Vol.2009-SLP-77 No.4
2009/7/17

調音運動 HMM に基づくワンモデル音声認識合成

新田恒雄[†]　武井匠[†]　木村優志[†]　桂田浩一[†]

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。本報告では，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する。音声認識エンジンは，3 段階の多層ニューラルネットから成る調音特徴抽出器を持ち，音声から調音特徴を高精度に抽出する。調音運動を表現する HMM は，1 名の学習にもかかわらず，他の話者に対して高い音素認識精度を達成した(実験は男性のみ)。また，音声合成エンジンでは，同じ HMM が出力する調音特徴系列を，声道パラメータ(PARCOR 係数)に変換することにより，明瞭な音声を生成することが可能になる。

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

Tsuneo Nitta[†], Takumi Takei[†], Masashi Kimura[†]
, and Kouichi Katsurada[†]

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

HMM に基づく音声認識は，近年，幾つかの分野で成功を収めたが，多くが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る結果，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，本文に示すように，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式を提案する。これまでに提案された標準的 HMM 音声合成[10]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)変換器に通した後，PARCOR 合成フィルタ [11]により合成音声を得る。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。

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[12],[13]，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。

[†] 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.2009-SLP-77 No.4
2009/7/17

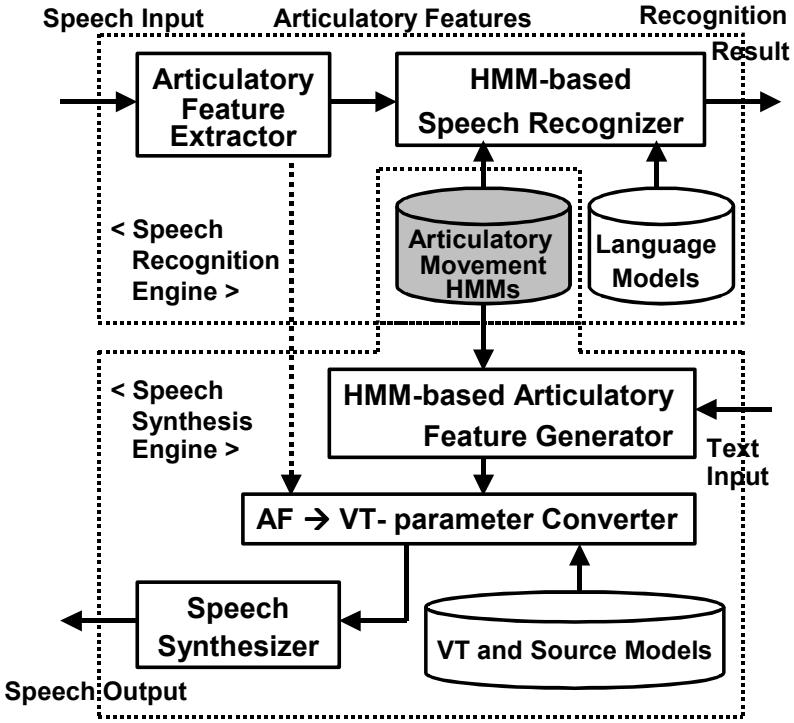


図 1　調音運動に基づく One-Model 音声認識・合成システム

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し，音源信号で駆動することで得られる。

提案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。

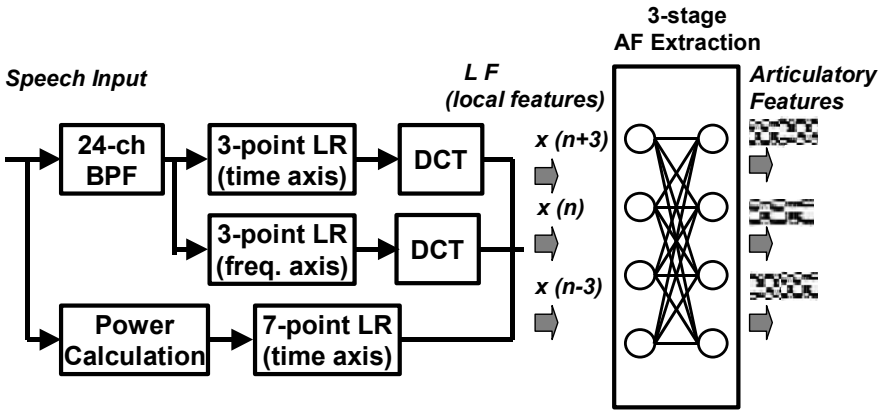


図 2 調音特徴抽出器の構成

3. 調音運動 HMMs に基づく音声認識

3.1 音声認識エンジンの構成

ワンモデル音声認識エンジンは，入力音声を AF 系列に変換する AF 抽出器（図 2 参照），および調音運動を表現した HMM（音素）分類器から成る。入力音声は，従来の音声認識処理と同様，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）メル尺度間隔で設計した 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。続いてパワースペクトル系列上の音響特徴抽出が行われる。