3B1-OS22c-2

調音運動の one-model を用いた音声認識・合成の改良

Improving Speech Recognition and Synthesis Based on One-model of Articulatory-Movement

新田 恒雄　小野田 高幸　荒木 厚太　木村 優志　入部 百合絵　桂田 浩一
Tsuneo Nitta　Takayuki Onoda　Kouta Araki　Masashi Kimura　Yurie Iribe　Kouichi Katsurada

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

Speech recognition (SR) and speech synthesis (SS) based on one-model of articulatory movement HMMs that are commonly applied to both an SR module and an SS module are described. The SR module has an articulatory feature (AF) extractor with multi-layer neural networks (MLNs) that output an AF sequence to HMMs. In the SS module, the speaker-invariant HMMs are applied to generate an articulatory feature (AF) sequence, and then, after converting AFs into vocal tract parameters by using a multi-layer neural network (MLN), a speech signal is synthesized by an LSP (Line Spectrum Pairs) digital filter. CELP coding technique is applied to improve sound quality when generating voice source from embedded codes in the corresponding state of HMMs. The proposed speech synthesis system separate phonetic information and speaker individuality. Therefore, target speaker's voice can be synthesized with a small amount of speech data. The experimental results show that the proposed system can produce good quality speech with only two-sentences.

1. はじめに

我々は，音声認識 (SR)と音声合成 (SS) に共通に適用可能な調音運動 HMM に基づく，ワンモデル SR – SS モジュールを開発している [1]。これまで SR と SS は独立したシステムとして設計され，近年の HMM ベース SR では話者独立の単音モデルを，また HMM ベース SS では話者依存の単音モデルが組込まれてきた[2]。これら二つの HMM ベースモジュールを一体化するには，次に示す三つの要求を満足しなければならない。すなわち，ワンモデル SR - SS モジュールでは:

(i) 音声信号から調音特徴(AFs) のような話者不変の特徴を抽出できること,
(ii) SR と SS に共通の HMM が話者不変特徴を用いて設計できること,
(iii) HMM から生成される特徴パラメータは，声道 (Vocal Tract; VT) パラメータのように，話者特有の合成パラメータに変換できること

が要請される。ワンモデル SR & SS における音声認識モジュールは，話者不変の特徴パラメータを HMM に導入したことで，効率の良い音声認識を実現できる。先の報告では，登録話者 1 名，混合数 1 の調音運動 HMM が，話者 100 名，混合数 16 の標準的 MFCC ベース HMM (monophone) を上回る音素認識性能を得ることを示した[1]。本報告では，triphone モデルの条件で提案する AF と，標準的な MFCC 特徴の性能比較を示す。

一方，音声合成モジュールでは，調音運動 HMM は脳の運動指令に相当する AF 系列を生成し，続いてこれらの AF 系列を多層ニューラルネット(MLN) に通すことで，VT パラメータに変換する。先の報告では，VT パラメータに PARCOR 分析から求められる k-parameter を使用し，分析時の残差信号を駆動音源として音声を生成した。また駆動音源を CELP 符号化[3] で用いられる符号帳(codebook)として設計し，HMM の各状態に符号を張りつける方式を提案した [4]。有声音のピッチ制御には PSOLA(Pitch Synchronous Overlap and Add)[5]方式を用いている。本報告では，VT パラメータに近年の音声符号化で主流となっている，LSP(Line Spectrum Pair) [6] を採用する。同時に，少量音声サンプルで AF→LSP 変換器の MLN を学習し，目標話者の VT に適応させる。また，音源の符号帳についても，VT 適応に用いたと同じ少量サンプルから残差信号を抽出し，母音と撥音に関する音源符号部分を目標話者に適応させる。実験では，音声品質と声質 (個人性) について，提案方式を評価する。

連絡先:新田恒雄，豊橋技術科学大学，〒441-8580 豊橋市天伯町雲雀ケ丘 1-1, 0532-44-6890, nitta@cs.tut.ac.jp

2. ワンモデル音声認識・合成システムの概要

図 1 に調音運動 HMM に基づく，音声認識・合成システムの概要を示す。図の上側が認識モジュール，下が合成モジュールである。二つのモジュールは共通の調音運動 HMMs を利用する。認識エンジンは，三段の多層ニューラルネット(MLN) で構成される調音特徴(AF)抽出器を持ち，AF 系列を調音運動 HMMs に送る。HMMs は単音ごとの調音ジェスチャの振舞いを確率的に表現している。

