3-3　多言語音声合成システム

3-3 Multilingual Speech Synthesis System

志賀芳則　河井　恒
SHIGA Yoshinori and KAWAI Hisashi

要旨

NICT で開発した音声合成システムは、音声研究の分野で注目を集める隠れマルコフモデルに基づく音声合成手法を採用している。この手法は音声コーパスを用いて学習した統計モデルから音声を合成するもので、言語依存性が比較的低く、数時間の音声データから声質・発話スタイルを学習し合成できる利点をもつ。こうした利点を活かし、NICT の音声合成システムは現在 7 ヵ国語に対応し、多言語音声翻訳システムや観光案内対話システムの音声出力に利用されている。

Adopting the Hidden Markov Model (HMM)-based technique that has become of major interest in the field of speech synthesis technology, we have developed a speech synthesis system with a high degree of flexibility. The technique, which generates speech from the models that were trained statistically on a speech corpus, is capable of acquiring voice characteristics and speaking style from a few hours of speech data, and is also applicable to new languages with relative ease. Exploiting such merits of flexibility, the system currently supports speech synthesis in seven languages and has been used as an output device in several applications such as multilingual speech translation systems and tourist guide spoken dialogue systems.

[キーワード]
音声合成，TTS，HMM，多言語，SSML
Speech synthesis, TTS, HMM, Multilingual, SSML

1　はじめに

文字で書かれた文書を音声に変換する技術を音声合成と言う。広義の音声合成には、録音した音声を要求に応じて再生する録音再生方式や、収録音声を単語や文節単位で蓄積しておき、指定の順番で連結して再生する録音編集方式も含まれる。前者の適用例として留守番電話の応答を挙げることができ、後者の例としては、列車の行き先などを知らせる駅構内アナウンス（行き先の駅名等をキャリア音声に埋め込む）がある。合成可能な音声の内容が限られるこれら方式に対して、本稿で取り扱う音声合成は任意文の読み上げが可能であり、そのため、より高度で複雑な処理を必要とする。この種の音声合成は、文字列（テキスト）からの音声合成であることを明示して、テキスト音声合成（Text-To-Speech synthesis：TTS）とも呼ばれる。

高音質化が進んだことで、今ではさまざまな場面で音声合成が利用されるようになった。カーナビやビデオゲームのインターフェースとして、また、電子辞書の単語や例文の読み上げに、そしてバスの車内放送や高速道路のラジオ放送に、音声合成を用いるケースが増えている。近年、スマートフォンの普及に伴って台頭してきた音声翻訳アプリケーションにおいても、言語翻訳の結果を音声化するために音声合成が使われている。最近では、国内の外国人旅行者の増加や、製品のグローバル化に伴って音声合成の多言語化の需要が増している。

本稿では、NICT が開発した多言語音声合成システムについて紹介する。このシステムは、音声合成の研究領域で近年最も注目を集める隠れマルコフモデル（Hidden Markov Model：HMM）に基づく手法[1]を採用し、7 ヵ国語の音声合成をサポートしている（2012 年 6 月現在）。本稿の

構成は次の通り。まず **2** では、NICT の音声合成システムのしくみについて概説する。続く **3** では、その特長である合成可能な音声の柔軟性・多様性を活かした応用事例を紹介する。最後に **4** では、現状の技術の問題点に言及し、今後の課題について述べる。なお本稿では、紙面の都合上、個々の技術の詳細には立ち入らない。適宜、文献を紹介してあるので詳しくはそちらを参照されたい。

2 NICT 多言語音声合成システム

NICT の多言語音声合成システムの構成を図 1 に示す。音声合成システムは概して、このようにテキスト処理部と音声信号処理部の 2 つの処理部から成り立っている。テキスト処理部の役割は、入力テキストに言語処理を施し、音声の音韻性（音色）と韻律性（抑揚）を制御するための言語的情報を付与した音素系列を生成することである。ここで音素とは、個々の言語における音声の最小単位である。一方、音声信号処理部の役割は、テキスト処理部が生成した音素系列に基づいて、音声の音響的な特徴を生成し、そこからシステムの最終的な出力となる音声波形を合成することである。以下では、NICT の音声合成システムについてそれぞれを詳細に見ていく。その後で、システムの多言語化について、また、W3C の音声合成記述言語への対応状況について触れる。

