調音運動に基づくワンモデル音声認識合成への CELP 適用

Applying CELP to the One-model Speech Recognition and Synthesis Based on Articulatory Movement

木村優志[*1] 小野田高幸[*1] 入部百合絵[*2] 桂田浩一[*1] 新田恒雄[*1]
Shinta SAWADA Masashi KIMURA Kouichi KATSURADA Tsuneo NITTA

豊橋技術科学大学 [*1]大学院工学研究科 [*2]情報メディア基盤センター
[*1] Graduate School of Engineering and [*2] Information & Media Center , Toyohashi University of Technology

Speech recognition and synthesis have been designed in the form of separate engines. In this paper, we propose one-model speech recognition (SR) and synthesis (SS) to which a common articulatory movement model is applied. The SR engine has an articulatory feature (AF) extractor with multi-layer neural networks (MLNs) that output an AF sequence to articulatory movement HMMs. The articulatory movement HMMs show high recognition performance even if the training data are limited to a single speaker. The SS engine, the same speaker-invariant HMMs generate AF sequences, and then they are converted into vocal tract parameters using a speaker-specific model. Synthesized speech is obtained by feeding the k-parameters into a PARCOR synthesizer. In this paper, CELP (Code Excited Linear Prediction) technique is applied to SS for improving the sound quality.

1. はじめに

音声認識と音声合成はこれまで別個のシステムとして開発されてきた。標準的な HMM ベース音声認識について考察すると，近年，幾つかの分野で成功を収めたが，多くが MFCC など音声スペクトル由来の特徴を使用するため，話者，音素コンテキスト，ノイズによる多様な変動を抱え，モデル近似に多くのデータと混合分布を要するという欠点を持つ。他方で人間の幼児は，親の声を通して不特定多数話者の音素体系を学習しており，音声認識システムのように多数話者の音声を学習する必要がない [1]。このような特殊な言語能力を可能にする機構を説明するために，人間の音声知覚が調音運動，すなわち調音ジェスチャを参照して行われるという説が古くから提唱されてきた [2]。調音ジェスチャを抽出して音声認識に利用しようとする試みは，古く 1970 年代の初めに販売された Threshold Technology 社の音声認識装置に見られたが (当時の装置技術資料による)，近年に至って数多くの方式が提案されるようになり [3], [4], [5], [6], [7] , 多数話者音声で学習した標準的 MFCC ベース HMM を上まわる性能も得られるようになっている。また，よく設計された調音特徴ベース HMM は，学習に 1 話者の音声データしか使用しない場合にも，従来方式を上まわる性能を得ることができることも示された [8]。

人間の音声生成と音声知覚が 1-system か 2-system かは，長年論争され未だ決着がついていないが [9]，近年の脳研究は 1-system 説を支持する結果を示しつつある [10]。我々は，音声認識のための調音運動モデルを HMM で実現し，同じモデルから音声を合成する方式を提案している。従来の標準的 HMM 音声合成 [11]は，スペクトル由来の特徴 (ケプストラム) を使用するため，特定話者の多量の音声を必要とし，また不特定話者の音声を認識することはできなかった。提案方式は，話者共通の調音運動を HMM で表現すると同時に，HMM から得られる調音特徴系列を，多層ニューラルネット(MLN)を用いて作成した声道パラメータ (PARCOR 係数)変換器に通し，PARCOR 合成フィルタ [12]を通して合成音声を得る。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，少量の音声資料で明瞭な音声を合成できる可能性がある。

連絡先：新田恒雄，豊橋技術科学大学 大学院工学研究科情報知能工学専攻，〒441-8580 愛知県豊橋市天伯町雲雀ヶ丘 1-1，0532-44-6890，nitta@cs.tut.ac.jp

