文音声中の基本周波数の時間変化に含まれる個人性に関する研究

大野 宏

北陸先端科学技術大学院大学 情報科学研究科

1998 年 2 月 13 日

キーワード： 基本周波数パターン, 個人性, 藤崎モデル.

1 はじめに

今日実用化が望まれている音声を用いたマンマシンインタフェース技術には音声認識、規則音声合成、話者認識等があるが、その実用化のためには音声に含まれる個人性をどのように扱うかが問題となる。そのため個人性に関する研究はかねてより行なわれてきた。基本周波数の動的変化については、過去に単語音声に関してはその個人性について報告されている。しかし、マンマシンインタフェース技術の実現のためには、文音声の個人性情報も明らかにする必要がある。

そこで本論文では、文音声中の基本周波数の時間変化に含まれる個人性について、藤崎モデルによる分析と聴取実験による検討を行なった。実験に使用する基本周波数変形合成音の作成にはSTRAIGHT [1] を用い、藤崎モデルによる基本周波数パターンの変形を施した。聴取実験の結果、文音声において基本周波数パターンが多くの個人性情報を含んでいることが確認できた。また、基本周波数パターンの中では基底周波数と時間構造に多くの個人性が含まれており、個人性知覚において基本周波数の高さを重視する被験者と基本周波数の時間構造を重視する被験者の 2 組が確認できた。そして、時間構造を含めたいくつかのパラメータを入れ換えることにより、話者性の判断を変化させることが出来ることが確かめられた。

2 藤崎モデル

本論文では、基本周波数パターンの操作を行なうために基本周波数記述モデルとして藤崎モデル [2] を採用する。F_0を基本周波数とすれば、藤崎モデルは以下のように定義される。

$$\ln F_0 = \ln F_{\min} + \sum_{i=1}^{I} A_{\mathrm{p}i} G_{\mathrm{p}i}(t - T_{0i}) + \sum_{j=1}^{J} A_{\mathrm{a}j}\{G_{\mathrm{a}j}(t - T_{1j}) - G_{\mathrm{a}j}(t - T_{2j})\}, \tag{1}$$

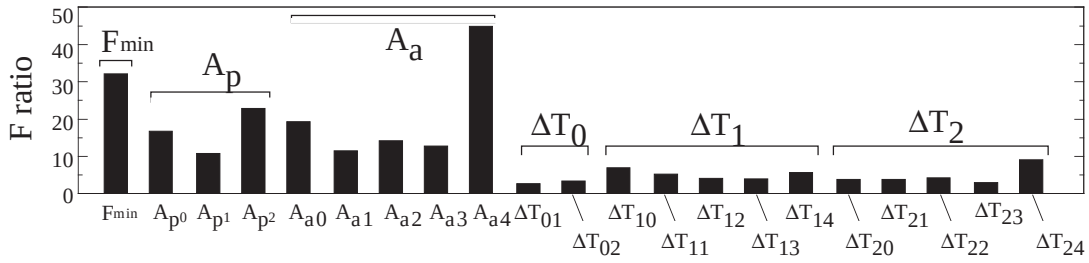

図 1: F 比による分析

$$G_{\mathrm{p}i}(t) = \begin{cases} \alpha_i^2 t \exp(-\alpha_i t) & (t \geq 0), \\ 0 & (t < 0), \end{cases} \tag{2}$$

$$G_{\mathrm{a}j}(t) = \begin{cases} \min[1 - (1+\beta_j t)\exp(-\beta_j t),\, \theta_j] & \\ & (t \geq 0), \\ 0 & (t < 0), \end{cases} \tag{3}$$

ここで $G_{\mathrm{p}i}(t)$ と $G_{\mathrm{a}j}(t)$ は、それぞれフレーズ制御機構のインパルス応答とアクセント制御機構のステップ応答である。式中のパラメータは、$F_{\min}$は基底周波数、I はフレーズコマンド回数 $(i = 0, 1, \cdots I-1)$、J はアクセントコマンド回数 $(j = 0, 1, \cdots J-1)$ である。また、$A_{\mathrm{p}i}$は i 番目のフレーズコマンドの大きさ、$A_{\mathrm{a}j}$は j 番目のアクセントコマンドの振幅 、T_{0i}は i 番目のフレーズコマンドの生成時刻、T_{1j}は j 番目のアクセントコマンドの開始時刻、T_{2j} は j 番目のアクセントコマンドの終了時刻、α_iは i 番目のフレーズ成分の固有角振動数、β_jは j 番目のアクセント成分の固有角振動数、θ_jは j 番目のアクセント成分の天井値である。

