Vol.2012-MUS-94 No.1
Vol.2012-SLP-90 No.1
2012/2/3

自動学習により人間のように歌う音声合成システム —Sinsy—

徳田 恵一[†1]　大浦 圭一郎[†1]

本稿では，HMM 音声合成の手法に基づいた歌声合成システム “Sinsy” について述べる．本システムは，歌声データと対応する歌詞付きの楽譜から，モデルパラメータを自動学習するものであり，学習後は，歌詞付きの楽譜を与えることにより，歌声データ提供者の声質，歌い方等を再現する形で，任意の曲を自動で歌わせることができる．まずはじめに HMM 音声合成について概説し，それがどのように歌声合成に拡張されるかについて述べる．また，2009 年 12 月に開設されたオンラインデモについて触れた上で，今後の技術開発に関してどのような展開が期待されるかについて議論する．

Sinsy —human-like singing voice synthesis system based on automatic learning

Keiichi Tokuda[†1] and Keiichiro Oura[†1]

This paper describes a singing voice synthesis system “Sinsy,” which is fully-based on the HMM-based speech synthesis approach: the system can automatically learn the model parameters from singing waveforms and corresponding musical scores, and then synthetic singing voice mimicking the original singer is generated from HMMs themselves. First, we summarize the basic structure of the HMM-based speech synthesis system and then explain how we can extend it to singing voice synthesis. We also describe a web-based on-line service which we started December 2009 and discuss future directions of the development.

†1 名古屋工業大学大学院工学研究科
Graduate School of Engineering, Nagoya Institute of Technology

1. はじめに

与えられた任意のテキストから対応する音声を合成することを，音声合成，あるいはテキスト音声合成 (Text-To-Speech synthesis; TTS) と呼ぶ．音声合成に関する研究は，計算機資源の増大とともに，人手によるルールに基づいたものから，データに基づいたものへと変化してきた．具体的には，90 年代以前には，人手によるルールに基づいたフォルマント合成が研究されたのに対し，90 年代以降は，波形接続型のアプローチが主流を占めるようになった．波形接続型のアプローチは，ダイフォン等の固定の音声単位を用いるものに始まり，複数の音声単位候補から合成時に動的に選択する単位選択型[1] と呼ばれるものに発展してきた．更に近年は，統計的パラメトリック音声合成と呼ばれる統計モデルに基づいた手法が広く研究されるようになってきた．[2] これらの中でも，統計モデルとして隠れマルコフモデル (Hidden Markov Model; HMM) を用いるものは，HMM 音声合成方式と呼ばれ，利用しやすい学習アルゴリズムやソフトウェアツールが普及しており，広く用いられるようになった．

HMM 音声合成方式には以下のような特徴がある．

(1) 与えられた音声データに基づいてモデルを自動学習することにより，元話者の声の特徴を再現する合成音声を得ることができる．
(2) 比較的少ない量の学習データで高品質な合成音声を得ることができる．
(3) 学習用の音声データをランタイムのシステムに蓄積する必要がない．
(4) HMM のモデルパラメータを適切に変換することにより，様々な声質及び発話スタイルの合成音声を得ることができる．

特に，(4) は他の手法では実現困難な特長であり，実際に「声を真似る」,[3] 「声を混ぜる」,[4] 「声をつくる」[5] 等の手法が開発されている．

テキスト音声合成と同様，歌声合成の研究に関しても長い歴史があり，様々な方式が検討されてきた．最近では，VOCALOID に代表される歌声合成ソフトウェアが市販され，広く利用されるようになってきた．一般の人々の認知度も高まっているが，それとともに，好きな歌い手の声で，好きな曲を簡単に歌わせたいという要請が高まっているものと思われる．実際，UTAU 等の歌声合成のためのフリーソフトウェアにおいては，ユーザーが作成した多くの歌手ライブラリ「UTAU 音源」が公開されている．また，VOCALOID, UTAU 等では，自然に歌わせるため，あるいは，様々な歌唱表現を行うために，後処理的な調整作業が行われることが多い．この作業は，作品制作上の創造的な部分ではあるが，一般のユーザ

