変換モデルと話者モデルの確率的統合に基づく声質変換法の検討*

☆齋藤大輔 (東大/NTT 研究所), 渡部晋治, 中村篤 (NTT 研究所), 峯松信明 (東大)

1 はじめに

声質変換は，入出力の対応関係を記述する変換モデルに基づいて，任意の文に対して入力音声の声質を所望の声質へ変換する技術である．声質変換は，テキスト音声合成における話者性の制御をはじめとして [1]，雑音環境下音声の音声強調や身体運動から音声への変換など多岐にわたる応用が検討されている [2, 3]．声質変換においては，話者性の再現という観点から，特に音声のスペクトルを表現する特徴が重要な役割を果たすため，多くの研究がスペクトル特徴の変換に主眼をおいている．

入出力の対応関係を記述する変換モデルの構築に関しては，多くの手法が提案されている．そのなかでも統計的変換手法は盛んに研究されており，コードブックマッピング法をはじめ [4]，ニューラルネットワークを用いた手法 [5] や混合正規分布モデル（GMM）に基づく変換法 [1, 6] など数多く提案されている．その中でも GMM に基づく変換法はその柔軟性から近年主流となっている．GMM に基づく変換法では，入力特徴ベクトル，もしくは入力と出力の特徴ベクトルを連結した結合ベクトルに対して，その確率密度分布を GMM によってモデル化する．この GMM を用いて，それぞれの正規分布に対応する線形変換を入力特徴ベクトルの事後確率で重み付けした重み付き線形和として，入出力間の対応関係を導出できる．特に結合ベクトルをモデル化する Kain らの手法においては，入出力ベクトルと各正規分布との対応を単一の特徴量空間で記述するため，入出力の対応関係をより精緻にモデル化することが可能となる [1].

しかし，これら GMM に基づく変換法では，基本的に同一発話内容の入出力音声対からなるパラレルデータを用いる必要がある．加えて，結合ベクトルを用いる場合，モデル化するベクトル空間の次元数が 2 倍になるため，学習に用いるパラレルデータの量が少なくなった場合，過学習の問題が顕著になる [1]．これらの問題に対処するため，上記のようなパラレルデータを必要としない手法もいくつか提案されている [7, 8]．これらの手法は，結合確率密度分布に含まれるパラメータを非パラレルデータを用いて適応するアプローチとなっている．これらの手法とは異なり，本研究では，結合ベクトルによって得られた変換モデルと，出力話者の特徴ベクトルをモデル化した話者モデルを確率的に統合する声質変換法を提案する．

提案法では声質変換の機能を，1) 入出力話者の言語内容の同一性の保証 および 2) 出力話者の話者性の表現 という二つに分離して考える．その上でこれらの機能のモデル化を，結合ベクトルによる変換モデルと出力話者の話者モデルをそれぞれ独立に用いて実現し，さらにこれらを確率的に統合することで所望の変換を実現する．提案法は分離したこれらのモデルを独立に学習することが可能であり，パラレルデータは前者のモデル化にのみ使用すればよい．提案法のさらなる利点として，それぞれのモデルに対して，固有声変換法 [9] や UBM に基づく話者モデル [10] など，より精緻なモデル化手法を導入することができる．また提案法は言語内容の同一性を保証する極少量のパラレルデータを用いるため，特に入力話者の負担を大きく軽減することができると考えられる．本稿では提案法の定式化について述べ，パラレルデータが極少量の場合における実験により，提案法の有効性を示す．

