統計的手法に基づくリアルタイム声質変換による音声生成機能拡張*

○戸田智基 (奈良先端大・情報)

1　はじめに

音声コミュニケーションでは，所望の言語情報を音声信号へと変換し，パラ言語情報および非言語情報と合わせて，同時に相手に伝達することができる．音声信号には多くの情報が埋め込まれるものの，その生成にかかる時間は短く，即時性は極めて高い．また，音声生成過程における物理的制約は，個人が生成できる声色の範囲を限定し，音声信号に個人性を与える一つの要因となる．一方で，この制約の強さ故に，時として，コミュニケーションにおける障壁が容易にもたらされる．例えば，音源生成器官や調音器官が正常に動作しなくなると，深刻な発声障害を患い，音声コミュニケーションに支障をきたす．仮に，物理的制約を超えた音声生成を可能とする機能が実現されれば，このような障壁を無くすことができる．さらには，各個人が意図的に制御できる声質の範囲が広がり，より多様な歌唱表現や発声表現も生まれると予想される．

声質変換は，言語情報を保持しながら，所望の非言語情報やパラ言語情報を変換する音声情報処理技術である．声質変換を実現する枠組みは様々であるが，中でも，入力された音声信号に対して変形処理を行う枠組みは，古くから研究されている．統計的手法により変形処理を求める枠組み [1] が主流であり，確率モデルを用いた手法 [2] が目覚ましい発展を遂げている．近年では，音声信号の時系列データとしての特徴を最大限に活用する変換法が提案され [3]，さらには，リアルタイム変換処理への拡張もなされている [4]．音声コミュニケーションにおいて本質的に重要な特徴である即時性を満たすことができるため，音声生成機能拡張を実現する可能性を大いに秘めた技術といえる．

本稿では，各種声質変換技術を概説した後で，統計的手法に基づくリアルタイム声質変換技術について詳しく述べる．また，リアルタイム声質変換による音声生成機能拡張の応用例について紹介し，今後の課題を述べる．

2　声質変換の枠組み

声質変換の出力は，所望の非言語情報もしくはパラ言語情報を持つ音声信号である．一方で，入力として用いる情報については，様々なものが考えられる．以下では，大きく三つの枠組みに分類する．

2.1　音声入力

入力された音声信号に対して変換処理を施す枠組みであり，声質変換は，この枠組みを指す用語として用いられることも多い．入力音声と出力音声の同一内容発話セットを学習データとして用いて，変換関数を求める統計的手法 [1] が主流であり，高度な変換処理が実現可能である．混合正規分布モデル（Gaussian mixture model: GMM）[2] やニューラルネットワーク [5] を初めとして，近年では，非負値行列分解に基づく手法 [6] や深い構造を持つネットワークに基づく手法 [7] など，様々な変換法が提案されている．

テキスト情報を入力として必要としないため，リアルタイム処理に適しているというのが最大の特徴である．また，言語依存性が低いという利点もある．なお，リアルタイム処理およびテキスト情報が未知という条件において，韻律的特徴量を高精度に変換するのは困難であるため，短時間特徴量である分節的特徴量を主な変換対象とする研究が多い．

2.2　テキスト入力

テキストを入力として音声を出力するテキスト音声合成（Text-to-Speech: TTS）において，出力音声の声質変換を行う処理は古くから研究されている．統計的パラメトリック音声合成 [8] の代表的手法である隠れマルコフモデル（hidden Markov model: HMM）に基づく音声合成法 [9] では，HMM のパラメータに対して変形処理を施すことで，出力音声の声質を変換することができる．最尤線形回帰 [10] や最大事後確率推定 [11]，平均声 [12]，固有声 [13, 14]，重回帰モデル [15] など，多種多様なモデル適応技術が適用されており，近年劇的な発展を遂げている技術である．