Vol.2012-MUS-94 No.1
Vol.2012-SLP-90 No.1
2012/2/3

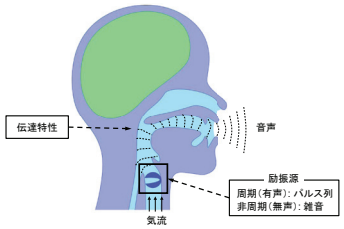


図 1 音声の生成過程
Fig. 1 Speech production process.

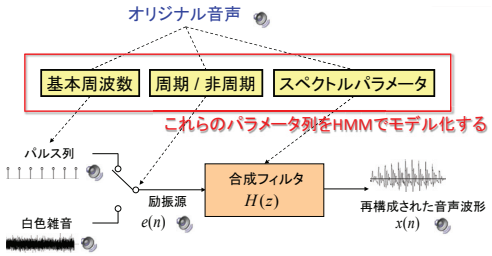


図 2 音声の分析と再構成
Fig. 2 Speech analysis and reconstruction.

にとっては敷居が高すぎるとの声もあり，自ら歌うことにより調整を行う VocaListener[6] 等の開発が行われた．我々は，このような流れの中で，自動学習による歌い手の声質・歌い方等を再現することが可能な歌声合成システム Sinsy[7] を HMM 音声合成に基づいて構築した．更に一般ユーザに広く利用されることを想定し，Sinsy オンラインデモ[8] として公開した．

以下，2 節で Sinsy のベースとなった HMM 音声合成方式について概説した上で，3 節で歌声合成に拡張するための要素技術について述べる．また，4 節で Sinsy オンラインデモについて構成や機能，公開前後の経緯について述べた上で，5 節で今後の課題等について議論する．

2. HMM 音声合成

2.1 音声の分析と再構成

図 1 に示す音声の生成過程は，**図 2** に示すようなデジタルフィルタにより模擬することができ，このようなモデルをソースフィルタモデルと呼ぶ．ソースフィルタモデルによれば，自然音声から抽出された 1) 周期/非周期情報，2) 基本周波数 (以下，F0)，3) メルケプストラム[9] などのスペクトルパラメータ，からなる音声パラメータの列を用いて，元の自然音声を聴感上良く近似する音声波形を再構成することができる．従って，これらの音声パラメータの列をテキストから推定することができれば，テキストから音声を合成することが可能となる．

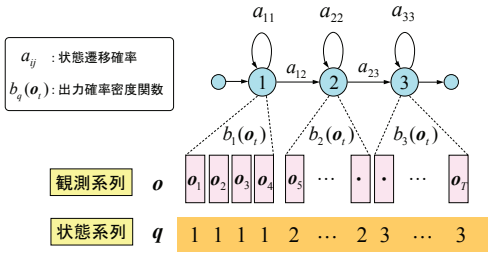


図 3 隠れマルコフモデル: HMM
Fig. 3 Hidden Markov model: HMM.

HMM 音声合成では，このような音声パラメータの列を，時系列の統計的モデルのひとつとして広く用いられている HMM によりモデル化する．**図 3** に HMM の例を示す．図中の観測系列が音声パラメータの列に対応する．従って，HMM 音声合成システムのブロック図は**図 4** に示すようなものとなる．

2.2 HMM による音声パラメータ列のモデル化

通常，各 HMM は音素等の音声単位に対応する長さの音声パラメータ列をモデル化する．HMM の学習は，尤度最大化基準に基づいた学習アルゴリズムによって行うことができる．以下，HMM による音声パラメータ列のモデル化において用いられる手法について述べる．

