Vol.2010-SLP-84 No.30
2010/12/21

調音運動 HMM 音声合成における調音特徴-声道パラメータ変換と音源の改良

小野田 高幸[†]　桂田 浩一[†]　新田 恒雄[†]

音声認識と合成を同じ調音運動モデルを用いて実現するシステムの開発を行っている．調音特徴を用いて HMM を設計することにより，音声認識と合成に共通な調音運動のワンモデルを実現している．合成では，HMM が生成する調音特徴系列を声道パラメータに変換し，駆動音源と組み合わせることで音声信号を得る．駆動音源は，CELP 符号化を応用した音源構築により音質改善を図る．提案する合成方式は,話者不変量である調音運動を HMM で表現することで音韻性を，また調音運動から声道パラメータへの変換を音韻共通の MLN で表現することで話者性を，各々独立したモジュールとして実装するため，少量の音声試料で特定話者の音声を合成できる可能性がある．評価実験では，被験者 10 名に対して受聴試験を行い，合成音の品質と声質（個人性）を評価する．その結果，二文の適応で品質・個人性ともに良好な結果を得ることができた．

Improvement of Articulatory-Feature to VT Conversion and Sound Source for HMM Speech Synthesis Based on Articulatory Movement

Takayuki Onoda[†], Kouichi Katsurada[†], and Tsuneo Nitta[†]

We are developing a speech recognition and synthesis system using common models of articulatory movement. By representing HMMs with articulatory features, we have achieved one-model speech recognition and synthesis. In the speech synthesis part, articulatory feature sequences generated by HMMs are converted into vocal tract parameters, and then they are combined with sound source. CELP coding technique is applied to improve sound quality when generating sound source. The proposed speech synthesis system separate phonetic information and speaker individuality. Therefore, it is expected to synthesize target speaker's speech using a small amount of voice data. In the experiments, we carried out listening tests for 10 subjects and evaluated both of sound quality and individuality of synthesized speech. As a result, we confirmed that the proposed synthesis system can produce good quality speech of target speaker with only two sentences.

1. はじめに

乳幼児は，身辺の僅かな人たちから音声を模倣するだけで音声言語を獲得している．親の音声と乳幼児の模倣音声は，声道形状等の違いから物理的には異なる音響信号になる．そのため，音韻知識が未確立の乳幼児がこれら二つの音声を同一と判断するには，何らかの不変量が内在することになる．この説明として，音声知覚が音響信号そのものではなく，調音器官に送られる運動司令を参照して行なわれるとし，音声の知覚と生成が一つのシステムで構成されているという見解が古くからある[1].

我々はその見解を基に，音声認識と音声合成の双方に，同一の音響モデル（AM）を使用する「ワンモデル音声認識合成方式」を開発している[2]．高精度音声認識の AM は，多数話者の音声コーパスが必要になる一方，HMM 音声合成では特定話者をターゲットにするため，認識と合成の双方を満足するには，話者不変の特徴量が必要になる．我々は先に，音声から話者不変な調音特徴を抽出し，調音運動を表現する AM を構成することで，高精度不特定話者音声認識が 1 名の話者データで達成可能なことを示した[2]．音声合成では，同じ AM から調音特徴系列を生成し，これを声道パラメータに変換（以後 VT 変換と表記）した後，音源符号帳で駆動して音声を合成する．この方式は，調音指令（motor command）と発声システムを分離できるため，少量の音声試料で特定話者の音声を合成できる可能性がある．

HMM 音声合成の音声品質の課題の一つに，パルス列と白色雑音で構成された音源の改良がある．本報告では，音源信号のうち有声区間に残差波形を利用することで合成音声の品質向上を試みる．残差を利用した音声合成手法はすでにいくつか提案されている[3][4]. 本文では, 音声符号化で用いられる CELP（Code-Excited Linear Prediction）方式[5]に基づく手法を検討する．

以下，2. では調音特徴系列から声道パラメータへの変換手法について解説する。続いて 3. で CELP 方式に基づく駆動音源の生成手法を説明した後，4. では参照話者 A の多量の文音声から VT 変換器と符号帳を設計し，これを目標話者 B の少量の音声で適応した際の，話者 B の音声品質と声質（個人性）を評価する．最後に 5. で本報告のまとめと今後の課題について述べる．

2. 調音特徴に基づく HMM 音声合成

2.1 調音特徴

調音特徴（Articulatory Feature; AF）[6]は，調音方法と調音位置から話者不変な音素属性を規定している．音声波形から AF を抽出する処理の流れを図 1 に示す.

