博士論文

複数人歌唱楽曲に対する音楽情報処理に関する研究

令和 3 年 12 月 1 日提出

指導教員　齋藤大輔准教授

東京大学大学院工学系研究科

電気系工学専攻

37-197069

須田仁志

本論文は東京大学大学院工学系研究科に博士号授与の要件として提出した博士論文である.

内容梗概

音楽は，有史以前から続く人類の重要な文化の 1 つであり，娯楽や宗教など様々な目的で演奏および鑑賞されている．情報技術の発展によって音楽を情報として扱うことができるようになり，様々な音楽情報処理技術が提案，研究されるようになった．音楽鑑賞においては，蓄音機や音楽のデジタル化によって，かつてその場で演奏される中で耳を傾け鑑賞されていた音楽は，好きな時間や場所で様々な目的で聴くことができるようになった．また，音楽サブスクリプションサービスの発展によって，膨大な音楽ライブラリをいつでもどこでも楽しむことができるようになった．音楽から様々な情報を抽出する技術は音楽理解技術とよばれ，音楽の検索や鑑賞に役立てられている．たとえば，コード進行や楽曲のジャンル，ムードなどといった音楽的情報の抽出技術は，膨大な楽曲に自動でメタデータを与え，曲名やアーティスト名だけでない様々なクエリによる音楽検索を可能にする．自動採譜技術は，特別な訓練なしに音楽を採譜することを可能にし，採譜された楽譜をもとに演奏を楽しむことを可能にする．音楽理解技術は音楽の可視化にも応用されており，楽曲を視覚的に楽しむことのできるインタフェースや，楽曲間の関係を可視化するインタフェースなどが提案されている．音楽を合成したり創作したりする技術も音楽情報処理に含まれる．たとえば，自動作編曲や自動作詞技術によって，音楽的な素養や訓練なしに好みの音楽を制作できる．また，歌声合成技術によって好みの歌手に既存の楽曲や制作した楽曲を歌わせることができ，こうした技術は音楽制作をより自由に豊かにしている．

複数人が歌唱する楽曲に対する音楽情報処理としては，各歌唱者やパートに分離する音源分離技術や，自然な同時歌唱を合成するための歌声合成技術などが研究されている．とくに複数人が楽曲を歌唱する際には，歌唱者が入れ替わりながら歌唱するパート割り構造を持つことも多く，それを対象とした研究もなされている．本論文ではとくに，パート割りのある楽曲から「誰がいつ歌唱しているか」を事前知識なく推定する歌唱者ダイアライゼーションとよばれる技術について検討する．これまでの歌唱者ダイアライゼーション研究では，会話音声を対象とした話者ダイアライゼーション技術を利用している．しかし，歌声は話声と比べて音高の幅が広く，音素継続長が長く，また複数人が同時に発声している区間が長い．したがって，話者ダイアライゼーション手法をそのまま用いては，高い精度のダイアライゼーションを実現できないことが懸念される．そこで本研究では，話者ダイアライゼーション手法を拡張し，高い精度で歌唱者ダイアライゼーションを行うことのできる手法を提案する．提案法には，各時刻で同時に何人が歌唱しているかを推定する同時歌唱者数推定技術や，ArcFace とよばれる埋め込み表現抽出法を用いた歌唱者表現を導入する．とくに同時歌唱者数推定では，Cosacorr スコアとよばれる新たに提案した音響特徴量を用いることで，高い精度での推定を可能にする．評価にはコンパクトディスク（compact disc; CD）に収録された日本のアイドルソングの音響信号をそのまま用い，現実的な条件で評価した．実験により，既存の話者ダイアライゼーション手法をもとにしたベースライン手法と比較して，提案する歌唱者ダイアライゼーション手法の有効性が示された．また，Cosacorr スコアや ArcFace の有効性についても確認された．

本論文では，歌声の声質を別の歌唱者の声質に変換する声質変換法についても検討する．声質変換は，入力話者の発話から抽出された音響特徴量を変換モデルによって変換し，変換された音響特徴量を用いて合成することで達成される．古典的な声質変換手法においては，入力話者と出力話者が同じ内容を発話したパラレルデータを必要とする．しかし，パラレルデータが利用できない条件では変換モデルを学習することができない．そこで，パラレルデータを必要とせず，入出力話者が異なる内容を発話した音声を用いて学習可能なノンパラレル声質変換法が広く検討されている．ノンパラレル声質変換法の多くは，音声から話者情報と言語情報を分離し，分離された話者情報のみを置換することで実現される．これまでのノンパラレル声質変換法は，膨大な話者の発話から構築した背景知識や，入出力話者による多数の発話を必要としており，システム全体を構築するための学習コーパスの用意が難しい．本研究では，非負値行列因子分解（non-negative matrix factorization; NMF）とよばれる行列分解法を利用して，背景知識なしに音声の話者情報と言語情報を分離することで，少量の入出力話者の発話のみを必要とする変換モデルの学習法を提案する．NMF は，スペクトログラムをはじめとした非負の物理量を，テンプレートである基底と，そのテンプレートの利用状態を表す生起状態に分解する手法である．本論文では，NMF の生起状態が，基底と音響特徴量間の連続的なアラインメントであることに着目する．このアラインメントは，INCA アルゴリズムとよばれるノンパラレルデータ間でアラインメントを得る手法と同様の手順で得られ，ノンパラレルな条件下でも変換モデルの学習に必要な生起状態を得られる．INCA アルゴリズムと異なり，得られるアラインメントが離散的ではなく連続的であるため，高い自然性を持つ変換音声が合成できることが期待される．話声を用いた評価実験により，既存のノンパラレル声質変換法である INCA アルゴリズムや CycleGAN-VC と比較して，提案法はより自然な変換音声を合成できることが確認された．本論文では，入力話者に関して事前に学習する必要のない one-shot 変換についても検討し，提案法を用いて one-shot 変換が実現できることを確認した．

複数人が歌唱する楽曲を楽しむことができるインタフェースの一例として，本論文では VocalRemixer とよばれる Web インタフェースを提案する．VocalRemixer は，楽曲からパート割りを推定する歌唱者ダイアライゼーション技術と，歌声を変換するノンパラレル声質変換技術の，本論文で議論する両者の技術を利用する．このインタフェースでは，パート割りのある楽曲に対してそのパート割りを自由に編集することができ，特定の歌唱者にソロで歌唱させたり，まったく異なるパート割りで歌唱させたりすることができる．被験者にインタフェースを利用させて行った評価実験では，インタフェースの新規性や魅力について高い評価を得た．また，被験者のコメントから，インタフェースをさらに拡張することで，より魅力的なインタフェースを構築できる可能性が示された．

