Vol.2011-SLP-85 No.10
2011/2/5

音声翻訳システムのための声質変換法と日中英語間における評価

服部信彦[†1] 戸田智基[†1] 河井恒[†2]
猿渡洋[†1] 鹿野清宏[†1]

異なる言語間でのコミュニケーションを可能にするため，音声翻訳システムの研究が行われている．このシステムでは，入力音声に対して音声認識，機械翻訳，テキスト音声合成（Text-To-Speech: TTS）の処理が行われ，翻訳された音声が出力される．しかし，出力音声は出力言語に応じた話者のものとなるため，その声質は入力音声とは異なるものとなり，入力話者の声の個人性は失われる．そこで，入力話者の声質による他言語音声出力を可能にするために，一対多固有声変換（Eigenvoice Conversion: EVC）および言語依存確率分布関数に基づく声質制御を提案する．一対多 EVC により，出力音声の言語情報を保存したまま，声質のみを入力音声のものへと変換する処理を実現する．また，言語依存確率分布関数を用いることで，個々の言語が持つ大局的な韻律的特徴を反映させた韻律変換を実現する．日本語，中国語，英語間における音声翻訳システムを対象として，提案法の有効性を示す．

Voice Conversion for Speech-to-Speech Translation System and Its Evaluation among Japanese, Chinese and English

NOBUHIKO HATTORI ,[†1] TOMOKI TODA ,[†1]
HISASHI KAWAI,[†2] HIROSHI SARUWATARI[†1]
and KIYOHIRO SHIKANO[†1]

To enable speech communication between different languages, there have been studied technologies for developing a speech-to-speech translation system. In this system, automatic speech recognition, machine translation, and text-to-speech (TTS) are sequentially performed to present translated voices. One weakness of this system is that speaker individuality is missed due to the output voice quality different from that of an input speaker since the TTS system needs to be developed with voices of another speaker in an output language. To address this issue, we propose voice quality control based on one-to-many eigenvoice conversion (EVC) and prosodic modification based on a language-dependent probability distribution function. The output voice quality is converted into that of the input speaker while keeping language contents unchanged by one-to-many EVC. Moreover, to further improve naturalness of the converted speech, prosodic parameters are globally modified considering their global differences between input and output languages. The effectiveness of the proposed method is demonstrated in experimental evaluations assuming speech-to-speech translation among Japanese, Chinese, and English.

1. はじめに

急激な国際化の進展により，異なる言語間でのコミュニケーションを可能にする翻訳システムへの期待が高まっている．それに伴い，音声翻訳システムの研究が盛んに行われている[1)]．このシステムは，入力音声を認識する音声認識部，認識結果を目標言語に変換する機械翻訳部，音声出力を行う音声合成部から構成される．これにより，入力音声に対して，翻訳された音声が出力される．しかし，従来の音声翻訳システムでは，TTS の出力音声は出力言語に応じた話者の声になるため，出力音声は入力音声とは異なる声質となってしまう．仮に，入力話者の声質を保持した出力言語音声を合成できれば，音声翻訳システムを用いて，より自然で円滑なコミュニケーションをとることができると期待される．

音声翻訳システムの出力音声をシステムユーザーの声にする手法として，隠れマルコフモデル（Hidden Markov Model: HMM）に基づく音声合成[2)] を用いて，モデル適応処理[3)4)] を行う手法が提案されている．文献 3) で提案されている方法では，デコーディング処理により音素系列を決定し，対応するトライフォン HMM に対する適応規則を求め，それを音声合成用のフルコンテキスト HMM へと適用する．また，文献 4) では，バイリンガルデータを用いて各言語の HMM を学習し，各モデル間のマッピングを定義することで，他言語間での適応処理を実現している．これら HMM の適応に基づく方法では，モデルが音韻的な制約を含むため，デコーディング処理が必須である．そのため，音声認識誤りの影響は避けられない．また，異なる言語のモデル間のマッピングを定義するためには，十分な量のバ

†1 奈良先端科学技術大学院大学
Nara Institute of Science and Technology
†2 独立行政法人 情報通信研究機構
National Institute of Information and Communications Technology