2 GMM に基づく声質変換法

今，入力話者の発話を表す特徴量系列を $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_{n_x}]$，同一発話内容の出力話者の特徴量系列を $\boldsymbol{Y} = [\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_{n_y}]$ とする．ただし $^\top$ は転置を表す．動的計画法を用いてこれらの系列をフレーム毎に対応づけることで，結合ベクトル $\boldsymbol{z} = [\boldsymbol{x}^\top, \boldsymbol{y}^\top]^\top$ の特徴量系列 $\boldsymbol{Z} = [\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_n]$ を得る．この特徴量系列を用いて，以下の式で表される GMM のパラメータを推定し，ベクトル $\boldsymbol{z}_t$ の確率密度をモデル化する．

$$P(\boldsymbol{z}_t|\boldsymbol{\lambda}^{(z)}) = \sum_{m=1}^{M} w_m \mathcal{N}(\boldsymbol{z}_t; \boldsymbol{\mu}_m^{(z)}, \boldsymbol{\Sigma}_m^{(z)}) \quad (1)$$

ここで $\mathcal{N}(\boldsymbol{z}_t; \boldsymbol{\mu}_m^{(z)}, \boldsymbol{\Sigma}_m^{(z)})$ は，平均ベクトルを $\boldsymbol{\mu}_m^{(z)}$，分散共分散行列を $\boldsymbol{\Sigma}_m^{(z)}$ とする m 番目の正規分布を表し，w_m は各分布の重みを表す．$\boldsymbol{\lambda}^{(z)}$ は結合ベクトルの GMM の一連のモデルパラメータを表すものとする．$\boldsymbol{z}_t$ のベクトル空間は，その部分空間として入力および出力話者の特徴ベクトルを含むため，これらの

---

* "Voice conversion based on probabilistic integration of joint density model and speaker model" by SAITO Daisuke (The Univ. of Tokyo / NTT Corporation), WATANABE Shinji, NAKAMURA Atsushi (NTT Corporation), and MINEMATSU Nobuaki (The Univ. of Tokyo)

モデルパラメータは以下のように表すことができる．

$$\boldsymbol{\mu}_m^{(z)} = \begin{bmatrix} \boldsymbol{\mu}_m^{(x)} \\ \boldsymbol{\mu}_m^{(y)} \end{bmatrix}, \boldsymbol{\Sigma}_m^{(z)} = \begin{bmatrix} \boldsymbol{\Sigma}_m^{(xx)} & \boldsymbol{\Sigma}_m^{(xy)} \\ \boldsymbol{\Sigma}_m^{(yx)} & \boldsymbol{\Sigma}_m^{(yy)} \end{bmatrix} \quad (2)$$

ただし $\boldsymbol{\mu}_m^{(x)}, \boldsymbol{\Sigma}_m^{(xx)}, \boldsymbol{\mu}_m^{(y)}, \boldsymbol{\Sigma}_m^{(yy)}$ は m 番目の正規分布における入力または出力話者の平均ベクトルおよび分散共分散行列である．また $\boldsymbol{\Sigma}_m^{(xy)}$ および $\boldsymbol{\Sigma}_m^{(yx)}$ は，入出力話者間の相互共分散行列を表す．

変換関数 $\mathcal{F}(\cdot)$ は，入力ベクトル $\boldsymbol{x}_t$ が与えられた場合の $\boldsymbol{y}_t$ の条件付き確率密度に基づいて導出することができる．この確率密度は上述の GMM のモデルパラメータ $\boldsymbol{\lambda}^{(z)}$ によって表現でき，以下のようになる．

$$P(\boldsymbol{y}_t|\boldsymbol{x}_t, \boldsymbol{\lambda}^{(z)}) = \sum_{m=1}^{M} P(m|\boldsymbol{x}_t, \boldsymbol{\lambda}^{(z)}) P(\boldsymbol{y}_t|\boldsymbol{x}_t, m, \boldsymbol{\lambda}^{(z)}) \quad (3)$$

