jaCappella コーパス v2：重唱分離・合成のための日本語アカペラ重唱コーパスの拡張

中村 友彦[1,a)]　高道 慎之介[2,b)]　丹治 尚子[2,c)]　深山 覚[1,d)]　猿渡 洋[2,e)]

概要：我々は計算機や人と相互作用可能な歌声合成・分析基盤の構築を目指し，歌唱データ資源として日本語アカペラ重唱コーパス（jaCappella コーパス）を構築してきた．jaCappella コーパスは，35 曲の著作権処理済み 6 声重唱曲の譜面と各声部（リードボーカル，ソプラノ，アルト，テナー，バス，ボーカルパーカッション）の歌唱音源からなる．楽曲はジャズや演歌など各ジャンルの典型的な特徴を持つ 7 つのサブセットに分かれており，様々なジャンルや歌唱スタイルの音源を含む．本発表では，より多様なジャンルや歌唱スタイルを包含するコーパスの作成を目指し，本コーパスを拡張した jaCappella コーパス v2 の構築に関して報告する．この構築のため，バラード，エレクトロニック・ダンス・ミュージック，ソウル・ファンクのジャンルに対応する計 3 つのサブセット，計 15 曲の重唱曲を新たに作成した．当該コーパスを重唱分離に適用し，既存の音源分離手法の性能評価を行う．

1. 序論

合唱は様々な文化圏で普及しているグループ歌唱形式であり，各声部を 1 人の歌唱者が歌う形式（重唱），同一声部を複数人で歌唱する形式（斉唱）などに細分される．合唱は音楽情報処理の重要な研究対象の 1 つであり，長年研究されてきた [1]．近年の機械学習の発展に伴い [2,3]，多重基本周波数（F_0）推定 [4,5]，重唱を対象とした音源分離（重唱分離）[6–8], 自動採譜 [9]，斉唱歌声生成 [10]，ダブルトラッキング歌唱の生成 [11] など合唱を対象とした様々な音楽情報処理タスクでもデータ駆動型アプローチが用いられている．このアプローチに基づく手法では，適した合唱データがどれだけ利用できるかが重要となる．

表 1 に既存の合唱データセットを示す．これらのデータセットの多くは，西洋音楽で広く用いられているソプラノ（S），アルト（A），テナー（T），バス（Bs）の 4 声部形式の合唱を収録している．同一声部で複数人が歌唱する斉唱を含むデータセットもあるものの [4,12]，主要な対象は重唱である [4–9]．歌唱曲を絞り練習時の音源も公開しているデータセット [4,12] や，40 曲以上の録音を含むデータセットもある [8]．他方ジャンルに関しては，讃美歌など伝統的な合唱が対象とされてきた．

より多様なジャンルを網羅するため，我々は日本語アカペラ重唱コーパス（jaCappella コーパス）を構築してきた [13][*1]．jaCappella コーパスは, 近年 Youtube や TikTok などのソーシャルメディアサービスで広まっている重唱スタイルを対象とし作成されている．この重唱スタイルでは S，A，T，Bs に加え，リードボーカル（Vo），ヴォーカルパーカッション（VP）を含む．VP は口や声を用いて打楽器音を模倣した発声をする歌唱法であり，ドラムなどの代わりにリズムを担うことができる．そのため，様々なジャンルの楽曲の重唱アレンジが歌われており，従来の合唱データセットの対象に比べジャンルや歌唱スタイルは多様である．そこで，jaCappella コーパスではそれぞれ異なるジャンルに対応する 7 つのサブセットを作成した．

本稿では，より多様なジャンルや歌唱スタイルを包含する重唱コーパスを構築するために行った，jaCappella コーパスの拡張について報告する．以降では，既存の jaCappella コーパスと本稿で報告する拡張を含んだ jaCappella コーパスを区別するため，前者を jaCappella コーパス v1，後者を jaCappella コーパス v2 と呼ぶ．jaCappella コーパス v2 では，jaCappella コーパスの構築方針に沿って新たに 15 曲を作成した．これらの曲は，5 曲毎にバラード（ballad），

[1] 産業技術総合研究所 人工知能研究センター
2-4-7 Aomi, Koto-ku, Tokyo, 135-0064, Japan
[2] 東京大学 大学院情報理工学系研究科
7-3-1 Hongo, Bunkyo, Tokyo 113-8654, Japan
[a)] tomohiko.nakamura.jp@ieee.org
[b)] shinnosuke_takamichi@ipc.i.u-tokyo.ac.jp
[c)] naoko_tanji@ipc.i.u-tokyo.ac.jp
[d)] s.fukayama@aist.go.jp
[e)] hiroshi_saruwatari@ipc.i.u-tokyo.ac.jp

[*1] jaCappella コーパスは，https://tomohikonakamura.github.io/jaCappella_corpus/ から入手可能である．

表 1: Specifications of our and conventional vocal ensemble corpus

Corpus/Dataset	Voice parts	Dur. [min]	# songs	Genre
Choral Singing [14]	S, A, T, Bs	7	3	Choral music
Dagstuhl ChiorSet [12]	S, A, T, Bs	55	2	Choral music
ESMUC Choir [4]	S, A, T, Bs	31	3	Choral music
Bach Chorales and Barbershop Quartets [8]	S, A, T, Bs	104	48	Choral and barbershop music
jaCappella v1 [13] (ours)	Vo, S, A, T, Bs, VP	34	35	Jazz, punk rock, bossa nova, popular, reggae, enka neutral (children's song)
jaCappella v2 (ours)	Vo, S, A, T, Bs, VP	54	50	Jazz, punk rock, bossa nova, popular, reggae, enka neutral (children's song), **ballad**, **EDM**, **soul/funk**

