調音特徴に基づく 1-model 音声認識-合成

One-model Speech Recognition and Synthesis Based on Articulatory Movement HMMs

新田 恒雄[*1]　武井 匠[*1]　木村 優志[*1]　桂田 浩一[*1]
Tsuneo Nitta　Takumi Takei　Masashi Kimura　Kouichi Katsurada

[*1] 豊橋技術科学大学 大学院 工学研究科
Graduate School of Engineering, 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 models are applied. The SR engine has an articulatory feature (AF) extractor with three-stage 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. In 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.

1. はじめに

近年の HMM ベース音声認識は, 幾つかの分野で成功を収めたが, 多くは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持ち, モデル近似に多くの音声データと混合数を要する欠点がある。

一方, 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することができ, 現在の音声認識システムのように多くの話者が発話した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として, 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提唱されている [2]。調音ジェスチャを抽出し音声認識に利用する研究は, 近年, 数多く提案され [3], [4], [5], [6], [7] 多数話者音声で学習した標準的 MFCC ベース HMM を上まわる性能が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, 本文 3 節に示すように, 従来方式を上まわる性能を得ることができる。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 [8], 近年の脳研究は 1-system 説を支持する結果を示しつつある [9]。本報告では, 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法を提案する。これまで提案された標準的 HMM 音声合成は, スペクトル由来の特徴を使用するため, 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。提案方法は, 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)変換器に通した後, PARCOR 合成フィルタ [11]により合成音声を得る。

2. ワンモデル音声認識合成

図 1 に調音運動モデルに基づく音声認識合成の概要を示す。図の上側が認識エンジン, 下が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,

連絡先：新田 恒雄, 豊橋技術科学大学, 豊橋市天伯町
雲雀ケ丘 1-1, (0532) 44-6890, (0532) 44-6873,
nitta@tutkie.tut.ac.jp

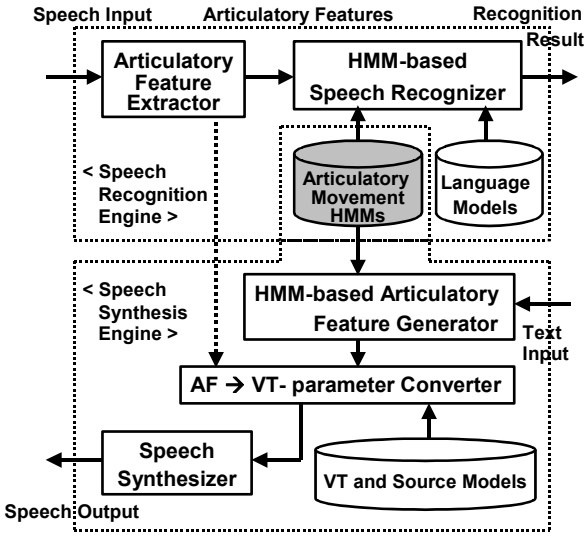

Fig. 1 One-Model Speech Recognition and Synthesis Based on Articulatory Movement Models.

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することで得られる (別に音源モデルが必要)。提案方式は, また, 図に示されているように, 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, 語学学習に利用できる。

3. 調音運動 HMMs に基づく音声認識

ワンモデルの音声認識エンジンは, 図 2 に示すように, 入力

音声を AF 系列に変換する AF 抽出器と, 調音運動を表現した HMM (音素) 分類器から成る。入力音声は 16kHz でサンプリングされた後, 25ms のハミング窓で 10ms 毎に, 512 点の FFT 処理を受ける。この結果はパワースペクトルの形で積分され, 中心周波数を (聴覚に近似した) メル尺度間隔に設計した 24-ch の BPF (Band Pass Filter) 出力にまとめられる。ここまでが分析処理である。続いて音響特徴抽出が行われる。パワースペクトルの時系列が構成する曲面は, 多様体として見ると時間と周波数方向の局所的な微分要素で表現できる (微分多様体)。そこで, BPF 出力を 3×3 の局所特徴に変換するため, 時間軸と周波数軸で 3 点の線形回帰 (Linear Regression; LR)演算を行い, 微分特徴としての局所特徴 (Local Feature; LF) を得る [7]。二つの局所特徴は各 24 次元であるが, 次に, 離散余弦変換 (Discrete Cosine Transform; DCT) の処理によって, 半分の 12 次元に圧縮される。これに対数パワー成分の微分を加えた 25 次元の特徴を, 以後局所特徴 LF と呼ぶ (Δt, Δf, ΔP)。