パワースペクトル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，時間軸と周波数軸上で各々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) を抽出する [7]。二つの局所特徴は各 24 次元であるが，続いて離散余弦変換 (Discrete Cosine Transform; DCT)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要素を加えた 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，MLN を適用する。音声の特徴抽出に利用される MLN では，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) を使用することが多い。Cepstrum は，BPF 出力の対数値を DCT することで計算される。MFCC は，BPF の出力値が互いに従

Vol.2009-SLP-77 No.4
2009/7/17

属しているという欠点を解消し，正規分布を仮定した HMM の尤度計算と相性がよく，現在，多くの音声認識装置が利用している。一方，先に説明した局所特徴 LF と MFCC を，調音特徴抽出器 MLN の入力信号として比較した結果によると，LF が MLN に対し優位であることが示されている[14]。

調音特徴 AF を抽出するために MLN を 3 段階に分けて使用する。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，前後の AF 情報を入れ context の制約を入れた MLN を組合わせている。図 3 はここまでの出力の例で，「人工衛星」に対する調音抽出の結果である。今回は調音特徴として，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，ほかを使用した。/N/は有声で鼻音，/k/は無声で破裂音，・・・といった動作が観測できる。MLN はここに述べた特徴が出せるようパラメータを調整して学習させている。

2 段目は，Inhibition と Enhancement の動作を利用している。具体的には調音動作の加速度成分から，調音点が目標に接近しているか（ΔΔが負），遠ざかっているか（ΔΔが正）を検出し，図 4 に示すシグモイド関数を用いて動作を制御(f (ΔΔ) を元の調音動作に乗算)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。

3.2 音声認識性能評価

調音特徴(45 次元) を MFCC(Δ，ΔΔ，ΔP，ΔΔP; 38 次元) と比較する。音声試料は次の 3 セットを用いた。

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) [15].

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) [16].
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).

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→ 100 名(MFCC のみ)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 5 に示す。

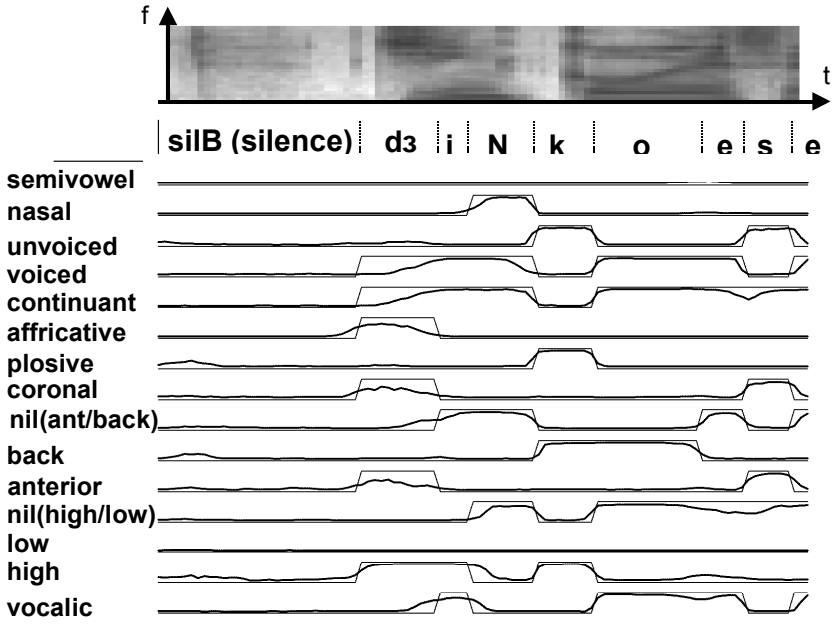


図 3 調音特徴系列: 「人工衛星」/jiNkoese (artificial satellite)/

$f(\Delta\Delta) = 1/[1 + \exp(\Delta\Delta - t_{d1})]$ $\Delta\Delta > t_{d1}$ 抑制

$f(\Delta\Delta) = 1/[1 + \exp(\Delta\Delta + t_{d2})]$ $\Delta\Delta < t_{d2}$ 強調

