Vol.2015-MUS-107 No.74
2015/5/24

混合ガウス分布を用いた自然音声への人工感の付与

小林 航也[1,a)]　齋藤 大輔[2,b)]　峯松 信明[1]　広瀬 啓吉[2]

概要：音楽や映像制作の場では人間の声を対象として様々な加工を施すことがある．それらは，音声の収録環境とは異なる音響環境を再現したり，音圧調整等によって聴感上の印象を改善するだけでなく，基本周波数の補正を強くかけることで自然音声では起こりえない音声の変化を与える事もある．本研究では新たな音声加工法として自然音声に人工感を与える技術構築を目的とする．基本周波数だけでなく音声の音色にあたるケプストラムも聴感を左右する重要な要素である．本報告では音声のケプストラム系列を混合ガウス分布を用いて離散的な系列として出力する手法を提案する．人工感に関する主観評価実験の結果，提案手法により発話内容の事前情報なしに，入力された自然音声に対して人工感を付与しうることが示された．

Koya Kobayashi[1,a)]　Daisuke Saito[2,b)]　Nobuaki Minematsu[1]　Keikichi Hirose[2]

1. はじめに

現在音楽や映像制作の場では人間の声を対象とした様々なエフェクターが存在する．それらにはコーラスやフェイザーなど声に揺れを与えるものや，リバーブやディレイなど残響感を与えるもの，ディストーションなど音を歪ませるものや，ピッチを補正する効果を持つものなどがある．これらのエフェクターは，音声の収録環境とは異なる音響環境を再現したり，音圧調整等によって聴感上の印象を改善するだけでなく，自然音声では起こりえない音声の変化を与える事も可能である．例えば，基本周波数を意図的に階段状に補正することによって，人間の生理機構では生成しえないような基本周波数パターンの音声となる．結果的に人間の音声が「加工されている」という印象を与える事になる．このように，音声が加工されたように聴取者が感じる印象を本研究では「人工感」と呼び，自然音声に人工感を付与する加工技術について議論する．本稿では特に，音声の音色／スペクトル情報の加工について述べる．

近年テキストを入力とし音声を自動的に合成する，テキスト読み上げ音声システム（Text-To-Speech; TTS）が実用段階に近づきつつある．現在音声合成で注目を集めているのは統計的手法の一つである隠れマルコフモデル（Hidden Markov Model; HMM）に基づく音声合成である．HMM音声合成はパラメトリック音声合成の一つであり，テキスト系列と音声の対応した学習用のコーパスから，各ラベルに対応する HMM のモデルパラメータを学習する．合成時には，入力されたテキストに対応する HMM から確率最大で出力される特徴量系列を用いて音声を合成する．HMM音声合成では確率分布のモデルパラメータに対して，話者適応や変換処理を行う事で合成する音声の話者性，感情，発話スタイル等を柔軟に制御可能である [1][2].

しかし，HMM 音声合成は波形接続型の音声合成と比べて，音声のパラメータ化や統計的モデリングに伴って，自然性が劣化してしまう．これまでに HMM 音声合成の品質向上のため様々な改良が行われてきた [3]．これらの手法は人工的な音声をより自然音声に近づけるプロセスと捉える事ができる．人工感の付与の定義から，本研究ではこれらの逆プロセスを考えることでその実現をはかる．

HMM 音声合成の原理並びに品質向上のための技術における重要な点として，1）状態単位での最尤パラメータの出力，および 2）動的特徴量の導入による時間的不連続性の緩和が挙げられる．1）についてはスペクトルの特徴量系列が原理的には離散的に出力される事になり，これが前述の人工感付与につながりうる．2）については，離散的な系列に動的特徴量の制約を加える事で，時間的な不連続性が緩和され自然性の向上につながるが，一方で時系列全体では系列内変動が縮退しうる．そこで本稿ではこれらをふまえ，事前に学習した混合ガウス分布を用いて，入力された自然音声を離散的な系列に変換する事によって，自然

[1] 東京大学大学院工学系研究科
113-8656　文京区本郷 7-3-1
[2] 東京大学大学院情報理工学系研究科
[a)] kobakoba@ginjo.t.u-tokyo.ac.jp
[b)] dsk_saito@gavo.t.u-tokyo.ac.jp