エレクトロニック・ダンス・ミュージック（EDM），ソウル・ファンク（soul/funk）の 3 つのサブセットからなり，従来の jaCappella コーパスに含まれるサブセットとは異なるジャンルの特徴をもつ．学術的な目的での研究結果例の公開を容易にするため，これらの曲は著作権保護期間の過ぎた楽曲を当該ジャンルの重唱曲へと編曲し作成されている．また，重唱分離や重唱を対象とした歌声合成（重唱合成）へと利用できるように，声部毎に歌唱音源を整備した．

2. jaCappella コーパス v1 [13]

本節では，jaCappella コーパス v1 の設計方針，重唱曲の作成，歌声収録方法に関して述べる．

2.1 コーパス設計

jaCappella コーパスの設計では，楽譜販売サイト[*2]から重唱アレンジ譜面を 60 曲分収集し，それらの曲の構成を基に構成を決定した．また，研究用途での自由な利用を可能とするために著作権や著作隣接権に関してどのように処理するか，どのようなデータフォーマットで整備するかも設計段階から決定した．以下に，設計に関する各項目について述べる．

声部構成：商用音源 60 曲のうち 46 曲が 6 声を採用しており，そのうち 42 曲が Vo，S，A，T，Bs，VP からなっていたため，jaCappella コーパスでも当該声部構成を採用した．以降はこの 42 曲を商用楽曲群と呼ぶ．

歌唱者の性別：歌唱者の性別は音域に関わるため，声部毎に決定した．商用楽曲群において，S，A は全て女性歌唱者，Bs，VP は全て男性歌唱者，T は 67 %が男性であった．これらの声部に関しては，商用楽曲群で優勢であった性別を採用した．一方，Vo は 81 %が男性であったものの，Vo の性別には歌詞や曲調も影響しうる．また，歌声合成においては女性の方が比較的合成しやすい傾向があるため，そちらを優先し Vo の性別は女性とした．

ジャンル：商用楽曲群は，ポップスから演歌まで幅広いジャンルの楽曲から編曲されて作成されていた．この特徴を反映するため，各ジャンルの特徴を捉えた複数のサブセットを作ることとした．

著作権と著作隣接権：商用楽曲を利用する方が実データに近いものの，著作権の関係からデータセットとしての配布や，それを用いて行った研究結果を Web 上で公開する際に制限が生じる．そこで，作詞・作曲の著作権保護期間が終了した楽曲を編曲することで，重唱曲を作成する方式を採用した．これにより，編曲に関する著作権と著作隣接権について適切に処理を行い，容易に研究利用できるように整備した．また，研究利用から商用利用へ円滑に移行できるようにも整備を行っている．

データフォーマット：重唱曲の譜面は，PDF および MusicXML 形式 [15] のファイルとして整備する．これは，MusicXML 形式は歌声合成 [16] や様々な音楽情報処理タスク [17] で用いられており，様々な重唱に関するタスクに利用しやすいためである．譜面に加え，重唱分離や重唱合成に利用するためには各声部の収録音源が必要である．音源ファイルは標本化周波数 48 kHz，ビット深度 24 bit，RIFF WAV 形式で整備する．

2.2 重唱曲への編曲

2.1 節の設計指針に基づき，著作権保護期間の終了した日本語の童謡・唱歌を編曲することで 35 曲の重唱曲を作成した．編曲元の楽曲は [18, 19] から選択した．また，編曲は 5 人のプロの編曲者が行った．Vo の歌詞と旋律は編曲元楽曲を基に作成されており，全ての曲で主旋律を担当する．それ以外の声部は編曲によって新たに作成されたものであり，VP 以外の声部には全て歌詞が付与されている．

作成した 35 曲は，5 曲毎に異なるジャンルに対応するサブセットに属する．表 2 に各サブセットの情報を示す．jaCappella コーパス v1 で作成したサブセットは jazz, punk rock，bossa nova，popular，reggae，enka，neutral の 7 つであり，各サブセットに含まれる重唱曲はサブセットの名称に対応するジャンルに編曲されている．ただし，neutral

[*2] https://elevato-music.com/?mode=cate&cbid=1727017&csid=0

表 2: Subset specifications of jaCappella corpus v1 and v2

Subset	Song titles (in Japanese)	Ensemble configuration	Duration [s]
Jazz	お玉じゃくし，ポプラ，七つの子，赤い靴，待ちぼうけ	Vo1, S1, A1, T1, Bs1, VP1	226.7
Punk rock	あの町この町，かえるとくも，しゃぼん玉，埴生の宿，蝶々	Vo2, S2, A2, T2, Bs1, VP2	310.7
Bossa nova	どんぐりころころ，浦島太郎，漁船，春よ来い，通りゃんせ	Vo3, S3, A2, T3, Bs2, VP3	334.5
Popular	こいのぼり，靴が鳴る，赤とんぼ，雪，茶摘	Vo1, S1, A1, T1, Bs1, VP1	352.5
Reggae	うさぎとかめ，おもちゃのマーチ，汽車，証城寺の狸囃子，村の鍛冶屋	Vo3, S3, A2, T3, Bs2, VP3	228.7
Enka	ひらいたひらいた，ふじの山，あおげば尊し，荒城の月，十五夜お月さん	Vo2, S2, A2, T2, Bs1, VP2	361.1
Neutral	かたつむり，こもりうた，鳩，春が来た，桃太郎	Vo1, S4, A3, T4, Bs1, VP4	260.1
Ballad	ひよこ，案山子，叱られて，動物園，仲よし小道	Vo1, S1, A1, T1, Bs1, VP1	403.5
EDM	めえめえ児山羊，黄金虫，紅葉，砂山，山寺の和尚さん	Vo1, S1, A1, T1, Bs1, VP1	391.7
Soul/funk	どこかで春が，揺籃のうた，犬，赤い帽子白い帽子，牧場の朝	Vo1, S1, A1, T1, Bs1, VP1	340.9

