

# 包絡と音源の独立操作による音声モーフィング

坂野 秀樹<sup>†</sup>

武田 一哉<sup>††</sup>

鹿野 清宏<sup>†</sup>

板倉 文忠<sup>††</sup>

## Speech Morphing by Independent Interpolation of Spectral Envelope and Source Excitation

Hideki BANNO<sup>†</sup>, Kazuya TAKEDA<sup>††</sup>, Kiyohiro SHIKANO<sup>†</sup>, and Fumitada ITAKURA<sup>††</sup>

あらまし スペクトル包絡と音源の独立操作により、ある話者の音声を別の話者の音声へと連続的に変化させる音声モーフィングを提案する。本手法では次の手順で音声モーフィングを実現する。1) 時間領域における DP マッチングにより単位波形の対応をとる。2) 単位波形をスペクトル包絡と音源に分離する。3) 周波数領域の DP マッチングにより周波数軸を非線形に伸縮し、スペクトル包絡間の対応付けを行う。4) スペクトル包絡および音源の補間を行う。5) 位相情報を付与し、単位波形を得る。6) PSOLA 法により合成する。この手法を用いることによって自然音声の時間的変化に比較的近い補間が可能となり、音声の調音結合部分をモーフィングにより生成する実験を行った結果、ケプストラム距離において従来法に比べ 1.9 dB ひずみを減少させることができた。また、対比較試験では男性から女性へのモーフィングにおいて 89%、女性から男性へのモーフィングでは 93% の割合で本手法の方が品質が良いと判断されており、本手法の有効性が示された。

キーワード 音声モーフィング、補間、スペクトル包絡、周波数軸の非線形伸縮

### 1. ま え が き

モーフィングはコンピュータグラフィックスなど映像の分野で用いられている技術であり、映像中に描かれた物体 A を物体 B へと変形させるものである。音合成研究の一環として、映像の代わりに音を対象としたサウンドモーフィングの研究がいくつか行われており [1]~[4]、これは二つの異なる音色をもつ原音間を連続的に補間するものである。我々は音声研究の立場から、対象を特に音声に限定した“音声モーフィング”について検討を行う。

“音声モーフィング”の定義は、ある音声 A を音声 B へと連続的に変化させるものであるということが出来る。この音声モーフィングを高品質に実現するためには、音声は滑らかに変化することと、音声は自然に変化することの二つの条件を満たす必要があると考えられ、本研究ではこれら二つの条件を満たすモーフィングシステムの構築を目的としている。但し、モーフィ

ングにより生成される中間的な音声（これを“モーフィング音声”と呼ぶことにする）は現実には存在しない音であるためその自然性を判断するのはそれほど簡単なことではない。そこでモーフィング音声の自然性をモーフィングによる音声の変化がどれだけ自然音声の変化に近いかという観点で評価し、滑らかで自然に変化するという条件を満たすモーフィング手法を提案する。

阿部らは次のようなスペクトルの入替えによるモーフィングを提案している [5]。まず、ピッチ同期で FFT 分析を行い、ある周波数を境にスペクトルの入替えを行う。この処理を FFT で得られたスペクトルの実部と虚部の両方に対して行い、最後に逆 FFT をして単位波形を得、PSOLA 方式により基本周波数を変更して合成することにより音声モーフィングを実現している。この方式を PSS (Progressive Substitution of Spectra) と呼ぶことにするが、ある周波数を境にスペクトルの入替えを行うこの手法により生成されるモーフィング音声は、“自然”であるかどうかは疑問である。そこで我々は自然なモーフィング音声を生成する音声モーフィング方式として、スペクトル包絡と音源の独立操作による音声モーフィング (Progressive

<sup>†</sup> 奈良先端科学技術大学院大学情報科学研究科，生駒市  
Graduate School of Information Science, Nara Institute of Science and Technology, Ikoma-shi, 630-0101 Japan

<sup>††</sup> 名古屋大学大学院工学研究科，名古屋市  
Graduate School of Engineering, Nagoya University, Nagoya-shi, 464-8603 Japan

Interpolation of Spectra: PIS) を提案する。本方式は以下の理由から自然なモーフィング音声の生成が可能である。

同一話者が同一音素を発声した場合でもさまざまな原因により声質が変わることが知られているが、その原因の一つとしてホルマント周波数の移動が考えられる。本方式ではスペクトル包絡と音源を分離し、周波数軸の非線形伸縮によりスペクトル包絡間の対応付けを行っている。このため、補間に伴うホルマントの消滅や分裂が生じず、自然音声のスペクトル包絡の変化に比較的類似したスペクトル包絡の補間が可能である。