状態出力分布 通常，音声認識等では，スペクトルパラメータ列のみを HMM によりモデル化するが，音声合成においては，音声波形再構成のために必要となるスペクトルパラメータ，F0 パラメータを連結したベクトルの列をモデル化する．F0 パラメータは，無声部（非周期部）で値がないという特殊な時系列であるため，このような時系列を扱うことのできる状態出力確率分布を用いる．[10]

動的特徴 音声パラメータ列には，動的特徴と呼ばれる各パラメータ列の時間方向の 1 次微分および 2 次部分に対応するパラメータ（デルタパラメータおよびデルタデルタパ

Vol.2012-MUS-94 No.1
Vol.2012-SLP-90 No.1
2012/2/3

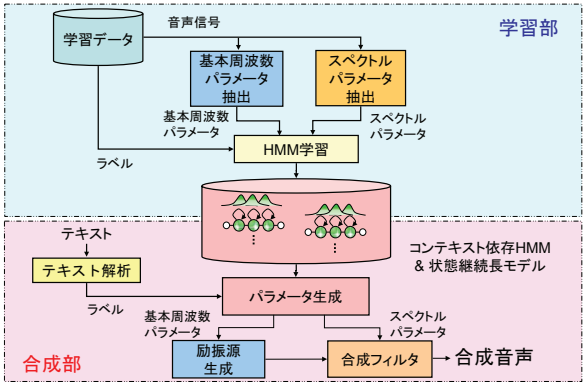


図 4 HMM 音声合成システムのブロック図
Fig. 4 Block diagram of HMM-based speech synthesis system.

ラメータ: Δ および Δ^2）を付加する．これは HMM が時系列の時間方向の相関関係をモデル化しにくい点を補うもので，音声認識でも広く用いられている手法である．

継続長モデル HMM では，時系列の時間方向の伸縮変動は，状態遷移確率によりモデル化されるが，音声の時間的な構造を精度よくモデル化するには不十分であるため，状態継続長モデルを導入する．[11)]

コンテキスト依存モデル 通常の音声認識においては，各音素 HMM は，先行・後続音素に依存したコンテキスト依存モデル（トライフォンモデル）を用いるが，音声合成では，それらのコンテキストに加えて，アクセント型，品詞，文長，文内位置等，言語的な情報をコンテキストとして考慮しなければならない．これは，スペクトルパラメータが主として音素コンテキストに影響を受けるのに対して，F0 パラメータ，継続長は言語的な情報に大きな影響を受けるためである．図 4 中の「ラベル」がコンテキスト依存モデルのモデル名に対応する．

パラメータの共有 コンテキストの組み合わせから，コンテキスト依存モデルの数は膨大なものとなるため，自動的に類似したモデルあるいは状態出力分布を統合するコンテキストクラスタリング[12)] と呼ばれる手法が用いられる．その際，スペクトルと F0 は，それぞれ別のコンテキストに強く影響を受けるため，状態出力分布のスペクトル部と F0 部，更に状態継続長分布は，それぞれ独立にクラスタリングされる．

以上により，スペクトル，F0，継続長のすべてがひとつの確率モデルによってモデル化されることになり，合成時に必要となるすべてのモデルパラメータを同時に自動学習する枠組みとなっていることがわかる．[13)]

2.3 音声の合成

図 4 の下側が音声の合成部である．まず，与えられたテキストに対応するラベル列（コンテキスト依存モデルのモデル名の列）に従って音声単位 HMM を連結することにより得られる HMM から，音声パラメータ列を生成する．音声パラメータの生成は，HMM からの出力確率を最大化するように行われる．この際，動的特徴を考慮することにより，時間的に滑らかに変化する音声パラメータ列を得ることができる．[14)] 生成された音声パラメータ列から，図 2 の合成フィルタを用いることにより，音声波形が再構成される．

