話者適応手法を用いた合成音声の個性化

佐々木　夏美†　佐藤　晴彦‡　小山　聡‡　栗原　正仁‡
北海道大学工学部†　北海道大学大学院情報科学研究科‡

1. はじめに

近年 Text-To-Speech(TTS)技術の発展により、スクリーンリーダーや公共交通機関のアナウンスなど多くの場面で合成音声が利用されるようになった。しかし、ロボットや CG キャラクターへの音声の付与などでは、読み上げ調だけでなく固有のキャラクターを持った音声が求められる。

また、多くの TTS が大量の音素 DB の中から適切な音素片データを繋ぎ合わせて音声を作る波形接続方式で、1話者の音声 DB の構築に膨大な収録・作業時間が必要となるため大量の話者モデルを作成することは難しい。様々な場面での要求に応えるためには少量のデータで多様な声質を実現することが望ましい。現状の音声合成システムから合成される音声を加工し、目標話者に適応させることができれば、容易に多様な個性をもつ音声をデザインすることが可能となる。

音声にはテキストとして表現できる言語情報の他に多くの情報が含まれている。我々は音声において個人性を表現・認識する際、話し方のスタイル・間の取り方・アクセント・感情表現など話者が意識的に変更可能であるパラ言語情報と、声の高さ・声質など身体的特徴によって決まる話者性情報を用いている。

北村による報告[1]では、母音間に共通する個人性情報が存在し、その情報は音声スペクトルに含まれることを示している。母音のスペクトル包絡を対象として個人性情報がどの周波数帯域に多く含まれるかを調べ、22ERBrate*1(2212Hz)以上の高周波数帯域に個人性情報を多く含むことを明らかとし、さらに基本周波数の動的成分・スペクトル概形も個人性知覚に寄与することも示している。

本論文では音声の個人性は母音の高周波数帯域特徴により表現されるという仮説を設け、母音の高周波帯域をパラメタとする話者適応手法を用いた個性を持つ合成音声の作成システムを提案する。

*1 ERB(Equivalent rectangular bandwidth)rate
Flether が提唱したヒトの聴覚フィルタ特性を考慮して作成された尺度. ERBS=21.4log(4.37f/1000+1)で表され、低周波数帯域では分解能が高く、周波数帯域が高くなるほど分解能は低くなる.

Personalization of Synthesized Speech with Speaker Adaptation
† Natsumi SASAKI; Faculty of Engineering, Hokkaido Univ.
‡ Haruhiko SATO; Satoshi OYAMA; Masahito KURIHARA; Graduate School of Information Science and Technology, Hokkaido Univ.

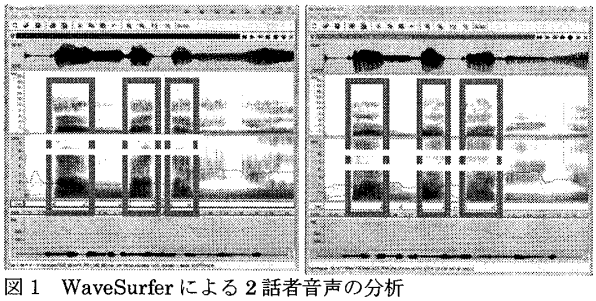
図1　WaveSurfer による2話者音声の分析

2. 仮説の検討

音声の個人性情報は母音の高周波数帯域に反映されるという仮説を検証するため、異なる2話者による同テキスト読み上げ音声を WaveSurfer[2]を用いて分析した。図 1 の赤枠は母音/a/の時間周波数表示である。テキスト中同位置母音の低次フォルマントの相対位置が話者間で良く似ているのに対し、黄枠で示すように高周波数帯のアンチフォルマントの位置は、話者内では類似性があるが話者間では大きく異なることが読み取れる。以上よりこの仮説の妥当性を確認した。

3. STRAIGHT による音声モーフィング

STRAIGHT[3]は音声を音源情報と伝達特性に分解し再合成する高品質な VOCODER であり、2音声試料を融合(モーフィング)することで話者変換・話者適応を可能とする技術である。

3.1 STRAIGHT パラメタ

合成に用いるパラメタは、基本周波数 pitch(有声/無声情報 vuv を含む)・非周期性指標 aperiodicity・時間周波数表現 spectrogram の3つである。

3.2 モーフィング手順

STRAIGHT 分析をして抽出したパラメタに関し、音声試料間で対応を取る。時間軸の特徴点として音声の開始・終了・フォルマントの変化点、周波数軸の特徴点としてフォルマントもしくはアンチフォルマントの位置などが用いられる。フォルマントとは周波数軸上でエネルギーの集中する位置、アンチフォルマントとは音が吸収されてしまう位置のことをさす。特徴点は知覚的に連続なモーフィングを行うために、時間周波数表現の座標をそろえる際に利用する。

