TR-IT-0208

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と学習データの選択方法

Spectral Mapping Method and Designing of Training Data Selection for Voice Conversion using Speaker Selection and Vector Field Smoothing Techniques

橋本 誠 Makoto Hashimoto
樋口 宜男 Norio Higuchi

1997.3.27

データベース音声を少量学習データで入力話者 (目標話者) 音声) に変換する声質変換のためのスペクトル写像法を提案した. 本方式では, 話者選択で複数話者の音声データベースから入力話者に近い話者を選択し, 選択話者空間から入力話者空間へのスペクトル写像を移動ベクトル場平滑化法によって行う. 1単語 /*uchiawase*/ のみの学習で, 別の50単語で写像を行った結果, 目標話者音声とのケプストラム距離は平均約25%, 最大約41%減少し, 有効性が示された. 聴取実験による主観評価においても約66%の割合で変換音声が目標話者に近いと認められた. また, 最適学習データの設定方法についても検討し, 学習量を表す尺度を定式化した. 提案尺度は写像精度との相関が比較的強く適切な学習データの選択に利用できることが示された.

もくじ

1 はじめに

本報告では，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（音声翻訳機能付きテレコンファレンスシステム）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することを可能と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[1, 2]．阿部らは，大量の学習データを用いて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による方法を提案している．また，岩橋らは，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．しかし，前者では，学習データを大量に必要とするため，目的話者が不特定であるようなシステムに応用しにくく，後者では，写像関数の最適化処理が複雑である．さらに後者では，複数話者のスペクトルパラメータ上での補間を用いており，音声生成過程を考慮した場合，あまり好ましくない．

一方，音声認識の分野では，精度向上のために話者適応について種々の検討が行われてきた[3, 4, 5, 6, 7, 8]．杉山らは，入力音声から認識された母音特徴に近いものを，予め記憶している複数の5母音の特徴パターンから距離最小規準で選択し，選択された特徴に修正を加えて入力話者音声に適応化させる教師なし話者適応化法を提案している[3]．服部らは，異話者間の音響特徴空間差分ベクトルを話者から話者への写像関数と考える移動ベクトル場平滑化法（Vector Field Smoothing method，以下では，VFS と略記）を提案し[4, 5]，少ない学習データで良好な結果を得ている[4, 5, 6, 7, 8]．話者適応の手法を声質変換に応用した例もあるが，男女声変換などスペクトル差の大きな話者間の変換が難しいことが報告されており[9]，高品質の合成音声を得るためには，写像元となる話者と目標話者がある程度類似している必要がある．

また，声質変換を行う方法としては，音源や声道モデルパラメータ上での写像を試みることも考えられるが，この場合，音声波形情報から音源や声道モデルパラメータを抽出するという問題を扱うことになり，複雑な処理が必要となる[10]．

以上のような観点から，本研究では，実用性が高く，比較的処理量が少ないということに主眼を置き，少ない学習データで，音響パラメータ上での1話者対1話者の声質変換を行うことを目的とした方法を提案する．音響パラメータ上での声質変換のためには，スペクトル，基本周波数，音素継続時間などの写像が必要であるが[11, 12, 13]，中でもスペクトルの写像は，聴覚的にも大きな影響があり[13]多次元空間を扱う難しい問題であるため，特に重要である．そのため，本報告ではスペクトル変換のみを扱うこととし，話者選択と VFS を用いたスペクトル写像法（Speaker Selection and VFS，以下では，SSVFS と略記）を提案すると共に，生成される音声の評価を行う[14, 15]．

この方法では，複数登録話者から，目標話者にある程度スペクトル距離の近い話者を１名選択し，選択された話者から目標話者への写像を行う．本報告では，１単語程度の少ない学習データで写像が行えることを確認し，本方式の有効性について検討する．なお，本報告では，極力少ない音声データを用いて学習を完了することに主眼を置いたため，１単語分の音声を用いた場合を中心に結果を示しているが，実際のシステムにおいては逐次的に学習を進め，ある程度学習が進んだ時点で学習処理を完了する方式が有効であると考えられる．このため，学習の進み具合を表す評価尺度の定義が必要であり，これも合わせて検討した [16].

