JAIST Repository
https://dspace.jaist.ac.jp/

Title	文音声中に含まれる個人性情報の知覚に関する研究
Author(s)	鈴木，教郎
Citation	
Issue Date	1999-03
Type	Thesis or Dissertation
Text version	author
URL	http://hdl.handle.net/10119/1258
Rights	
Description	Supervisor:赤木 正人，情報科学研究科，修士

JAIST
JAPAN
ADVANCED INSTITUTE OF
SCIENCE AND TECHNOLOGY

Japan Advanced Institute of Science and Technology

文音声中に含まれる個人性情報の知覚に関する研究

鈴木 教郎

北陸先端科学技術大学院大学 情報科学研究科

1999 年 2 月 15 日

キーワード： 話者の個人性, STRAIGHT, テンポラルデコンポジション.

1 はじめに

音声には話者の特徴すなわち個人性が含まれており、人間は、これを話者の識別に利用している。しかし、人間が音声中のどのような物理量を個人性として知覚するのか、ほとんどあきらかになっていない。

話者知覚に利用される物理量が明らかになれば、text-to-speech 合成音声や音声モーフィングなどさまざまな音声処理技術に応用することができる。

個人性に関する研究は主にスペクトルと基本周波数の 2 つの物理量を中心に行われてきた。家永ら [1] は基本周波数の時間変化パターンに着目し、基本周波数の時間変化パターンに個人性が多く含まれていることを報告し、北村ら [2] はスペクトル遷移パターンの個人性知覚に与える影響を調べ、スペクトル遷移パターンが話者識別に与える影響は少ないことを報告した。

これらの研究では、基本周波数の時間変化とスペクトルの動きの両方統合された変化に関して検討されていない。スペクトル、基本周波数、これらの変化を同時に制御して、話者知覚への関連性を調べる必要がある。

そこで、本研究では、文音声中に含まれる個人性とその物理関連量を調べるために、文音声のスペクトル、基本周波数、これらの変化について、STRAIGHT と S^2BEL-TD を用いて分析し、物理量を他話者のものと入れ替えた刺激音を用いて聴取実験による検討をおこなった。用いた物理量はスペクトル、基本周波数、そしてそれらの変化情報であり、これらを同時に変形することにより関連を調べた。

2 STRAIGHT

分析合成には STRAIGHT[3] を用い、分析条件は、表 1 に示す通りである。

STRAIGHT で得られるパラメータはスペクトル包絡であるため、これを補間特性のよいパラメータに変換する必要がある。そこで、線形補間性に優れている Line Spectral Frequncies(LSF) を用いる。

Copyright © 1999 by Suzuki Norio

表 1: 分析条件

分析窓長	40ms
分析シフト幅	1ms

3 S^{2}BEL-TD

時間変化を記述するためにスペクトルの時間変化パターンをモデル化する必要がある。そこで下記に示す理由により S^{2}BEL-TD[4] を採用した。

- S^{2}BEL-TD はスペクトルパラメータをスペクトルの時間変化パターンとイベント位置におけるスペクトル情報に置き換えることができる。

ここで、イベント位置とは、スペクトル変化の安定点である。

S^{2}BEL-TD で時間分解されるスペクトルパラメータ$\hat{\mathbf{y}}(n)$ は次式であらわされる。

$$\hat{\mathbf{y}}(n) = \sum_{k=1}^{K} \mathbf{a}_k \phi_k(n), \quad 1 \leq n \leq N \tag{1}$$

ここで、$\mathbf{a}_k$と$\phi_k(n)$ は、それぞれ k番目のイベントダーゲット、とイベントファンクションである。これにより、スペクトルパラメータは、イベント位置のスペクトル $\mathbf{a}_k$を重みとした時間変化パターン$\phi_k(n)$ の線形和として表される。

イベントファンクション$\phi_k(n)$ の意味するところは、スペクトルの変化の安定点 kから次の安定点 $k+1$ に時間が移動する時に、前後のスペクトルの混合する割合を時間的に示したものである。