テキスト情報を入力とするため，分節的特徴量のみでなく韻律的特徴量も対象とした声質変換処理を容易に実現できる．また，音声合成モデルは大量のデータを用いて事前に学習されるため，汎化処理により，音声分析誤差が合成音声に与える影響を低減できる．一方で，音声入力と比べると即時性に乏しい．

2.3　音声・テキスト入力

音声とテキストの両方を入力として，所望の出力音声を得る枠組みも研究されている．その一例として，TTS で出力される合成音声に対する後処理として，声質変換を行う枠組みが挙げられる．波形接続方式などに基づく TTS に対しても声質変換を容易に適用できるなど，移植性は高い．また，テキスト情報を用いて変換関数を記述することができるため，韻律的特徴量の変換にも対応しやすい [16–18]．

TTS における合成処理の時点で入力音声を活用することで，所望の合成音声を実現する枠組みもある．例えば，歌声合成における VocaListener[19] に代表されるように，入力音声の韻律的特徴を持つ合成音声が得られるように合成処理を最適化することで，合成音声を入力音声で調整する機能を実現できる．HMM を用いた声質変換法 [20] も，類似のものとみなせる．

---

*Augmented speech production based on real-time statistical voice conversion. by TODA, Tomoki (Nara Institute of Science and Technology)

3　統計的手法に基づくリアルタイム変換

声質変換による音声生成機能拡張では，リアルタイム変換処理が必要不可欠となるため，音声入力に基づく声質変換の枠組みが有効である．一例として，GMM を用いた系列単位の声質変換法 [3] に基づくリアルタイム変換処理 [4] について述べる．

3.1　特徴量抽出処理

入力音声の各フレームにおいて，簡易なスペクトル分析を行い，スペクトル包絡パラメータベクトル $\boldsymbol{x}_t$ を抽出する[1]．ここで，t はフレーム番号を表す．フレーム t におけるスペクトルセグメント特徴量として，当該フレームおよび前後 C フレームのスペクトル包絡パラメータベクトルを用いて，$D^{(x)}$ 次元入力特徴量ベクトル $\boldsymbol{X}_t$ を次式にて計算する．

$$\boldsymbol{X}_t \quad = \quad \boldsymbol{A}\left[\boldsymbol{x}_{t-C}^\top, \cdots, \boldsymbol{x}_t^\top, \cdots, \boldsymbol{x}_{t+C}^\top\right]^\top + \boldsymbol{b} \quad (1)$$

ここで，$^\top$ は転置を示す．行列 $\boldsymbol{A}$ およびベクトル $\boldsymbol{b}$ はセグメント特徴量抽出のためのパラメータを示す[2]．

出力特徴量として，結合静的・動的特徴量ベクトル $\boldsymbol{Y}_t = \left[\boldsymbol{y}_t^\top, \Delta\boldsymbol{y}_t^\top\right]^\top$ を用いる．ここで，$\boldsymbol{y}_t$ はフレーム t における $D^{(y)}$ 次元の音声パラメータベクトルとし，動的特徴量ベクトル $\Delta\boldsymbol{y}_t$ は次式にて計算する．

$$\Delta\boldsymbol{y}_t \quad = \quad \boldsymbol{y}_t - \boldsymbol{y}_{t-1} \quad (2)$$

各種応用例に応じて，スペクトル包絡パラメータや非周期成分パラメータ，基本周波数パラメータなどを，音声パラメータとして用いる．なお，出力音声に対するパラメータ抽出には，STRAIGHT[21] などの高精度な音声分析系を用いることが重要である．

3.2　学習処理

学習データとして，入力音声と出力音声の同一発話対で構成されるパラレルデータを用いる．パラレルデータに対して，フレーム間の対応付けを行うことで得られる結合入力・出力特徴量ベクトル $\left[\boldsymbol{X}_t^\top, \boldsymbol{Y}_t^\top\right]^\top$ を用いて，入力特徴量ベクトルと出力特徴量ベクトルの結合確率密度関数を，次式に示す GMM によりモデル化する [22].