3. HMM 歌声合成システム Sinsy

3.1 歌声合成への拡張

テキスト音声合成では，テキストと音声波形の間の関係を HMM でモデル化するが，歌声合成では，楽譜と音声波形の間の関係をモデル化することとなる．ここでは，HMM 音声合成を歌声合成に拡張するために必要な各手法について述べる．

ビブラートのモデル化 ビブラートは主として音高を周期的に揺らす歌唱表現である．ビブラートのかかるタイミングやその周期，振幅の変化は歌い手毎に異なるため，ビブラートはモデル化すべき歌唱表現の一つであると考えられる．そこで，ビブラートの周期と振幅を表す特徴パラメータを音声パラメータベクトルに加えることにより，ビブラートをモデル化している．[15)] これにより，ビブラートに関しても，学習用の歌声データから自動でモデル学習を行うことができる．

タイミングモデル テキスト音声合成と異なり，歌声合成については，曲のテンポやリズムに適合した音声が合成される必要がある．しかし，楽譜から計算される音符と実際の発声との間には時間的なずれがあり，これは「ノリ」などと呼ばれる発声タイミングによる歌唱表現と捉えることができる．Sinsy では楽譜から計算される音符の絶対的な時間と実際の発声とのずれに関しても統計モデルの枠組みでモデル化している．[16)]

歌声モデル化のためのコンテキスト 歌声合成においても，テキスト音声合成と同様，コンテキスト依存モデルを用いるが，考慮すべきコンテキストは当然，テキスト音声合成とは異なったものとなってくる．歌声には音高やテンポ，調，拍子などのコンテキストが考えられるため，Sinsy では新しく歌声合成に特化したコンテキストを定義した．[7)] それらは，音高やテンポ，調，拍子などからなり，具体的には，「音高が C3，音符長が 0.8

Vol.2012-MUS-94 No.1
Vol.2012-SLP-90 No.1
2012/2/3

図 5 Sinsy オンラインデモページ (http://www.sinsy.jp/)
Fig. 5 Sinsy on-line demo page (http://www.sinsy.jp/).

秒の 4 分音符，歌詞は「あ」，小節内で 4 番目で 3 拍目の音符，ひとつ前の音符の高さは D3，…」というように表現されるものである．状態出力分布のスペクトル部，F0 部，ビブラート部はもとより，状態継続長分布，タイミングモデルの分布に関しても，コンテキスト依存のモデル化がされる．

以上により構築される歌声合成システムは，声の質を表わすスペクトル情報，声の高さを表す基本周波数情報，タイミング等を表す時間情報，ビブラート情報をすべて同時にモデル化する方式となっており，声質や音量は元より，基本周波数の変化パターンによって表わされるプレパレーション，オーバーシュート，ビブラート等に関する特徴だけでなく，音符に対するタイミングも自動学習するため，時間構造によって表現される歌唱表現についても自動学習によりモデル化・再現することができる枠組みとなっている．

3.2 歌声データベース

学習時には，特定の歌い手によって歌われた歌声データと対応する歌詞付きの楽譜を数十曲分を用意する．歌い間違い等もモデル化され，合成された歌声に影響を及ぼす可能性があるため，学習用の楽譜と歌声は正しく対応が取れている必要がある．

4. Sinsy オンラインデモ

4.1 オンラインデモページ

Sinsy オンラインデモページ（**図 5**）では，MusicXML で記述された歌詞付きの楽譜をアップロードすることにより，自動合成された歌声をダウンロードすることができる．短い曲ならば，数秒から数十秒で合成が完了する．今のところ，f001j と f002j のふたつの歌声モデルを選択して利用可能である．また，合成時に「声質」,「ビブラート強度」,「ピッチシフト」を指定することもできる．

図 6 に楽譜と対応する MusicXML の例を示す．measure, note, pitch, duration, lyric, それぞれが，小節，音符，音高，音長，歌詞に対応する．f（フォルテ），p（ピアノ）などの強弱記号等も記述することができる．楽譜の表現に MusicXML を採用したのは，オープンなフォーマットであること，楽譜を正確に表現することができること等による．MusicXML を人手で直接，書くのは手間がかかり簡単ではないが，MusicXML ファイルのエクスポート機能をもった市販およびフリーの楽譜編集ツールを用いることにより，容易に MusicXML ファイルを作成することができる．