3　文音声の基本周波数パターンの個人差の分析

男性 5 名の発話「青い葵が青い屋根の上にある」という文音声について、基本周波数パターンを抽出し藤崎モデルによる分析を行なう。

このように抽出されたパラメータのうち、話者間で個人差の大きいパラメータを調べるためにF比による分析を行なう。F 比は、カテゴリの数を n 、サンプル数を N、第 i カテゴリの第 j サンプルの特徴量を c_{ij} とすると、級間分散と級内分散の比

$$F = \frac{\sum_i^n \left(\bar{c_i} - \frac{1}{n}\sum_i^n \bar{c_i}\right)^2}{\frac{1}{N}\sum_i^n \sum_j^N (c_{ij} - \bar{c_i})^2}, \quad \left(\bar{c_i} = \frac{1}{N}\sum_j^N c_{ij}\right) \tag{4}$$

で与えられ、その値が大きいほどそのパラメータがカテゴリを分類する尺度として有用である。

図 3は、各パラメータの F 比による分析結果である。ここで ΔT_{0i} 、 ΔT_{1j} 、 ΔT_{2j} はそれぞれフレーズコマンドの立上り時間 T_{00} からの相対時間である。

4　スペクトル包絡変換合成音とその個人性知覚

スペクトル包絡変換合成音声を用いた聴取実験により、被験者が基本周波数パターンの違いをどの程度聞き分けられるか調べる。また、基本周波数パターンとして藤崎モデルにより近似した

表 1: 実験条件

話者	5 名 (既知)
被験者	5 名
ヘッドフォン	SENNHEISER HDA 200 (両耳受聴)
ヘッドフォンアンプ	SANSUI AUα-907MR
受聴レベル	約 76 dB (A)

表 2: 実験結果の t 統計量 (合成音声間)

音声サンプル	同じ音声	同じ話者	異なる話者
原音声,TEMPO	1.424	4.079	9.111
原音声, 藤崎モデル	1.585	3.654	9.199
TEMPO, 藤崎モデル	1.187	0.115	0.265

$t_{0.05} = 1.960, t_{0.01} = 2.576$

パターンを用いた場合の個人性知覚への影響を調べる。

実験音声としては、次の 3 種類の音声を用いる。

1. 原音声
2. スペクトル包絡変換音声 (基本周波数パターンは TEMPO により抽出したもの)
3. スペクトル包絡変換音声 (基本周波数パターンは藤崎モデルにより記述したもの)

聴取実験は、聴き直しを許さない環境で対比較の 5 段階評価により行なった。

表 2 は実験に使用した 3 種類の合成音声間に対して、表 3 は呈示組合せ間に対してそれぞれ t 検定を行なった結果である。実験結果より、次の事が確認できた。

- 文音声において、スペクトル包絡を変換した音声でも話者が知覚でき、基本周波数パターンに個人性情報が多く含まれているといえる。
- 藤崎モデルを用いてもその個人性情報はほとんど失われない。

5 各パラメータの個人性知覚への影響

実験には、藤崎モデルによるスペクトル包絡変換音声で話者 a の藤崎モデルのパラメータの一部を話者 b のものに変換した基本周波数変形合成音声を使用する。変換する藤崎モデルのパラメータとしては、以下のものを考える。

1. 基底周波数 F_{min}
2. フレーズ成分 A_{pi}
3. アクセント成分 A_{aj}
4. 時間構造 T_{0i}, T_{1j}, T_{2j}

話者 a・b による藤崎モデルのパラメータの変換パターンには、表 4 のような 8 通りを考えることとする。

表 3: 実験結果の t 統計量 (呈示刺激音声の組合せ間)

