声の類似性から個人性知覚を探る*

○北村達也 (甲南大), 出水田剛志 (北陸先端大), 橘亮輔 (甲南大)

1　はじめに

音声の個人性を知覚する能力は，ヒトの音声コミュニケーションの基盤である．音声の個人性は，話者が誰かを知るためのみでなく，話者による音響的差異に適応し (もしくは音響的差異を正規化し) 言語情報を得るためにも用いられる．したがって，個人性知覚のメカニズムの解明は音声科学の基本的課題の 1 つといえる．

個人性知覚研究における第 1 の問いは，個人性に対応する音声中の特徴は何かである．それがどのような方法で抽出され，どのような形で脳にコードされているかも明らかにする必要がある．第 2 の問いは，その特徴を用いて未知話者の個人性を学習し，既知話者の音声を照合するアルゴリズムはどのようなものかである．ただし，個人性が現れやすい特徴は話者間で異なり，また個人性として用いる特徴が個人間で異なることを考慮する必要がある．そして第 3 の問いは，上記のプロセスの神経メカニズムはいかなるものかである．個人性知覚メカニズムの解明には少なくともこの 3 つの問い答える必要があると考えている．言うまでもなく，これらの知見は工学的応用が可能である．

個人性に対応する特徴に関してはこれまでに多数の研究が行われ，様々な物理量の話者識別への寄与が指摘されてきた．しかし，個人性を表すために必要十分な物理量のリストは示されていない．個々の話者の音声においては様々な物理量が異なっているため，それら全てに関して個人性知覚に与える影響を調べ上げることは容易ではない．

著者らは，個人間の差異ではなく類似性に着目することによってこの問題の解決を試みている [1]．個人性が類似している 2 人の音声を分析し，その間で値が近い物理量を抽出すれば，個人性知覚に本質的な物理量のリストが特定できると考えられる．逆に，この 2 人の間で値が大きく異なる物理量は，個人性知覚にとって本質的ではないといえる．現段階ではまだリストの特定に至っていないが，その途中経過について紹介する．

2　個人性が類似した話者ペアの選定

2.1　音声データ

話者を混同されることがある成人男性 2 名 (MT，TK) に加え成人男性 3 名 (HF，HH，HT) の文章音声を用いた．防音室にて話者に 5 文を 5 回ずつ読み上げさせ，その音声を DAT (Sony TCD-D10 Pro II) を用い，標本化周波数 48 kHz，量子化 16 bit にて録音した．録音した音声データを PC にて 16 kHz にダウンサンプリングした．

2.2　類似度評定実験

1 文 (「あらゆる現実を全て自分の方へねじ曲げたのだ」) の音声データを対象にして実験協力者 10 名に個人性の類似度を評定させた．実験協力者は実験前に音声データの話者の声を聞いたことがなかった．2 話者の音声をランダムに提示し，「似ている」「やや似ている」「あまり似ていない」「似ていない」の 4 段階で評価させた．

実験結果を表 1 に示す．話者 TK-MT 間に高い類似性，話者 HF-HT 間に類似性があると回答された．

Table 1　Results for the four-level rating of perceptual similarity of pairs with the scale “similar”(A), “rather similar”(B), “not very similar”(C), “dissimilar”(D).

pairs of speakers	similarity			
	A	B	C	D
HF-HH	0	1	5	4
HF-HT	2	6	2	0
HF-MT	0	2	5	3
HF-TK	0	1	2	7
HH-HT	0	5	2	3
HH-MT	0	1	0	9
HH-TK	0	1	4	5
HT-MT	0	1	2	7
HT-TK	0	0	3	7
MT-TK	7	3	0	0

*Exploring mechanism of perception of speaker individualities from voice similarity between speakers by KITAMURA, Tatsuya (Konan Univ.), IZUMIDA, Tsuyoshi (JAIST), TACHIBANA, Ryosuke (Konan Univ.)

