話者空間のテンソル表現を用いた一対多声質変換*

◎齋藤大輔, △山本敬介, 峯松信明, 広瀬啓吉 (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，広義には異なる二つの特徴量空間のマッピング技術と考えられ，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]．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．

しかし，変換モデルの構築の際には，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．また，変換モデルの利用は学習時の入出力話者対に限定される．すなわち話者性を柔軟に制御することは声質変換における重要な課題といえる．そのためには他の話者の音声データを事前知識として用いることが有効と考えられ，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．一方，多数の話者の音声データをより効果的に用いる手法として，固有声変換法（Eigenvoice conversion; EVC）が提案されている [9]．EVC では，参照話者と多数の事前収録話者との間の複数のパラレルデータを用いて，固有声に基づく混合正規分布モデル（EV-GMM）を構築する．複数のパラレルデータより得られた結合確率密度分布から，事前収録話者の話者 GMM をそれぞれ抽出し，ガウス分布の平均ベクトルを連結した GMM スーパーベクトルを用いて話者空間を構築する．話者認識の場合と同様に，任意の話者はこの話者空間の一点で表され，基底に対する少数の重みパラメータを推定することで，話者性を柔軟に制御することができる．

しかし，GMM スーパーベクトルによる話者空間表現は，複数要因からの音響的な変動を一つの特徴量空間に含んでいる．すなわち，GMM のガウス分布の要素と平均ベクトルの次元が混在した高次の特徴量空間となっている．本研究では，話者性を柔軟に制御するという観点から，話者空間をテンソルで表現した一対多声質変換を提案する．提案法では，任意の話者はスーパーベクトルではなく，行および列がそれぞれ GMM の要素と平均ベクトルに対応するような行列の形で表現される．このような話者表現を用いることで事前収録話者のデータセットが 3 階のテンソルで表現でき，テンソル解析を導入することで話者空間を構築することができる．テンソル解析は複数要因からの変動を適切に扱うことが可能であり [10]，提案する話者空間表現による性能の向上が期待できる．また，本稿では一対多声質変換にテンソルを用いた話者空間表現を導入するが，提案法は多対一声質変換や話者認識にも応用することができる．また，提案法は主に話者空間表現に着眼しているため，話者正規化学習を用いた EVC [11] や非パラレルデータを用いた学習 [12] との統合も可能である．本稿では提案法の定式化について述べ，従来の EVC と比較して提案法の有効性を示す．

2　固有声変換法（EVC）

2.1　固有声に基づく混合正規分布モデル

本章では，一対多 EVC について概説する．今，参照話者の音響特徴量を $\boldsymbol{X}_t = [\boldsymbol{x}_t^\top, \boldsymbol{\Delta x}_t]^\top$，$s$ 番目の事前収録話者の音響特徴量を $\boldsymbol{Y}_t^{(s)} = [\boldsymbol{y}_t^{(s)^\top}, \boldsymbol{\Delta y}_t^{(s)^\top}]^\top$ と表す．ただし $^\top$ は転置を表す．ここで，音響特徴量は D 次元の静的および動的特徴量を結合した $2D$ 次元の音響特徴量となる．参照話者と事前収録話者の結合確率密度は，EV-GMM として以下のようにモデル化される．

$$P(\boldsymbol{X}_t, \boldsymbol{Y}_t^{(s)} | \boldsymbol{\lambda}^{(EV)}, \boldsymbol{w}^{(s)}) = \sum_{m=1}^{M} \alpha_m \mathcal{N}([\boldsymbol{X}_t^\top, \boldsymbol{Y}_t^{(s)^\top}]^\top; \boldsymbol{\mu}_m^{(Z)}(\boldsymbol{w}^{(s)}), \boldsymbol{\Sigma}_m^{(Z)}) \quad (1)$$

$$\boldsymbol{\mu}_m^{(Z)}(\boldsymbol{w}^{(s)}) = \begin{bmatrix} \boldsymbol{\mu}_m^{(X)} \\ \boldsymbol{B}_m \boldsymbol{w}^{(s)} + \boldsymbol{b}_m^{(0)} \end{bmatrix}, \boldsymbol{\Sigma}_m^{(Z)} = \begin{bmatrix} \boldsymbol{\Sigma}_m^{(XX)} \boldsymbol{\Sigma}_m^{(XY)} \\ \boldsymbol{\Sigma}_m^{(YX)} \boldsymbol{\Sigma}_m^{(YY)} \end{bmatrix} \quad (2)$$

