CE-004

レアなモーラを含む日本語歌唱データベースの構築と基礎評価
Building a Japanese Singing Database Including Rare Moras and Its Evaluation

森勢　将雅†　　藤本　健‡　小岩井　ことり*
Masanori Morise　Ken Fujimoto　Kotori Koiwai

1. はじめに

テキストから音声を生成するテキスト音声合成（Text-to-Speech: TTS）の研究は現在盛んに進められており，その品質はすでに人間と等価な水準に達している．TTS の研究においては，音声データベース（データベースは以下 DB と表記する）が品質に直結する重要な役割を担う．音声 DB は言語情報を含むため言語毎に構築する必要があり，日本語では ATR 音素バランス文[1]が代表的である．とりわけ，音素環境をバランスさせた 503 文は標準的なテキストセットであり，日本語の TTS では標準的なものとして利用されている．TTS では発話者の話者性や発話スタイルが合成結果に反映されるため，それらを変化させるためには，一般に話者や発話スタイルを変化させた，別の DB を用いる必要がある．これは，音声 DB の拡充が TTS 研究の発展に重要であることを意味する．

コンテンツ制作の現場では，現在も VOCALOID [2]が広く利用されているが，統計的手法に基づく歌声合成のソフトウェアもシェアを伸ばしつつある．歌詞と譜面から歌声を生成する統計的歌声合成の場合，歌詞に相当するテキストに加え，譜面に関する音楽的な情報も必要である．歌唱 DB の構築では，音声 DB と異なり楽曲を扱うため，著作権の問題が生じるという歌声合成固有の問題も考慮する必要がある．Sinsy は統計的歌声合成における先駆けであり，学習には童謡などの著作権切れの楽曲が利用されている[3]．TTS において発話者の話者性や発話スタイルが合成結果に反映されるように，統計的歌声合成においても歌唱者の話者性と歌唱スタイルが結果に反映される．童謡で歌唱 DB を構築した場合は歌唱表現も童謡的になりやすく，したがって合成結果も同様に近い歌い方となる．合成対象のジャンルが童謡であれば高い品質を達成できるが，例えば J-POP などの歌い方が異なる楽曲を対象とした場合，常に高い品質が得られるとは言い難い．すなわち，歌わせたい楽曲のジャンルに応じて，該当するジャンルの楽曲で構成された歌唱 DB を構築することが望ましいと言える．一方，既存の楽曲の大半は著作権の問題が生じるため，研究用の DB として公開することは困難な状況にあった．

筆者らは，この問題への対応として，2019 年の著作権法改正に基づき，限定した条件下で利用が可能な歌唱 DB として，J-POP を中心とした既存の楽曲から構成される東北きりたん歌唱 DB （きりたん歌唱 DB）[4]を構築した．その結果，既存の楽曲を用いた場合，歌詞が既存の楽曲に縛られるため，特定のモーラが存在しないことを確認した．そこで本研究では，著作権に関する問題をクリアし，さらにレアなモーラまで含み研究・開発者が自由に利用できる歌唱 DB を構築することでこの問題の解決を図る．

† 明治大学　Meiji University
‡ フラクタル・デザイン　Fractal Design Inc.
* ピアレスガーベラ　Peerless Gerbera

本論文は以下の内容で構成される．まず 2 章では，統計的手法に基づく音声合成や歌声合成について説明し，必要とされる音声・歌唱 DB の条件と本研究の位置づけについて述べる．3 章では，提案する DB のコンセプトと基本データについて説明する．4 章では，構築した DB について統計的な解析を実施し，エントロピーを用いた評価により他の歌唱 DB と比較することで性能を確認する．

2. 歌声合成に関する関連研究

歌唱 DB を必要とする現在の主な研究領域は，統計的手法に基づく歌声合成である．歌声合成研究は話し声を合成する TTS にも関連するため，ここでは両方について関連する研究について説明する．

2.1 歌声合成手法の発展と現状

TTS にも様々な手法が存在するが，Hidden Markov model (HMM)に基づく統計的パラメトリック音声合成 [5]が 2000 年代から幅広く利用されてきた．その後 2013 年には Deep neural network (DNN)を用いた方式が提案され[6]，より品質の高い音声合成を可能にした．これらの方式は，音響特徴量を出力し Vocoder により波形を生成していたが，2016 年に提案された WaveNet [7]は，音声波形を直接生成することで品質の更なる向上を達成した．現在の TTS は End-to-End が主流であり，すでに人間とほぼ等価な品質の TTS が実現されている．

