音声言語情報処理 35－2
（2001．2．2）

声道模擬型規則合成における音源の改善について

神谷賢，雨宮沙織，有泉均

mkamiya@pop16.odn.ne.jp，saori@alps1.esi.yamanashi.ac.jp，
ha@esi.yamanashi.ac.jp

山梨大学工学部コンピュータメディア工学科

〒400-8511　山梨県甲府市武田 4-3-11

あらまし

筆者らは声道模擬型の音声合成器の研究をしている．人間の放射音声/a/から求めた逆フィルタ波形を1波形に切り出さないで連続したまま音声合成器へ入力した．すると，その人の個人性を保存した肉声レベルの/aiueo/を合成できた．このような発見は過去にもなされている [1,2,3,4,5] が，筆者らの合成方法はターミナルアナログ型であるため，人間の発声機構の理解に役立つ上にさまざまな用途に応用することもできる．今後は逆フィルタ波形に載っている肉声感と個人性の情報を解明し，声道模擬型規則音声合成器や話者認識システムに応用することを目標としている

About the improvement of the sound source in the vocal tract simulation type speech synthesis system by rule.

Kamiya Masaru and Amemiya Saori and Ariizumi Hitoshi

E-mail : mkamiya@pop16.odn.ne.jp，saori@alps1.esi.yamanashi.ac.jp，
ha@esi.yamanashi.ac.jp

Yamanashi-University Computer media engineering department

4-3-11, takeda, kofu-shi, yamanashi-ken 400-8511, Japan

Abstract

Authors are researching the voice synthesis system of a vocal tract control type by rule. The sound source was requested from the radiation voice by the reverse-filter method to request the sound source of human for the improvement of tone quality this time. When it was input to the voice synthesis machine while having continued directly, it was found to be able to obtain a synthetic voice at the voice level by which the person's individual was preserved. It is possible to apply this report to useful for the understanding of a vocal mechanism of actual man because it is a terminal analog type, and various usages.

1. はじめに

当研究室の音声合成方式は，後述するターミナルアナログ合成方式である．これまで，このタイプの方式による音源には三角波のような，人工の数式モデルを用いてきた．しかし，これらの音源の音質は，男声，女声，子供の声を判別できる程度であり，個人の音声を再現できるまでには至っていない．

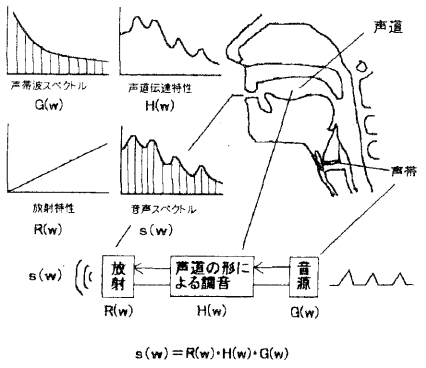

図１：音声の生成

そこで，今回は，どの母音から得られた逆フィルタ波形がいいのか，その中に肉声感や個人性の情報がどこに含まれているのか，それをモデル化できないか，について検討した．

2. 実験準備

2.1. 簡易版音声合成システム

当研究室の音声合成は図１のように人間の発声機構を反映した音源フィルタ理論に基づいている．回路網理論の見地から声道伝達関数をより厳密に近似するターミナルアナログ型を採用している．そのモデルは式１のように極とゼロが無限につながったものである．

当研究室の音声合成器は，図２のように，音韻情報を入力とすると，それから声道を模擬してフォルマント周波数を連続的に計算する．それを波形合成部において，ピッチ情報を音源波形に付加して共振フィルタへ入力して目的の音声を得る．

ディジタルフィルタの構成は図３のようになっており，その中の極回路は図４に示した通りである．ディジタルフィルタは極回路Ｐを直列に接続してあり，それぞれの極回路Ｐに第一第二とフォルマント周波数や帯域幅を設定する．これに音源波形を左から入力しその出力を次の右の段の極回路Ｐの入力とする．