以下，スペクトル写像方式の概要と，学習データに１単語を用いた場合の評価実験結果について述べ，さらに，学習の進み具合を表す評価尺度の定式化と，この評価尺度を用いた学習データ設定法について述べる．

2　方式の概要

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，とよぶこととする．

本方式は，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）登録話者の中から選択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空間へ写像するための話者間移動ベクトル（以下，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ィベクトル量子化と，移動ベクトルを用いた目標話者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．

以下では，上記の各ステップについて説明する．

2.1　話者選択

このステップでは，各登録話者に対して，DTW で時間整合された目標話者の学習音声スペクトル時系列と登録話者の学習音声スペクトル時系列との距離を求め，２乗誤差最小規準により最も距離の小さい登録話者を１名選択する．

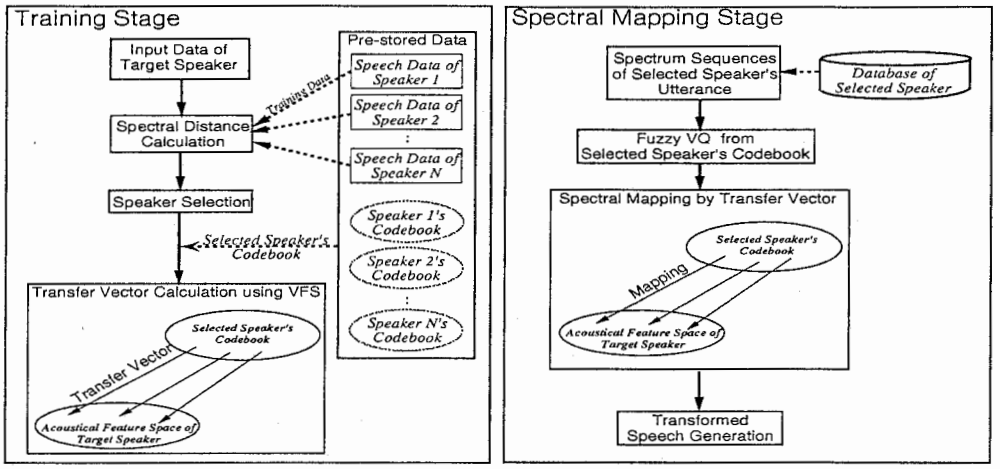

図 1: 本方式のブロック図

2.2 移動ベクトル計算

このステップでは，選択話者のスペクトルコードブック C^s を目標話者空間に写像してスペクトルを変換するための移動ベクトルを求め，これによって，目標話者空間に写像されたコードブックを得る．ここで，目標話者空間に写像されたコードブックを，写像コードブック C^t と定義する．C^t の生成には VFS を用いる [5, 6]．VFS は，音響空間の話者間の差のベクトルは連続的に変化するという仮定のもとに，ある話者の音響空間を他話者の音響空間に写像する手法である．この手法では，補間と平滑化処理により移動ベクトルを求める．以下に，本ステップのアルゴリズムを示す．

移動ベクトルの補間

(1) C^s を C^t の初期状態とする．

(2) 選択話者の学習音声スペクトル時系列を C^s でベクトル量子化する．

(3) 量子化されたベクトルと目標話者音声の入力スペクトル時系列とを DTW により対応付ける．

(4) m 番目のベクトル C^s_m と，これに対応付けられた入力スペクトル x の平均ベクトルとの差分ベクトル V_m を求める．これを移動ベクトルとする．

$$V_m = \frac{1}{N_m} \sum_{x \in M} x - C^s_m$$

ここで，N_m は C^s_m に対応付けられた入力スペクトルベクトルの個数，M は C^s_m に対応付けられた入力スペクトル時系列のベクトルの集合である．

(5) 学習で対応付けが行われなかったコードベクトル C^s_n に対する移動ベクトル V_n を，その近傍にある k 個の対応付けが行われたコードベクトル C^s_k に対する移動ベクトルの補間によって求め，その後，写像コードブックのすべてのベクトルを更新する．

$$V_n = \sum_{k \in K} \mu_{n,k} V_k$$

$$C^t = C^s + V$$

ここで，移動ベクトル V_k への重みである $\mu_{n,k}$ は，C^s_n と C^s_k から求められ，本報告では以下のファジー級関数を用いる．

$$\mu_{n,k} = 1/\{\sum_{j\in K}(d_{n,k}/d_{n,j})^{1/(f-1)}\}$$

ここで，$d_{n,k}$ は C^s_n と C^s_k との距離，f は $\mu_{n,k}$ の制御パラメータとなるファジネス，K は学習で対応付けが行われたベクトルの番号の集合である．