合成エンジンでは，認識と同じ話者不変の HMMs が，単音モデルを結合しながら AF 系列を生成し，これらを話者特有の

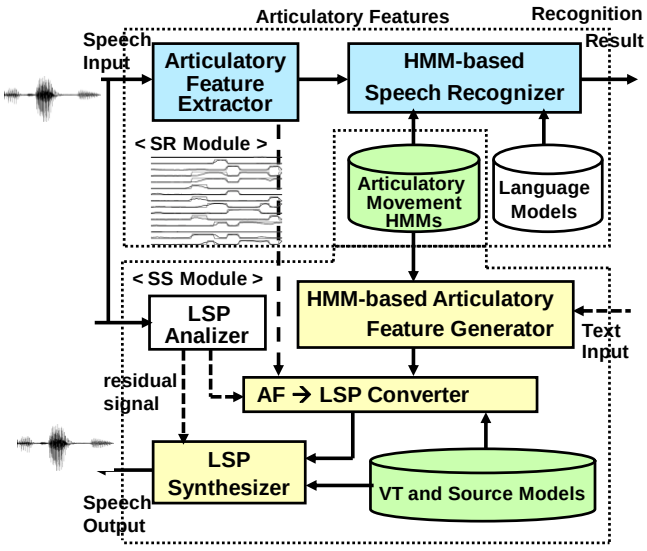


図 1 ワンモデル音声認識合成システム

LSP 声道パラメータ系列に変換する。音声は，LSP 系列と音源信号を入力として，LSP デジタルフィルタで構成される合成器から生成される。音源信号は，HMM から音源符号が読みだされ，PSOLA 方式を用いて，ピッチの音調曲線(pitch contour; 現在は音声から直接抽出したものを使用) に沿った制御を行う。提案方式は，また，図に点線で示されているように，調音特徴抽出器の出力を直接，AF→ LSP 変換器に加えることで音声を合成することができる。この機能は，対話システムで未知語を確認する際の talk-back や，語学学習に利用することができる。

3. 調音運動 HMM に基づく音声認識

3.1 調音特徴

調音特徴(AF)は，音声生成時の調音に関する素性を表現する特徴量で，調音位置(高舌性，前方性，前端性，…)や調音様式(母音性，連続性，摩擦性，…)といった，音声の持つ構造特徴から成る。AF ではあらゆる音素をその属性の有無(＋/－)から表現する。

3.2 調音特徴の抽出

AF の抽出過程を以下に説明する[7], [8]。

(i) 24ch.の BPF 出力(時間-周波数パターン)から 3 点線形回帰(LR)係数を時間軸，周波数軸に沿って求め局所特徴(Local Feature: LF)を抽出する(音声パワーについても，5 点回帰係数ΔP を抽出して LF に組込む)，
(ii) LF から AF への変換を 1 段目の MLN 学習により求める，
(iii) 前後の AF コンテキストの違いを考慮した AF 系列の整形を 2 段目の MLN で行う，
(iv) AF 系列から加速度係数を求め，調音運動の強調と抑制を 3 段目の MLN 学習により行う，
(v) HMM への入力特徴に要請される無相関化を，AF 時系列間の Gram-Schmidt 直交化により行う。

3.3 調音運動モデルを表現する HMM

3.2 に説明した抽出過程を経た AF は，局所的な振幅分布を持つため，混合数を増やしても認識性能が改善されない傾向がある[1]。以下では，triphone モデルで比較実験を行うにあたり，状態数を 3-loop (以下 3 状態と呼ぶ) から 5-loop に増加させ，時間方向にモデルを精緻化することで，認識性能(音素正解率，音素正解精度)を改善することを検討する。なお，今回の評価は 3.2 の処理から(iv) の加速度係数による，調音運動の強調・抑制制御を除いて行っている。

3.4 評価実験

triphone モデルでの状態数の違い(3-loop ←→ 5-loop) を比較すると共に，AF と MFCC の性能比較を行う。

3.4.1 音声試料

- 音響モデル(AM)学習データセット
  新聞記事読み上げ音声コーパス(16kHz,16bit) のうち男性話者 33 名，5000 文 [9]。
- 評価データセット
  AM 学習データとは異なる，新聞記事読み上げ音声コーパス(16kHz,16bit) のうち男性話者 17 名，2719 文。