また、STRAIGHT で得られた基本周波数 $p(n)$、振幅成分 $g(n)$ はイベントファンクション$\phi_k(n)$ と基本周波数ターゲット、振幅ターゲットを用いて次のように再現することができる。ただし、基本周波数、振幅成分は対数変換をおこなったものを用いた。

$$\hat{p}(n) = \sum_{k=1}^{K} p_k \phi_k(n), \quad 1 \leq n \leq N \tag{2}$$

$$\hat{g}(n) = \sum_{k=1}^{K} g_k \phi_k(n), \quad 1 \leq n \leq N \tag{3}$$

ここで、$\hat{p}(n)$ と p_kは、それぞれ再現された基本周波数と k番目の基本周波数ターゲットであり、$\hat{g}(n)$ と g_kは、それぞれ再現された振幅成分と k番目の振幅ターゲットである。

さらに S^{2}BEL-TD で得たパラメータを話者間で入れ替えを行う為また物理的距離を求める為に、話者間でイベントの発生位置の対応付けが必要である。

このため、1 つの音韻に対して 1 つのイベントを取ることにする。このことにより、話者が違っても同じ文章なら同じ音韻数を取ることができ、イベントの対応がつくことになる。

4 分析

文音声についてのスペクトルの時間変化パターンを S^{2}BEL-TD により抽出し、抽出された各パラメータに現れる話者間の物理的距離について分析を行う。

音声データ　音声データに用いた文章は、声帯振動の伴う母音または有声子音で構成した。分析に用いた音声データの文章には「いいえ、うえにある」を採用した。発話者は男性 5 名である。

録音は、防音室で行った。マイクロホン (SONY C-536P) からの距離を約 15cm に保って発話させた音声を DAT レコーダ (SONY TCD-D10 PRO II) に入力しサンプリング周波数 48kHz で録音した。この音声を 8kHz にダウンサンプリングして WS に保存した。

分析パラメータ　分析に使用するパラメータは、スペクトルの時間変化パターン$\phi_k(t)$ とイベント位置のスペクトルパラメータ LSF(30 次)、基本周波数、振幅成分 (LSF 残余) である。

4.1　分析の手法

文音声データから抽出した各パラメータから各話者間での距離を求め、それを元に多次元尺度構成法 (MDS) により 3 次元または 2 次元分布を求める。

スペクトルパラメータの話者間の物理的な距離 **CD** については、LSF を LPC 係数に、LPC 係数を LPC ケプストラムに変換し、イベント位置のケプストラム距離 ($\mathbf{cd}_k$) を求め、全イベント数 K の平均として求めた。

$$\mathbf{cd}_k = D_b\sqrt{2\sum_{i=1}^{p}(c_{ik}^{(x)} - c_{ik}^{(y)})^2} \tag{4}$$

$$\mathbf{CD} = \frac{1}{K}\sum_{k=1}^{K}\mathbf{cd}_k \tag{5}$$

ここで、p はケプストラム次数 (30 次)、$c_{ik}^{(x)}, c_{ik}^{(y)}$は k番目のイベントに対する話者 x と話者 y のケプストラム係数、D_bは距離尺度をデシベルに変換するための定数で、$D_b = 10/\ln 10$ である。

また、基本周波数、振幅も同様に話者間の物理的距離 $\mathbf{DIST_{F0}}$、$\mathbf{DIST_g}$を求める。

$$\mathbf{DIST_{F0}} = D_b\sqrt{\sum_{i=1}^{K}(p_i^{(x)} - p_i^{(y)})^2} \tag{6}$$

$$\mathbf{DIST_g} = D_b\sqrt{\sum_{i=1}^{K}(g_i^{(x)} - g_i^{(y)})^2} \tag{7}$$

ここで、$p_i^{(x)}, p_i^{(y)}, g_i^{(x)}, g_i^{(y)}$はそれぞれ話者 x と話者 yの基本周波数、振幅の係数である。

時間変化パターンの話者間の距離は、文音声中の各音韻長の話者間の距離 $\mathbf{DIST_t}$を求め (式 (8))、それを元に MDS から 2 次元分布を求める。

$$\mathbf{DIST_t} = \sqrt{\sum_{i=1}^{K}(s_i^{(x)} - s_i^{(y)})^2} \tag{8}$$

ここで、$s_i^{(x)}, s_i^{(y)}$は話者 x、話者 y の i 番目の音韻長である。

分析結果　図 1 にスペクトルと基本周波数のの 3 次元および 2 次元により付置図を示す。

図 1 のスペクトル距離の付置図 (左) から、各話者から離れている話者は話者 a、e であった。また、話者 b、d はほぼ位置にあることがわかる。図 1 の基本周波数の距離の付置図 (右) では、平均から最も遠い話者は話者 a であり、話者 b、d はほぼ同じ位置にあることがわかる。