本論文では，本研究における究極の目的である複数人歌唱楽曲に対する音楽情報処理において，とくに音楽理解技術を起点とした技術に着目して述べる．音楽情報処理には様々な方向性の研究が含まれており，それぞれの技術が相互に作用している．これは，複数人が歌唱する楽曲に対する音楽情報処理においても同様である．本論文で述べる技術は複数人歌唱楽曲に対する音楽情報処理のほんの一端であり，他の様々な音楽情報処理技術や未検討の音楽情報処理技術と相交わることで，音楽情報処理がより発展するものである．

Abstract in English

Music has been one of the most important parts of human culture and is performed and listened to for various purposes such as entertainment and religion. The development of information technology has enabled people to treat music as information, and various music information processing technologies have been proposed and studied. Before music information processing was introduced, people listened to music only at the spot where the music was performed. In the field of music information processing for music listening, the digitalization of phonographs and music has enabled people to listen to music for various purposes at any time and place. In addition, music subscription services have enabled people to enjoy enormous music libraries anytime and anywhere. The technologies to extract various information from music are called music understanding technologies and are useful for music search and listening. For example, technologies for extracting musical information such as chord progression, song genres, and moods automatically provide metadata to a huge amount of music and enable people to search music by various queries other than just titles and artist names. Automatic music transcription techniques enable people to transcribe music without training and to enjoy playing music using the transcribed scores. Music understanding technologies have also been applied to music visualization, and several interfaces that visualize songs or the relationships between songs have been proposed. Music information processing also includes technologies for synthesizing and creating music. For example, automatic music composition and lyrics generation techniques allow users to create their preferred music without any musical background and training. In addition, singing voice synthesis techniques enable users to make their favorite virtual singers sing songs that they like or created, thus the techniques facilitate and enrich the music production.

As musical information processing for songs sung by multiple singers, sound source separation techniques to separate each singer or part and singing voice synthesis techniques to synthesize natural simultaneous singing have been studied. In particular, when multiple singers sing within a song, the singers often sing alternatively, and this paper calls such a music structure *song division*. This paper focuses on a technique called *singer diarization*, which recognizes *who sings when* from songs without prior knowledge. The traditional singer diarization techniques have adopted methods for speaker diarization, which is intended to recognize conversational speech. However, in singing voices, the range of pitch is wider, the duration of phonemes is longer, and the segments where multiple singers are singing at the same time are longer compared to conversational speech. Therefore, the traditional speaker diarization techniques cannot achieve high accuracy in singer diarization. This paper introduces a new technique to achieve highly accurate singer diarization by extending an existing speaker diarization framework. The proposed framework utilizes a voice activity and overlap detection component that estimates how many singers are simultaneously singing at each time and singer representations based on ArcFace, an embedding extractor. Particularly in the voice activity and overlap detection, this paper introduces a new acoustic feature named Cosacorr scores to achieve high accuracy in the detection. For the evaluation, this paper used the acoustic signals of Japanese idol songs recorded on compact discs (CDs) and evaluated the proposed framework under realistic conditions. The experimental results showed the effectiveness of the proposed framework compared to a baseline method based on an existing speaker diarization framework. The results also showed the effectiveness of Cosacorr scores and ArcFace in the proposed framework.

This paper also discusses voice conversion to convert the voice quality of singing voice. Voice conversion is achieved by extracting acoustic features from utterances uttered by a source speaker, converting the acoustic feature using a conversion model, and synthesizing utterances from the converted features. Traditional techniques require parallel corpora,

in which the source and target speakers uttered the same linguistic contents. However, the conversion models cannot be constructed if the parallel corpora are unavailable. Therefore, some studies have proposed nonparallel conversion methods to utilize nonparallel corpora, in which the source and target speakers uttered different contents. Most of the nonparallel conversion methods are achieved by disentangling speaker and linguistic information from utterances and replacing only the speaker information. Traditional nonparallel techniques require either huge background knowledge trained using a large number of utterances spoken by many speakers or a huge number of utterances of source and target speakers, and thus it is difficult to prepare the training corpora to construct conversion models. This paper proposes a new training method of nonparallel conversion models by separating speaker and linguistic information without any background knowledge by utilizing non-negative matrix factorization (NMF). NMF is a method to decompose non-negative physical quantities such as spectrograms into bases, or templates, and activation, which determines the usage of the templates. This paper focuses on the fact that the activation of NMF is a continuous alignment between the bases and the decomposed acoustic features. The alignment can be obtained similarly to the INCA algorithm, which is a technique to obtain alignment between nonparallel corpora, and thus the proposed method can acquire the activation required to train conversion models without parallel corpora. In contrast to the INCA algorithm, the alignment is continuous, not discrete, and the method is expected to be able to synthesize more natural utterances than the methods based on the INCA algorithm. The results of experiments that used speech utterances showed that the proposed method can generate more natural utterances than the INCA algorithm and CycleGAN-VC, which are the existing nonparallel conversion methods. This paper also examined the quality of one-shot conversion, which does not require information about the source speakers, and the experimental results showed that the proposed method can be utilized for one-shot conversion.

As an example of an interface that allows users to enjoy songs sung by multiple singers, this paper proposes a new web interface named VocalRemixer. VocalRemixer utilizes both singer diarization techniques and nonparallel voice conversion techniques. Users can edit the structures of the vocal parts of the songs with song divisions, and users can make a specific singer sing solo or make the singers sing with completely different divisions. The subjective experiments are conducted by making users use VocalRemixer, and the results showed the novelty and attractiveness of VocalRemixer. In addition, comments from the subjects indicated the possibility of constructing more attractive interfaces by further extending the proposed interface.

The ultimate goal of this study is to establish music information processing for songs sung by multiple singers. This paper particularly focuses on the techniques that begin with music understanding technologies for the songs. Music information processing includes various directions of studies, and the studies interact with each other. This is also true for music information processing for songs sung by multiple singers. The techniques described in this paper are only a part of the music information processing for the songs. Music information processing itself will be further developed by entwining the techniques described in this paper with other various music information processing technologies including unexplored technologies.

