TR-IT-0215

話者選択と移動ベクトル場平滑化を用いた声質変換における写像元話者の選択方法

Selection of Reference Speaker for Voice Conversion Using SSVFS Spectral Mapping with Consideration of Vector Field Smoothing Algorithm

橋本 誠 Makoto Hashimoto
樋口 宜男 Norio Higuchi

1997.3.27

話者選択と移動ベクトル場平滑化法（以下、VFS）を用いた声質変換法SSVFSにおける写像元話者選択方法を提案した．SSVFSでは，まず話者選択により複数話者の音声データベースから1名を選択し，次に選択話者空間から目標話者空間へのスペクトル写像をVFSによって行う．これまでに，1単語程度の少ない学習データでもデータベース音声を目標話者音声に近づけられることを示したが，話者選択は，VFSのアルゴリズムに対する適／不適を特に考慮したものではなかった．これは，話者によって写像精度に差が生じる原因ともなっていた．本報告ではVFSに適した話者を選択するための尺度として，移動ベクトルの向きのばらつきを反映した尺度を提案し，従来尺度よりも写像精度との相関が強いことが示された．

もくじ

1 まえがき

本報告では，話者選択（SS：Speaker Selection）と移動ベクトル場平滑化法（VFS：Vector Field Smoothing method，以下ではVFSと略記）[1][2]を用い，コードブックマッピングを基本とした，極めて少ない学習データによる声質変換法SSVFS [3][4]における話者選択方法について述べる．声質変換は，多数話者間での音声翻訳システムにおける話者識別のために必要不可欠であると同時に，利用者のニーズや好みに合った音質を利用者自身が選択できるという機能を有する音声合成システムの実現のためにも非常に重要である．

声質変換に関しては，これまでにも種々の試みがなされてきた[5, 6, 7]．文献[5]では，2話者間のコードブックの対応をとるコードブックマッピング法を提案しているが，学習データが大量に必要となる．また，文献[6]では，複数話者のスペクトルを補間して目標話者に近いスペクトルを得る方法を提案しているが，複数話者のスペクトルパラメータ上での補間であるため，音声生成過程を考慮した場合，あまり好ましくない．更に，文献[7]では音源や声道モデルパラメータ上での制御を試みているが，目標話者の声質に近づけるという手法ではない．

筆者らは，目標話者の少量学習データを入力するだけで，データベース音声を目標話者音声に変換することが可能な声質変換法の確立を試みており，これまで，声質変換のためのスペクトル写像法SSVFSを提案し，1単語程度の少ない学習データでもスペクトルを目標話者に近づけられることを示すと共に[3][4]，SSVFSの学習過程を考慮した適切な学習データの決定指標についても提案してきた[8][9]．

SSVFSは，音声データベースとしてあらかじめ記憶された複数話者の中から1名を選択し，選択された話者を写像元話者として，VFSによって目標話者の声質に近い音声へ変換する方式である．従来のSSVFSでは，写像元話者の選択方法として，学習音声のスペクトル距離最小規準に基づき目標話者の学習音声とのスペクトル距離が最も小さい話者を選択していた．しかし，話者によって写像精度に差があり，任意の話者に対してロバストな性能を得ることが課題であった．この原因として，話者選択に用いていた尺度が写像に用いているVFSのアルゴリズムを考慮したものではなく，VFSに適さない写像元話者の選択を許していたことが挙げられる．

スペクトル写像を行う場合，音響的構造の類似した話者間では比較的精度良く写像が行えると考えられる．VFSは，話者間移動ベクトル（以下，移動ベクトル）の補間と平滑化処理によって写像を行うため，2話者の音響空間で対応づけられる移動ベクトル場の構造が単純な程，精度が向上すると考えられる．少量学習データの場合，従来のようなスペクトル距離のみの尺度では，移動ベクトル場の構造を考慮することはできないため，VFSの前処理となる話者選択としては最適な手法であるとは言えなかった．

これに対して，目標話者・写像元話者間の全移動ベクトルは，学習音声のDTWによる対応づけと，対応づけられた移動ベクトルを用いた未対応コードに対する移動ベクトルの推定を行