イリンガルデータが必要となるが，様々な言語対に対して，そのようなデータを収録するのは容易ではない．

本稿では，音声認識処理やバイリンガルデータを必要とせずに，合成音声の声質を入力話者のものへと変換する手法を提案する．異なる言語間における声質変換を実現するために，音韻情報を用いずに音響パラメータの確率密度をモデル化して変換を行う混合正規分布（Gaussian mixture model: GMM）に基づく声質変換法[5)6)]に注目し，その応用技術である一対多固有声変換（Eigenvoice Conversion: EVC）[7)]を音声翻訳システムに導入する[8)]．一対多 EVC は，ある特定の話者から任意の話者へと声質を変換する技術であり，目標話者による極少量かつ任意の発声から得られる音声特徴量のみを用いて，事前に学習されたモデルを適応することで，特定話者から目標話者への変換モデルを作成する．これにより，TTS 出力話者からユーザーの声質への変換が可能となる．また，出力言語音声の自然性を改善するために，個々の言語が持つ韻律パラメータの大局的な特徴に着目し，言語依存確率分布関数を用いた韻律変換法を提案する[9)]．本稿では，日本語，中国語，英語の全組み合わせに対する実験的評価により，一対多 EVC および言語依存確率分布を用いた変換法の有効性を示す．

2. 一対多固有声変換法（一対多 EVC）

2.1 固有声 GMM（EV-GMM）

$\boldsymbol{X}_t = [\boldsymbol{x}_t^\top, \Delta\boldsymbol{x}_t^\top]^\top$，$\boldsymbol{Y}_t^{(s)} = [\boldsymbol{y}_t{}^{(s)\top}, \Delta\boldsymbol{y}_t{}^{(s)\top}]^\top$ を，t フレーム目における元話者の特徴量および s 番目の事前収録目標話者の特徴量とする．また，$\boldsymbol{Z}_t^{(s)} = [\boldsymbol{X}_t^\top, \boldsymbol{Y}_t^{(s)\top}]^\top$ を，元話者と目標話者の特徴量をフレーム毎に対応付けた結合ベクトルとする．次式により，結合確率密度を EV-GMM $\lambda^{(EV)}$ でモデル化する．

$$P\left(\boldsymbol{Z}_t^{(s)}|\lambda^{(EV)}, \boldsymbol{w}^{(s)}\right) = \sum_{i=1}^{M} \alpha_i \mathcal{N}\left(\boldsymbol{Z}_t^{(s)}; \boldsymbol{\mu}_i^{(Z)},\ \boldsymbol{\Sigma}_i^{(ZZ)}\right)$$

$$\boldsymbol{\mu}_i^{(Z)} = \begin{bmatrix} \boldsymbol{\mu}_i^{(X)} \\ \boldsymbol{\mu}_i^{(Y)} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\mu}_i^{(X)} \\ \boldsymbol{B}_i^{(Y)}\boldsymbol{w}^{(s)} + \boldsymbol{b}_i^{(Y)}(0) \end{bmatrix}, \boldsymbol{\Sigma}_i^{(ZZ)} = \begin{bmatrix} \boldsymbol{\Sigma}_i^{(XX)} & \boldsymbol{\Sigma}_i^{(XY)} \\ \boldsymbol{\Sigma}_i^{(YX)} & \boldsymbol{\Sigma}_i^{(YY)} \end{bmatrix} \quad (1)$$

ここで，$\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ は平均ベクトル $\boldsymbol{\mu}$，共分散行列 $\boldsymbol{\Sigma}$ の正規分布であり，α_i は i 番目の分布に対する分布重み，M は混合数を表す．EV-GMM では，目標話者の平均ベクトルは，バイアスベクトル $\boldsymbol{b}_i^{(Y)}(0)$ と $\boldsymbol{B}_i^{(Y)} = [\boldsymbol{b}_i^{(Y)}(1), \boldsymbol{b}_i^{(Y)}(2), \cdots, \boldsymbol{b}_i^{(Y)}(J)]$ で示される J 個の固有ベクトル $\boldsymbol{b}_i^{(Y)}(j)$ の線形結合で表わされる．目標話者の声質は，J 次元の重みベクトル $\boldsymbol{w}^{(s)} = [w_1^{(s)}, w_2^{(s)}, \cdots, w_J^{(s)}]^\top$ を用いて制御される．

