音声モーフィングにおける基準点付与の自動化

川本 真一 *　　鶴見 智 *　　滝澤 照太 **

(2019 年 1 月 7 日受理)

1　はじめに

音声は重要なコミュニケーション手段の 1 つであり，カーナビゲーションシステムやスマートフォンなどの工業製品においてインタフェースとして採用されている．音声が与える印象はその製品の印象に繋がるため，想定する使用者層，使用環境などに適した音声が要求される．

音声モーフィングは複数の話者の音声を混合する音声加工技術であり，中間的な特徴を持つ音声を生成することができる．モーフィング手法として，メル周波数ケプストラム係数 [1] や STRAIGHT 分析 [2] など特徴量の変形に基づくものが提案されている．また，統計モデル [3]，深層学習 [4] などを用いて特定の声の特徴への写像を学習するアプローチが提案されている．しかし，特徴量を変形する手法では基準点付与などの前処理を手作業で行う必要があり，膨大な手間や時間を要する．また，統計モデルや深層学習を使用する手法は，学習のために音声データを相当量用意する必要がある．

本研究では，特徴量の変形に基づくモーフィング手法の課題である，基準点を自動的に付与する手法を提案する．また，話者性を強く反映していると考えられている母音 [5] に焦点を当て，提案手法により付与した基準点を用いて音声モーフィングを実現し，動作を検証する．

2　基準点付与の自動化

2.1　概要

縦軸に周波数，横軸に時間を配置して，音声の周波数成分の時間変化を可視化した図をスペクトログラムと呼ぶ．スペクトログラムを時間方向に見たときに変化が大きな点を基準点とし，これを同一内容の発話を収録した複数の音声間で対応付けることで，時間領域でモーフィングを行うことができる．また，スペクトルは音素や個人性に関する情報を含んでいる．スペクトルの特徴的な点を基準点とし，これを複数の音声間で対応付けることで，周波数領域でモーフィングを行うことができる．

本研究では，音声分析合成系 WORLD [6,7] を用いて，音声の基本周波数，スペクトル包絡，非周期性指標を分析し，スペクトル包絡に注目して基準点の決定を行う．入力する音声はサンプリング周波数 16kHz，量子化ビット数 16bit のモノラル音声を想定し，WORLD による音声処理はフレーム周期 80 点，フレーム長 (FFT サイズ) 1024 点を用いる．その他の WORLD に与えるパラメータについては，WORLD に付属するサンプルにて提供される標準的な値を使用する．本論文では，音声の基本的な処理単位としてフレームを使用する．

2.2　時間領域

スペクトログラムの時間変化を表現する手法として，調音結合の解析に用いられる Temporal Decomposition (TD) [8] と呼ばれる手法がある．TD は，音声中の特徴的なフレームにイベントを設定し，そのフレームのスペクトルの重み付き線形和でスペクトログラムをモデル化する手法である．イベントがある時刻のスペクトルをイベントベクトルと呼ぶ．スペクトルを Line Spectral Frequency (LSF) [10] と呼ばれるパラメータで表現すると，スペクトログラムは時間に関するベクトル関数 $\mathbf{y}(n)$ とみなすことができる．式 (1) は，イベントベクトルの本数を表す次数が m のスペクトログラム $\mathbf{y}(n)$ の TD モデル $\hat{\mathbf{y}}(\mathbf{n})$ である．$\mathbf{a}_k$ は k 番目のイベントベクトル，$\phi_k(n)$ は n フレームにおける k 番目のイベントベクトルの重みを表現するイベント関数である．

$$\hat{\mathbf{y}}(n) = \sum_{k=1}^{m} \mathbf{a}_k \phi_k(n), \quad 1 \leq n \leq N \tag{1}$$

本研究では，TD を改良した Modified Restricted Temporal Decomposition (MRTD) [9] を使用し，WORLD で分析したスペクトル包絡をモデル化する．イベントが設定されているフレームを時間領域の基準点とする．図 1 に時間領域の基準点決定の流れを示す．MRTD モデルの作成には，Nguyen らが提案している手法 [9] を使用している．SFTR はフレームに関する関数であり，n 番目のフレームの SFTR は式 (2) で表すことができる．N は音声の全フレーム数，P は LSF の次数，$\mathbf{y}(\mathbf{k})$ は k フレーム目の LSF，M は MRTD