$f(\Delta\Delta) = 0.5$ $t_{d2} \le \Delta\Delta \le t_{d1}$

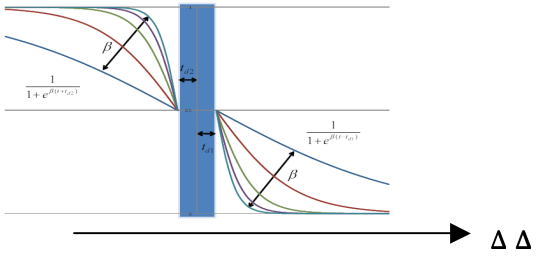


図 4 調音動作の加速度による制御

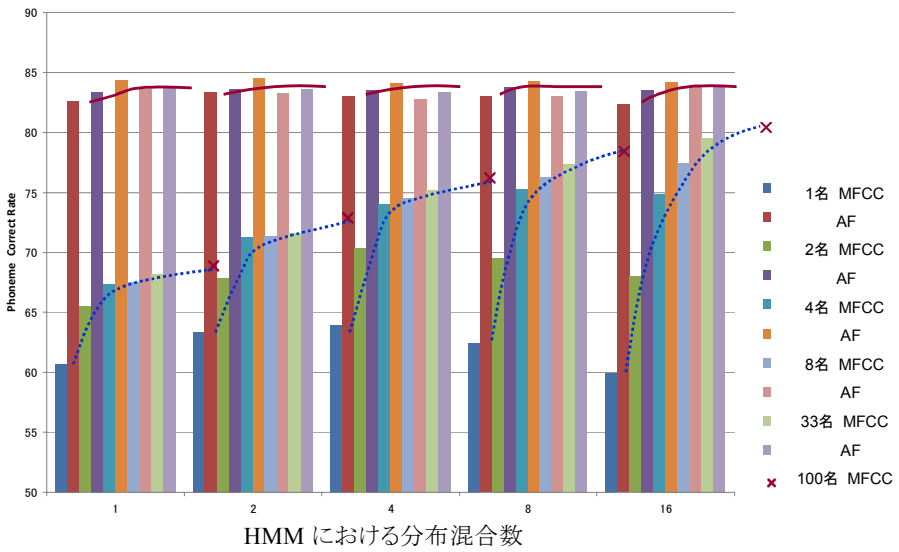


図 5　登録話者数と音素認識性能の比較
（破線：MFCC，実線：AF (調音特徴)）

調音特徴 AF は登録人数に関係なく，混合数も 1 混合で高い音素認識を達成している。これに対して MFCC は，登録人数を増やし，同時に混合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利点が大きいと考えられる。

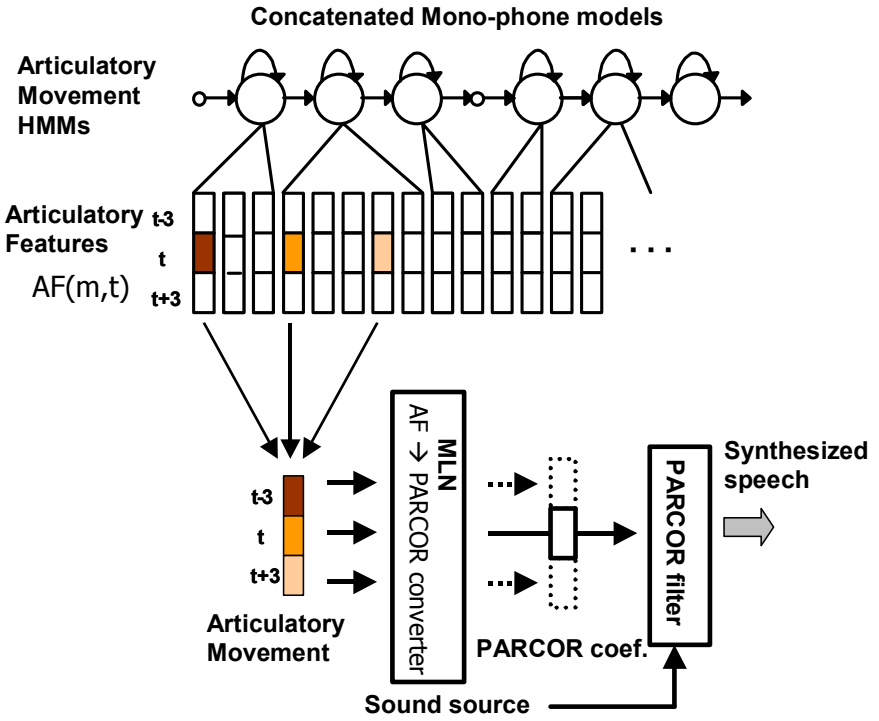


図 6 調音運動 HMM に基づく音声合成

4.1　HMM ベース音声合成

図 6 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 6 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) [17]。