は原曲の雰囲気を損なわずに編曲を行ったものである．各ジャンルの特徴として，例えば jazz に含まれる曲は Bs にウォーキングベースラインを含む．ウォーキングベースラインは音高に関して上昇と下降を繰り返しながら並ぶ同一音価の音符列であり，ジャズの典型的な特徴の 1 つである．

非日本語話者も jaCappella コーパスを利用しやすくするため，日本語歌詞の譜面に加え，Hepburn 式ローマ字表記の歌詞の譜面も整備した．この整備の際には，譜面上の表記と異なる発音となる文字（例えば，助詞の「は」を「わ」と発音する場合）は人手で発音通りのローマ字表記へと変換した．

2.3 歌声収録

重唱曲を作成後，各声部の歌唱音源の収録を行った．COVID-19 の感染対策のため，収録は歌唱者毎にレコーディングスタジオで実施した．マイクロホンには Shure SM58 を用い，チャネル数は 1，標本化周波数は 48 kHz，ビット深度は 24 bit とした．歌唱収録中，各歌唱者には対象の参考音源（俗に言う仮歌）や当該収録時までに収録済みの他声部の歌唱音源，拍位置を示すティック音を必要に応じてヘッドホンを通し提示した．歌唱者は計 20 人のセミプロ歌手であり，サブセット毎に歌唱者の割り当てを決定した．表 2 の Ensemble configuration の列に，サブセット毎の歌唱者の識別子を示す．Vo は 3 人，S は 4 人，A は 3 人，T は 4 人，Bs は 2 人，VP は 4 人であった．

収録後は，通常の音楽制作と同様に複数テイクを組み合わせベストテイクを作成した．このとき，歌声の音高が楽譜の音高と半音以上異なる場合にはピッチ修正を行った．ピッチ修正は，プロの音響エンジニアが Melodyne を用いて行った．リバーブやイコライザーなども通常の音楽制作で使用されるが，歌声合成性能を低下させる可能性があるため，ピッチ修正以外の処理は行わなかった．表 2 の Duration の列に各サブセットの合計時間を示す．各曲の全声部の歌唱音源は時間的に同期されており，重唱分離や重唱合成で利用しやすくなっている．当該コーパスが重唱分離のベンチマークとして利用できるように，モノラルとステレオの混合音も含め配布している．

3. jaCappella コーパスの拡張

jaCappella コーパス v1 では，従来の重唱コーパスやデータセットよりも広範なジャンルの重唱曲を収録した．この方針を推し進め，さらに 3 つの新たなサブセットを当該コーパスに追加し jaCappella コーパス v2 を構築した．

3.1 重唱曲への編曲

jaCappella コーパス v2 でも，2.1 節で述べた設計方針に沿って重唱曲を作成した．作成したサブセットは ballad，EDM，soul/funk の 3 つであり，これまでに作成したサブセットに含まれる曲と重複がないように編曲元楽曲を選択した（表 2 参照）．

Ballad：バラードに編曲された重唱曲からなり，他のサブセットと比べ比較的テンポが遅く，平均 beats per minutes（BPM）は 72.8 である．また，他のサブセットも含めた全ての曲の中で唯一「仲よし小道」にのみフィンガースナップが 7 つ目のパートとして含まれる．フィンガースナップは譜面上は別の声部として表記した．

EDM：エレクトリック・ダンス・ミュージックを模した重唱曲からなり，当該ジャンルの特徴である短いフレーズの繰り返しが多い．主に，Bs と VP にその特徴が現れており，Bs にはグリッサンドも含まれる．

Soul/funk：ソウルとファンクの特徴を備えた重唱曲からなり，他のサブセットに比べ Bs が他の声部に比べて目立ちやすいリズムとなっている．

図 1 は作成した譜面の例である．基本的には Vo のみが原曲の歌詞を歌うものの，2 小節目のように複数声部で同時に歌う部分もある．また，4 小節目のように特定声部が独自に担当する部分もある．さらに，通常の歌詞に加え，図 1 で紫色で示される「ha」や「tu」などの言語的に無意味な語句も含む．本稿では，前者を語彙的シラブル，後者を非語彙的シラブルと呼ぶ．これらの特徴は商用楽曲群で