3　基本周波数パターンの分析

上記の実験後，実験協力者に類似性の判断材料を質問したところ,「韻律」との回答が多数であった．そこで，上記の実験に用いた文音声を対象にして基本周波数パターンの分析を行った．

3.1　基本周波数パターンの比較

WaveSurfer[2] の Pitch Contour 機能により基本周波数を抽出した．分析手法は ESPS[3]，フレーム長 7.5 ms，フレーム周期 10 ms である．異常値は目視にて修正した．

話者 MT と TK，HF と HT の基本周波数を図 1 に示す．これらの基本周波数パターンはよく似ており，基本周波数パターンが声の類似性に寄与することを示唆している．ただし，基本周波数パターンを個人性知覚の特徴として用いているにしても，何らかの発話内容の変化に対して不変な特徴を抽出しているはずである．

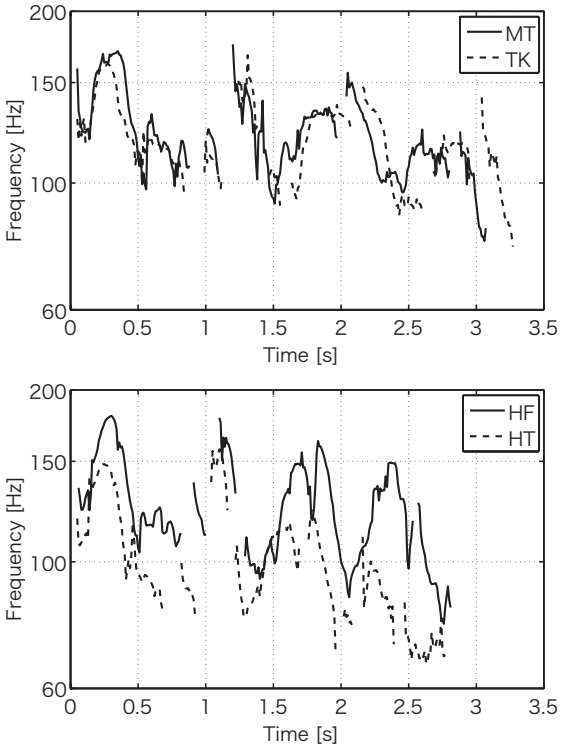

Fig. 1　Pitch frequency contours of speakers MT and TK (upper panel) and HF and HT (lower panel). (The frequency scale is logarithmic.)

3.2　「落ち着き」スケールの比較

山本と松本 [4] は，文音声の個人性知覚に関する検討の中で以下の「落ち着き」スケールという尺度を導入している．

$$\frac{\sigma(\tilde{F}_0)}{\text{mora per second}} \tag{1}$$

ここで，$\sigma(\tilde{F}_0)$ は有声区間の対数基本周波数の標準偏差である．上式は文内の抑揚の程度を話速で割った値になっており，抑揚をつけてゆっくり発話された音声において高い値を示す．

上記の文音声の有声区間における基本周波数の平均値および標準偏差，時間長,「落ち着き」スケールの値を表 2 に示す．類似度評定実験にて類似性が高かった話者 MT と TK，話者 HF と HT のペアで「落ち着き」スケールの値が近い．この尺度は発話内容に依存しないため，基本周波数の変化に関連する個人性を表す特徴として期待できる．

ただし，話者 HF と TK の「落ち着き」スケールの値が近いが，類似度評定実験における類似性は低いので,「落ち着き」スケールだけで類似性が決まるわけではない．

Table 2　The mean ($\bar{F}_0$) and standard deviation ($\sigma(F_0)$) of the pitch frequency of voiced segment in Hz, the length of speech data (Len.) in second, and the calm scale.

	$\bar{F}_0$	$\sigma(F_0)$	Len.	Calm
HF	124.8	24.4	2.868	0.0097
HH	121.0	22.9	3.737	0.0120
HT	100.8	23.3	2.794	0.0108
MT	121.8	19.8	3.176	0.0087
TK	118.4	18.6	3.380	0.0093

4　基本周波数パターンの個人性知覚への寄与

上記の分析結果を受けて，基本周波数パターンが個人性知覚に与える影響を調査した．

4.1　刺激音

話者 HF，HH，TK の 2 つの文 (「あらゆる現実を全て自分の方へねじ曲げたのだ」,「一週間ばかりニューヨークを取材した」) の音声データから以下の 3 種類の刺激音を作成した．これらの 3 名は類似度評定実験において類似度が低いと判定されていた．刺激音の作成には STRAIGHT 分析合成系 [5] を用いた．

刺激音 1 STRAIGHT 分析合成音声

刺激音 2 原話者の音声において音素継続時間長を目標話者のもので置換した合成音声

刺激音 3 刺激音 2 の処理に加え，基本周波数パターンも目標話者のもので置換した合成音声

音素継続時間長の置換は DP マッチングに基づいて行った．刺激音の最大振幅値は正規化した．

4.2 実験協力者

正常な聴力を有する成人 20 名 (男性 10 名，女性 10 名) が実験に参加した．これらの実験協力者は，実験前に刺激音の話者の音声を聞いたことがなかった．

4.3 実験方法

XAB 法により実験を行った．実験協力者に 3 つの刺激音 X，A，B を 0.5 s 間隔で提示し，X の話者が A と B のどちらの話者に似ているかを強制判断させた．順序効果を打ち消すため，XBA の順でも提示した．文献 [1] の実験とは異なり，X の文と A，B の文は別のものを用いた．文に依存しない個人性について調査するためである．