$$P\left(\boldsymbol{X}_t, \boldsymbol{Y}_t | \boldsymbol{\lambda}^{(X,Y)}\right) = \sum_{m=1}^{M} \alpha_m \mathcal{N}\left(\begin{bmatrix}\boldsymbol{X}_t \\ \boldsymbol{Y}_t\end{bmatrix}; \begin{bmatrix}\boldsymbol{\mu}_m^{(X)} \\ \boldsymbol{\mu}_m^{(Y)}\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Sigma}_m^{(XX)} & \boldsymbol{\Sigma}_m^{(XY)} \\ \boldsymbol{\Sigma}_m^{(YX)} & \boldsymbol{\Sigma}_m^{(YY)}\end{bmatrix}\right) \quad (3)$$

ここで，$\mathcal{N}(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ は平均ベクトル $\boldsymbol{\mu}$ および共分散行列 $\boldsymbol{\Sigma}$ の正規分布を示す．また，m は分布番号，M は分布数を表す．GMM のパラメータセットは $\boldsymbol{\lambda}^{(X,Y)}$ で示され，個々の分布に対する分布重み α_m，入力および出力に対する平均ベクトル $\boldsymbol{\mu}_m^{(X)}$，$\boldsymbol{\mu}_m^{(Y)}$ と共分散行列 $\boldsymbol{\Sigma}_m^{(XX)}$，$\boldsymbol{\Sigma}_m^{(XY)}$，$\boldsymbol{\Sigma}_m^{(YX)}$，$\boldsymbol{\Sigma}_m^{(YY)}$ で構成される．

[1]例えば，固定分析窓を用いた高速フーリエ変換と 1 次の全域通過フィルタリングで得られるメルケプストラム係数を用いる．

[2]例えば，結合静的・動的特徴量ベクトルを抽出する回帰行列や，主成分分析により得られる固有ベクトルおよび平均ベクトルを用いて設定する．

3.3　変換処理

高い変換精度を得るには，フレーム間の相関や高次統計量など，時系列データとしての特徴を，適切にモデル化することが重要である．変換対象となる入力音声が，フレーム数 T の時系列データで表されるとし，入力特徴量系列ベクトルを $\boldsymbol{X} = \left[\boldsymbol{X}_1^\top, \cdots, \boldsymbol{X}_T^\top\right]^\top$，出力特徴量系列ベクトルを $\boldsymbol{Y} = \left[\boldsymbol{Y}_1^\top, \cdots, \boldsymbol{Y}_T^\top\right]^\top$ とする．最尤系列変換法 [3] では，変換静的特徴量系列ベクトル $\hat{\boldsymbol{y}} = \left[\hat{\boldsymbol{y}}_1^\top, \cdots, \hat{\boldsymbol{y}}_T^\top\right]^\top$ は次式で求められる．

$$\hat{\boldsymbol{y}} = \underset{\boldsymbol{y}}{\operatorname{argmax}} P\left(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\lambda}^{(X,Y)}\right) P\left(\boldsymbol{v}^{(\boldsymbol{y})}|\boldsymbol{\lambda}^{(v)}\right)^\omega \quad (4)$$

$$\text{subject to} \quad \boldsymbol{Y} = \boldsymbol{W}\boldsymbol{y} \quad (5)$$