微分多様体としての音声パターンから大域的な (調音) 特徴を取り出すためにニューラルネットを利用する。MLN の入力として, パワースペクトルの濃度情報から計算する MFCC (Mel-Frequency Cepstrum Coefficients) を利用することが多い。Cepstrum は, BPF 出力の対数値を DCT することで計算される。BPF 出力値が互いに従属しているという欠点を解消できるため, HMM の計算とも相性がよく, ほとんどの音声認識装置は, 現在, MFCC を使用している。先に説明した局所特徴 LF と MFCC を, 調音特徴抽出器の MLN 入力で比較すると LF が勝つ。この理由は MLN の使い方にも依るため, ここではこれ以上述べない。

調音特徴 AF を抽出するために MLN を 3 段階に分けて使用した。1 段目は単純に注目フレームの調音特徴を抽出する MLN と, 音素境界で目に付く分類誤りを補正するために, 少し長い時間の情報を入れ AF context の制約を使えるようにした MLN を組合わせている。図 3 はこの出力の例で,「人工衛星」に対する調音抽出の結果である。ここでは, 調音特徴として, 半母音, 鼻音, 無声音, 有声音, 持続性, 破擦性, 破裂性, 舌端性, 後舌母音, 前方性, 低母音, 高母音, ほかを使用している。/N/は有声で鼻音, /k/は無声で破裂音, ・・・ということが分かる。MLN はこうした特徴が出るように学習させている。

2 段目は, Inhibition と Enhancement の動作を利用しており, 調音動作の加速度成分により, 調音点が目標に接近しているか, 遠ざかっているかによって制御している。最後に 3 段目は, 特徴間の独立性 (直交性) を保持する処理で, Gram-Scmidt の直交化を利用している。

■ 調音特徴の評価

< 音声試料 >

D1: 学習セット-1 （MLNs 学習用）

日本音響学会 (ASJ)の連続音声データベース

4,503 文, 男声 30 名 (16 kHz, 16 bit) [14].

D2: 学習セット-2 （HMMs 学習用）

日本音響学会新聞記事読み上げコーパス (JNAS) [15].

5,000 文, 男声 33 名 (16 kHz, 16 bit).

D3: 評価セット

日本音響学会新聞記事読み上げコーパス (JNAS)

2,719 文, 男声 17 名 (16 kHz, 16 bit).