ここで $\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ は，平均ベクトルを $\boldsymbol{\mu}$，分散共分散行列を $\boldsymbol{\Sigma}$ とする正規分布を表す．m 番目の要素の

* “One-to-many voice conversion using tensor representation of speaker space” by SAITO Daisuke, YAMAMOTO Keisuke, MINEMATSU Nobuaki, and HIROSE Keikichi (The Univ. of Tokyo)

重みは α_m で表し，混合数を M とする．EV-GMM では，S 人の事前収録話者を利用して，出力話者の平均ベクトル $\boldsymbol{\mu}_m^{(Y)}$ をバイアスベクトル $\boldsymbol{b}_m^{(0)}$ と $K(< S)$ 個の表現ベクトルの線形結合で表す．このとき，出力話者の話者性は K 次元の重みベクトル $\boldsymbol{w}^{(s)}$ で制御できる．すなわち話者空間が K 個の基底スーパーベクトル $\boldsymbol{B} = [\boldsymbol{B}_1^\top, \boldsymbol{B}_2^\top, \ldots, \boldsymbol{B}_m^\top]^\top \in \mathcal{R}^{2DM \times K}$ とバイアススーパーベクトル $\boldsymbol{b} = \left[\boldsymbol{b}_1^{(0)\top}, \boldsymbol{b}_2^{(0)\top}, \ldots, \boldsymbol{b}_m^{(0)\top}\right]^\top \in \mathcal{R}^{2DM \times 1}$ によって構築される．

2.2 EVC における話者空間構築

主成分分析（PCA）に基づいて，EV-GMM を構築する場合，最初に出力話者非依存の GMM（TI-GMM）を，全ての参照話者と事前収録話者とのパラレルデータを用いて学習する．次に，対応するパラレルデータを用いて TI-GMM の出力話者の平均ベクトルを更新することで，話者依存のモデルを得る．話者空間の特徴量ベクトルとして，事前収録話者の GMM の各要素の平均ベクトルを連結し，スーパーベクトルを生成する．得られたスーパーベクトルを用いて PCA を行うことで，バイアスベクトル $\boldsymbol{b}$ と表現ベクトル $\boldsymbol{B}$ を得ることができる．

2.3 EV-GMM の教師なし適応

任意の話者に対する EV-GMM は，出力話者の音声データを用いて，最尤基準に基づいて重みベクトル $\boldsymbol{w}$ を推定することで適応できる [9]．今，出力話者の音響特徴量系列を $\boldsymbol{Y}^{(tar)}$ とすると，$\boldsymbol{w}$ は以下のように推定できる．

$$\hat{\boldsymbol{w}} = \underset{\boldsymbol{w}}{\operatorname{argmax}} \int P(\boldsymbol{X}, \boldsymbol{Y}^{(tar)} | \boldsymbol{\lambda}^{(EV)}, \boldsymbol{w}) d\boldsymbol{X} \quad (3)$$

EM アルゴリズムを用いて，下記の更新式を得る．

$$\hat{\boldsymbol{w}} = \left\{ \sum_{m=1}^{M} \overline{\gamma}_m^{(tar)} \boldsymbol{B}_m^\top \boldsymbol{\Sigma}_m^{(YY)^{-1}} \boldsymbol{B}_m \right\}^{-1} \sum_{m=1}^{M} \boldsymbol{B}_m^\top \boldsymbol{\Sigma}_m^{(YY)^{-1}} \overline{\boldsymbol{Y}}_m^{(tar)} \quad (4)$$

$$\overline{\gamma}_m^{(tar)} = \sum_{t=1}^{T} \gamma_{m,t}, \quad \overline{\boldsymbol{Y}}_m^{(tar)} = \sum_{t=1}^{T} \gamma_{m,t} (\boldsymbol{Y}_t^{(tar)} - \boldsymbol{b}_m^{(0)}) \quad (5)$$

$$\gamma_{m,t} = P(m | \boldsymbol{Y}_t^{(tar)}, \boldsymbol{\lambda}^{(EV)}, \boldsymbol{w}) \quad (6)$$

式 (4) はおよそ話者空間の基底ベクトルへの射影重みを推定していることに相当する．なお，式 (6) の初期化に際しては，TI-GMM を用いる．適応後のパラメータ生成については [13] と同様である．