(6) 変換後のベクトル列と目標話者音声のスペクトル時系列との間の DTW 時の距離が収束するまで (3) から (5) を繰り返す．

移動ベクトルの平滑化

ここまでのステップでは，学習データが少ない場合に異話者間の真の対応関係を表せないため，特に学習で対応付けが行われなかった移動ベクトルについては，大きな誤差を含んでいると考えられる．そこで，移動ベクトルに連続性の拘束条件を入れ，以下に述べるように平滑化を行って，誤差を吸収させる．

(1) C^s_l とその k- 近傍にある C^s_k との間のファジー級関数 $\mu_{l,k}$ を求める．

(2) $\mu_{l,k}$ を用いて平滑化移動ベクトル V_l を求める．

$$V_l = \sum_k \mu_{l,k} w V_k / \sum_k \mu_{l,k} w$$

ここで，w は重み係数である．k=l のとき $\mu_{l,k} = 1$ とする．

(3) V_l を用いて写像コードブックのすべてのベクトルを更新する．

$$C^t_l = C^s_l + V_l$$

2.3 スペクトル写像

本ステップでは，選択話者コードブックを用いて，選択話者音声スペクトル時系列 X^s をファジィベクトル量子化する．このときも，移動ベクトル計算時と同様に，X^s_p (p：フレーム番号) と，その近傍にある k 個の C^s_q ($q = 1 \sim k$) との間のファジィ級関数 $\mu_{p,q}$ を求め，C^s_q に対応付けられた移動ベクトル V_q から求まる C^t_q と $\mu_{p,q}$ とにより，変換後ベクトル X^t_p を求める．この結果，選択話者から目標話者に写像された音声スペクトル時系列を得る．以下では，スペクトル写像後の音声を変換音声とよぶこととする．

3 評価実験条件

本方式によるスペクトル写像性能を客観尺度と主観尺度の両面から評価した．

表 1: 実験条件

音声試料	ATR 音声データベース
学習データ	1 単語
評価データ	50 単語
サンプリング周波数	12kHz
分析窓	ブラックマン窓
フレーム周期	5ms
FFT ポイント数	256
登録話者	男女各 4 名
目標話者	登録話者以外の男女各 4 名
コードブック作成データ	音素バランス 503 文 [18]
クラスタ数	512
特徴量	30 次 FFT ケプストラム
VFS $k-$ 近傍数	4
平滑化重み係数 w	1.0

本実験では，音声試料として ATR 音声データベース [17, 18] を用いた．学習データは 1 単語とし，母音種類が比較的多く，かつ無声化母音がなく，子音もすべて異なる */uchiawase/* を採用した．評価用には 50 語を使用した．アナウンサまたはナレータである男女各 4 名を登録話者とし，登録話者以外の男女各 4 名を目標話者とした．距離計算には次式を用いた．

$$D = \frac{1}{N_{fr}} \sum_{ij} (c_{ij}^r - c_{ij}^t)^2$$

ここで，c_{ij}^r は DTW 後の写像元話者の第 i フレーム j 次ケプストラム係数，c_{ij}^t は目標話者の第 i フレーム j 次ケプストラム係数，N_{fr} はフレーム数である．表 1に実験条件を示す．

4 評価実験結果と考察

4.1 ケプストラム距離による客観評価

本方式の基本性能を調べるため，変換音声と目標話者音声および選択話者音声と目標話者音声のケプストラム距離を比較した．

[実験1]

まず，ファジネスの値の違いによる写像性能の差を見るため，目標話者男女各 1 名（MAU, FAF）に対して，平滑化時のファジネスを 1.5 〜 5.0 の間で変化させて距離を求めた．補間時

