	2.9	3.1	3	0
10	1.2	3.2	1.2	1	0.5	0.5	100	2.7	2.4	3.2	3.6	3.1	0
30	0	0.1	2.6	2.4	2.3	2.4	60	2.7	3	3.4	3.2	3.6	0
60	0.3	0.3	1.4	2.9	2.8	2.2	30	2.9	2.6	2.7	2.9	3.2	0
100	0.3	0	0.1	2.9	2.6	2.6	10	3.7	2.3	2.5	2.9	3	0
L	0.4	0.8	0.7	1.8	2.5	2.7	L	0.1	0	0	0	0	3.2

この心理距離行列から、各合成音間の心理距離を把握することは難しい。そこで、上述の心理距離行列をもとに多次元尺度構成法 [16] を用い、各合成音間の距離を算出することにした。合成音群が全体で 6 種類であるので、多次元尺度構成により得られる布置図も 6 種類である。図 4.2から図 4.7 の図がそれぞれ話者 A(第 1 集団) の LPF セット、HPF セット、話者 B(第 2 集団) の LPF セット、HPF セット、話者 C(第 3 集団) の LPF セット、HPF セットを用いた場合の実験結果から得られる、2 次元対象布置図である。なお、2 次元で多次元尺度構成した場合の適合度 (Stress 値) は各図の表題に併記する。

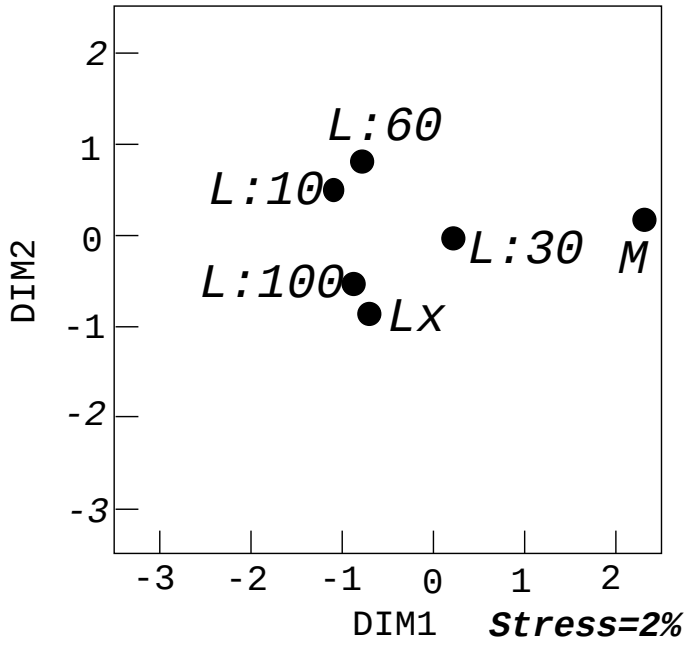

図 4.2: 話者 A(第 1 集団):LPF セットの 2 次元対象布置図 (適合度= 2%)

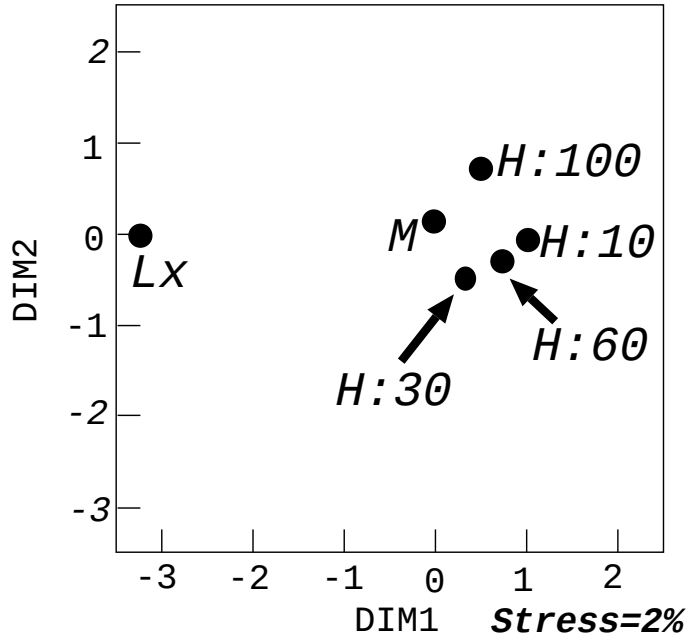

図 4.3: 話者 A(第 1 集団):HPF セットの 2 次元対象布置図 (適合度= 2%)