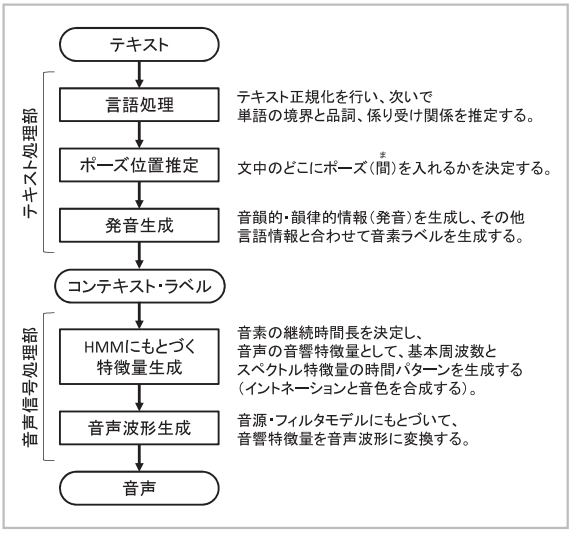

図 1 NICT 多言語音声合成システムの構成

2.1 テキスト処理部

テキスト処理部は、言語処理、ポーズ位置推定、発音生成を行う 3 つのモジュールで構成される。言語処理モジュールではまず初めにテキスト正規化を行う。ここでは、数字（位取り、電話番号、時刻など）や単位・シンボル（cm、“$”記号など）を、さらに英語等の場合は短縮形や省略形（Mr、Ltd、Co など）を適切なテキストに変換する。テキスト正規化に続いて、わかち書き（語の区切りに空白を入れる記法）をしない日本語や中国語では、単語や文節の範囲を定めるため形態素解析を行う。形態素解析とは、それぞれの言語において、意味をもつ最小単位である形態素に文を分割する処理である。英語等のわかち書きされる言語においても、連続する単語がいくつかまとまって特定概念を表すことがあり、そのために形態素解析と同様の処理を行う。

ポーズ位置推定モジュールでは、係り受け解析によって隣接単語間または文節間の接続強度を解析し、その結果をもとに文中の間（ま）の挿入位置を決定する。挿入ポーズの時間的な長さは話者によって異なるため、この段階では決定せず、後段の音声信号処理部において音素の継続時間長と同様、統計的に決定している。

発音生成モジュールでは、辞書を参照して読みと、アクセント（日本語など）、ストレス（英語など）、声調（中国語など）などの情報を生成する。辞書に含まれない語については、ルールベースで表記を発音に変換する。例えば日本語であれば、あらゆる漢字のそれぞれに読みを付した単漢字辞書と読み・アクセント付与規則を利用して未知語に対して読みとアクセントを与える。発音生成にはその他にも、言語に応じてさまざまな処理が伴う。再び日本語を例に挙げると、文節境界の決定、同形語（読みが複数ある単語。「市場」→「いちば」「しじょう」）の読み決定、音便化処理、母音の無声化処理、各形態素のアクセント型・アクセント結合規則に基づく文節のアクセント型の決定などがある。そして最終的に、発音生成モジュールはさまざまな言語情報からなるコンテキストを含む音素ラベル（以下、コンテキストラベルと呼ぶ）の系列を生成する。音素のコンテキストについては **2.2.1** で詳しく述べる。

なお、日本語音声合成のための一般的なテキス

ト処理については文献[2]が詳しい。

2.2 音声信号処理部

テキスト処理部が出力する音素ラベル系列をもとに、音声信号処理部では、(1) 各音素の継続時間長を決定し、音声の (2) 基本周波数 (F_0) の時間パターンと (3) スペクトル特徴量の時間パターンを生成する。これら音響特徴量の生成にあたって HMM に基づく手法を採用している。この音声合成手法では、確率モデルである HMM によって上記 (1) から (3) が同時にモデル化されている。

2.2.1 音声コーパスを用いた HMM の学習

HMM は音声コーパスを用いて予め統計的に学習しておく。図 2 に HMM 学習のブロック図を示す。1 つの HMM はそれぞれの言語の音素にほぼ対応し、上記 3 つの音響特徴量をモデル化するため、個々の音素に対してそのコンテキストに依存した HMM を複数用意する。ここでコンテキストとは、当該音素の音響特徴量に影響を与える可能性のある言語的情報の集まりである。例えば、当該音素のスペクトル特徴は周辺音素の影響を受けるため、先行・後続音素の種別をコンテキストの一部として扱う。また、各音素の継続時間長や F_0 時間パターンをモデル化するために、アクセント／ストレス／声調などとともに、文や節、句内の当該音素の位置などもコンテキストとして考慮する。