図 2 に、振幅、時間変化パターンの 3 次元および 2 次元により付置図を示す。

図 2 の振幅距離の付置図 (左) より、平均から一番近い話者は話者 c である。時間変化パターンの分析結果 (図 2) では、各話者とも分散が大きく混在している。

図 2 の音韻長の距離の付置図 (左) では、話者が混在している。

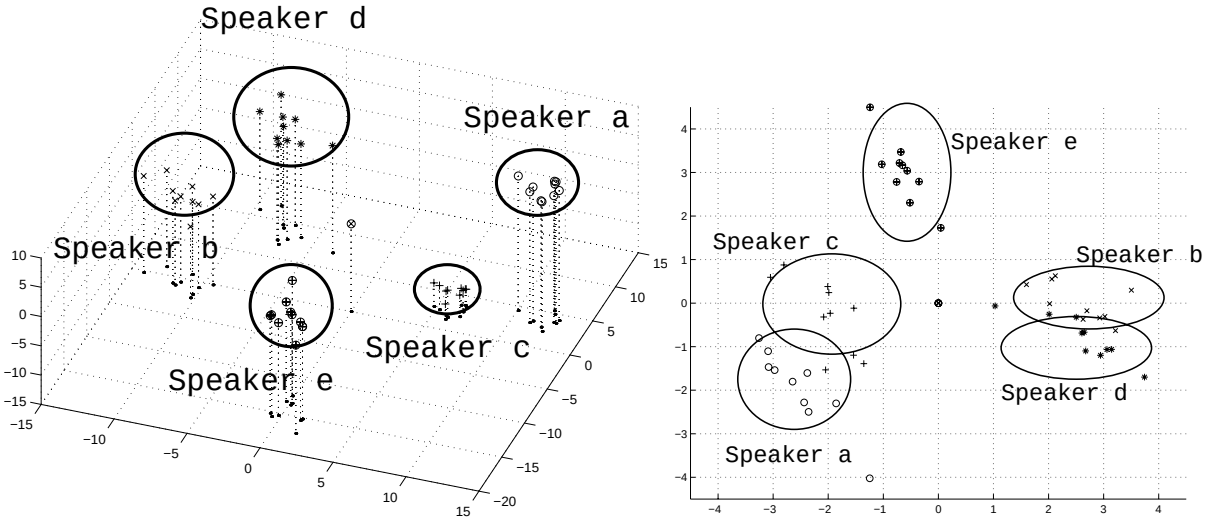

図 1: スペクトル距離の付置図 (左)、基本周波数 (右) の付置図 (話者 a:○、話者 b:×、話者 c:+、話者 d:∗、話者 e:⊕、話者平均:⊗)