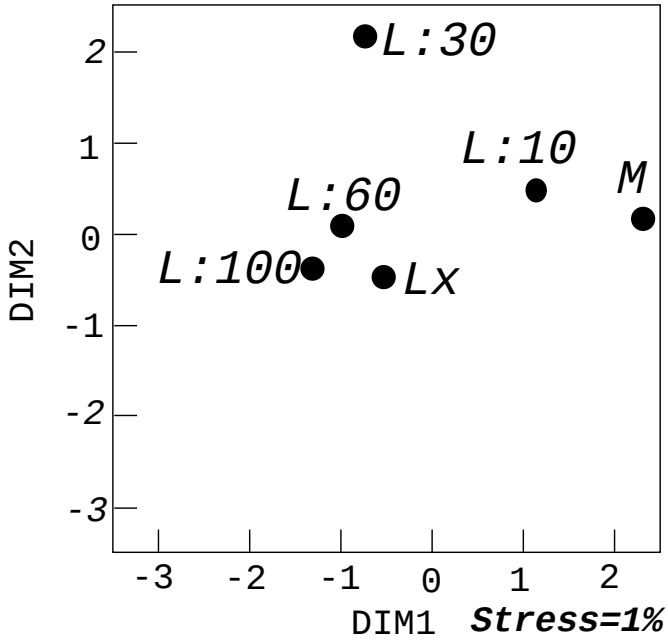

図 4.4: 話者 B(第 2 集団):LPF セットの 2 次元対象布置図 (適合度= 1%)

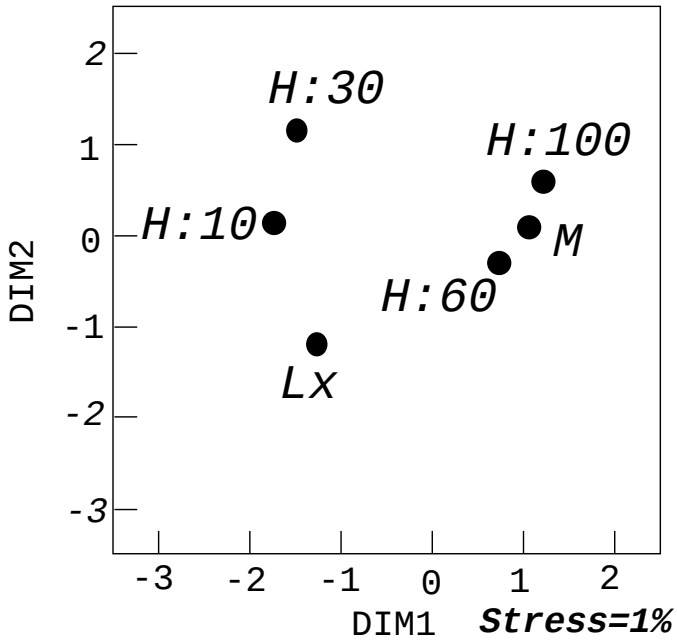

図 4.5: 話者 B(第 2 集団):HPF セットの 2 次元対象布置図 (適合度= 1%)

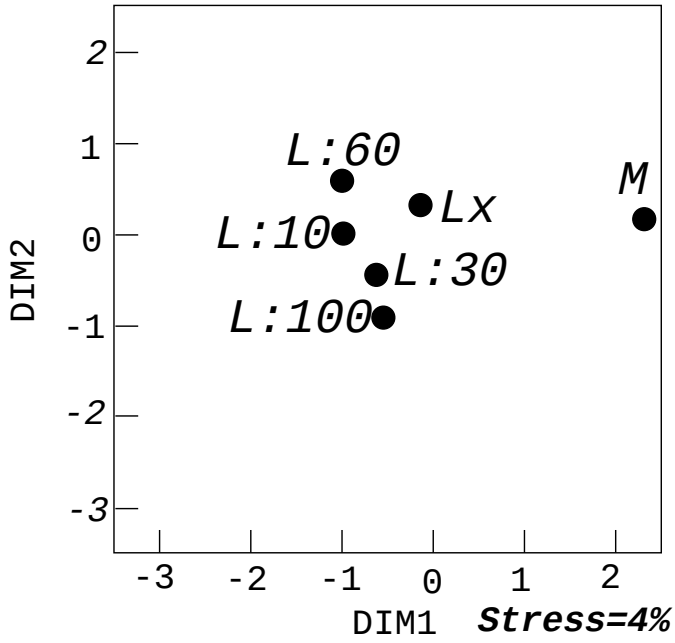

図 4.6: 話者 C(第 3 集団):LPF セットの 2 次元対象布置図 (適合度= 4%)