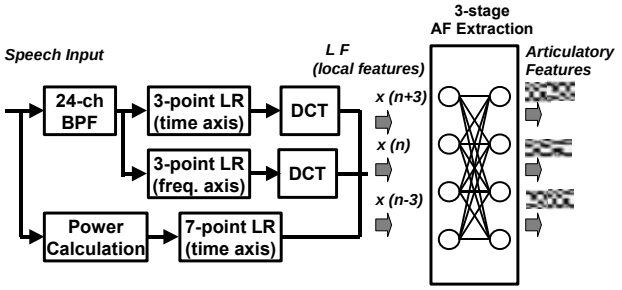

Fig. 2 Articulatory Feature Extraction

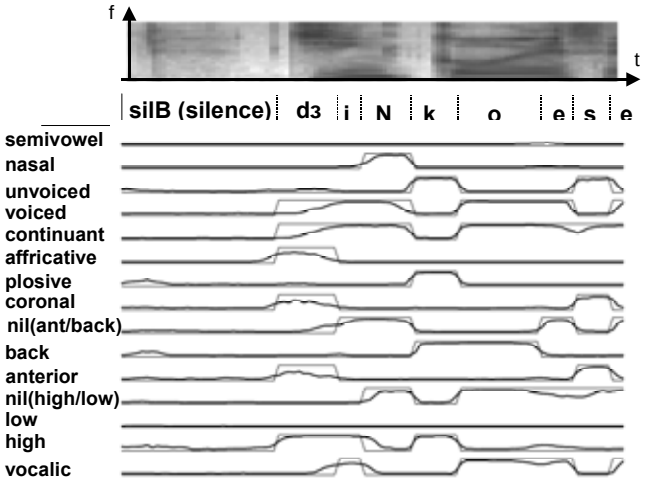

Fig. 3 Articulatory Feature Sequence

: /jiNkoese (artificial satellite)/

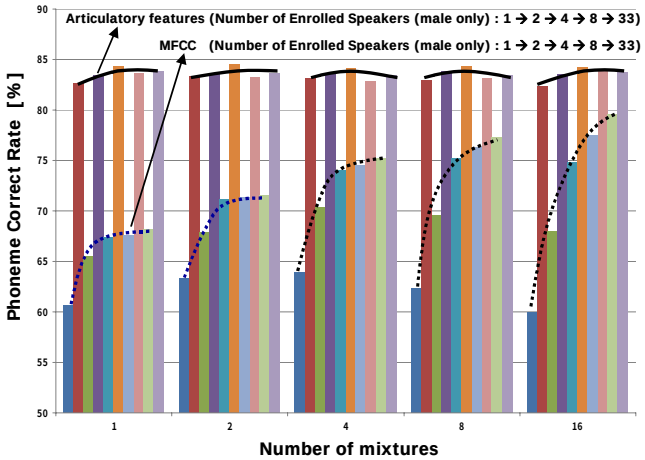

Fig. 4 Phoneme Correct Rate

vs. Number of Mixtures and Enrolled Speakers.

< 評価実験結果 >

音素認識率を評価した。HMM は 5 ステート 3 ループの標準的な left-to right 型を使用した。単音(mono-phone)単位で, 混合数を 1, 2, 4, 8, 16 とし, 学習に使用した話者は 1 名→ 2 名

→ 4 名→ 8 名→ 33名と増加させながら, D3 セットの音素認識性能を調べた。結果を図 4 に示す。

調音特徴は登録人数に関係なく, また当然, 混合数にも無関係である。これに対して MFCC は, 登録人数を増やし, 同時に正規分布の数を増やすほど向上する。この結果から, 調音特徴は話者不変のパラメータであることが示唆される。

4. 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[10]。このため, 近年は効率をよくするための工夫が話者適応など種々行われている。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, 調音特徴は前節でみたように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は利点があると考えられる。

4.1 HMM ベース音声合成

図 5 は調音特徴を使用した音声合成を示している。HMM は音声認識のため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。これによって, 滑らかな音声が生成できる。

4.2 調音特徴から声道パラメータへの変換と評価

図 5 で, 調音パラメータは PARCOR 係数に変換され, 結果が PARCOPR 合成器（フィルタ）に送られる。変換に用いる MLN は, 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)で, 隠れ層のユニット数は 450 である。学習には ATR 音素バランス文の中の 1 話者を使用している (使用した読み上げ文の数は 50)。

図 6 に(a) 元の音声, (b) 調音抽出器の出力から採った調音特徴系列を MLN に入力して得た音声, (c) 調音運動 HMM から合成された音声のスペクトル（PARCOR 分析）を示した。(b), (c)は(a)の元の音声と比較すると, 平滑されているが, スペクトル上のホルマントなどの特徴は保存されていることがわかる。今回は, 音源としてパルス列と白色雑音を使用した。11 名の被験者に音質を確認しもらったところ, 音節の違いは十分確認できた。今後は, MOS 値などの定量的な評価を行うと共に, 音源の改良を進めたい。