以下 2. で本方式のアルゴリズムについて述べ、3. でモーフィングの性能を主観・客観の両面から評価する。最後に 4. において本論文の結論を述べる。

## 2. 音声モーフィング処理

提案する手法を以下に述べる。

### 2.1 DP マッチングによる時間方向の対応付け

まず、ある話者 A と別の話者 B の音声データを用意し、あらかじめ有声音の波形のピークの位置に基本周期を示すピッチマークを付与しておく。ここでは文献[6]の方法を用いている。そして、ピッチマークを中心に 2 ピッチ周期分の長さで波形を切り出す。以下、このように切り出した波形を単位波形と呼ぶことにする。更に、時間領域において、単位波形の LPC ケプストラムを距離尺度とした DP マッチングを行い、話者 A, B の単位波形間の対応付けを行う。この処理によって以降の処理は対応づけられた単位波形間で行われる。なお、無声音についてはピッチとは無関係に処理を行う。

また、ここでモーフィング音声に話者 B の音声を含む割合（モーフィング率） $r$  ( $0 \leq r \leq 1$ ) を単位波形ごとに決定しておく。

### 2.2 単位波形の切出し

対応づけられた単位波形のピッチ間隔 ( $T_l^{(o)}$ ) とモーフィング率から合成時のピッチ間隔 ( $T_l^{(s)}$ ) を求めた上で、合成時を考慮し以下のように再度波形の切出しを行う。

今、あるピッチマークを中心に波形を切り出すとする。このとき、そのピッチマークと先行するピッチマークとの間隔  $T_l^{(o)}$  が、合成時におけるその時点のピッチ間隔  $T_l^{(s)}$  よりも大きい場合は長さ  $T_l^{(s)}$  で切り出し、小さい場合は長さ  $T_l^{(o)}$  で切り出す。同様に、後続のピッチマークとの間隔  $T_r^{(o)}$  が、合成時における

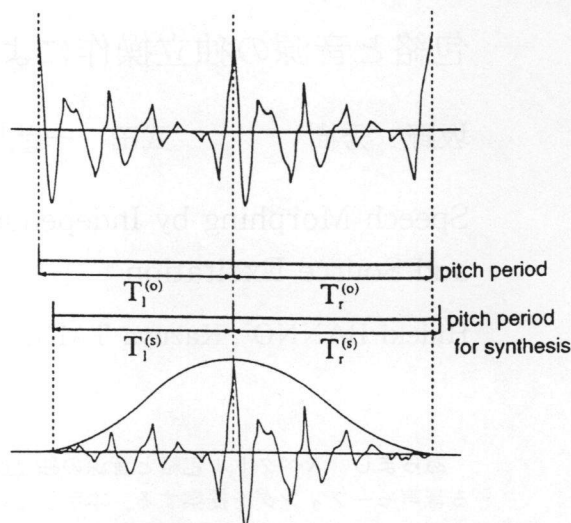


図 1 単位波形の切出し  
Fig. 1 Windowing of unit waveform.

その時点のピッチ間隔  $T_r^{(s)}$  よりも大きい場合は  $T_r^{(s)}$ 、小さい場合は  $T_r^{(o)}$  で切出しを行う (図 1)。

このようにピッチに同期した非対称な窓で波形の切出しを行うことによって、スペクトルにはピッチ構造と呼ばれる周期性は現れなくなる。そこで、以下ではスペクトルの微細構造を取り除き平滑化したものをスペクトル包絡成分、スペクトルの微細構造を音源成分として説明を行う。

### 2.3 スペクトル包絡成分と音源成分の分離

切り出した単位波形のスペクトルをスペクトル包絡成分と音源成分とに分離する。ここでは FFT ケプストラムをリフタリングすることにより、スペクトル包絡の抽出を行っている。

まず、FFT ケプストラム  $c(n)$  は

$$X(k) = \sum_{t=0}^{N-1} x(t) e^{-j \frac{2\pi}{N} kt} \quad (0 \leq k \leq N-1) \quad (1)$$

$$c(n) = \frac{1}{N} \sum_{k=0}^{N-1} \log |X(k)| e^{j \frac{2\pi}{N} kn} \quad (0 \leq n \leq N-1) \quad (2)$$

により求める。但し、 $x(t)$  は単位波形に対し零づめを行ったものを用いる。この FFT ケプストラムを  $n_\alpha$  次を境に分離し、スペクトル包絡のケプストラム表現  $e(n)$  および音源のケプストラム表現  $s(n)$  を求める。



$$e(n) = \begin{cases} c(n) & (0 \leq n < n_\alpha, N - n_\alpha \leq n \leq N - 1) \\ 0 & (n_\alpha \leq n < N - n_\alpha) \end{cases} \quad (3)$$

$$s(n) = \begin{cases} 0 & (0 \leq n < n_\alpha, N - n_\alpha \leq n \leq N - 1) \\ c(n) & (n_\alpha \leq n < N - n_\alpha) \end{cases} \quad (4)$$

