(1993年8月31日受付)

【参 考 文 献】

- 1) 佐藤, 勾坂, 小暮, 嵯峨山: “日本語テキストからの音声合成”, 通研実報, **32**, 11, pp. 2243-2252 (1983)
- 2) 長尾監修: “日本語情報処理”, 電子通信学会 (1984)
- 3) 池原, 村上, 宮崎, 大山: “日本文音声出力システムの構成”, 通研実報, **35**, 2, pp. 145-155 (1986)
- 4) NHK 編: “日本語発音アクセント辞典”, 日本放送出版協会 (1985)
- 5) 勾坂, 佐藤: “日本語単語連鎖のアクセント規則”, 信学論 (D), **J66-D**, 7, pp. 849-856 (1983)
- 6) D. H. Klatt: “Software for a Cascade/Parallel Formant Synthesizer”, J. Acoust. Soc. Am., **67**, 3, pp. 971-995 (1980)
- 7) 北脇, 小林, 山本: “PARCOR 形音声分析合成系における符

- 号構成と回路構成”, 音響学会音声研資, S 75-13 (1975)
- 8) 管村, 板倉: “線スペクトル対 (LSP) 音声分析合成方式による音声情報圧縮”, 信学論 (A), **J64-A**, 8, pp. 599-606 (1981)
- 9) 酒寄, 佐々部, 北川: “数量化 I 類を用いた文韻律制御”, 音響学会講演論文集, pp. 195-196, Mar. (1987)
- 10) 水野, 広川, 中島, 箱田: “COC 音声合成ボード”, NTT R & D, **41**, 11 (1992)

きたがわ ひろお
北川 博雄 昭和 59 年, 慶應義塾大学大学院工学研究科博士課程電気工学専攻修了。同年, (株)リコーに入社。日本語音声情報処理, 3 次元物体認識, 医用画像処理システムなどの研究に従事。工学博士。

4-3 声 質 変 換

—自然な品質を保ちながら, 声の特徴を加工する技術—

都 木 徹†

1. ま え が き

声には, そのまま文字列に置き換えた言語としての情報以外に, 個人性 (その人らしさ) や感情といった非言語情報も含まれている。音声合成技術を用いて, これらの多様な情報を人間の発声と同等の自然性をもって制御できれば, 将来のマルチメディア等をより身近で暖か味のあるものにしていくことができる。さらに, 聴く人の好みや聴覚特性に合わせて放送等の音声の聴き易さを改善する装置 (例えば, 早口の聴取が苦手なお年寄り向けに, ゆっくりに変換する) をはじめ, ドラマや映画の新しい制作手法の開発にも寄与できよう。

「声質変換」とは, 上記のことを目的として, 音声に含まれる主に非言語的な特徴を加工して出力する技術である。本稿では, 最近開発された品質劣化の少ない変換手法の一例とその応用を中心に紹介する。

声質変換の対象となる声の特徴には, 声の高さや響きなど, いわゆる「声の質」ばかりでなく, 抑揚 (アクセントやイントネーション, 振幅, 話速) など含まれている。これら個々の特徴を独立または組合せて

制御することにより, ことばの聴きやすさ, 個人性, 感情などの制御が可能となる。

この際, 「音がひずむ」, 「ゴロゴロといった異音が混じる」, 「鼻づまりのような声になる」といった品質の劣化が生じると, その劣化自身が声の特徴変化の一要素となり, 不要な声質の変化が加わることになる。したがって, 声質変換においては, 変換音声の品質劣化を極力少なくすることが, 重要な技術課題である。

2. 声質変換の原理

2.1 発声の原理に基づいた変換

声質変換の技術的基礎となっているのは, 音声の「分析合成」である。これは, 自然の音声进行分析し, 声帯の振動や声道 (口や喉, 鼻で形作られる空間) での共鳴など。音声の性質を決定する特徴パラメータをそれぞれ独立に時系列の形で抽出し, その特徴パラメータに従って合成装置を動作させるものである。

声は, 一部の子音を除き, 肺から押し出された空気が声帯の振動によって周期的に断続された後, 声道において共鳴を受けることによって生成される。ただし, 鼻腔が関与する場合は反共鳴 (共鳴と反対に, ある周波数のエネルギーが吸収される) も加わる。声の高さは主にピッチ周波数 (声帯の振動周波数) によって決まり, その時間変化で抑揚が形成される。母音や子音の違いや個人性の主要な部分は, 声道の体積や形

† NHK 放送技術研究所

“Voice Quality Conversion Keeping Naturalness of Human Voice” by Tohru Takagi (NHK Science and Technical Research Laboratories, Tokyo)

