人まね音声合成のための個人性情報の評価*

比屋根廣紀　高良富夫（琉球大工）

1　はじめに

個人性のはっきりした音声の合成は，なつかしい故人の音声の再現等、有用な応用がある。

音声に含まれる個人性は，多くはスペクトルが担い，基本周波数（F0）[1]等，韻律的特徴パラメータにも含まれている。特にプロの人まねでは，韻律的特徴を強調していると思われる。

しかし，これらのパラメータのうちどれが個人性に関して重要な要因であるかは明らかでない。

そこで本研究では，文の合成音声を用いて，スペクトル, F0, タイムアラインメント, パワーのうちどれが，個人性に関して，どの程度重要であるかを検討する。また個人性の重要度に関して等価となるパラメータのそれぞれの値を決定する。

2　個人性要因の検討

本研究では、文音声に含まれる個人性の要因を検討する。例えば、単語音声のような短い音声にも個人性は含まれているが、それよりも長い文音声の方に個人性が多く含まれていると考えられる。そこで文音声における個人性要因の検討を行うことにする。

2.1 音の三要素とタイムアラインメント

個人性としては，基本周波数パターン・文音声における音の強弱・スペクトルに加え，時間上の音素の伸縮などに対応するタイムアラインメントを考える。そして，この４要素(スペクトル・基本周波数・パワー・タイムアラインメント)それぞれについて，特徴距離を決定する。この特徴距離に関しては 2.3 以降で述べる。

2.2 時間軸の不整合

本研究では，２話者間の文音声における個人性要因を検討する。そして，２話者の特徴パラメータを交換した音声を合成し，その合成音声の個人性の入れ換わりにより個人性要因の重要度を検討する。

特徴パラメータを交換する時は，時間軸の不整合の問題が生じる。例えば、２話者が同じ文を話したとしても，これらの音声は時間軸上で発声タイミングのずれがある。時間軸上の整合を行うため DP マッチング法を用いて，２つの音声の時間軸上の対応表を作成し，合成を行う。DP マッチングの局所的距離としてはスペクトル距離を用いる。

2.3 特徴距離

個人性の重要度に関して等価なパラメータ値を決定するため，前に述べた４要素(基本周波数・パワー・スペクトル・タイムアライメント)から特徴距離を決定する。

2.3.1 スペクトル

スペクトルにおける特徴距離としては，DP マッチングで求めた時間正規化距離である，平均スペクトル距離を用いる。

2.3.2 基本周波数

基本周波数における特徴距離としては，平均基本周波数距離と平均基本周波数概形距離の２つを用いる。

平均基本周波数距離は，２話者の基本周波数の平均値の差である。

平均基本周波数概形距離は，各話者の基本周波数概形からそれぞれの平均基本周波数を引きさることにより正規化し，２つの概形の差の平均値として求める。

---

*"Evaluation of personality parameter of speech for Speech Mimicking System", by Tomio TAKARA and Hiroki HIYANE (Department of Information Engineering, University of the Ryukyus).

2.3.3 パワー

パワーの特徴距離に関しては，2.3.2 と同様に，平均パワー距離と平均パワー概形距離を求める。

平均パワー距離は，２話者の音声パワーの平均値の差である。

平均パワー概形距離は，２話者の音声パワーの概形間の差の平均値である。

2.3.4 タイムアラインメント

タイムアラインメントの特徴距離としては、総フレーム差・最大フレーム伸縮差・有声部最大フレーム伸縮差の３つを求める。なお、ここで１フレームは 10[ms]である。

総フレーム差は、２話者の音声の総フレーム数の差である。

最大フレーム伸縮差は、DP マッチングにおけるフレームの伸びの最大値と縮みの最大値の差である。

有声部最大フレーム伸縮差は、最大フレーム伸縮差を有声部だけで求めた値である。

3　聴取実験[2]

個人性の検討のため，前の節で示した４つのパラメータを２人の話者の音声間で交換し，全ての組み合わせである 16 種の合成音声を作成する。そして，この合成音声を利用し、後に説明する実験 1～実験 9 の聴取実験を行った。

3.1 刺激音

パラメータの交換をした 16 種の合成音声は、量子化精度 16 bit，標本化周波数 10 kHz である。実験１～実験６では，誰もが電話で話すような文を使用し，３人の男性が発声したものを録音した。文は次の２つである。

- もしもし，今日どこに行く？
- もしもし，今日はありがとう。

３人の話者の中から２人を選び，パラメータの交換を行い，それを２つの文で行ったので計 96 $(16\times{}_3C_2\times2)$の合成音を作成した。

次に実験 7～実験 9 では，参考文献[3]で使われていた文を利用して，男性３人が発声したものを録音した。文を以下に示す。

- “それをかばおうとして，右膝とふくらはぎをやられた”