次に、スペクトル包絡のケプストラム表現  $e(n)$  を FFT することで、対数スペクトル包絡  $E(k)$  ( $e(n)$  が偶関数であるから実数部のみとなる) を求める。以後これを単にスペクトル包絡と呼ぶことにする。

$$E(k) = \sum_{n=0}^{N-1} e(n) e^{-j \frac{2\pi}{N} kn} \quad (0 \leq k \leq N-1) \quad (5)$$

なお、リフタ長  $n_\alpha$  を長くとると、スペクトル包絡に微細構造を多く含むこととなるが、その場合、次節で説明するスペクトル包絡の対応が単位波形ごとに大きく異なるという現象が発生し、聴感上不連続な音声合成される。そこで、そのような不連続の発生を抑えるために、リフタ長をやや短めの 1.25 ms (10 ポイント) に設定した。

#### 2.4 スペクトル包絡成分のモーフィング

スペクトル包絡を補間する場合、補間方法によってはホルマントが二つ現れたり、ホルマントがなくなったりするため、以下のような手順で処理を行う。

まず、周波数領域における DP マッチングにより周波数軸を非線形に伸縮し、話者 A のスペクトル包絡  $E_A(k)$  と話者 B のスペクトル包絡  $E_B(k)$  の対応付けを行う (図 2)。DP マッチングは周波数 0 を始点、 $N/2$  を終点とする始点・終点固定の単純なものであり、この処理により話者 A の周波数  $k_A$  に対応する話者 B の周波数  $k_B$  がパス関数  $\theta$  を用いて

$$k_B = \theta(k_A) \quad (0 \leq k_A, k_B \leq N-1) \quad (6)$$

と求められる。得られた周波数間の対応に基づき、モーフィング率  $r$  に応じたモーフィング音声のスペクトル包絡  $E_M(k)$  を以下により求める。

$$E_M(k) = (1-r)E_A(k_A) + rE_B(\theta(k_A)) \quad (0 \leq k \leq N-1) \quad (7)$$

ここで、周波数  $k$  はモーフィング率  $r$  に応じて

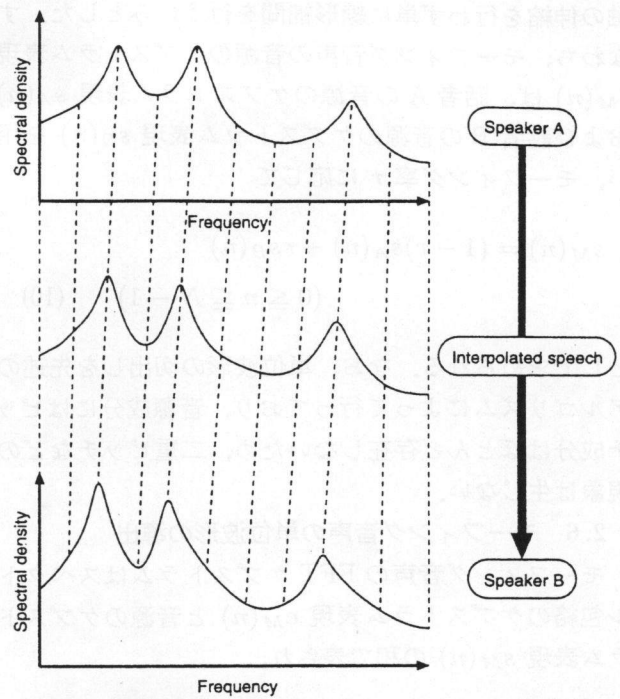


図2 周波数軸の非線形伸縮によるスペクトル包絡の補間  
Fig.2 Interpolation of spectral envelope based on nonlinear frequency warping.

$$k = (1-r)k_A + r\theta(k_A) \quad (0 \leq k_A \leq N-1) \quad (8)$$

により求められる。

一般に  $k$  は連続な値をとるとは限らないため、 $E_M(k)$  に不連続点が生じた場合は線形補間によって補う。

このモーフィング音声のスペクトル包絡  $E_M(k)$  を逆 FFT することにより、モーフィング音声のスペクトル包絡のケプストラム表現  $e_M(n)$  が求められる。

$$e_M(n) = \frac{1}{N} \sum_{k=0}^{N-1} E_M(k) e^{j \frac{2\pi}{N} kn} \quad (0 \leq n \leq N-1) \quad (9)$$

これら一連の処理は原音声におけるケプストラムの低次成分のみに対して行うため、このスペクトル包絡のケプストラム表現  $e_M(n)$  の高次成分はほぼ零の値をとる。

#### 2.5 音源成分のモーフィング

音源成分の補間についても、スペクトル包絡の場合と同様にケプレンシー軸の非線形伸縮により対応付けを行うことができるが、単位波形ごとに対応が大きく異なってしまうため、合成された音声はかなり不連続なものになってしまう。そこで、音源成分については、