歌声合成に関しては，HMM ベースの Sinsy [3]や DNN ベースの歌声合成[8]が提案されており，音声合成と同様の流れで品質向上が達成されている．Neural Parametric Singing Synthesizer [9, 10]は，WaveNet の考え方を取り入れて音声特徴量を出力する方式であり，シンプルな DNN を用いた方法よりも高い品質を達成している．WaveNet とは異なる方法として，Convolutional Neural Network (CNN)を用いた方法が提案されるなど[11]，品質を高めるための検討が進められている．歌声合成に関しては一般向けのソフトウェアの配布・販売も進められており，特に国内ではニコニコ動画や YouTube において，CeVIO [12]と Synthesizer V [13]を用いたコンテンツが多数投稿されている．フリーソフトでは，2020 年 2 月に公開・配布を開始した NEUTRINO [14]が有名であり，2021 年 6 月現在，「NEUTRINO (歌声合成エンジン)」をタグのキーワードとして検索すると，7,000 件以上の動画が投稿されていることを確認できる．

製品ベースでは，ソフトウェアの評価はユーザに委ねられるが，研究成果として公開する場合には，関連する技術との比較により品質の向上など有効性を示すことが望まれる．一方，歌声合成結果の品質を評価する難しさは，歌声合成の技術的，学習に用いた歌唱 DB の違いなど，品質に影響を与える要因が多岐にわたることにある．同じ手法でも学習に用いる歌唱 DB により合成品質に差が生じる以上，

歌唱 DB も手法も異なる場合，品質で生じた差がどちらの影響なのかを分離して議論することが困難となる．学習用に音声・歌声の統一した歌唱 DB を用いることで，品質に与える要因を絞り込むことが可能となる．

2.2 音声・歌唱データベース

以下では，DB に加えコーパスという単語が複数出てくるが，本論文においては言語に関する情報を集積したものをコーパスとし，言語以外の，例えば譜面や歌声まで含むものは DB として使い分ける．ただし，提案者が命名したものについては，その名称をそのまま利用することとする．

日本語で構成される音声コーパスの先駆けとしては，音素バランス文からなる ATR 音声・言語データベース[1]が挙げられる．日常的に利用される言葉を無作為に選ぶ場合，音素の出現頻度に偏りが生じるため，学習データとして幅広い音素を不足なくカバーするためには膨大なコーパスが必要となる．音素バランス文はこの問題を解決し，相対的に少量のデータから品質の高い音声合成を実現するために貢献している．同様のコンセプトでは，声優統計コーパス[15]が構築されている．高道らによる JSUT と JVS [16]も音素バランス文を含む日本語のコーパスであり，テキスト音声合成研究を加速する目的で構築されている．TTS は言語単位で構築する必要があり，英語など多言語の音声 DB が公開されている．大規模なものでは LibriTTS があり，これは 2,456 話者からなる合計 586 時間もの音声が収録された，TTS 研究のための大規模な音声 DB である．このように，研究目的に応じて多様なコーパス・DB が存在することは，音声合成研究を加速させるために重要な役割を担う．

統計的歌声合成に向けた歌唱 DB の問題は，楽曲には著作権があるため歌声の配布が困難なことである．1 つの解決策は，童謡などの著作権切れの楽曲で構成された歌唱 DB の構築であり，国内では JSUT-song [18]が提案されている．英語の歌唱 DB では，VocalSet [19]が 20 名の歌手から構成される約 10 時間の歌唱データを公開している．歌唱 DB は歌わせる言語毎に必要となり，中国語や韓国語の歌唱 DB も公開されている[20, 21]．楽曲のジャンルにも幅があり，文献[21]では韓国語の童謡を用いている．

著作権については国による差があるため，以下では日本語の歌唱 DB に対象を絞って議論する．著作権切れではない楽曲を用いるアプローチとしては，2019 年に改定された著作権法第三十条の四に基づき，条件付で既存の楽曲を公開した東北きりたん歌唱 DB [4]が挙げられる．既存の楽曲を無作為に用いた場合は，話し声と同様に特定の音素の出現頻度が低いことが問題となる．この問題の解決策として，音素バランス文にメロディを与えた PJS [22]や，既存の楽曲リストを対象に，音素バランスに加え音高や音高差，各音符の継続長まで含めて選定し，1 名の歌手で合計 5 時間もの歌声を収録した LJSing が構築されている[23]．

2.3 歌唱データベースの問題と本研究の位置づけ

日本語の歌唱 DB の中で入手可能な東北きりたん歌唱 DB と PSJ を対象に解析すると，どちらも「にぇ」のように日本語の単語が極端に少ないモーラが含まれない．あらゆる歌を歌えるようにするためには，このようなレアなモーラもカバーした歌唱 DB が必要になる．そのためには，既存の楽曲に縛られず該当するモーラを含む歌詞で構成され，著作権の問題がクリアされた楽曲であることが望ましい．