2.2 EV-GMM の学習

本稿では，話者正規化学習（Speaker Adaptive Training: SAT）[10)]を行うことで EV-GMM を学習する．SAT では，適応後のモデルの尤度が最大になるように，以下の式に従い適応元モデルを学習する．

$$\hat{\lambda}^{(EV)}, \hat{\boldsymbol{w}}^{1:S} = \mathrm{argmax} \prod_{s=1}^{S} \prod_{t=1}^{T_s} P(\boldsymbol{Z}_t^{(s)}|\lambda^{(EV)}, \boldsymbol{w}^{(s)}) \quad (2)$$

ここで，$\hat{\lambda}^{(EV)}$ は更新された適応元モデル，$\hat{\boldsymbol{w}}^{(1:S)}$ は全ての目標話者に対する重みベクトルのセットを示す．

2.3 EV-GMM の適応と変換

EVC では，所望の目標話者の音声データのみを用いて，次式に従って EV-GMM の重みベクトル $\boldsymbol{w}$ を最尤推定することができる．これにより，ある話者から所望の目標話者への変換モデルが構築される．

$$\hat{\boldsymbol{w}} = \mathrm{argmax} \prod_{t=1}^{T} \int P(\boldsymbol{X}_t, \boldsymbol{Y}_t^{(tar)} \mid \lambda^{(EV)}, \boldsymbol{w}) d\boldsymbol{X}_t \quad (3)$$

ここで，$\boldsymbol{Y}_t^{(tar)}$ は t フレーム目の目標話者の特徴量を表す．

適応された EV-GMM を用いて声質変換を行う．元話者の特徴量系列 $\boldsymbol{X} = [\boldsymbol{X}_1^\top, \cdots, \boldsymbol{X}_T^\top]^\top$ から，目標話者の静的特徴量系列 $\boldsymbol{y} = [\boldsymbol{y}_1^\top, \cdots, \boldsymbol{y}_T^\top]^\top$ への変換は，次式に従って，目標話者の特徴量系列 $\boldsymbol{Y} = [\boldsymbol{Y}_1^\top, ..., \boldsymbol{Y}_T^\top]^\top$ の条件付き確率密度を最大化することで求められる．

$$\hat{\boldsymbol{y}} = \mathrm{argmax} P(\boldsymbol{Y}|\boldsymbol{X}, \hat{\boldsymbol{m}}, \lambda^{(EV)}, \hat{\boldsymbol{w}}) \quad (4)$$

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

ここで，$\boldsymbol{W}$ は静的特徴量系列 $\boldsymbol{y}$ を静的・動的特徴量系列 $\boldsymbol{Y}$ に拡張する変換行列を表し，$\hat{\boldsymbol{m}}$ は入力特徴量系列 $\boldsymbol{X}$ に対して次式により求められる最尤分布系列である．

$$\hat{\boldsymbol{m}} = \mathrm{argmax} P(\boldsymbol{m}|\boldsymbol{X}, \lambda^{(EV)}) \quad (6)$$

なお，本稿では，変換性能を改善するために，系列内変動を考慮した変換処理[11)]を行う．

3. 一対多固有声変換に基づく音声翻訳システムの出力声質制御

3.1 一対多 EVC を用いた音声翻訳システム

提案する音声翻訳システムの構成を図 1 に示す．TTS の出力音声に対して一対多 EVC を行うことで，声質制御を行う．まず，TTS の出力話者を入力話者とした一対多 EV-GMM を事前に学習する．システム使用時には，音声翻訳システムへの入力音声を適応データとして用いて一対多 EV-GMM の教師無し適応を行うことで，TTS 出力話者からシステム入力話者への変換モデルを構築する．得られた変換モデルを用いて，TTS 出力音声の声質をシステム入力話者の声質へと変換する．なお，本稿では，TTS の音声合成処理方式として，HMM に基づく音声合成方式を用いる．

音声合成処理では，音声認識及び機械翻訳の結果に基づいて出力文 HMM $\lambda^{(\mathrm{HMM})}$ を決定し，次式に基づき出力音声特徴量系列を生成する．

$$\hat{\boldsymbol{q}} = \mathrm{argmax} P(\boldsymbol{q}|\lambda^{(\mathrm{HMM})}) \tag{7}$$

$$\hat{\boldsymbol{x}} = \mathrm{argmax} P(\boldsymbol{W}\boldsymbol{x}|\lambda^{(\mathrm{HMM})}, \hat{\boldsymbol{q}}) \tag{8}$$