軸の伸縮を行わず単に線形補間を行うのみとした。すなわち、モーフィング音声の音源のケプストラム表現  $s_M(n)$  は、話者 A の音源のケプストラム表現  $s_A(n)$  および話者 B の音源のケプストラム表現  $s_B(n)$  を用い、モーフィング率  $r$  に応じて

$$s_M(n) = (1-r)s_A(n) + rs_B(n) \quad (0 \leq n \leq N-1) \quad (10)$$

として求められる。なお、単位波形の切出しを先述のアルゴリズムによって行っており、音源成分にはピッチ成分はほとんど存在しないため、二重ピッチなどの現象は生じない。

## 2.6 モーフィング音声の単位波形の導出

モーフィング音声の FFT ケプストラムはスペクトル包絡のケプストラム表現  $e_M(n)$  と音源のケプストラム表現  $s_M(n)$  の和で表され、

$$c_M(n) = e_M(n) + s_M(n) \quad (0 \leq n \leq N-1) \quad (11)$$

となる。更にこれを FFT し、指数をとることでモーフィング音声の振幅スペクトル  $|X_M(k)|$  が求められる。

$$|X_M(k)| = \exp \left( \sum_{n=0}^{N-1} c_M(n) e^{-j \frac{2\pi}{N} kn} \right) \quad (0 \leq k \leq N-1) \quad (12)$$

この振幅スペクトルを用いて零位相や最小位相で音声を合成することも可能であるが、位相情報が失われることによる音質の劣化が生ずる。そこで、今回は比較のため、次のような手順で PSS 法と同等の位相情報を付与することにした。

モーフィング音声の振幅スペクトル  $|X_M(k)|$  に対し、周波数  $k_\beta$

$$k_\beta = rN/2 \quad (13)$$

を境界として低域に話者 B の位相を、高域に話者 A の位相を付与し、複素スペクトル  $X_M(k)$  を得る。

$$X_M(k) = \begin{cases} |X_M(k)| \frac{X_B(k)}{|X_B(k)|} & (0 \leq k < k_\beta, N - k_\beta \leq k \leq N-1) \\ |X_M(k)| \frac{X_A(k)}{|X_A(k)|} & (k_\beta \leq k < N - k_\beta) \end{cases} \quad (14)$$

これを逆 FFT することにより、モーフィング音声の単位波形  $x_M(t)$

$$x_M(t) = \frac{1}{N} \sum_{k=0}^{N-1} X_M(k) e^{j \frac{2\pi}{N} kt} \quad (0 \leq t \leq N-1) \quad (15)$$

が求まる。

## 2.7 モーフィング音声の合成

モーフィング音声の単位波形を所望のピッチ間隔で配置することにより (PSOLA 法 [7], [8]) モーフィング音声を合成する。このとき、波形の繰返し・間引きにより話速も連続的に変化させる。

## 3. 評価実験

本章では提案するモーフィング手法の性能を PSS 法との比較で評価する。

### 3.1 モーフィング音声のスペクトル包絡

モーフィング音声のスペクトル包絡を本手法の場合と PSS 法の場合について示し、それらを比較する。

#### 3.1.1 実験条件

音声データには「あらゆる現実をすべて自分の方へねじ曲げたのだ」を用い (日本音響学会研究用連続音声データベースより)、話者 A を男性話者、話者 B を女性話者とした。

本手法によるモーフィングの実験条件を表 1 に示す。PSS 法についても同様の実験条件であるが、PSS 法ではスペクトル包絡の分離などは行っていないため、リフタ長の条件は本手法のみで有効である。なお、PSS 法については予備実験によって低域に話者 B のスペクトルを、高域に話者 A のスペクトルを用いて合成した方が良い結果が出たため、以下の実験でも同様の方法で音声を合成する。

#### 3.1.2 実験結果

モーフィング音声のスペクトル包絡を本手法と PSS 法の場合についてそれぞれ図 3、図 4 に示す。これらの図に示すのはモーフィング率を 50% としたときの /arayuru/ の /a/ の部分の単位波形におけるスペクトル

表 1 モーフィング処理における分析条件  
Table 1 Analysis conditions for experiments.

| 標準化周波数    | 8 kHz              |
|-----------|--------------------|
| 分析窓       | 非対称ハニング窓           |
| 分析窓長 (可変) | ~32 ms (256 point) |
| FFT ポイント数 | 256 point          |
| リフタ長      | 1.25 ms (10 point) |



ル包絡であり、リフタ長 2.5 ms (20 ポイント) で平滑化してある。

図 4 より、PSS 法を用いた場合は、低域が話者 B のスペクトル、高域が話者 A のスペクトルとなっており、スペクトルに不連続が生じていることがわかる。それに比べ本手法を用いた場合では、図 3 よりスペクトル包絡の山をなすホルマント周波数がある程度移動するような形でモーフィングが行われている。

一般に、人間の声質はホルマント周波数の移動によって変化と言われているが、本手法を用いた場合のスペクトル包絡の変化の様子はこれをよく近似している。このことは、本手法が PSS 法と比較してより自然なモーフィング音声を生成可能であることを示唆している。

