論　文

隠れマルコフモデルに基づく英語歌声合成

中村　和寛[†a)]　大浦圭一郎[†]　南角　吉彦[†]　徳田　恵一[†]

Hidden Markov Model-Based English Singing Voice Synthesis

Kazuhiro NAKAMURA[†a)], Keiichiro OURA[†], Yoshihiko NANKAKU[†], and Keiichi TOKUDA[†]

あらまし　本論文では隠れマルコフモデル (Hidden Markov Model; HMM) に基づく英語歌声合成について述べる．HMM 歌声合成システムは，学習用の歌声データに基づいて，あらかじめスペクトル，基本周波数，ビブラートを HMM により同時にモデル化しておき，合成時には合成したい歌声の楽譜に合わせて HMM を連結し，歌声を生成する．これまでに，日本語の楽譜から歌声を合成するシステムが提案され，一般ユーザによる楽曲作成の際のボーカルとして利用されてきている．本論文ではこのシステムを，英語の歌声を合成できるように拡張するために，英語歌声合成のコンテクストを定義し，楽譜の音符と実際の発音を対応付ける手法を提案する．客観・主観評価実験により効果を確認し，また，日本語歌声合成との比較実験も行う．

キーワード　英語歌声合成，HMM 音声合成，HMM 歌声合成

1. まえがき

コンピュータによる歌声合成の研究は古くから行われており，最近では，VOCALOID [1] の技術を用いた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ている．また，一般の人々の認知度が高まるにつれて，より簡単に，好きな歌い手の声で，好きな曲を歌わせることのできる柔軟なシステムの需要が高まっている．実際，UTAU [2] 等の歌声合成のためのフリーソフトウェアにおいては，ユーザが作成した多くの歌手ライブラリが公開されている．そうした中で，我々の研究グループは，HMM に基づく歌声合成方式 [3] を提案した．デモンストレーション用の WEB サービス [4], [5] を一般公開しており，これまでに多くのユーザに利用されている．

現在，一般に用いられている歌声合成システムの多くは，単位選択型あるいは素片連結型と呼ばれる音声合成方式と同じアイディアに基づいている．素片連結型の音声合成方式では，あらかじめ蓄積された音声波形データを素片に分割し，合成したいテキストに応じて素片を選択，連結することにより，音声を合成する．これに対し，HMM 歌声合成の元となった HMM 音声合成方式は，素片連結型にはない以下のような特徴をもつ．

(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，元話者の声の特徴や発話スタイルを精度良く再現する音声を合成することができる．
(2)　比較的少ない量の学習データで高品質な音声を合成することができる．
(3)　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ため，軽量である．
(4)　HMM のモデルパラメータを適切に変更することにより，様々な声の合成音声を得ることができる．

特に (4) は HMM 音声合成方式の重要な優位点の一つであり，実際に「声を真似る」話者適応手法 [6],「声を混ぜる」話者補間手法 [7],「声を作る」固有声手法 [8] などが提案されている．

HMM 歌声合成では，譜面とそれに対応した歌声の関係を HMM によりモデル化する．スペクトル，基本周波数，ビブラートを同時にモデル化する方式となっており，声質は元より，基本周波数の変化パターンによって表されるプレパレーション，オーバシュート等の特徴だけでなく，音符に対する発声のタイミングも自動学習するため，前ノリ，後ノリ等の歌唱スタイルさえも再現することができる．

日本語の歌声合成技術は広く利用されるようになり

[†] 名古屋工業大学，名古屋市
Nagoya Institute of Technology, Gokiso-cho, Showa-ku, Nagoya-shi, 466–8555 Japan
a) E-mail: nkazu@sp.nitech.ac.jp

つつあるが，日本語以外の歌詞も歌わせることができるようになればユーザの表現の幅が広がり，また，世界中の人々が歌声合成を利用できるようになると期待できる．そのためには歌声合成技術の多言語対応が必要であり，その第一歩として，世界共通語として普及している英語に対応することは，非常に有用性が高いと考えられる．本論文では英語の高品質な歌声合成を実現するために，言語依存性が低い統計的な手法である HMM 歌声合成の枠組みに基づく英語歌声合成手法を提案する．

以下，**2.** で HMM 歌声合成システムについて，**3.** では歌声合成特有の手法について，**4.** では提案法となる英語歌声合成について，**5.** では評価実験について述べ，最後に **6.** で全体をまとめる．

2. HMM 歌声合成システム

HMM 歌声合成システムの概要を図 1 に示す．このシステムは学習部・合成部の二つのパートで構成される．

2.1 学 習 部

HMM の学習のために，歌声データベースから各種特徴量を抽出する．特徴ベクトルにはメルケプストラム，対数基本周波数，ビブラートパラメータ（振幅，周波数）[9]，これらのパラメータの動的特徴量を用いており，HMM には多空間上の確率分布に基づいた HMM (Multi-Space Probability Distribution HMM; MSD-HMM) [10] を用いている．モデルの学習は音素単位で行うが，譜面情報や前後の音素などを考慮したコンテクスト依存モデルを用いて詳細なモデリングを行う．コンテクスト全ての出現を考慮した組合せは膨大な数になり，限られた音声データでそれらの全てをカバーすることは不可能であるため，決定木に基づくコンテクストクラスタリング [11] を用いて頑健なモデルを構築している．

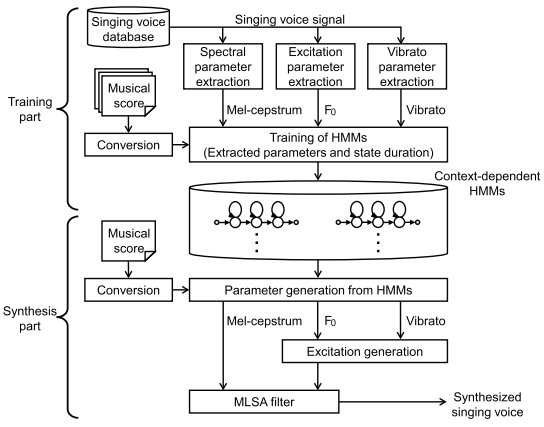

図 1　HMM 歌声合成システムの概要
Fig. 1　Overview of the HMM-based singing voice synthesis system.