なお，図 2，図 4 にある通り，音声波形を生成する直前には，歌声はスペクトル情報，基本周波数情報等で表現されているため，必要があれば，この段階で調整を行うことは容易である．Sinsy オンラインデモでは，歌声合成時に，波形だけでなく，波形生成の際，使われた音声パラメータをダウンロードすることができる（図 5）．これらの音声パラメータから，SPTK (Signal Processing Toolkit)[17] を用いることにより音声波形を再構成することが可能である．

その他，Sinsy オンラインデモページには，動画投稿サイトにアップロードされたデモ動画，Sinsy を用いたユーザー作品，関連情報等がまとめてリストされている．

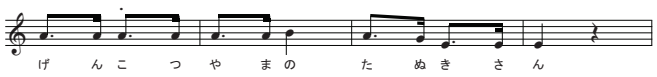
4.2 オープンソースによるシステム構成

Sinsy は，HTS (HMM-based Speech Synthesis System),[18] hts_engine API,[19] SPTK (Speech Signal Processing Toolkit)[17] を軸としながら，HTK (HMM Toolkit),[20] STRAIGHT,[21] Snack Sound Toolkit,[22] ESPS Toolkit,[23] SWIPE[24] 等のオープンソースソフトウェアを利用して構築されていることもその特徴のひとつである．従って，比較的容易に，Sinsy を独自に改良したり，他のソフトウェアツールの一部として組み込んだりすることができる．

4.3 利 用 状 況

Sinsy のリリースをきっかけに UTAU, VOCALOID, Cadencii 等のスコア編集機能をもったツールのファイル形式を MusicXML 形式に変換するためのツールが複数制作されたり，ツール自体から MusicXML 形式のファイルを出力する機能が追加されたりした．更には，Sinsy の使い方を解説する動画なども投稿された．2009 年 12 月のウェブサイト開設

Vol.2012-MUS-94 No.1
Vol.2012-SLP-90 No.1
2012/2/3




```
.
.
.
<measure number="5" width="311">
  <print new-system="yes">
    <system-layout>
      <system-distance>137</system-distance>
    </system-layout>
  </print>
  <note default-x="55">
    <pitch>
      <step>A</step>
      <octave>4</octave>
    </pitch>
    <duration>3</duration>
    <voice>1</voice>
    <type>eighth</type>
    <dot/>
    <stem default-y="10">up</stem>
    <beam number="1">begin</beam>
    <lyric default-y="-77" number="1">
      <syllabic>single</syllabic>
      <text font-size="9.9">げ</text>
    </lyric>
  </note>
  <note default-x="140">
    <pitch>
      <step>A</step>
      <octave>4</octave>
    </pitch>
    <duration>1</duration>
    <voice>1</voice>
    <type>16th</type>
    <stem default-y="10">up</stem>
    <beam number="1">end</beam>
    <beam number="2">backward hook</beam>
    <lyric default-y="-77" number="1">
      <syllabic>single</syllabic>
      <text font-size="9.9">ん</text>
    </lyric>
  </note>
  <note default-x="180">
    <pitch>
      <step>A</step>
      <octave>4</octave>
            .
            .
            .
```

図 6 MusicXML による歌詞付き楽譜の表記
Fig. 6 Description of musical scores with lyrics using MusicXML.