図1 従来の声質変換方式（線形予測分析に基づく残差駆動方式）

状によって決まる。なお、個人性の一部は声帯の振動特性にも関連している。

声質変換では、入力音声波形から数ミリ秒前後の短い時間間隔で、「声帯の振動特性」と「声道の共鳴特性を表すスペクトル包絡」を別々に抽出し、それらの時系列を独立にまたは組合せて変更した後、再び音声波形を合成する。すなわち、発声の原理に基づいて音声波形の加工を行うのである。

例えば声の高さを変えるには、アナログテープレコーダの回転数を変えたり、デジタル処理で入出力のサンプリング周波数を変える方法も考えられる。しかし、これら従来法の欠点は、声帯の振動特性と声道の共鳴特性が独立に制御できないため、声の高さだけでなく個人性なども同時に変化することにある。

2.2 従来の分析合成法による声質変換

まず、従来の典型的な方法である線形予測分析に基づく手法を例に、分析合成による声質変換を説明する。図1は残差駆動方式と呼ばれるもので、合成装置の音源として、分析時に入力音声から分離された残差波形を保存しておき、これを用いる方法である。

分析部においては、まず入力音声に対して10ms前後の間隔で線形予測分析を行い、短時間内の音声波形中に現れた声道の共鳴特性を近似したスペクトル包絡 $H(\omega)$ を抽出する。線形予測分析では、 $H(\omega)$ を数個の共振点を持ったデジタルフィルタの伝達特性と仮定しており、 $1/H(\omega)$ 、すなわち声道の逆特性のフィルタで入力波形をろ波することにより残差波形 $e(n)$ が得られる。フィルタの係数はこの残差波形のパワーが最小になるように決定される。したがって、残差波形のスペクトル包絡はほぼ平坦なものとなり、入力音声に有声音（声帯振動を伴う音声）であればピッチ周期（声帯の振動周期）に対応したピークを持つパルス列状の周期波形となり、無声音（声帯振動を伴わない音声）であれば白色雑音に近くなる。一般の音声合成装置では、この残差波形を単純なパルス列や白色雑音等で置き換えることにより、情報量の大幅な削減をはかって

いる。しかし、この置き換えにより、自然さの情報が欠落するため合成音声の品質が低下してしまう。一方、残差駆動方式では、残差波形を保存して用いることで、品質の向上を図っている。

声の高さを変換するには、有声音部の残差波形の周期を変更する。また、声道の特徴を変換するには $H(\omega)$ の形を変形した新しい声道特性を持ったフィルタを構成する。変更された残差波形を、変更した声道フィルタに入力すれば、声質変換された音声波形が得られる。

この方式では、残差と $H(\omega)$ の両者に一切手を加えなければ原理的に入力波形が再現できるので、ピッチ周期やスペクトル包絡の変化量が少ない範囲では、比較的高い品質の変換音声を得られる。しかし、ピッチ周期を10%以上変化させると、特に女性の声の場合に大きな品質劣化を生じてしまう。残差波形の分離に用いるフィルタの特性 ($1/H(\omega)$) はあくまでも近似であるため、実際の残差波形の中には声道特性の一部が残っている。残差波形の周期を変更すると、その声道特性の一部にも変化を与えることになり、これがスペクトル的なひずみとなって出力音声の品質劣化につながると考えられる。

2.3 スペクトルを修正する高品質声質変換方式¹⁾

この方式の特徴は、従来の分析合成方式では避けられなかった有声音のピッチ周期変更に伴う品質劣化を、その原因であるスペクトル的なひずみを修正することにより低減させる点にあり、その原理を図2に示す。

分析部では、まず入力音声を、有声/無声/無音の各部分に分割し、有声音部に対しては、各ピッチ区間（声帯の1回の振動で生じる音声波形）を定め、個々のピッチ区間毎に線形予測分析等によりスペクトル包絡 $H(\omega)$ を抽出する。

次にピッチ制御部において、声の高さや抑揚に変化を与えるためピッチ周期を変更する。これは、各ピッチ区間毎の波形に対し、その区間長を短縮/延長する

図2 新たに開発された高品質な声質変換方式
(ピッチ周期変更時のスペクトルひずみを修正することを特徴とする)

図3 ピッチ区間長の延長と短縮

ことにより行う。短縮する場合には波形を途中で打ち切り、延長する場合には当該ピッチ区間と同じスペクトル包絡を持った後続波形を推定して接続する。発声速度を保つため、変更後の各ピッチ区間の波形を適宜間引くかまたは繰返して接続して有声音部全体の時間長を調整し、ピッチ制御部の出力とする。ただし、発声速度を変化させる場合は、間引きや繰返しの比率を増減する。これらの操作により、隣接するピッチ区間の接続部は波形が不連続となるので、接続点付近において曲線近似を行い、連続的に接続する。このようにしてピッチ区間長を変更した例を図3に示す。