これも同様に，３人の話者の中から２人を選びパラメータの交換を行い，計 48 $(16\times{}_3C_2)$の合成音を作成した。

3.2 実験手順

実験 1～実験 6 ではクローズドテストを行った。

聴取実験は，十分音声が聞き取れる環境でヘッドホンを用いて実施した。実験 1～実験 6 のそれぞれの実験では，2 人の話者間で 4 つのパラメータを交換した 16 の合成音と原音声，計 18 の音声を聴取させた。聴取実験を行う前に，被験者は訓練用音声で聴取訓練を行った。訓練用音声として，聴取実験で利用する 2 人(男性 A，男性 B)の原音声を用いた。訓練は，被験者が個人の声を識別することが可能であると認められた時点で、十分な時間をとって終了する。訓練後，合計 18 個の音声を被験者に３秒おきにランダムに聴取させ，男性 A と男性 B のどちらが話したものに似ているかを強制判断させた。刺激音は１度だけ再生される。被験者は，５セットの実験を行った。

実験 7～実験 9 では，実験 1～実験 6 と同様の手順だが，オープンテストを実施した。すなわち訓練用音声としては，聴取実験で利用するもの以外の文音声(原音声)をそれぞれ 2 つずつ用いた。被験者は，３セットの実験を行った。

3.3 実験結果

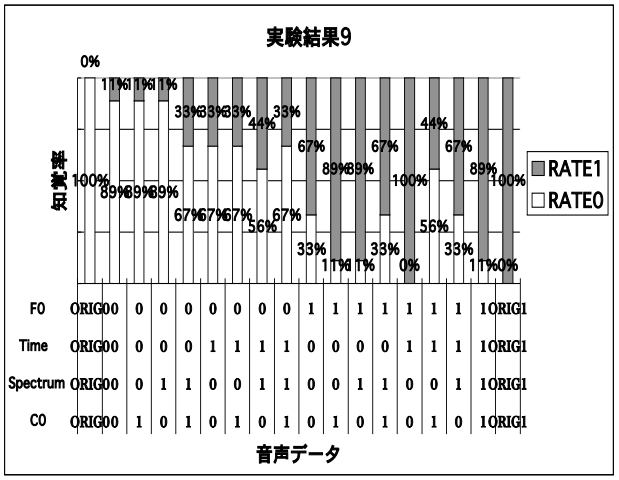

図.1 実験９の実験結果

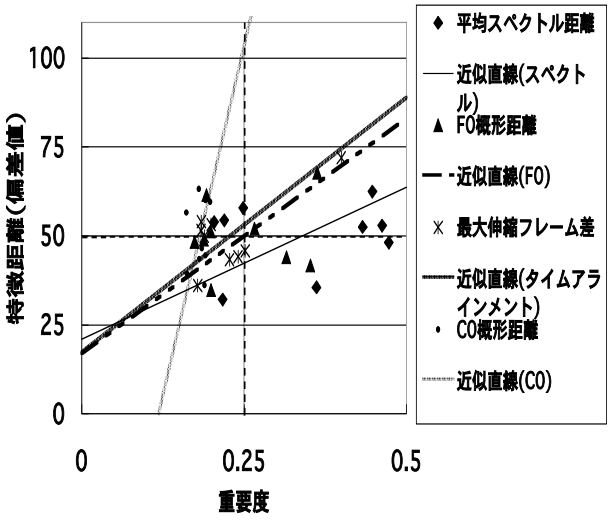

図.2 各パラメータの重要度と偏差値

図 1 に、例として聴取実験９の結果を示す。図１の横軸は音声データ，縦軸は知覚率を示している。図 1 では,その実験で重要度が高かった順にパラメータを並べてある。

図 2 に, 結果を散布図にして,それぞれパラメータに回帰直線を引いたものを示す。横軸は,重要度を表している。縦軸は,各パラメータの実際に使用したデータの中での偏差値を示している。

3.4 重要度

重要度とは, 被験者の選択率から求めた４つのパラメータ(F0, タイムアライメント, スペクトル, パワー)の重要さを表す量である。重要度は, この 4 つパラメータの選択割合の平均値である。また, ４パラメータの重要度の和が１となるように正規化をしている。