* 電子情報工学科
** 元 専攻科生産システム工学専攻

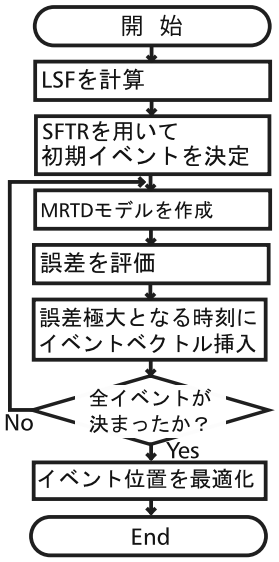

図 1　MRTD による時間領域基準点決定のフローチャート

の窓幅であり，今回は $M = 8$ を用いた．

$$s(n) = \sum_{i=1}^{P} {c_i}^2(n), 1 \leq n \leq N \tag{2}$$

$$c_i(n) = \frac{\sum_{m=-M}^{M} m\mathbf{y}_i(n+m)}{\sum_{m=-M}^{M} m^2}, 1 \leq i \leq P \tag{3}$$

また，誤差の評価には，式 (4) によって計算できる Mean Squared Error (MSE) を用いる．$y_i, \hat{y}_i$ はベクトル $\mathbf{y}, \hat{\mathbf{y}}$ の i 番目の要素を抜き出したものである．

$$E = \sum_{n=1}^{N} \sum_{i=1}^{P} (\hat{y}_i(n) - y_i(n))^2 \tag{4}$$

音素ごとにこの処理を適用し，音声の全区間に対して基準点を決定する．

本研究ではさらに，通常の MRTD のイベントベクトル決定処理後に，イベント時刻の最適化処理を追加している．イベントベクトルの設定が完了後，各フレームの LSF と MRTD モデルの間の MSE が最小になる位置にイベントを移動させる．

2.3　周波数領域

音声の音韻性，個人性を表現する重要な要素は，フォルマントと呼ばれるスペクトル包絡のピークである．LSF は周波数領域のパラメータで，スペクトル包絡においてパワーが集中する周波数付近に多くのパラメータが配置される特性を持つ．そこで本研究では LSF を周波数領域の基準点に用いる．

さらに，LSF はスペクトル包絡のピークを挟み込むように配置される特性を持つため，パラメータ間で最大または最小となる周波数を探索することでピークやノッチを得ることができる．そこで，隣接する LSF 区間につき 1 つ，ピーク (ピークが見つからない場合はノッチ) を探索し，ピーク (もしくはノッチ) に対応する周波数も基準点として追加する．

なお，処理を容易にするため，実際には LSF と隣接 LSF 間のピーク (もしくはノッチ) 周波数を低域側から並べたベクトルを周波数領域の基準点として使用する．

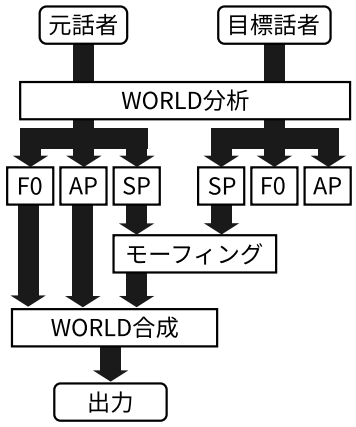

図 2　評価のためのモーフィングにおけるデータの流れ

3　二話者間のモーフィング

決定した基準点を用いて音声モーフィングを実施し，基準点の妥当性を確認した．モーフィングにおけるデータの流れを図 2 に示す．図中の F0 は基本周波数，SP はスペクトログラム，AP は非周期性指標を表す．今回はスペクトログラムのみを混合し，F0 と AP は元話者のものをそのまま使用する．音声の長さも元話者のものを基準とし，目標話者の音声を線形に伸縮する．

