HMM 音声合成の話者モデル作成の効率化に関する検討*

◎酒向 慎司，西本 卓也，嵯峨山 茂樹（東大・情報理工）

1 はじめに

音声アプリケーションの活用の場が広がりつつある中，隠れマルコフモデル (HMM) に基づいた音声合成手法 [1, 2] では，波形接続方式では難しいとされる声質の変換や適応が可能であるという利点があり，合成品質についても大きく改善されるなど，盛んに研究が行われている．また，それらの成果であるモデル学習のためのツール類の配布も積極的に進められており，IPA の元で開発された擬人化音声対話ツールキット（Galatea）のテキスト音声合成モジュール GalateaTalk[3] では，波形生成部において HMM 音声合成手法が導入されている．

このように，テキスト音声合成ソフトウエアの開発・利用環境が整いつつある一方で，独自の話者モデルを作成するための枠組みはまだ確立しているとはいえない．話者モデルの構築は学習データである波形データと言語，韻律情報などを記述したラベルデータが与えられれば可能であるが，ラベルデータの作成が大きな課題となっている．ラベルデータ修正作業の支援環境の検討，アクセント位置・境界の推定などの研究が行われているものの，まだ実用的レベルには至っていない．

我々は，一定量の文章を読み上げるだけで，独自の話者モデルを自動構築する効率的な手法について検討を行っている．ここでは，音声対話技術コンソーシアム（ISTC）が開催する音声対話技術講習会で行われている実習で実施されている事例を中心に，少量の音声データから音声合成用モデルを自動構築する試みについて報告する．

2 HMM に基づいた音声合成

日本語の漢字仮名混じり文を入力とした音声合成システムでは，信号としての音声波形を生成する一方で，テキストの読みの付与やアクセントパタンの生成など主にテキスト解析を行うパートが必要となり，GalateaTalk でもそれぞれの機能ごとに実装されている（図 1）．

2.1 モデル学習と波形生成

HMM 音声合成におけるモデル学習では，波形データとそれに付随するトランスクリプション（音素境界，アクセント，品詞や係り受けなどの言語情報，以下ラベルデータと読む）を元に，ラベルデータに記述された様々な環境依存要因（以下，コンテキストと呼ぶ）に基づき，スペクトル，基本周波数（F_0），音素状態継続長を統計的な枠組みによってモデル化を行っている．

HMM の各状態では，フレーム単位のスペクトルパラメータと F_0 の系列を特徴量として，それらの系列を再現する生成系と見ることができる．モデルの単位としては音素を基本単位として，前後の音素環境やアクセントなどのコンテキストを考慮することにより，詳細なモデルによって分類される．時間構造のモデル化は，HMM の学習過程で得られる，各状態における滞在確率を多次元ガウス分布によってモデル化することで，学習データに基づいた継続長モデルを得ることができる．

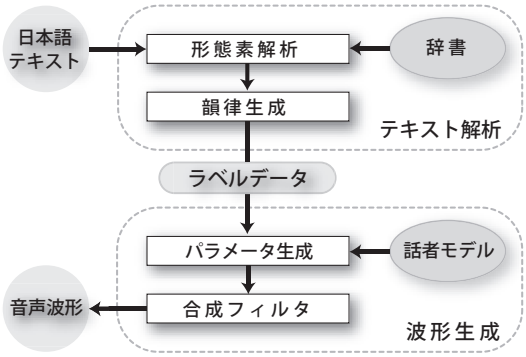

Fig. 1 GalateaTalk の基本構成

合成時には，与えられたラベルデータに沿って状態継続長分布を選択し HMM の各状態の時間構造を決定する．その状態系列のもとで HMM の各状態からパラメータ生成を行う．最後に生成されたパラメータ系列から MLSA フィルタによって音声を合成する．

学習に必要な音声データ量について，これまでに発表されている文献では，450 文章程度の音声データからモデル学習を行う場合が一般的だが，比較的少量の音声データからも，了解性のある音声合成が可能であることが示されている [4].

2.2 テキスト音声合成システムの構成

前節の波形生成では，合成したい発話内容に応じたラベルデータが与えられることを前提としている．日本語テキストから音声を合成するためには，入力されたテキスト列から，相当する HMM の系列を決定するためのラベル系列へと変換する過程が必要となる．この過程では，テキストの読みの付与，アクセント境界およびアクセント型の決定などが含まれ，GalateaTalk では，形態素解析やアクセント結合などの処理を行うテキスト解析部が実装されている．

3 話者モデルの自動学習

特定の話者をターゲットとしてモデルを学習する場合，中性的なモデルを用意して，話者適応を行うことで特定話者モデルを作成することも考えられるが，ここでは，とくに初期モデルを用意することはせず，特定話者の音声データのみを用いてモデル学習を行う方法ついて検討する．

---

*A study on developing acoustic model efficiently for HMM-based speech synthesis. by Sako Shinji, Nishimoto Takuya, Sagayama Shigeki (Graduate School of Information Science and Technology, The University of Tokyo).