5. おわりに

調音特徴を抽出し, 音声認識と合成に共通して利用可能なモデルの構築を検討した。今後は, 合成エンジンの音質改良とともに, 認識エンジンの頑健化を進め, エンジンソフトを完成させたい。

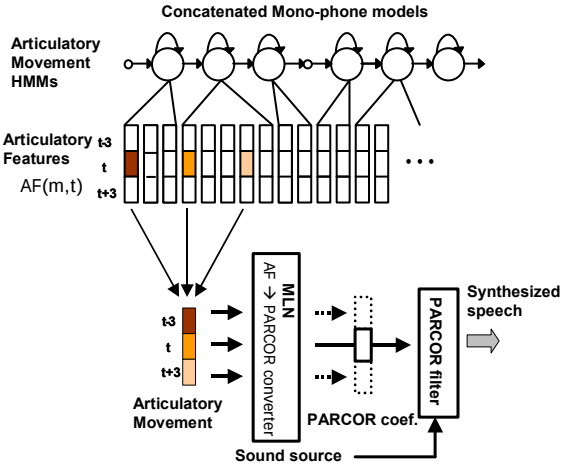

Fig. 5 Proposed HMM-based Speech Synthesis with Articulatory Movement Models.

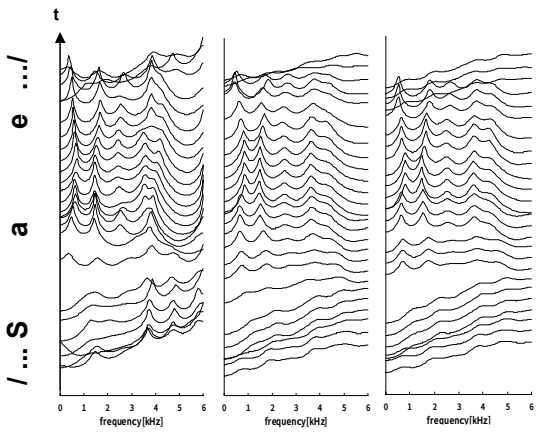

Fig. 6 Comparison of Spectrum Envelope:/…sae(s)…/

参考文献

[Miller 96] Miller, J. L. and Eimas, P. D., Internal structure of voicing categories in early infancy, Percept. Psychophys., 58, 1157-1167 (1996).

[Liberman 45] Liberman, A. M. and Mattingley, I. G.: The motor theory of speech perception revised, Cognition, 21, 1-36 (19845).

[King 00] 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).

[Eide 01] Eide, E, Distinctive features for use in an automatic speech recognition system, Proc. Eurospeech 2001, vol.III, pp.1613-1616 (2001).

[Kirchhoff 02] Kirchhoff, K. Combining acoustic and articulatory feature information for robust speech recognition, Speech Commun., vol. 37, pp.303-319 (2002).

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

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

[Miller 91] Miller, G. A.: The science of word, Scientific American Library (1991).

[Wilson 04] 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).

[Masuko 96] Masuko, T., Tokuda, K., Kobayashi, T. and Imai, S., Speech synthesis from HMMs using dynamic features, Proc. of ICASSP1996, pp.389-392 (1996).

[Itakura 68] Itakura, F. and Saito, S., Analysis Synthesis Telephony based on the Maximum Likelihood, 6th ICA, C-5-5 (1968).

[Huda 08] 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).

[Huda 09] 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).

[Kobayashi 92] Kobayashi, T., Itahashi, S., Hayamizu, S. and Takezawa, T. "ASJ Continuous Speech Corpus for Research," Acoustic Society of Japan Trans. Vol.48, No.12, pp.888-893 (1992).

JNAS: Japanese Newspaper Article Sentences.
http://www.milab.is.tsukuba.ac.jp/jnas/instruct.html

[Abe 90] Abe, M., Sagisaka,Y., Umeda, T. and Kuwabara, H., Speech Database User's Manual. *ATR Technical Report*, TR-I-0116 (1990). (in Japanese)