社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2003－ＳＬＰ－49 (37)
2003／12／18

F0・パワー・フォルマントを用いた合成音声の明瞭性制御に関する検討

藤原 敬記† 広重 真人† 荒木 健治† 栃内 香次‡

†北海道大学工学研究科 〒060-8628 札幌市北区北 13 条西 8 丁目
‡北海学園大学経営学部 〒062-8605 札幌市豊平区旭町 4 丁目 1-40
E-mail: †{fujiwara, hiro, araki}@media.eng.hokudai.ac.jp, ‡tochinai@econ.hokkai-s-u.ac.jp

あらまし 自由会話では音声の全ての部分において常に高い明瞭性を保っているわけでない。例えば、重要な情報の含まれない部分や文末はそれほど明瞭に発話されない。人間は基本周波数、パワー、発話速度、調音位置などを使い明瞭性を制御していると考えられる。明瞭性は連続的に変化し、この変化が発話に流暢なリズムを生むと考えられる。本研究の目的は合成音声へ明瞭性の交替を導入することである。本稿ではF0、パワー、フォルマントを後処理加工することにより、合成音声の明瞭性を制御することを試みる。またSD法を用いた聴取実験により明瞭性の制御を導入した合成音声を評価した。その結果、明瞭性制御を行った音声は行わなかった音声に比べ、より「落ち着いた」「丸みのある」印象を感じることがわかった。
キーワード 明瞭性, 合成音声, フォルマント, F0, パワー

A Study on Clarity Control of Synthesized Speech with F0, Power and Formants

Noriki Fujiwara†, Makoto Hiroshige†, Kenji Araki†, Koji Tochinai‡

†Graduate school of Engineering Hokkaido University N13W8, Kita-ku, Sapporo, 060-8628 Japan
‡Graduate school of Bussiness Aministration Hokkai Gakuen University 4-1-40 Asahimachi, Toyohira-ku, Sapporo, 062-8605 Japan
E-mail: †{fujiwara, hiro, araki}@media.eng.hokudai.ac.jp, ‡tochinai@econ.hokkai-s-u.ac.jp,

Abstract In spontaneous conversational speech, all portions of speech do not always have high clarity. For example, the portions not having important information or the end of a sentence are not very clear. We consider that clarity of speech is controlled by fundamental frequency, power, speech rate, place of articulation and so on. We consider that the clarity changes continuously, and change of clarity of speech produce a fluent rhythm in human speech. The purpose of our research is introducing the change of clarity into synthesized speech. In this paper, we try to control clarity of synthesized speech by post-processing of fundamental frequency, power and formants. We evaluate the synthesized speech by auditory tests using SD method. The synthesized speech with control of clarity is better than the synthesized speech without control of clarity in several speech properties, e.g., calmness and smoothness.
Keyword clarity , synthesized speech, formants, F0, speech power

1. はじめに

人間にとって音声は最も身近で効率のよい情報伝達手段のひとつである。近年コンピュータの高速化などに伴い、音声による情報の入出力が普及してきたが[1][2]、音声による入出力の際に必要となる技術のひとつに合成音声技術がある。音声合成の研究は古くから行われており、現在では高品質の音声を合成できる合成システムも存在する[3]。それらのシステムの多くはより明瞭な音声を合成することを目指している。しかし人間の発話する自然音声を考えると、常に高い明瞭性を持って発話されるわけではない。重要な情報の含まれない部分や文末などは、それほどはっきりと発話されないことがある。発話内容や発話様式により韻律要素・音韻要素が変化し、それにより音声の持つ明瞭性も変化すると考える[4]。この明瞭性の変化は連続的に起こり音声に自然で流暢なリズムやテンポを与えると考え、本研究ではこの明瞭性の変化を合成音声へ導入することを目指す。本稿では、基本周波数（以下F0）、パワー、フォルマントを後処理加工することで合成音声の明瞭性を制御することを試みる。

