2007－SLP－67 (16)
2007／7／21

楽曲検索システムにおけるプレイリストに適応した音響モデル構築手法に関する検討

原 直† 宮島千代美† 北岡 教英† 伊藤 克亘†† 武田 一哉†

† 名古屋大学大学院 情報科学研究科 〒464–8603 名古屋市千種区不老町 1
†† 法政大学 情報科学研究科

あらまし 本論文では楽曲検索システムの音声インタフェースに適用するための与えられた認識語彙集合に最適な HMM 音響モデルを学習するための手法について述べる。本論文が対象とする楽曲検索アプリケーションにおいては各ユーザ毎に HMM 音響モデルをカスタマイズすることが重要である、なぜなら、1) 楽曲名やアーティスト名には一般的なテキスト読み上げコーパスには出現しないような音韻コンテキストが存在すること、2) ユーザによって蓄積している音楽が異なっていること、が理由としてあげられる。特に、認識語彙集合に対して最適な状態共有構造を探すということは音響モデルの学習における新しい問題である。そこで本研究では 100 名以上の話者による合成音声を用いてタスクに関連した語彙発話を生成しタスク依存音響モデルを構築する手法を提案する。フィールドテストによる評価実験の結果、提案手法により作成したタスク依存音響モデルはタスク非依存音響モデルに比べて約 10%の単語誤り削減率を達成することを確認した。

キーワード 音響モデル，タスク適応，HMM 音声合成

Constructing Acoustic Model for User-specific Song List in a Music Retrieval System

Sunao HARA†, Chiyomi MIYAJIMA†, Norihide KITAOKA†, Katsunobu ITOU††, and Kazuya TAKEDA†

† Graduate School of Information Science, Nagoya University
†† Graduate School of Information Science, Hosei University

Abstract This paper discusses a training method for the HMM acoustic model that efficiently cover the given vocabulary in order to apply it to the speech interface of a music retrieval system. Customizing the acoustic model to each user is important in this application because 1) song titles and artist names contain many phonetic contexts that are rare in general, e.g. text reading corpora, and 2) the songs stored in a device are different among users. In particular, finding an optimal state-tying structure for the given vocabulary is a new problem in acoustic model training. We propose a method for building a task-dependent acoustic model that uses task-related synthetic utterances of more than one hundred speakers by means of HMM-based speech synthesis. From the experimental evaluation using field test data, we confirmed that the task-dependent acoustic model trained by the proposed method can reduce word error rate by 10% compared to a task-independent model.

Key words Acoustic model, Task adaptation, HMM-based speech synthesis

1. はじめに

ガウス分布の数などの HMM 音響モデルの複雑さは音声認識システムの性能に大きな影響を与える。一般に、複雑なモデルを利用することで高い音声認識精度を得ることができるが、複雑なモデルを学習するためには大量の学習用データが必要となる。また、複雑なモデルを扱うためにはメモリリソースが多く要求されることから、小型携帯端末など一部のプラットフォーム上では利用することは困難である。これらのトレードオフを解決し効率的な音響モデルを構築する手法として、状態共有モデル [1] [2] や分布共有モデル [3] などが提案されている。これらの手法は学習用コーパスに出現したトライホンの分布に最適

な状態共有構造を見つけ出す手法である。

本論文が対象とする音声対話楽曲検索システムでは、認識対象語彙はユーザが所有する楽曲により決まるため、辞書や文法だけではなく最適な音響モデル構造もユーザ毎に異なると考えられる。しかし、これまでにモデルの構造の観点からタスク依存音響モデルを構築する方法について述べた研究はあまり見られない。最も単純なタスク依存音響モデルを構築する手法は、タスクに出現する語彙を含んだ発話を収集し学習コーパスとして利用することだが、これは明らかに新規ユーザに対して有効な手法ではない。本論文の目標は、楽曲検索システムにモデルを適応するために認識対象語彙が既知の条件下でHMM音響モデルを構築する手法を確立することである。