### 3.2 スペクトルひずみによる客観評価

本手法によるモーフィングがどれだけ自然音声の変化に近いかを判断するため、音声の調音結合の部分とモーフィング音声とのスペクトルひずみを測定した。

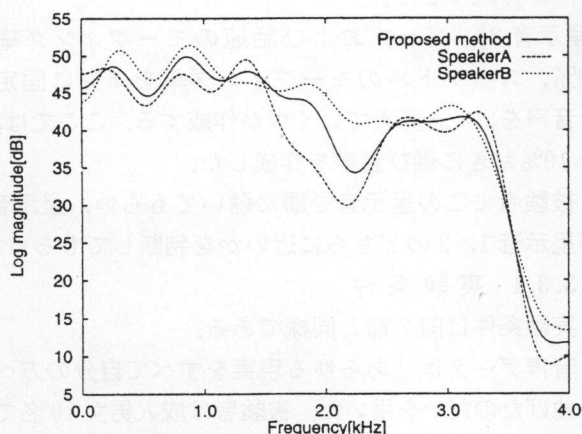


図 3 本手法によるスペクトル包絡  
Fig. 3 Spectral envelope of the proposed method.

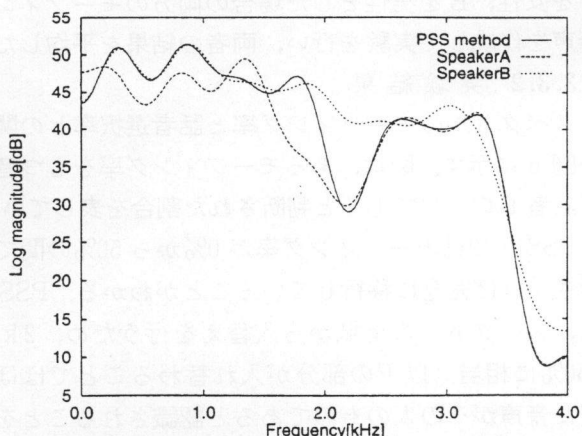


図 4 PSS 法によるスペクトル包絡  
Fig. 4 Spectral envelope of the PSS method.

以下にその方法を述べる。

- (1) 音声の調音結合の部分を選択により切り出す。
- (2) 切り出した部分の最初の単位波形と最後の単位波形を切り出す。
- (3) 最初の単位波形から最後の単位波形へのモーフィングを行う。ここではモーフィング率を 10% おきに連続的に変化させている。
- (4) モーフィング音声と調音結合部分との時間的整合をとるため DP マッチングにより対応付けを行う。
- (5) 対応した単位波形間のスペクトルひずみを測定する。ここでは次式で定義される LPC ケプストラム距離尺度によるスペクトルひずみを用いた [9]。但し、 $c_o(n)$  は自然音声の、 $c_m(n)$  はモーフィング音声の LPC ケプストラムを示す。

$$CD = (10 / \ln 10) \sqrt{2 \sum_{n=1}^p (c_o(n) - c_m(n))^2} \quad (16)$$

#### 3.2.1 実験条件

モーフィング処理における実験条件は前節と同様である。また、スペクトルひずみ測定における分析条件は表 2 に示すとおりである。

音声データとして男性話者の「あらゆる現実をすべて自分の方へねじ曲げたのだ」における /geNjitsu/ の /u/ から /o/ へと変化する部分を用いた。

#### 3.2.2 実験結果

測定したスペクトルひずみを表 3 に示す。また、図 5 に (a) 自然音声、(b) 本手法によるモーフィング音声、(c) PSS 法によるモーフィング音声のスペクトログラムを示す。

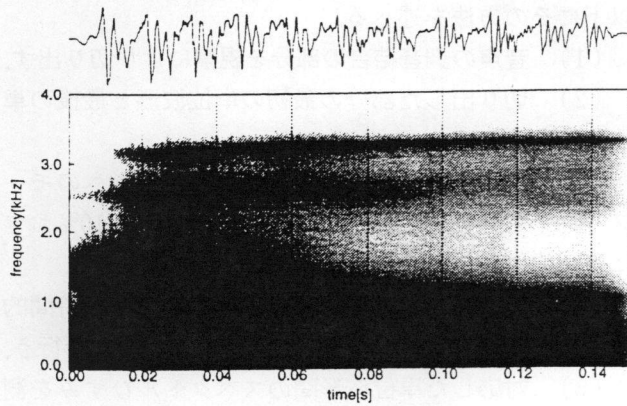
表 3 から本手法により生成されたモーフィング音声で、よりスペクトルひずみが小さいことがわかる。これは本手法によるスペクトル包絡の変化が調音結合を

表 2 スペクトルひずみ測定における分析条件  
Table 2 Analysis conditions for measuring spectral distortion.