2.2 合 成 部

合成したい曲の譜面データから抽出した歌詞や音高等のコンテクストを元にモデルを連結する．次に譜面データの音符長情報と学習した継続長モデルから音素継続長及び各状態の継続長を求め，連結したモデルからパラメータ生成アルゴリズム [12] によりパラメータ系列を生成する．そして，ビブラートパラメータから計算した正弦波を対数基本周波数系列に重ね合わせてビブラートを再現し，生成したパラメータに基づいて，MLSA フィルタ [13] を駆動することで歌声を合成する．

3. 歌声合成特有の手法

HMM 歌声合成システムでは，通常の HMM 音声合成では用いられない，発声タイミングモデル [14] や音高正規化学習 [15] の導入により，自然な歌声の合成を可能としている．

3.1 発声タイミングモデル

HMM 歌声合成の発声タイミングは譜面の音符情報から得られるが，譜面通りのタイミングで歌い出すと，得られる合成音声はリズム感のない不自然なものとなってしまう．これは，人間の歌声において，実際の発声タイミングと譜面上のタイミングの間に微妙なずれが存在するためである．その例を図 2 に示す．このずれは無意識のうちに生じることも多いが，その一方で意識的にずれをもたせる歌唱表現もあり，いずれの場合においても歌い手の特性をモデル化する上でも重

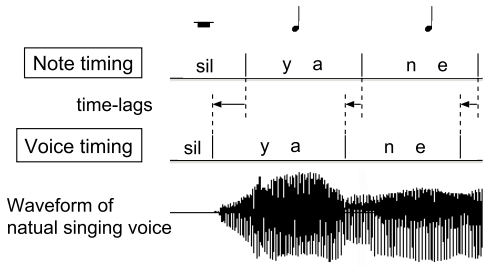

図 2　発声タイミングのずれの例
Fig. 2　An example of time-lag.

要な要素である．そこで，実際の人間の歌声の発声タイミングのずれを再現するために，発声タイミングモデル [14] が提案された．発声タイミングモデルは，ある音符の中に含まれる先頭の音素の開始時間と譜面上のその音符の開始時間の差がどれくらいになるのかという情報を保持する．学習の際には，あらかじめ作成しておいた歌声モデルを用いて学習データの音素アライメントをとることで各音素の時間境界を取得し，譜面上の音符の開始時間とその音符内の先頭の音素の開始時間の差を 1 次元のガウス分布でモデル化する．合成時には，発声タイミングモデルからタイミングのずれを求め，それを踏まえて状態継続長を決定する．発声タイミングモデルを用いることで，リズム感のある自然な歌声を合成することが可能になる．

3.2 音高正規化学習

HMM 歌声合成は統計的手法であるため，学習データに含まれない音高は合成することができない．あらゆる音高を合成するためには，あらゆる音高を含む学習データベースを用意する必要があるが，実際に歌声に含まれる音高には偏りがあるため，あらゆる音高を偏りなく十分に含むデータベースを用意することは困難である．そこで音高を直接モデル化するのではなく，歌声の対数基本周波数系列と音符の音高の差分をモデル化する音高正規化学習 [15] が提案された．音高 p のコンテクストをもつある状態 i における対数基本周波数の静的特徴量の平均の推定値 $\hat{\mu}_i^{(p)}$ は次式で表される．

$$\hat{\mu}_i^{(p)} = \mu_i + b_i^{(p)} \tag{1}$$

ここで，μ_i は歌声の対数基本周波数と音符の音高の差分の平均であり，$b_i^{(p)}$ は音符の対数基本周波数である．音符の音高は固定のため推定する必要はなく，歌声の対数基本周波数と音符の音高の差分を表す分布のみを推定する．音高正規化学習を用いることであらゆる音高を合成可能なモデルを生成することができる．

4. 英語歌声合成

4.1 英語譜面の歌詞情報

一般的に日本語の譜面は歌詞をかな表記で記述するため，モーラ単位のかなと音素の対応テーブルを用意することにより，歌詞を音素列に変換することが可能である．それに対して，英語の譜面は，一般的に図 3 に示すように通常のテキストで歌詞を記述するため，単語辞書と音素の対応テーブルを用意するだけでは，“the” や “lead” のように文脈によって発音が変わる単語をうまく扱うことができない．そこで，歌詞の文字列から音素列に変換する形態素解析処理が必要となる．休符で区切られたフレーズと呼ばれる区間を文とみなして形態素解析することにより，歌詞の文字列を音素列に変換し，母音を核とするシラブルという単位に分割する．日本語の文字と発音の関係の例を表 1 に，英語の文字と発音の関係の例を表 2 に示す．表 2 における単語列を形態素解析することで，発音のシラブル列と音素列が得られる．ここでは母音と子音を区別するために，母音を太字で示している．