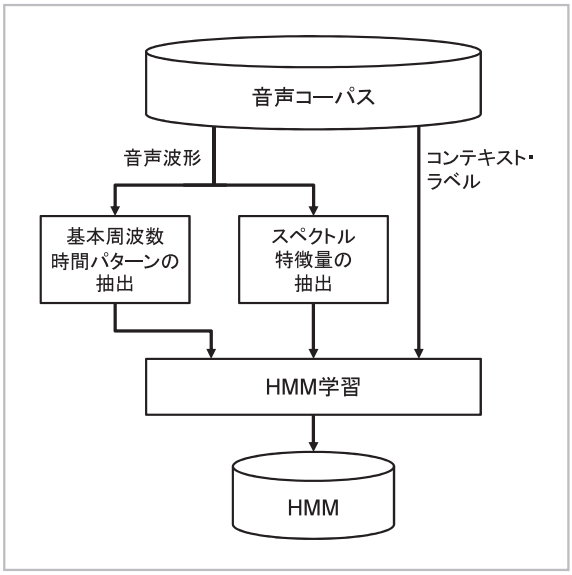

図 2 音声コーパスに基づく HMM の学習

しかし、様々な種類の言語的情報を多数考慮するため（例として、日本語 HMM のコンテキストは 50 種類の言語的情報から成る）組み合わせによってコンテキスト数が膨大になってしまい、すべてのコンテキスト依存 HMM に対して十分な学習データを用意することが困難になる。そこで、コンテキストのクラスタリングを行って、クラスタ毎に学習データ量を確保しつつ HMM を学習する[3]。

2.2.2 HMM からの音声合成

音声合成の際は、まず、前段のテキスト処理部が出力するコンテキストラベルに従って、上記で学習したコンテキスト依存 HMM を連結する。次に、HMM からの音響特徴量生成アルゴリズム[4]を用いて前述の (1) ～ (3) の特徴量を合成する。最後に、人間の発声機構を模した音源・フィルタモデルを用いて特徴量を音声波形に変換する[5]。

2.3 多言語対応

NICT の音声合成システムは日本語、英語、中国語、韓国語のほか、インドネシア語、マレー語、ベトナム語に対応している（2012 年 6 月現在）。今後さらに、アジア圏の言語を中心に対応言語を増やしていく予定である。

言語を拡張する際には、主に 2 つの開発が伴う。1 つ目は、当該言語のためのテキスト処理部の開発である。コーパスベースで新たに開発する場合には、(1) 形態素境界、品詞情報が付与されたテキストコーパスの構築、(2) テキストコーパスに基づく形態素解析用の文法モデルの作成、(3) 発音生成規則の構築、が必要となる。

2 つ目は、拡張する言語の HMM の開発である。そのためには、(1) 言語の特徴に合わせたコンテキストの設計、(2) コンテキストラベルを付与した音声コーパスの構築、が必要となる。最低限の品質の音声を合成するのに必要な音声コーパスの規模は、経験的に、音声データ約 2 時間を含む規模である。さらに、実用レベルの高品質な合成音声を得るには、5 時間以上の規模のコーパスが必要となる。この時間数は音声データの正味時間であって、音声の収録には通常この 3 ～ 4 倍の時間を要する（発声と発声のあいだにポーズが

入るうえ、発声者の喉の疲労に配慮して休憩を挿まなければならないため)。例えば、5時間規模の音声コーパスを構築するには、休憩を含めて15～20時間のスタジオ収録が必要となる。

2.4 音声合成記述言語への準拠

NICTの音声合成システムは音声合成記述言語(Speech Synthesis Markup Language: SSML) Version 1.1 [6] を入力として受け付ける。SSMLとは、Webやその他のアプリケーションにおける音声合成を支援する記述言語でXMLに基づく。W3Cにより勧告され、広範な言語への対応が進められている。SSMLを入力として用いることで、例えば、文書を読み上げている途中で音声合成の言語や話者を切り替えることができ、さらには、文中のポーズを挟む位置や長さなどを記述して出力音声をきめ細かく制御できる。

NICT音声合成システムで利用可能なSSMLタグの一覧を表1と表2に示す(2012年6月現在)。今後、利用できるタグを増やしていく予定である。SSMLの使用例として、日本人2名の会話を表すSSMLを図3に示す。

3 応用

前述のようにHMMに基づく音声合成法は、数時間規模の音声コーパスから発声者の声質や発話スタイルを獲得できる柔軟性を利点としてもつ。加えて、話者1人分のモデルセット(以下、