本研究では，これらの条件を満たすような歌詞の作詞，および著作権の問題が生じないよう作曲したメロディで構成される歌唱 DB を構築する．音高は歌唱者のキーにより調整できることから，音高差と音符の継続長についてバランスを取ることで，統計的歌声合成に適した歌唱 DB の構築を目指す．

3. 構築したデータベース「No.7」

構築した歌唱 DB は，合計 51 曲からなる約 60 分の歌唱データから構成される．VOCALOID などの歌声合成においても歌声にキャラクターを与えることが標準になりつつあるため，この歌唱 DB を統計的歌声合成に組み込むことを想定し「No.7」というキャラクターを設定した．

3.1 設計コンセプト

本歌唱 DB は，作詞・作曲・歌唱を全て同一人物（第三著者である小岩井ことり）が行うこととし，以下の条件を満足するように作業を進めた．

1. 合計で約 1 時間程度の量にする
2. レアなモーラを可能な限り含めるようにする
3. 楽曲のテンポ，音高差，継続長についてもある程度バランスを取る
4. 歌いやすさを重視する

1については，きりたん歌唱 DB を用いたフリーソフトである NEUTRINO により高品質な歌声合成が可能であり，収録時間が約 1 時間であったことに由来する．4 については，歌唱スタイルが合成結果に反映されるため，音素バランスなどを重視しすぎることで歌いやすさが損なわれることが内容に調整する意図である．本歌唱 DB の構築においては，作曲者自身が歌いながら作詞・作曲することで対応することとした．

構築した歌唱 DB の収録条件を表 1 に示す．収録はレコーディングスタジオで実施し，収録後にノイズの除去とサウンドエンジニアによるタイミング・ピッチの補正を実施した．歌唱時間については，作曲に用いた MIDI データに基づいて各音符の長さを算出し，総和することで求めた．歌唱 DB に収録される歌声にはブレスのデータも含まれるが，譜面から算出しているためブレスは歌唱時間に含まれていない．また，収録された歌声は，音符の長さに必ずしも忠実ではないことから，表 1 の数字に示された歌唱時間は譜面に基づく目安である．

表 1：本歌唱 DB の収録条件

歌手	小岩井ことり
曲数	51 曲
歌唱時間（楽譜から概算）	約 3808 秒
レコーディング場所	レコーディングスタジオ
使用マイク	NEUMANN U 87 Ai
サンプリング	96 kHz/32 bit

表 2：他の歌唱 DB との基本情報の比較

	きりたん	PJS	No.7
モーラ種類	103	111	140
モーラ数	11,028	3,899	8,491
継続長 (秒)	3,424	1,137	3,808
周波数幅 (cent)	2,300	3,000	3,500

3.2 基本データの比較分析

はじめに，設計コンセプトに関係する基本情報を他の歌唱 DB と比較分析する．この比較には，歌唱 DB の詳細を入手可能な，きりたん歌唱 DB（表中ではきりたん）[4]と PJS [22]を用いた．LJSing は継続長が 5 時間であることが文献[23]に示されているが，現在入手することが不可能なため本比較には含めていない．統計的な解析は 4 章で実施することとし，ここでは数値的なデータのみ比較する．

表 2 が基本情報の比較である．カバーしているモーラの種類に注目すると，既存の楽曲で構成されるきりたん歌唱 DB は 103 種類であることに対し，声優統計コーパスに基づく PJS では 111 種類である．構築した歌唱 DB（No.7）は 140 種類であり，既存の歌唱 DB よりも多くのモーラをカバーできていることが確認できる．PJS では声優統計コーパスに基づくため音素バランスに優れていることが特色であるが，これは音素単位での出現確率が対象である．レアなモーラの種類で比較すると，本歌唱 DB が最も優れているといえる．

表 3 に全モーラの出現回数をまとめた結果を示す．本表に示すモーラの種類は，Sinsy のリファレンスに基づく．本歌唱 DB では，Sinsy のリストにおける全てのモーラをカバーできていない．具体的には「くぁ（/kwa/）」，「ぐぁ（/gwa/）」などの/kw/と/gw/の音素を含むモーラは，歌いやすい適切な単語を割り当てることが困難なため，候補から外した．同様に，表中では「てゃ（/tya/）」と「でぇ（/dye/）」が 0 となっており，この 2 つについても同様の理由により導入を断念した．これらのモーラ以外については，最低 1 回は含まれている．また，表中では「しぃ」と「すぃ」を区別しているが，Sinsy における音素表記ではどちらも/si/で記述しているため，音素としては同一である．本リストは歌詞のテキストデータに基づいて作成しており，実際の歌手が歌った場合には，歌詞とは異なるモーラとして歌われる可能性があることには注意する必要がある．