今回は声道を制御せず，与えられたフォルマント周波数を用いる簡易版の音声合成システムを構成した．

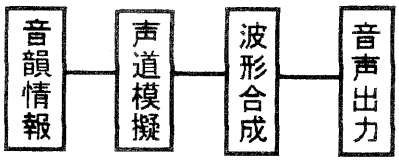

図２音声合成システム

$$H(S)=\frac{\prod_{z=1}^{\infty}(S-S_z)(S-S_z^*)}{\prod_{p=1}^{\infty}(S-S_p)(S-S_p^*)}$$

S_p：**極**　，　S_z：**零点**，*：**共役**

式１：声道伝達関数のモデル

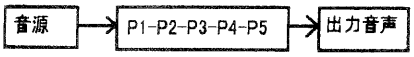

図３：極零回路の直列接続

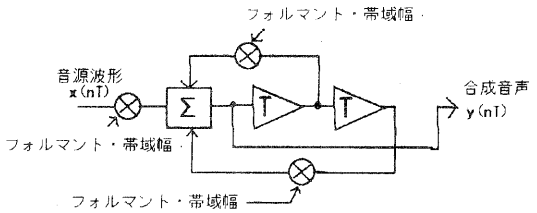

図４　極回路Ｐの構成

$$y_a(nT)=\frac{x_a(nT)-2r_i\cos b_iTx_a(nT-T)+r_i^2x_a(nT-2T)}{1-2r_i\cos b_iT+r_i^2}$$

式２　極回路の逆回路の数式

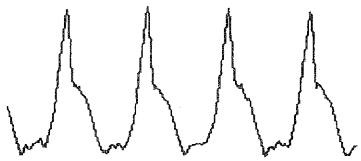

図５：逆フィルタ波形

2.2. データ収集

音声波形は普通に/a//i//u//e//o/を発声

したものをマイクロフォン AIWA DM-H300 などを用いて録音した. AD 変換は Sound Blaster を用いた. 16 ビット, 11025Hz でサンプリングし, これを 12 次の LPC 分析をしてフォルマントと帯域幅を求めた. これを大学生女子5名から5母音を集めた.

2.3. 逆フィルタ波形の求め方

ここでは, 先に述べたように, 共振フィルタの逆特性から求める方法を説明する.

共振周波数のディジタルフィルタは極回路であり, その逆回路は式2で表される.

この数式は図 4 の極回路 P の逆回路を数式で表現したものである. ここで, x(nT)は音声波形, y(nT)は音源波形, b はフォルマント周波数, r はフォルマント帯域幅の情報である. 第5フォルマント(P5)の逆回路に音声波形を入力しこれを順に第1フォルマント(P1)まで計算する. P1 の出力が音源波形になる.

これを音声合成器の逆フィルタに値を設定し, 放射音声の逆フィルタ波形を求めた.

求まった逆フィルタ波形を図 5 に示す. この5つの母音から得られた逆フィルタ波形を次の章で述べる評価実験に用いた.

2.4. 評価方法

評価実験はいずれも以下の方法で行われた.

各実験の条件で合成音声を作成した. 合成音声はいずれも/aiueo/と発声している.

評価方法は, ABX 法[1]を参考にして実験を行った.

A に基準音声としてその人の自然音声を聞かせ, B に呈示音声として合成音声を聞かせる. B の合成音声はそれぞれ, 5 母音から逆フィルタを用いて得られた音源より合成音声を5つ作成した.

音声はカセットテープに録音し, 音質, 個人性, 音韻性の類似度および総合評価を5段階で評価してもらった. A と B の音声の聞かせ方は ABAB の順で聞かせた.

判定は5段階で個人性, 音韻性, 音質, 総合評価についてAからEまでで採点してもらった.

A. 似ている(5点)
B. 結構似ている(4点)
C. どちらともいえない(3点)
D. あまり似ていない(2点)
E. 似ていない(1点)

である.

3. 実験1(逆フィルタ波形の性能評価)

5母音から求まった逆フィルタ波形のどれで音声を合成するのがもっとも良いのかを客観的に実験で求めた.

3.1. 方法

5人の女性から5母音を収集した. その5母音の逆フィルタ波形を用いて音声を合成し, 先に述べた ABX 法で個人性・音韻性, 音質総合評価について聞き取り実験をした.