うことにより求められるため，得られた移動ベクトルは，話者間の音響空間の対応を表しており，各移動ベクトルの向きのばらつきが2話者の音響的構造の類似性を反映するものと考えられる．

以上のような観点から，本報告では，話者に対するSSVFSのロバスト性を向上させるために，移動ベクトルの向きのばらつきに着目した適切な写像元話者選択のための尺度を提案し，提案尺度・従来尺度それぞれについて，50単語の変換音声の写像精度と尺度の出力値との相関を調べることにより，有効性を検討した．

2　SSVFSの概要

本報告では，あらかじめ音声データを用意しておく複数の話者を登録話者，変換ターゲットとなる話者を目標話者，登録話者から選ばれた話者を選択話者あるいは写像元話者，と呼ぶこととする．

SSVFSは，学習過程とスペクトル写像過程に分けることができる．学習過程では，

1. 登録話者の中から写像元話者を決定する話者選択
2. 選択話者音声スペクトルを目標話者音響空間へ写像するための話者間移動ベクトルを求める移動ベクトル計算

を行い，スペクトル写像過程では，

3. 発話内容に応じた選択話者スペクトル時系列のファジーベクトル量子化と，移動ベクトルを用いた目標話者音響空間へのスペクトル写像

を行う．

本方式のブロック図を図1に示す．なお，各登録話者に対する学習音声スペクトル時系列，合成用音声データおよび大量のスペクトルデータを用いて作成されるコードブックはあらかじめ準備しておくものとする．

3　VFS処理を考慮した写像元話者選択尺度の定式化

従来のSSVFSでは，学習音声のみでの目標話者とのスペクトル距離最小規準により，写像元話者を選択していた．しかし，

1. 少量学習データの場合，学習音声空間のみを考慮した尺度は信頼性が低下する可能性がある
2. スペクトル写像に用いるVFSの前処理としての話者選択であるためVFSのアルゴリズムに適した話者を選択することが望ましい

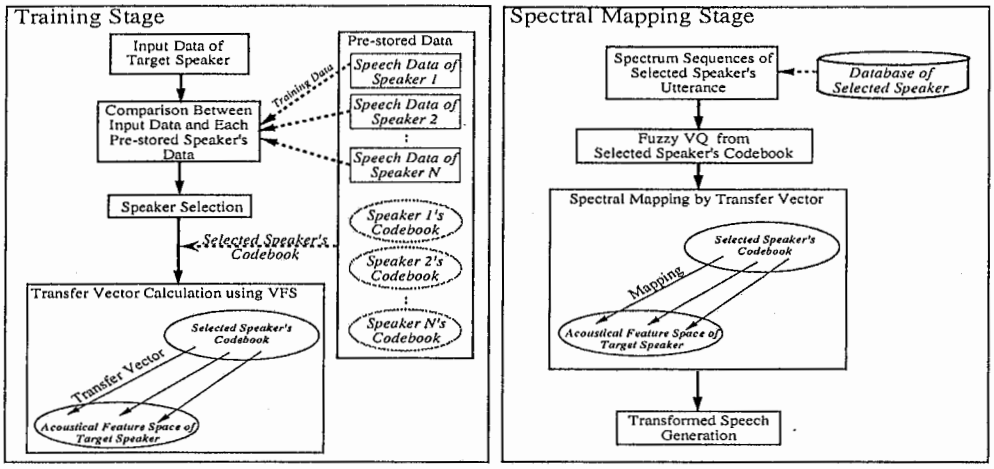

図 1: SSVFS スペクトル写像方式のブロック図

などの理由から，これまで用いてきた学習音声空間のみでのスペクトル距離尺度に基づく話者選択は，任意の話者に対してロバストなシステムを構築するための最適な尺度とは言えなかった．