モーラ数と継続長を比較すると，PJS が約 19 分であることに対し，きりたん歌唱 DB と本歌唱 DB では 1 時間前後のデータ量を確保している．一方，モーラ数は本歌唱 DB がきりたん歌唱 DB より 33%ほど少ない．これは，きりたん歌唱 DB ではハイテンポの曲が多く 1 モーラあたりの継続時間が短いことに対し，本歌唱 DB ではローテンポの曲も多く，ロングトーンの音符が多く存在していることを示す．歌唱 DB の継続長と音質との関係については，例えば童謡を用いた Neural parametric singing synthesizer [9, 10]では，30 分程度の歌声から高品質な歌声を生成している．約 1 時間という分量は，過去の合成結果から一定の品質を達成可能な目安としている．この量が高品質な歌声合成に十分であるか否かについては継続的な議論が必要となるが，本論文では対象としない．

表 3：各モーラの出現回数．この表には含まれいないが，「ん」は 305 回，「っ」は 198 回出現している．

あ	188	い	603	う	244	え	149	お	208
か	284	き	219	く	218	け	89	こ	177
さ	83	し	209	す	120	せ	54	そ	83
た	311	ち	76	つ	114	て	220	と	290
な	360	に	181	ぬ	18	ね	44	の	222
は	63	ひ	48	ふ	45	へ	11	ほ	44
ま	150	み	165	む	46	め	79	も	170
や	34			ゆ	51	いぇ	7	よ	67
ら	221	り	125	る	214	れ	128	ろ	41
わ	224								
が	132	ぎ	18	ぐ	20	げ	29	ご	35
ざ	25	じ	54	ず	68	ぜ	12	ぞ	12
だ	134					で	125	ど	109
ば	33	び	39	ぶ	21	べ	15	ぼ	28
ぱ	27	ぴ	15	ぷ	15	ぺ	12	ぽ	15
きゃ	12			きゅ	9	きぇ	2	きょ	28
しゃ	12	しぃ	1	しゅ	17	しぇ	6	しょ	29
ちゃ	12			ちゅ	10	ちぇ	5	ちょ	9
てゃ	0	てぃ	15	てゅ	2			てょ	2
にゃ	7			にゅ	5	にぇ	2	にょ	8
ひゃ	5			ひゅ	2	ひぇ	2	ひょ	6
みゃ	4			みゅ	6	みぇ	3	みょ	1
りゃ	7			りゅ	9	りぇ	5	りょ	10
ぎゃ	5			ぎゅ	5	ぎぇ	2	ぎょ	6
じゃ	17			じゅ	14	じぇ	8	じょ	13
びゃ	3			びゅ	3	びぇ	3	びょ	3
ぴゃ	4			ぴゅ	2	ぴぇ	4	ぴょ	4
でゃ	2	でぃ	13	でゅ	2	でぇ	0	でょ	1
		うぃ	10			うぇ	7	うぉ	5
		すぃ	4						
つぁ	6	つぃ	7			つぇ	9	つぉ	3
				とぅ	5				
ふぁ	17	ふぃ	7			ふぇ	4	ふぉ	8
		ずぃ	4						
				どぅ	5				
ヴぁ	7	ヴぃ	8	ヴ	3	ヴぇ	4	ヴぉ	6

周波数幅は，歌声合成システムを実装した際に品質を担保可能な音域に相当する．きりたん歌唱 DB が 2 オクターブ未満であり，本歌唱 DB が最も広く 3 オクターブ近い範囲をカバーしている．本歌唱 DB における音域は，歌手が作曲しながら自身が歌いやすい音域となるよう調整した結果であり，これを歌唱 DB の標準としている．ただし，歌手によっては音域が 3 オクターブに至らない場合もあることから，別の歌手が歌う場合は，キーを変えることで歌いやすい音域へ調整することが想定される．本論文における評価では，この影響までは加味しないこととする．

4. 構築した歌唱データベースの統計解析

構築した歌唱 DB の統計解析により，他の歌唱 DB との性質について議論する．3 章と同様に，比較にはきりたん歌唱 DB と PJS を用いることとする．基本的な統計解析では，譜面に基づく情報のヒストグラムに用いて議論する．