ここで，$\boldsymbol{W}$ は静的特徴量系列ベクトルを結合静的・動的特徴量系列ベクトルへと拡張する $2D^{(y)}T \times D^{(y)}T$ の行列 [23] である．式 (3) より，$\boldsymbol{X}$ が与えられた際の $\boldsymbol{Y}$ の条件付き確率密度関数は次式でモデル化される．

$$P\left(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\lambda}^{(X,Y)}\right) = \prod_{t=1}^{T} P\left(\boldsymbol{Y}_t|\boldsymbol{X}_t, \boldsymbol{\lambda}^{(X,Y)}\right) \quad (6)$$

$$P\left(\boldsymbol{Y}_t|\boldsymbol{X}_t, \boldsymbol{\lambda}^{(X,Y)}\right) = \sum_{m=1}^{M} \alpha_{m,t}^{(Y|X)} \mathcal{N}\left(\boldsymbol{Y}_t; \boldsymbol{\mu}_{m,t}^{(Y|X)}, \boldsymbol{\Sigma}_m^{(Y|X)}\right) \quad (7)$$

ここで，式 (7) も GMM で表され，そのパラメータは

$$\alpha_{m,t}^{(Y|X)} = P\left(m|\boldsymbol{X}_t, \boldsymbol{\lambda}^{(X,Y)}\right) \quad (8)$$

$$\boldsymbol{\mu}_{m,t}^{(Y|X)} = \boldsymbol{\mu}_m^{(Y)} + \boldsymbol{\Sigma}_m^{(YX)}{\boldsymbol{\Sigma}_m^{(XX)}}^{-1}\left(\boldsymbol{X}_t - \boldsymbol{\mu}_m^{(X)}\right) \quad (9)$$

$$\boldsymbol{\Sigma}_m^{(Y|X)} = \boldsymbol{\Sigma}_m^{(YY)} - \boldsymbol{\Sigma}_m^{(YX)}{\boldsymbol{\Sigma}_m^{(XX)}}^{-1}\boldsymbol{\Sigma}_m^{(XY)} \quad (10)$$

で与えられる．また，式 (4) において，ベクトル $\boldsymbol{v}^{(\boldsymbol{y})} = \left[v_1^{(\boldsymbol{y})}, \cdots, v_{D^{(y)}}^{(\boldsymbol{y})}\right]^\top$ は系列内変動（global variance: GV）を表し，次式で計算される．

$$v_d^{(\boldsymbol{y})} \quad = \quad \frac{1}{T}\sum_{t=1}^{T}\left(y_{t,d} - \frac{1}{T}\sum_{\tau=1}^{T} y_{\tau,d}\right)^2 \quad (11)$$

ここで，$y_{t,d}$ はフレーム t における出力静的特徴量ベクトルの d 次元目の要素を表す．GV の確率密度関数は正規分布でモデル化され，パラメータセット $\boldsymbol{\lambda}^{(v)}$ は平均ベクトル $\boldsymbol{\mu}^{(v)}$ および共分散行列 $\boldsymbol{\Sigma}^{(v)}$ から成る．これらは，GMM と同様に学習データから事前に求める．定数 ω は GV 尤度重みを表す．式 (4) において，変換静的特徴量系列ベクトルは，系列単位のバッチ処理による反復演算により求められるが，リアルタイム変換処理では，さらに，再帰的解法 [23, 24] に基づく短遅延変換法 [25] と GV ポストフィルタ [4] を導入することで，フレーム処理による近似解を求める．

まず，$P\left(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\lambda}^{(X,Y)}\right)$ の最大化を行う．各フレームにおいて，式 (8) の事後確率に基づき，式 (7) の条件付き確率密度関数を単一分布 $\hat{m}_t$ で近似する．

$$\hat{m}_t \quad = \quad \arg\max_m \alpha_{m,t}^{(Y|X)} \quad (12)$$

また，共分散行列 $\mathbf{\Sigma}_m^{(Y|X)}$ の対角要素のみを用いて，次元毎に独立に変換処理を行う．この時，変換静的特徴量系列は T 個の連立一次方程式の解として与えられるが [26]，カルマンフィルタによる近似処理を導入することで，短遅延処理を実現する[3]．$(L+1)$ 次元の出力静的特徴量セグメントベクトル $\boldsymbol{y}_d = [y_{t-L,d}, \cdots, y_{t,d}]^\top$ に対する以下の状態空間モデルを用いる．

$$\boldsymbol{y}_d^{(t)} = \boldsymbol{J}_L \boldsymbol{y}_d^{(t-1)} + \left[\boldsymbol{0}_{1\times L}, \mu_{\hat{m}_t,t,d}^{(y|X)} + n_{\hat{m}_t,t,d}\right]^\top \quad (13)$$

$$\mu_{\hat{m}_t,t,d}^{(\Delta y|X)} = \boldsymbol{w}_L \boldsymbol{y}_d^{(t)} + e_{\hat{m}_t,t,d} \quad (14)$$