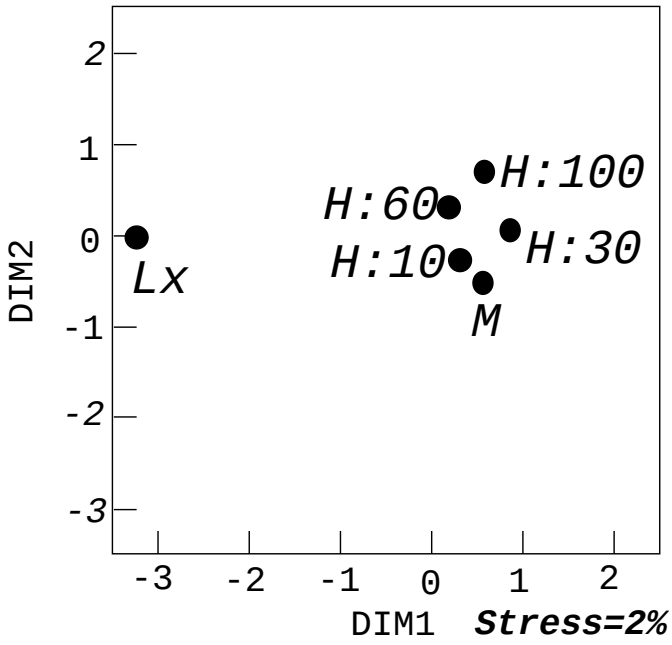

図 4.7: 話者 C(第 3 集団):HPF セットの 2 次元対象布置図 (適合度= 2%)

4.3.4 考察

まず話者 A(第 1 集団) の布置図について考察する。図 4.2の布置からは、「Mean」は他のどの合成音とも距離があり、「$LPF : F_c = 60$」と「$LPF : F_c = 10$」の距離、「$LPF : F_c = 100$」と「Laryngograph」の距離が近く、「$LPF : F_c = 30$」がこの両者のほぼ中間に位置する。「$LPF : F_c = 10$」と「$LPF : F_c = 30$」の距離よりも「$LPF : F_c = 10$」と「$LPF : F_c = 60$」の距離が近く、「$LPF : F_c = 60$」と LPF:「$F_c = 100$」が離れている。これらより、被験者は 10Hz 以下の成分と 60Hz〜100Hz の成分に着目したと考えられる。一方、図 4.3の布置から、「Laryngograph」が他の全ての合成音と距離があり、「Laryngograph」以外の合成音同士が比較的距離が近いところに位置している。「$HPF : F_c = 10$」、「$HPF : F_c = 30$」、「$HPF : F_c = 60$」は距離が近く、これらの一群と「$HPF : F_c = 100$」とはある程度の距離があるが、この両者の「Laryngograph」との距離がほとんど同じである。したがって、被験者が知覚する際、60Hz〜100Hz の成分も着目するものの、10Hz 以下の成分の方により着目したことになる。以上より、人が話者 A「あ」の基本周波数変動を知覚する時、10Hz 以下の成分を主に、60Hz〜100Hz の成分も多少は手がかりとしていると考えられる。

次に話者 B(第 2 集団) について考察する。図 4.4の布置より、「$LPF : F_c = 60$」、「$LPF : F_c = 100$」、「Laryngograph」はおたがいに距離が近い。また、「Mean」が「Laryngograph」