3 つの刺激音の組み合わせには，X，A，B がいずれも刺激音 1 の場合と X が刺激音 2 または 3 で A と B がともに刺激音 1 の場合がある．前者の場合，X は A の話者が発話した音声であり，後者の場合，X は A と B の話者の音声から合成した刺激音である．

実験協力者は刺激音をヘッドホン (Sennheiser HDA200) で両耳受聴し，A または B を回答した．刺激音の聴き直しは許さなかった (詳しくは文献 [1] を参照のこと)．

4.4 実験結果

刺激音 X の話者が B の話者に似ていると回答された割合を図 2 に示す．実験結果は実験協力者間で平均した．実験結果を分散分析で検定したところ ($F(1,238;0.05) = 3.881$)，刺激音 1 と 2 の間 ($F(1,238) = 9.293$)，刺激音 2 と 3 の間 ($F(1,238) = 8.653$) に有意差があった．これらはそれぞれ音素継続時間長と基本周波数が個人性知覚に寄与することを示している．音素継続時間長は話速と関連するため，この結果は上述の「落ち着き」スケールと個人性の類似性判定の対応がよいことと関係している可能性がある．また，平均基本周波数や基本周波数パターンが個人性知覚に寄与するとした先行研究 [6, 7, 8, 9] とも矛盾しない．しかし，刺激音 2 と 3 の値は 20%以下であることからこれらの特徴量の寄与は小さく，主な個人性知覚要因はスペクトルであるといえる．

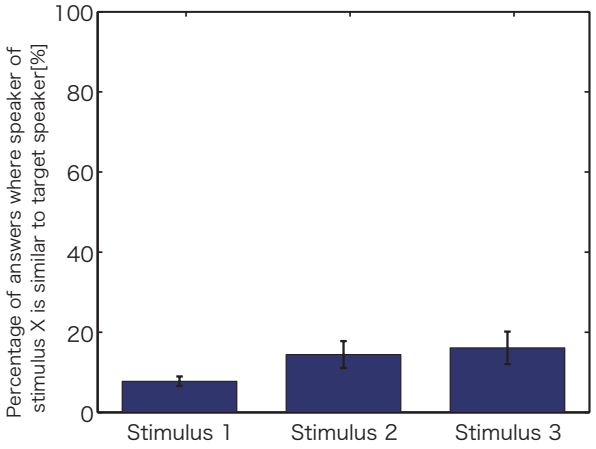

Fig. 2 Percentages of answers where the speaker of stimulus X is similar to the target speaker.

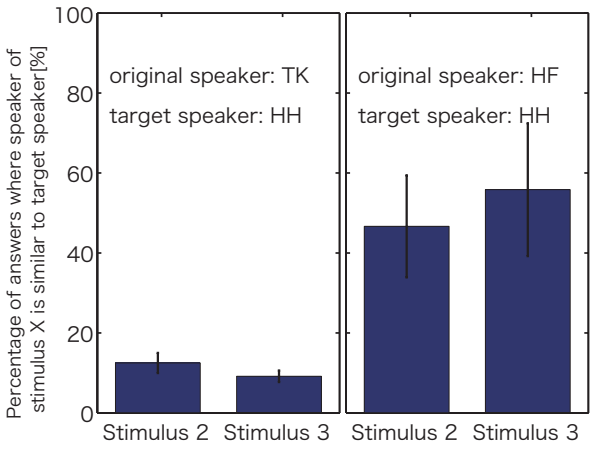

Fig. 3 Percentages of answers when the target speaker is speaker HH. The original speaker is speaker TK (left) or HF (right).