ここで

$$\begin{aligned} P(m|\boldsymbol{x}_t, \boldsymbol{\lambda}^{(z)}) &= \frac{w_m \mathcal{N}(\boldsymbol{x}_t; \boldsymbol{\mu}_m^{(x)}, \boldsymbol{\Sigma}_m^{(xx)})}{\sum_{m=1}^{M} w_m \mathcal{N}(\boldsymbol{x}_t; \boldsymbol{\mu}_m^{(x)}, \boldsymbol{\Sigma}_m^{(xx)})} \\ P(\boldsymbol{y}_t|\boldsymbol{x}_t, m, \boldsymbol{\lambda}^{(z)}) &= \mathcal{N}(\boldsymbol{y}_t; \boldsymbol{E}_{m,t}^{(y)}, \boldsymbol{D}_m^{(y)}) \\ \boldsymbol{E}_{m,t}^{(y)} &= \boldsymbol{\mu}_m^{(y)} + \boldsymbol{\Sigma}_m^{(yx)} \boldsymbol{\Sigma}_m^{(xx)-1} (\boldsymbol{x}_t - \boldsymbol{\mu}_m^{(x)}) \\ \boldsymbol{D}_m^{(y)} &= \boldsymbol{\Sigma}_m^{(yy)} - \boldsymbol{\Sigma}_m^{(yx)} \boldsymbol{\Sigma}_m^{(xx)-1} \boldsymbol{\Sigma}_m^{(xy)} \end{aligned} \quad (4)$$

となる．最小平均二乗誤差基準に基づく変換関数は以下のようになる．

$$\mathcal{F}(\boldsymbol{x}_t) = \sum_{m=1}^{M} P(m|\boldsymbol{x}_t, \boldsymbol{\lambda}^{(z)}) \boldsymbol{E}_{m,t}^{(y)} \quad (5)$$

一方，最尤変換に基づくパラメータ生成を導入した場合は，式 (4) における分散共分散行列を考慮し，以下のようなパラメータ生成のための更新式を得る [11]．

$$\begin{aligned} \hat{\boldsymbol{y}}_t &= \left(\sum_{m=1}^{M} \gamma_{m,t} \boldsymbol{D}_m^{(y)-1} \right)^{-1} \left(\sum_{m=1}^{M} \gamma_{m,t} \boldsymbol{D}_m^{(y)-1} \boldsymbol{E}_{m,t}^{(y)} \right) \\ \gamma_{m,t} &= P(m|\boldsymbol{x}_t, \boldsymbol{y}_t, \boldsymbol{\lambda}^{(z)}) \end{aligned} \quad (6)$$

3 モデルの確率的統合による声質変換

3.1 変換機能の独立なモデル化

基本的に，声質変換は入力話者の音声の特徴を「言語内容を変化させずに」出力話者のそれへと変換する技術として考えることができる．一方，声質変換を出力話者の音声を出力するシステムとして捉えると，入力話者の発声は言語内容を提供する入力として考えることもできる．このような視点から，声質変換における変換モデルは二つの機能を有していると考えられる．一つは入出力話者の言語内容の同一性を保証する機能，もう一つは出力話者の話者性を表現する機能である．前者の機能をモデル化するためには，同一発話内容を含むパラレルコーパスが必要となる．しかし後者については，出力話者の音声データのみを用いてモデル化することが可能である．よって提案法では前者に変換モデル，後者に話者モデルを用いてこれらの機能を実現する．変換に際しては，出力ベクトルに対する条件付き最尤基準を考え，これらのモデルを確率的に統合する．

3.2 提案法の定式化

まず，従来の声質変換法と同様に入力ベクトル $\boldsymbol{x}_t$ が与えられた場合の $\boldsymbol{y}_t$ の条件付き確率密度に着目する．条件付き最尤基準に基づき，最適な出力ベクトルは以下の形で与えられる．

$$\hat{\boldsymbol{y}}_t = \underset{\boldsymbol{y}_t}{\operatorname{argmax}} P(\boldsymbol{y}_t|\boldsymbol{x}_t) \quad (7)$$

ベイズの定理により，音声認識や統計的機械翻訳の定式化と同様に式 (7) は以下の形で表すことができる．

$$\hat{\boldsymbol{y}}_t = \underset{\boldsymbol{y}_t}{\operatorname{argmax}} \underbrace{P(\boldsymbol{x}_t|\boldsymbol{y}_t)}_{\text{from joint density model}} \underbrace{P(\boldsymbol{y}_t)}_{\text{from speaker model}} \quad (8)$$