音声に人工感を付与する手法を提案する．またその際に動的特徴量 [3] を考慮した場合の影響についても検証する．

2. HMM 音声合成

2.1 HMM 音声合成の概要

HMM 音声合成はテキスト音声合成（Text-to-speech; TTS）の一つであり，統計的パラメトリック合成手法である．HMM 音声合成では学習用音声からスペクトル包絡特徴量と基本周波数を抽出し，それぞれの特徴量を連結したベクトルデータ $\boldsymbol{O}$ とテキストから抽出したラベルデータ W から式 (1) の最尤基準で HMM のモデル λ を学習する．

$$\hat{\lambda} = \underset{\lambda}{\operatorname{argmax}}\, p(\boldsymbol{O}|W, \lambda) \tag{1}$$

合成時は，入力テキスト w から抽出されたラベルに対して式（2）を用いて HMM から尤度最大化基準によって生成された特徴量を用いて合成音声 $\boldsymbol{o}$ を作成する．

$$\hat{\boldsymbol{o}} = \underset{\boldsymbol{o}}{\operatorname{argmax}}\, p(\boldsymbol{o}|w, \hat{\lambda}) \tag{2}$$

2.2 隠れマルコフモデル

隠れマルコフモデル（Hidden Markov Model: HMM）は，各単語や音素を標準的な確率状態遷移で表すモデルである．このモデルにはスペクトル時系列の統計的変動をモデルのパラメータに反映させることができる特徴がある．a_{ij} は状態 S_i から状態 S_j への状態遷移確率を表し，状態数を S とすると $S \times S$ の行列で表現できる．通常音声パターンは非可逆なので $i > j$ なら $a_{ij} = 0$ である．また，$b_i(o_t)$ は各状態 i での特徴量 o_t を出力する確率で，通常ガウス分布でモデル化される．HMM 音声合成では o_t が各フレームにおける音声の特徴量の系列を表す．

また HMM がメルケプストラムなどの静的なパラメータだけで構成されていると，ガウス分布を出力分布とする HMM において各時刻毎に最尤な出力は，現在の状態の平均ベクトルとなるため，最尤な出力系列が階段状の不連続な系列となってしまう．そのため Δ ケプストラムパラメータを組み合わせるなどして時間的な制約を加えることで，滑らかで連続的なケプストラム系列を得るための研究が行われている [3]．

2.3 動的特徴量

先に述べたように HMM が静的なパラメータのみで構成されていると，出力ベクトル系列は階段状の不連続な系列になる．この問題を解決するために用いられるのが動的特徴量である．以下の式で示すような 1 次微分，2 次微分を特徴量として加える．

$$\boldsymbol{o_t} = [\boldsymbol{c_t}^\top, \Delta\boldsymbol{c_t}^\top, \Delta^2\boldsymbol{c_t}^\top] \tag{3}$$

$$\Delta\boldsymbol{c_t} = \tfrac{1}{2}(\boldsymbol{c_{t+1}} - \boldsymbol{c_{t-1}}) \tag{4}$$

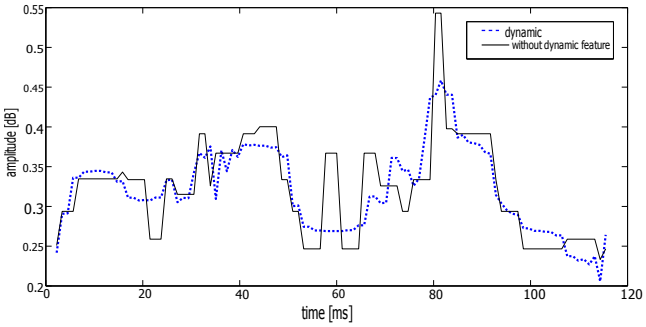

図 1 動的特徴量の有無による出力メルケプストラム系列の変化

$$\Delta^2\boldsymbol{c_t} = \boldsymbol{c_{t-1}} - 2\boldsymbol{c_t} + \boldsymbol{c_{t+1}} \tag{5}$$

動的特徴量を加えたパラメータ生成は式 (6) のようになる．

$$\hat{\boldsymbol{c}} = \underset{\boldsymbol{c}}{\operatorname{argmax}}\, P(\boldsymbol{W}\boldsymbol{c}|\lambda) \tag{6}$$

ここで $\mathbf{c} = [\mathbf{c}_1^\top, \cdots, \mathbf{c}_T^\top]^\top$ は T フレームまでの音声パラメータ系列，$\mathbf{c}_t = [c_t(1), \cdots, c_t(D)]^\top$ は時刻 t における D 次元の音声パラメータ，W は動的特徴量の計算に用いる重み係数によって決定される行列，λ は HMM のパラメータセットを表す．