3 話者空間のテンソル表現

3.1 多重線形解析

本章では，提案するテンソル解析に基づく話者空間の構築について述べる．まず，提案法に関連する多

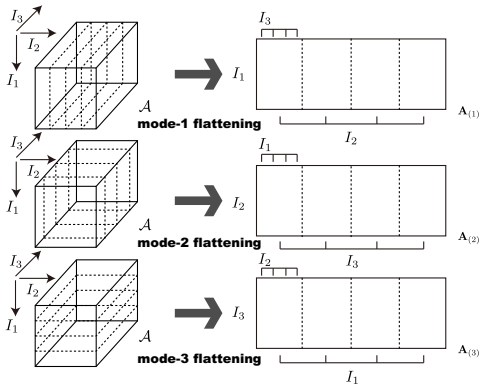

Fig. 1 $I_1 \times I_2 \times I_3$ テンソル $\mathcal{A}$ の平坦化

重線形解析について説明する [14]．テンソルは，行列表現を一般化した多次元配列表現である．テンソルにおける個々のインデックスはモードと呼ばれる．今，$\mathcal{A} \in \mathcal{R}^{I_1 \times I_2 \times I_3}$ を 3 階のテンソルとする．一般に高階のテンソルは Fig. 1 に示すようなモード n の平坦化操作によって行列の形で表現できる．モード n の平坦化では，モード n のインデックスに沿うようにテンソルをスライスし，行列の形に連結する．この平坦化操作を用いることで，テンソルと行列の間の積を定義できる．モード n 積は $\mathcal{A} = \mathcal{G} \times_n \boldsymbol{B}$ と表現し，モード n の平坦化行列を用いて $\boldsymbol{A}_{(n)} = \boldsymbol{B} \cdot \boldsymbol{G}_{(n)}$ のように演算できる．

行列代数において最も重要な演算として特異値分解（SVD）がある．ここで，行列は 2 階のテンソルと見なすことができ，$\boldsymbol{A}$ の特異値分解は，モード n 積を用いて以下のように表される．

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^\top = \mathcal{S} \times_1 \boldsymbol{U} \times_2 \boldsymbol{V} \quad (7)$$

この SVD の表現を高階テンソルに拡張することで，テンソルに対する下記の分解を得る．

$$\mathcal{A} = \mathcal{S} \times_1 \boldsymbol{U}_1 \times_2 \boldsymbol{U}_2 \times_3 \boldsymbol{U}_3 \quad (8)$$

行列の SVD と異なり，ここで $\boldsymbol{U}_1, \boldsymbol{U}_2$ および $\boldsymbol{U}_3$ が直交行列の場合，テンソル $\mathcal{S}$ は密なテンソルとなる．このとき式 (8) を高階特異値分解または Tucker 分解とよぶ [14, 15]．PCA はデータ行列に対する SVD と見なせるため，データテンソルを導入した場合，特徴空間の構築は Tucker 分解によって拡張できる．

3.2 Tucker 分解による話者空間の構築

Tucker 分解に基づいて話者空間を構築するため，各事前収録話者を $M \times D'$ の行列で表現する [16]．ここで M は混合数であり，$D' = 2D$ とする．まずはじめに全データ行列の平均を求め，バイアス行列 $\boldsymbol{b}' = \left[\boldsymbol{b}_1^{(0)}, \boldsymbol{b}_2^{(0)}, \ldots, \boldsymbol{b}_m^{(0)}\right]^\top$ とする．これを各事前収録話者の行列からあらかじめ減算しておく．ここで事前収録話者の話者数を S とすると，話者空間を構築

するデータセットは 3 階のテンソル $\mathcal{M} \in \mathcal{R}^{M \times D' \times S}$ で表される．このデータテンソルを Tucker 分解すると以下のように表される．

$$\mathcal{M} = \mathcal{G}^{M \times D' \times S} \times_1 \boldsymbol{U}^{(M)} \times_2 \boldsymbol{U}^{(D')} \times_3 \boldsymbol{U}^{(S)} \quad (9)$$

ここで，$\boldsymbol{U}^{(M)} \in \mathcal{R}^{M \times M}$, $\boldsymbol{U}^{(D')} \in \mathcal{R}^{D' \times D'}$, $\boldsymbol{U}^{(S)} \in \mathcal{R}^{S \times S}$ である．これらの行列は，それぞれ GMM の混合要素，平均ベクトルの次元，話者インデックスの効果を捉えており，テンソル $\mathcal{G}$ がこれらを統合している．ここで，第 3 モードのインデックスを固定することで，話者 n を表す行列を表現することができる．

$$\boldsymbol{\mu}^{(n)} = \mathcal{G} \times_1 \boldsymbol{U}^{(M)} \times_2 \boldsymbol{U}^{(D')} \times_3 \boldsymbol{U}^{(S)}(n,:) \quad (10)$$