VFS における標準話者（写像元話者）の選択方法としては，話者適応に関連した文献 [10] がある．これは，学習単語に対する話者適応後の HMM 出力尤度が最大になるような話者を標準話者とする方式である．しかし，この方式も，学習音声空間のみに着目した尺度に基づいた話者選択を行うものであり，少量学習データの場合における話者ロバスト性を保証するには十分ではない．

VFS では，学習データによる対応づけ，未対応コードに対する移動ベクトルの補間などにより話者間の移動ベクトルを求めることで写像を行う．本報告では，求められた移動ベクトルが話者間の音響空間の対応を表しており，各移動ベクトルの向きのばらつきが 2 話者の音響的構造の類似性を表現するものととらえ，各移動ベクトルの向きのばらつきに着目した写像元話者選択のための尺度を定式化した．

以上の観点から，移動ベクトルの向きのばらつきを表す尺度は，全移動ベクトルの平均ベクトル（以下，平均移動ベクトル）と個々の移動ベクトルとの関係に基づいたものとし，まず，学習音声空間のみを考慮した尺度と，音響空間全域を考慮した尺度の優劣を評価するため，以下の 2 つの定式化を行った．

- 平均移動ベクトル・被学習移動ベクトル間の差のノルムに基づく尺度 D_t（以下，被学習移動ベクトル平均偏差）

$$D_t = \frac{1}{F}\sum_{i=1}^{F} d_i(T_i, T_{mean})$$

- 平均移動ベクトル・全移動ベクトル間の差のノルムに基づく尺度 D_a（以下，全移動ベクトル平均偏差）

$$D_a = \frac{1}{C}\sum_{j=1}^{C} d_j(T_j, T_{mean})$$

ここで，T_i は選択話者コードブックにおいて学習で対応づけられたコードに対応する移動ベクトル，F は選択話者の学習音声フレーム数，T_j は j 番目のコードに対する移動ベクトル，C はコードブックのクラスタ数である．また，T_{mean} は平均移動ベクトル，d_iおよびd_j は平均移動ベクトルと各移動ベクトルとの差のノルムを表し，それぞれ次式で定義した．

$$d_k(T_k, T_{mean}) = \sqrt{\sum_{l=1}^{L}\{T_k(l) - T_{mean}(l)\}^2} \tag{1}$$

$$T_{mean} = \frac{1}{C}\sum_{n=1}^{C} T_n \tag{2}$$

ここで，L はベクトルの次数である．これらの定式化において，D_t は，学習音声空間のみに着目した尺度を表し，D_a は，話者の音響空間全域に着目した尺度を表している．

4　対象音響空間の違いによる影響

尺度の評価としては，それぞれの移動ベクトル平均偏差と写像精度との相関係数を用いることとし，尺度の定式化として着目した音響空間が，学習音声空間のみである場合と全空間である場合の違いを調べた．

4.1　実験手順

以下に実験手順を示す．

1. それぞれの話者に対する移動ベクトル平均偏差を求める
2. 変換音声を生成し，写像精度を数値化する
3. それぞれの移動ベクトル平均偏差と写像精度との相関係数を算出する

4.2　実験条件

本実験では，音声試料として ATR 音声データベース [11][12] を用いた．変換音声を 50 単語作成し，写像精度を目標話者音声と変換音声との 50 単語の平均ケプストラム距離で表した．この時，移動ベクトル平均偏差と平均ケプストラム距離との間で正の相関が強い程，より適切な話者選択尺度であることを表している．また，学習データ量の違いによる影響を同時に観察するため，学習音声を 1 単語，3 単語，5 単語とした場合について，それぞれ相関係数を求め

表 1: 学習データ [11]

1 単語	/uchiawase/
3 単語	/boNyari/+/uchiawase/+/dekigoto/
5 単語	/boNyari/+/uchiawase/+/dekigoto/ +/hyoujou/+/puroguramu/

表 2: 実験条件

音声試料	ATR 音声データベース
サンプリング周波数	12kHz
分析窓	ブラックマン窓
分析窓長	21.3ms
フレーム周期	5ms
目標話者	男女各 4 名
登録話者	目標話者以外の男女各 4 名
コードブック作成データ	音素バランス 503 文 [12]
クラスタ数	512
特徴量	30 次 FFT ケプストラム
VFS $k-$ 近傍数 [3]	4
VFS 平滑化重み係数 [3]	1.0