[†] 豊橋技術科学大学 大学院工学研究科
Graduate School of Engineering, Toyohashi University of Technology

Vol.2010-SLP-84 No.30
2010/12/21

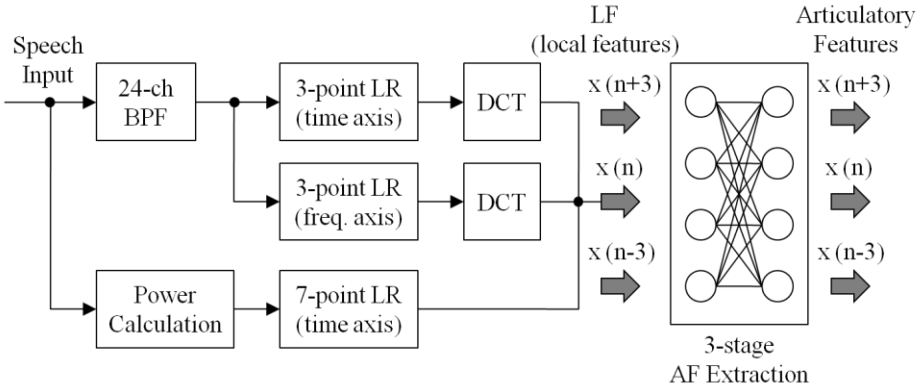


図 1　調音特徴抽出の処理の流れ

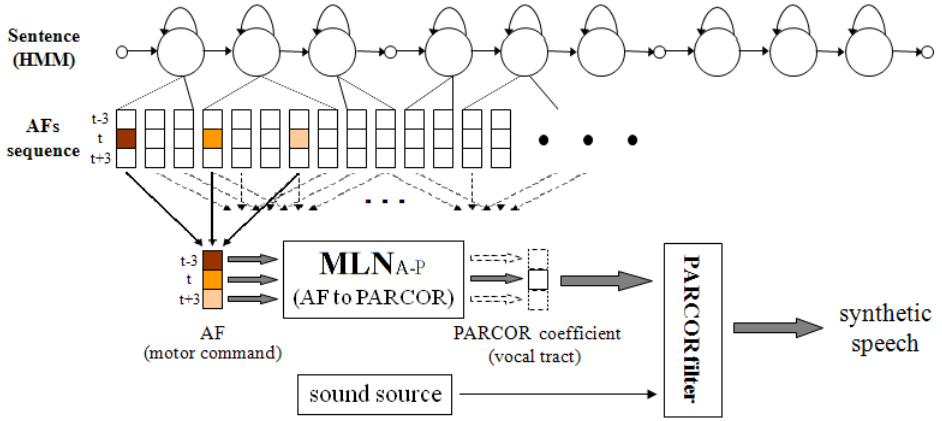


図 2　調音特徴に基づく HMM 音声合成

AF は，音声スペクトル系列の時間微分と周波数微分から求められる局所特徴（Local Feature; LF）を多層ニューラルネット（Multi-Layer Neural network; $MLN_{LF\text{-}AF}$）に入力して得られる．$MLN_{LF\text{-}AF}$ は前後の音素環境の違いを学習しており，抽出された 3 フレーム分の AF は調音の運動を表現したものに相当し，音声生成過程における運動司令（motor command）を構成する．