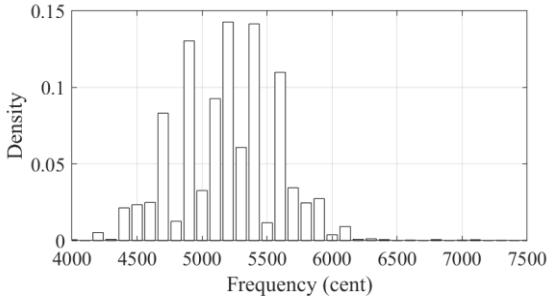

図 1：音高のヒストグラム

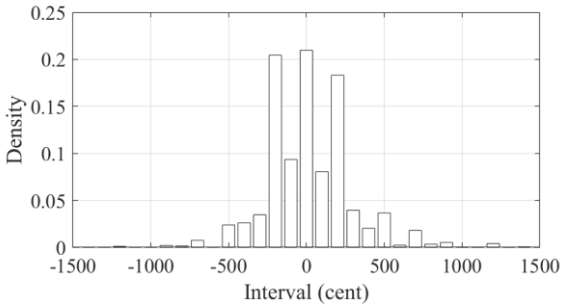

図 2：音高差のヒストグラム

他の歌唱 DB との比較では，エントロピーを用いて評価する．

4.1 基本データのヒストグラム

はじめに，全ノートに対する音高，音高差，継続長を対象としたヒストグラムにより，構築した歌唱 DB の基礎データを解析する．図 1, 2, 3 が，それぞれ算出した音高，音高差，継続長のヒストグラムである．全ての図において，縦軸は出現確率に相当する．音高，音高差の横軸は，それぞれ cent を単位とした音高の絶対値，音高差に相当する．図 3 の横軸は，秒を単位とした継続長に対応する．なお，音高の単位を周波数から cent に変換する際には，以下の式を利用した．

$$f_{\mathrm{cent}} = 1200 \log_2 \left(\frac{f_{\mathrm{Hz}}}{f_c} \right) + 4800 \qquad (1)$$

この式におけるf_{Hz}は変換元となる Hz を単位とした音高を示す．f_cは基準となる音高であり，本論文では C4 に対応する約 261.62 Hz を採用した．この基準であれば，C4 が 4,800 cent となり，1 オクターブが 1,200 cent で記述されることとなる．

図 1 のヒストグラムからも明らかに，音高の出現確率は均一ではなく，偏りが生じていることが確認できる．特に，4,800 cent や 5,500 cent など，一部の音高の出現確率が低くなっており，これは，作曲者が 1 名であり好みとなるキーに偏りが生じていたことが原因と考えられる．本歌唱 DB では，歌手が最も歌いやすいキーで調整しているが，これは，ピッチシフトにより補正することが可能である[23]．ピッチシフトによる偏りを是正する効果については，4.2 節で議論する．

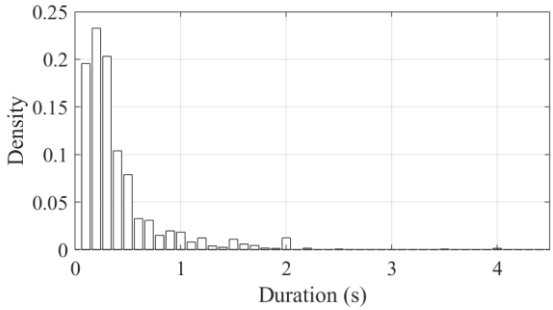

図 3：継続長のヒストグラム

図 2 に示す音高差については，−500 から 500 cent の範囲に多くが集中していることが確認できる．きりたん歌唱 DB でも同様の傾向であり，特に±600 cent の音高差が極端に少ないことも共通した傾向である．ただし，本歌唱 DB では，音高差として±600 cent を最低限入れることを目指しており，−600 cent は 2 回，600 cent は 16 回出現している．音高差は楽曲のキーの変更では変化しないため，音高とは異なりこのヒストグラムが歌唱 DB の性能に直結することとなる．

最後に継続長については，文献[23]と同様に 0.1 秒単位で量子化した結果に基づきヒストグラムを求めている．図 3 からも，0.5 秒以下の継続長が多いことは明らかであるが，1 秒を超えるロングトーンも一定量存在していることが確認できる．収録にあたっては，ビブラートをかけやすいスローテンポの楽曲も存在しているため，ロングトーンが一定量存在することは，特にビブラートなどの表現において有利に働く可能性が考えられる．

ここまで，構築した歌唱 DB の基本的な統計データについて議論してきた．この分布が他の歌唱 DB と比べどの程度偏っているかについては，エントロピーを用いることで数値化して評価する．