表1 全言語で対応しているSSMLタグ

要素	属性	説明
speak	version	SSML仕様のバージョン(1.1)
	xml:lang	文書内で使用する言語を指定
voice	languages	タグで囲まれた範囲の言語を指定
	name	音声合成に使用する話者モデル名を指定
	gender	性別を指定(M/F)
	f0-mean	基本周波数(F_0)の平均を指定(NICT独自仕様)

表2 日本語のみ対応しているSSMLタグ

要素	属性	説明
P	-	段落境界を示す
S	-	文境界を示す
token (w)	-	1単語として処理する複数形態素を指定
say-as	interpret-as	テキストの解釈方法を指定する(数詞句に対応) (telephone, date, time, characters, cardinal)
break	strength	区切りの強さを指定。6段階
	time	ポーズの時間長を指定

```
<speak version="1.1" xml:lang="ja-JP"> ← SSML ver 1.1, 言語は日本語
  <voice name="JF009">          ← 日本語話者モデル JF009 を指定
    <p>                 ← 段落の始まり
      <s>               ← 文の始まり
        よろしければ、<break strength="strong"/>電話番号を教えていただけますか？
      </s>              ← 文の終わり    ↑ 文中で大きく区切る
    </p>                ← 段落の終わり
  </voice>
  <voice name="JM001">          ← 日本語話者モデル JM001 に切り替え
    <p>
      <s>
        良いですよ。
      </s>
      <s>
        <say-as interpret-as="telephone">123-4567</say-as>です。
      </s>                          ↑ 電話番号として読みあげる
    </p>
  </voice>
</speak>
```

図3 SSMLによる記述例

単に話者モデルという）を格納する容量が、非圧縮時で数Ｍ～十数Ｍバイトと比較的小さい（製品レベルで主流の波形接続方式の場合、数百Ｍ～数Ｇバイト必要）。以下では、こうした長所を活かした応用事例を紹介する。

3.1　対話音声合成

2010年度に開発した話者モデル「HANNA」は、京都観光案内対話システム向けに、観光ガイドと客との模擬対話から構築した対話音声コーパスから作成されている。HMMに基づく音声合成の柔軟性を活用した一例である。

プロの観光ガイドと客との自発的対話を書き起こした台本をベースに、プロ声優2人が模擬的に行う対話を収録した。そこから観光ガイド役の発話のみを抽出し、約4時間規模の音声コーパスを構築しHMMの学習に用いた。観光ガイドの受け答えに適した対話風の発話スタイル、発話内容に応じたイントネーションをもち、プロ声優による明瞭な発音を反映した合成音声を得ることができる。こうして作られた合成音声が、ユーザの聞き手としての自然な反応を引き出すことを実験により確認している[7]。

3.2　声の個人性を保持した異言語音声合成

音声翻訳において、翻訳結果をユーザ（翻訳元言語の発声者）の声で合成したいという要求がある。これが可能になると、音声翻訳を多人数で利用する際、合成される声の違いから個々のユーザの特定が容易になる。実現方法として次のようなアプローチが考えられる。まず、HMMに基づく音声合成の柔軟性と話者モデル格納容量の小ささを活かし、翻訳先言語の話者モデルを多数作成してシステムにまず格納しておく。音声翻訳時には、何らかの尺度を用いて、入力音声（ユーザの声）に対して音響的に最も近い話者モデルをその中から選択する。そして、選択した話者モデルを用いて翻訳結果を音声合成する。

2011年度に試作したシステムは、日本語から英語への音声翻訳を行うもので、入力された日本語音声にもっとも類似した英語音声の話者モデルを、男声50、女声50のモデルの中から選択するとともに、F_0基準値（声の高さ）を入力音声に一致させて音声を合成する。話者モデルの選択法については文献[8]を参照してほしい。異言語間の話者類似性を評価する知覚実験によって本手法の有効性が示されている[8]。

4　おわりに

NICTの音声合成システムが採用しているHMMに基づく音声合成手法は、これまで見てきたように、多言語化や発話スタイルの多様化に適した音声合成手法である。しかし反面、出力される音声が、いわゆるボコーダ声と呼ばれる鼻声のようなくぐもった音声になりやすいという欠点がある。音響特徴量から音声波形を再合成するプロセスを経ること、そして、統計的学習に伴う平均化に起因して音響特徴量の過平滑化が生じることが、こうした音質劣化の主原因となっている。