式 (8) において，第 1 項 $P(\boldsymbol{x}_t|\boldsymbol{y}_t)$ は，同一発話内容の音声データ対によってモデル化することで，入力発話と出力発話の同一性を担保するモデルとして捉えることができる．一方，第 2 項 $P(\boldsymbol{y}_t)$ は，出力話者の話者性を表現するモデルに対応する．本研究では第 1 項のモデルをパラレルデータを用いて学習し，第 2 項は話者認識で用いられるような話者 GMM を用いてモデル化する．話者 GMM は出力話者の音声データのみで学習することが可能である．

以下では，式 (8) の最適化による声質変換法を導出する．変換モデルおよび話者モデルは共に GMM を用いてモデル化し，$\boldsymbol{\lambda}^{(z)}$ および $\boldsymbol{\lambda}^{(s)}$ をそれぞれのモデルパラメータとする．式 (8) に基づき，以下のような尤度関数を定義する．

$$\mathcal{L}(\boldsymbol{y}_t; \boldsymbol{x}_t, \boldsymbol{\lambda}^{(z)}, \boldsymbol{\lambda}^{(s)}) \triangleq P(\boldsymbol{x}_t|\boldsymbol{y}_t, \boldsymbol{\lambda}^{(z)}) P(\boldsymbol{y}_t|\boldsymbol{\lambda}^{(s)})^{\alpha} \quad (9)$$

ここで α は話者モデルに対する重みを表す定数であり，音声認識における言語モデル重みに相当する．$\mathcal{L}$ に対する最適解 $\hat{\boldsymbol{y}}_t$ を得るため，$\hat{\boldsymbol{y}}_t$ に対する補助関数を導入する．以下の導出では，従来法と同様の形で，n および N を話者 GMM における分布インデックスおよび混合数とする．このとき，

$$\begin{aligned} &\log \mathcal{L}(\hat{\boldsymbol{y}}_t) \\ &= \log \sum_{m=1}^{M} P(\boldsymbol{x}_t, m|\hat{\boldsymbol{y}}_t, \boldsymbol{\lambda}^{(z)}) + \alpha \log \sum_{n=1}^{N} P(\hat{\boldsymbol{y}}_t, n|\boldsymbol{\lambda}^{(s)}) \\ &\geq Q_{z_1}(\boldsymbol{y}_t, \hat{\boldsymbol{y}}_t) + Q_{z_2}(\boldsymbol{y}_t, \hat{\boldsymbol{y}}_t) + \alpha Q_s(\boldsymbol{y}_t, \hat{\boldsymbol{y}}_t) \end{aligned} \quad (10)$$

となる．Q_{z_1}, Q_{z_2} および Q_s は，以下のように表される補助関数である．

$$Q_{z_1}(\boldsymbol{y}_t, \hat{\boldsymbol{y}}_t) = \sum_{m=1}^{M} \gamma_{m,t} \log P(m|\hat{\boldsymbol{y}}_t, \boldsymbol{\lambda}^{(z)}) \quad (11)$$

$$Q_{z_2}(\boldsymbol{y}_t, \hat{\boldsymbol{y}}_t) = \sum_{m=1}^{M} \gamma_{m,t} \log P(\boldsymbol{x}_t|m, \hat{\boldsymbol{y}}_t, \boldsymbol{\lambda}^{(z)}) \quad (12)$$

$$Q_s(\boldsymbol{y}_t, \hat{\boldsymbol{y}}_t) = \sum_{n=1}^{N} \gamma_{n,t} \log P(\hat{\boldsymbol{y}}_t, n|\boldsymbol{\lambda}^{(s)}) \quad (13)$$

$$\gamma_{m,t} = P(m|\boldsymbol{y}_t, \boldsymbol{\lambda}^{(z)}), \gamma_{n,t} = P(n|\boldsymbol{y}_t, \boldsymbol{\lambda}^{(s)}) \quad (14)$$