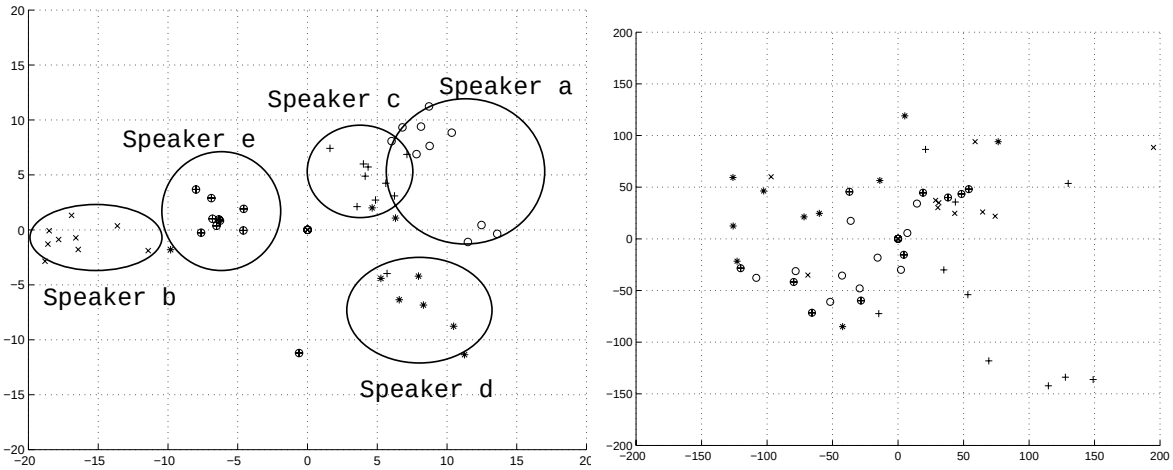

図 2: 振幅距離の付置図 (左)、音韻長距離の付置図「いいえ、うえにある」(右)(話者 a:○、話者 b:×、話者 c:+、話者 d:∗、話者 e:⊕、話者平均:⊗)

5 聴取実験

5.1 実験 1 存在確認

時間変化パターンを考慮した合成音声に個人性が存在することを確認する。

5.1.1 実験方法

音声データ

4 節で分析した音声データのうち、各話者 3 データを利用した。このデータを実験 2、3 でも用いる。

刺激音

聴取実験には以下の 4 種類の刺激音を用いた。

1-A. 原音声
1-B. STRAIGHT 分析合成音声
1-C. STRAIGHT で得られたスペクトルをスペクトルパラメータ (LSF30 次) まで分解し合成した合成音声
1-D. 1-C において、スペクトルパラメータを S^2BEL-TD を用いて時間構造、イベント位置に対するスペクトルパラメータや基本周波数などに分解し合成した合成音声

被験者

被験者は正常聴力を有し、音声データの収録の対象とした話者と日頃接している 22 歳から 36 歳の男性学生 10 名とした。この被験者 10 名は実験 2、3 でも同じである。

実験方法

実験は Naming 法により行った。一刺激につき 1 セットとし、計 4 セット行った。刺激音 1-A は、一話者につき 6 回、計 30 回でランダムに呈示した。刺激音 1-B〜D については、一話者につき 9 回、計 45 回をランダムに呈示した。被験者は防音室内でヘッドホンから各話者の聞きやすいレベルで両耳受聴した。被験者には聞き直しを許し、強制判断させた。実験方法は実験 2、3 でも同様におこなった。

結果と考察　実験結果は、話者知覚できた割合を知覚率として図 3 に示す。

刺激音 1-B だけ知覚率が 99.8%と低くなった (他は 100%) が、刺激音 1-A と刺激音 1-B の話者知覚率に有意差があるか否かを有意水準 5%で F 検定を行ったところ、有意差が認められなかった (図 3)。これから刺激音 1-D は話者聴取実験に用いるのに十分な品質を有しているといえる。