目次

図目次

表目次

第１章

序論

1.1　研究の背景

1.1.1　音楽とは

音楽は，有史以前から続く人類の文化の 1 つである．音楽は音にまつわる芸術全般を指すが，明確な定義は存在しない．音楽は音を介した芸術であって，そこに歌声や楽器音がある必要はない．たとえば，ジョン・ケージにより 1952 年に作曲された『4 分 33 秒』は，歌声や楽器音ではなく自然に発生する環境音に焦点を当てた音楽で，無音の音楽として知られている [1]．また，単なる意思疎通が音楽と捉えられることもある．たとえば，シジュウカラなどの鳥の鳴き声などは，その動物たちにとっては意思疎通の目的で発されているが，我々人にとっては歌と感じられる．したがって，音楽の定義は明確ではなく，聴き手が音楽と感じれば音楽であると定義できる [2]．

現代，音楽は娯楽の目的や讃美歌をはじめとした宗教的な目的など様々な目的で演奏されているが，音楽の起源，すなわちなぜ音楽が生まれたかについては定説がなく議論するに値しない [3]．発見されている人類の作った楽器のうちもっとも古い楽器として，動物の骨で作られたおよそ 43000 年前の笛が 2 つ知られている．Divje Babe flute は，1995 年にスロベニアで発見されたもので，ネアンデルタール人が使っていたと推測されている [4, 5]．ドイツの Geißenklösterle 洞窟で発見された笛は，オーリニャック文化のものとされる [6]．これらの楽器の出土から，少なくとも旧石器時代から人類は音楽を演奏していたことが明らかになっている．どのように音楽が発生，伝来したかは定かではないが，紀元前数千年ごろには様々な地域で音楽が演奏されていたことが知られている [3]．日本においても，人類が移り住んだのに合わせて朝鮮半島やインドシナ方面から音楽が伝来し，祭祀などに用いられていたとされる．

本来音楽は，演奏されているものをその場で鑑賞するもので，音楽に集中して耳を傾けるのが鑑賞のあり方であった．しかし現在では情報技術の発展によって，様々な場所，目的で，多様な音楽を鑑賞することが可能になっている．大学生を対象とした調査では，移動中に携帯型音楽プレーヤを用いて音楽を聴く学生が多く，音楽のみに集中して聴く本来のあり方とは異なる [7]．音楽を聴く目的としても，表 1.1 に示すように，「高揚・元気になるため」「リラックスするため」といったメンタル面での目的のほかに，「暇つぶしのため」「音がないと寂しい」「集中するため」などの音楽を環境音として用いる学生もみられる．さらなる情報技術の発展によって，音楽の楽しみ方はさらに多様化するものと思われる．

1.1.2　音楽情報処理の現在

音楽情報処理とは

音楽情報処理とは，音楽を対象とする情報処理をいい，音楽の分析，生成，演奏，発見技術や，音楽を起点とした脳科学など，広範な情報処理技術を含む [8]．膨大な音楽が氾濫し，音楽を聴く目的が多様化している中で，目的に応じた楽曲を発見する技術や，さらに目的を多様化できる楽曲分析，鑑賞技術が求められる．音楽情報処理は，こうした課題を解決する技術でもある．

表 1.1　音楽を聴く目的 [7]．大学生 44 人に自由記述で回答させた結果を集計したもの．

記述（集約後）	出現率（%）
高揚・元気になるため	29.5
リラックスするため	22.7
暇つぶし	18.2
気分転換	15.9
演奏のため	13.6
音がないと寂しい	6.8
すっきりする	6.8
集中するため	6.8
音楽が好き	6.8
何となく	4.5
眠るため	4.5
その他	13.6

記録する技術，鑑賞する技術

最初の音楽情報処理技術は，エジソンによる蓄音機である．発明当時の蓄音機は手回し式で，音声の録音には蝋管が用いられていた．蓄音機の登場によって，これまでその場で演奏されなければ鑑賞できなかった音楽を，蓄音機さえあればいつでも鑑賞できるようになった．その後，ビニルなどで作られたレコードが登場し，蓄音機は一般家庭に浸透した．

音楽鑑賞の歴史において重要な技術として，コンパクトディスク（compact disc; CD）が挙げられる．それまで音楽はレコードやカセットテープなどのアナログ媒体に収録されていたが，CD の登場によって広くデジタル化された．コンピュータが一般家庭に広まると，CD に収録された音楽はコンピュータに取り込まれてポータブルオーディオプレーヤなどの小型デバイスを通して持ち運ぶことができるようになり，何万時間もの音楽を小さなデバイスに携帯できるようになった．

また現在では Spotify[*1] や Apple Music[*2] といった音楽サブスクリプションサービスも広く利用されている．Apple Music の会員数は 2019 年 6 月現在で 6000 万人，Spotify のアクティブユーザ数は 2020 年第 4 四半期に 3 億 4500 万人を超え，いまや世界中のユーザが音楽サブスクリプションサービスを利用している [9]．これらのサブスクリプションサービスによって，CD などを買わずとも，膨大な音楽ライブラリから好みの楽曲を聴くことができるようになった．サブスクリプションサービスの中には，ユーザの嗜好を分析することでユーザに好みの音楽を推薦する機能を備えているサービスもあり，こうしたサービスは新たな音楽の発見にも役立てられている．

[*1] https://www.spotify.com/
[*2] https://www.apple.com/apple-music/

認識する技術，理解する技術

音楽情報処理として，音楽から様々な情報を抽出しそれを活用する技術が研究されている．とくに音楽から音楽的情報を抽出する技術は音楽理解技術とよばれる．たとえば，楽曲のコード進行などを推定するコード認識 [10]，楽曲の拍などを推定するビート認識 [11]，楽曲のジャンルを認識するジャンル認識 [12]，楽曲の持つ雰囲気や聴取によりもたらす感情などを推定するムード認識 [12]，楽曲の演奏するアーティストを推定するアーティスト識別 [12] などが音楽理解技術に含まれる．音源分離による楽器音の分離技術も音楽理解技術の 1 つである [13, 14]．音源分離によって，各楽器の演奏に対して音楽理解技術を適用したり，楽曲のリミックスなどに活用したりすることができる．とくにポップスの歌声分離は，類似歌手検索や歌詞認識などにも応用できる [15]．音楽の自動採譜も音楽理解技術の 1 つである [16, 17]．音楽を聴いて採譜を行うには訓練が必要であり，自動採譜技術によってこうした訓練なく採譜を行い演奏を楽しむことができる．歌詞認識も音楽理解技術の 1 つとして挙げられる [18]．歌詞を手作業で与えることなくメタデータとして利用できるほか，楽曲のテーマやムードの推定にも応用できる．