た．学習データは，母音・子音の種類が比較的多くなるように選択した．アナウンサまたはナレータである男女各 4 名を目標話者，別の男女各 4 名を登録話者とし，各目標話者に対する写像元話者は，従来尺度で選択された話者とした．なお，移動ベクトルの補間に使用する近傍数は 4 とし，平滑化する際に用いている重み係数の値は従来法 [3] と同じ 1.0 とした．表 1に実験に用いた学習データを，表 2に実験条件を示す．

4.3 実験結果

図 2に，実験結果を示す，1 単語学習，3 単語学習，5 単語学習いずれの場合においても D_t よりも D_a の方が正の相関が強いということが明らかとなり，移動ベクトルの向きのばらつきに着目した尺度の定式化を行う場合，学習される音響空間のみをカバーするよりも，音響空間全体をカバーした方が良いということが示された．なお，学習単語数が増えるに従って，D_t と D_a との間の相関係数の差が小さくなる傾向が観られるが，これは，学習データ量が増加することによって，学習過程で対応づけられるコード数がコードブックのクラスタ数に近づき，D_t が

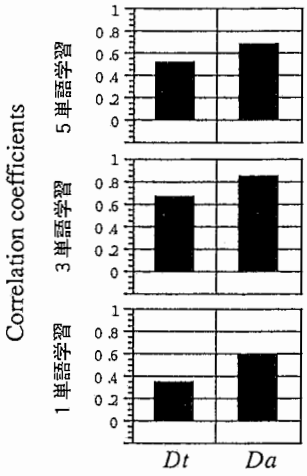

図 2: 提案尺度と写像精度（平均ケプストラム距離）との相関係数

音響空間全体を考慮する尺度に近づくためであると考えられる．次節では，更に多くの話者の組合せを用い，従来尺度との比較により提案尺度のうち D_a の有効性を述べる．

5　提案尺度の評価

前節の実験では，D_t よりも D_a の方が話者選択尺度として良いことが示された．しかし，前節の実験では，写像元話者を従来尺度（学習データのみのスペクトル距離）で選択された話者にしたことにより，1名の目標話者に対する写像元話者が固定され，話者の組合せが少なかった．尺度の有効性を評価するためには，より多くの話者の組合せで評価する必要がある．

そこで次に，準備した目標話者・登録話者のすべての組合せに対して，提案尺度 D_a を用いた場合の相関係数と，従来尺度（学習データのみのスペクトル距離）を用いた場合の相関係数とを比較することにより，提案尺度の有効性の評価を行った．