式 (10) から，話者空間の基底として複数の候補が考えられるが，本研究では [16] と同様に以下のようにグルーピングする．

$$\boldsymbol{\mu}^{(n)} = \boldsymbol{U}^{(M)} \left\{ \mathcal{G} \times_2 \boldsymbol{U}^{(D')} \times_3 \boldsymbol{U}^{(S)}(n,:) \right\}^\top = \boldsymbol{U}^{(M)} \boldsymbol{W}_n^\top \quad (11)$$

ここで $\boldsymbol{U}^{(M)}$ を基底とし，$\boldsymbol{W}_n \in \mathcal{R}^{D' \times M}$ を重み行列とする．次元圧縮の観点から，基底を縮約することで，任意の話者は以下のような行列で表される．

$$\boldsymbol{\mu}^{(new)} = \boldsymbol{U}^{(M)} \boldsymbol{W}_{(new)}^\top + \boldsymbol{b}' \quad (12)$$

$\boldsymbol{U}^{(M)} \in \mathcal{R}^{M \times K} (K \leq S)$, $\boldsymbol{W}_{(new)} \in \mathcal{R}^{D' \times K}$ はそれぞれ，表現行列および重み行列となる．ゆえに提案法では，K 次元の重みベクトルを推定する EVC と異なり，$D' \times K$ の重み行列を推定することになる．

[16] では，最小平均二乗誤差に基づく適応アルゴリズムが提案されているが，本研究では最尤基準を導入し，以下の更新式により重み行列を推定する．

$$\mathrm{vec}(\boldsymbol{W}) = \left[\sum_{m=1}^{M} \bar{\gamma}_m^{(tar)} \boldsymbol{U}_m^\top \boldsymbol{U}_m \otimes \boldsymbol{\Sigma}_m^{(YY)^{-1}} \right]^{-1} \mathrm{vec}(\boldsymbol{C}) \quad (13)$$

$$\boldsymbol{C} = \sum_{t=1}^{T} \sum_{m=1}^{M} \gamma_{m,t} \boldsymbol{\Sigma}_m^{(YY)^{-1}} (\boldsymbol{Y}_t^{(tar)} - \boldsymbol{b}_m^{(0)}) \boldsymbol{U}_m \quad (14)$$

$$\boldsymbol{U}_m = \boldsymbol{U}^{(M)}(m,:) \in \mathcal{R}^{1 \times K} \quad (15)$$

ここで，vec() は行列を列ベクトルに展開する演算子である．式 (4) と比較すると，基底の組み合わせは異なるものの近い形をしている．しかし EVC では K 個のパラメータを推定するが，提案法では $D' \times K$ 個のパラメータを推定することになり，提案法はより柔軟にデータに対して適応しうると考えられる．

4 実験

4.1 実験条件

提案手法の有効性を確かめるため，一対多声質変換の実験を行った．参照話者として ATR 日本語音声データベース [17] から男性 1 名のデータを用いた．また事前収録話者として JNAS から男性話者 137 名，女性話者 136 名の計 273 名の発声を用いた [18]．各事前収録話者は 50 文を読み上げている．評価対象話者として男女 3 名ずつを選んだ．適応文数を 1 文から 16 文で変化させ，各話者 21 文を評価に用いた．

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 [19]．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． また GMM の混合数（M）は 128 とした．

変換性能について，提案法に基づく一対多声質変換および一対多 EVC を比較した．また参考として従来のパラレル学習に基づく声質変換の性能についても検証した [13]．なお本研究では話者適応学習は導入しなかった．

