Vol.2015-MUS-107 No.73
2015/5/24

声真似音声を利用した低コストで付加価値の高い音声合成の検討

本郷 康貴[1,a)]　能勢 隆[1]　伊藤 彰則[1]

概要：芸能人などの知名度が高い話者の音声を音声合成アプリケーションに利用できれば高い付加価値を持つと考えられるが，波形接続合成方式，HMM 音声合成方式では収録のコスト面が問題となる．そこで話者適応技術を用いることで収録コスト面の問題に対処できると考えられる．これまでの話者適応を用いた音声合成では平均声から適応を行っていたが，これに対し平均声に比べ目標話者に近い声質を持つと考えられる声真似話者から適応を行うことで，より適応の精度が向上すると考えられる．本研究ではこの手法の有効性を検証するため，実際に設定した目標話者に対して声真似した音声を収録し，それを用いて適応を行った合成音声の客観評価実験と主観評価実験を行った．本稿では以上の結果と，それを踏まえた考察を報告する．

1. はじめに

近年，情報案内やテキスト自動読み上げなどの音声合成アプリケーションは，情報端末の普及等により益々一般に浸透している．ここでアイドルや俳優など，知名度の高い話者の音声を音声合成アプリケーションに利用できれば，商業的に高い価値を持つと考えられるが，その際にはコストが問題となる．現在主流の音声合成方式には，録音した音声の素片を接続する波形接続音声合成や，音声のスペクトルなどを隠れマルコフモデル（HMM）によりモデル化する HMM 音声合成 [1] などがある．十分な品質の合成音声を得るためには，波形接続合成では数時間〜数十時間，HMM 音声合成では数十分〜数時間の大量の音声データが必要となる．知名度の高い話者の大量の音声データを収録するには高いコストがかかると考えられるため，音声合成アプリケーションに利用する際に問題となる．

HMM 音声合成では，HMM のパラメータを変換することで，容易に合成音声の声質を変換することが可能である．特に MLLR[2] や MAP[3] などの話者適応手法では，他の話者の HMM パラメータを目標話者の少量の音声データで目標話者の声質・韻律に近づけることにより，任意の話者に近い声を合成できることが示されている．話者適応を利用すれば，知名度の高い話者の音声データを少量用意するだけでその話者の声を合成できるため，収録コストを抑えることができる．

これまでの話者適応を用いた HMM 音声合成の先行研究では，適応を行う前の話者モデルとして，複数話者の音声から HMM の学習を行った平均声モデルを用いているものが多い [4]．適応を行う際に，ある特定の話者の音声から学習した HMM を初期モデルとして利用すると，初期モデルの話者と目標話者の組み合わせによって適応の精度が変化すると考えられる．この影響を低減するために，複数の話者の平均的な声質を持つ平均声から適応を行うことで，どの話者に適応をしても安定した性能で適応されることが示されている．

2. 提案手法

平均声モデルから適応する方法に対し，初期話者と目標話者の組み合わせが適応の精度に影響するならば，初期モデルを学習する話者として平均声よりも目標話者との声の特徴が似ている話者を選び，その話者モデルから適応を行うことで精度が向上すると考えられる．本研究では，話者適応の精度向上を目的として，目標話者の声真似をしてもらった音声で学習した声真似話者モデルから適応を行う手法を提案する．

以下では，提案手法の有効性を検討するために，実際に設定した目標話者に対して声真似した音声を収録し，それを用いていくつかのパターンで適応した合成音声の客観評価実験と主観評価実験を行い，その結果を分析する．

[1] 東北大学 大学院工学研究科
Graduate School of Engineering, Tohoku University
[a)] kouki.hongou.p4@dc.tohoku.ac.jp

表 1: 収録条件

話者	女性話者 1 名 （目標話者の声真似をして発話）
ファイル フォーマット	リニア PCM サンプリング周波数 16kHz 16bit 量子化
文章数	計 3003 文 セリフ調 2500 文，ATR 503 文
モーラ数	49435 モーラ セリフ調 33896 モーラ，ATR 15539 モーラ
収録時間	157 分 58 秒 セリフ調 114 分 53 秒，ATR 43 分 5 秒