英語歌声合成に用いるコンテクストの設計のために，日本語コンテクスト [4] を元に (i) 言語依存コンテクストと被依存コンテクストを明確に切り分け，(ii) 英語のシラブルと日本語のモーラの情報格納領域を共有可能な形にすることでコンテクストを共通化し，(iii) 言語特有の情報をもつ部分を定義することで共通化できない言語依存のコンテクストを扱えるようにする．これにより，英語のみに含まれるアクセント及びストレス情報も扱うことが可能になる．言語依存部分と言語非依存部分を切り分けた提案コンテクストを表 3 に示す．言語特有の情報をもつ部分を太字とした．

本論文では形態素解析に Flite [16] を，単語辞書に CMU Pronouncing Dictionary [17] を採用した．こ

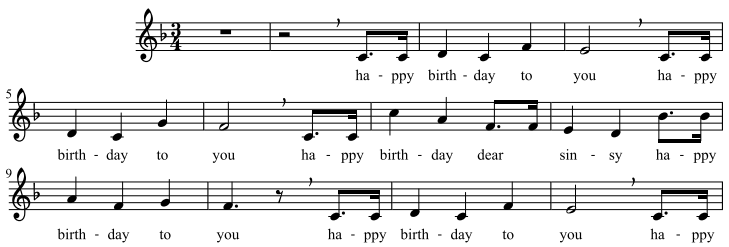

図 3 英語の譜面の例
Fig. 3 An example of English score.

表 1　日本語の文字と発音の関係の例
Table 1　An example of the relationship between Japanese characters and pronunciations.

String	Mora	げ		ん	こ		つ		や		ま		の		た		ぬ		き		さ		ん
Pronunciation	Mora	ge		N	ko		tsu		ya		ma		no		ta		nu		ki		sa		N
	Phoneme	g	**e**	**N**	k	**o**	ts	**u**	y	**a**	m	**a**	n	**o**	t	**a**	n	**u**	k	**i**	s	**a**	**N**

表 2　英語の文字と発音の関係の例
Table 2　An example of the relationship between English characters and pronunciations.

String	Word	rhythm						of		the		classical									music					
	Syllable	rhy		thm				of		the		clas			si		cal				mu			sic		
Pronunciation	Syllable	rih		dhaxm				ahv		dhax		klae			sih		kaxl				myuw			zihk		
	Phoneme	r	**ih**	dh	**ax**	m		**ah**	v	dh	**ax**	k	l	**ae**	s	**ih**	k	**ax**	l		m	y	**uw**	z	**ih**	k

表 3　コンテクストの一覧
Table 3　Proposed context design.

Phoneme	Quinphone. (Phoneme within the context of two immediately preceding and succeeding phonemes)
Syllable (Mora)	Number of phonemes in {previous, current, next} syllable.
	Position of {previous, current, next} syllable in note.
	Language dependent context in {previous, current, next} syllable. (English: with or without {accent, stress}, Japanese: undefined)
Note	Musical {tone, key, beat, tempo, and length} of {previous, current, next} note.
	Position of current note in {measure, phrase}.
	With or without a slur between current and {previous, next} note.
	Dynamics to which current note belongs.
	Difference in pitch between current note and {previous, next} note.
	Distance between current note and {next, previous} {accent, staccato}.
	Position of current note in current {crescendo, decrescendo}.
Phrase	Number of {syllables, notes} in {previous, current, next} phrase.
Song	Number of {syllables, notes} / Number of measures.
	Number of phrases.

れらに，長い無音を表す “sil”，発声の前後の無音を表す “pau” と，ブレスを表す “br” を加えたものを，英語歌声合成の音素セットとした．

4.2　シラブルの割り当て手法

英語の譜面では，図 3 の “happy” や “birthday” のように，一つの単語が複数の音符に分割されて記述されていることも多い．しかし，ユーザが譜面に歌詞を記述する際に，シラブルの境界位置で正しく分割して記述するとは限らない．例えば，“everything” という単語は “e, verything”, “ev, erything”, ···, “everythin, g” のように 9 通りに分割される可能性があり，どのように分割された場合にも歌声を合成できるように，音符とシラブルの対応を推定する必要がある．一方で，ユーザによる単語の分割を歌詞入力時にチェックし，シラブル境界以外における分割を制限する手法も考えられるが，形態素解析器が用いている辞書におけるシラブル境界をユーザが意識する必要があり，歌詞入力の利便性が低下する可能性がある．また，人名や地名等の固有名詞やユーザによる新造語といった辞書に登録されていない単語が歌詞として入力された場合には，形態素解析器は単語の表記からシラブル列を推定する必要があるが，ユーザが意識するシラブル境界と形態素解析器により推定されたシラブル境界が一致するとは限らないため，そのような場合にも音符とシラブルの対応を推定する手法が必要となる．どの音符にどのシラブルを割り当てるかを決定するために，本論文では次の 2 手法を提案する．

1: 先頭から割り当て　先頭の音符から順にシラブルを一つずつ割り当て，足りない分は最後のシラブルを伸長する．余った場合は最後の音符に余ったシラブルを全て割り当てる．

2: スコアに基づく割り当て　音符のスコアを用いてシラブルを割り当て，足りない部分は一つ前のシラブルを伸長する．手法の流れを以下に示す．

ステップ 1: 各音符の文字数をカウント　手法 2 では，音符を以下のようにスコアリングする．まず，各音符に割り当てられている歌詞の文字数をカウントする．文字は，歌詞のアルファベット一つ一つに対応し

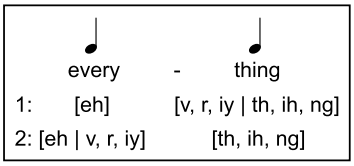

図 4　2 種類のシラブルの割り当て手法
Fig. 4　Two methods for syllable allocation.