ピッチ制御部の出力波形に対し、再びスペクトル包絡 $H'(\omega)$ を求めると、この包絡はピッチ変更の影響で変化しており、入力音声のスペクトル包絡 $H(\omega)$ とは異なる。この2つのスペクトル包絡の差異をピッチ

変更によるスペクトルのひずみと考える。

スペクトル制御部では、上記のピッチ変更によるスペクトルのひずみの修正を行うと共に、個性や明瞭さを制御するため、入力の実スペクトル包絡に現れている共振周波数等に変更を加える。これらのスペクトル包絡の修正や変更は、FFT（高速フーリエ変換）後のスペクトルの絶対値を制御することによって行う。すなわち、ピッチ制御部の出力波形をFFTによりスペクトル領域に変換して $X'(k)$ とし、次式により各周波数成分毎に絶対値を修正・変更して $X''(k)$ とする。

$$X''(k) = \frac{H(\omega_k)}{H'(\omega_k)} \cdot \frac{H''(\omega_k)}{H(\omega_k)} \cdot X'(k) \quad (1)$$

ここで、 $\omega_k = k/TN$ であり、 T はサンプリング周期、 N はFFTのポイント数、 $k=0, \dots, N-1$ である。式中の右辺の1番目の要素 $H(\omega_k)/H'(\omega_k)$ によってピッチ周期変更に伴うスペクトルのひずみが修正され、2番目の要素 $H''(\omega_k)/H(\omega_k)$ により共振特性が変更される。 $X''(k)$ は逆FFTによって時間領域の波形に戻される。

必要に応じて無声音区間の強調や抑圧、無声音区間の伸縮等も行い、変換された有声音区間と結合して、最終的な変換波形とする。

この方法により、元の音声のピッチ周波数を ± 1 オクターブ程度変化させてもかなり自然な品質が保てる。聴取実験による品質評価の結果、開発した方式が従来の残差駆動方式に比べ、ピッチ周期変更時の品質劣化が少ないことが確認された。

高品質な声質変換を目的とした方式として、短時間フーリエスペクトルを用いてピッチ周期をスペクトル領域における線形伸縮で変更する分析合成法²⁾³⁾も提案されている。

写真1 話速変換型受聴システム

写真2 抑揚変換装置

表1 無感情に発声された音声の物理パラメータを“悲哀”のものに変換した音声に対する“悲哀”と判定された割合
S：スペクトル包絡，T：時間構造，P：ピッチと振幅，T'：他人の時間構造，P'：他人のピッチと振幅

変換した パラメータ	S	T	P	S+T	S+P	T+P	S+T+P	S+T'+P'
判定率(%)	29	19	55	50	94	91	100	93

3. 声質変換の応用例

3.1 話速変換型受聴装置⁴⁾

お年寄りの聴覚特性の特徴のひとつに，言語識別速度の低下により，早口の音声聞き取りにくくなる現象がある．お年寄りの聴覚特性を最適に補償できるシステムを用いて，放送等の音声サービスを楽しんで頂くことを目的に，前述の高品質な声質変換方式を応用して，早口で話された音声を「ゆっくり」した音声に変換する装置（写真1）が開発されている．

本装置は，並列演算用LSI（FPU性能：2.3 MFlops）とDSP（同：80 MFlops）を複数縦並列接続して構成することにより，実時間処理が可能である．また，話速を外部から好みの速度に制御するためのボリュームを設け，「有声音部」と「ポーズなど比較的長い無音部の長さ」がそれぞれ独立に調節可能である．声の高さや個人性は保ったまま，高品質に話す速さのみを変換できる．現在，より小型な装置にするための検討が進められている．

3.2 音声の感情制御と抑揚変換装置⁵⁾⁶⁾

声質変換を用いて，感情表現に寄与する物理量を調べることができる．女性が無感情に発声した「あーそうですか」という声の，①スペクトル包絡，②時間構造（各母音や無音区間の時間長），③ピッチ周波数（振幅も同時に制御）の3種の特徴量を，本人または他人

が同一文を感情を伴って発声した音声のものに変換した．“悲哀”の感情についての主観評価結果を表1に示す．数字は，変換音声か“悲哀”の感情であると判断された割合(%)である．

ピッチ周波数単独の変換では，判断の割合が55%で半分程度の効果がある．ピッチ周波数と時間構造の組合せ（いわゆる抑揚）では91%とかなり効果があり，スペクトル包絡も含めすべてを組合せると完全となり，各特徴量の相乗効果がある．また，抑揚については他人のものに変換しても十分な効果がある．

この結果からもわかるように，抑揚は感情表現等において重要な役割を演じている．そこで，小型のワークステーションを用いて，簡単な対話形式の操作で入力音声の抑揚を変換する装置（写真2）を製作した．