ここで，$\boldsymbol{0}_{N\times M}$ は $N \times M$ の零行列を示し，$(L+1)$ 次元行ベクトル $\boldsymbol{w}_L$ と $(L+1)\times(L+1)$ 行列 $\boldsymbol{J}_L$ は

$$\boldsymbol{w}_L = \left[\boldsymbol{0}_{1\times(L-1)}, -1, 1\right] \quad \boldsymbol{J}_L = \begin{bmatrix} 0 & \boldsymbol{I}_{L\times L} \\ 0 & \boldsymbol{0}_{1\times L} \end{bmatrix} \quad (15)$$

で表される．また，$\boldsymbol{I}_{L\times L}$ は $L\times L$ の単位行列を示す．雑音成分 $n_{\hat{m}_t,t,d}$ および $e_{\hat{m}_t,t,d}$ は，各々，平均 0 で分散が $\Sigma_{\hat{m}_t,d}^{(y|X)}$ および $\Sigma_{\hat{m}_t,d}^{(\Delta y|X)}$ の正規分布に従う．ここで，$\mu_{\hat{m}_t,t,d}^{(y|X)}$，$\mu_{\hat{m}_t,t,d}^{(\Delta y|X)}$，$\Sigma_{\hat{m}_t,d}^{(y|X)}$，$\Sigma_{\hat{m}_t,d}^{(\Delta y|X)}$ は，各々，式 (9) の $\boldsymbol{\mu}_{m,t}^{(Y|X)}$ および式 (10) の $\mathbf{\Sigma}_m^{(Y|X)}$ における d 次元目の静的・動的特徴量に対する要素である．状態空間の共分散行列を $\boldsymbol{P}_d^{(t)}$，平均ベクトルを $\hat{\boldsymbol{y}}_d^{(t)}$ とする．各々の初期値 $\boldsymbol{P}_d^{(0)}$, $\hat{\boldsymbol{y}}_d^{(0)}$ を零行列および零ベクトルとして，各フレームにおいて，式 (16) と式 (17) の予測処理と，式 (18) と式 (19) の更新処理を行う．

$$\boldsymbol{P}_d'^{(t-1)} = \boldsymbol{J}_L \boldsymbol{P}_d^{(t-1)} \boldsymbol{J}_L^\top + \mathrm{diag}\left[\boldsymbol{0}_{1\times L}, \Sigma_{\hat{m}_t,d}^{(y|X)}\right] \quad (16)$$

$$\hat{\boldsymbol{y}}_d'^{(t-1)} = \boldsymbol{J}_L \hat{\boldsymbol{y}}_d^{(t-1)} + \left[\boldsymbol{0}_{1\times L}, \mu_{\hat{m}_t,t,d}^{(y|X)}\right]^\top \quad (17)$$

$$\boldsymbol{P}_d^{(t)} = \left(\boldsymbol{I} - \boldsymbol{k}_d^{(t)} \boldsymbol{w}_L\right) \boldsymbol{P}_d'^{(t-1)} \quad (18)$$

$$\hat{\boldsymbol{y}}_d^{(t)} = \hat{\boldsymbol{y}}_d'^{(t-1)} + \boldsymbol{k}_d^{(t)} \left(\mu_{\hat{m}_t,t,d}^{(\Delta y|X)} - \boldsymbol{w}_L \hat{\boldsymbol{y}}_d'^{(t-1)}\right) \quad (19)$$