3.2. 結果

最も成績の良かった話者 risa についての総合判定結果を図 6 に示す. すべての判定項目における結果の平均点である. グラフより/a/や/e/の逆フィルタ波形が良く聞こえたことが分かった. これはすべての話者に共通していた結果である.

3.3. 考察

/i/や/u/では調音の段階で音源の情報が消えてしまったところを無理やり復元しようとするのでノイズが発生したり, 音源波形のスペクトルに山が残っていたり, 完全に適度な傾斜で単調減少する音源波形が得られなか

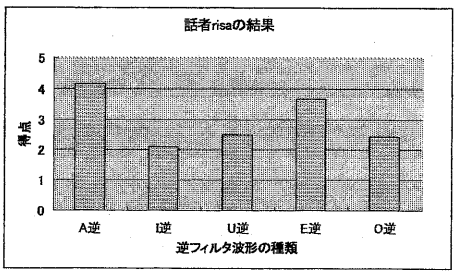

図6：実験1の結果

ったときに，音質の劣化が起こっている．

4. 実験２（肉声感の所在について）

ここでは，実験１で最も良い結果が得られた/a/の逆フィルタ波形を使って，肉声音質の音源のどこに肉声感があるのかを調べた．

4.1. 方法

肉声感はピッチ，振幅，波形のゆらぎにあると仮定し，以下の音声資料を作成した．

図７に示すように，/a/の逆フィルタ波形①を１ピッチ切り出して②連結し③，自然音声の基本周波数パターンを TD-PSOLA 法によって与える④．次に，音源波形同士を１ピッチずらして引くことで得られるノイズ④をその波形に重畳した⑤．これも話者 risa とそれぞれ個人の話者の音源を用いて音声資料を作成した．基本周波数およびフォルマント，音節時間長はそれぞれの話者のものを用いた．これを５人分の話者について合成音声を作成して，聞き取り実験を行った．

4.2. 結果

結果を表 1,2 に示す．表１は最も評価の高かった話者 risa の音源波形を用いたものである．表２はそれぞれの話者の音源波形を用いた場合の聞き取り実験結果である．

これより，どちらの音源を用いても個人性の認識結果はほとんど同じであった．また，５点満点のうち，４点である，「結構似ている」以上の成績が得られた．

4.3. 考察

本実験より，肉声感の原因は基本周波数パターンおよび，それに含まれる変動成分と，波形に重畳しているノイズが有力であることが分かる．また，こちらも音源が他人のものでも良好に個人性を判別することができたため，「地声」の場合，個人性の影響は音源波形のスペクトル包絡にはあまり載っていないと考えられる．

5. 実験 3(個人性の所在について)

逆フィルタ波形のどこに個人性は含まれているのかを以下の実験により調べた．

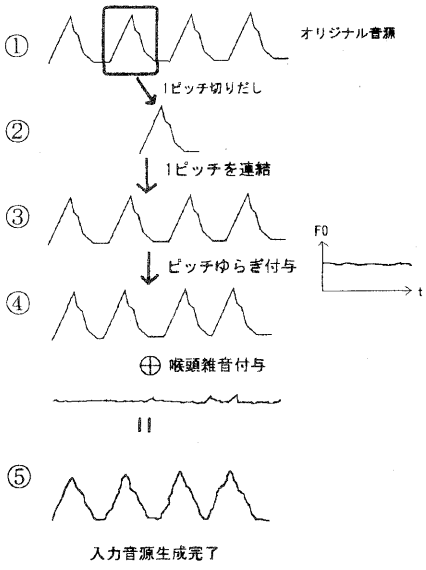

図７：１ピッチからの音源生成

表１実験２の結果 1(話者 risa の音源使用)

話者	個人性	音韻性	音質	総合
Li	4.3	4.0	3.9	4.1
saori	3.4	3.4	3.4	3.4
risa	4.0	4.0	3.7	3.9
mori	4.7	4.5	4.5	4.7
wada	3.1	3.3	2.4	2.9

表 2:実験２の結果２(個人の音源波形使用)