2章では本研究における基本的な考えと、いくつかの発話様式のF0、パワー、フォルマントの特徴について述べる。3章ではF0、パワー、フォルマントの後処理加工の方法について述べる。4章、5章では、明瞭性の制御を目指し後処理加工した合成音声を使った聴取実験について述べる。

2. 基本的な考え方

2. 1 明瞭性制御と発話生成に関する考え方

この章では、本研究における2つの基本的な考え方を述べる。ひとつは明瞭性の制御に関する考え方であり、もうひとつは発話生成に関する考え方である。

発話生成に関する考え方を図1に示す。従来から音声の特徴は、大まかには分節的な音韻要素と超分節的な韻律要素に分類される。しかし、音韻の調音結合などは、分節にまたがって発生しており、音韻要素のなかの韻律的な要素と考えることができる。韻律要素にはF0、パワー、発話速度などがあるが、これらに上記の韻律的な音韻要素を加え、音韻要素にも韻律的な制御を行うことにより、発話全体としての了解度を損なわずに滑らかな音声を実現できる可能性があると考える。

明瞭性の制御に関しては、人間の発話において発話内容や発話様式によって韻律要素・音韻要素は変化し、さらに明瞭性も変化すると考える。つまり、韻律的な緊張の高い状態として大きな声、高い声、発話速度の変化が大きく、音韻のはっきりした発話を考え、韻律的な緊張の低い状態として小さな声、低い声、発話速度の変化が小さく、音韻のはっきりしない発話を考える。この韻律的な緊張の変化が流暢なリズムやテンポを生み、この韻律的な緊張の高低遷移が人間の音声の自然なリズム構成の一つの要因であると考える。

本研究では明瞭性の制御を目指し、韻律的緊張の遷移と音韻要素の中の韻律的特徴を合成音声へ導入することを試みる。

2. 2 予備調査

発話様式などによる韻律要素・音韻要素への影響を確認するため実際に音声を収録し、その特徴を調査した。

2. 2. 1 音声収録

日本語を母国語とする大学院生2名に、短文5文（表1）を3種類の発話様式で発話させた。発話者は意図的に発話様式を変えることに慣れている演劇経験者を選んだ。防音室にて48kHzで収録し、後に16kHzにダウンサンプリングした。短文ごとに文中のある文節を主題部分と定め、それ

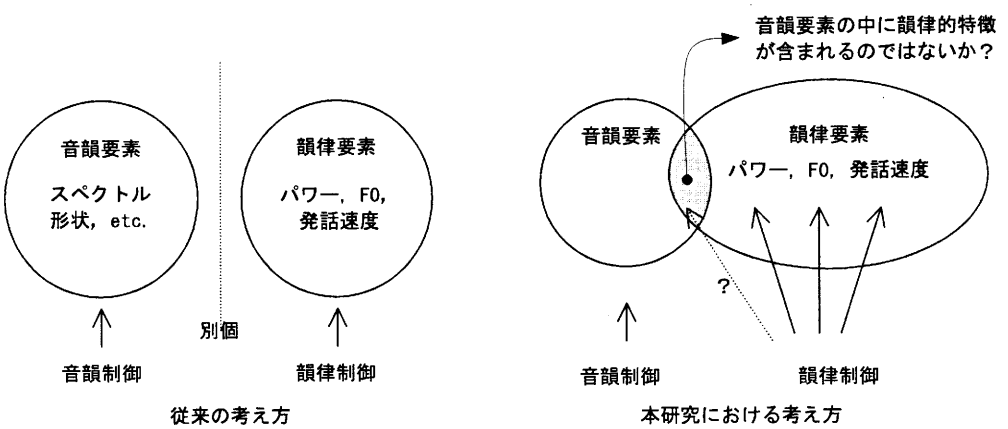


図1. 発話生成に関する考え方

表 1.　収録に用いた短文

文頭部分 （非主題部分）	文中部分 （主題部分）	文尾部分 （非主題部分）
次の停車駅は	北広島	です。
国内線ですから成田空港ではなく	羽田空港	です。
来週の月曜日は	開校記念日なので	休校です。
年末に家族で	海外旅行に	行った。
机の上の	青い本を	取って下さい。