5.2 実験 2 各パラメータと個人性知覚

スペクトル、基本周波数、振幅のうちどのパラメータが話者知覚に大きく影響するかを調べる。時間変化パターンについては個人性があるという前提でそのまま使用する。

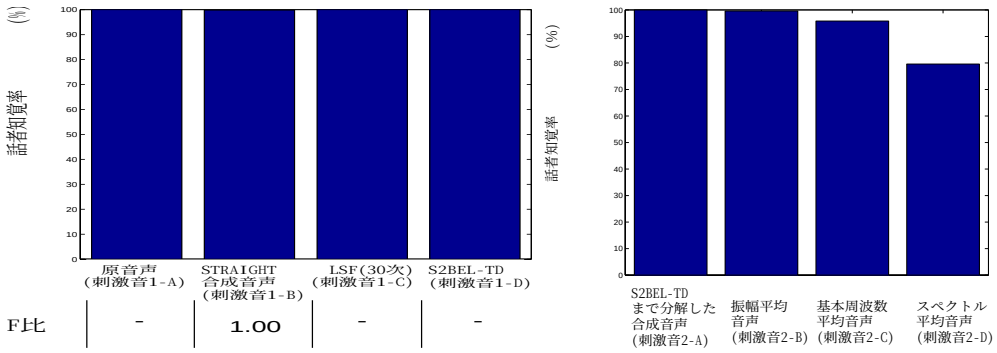

図 3: 話者知覚率 { 実験 1(左)}、{ 実験 2(右)}{F(1,18:0.05)=4.41}

表 2: F 検定

	刺激音 2-B	刺激音 2-C	刺激音 2-D
F 比 (刺激音 2-A)	1.00	29.8	43.4
F 比 (刺激音 2-B)	-	17.8	40.7

{F(1,18:0.05)=4.41,F(1,18:0.01)=8.28}

5.2.1 実験方法

刺激音 刺激音には以下の合成音声を用いた。パラメータの平均は全データの平均をおこなった。

2-A. 刺激音 1-D
2-B. 刺激音 1-D の振幅を話者間で平均したもの。
2-C. 刺激音 1-D の基本周波数を話者間で平均したもの。
2-D. 刺激音 1-D のスペクトルパラメータ (LSF30 次) を話者間で平均したもの。

実験方法
実験 1 と同じ。すべてのセットで、一話者につき 9 回、計 45 回をランダムに呈示した。

5.2.2 結果と考察

実験結果は、図 3 のようになった。平均を行わないパラメータの話者と知覚した割合は、刺激音 2-A で 100%、刺激音 2-B で 99.6%、刺激音 2-C で 95.8%、刺激音 2-D で 79.6%である。

有意水準 5%で F 検定を行った結果 (表 2)、刺激音 2-A と 2-B の間には有意差がなく、刺激音 2-A と 2-C、刺激音 2-A と 2-D には有意差が認められた。

刺激音 2-D の知覚率が一番低いことがわかる。このとき元の話者以外だと知覚した率 20.4%のうち、77.2%が話者 c と回答した。スペクトルの分析 (図 3) の結果でも話者 c が平均から最も近いことから、このような結果が得られたと思われる。

また、有意水準 5%で検定を行った (表 2) 結果、刺激音 2-B と刺激音 2-C、刺激音 2-B と 2-D の話者知覚率の間に有意差が認められた。この結果より、振幅パラメータが話者知覚に影響がないということがわかった。今後の実験では話者間で振幅は平均したものを用いる。

5.3 実験 3 時間変化パターンと話者知覚

時間変化パターンを含めた 3 つの要素 (時間変化パターン、スペクトル、基本周波数) が個人知覚にどのように影響を及ぼすかを調べる。そのために、3 つ要素を話者間で入れ替えを行い、聴取実験を行う。

5.3.1 実験方法

刺激音

刺激音は、時間変化パターン (動き)、スペクトルパラメータ (LSF)、基本周波数 (F0) の 3 つを話者間で入れ替えを行った合成音声を用いた。刺激音は 3 種類に分けることができる。括弧 () 中の数字は実験一セット当りの各刺激音の種類である。

1-A. 3 つのパラメータがすべて同じ話者のもの (5 通り)

1-B. 2 つのパラメータが同一話者のもの (60 通り)

1-C. 3 つのパラメータ全てが異なる話者のもの (60 通り)

実験方法

実験は計 4 回行った。一回につき各話者の文音声の 1 データを用いて行った。1 回目と 4 回目は同じ音声データを用いた。2 回目、3 回目については、各々音声データの種類を 1 回目と違うものを用いた。また、実験結果は、1 回目は分散が大きいとして除き、3 回の実験の平均を示す。

各話者の文音声の 1 データのパラメータ 3 種類を話者間で入れ替えを行い、一セットの合計を 125 回とし刺激音をランダムに被験者に呈示した。また、刺激音呈示レベルは約 75〜80dB(A) の範囲で、両耳にモノラルで呈示した。他、実験 2 と同様に行った。

5.4 実験結果

結果 1：全て同じ話者の場合　刺激音 3-A の知覚率は、100%であった (表 3)。ここで、知覚率は、被験者がその話者であると答えた割合を示してある。

結果 2：2 つが同じパラメータの場合　被験者が 2 つのパラメータの話者であると答えた割合を知覚率として、刺激音 3-B の知覚率を表 3 に示す。

また、表 4 、表 5 には、刺激音 1-B で LSF と動き、F0 と動きが同じ時にその話者だと知覚した回答の内訳を示してある。2 つの話者のとき (表中左の話者 a〜e)、どの話者 (表中上の話者 a〜e) に知覚したかを示してある。other は、音声を構成する話者以外と答えた時の割合である。