2.2 調音特徴から VT パラメータへの変換

話者に依存しない AF 系列から，話者固有の音声信号を得ることを考える．即ち，脳から送られる運動指令によって，発話者毎の声道（VT）を駆動するように，VT の音響フィルタを構成して音源信号で駆動したい．VT パラメータには，PARCOR（PARtial auto-CORrelation）係数[7]を用いた．

2.2.1 PARCOR 係数

PARCOR 係数は，線形予測における前向き予測誤差と後ろ向き予測誤差の相関係数として定義される．音声波形からは，自己相関係数を求め，Levinson-Durbin 法[8]を用いることで抽出することができる．PARCOR 係数は，声道の反射係数に関連した特徴とされており，声道形状と深い関係を持つ．そのため，AF と PARCOR 係数の間にはある程度の相関が保たれていると推測され，実際，AF から得た PARCOR 係数と原音声から得た PARCOR 係数は高い相関を持つ[2]．

音声信号を PARCOR 係数から構成した逆フィルタに通すことで，音源に相当する残差信号が得られる．この残差信号と PARCOR 係数を PARCOR フィルタに通すことで，音声を合成することができる．

2.2.2 調音特徴から VT パラメータへの変換

提案方法では，音声信号から AF と PARCOR 係数を直接計算し，対応関係を多層ニューラルネットワーク（以下，$MLN_{A\text{-}P}$ と表記）から導出することで，話者依存の VT モデルを得る．一方，AF は調音器官の実際の形状を抽出している訳ではない．つまり，ある時点での AF と調音器官形状の関係は一対一ではないため，AF から声道形状を推定する際，前後のコンテキスト情報を合わせて入力し，これを，調音結合を含む調音運動への制約として利用する．これにより $MLN_{A\text{-}P}$ が PARCOR 係数のパラメータ空間上で，滑らかな調音運動を実現することができる．

2.3 調音特徴に基づく音声合成システム

調音特徴に基づく合成システムの構成を図 2 に示す．まず，AF 系列を学習した HMM から，音素列と状態継続長を得た後，各状態の AF 平均ベクトルを得る．次に得られた AF から，$MLN_{A\text{-}P}$ によって PARCOR 係数を推定する．最後に PARCOR 合成器を音源信号で駆動し，合成音声を得る．

ここで，$MLN_{A\text{-}P}$ は予め，大量の文音声で学習した後，目標話者の少量の音声で適応化を行う．これにより，少量文学習で目標話者に近い PARCOR 係数が推定されることを期待している．駆動音源についても，大量の音声を使用した初期符号帳を，目標話者の少量の音声で適応する．

Vol.2010-SLP-84 No.30
2010/12/21

3. CELP 方式による駆動音源生成

一般的な HMM 音声合成では，パルス列と白色雑音を音源とするため，音質劣化が問題になる．そこで，この問題の解決法として CELP 符号化[5]を用いた駆動音源改善を提案する．

3.1 CELP 符号化

CELP 符号化は，人間の発声機構を音源成分とスペクトル包絡成分に分離してモデル化する vocoder 方式に属する．二つの成分を合成フィルタに供給して音声を生成する際，駆動音源成分を符号帳から探索し，入力波形に最も近いものを決定する．A-b-S（Analysis by Synthesis）法に基づく閉ループ探索を実装したことで，高音質音声符号化を実現している．

CELP 符号化の流れを図 3 に示す．まず，残差符号帳として残差波形データベースを構築しておく．符号化器では，入力音声を声道パラメータに変換を行い，残差符号帳内の残差素片の組み合わせで構成された音源と逐次合成を行う．そして，音声波形レベルでの誤差が最小となる残差の組み合わせを選択する．その選択された残差のインデックスと声道パラメータを複合化器へ伝送することで，音声の再合成を行う．

3.2 駆動音源の生成

今回提案する合成方式では，学習データから抽出した残差素片を CELP 符号化に基づく閉ループ探索を適用して，HMM の各状態に割り当てる．この手順を図 4 に示す．