ここで，$\boldsymbol{q}$ は状態系列を表わす．$\hat{\boldsymbol{x}}$ は HMM により生成される音声特徴量を表わす．次に，式 (3) に基づき，音声翻訳システムへの入力音声に対して，EV-GMM の重みベクトル $\boldsymbol{w}$ を推定する．なお，適応データ量が極端に少ない場合に過剰な適応によって変換精度が劣化するのを防ぐため，本稿では最大事後確率推定を用いる[12]．得られた適応 EV-GMM を用いることで，次式により，式 (5) に示す条件の下で，出力音声特徴量系列からシステム入力話者の音声特徴量系列へと変換する．

$$\hat{\boldsymbol{m}} = \mathrm{argmax} P(\boldsymbol{m}|\boldsymbol{W}\hat{\boldsymbol{x}}, \lambda^{(EV)}) \tag{9}$$

$$\hat{\boldsymbol{y}} = \mathrm{argmax} P(\boldsymbol{Y}|\boldsymbol{W}\hat{\boldsymbol{x}}, \hat{\boldsymbol{m}}, \lambda^{(EV)}, \hat{\boldsymbol{w}}) \tag{10}$$

また，基本周波数 F_0 については，次式にて変換を行う．

$$\log \hat{F_0}^{(Y)} = \frac{\sigma^{(Y)}}{\sigma^{(X)}}(\log {F_0}^{(X)} - \mu^{(X)}) + \mu^{(Y)} \tag{11}$$

ここで，$\mu^{(X)}, \sigma^{(X)}$ は，元話者の音声の対数 F_0 の平均，標準偏差を表し，HMM により生成される特徴量より計算する．また，$\mu^{(Y)}$，$\sigma^{(Y)}$ は目標話者（システム入力話者）の対数 F_0 の平均及び標準偏差であり，適応データから計算する．

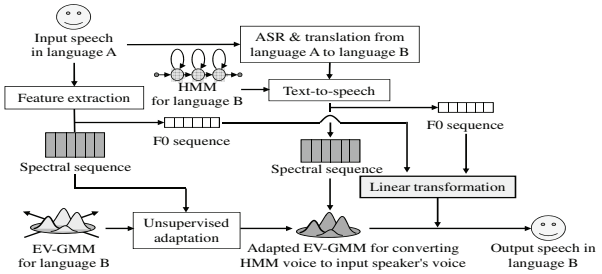

図 1 一対多 EVC を用いた音声翻訳システム．
Fig. 1 Speech-to-speech translation system with one-to-many eigenvoice conversion.

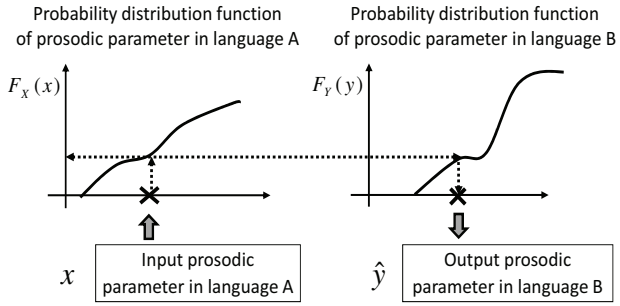

図 2 言語依存確率分布関数に基づく韻律パラメータ変換法．
Fig. 2 Prosodic parameter conversion method based on language-dependent probability distribution functions.

3.2 音声翻訳システムのための EV-GMM 学習法

EVC では複数話者のパラレルデータから，EV-GMM を事前に学習する．そのため，入力話者と各出力話者による同一発話内容の自然音声が数多く必要になってしまう．しかし，そのような音声データを得ることは容易ではない．そこで，EV-GMM の学習データに用いる入力話者の音声特徴量として，TTS により生成された音声特徴量を用いる[8]．TTS で任意の発話内容に対する音声特徴量を容易に生成できるため，パラレルデータ構築のために再度 TTS の出力話者による音声収録を行う必要がなくなる．結果，既存の音声データベースに含まれる様々な話者の音声データを用いて，EV-GMM を学習することが可能となる．