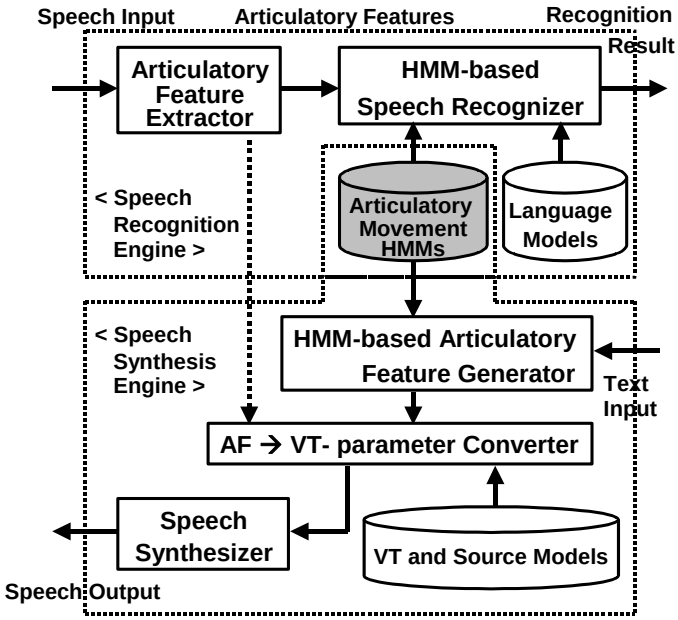

図 1 調音運動に基づく One-Model 音声認識・合成システム

本報告では，先に提案した調音運動 HMM に基づくワンモデル音声認識・合成の合成部に対して，CELP 符号化技術を適用し品質改善を図る。以下ではまず，2 節で One-model 音声認識合成システムの概要を説明した後，3 節で調音運動 HMMs に基づく音声合成の音質改善について検討した結果を述べる。

2. One-model 音声認識合成システム

図 1 に調音運動モデルに基づく音声認識合成システムの概要を示す。図の上側が音声認識エンジン，下が音声合成エンジンである。二つのエンジンは共通の調音運動 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) で構成した調音特徴(AF)抽出器を持ち[13],[14], 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，これらを話者依存の声道パラメータ（k-parameter）に変換する。合成音声は，この k-parameter 系列を PARCOR 合成フィルタに供給し，音源信号で駆動することで得られる。この方式は図に示すように，調音特徴抽出器の出力を直接，AF→ VT (Vocal Tract；声道)パラメータ変換器に加えることで音声を合成することもできる。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 HMMs に基づく音声合成

HMM 音声合成方式は，一般に特定話者の音声データを元に HMM のモデルを制作する[11]。このため，近年は効率をよくする工夫が，話者適応を初め種々行われている。効率を悪くしている理由の一つは，スペクトラム情報を扱っていることからきている。これに対して，調音特徴は話者に関して不変なパラメータのため，話者にカスタマイズしたい用途で利点が大きいと考えられる。

図 2 は調音特徴を使用した音声合成の処理過程を示している。HMM は音声認識用に作成したものをそのまま使用している。HMM は単音モデルを連結しながら調音特徴を生成する。各状態の平均ベクトルが，AF→ PARCOR 変換器に送られるが，この時，前後の少し離れたフレームの値も同時に利用することで，滑らかな音声が生成できる。

3.1 調音特徴から声道パラメータへの変換

図 2 に示す調音パラメータ→PARCOR 係数変換器は，MLN で構成されており，入力ユニット 45 (15×3 フレーム)，出力ユニット 39 (13×3 フレーム)，隠れ層ユニット数は 450 である。学習には ATR 音素バランス文 [16] から 30 名 1508 文を使用した。また，話者を特定した音声合成を評価するため，これと別に 1 名の 2 文を用いて，調音パラメータ→PARCOR 係数変換器を適応学習した。図 3 にスペクトルパターンの比較を示す（発話は「受賞者は次の通り」）。図の(C) が原音，(A) が不特定多数話者で学習した MLN から得た合成音声，(B)が 2 文で適応した際の合成音声である。なお，音声合成の音源は，PARCOR 分析の残差信号を用いた。図から，2 文程度で特定話者の音声に近いスペクトルが得られることが分かる。 図 4 は，主観評価実験(MOS 値) により，2 文と 32 文の適応効果を比較した結果である。被験者は 9 名で，対象話者 3 名について評価している。目安となる MOS 値 4 にはまだ達しておらず，一層の改良が必要である。

3.2 駆動音源の改良

学習データから抽出した残差素片を CELP 符号化の手法を用いて， HMM の各状態に割り当て，残差素片選択して音声品質を改善する。この手順を図 5 に示す。まず，学習データから残差波形を抽出すると共に，ピッチマークを付与する。次に，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。音声合成の際は，HMM から AF 系列を生成した後，PARCOR 係数に変換する。PARCOR 係数と予