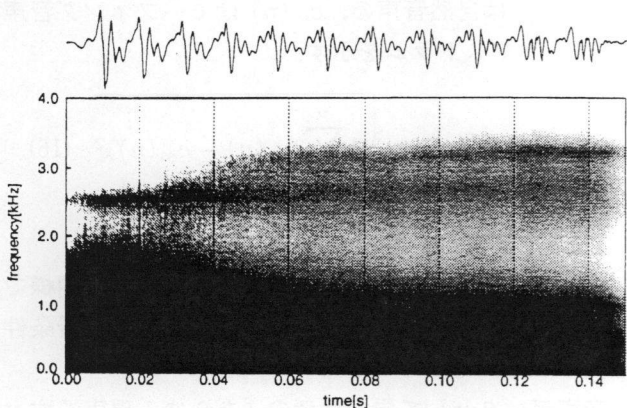
|           |                    |
|-----------|--------------------|
| 標本化周波数    | 8 kHz              |
| 分析窓       | 非対称ハニング窓           |
| 分析窓長 (可変) | ~32 ms (256 point) |
| 線形予測次数    | 12 次               |
| ケプストラム次数  | 16 次               |

表 3 スペクトルひずみ  
Table 3 Spectral distortion.

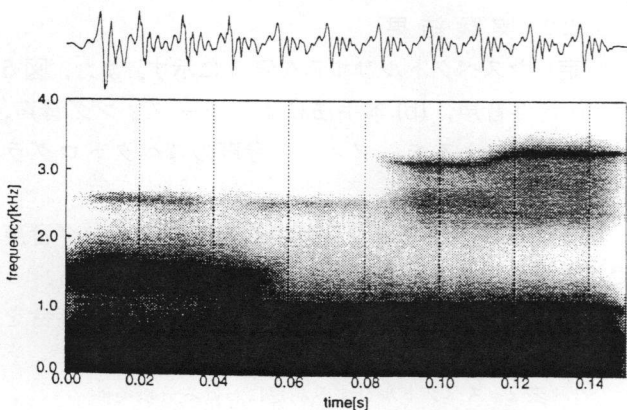
|       |        |
|-------|--------|
| 本手法   | 2.7 dB |
| PSS 法 | 4.6 dB |



(a) 自然音声の音素間遷移  
(a) Co-articulation in natural utterance.



(b) 本手法によるモーフィング音声  
(b) Morphing speech of the proposed method.



(c) PSS 法によるモーフィング音声  
(c) Morphing speech of the PSS method.

図5 スペクトログラム  
Fig.5 Spectrograms.

よく近似していることを示しており、本手法によるスペクトル包絡のモーフィングがより“自然”であると言える。また、図5からも本手法によるスペクトル包絡の変化が調音結合による自然音声の変化により類似していることがわかる。

### 3.3 スペクトルのモーフィング率と話者性の関係

ここでは、モーフィング率をどのように選べば最も滑らかに話者性を変化させることができるかを調査するため、モーフィング率と話者性の関係を本手法とPSS法の場合について示す。但し、ピッチおよび話速は固定し、スペクトルのモーフィング率のみによる話者性との関係を調べる。

呈示音声は以下の手順で作成した。今、モーフィングは話者Aから話者Bへと行うとする。

[呈示音1] ピッチおよび話速のモーフィング率を50%、スペクトルのモーフィング率を0%に固定し、モーフィングを行う。この処理により、スペクトルは話者Aのもので、スペクトル以外は中間的である音声を作成できる。

[呈示音2] ピッチおよび話速のモーフィング率を50%、スペクトルのモーフィング率を100%に固定し、モーフィングを行う。この処理により、スペクトルは話者Bのもので、スペクトル以外は中間的である音声を作成できる。

[呈示音3] ピッチおよび話速のモーフィング率を50%、スペクトルのモーフィング率を $r$ %に固定した音声を、 $r$ を変えていくつか作成する。ここでは、 $r$ を10%おきに選び音声を作成した。

被験者にこの呈示音を順に聴いてもらい、呈示音3が呈示音1, 2のどちらに近いかを判断してもらった。

#### 3.3.1 実験条件

実験条件は前2節と同様である。

音声データは「あらゆる現実をすべて自分の方へねじ曲げたのだ」を用いた。被験者は成人男女10名である。なお、PSS法では話者A, Bの選び方によって合成音は異なるため、Aを男性、Bを女性とした場合と、Aを女性、Bを男性とした場合の両方のモーフィング音声を作成して実験を行い、両者の結果を平均した。

#### 3.3.2 実験結果

スペクトルのモーフィング率と話者選択率との関係を図6に示す。図は、あるモーフィング率をもつ音声の話者Bのものであると判断された割合を表している。

PSS法ではモーフィング率が0%から50%の間で話者性がほぼ完全に移行していることがわかる。PSSでは、スペクトルの低域から入替えを行うため、2kHz(50%に相当)以下の部分が入れ替わることによって完全に音声がその人のものであると認識されることを示している。それに対し本手法では、話者選択率がモーフィング率の変化に伴い滑らかに変化していることが