Vol.2011-SLP-85 No.10
2011/2/5

4. 言語依存確率分布関数を用いた韻律変換

入力話者に適応した出力言語音声の自然性をさらに改善するためには，異なる言語間における韻律変換の導入が効果的であると予想される．しかしながら，話者間（翻訳システム入力話者と HMM 出力話者）における正確な韻律パラメータ変換規則を抽出するためには，大量のバイリンガルデータが必要となる[4]．また，継続長等のように，パラメータ自体が言語依存である韻律特徴量（例えば，日本語ではモーラ単位を用いるのに対し，英語では音節単位を用いる場合など）に関しては，異なる言語間でパラメータ変換を行うのは容易ではない．この問題に対して，ある言語の韻律パラメータにおける相対的な話者間の関係は，別の言語でも保存されると仮定し，言語依存確率分布関数を用いた変換法を提案する．

4.1 変 換 法

言語依存確率分布関数に基づく変換法を図 2 に示す．事前に収録された多数の入出力言語話者の音声データを用いて，各言語に対して独立に事前収録話者の韻律パラメータに関する確率分布関数をもとめる．

$$F_X(x) = P(X \leq x) = \int_{-\infty}^{x} f_X(x\prime)dx\prime \tag{12}$$

$$F_Y(y) = P(Y \leq y) = \int_{-\infty}^{y} f_Y(y\prime)dx\prime \tag{13}$$

ここで，x 及び X は入力言語における話者依存韻律パラメータ及びその確率変数を表し，y 及び Y は出力言語における話者依存韻律パラメータ及びその確率変数を表す．また，f_X および f_Y は，入出力言語における韻律パラメータの確率密度関数を表す．ある話者において，入力言語の韻律パラメータと出力言語の韻律パラメータ間には以下の関係が成り立つと仮定する．

$$P(Y \leq y) = P(X \leq x) \tag{14}$$

これは，例えば，入力言語において他の話者と比べて相対的に声の高さが低いのであれば，出力言語においても他の話者と比べて相対的に声の高さが低くなると仮定することを意味する．この場合，入力言語の韻律パラメータから出力言語の韻律パラメータへの変換は，次式で表される．

$$\hat{y} = F_Y^{-1}(F_X(x)) \tag{15}$$

提案法は，個々の言語において多数話者の音声データを必要とする．音声認識等の研究を通じて，そのようなデータの整備は広く行われており，入出力言語のバイリンガルデータより も容易に入手可能である．また，話者依存の韻律パラメータとして，コンテキストに対する依存性が低く，少量の音声データから容易に計算できるものを用いることで，音声翻訳システムにおいて入力音声のみを用いた教師なし適応が可能となる．なお，本手法は基本周波数と継続長に対して適用可能だが，継続長に対しては改善が得られなかったため[9]，本稿では基本周波数に対してのみ適用する．

4.2 基本周波数の変換

F_0 変換では，入力言語（翻訳システム入力言語）と出力言語（HMM の出力言語）の対数 F_0 の平均と標準偏差を韻律パラメータとする．事前に，各言語において，多数話者の音声データから，両パラメータの確率分布関数を求めておく．目標話者（翻訳システム入力話者）の適応データから算出される入力言語における対数 F_0 の平均及び標準偏差を，式 (15) により出力言語における対数 F_0 の平均及び標準偏差へと変換し，得られた値を式 (11) の $\mu^{(Y)}$ および $\sigma^{(Y)}$ とすることで，F_0 を変換する．

4.3 確率分布関数のモデリング

韻律パラメータの確率分布関数を精度良く求めるためには，膨大な数の話者数が必要となるが，実際に使用できる話者数は限られる．そこで，より頑健に確率分布関数を求めるために，確率分布関数のモデリングを行う．対数 F_0 の平均に関しては，男性話者及び女性話者による分布を考慮して，2 混合の GMM で確率密度関数をモデル化する．この時，全ての言語，全ての正規分布において，等混合重み，等分散という制約を用いる．さらに，異なる言語間において，一番目の正規分布の平均値の差分と二番目の正規分布の平均値の差分は等しいという制約も用いる．一方で，対数 F_0 の標準偏差に関しては，各言語において，等分散の正規分布で確率密度関数をモデル化する．この場合，入力言語の韻律パラメータから出力言語の韻律パラメータへの変換は，次式のように簡略化できる．

$$\hat{y} = x + (\mu_{output} - \mu_{input}) \tag{16}$$

ここで，x は入力言語における話者依存韻律パラメータを示し，μ_{input}，μ_{output} はそれぞれ入力言語に対する正規分布の平均値と出力言語に対する正規分布の平均値を示す．