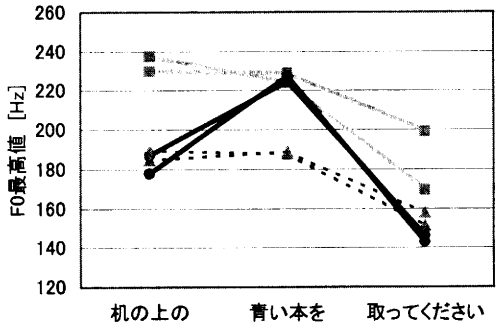


図 2. F0 の最高値

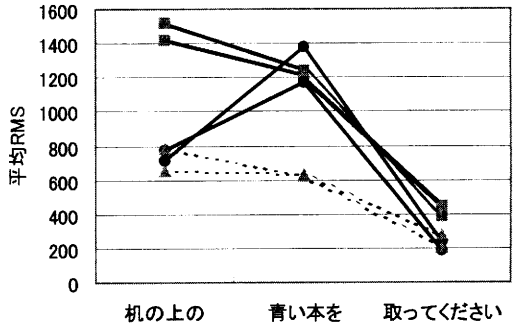


図 3. RMS パワーの平均

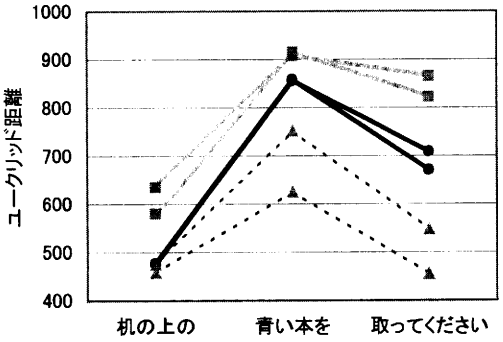
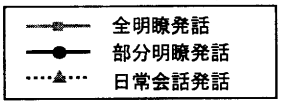


図 4. 中立母音との距離



以外の部分を非主題部分とした。主題部分・非主題部分は全ての文で「非主題部分・主題部分・非主題部分」の順になるように定め、以下各部分を「文頭部分」、「文中部分」、「文尾部分」と呼ぶ。3 種類の発話様式は、(1) 文全体をはっきりと発話する。(2) 主題部分のみをはっきりと発話し、非主題部分は日常会話程度のくだけた口調で発話する。(3) 文全体をくだけた口調で発話するものである。各発話様式をそれぞれ、(1)「全明瞭発話」、(2)「部分明瞭発話」、(3)「日常会話発話」と呼ぶことにする。例えば収録した 1 文の「次の停車駅は北広島です。」の場合、表 1 に示すように「北広島」を主題部分としており、文頭部分は「次の停車駅は」、文中部分は「北広島」、文尾部分は「です。」となる。

2. 2. 2　音声分析

各文の文頭・文中・文尾部分ごとに、「F0 の最高値」・「RMS パワーの部分内平均」・「中立母音からの距離」を求めた。「RMS パワーの部分内平均」とは、各部分においてフレームごとに求めた RMS パワーの合計値をフレーム数で割ったものである。「中立母音」とは損失のある一様音響管の周波数応答[5]と明瞭性を低く発話した実音声を参考にして設定したフォルマント周波数(F1=481, F2=1342, F3=2203, F4=33 16[Hz])を持つ中立的な母音である。この周波数は「schwa」のフォルマント周波数に近い特性を持つ。「中立母音からの距離」とは、中立母音の F1～F4 の周波数と当該フレームの F1～F4 の周波数とのユークリッド距離をフレームごとに求め、部分内での合計値をフレーム数で割ったものである。なお、F0 とフォルマント周波数は、明らかに抽出誤りと見られる箇所については手動による修正を行っている。F0 の修正は、倍ピッチを抽出していると思われ