話者	個人性	音韻性	音質	総合
Li	4.9	4.4	3.8	4.2
saori	3.2	3.9	3.4	3.4
risa	4.0	3.7	3.2	3.5
mori	4.4	4.3	4.1	4.3
wada	3.0	3.5	2.5	3.1

5.1. 方法

ここでは，個人性の情報は図８に示すように，音源の調和成分の組み合わせ比率全体に，個人性の情報が載っていると仮定した．

そこで，/a/の逆フィルタ波形を用いてその人の自然音声のピッチと変動を TD-PSOLA 法により音源に与え，フォルマントもその人個人のものを用いて音声を/aiueo/と合成した．比較のため，音源波形を話者 risa の音源を用いた場合(A グループ)，その人の音源

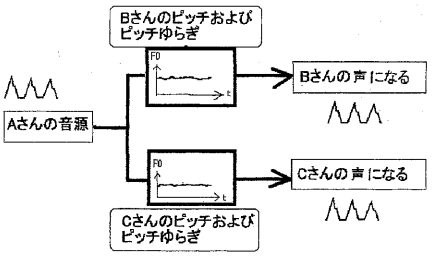

図 8：音源の性質

表 3：実験 3 の結果 1 (A グループ)

話者	個人性	音韻性	音質	総合
li	4.6	4.4	4.1	4.3
saori	4.5	4.5	4.2	4.2
risa	3.6	3.7	3.0	3.2
mori	4.5	4.4	3.8	4.2
wada	2.4	3.1	2.3	2.4

表 4 ：実験 3 の結果 2 (B グループ)

話者	個人性	音韻性	音質	総合
li	4.0	3.8	3.8	3.8
saori	3.9	4.0	3.8	3.8
risa	4.6	4.2	4.3	4.3
mori	4.3	4.3	4.1	4.1
wada	2.8	2.9	2.8	2.8

を使いた場合（B グループ）の 2 セット合成音声を作り，それがどの程度その人に聞こえるかを聞き取り実験で調べた．被験者は大学 2 年生 3 6 人である．評価は 5 段階評価とする．

5.2. 結果

結果を表 3,4 に示す．表 3 は評価の高かった話者 risa の音源をすべての人に使い基本周波数はその人の自然音声のものを使用したものである．表 4 は個人それぞれの音源を用いたが，その音源の基本周波数パターンを変更せず，そのまま使用したものである．この表 3,4 を見れば分かるが，どちらの波形を用いても個人を認識することが良好にできている．表 3 と表 4 を見比べると，話者 Risa, Wada では，他の話者と比較して成績が相反していることに気がつく．これは，TD-PSOLA 法でピッチを変更した際に音質の劣化が大きく，逆に，表 4 では，本人の無加工な音源を用いた場合に音質の劣化が起きず，その分結果が良かったと考えられる．

5.3. 考察

この結果から，個人性は，基本周波数によって決まる音源波形の調和成分の組み合わせ及び，その人のピッチゆらぎおよびフォルマント周波数で個人性が決定すると考える．音源波形の役割は，5 母音の源と，人間らしい声を保証するためだと考えられる．

図 8 に示すように,ある人の音源波形をベースにして，ある人の基本周波数およびその変動成分を与えた結果，その人の音源と聴覚上等価になることが分かった．

6. 基本周波数ゆらぎのモデル化

ピッチゆらぎを図 9 に示す．ピッチのゆらぎは話者により異なることが分かっており，それが話者の個人性を反映していると思われる．そのため，基本周波数に見られる変動特性をモデル化して個人性を統一的に記述し，音声合成や話者認識に応用していく．

図 10 は基本周波数の変動を隣り合うフレーム区間の差分を取ったものである．この時系列の情報を確率を用いて表現する．

図 11 はそのモデルパラメータを表したものである．図 12 はそれをプログラム化するために，オートマトンでモデル化したものである．ゆらぎはプラスとマイナスがほぼ同数対になって現われるため，プラス区間とマイナス区間およびゼロ区間の 3 つの状態に分ける．プラスのゆらぎが連続しているときは状態 S0，マイナスのゆらぎが連続している場合 S2，ゼロが連続しているは場合 S1 に遷移する．各パラメータの説明は以下の通りである．