動的特徴量を考慮したパラメータ生成の効果を図 1 に示す．動的特徴量を考慮する事で，特徴量系列の不連続が解消され，滑らかな軌跡となっていることがわかる．しかし，式 (6) により生成されるパラメータは過剰に平滑化される傾向にあり，HMM 音声合成において生成されるパラメータ系列の過剰なスムージングが音質劣化の一因となることが知られている．

3. GMM を用いたメルケプストラムの離散化

3.1 混合ガウス分布モデル (Gaussian Mixture Model; GMM)

GMM は複数のガウス分布の重み付き和で表される確率分布であり d 次元のベクトル $\boldsymbol{x}$ が混合数 K の GMM から生成されるとき，その確率密度は

$$P(\boldsymbol{x}) = \sum_{k=1}^{K} P(k, \boldsymbol{x_k}) = \sum_{k=1}^{K} P(k)P(\boldsymbol{x_k}|k) \tag{7}$$

$$= \sum_{k=1}^{K} \pi_k N(\boldsymbol{x}; \mu_k, \Sigma_k) \tag{8}$$

のように表現される．$K, \pi_k, \mu_k, \Sigma_k$ はそれぞれガウス分布の混合数，混合重み，平均，分散である．

3.2 提案手法

本節では，GMM を用いた特徴量系列の離散化について述べる．音声の音韻性はメルケプストラムに現れるため，適当な混合数のもとで，作成されるガウス分布の一つ一つがおよそ音素や単音などの音響的な単位を表す事が期待できる．提案手法の流れを図 2 に示す．提案手法では，ある音声の特徴量系列 $\boldsymbol{o}$ を入力し，時刻 t の特徴量 $\boldsymbol{o_t}$ について学習された K 混合のガウス分布に対する事後確率を式

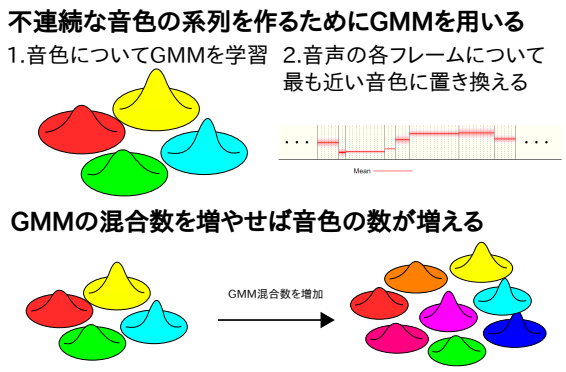

図 2　メルケプストラム離散化の流れ

(9) を用いてすべてのインデックス k について計算する．

$$P(k|\boldsymbol{o_t}) = \frac{P(k, \boldsymbol{o_t})}{P(\boldsymbol{o_t})} = \frac{\pi_k N(\boldsymbol{o_t}; \mu_k, \Sigma_k)}{\sum_{k=1}^{K} \pi_k N(\boldsymbol{o_t}; \mu_k, \Sigma_k)} \qquad (9)$$

このとき最も事後確率の高かったインデックス k の平均ベクトルで特徴量ベクトルの系列を置き換えるという処理をすべての時刻について計算し，特徴量系列 $\acute{\boldsymbol{o}}$ を得る．また動的特徴量を用いる場合，時刻 t における観測ベクトルは D 次元の静的特徴量ベクトル $\boldsymbol{c_t} = [c_t(1), \cdots, c_t(D)]^\top$ と，その 1 次の動的特徴量ベクトル $\Delta \boldsymbol{c_t}$ の結合ベクトル $\boldsymbol{o_t} = [\boldsymbol{c_t}^\top, \Delta \boldsymbol{c_t}^\top]$ で表す． これを用いて GMM を学習し，式 (9) から特徴量系列 $\acute{\boldsymbol{o}}$ を得る．これを用いて静的特徴量ベクトル系列 $\hat{\boldsymbol{c}}$ を次式にて生成する．

$$\hat{\boldsymbol{c}} = \underset{\boldsymbol{c}}{\operatorname{argmax}} P(\acute{\boldsymbol{o}}|\lambda) \quad \text{subject to } \boldsymbol{o} = \boldsymbol{W}\boldsymbol{c} \qquad (10)$$

ここで，$\boldsymbol{W}$ は静的特徴量系列から静的・動的特徴量系列へ変換するための行列を表す．最尤系列 $\hat{\boldsymbol{c}}$ は以下に示す通り求められる．

$$\hat{\boldsymbol{c}} = \left(\boldsymbol{W}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{W}\right)^{-1} \boldsymbol{W}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}^\top \qquad (11)$$