ている．ただし，シラブルを構成するために必要な母音が多く含まれる音符には優先的にシラブルを割り当てるべきであると考えられるため，母音となる可能性が高い文字 “a”，“e”，“i”，“o”，“u” は他の文字の 2 文字分となるようにカウントする．例えば，表 2 において，“classical” という単語は文字 “a” を二つと文字 “i” を一つ含んでおり，それらの文字は異なるシラブルに分けられ，発音の音素において母音となっている．このように文字 “a”，“e”，“i”，“o”，“u” は多くの場合に母音の音素となるが，例外も存在する．その一つが表 2 の “rhythm” であり，歌詞の文字には “a”，“e”，“i”，“o”，“u” のいずれの文字も含まないが，発音の音素には母音を含む．

ステップ 2: 各音符のスコアを計算　単語内の音符数を N，形態素解析結果のシラブル数を S，音符 n における文字カウントを c_n としたとき，音符 n のスコア w_n を，スコアの合計がシラブル数と等しくなるように次式で定義する．

$$w_n = \frac{Sc_n}{\sum_{n'=1}^{N} c_{n'}} \tag{2}$$

ステップ 3: 各音符に割り当たるシラブル数を決定

最後に，各音符に割り当たるシラブルの数 k_n を決定する．あらかじめ全ての n について $k_n = 0$ と初期化しておく．そして，最もスコアの高い音符 $\hat{n}$ を一つ選択し，$k_{\hat{n}} = k_{\hat{n}} + 1$, $w_{\hat{n}} = w_{\hat{n}} - 1$ とする．これを S 回繰り返すことで，各音符のシラブル数が決定する．ただし，単語の先頭の音符には，必ずシラブルが一つ以上割り当たることとする．

図 4 にシラブルの割り当て手法の例を示す．“everything” という単語は CMU Pronouncing Dictionary で “eh | v, r, iy | th, ih, ng” と表され三つのシラブルとなる．ただし，“|” はシラブルの境界を表す記号である．この単語が二つの音符に割り当てられた場合，手法 1 では先頭の音符からシラブルを一つずつ割り当て残りを最後の音符に割り当てるため，“eh”，“v, r, iy | th, ih, ng” となる．手法 2 では $S = 3$，$c_1 = 7$，$c_2 = 6$ より，各音符のスコアは

$$w_1 = 3 \times 7/(7+6) \simeq 1.62 \tag{3}$$

$$w_2 = 3 \times 6/(7+6) \simeq 1.38 \tag{4}$$

となり，1 番目の音符に “eh | v, r, iy” が，2 番目の音符に “th, ih, ng” が割り当てられる．

表 4　二重母音の複製ルール
Table 4　Diphthong duplication rules.

Original	ey	ay	ow	aw	oy
Duplicated	eh, ey	aa, ay	ao, ow	aa, aw	ao, oy

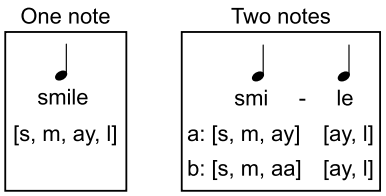

図 5　2 種類の二重母音の複製手法
Fig. 5　Two methods to duplicate diphthong.

4.3　二重母音の複製手法

シラブル数が音符数より多い場合には，一つの音符に複数のシラブルが割り当てられることがあるが，逆に，音符数よりもシラブル数が少ないような場合には，シラブルが一つも割り当てられない音符が存在することになる．シラブルが割り当てられない音符も音声データにおいては歌声が存在するため，何らかの手法でシラブルを割り当てる必要があり，そのための手法を二つ提案する．

a: そのまま複製　一つ前の音符のシラブル核を複製し，複製したシラブル核から後ろを当該音符にシフトする．

b: 複製ルールに従って複製　シラブル核が二重母音だった場合，そのまま複製すると連続性が失われる可能性があるため，手法 b では表 4 のような複製ルールを定める．

図 5 に二重母音の複製手法の例を示す．“smile” という単語は CMU Pronouncing Dictionary で “s, m, ay, l” と表され一つのシラブルとなるが，これが二つの音符に割り当てられた場合には，手法 a では “s, m, ay”，“ay, l” となり，“ay” が連続した発音となる．手法 b では “s, m, ah”，“ay, l” のように前の音符の “ay” が “ah” に変換される．

5. 実　　験

英語歌声合成特有の手法の評価と，日英間の歌声合成の比較を行うために客観・主観評価実験を行った．主観評価実験では，女性 1 名による英語楽曲の歌声を，学習に 20 曲，評価に 5 曲用いた．また，比較のために日本語楽曲の歌声を学習に 17 曲，評価に 5 曲用いた．英語と日本語の学習用楽曲は，有声部分の長さがほぼ等しくなるように選択した．有声部分の長さは各言語約 30 分間である．サンプリング周波数は 48kHz，量子化ビット数は 16bit，モノラルである．スペクトル，基本周波数及びビブラートパラメータとして，STRAIGHT [18] によって抽出されたスペクトルに，メルケプストラム分析 [19] を適用することにより得られた 49 次元のメルケプストラム，対数基本周波数，ビブラートの振幅と周波数，またそれらの Δ，Δ^2 を用いた．フレーム周期は 5ms である．モデルには 5 状態の left-to-right 型 HSMM [21] を用いた．学習用ラベルの音素境界情報の初期値には，確定的アニーリング EM (Deterministic Annealing EM; DAEM) [20] アルゴリズムにより求めた学習データの音素アライメント結果を用いた．コンテクストクラスタリングの分割停止基準には，MDL 基準 [22] の式 (1) の第 2 項に 3.0 をかけたものを用いた(注1)．主観評価実験は，英語被験者 10 名若しくは日本語被験者 10 名により行った．英語被験者は英語を母国語とする者若しくは英語で大学の学部を卒業した者とし，日本語被験者は日本語を母国語とする者とした．英語被験者は英語楽曲の評価を行い，日本語被験者は英語楽曲と日本語楽曲の評価を行った．各被験者に 30 フレーズの中から被験者ごとにランダムに選ばれた 10 フレーズを聞かせ，歌声の自然性について 5 段階 MOS で評価させた．1 フレーズの平均の長さは 8.1 秒であった．また，英語被験者は騒音が 35dB 以下の静かな部屋においてヘッドフォン (SONY 製 MDR-Z900HD) を装着した状態で，日本語被験者は防音室 (高橋建設製サイエンスキャビン SC-3) においてヘッドフォン (SONY 製 MDR-CD900ST) を装着した状態で受聴し，被験者が聴きやすいレベルになるよう被験者本人に音圧を調整させた．