3.1 データの収録とラベルデータ

音声データの録音は，理想的には遮音設備などを利用し，クリーンな環境で行われることが望ましいが，一般的な環境での作業を想定し，PC に搭載されているサウンド機能を利用し，ヘッドセットマイクで音声を収録する．

個々の発声データについて，想定された読みに従って正確に読み上げられているとすると，音素系列や形態素情報など構文解析によって一意に定まる．しかし，音素の時間境界，ポーズ位置，アクセント位置などの韻律情報は発話ごとに依存するため，一律に同じものを用いることはできない．

本来は，発声データごとにアクセント情報の修正を行うことが望ましいが，現時点では自動推定や効率的な修正作業は難しい．そこで，韻律情報が整備された音声データベースを基本発声として利用することにする．つまり，韻律情報の付与された正しい発声パタンのサンプルを提示し，それに倣った発声を録音することで，サンプルのもつ韻律情報に沿った発声データを得ることが期待できる．また，提示音声によって，漢字の読みの不一致や，単純な読み誤りを防ぐ効果も期待できる．

3.2 学習データの作成

音素境界情報は，連続音声認識システム Julius を利用し，文章ごとに音素系列を固定した上で viterbi alignment を求めた．F_0 系列は，音声データから自動 F_0 抽出プログラム[1]を利用して求めたものをそのまま使用する．自動推定された音素境界，F_0 の結果には，音素境界のずれ，倍ピッチや，有声と無声判別の誤りなどが含まれるため，それらをできるだけ修正することが望ましいが，ここでは自動推定された結果をそのまま利用する．

3.3 モデル学習

公開されている HTS の学習プログラムと学習用サンプル[2] を利用し，収録された音声データから学習データの作成とモデル学習を行う．メルケプストラム分析の他，学習データ作成に必要なプログラム類は主に SPTK を使用する[3]．この過程で学習された HMM は，HTK 形式の記述ファイルから，GalateaTalk で利用可能な話者モデルデータであるメルケプストラム・対数基本周波数・状態継続長分布のモデルパラメータとツリー形状を記述したファイルへと変換される．

4 モデル作成法の評価

本研究で提案する音声収録からモデル学習までの工程は，2004 年度以降の ISTC の開催する技術講習会において，音声合成の実習課題として実施されている．作成手法は，段階的に修正を施しながら改善しているが，ここでは 2005 年度に実施した実習内容を元に報告する．

ATR 音声データベース B セットから 50 文章の音声データを収録し，発声の際には，基準となるサンプル音声を提示することで，アクセントなどの韻律情報の一致を図った．録音時には，文章ごとにサンプル音声の再生，録音，録音された音声の再生というサイクルを繰り返し，録音された音声を確認して，必要に応じて再録音を行った．個人差はあるものの，15〜30 分程度で収録することができた．

音声の収録後は，録音された波形データから音素境界と F_0 の自動推定を行い，その結果に基づいてメルケプストラム分析など学習データを作成する．モデル学習では，GalateaTalk で使用されているモデル構造，コンテキスト要因など同等の形式に準じ，収録された音声データのみを使用して HMM を学習した．50 文章の発話データから，GalateaTalk の話者モデルの生成に要した時間は，発話速度による音声データの長さに依存するが，概ね 20 分程度ですべての作業が完了した[4]．

4.1 結果と考察

2005 年度の実習では，全体的には前年度から向上したものの，了解性を満足する合成品質まで達しないケースも見られた．モデル学習は，ラベルデータに記述されている情報に基づいて行われるため，自動推定の誤りなどの波形データとラベルデータとの不一致が大きな悪影響を与えると考えられる．

実習の結果から得られた問題点を踏まえて，音素境界を求める際に，読みが大きく異なる，正しく録音できていない，等の理由によりセグメンテーションに失敗する場合は，その発声データを排除する．また，F_0 推定の条件を適切に設定することで，より安定した合成音声が得られることが確認できている．

5 むすび

少量の音声データの収録から，HMM 音声合成システムの話者モデル作成を自動化する試みについて検討した．本稿では，技術講習会の実習として実施できるよう，簡易的なモデル手法という形で報告しているが，より一般的に，HMM 音声合成のモデル作成を支援する手法，あるいは環境整備を主眼に研究を進めている．なお，これらの成果を元にしたモデルの自動作成のツール・ドキュメント類は，改善を進める一方で，GalateaTalk とともに公開していく予定である．今後の課題としては，発話データに応じたラベルデータの自動推定，発話データ量や文章内容の最適化などの検討がある．

謝辞　音声対話技術講習会の実習において，有益な助言を頂いた関係者，並びに受講生各位に感謝する．

参考文献

[1] 益子 貴史，他，動的特徴を用いた HMM に基づく音声合成，信学論 (D-II), vol.J79-D-II, no.12, pp.2184–2190, 1996.
[2] 吉村 貴克, 他, HMM に基づく音声合成のためのスペクトル，ピッチ，状態継続長のモデル化, 信学技報, vol.99, no.255, pp.33–38, 1999.
[3] 山下 洋一 他, マルチモーダルコミュニケーションのための音声合成プラットホーム, 情報処理学会研究報告, SLP-40-12, pp.67–72, 2002.
[4] 高御堂 雄三, 他, HMM を用いた音声合成による学習データ量と音質の調査, 日本音響学会講論集, pp.291–292, 2002.

[1] ESPS get_f0 コマンドと機能的に同等のプログラムを利用した
[2] HTK のパッチとして配布されている http://hts.ics.nitech.ac.jp/
[3] 音声信号処理ツールキット http://kt-lab.ics.nitech.ac.jp/tokuda/tokuda/SPTK/
[4] 実習では Pentium4 3GHz クラスの PC を使用