から最も距離があり、カットオフ周波数が高い合成音ほど、「Laryngograph」との距離が近くなっていく。しかし、「$LPF : F_c = 30$」までは「Laryngograph」から比較的離れているものの、「$LPF : F_c = 60$」からは急激に「Laryngograph」との距離が近くなる。つまり被験者は、この音声の周波数変動を知覚する時、30Hz〜60Hz の成分に着目したと考えられる。同様なことが図 4.5からも見てとれる。同図では「Mean」、「$HPF : F_c = 100$」、「$HPF : F_c = 60$」が非常に接近していて、カットオフ周波数が低い合成音ほど「Laryngograph」との距離が縮まっていく。つまり、図 4.4の場合と同様な考察ができる。したがって、被験者が話者 B「あ」の基本周波数変動を知覚する時、30Hz〜60Hz の成分に特に着目していたことになる。

最後の話者 C(第 3 集団) について考察する。図 4.6より、布置が「Mean」と他全ての合成音とで2 分されていることがわかる。また、「Mean」以外の合成音同士は、特にカットオフ周波数によって関連づけられるような布置ではない。一方、図 4.7の布置も、「Laryngograph」と他全ての合成音とで 2 分されていて、「Laryngograph」以外の合成音同士は、カットオフ周波数による関係が見られない。したがって、被験者が話者 C「あ」の基本周波数変動を知覚する時、10Hz 以下の成分に特に着目していると考えられる。

以上より、3 種類の基本周波数変動を知覚する際に被験者が着目した周波数成分は以下のようになる。

- 話者 A「あ」: 特に 10Hz 以下の成分で、60Hz〜100Hz も多少手がかりとした
- 話者 B「あ」: 30Hz〜60Hz の成分を手がかりとした
- 話者 C「あ」: 10Hz 以下の成分を手がかりとした

この結果と前章の分析との対応について考察すると以下のようになると考えられる。

- 細かい変動がほとんどない基本周波数 (第 1、3 集団) の場合、低域 (実験からは 10Hz 以下) の成分が重要で、この帯域の成分のある／なし、ということが基本周波数の変動の知覚に強く影響する
- 細かい変動が大きい基本周波数 (第 2 集団) の場合、相対的に高域 (実験からは 30Hz〜60Hz) の成分が重要で、この帯域の成分がある／なし、ということが、基本周波数の変動の知覚に強く影響する

- 全くの推測であるが、第 4 集団に属する基本周波数は 10Hz 以下の成分と、30Hz〜60Hz の成分が、この場合の基本周波数の変動を知覚する上で強く影響すると考えられる

また、基本周波数変動の 60Hz 以上の成分が削除されていても、基本周波数の変動を知覚する時にはあまり影響しないようである。

第 5 章

結論

5.1 本研究で明らかになったこと

本研究では連続発話母音の基本周波数に存在する細かい変動に含まれる個人性に関する検討を行なった。そのため、基本周波数の推定にあたり、Laryngograph より EGG を得て、平滑化、微分、いき値処理により瞬時基本周波数を抽出、、その後メディアンフィルタによる異常値修正、線形補間を用いた内挿、という処理を経て細かい変動を含んだ、任意の時刻における基本周波数を推定することができた。

このようにして得た基本周波数の推定値を用いて、細かい変動が個人ごとに特徴のある分布をしているかということをヒストグラムを用いて分析した。しかし、ヒストグラムの分布は個人ごとに共通の形状をしていることは認められなかった。ただし、話者や母音の種類に関係ないが、似たような分布になっている基本周波数が存在することが明らかになった。

そこで、基本周波数を「細かな変化」と「緩やかな変化」の 2 つに分離し、双方の標準偏差と平均値より統計学で用いられる変動係数を求め、これを用いると基本周波数をその変動の様相から 4 つの集団に分離することが可能であることが明らかになった。また、分類した集団から 1 つずつサンプルとなる基本周波数を取り出し合成音を作ると、合成音に変動の差による音質の差があることが確認できた。

次に聴取実験より、各集団から 1 つずつサンプルとなる基本周波数を取り出し、基本周波数の変動以外の条件を同一にして合成音を作る。これを被験者に呈示し、同じ音、あるいは違う音に聞こえたかを判別させた。この結果より、サンプルとして選んだ基本周波数

の変動の差は、人が知覚することができるということを示すことができた。さらに、人が基本周波数の揺れを聞き分ける時、どの帯域の周波数成分に着目するかということを実験より明らかにした。その結果、基本周波数の低周波成分の変動が大きい場合は 10Hz 以下の成分に、高周波成分の変動が大きい場合は 30Hz〜60Hz の成分に、人が基本周波数の変動を知覚する際には重要であるという知見が得られた。また、この考察より、音声合成時には基本周波数の 60Hz 以上の成分は合成された音質にあまり影響を与えず、10Hz 以下、あるいは 30Hz〜60Hz の成分は人が知覚できるくらいの差を音質に与えるということが考えられる。