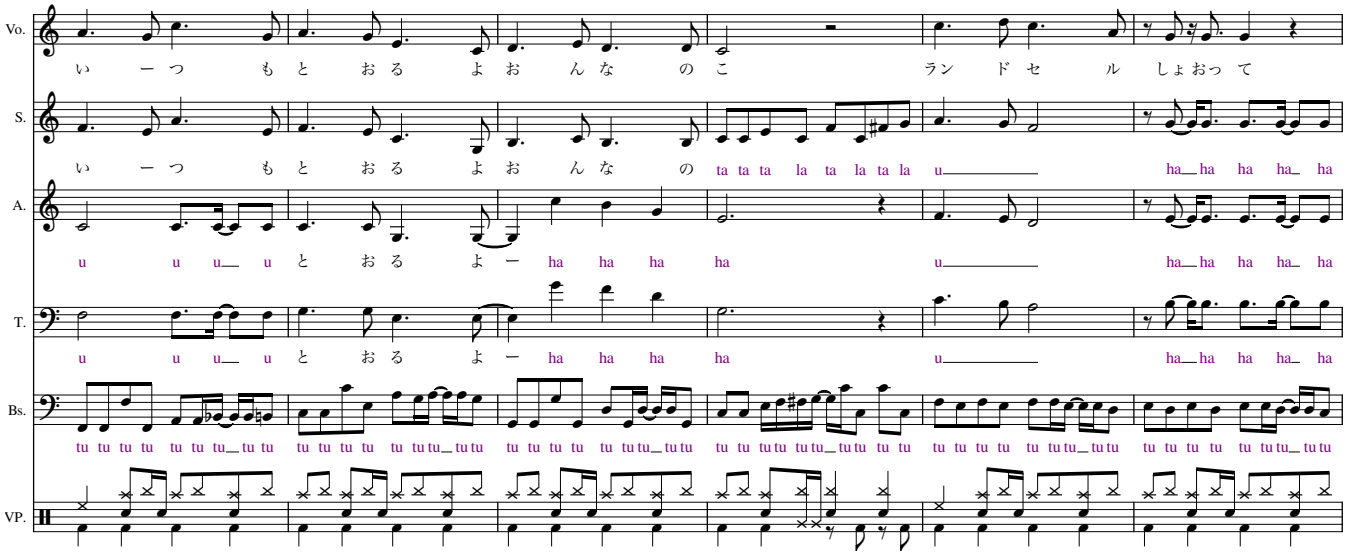


図 1: Musical score excerpt of “Akaiboushishiroiboushi (赤い帽子白い帽子)”. Non-lexical syllables are colored in purple.

も見られる特徴であり，作成した重唱曲が実際の重唱曲と類似した傾向を持つことを示している．

3.2 歌声収録

2.3 節と同様に歌声収録を行った．jaCappella コーパス v2 で追加したサブセットの歌唱者は，jazz と popular のサブセットと同一である．EDM の音響的特徴を反映するため，当該サブセットの VP の収録の際にはフィルタのかかったような演奏となるように歌唱者に指示した．第 1 著者が聞く限り他のサブセットの VP とは異なる音色となっており，VP の音響的な差異も反映したデータを収録できた．フィンガースナップに関しても個別に収録を行い，当該音源の有無を変更できるように作成した．ベストテイクの作成に関しても 2.3 節と同様に行った．

図 2 は各声部のスペクトログラムの例である．音響信号からスペクトログラムへの変換には定 Q フィルタバンクを用いた．Vo から Bs までの声部では調波構造が明瞭である一方，VP は多くの部分が打楽器音のスペクトログラムに類似した構造を持つ．

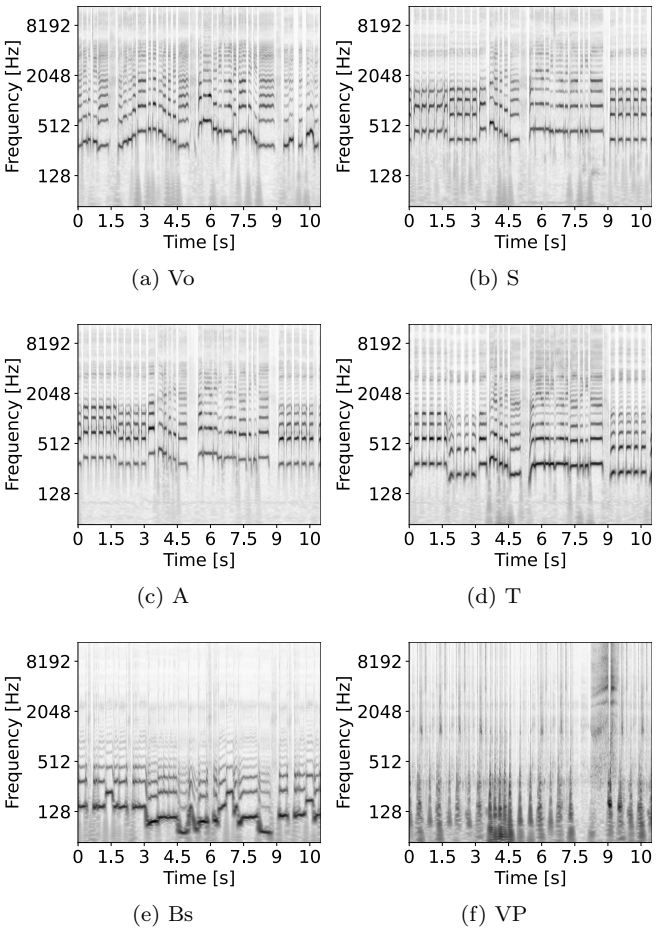
4. jaCappella コーパス v2 の分析

本節では，歌詞と音高に関して jaCappella コーパス v2 のデータの分析を行う．

4.1 歌詞

歌詞におけるシラブルの頻度の偏りは重唱曲の重要な特徴の 1 つである．シラブルの出現頻度には偏りがあり，各声部における語彙的シラブルと非語彙的シラブルの頻度も異なる．本節では，これら 2 点に関し商用楽曲群と比較をしつつ jaCappella コーパス v2 について分析を行う．

図 3 は jaCappella コーパス v2 と商用楽曲群の各シラブ



(a) Vo (b) S (c) A (d) T (e) Bs (f) VP

図 2: Spectrogram examples of all voice parts. These examples are excerpts of “Koganemushi (黄金虫)”.

ルに関する頻度を表す．ただし，シラブルのインデックスは降順に並び替えられている．jaCappella コーパス v2 の方が商用楽曲群より曲数が多いのにも関わらずシラブルの合計頻度が少ないのは，商用楽曲群の各曲長が jaCappella

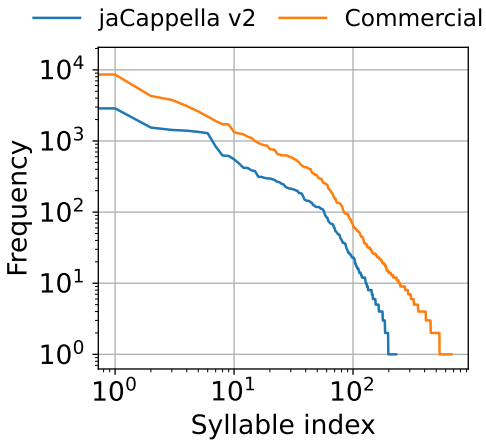


図 3: Syllable frequency of jaCappella corpus v2 and Commercial. Syllable indices are sorted in descending order.