本論文では、認識対象語彙の合成音声を用いた学習によるタスク依存音響モデル構築を提案する。HMMによる音声合成手法[4]を用いることで、数百発話から学習された話者依存HMMモデルによる話者毎の合成音声を作り出すことが可能となる。この合成音声コーパスを用いることでタスク非依存コーパスでは出現しない音韻コンテキストも含めたトライホンHMMモデルを学習することができる。また、認識タスクに応じて学習音声を合成しているため、認識タスクで使われないような冗長なトライホンは出現しない。従って、学習されたモデルの状態共有構造は、状態数を削減しても必要な音韻コンテキストをすべて含んでいると考えられる。この特徴により、話者及び音響環境に適した音響モデルが作成され、高い性能を得ることができるだろう。

本論文は以下の節から構成される。第2節では本論文で使用する楽曲検索システム *MusicNavi* について概説し、実用アプリケーションを想定したタスク依存音響モデルの利用方法を述べる。第3節では、本論文で提案するモデル構築手法を詳細に説明する。第4節では、提案手法を用いた評価実験について述べる。最後に第5節で本論文の結論を述べる。

2. *MusicNavi* – 音声対話楽曲検索システム

MusicNavi[5]はPCに収録された楽曲を検索・再生するためのインタフェースである。*MusicNavi*は起動時にユーザが指定したフォルダ内の楽曲（MP3ファイルなど）を探索し、楽曲ファイルに付けられているメタ情報（ID3タグ[6]など）からアーティスト名と楽曲名の情報を読み取り、プレイリストを作成する。作成されたプレイリストはリモートサーバに送信され、認識に必要な認識単語辞書、認識文法、音声合成辞書、音響モデルをサーバ上で構築し、*MusicNavi*にダウンロードを促す。この機能により、タスク遂行に必要となる最小構成の音声認識・合成機能を備えた音声対話インタフェースをユーザのPC上に構築することができる。*MusicNavi*はサーバ上で構築された辞書等を組み込むと、音声認識を開始しユーザの声に応じて楽曲を検索・再生する。*MusicNavi*終了時にはデータアップロードのためのプログラムが起動する。このプログラムは、ユーザが*MusicNavi*を利用した際の音声データ、認識結果、認識辞書、*MusicNavi*の動作ログをリモートサーバに送信する。送信されたデータはユーザIDにより一意に識別することが可能なので、音声と認識結果を元に音響モデルの適応を行い、ユーザ専用の話者依存音響モデルを再構築することができる。

楽曲検索というタスクの特性とシステム構成のため、*MusicNavi*の音声認識器は次のような特徴を持っている。

- アーティスト名や楽曲名と幾つかのコマンド語を認識対象とした孤立単語音声認識タスク

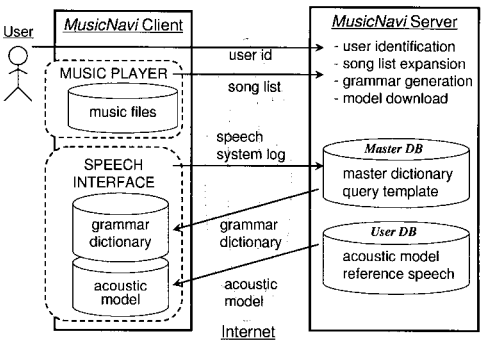

図1　楽曲検索システム MusicNavi の概要

- PCに保存された楽曲のおおよそ1.2倍から1.5倍程度の語彙数（評価実験では約1000単語規模）
- ユーザは認識対象語彙となるPC内の楽曲名などを覚えているため、未知語発生率は少ない

一方で、*MusicNavi*の認識対象語彙であるアーティスト名や楽曲名には外来語や当て字などが多く、新聞記事の読み上げなどの一般的なタスクでは利用されないような音韻コンテキストも含まれている。従って、音響モデルを構築する際には様々な音韻コンテキストの存在するコーパスを利用して学習することが望ましい。

コンテキスト志向の状態共有音響モデル[2]では一定数のトライホンにより学習コーパスに出現するすべての音韻結合を網羅することが可能となる。一般に、状態共有構造の最適化処理は与えられた学習コーパスに依存する。例えば、HTK[7]における実装では、学習発話内で出現頻度の高い状態ほど、より細かなコンテキストに分割される。従って、一般的な文章により学習された状態共有音響モデルは*MusicNavi*システムに対しては冗長なトライホンを含んでおり、認識精度を落とす原因ともなりうる。*MusicNavi*システムに最適な音響モデルを学習するためには、出現しうる音韻コンテキストをすべて含んだ音響モデルを構築する手法を確立する必要がある。