モーフィング率 0 は元音声の、1 は目標音声の STRAIGHT 分析合成音に等しい。モーフィングを行って得られたパラメタを用いて STRAIGHT 分析

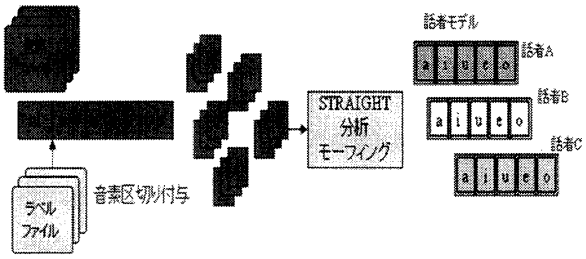

図 2　提案システムの学習部の流れ

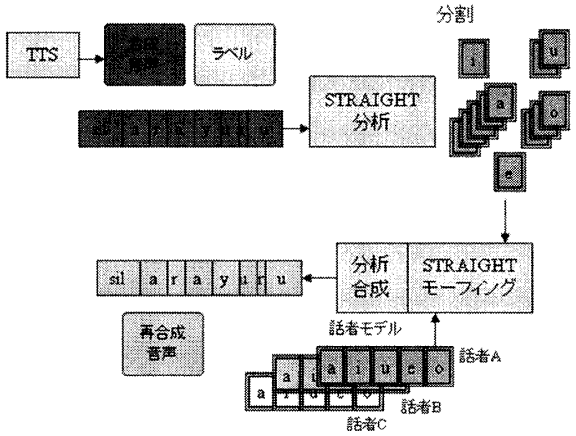

図 3　提案システムの合成部の流れ

合成を行い、モーフィング音声を作成する。なお、時間周波数の座標軸や強度、基本周波数などモーフィングに関するパラメタの一部を固定した部分モーフィングも可能である。

4. 提案システム

4.1　提案システムの構成

本提案システムは、話者ごとの特徴を分析・保存する学習部と、TTS システムなどから出力した音声ファイルを STRAIGHT 分析し学習済み話者のモデルとのモーフィングを行い話者適応した音声を出力する合成部の二つに大別される。

4.2　学習部

学習部の処理の流れを図 2 に示す。多数の話者の複数音声ファイルおよびその音素区切りを保存したラベルファイルを学習データとし、各音声の母音部分のみを取り出し STRAIGHT 分析を行う。

次に特徴点を付与し、母音ごとに 2 つずつモーフィング率 0.5 で順にモーフィングしていき最終的に各母音につき 1 つの平均スペクトルを得る。また、有声部分の平均基本周波数と平均話速(秒/モーラ)も求め話者の特徴量として保存する。

4.3　合成部

合成部の処理の流れを図 3 に示す。まず変換対象である元音声を STRAIGHT 分析しパラメタを抽出する。元音声中の位置を保持しながら分析データを母音ごとに分け、学習部で設定した点に対応するよう特徴点を設定する。その母音データを元音声中での継続時間長に合わせて話者モデルとモーフィングする。そして変換された各母音を元音声の該当位置へ挿入する。また、モデルの平均基本周波数・平均話速との差分を利用して全体に変換を行う。最後に合成部で再合成音声を出力する。

5. 実験

提案システム実装のための基本実験として以下の実験を行った。

5.1　音声試料

音声試料として発話のプロフェッショナルによる ATR25 文の読み上げ 8[モーラ/秒]の女性 2 名の音声[4]を使用した。

5.2　実験方法

まず目標話者 1 名について 1 音声ファイルを用いて話者モデルを作成した。その際、時間軸特徴点を開始・終了の 2 点、周波数軸特徴点を第 1～第 5 フォルマント位置の 5 点に手動で設定した。次に異なる話者の同テキスト音声ファイルを元音声として全周波数帯域について目標話者への母音モーフィング、基本周波数の変換を行った。

5.3　実験結果

合成されたモーフィング音声はある程度目標話者の声質に近づいていることを確認した。また、高周波数帯のアンチフォルマントの位置など現時点で十分反映されていない特徴が存在することを確認できた。なお、音素境界で少し不自然な接続が起こる部分があったが、これは音素境界部分のモーフィング率を低くし、徐々に変換率を高くすることで改善することができると考えられる。

6.おわりに

本論文では STRAIGHT 音声モーフィングを用いた合成音声の個性化手法について述べた。基本周波数および母音のみのスペクトル変換により合成音声から受ける印象を変化させられることが確認できた。今後の課題として、個人性をうまく捉える特徴点設定の仕方やモーフィング手法の検討、特徴点の自動設定方法、新たに高周波数帯特徴・基本周波数の動的成分・話速を特徴量して追加することなどが挙げられる。

謝辞

MATLAB 版 STRAIGHT の研究使用を許可してくださった和歌山大学河原教授に感謝いたします。

参考文献

[1] 北村: “音声における個人性の知覚と生成について,” 甲南大学紀要. 知能情報学編, pp.141-155 (2008).
[2] WaveSurfer: http://www.speech.kth.se/wavesurfer/index.html
[3] 河原: “聴覚の情景分析が生んだ高品質 VOCODER: STRAIGHT,”日本音響学会誌,54 巻,7 号,pp.521-526 (1998).
[4] 電気通信大学情報通信工学科/情報通信工学専攻高橋弘太研究室の話速バリエーション型音声データベース公開ページ: http://www.it.ice.uec.ac.jp/SRV-DB/SRV-DB_release.html