5.1　実験手順

以下に実験手順を示す．

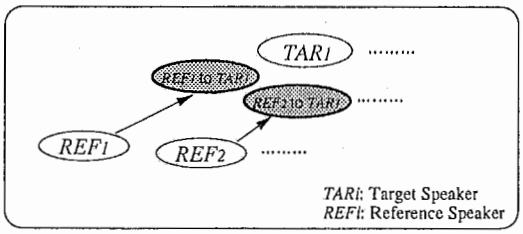

図 3: 実験に用いた話者の組合せ

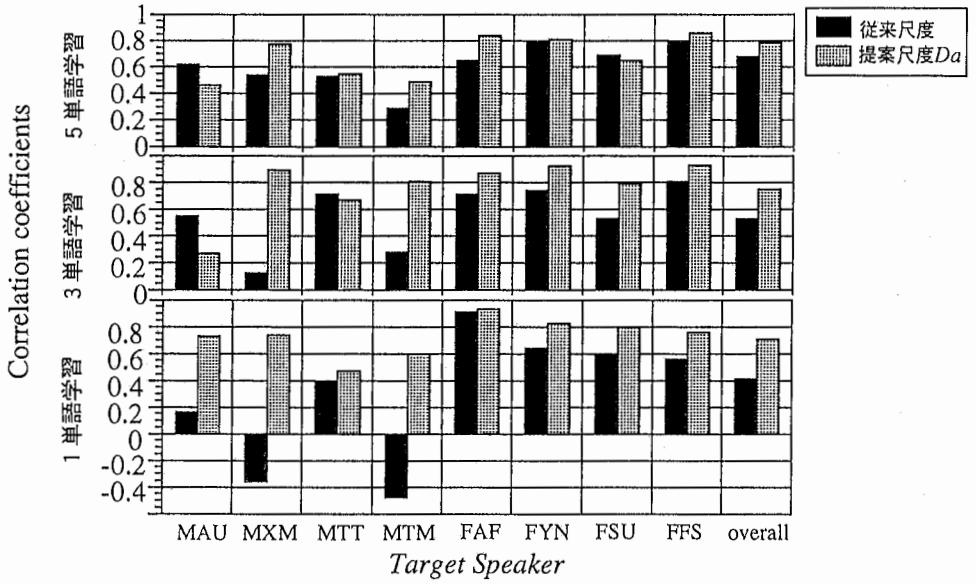

図 4: 従来尺度と提案尺度 D_a の写像精度（平均ケプストラム距離）に対する相関係数

1. 図 3に示すように，1名の目標話者に対して，各登録話者を写像元話者にしたときの尺度量（従来尺度と提案尺度 D_a）をそれぞれ求める

2. 変換音声を 50 単語生成し，写像精度を数値化する

3. 尺度量と写像精度との相関係数を算出する

4. すべての目標話者に対して（1）～（3）を繰り返す

5.2 実験条件

話者の組合せを除いて，前節の実験と同じ条件とした.

5.3 実験結果と考察

図 4に，実験結果を示す. 図 4では，目標話者を固定した状態での各登録話者からの変換精度と尺度量との相関，および全目標話者・全登録話者の組合せ（総組合せ 64）に対する変換精度と尺度量との相関を示しており，それぞれ左から，従来尺度・提案尺度 D_a の場合の平均ケプストラム距離との相関係数を表す.

この結果,

（1）従来尺度の場合は話者によって逆相関となるケースが存在するなど，話者による変動が大きかったが，提案尺度の場合は話者変動が大幅に改善されており相関も従来尺度より強いこと,