A. 状態 S0, S1 について

(1)大きさ Xi のゆらぎが出現する確率[Sxi]

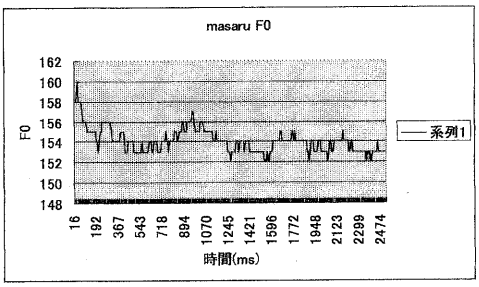

図 9：話者 Masaru のピッチゆらぎ

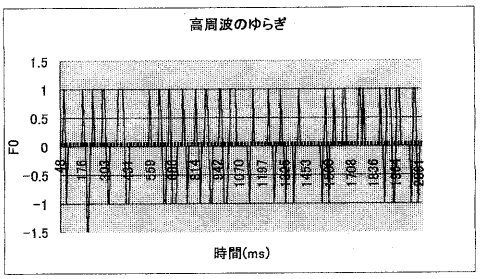

図 10：F0 の差分（変動成分）

(2) n 本のゆらぎが連続発生する確率[Snx]

(3)連続発生している状態でゼロが出現する確率[Szx]

(4)状態 S0 から状態 S2 に直接遷移する確率[S02]

(5)状態 S2 から状態 S0 に状態 S1 を介さないで直接遷移する確率[S20]

(6)S0 から S1 に遷移する確率，S2 から S0 に遷移する確率[S01,S20]

B. 状態 S1 について

(7) n 個のゼロが連続して出現する確率[Szx1]

(8) S0→S1→S0 に遷移する確率，S2→S1→S2 に移る確率 [S010,S212]

7. 今後について

今後はこのモデルの検証を行い，改良を重ねていく．

また，今回は地声のみを扱ったが，今後は様々な声質の特徴を調べていく．

謝辞

本研究の応用として，人工喉頭で実験をしてくださったセコム IS 研究所の今村俊樹様に感謝いたします．

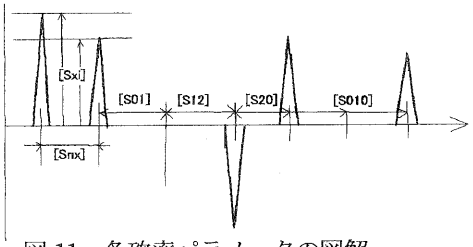

図 11：各確率パラメータの図解

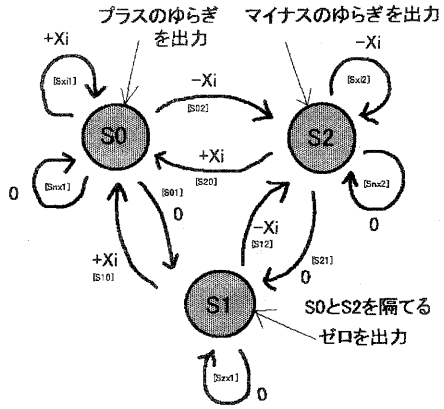

図 12：ゆらぎのモデル

参考文献

[1] 青木直史，伊福部達，"持続発声母音における振幅ゆらぎ及びピッチゆらぎの周波数特性とその音響心理的効果"，信学論，A, Vol. J82-A, No5 pp. 649–657，1999 年 5 月

[2] 皆川知也，赤木正人，"連続発話母音の基本周波数変動とその知覚"，信学技報，pp29—pp36，SP-134 (1998-3)

[3] 小室修，粕谷英樹，"基本周期のゆらぎの性質とそのモデル化に関する検討" 音響誌，47 巻 12 号，pp928—934，1991

[4] 遠藤康男，粕谷英樹，"周期ごとのゆらぎを考慮した音声の分析・変換・合成システム"，信学論 A Vol.J81-A No.7 pp1031—1041 1998 年 7 月

[5] 青木直史，伊福部達，"合成持続発声母音の自然性改善を目的とした音源波形揺らぎの生成とその主観及び客観評価"，信学論，D-II，vol. J82-D-II，No5，pp843-852，1999 年 5 月