表 3: Frequency of lexical and non-lexical syllables in jaCappella corpus v2 and commercial music collection (Commercial)

Voice part	jaCappella v2 [%]		Commercial [%]	
	Lexical	Non-lexical	Lexical	Non-lexical
Vo	79.5	20.5	91.1	8.9
S	38.2	61.8	34.2	65.8
A	31.1	68.9	35.1	64.9
T	30.4	69.6	36.3	63.7
Bs	0.9	99.1	2.9	97.1

コーパス v2 の曲長よりも長いためである．具体的には，jaCappella コーパス v2 の平均曲長は 64 s，商用楽曲群の平均曲長は 220 s である．シラブルの出現頻度は特定のシラブルに集中する傾向があり，出現頻度の高いシラブルは非語彙的シラブルであった．曲によっては 1 つの非語彙的シラブルが 9 割以上出現頻度を占めているものもあった．この傾向は両データで見られており，jaCappella コーパス v2 がシラブルの出現頻度に関する実データの特性を反映できていることを示す．

表 3 に VP 以外の各声部での語彙的シラブルと非語彙的シラブルの頻度を示す．この頻度は声部によって大きく異なり，特に Bs では非語彙的シラブルが多い．割合として多少の差異はあるものの，この傾向は両データで共通しており，この点でも jaCappella コーパス v2 が実データの特性を反映できていることが確認できる．Vo では 10%以上差があるものの，これは jaCappella コーパス v2 の作成の際に編曲元の楽曲の長さが 30 s よりも短い場合に曲の最初と最後に非語彙的シラブルからなる部分を付加したためである．

4.2 音高

図 4 に VP 以外の声部毎の音高の分布を示す．Vo，T に関して，jaCappella コーパス v2 はそれぞれ女性，男性歌唱者のみ，商用楽曲群は男女共に含むことに注意されたい．商用楽曲群の Vo は男女の歌唱者が含まれるため音域の差異は大きいものの，声部毎の音高の分布は両データとも類似しており，jaCappella コーパス v2 は実データの音高分布の傾向をある程度反映していることが確認できる．

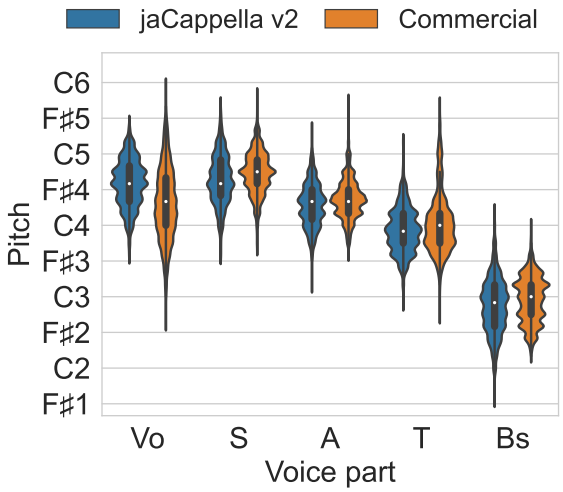


図 4: Pitch distribution of jaCappella corpus v2 and Commercial per voice part.

jaCappella コーパス v2 に関して，Vo と S の音高分布は大部分が重複しており，次いで A，T とも部分的に重複している．一方，Bs は他の声部と重複が少ない．そのため，特に Vo，S，A，T の重唱分離は難しいことが予期される．一方，Bs の音高は通常の歌唱に比べて低周波数帯域に偏っており，周波数解像度によっては分離が困難となることも予期される．これらに関しては，5.3 節で重唱分離性能と照らし合わせ議論する．

5. 重唱分離への適用

本節では，jaCappella コーパス v2 の応用例の 1 つとして重唱分離実験を行い，既存の音源分離手法の性能について議論する．

5.1 実験条件

対象となる重唱分離のタスクでは，48 kHz の標本化周波数のモノラル混合音を 6 つの各声部の音源信号に分離する．学習，検証，テストデータに関しては，5.2 節と 5.3 節で異なるため各々の節で述べる．評価指標として，scale-invariant source-to-distortion ratio（SI-SDR）[20] の改善量（ΔSI-SDR）を用いた．

比較手法は [13] と同様に以下の 3 手法を用いた．これらの実装は著者らの GitHub リポジトリ[*3]で公開しており，パラメータは [13] と同一とした．

[*3] https://github.com/TomohikoNakamura/asteroid_jaCappella/tree/jaCappella/egs/jaCappella

DPTNet：時間波形領域で音声分離を行う DPTNet [21] を基に，S，A，T，Bs の 4 声部の重唱分離のために提案された手法である [8]．当該手法は 22.05 kHz の標本化周波数用に設計されていたため，最初の畳み込み層と最後の畳み込み層のカーネルサイズとストライドをそれぞれ 32, 16 とすることで 48 kHz 用対応させた．エポック数は 600 とし，その他のパラメータは著者による実装[*4]と同一の値を用いた．学習に音源の出力順が任意となることをを許容した permutation invariant training（PIT）を用いているため，分離音の SI-SDR の計算の際には最も当該指標が高くなる分離音と正解音源の割り当てを算出し用いた．

MRDLA：時間波形領域で分離を行う楽音分離手法であり，ウェーブレット変換にヒントを得たダウンサンプリング層を持つことが特徴である [22]．Haar ウェーブレットを用いたモデルを基に，チャネル数，カーネルサイズ，活性化関数を変更し重唱分離用のモデルを作成した．原論文 [22] の表記を用いて，$C^{(\mathrm{e})}$ を 18 から 32 に，$f^{(\mathrm{e})}$ と $f^{(\mathrm{d})}$ を 21 に変更した．また，活性化関数として parameteric rectified linear units の代わりに Gaussian error linear units を用いた．さらに，ロス関数として [23] で提案された時間領域のロス関数と時間周波数領域のロス関数を統合したものを用いた．パラメータの学習には学習率 1.0×10^{-4} の Adam を用い，エポック数は 1000 とした．