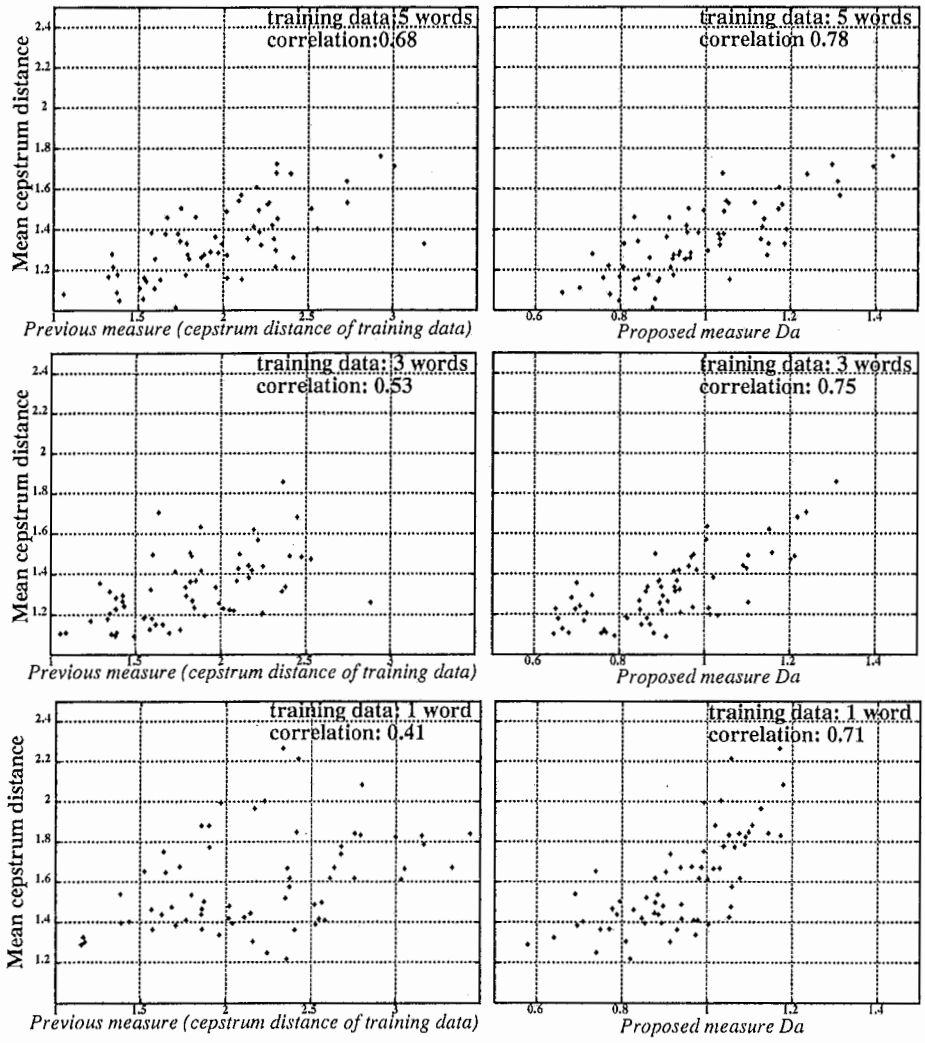

図 5: 尺度量と写像精度との関係

（2）従来尺度では，学習データの違いによる相関係数の変動が大きいが，提案尺度では，MAU を目標話者とした場合を除き，学習データの違いによる影響が比較的小さく，少ない学習データでも相関が強いこと

が示された.

これらから，平均移動ベクトル・全移動ベクトル間の差のノルムに基づいた提案尺度 D_a が，話者選択尺度として有効であることが明らかとなった．つまり，VFS は平均移動ベクトルからの各移動ベクトルの向きのばらつきが小さい話者間の写像に適した手法であると言える.

但し，相関係数の値だけでは，尺度量に対する写像精度の分布を知ることができない．そこで次に，提案尺度 D_a について，すべての話者の組合せに対する尺度量と写像精度をグラフ上にプロットし，実際の尺度量と写像精度との関係を調べた．比較のため，従来尺度の場合の分布もグラフ化した．図 5に結果を示す.

図 5は，下から，1単語学習，3単語学習，5単語学習の場合の分布を示したもので，左に

従来尺度，右に提案尺度 D_a の場合を示す．図5から，従来尺度では学習データ量によって相関が大きく変化して不安定であるが，提案尺度 D_a では学習データ量に依存せずに少ない学習データでも強い正の相関を示していることがわかる．従って，実際の尺度量と写像精度の分布の上でも提案尺度 D_a が話者選択尺度として有効であることが示された．特に，実用性を考慮して極力少ない学習データ量でスペクトル写像を実現することを目指した制約の上では，提案尺度 D_a を用いる効果が大きいと言える．

なお，本報告で提案した尺度 D_a によって選択された話者を観察すると，男性が目標話者であるとき，女性が選択された場合が12組合せ中7ケース存在していた．しかし，文献[4][13]などのように，スペクトル距離の大きな話者間の写像性能は良くないという報告もあることから，本報告での提案尺度 D_a と従来尺度であるスペクトル距離との和を最小化する方法や，木構造話者クラスタリング[14]の利用などによって更に精度を向上させることが可能であると考えられる．

6 むすび

少ない学習データで声質変換を実現することを目的としたスペクトル写像法SSVFSにおいて，任意の話者に対してロバストな写像を実現するための適切な写像元話者の選択法として，移動ベクトルの向きのばらつきに着目した尺度を提案し，学習で求まった全移動ベクトルの平均ベクトルと個々の移動ベクトルとの差のノルムに基づく定式化を行った．