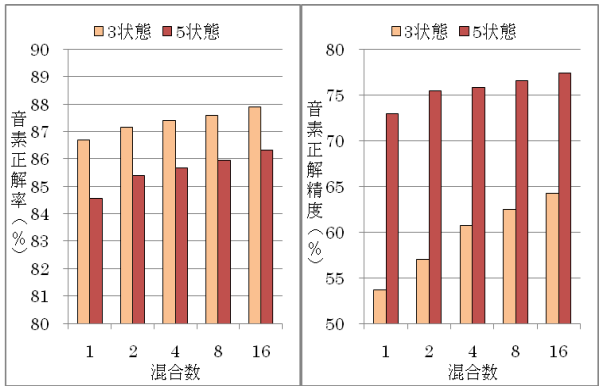


図 2 調音特徴(AF) における 3 状態と 5 状態の比較

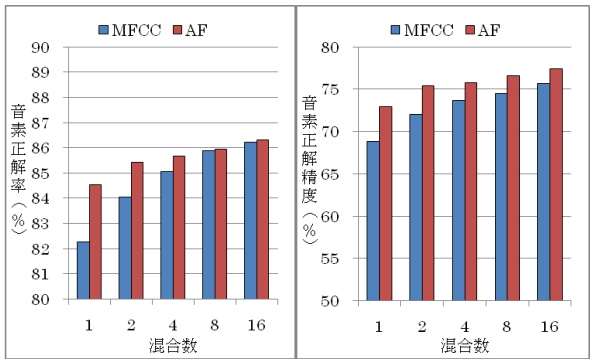


図 3 調音特徴(AF) と MFCC との比較 (5 状態)

3.4.2 音響モデル(AM)の諸元

音響モデルには，日本語 38 音素の triphone モデル(状態数が約 3,000 となるよう状態を共有化)を作成し，混合数は 1, 2, 4, 8, 16 を比較する。特徴パラメータの次元数は以下。

MFCC: MFCC+$\Delta t+\Delta\Delta t+p+\Delta p+\Delta\Delta p$ (39 次元)
AF : AF+AF_{t-3}+AF_{t+3} (45 次元)

3.4.3 評価結果

図 2 に triphone AM を 3 状態から 5 状態に増やし，時間方向にモデルを精緻化した際の比較結果を示す。5 状態は 3 状態に比べて，正解率が微減するが，正解精度が大きく改善される。調音特徴系列はカテゴリカル(範疇的)なパターンを示すことから，時間区分を詳細に表現した 5 状態モデルが性能向上に寄与したと考えられる。次に，triphone モデルにおいて高い正解精度を示した 5 状態 HMM で，AF と MFCC 間の性能比較を行った結果を図 3 に示す。MFCC の場合も 3 状態から 5 状態としたことで，AF の場合と同じく正解率が 1% 程度低下したが，正解精度は大きく改善された。また，AF の MFCC に対する優位は，monophone の場合ほどではないが変わらなかった。特に低混合数で AF の優位は大きい。

4. 調音運動 HMM に基づく音声合成

4.1 調音特徴から VT パラメータへの変換

調音運動を表現する HMM から生成される，AF 系列を特定話者の VT パラメータ (LSP 係数)に変換し，音源符号帳で駆動する。AF から VT への変換には，MLN を用いた(以下 MLN_{A-L})。

MLN_{A-L} の入力は，滑らかな調音動作を実現するため，AF 系列で前後のコンテキスト情報を合わせて入力する。

4.2 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 4 に示す。まず，AF 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-L} によって LSP 係数を推定する。ここで，MLN_{A-L} 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 LSP 係数が推定されることを期待している。最後に LSP 合成器を音源信号で駆動し，合成音声を得る。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。

図 5 に原音声から求めた LSP 係数と，HMM から生成された調音特徴を MLN_{A-L} に通して得た LSP 係数間の相関係数を示す。MLN_{A-L} の適応学習には二文を使用した。二つの LSP 係数は高い相関値を示すことが分かる。

4.3 CELP 方式による駆動音源の生成

CELP 符号化[3]は，人間の発声機構を音源生成部とスペクトル包絡成分に相当する声道部からモデル化する。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符号を選択する。AbS (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再現している。

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に，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 6 に示した。