る箇所はその値を 1/2 にし、また隣接する値と25Hz以上の差がありF0軌跡の連続的変化を損なう箇所は隣接する値に等しくなるようにしている。フォルマント周波数は、F2やF3を抽出できず、F3をF2の周波数、F4をF3の周波数としている箇所が多くあったので、これらのフォルマントのずれを修正し、抽出できていないフォルマントについては前後から線形補間を行い求めた。

2. 2. 3 結果と考察

結果の1例を図2、図3、図4に示す。これらの図より、部分明瞭発話のF0の最高値・RMSパワーの部分内平均・中立母音との距離は、文頭部分と文尾部分では全くだけ発話の値に近く、文中部分では全明瞭発話に近い値になることがわかる。中立母音との距離はF0やパワーの特徴に比べ、変化の差は大きくはないが、F0、パワーといった韻律要素と同様に、発話様式による違いが見られた。これにより発話様式の違いによっては音韻要素・韻律要素の特徴が変化することが確認された。

3. 合成音声加工

3. 1 加工する韻律要素・音韻要素の選択

従来あいまいな発話には、パワーの減少、母音の中立化、有声無声の切り替えの省略などが起こると言われている。合成音声の明瞭性を加工する要素を決定するために、数段階の明瞭性で発話した音声の特徴について調査した。日本語を母国語とする成人男性1名に数種類の句・短文を数段階の明瞭さで発話させた[4]。その結果、明瞭さを低く発話した音声には、「F0の平坦化」、「パワーの減少」、「スペクトルの中立化」が顕著に見られた。これにより合成音声の明瞭性を制御するための韻律要素・音韻要素として、F0・パワー・フォルマントを加工することにする。

3. 2 合成音声処理

合成音声の後処理加工は、まず合成音声を作成しF0・音源rmsパワー・フォルマントについて分析・再合成を行う。再合成の際、F0、パワー、フォルマントをそれぞれ独立に以下のように加工する。なお本稿では、合成音声の作成には「Visual Speech Creator」(NTT-IT(株)製)[6]を使用し、分析・再合成には、ESPS/waves+ [7]の簡単なフォルマント分析・合成器を用いた。パワー、フォルマントの加工は元の合成音声から求めたF0(以下、原F0)が低い部分にだけ原F0に同期させて行う。F0が低い部分とは、原F0の最大値 $F0_{max}$[Hz]と最小値 $F0_{min}$[Hz]の範囲において、

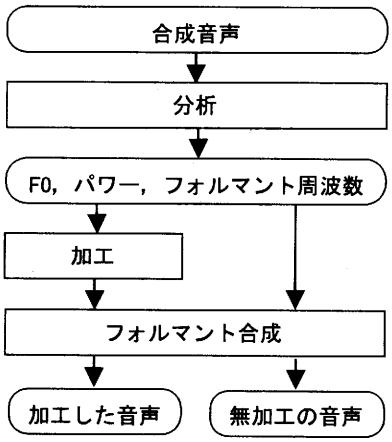


図5. 合成音声処理の流れ

原F0が式(1)で求まる $F0_{flag}$[Hz]よりも小さな部分を指す。

$$\log F0_{flag} = (\log F0_{\max} - \log F0_{\min}) \times 2/3 + \log F0_{\min} \quad (1)$$

3. 2. 1 F0 加工

F0の加工は、$F0_{min}$を固定して $F0_{max}$と $F0_{min}$ の範囲を r_{f0} に縮小させる。r_{f0}は音声サンプル全体にわたって一定の値とする。

3. 2. 2 パワー 加工

パワーの加工は、オリジナルの音源パワーに加工乗数 r_{pow} を乗じて求める。この加工乗数 r_{pow} は上述の原F0が低い部分で原F0の上下動に比例して変化するように定められ、音声サンプル中で原F0が $F0_{flag}$[Hz]以上のとき1.0、原F0が $F0_{min}$[Hz]のとき r_{pow_min} をとるようにする。