式 (10) の導出に際しては，Jensen の不等式を用いた．ここで，声質変換においては言語内容を変化させないことから，$\hat{\boldsymbol{y}}_t$ の変化に対して式 (11) が急激には変化しないと仮定する．すなわち，Q_{z_1} の $\hat{\boldsymbol{y}}_t$ に対する微係数を無視する．結果として，以下の関数を逐次的に最大化することで最適解 $\hat{\boldsymbol{y}}_t$ を得る．

$$Q'(\boldsymbol{y}, \hat{\boldsymbol{y}}_t) = Q_{z_2}(\boldsymbol{y}_t, \hat{\boldsymbol{y}}_t) + \alpha Q_s(\boldsymbol{y}_t, \hat{\boldsymbol{y}}_t). \quad (15)$$

式 (15) の $\hat{\boldsymbol{y}}_t$ に対する微係数を 0 とすることで，以下の更新式を得る．

$$\hat{\boldsymbol{y}}_t = \left(\sum_{m=1}^{M} \gamma_{m,t} \boldsymbol{D}_m^{'(y)-1} + \alpha \sum_{n=1}^{N} \gamma_{n,t} \boldsymbol{\Sigma}_n^{-1} \right)^{-1} \times \left(\sum_{m=1}^{M} \gamma_{m,t} \boldsymbol{D}_m^{'(y)-1} \boldsymbol{E}_{m,t}^{'(y)} + \alpha \sum_{n=1}^{N} \gamma_{n,t} \boldsymbol{\Sigma}_n^{-1} \boldsymbol{\mu}_n \right) \quad (16)$$

ここで $\boldsymbol{\mu}_n$ および $\boldsymbol{\Sigma}_n$ は話者 GMM の n 番目の分布の平均ベクトルおよび分散共分散行列である．また $\boldsymbol{E}_{m,t}^{'(y)}, \boldsymbol{D}_m^{'(y)-1}$ は以下のように表される．

$$\boldsymbol{E}_{m,t}^{'(y)} = \boldsymbol{\mu}_m^{(y)} + \boldsymbol{\Sigma}_m^{(yy)} \boldsymbol{\Sigma}_m^{(xy)+} (\boldsymbol{x}_t - \boldsymbol{\mu}_m^{(x)}) \quad (17)$$

$$\boldsymbol{D}_m^{'(y)-1} = \boldsymbol{D}_m^{(y)-1} - \boldsymbol{\Sigma}_m^{(yy)-1} \quad (18)$$

ただし，$(\cdot)^+$ は一般化逆行列を表す．最適解 $\hat{\boldsymbol{y}}_t$ は式 (14) および (16) を相互に用いることで得られる．式 (14) の初期値について，本研究では以下の二通りについて考える．

1. 入力話者の $\boldsymbol{x_t}$ を用いて，$\gamma_{m,t} = P(m|\boldsymbol{x}_t, \boldsymbol{\lambda}^{(z)})$ および $\gamma_{n,t} = P(n|\boldsymbol{x}_t, \boldsymbol{\lambda}^{(s)})$ とする方法（以下，INIT1 と呼ぶ）．
2. 入力話者の $\boldsymbol{x_t}$ を従来法による式 (5) により変換し，$\boldsymbol{y_t} = \mathcal{F}(\boldsymbol{x_t})$ として式 (14) を適用する方法（以下，INIT2 と呼ぶ）．

式 (16) は，従来法で最尤基準を導入した式 (6) と同様の形をしている．しかし，提案法の更新式は結合ベクトルに基づく変換モデルおよび話者モデルからの，それぞれの影響の重み付き線形和の形になっている．すなわち提案法においては，結合ベクトルに基づく変換モデルにおける過学習の影響を，出力話者の話者モデルによって緩和しうると考えられる．