学習データから予め残差波形を抽出すると共に，ピッチマークを付与する．続いて，ピッチマークを中心に基本周期の約 2 倍の領域を抽出し，一つの残差素片とする．こうして得た残差素片をデータベース化し，残差符号帳を構築する．その後，PARCOR 係数と予め付与したピッチマークを用いて，元の音声とピッチパルスの位置を合わせた後，閉ループ学習により残差素片を選択して，HMM の各状態に割り当てる．

ここで，子音には大量の音声から得た残差素片を，母音・撥音には目標話者の少量の残差素片を割り当てる．これにより，少ない学習データから目標話者に近い駆動音源が生成できることを期待している．

さらに，前後音素を考慮した残差素片選択を行い，各音素の HMM に，前後音素によって異なる最適な残差素片を複数持たせるようにした．これにより，滑らかな音源を実現することができる．

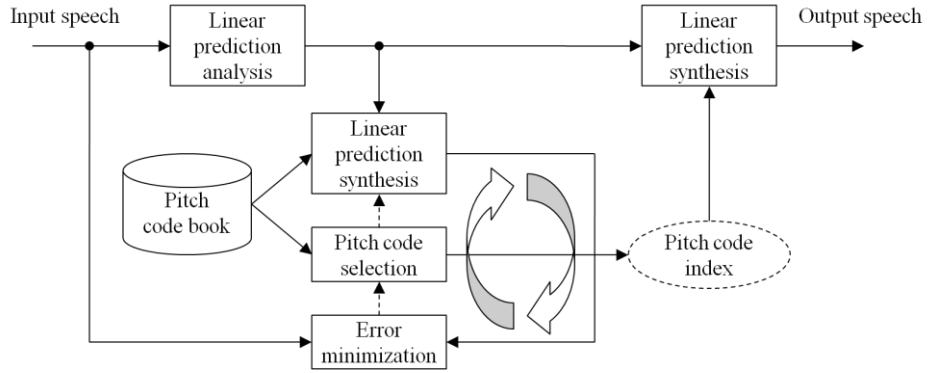


図 3　CELP 符号化の流れ

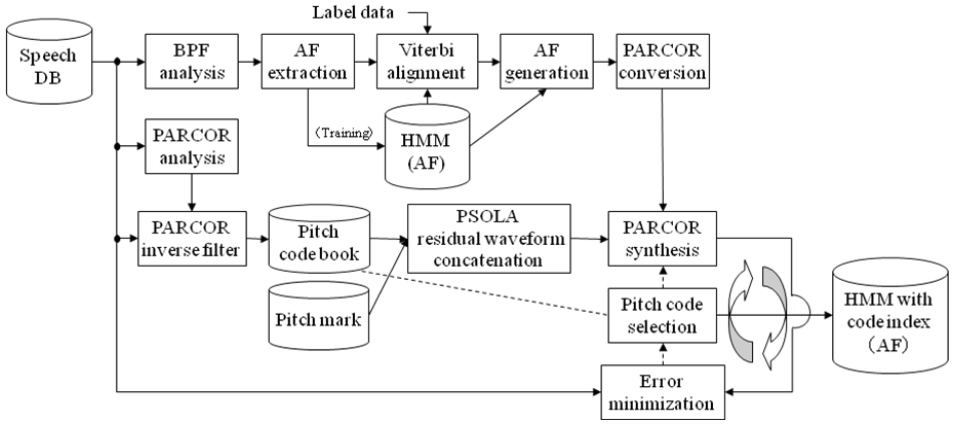


図 4　CELP 方式による駆動音源の生成

Vol.2010-SLP-84 No.30
2010/12/21

4. 評価実験

4.1 評価環境

今回の音声合成システムで使用した HMM の仕様を表 1 に，AF-PARCOR 係数変換器を表 2 に示す.

音源符号帳作成の学習データは，AF-PARCOR 係数変換器と同様の音声コーパス，話者，文数を用いた．また，分析窓長は 25ms，分析周期は 10ms である．なお，今回の実験では，音素列と状態継続長を，音声から直接抽出している.