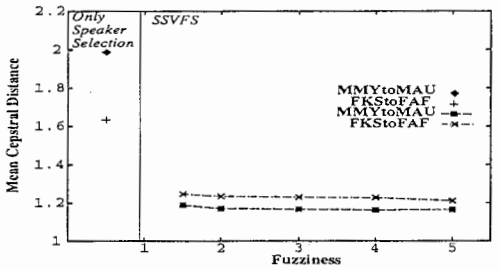

図 2: 変換音声と目標話者音声のケプストラム距離に対するファジネスの影響

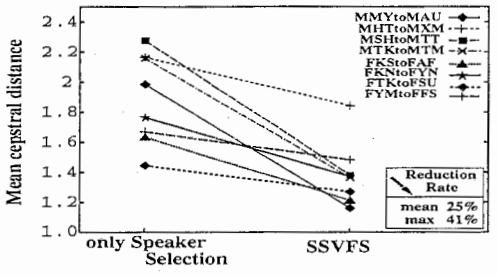

図 3: 話者選択のみを用いた場合と話者選択と VFS を用いた場合のケプストラム距離の比較

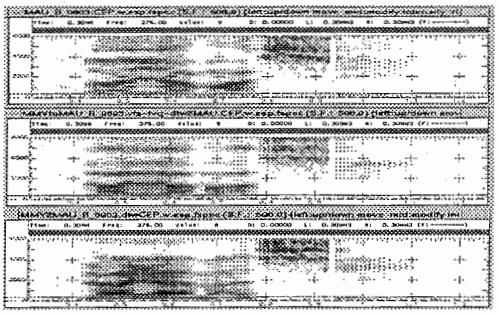

図 4: /urayamashii/ のスペクトログラム（上から目標話者音声，変換音声，選択話者音声）

のファジネスは平滑化時のファジネスと同じ値とし，スペクトル写像ステップでのファジネスは 1.5 に固定した．ケプストラム距離の 50 単語平均値を，図 2に示す．

この結果，各目標話者に対して，変換音声と目標話者音声との距離は，選択話者音声と目標話者音声との距離よりも平均で約 33% 小さくなり，本方式でスペクトルの目標話者への写像が行えることが確認できた．なお，ファジネスを変えることにより，変換音声と目標話者音声との距離を若干減少させることができるが，その減少幅は小さかった．以下の実験では，補間時，平滑化時のファジネスを 5.0，スペクトル写像時のファジネスを 1.5 とした．

[実験2]

次に，目標話者を 8 名（男性 4 名，女性 4 名）に増やして，写像性能を調べた．ケプストラム距離の 50 単語平均値を図 3に，スペクトログラムの一例を図 4に示す．

この結果，全組合せに対して，目標話者とのケプストラム距離が減少することが確認され，

本方式の有効性が示された．変換音声と目標話者音声との距離は，選択話者音声と目標話者音声との距離よりも平均で約25%，最大で約41%減少した．また，図4からも，スペクトルが目標話者に近付いていることが確認された．

4.2 聴取実験による主観評価

次に，聴覚的に本方式の効果があるかどうかを調べるため，男女各1名の目標話者（MAU, FAF）に対する聴取実験を，ABX法により行った．

[聴取実験手順]

A，Bには，それぞれ目標話者の分析合成音声，選択話者の分析合成音声，Xには変換音声を用いた．また，話者選択のみを用いた場合と比較するため，Xを選択話者の分析合成音声とするパターンも作成した．なお，順序効果の影響を抑えるため，半数についてはA，Bを逆にした．音声の呈示パターンを図5に示す．

ここでは，変換音声として，50単語のうちケプストラム距離の減少比が50単語平均値よりも小さかった単語，大きかった単語，同程度であった単語を，それぞれ話者毎に1サンプルずつ抽出したものを用いた．図5に示された4パターンを被験者にランダムに呈示して，「Xの話者が，A，Bどちらの話者に近いか」という規準で強制判定させた．被験者は6名で，呈示回数は1パターンあたり2回である．なお，スペクトル写像精度のみを評価するために，基本周波数，音素継続時間，パワーは目標話者と一致させた．音素継続時間はDTWにより一致させた．

[聴取実験結果]

聴取実験結果は，次式にしたがって判定率Rを求め，この値で比較した．

$$R = \frac{P_j}{P_{all}} \times 100\ [\%]$$

ここで，P_jは「'Xが目標話者に近い'と判定された回数」，P_{all}は「呈示回数」である．図6に結果を示す．

この結果，変換音声が目標話者に近いと判定された割合は，目標話者MAUに対しては約67%，FAFに対しては約65%であり，選択話者が目標話者に近いと判定された割合は，MAUに対して約18%，FAFに対しては約25%であった．

本評価手法では，話者選択で目標話者に近い話者が選択される程，選択話者と目標話者の識別が難しくなる．このため，目標話者および選択話者の分析合成音声を刺激Xとした場合の判定率を$(50 \pm a)\%$で表わすと，両者が十分離れている場合にはaはほぼ50と考えられるが，MMYとMAUおよびFKSとFAFの組合せではaはぞれぞれ約32および約25となっており，仮に理想的なスペクトル変換が行われたとしても，判定率は約82%および約75%にと