(注1)：頑健なモデル学習を行うために，MDL 基準のペナルティ項に係数をかけてパラメータ数を調整することはよく行われる．今回は経験的に係数を 3.0 としたが，適切なパラメータ数の自動調整手法の確立は今後の課題である．

表 5　各音符に対する生成された音素列の正解率 (A)，正解音素数 (C)，挿入誤り音素数 (I)，削除誤り音素数 (D)，置換誤り音素数 (S) の比較

Table 5　The comparison of (A)ccuracy rate, (C)orrect, (I)nserted, (D)eleted, and (S)witched phonemes of phoneme sequence for each note.

Method	A (%)	C	I	D	S
1-a	92.19	3149	33	35	196
1-b	**95.98**	**3277**	33	35	**68**
2-a	**95.56**	**3230**	**0**	**2**	**148**
2-b	**99.41**	**3360**	**0**	**2**	**18**

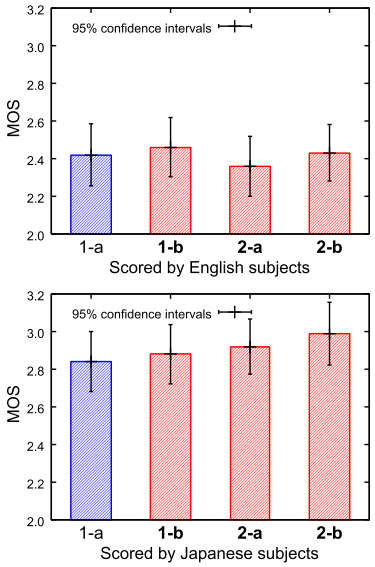

図 6　シラブルの割り当て手法と二重母音の複製手法の影響

Fig. 6　The effect of syllable allocation methods and duplication methods of diphthongs.

5.1 実　験　1

シラブルの割り当て手法と二重母音の複製手法を評価した．シラブルの割り当て手法は次の 2 手法である．

1: 先頭から割り当て

2: スコアに基づく割り当て

また，シラブルを伸長する場合の二重母音の複製手法は次の 2 手法である．

a: そのまま複製

b: 複製ルールに基づいて複製

これらを組み合わせた 1-a, 1-b, 2-a, 2-b の 4 手法に関して評価を行った．

まず，客観評価として，評価に用いた 5 曲に関して，人手による音素列を正解とし，各手法により生成されたラベルの音素列を音符ごとに評価した．結果を表 5 に示す．音素の挿入誤りと削除誤りは，シラブルを正

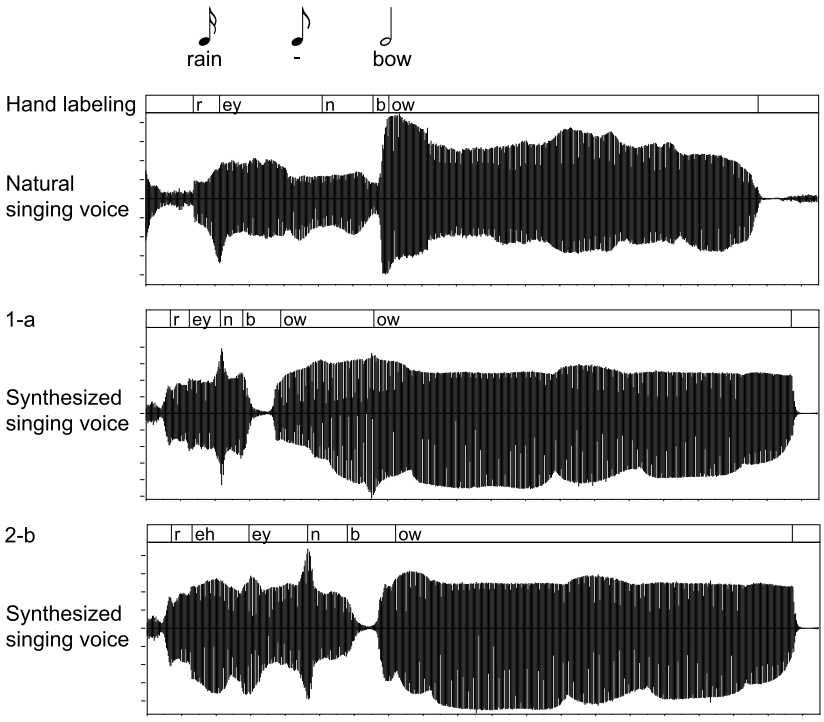

図 7　シラブルの割り当て手法，複製手法の違いによる合成音声の比較
Fig. 7　The comparison of synthesized speech waveforms due to the difference of syllable allocation methods and syllable duplication methods.