3. 声真似音声の収録

3.1 目標話者の設定

有名人や芸能人といった知名度の高い人物の声や，恋人の声，亡くなった方の声等には，「この人の声を聞いてみたい」と思わせる付加価値があると考えられる．本研究では，知名度の高いアニメのキャラクターの声にもこのような付加価値があると考え，アニメ「新世紀エヴァンゲリオン」に登場する綾波レイ（CV：林原めぐみ）を合成の目標話者とした．アニメのセリフ等 10 文を発話した音声データを適応に用いるデータとして用意した．

3.2 音声収録

音声収録には，目標話者とは異なる女性話者 1 名に参加してもらった．この 1 名は，3 名の候補者を用意し，最も声真似が似ている話者を複数被験者に主観的に評価してもらい選出した．収録は防音室内で実施した．まず，綾波レイのセリフ調の文章 2500 文を事前に作成した．次に，収録対象の女性話者に，目標話者の声の特徴に似せて発話をしてもらい，その文章の収録を行った．また，ATR 音韻バランス文 [5]503 文についても収録を行った．2 種類の音声の収録を行ったのは，セリフ調の文を会話や音声対話，朗読調の ATR 文を情報案内やナレーションなど，幅広い発話様式の合成に利用することを目的としている．またセリフ調の文は音素バランスを考慮していないため，ATR 文で音素バランスを補完するという目的もある．収録条件を表 1 に示す．収録した声真似音声が実際に目標話者に似ているかについては，以降の実験で確認する．

4. 実験

4.1 実験条件

本研究では，HTS [6] を用いて音声の学習と合成をおこなった．HMM の学習データは無音を含む 31 種類の音素を単位とし，コンテクスト情報の含まれるラベルを作成して用いた．音声特徴量の抽出には STRAIGHT [7] を用いた．サンプルレート 16kHz の音声データを，フレーム周期 5ms でメルケプストラム分析し，0 次から 39 次のメルケプストラムを求めた．ピッチは対数基本周波数を特徴パラメータとした．また帯域非周期成分も特徴パラメータとして用いた．これらのパラメータに，各々の Δ，Δ^2 パラメータを加えた 138 次元のベクトルを特徴ベクトルとし，モデルは対角共分散単一ガウス分布を持つ 5 状態の left-to-right HSMM [8] を用いた．

4.2 目標話者への適応実験

提案法の有効性を確認するため，表 2 に示すように従来法と提案手法，また平均声から一度声真似話者に適応し，さらに目標話者に適応する方法，またそれらの各初期モデル計 6 パターンのモデルについて合成を行い，合成音声の自然性と目標話者との類似度を評価した．ただし，AVR は平均声，MIM は声真似話者，T は目標話者を表す．

一般的に適応データ量は初期モデルを学習するデータ量よりも少量であるため，複数の状態の分布で回帰行列を共有することで適応データに存在しない状態についても適応を行う．平均声から直接目標話者に適応するよりも，比較的データ量が多い声真似話者に一度適応を行うことで回帰行列の共有が減り，より適応の精度が向上すると考えられる．そのため従来法，提案法に加え平均声から一度声真似話者に適応し，さらに目標話者に適応する方法も検討した．

表 3 に学習条件を示す．声真似話者 HMM は 3 節で収録したコーパスのうち ATR 文から 100 文，セリフ調の文から 200 文計 300 文を学習データとして作成した．平均声HMMは，ATR データベース B セットに含まれる女性話者 4 名，各 450 文を学習データとして作成した．学習の際は山岸らの手法 [9] に倣い共有決定木コンテキストクラスタリングと話者適応学習を行っている．話者適応は CSMAPLR[10] によるアフィン変換と MAP 推定法によりスペクトルと F0 について適応を行い，適応モデルの状態継続長分布は声真似話者モデルの分布をそのまま用いている．これは状態継続長分布も目標話者に適応したところ，目標話者との RMS 誤差が悪化したためである．また声真似話者の分布を用いたのは，初期話者のうち目標話者との継続長の RMS 誤差が最も小さかったためである．ただし平均声から声真似話者に適応する場合のみ状態継続長分布も適応を行った．声真似話者に適応する際には前述の 300 文を適応データとした．目標話者に適応する際には，適応データをなるべく多く確保するために，目標話者の音声データ 10 文のうち 9 文を適応データとし，残り 1 文を合成する文章として，10-fold クロスバリデーションにより実験を行った．

