スペクトルと基本周波数のイベント操作による音声モーフィングに関する研究

藤野 善行

北陸先端科学技術大学院大学 情報科学研究科

2001 年 2 月 15 日

キーワード: 音声モーフィング、サブセットデータ、母音スペクトル.

1 はじめに

モーフィングとは、画像処理の分野で用いられている処理であり、映像中に描かれた物体を他の物体に変形させるものである。これと類似したことを音声処理で行なうことが音声モーフィングである。

音声モーフィングにも、その対象とする音響的特徴量によって様々な見方ができるが、本研究では、個人性・話者性に着目し、ある話者の音声の声質を別の話者の音声の声質に変化させる、複数話者間における声質変換を音声モーフィングと定義する。

これまでの音声モーフィングに関する研究では、モーフィング音声を作成するためには元音声だけなくモーフィング対象話者（目標話者）音声に関する膨大な量の音声データ（フルサイズデータ）が必要であり、その処理も複雑なものであった [1, 2]。

そこで、本研究では、任意の話者に関するフルサイズデータが存在すると仮定し、それらを目標話者に関する少量の情報（サブセットデータ）を用いて音声モーフィングを試みる。また、用いたモーフィングパラメータがモーフィング音声に与える影響を検討する。

2 音声分析合成系

本研究では、主に母音スペクトルと基本周波数を中心とした個人性情報を取り扱っていく。そのような合成音声を作成するためには、音声から個人性情報を抽出し、また母音成分の情報を独立に扱える音声分析合成系を用いる必要がある。このための音声分析合成系として、音声をスペクトルと基本周波数に分解でき、高品質な合成音声を合成できる

STRAIGHT（Speech Transformation and Representation basedon Adaptive Interpolation of weiGHTed spectrogram）[3] を、スペクトルを子音成分と母音成分に分解できるテンポラルデコンポシション S^2BEL-TD (Spectral Stability Based Event Localizing Temporal Decomposition)[4] を採用する。

3　モーフィングパラメータ

本研究では、サブセットデータによる音声モーフィングを試みる。より少量のサブセットデータで効果的な音声モーフィングを行なうためには、どのモーフィングパラメータをサブセットデータとして選択するかが重要な問題となってくる。

本節では、音声モーフィングを行なうための具体的なパラメータを取り上げ、その内容を説明する。

3.1　モーフィングを行なう音声データ

3.1.1　話者 A（フルサイズデータ）

モーフィングに使用する話者 A のフルサイズデータとしては、ATR 音声データベース男性話者 mms に関する情報を用いた。音声データは半母音や拗音を含まず、母音または有声・無声子音で構成された単語「そびえる」を採用した。

3.1.2　目標話者 B（サブセットデータ）

モーフィングに使用するサブセットデータを得るための話者 B には、24 歳の男性の大学院学生を用いた。

3.2　単独発話母音イベント

文の構成単位は単語であり、単語は音節からなる。日本語の音節は、通常、子音と母音の組からなる。また、母音は子音に比べて通常長い継続時間長を持ち、スペクトルも比較的明確である。よって、母音は通常容易にかつ確実に認識されることができるので、人間による音声認識でも機械による音声認識でも、重要な役割を果たしている。

よって、最小限の情報として、目標話者 B の単独発話 5 母音（/a/,/i/,/u/,/e/,/o/）をサブセットデータとして用いた音声モーフィングを試みる。

3.3　なまけ情報

連続音声中では、単音の物理的性質はその置かれた環境によって単独に発声された場合とは異なってくる現象、調音結合が生じている。この調音結合の影響により、なまけ ’の

現象が現れる。すなわち、サブセットデータとして、さらに‘ なまけ ’情報を考慮することにより、より自然なモーフィング音声を生成することが期待できる。

この‘ なまけ ’情報を付与する手法として、3 連続母音における第 2 母音イベントによる音声モーフィングを試みた。

3.3.1 音声データ（3 連続母音）

録音条件は前説で説明した通りである。目標話者 B に関するサブセットデータとして採取した音声データは、目標話者 B による 3 連続母音/oie/、/ieu/である。これらは、モーフィグ対象音声データ「そびえる」に対応した 3 連続母音である。これらの第 2 母音/oie/、/ieu/を用いた。また、「そびえる」における第 1 母音/o/、第 4 母音/u/ に関しては、3 連続母音における/oie/、/ieu/を用いた。

3.4 基本周波数

基本周波数の時間的変化パターンには個人性が多く含まれる [5]。よって、本研究では、この基本周波数に関連した個人性情報をモーフィングパラメータとして採用した。

3.4.1 平均基本周波数

目標話者 B に関する最小限の基本周波数情報として、目標話者 B の平均基本周波数を採用した。

3.4.2 基本周波数イベント

話者によって基本周波数の時間的変化パターンは異なるが、この基本周波数の立ち上がりの違い、すなわちアクセントの強さの違いによって音色は変化する。

そこで、同話者では単語におけるアクセントの強さには違いがないものと仮定し、代表的なアクセント型の音声データをサブセットデータとして用いる。

4 聴取実験

モーフィングを行なった音声が、どれだけ目標話者に近づいたかを確認するために、また、入れ替えたパラメータによるモーフィング音声への影響を調べるために、ABX 法による聴取実験を行う。