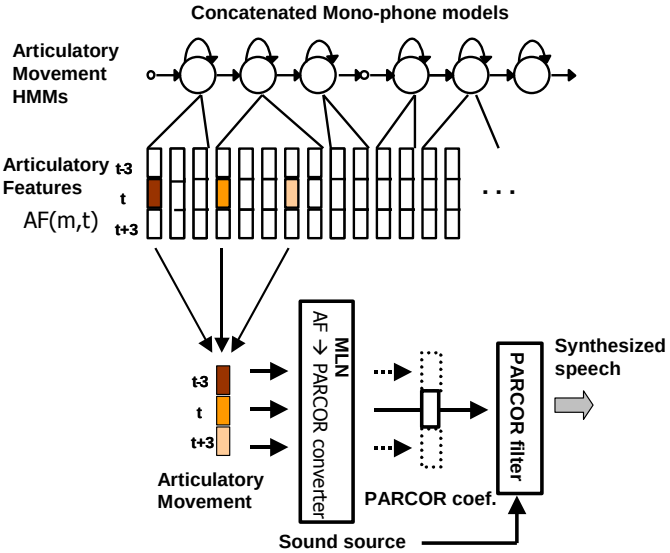

図 2 調音運動 HMM に基づく音声合成

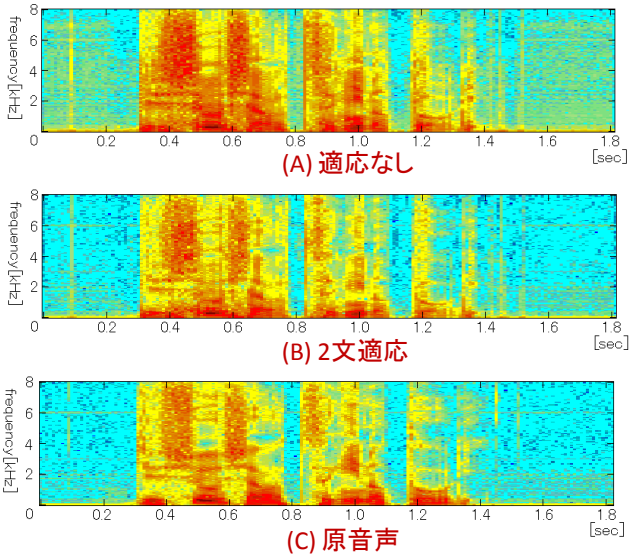

図 3 合成音声のスペクトルパターン比較

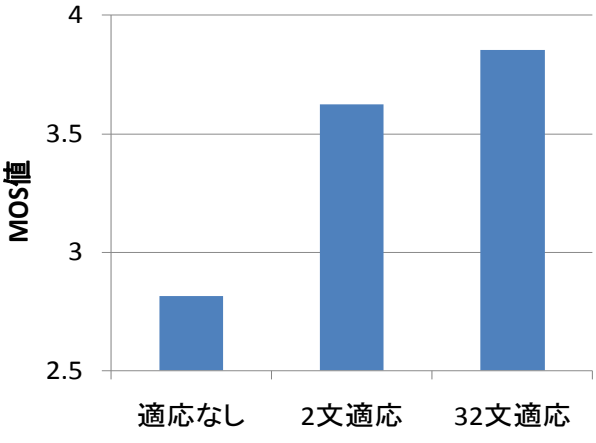

図 4 合成音声に対する主観評価（MOS）比較

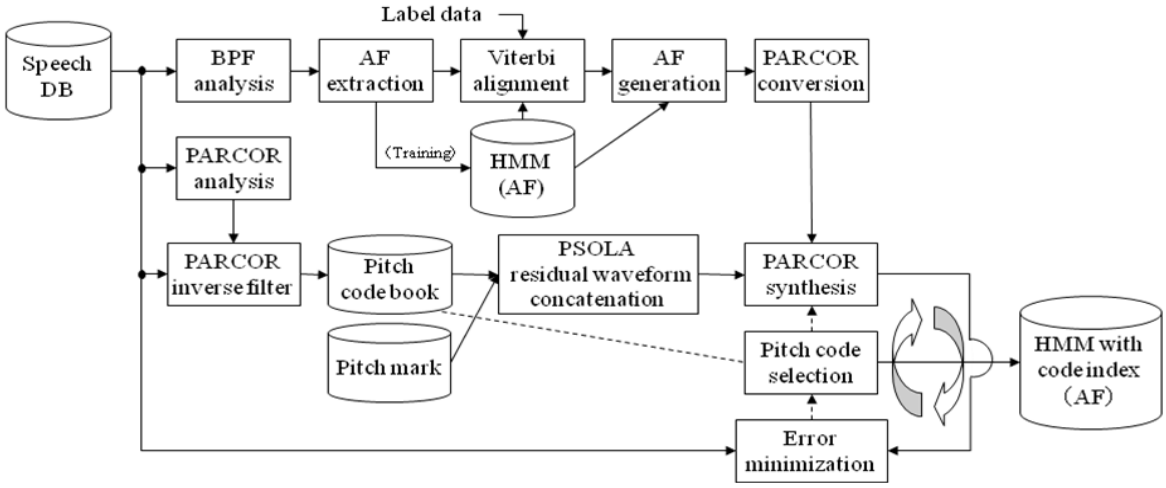

図 4 CELP に基づく音源改良

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する。

音源作成の学習データには，ATR 音素バランス文（話者 MHT 14 文）を使用した。残差符号帳は，同じ話者の 14 文（特定話者）を用いて作成したものと，不特定話者 5 名の計 70 文を用いて作成したものを比較する。符号帳のサイズは 3600 である。また，調音運動から PACOR 係数へ変換する MLN_{A-P} の学習データは，ATR 音素バランス文 50 文（同じ話者 MHT）を使用した。なお，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。

提案手法の主観評価実験を行った。(a) 原音声，(b) 従来音源（パルス&ノイズ）による合成音声，(c) 提案音源（特定話者の残差符号帳を使用）による合成音声，(d) 提案音源（不特定話者の残差符号帳を使用）による合成音声，の 4 つを比較している。被験者 11 名が音声品質を評価した結果を図 5 に示す。提案音源 (c) はパルス音源 (b) と比べて高い MOS 値を得ているが，まだ十分な値ではなく音質改善が必要である。

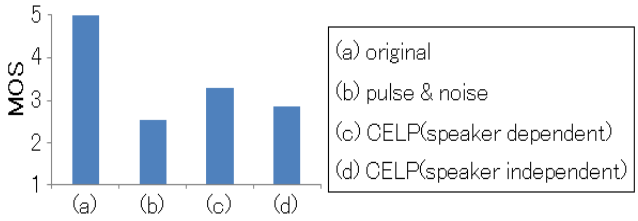

図 5 CELP に基づく音源改良の主観評価結果

4. おわりに

調音特徴を抽出し，音声認識と合成に共通に利用できる HMM の調音運動モデルから音声を合成する方式において，音声品質を改善する方策を検討し，評価実験を行った。今後，調音運動→ PARCOR 変換器と，CELP ベースの音源の双方で，一層の改良を進めたい。

参考文献

[1] Miller, J. L. and Eimas, P. D., Internal structure of voicing categories in early infancy, Percept. Psychophys., 58, 1157-1167 (1996).

[2] Liberman, A. M. and Mattingley, I. G.: The motor theory of speech perception revised, Cognition, 21, 1-36 (19845).

[3] King, S.and Taylor, P., Detection of phonological features in continuous speech using neural networks, Comput. Speech Lang., vol.14, no.4, pp.333-345 (2000).