図 2 の結果のうち，話者 HH を目標話者として合成した刺激音に関する結果を図 3 に示す．原話者を話者 TK とした場合と話者 HF とした場合の間に有意差が見られた ($F(1,38;0.05) = 4.098$, 刺激音 2: $F(1,38) = 19.477$, 刺激音 3: $F(1,38) = 7.961$)．この結果から，音素継続時間長と基本周波数の個人性知覚への寄与は話者の組み合わせによって異なることがわかる．

話者 HH を原話者，話者 HF を目標話者とした場合 (図 3(右))，目標話者として回答された割

合は 50%前後，すなわちほぼチャンスレベルであった．よって，これらの話者はスペクトルが類似していることが示唆される．一方で，この 2 話者の間の類似性は低く (表 1)，原音声を用いた実験 (X を刺激音 1 とした実験) では話者 HH と HF が正しく識別されている．従って，2 話者間でスペクトルが類似しており話者識別の手がかりにならない場合には，音素継続時間長や基本周波数を手がかりとして話者が識別されると考えられる．

5 考察

上記の結果に基づき，知覚平面上での話者 HF, HH, TK の関係を推測したものを図 4 に示す．この平面はスペクトルと韻律 (音素継続時間長および基本周波数) に関する知覚上の距離を軸としている．

話者 HF-HH 間ではスペクトルに関する知覚上の距離が小さいため，韻律が話者識別の手がかりとして用いられる．そのため，韻律を置換した刺激音 3 では話者を識別できなかったと考えられる (図 3(右))．今後これらの話者のスペクトルの類似性を分析することにより，スペクトル中の何が個人性知覚要因かを解明できるはずである[1]．

一方，話者 HH-TK 間ではスペクトルに関する知覚上の距離が大きかったため，刺激音 3 でもスペクトルの情報を用いて話者を識別できたと考えられる (図 3(左))．

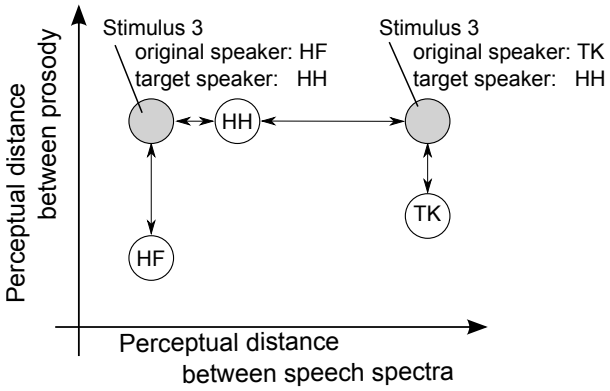

Fig. 4 Relationship among speakers HF, HH, and TK in perceptual space spanned by axes of perceptual distances between the speech spectra and prosody.

[1]MFCC による比較を試みたが，これらの話者の類似性を示すデータは得られていない

6 おわりに

本稿では，始めに個人性知覚に関する研究が解くべき 3 つの問いを挙げ，次いで個人性の類似性に着目した研究を紹介した．先行研究 [6, 7] と同様，スペクトルが個人性知覚に最も寄与することが示されたが，話者の組み合わせによっては韻律が主な知覚要因になることも明らかになった．

心理学的な話者間距離から個人性知覚要因を探る研究は過去にも行われており [10, 11, 12]，今後これらの研究結果と比較していく必要がある．

ただし，我々の研究も含め，それぞれ研究の結果は実験に用いた話者セットに少なからず依存する．今後この分野の研究を進めていくためには，各研究者が共通の音声データベースの上で研究し，結果を比較検証できるようにしていくことが不可欠である．

謝辞 本研究の一部は平成 23 年度科学研究費補助金 (21300071, 21330170) の支援を受けた．聴取実験の一部は甲南大学知能情報学部の原田江里子さんの協力を得た．STRAIGHT を利用させていただいている和歌山大学システム工学部の河原英紀先生に感謝します．

参考文献

[1] Izumida, Kitamura, Acoust. Sci. Tech. (printing).
[2] WaveSurfer, `http://www.speech.kth.se/wavesurfer/`
[3] Talkin, in “Speech Coding and Synthesis,” 495–518, Elsevier, 1995.
[4] 山下, 松本, 音響誌, 62(12), 856–864, 2006.
[5] Kawahara *et al.*, Speech Commun., 27, 187–207, 1999.
[6] 伊藤, 斉藤, 信学論, J65-A(1), 101–108, 1982.
[7] 橋本, 北川, 樋口, 音響誌, 54(3), 169–178, 1998.
[8] Akagi, Ienaga, JASJ(E), 18(2), 73–80, 1997.
[9] 大野, 赤木, 信学技報, SP97-128, 1998.
[10] Matsumoto *et al.*, IEEE Trans. Audio Electroacoust., AV-21, 428–436, 1973.
[11] Murry, Singh, JASA, 68(5), 1294–1300, 1980.
[12] 津崎ら, 信学技報, SP2010-116, 2011.