ここで，$(L+1)$ 次元カルマンゲインベクトル $\boldsymbol{k}_d^{(t)}$ は

$$\boldsymbol{k}_d^{(t)} = \boldsymbol{P}_d^{(t-1)} \boldsymbol{w}_t^\top \left(\Sigma_{\hat{m}_t,d}^{(\Delta y|X)} + \boldsymbol{w}_L \boldsymbol{P}_d^{(t-1)} \boldsymbol{w}_L^\top\right)^{-1} \quad (20)$$

で与えられる．各フレームにおいて，更新後の $\hat{\boldsymbol{y}}_d^{(t)}$ の一つ目の要素を，$t-L$ フレームにおける変換特徴量ベクトル $\hat{\boldsymbol{y}}_{t-L}$ の d 次元目の要素 $\hat{y}_{t-L,d}$ とする．

次に，$P\left(\boldsymbol{v}^{(\boldsymbol{y})}|\boldsymbol{\lambda}^{(v)}\right)$ の最大化を行うために，GV を考慮したポストフィルタ処理を行う．

$$\hat{y}_{t-L,d}^{(\mathrm{GV})} = \mu_d^{(v)^{\frac{1}{2}}} \hat{\mu}_d^{(v)^{-\frac{1}{2}}} \left(\hat{y}_{t-L,d} - \langle \hat{y}_d \rangle\right) + \langle \hat{y}_d \rangle \quad (21)$$

ここで，$\mu_d^{(v)}$ および $\hat{\mu}_d^{(v)}$ は，出力静的特徴量および変換静的特徴量の GV 平均ベクトルの d 次元目の要素を表し，$\langle \hat{y}_d \rangle$ は変換静的特徴量の d 次元目のバイアス値の平均を表す．これらの値は，学習データを用いて事前に求めておく．

[3]連立一次方程式で取り扱う $T\times T$ の逆行列を帯行列で近似し，フレーム毎に帯行列の要素（式 (18) における $\boldsymbol{P}_d^{(t)}$ に対応）を再帰的に更新する処理となる．

図 1 に，リアルタイム変換処理の流れを示す．各フレームにおいて，特徴量抽出処理，短遅延変換処理が行われる．波形合成処理においては，PSOLA（pitch synchronous overlap and add）[27] により励振源波形を生成した後に，スペクトル包絡パラメータを畳み込む．分析窓長を 25 ms とし，入力セグメント特徴量抽出時のパラメータ C を 2，短遅延変換時のパラメータ L を 2，F_0 の最小値を 70 Hz（基本周期 <15 ms）とした際には，最大遅延時間は 50 ms となる．

4 音声生成機能拡張

リアルタイム声質変換処理により，音声コミュニケーションにおける音声生成機能の拡張を行う．以下では，応用例をいくつか紹介する．

4.1 無喉頭音声強調

喉頭摘出者は声帯の消失に伴い，食道発声や電気式人工喉頭を用いた発声に代表される代替発声法を用いる必要がある．しかしながら，生成される音声（無喉頭音声）の品質は，通常音声と比較して大きく劣化する．そこで，リアルタイム声質変換を用いて，無喉頭音声を通常音声へと変換することで，喉頭摘出者の音声生成機能を拡張する [28, 29]．入力特徴量として無喉頭音声のスペクトルセグメント特徴量を使用する．出力特徴量として，通常音声のスペクトル特徴量のみでなく，F_0 や非周期成分などの音源特徴量の推定も行う．本処理により，無喉頭音声の自然性を大幅に改善することが可能となる．