[4] Eide, E, Distinctive features for use in an automatic speech recognition system, Proc. Eurospeech 2001, vol.III, pp.1613-1616 (2001).

[5] Kirchhoff, K. Combining acoustic and articulatory feature information for robust speech recognition, Speech Commun., vol. 37, pp.303-319 (2002).

[6] Sivadas, S and Hermansky, H., Hierarchical tandem feature extraction, ICASSP'02, vol.I, pp.809-812 (2002).

[7] Fukuda, T, Yamamoto, W. and Nitta, T, Distinctive phonetic feature extraction for robust speech recognition, Proc. ICASSP'03, vol.II, pp.25-28 (2003).

[8] 新田，武井，木村，桂田：調音運動 HMM に基づくワンモデル音声認識合成，情処研究報告,Vol. 2009-SLP-77 No.4. (2009).

[9] Miller, G. A.: The science of word, Scientific American Library (1991).

[10] Wilson, S.M., Saygm, A.P., Sereno, M.I. and Iacoboni, M., Listening to speech activates motor areas involved in speech production, Nat. Neurosci., 7, 701-702 (2004).

[11] Masuko, T., Tokuda, K., Kobayashi, T. and Imai, S., Speech synthesis from HMMs using dynamic features, Proc. of ICASSP1996, pp.389-392 (1996).

[12] Itakura, F. and Saito, S., Analysis Synthesis Telephony based on the Maximum Likelihood, 6th ICA, C-5-5 (1968).

[13] 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).

[14] 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).

[15] 福田, 山本, 新田:弁別的特徴ベクトルを用いた音声認識に関する検討, 音学講論, Vol. I, No. 1-9-1, pp. 1 – 2 (2002).

[16] Abe, M., Sagisaka, Y., Umeda, T. and Kuwabara, H., Speech Database User's Manual. *ATR Technical Report*, TR-I-0116 (1990). (in Japanese)