3. ユーザ固有の語彙や声に適した音響モデルのカスタマイズ手法

前節で述べたように、アプリケーションに最適な音響モデルを作成するためには、出現する語彙に存在するすべてのトライホン集合が含まれている必要がある。語彙依存音響モデルを構築するための最も単純な方法は認識対象となる語彙の発話を持つ学習コーパスを利用することである。しかし、ユーザが所有する楽曲はユーザ毎に異なっており、アプリケーションを利用する際にユーザが楽曲リストを変更する可能性もあるため、対象語彙を特定した発話を大量に収集することは難しい。

そこで本論文では、HMM音声合成技術[4]を利用した語彙依存音響モデル構築を提案する。HMM音声合成システムは数百発話を用いることで話者依存音響モデルを作成することができるため、複数の話者による特定の語彙集合を含んだ学習用音声コーパスを容易に作成することができる。HMM音声合成システムは、音声特徴パラメータの混合分布 x とその動的特徴量 Δx から、特徴パラメータ系列を生成する。生成された音韻系

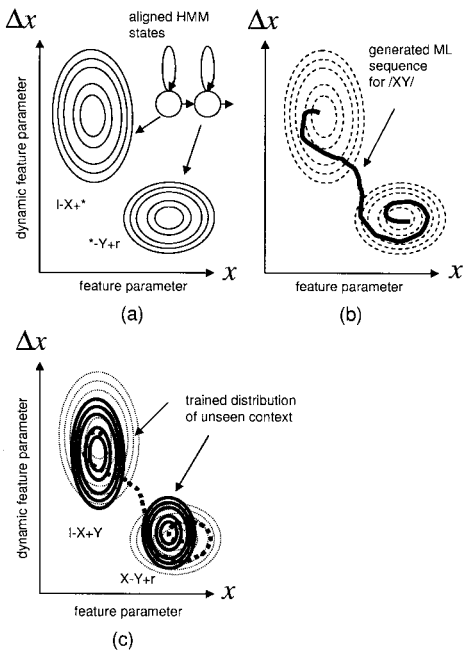

図 2　音声合成による学習の概略図

列はガウス分布で特徴づけられた HMM 状態を結合しているため、未知の系列であっても自然な状態遷移系列が生成される。

図 2 は、学習コーパスに未知のコンテキストである `/...lXYr.../`を音響モデルとして学習する際の考え方を示している（ここで、`/l/`,`/X/`,`/Y/`,`/r/`はそれぞれ音素を示す）。

(a)　各話者依存 HMM に対して、二つのガウス分布がトライホン `l-X+*` と `*-Y+r` の確率分布を表しているとする。ここで、トライホン `l-X+*`とは音素`/X/`の前に`/l/`を持つが後ろは任意の音素となるコンテキストを学習したトライホンを意味している。この例では、コンテキスト`/lXY/`も`/XYr/`も学習コーパスには含まれていなかったと仮定する。

(b)　音声合成器は与えられた状態継続長を元に、音素境界上で最尤の HMM 状態遷移系列を生成する。すべての話者に対してこの生成を繰り返すことで様々な状態遷移パターンが得られる。

(c)　ガウス分布をその遷移パターンにフィッティングすることで、与えられたコンテキストをより正確に表現したトライホン `l-X+Y` と `X-Y+r` が学習される。

しかし、特定の話者とコンテキストからは単一の最尤系列しか出現しないため、合成音声コーパスからは話者内の分散を学習することができない。そこで現在の実装では自然発話コーパスから学習した話者非依存音響モデルの分散をタスク依存音響モデルの分散として利用した。

モデルの構築手順を図 3 に示す。話者毎に作成された話者依存 (SD)HMM と全話者について学習した話者非依存 (SI)HMM は自然発話コーパスを用いて学習された音声合成用の HMM 音響モデルである。作成した音響モデルを用いて HMM 音声合成を行い、複数話者によるタスク語彙の音声特徴パラメータ系列