解とは異なる音符に割り当てたことにより発生するが，シラブルの複製手法には影響を受けないため，1-a と 1-b，及び，2-a と 2-b の挿入誤り，削除誤り数は同じである．手法 2 により挿入・削除誤りが低減され，手法 b により置換誤りが低減されており，最大で 92%の誤り削減率を達成することができた．

次に，主観評価実験の結果を図 6 に示す．シラブルの割り当て手法，シラブルの複製手法共に，有意差は確認できなかったが，英語被験者に対しては手法 b に，日本語被験者に対しては手法 2 と手法 b に，音質改善の傾向が見られた．図 6 の英語被験者の結果において，手法 1 と手法 2 の違いによる音質への影響が小さかったのは，表 5 のとおり，置換誤りよりも挿入・削除誤りの方が少なかったためだと考えられる．また，被験者が異なる MOS 試験に関しては，被験者が合成音声を聞き慣れているかどうか等により MOS の絶対値に大きな差が出ることが知られており [23]，結果を直接比較することは適切ではない．本実験では日本語被験者は合成音声を聞き慣れており，英語被験者よりも高いスコアを付与した可能性がある．実際に，数名の英語被験者からは，合成音声特有の不自然さが気になったという感想が寄せられている．

合成音声の違いを確認するため，自然音声と人手で付与した音素アライメント，1-a，2-b の合成音声と合成時の音素アライメントを図 7 に示す．これらの音声に対応する譜面上の歌詞は “rainbow” となっており，シラブルは “r, ey, n” と “b, ow” の二つである．1-a では一つ目と二つ目の音符にシラブルが割り当てられ，“b, ow” が三つ目の音符に伸長されて “b, ow” と “ow” となる．一方，2-b では一つ目と三つ目の音符にシラブルが割り当てられ，“r, ey, n” が二つ目の音符にルールに基づき伸長されて “r, eh” と “ey, n” となり，人手で付与した音素アライメントにより近い結果になっている．

また，英語歌声合成で用いた言語依存コンテクストの影響を調査するために，アクセント・ストレスのコンテクストを用いない場合と用いた場合の合成音声について，収録音声とのメルケプストラムひずみを測定した．その結果，学習に用いた 20 曲に対するフレーム平均メルケプストラムひずみは，アクセント・スト

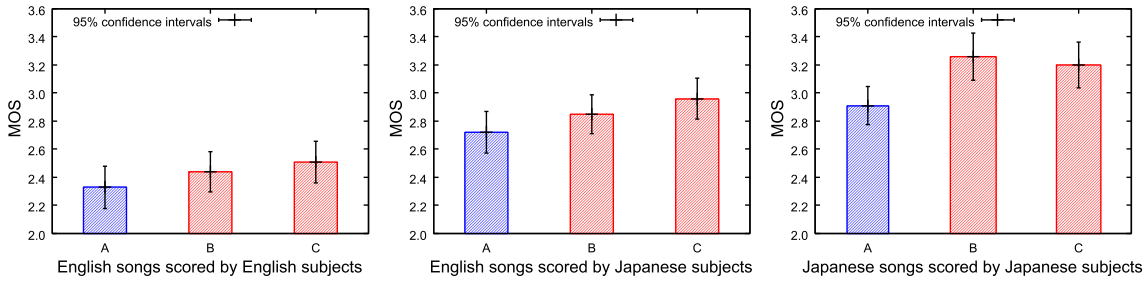

図 8　発声タイミングモデルの影響．A: 発話タイミングモデルなし，B: 先頭音素位置を基準にずれをモデル化，C: シラブル核位置を基準にずれをモデル化

Fig. 8　The effect of the time-lag modeling. A: Without time-lag models, B: With head-phoneme-based time-lag models, C: With syllable-nucleus-based time-lag models.

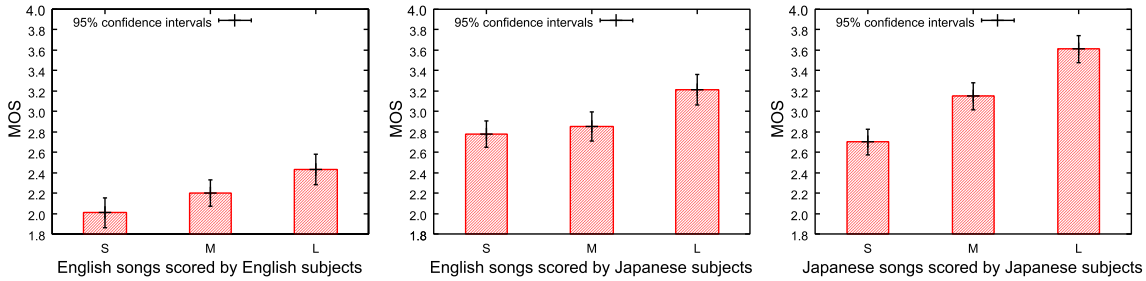

図 9　学習データ量の影響．S: 8 分，M: 15 分，L: 30 分

Fig. 9　The effect of the amount of training data. S: 8 minutes, M: 15 minutes, L: 30 minutes.