音声	同じ音声と異なる話者	同じ話者と異なる話者
TEMPO	41.024	61.221
藤崎モデル	37.722	57.52

$t_{0.05} = 1.960, t_{0.01} = 2.576$

表 4: パラメータの組み合わせ

組み合わせ	A	B	C	D	E	F	G	H
基底周波数	a	b	a	a	a	b	b	a
フレーズ	a	a	b	a	a	b	a	b
アクセント	a	a	a	b	a	a	b	b
時間構造	a	a	a	a	b	a	a	a

実験は、聞き直しを許す環境で ABX 法による聴取実験を行なった。基本周波数に藤崎モデルにより近似した基本周波数パターンを用いたスペクトル包絡変換音声 a・b と藤崎モデルのパラメータの一部を a から b に入れ換えたパラメータ変換音声 x を聴いてもらい、x の音声が a と b どちらの話者の発話による音声と感じるかを強制判断してもらった。

聴取実験の結果を表 5 に示す。これは、話者 a の基本周波数パターンの一部を話者 b のものに変換した基本周波数変形音声 x を聴いて b の話者の音声に近いと答えた割合から、被験者毎に全ての話者の組合せの知覚率を求め平均した結果である。

ただし、×, △, ○, ◎はそれぞれ知覚率が 5 %未満,5〜20 %,20〜40 %,40 %以上を表す。実験結果から次のようなことが分かった。

その結果、次のようなことが確認できた。

- 話者間で極端に大きな差を持ったパラメータが存在するとき、知覚に強く影響する。
- 文音声においては非常に多くの個人性情報を含んでおり、話者を知覚する上で重要な要素となっている。
- 被験者により各パラメータが個人性知覚に影響を与える大きさが異なり、被験者は主に基本周波数パターンの高さや変化を重視するグループと基本周波数パターンのタイミングを重視するグループの 2 組に分けられる。
- 時間構造を含めた 3 つのパラメータを変換することにより、話者の知覚を変化させることが出来る。

これは、単語音声では基本周波数の変化の大きさが被験者の知覚に影響を及ぼしていた [3] のに対し、文音声では基本周波数の高さと変化の大きさも重要であるものの、時間構造の重要性が増し基本周波数の動きが知覚に強く影響しているものと考えられる。また、音響特徴の差が個人性の知覚に反映されるという結果も過去の報告 [4] とも一致している。

表 5: 実験結果 (パラメータの組み合わせ毎の知覚率)

組み合わせ	A	B	C	D	E	F	G	H
被験者 1	×	△	○	△	○	◎	○	◎
被験者 2	×	○	○	△	○	◎	○	◎
被験者 3	×	○	△	△	◎	◎	○	◎
被験者 4	△	○	△	△	◎	○	○	○
被験者 5	×	△	×	△	◎	△	○	○
合計	×	○	△	△	◎	○	○	◎

6 おわりに

文音声の基本周波数パターンに含まれる個人性情報を調べるため、藤崎モデルを用いてパラメータの抽出を行ない、そこに現れる個人差について分析を行なった。また、文音声において基本周波数パターンが多くの個人性情報を含むことを聴取実験により確かめた。そして、実際に基本周波数パターンのパラメータを変更することにより、個人性知覚に及ぼす影響について調べ、検討を行なった。

その結果、基本周波数パターンには個人性が含まれていることが確かめられた。また、単語の場合と同様に高さとその変化に個人性が含まれているが、その時間構造にも多くの個人性が含まれていることが明らかになった。

今後は、より多くの音声、被験者に対して同様の検討を行なう必要がある。

参考文献

[1] 河原：“聴覚の情景分析と高品質音声分析変換合成法 STRAIGHT”，音響学会論文集，pp.189-192 (1997)

[2] H. Fujisaki and K. Hirose: “Analysis of voice fundamental frequency contours for declarative sentences of Japanese”, J. Acoust. Soc. Jpn. (E) **5**, 4 (1984)

[3] M. Akagi and T. Ienaga:“Speaker individualities in fundamental frequency contours and its control”, J. Acoust. Soc. Jpn. (E) **18**, 2 (1997)

[4] 橋本, 樋口：“ Analysis of acoustic features affecting speaker identification”, Eurospeech'95, pp.435-438 (1995)