4.2 サイレント音声通話

音声コミュニケーションに内在する本質的な問題として，相手に聴取可能な音声を発声しなければならないという点がある．結果として，周囲に第三者がいる際には秘匿性の高い会話が困難となったり，発声により周囲に迷惑をかけるなどといった状況が生じる．この問題に対して，体内伝導マイクロホンの一つである非可聴つぶやき（nonaudible murmur: NAM）マイクロフォンを用いて，周囲が聴取困難なほど小さな声を収録する枠組みが提案されている [30]．しかしながら，収録される体内伝導音声の品質および明瞭性は著しく劣化するために，リアルタイム声質変換を用いて，体内伝導音声を通常音声へと変換する [31]．これにより，声を出さずに特定の相手に対してのみ意思伝達を行う一種のテレパシーのような音声コミュニケーション形態が実現される．

4.3 ボイス／ボーカルエフェクタ

リアルタイム声質変換により，自身の身体的制約を超えた声質による発声が可能となる．さらに，固有声変換技術 [32] を導入することで，特定の声質を持つ話者の声に変換するだけでなく，声質を制御するパラメータを手動で操作することで，所望の声質を生み出すことも可能となる．歌声変換にも応用できるため [33]，新たな歌唱表現が生み出される可能性がある．

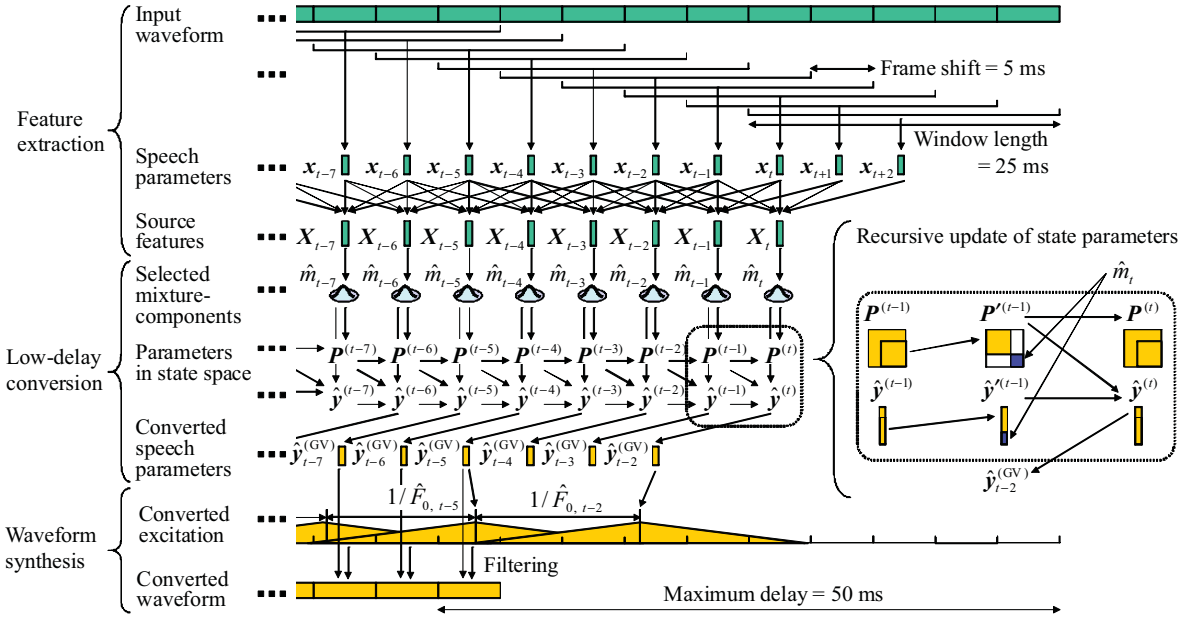

Fig. 1　Frame-by-frame processing in real-time voice conversion ($C = 2$, $L = 2$).