4.2.1 客観評価結果

各モデルの合成音声と目標話者との類似度の客観評価を行った．評価尺度として目標話者の音声と合成音声の平均

表 2: 実験で比較した話者モデル

モデル	詳細	状態継続長
AVR	平均声（女性話者 4 名）	–
MIM	声真似話者の特定話者モデル（SD）	–
AVR-MIM	平均声から声真似話者に適応したモデル	適応有り
AVR-T	平均声から目標話者に適応したモデル	適応無し
MIM-T	声真似話者から目標話者に適応したモデル	適応無し
AVR-MIM-T	平均声から声真似話者に適応し，さらに目標話者に適応したモデル	適応無し

表 3: 学習条件

学習データ	声真似話者…100，300，2000 文 平均声…女性話者 4 名 各 450 文 計 1800 文
適応データ	声真似話者…100，300，2000 文 目標話者…10 文

表 4: 客観評価結果（声真似音声 300 文）

モデル	Spec.[dB]	log F0 [cent]	Dur.[msec]
AVR	7.16	384.9	33.6
MIM	5.99	232.7	25.8
AVR-MIM	5.86	195.6	26.6
AVR-T	5.65	181.5	25.8
MIM-T	5.62	202.3	25.8
AVR-MIM-T	5.64	185.8	25.8

メルケプストラム距離，対数 F0 と音素継続長の RMS 誤差を算出した．結果を表 4 に示す．平均声（AVR）と声真似話者（MIM）を比較すると，MIM はどの指標においても AVR よりも小さい値となっており，声真似話者は平均声よりも目標話者に近い声をしていると言える．AVR-MIM では AVR から MIM に適応することで MIM からの改善が見られた．平均声から適応を行ったことでモデルの品質が向上したためであると考えられる．目標話者への適応モデルを比べると，全体的に適応の効果が見られた．メルケプストラムは MIM-T が最も良い値となっており，提案手法の効果が表れたと言える．対数 F0 は AVR-T が最も良い値となった．声真似音声が 300 文の場合は，声真似話者モデルのイントネーションが不安定になる部分があることが理由として考えられる．

4.2.2 主観評価実験による自然性の評価

主観評価実験により，各モデルから合成した音声の自然性を評価した．被験者は 10 名で，防音室内でヘッドホンによる両耳聴取により実験を行った．刺激には 6 モデル 10 文，計 60 文の音声データを用いた．刺激は順番無作為に提示し，被験者は表 5 の 5 段階の評価値と評価語により評価した．

図 1 に自然性の主観評価結果を示す．横軸は評価スコアの平均値を表し，エラーバーは標準偏差を表す．Bonferroni 補正による多重比較検定を行った結果を表 6 に示す．初期モデルのうちでは，MIM は AVR に比べ自然性が高いと評価された．平均声モデルの学習データは読み上げ調で，評価文はセリフ調であったことが合成音声の自然性に影響したと考えられる．また声真似話者の学習データにはセリフ調の文が含まれていたため，平均声に比べ学習データ量が少ないものの良い結果が得られたと考えられる．目標話者への適応モデルのうちでは，MIM-T が AVR-T，AVR-MIM-T より良い結果となった．初期モデルの自然性の高さが反映されたと言える．