5. 評 価 実 験

5.1 実 験 条 件

TTS の日本語，中国語，英語出力話者として各々女性 1 名を用いる．EV-GMM 学習時に用いる事前収録目標話者として，JNAS データ[13]，ATRPTH データ[14]，BTEC データ[15] に含まれる男女各 50 名の話者を用いて，それぞれ日本語，中国語，英語の EV-GMM を

作成する．また，EV-GMM 学習時に用いる入力話者の音声特徴量として，TTS により生成された音声特徴量を学習データとして用いる．確率分布関数の作成に用いる事前収録話者として，日本語に関しては BTEC データに含まれる男女各 163 名の計 326 名の話者を用いる．中国語に関しては ATRPTH データに含まれる男女各 270 名の計 540 名の話者を用いる．英語に関しては BTEC データに含まれる男女各 100 名の計 200 名の話者を用いる．主観評価実験では，目標話者（システム入力話者）として，JNAS データ（日本語），ATRPTH データ（中国語），BTEC データ（英語）から，確率分布関数作成話者に含まれない男女各 2 名の計 4 名の話者を用いる．各目標話者の適応データとして，日本語，中国語，英語それぞれ 2 文を用いる．評価データとして，学習データに含まれていない 40 文を用いる．また，客観評価実験では，目標話者として，日本語・英語バイリンガル話者 4 名，日本語・中国語バイリンガル話者 2 名を用いる．各目標話者の適応データとして，日本語，中国語，英語それぞれ 1〜32 文を用いる．評価データとして，学習データに含まれていない 20 文を用いる．

スペクトル特徴量として，STRAIGHT[16] により得られるメルケプストラム係数を用いる．メルケプストラムの分析次数は 24 とする．EV-GMM の混合数は 128 とし，固有ベクトル数は 99 とする．確率分布関数を求める韻律パラメータとして，F_0 に関しては，日本語話者，中国語話者，英語話者共に，対数 F_0 の平均及び標準偏差を用いる．

客観評価実験では，適応時に用いる音声の言語とモデル学習時に用いる音声の言語が異なる場合および同じ場合において，変換音声と目標音声間のメルケプストラム歪みを求め，スペクトル変換精度を評価する．

主観評価実験では，話者性と自然性に関して，対比較評価（XAB テスト）を行う．初めに，目標話者であるシステム入力話者の分析合成音声を提示し，次に各種手法による出力音声のペアをランダムな順で提示する．話者性の評価では，どちらの変換音声が目標話者の音声に近いかを判断する．自然性の評価では，どちらの変換音声が目標話者が発声した出力言語として自然であるかを判断する．被験者は，各出力言語で 10 名（日本語：日本人 10 名，中国語：中国人 10 名，英語：アメリカ人 2 名およびフィリピン人 8 名）であり，出力言語を母国語または公用語として使用している国の出身者で行う．評価対象を表 1 に，評価に用いた言語の組み合わせを表 2 に示す．

5.2 バイリンガルデータを用いたスペクトル変換精度の評価

図 3 に客観評価実験の結果を示す．図 3 中のラベルは表 3 に基づく．一対多 EVC を行わない場合（すなわち，変換前）のメルケプストラムひずみは，それぞれ日本語で 8.25 [dB]，中国語で 8.19 [dB]，英語で 8.31[dB] である．図 3 から，適応時とモデル学習時に用いる音声の言語が同一の場合も異なる場合も，一対多 EVC により変換前より歪みが大幅に下がっており，1 文といった極少量の適応データでも大きな話者性改善効果が得られることがわかる．また，適応時とモデル学習時に用いる音声の言語が異なる場合は，同一の場合と比べて，メルケプストラム歪みが若干大きくなる傾向が見られる．

表 1 評価対象
Table 1 Methods to be evaluated

Label	Spectrum	F_0
w/o conversion	Generated from HMM	Generated from HMM
EV-GMM + LT	Converted with EV-GMM	Transformed linearly
EV-GMM + LT-PDF	Converted with EV-GMM	Transformed linearly with language-dependent PDFs