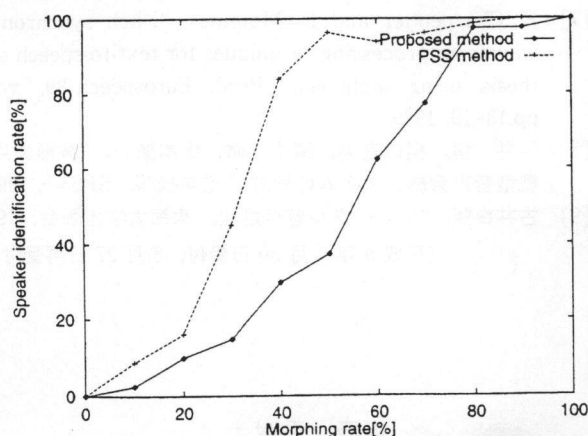


図6 モーフィング率と話者選択率との関係  
Fig.6 Relation between the morphing rate and the speaker identification rate.

わかる。

以上の結果を用いて、スペクトルのモーフィング率を、話者選択率が線形に移動するように選べば、滑らかなモーフィングの実現が可能となる。次節では、この結果を用いて実際にモーフィングを行う際のモーフィング率を決定し、評価実験を行う。

### 3.4 対比較試験による PSS 法との比較

本手法によるモーフィング音声を PSS 法によるものと比較する。ここでは対比較試験による主観評価でこれらの品質を評価する。なお、スペクトルのモーフィング率については前節の実験の結果を用いて、次のように決定する。

#### 3.4.1 スペクトルのモーフィング率の決定方法

まず、モーフィング率と話者選択率との関係の曲線を近似する。モーフィング率を  $r$ 、話者選択率を  $r_i$  とすると、本手法については、

$$r_i = 1 / (1 + \exp(-9(r - 0.5))) \quad (0 < r < 1) \quad (17)$$

を用いて近似し (図 7)、PSS 法については、

$$r_i = 1 / (1 + \exp(-15(r - 0.3))) \quad (0 < r < 1) \quad (18)$$

を用いて近似した (図 8)。

実際にモーフィングを行う際のモーフィング率は、話者選択率を線形に変化させたときの近似曲線の逆関数により求める。すなわち、本手法については、

$$r = -(1/9) \ln(1/r_i - 1) + 0.5 \quad (0 < r_i < 1) \quad (19)$$

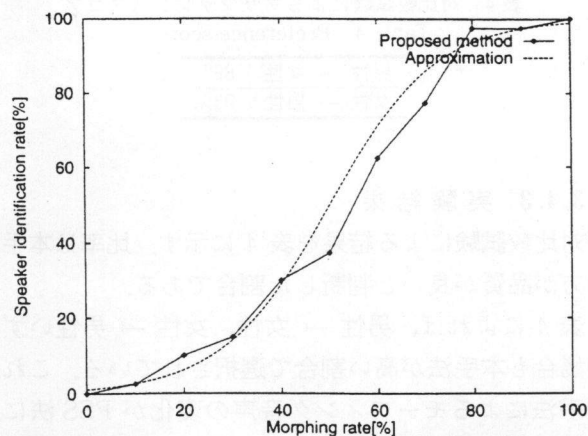


図7 本手法におけるモーフィング率と話者選択率との関係の近似

Fig.7 Approximation of the relation between the morphing rate and the speaker identification rate of the proposed method.

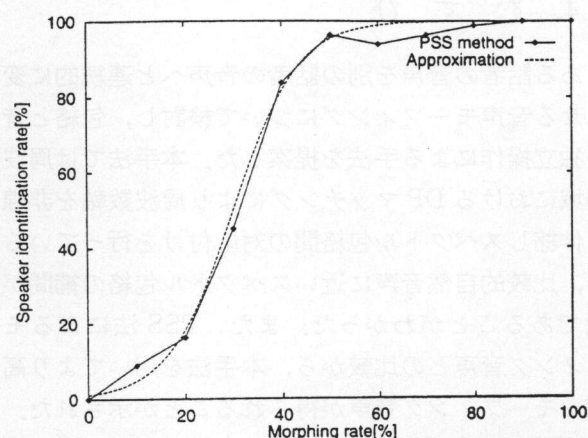


図8 PSS 法におけるモーフィング率と話者選択率との関係の近似

Fig.8 Approximation of the relation between the morphing rate and the speaker identification rate of the PSS method.