X-UMX：当該手法は Music Demixing Challenge 2021 [24] で楽音分離のベースラインとして用いられたものを，重唱分離に転用したものである．エポック数は 1000 とし，その他のパラメータは公式の実装[*5]と同一とした．

これらの手法を学習し，それぞれ検証データが最も低くなったエポックのモデルを用いた．

5.2 追加データに対する分離性能

本節では，jaCappella コーパス v1 のデータで学習されたモデルので性能を比較する．学習済みモデルとして [13] で用いられたものを使用した．このモデルは jaCappella コーパス v1 の 35 曲のうち，各サブセットから 1 曲ずつ取り去った計 28 曲のデータを学習，検証データに利用したものである．学習と検証データは 28 曲を [25] で提案されたデータ拡張法を適用し作成した．この手法では，異なる曲同士の各声部の音源信号をテンポと音高を適切に変更して組み合わせ，新たな混合音と正解音源信号の組を作成する[*6]．最終的に得られた学習，検証データの時間長は 7650.7，2811.7 s であった．このデータ拡張に加え，

[*4] https://github.com/saurjya/asteroid/tree/4e00daa4c4da77bbee6c0109fa4e2c3611217e72
[*5] https://github.com/asteroid-team/asteroid/tree/master/egs/musdb18/X-UMX
[*6] データ拡張の実装は https://github.com/TomohikoNakamura/asteroid_jaCappella/blob/jaCappella/egs/jaCappella/automix.py で公開している．

表 4: Average ΔSI-SDRs [dB] of vocal ensemble separation methods for each subset. Results were obtained with methods trained with jaCappella corpus v1

Subset	Method	Vo	S	A	T	Bs	VP
Ballad	DPTNet	13.5	12.3	12.7	16.9	19.5	23.8
	MRDLA	12.2	8.9	8.8	14.2	17.1	21.0
	X-UMX	8.7	6.6	8.4	13.5	15.0	22.3
EDM	DPTNet	13.5	12.2	11.4	13.4	18.6	19.7
	MRDLA	9.1	7.3	6.4	12.0	17.1	17.2
	X-UMX	7.1	6.2	5.3	9.3	14.6	17.5
Soul/funk	DPTNet	13.1	11.7	12.1	14.6	16.4	17.7
	MRDLA	11.9	9.7	9.5	13.2	10.3	9.0
	X-UMX	10.6	8.6	8.0	10.8	9.5	9.5

表 5: Average ΔSI-SDRs [dB] of DPTNet, MRDLA, and X-UMX trained with jaCappella corpus v1 and v2

Method	Vo	S	A	T	Bs	VP
DPTNet v1	10.5	9.9	12.7	15.3	19.7	21.2
DPTNet v2	9.5	9.0	9.7	12.8	18.2	19.0
MRDLA v1	10.1	9.2	9.0	13.3	17.5	19.4
MRDLA v2	11.3	10.0	9.8	13.7	17.9	20.9
X-UMX v1	9.1	8.3	8.6	11.9	15.7	19.7
X-UMX v2	9.0	8.4	7.2	11.9	15.9	20.0

バッチ生成の際に対象の曲からランダムに切り出した後，$[0.25, 1.25]$ 上の一様分布からサンプルしたゲインを適用し，混合音を作成した．テストデータとしては，新たに追加した ballad，EDM，soul/funk のサブセットの計 15 曲を用いた．テストデータの歌唱者は学習データにも含まれているものの，サブセットに関してはオープンな状況である．

表 4 はサブセット毎の各手法の平均 ΔSI-SDR である．jaCappella コーパス v1 で学習に使用しなかった 5 曲に関しては DPTNet と MRDLA が同程度の性能であったが [13]，未知のサブセットの曲に関しては DPTNet の ΔSI-SDR が高かった．次いで，MRDLA の ΔSI-SDR が高く，時間波形領域の分離手法が有効であることが確認できる．DPTNet は特に Vo と Bs において他の手法に比べ優位に高く，PIT による学習が未学習のジャンルに対して有効である可能性がある．

5.3 追加データの有無による分離性能の比較

追加したサブセットの学習への効果を検証するため，jaCappella コーパス v2 の全サブセットを使った学習モデルを作成した．5.2 節と同様に，各サブセットから 1 曲ずつ取り去った計 40 曲をデータ拡張した．取り去った計 10 曲をテストデータとして用いた．作成した学習，検証データの時間長はそれぞれ 10938.0，4396.7 s であった．以下