式 (5) に時間領域基準点のモーフィングを示す．今回は元話者の音声を基準とするため，元話者の基準点をそのまま使用する．$r_s(k)$，$r_t(k)$ は元話者と目標話者の k 番目の時間領域基準点となるフレーム，$r'(k)$ はモーフィング後の音声の k 番目の時間領域基準点となるフレームである．

元話者と目標話者の音声は長さが異なるため，時間領域基準点を対応させるように基準点間のフレームをマッピングして対応する必要がある．時間領域基準点上では 2 つの音声が対応しているため，式 (6) のように周波数領域のモーフィングができる．$\mathbf{a}^s_{r_s(k)}$，$\mathbf{a}^t_{r_t(k)}$ は元話者と目標話者の k 番目の周波数領域基準点のベクトルであり，$\mathbf{a}'_{r'(k)}$ はモーフィング後の音声の k 番目の周波数領域基準点のベクトルである．R は混合割合であり，目標話者 : 元話者 $= R : (1 - R)$ の割合で混合を行う．

$$r'(k) = 1 \times r_s(k) + 0 \times r_t(k) \tag{5}$$

$$\mathbf{a}'_{r'(k)} = (1 - R) \times \mathbf{a}^s_{r_s(k)} + R \times \mathbf{a}^t_{r_t(k)} \tag{6}$$

$r'(k)$ から $r'(k+1)$ までの基準点間では，区間中の n 番目のフレームに対して式 (7) のように，目標話者のフレームをマッピングしてモーフィングを適用する．$\mathbf{a}^s_n$ は $r'(k)$ から n 番目のモーフィング後のスペクトル包絡，$\mathbf{a}^s_n$ は $r'(k)$ から

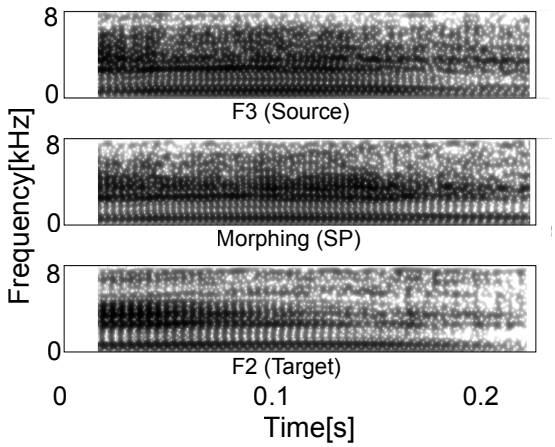

図 3　元話者-モーフィング音声-目標話者のスペクトログラム

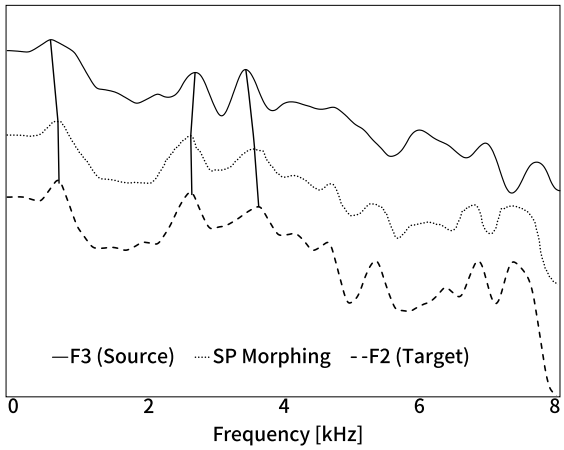

図 4　モーフィング音声のフォルマント位置確認

n 番目の元話者のスペクトル包絡，$\mathbf{a}^t_m$ はマッピングにより対応づけた位置の目標話者のスペクトル包絡である．

$$\mathbf{a'}_n = (1-R) \times \mathbf{a}^s_n + R \times \mathbf{a}^t_m \tag{7}$$

$$m = \frac{(r'(k+1) - r'(k))}{(r_t(k+1) - r_t(k))} \times n \tag{8}$$

但し，$1 \leq k \leq N-1$，$1 \leq n \leq r'(k+1) - r'(k+1)$ である．

北陸先端科学技術大学院大学研究用日本語感情音声データベース (JAIST-ESD) の女性話者 F2，F3 の 2 名が日本語母音/e/を発音した音声に対して，F3 を元話者，F2 を目標話者として，混合割合 $R = 0.5$ でモーフィングを行った．MRTD の次数は 5，LSF の次数は 36 とした．