から 2011 年 12 月までで一万回近く MusicXML ファイルがアップロードされており，また，Sinsy を利用したユーザーの楽曲も継続的に動画投稿サイト等に投稿されており，ある程度の浸透が進んでいるものと思われる．今後，Sinsy が，このような「UGC 的」あるいは「オープンソース的」なアクティビティの更なる活性化の一助となることを期待したい．

5. HMM 歌声合成の柔軟性

HMM 音声合成の特長のひとつは，比較的容易に多様な声質，発話スタイルの音声が得られることである．これらの特徴は，歌声合成でも活かすことができる．本節では，このような観点からいくつかの関連事項について議論する．

話者適応（声を真似る） テキスト音声合成では，話者適応の手法を用いることにより，特定の話者の声質あるいは発話スタイルを真似ることが可能となる．[3] 同様の手法を歌声合成に適用することにより，歌い手の声質，歌い方等を「真似る」ことができる．[7]

話者補間（声を混ぜる） HMM 音声合成では，各話者の特徴は，HMM のモデルパラメータとして表現されているため，二つの話者モデルがあったとき，それらのモデルパラメータを適切な方法で補間することにより，二人の話者の中間的な性質をもったモデルを得ることができる．[4] 複数の話者モデルがあった場合，あるいは，各特定話者モデルを特定の発話スタイルのモデルに置き換えた場合も同様である．[25] 同様のことを歌声合成に適用することにより，声を混ぜたり，歌唱スタイルを補間したりすることができる．

固有声手法（声をつくる） 主成分分析等の手法を話者モデルのパラメータ集合に適用することにより，全話者空間を少数のパラメータで表そうとするのが，固有声手法である．[5] 音声合成での応用では，次元圧縮された話者空間で重み係数を設定することにより，ユーザーは容易に所望の「声」の話者モデルを生成することが可能となる．歌声についても，同様のことが可能と予想される．

多言語音声合成 HMM 音声合成システムは，言語に依存する部分がほとんどないという特徴があるため，様々な言語へ容易に適用することができる．更に，言語間話者適応と呼ばれる手法[26] を用いることにより，元話者が話さない言語の音声合成システムを構築することも可能である．これらはいずれも，歌声合成に関しても同様に適応することができる．更に，テキスト音声合成用の読み上げ音声データから歌声音声システムを構築したり，逆に歌声データからテキスト音声合成システムを構築したりといったことも可能と予想される．

フットプリント 波形接続型の音声合成方式あるいは歌声合成方式では，大量の音声データをランタイムのシステムに蓄積する必要があり，通常，10MB から数百 MB，場合によっては数 GB のメモリ容量を必要とする．HMM 音声合成および HMM 歌声合成では，HMM 等の統計モデルのモデルパラメータのみを蓄積するため，特別な工夫をしないままでも 2MB 程度のメモリしか必要としない利点がある．冗長なパラメータを削除

する等の工夫をすることにより，100KB 程度でも明瞭性のある合成音声を得ることができるため，携帯端末，情報家電等，組み込み用途への応用が容易である．

6. 今後の展開

MusicXML で記述された楽譜とそれに対応した歌声データのセットが，数十曲分あれば，Sinsy の歌声モデルの学習は，ほぼ自動で行うことができる．近い将来，このような歌声データをアップロードしただけで自動的にモデル学習を行い，Sinsy オンラインデモページに歌声モデルが追加されるというサービスにチャレンジしたいと考えている．次々と歌声モデルが増え，楽曲の創作者が好みの歌声モデルを自由に選べるようになればと考えている

また，今後は，Sinsy の合成音声品質を向上させるとともに，前節で議論した歌声合成の柔軟性を高める試みに関しても進めていきたい．ひいては Sinsy を利用した多様な楽曲が動画投稿サイト等に多く投稿され，2 次/n 次創作の輪が更に広がることを期待したい．