4.2 駆動音源に関する客観評価テスト

4.2.1 実験内容

提案手法による駆動音源が，従来手法と比べて改善されているかを確認するため，

(a) パルスと白色雑音から成る音源を用いた合成音声（従来手法）
(b) CELP 符号化による駆動音源を用いた合成音声（提案手法）
(c) 目標話者の原音声

の音声波形を比較する．なお，目標話者の文数は二文使用した.

4.2.2 実験結果

得られた音声波形（発話：「一週間ばかり，ニューヨークを取材した.」）を図 5 に示す．図 5(b)より，CELP 符号化による駆動音源を用いることで目標話者の音声波形に近づいたことが分かる.

4.3 話者適応に関する客観評価テスト

4.3.1 実験内容

目標話者の音声が正しく適応されたかを確認するため，

(a) 合成音声（目標話者の音声の適応なし）
(b) 合成音声（目標話者の音声を 2 文だけ適応）
(c) 目標話者の原音声

のスペクトログラムを比較する.

4.3.2 実験結果

得られたスペクトログラム（発話：「一週間ばかり，ニューヨークを取材した.」）を図 6 に示す．図 6(b)より，目標話者の音声を 2 文使用しただけで目標話者のスペクトルに近づいたことが分かる.

表 1 実験で使用した HMM

HMM	monophone-HMM（38 音素）, 5-state 3-loop, left-to-right
学習コーパス	JNAS（男性 38 名, 5000 文; 16bit, 16kHz）
特徴量	AF15 次元×3 フレーム（計 45 次元）

表 2 実験で使用した AF-PARCOR 係数変換器

MLN	3 層（入力層 45, 中間層 450, 出力層 39）
学習コーパス	ATR 音素バランス文（16bit, 12kHz） ● 大量音声話者 ： MHT（男性, 503 文） ● 目標話者 ： MMY（男性, 2, 10, 30 文）
入力	AF15 次元×3 フレーム（計 45 次元）
出力	PARCOR 係数 13 次元×3 フレーム（計 39 次元）