音楽の可視化技術も音楽情報処理の 1 つと見なすことができる．音響特徴量を可視化するビジュアライザや，楽曲の音楽的構造を可視化するインタフェース [19]，音楽と音楽との関係を可視化するインタフェース [20] などが提案されている．音楽鑑賞サービス Songle は，YouTube やニコニコ動画上にアップロードされた音楽に対して自動的に楽曲構成，ビート構造，コード，旋律を分析し可視化することで，音楽理解を深めることが可能なインタフェースを提供している [19]．音楽関連コンテンツ可視化サービス Songrium では，ニコニコ動画などの動画投稿サービス上に投稿された様々なクリエイターが制作した楽曲を「星図」として可視化できる [20]．音楽鑑賞インタフェース TextAlive は歌詞を活用した音楽鑑賞インタフェースで，音楽音響信号と歌詞を自動で対応付け，現在歌唱されている歌詞を kinetic typography とよばれる効果的に歌詞を示す手法で表示できる [21]．音楽の可視化だけでなく，SOUND HUG[*3] などの音楽を振動などに変換し触覚で音楽を楽しむことのできるデバイスも提案されている．

音楽検索技術も音楽情報処理の 1 つと見なすことができる．音楽の検索には，曲名やアーティスト名などのメタデータだけでなく，様々な音楽要素をクエリにした音楽検索が考えられる．たとえば，ジャンル認識やムード認識によって得られた情報や，楽曲を構成する楽器などもクエリとして利用可能である．類似歌手検索システム VocalFinder は，歌声の声質をクエリとして類似の声質の歌手が歌唱する楽曲を検索することができ，同様の技術が音楽鑑賞サービス Songle にも導入されている [19, 22]．また，歌声やハミングをクエリにして旋律から楽曲を検索する技術は query-by-singing/humming（QBSH）とよばれ，広く研究されている [23]．音楽検索技術を利用した音楽推薦機能も音楽情報処理の 1 つである [24]．Apple Music をはじめとした音楽サブスクリプションサービスでは，多数のユーザの嗜好の分析による楽曲推薦機能が導入されている．

[*3] https://pixiedusttech.com/soundhug/

合成する技術，創作する技術

音楽を自動で創作する技術として，自動作編曲 [25]，自動作詞 [26] などが知られている．とくに日本語の歌詞に曲を与える自動作曲システムとしては Orpheus [27] が公開されており，インターネット上のコミュニティなどで活発に利用されている．全自動で作曲を行うのではなく，作曲や作詞の支援を行い，音楽制作の補助となるようなソフトウェアも提案されている [28, 29].

歌声合成も音楽情報処理において重要な技術の 1 つである [30]．歌声合成は，話声の音声合成の延長として構築されるが，ビブラートや裏声などの歌声特有の表現法が存在するため，歌声に特化した合成技術が求められる．VOCALOID[*4] をはじめとした波形接続型の手法や，Sinsy [31] をはじめとした隠れマルコフモデル（hidden Markov model; HMM）を用いた手法，CeVIO AI[*5] や NEUTRINO[*6] をはじめとしたニューラルネットワークを用いた手法などが，歌声合成アプリケーションに利用されている．

音楽を演奏するロボットも，音楽情報処理の 1 つと見なすことができる [32]．1800 年ごろに開発された自動ピアノをはじめとして，種々の楽器を自動演奏する手法が研究されている．鍵盤楽器や打楽器だけでなく，弦楽器や木管楽器などについてもロボットによる演奏が実現されつつある．

1.1.3　複数人が歌唱する楽曲に対する音楽情報処理

複数人が歌唱する楽曲は，1 つの同一の声部を全員が歌唱する斉唱（ユニゾン），複数の声部をそれぞれ 1 人の歌唱者が歌唱する重唱，複数の声部をそれぞれ複数の歌唱者が歌唱する合唱に分類される．キリスト教の讃美歌やベートーヴェンによる交響曲第 9 番をはじめとして，様々な音楽に複数人歌唱がみられる．とくに現在の日本では AKB48 などのアイドルが歌唱する楽曲やクラス合唱など様々なシーンで斉唱や合唱がみられる．こうした複数人が歌唱する楽曲を対象とした音楽情報処理技術としては，音源分離や歌声合成などが研究されている．

音源分離を目的とした技術では，複数の声部からなる歌声を，深層学習を用いて分離する手法が提案されている [33]．また，ユニゾンを分離する手法についても提案されている [34]．複数人が歌唱する歌声の分離は，複数人が調和して歌唱することが多く，音韻により音色が時間的に変化するため，楽器音の分離と比較して困難である．とくにユニゾンの歌声については，複数の歌唱者が非常に近い音高で歌唱するためきわめて分離が難しく，既存のスペクトログラムを利用した信号処理的な手法では高精度な分離が望めない．ユニゾンの分離手法の研究においては，収録された自然な歌声，とくに 3 人以上の歌声については実験的評価がなされておらず，ユニゾンに対する分析技術のさらなる検討が必要である．

複数の歌唱者による楽曲特有の分析技術として，特定の歌唱者がその楽曲に含まれているかを推定する target-singer detection や，各時刻でその推定を行い特定の歌唱者がその時刻で歌唱しているかを推定する target-singer tracking といった技術も検討されている [35, 36]．また，歌唱者の事前知識なく，楽曲を歌唱している歌唱者を認

[*4] https://www.vocaloid.com/
[*5] https://cevio.jp/
[*6] https://n3utrino.work/

識し各歌唱者に対して各時刻で歌唱しているかを推定，すなわち「誰がいつ歌唱しているか」を推定する歌唱者ダイアライゼーションとよばれる技術についても検討されている [37]．歌唱者ダイアライゼーションは，歌唱者が何人いるか，どのような声質の歌唱者がいるか，それぞれの歌唱者がいつ歌唱しているかを推定できる技術であり，複数人が歌唱する楽曲を分析する技術として基礎的であるにも関わらず，詳細な分析や評価がなされていない．本論文では複数人歌唱楽曲の重要な基礎的分析技術として，この歌唱者ダイアライゼーションについて扱う．

複数人の同時歌唱，とくに複数人が同一の音高で歌唱する場合の歌声合成についても様々な検討がなされている [38]．単純に既存の歌声合成を複数人歌唱に拡張する場合，音高が機械的に一致するため，たとえ複数の歌声を重ね合わせても 1 人が歌唱しているように聞こえてしまう．複数人が同一の音高で歌唱する場合には，楽譜上の音高が同一であっても，声の高さのわずかな違いや音の開始時刻（オンセットタイム）の違い，ビブラートの周波数，位相差などによって完全に同一の発声にならず，このわずかな違いに複数人歌唱らしさが現れる．こうした違いをどのように制御することで自然な斉唱音声が合成できるかについては，主観評価実験により検討されている [39]．また，複数人が同時に同一音高で歌唱したとき次第に声の高さが引き込まれる引き込み現象が確認されており [40]，引き込み現象を再現した合成音声の評価もなされている [41]．さらに，複数人が合唱などで歌唱する際の声質の調和についても研究がなされている．複数人が同時に歌唱する際，歌唱者の声質の相性が聴感上の声の調和度に影響を与えることが指摘されており，調和度の高い歌唱者の組み合わせを選ぶ手法や，調和的な歌声を声質変換によって合成する手法が検討されている [42,43]．とくに複数人が合唱などで歌唱するとき，それぞれの歌唱者は単独で歌唱する際と比較してより調和するように声質を変化させることが知られている [44]．より自然な複数人歌唱を合成するためには，単独で歌唱した歌声を単に重ね合わせるだけでなく，楽曲全体としての声の響きやバランスを制御する必要がある．