参 考 文 献

1) A. Hunt, A. Black, “Unit selection in a concatenative speech synthesis system using a large speech database,” Proc. ICASSP, pp.373–376, May 1996.
2) H. Zen, K. Tokuda, A. Black, “Statistical parametric speech synthesis,” Speech Communication, vol.51, no.11, pp.1039–1064, November 2009.
3) J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata, J. Isogai, “Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm,” IEEE Audio, Speech, & Language Processing, vol.17, no.1, pp.66-83, January 2009.
4) T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, K. Kitamura, “Speaker interpolation for HMM-based speech synthesis system,” The Journal of the Acoustical Society of Japan (E), vol.21, no.4, pp.199-206, Apr. 2000.
5) K. Shichiri, A. Sawabe, K. Tokuda, T. Masuko, T. Kobayashi, T. Kitamura, “Eigenvoices for HMM-based speech synthesis,” Proc. ICSLP, pp.1269-1272, Sept. 2002.
6) 中野 倫靖, 後藤真孝, “VocaListener：ユーザ歌唱の音高および音量を真似る歌声合成システム,” 情報処理学会論文誌, vol.52, no.12, pp.3853–3867, December 2011.
7) K. Oura, A. Mase, T. Yamada, S. Muto, Y. Nankaku, K. Tokuda, “Recent Development of the HMM-based Singing Voice Synthesis System - Sinsy,” Proc. of 7th ISCA Speech Synthesis Workshop (SSW7), pp.211–216, Sep. 2010.
8) http://www.sinsy.jp/.
9) 徳田恵一, 小林隆夫, 深田俊明, 斎藤博徳, 今井 聖, “メルケプストラムをパラメータとする音声のスペクトル推定,” 電子情報通信学会論文誌(A), vol.J74-A, no.8, pp.1240-1248, Aug. 1991.
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,” 電子情報通信学会論文誌 (D-II), vol.J83-D-II, no.7, pp.1579-1589, July 2000.
11) H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, T. Kitamura, “Hidden semi-Markov model based speech synthesis system,” IEICE Transactions on Information Systems, vol.E90-D, no.5, pp.825-834, May 2007.
12) J. J. Odell, “The use of context in large vocabulary speech recognition,” PhD thesis, Cambridge University, 1995.
13) 吉村貴克, 徳田恵一, 益子貴史, 小林隆夫, 北村 正, “HMM に基づく音声合成におけるスペクトル・ピッチ・継続長の同時モデル化,” 電子情報通信学会論文誌(D-II), vol.J83-D-II, no.11, pp.2099-2107, Nov. 2000.
14) K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, T. Kitamura, “Speech parameter generation algorithms for HMM-based speech synthesis,” Proc. ICASSP, vol.3, pp.1315–1318, June 2000.
15) 山田知彦, 武藤聡, 南角吉彦, 酒向慎司, 徳田恵一, “HMM に基づく歌声合成のためのビブラートモデル化,” 情報処理学会研究報告, vol.2009-MUS-80, no.5, pp.309–312, 2009.
16) K. Saino, H. Zen, Y. Nankaku, A. Lee, and K. Tokuda, “An HMM-based Singing Voice Synthesis System,” Proc.of ICSLP, pp.1141–1144, 2006.
17) http://sp-tk.sourceforge.net/.
18) http://hts.sp.nitech.ac.jp/.
19) http://hts-engine.sourceforge.net/.
20) http://htk.eng.cam.ac.uk/.
21) http://www.wakayama-u.ac.jp/~kawahara/STRAIGHTadv/.
22) http://www.speech.kth.se/snack/.
23) http://www.speech.kth.se/software/#esps.
24) http://ling.upenn.edu/kgorman/c/swipe/.
25) M. Tachibana, J. Yamagishi, T. Masuko, T. Kobayashi, “Speech synthesis with various emotional expressions and speaking styles by style interpolation and morphing,” IEICE Transactions on Information and Systems, vol.E88-D, no.11, pp.2484–2491, 2005.
26) Y.J. Wu, Y. Nankaku, K. Tokuda, “State mapping based method for cross-lingual speaker adaptation in HMM-based speech synthesis,” Proc. INTERSPEECH, pp.528-531, Sept. 2009.