ここで $\boldsymbol{\mu}$ はガウス分布の平均ベクトル，$\boldsymbol{\Sigma}$ は共分散行列を全ての時刻について並べたものを表す．

提案法においては，自然音声を入力として，事前学習された GMM に基づいて離散的なパラメータ系列を生成する．一方 HMM 音声合成では，テキストを入力としてあらかじめ学習されたモデルのもと音声パラメータ系列を生成する．提案法および HMM 音声合成を，人工感のある音声の生成器として捉えた場合，HMM 音声合成はテキストを入力とするパラメータ生成系，また提案法は音声を入力とするパラメータ生成系である．このとき，提案法では音声を入力とするため，発話内容に関する事前情報を必要としない点が，HMM 音声合成と大きく異なる．

4. 実験

4.1 実験条件

音声データとして ATR 音素バランス文 [4] を女性 1 名が発話したものを用いた．各データから STRAIGHT[5] を用いて，サンプリング周波数 16kHz，フレーム周期 1ms の条件でピッチ，STRAIGHT スペクトル，非周期性指標を得た．この STRAIGHT スペクトルから得られる 40 次元のメルケプストラムとその Δ パラメータ 40 次元を音響特徴量とした．このうちサブセット A–I の 450 文を用いて GMM を学習した．GMM の作成には Hidden Markov Model Toolkit (HTK)[*1] を用い，4,8,16,32,64 混合の GMM を作成した．HMM 音声合成ではサブセット A–I の 450 文を用い，音響的特徴量としてメルケプストラム 40 次元，平均非周期性指標 5 次元，対数基本周波数 1 次元および，それぞれの Δ,Δ^2 パラメータを用いて 5 状態 left-to-light 型の HMM を学習し，J セット 53 文を合成した．

今回作成した音声の種類は (1) 提案法において動的特徴量を用いずに GMM によって系列を離散化した音声，(2) 提案法において動的特徴量を考慮し，系列を離散化した上で式 (11) でパラメータを生成した音声，(3) (2) で作成したメルケプストラム系列をスペクトル情報とし，基本周波数について 60Hz から 400Hz まで 12 段階に離散化した音声，(4) HMM 合成音声，(5) STRAIGHT により分析，再合成のみを行い他に加工をしない音声である．

これらに対して 10 人の被験者によって音声から受ける人工感について，1: 人工感が感じられない，2: 人工感があまり感じられない，3: どちらともいえない，4: 人工感が少し感じられる，5: 人工感が感じられる，の 5 段階で主観評価を行った．また発話内容の了解性について，1: 発話内容が聞き取れない，2: 発話内容があまり聞き取れない，3: どちらともいえない，4: 発話内容が少し聞き取れる，5: 発話内容が十分聞き取れる，の 5 段階で主観評価を行った．

4.2 実験結果

図 3 に GMM の混合数を変化させた場合における提案法の人工感に関する結果を示す．図 3 より，提案法において動的特徴量を用いない場合は，混合数に関わらず中程度の人工感が知覚されていることがわかる．一方，動的特徴量を考慮する事で，混合数が 16 以上の場合に，人工感が減少していることが分かる．このことから時間的な不自然性と人工感との関係が示唆される．

図 4 は混合数 64 における提案法と，参照となる HMM 音声合成並びに分析再合成音を比較したものである．分析再合成音に比べると，その他の音声は統計的モデリングの導入に伴って，人工感が現れている事がわかる．また，動的特徴量を考慮していない場合の提案法によって，スペクトルの加工のみによってもある程度の人工感を付与できている．一方 HMM 音声合成および基本周波数の離散化による人工感が大きい事から，韻律に起因する自然音声との乖離が，人工感に大きな影響を与えていることが示唆される．