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，LSP 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の残差素片，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。

4.4 評価実験

MLN_{A-L} の学習データとして，ATR 音素バランス文を使用した[10]。大量の音声には話者 MHT の 503 文を，目標話者の音声には話者 MMY の 2, 10, 30 文をそれぞれ使用した。音源符号帳作成の学習データも同様，ATR 音素バランス文を用いて，大量の音声，目標話者の音声共，MLN_{A-L} の学習データと同じ話者，文数を用いた。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。

4.4.1 スペクトルパターン比較による評価

適応後の目標話者の音声を評価するため，以下の三つの音声を比較する。

(a) 合成音声（目標話者の音声未使用），
(b) 合成音声（二文で適応），
(c) 目標話者の原音声

図 7 に三つの音声に対するスペクトルパターン（発話は /kaN/ ）を示した。(c) の目標話者の原音は，(a) の話者適応なしの音声は母音部，撥音部とも大きく異なる。一方，(b)の二文適

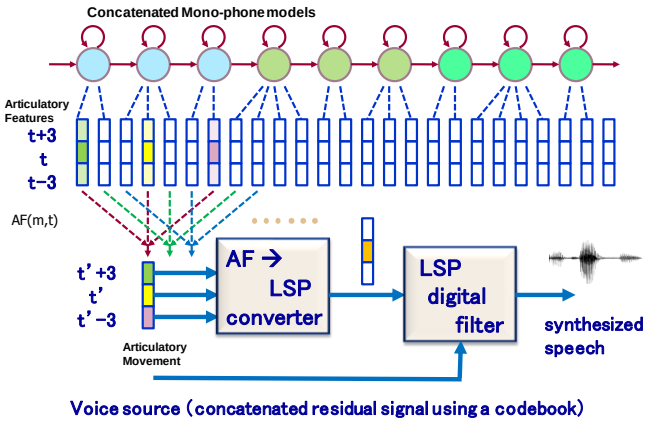


図 4 調音運動 HMM を用いた音声合成

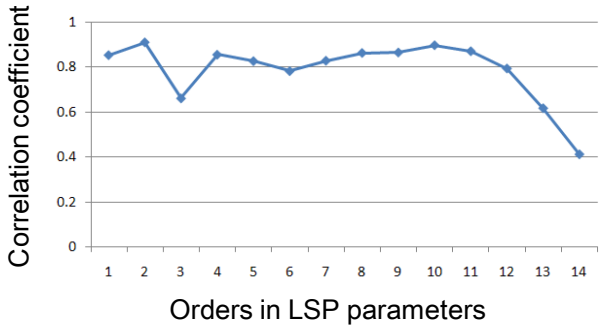


図 5 原音と合成音声間の LSP 係数相関値

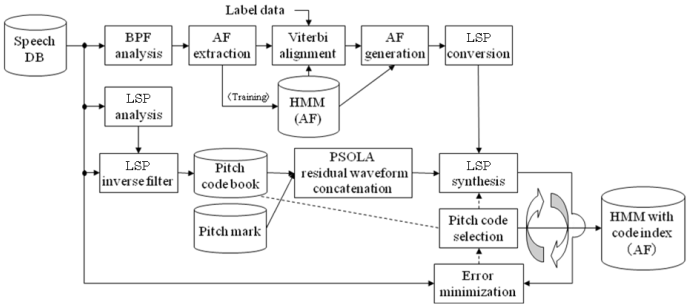
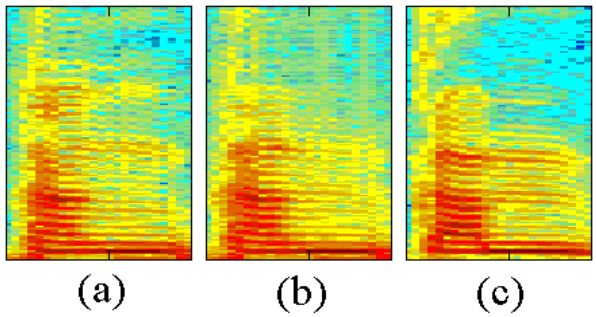


図 6 CELP 方式の閉ループ学習を適用した音源改良



(a) 合成音声（目標話者の音声による適応なし）
(b) 合成音声（目標話者の二文の音声で適応）
(c) 目標話者の原音声

図 7 スペクトルパターン比較 (/kaN/)