1.2　研究の目的と本論文のあらまし

本研究の究極の目的は，複数人が歌唱する楽曲に対する音楽情報処理を確立することである．その一部として，本論文では，複数人が歌唱する楽曲に対する基礎的な音楽理解技術を構築し，複数人歌唱楽曲に対するさらに豊かな鑑賞の実現に繋げることを目的とする．

複数人が歌唱する楽曲には，1 曲の中で歌唱者が入れ替わりながら交互で歌ったり全員で歌ったりなどする，パート割り[*7]とよばれる構造を持つ楽曲が多数存在する．音楽理解技術の 1 つとして，本論文では複数人が歌唱しているパート割りのある楽曲から「誰がいつ歌唱しているか」を推定する歌唱者ダイアライゼーション技術に着目する．歌唱者ダイアライゼーションは，パート割りのある楽曲に対してその構造を推定する技術であるものの，ダイアライゼーションの性質上，どのような声質の歌唱者がおり，総じて何人の歌唱者がいるか，といった，複数人歌唱楽曲の分析において基礎的な情報を明らかにすることも可能である．歌唱者ダイアライゼーションによって得られた情報は，楽曲構造の分析や，自動楽曲演出など，様々な応用可能性を持つ．これまで歌唱者ダイアライゼー

[*7] 歌割り，歌い分け，パート分けなどともよばれる．

ション技術や類似技術である target-singer tracking などが提案および考察されているものの，現代のポップスを十分に分析できる技術は構築されていない．本論文では，このような歌唱者ダイアライゼーションの重要性と意義を踏まえ，基礎的な歌唱者ダイアライゼーション手法を確立する．市販の CD に収録されている楽曲を利用して評価することで，現実的な実験条件で提案法の有効性を検証する．

また，複数人が歌唱する楽曲を有効に活用した音楽鑑賞技術を実現するため，新たな音楽鑑賞インタフェースを提案する．音楽理解技術を活用したインタフェースは様々に提案されているが，とくに複数人が歌唱する楽曲を有効に活用できるインタフェースは提案されていない．複数人の歌唱する楽曲には，歌に時間方向と歌唱者方向の情報が含まれ，その情報を活用した音楽の可視化や加工を行うことのできるインタフェースに幅広い可能性がある．本論文ではその一例として，パート割りを編集できるインタフェース VocalRemixer を提案および実装し，被験者に利用させ意見を集めることで主観的な評価を行う．VocalRemixer は，歌声のパート割りを編集できるという 1 つの機能のみを持つインタフェースであるが，歌声を複数トラックに分解するという性質上，音楽可視化や楽器音の加工といった様々な音楽鑑賞インタフェースと融合できる可能性を持つインタフェースである．

VocalRemixer の実現には，歌唱者ダイアライゼーション技術と声質変換技術が必要である．楽曲中，複数の歌唱者が同一のフレーズをソロで歌唱することは仮定できないため，入出力歌唱者が異なる内容を歌唱したノンパラレルデータを利用する必要があり，ノンパラレル声質変換法の導入が必要である．とくに VocalRemixer では，学習に利用できる音声が短時間のため，少量の歌声を有効に活用できる声質変換法が必要である．一方，既存のノンパラレル声質変換法では，多数の話者の発話から学習した大規模な背景知識や大量の入出力話者の音声が必要で，小規模なシステムを構築することが難しい．そこで，本論文では既存手法と比較して短時間の音声で学習可能な効率的なノンパラレル声質変換法を提案する．少量の発話を用いて実験を行い，主観評価によって提案法の有効性を評価する．また，VocalRemixer で必要な，入力話者について学習することなく声質変換を実現する one-shot 変換についても検討し評価を行う．

1.3　本論文の構成

第 1 章では，本研究の背景および目的を説明した．第 2 章では，本論文の理解に必要な音声の分析合成に関する基礎知識を述べる．第 3 章では，提案する歌唱者ダイアライゼーション手法について解説し，提案法の評価実験およびその結果について述べる．第 4 章では，VocalRemixer の実装を目的として提案するノンパラレル声質変換法について解説し，評価実験について述べ，その結果について深く考察する．第 5 章では，音楽鑑賞インタフェース VocalRemixer の内容とその実装について説明し，主観評価実験の内容およびその結果を述べる．加えて，VocalRemixer の拡張の構想についても提案する．第 6 章では，本論文を総括し，今後の展望を述べる．付録 A では，歌唱者ダイアライゼーション手法の評価実験のために構築したデータセットについて詳細に説明する．

第２章

基礎知識

本章では，発声の機構および音声のパラメータ表現について説明する．まず，人が発声する際の機構を，有声音，無声音に区別して説明する．次に，この機構にもとづいた音声のモデル化であるソースフィルタモデルについて説明する．さらに，音声のパラメータ表現である音響特徴量について説明する．音声の信号そのものを扱うことは難しいため，音声の認識や合成においては音響特徴量がよく用いられる．こうした音響特徴量は，音声の生成や知覚の機構にもとづいて設計されている．また，音響特徴量の一つである基本周波数の推定手法について説明する．

2.1　発声の機構

人の発声は有声音と無声音の 2 種類に区別され，それぞれの発声機構は異なる．音声のパラメータ表現を設計する上では，両者の発声機構の理解が必要である．これらの発声機構にもとづいた，ソースフィルタモデルとよばれる音声のモデル化手法が音声の分析や合成に用いられる．

2.1.1　有声音

有声音は，次の過程を経て聴取される．まず，声帯が周期的に振動することで縦波の音波が生じる．次に，その音波が喉頭，咽頭，鼻腔，口腔などからなる声道を通ることにより共振し，音色付けされる．この音色付けによって，音韻性や話者性といった情報が音声に保持される．共振周波数はフォルマント周波数とよばれ，とくに音韻性や話者性が強く現れる．さらに，口から放たれた音波が空気を通ることで，最終的に聴取者の耳に届き，知覚される．