図 3 に元話者 (Source) とスペクトログラムモーフィング (Morphing SP)，目標話者 (Target) の音声のスペクトログラムを示す．濃淡の濃い部分は強い周波数成分を表す．音声冒頭部分に注目すると，元話者 F3 よりも目標話者 F2 のほうが低域に集中している．モーフィング音声は，この中間的な値をとっている．

図 4 は元話者と目標話者，モーフィング音声の対応するフレームにおけるスペクトル包絡の第 3 フォルマント (低域から 3 つ目までのピーク) を線で結んだものである．式 (7) によるマッピングを考慮し，元話者とモーフィング音声は 31 フレーム目，目標話者は 11 フレーム目を抽出している．モーフィング音声の低域部分において，スペクトル包絡のピークは元話者と目標話者の中間的な位置に出現していることがわかる．

以上の点から，モーフィング音声のスペクトログラムは 2 話者の中間的なものとなっていることがわかる．

4　評価実験

4.1　聴取実験の準備

2 種類のモーフィング手法の性能を比較するため，主観評価聴取実験を行った．

聴取実験に用いる音声素材は，音声データベース JAIST-ESD の男性話者 5 名 (M1〜M5) と女性話者 5 名 (F1〜F5) が単語 10 種類を発音する音声を，サンプリング周波数 16kHz，量子化ビット数 16bit にリサンプリングしたものを用いた．基準点付与は音素ごとに行い、音素境界は汎用大語彙連続音声認識エンジン Julius [11] を用いて音素セグメンテーション結果を用いた。これらの音声に対してモーフィングを施して聴取実験用音声を作成する．また，個人性に強く影響する母音 [5] のみをモーフィングの対象とし，子音部分は元話者のものをそのまま使用する．母音と子音の区別には Julius の出力を使用する．

比較対象として，LSF パラメータ空間上でモーフィングを行なう手法 (LSF モーフィング) [12] を取り上げる．本稿で報告するスペクトログラムモーフィングとの違いは，1)LSF パラメータ空間上でモーフィング (パラメータの混合) を行った後にスペクトル包絡に変換する点と，2) スペクトル包絡のピークを明示的に基準点として扱わない点である．

MRTD 次数および LSF 次数の設定については，予備実験において，良好な結果であったパラメータを用いた。LSF モーフィングでは MRTD 次数を 5，LSF 次数を 36 に設定し，スペクトログラムモーフィングでは MRTD 次数を 3，LSF 次数を 36 に設定したものを用いた．

4.2　音質評価

LSF モーフィング，スペクトログラムモーフィングの 2 手法において，どちらが音質面で有利であるかどうかを調べる主観評価聴取実験を実施した．聴取実験用音声から，同性話者，同一単語の音声を 2 つ抽出し，$R = 0.5$ として 2 つの手法でモーフィングをした音声をペアとする．被験者は 120 ペアの音声を聴取し，各ペア中での音質が良いと感じた音声を選択する．5 名の被験者の実験結果より得られた結果からプリファレンススコアを算出した．

図 5 に聴取実験の結果から得られたプリファレンススコアを示す．有意水準 1% で有意差が見られ，スペクトログラムモーフィングのほうが音質面で良いという結果を得た．