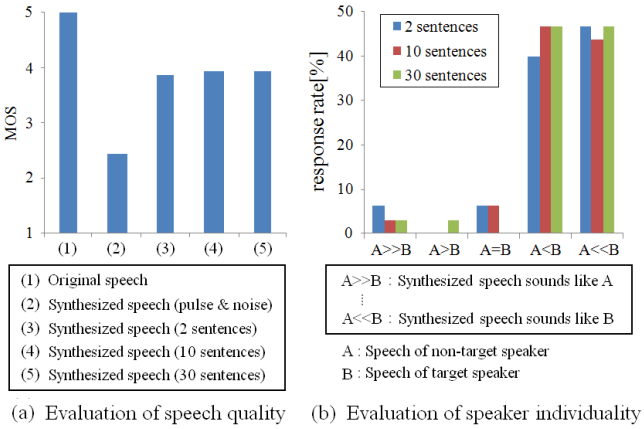
応後のスペクトルは，目標話者と母音部/a/，撥音部 /N/ とも近付いていることが分かる。

4.4.2 主観評価テスト

目標話者の音声が正しく適応されたかを主観評価から確認するため，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。評価は次の二項目で行う。

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)で評価する。
② ABX 法を用いた個人性評価実験
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 文で適応学習した音声
被験者ごとに A と B を入れ替えて受聴し評価する。

受聴試験の結果を図 8 に示す。音声試料を増やしても MOS 値の変化はなかったが，2 文使用時 (3)でも従来音源(2)を使用した時と比べ MOS 値が大幅に向上した。また，
2 文使用時でも約 87%の割合で声質が目標話者に近いと判断され，適応手法の有効性が示された。



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

図 8 音質と個人性に対する評価結果

5. おわりに

調音運動の共通モデルを音声認識と音声合成に適用する方式を提案した。音声認識では，従来使用していた 3 状態のモデルを 5 状態とすることで，音素正解精度が大きく向上することを示した。音声合成では，二文程度と少ない文数で目標話者に近い音声を合成できた。また，CELP 符号化の手法を応用することにより，高品質な音声を再生できた。今後は，ピッチや状態継続長についてもモデル化し，テキスト音声合成を実現したい。

参考文献

[1] Nitta, T., Onoda, M., Kimura, M., Iribe, Y., and Katsurada, K., One-model speech recognition and synthesis based on articulatory movement HMMs, Proc. Of INTERSPEECH 2010, pp.2970-2973 (2010-9).
[2] Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T., and Kitamura, T., Speech parameter generation algorithms for HMM-based speech synthesis, Proc. of ICASSP, pp.1315- 1318 (2000-6).
[3] Schroeder, M. R., Atal,B. S., Code-excited linear prediction (CELP): high-quality speech at very low bit rates, Proc. of ICASSP'85, vol.10, pp.937–940 (1985).
[4] 木村，小野田，入部，桂田，新田，調音運動に基づくワンモデル音声認識合成への CELP 適用，人工知能学会全国大会，OS-13-2 (2010).
[5] Charpentier, F. J., and Stella, M. G., Diphone synthesis using an overlap-add technique for speech waveforms concatenation, Proc. of ICASSP'86, pp. 2015-2018 (1986).
[6] Itakura, F., Line spectrum representation of linear predictor coefficients of speech signals, J. Acoust. Soc. Am. Vol. 57, Issue S1, pp.35-35 (1975).
[7] Huda, M.N., Katsurada, K. and Nitta, T., Phoneme recognition based on hybrid neural networks with inhibition/ enhancement of Distinctive Phonetic Feature (DPF) trajectories, Proc. Interspeech'08, pp.1529-1532 (2008).
[8] Huda, M.N., Kawashima, H. and Nitta, T., Distinctive Phonetic Feature (DPF) extraction based on MLNs and Inhibition/ Enhancement Network, IEICE Trans. Inf. & Syst., Vol.E92-D, No. 4, pp.671-680 (2009).
[9] JNAS: Japanese Newspaper Article Sentences. http://www.milab.is.tsukuba.ac.jp/jnas/instruct.html
[10] Abe, M., Sagisaka,Y., Umeda, T. and Kuwabara, H., SpeechDatabase User's Manual. *ATR Technical Report*, TR-I-0116 (1990). (in Japanese)