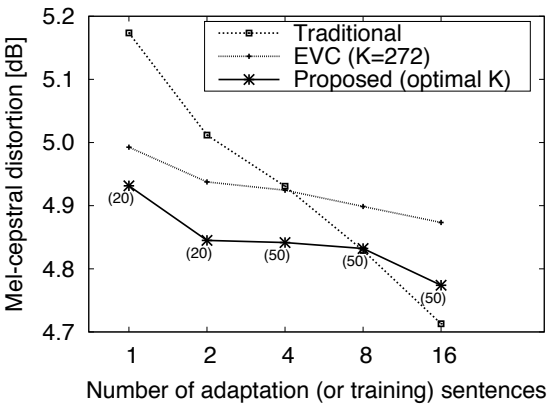

Fig. 2　客観評価実験結果; 提案法の () は最適な K を表す．

4.2 客観評価実験

メルケプストラム歪みに基づく客観評価の結果を Fig. 2 に示す．Fig. 2 は適応データ数に対するメルケプストラム歪みの変化を示している．従来のパラレル学習の結果は，それぞれのデータ数で最適な混合数を選択している．適応データ数が少ない場合は提案法，EVC ともに従来のパラレル学習の結果を上回っている．すなわち事前収録話者のデータが効果的に作用しているといえる．提案法は，いずれの適応文数でも EVC の性能を上回っている．これは提案する話者空間表現がスーパーベクトルによる表現に比べて，より声質変換において有効であるといえる． また EVC ではいずれの適応文数でも全ての表現ベクトルを使用するのが最適となった．一方，提案法では，最適な表現行列のサイズは元の混合数よりも小さく

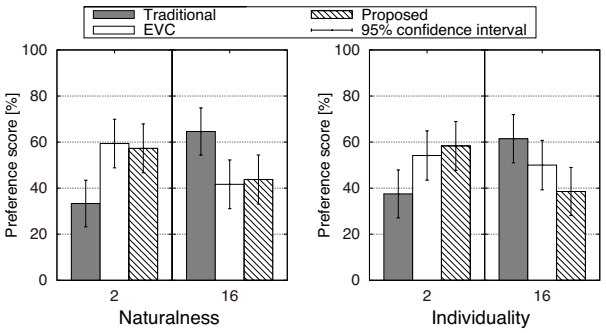

Fig. 3　主観評価実験結果

なっており，また適応文数が少ない場合は K が小さくなっている．すなわち提案法は，話者空間を効果的に表現しているといえる．

4.3　主観評価実験

聴取実験により変換音声の自然性と話者性について評価した．被験者は 8 名で，自然性の評価には一対比較法，話者性の評価には RAB 法を用いた．一対比較法では被験者は 2 種類の変換音声を聴取し，どちらの音声がより自然かを評価した．RAB 法では，2 種類の変換音声を目標話者の参照音声の後に提示し，どちらが参照音声に近いかを評価した．各被験者はランダムに選ばれた 36 文を評価した．

Fig. 3 に主観評価実験の結果を示す．適応文数が 2 文の場合は，提案法，EVC ともに従来のパラレル学習より自然性が向上していることがわかる．EVC と提案法を比較すると適応文数が 16 文の話者性評価を除いて，提案法が EVC と同等か若干よい性能が得られることがわかる．

5　おわりに

本稿では，話者空間のテンソル表現を用いた声質変換における適応手法を提案した．提案法では各話者を行列の形で表現し，事前収録話者のデータテンソルに対してテンソル解析を適用することでより柔軟に話者を表現することが可能となる．今後の課題として，話者適応学習や非パラレルデータの利用等，他の有用な手法と提案法を統合することが上げられる．またテンソルの基底と重みの選択についても検討の余地がある．また表現行列のサイズの最適化についても検討していく予定である．

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] M. A. O. Vasilescu and D. Terzopoulos, “Mutilinear analysis of image ensembles: TensorFaces,” Proc. ECCV, pp. 447–460, 2002.

[11] Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano, “Speaker adaptive training for one-to-many eigenvoice conversion based on Gaussian mixture model,” Proc. INTERSPEECH, pp. 1981–1984, 2007.

[12] Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano, “Non-parallel training for many-to-many eigenvoice conversion,” Proc. ICASSP, pp. 4822–4825, 2010.

[13] 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.

[14] L. De Lathauwer, B. De Moor and J. Vandewalle, “A multilinear singular value decomposition,” SIAM Journal on Matrix Analysis and Applications, vol. 21, No. 4, pp. 1253–1278, 2000.

[15] L. R. Tucker, “Some mathematical notes on three-mode factor analysis,” Psychometrika, vol. 31, no. 3, pp. 279–311, 1966.

[16] Y. Jeong, “Speaker adaptation based on the multilinear decomposition of training speaker models,” Proc. ICASSP, pp. 4870–4873, 2010.

[17] 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.

[18] “Jnas: Japanese newspaper article sentences,” http://www.milab.is.tsukuba.ac.jp/jnas/instruct.html

[19] 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.