2.1.2　無声音

無声音は有声音と異なり，様々な過程で生成される．無声音はおもに発声過程の違いにより摩擦音と破裂音に区別される．たとえば，摩擦音の /s/ は，舌端と歯茎の間の狭い隙間を空気が通ることによって発声される．破裂音の /p/ の場合には，上下の唇で作られた閉鎖が一気に開放されることによって発声される．無声音も有声音と同様に，最終的に口から放たれた音波が空気を通って聴取者の耳に届き，知覚される．

2.1.3　ソースフィルタモデル

ソースフィルタモデルは，声帯振動による音波を音源波，声道による音色付けをフィルタと仮定する，音声の生成過程のモデルである．ソースフィルタモデルでは，最終的な音声の信号 $y(t)$ を次のようにモデル化する．

$$y(t) = h(t) * x(t) + n(t) \tag{2.1}$$

ここで，$h(t)$ は声道によるフィルタ，$x(t)$ は声帯振動により作られる音源波であり，$*$ は畳み込み積を表す．$n(t)$ は無声音に対応する雑音源である．ソースフィルタモデルでは，$x(t)$ はインパルス列として近似されることが多い．

式 (2.1) の両辺にフーリエ変換を施すと，次式のように畳み込み積が積に変換される．

$$Y(f) = H(f)X(f) + N(f) \tag{2.2}$$

すなわち，音声のスペクトルは，声道特性 $H(f)$ と音源波スペクトル $X(f)$ の積および雑音源の特性 $N(f)$ の和によって表される．声道特性 $H(f)$ は，音源波のスペクトルと比較してなだらかな形状をもち，音声のスペクトル $Y(f)$ の包絡として現れるため，スペクトル包絡とよばれる．

2.2　音響特徴量

音響特徴量は，音響信号から抽出される，扱いやすい形に加工された特徴量である．音声合成や音声の認識においては，音響信号をそのまま扱うことは難しいため，音響特徴量を代わりに用いることが多い．音響特徴量は，フレームとよばれる短時間の信号から抽出され，時系列特徴量として扱われることが多い．一般的な音響特徴量抽出の手順を図 2.1 に示す．利用する目的によってどのような音響特徴量を用いるかは異なり，たとえば音声合成にはソースフィルタモデルを表現できる音響特徴量が用いられる．

2.2.1　短時間フーリエ変換によるスペクトログラム

音響信号は，時間によって変化する信号である．ある時刻付近の信号のみを取りだしてこれをフーリエ変換することで，各時刻における信号の性質を抽出できる．これを短時間フーリエ変換といい，時系列にわたったスペクトル系列，すなわち信号の時間周波数特性を表した特徴量をスペクトログラムとよぶ．信号 $x(t)$ のスペクトログラム $X_t(\omega)$ は次式で定義される．

$$X_t(\omega) = \frac{1}{\sqrt{2\pi}} \int e^{-j\omega\tau} x(\tau) w(\tau - t) d\tau \tag{2.3}$$

ここで，t は着目する時刻，ω は角周波数である．また，$w(t)$ は窓関数とよばれる時刻 0 付近に値を持つ関数である．短時間フーリエ変換では，窓関数と信号の時間領域での積に対してフーリエ変換を適用するため，周波数領域では信号の周波数特性に対して窓関数の周波数特性が畳み込まれる．したがって，窓関数の選択によって得られるスペクトログラムが異なる．窓関数には，ハミング窓やハン窓など様々な種類の関数が提案されており，目的に応じて適切な窓関数を選ぶ必要がある．窓関数には通常左右対称で山状の関数が選ばれるが，音響信号の符号化においては左右非対称の窓関数が用いられる場合もある．

音声の分析や合成においては，人の聴覚が位相に鈍感であることから，スペクトログラムの振幅やパワー，すなわち振幅スペクトログラムやパワースペクトログラムが利用されることが多い．位相情報を失った振幅スペクトログラムやパワースペクトログラムから音声を合成する場合には位相情報を付与する必要があるが，これには Griffin–Lim 法 [45] などの位相復元法や，WaveNet [46] などの深層学習による波形生成法などが用いられる．また音声の分析においては，短時間フーリエ変換を行う周期をフレーム周期やフレームシフト，窓関数の長さを窓長と