	A	B	X
Pattern 1	S	T	Tr
Pattern 2	T	S	Tr
Pattern 3	S	T	S
Pattern 4	T	S	S

S: resynthesized speech of the selected speaker
T: resynthesized speech of the target speaker
Tr: transformed speech

図 5: 聴取実験音声パターン

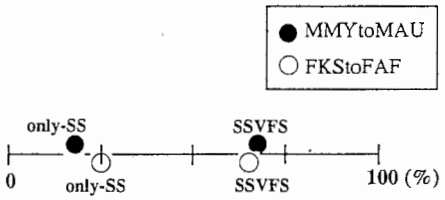

図 6: 聴取実験結果

どまるものと考えられる．この点を考慮すると，約 67% および約 65% という値は共にかなり高い割合と考えられ，聴覚的にも本方式が有効であることが示されたといえる．

以上から，本方式によって，1 単語程度の少ない学習データでも精度良くスペクトル写像を行うことが可能であり，実用的な声質変換法の実現が期待できるものと考えられる．

4.3　話者選択の効果

従来提案されている声質変換のためのスペクトル写像法では，前もって目標話者の声質に類似した話者を写像元話者として選択するということはしていない．音声認識における話者適応化では，話者選択を行っているものがあるが [3],[8]，音声認識率などによって評価されるのが通常であり，写像後の音響特徴を用いた音声の評価は行われない．

ここでは，本提案方式の特徴である話者選択の効果を調べるため，選択されなかった話者からの写像を行った場合の写像精度を求め，話者選択を行った場合と比較した．

[実験結果]

松本らの研究 [9] では男声女声変換を試みており，これと比較するため，2 名の女性話者を目標話者とした．目標話者に対して，各登録話者を写像元話者としてスペクトル写像を行い，写像精度を求めた．写像精度は，変換音声と目標話者音声とのケプストラム距離の 50 単語平均値である．結果を表 2に示す．

表 2から，話者選択を行った場合の方が，話者選択を行わない場合よりも変換音声と目標話者音声との平均ケプストラム距離が小さくなる傾向にあるという結果が得られ，話者選択の効果が示された．また，男声女声変換よりも女声女声変換の方が精度が良いことが示された．これらは，スペクトル差の大きな話者間の写像が容易ではないことを示しており，松本らの知見

表 2: 変換音声と目標話者音声とのケプストラム距離（50 単語平均値，＊印は選択話者）

変換音声 (写像元→目標)	目標話者との距離	変換音声 (写像元→目標)	目標話者との距離
FKS* → FAF	1.21	FKS → FYN	1.43
FKN → FAF	1.45	FKN* → FYN	1.37
FTK → FAF	1.40	FTK → FYN	1.32
FYM → FAF	1.31	FYM → FYN	1.32
MMY → FAF	1.57	MMY → FYN	1.55
MHT → FAF	1.84	MHT → FYN	1.84
MSH → FAF	1.59	MSH → FYN	1.58
MTK → FAF	1.66	MTK → FYN	1.53

[9] とも一致する．但し，目標話者 FYN に対しては，選択話者からの変換音声と目標話者音声との距離が最小ではない．これは，本報告で用いた距離最小規準以外にも考慮しなければならない要因があり得ることを示唆しており，今後，話者選択尺度の詳細な検討が必要であると考えられる．

5 学習量の評価尺度

SSVFS では，1 単語程度の学習データでもスペクトル写像が可能であることが確認されたが，その際，適切な学習データをどのようにして決定するかが重要となる．また，逐次的に学習を進めることを考慮した場合，その進み具合を定量的に把握する必要がある．そこで，次に筆者らは，学習量と写像精度に着目した適切な学習データ決定手法について検討を行った．

本報告では，適切な学習データを決定するために，ある学習データを使用した時の学習量を評価尺度として扱い，写像精度との関係を明らかにすることによって，評価尺度の妥当性を検証した．