4.2 エントロピーによる比較評価

本論文では，以下の式によりエントロピーを計算し評価することとした．

$$S = -\sum_{n=0}^{N-1} p_n \log_2 p_n \qquad (2)$$

$$\sum_{n=0}^{N-1} p_n = 1 \qquad (3)$$

ここで，Nは図 1, 2, 3 それぞれのヒストグラムにおけるバーの総数に相当する．表 3 のモーラについては，モーラの種類に相当する 140 に設定した．p_nは，各事象の出現確率に対応する．一般に，歌詞やテキストのエントロピー計算においては，文献[23]のように音素を単位とする．本論文では，レアなモーラを出現させることを意図しているため，音素単位ではなくモーラ単位でエントロピーを計算した．歌詞をベースに算出しているため，ブレス記号は計算に含まれないが，「っ」はエントロピー計算に含めている．

表 4：エントロピーの算出結果．表中の数値の単位は bit である．

	きりたん	PJS	No.7
モーラ	5.78	5.92	5.91
音高	3.93	4.31	3.70
音高差	3.24	3.18	3.25
継続長	2.54	2.38	3.18

計算されたエントロピーをまとめた結果を表 4 に示す．LJSing については歌詞や譜面が公開されていないが，文献[23]に計算結果が示されている． SongSet と PhraseSet を統合した All について数値を引用すると，音高，音高差，継続長それぞれのエントロピーは 4.33, 3.30, 3.06 bit であった．歌詞については音素単位で計算しているため，モーラでの計算結果は存在しない．

モーラのエントロピーについて比較すると，音素バランスに優れた PJS の結果が最も優れている一方，本歌唱 DB の結果もほぼ等価なエントロピーであることが確認できる．モーラ数は PJS よりも 29 種類多いため，幅広いモーラをカバーできており，モーラ単位での出現頻度もある程度高いことが確認できた．

音高については，図 1 のヒストグラムにおいても偏りが生じていたように，比較した 2 種類の歌唱 DB と比較して本歌唱 DB は最も低いエントロピーである．このエントロピーは，楽曲のキーを以下の手順で図 4 のようにヒストグラムを変化させることにより向上させることが可能である．本論文では，音域を変化させないように，音高の最低値を含む楽曲はキーを下げることはせず，最高値を含む楽曲のキーを上げることはしないように条件を設定した．この条件を満たす範囲で

1. 特定楽曲のキーを標準に対し±1 し，全曲に対してエントロピーを計算
2. もっともエントロピーが高いキーをその楽曲の制御後のキーとして設定

という手順により曲単位でキーを制御し，全楽曲について±1 の範囲でキーを制御する．この際に処理を実施する楽曲の順番はランダマイズし，エントロピーが収束するまで再帰的に処理を繰り返した．今回の処理では，収束しやすくするため標準キー±1 の範囲で固定しているが，この範囲を広げることでさらにエントロピーを向上させることも可能である．処理の結果，本歌唱 DB のエントロピーは 4.11 bit まで向上させることが可能であった．この結果は，きりたん歌唱 DB (3.93 bit)よりは良好であるが，PJS (4.31 bit)や LJSing (4.33 bit)には及ばない数値である．

音高差については PJS が最も低く，きりたん歌唱 DB と本歌唱 DB がほぼ等価な数値であることが確認できる．PJS については，音高に対するエントロピーはメロディ生成の際にばらつかせていた一方，音高差については調整していなかった可能性が示唆される．本歌唱 DB については±600 cent を意図的に含めるようにするなど，音高差が重要であることを作曲時に意識しており，その効果が表れた結果であると解釈できる．LJSing は最も高いエントロピーである

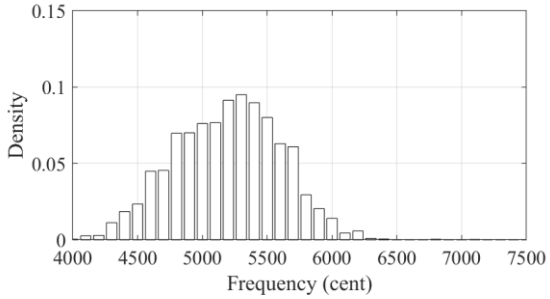

図 4：ピッチシフトにより最適化した本歌唱 DB の音高のヒストグラム．

3.30 bit であり，これは楽曲選択において適切な候補が選ばれていたことを示す．

継続長については，本歌唱 DB が LJSing を含めた全歌唱 DB と比較して最も高い結果が得られた．きりたん歌唱 DB では J-POP やアニメソングを中心としたハイテンポの曲が中心であることが原因であるが，PJS が低い値であることについても，楽曲のジャンルやテンポが偏っていることに原因があると考えられる．