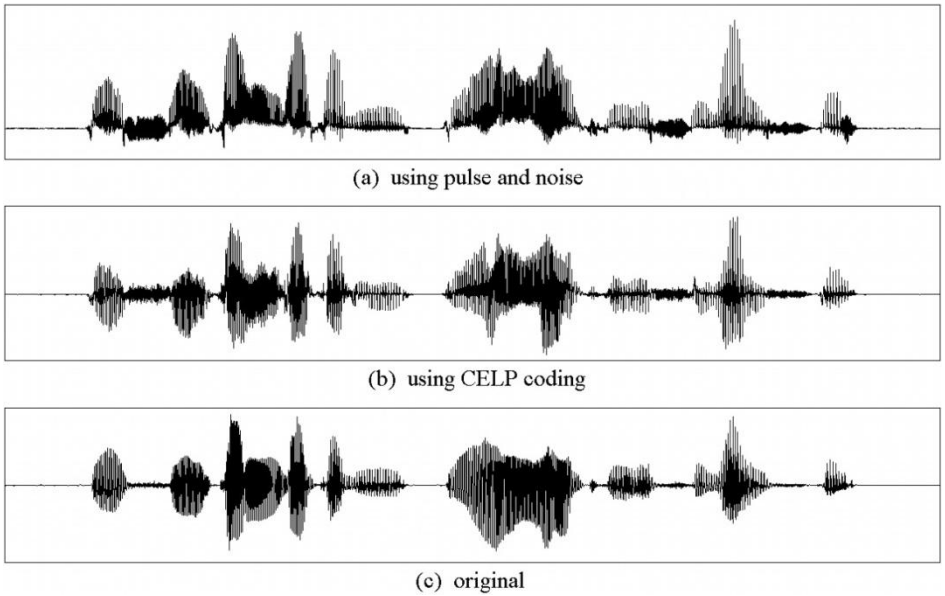


図 5 音声波形比較

Vol.2010-SLP-84 No.30
2010/12/21

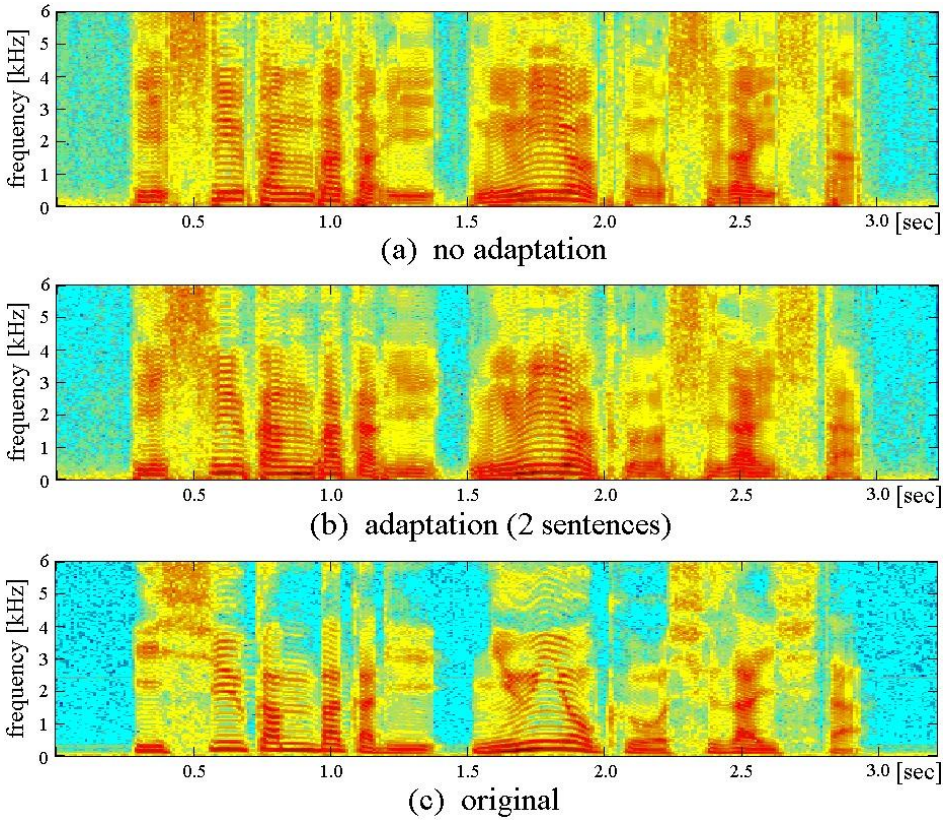


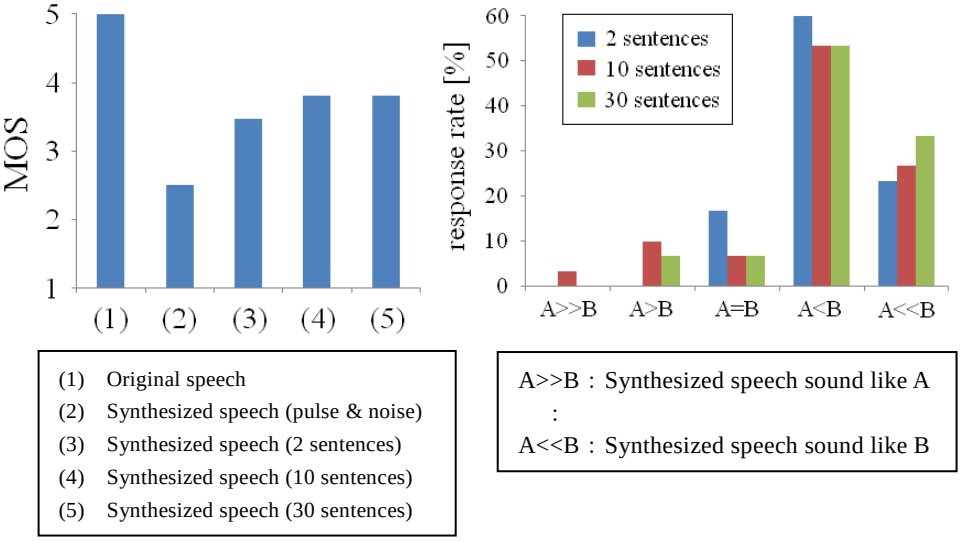
図 6　スペクトログラム比較

4.4　主観評価テスト

4.4.1 実験内容

目標話者の音声が正しく適応されたか，品質が向上したかを確認するため，被験者10名に対して以下の受聴試験を行った. なお, 合成音声は各実験とも9文を使用した.