表 5: 評価値と評価語

評価値	評価語
1	bad
2	poor
3	fair
4	good
5	excellent

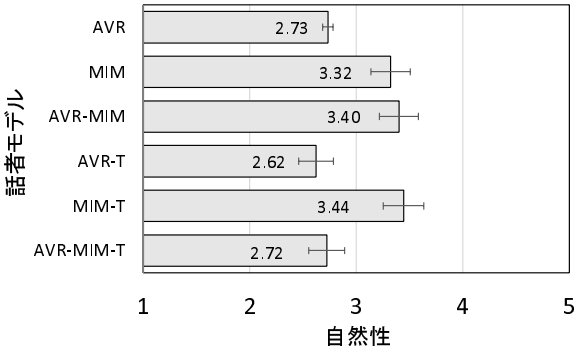

図 1: 自然性の主観評価結果（声真似音声 300 文）

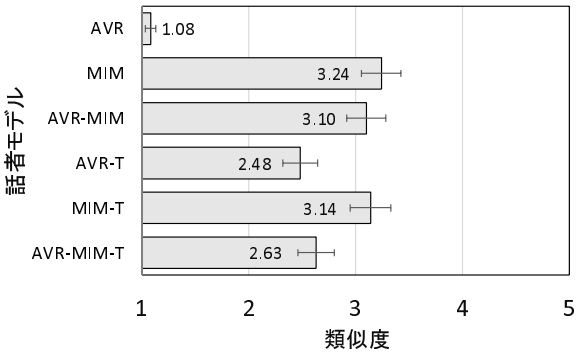

図 2: 類似度の主観評価結果（声真似音声 300 文）

表 6: Bonferroni 補正による自然性の多重比較検定の結果（p 値）（声真似音声 300 文）

	MIM	AVR-MIM	AVR-T	MIM-T	AVR-MIM-T
AVR	< 0.001	< 0.001	0.414	< 0.001	0.935
MIM		0.568	< 0.001	0.369	< 0.001
AVR-MIM			< 0.001	0.774	< 0.001
AVR-T				< 0.001	0.462
MIM-T					< 0.001

表 7: Bonferroni 補正による類似度の多重比較検定の結果（p 値）（声真似音声 300 文）

	MIM	AVR-MIM	AVR-T	MIM-T	AVR-MIM-T
AVR	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001
MIM		0.185	< 0.001	0.343	< 0.001
AVR-MIM			< 0.001	0.705	< 0.001
AVR-T				< 0.001	0.156
MIM-T					< 0.001

4.2.3 主観評価実験による類似度の評価

主観評価実験により，各モデルから合成した音声と目標話者との類似度を評価した．自然性の評価と同様の条件，刺激を用い，各刺激の提示前に目標話者の分析合成音を提示して，評価者はその音声との話者性の類似度を表 5 の 5 段階の評価語と評価値により評価した．

図 2 に目標話者との類似度の主観評価結果を示す．横軸は評価スコアの平均値を表し，エラーバーは標準偏差を表す．Bonferroni 補正による多重比較検定を行った結果を表 7 に示す．初期モデルでは MIM と AVR-MIM が同程度に類似度が高いと評価された．適応モデルでは MIM-T が最も類似度が高いと評価されたものの，MIM と同程度の評価となった．AVR-T と比べると改善が見られる結果となった．客観評価と主観評価を通して，声真似話者から適応することにより，平均声から適応を行うよりもより目標話者に近い合成音声が得られることが示された．

4.2 節の実験から，

- 収録した声真似音声は平均声よりも目標話者に近い声質および韻律を持つ．
- 声真似音声から適応することで，平均声から適応するよりも適応の精度が向上する．
- 目標話者と同じ発話スタイルの音声を学習・適応に利用することで，合成音声の品質が向上する．

以上のことが確認された．

4.3 声真似話者のデータ量の違いによる性能の評価

前節までの実験では，声真似話者モデルの学習データ量を 300 文としていたが，声真似話者モデルの学習データ量の違いは適応の精度や合成音声の自然性に影響すると考えられる．そこで 100 文に減らした場合と 2000 文に増やした場合に，合成音声の自然性や目標話者との類似度がどのように変化するかを評価した．比較するモデルは表 2 に示す 6 モデルである．

声真似話者モデルの学習データ量，声真似話者に適応する際の適応データ量をそれぞれ 100 文，2000 文と変えて各モデルを学習した．100 文の場合は音素バランスを重視し ATR 音韻バランス文を用い，2000 文の場合は 3 節のセリフ調のデータを用いた．その他の条件については 4.2 に倣った．