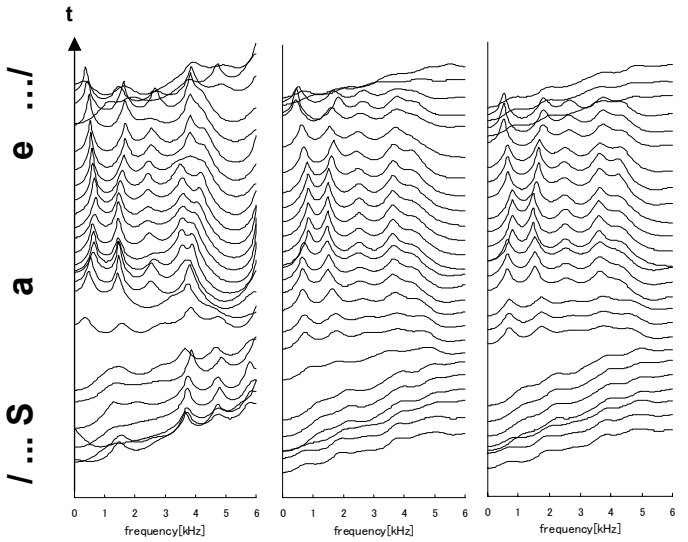
Vol.2009-SLP-77 No.4
2009/7/17

図 7 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, 音源にパルス列と白色雑音を使用した。図 8 に, 元の音声(1 発話)から抽出した PARCOR 係数と, 調音特徴系列を MLN で変換した係数の相関値を示す。 (1) 調音特徴系列から MLN により変換した PARCOR 係数と, (2) AF(図では DPF) でモデル化した HMM から生成した調音特徴系列を MLN に通して得た PARCOR 係数との差は小さいと言える。ATR 音素バランス文から話者 1 名(MHT(B)) の 50 文を使用し, MLN を学習して 11 名の被験者に音質を確認してもらったところ, 音節の違いは十分確認できた。ただし, MOS 値は原音声の 5 に対して平均 3 程度とまだ低い。今後, MLN と音源の改良に注力する必要がある。

5. 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 の調音運動モデルを検討した。音声認識エンジンでは, 音声から調音特徴を高精度に抽出することにより, HMM の学習が 1 名でも高い音素認識精度を達成できることを示した。また, 音声合成エンジンでは, 同じ HMM が出力する調音特徴系列を, 声道パラメータ(PARCOR 係数)に変換することにより, 明瞭な音声を生成することが可能なことを示した。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(対音素コンテキスト, および対騒音) を進めたい。また現在は, 調音特徴として音韻論に基づく弁別的素性に, 音声学に基づく調音方法と調音位置を加味した特徴を用いているが, 語学学習などマルチリンガル対応が要請されており, 言語に依らず導入が容易な調音特徴に統一する必要があると考えている。



(a) Original speech　(b) Converted from AF seq. directly　(c) Converted from AF seq. Output by HMM

図 7 音声スペクトラム包絡の比較　:/…sae(s)…/

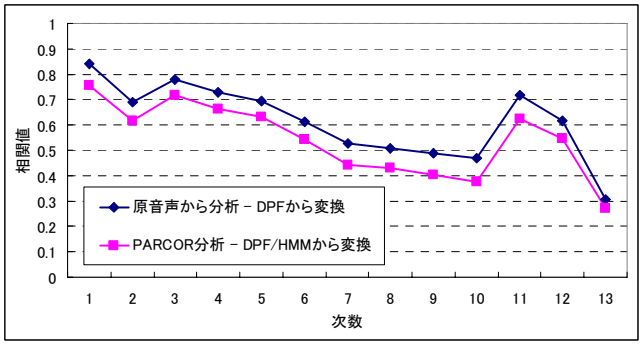


図 8 PARCOR 係数間の相関比較

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] Miller, G. A.: The science of word, Scientific American Library (1991).
[9] 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).
[10] Masuko, T., Tokuda, K., Kobayashi, T. and Imai, S., Speech synthesis from HMMs using dynamic features, Proc. of ICASSP1996, pp.389-392 (1996).
[11] Itakura, F. and Saito, S., Analysis Synthesis Telephony based on the Maximum Likelihood, 6th ICA, C-5-5 (1968).
[12] 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).
[13] 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).
[14] 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).
[15] 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).
[16] JNAS: Japanese Newspaper Article Sentences. http://www.milab.is.tsukuba.ac.jp/jnas/instruct.html
[17] Abe, M., Sagisaka,Y., Umeda, T. and Kuwabara, H., Speech Database User's Manual. *ATR Technical Report*, TR-I-0116 (1990). (in Japanese)