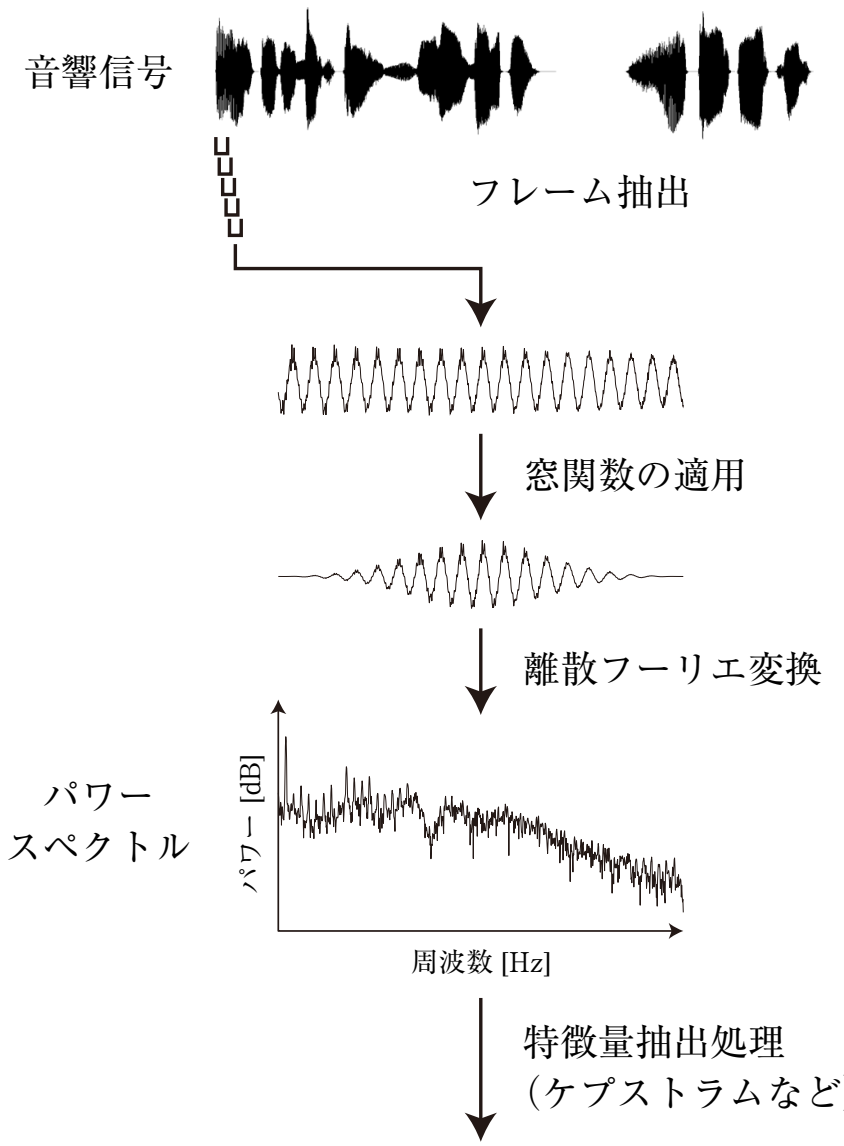


図 2.1　一般的な音響特徴量抽出の手順．短時間フーリエ変換によって得られたスペクトルに対して特徴量抽出処理を適用することで，音響特徴量の系列を得る．

よぶ．窓長が長いほど周波数分解能が高まるが，着目する時刻周辺の信号に影響されやすくなるため，適切な窓長を選択する必要がある．

2.2.2　ボコーダに用いる音響特徴量

ボコーディング（vocoding），すなわち音声の合成には，式 (2.1) によれば，3 つのパラメータが必要である．音源波 $x(t)$ はインパルス列であるため，その周波数のみによって表現でき，この周波数を基本周波数とよぶ．無声音の特徴 $N(f)$ は周波数領域で表現され，周期的な有声音と比較して非周期性指標とよばれる．非周期性指標は，詳細な周波数特性で表現せず，10 程度の帯域に分割し，それぞれの帯域での有声音と雑音との比率で表現することが多い．すなわち，基本周波数、スペクトル包絡，非周期性指標の 3 パラメータで音声を表現することができ，これらのパラメータを利用して音声を合成できる．

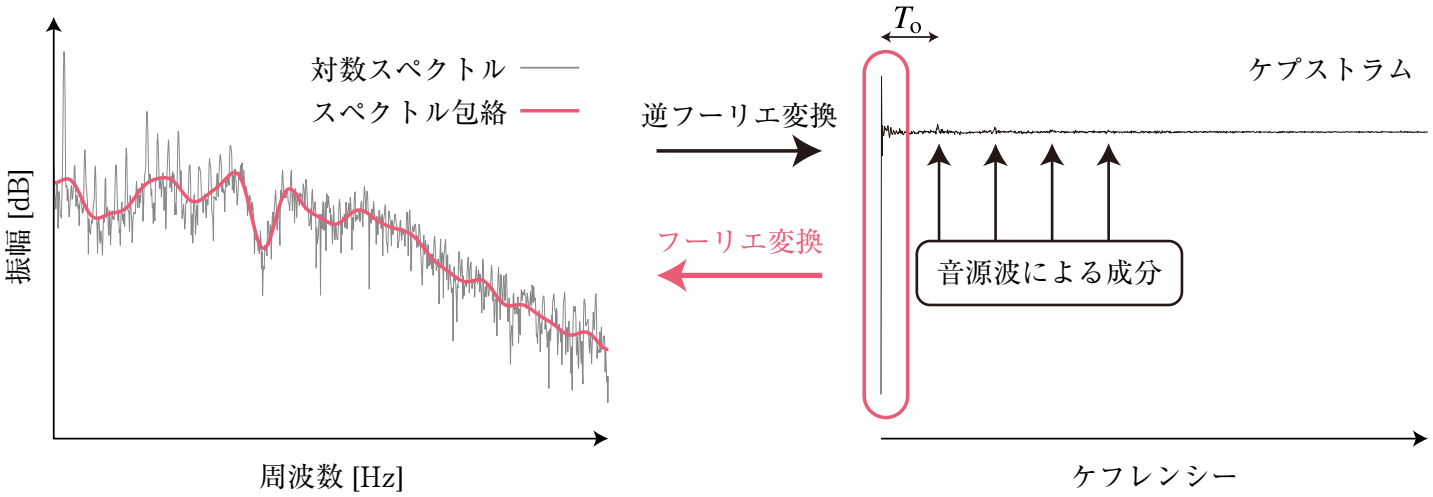


図 2.2　リフタリングによりスペクトル包絡を求めた例．高いケフレンシーに現れる音源波の成分をリフタリングによって除去することで，スペクトル包絡のみを選択できる．

2.2.3　ケプストラム

ケプストラムは対数スペクトルの逆フーリエ変換と定義され，時間にあたる軸はケフレンシーとよばれる．音声のすべてが有声音と仮定したとき，信号 $y(t)$ のケプストラム $c_y(t)$ は，次式のように音源波のケプストラムと声道特徴のケプストラムの和で表される．

$$c_y(t) = \mathcal{F}^{-1}(\log H(f)) + \mathcal{F}^{-1}(\log X(f)) \tag{2.4}$$

$$= c_h(t) + c_x(t) \tag{2.5}$$

ここで，$\mathcal{F}$ はフーリエ変換を表す．対数スペクトル上では，音源波は櫛状の短い周期のスペクトルを持つ一方，声道特徴はなだからなスペクトルを持つ．そのため，ケプストラム上では，低いケフレンシーに声道特徴が現れ，高いケフレンシーに音源波の特徴が現れる．したがって，図 2.2 のように，低いケフレンシーのみを抽出してフーリエ変換を適用することで，スペクトル包絡のみを抽出できる．すなわち，低次のケプストラムのみを抽出することで，声道特徴に対応する音響特徴量を得ることができる．このようなケプストラム領域でのフィルタリングをとくにリフタリングとよぶ．

2.2.4　メルスペクトルとメルケプストラム係数

人の聴覚において，周波数の分解能は低域では細かく高域では粗いことが知られている．人の聴感上，線形に感じられるように設計された周波数の単位にメルがあり，メル $f_{\mathrm{mel}}[\mathrm{mel}]$ と周波数 $f[\mathrm{Hz}]$ とは次式の関係にある．

$$f_{\mathrm{mel}} = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \tag{2.6}$$

聴覚上より自然に合成できるように，周波数軸をメル軸に変換した領域でのスペクトルであるメルスペクトルや，同様のケプストラムであるメルケプストラム係数も，おもに音声合成において音響特徴量として広く用いられてい

る．メルケプストラム係数を計算する際には，ケプストラムから再帰的に変換する手法 [47] や，対数スペクトルとの距離を最小化するように反復的に最適化する手法 [48] が用いられる．