表 8: 客観評価結果（声真似音声 100，2000 文）

文章数	モデル	Spec.[dB]	log F0 [cent]	Dur.[ms]
100	AVR	7.16	384.9	33.6
	MIM	6.17	217.6	26.1
	AVR-MIM	6.01	206.1	29.6
	AVR-T	5.70	192.5	26.1
	MIM-T	5.75	199.6	26.1
	AVR-MIM-T	5.68	197.8	26.1
2000	AVR	7.16	384.9	33.6
	MIM	5.84	212.4	23.7
	AVR-MIM	5.84	212.6	25.7
	AVR-T	5.39	175.1	23.7
	MIM-T	5.08	152.4	23.7
	AVR-MIM-T	5.38	170.8	23.7

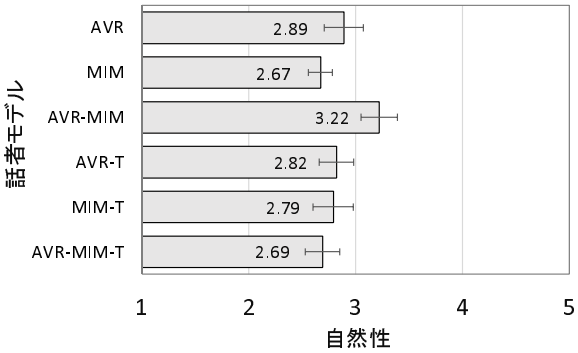

図 3: 自然性の主観評価結果（声真似音声 100 文）

4.3.1 客観評価結果

各モデルの合成音声と目標話者との類似度の客観評価を行った．評価尺度として目標話者の音声と合成音声の平均メルケプストラム距離，対数 F0 の RMS 誤差を算出した．

結果を表 8 に示す．100 文の場合，初期モデルではメルケプストラム，対数 F0 いずれも AVR-MIM が MIM よりも良い値になっている．また適応モデルではメルケプストラム，対数 F0 いずれも MIM-T は他の 2 つよりも悪い結果になった．これより声真似話者の学習データ量が少ない時には，平均声から一度声真似話者に適応し，さらに目標話者に適応する方法が良いことが示唆された．2000 文まで増やすと，初期モデルでは MIM のメルケプストラムと対数 F0 の値が AVR-MIM と同等になり，適応モデルでは MIM-T が AVR-T，AVR-MIM-T より良い値となった．これより声真似話者の学習データ量が増えるに従い，声真似話者モデルからの適応の精度は向上していくと言える．

4.3.2 主観評価実験による自然性の評価

主観評価実験により，各モデルから合成した音声の自然性を評価した．声真似音声 100 文の場合，2000 文の場合，4.2 節の 300 文の場合の実験は各々別日に分けて実施した．被験者は実施日ごとに異なる 10 名で，防音室内でヘッドホンによる両耳聴取により実験を行った．刺激には各場合で 6 モデル 10 文，計 60 文の音声データを用いた．刺激は