レスのコンテクストを用いない場合に 7.533，用いた場合に 7.526 となり，評価に用いた 5 曲に対しては，それぞれ 8.151，8.136 となった．どちらの場合も，アクセント・ストレスのコンテクストを用いることによりメルケプストラムひずみの改善が見られた．

5.2 実　験　2

発声タイミングモデルの有効性と，ずれの基準位置による影響を日本語歌声合成と比較した．ただし，異言語間で音質を比較するのは容易ではないため，日本語楽曲と英語楽曲の MOS 試験は別々に行っており，結果を直接比較することは適切ではない．ずれの基準位置に関しては次の 3 パターンを比較した．

A:　発話タイミングモデルなし
B:　先頭音素位置を基準にずれをモデル化
C:　シラブル核位置を基準にずれをモデル化

シラブルの割り当て手法と二重母音の複製手法に関しては実験 1 で最も評価の高かった 2-b を用いている．実験 2 のみ，被験者がタイミングの指標とするために，歌声と同時に 4 分音符 1 回ごとに 1 回鳴るクリック音を再生した．

主観評価実験の結果を図 8 に示す．日本語と同様に，英語においても発話タイミングモデルありの手法は無しの手法と比較して自然性が高い傾向が見られた．英語においては基準位置はシラブル核位置が適していると考えられるが，日本語においては先頭音素位置がやや高い結果となった．この原因として，英語は子音が二つ以上連続することもあり，子音部分の時間が長くなることが多いため，シラブル核でタイミングをとる方が自然に聞こえる可能性が考えられる．

5.3 実　験　3

学習データ量を変化させたときの合成歌声の自然性の変化を日本語歌声合成と比較した．学習データ量は有声部分の長さを基準として次の 3 種類とした．

S:　8 分（英語 5 曲，日本語 5 曲）
M:　15 分（英語 10 曲，日本語 9 曲）
L:　30 分（英語 20 曲，日本語 17 曲）

シラブルの割り当て手法と二重母音の複製手法に関しては実験 1 で最も評価の高かった 2-b を，発話タイミングモデルは実験 2 で英語において最も評価が高く，日本語においても高い評価を獲得した C のシラブル核

位置を基準にずれをモデル化した手法を用いた．

主観評価実験の結果を図 9 に示す．学習データ量が増えると自然性が向上するという傾向はどちらの言語においても見られた(注2)．

6. むすび

本論文では，英語の歌声合成システムを構築した．言語依存のコンテクストと非依存のコンテクストを明確に切り分け，英語歌声合成のコンテクストを定義した．譜面の音符と歌詞のシラブルをマッチングする必要があるため，シラブルの割り当て手法と二重母音の複製手法を提案し，客観評価実験の結果，音符ごとの音素割り当て精度の向上を確認した．また，発声タイミングモデルの有効性と学習データ量と自然性の関係について日本語歌声合成と比較した．どちらの言語においても発声タイミングモデルは有効であり，学習データ量の増加に伴い合成音声の自然性が改善した．今後の課題として，複数の歌い手の歌声データを用いた評価実験，中国語等の他の言語への拡張，言語依存コンテクストが歌声の自然性に与える影響の評価が挙げられる．

謝辞 本論文で述べた研究開発の一部は，JST CREST 及び堀科学芸術振興財団の支援を受けた．

文　献

[1] H. Kenmochi and H. Ohshita, “VOCALOID – Commercial singing synthesizer based on sample concatenation,” Proc. Interspeech, Special session, Antwerp, Belgium, Aug. 2007.

[2] “歌声合成ツール UTAU,” http://utau2008.web.fc2.com/, 参照 May 31, 2013.

[3] 酒向慎司，宮島千代美，徳田恵一，北村　正，“隠れマルコフモデルに基づいた歌声合成システム，” 情処学論，vol.43, no.3, pp.719–727, March 2004.

[4] K. Oura, A, Mase, T. Yamada, S. Muto, Y. Nankaku, and K. Tokuda, “Recent development of the HMM-based singing voice synthesis system – Sinsy,” Proc. Speech Synthesis Workshop, pp.211–216, Kyoto, Japan, Sept. 2010.

[5] “Sinsy – HMM-based Singing Voice Synthesis System,” http://www.sinsy.jp/, 参照 May 31, 2013.

[6] K. Yutani, Y. Uto, Y. Nankaku, A. Lee, and K. Tokuda, “Voice conversion based simultaneous modeling of spectrum and F0,” Proc. ICASSP, pp.3897–3900, Taipei, Taiwan, April 2009.

[7] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, “Speaker interpolation in HMM-based speech synthesis system,” Proc. EUROSPEECH, vol.5, pp.2523–2526, Rhodes, Greece, Sept. 1997.

[8] K. Shichiri, A. Sawabe, T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, “Eigenvoice for HMM-based speech synthesis,” Proc. ICSLP, vol.1, pp.1269–1272, Colorado, USA, Sept. 2002.

[9] 山田知彦，大浦圭一郎，南角吉彦，徳田恵一，“HMM 歌声合成システムのためのビブラートモデルの導入，” 音響秋季講論集，2-2-11, pp.309–312, Sept. 2009.

[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（D-II），vol.J79-D-II, no.7, pp.1579–1589, July 2000.

[11] 篠田浩一，渡辺隆夫，“情報量基準を用いた状態クラスタリングによる音響モデルの作成，” 信学技報，SP96-79, Dec. 1996.

[12] K. Tokuda, T. Kobayashi, and S. Imai, “Speech parameter generation from HMM using dynamic features,” Proc. ICASSP, pp.660–663, Detroit, Michigan, USA, May 1995.

[13] 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，” 信学論（A），vol.J66-A, no.2, pp.122–129, Feb. 1983.

[14] K. Saino, H. Zen, Y. Nankaku A. Lee, and K. Tokuda, “An HMM-based singing voice synthesis system,” Proc. Interspeech, pp.1141–1144, Pittsburgh, PA, USA, Sept. 2006.