5.1 学習量の定義

本方式では，VFS によって異話者間の写像関数となる移動ベクトルの学習を行う．図 7 に VFS の学習原理を示す．VFS では，まず学習時に目標話者音声と選択話者音声コードブックとの対応をとり，対応関係の見つかったコードに対して移動ベクトルが求められる．未学習コードに対する移動ベクトルは，近傍の学習済移動ベクトルから補間によって求められる．つまり，学習時に対応関係の求まったコードに対する移動ベクトルに存在する誤差は，未学習コードに対する移動ベクトルに存在する誤差よりも少ないと考えられる．このことから，学習時に対応

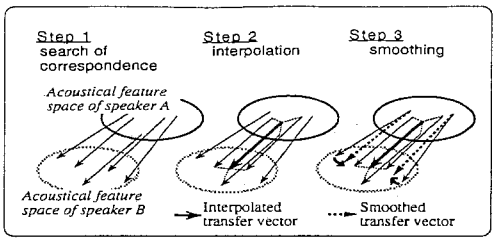

図 7: VFS の学習原理

関係の求まったコードについての学習量が基準となり，未学習のコードに対しては，近傍の学習済コードとの距離に応じた学習量が得られるものと考えられる．また，学習で対応付けられた目標話者音声のフレーム数が多いほど個々の学習済コードの信頼性が増すことによって学習量は大きくなると考えられる．

これらを考慮して，学習量 T は，以下に示すような関数で定義した．

$$T = \frac{1}{N_{cb}}\{\sum_i C_{trained}(i) + \sum_j C_{untrained}(j)\}$$

ここで，N_{cb} はコードブックサイズである．$C_{trained}$ は学習時に対応付けの行われたコードの学習量，$C_{untrained}$ は対応付けの行われなかった未学習コードの学習量であり，それぞれ次式で表す．

$$C_{trained}(i) = 1 - exp(-\alpha N_i)$$

$$C_{untrained}(j) = \max_k [exp(-\beta D_{j,k}) \times C_{trained}(k)]$$

ここで，N_i は i 番めのコードに対応付けられた目標話者音声のフレーム数，$D_{j,k}$ は j 番めのコードベクトルとその $k-$ 近傍にあるコードベクトルとのユークリッド距離，α，β は学習量 T を制御する重み係数である．

学習で対応付けられたコード群が音響空間内で集中している場合には，複数のベクトルからの学習効果にオーバーラップが生じ，結果的に全体としての学習量を抑制する効果が生じる．従って，本定式化により，音響空間内の広い範囲をカバーするような音声サンプルでは学習量 T が増加し，そのサンプルが学習に適していることがわかる．

なお，$\alpha = \infty$，$\beta = \infty$ のときは，信頼性や空間内位置の影響を全く考慮しない場合，つまり，学習で対応付けられたコード数の割合のみを学習量とすることと等価となる．

5.2　実験条件

本実験では，学習量 T の異なるデータでの比較を行うため，学習データとして表 3に示すようなモーラ数や母音種類，子音の有無が異なる 7 種類の単語を用い，写像精度として，学習に使用していない 50 語に対する，スペクトル写像後の音声と目標話者音声との平均距離 D を用いた．なお，/uchiawase/ を学習データとしたときの選択話者を，それぞれの写像元話者とした．

表 3: 学習データ

/ai/, /ou/, /megane/, /ukeau/
/wagamama/, /kewashii/, /uchiawase/

表 4: 学習量と変換音声・目標話者音声間距離との相関

写像元話者	目標話者	相関係数	
		$\alpha = 0.01, \beta = 50$	$\alpha = \infty, \beta = \infty$
MMY	MAU	-0.96	-0.98
MHT	MXM	-0.60	-0.27
MSH	MTT	-0.84	-0.67
MTK	MTM	-0.77	-0.75
FKS	FAF	-0.96	-0.56
FKN	FYN	-0.62	-0.48
FTK	FSU	-0.76	-0.89
FYM	FFS	-0.43	-0.14

5.3 実験結果と考察

提案した尺度の有効性を検証するため，各学習データの学習量 T と写像精度を求め，その相関を調べた．比較のために，学習で対応付けられたコード数の割合のみを学習量とした場合（$\alpha = \infty$, $\beta = \infty$ と等価）の相関も求めた．

重み係数 α, β の値を変化させて実験を行った結果，$\alpha = 0.01$, $\beta = 50$ のときが，相関の強い組合せが多いことが明らかになった．相関係数の値を表 4に，平均ケプストラム距離と学習量 T との関係を図 8に示す．