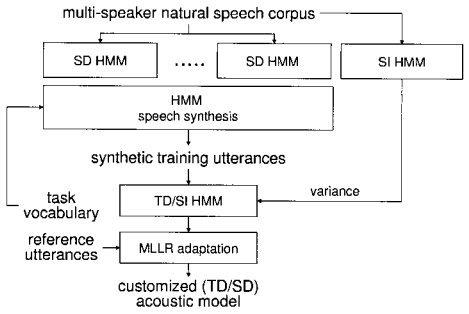

図 3　タスク語彙とユーザに適応したモデルカスタマイズの手順。自然発話コーパスから音声合成用のモデルとして話者非依存 (SI) 音響モデルと複数話者依存 (SD) 音響モデルを作成する。次に HMM による合成音声を精製しタスク依存・話者非依存 (TD/SI) モデルを作成する。最後に MLLR 適応によりタスク依存・話者依存 (TD/SD) モデルを作成する。

を生成する。生成された特徴パラメータ系列を用いて話者非依存・タスク依存 (SI/TD) 音響モデルを作成する。最後にユーザの声や利用環境に合わせるために教師無し MLLR 適応 [8] を行い、話者依存・タスク依存 (SD/TD) 音響モデルを作成する。

4. 評価実験

提案手法により作成されたタスク依存音響モデルの有効性を評価するために、*MusicNavi* システムによるフィールドテストを行った。フィールドテストでは、ユーザが *MusicNavi* システムを利用して自由に楽曲を検索して再生する際の音声を収録した。従って、収集されたデータには音楽再生中の発話も含まれていた。

認識対象語彙は 1352 単語であり、一部のコマンド発話を除いてすべてアーティスト名や楽曲名である。音声合成用モデルを学習するための自然発話コーパスとして新聞記事読み上げ（JNAS）コーパス [9] を利用した。JNAS コーパスから男性 145 名による 22,395 発話を抽出し、話者非依存音響モデルを学習した。このモデルをベースラインのタスク非依存 (TI) モデルとして評価を行う。同じ JNAS コーパスを用いて男性 145 名分の音声合成用の話者依存 HMM 音響モデルを作成した。それぞれの話者依存音響モデルは 150 発話で学習されており、発話内容はすべての話者が同じ文章を発声している [10]。この 145 名の話者依存 HMM 音響モデルによって合成音声の特徴パラメータ系列を合計で 196,040 (= 145 × 1352) 発話を生成した。生成した 196,040 発話によって、提案手法によるタスク依存音響モデル (TD) を学習した。本実験では、音声合成システムとして HMM 音声合成ツールキット HTS [11] を利用した。

HMM モデルの構造については、TI モデルも TD モデルも HMM 状態数及び状態あたりの混合分布数を同数とした。評価用発話としては、*MusicNavi* フィールドテストで得られた 8 名の男性話者による 195 発話を利用した。その他の実験条件を表 1 に示す。

図 4 に TI モデルと TD モデルの単語誤り率 (WER) を示す。図より、すべての混合数条件において TD モデルが TI モデルよりも高い性能を示していることが分かる。また、TI モデルが混合数 8 ほどで下限を示しているのに対して、TD モデルで

表 1 実験条件

音声合成	
特徴量	12 MFCC + Δ +ΔΔ + Power + Δ Power + ΔΔ Power
HMM 状態数	666 (average)
音声認識	
標本化周波数	16kHz
特徴量	12 MFCC + Δ + Δ Power
フレーム幅	25 ms (ハミング窓)
フレームシフト	10 ms
HMM 状態数	3000
混合数	1, 2, 4, 8, 16

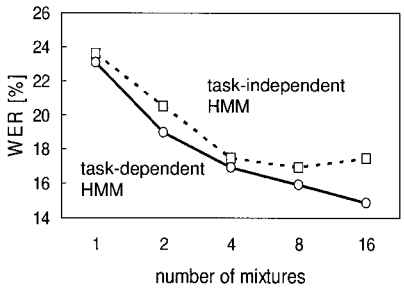

図 4 タスク依存 (TD) モデルとタスク非依存 (TI) モデルの単語誤り率

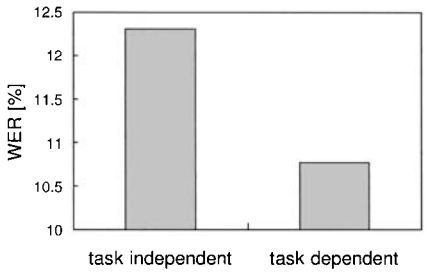

図 5 MLLR 適応後のタスク依存 (TD) モデルとタスク非依存 (TI) モデルの単語誤り率

は混合数 16 でも下限を示していない。認識語彙に存在するトライホンは元の新聞記事コーパスにはほとんど含まれていないため TI モデルでは複雑なモデルを学習するのが困難であったと考えられる。