画面に表示された音声波形の任意の範囲のピッチ周波数や無音区間長を，マウスを用いて所望の形や長さに自由に変形することにより，それに応じた変換音声か，元の話者の個人性は保ちつつ得られる．また，目標とする抑揚で発声した話者のものに機械的に似せて変換することもできる．今後は，使用目的に合わせた変換アルゴリズムの改良や使い易さの向上を図っていく予定である．

4. む す び

ピッチ周波数の変更によるスペクトル的なひずみを

修正することを特徴とした方式を中心に、音質変換技術について解説した。

近年のデバイス等の発達によりテキスト合成（文字情報を音声に変換する）技術も、情報量を削減することが前提の方式だけでなく、より品質を重視する方向へも発展しており、波形編集型⁷⁾等も提案されている。本稿で取り上げた声質変換方式は情報の削減を行わないので、波形編集などと組合せてテキスト合成に用いた場合、高品質化が期待できる。

より自然音声に近い品質で声質変換を行う手法は、音声研究を行う上で有力な道具になると共に、放送その他における高度な音声処理技術の基礎となるもので、今後一層の発展が期待される。

(1993年8月9日受付)

【参 考 文 献】

- 1) 都木, 梅田: “ピッチ変更時のひずみをスペクトル領域で修正する声質変換方式とその品質の心理評価”, 信学論, **J73-A**, 3, pp. 387-396 (Mar. 1990)

- 2) 阿部, 田村, 桑原: “FFT スペクトルからの Signal Reconstruction による音声変換手法”, 信学技報, SP88-48, pp. 25-32 (Sep. 1988)
- 3) F. J. Charpentier and M. G. Stella: “Diphone Synthesis Using an Overlap-add Technique for Speech Waveforms concatenation”, Proc. IEEE ICASSP86, **38**, 5, pp. 2015-2018 (Apr. 1986)
- 4) 中村, 清山, 池沢, 都木, 宮坂: “高品質話速変換型受聴システム”, 音響学会聴覚研資, H-92-54, pp. 1-10 (Nov. 1992)
- 5) 都木, 桑原, 梅田: “声質変換による, 感情表現制御のための物理的要因の解析”, 音響学会講演論文集, pz. 275-276 (Oct. 1989)
- 6) T. Takagi, E. Miyasaka: “A Speech Prosody Conversion System with a High Quality Speech Analysis-Synthesis Method”, EUROSPEECH'93, 31-3 (Sep. 1993)
- 7) 広川: “波形辞書を用いた規則合成法”, 信学技報, SP88-9, pp. 65-72 (May 1988)

高木 都木 とおる 昭和56年、電気通信大学通信工学専攻修士課程修了。同年、NHK入局。昭和59年より同放送技術研究所音響聴覚研究部勤務。声質変換、音声知覚、音声合成などの研究に従事。

4-4 話 者 認 識

古 井 貞 熙†

1. ま え が き

音声に含まれる個人性情報を用いて、誰の声であるかを自動的に判定することを、話者認識 (Speaker Recognition) という。これが実現できれば、電話によるバンキング、買い物、および情報案内サービス、音声メール、コンピュータのリモートアクセスなどにおいて、音声によって誰であるかが確認できるので、極めて便利になると期待されている。最近公開されたアメリカ映画「スニーカーズ」では、話者認識装置が建物への非登録者の侵入防止に用いられている。話者認識の詳しい原理や基本的方法などに関しては、すでに種々の解説があるので^{1)~4)}、本解説では、2~3節で基本的原理を説明した後、4節以降で最近の技術動向に絞って紹介したい。

2. 話者認識の原理

話者認識の形態は、話者識別 (Speaker Identification) と話者照合 (Speaker Verification) に分けられる。話者識別とは、入力音声は、あらかじめ登録されている誰の声であるかを判定することである。一方、話者照合では、入力音声と同時に自分が誰であるかを名乗り、その音声は本当にその話者の声であるか否かを判定する。いずれの場合も、入力音声と登録されている各話者の標準パターンあるいはモデルとの類似度を調べ、その値によって判定を行う。話者識別の場合は、多数の登録話者の中から最も類似度の高い話者を選び、話者照合の場合は、類似度が一定の閾値よりも大きければ名乗った本人の音声であるとして受理し、そうでない場合は他人の音声と判定して棄却する。音声を本人確認の一種の鍵として用いる多くの応用のほとんどは、話者照合に該当する。話者識別の性能は、登録話者の数が大きくなると低下するが、話者照合では、登録話者の数がある程度以上大きければ、性能はその数には依存しない。

† NTT ヒューマンインタフェース研究所
“Speaker Recognition” by Sadaoki Furui (NTT Human Interface Laboratories, Tokyo)