この結果，

- 未学習コードの学習量や学習コードの信頼性などを考慮した方が相関が強くなる．
- 大部分の話者の組合せで，学習量 T と写像精度との間に強い相関がある．

という現象が見られ，提案した手法により，学習量 T から高い写像精度を実現する適切な学習データの決定が可能であることが示された．しかし，相関があまり強くない話者の組合せも見られたことから，α, β の最適化や学習量の定式化についての詳細な検討および話者の組合せによる影響の分析などが必要であると考えられる．

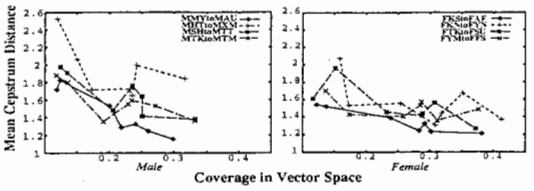

図 8: 学習量 T と変換音声・目標話者音声間のケプストラム距離との関係（$\alpha = 0.01$, $\beta = 50$）

6　むすび

本報告では，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，話者選択と移動ベクトル場平滑化によるスペクトル写像法（SSVFS）を提案した．本方式の有効性を評価するため，スペクトル距離による客観評価および ABX 聴取実験による主観評価を行った．ATR 音声データベース音素バランス単語中の 1 単語（/*uchiawase*/）のみを学習に用い，50 単語で評価を行った結果，変換音声と目標話者音声とのスペクトル距離は，選択話者音声と目標話者音声との距離より平均で約 25%，最大で約 41% 小さくなり，また，聴取実験でも良好な結果が得られ，本方式の有効性が示された．

さらに，効率よい写像を実現するための適切な学習データ決定手法として，学習量に着目した評価尺度を提案し，未学習コードに対する VFS の影響，および学習コードの信頼性を考慮することにより，間接的に学習コード群の空間内位置の影響をも考慮した学習量の定式化を行った．本手法の有効性および定式化した学習量 T の妥当性を評価するため，学習データを複数用意し，各学習データを使用したときの 50 単語に対する目標話者音声とのスペクトル距離と学習量 T との関係を調べた．その結果，大部分の話者の組合せで，本報告で定式化した学習量 T とスペクトル距離との間に強い相関があることが示され，提案手法で適切な学習データの決定が可能であることが確認できた．しかし，相関があまり強くない話者の組合せも存在したため，学習量の詳細な検討や話者の相性について調べる必要があると考えられる．

今後は，目標話者との相性のよい最適な写像元話者の選択法，最適な登録話者群の設計法，基本周波数など他の音響特徴パラメータの写像方法などについて検討する予定である．

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および ATR 諸氏，VFS のプログラムについて御指導頂いた大倉計美氏（現三洋電機（株））に感謝致します．

参考文献

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

[2] 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.

[3] 杉山雅英, 好田正紀, "認識結果を利用した母音標準パターンの教師なし話者適応化法, " 信学論, **vol.J69-D**, No.8, pp.1197-1204, 1986.

[4] 服部浩明, 嵯峨山茂樹 : "少量学習データを用いたコードブックマッピングによる話者適応化", 音講論, pp.49–50, Mar. 1991.

[5] 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.

[6] 大倉計美, 杉山雅英, 嵯峨山茂樹 : "混合連続分布 HMM を用いた移動ベクトル場平滑化話者適応方式", 信学技報, SP92-16, pp.23–28, Jun. 1992.

[7] 小坂哲夫, 鷹見淳一, 嵯峨山茂樹 : "話者混合逐次状態分割法による不特定話者音声認識と話者適応", 信学論 (A) , **J77-A**, pp.103–111, Feb. 1994.

[8] 外村政啓, 小坂哲夫, 松永昭一, "最大事後確率推定法を用いた移動ベクトル場平滑化話者適応方式, " 音響学会講論集, pp.77-78, Oct. 1994.

[9] 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.

[10] 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.

[11] 橋本 誠, 樋口宜男 : "個人性の知覚に影響を及ぼす音響的特徴の分析", 音講論, pp.323–324, May. 1995.

[12] N. Higuchi and M. Hashimoto : "Analysis of acoustic features affecting speaker identification", Proc. EUROSPEECH'95, pp.435–438, Sept. 1995.

[13] 伊藤憲三, 斉藤収三 : "音声の音響的特徴パラメータが個人性の知覚に及ぼす影響", 信学論, **J65-A**, pp.101–108, 1982.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.