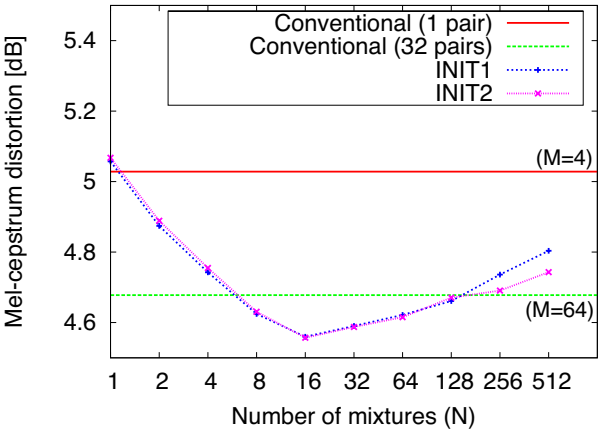
3.3 他手法との関連

声質変換において，学習時に際してある程度のパラレルデータを必要とするという要件は柔軟な声質変換アプリケーションを構築する上では大きな問題となる．そのため，多量のパラレルデータを必要としない声質変換法が検討されている．Mouchtaris らは，異なる話者間で構築された変換モデルのパラメータについて，制約付きの線形回帰によって適応するアプローチを提案している [7]．また Lee らにより，MAP 適応によるアプローチも検討されている [8]．提案法を比較すると，これらの手法は主に出力話者の話者性を柔軟に制御することに主眼をおいており，一度十分量のパラレルデータを用いて変換モデルを構築する必要があり，変換モデルの構築に際して入力話者はある程度の量の文章を読むことになる．一方，提案法では変換モデルの学習データが極少量であっても有効に機能することが期待できる．すなわち，声質変換アプリケーションのユーザである入力話者の負担を大きく軽減することが可能となる．さらに提案法は変換モデルと話者モデルを独立に構築するため，それぞれの研究で培われた精緻なモデル化を容易に導入する余地がある．

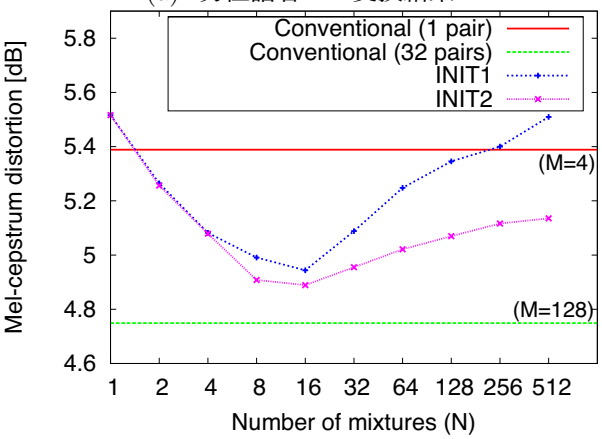
4 実験

提案手法の有効性を確かめるため，日本語文章を用いた声質変換実験を行った．音声データとして，ATR 日本語音声データベース [12] の B セットから 5 名の話者の選定した（MSH, MMY, MTK, FKS, FTK）．このうち MSH を入力話者とし，その他を出力話者とした．評価データはサブセット J の 53 文とした．変換モデルは 1 文のパラレルデータを用いて学習し，GMM の混合数（M）は 4 とした．一方，提案法で用いる話者モデルについては，サブセット I の 50 文を用いて学習した．話者 GMM の混合数（N）について，1 から 512 まで変化させた．話者モデルの重み（α）は 1 とした．提案法における最適化の反復回数は 5 とした．

スペクトル特徴量として，STRAIGHT 分析に基づくスペクトルから得られた 24 次のメルケプストラムを用いた [13]．STRAIGHT による合成に用いる非周期性指標については全周波数において-30 dB とした．パワーおよび基本周波数については平均と標準偏差を考慮した単純な線形変換によって変換した．本実験では特に，従来の変換モデルのみを用いた手法と提