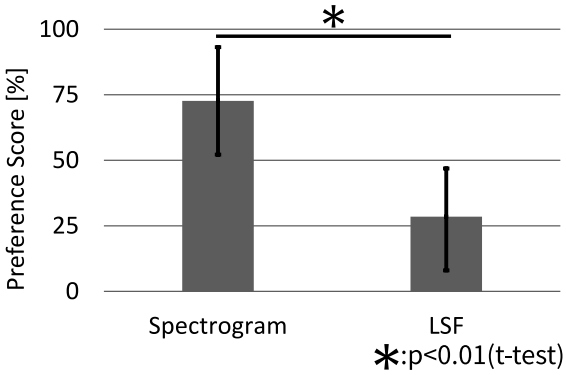

図 5　音質評価結果（エラーバー：99% 信頼区間）

5　おわりに

本研究では，モーフィングに用いる基準点の自動的に付与する手法，およびスペクトログラムモーフィングへの適用についてを報告した．二話者間のスペクトログラムモーフィングにおいて，MRTD と LSF を利用した基準点付与手法の動作の妥当性と，モーフィング音声の音質面から見た有効性を確認した．本手法を用いることにより，音声モーフィングを利用した音声加工を短時間で系統的に行うことが可能となる．

謝辞

本研究は JSPS 科研費 JP25240026, JP15K21024 の助成を受けたものです．

参考文献

1) M. Slaney, M. Covell, and B. Lassiter, “Automatic audio morphing,” Proc. ICASSP1996, VOL. 2, pp. 1001-1004, 1996.
2) H. Kawahara and H. Matsui, “Auditory morphing based on an elastic perceptual distance metric in an interference-free time-frequency representation,” Proc. ICASSP2003, VOL. 1, pp. 256-259, 2003.
3) T. Toda, A.W. Black and K. Tokura, “Voice conversion based on maximum likelihood estimation of spectral parameter trajectory,” Proc. IEEE TASLP, VOL. 15, Issue 8, pp. 2222-2235, 2007.
4) S.H. Mohammadi and A. Kain, “Voice conversion using deep neural networks with speaker-independent pre-training,” Proc. IEEE SLT, pp. 19-23, 2014.
5) T. Kitamura and P. Mokhtari, “Effects of vowel types on perception of speaker characteristics of unknown speakers,” Proc. NCSP2006, pp. 45-48, 2006.
6) M. Morise, F. Yokomori, and K. Ozawa, “WORLD: a vocoder-based high-quality speech synthesis system for real-time applications,” IEICE TRANS. INF. & SYST., VOL. E99-D, NO. 7, pp. 1877-1884, 2016.
7) M. Morise, “D4C, a band-aperiodicity estimator for high-quality speech synthesis,” Speech Communication, VOL. 84, pp. 57-65, 2016.
8) B.S. Atal, “Efficient coding of LPC parameters by temporal decomposition,” Proc. ICASSP1983, pp. 81-84, 1983.
9) P.C Nguyen, T. Ochi, M. Akagi, “Modified restricted temporal decomposition and its application to low rate speech coding,” IEICE TRANS. INF. & SYST., VOL. E86-D, NO. 3, pp. 397-404, 2003.
10) F. Itakura, “Line spectrum representation of linear predictor coefficients of speech signals,” J. Acoust. Soc. Amer., VOL. 57, S35(A), 1975.
11) A. Lee, T. Kawahara, and K. Shikano, “Free software toolkit for Japanese large vocabulary continuous speech recognition,” Proc. ICASSP2000, VOL. 4, pp. 476-479, 2000.
12) S. Takizawa, and S. Kawamoto, “Automatic reference point assignment technique for voice morphing,” Proc. GCCE2017, pp. 1-3, 2017.

Automatic Reference Point Placement Technique for Voice Morphing

Shinichi KAWAMOTO, Satoshi TSURUMI, Shota TAKIZAWA

Automatic reference point placement method for voice morphing is reported in this paper. Voice morphing is one of fundamental voice editing methods to blend feature vector sequences of two voices based on corresponding reference points. Reference points are basically assigned by hands, and depends on the quality of voice morphing output. Moreover, assigning reference points is a time-consuming task. The proposed method realizes to assign reference points on spectrogram in time- and frequency-domain automatically based on temporal decomposition (TD) and line spectral frequency (LSF). As results of two-speakers’ voice morphing, the proposed method was worked well by using voice and its transcription as inputs.