この結果は実際のユーザ利用環境における自然発話認識の難しさも示唆している。クリーンな音響環境を対象とした事前評価実験では、単語誤り率は 3.55%であった [12]。実環境での誤り率増大の主な原因としては、背景音楽が鳴っていたことや発話の怠けの存在が考えられる。

図 5 に混合数 16 の TI モデルと TD モデルの MLLR 適応後の単語誤り率を示す。この実験では各話者毎に 5–10 発話を用いて教師無し MLLR 適応を行った。図より、どちらのモデルも MLLR 適応による性能向上が見られるが、TD モデルがより高い性能を示しており、*MusicNavi* のような楽曲検索システムにおいて有効なモデルであることを示している。

5. まとめと今後の課題

本論文ではユーザ毎に異なる認識語彙を必要とする楽曲検索システムのためのタスク依存音響モデル構築手法を提案した。提案手法では、認識語彙による学習用発話を大量に生成するために HMM 音声合成技術を用いた。実環境でのフィールドテストで得られた音声を用いた評価実験により、本論文の提案手法が有効であることを示した。

提案手法では大量の学習用データ生成し利用するため、学習に多くの時間を必要とする。そこで今後はより短時間で学習する手法を検討する必要があるだろう。また、話者内分散を学習することができないため、モンテカルロ法を用いた合成 [13] などの手法により話者内分散も考慮した学習方法を検討する必要がある。また、HMM 音声合成のための TD モデルを音響モデルの学習に再利用する手法の検討も今後の課題としてあげられる。

謝辞 本研究の一部は文部科学省リーディングプロジェクト「e-Society 基盤ソフトウェアの総合開発」によるものである。

文 献

[1] H.W. Hon, and K.F. Lee, "On vocabulary-independent speech modeling," Proc. of International Conference on Acoustic Speech and Signal (ICASSP 1990), vol.2, pp.725–728, 1990.

[2] S. Young, J. Odell, and P. Woodland, "Tree based state tying for high accuracy modeling," Proc. of ARPA Workshop on Human Language Technology, Princeton, March 1994.

[3] A. Lee, T. Kawahara, K. Takeda, and K. Shikano, "A new phonetic tied-mixture model for efficient decoding," Proc. of International Conference on Acoustic Speech and Signal (ICASSP 2000), vol.3, pp.1269–1272, 2000.

[4] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for hmm-based speech synthesis," Proc. of International Conference on Acoustic Speech and Signal (ICASSP 2000), vol.3, pp.1315–1318, 2000.

[5] 原直, 宮島千代美, 伊藤克亘, 武田一哉, "汎用 pc 上で利用された音声対話システムによる音声収集と評価," 情処学研報, vol.2006, no.136, pp.167–172, Dec. 2006.

[6] "ID3.org," `http://www.id3.org/`.

[7] "The HTK Book," `http://htk.eng.cam.ac.uk/`.

[8] C. Leggetter, and P. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models," Computer Speech and Language, vol.9, no.2, pp.171–185, 1995.

[9] K. Itou, M. Yamamoto, and K.T. et al, "The design of the newspaper-based japanese large vocabulary continuous speech recognition corpus," Proc. of International Conference on Spoken Language, vol.7, pp.3261–3264, 1987.

[10] 徳田恵一, 益子貴史, 宮崎昇, 小林隆夫, "多空間上の確率分布に基づいた HMM," 信学論, vol.J83-D-II, no.7, pp.1579-1589, July 2000.

[11] "HTS," `http://hts.sp.nitech.ac.jp/`.

[12] 石原正光, 宮島千代美, 北岡教英, 伊藤克亘, 武田 一哉, "認識対象語彙に応じた音響モデルの構築に関する検討," 日本音響学会講演論文集, 1-P-15, pp.153–154, March 2007.

[13] 片桐章宏, 宮島千代美, 伊藤克亘, 武田一哉, "学習データの分布に従う揺らぎのある HMM 音声合成," 2007 年電子情報通信学会総合大会, D-14-12, p.148, March 2007.