1. 目標話者の音声を 2, 10, 30 文使用した時の合成音声をランダムに聴かせ，音質をそれぞれ 5 段階（5: 良い～1: 悪い）で主観評価
2. ABX 法による受聴試験 (A: 大量音声話者の原音声, B: 目標話者の原音声, X: 目標話者の音声を 2, 10, 30 文使用した際の合成音声．被験者ごとに A と B を入れ替え)



(a) Evaluation of speech quality　　(b) Evaluation of speaker individuality

図 7　主観評価テスト結果

4.4.2 実験結果

受聴試験の結果を図 7 に示す. 図 7(a)より, CELP 符号化による駆動音源（(3)～(5)）を用いることで，パルスと白色雑音から成る音源（(2)）を用いた時と比べて MOS 値が大幅に向上した．また，目標話者の音声試料を増やすことで MOS 値が向上した．図 7(b)からは，2 文使用時でも目標話者の音声に近いと判断される割合が約 83%となり，音声試料を増やすことで確度が高くなった．しかし，2 文使用時の MOS 値はまだ低く，原因として，$MLN_{A\text{-}P}$ の追加学習が不十分であったと考えられる．

5.　まとめ

本報告では，運動司令に相当する調音特徴系列から，音韻性と話者性を保持した音声を合成できることを示した．また，CELP 符号化の手法を応用することにより，従来の音源（パルス＋ノイズ）と比較して高品質な音声を再生できた．さらに，少ない文で目標話者に近い音声を合成することができ，評価テストでは二文適応でも約 83%の割合で目標話者に近いと判断された.

今後は, AF-PARCOR 変換器と音源のさらなる改良を行い, 品質向上を目指したい.

Vol.2010-SLP-84 No.30
2010/12/21

なお，今回はピッチや状態継続長を原音声から直接抽出したものを用いた．今後，ピッチや状態継続長についてもそれぞれモデル化し，テキスト音声合成を実現したい．

参考文献

[1] A. M. Liberman, I. G. Mattingley：The motor theory of speech perception revised, Cognition, 21, pp.1-36 (1985).
[2] 新田恒雄，武井匠，木村優志，桂田浩一：調音運動 HMM に基づくワンモデル音声認識合成，情報処理学会研究報告 SLP, Vol.2009-SLP-77, No.4, pp.1-6 (2009).
[3] 斎藤隆：圧縮した残差を用いた規則音声合成法，情報処理学会全国大会講演論文集, Vol.45, No.2, pp.339-340 (1992).
[4] 小池宗幸，岩野公司，古井貞熙：HMM 音声合成における残差駆動による自然性向上の検討，日本音響学会春季研究発表会講演論文集, vol.1, 1-6-10, pp.241-242 (2003).
[5] M. R. Schroeder, B. S. Atal：Code-excited linear prediction (CELP): high-quality speech at very low bit rates, ICASSP'85, vol.10, pp.937–940 (1985).
[6] ムハマド ヌルル フダ，河嶋宏明，新田恒雄：3 ステージ MLN と抑制／強調処理に基づく調音特徴抽出，情報処理学会研究報告 SLP, Vol.2008, No.123, pp.149-154 (2008).
[7] 板倉文忠，斎藤収三：偏自己相関関数による音声分析合成系，日本音響学会誌, Vol.25, No.5, pp.306 (1969).
[8] N. Levinson：The Wiener RMS (Root Mean Square) error criterion in filter design and prediction, Journal of Mathematics and Physics, Vol.25, pp.261-278 (1947).