表 2 評価に用いた言語対
Table 2 Language-pairs used in experimental evaluations

Label	Input and output languages of speech-to-speech translation system
CHI-ENG	Chinese to English and English to Chinese
JPN-ENG	Japanese to English and English to Japanese
JPN-CHI	Japanese to Chinese and Chinese to Japanese

表 3 客観評価における評価対象
Table 3 Methods in objective evaluation

Label	Training	Adaptation/Conversion
JPN-same	Japanese	Japanese
JPN-cross	Japanese	Chinese and English
CHI-same	Chinese	Chinese
CHI-cross	Chinese	Japanese and English
ENG-same	English	English
ENG-cross	English	Japanese and Chinese

5.3 一対多 EVC および言語依存確率分布を用いた変換法の評価

各言語（日本語，中国語，英語）における個々の韻律パラメータの確率分布関数を図 4 に示す．対数 F_0 の標準偏差において言語間に大きな違いが見られ，特に，日本語と英語間では大きな違いがみられる．このことから，例えば英語と日本語間の変換を考えた場合，英語発声から得られる F_0 の標準偏差を日本語発声に直接適用すると，日本語の F_0 としては標準偏差が小さくなりすぎると予想される．

図 5 に主観評価結果を示す．図 5 から，全ての言語の組み合わせにおいて，一対多 EVC を用いることで，より目標話者の声質に近い出力音声を合成できることがわかる．また，日

Vol.2011-SLP-85 No.10
2011/2/5

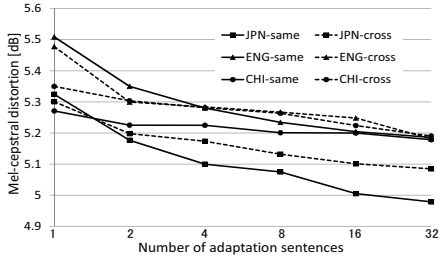

図 3 各適応文数におけるメルケプストラムひずみ．同一言語間における変換時と異なる言語間における変換時の比較.

Fig. 3 Mel-cepstral distortion as a function of the number of adaptation sentences in same-language voice conversion and cross-language voice conversion.

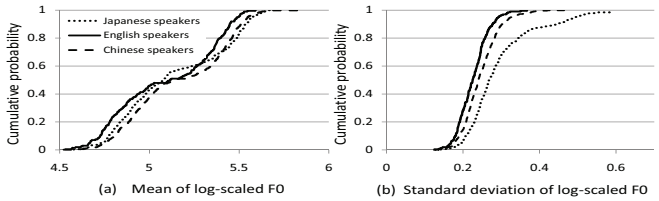

図 4 各韻律パラメータの確率分布関数.

Fig. 4 Probability distribution function of each prosodic parameter.

本語と英語間，日本語と中国語間の変換で，F_0 変換に確率分布関数を用いた提案法を導入することで，出力音声の自然性を改善できることがわかる．これは，図 4 に見られる言語間の対数 F_0 の標準偏差の違いを考慮することで，出力言語においてより自然な F_0 へと変換できるためである．中国語と英語間の変換で自然性の向上が見られないのは，図 4 に見られるように，中国語と英語間では対数 F_0 の標準偏差に大きな違いが見られないためである．

6. ま と め

本稿では，音声翻訳システムにおいて，個人性に優れた出力音声の合成を行うために，一対多固有声変換法（Eigenvoice Conversion: EVC）を音声翻訳システムに適用した．さらに，変換音声の自然性を改善するために，言語依存確率分布関数に基づく韻律パラメータ変換法を提案した．日本語，中国語，英語間における実験的評価結果から，提案法の高い有効性を示した．

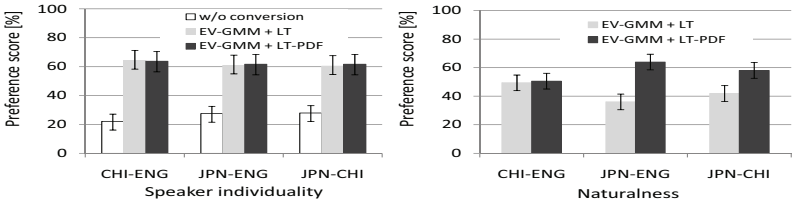

図 5 主観評価結果.