5.2 今後の課題

今後の課題を以下で列挙する。

1 つ目は分析対象のデータ総数の増加である。分析より基本周波数を 4 つの集団に分類することが可能となったが、実際に分類すると 3 つの集団 (第 1、2、3) に全てのデータが分類されてしまい、残りの第 4 集団に属する基本周波数を用いての実験が行なえなかった。また、話者が全員男性であったので女性の声についても同様の分析、実験を行なう必要がある。

2 つ目はより多くのデータを使っての聴取実験である。今回得られた結果が、用いたデータに依存している可能性は否定できない。したがって、一般的となるくらいのデータ数についての結果が必要である。

3 つ目は基本周波数の推定方法の改良である。今回用いた方法ではいき値処理で微分波形のパルス列検出が行なえないと基本周波数の推定ができない。実際に推定ができなかった話者のデータも存在した。今後、女性のデータを扱うことになれば、このようなケースが多くなることが予想される。これについて何かの対策が必要である。

謝辞

常日頃から数多くの有益な御助言、御指導を頂きました赤木正人助教授、岩城護先生、ならびに赤木研究室の皆様に深く感謝いたします。

また、本研究を進めるにあたって、音声と EGG を採取させていただいた本学 9 名の皆様、ならびに聴取実験に協力していただいた 5 名の皆様に厚くお礼申し上げます。

最後に、3 年間の研究生活を支えて下さった家族、友人、私と関わりのあった方皆様に感謝致します。

参考文献

[1] H,Fujisaki and K.Hirose, 'Analysis of voice fundamental frequency contours for declarative sentences of japanese', J.Acoust.Soc.Jpn.(E)5,4,1984

[2] 桑原 尚夫, ' 個人性の音響的特徴とその制御', 音響論集 1-7-11,1993

[3] 桑原 尚夫, 大串 健吾, ' ホルマント周波数・バンド幅の独立制御と個人性判断', 信学論 J69-A,4,pp.509-517,1986

[4] 伊藤 憲三, 斉藤 収三, ' 音声の音響的パラメータが個人性の知覚に及ぼす影響', 信学論 J65-A,1,pp.101-108,1982

[5] A.J.Fourcin 'Normal and pathological speech:phonetic,acoustic and laryngographic aspects', Laryngograph 添付資料

[6] Koike,Y. 'Aplplication of some acoustic measures for the evaluation of laryngeal dysfunction', Studia Phonolgica(Kyoto Univ.), 7:pp17–23, 1973.

[7] Hideki Kasuya,Shigeki Ogawa and Yoshinobu Kikuchi, 'An acoustic analysis of pathological voice and its application to the evaluation of laryngeal pathlogy', Speech Communication, 5, pp171–181, 1986.

[8] 木戸 博, 粕谷 英樹, ' 個人的特徴に関連した声質の表現語', 日本音響学会聴覚研究会資料,H-97-75,1997

[9] 伊福部 達, 橋場 参生, 松島 純一, ' 母音の自然性における「波形ゆらぎ」の役割', 日本音響学会誌 47 巻 12 号,1991

[10] 江原 義郎, ユーザーズディジタル信号処理, 東京電機大学出版局, 1991.

[11] L.R.Rabiner,R.W.Schafer, (鈴木 久喜 訳), 音声のディジタル信号処理 (上), コロナ社, 1983

[12] L.R.Rabiner,R.W.Schafer, (鈴木 久喜 訳), 音声のディジタル信号処理 (下), コロナ社, 1983

[13] Alan V.Oppenheim,Ronald W.Schafer, (伊達 玄 訳) , ディジタル信号処理 (上), コロナ社, 1978

[14] Alan V.Oppenheim,Ronald W.Schafer, (伊達 玄 訳) , ディジタル信号処理 (下), コロナ社, 1978

[15] 中田 和男, 音声, コロナ社, 1995

[16] 橋本 智雄, 基礎課程 統計学, 共立出版株式会社, 1989

[17] 林 知己夫, 飽戸 弘, 多次元尺度解析法, サイエンス社, 1989