[*1] http://htk.eng.cam.ac.uk/

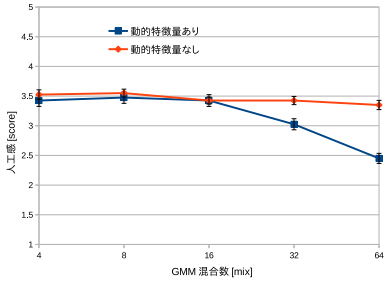

図 3 GMM 混合数を変化させたときの音声の人工感

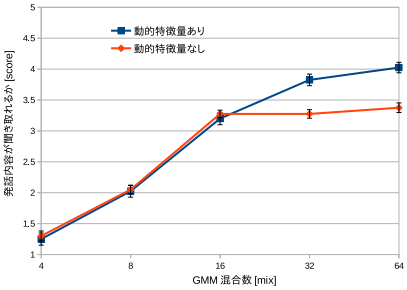

図 5 GMM 混合数を変化させたときの音声の了解性

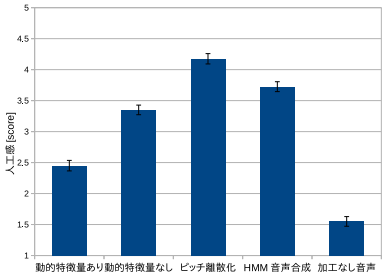

図 4 手法ごとの音声の人工感

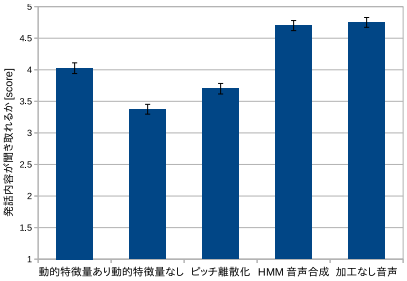

図 6 手法ごとの音声の了解性

図 5 に GMM の混合数を変化させた場合における提案法の了解性に関する結果を示す．了解性については動的特徴量なしの場合 GMM の混合数が 16 以上のとき，了解性の結果が飽和している．これは日本語の音素数はおよそ二十数個といわれており，GMM の混合数が 16 程度になると日本語音声の言語内容の知覚が可能な表現力が得られたためと考えられる．また混合数 64 の場合を比べてみると，動的特徴量を用いメルケプストラムの時間的不連続性を解消することで了解性が改善することがわかる．

図 6 は了解性について，混合数 64 における提案法と，参照となる HMM 音声合成並びに分析再合成音を比較したものである．図 6 から，HMM 音声合成は分析再合成音と同程度の了解性を達成している．一方で離散化を伴う提案法においては，一定程度の了解性の低下が見られる．特にスペクトル特徴量の離散化である提案法は，基本周波数の離散化よりも了解度低下に影響を与えている．

以上の結果から，スペクトル情報を離散的にすることで一定程度の人工感を付与可能であることが分かった．一方で韻律の自然性が人工感に強く影響している可能性も示唆された．また了解性については，テキスト情報を入力する HMM 音声合成の高い了解性が示された一方，64 混合 GMM のような，HMM 音声合成に比べてよりシンプルなモデルを用いても了解性の低下はそれほど大きくないとも解釈しうる．今後この点については検証を行う必要がある．

5. おわりに

本論文では自然音声に人工感を与えるために GMM を用いてメルケプストラムを離散化する手法を提案した．主観評価実験により，動的特徴量を用いずメルケプストラムを離散化する提案手法で STRAIGHT による分析再合成に比べて音声に人工感を与えられることを確認した．

今回は HMM 音声合成に着目してケプストラムを離散化した音声を作成した．HMM 音声合成の品質を改善するための手法として，変調スペクトル上でのポストフィルタリングを行うものがある [6]．今回作成された音声はメルケプストラムを離散化しており，この処理の影響が変調スペクトル上に現れると考えられる．今後は今回提案された手法を用いて作成された音声を用いて変調スペクトル上でのポストフィルタを作成し，自然音声に人工感が与えられるか検討していきたい．

参考文献

[1] J. Yamagishi, K. Onishi, T. Masuko, and T. Kobayashi, “Acoustic modeling of speaking styles and emotional expressions in HMM–based speech synthesis,” IEICE Trans.Inf. & Syst., vol. E88-D, no. 3, pp. 503–509, 2005.

[2] T. Nose, Y. Kato, and T. Kobayashi, “A speaker adaptation technique for MRHSMM–based style control of synthetic speech,” Proc. ICASSP, pp. 833–836, 2007.

[3] T. Toda, A. Black, and K. Tokuda, “Voice Conversion Based on Maximum Likelihood Estimation of Spectral Parameter Trajectory,” IEEE Signal Processing Society, pp. 2222–2235, 2007.

[4] 磯健一, 渡辺隆夫, 桑原尚, “音声データベース用文セットの設計,” 日本音響学会講演論文集 (春), pp. 89–90, 1988.

[5] 河原英紀, “高品質音声分析変換合成法 STRAIGHT の出自・経歴・前途,” 電子情報通信学会技術研究報告, 音声 105(571), pp. 13–18, 2006.

[6] 高道慎之介, 戸田智基, Graham Neubig, Sakriani Sakti, 中村哲, “変調スペクトルを考慮した HMM 音声合成,” 日本音響学会講演論文集, 2-7-10, 2013.