Fig. 5 Results of subjective evaluations.

謝辞 本研究の一部は，科研費補助金若手研究（A）及び総務省 SCOPE により実施したものである．

参 考 文 献

1) S. Nakamura, K. Markov, H. Nakaiwa, G. Kikui, H. Kawai, T. Jitsuhiro, J. -S. Zhang, H. Yamamoto, E. Sumita, and S. Yamamoto, “ATR Multi-lingual Speech-To-Speech Translation System,”, *IEEE Trans. ASLP*, Vol. 14, pp. 365–376, 2006.
2) 吉村 貴克, 徳田 恵一, 益子 貴史, 小林 隆夫, 北村 正, “HMM に基づく音声合成におけるスペクトル・ピッチ・継続長の同時モデル化,”, 信学論 (D-II), Vol. J83–D-II, No. 11, pp. 2099–2107, 2000.
3) S. King, K. Tokuda, H. Zen, and J. Yamagishi, “Unsupervised adaptation for HMM-based speech synthesis,”, *Proc. INTERSPEECH*, pp. 1869-1872, Brisbane, Australia, 2008.
4) Y. J. Wu, Y. Nankaku, and K. Tokuda, “State mapping based method for cross-lingual speaker adaptatio in HMM-based speech synthesis,”, *Proc. INTERSPEECH*, pp. 528–531, 2009.
5) Y. Stylianou, O. Capp´e and E. Moulines, “Continuous Probabilistic Transform for Voice Conversion,”, *IEEE Trans. SAP*, Vol. 6, No. 2, pp. 131–142, 1998.
6) T. Toda, A. W. Black, and K. Tokuda, “Voice Conversion Based on Maximum Likelihood Estimation of Spectral Parameter Trajectory,”, *IEEE Trans. ASLP*, Vol. 15, No. 8, pp. 2222–2235, 2007
7) T. Toda, Y. Ohtani, and K. Shikano, “One-to-many and many-to-one voice conversion based on eigenvoices,”, *Proc. ICASSP*, pp. 1249–1252, 2007.
8) 服部 信彦, 戸田 智基, 河井 恒, 猿渡 洋, 鹿野 清宏, “音声翻訳システムのため の一対多固有声変換に基づく声質制御,”, 音講論, pp. 321–322, 2010.
9) 服部 信彦, 戸田 智基, 猿渡 洋, 鹿野 清宏, “音声翻訳システムのための言語依存確率分布関数に基づく韻律変換,”, 音講論, pp. 325–326, 2010.
10) Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano, “Adaptive training for voice conversion based on eigenvoices,”, *IEICE Trans. Information and Systems*, Vol. E93-D, No. 6, pp. 1589–1598, 2010.
11) T. Toda, A. W. Black, and K. Tokuda, “Voice Conversion Based on Maximum Likelihood Estimation of Spectral Parameter Trajectory,”, *IEEEE Trans. ASLP*, Vol. 15, No. 8, pp. 2222–2235, 2007.
12) D. Tani, T. Toda, Y. Ohtani, H. Saruwatari, and K. Shikano, “Maximum A Posteriori Adaptation for Many-to-One Eigenvoice Conversion,”, *Proc. INTERSPEECH*, pp. 1461–1464, 2008.
13) JNAS:JapaneseNewspaperArticleSentences.
http : //www.mibel.cs.tsukuba.ac.jp/jnas/instruct.html
14) J. S. Zhang, M. Mizumachi, F. K. Soong, and S. Nakamura, “ATRPTH の紹介:音韻カバレッジを考慮した中国語音声データベース,”, 音講論, pp. 167–168, 2003.
15) T. Takezawa, E. Sumita, F. Sugaya, H. Yamamoto, and S. Yamamoto, “Toward a broad-coverage bilingual corpus for speech translation of travel conversations in the real world,”, *Proc. LREC*, pp. 147–152, 2002.
16) H. Kawahara, I. Masuda-Katsuse, and A. Cheveigne, “Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneousfrequency-based F0 extraction,”, *Proc. Speech Communication*, Vol. 27, No. 3-4, pp. 187–207, 1999.