重要度の計算例を以下に示す。

表 1 実験結果例

Spectrum	F0	Time	C0	Speech(0)	Speech(1)
0	0	1	1	6	4

例えば, 表１に，ある合成音声に対する聴取実験の結果の例を示した。話者 A が０, 話者 B が１を表すものとする。この例は, (Spectrum,F0,Time,C0)=(0,0,1,1)の合成音について，話者 A と答えた人数が 6 人, 話者 B と答えた人数が 4 人であることを表している。この合成音における話者 A の要素が Spectrum と F0 であり, かつ話者 A と答えた人数が 6 であるので, この全体の 60%を Spectrum と F0 に分割した値がそれぞれの重要度となる。つまり, Spectrum 重要度と F0 重要度は共に 0.30 となる。同様に, Time と C0 の重要度を求める。Time と C0 は話者 B の要素で合成を行い、かつ全体の 40%が話者 B だと知覚しているので、それを２分割した 0.20 が Time と C0 それぞれの重要度となる。

次に，１つの実験全体の重要度の計算例について述べる。合成音声は，２話者間において全部で 16 種類作ることができる。重要度に関しては，(0,0,0,0) と (1,1,1,1)の 2 種類を除いた合計 14 種類の合成音について考える。この 14 種類(0,0,0,1)から(1,1,1,0)までの合成音声の各重要度の平均値を求める。その求めた平均値をさらに４つの重要度の和が１.00 となるように正規化を行った値がその実験における重要度となる。

3.5 検討

図１は, 3.3 でも述べたように，重要度の高い順に並べてある。つまり F0・Time・Spectrum・C0 の順番に重要であるという結果が得られた。横軸の F0 をみると, F0 が 0 から 1 に変わる時に, 知覚率も 0 か

ら 1 に変化しているのが分かる。また，個人性の多くを担っている Spectrum よりも F0 や Time が重要であることもあるということが，この実験結果から得られた。

図 2 で，特徴距離の偏差値 50 の線と重要度 0.25 の線に着目すると，重要度の順位と各特徴距離の等価値を求めることができる。例えば偏差値 50 の線では，特徴距離が平均的である場合、どの要素が一番重要であるかが分かる。図 2 から，スペクトル，F0，タイムアラインメント，C0 の順で重要であることが分かる。

表 2 は，図 2 で求めた近似直線から求めた重要度をまとめたものである。これは，各パラメータの値が平均的である場合の重要度である。表 2 を見て分かるように，スペクトル距離，F0 概形距離，最大フレーム伸縮差，C0 概形距離の順に重要なパラメータであることがわかる。表 3 は，そのときの値(平均値)である。

表 2 偏差値 50 における各パラメータの重要度

スペクトル距離	**0.34**
F0 概形距離	**0.25**
最大フレーム伸縮差	**0.23**
C0 概形距離	**0.18**

表 3 各パラメータの値

スペクトル距離	**7.54 [dB]**
F0 概形距離	**4.87 [Hz]**
最大フレーム伸縮差	**18.11 [10ms]**
C0 概形距離	**0.7 [dB]**

表 4 重要度 0.25 における各パラメータ

スペクトル距離(偏差値)	**42.3**
スペクトル距離	**6.71 [dB]**
F0 概形距離(偏差値)	**50.0**
F0 概形距離	**4.88 [Hz]**
最大フレーム伸縮差(偏差値)	**53.1**
最大フレーム伸縮差	**22.0 [10ms]**
C0 概形距離(偏差値)	**103.6**
C0 概形距離	**1.46 [dB]**

4 つの重要度の合計は 1 になるようにしていることは，前に述べた。重要度が均等になるとき，すなわち重要度が 0.25 となるときの各パラメータの等価値を今回の実験から表 4 のように求めることができた。また，表 4 の等価値において C0 概形距離の偏差値がかなり大きな値になっていることが確認できる。このように，C0 で個人性の特徴を強調するのは難しいといえる。

4 むすび

音声に含まれる個人性を担う音声パラメータについて，その聴感上の重要度を合成音声の聴取実験により調べた。その結果，スペクトル，F0，タイムアラインメント，パワーの順に重要であり，その重要度は，それぞれ 0.34, 0.25, 0.23, 0.18 であることが分かった。また重要度が同じになる各パラメータの差は 6.71[dB], 4.88[Hz], 22.0[フレーム], 1.46 [dB]であった。また C0 に関しては，今回求められたた等価値の偏差値が 103.6 と高いことから，C0 の強調による個人性の入れ換えは難しいことが分かった。

参考文献

[1] Masato Akagi, Taro Ienaga, “Speaker individuality in fundamental frequency contours and its control”, J. Acoust. Soc. Jpn (E) 18, 2 pp.73-80, 1997.

[2] 高良富夫，”琉球方言の声門破裂音の音韻性”，日本音響学会誌，51 巻 8 号，pp. 599-605，1995.

[3] Masanobu ABE, “Speech morphing by gradually changing spectrum parameter and fundamental frequency”, IEICE technical report.Speech, Vol.96 No160, pp.25-32, 1996.