3. 2. 3 フォルマント 加工

フォルマントは、原F0の低下に伴って、2.2.2節で述べた中立母音のフォルマントに引き寄せられるように加工する。各時点で観測された各フォルマントを f_i 、加工後の各フォルマントを f_i' 、中立母音の各フォルマントを $f_{i\cdot lazy}$ (共に i=1,2,3,4)として、式(2)によりフォルマント加工を行う。r_{for} は r_{pow} と同様に原F0の上下動に比例して変化する加工乗数で、原F0が $F0_{flag}$[Hz]以上のとき1.0、原F0が最小値 $F0_{min}$[Hz]のとき r_{for_min} をとるように定める。

$$f' = f_{i\cdot lazy} + (f_i - f_{i\cdot lazy}) \times r_{for\ (i=1,2,3,4)} \quad (2)$$

今回各設定値は、(r_{f0} , r_{pow_min} , r_{for_min}) = (0.5, 0.5, 0.3) とした。

4. 聴取実験

F0、パワー、フォルマントの加工を非主題部分に行った音声（以下、部分加工音声）と、加工を行っていない合成音声（以下、無加工音声）（図5)を用いて聴取実験を行った。部分加工音声は分析・再合成を行っているが、簡易なフォルマント分析・合成器を用いたため、元の合成音声に比べ音質は明らかに劣化している。この音質劣化が実験に影響するのを除くために、図5に示すように無加工音声にも一度分析・再合成を行っている。

実験手法としてある事象に対する人間のイメージを評価するために用いられる[8][9]SD 法を用いた。被験者は日本語を母国語とする男子大学院生 13 名、雑音の少ない部屋でヘッドフォンを用いて音声を呈示した。回答用紙として CGI と Web ブラウザを用いた。短文5文について無加工音声と部分加工音声の2種類をそれぞれ作成し、計 10 種類の音声刺激を用意した。10 文の音声の順序をランダムにしたものを1セットとする。練習時には音声刺激を1セット呈示し、その後の本番時には3セット呈示した。尺度は[10]を参考に15尺度(表2)を選び、尺度呈示順は被験者ごとにランダムに変更した。音声刺激の間隔は練習時には 60 秒、本番時には被験者の慣れを考慮し 50 秒とした。SD 法の尺度は一対の形容詞対であるが、一般的に良い印象を持つ形容詞を「positive」側、もう一方の形容詞を「negative」側とし、7段階で評価させた。

5. 結果と考察

尺度ごとに平均を求め、被験者ごとに5種類の文で部分加工音声の方を positive に選んだ数を図5に示す。これを見ると、部分加工した音声は無加工の音声に比べより「静かな」、「落ち着いた」、「丸みのある」、「やわらかい」印象を持つことがわかる。「丸みのある」「やわらかい」という形容詞は人間性に関係する言葉であり、合成音声を加工することにより、自然音声に近づけられたと考える。

一方、「はっきりした」、「引き締まった」といった形容詞について極端に低い。この原因として、

表 2.　聴取実験に用いた形容詞対

	positive	negative		positive	negative		positive	negative
1	人間的	機械的	6	丸みのある	とげとげしい	11	明るい	暗い
2	静かな	騒々しい	7	好ましい	好ましくない	12	美しい	汚い
3	落ち着いた	甲高い	8	はっきりした	ぼんやりした	13	冷静な	感情的な
4	溶け合った	ばらばらな	9	調和のとれた	不調和な	14	分かりやすい	分かりにくい
5	快い	不快な	10	やわらかい	かたい	15	引き締まった	たるんだ