(a): 男性話者への変換結果



(b): 女性話者への変換結果

Fig. 1　客観評価結果: M は従来法での変換モデルの混合数，N は話者モデルの混合数

案手法との比較を行った．

メルケプストラム歪みに基づく客観評価の結果を図 1 に示す．図 1 は変換対象話者の性別毎に表示している．また参考のため，32 文パラレルデータを用いた従来法の結果を示している．なお混合数は最良の結果のものを選んでいる．(a)(b) いずれの場合にも，従来法の 1 文による変換モデルを上回っていることがわかる．これは 1 文のパラレルデータでは言語内容の同一性保証および話者性のモデル化の二つを共に充足するほど十分なモデル化ができていないことを意味する．加えて (a):男性話者への変換の場合には，32 文のパラレルデータを用いた場合を上回る性能になっている．これは出力話者の特徴量空間のモデリングに特化した形で話者モデルを利用できる効果と考えられる．一方，3.2 における初期化法の違いとして，特に (b):女性話者への変換において，事前に従来の変換モデルを用いて初期値を設定することで，特に男性から女性への変換のような，より両特徴の離れた場合の変換に対して有効に機能していると考えられる．以上より提案手法は特にパラレルデータが極少量の場合に有効性を持っているといえる．

5　おわりに

本稿では，言語内容の同一性を保証する変換モデルと出力話者の話者性を表現する話者モデルを確率的に統合した声質変換法について述べた．提案法はこれらのモデルを独立に学習することができ，特にパラレルデータが極少量の場合でも有効に機能することを実験的に示した．今後の課題として，聴取実験によって本手法の有効性を示すこと，およびそれぞれのモデルに対してより精緻なモデル化を導入することがあげられる．

参考文献

[1] A. Kain and M. W. Macon, “Spectral voice conversion for text-to-speech synthesis,” Proc. ICASSP, vol. 1, pp. 285–288, 1998.

[2] L. Deng, A. Acero, L. Jiang, J. Droppo, and X. Huang, “High-performance robust speech recognition using stereo training data,” Proc. ICASSP, pp. 301–304, 2001.

[3] A. Kunikoshi, Y. Qiao, N. Minematsu, and K. Hirose, “Speech generation from hand gestures based on space mapping,” Proc. INTERSPEECH, pp. 308–311, 2009.

[4] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, “Voice conversion through vector quantization,” Proc. ICASSP, pp. 655–658, 1988.

[5] S. Desai, E. V. Raghavendra, B. Yegnanarayana, A. W. Black, and K. Prahallad, “Voice conversion using artificial neural networks,” Proc. ICASSP, pp. 3893–3896, 2009.

[6] Y. Stylianou, O. Cappé, and E. Moulines, “Continuous probabilistic transform for voice conversion,” IEEE Trans. on Speech and Audio Processing, vol. 6, no. 2, pp. 131–142, 1998.

[7] A. Mouchtaris, J. V. der Spiegel, and P. Mueller, “Non-parallel training for voice conversion based on a parameter adaptation approach,” IEEE Trans. on Audio, Speech, and Language Processing, vol. 14, no. 3, pp. 952–963, 2006.

[8] C. H. Lee and C. H. Wu, “Map-based adaptation for speech conversion using adaptation data selection and non-parallel training,” Proc. INTERSPEECH, pp. 2254–2257, 2006.

[9] T. Toda, Y. Ohtani, and K. Shikano, “Eigenvoice conversion based on Gaussian mixture model,” Proc. INTERSPEECH, pp. 2446–2449, 2006.

[10] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, “Speaker verification using adapted Gaussian mixture models,” Digital Signal Processing, vol. 10, no. 1–3, pp. 19–41, 2000.

[11] T. Toda, A. W. Black, and K. Tokuda, “Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory,” IEEE Trans. on Audio, Speech, and Language Processing, vol. 15, no. 8, pp. 2222–2235, 2007.

[12] A. Kurematsu, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara, and K.Shikano, “ATR Japanese speech database as a tool of speech recognition and synthesis,” Speech Communication, vol.9, pp.357–363, 1990.

[13] H. Kawahara, I. Masuda-Katsuse, and A.de Cheveigné, “Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds,” Speech Communication, vol.27, pp.187–207, 1999.