を用い、PSS 法については、

$$r = -(1/15) \ln(1/r_i - 1) + 0.3 \quad (0 < r_i < 1) \quad (20)$$

により求める。

なお、スペクトル以外の、ピッチおよび話速のモーフィング率は線形に変化させた。

#### 3.4.2 実験条件

実験条件は前 3 節と同様である。

音声データは「あらゆる現実をすべて自分の方へねじ曲げたのだ」を用い、男性から女性および女性から男性の 2 種類のモーフィング音声について対比較試験を行った。被験者は成人男女 10 名である。

表 4 対比較試験によるプリファレンススコア  
Table 4 Preference score.

|         |     |
|---------|-----|
| 男性 → 女性 | 89% |
| 女性 → 男性 | 93% |

### 3.4.3 実験結果

対比較試験による結果を表 4 に示す。比率は本手法の方が品質が良いと判断した割合である。

表 4 によれば、男性 → 女性、女性 → 男性いずれの場合も本手法が高い割合で選択されている。これは本手法によるモーフィング音声の変化が PSS 法に比べ聴覚的により自然であること、PSS 法では低域と高域がそれぞれ異なった話者のスペクトルをもっているため二人の話者の音声のように聞こえてしまうことによると考えられる。

## 4. む す び

ある話者の音声を別の話者の音声へと連続的に変化させる音声モーフィングについて検討し、包絡と音源の独立操作による手法を提案した。本手法では周波数領域における DP マッチングにより周波数軸を非線形に伸縮しスペクトル包絡間の対応付けを行っているため、比較的自然音声に近いスペクトル包絡の補間が可能であることがわかった。また、PSS 法によるモーフィング音声との比較から、本手法を用いてより高品質なモーフィング音声を得られることが示された。

謝辞 PSS法に関して貴重なコメントを頂いた NTT ヒューマンインタフェース研究所の阿部匡伸博士に感謝致します。また、音声データを提供して下さいった KDD 研究所の河井恒博士に感謝申し上げます。

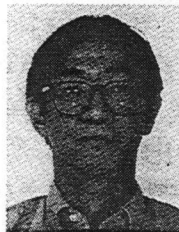
## 文 献

- [1] E. Tellman, L. Haken, and B. Holloway, "Timber morphing using the lemur representation," ICMC94 Proceeding, pp.329-330, 1994.
- [2] E. Tellman, L. Haken, and B. Holloway, "Timber morphing of sound with unequal numbers of features," Journal of the AES, vol.43, no.9, pp.678-689, 1995.
- [3] M. Slaney, M. Covell, and B. Lassiter, "Automatic Audio Morphing," Proceeding of 1996 ICASSP, vol.2, pp.1001-1004, 1996.
- [4] 小坂直敏, "Sinusoidal model を用いた母音の声質補間," 日本音響学会講演論文集, 2-1-10, pp.263-264, 1995.
- [5] 阿部匡伸, "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.
- [6] 河井 恒, 山本誠一, "基本周波数および音素持続時間を考慮した音声合成用波形素片データセットの作成," 信学技報, SP95-7, 1995.
- [7] F. Charpentier and E. Moulines, "Pitch synchronous waveform processing techniques for text-to-speech synthesis using diphones," Proc. Eurospeech '89, vol.2, pp.13-19, 1989.
- [8] 河井 恒, 樋口宜男, 清水 徹, 山本誠一, "波形素片接続型音声合成システムの検討," 信学技報, SP93-9, 1993.
- [9] 古井貞熙, "デジタル音声処理," 東海大学出版会, 1985.  
(平成 9 年 5 月 30 日受付, 8 月 27 日再受付)



坂野 秀樹

1996 名大・工・電子情報卒。同年、奈良先端大博士前期課程入学。音声合成に関する研究に従事。日本音響学会会員。



武田 一哉 (正員)

会員。

1983 名大・工・電気卒。1985 同大大学院博士(前期)課程了。同年国際電信電話株式会社入社, ATR 自動翻訳電話研究所, KDD 研究所において音声合成・認識システムの研究を行う。1994 名大・工・助教授。工博。IEEE, 音響学会, 情報処理学会各



鹿野 清宏 (正員)

1970 名大・工・電気卒。1972 同大大学院修士課程了。同年電電公社(現 NTT) 武蔵野通研入所。1984~1986 カーネギーメロン大客員研究員。1986~1990 ATR 自動翻訳電話研究所音声情報処理研究室長。1992 NTT ヒューマンインタフェース研究所主席研究員。1994 奈良先端大・教授。工博。音声・音情報処理の研究および研究指導に従事。1975 本会米沢賞。1991 IEEE SP 1990 Senior Award, 1994 日本音響学会技術開発賞。IEEE, 音響学会, 情報処理学会各会員。



板倉 文忠 (正員)

1963 名大・工・電子卒。1968 同大大学院博士課程了。同年電電公社(現 NTT) 武蔵野通研入所。音声処理の研究に従事。工博。1973~1975 ベル研究所にて音声認識・音声分析の研究を行う。1984 名大・工・教授。昭 45, 53, 56 年度論文賞, 昭 47, 56 年度業績賞受賞。IEEE, 日本音響学会各会員。