学習データを1単語，3単語，5単語とした場合について，変換音声と目標話者音声との間の平均ケプストラム距離で表した写像精度と定式化した尺度の出力である尺度量との相関係数を求め，学習音声のみのケプストラム距離で定義した従来尺度と比較することにより評価を行った．その結果，

（1）従来尺度の場合は話者によって逆相関となるケースが存在するなど，話者による変動が大きかったが，提案尺度の場合は話者変動が大幅に改善されており相関も従来尺度より強いこと，

（2）従来尺度では，学習データの違いによる相関の変動が大きいが，提案尺度では，学習データの違いによる影響が比較的小さく，少ない学習データでも相関が強いこと

が明らかとなり，提案尺度が適切な写像元話者の選択に有効であることが示された．

今後は，本報告で提案した尺度 D_a を従来尺度や木構造話者クラスタリング手法[14]と融合するなどの方法により，音響的構造の類似性表現能力の向上を図り，高精度な声質変換法の確立を目指す予定である．

謝辞

研究の機会を与えて頂いた，ATR 音声翻訳通信研究所山崎泰弘社長に感謝致します．また，日頃から討論頂くニック・キャンベル主幹研究員ならびに ATR 諸氏に感謝致します．

参考文献

[1] 服部浩明，嵯峨山茂樹：“移動ベクトル場平滑化話者適応の原理とアルゴリズム”, 信学技報, SP92-15, 1992.

[2] K. Ohkura, M. Sugiyama and S. Sagayama : “Speaker Adaptation based on Transfer Vector Field Smoothing with Continuous Mixture Density HMMs”, Proc. ICSLP92, pp.369–372, 1992.

[3] M. Hashimoto and N. Higuchi : “Spectral Mapping for Voice Conversion Using Speaker Selection and Vector Field Smoothing”, Proc. EUROSPEECH'95, pp.431–434, Sept. 1995.

[4] 橋本 誠，樋口宜男：“話者選択と移動ベクトル場平滑化による声質変換のためのスペクトル写像”, 信学論, **J80-D-II**, No.1, 1997.

[5] M.Abe, S.Nakamura, K.Shikano and H.Kuwabara : “Voice conversion through vector quantization”, Proc. ICASSP'88, pp.565–568, 1988.

[6] N.Iwahashi and Y.Sagisaka : “Speech spectrum conversion based on speaker interpolation and multi-functional representation with weighting by radial basis function networks”, SPEECH COMMUNICATION, **16**, pp.139–151, 1995.

[7] W. Ding, H. Kasuya, and S. Adachi : “Simultaneous estimation of vocal tract and voice source parameters based on an ARX model”, IEICE Trans. Inf. & Syst., **E78-D**, no.6, pp.738–743, 1995.

[8] 橋本 誠，樋口宜男：“話者選択と VFS を用いたスペクトル写像のための学習データ決定法”, 音講論, pp.261–262, Sept. 1995.

[9] M. Hashimoto and N. Higuchi : “Training Data Selection for Voice Conversion Using Speaker Selection and Vector Field Smoothing”, Proc. ICSLP96, pp.1397–1400, Oct. 1996.

[10] 宮沢康永，嵯峨山茂樹：“移動ベクトル場平滑化話者適応方式における標準話者選択方式の検討”, 音講論, pp.121–122, Oct. 1992.

[11] 武田一哉，匂坂芳典，片桐 滋，阿部匡伸，桑原尚夫：“研究用日本語音声データベース利用解説書，”, Tech. report of ATR, TR-I-0028, May. 1988.

[12] 阿部匡伸，匂坂芳典，梅田哲夫，桑原尚夫：“研究用日本語音声データベース利用解説書（連続音声データ編）”, Tech. report of ATR, TR-I-0166, Sept. 1990.

[13] 松本 弘，丸山靖史，井上博夫："教師あり / 教師なしスペクトル写像による声質変換"，音響学会誌，**50**, No.7, pp.549–555, July 1994.

[14] 小坂哲夫，松永昭一，嵯峨山茂樹："木構造話者クラスタリングを用いた話者適応"，信学論（D-II），**J78-D-II**, no.1, pp.1–9, 1995.