5　おわりに

リアルタイム声質変換技術と，その応用例である音声生成機能拡張について述べた．未だ解決すべき問題は多く残されており，外部雑音環境や発話様式の変化への対応，声質制御の操作性改善，遅延時間の削減および変換精度の向上などに取り組む必要がある．なお，音声生成機能拡張により，これまで非言語情報として取り扱われていたものも意図的に制御可能となる．より豊かな発声表現の実現が期待される反面，なりすまし等の危険性は増す．本技術がもたらす利点と危険性の両面に真摯な目を向け，社会的に正しく技術が認知されるよう努めることが重要である．

謝辞　本研究の一部は，科研費補助金若手研究（A）により実施したものである．

参考文献

[1] Abe *et al.*, J. Acoust. Soc. Jpn. (E), **11**(2), 71–76, 1990.
[2] Stylianou *et al.*, IEEE Trans. Speech & Audio Process., **6**(2), 131–142, 1998.
[3] Toda *et al.*, IEEE Trans. Audio, Speech & Lang. Process., **15**(8), 2222–2235, 2007.
[4] Toda *et al.*, Proc. INTERSPEECH, 2012.
[5] Narendranath *et al.*, Speech Commun., **16**(2), 207–216, 1995.
[6] Takashima *et al.*, Proc. SLT, 313–317, 2012
[7] 中鹿 他, 音講論, 517–520, Mar. 2013.
[8] Zen *et al.*, Speech Commun., **51**(11), 1039–1064, 2009.
[9] 吉村 他, 信学論（D–II）, **J83–D–II**(11), 2099–2107, 2000.
[10] Gales, Computer Speech & Lang., **12**(2), 75–98, 1998.
[11] Gauvain and Lee, IEEE Trans. Speech & Audio Process., **2**(2), 291–298, 1994.
[12] Yamagishi and Kobayashi, IEICE Trans. Inf. & Syst., **E90-D**(2), 533–543, 2007.
[13] Kuhn *et al.*, IEEE Trans. Speech & Audio Process., **8**(6), 695–707, 2000.
[14] Shichiri *et al.*, Proc. INTERSPEECH, 1269–1272, 2002.
[15] Nose *et al.*, IEICE Trans. Inf. & Syst., **E90-D**(9), 1406–1413, 2007.
[16] Tao *et al.*, IEEE Trans. Speech & Audio Process., **14**(4), 1145–1154, 2006.
[17] Wu *et al.*, IEEE Trans. Speech & Audio Process., **18**(6), 1394–1405, 2010.
[18] Inanoglu *et al.*, Speech Commun., **51**(3), 268–283, 2010.
[19] 中野, 後藤, 情処学論, **52**(12), 3853–3867, 2011.
[20] Nose *et al.*, IEICE Trans. Inf. & Syst., **E93-D**(9), 2483–2490, 2010.
[21] Kawahara *et al.*, Speech Commun., **27**(3–4), 187–207, 1999.
[22] Kain and Macon, Proc. ICASSP, 285–288, 1998.
[23] 徳田 他, 音響誌, **53**(3), 192–200, 1997.
[24] Koishida *et al.*, IEICE Trans. Inf. & Syst., **E84-D**(10), 1427–1434, 2001.
[25] Muramatsu *et al.*, Proc. INTERSPEECH, 1076–1079, 2008.
[26] Tokuda *et al.*, Proc. ICASSP, 1315–1318, 2000.
[27] Moulines and Charpentier, Speech Commun., **9**(5–6), 453–467, 1990.
[28] Doi *et al.*, IEICE Trans. Inf. & Syst., **E93-D**(9), 2472–2482, 2010.
[29] Nakamura *et al.*, Speech Commun., **54**(1), 134–146, 2012.
[30] 中島 他, 信学論, **J87-D-II**(9), 1757–1764, 2004.
[31] Toda *et al.*, IEEE Trans. Speech & Audio Process., **20**(9), 2505–2517, 2012.
[32] 戸田, 音講論, **1-8-11**, 257–260, Sep. 2011.
[33] Doi *et al.*, Proc. APSIPA ASC, 2012.