表 1にモーフィング手法を、表 2は実験で用いたモーフィング音声の一覧を示す。

結果は図 1のようになった。−3 に近づくほど話者 A に、3 に近づくほど目標話者 B に似ていることを表している。

図 1から、操作 X-1 と操作 X-2 における音声は平均値にもその差が認められる。

表 1: モーフィング手法

操作	内容
Fev	話者 A の基本周波数イベント変化を目標話者 B に合わせる
Fav	話者 A の平均基本周波数を目標話者 B の平均基本周波数に合わせる
X-1	話者 A の母音イベントを話者 B の単独発話母音イベントに入れ替える
X-2	話者 A の母音イベントを話者 B の 3 連続母音中の第 2 母音イベントに入れ替える
X-3	話者 B の母音イベントを話者 B の 3 連続母音中の第 2 母音イベントに入れ替える

表 2: 実験で用いたモーフィング音声

音声	a	b	c	d	e	f	g	h	i
Fev	○				○			○	
Fav		○		○			○		
X-1			○	○	○				
X-2						○	○	○	
X-3									○

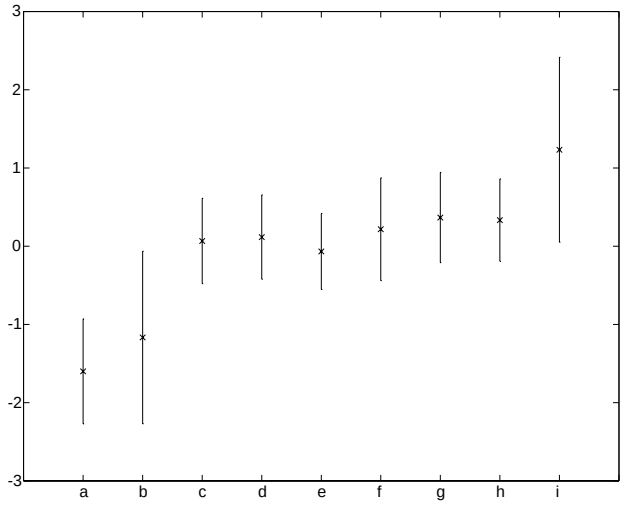

図 1: モーフィング音声「そびえる」の付置関係とその平均、標準偏差

また、得られたモーフィング音声に対し、F 検定および T 検定を行なった。その結果、母音イベント操作の違い（X-1 と X-2）において、音声 e と h にのみ有意な差が認められたが、その差は大きなものではない。

しかし、単独発話母音イベントを用いたモーフィング音声よりも 3 連続母音イベントを用いたモーフィング音声が、よりなめらかで自然な音声だと知覚されたことから、調音結合により生成する‘ なまけ ’は合成音声の自然性を表わすことができるといえる。すなわち、個人性のみ着目した音声モーフィングを行なうためには、操作 X-1 でも十分であるが、さらに音声の自然性を保った音声モーフィングを行なうためには、操作 X-2、すなわち 3 連続母音イベントは重要なパラメータであることがいえる。

基本周波数の変化による声質の変化に関しては、スペクトルの変化によるものほど顕著には表れなかった。

以上のことを考慮した結果、サブセットデータで効果的な音声モーフィングを行なうにあたって、サブセットとして盛り込むべきモーフィングパラメータを母音イベントと平均基本周波数および基本周波数イベントとすることは必要条件であることがいえた。

5 おわりに

本論文では、目標話者に関するサブセットデータを用いて音声モーフィングを行なった。用いたモーフィングパラメータがモーフィング音声に与える影響を検討し、効果的な音声モーフィングを行なうためにサブセットデータとして取り入れるべきパラメータを求めた。

その結果、3 連続母音イベント、平均基本周波数および基本周波数イベントは効果的なモーフィングを行なうためには重要なパラメータであることがわかった。すなわち、目標話者に関するこれらの情報が含まれた音声、『3 連続母音全ての組み合わせを中心に構成され、かつ多くのアクセント情報が盛り込まれた文章』をサブセットデータとして用いることが、サブセットデータでフルサイズデータの音声モーフィングを行なうための必用条件であることを示した。

得られたモーフィング音声は完全なものとはいかなかったが、サブセットデータで音声モーフィングを行なうための目安を明らかにした。

参考文献

[1] 阿部, “基本周波数とスペクトルの漸次変形による音声モーフィング,” 日本音響学会講演論文集, 2-1-8, pp.259-260, 1995.

[2] 坂野秀樹, 武田一哉, 板倉文忠, “包絡と音源の独立操作による音声モーフィング,” 信学技報, SP96-6, May 1996.

[3] 河原英紀, “聴覚の情景分析と高品質音声分析変換合成法 STRAIGHT,” 音学講論, 1-Q-21, pp.183-184, Oct.1994.

[4] A.C.R.Nandasena and M.Akagi, “Spectral Stability Based Event Localizing Temporal Decomposition,” Proc.ICASSP98, II, 957-960

[5] 家永太郎, 赤木正人, “音声のピッチ周波数の時間変化パターンに含まれる個人性とその制御,” 信学技報, SP94-104, May, 1995.