前者の再合成プロセスから生じる音質劣化については、候補話者の音声に対して同プロセスを適用し、劣化の小さい話者を選択することで軽減できる。一方、後者の過平滑化を防ぐ有効な方法はまだ見つかっていない。ただ、コンテキストによって表現できない変動が音響特徴量に多く含まれると平滑化が起こりやすいことから、発声の安定した話者を選ぶことで問題をある程度緩和できると考えられる。そこで現状は、収録に先立ってプロのナレーターや声優を数十人集めてオーディションを行い、録音したサンプル音声の音響特徴量から音声を再合成し、その際の音質劣化の度合いを候補話者毎に評価している。同時に声の安定性を研究者がチェックし、総合的に最適と判断した話者を選択している。

しかしこれは裏返せば、現在の技術は万人の声を高品質に合成できるレベルにはまだ達していないことを意味する。この問題を解決するにはまず、前述の音質劣化原因に対処しなければならない。そこでNICTでは、声に対する向き不向きのない音響特徴量と音声波形の再合成プロセス、そして過平滑化に起因する劣化の少ない音声のモデル化手法の研究に取り組んでいる[9]-[11]。さらに、こうした取り組みと併せて、短時間の収録から高品質に音声合成可能な手法を確立する必要がある。声のプロでない人に対して、収録を長時間・長期間にわたって安定して行うのは困難である。合成したい話者の少量の音声に対して、大規

模コーパスから学習した異なる話者のモデルを適応する話者適応技術[12]がこれには有効であると考えられ導入を図っているが[13]、原理的に音響特徴量の変形を伴うことなどから合成音声の品質が大きく劣化する問題がある。「誰のどんな声であっても、数分から十数分程度の収録で実用レベルの品質の音声を合成できる」そうした技術の確立が今後の課題となる。

参考文献

1 H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," Speech Communication, Vol. 51, No. 11, pp. 1039–1154, Nov. 2009.

2 匂坂芳典, "音声合成における自然言語処理," 情報処理, Vol. 34, No. 10, pp. 1281–1286, Oct. 1993.

3 T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," in Proc. EUROSPEECH'99, pp. 2347–2350, Sept. 1999.

4 K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in Proc. ICASSP, pp. 1315–1318, June 2000.

5 今井聖, 住田一男, 古市千枝子, "音声合成のためのメル対数スペクトル近似（MLSA）フィルタ," 電子通信学会論文誌（A）, Vol. J66-A, No. 2, pp. 122–129, Feb. 1983.

6 http://www.w3.org/TR/speech-synthesis11/

7 翠輝久, 水上悦雄, 志賀芳則, 川本真一, 河井恒, 中村哲, "ユーザの相づち・うなずきを喚起する音声対話システム," 電子情報通信学会論文誌, Vol. J95-A, No. 1, pp. 16–26, 2012.

8 津崎実, 徳田恵一, 河井恒, 志賀芳則, 倪晋富, 大浦圭一郎, 塩田さやか, "個人性を考慮した異言語音声合成に対する知覚評価," 電子情報通信学会技報, Vol. 112, No. 81, SP2012-39, pp. 33–38, June 2012.

9 R. Maia, T. Toda, K. Tokuda, S. Sakai, and S. Nakamura, "A decision tree-based clustering approach to state definition in an excitation modeling framework for HMM-based speech synthesis," in Proc. Interspeech2009, pp. 1783–1786, Sept. 2009.

10 Y. Shiga, T. Toda, S. Sakai, and H. Kawai, "Improved training of excitation for HMM-based parametric speech synthesis," in Proc. Interspeech2010, pp. 809–812, Sept. 2010.

11 Y. Shiga, "Pulse Density Representation of Spectrum for Statistical Speech Processing," in Proc. Interspeech2009, pp. 1771–1774, Sept. 2009.

12 J. Yamagishi and T. Kobayashi, "Average-voice-based speech synthesis using HSMM-based speaker adaptation and adaptive training," IEICE Trans. Inf. & Syst., E90-D(2), pp. 533–543, Feb. 2007.

13 倪晋富, 河井恒, "On effects of speaker similarity in average voices on adapted web-based HMM voices," 日本音響学会 2011 年春季研究発表会, Vol.I, 3-7-3, pp. 303–306, Mar. 2011.

（平成 24 年 6 月 14 日 採録）

志賀芳則（しがよしのり）
ユニバーサルコミュニケーション研究所
音声コミュニケーション研究室
主任研究員
Ph.D. in Speech Technology
音声信号処理、音声合成
yoshinori.shiga@nict.go.jp

河井 恒（かわい ひさし）
株式会社 KDDI 研究所主幹研究員／
元ユニバーサルコミュニケーション研究所
音声コミュニケーション研究室
上席研究員
工学博士
音声情報処理、音声翻訳
hi-kawai@kddilabs.jp