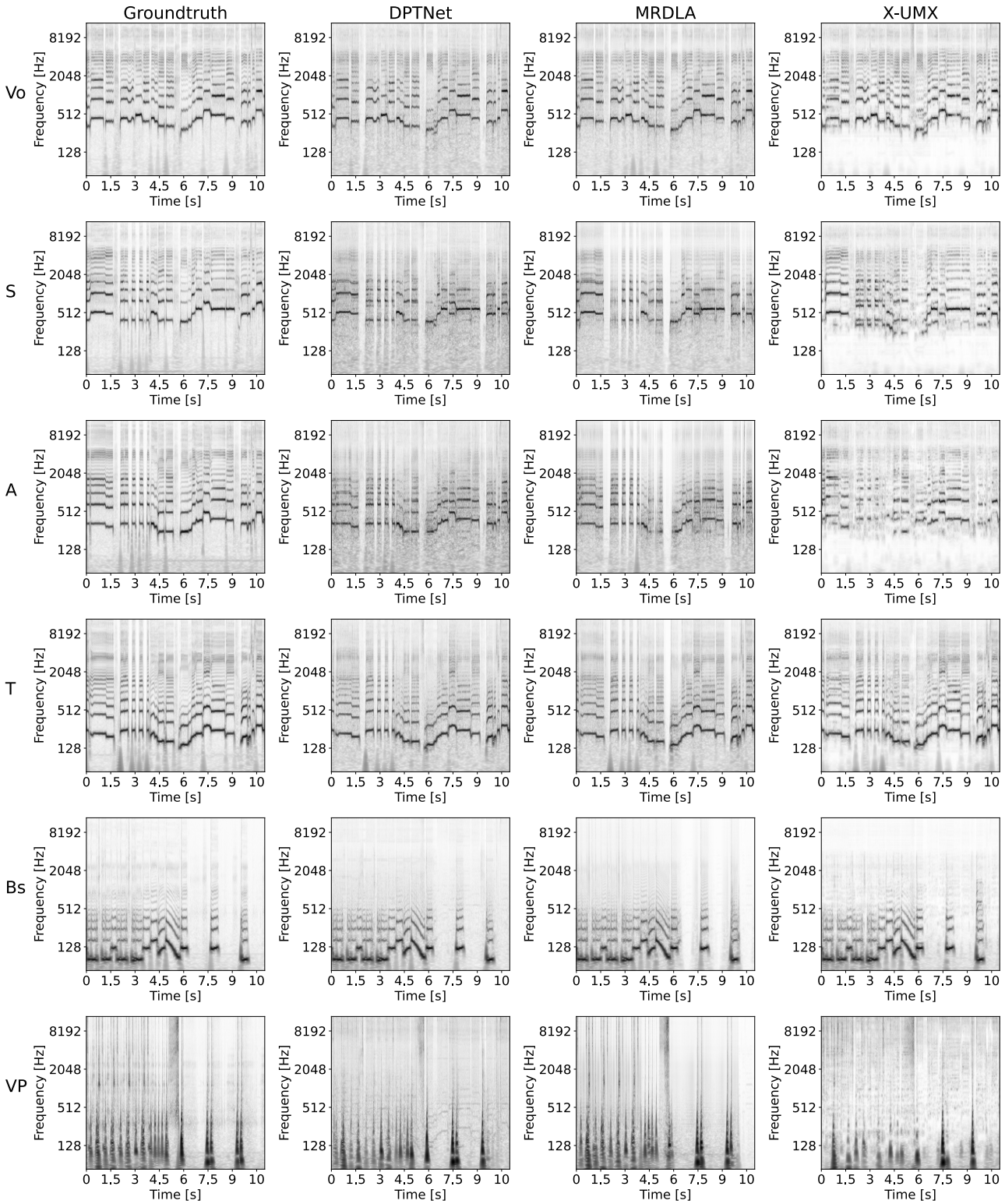


図 5: Spectrogram examples of groundtruth signals and separated signals obtained with DPTNet, MRDLA, and X-UMX. Each row and column correspond to voice part and method, respectively.

では，5.2 節で用いた学習データを用いたモデルと本節の学習データを用いたモデルを区別するため，手法名の後ろにそれぞれ v1，v2 をつけて表記する．

表 5 に各手法での平均 ΔSI-SDR を示す．比較のため, 5.2 節で用いた学習データを用いたモデルでの平均 ΔSI-SDR も記載した．MRDLA ではデータを増やすことで全ての声部に関して ΔSI-SDR が向上しており，データの追加が有効に機能した．一方，DPTNet では 1 dB 以上 ΔSI-SDR が低下する声部もあり，学習時のハイパーパラメータの変更が必要である可能性がある．また，X-UMX では A に関

して 1 dB 程度 ΔSI-SDR が低下したが，他の声部では同程度であった．X-UMX でも多少の学習時のハイパーパラメータの変更により性能が向上する可能性はあるものの，時間波形領域のモデルに比べスペクトログラム領域のモデルは少ないデータでも学習できることが多いため，データの追加が大幅な性能向上に繋がるかは調査が必要である．

聴取した際の印象も [13] で報告された傾向と同一であった．図 5 に正解音源信号と推定音源信号の声部毎のスペクトログラムの例を示す．第一著者が聴取した限り，MRDLA と DPTNet は分離音に比較的アーティファクトが含まれており，特に DPTNet は顕著であった．図 5 でも DPTNet や MRDLA のスペクトログラムは，他のスペクトログラムに比べエネルギーが全周波数帯域に満遍なくしている．一方，X-UMX はそのようなアーティファクトはないものの，他の声部の音が混入しやすい傾向があり，Bs では高周波成分がほとんど欠落していた．図 5 の Bs のスペクトログラムを見ると，X-UMX の分離結果は 4 kHz 以上の周波数帯域のエネルギーが極端に低いことが分かる．また，Vo，S,A では逆に低い周波数帯域のエネルギーが低くなっている傾向も見られた．これらの傾向は時間波形領域とスペクトログラム領域のモデルの差異として楽音分離でも報告されており [26]，楽音分離での知見を導入することで軽減できる可能性がある．

6. 結論

本稿では，日本語アカペラ歌唱コーパスである jaCappella コーパスの拡張について報告した．これまで構築してきた 35 曲に加え，新たに 15 曲の著作権処理済み重唱曲を作成しコーパスを拡張した．これらの曲は 3 つのジャンルの特徴を備えたサブセットに分かれており，既に作成していた 7 つのサブセットとは異なるジャンルの特徴をもつ．商用楽曲群との比較によって，歌詞や音高の頻度分布の観点で jaCappella コーパス v2 は実データをよく捉えていることを示した．また，重唱分離実験により，追加データを使用することで音源分離モデルによっては分離性能が向上することを示した．

今後は当該コーパスのアノテーションを増強し，重唱に関連する他のタスクにも利用できるように整備を進める．例えば，F_0 のアノテーションを整備することで多重 F_0 推定に利用できる．収録した音源は静音下での収録のため，既存の F_0 抽出器で高精度に F_0 が抽出できることを確認しているものの，一部にオクターブ誤りが含まれる場合があり，目視での確認を行いつつ整備を進める予定である．