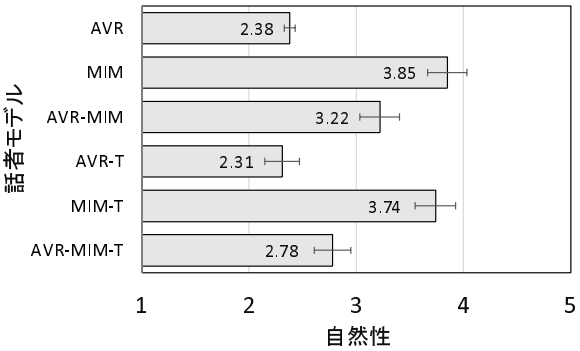

図 4: 自然性の主観評価結果（声真似音声 2000 文）

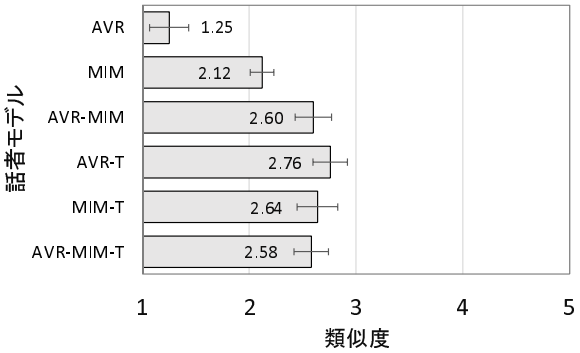

図 5: 類似度の主観評価結果（声真似音声 100 文）

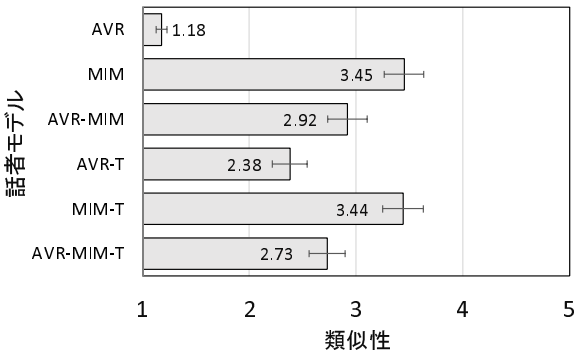

図 6: 類似度の主観評価結果（声真似音声 2000 文）

順番無作為に提示し，被験者は表 5 の 5 段階の評価値と評価語により合成音声の自然性を評価した．結果を図 3，図 4 に示す．横軸は評価スコアの平均値を表し，エラーバーは標準偏差を表す．100 文の場合と 2000 文の場合の各モデルについて Bonferroni 補正による多重比較検定を行った結果は表 9 に示す．100 文の場合，MIM は AVR よりも値が低く，AVR-MIM が最も自然性が高いと評価された．声真似話者の学習データ量が 100 文と少ない場合は，学習データ量が多い平均声から適応を行うことで，品質の高い声真似話者モデルが得られると考えられる．一方で適応モデルの値は有意差は無く，ほぼ横並びとなっている．2000 文に増えると，初期モデルでは MIM の値が他の 2 つを追い越し，適応モデルでも MIM-T が最も良い値となった．2000 文では声真似話者の特定話者モデルが十分に精度良く学習されていると考えられる．

表 9: Bonferroni 補正による自然性の多重比較検定の結果（p 値）

(a) 100 文

	MIM	AVR-MIM	AVR-T	MIM-T	AVR-MIM-T
AVR	0.124	0.021	0.624	0.484	0.162
MIM		< 0.001	0.294	0.401	0.889
AVR-MIM			0.005	0.002	< 0.001
AVR-T				0.834	0.363
MIM-T					0.484

(b) 2000 文

	MIM	AVR-MIM	AVR-T	MIM-T	AVR-MIM-T
AVR	< 0.001	< 0.001	0.556	< 0.001	<0.001
MIM		< 0.001	< 0.001	0.354	< 0.001
AVR-MIM			< 0.001	< 0.001	< 0.001
AVR-T				< 0.001	< 0.001
MIM-T					< 0.001

表 10: Bonferroni 補正による類似性の多重比較検定の結果（p 値）

(a) 100 文

	MIM	AVR-MIM	AVR-T	MIM-T	AVR-MIM-T
AVR	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001
MIM		< 0.001	< 0.001	< 0.001	< 0.001
AVR-MIM			0.170	0.731	0.864
AVR-T				0.303	0.123
MIM-T					0.607

(b) 2000 文

	MIM	AVR-MIM	AVR-T	MIM-T	AVR-MIM-T
AVR	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001
MIM		< 0.001	< 0.001	0.926	< 0.001
AVR-MIM			< 0.001	< 0.001	0.766
AVR-T				< 0.001	< 0.001
MIM-T					< 0.001

4.3.3 主観評価実験による類似性の評価

主観評価実験により，各モデルから合成した音声と目標話者との類似度を評価した．自然性の評価と同様の条件，刺激を用い，各刺激の提示前に目標話者の分析合成音を提示して，評価者はその音声との話者性の類似度を表 5 の 5 段階の評価語と評価値により評価した．結果を図 5，図 6 に示す．横軸は評価スコアの平均値を表し，エラーバーは標準偏差を表す．100 文の場合と 2000 文の場合の各モデルについて Bonferroni 補正による多重比較検定を行った結果は表 10 に示す．100 文の場合，初期モデルでは MIM は AVR-MIM よりも値が小さくなっている．これは自然性の低さが類似度の評価にも影響を与えたと考えられる．一方で適応モデルの類似度はほぼ横ばいとなった．2000 文に増えるに従い，MIM と MIM-T は値が向上した．声真似話者モデルの学習データ量が増えるほど，初期モデルとしての性能が向上していくと言える．