4.3 考察

本歌唱 DB は，レアなモーラを含む 140 種類もの幅広いモーラをカバーしている．エントロピーの評価においても，全てにおいて最高の評価とはならないが，音高差と継続長については，他の歌唱 DB と比較して高い数値が得られた．以下では，いくつかの観点から，本歌唱 DB の有効性について議論する．

歌声のデータ量については，本歌唱 DB がきりたん歌唱 DB よりもやや多いため，NEUTRINO で公開された東北きりたんの歌声と等価な品質が期待される．音域も広く，カバーしているモーラ数も多いことは，相対的に様々な入力に対し破綻しにくくなることが予想される．一方，全体のモーラ数はきりたん歌唱 DB より 33%程少ない．モーラ数が合成結果の品質に与える影響については，今後比較検討することが必要である．今回は，収録された歌声に対してピッチ・タイミングの補正とノイズ除去も実施しているため，合成結果におけるこれらのズレは相対的に少ないことが予想される．ただし，そのずれが「人間らしさ」として評価されることもあるため，補正することが歌声としての魅力に与える影響については，出力された歌声を用いて検討する必要がある．

エントロピーの評価結果から，本歌唱 DB は既存の歌唱 DB と比較しても概ね良好な数値が得られていると言える．一般に，分類問題においては，不均衡データに対する学習において，特に工夫しない場合は少数派の分類精度が弱くなることが知られている．ただし，音声合成においてこの影響がどのようになるかについては，明らかとは言い難い．本歌唱 DB については，エントロピー面だけではなく，カバーするモーラ数そのものは，他の歌唱 DB よりも明らかに多い．このことから，歌唱データのエントロピーとモーラのカバー率の両面から，バランスが取れている歌唱 DB であると考えられる．

4.4 歌唱データベース配布に向けて

本歌唱 DB は，きりたん歌唱 DB と同様に，研究・開発者を対象に配布する．きりたん歌唱 DB では加工を含まない歌唱のみが公開対象であったが，今回の歌唱 DB では，歌声の生データに加え，スタジオのサウンドエンジニアにより調整された歌声も記録している．具体的には，

1. 収録された歌声の生データ
2. 生データからノイズを除去した歌声
3. 2. に対しピッチ補正を実施した歌声
4. 2. に対しタイミング補正を実施した歌声
5. 2. に対し 3.と 4.両方の処理を施した歌声

の 5 パターンを記録している．これらのデータを用いることで，例えば学習データに対し補正をかけることが品質に与える影響の調査が可能となる．

統計的歌声合成に向けては，譜面に相当するデータが必要であり，学習用に音素ラベル付けも必要となる．本歌唱 DB では，最も品質が高いことが想定される 5. の歌声を対象に譜面・ラベル情報を付与し，歌唱 DB に含めて配布する予定である．譜面情報については MusicXML を利用する．

本歌唱 DB が類似したきりたん歌唱 DB と異なる点として，著作権法の例外規定に縛られず利用できることが挙げられる．著作権法第三十条の四では人間が聴取することは認められないため，歌唱 DB に収録された歌声の聴取実験等は，DB 製作者などの一部例外を除き認められなかった．本歌唱 DB にはこの問題が生じないため，統計的歌声合成以外の用途に対しても，研究用であれば利用することが可能となる．

5. おわりに

本論文では，統計的手法に基づく歌声合成のために利用可能な歌唱 DB を構築し，有効性について検証した．本歌唱 DB の特色は，日本語では出現しにくいレアなモーラを含んでおり，既存の音素バランス文に基づく音声 DB と比較しても，より多くのモーラをカバーしていることを示した．音高差や継続長に関するエントロピー評価では，既存の歌唱 DB と比較しても，概ね等しいかそれ以上の性能を有することが確認された．

今後の課題として，本歌唱 DB を用いた統計的歌声合成の品質の検証が挙げられる．本歌唱 DB や既存の歌唱 DB を用いることで，同じアルゴリズムを用いて歌唱 DB による差を見極めることや，同じ歌唱 DB を用いて歌声合成アルゴリズムの差などの検証が容易になる．これらの検討から，統計的歌声合成において妥当となる歌唱 DB の規模について検証を続けていくことを計画している．

謝辞

本研究は，JST さきがけ JPMJPR18J8，JSPS 科研費 JP21H04900，JP19K12736 の支援を受けました．また，作成において協力頂いた SHACHI 氏（STUDIO NEUTRINO），および Studio KSP の皆様に深謝いたします．