結果 3：すべて違うパラメータの場合　表 3 に結果を示す。知覚率は、そのパラメータの話者であると知覚した割合を示したものである。また、表 6 には、刺激音 3-C を回答した内訳を示す。LSF の話者のとき (表中左の話者 a〜e)、どの話者 (表中上の話者 a〜e) に知覚したかを示してある。other は、音声を構成する話者以外と答えた時の割合である。

表 3: 話者知覚率 (結果 1,2,3)

		知覚率 (%)
結果 1(刺激音 1-A)	すべて同じ	100
結果 2(刺激音 1-B)	LSF と F0	98.8
	LSF と動き	83.8
	F0 と動き	20.3
結果 3(刺激音 1-C)	LSF	71.1
	F0	11.1
	動き	7.0

表 4: 回答の内訳 (動きと LSF が同一話者)

	話者					
話者	a	b	c	d	e	other
a	95.0*	0	0	1.7	1.7	1.7
b	0	78.3*	5.8	6.7	0.8	8.3
c	0	3.3	84.2*	1.7	0	10.8
d	0	5.0	5.8	67.5*	0.8	20.8
e	0	0	1.7	1.7	94.2*	2.5

5.5 考察

前節の分析の結果と実験の結果であきらかになったことを示す。すべての話者についていえることは、次のようになった。

- 時間変化を考慮した合成音声に個人性が存在すること (実験 1)
- 時間変化以外の 3 つの要素 (スペクトル、基本周波数、振幅) のうち、個人性の関与しないものは振幅であること (実験 2)
- LSF が最も個人性に関与する (実験 2、実験 3 の結果 3)
- 時間変化は話者知覚にあまり影響を与えない (実験 3 の結果 3)

知覚に関して 5 話者は、大まかにわけて 2 パターンにわかれる (実験 3 の結果 2、3 とスペクトルの分析より)。

表 5: 回答の内訳 (F0 と動きが同一話者)

	話者					
話者	a	b	c	d	e	other
a	0.8*	12.5	25.0	5.8	24.2	31.7
b	23.3	29.2*	14.2	7.5	24.2	1.7
c	25.0	12.5	25.0*	9.2	22.5	5.8
d	23.3	3.3	8.3	40.0*	22.5	2.5
e	23.3	19.2	22.5	20.8	6.7*	7.5

表 6: 回答の内訳 (全てが違う話者の場合)

	話者					
話者	a	b	c	d	e	other
a	95.6*	0	0	2.8	1.1	0
b	0	62.5*	21.1	13.9	2.5	7.5
c	6.1	6.9	76.9*	9.7	0	5.0
d	0	28.6	25.6	44.7*	1.1	19.7
e	0	1.1	2.8	0	95.6*	1.7

- LSF(LSF と時間) のみで知覚できる話者
- F0 と時間変化の影響を受けやすい話者

6 おわりに

本研究では、STRAIGHT と S^2BEL-TD を用い文音声からスペクトル、基本周波数、動きを抽出し、各パラメータに現れる話者間の物理的距離から付置図を作成し、聴取実験を通じて個人性知覚に及ぼす影響について検討をおこなった。

その結果、スペクトルの静的成分が個人性を最も多く含むことを確認した。また、同一話者のスペクトルと基本周波数の静的成分を使用すればその話者に知覚することを確認した。さらに、ある話者と他の話者とのスペクトル距離が大きい場合、スペクトルだけでその話者と知覚されることが明らかとなった。この音響特徴の差が個人性の知覚に反映されるという結果は、ABX 法で行った報告 [5] と一致する。

参考文献

[1] T.Ienaga and M.Akagi "Speaker individuality in fundamental frequency contours and its control" J. Acoust. Soc. Jpn. (E)、**18**, 2, 1997

[2] 北村達也、赤木正人、北沢茂良"スペクトル遷移パターンが個人性知覚に与える影響について" 聴覚研究会, H98-97, 1998

[3] 河原英紀 "音声分析・変換・合成方法 STRAIGHT-TEMPO における相補的な時間窓の利用について" 聴覚研究会, H97-47, 1997

[4] A. C. R. Nandasena and M. Akagi "Spectral stability based event localizing temporal decomposition" Proc. ICASSP98, II, 957-960

[5] 橋本誠、北川敏、樋口宣男 "音声の個人性知覚に影響を及ぼす音響的特徴の定量的分析" 音響学会誌 54 巻 3 号, 1998