謝辞 本研究は，国立情報学研究所（NII）CRIS と LINE 株式会社とが推進する NII CRIS 共同研究および JSPS 科研費 21H04900，21K12202，23K18474 の助成を受けた．

参考文献

[1] Sundberg, J.: 歌声の科学，東京電気大学出版局，東京 (2007).
[2] Purwins, H., Li, B., Virtanen, T., Schlüter, J., Chang, S.-Y. and Sainath, T.: Deep Learning for Audio Signal Processing, *IEEE J. Selected Topics Signal Process.*, Vol. 13, No. 2, pp. 206–219 (2019).
[3] Briot, J.-P., Hadjeres, G. and Pachet, F.-D.: *Deep Learning Techniques for Music Generation*, Springer (2020).
[4] Cuesta, H.: Data-driven Pitch Content Description of Choral Singing Recordings, PhD Thesis, Universitat Pompeu Fabra (2022).
[5] Cuesta, H., McFee, B. and Gómez, E.: Multiple F0 Estimation in Vocal Ensembles Using Convolutional Neural Networks, *Proc. Int. Soc. Music Inf. Retrieval Conf.*, pp. 302–309 (2020).
[6] Gover, M. and Depalle, P.: Score-informed Source Separation of Choral Music, *Proc. Int. Soc. Music Inf. Retrieval Conf.* (2020).
[7] Petermann, D., Chandna, P., Cuesta, H., Bonada, J. and Gómez, E.: Deep Learning Based Source Separation Applied to Choir Ensembles, *Proc. Int. Soc. Music Inf. Retrieval Conf.*, pp. 733–739 (2020).
[8] Sarkar, S., Benetos, E. and Sandler, M.: Vocal Harmony Separation Using Time-Domain Neural Networks, *Proc. INTERSPEECH*, pp. 3515–3519 (2021).
[9] McLeod, A., Schramm, R., Steedman, M. and Benetos, E.: Automatic Transcription of Polyphonic Vocal Music, *Appl. Sci.*, Vol. 7, No. 12, 1285 (2017).
[10] Chandna, P., Cuesta, H., Petermann, D. and Gómez, E.: A Deep-Learning Based Framework for Source Separation, Analysis, and Synthesis of Choral Ensembles, *Frontiers Signal Process.*, Vol. 2 (2022).
[11] Tamaru, H., Saito, Y., Takamichi, S., Koriyama, T. and Saruwatari, H.: Generative Moment Matching Network-Based Neural Double-Tracking for Synthesized and Natural Singing Voices, *IEICE Trans. Inf. Systems*, Vol. E103.D, No. 3, pp. 639–647 (2020).
[12] Rosenzweig, S., Cuesta, H., Weiß, C., Scherbaum, F., Gómez, E. and Müller, M.: Dagstuhl ChoirSet: A Multitrack Dataset for MIR Research on Choral Singing, *Trans. Int. Soc. Music Inf. Retrieval*, Vol. 3, No. 1, pp. 98–110 (2020).
[13] Nakamura, T., Takamichi, S., Tanji, N., Fukayama, S. and Saruwatari, H.: jaCappella Corpus: A Japanese A Cappella Vocal Ensemble Corpus, *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* (2023).
[14] Cuesta, H., Gómez, E., Martorell, A. and Loáiciga, F.: Analysis of Intonation in Unison Choir Singing, *Proc. Int. Conf. Music Perception Cognition* (2018).
[15] Good, M.: MusicXML for Notation and Analysis, *The Virtual Score: Representation, Retrieval, Restoration* (Walter, B. H. and Eleanor, S.-F., eds.), MIT Press, pp. 113–124 (2001).
[16] Hono, Y., Hashimoto, K., Oura, K., Nankaku, Y. and Tokuda, K.: Sinsy: A Deep Neural Network-Based Singing Voice Synthesis System, *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, Vol. 29, pp. 2803–2815 (2021).
[17] Müller, M.: *Fundamentals of Music Processing: Audio, Analysis, Algorithms, Applications*, Springer, first edition (2015).
[18] 野ばら社編集部, 久保昭二 編：童謡，野ばら社，改版 edition (2010).

[19] 野ばら社編集部 編: 唱歌：明治・大正・昭和, 野ばら社 (2009).
[20] Le Roux, J., Wisdom, S., Erdogan, H. and Hershey, J. R.: SDR – half-baked or well-done?, *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, pp. 626–630 (2019).
[21] Chen, J., Mao, Q. and Liu, D.: Dual-Path Transformer Network: Direct Context-Aware Modeling for End-to-End Monaural Speech Separation, *Proc. INTERSPEECH*, pp. 2642–2646 (2020).
[22] Nakamura, T., Kozuka, S. and Saruwatari, H.: Time-Domain Audio Source Separation With Neural Networks Based on Multiresolution Analysis, *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, Vol. 29, pp. 1687–1701 (2021).
[23] Kong, Z., Ping, W., Dantrey, A. and Catanzaro, B.: Speech Denoising in the Waveform Domain with Self-Attention, *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, pp. 7867–7871 (2022).
[24] Mitsufuji, Y., Fabbro, G., Uhlich, S., Stöter, F., Défossez, A., Kim, M., Choi, W., Yu, C.-Y. and Cheuk, K.-W.: Music Demixing Challenge 2021, *Frontiers Signal Process.*, Vol. 1 (2022).
[25] Défossez, A.: Hybrid Spectrogram and Waveform Source Separation, *Proc. Music Demixing Challenge Workshop* (2021).
[26] Schaffer, N., Cogan, B., Manilow, E., Morrison, M., Seetharaman, P. and Pardo, B.: Music Separation Enhancement with Generative Modeling (2022). arXiv: 2208.12387.