参考文献

[1] 匂坂芳典, 浦谷則好：ATR 音声・言語データベース，音響誌, vol. 48, no. 12, pp. 878-882, 1992.
[2] H. Kenmochi and H. Ohshita: VOCALOID – Commercial singing synthesizer based on sample concatenation, in Proc. INTERSPEECH 2007, pp. 4009-4010, 2007.
[3] K.Oura, A.Mase, T.Yamada, S.Muto, Y.Nankaku, and K.Tokuda: Recent development of the HMM-based singing voice synthesis system — Sinsy, in Proc. SSW7, pp. 211-216, 2010.
[4] I. Ogawa and M. Morise: Tohoku Kiritan singing database: A singing database for statistical parametric singing synthesis using Japanese pop songs, Acoustical Science and Technology, vol. 42, no. 3, pp. 140-145, 2021.
[5] H. Zen, K. Tokuda, and A. W. Black: Statistical parametric speech synthesis, Speech Communication, vol. 51, no. 11, pp. 1039-1064, 2009.
[6] H. Zen, A. Senior, and M. Schuster: Statistical parametric speech synthesis using deep neural networks, in Proc. ICASSP 2013, pp. 7962-7966, 2013.
[7] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior and K. Kavukcuoglu: WaveNet: A generative model for raw audio, arXiv 1609.03499, 2016.
[8] M. Nishimura, K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda: Singing voice synthesis based on deep neural networks, in Proc. INTERSPEECH, pp. 2478-2482, 2016
[9] M. Blaauw and J. Bonada: A neural parametric singing synthesizer, in Proc. INTERSPEECH 2017, pp. 4001-4005, 2017.
[10] M. Blaauw and J. Bonada: A neural parametric singing synthesizer modeling timbre and expression from natural songs, Appl. Sci., vol. 7, no. 12, pp. 1-23, 2017
[11] K. Nakamura, S. Takaki, K. Hashimoto, K. Oura, Y. Nankaku, K. Tokuda: Fast and high-quality singing voice synthesis system based on convolutional neural networks, in Proc. ICASSP 2020, pp. 7239-7243, 2020.
[12] CeVIO: https://cevio.jp/ (2021 年 5 月 18 日閲覧)
[13] Synthesizer V: https://dreamtonics.com/synthesizerv/ (2021 年 5 月 18 日閲覧)
[14] NEUTRINO: Neural singing synthesizer: https://n3utrino.work/ (2021 年 5 月 18 日閲覧)
[15] 声優統計コーパス：https://voice-statistics.github.io/ (2021 年 5 月 18 日閲覧)
[16] S. Takamichi, R. Sonobe, K. Mitsui, Y. Saito, T. Koriyama, N. Tanji, and H. Saruwatari: JSUT and JVS: Free Japanese voice corpora for accelerating speech synthesis research, Acoust. Sci. & Tech., vol. 41, no. 5, pp. 761-768, 2020
[17] H. Zen, V. Dang, R. Clark, Y. Zhang, R. J. Weiss, Y. Jia, Z. Chen, and Y. Wu: LibriTTS: A corpus derived from LibriSpeech for text-to-speech, in Proc. INTERSPEECH 2019, pp. 1526-1530, 2019.
[18] JSUT-song https://sites.google.com/site/shinnosuketakamichi/publication/jsut-song (2021 年 5 月 18 日閲覧)
[19] J. Wilkins, P. Seetharaman, A. Wahl, and B. Pardo: VocalSet: A singing voice dataset, in Proc. ISMIR 2018, pp. 468-474, 2018.
[20] P. Lu, J. Wu, J. Luan, X. Tan, and L. Zhou: XiaoiceSing: A high-quality and integrated singing voice synthesis system, in Proc. INTERSPEECH 2020, pp. 1306-1310, 2020.
[21] J. Kim, H. Choi, J. Park, M. Hahn, S. Kim, and J. Kim: Korean singing voice synthesis system based on an LSTM recurrent neural network, in Proc. INTERSPEECH 2018, pp. 1551-1555, 2018.
[22] J. Koguchi, S. Takamichi, and M. Morise: PJS: phoneme-balanced Japanese singing-voice corpus, in Proc. APSIPA ASC 2020, pp. 487-491, 2020.
[23] 藤村拓憧，能勢隆，伊藤彰則：LJSing: 単一歌唱者による大規模日本語歌声コーパス，音講論（秋），pp. 871-874, 2020.
[24] 大浦圭一郎，間瀬絢美，南角吉彦，徳田恵一：HMM 歌声合成における音高正規化学習の検討，情報処理学会研究報告，vol. 2012-MUS-94, no. 15, pp. 1-6, 2012.