[15] K. Oura, A. Mase, Y. Nankaku, and K. Tokuda, “Pitch adaptive training for HMM-based singing voice synthesis,” Proc. ICASSP, pp.5377–5380, Kyoto, Japan, March 2012.

[16] “flite,” http://www.festvox.org/flite/, 参照 May 31, 2013.

[17] “CMU Pronouncing Dictionary,” http://www.speech.cs.cmu.edu/cgi-bin/cmudict, 参照 May 31, 2013.

[18] H. Kawahara, M.K. Ikuyo, and A. Cheneigne, “Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F_0 extraction: Possible role of a repetitive structure in sounds,” Speech Commun., vol.27, pp.187–207, April 1999.

[19] 徳田恵一，小林隆夫，千葉健司，今井　聖，“メル一般化ケプストラム分析による音声のスペクトル推定，” 信学論（A），vol.J75-A, no.7, pp.1124–1134, July 1992.

[20] N. Ueda and R. Nakano, “Deterministic annealing EM algorithm,” Neural Netw., vol.11, pp.271–282, March 1998.

[21] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, and

(注2)：波形接続型音声合成 [24] が数十時間のデータを用いることが多いのに対し，HMM 音声合成では数十分のデータでモデル学習を行うことが多い．ただし，扱うことができるデータの量は年々増加しており，音声合成の国際的な評価会である Blizzard Challenge ワークショップでは，データ量が 2005 年から 2011 年の 6 年間で約 10 倍になっている [25], [26]．データ量の増加と技術の向上に伴い，合成音声の品質はどの方式においても向上しているものの，まだデータ量の増加による性能の飽和点は明らかになっていない．

T. Kitamura, “A hidden semi-Markov model-based speech synthesis system,” IEICE Trans. Inf. & Sys., vol.E90-D, no.5, pp.825–834, May 2007.
[22] K. Shinoda and T. Watanabe, “MDL-based context-dependent subword modeling for speech recognition,” J. Acoust. Soc. Jpn. (E), vol.21, no.2, pp.76–86, March 2000.
[23] C.L. Bennett, “Large scale evaluation of corpus-based synthesizers: Results and lessons from the blizzard challenge 2005,” Proc. Interspeech, pp.105–108, Lisbon, Portugal, Sept. 2005.
[24] 河井　恒，戸田智基，山岸順一，平井俊男，倪　晋富，西澤信行，津崎　実，徳田恵一，“大規模コーパスを用いた音声合成システム XIMERA,” 信学論（D-II），vol.J89-D-II, no.12, pp.2688–2698, Dec. 2006.
[25] A.W. Black and K. Tokuda, “The blizzard challenge 2005: Evaluating corpus-based speech synthesis on common datasets,” Proc. Interspeech, pp.77–80, Lisbon, Portugal, Sept. 2005.
[26] S. King and V. Karaiskos, “The blizzard challenge 2011,” Proc. Blizzard Challenge 2011, Turin, Italy, Sept. 2011.

（平成 26 年 2 月 12 日受付，5 月 24 日再受付）

中村　和寛

2005 名工大・工・知能情報システム学科卒．2007 名工大大学院工学研究科博士前期課程情報工学専攻修了．同年ブラザー工業株式会社入社．2011 名工大大学院工学研究科博士後期課程創成シミュレーション工学専攻入学．音声認識，音声合成の研究に従事．2007 日本音響学会ポスター賞受賞．現在，日本音響学会会員．

大浦圭一郎

2005 名工大・工・知能情報システム学科卒．2007 ATR 音声言語コミュニケーション研究所有期補佐員．2008 FP7 EMIME project 研究員．2009 総務省 SCOPE 研究員．2010 名工大大学院工学研究科博士過程情報工学専攻修了．同年名工大大学院工学研究科情報工学専攻特任助教．2012 科学技術振興機構 CREST 研究員．音声認識，音声合成の研究に従事．2008 ISCSLP2008 最優秀学生論文賞，2012 情報処理学会山下記念研究賞，2013 日本音響学会独創研究奨励賞板倉記念，2013 情報処理学会喜安記念業績賞，各受賞．現在，日本音響学会，情報処理学会各会員．

南角　吉彦　（正員）

1999 名工大・工・知能情報システム学科卒．2004 名工大大学院工学研究科博士課程電気情報工学専攻修了．同年名工大テクノイノベーションセンター大学院 VBL 部門中核的研究機関研究員．2005 名工大大学院工学研究科助手．2007 名工大大学院工学研究科助教（法改正による呼称変更）．2012 名工大大学院工学研究科准教授．統計的学習，画像認識，音声認識，バイモーダル音声認識，音声合成の研究に従事．現在，電子情報通信学会，日本音響学会各会員．

徳田　恵一　（正員）

1984 名工大・工・電子卒．1989 東工大大学院博士課程修了．同年東工大電気電子工学科助手．1996 名工大知能情報システム学科助教授．2004 名工大大学院情報工学専攻教授．工博．音声言語情報処理，マルチモーダル情報処理，統計的学習理論の研究に従事．2001 及び 2008 電気通信普及財団賞，2001 電子情報通信学会論文賞及び猪瀬賞，2008 電子情報通信学会ソサイエティ論文賞，各受賞．IEEE 音声技術委員会委員，電子情報通信学会編集委員，日本音響学会編集委員，日本音響学会代議員・評議員等を歴任．現在，ISCA Advisory Council，IEEE 編集委員．電子情報通信学会，日本音響学会，人工知能学会，情報処理学会，IEEE，ISCA 各会員．