4.3 節の実験から，

- 声真似音声のデータ量が少ない場合には提案手法の効果は小さく，平均声から一度声真似話者に適応したモデルを初期モデルとする方法をとる必要がある．
- 声真似音声のデータ量が十分に多い場合には提案手法の効果は大きく現れる．

以上のことが確認された．

5. まとめと今後の課題

本稿では，話者適応を用いたHMM音声合成により収録コストが高い話者の音声の合成を低コストで実現することを目的とし，特に話者適応の精度を向上させるため，声真似音声を利用する手法を検討した．提案法では，従来話者適応を用いたHMM音声合成で適応前の初期モデルとして利用されていた平均声モデルに対し，目標話者に近い声の特徴を持つ声真似話者音声から学習したモデルを初期モデルとして適応を行うことで，より目標話者に近い声を合成することが可能である．実験では，提案法と従来の平均声を用いる方法，また平均声モデルから一度声真似話者モデルに適応したものを初期モデルとする方法を客観評価と主観評価により比較し，また声真似話者の学習データ量を増減させた場合のモデルの性能評価も行った．声真似話者の学習データ量が十分な場合には提案法の有効性が示された．その一方で，声真似話者の学習データ量が少ない場合には，従来法からの改善が見られなかったため，平均声から一度声真似話者に適応する方法をとるなどの対策が必要である．今後の課題としては，目標話者の音声が増減した場合の検討などが挙げられる．

参考文献

[1] 益子貴史，徳田恵一，山田哲也，小林隆夫，今井　聖：HMMを用いた音声合成法に関する検討，日本音響学会研究発表会講演論文集，Vol. 1995, No. 2, pp. 253–254 (1995).
[2] C. J. Leggetter and P. C. Woodland: Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models, *Computer Speech & Language*, Vol. 9, No. 2, pp. 171–185 (1995).
[3] J. L. Gauvain and C. H. Lee: Speaker adaptation based on map estimation of hmm parameters, *ICASSP 93*, Vol. 2, pp. 558–561 (1993).
[4] 田村正統，益子貴史，徳田恵一，小林隆夫：HMMに基づく音声合成におけるピッチ・スペクトルの話者適応，電子情報通信学会論文誌，Vol. J85-D2, No. 4, pp. 545–553 (2002).
[5] ATR音声データベース，ATR-Promotions（オンライン），入手先 ⟨http://www.atr-p.com/products/sdb.html⟩（参照 2015-3-25）.
[6] The HMM-based Speech Synthesis System（HTS），入手先 ⟨http://hts.sp.nitech.ac.jp⟩（参照 2015-3-25）.
[7] H. Kawahara, I. Masuda-Katsuse and A. Cheveigne: Restructuring speech representations using a pitch adaptive timefrequency smoothing and an instantaneous frequency based F0 extraction, *Speech Communication*, Vol. 27, No. 3–4, pp. 187–207 (1999).
[8] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi and T. Kitamura: A hidden semi-Marcov model-based speech synthesis system, *IEICE Transactions on Information and Systems*, Vol. E90-D, No. 5, pp. 825–834 (2007).
[9] 山岸順一，田村正統，益子貴史，小林隆夫，徳田恵一：平均声モデル構築におけるコンテキストクラスタリングと話者適応学習の検討，電子情報通信学会技術研究報告，Vol. 音声 102, No. 292, pp. 5–10 (2002).
[10] J. Yamagishi and T. Kobayashi: Average-voice-based speech synthesis using HSMM-based speaker adaptation and adaptive training, *IEICE Transactions on Information and Systems*, Vol. E90-D, No. 2, pp. 533–543 (2007).