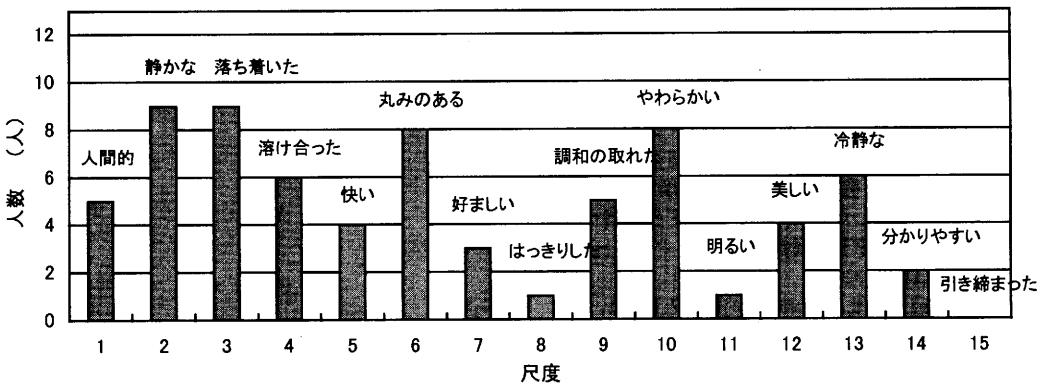


図 6.　無加工音声より部分加工音声の得点が高かった人数

パワーとフォルマントの加工乗数がF0の低下に伴い大きくなるので、文末のF0低下によってフォルマント・パワーの制御が強く行われたためと考えられる。実験後の内観調査では「文末が聞き取りにくく、文末の印象で選んだ」との意見もあった。この文末の影響が、被験者に部分加工音声に対してnegativeな印象をいだかせたひとつの理由と考えられるので、制御に関する設定値を再検討する必要がある。

6. まとめ

F0・パワー・フォルマントの後処理加工により合成音声の明瞭性を制御することを試みた。また後処理加工した合成音声をSD法による聴取実験で評価した。その結果、加工した合成音声は、「丸みのある」、「やわらかい」といった人間性に関係する形容詞について無加工の音声より強い印象を持つことがわかった。人間性を表す形容詞について強い印象を与えることより、F0・パワー・フォルマントを後処理加工することで人間の発話に一部近づけられたと考える。一方、「はっきりした」、「ひきしまった」と言った印象は加工した音声の方が極端に低いので、実音声を参考にするなどにより今後加工量等の検討が必要である。

謝辞

音声合成の際、日本音響学会第68回技術講演会で使用されたNTT-IT(株)製「Visual Speech Creator」を使わせていただきました。深く感謝致します。

参考文献

[1] 日本語音声認識/合成ソフト"LaLaVoice 2001", http://www3.toshiba.co.jp/pc/lalavoice/index_j.htm
[2] 連続音声認識プログラム"ViaVoice", http://www.scansoft.co.jp/viavoice/
[3] ニック・キャンベル, アラン・ブラック, "CHATR:自然音声波形接続型任意音声合成システム", 信学技法, SP96-7, pp.45-53
[4] 藤原敏記, 広重真人, 荒木健治, 栃内香次"明瞭度を考慮した規則合成の提案と基礎的検討", 日本音響学会講演論文集, 1-P-15, pp.381-382(2002.3)
[5] L.R.Rabiner/R.W.Schafer(鈴木久喜訳), 「音声のディジタル信号処理(上)」, コロナ社(1983)
[6] NTT-IT(株), "Visual Speech Creator 操作説明書"
[7] Software manuals of ESPS/waves+ with EnSigTM. 1997. Entropic Research Laboratory, Inc.
[8] 津崎実, 河井恒, "声質の長期的変動に関する印象空間 -方法論と実験結果-", 日本音響学会2003年秋季研究発表会講演論文集, pp.237-238
[9] 宮川雅充, 鈴木真一, 青野正二, 高木興一, "視覚情報が種々の環境音の印象に与える影響", 日本音響学会誌56巻6号(2000), pp.427-436
[10] 難波精一郎, 桑野園子, 「音の評価のための心理学的測定法」, コロナ社(1993)