2.2.5　メル周波数ケプストラム係数（MFCC）

音声認識においては，メル周波数ケプストラム係数（mel-frequency cepstral coefficients; MFCC）とよばれる音響特徴量が広く用いられる．人の聴覚機構においては，内耳中の蝸牛にある基底膜が音の周波数に応じて異なる振動をするフィルタバンクのような働きを持つことが知られている．MFCC の計算には，このような基底膜の振動を模した，メル軸上で等間隔の三角状のフィルタバンクが用いられる．MFCC は，振幅スペクトルに対してフィルタバンクを適用し，その系列の対数を計算し，それに対して離散コサイン変換を行うことで計算される．MFCC の計算の際には，人の声帯振動の周波数特性 $-12\,\mathrm{dB/oct}$ と放射特性 $6\,\mathrm{dB/oct}$ を合わせた $-6\,\mathrm{dB/oct}$ を補償するようなハイパスフィルタを，1 次の有限インパルス応答（finite impulse response; FIR）フィルタを用いて適用することが多い．この MFCC の計算手順を図 2.3 に示す．

2.2.6　Δ 特徴量と ΔΔ 特徴量

メルケプストラム係数や MFCC は，着目した時刻のみから抽出される静的な音響特徴量である．これらの特徴量が時間的にどのように遷移したかを考慮することで，音声の合成や認識の品質が向上することが知られている．Δ 特徴量や ΔΔ 特徴量は，このような動的成分を扱うために導入される特徴量で，次式のように定義されることが多い．

$$\Delta \boldsymbol{x}_t = \frac{\boldsymbol{x}_{t+1} - \boldsymbol{x}_{t-1}}{2} \tag{2.7}$$

$$\Delta\Delta \boldsymbol{x}_t = \frac{\boldsymbol{x}_{t+1} - 2\boldsymbol{x}_t + \boldsymbol{x}_{t-1}}{4} \tag{2.8}$$

ここで，$\boldsymbol{x}_t$ は t 番目のフレームにおける静的特徴量である．Δ 特徴量は特徴量の一次微分（速度成分），ΔΔ 特徴量は二次微分（加速度成分）にあたる．音声認識や音声合成では，静的特徴量と動的特徴量を連結した静的動的特徴量を音響特徴量として用いる場合がある．

2.3　基本周波数の推定手法

基本周波数は，声帯振動の周波数に相当する音響特徴量である．音声から基本周波数を推定する手法として，様々な手法が提案されている．本節では代表的な手法を示す．ここでは基本周波数を f_o，基本周波数に対応する周期を T_o と示す．

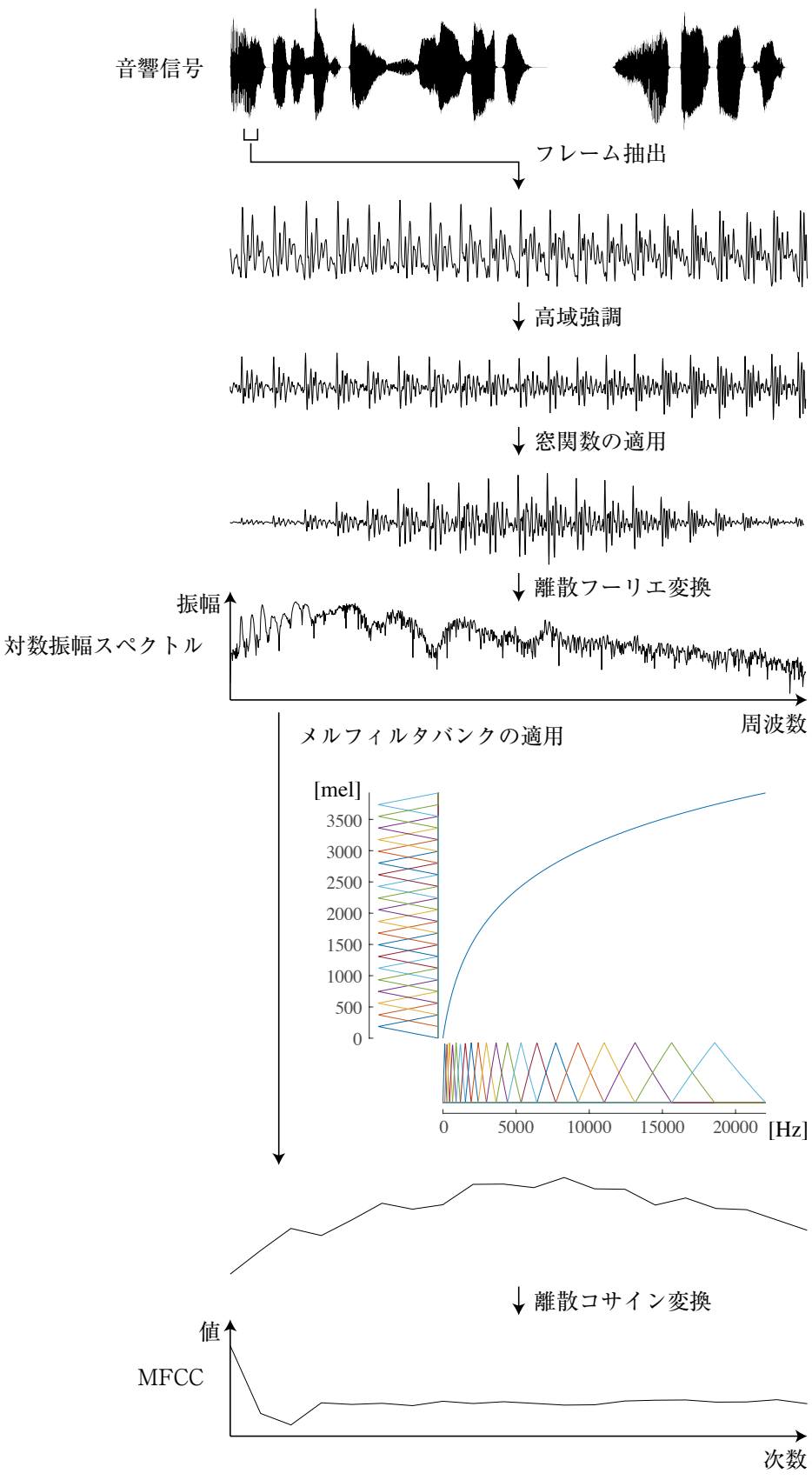


図 2.3　MFCC の計算手順の模式図．スペクトルに対してメル軸状で等間隔のフィルタバンクを適用し，それを離散コサイン変換することで得る．

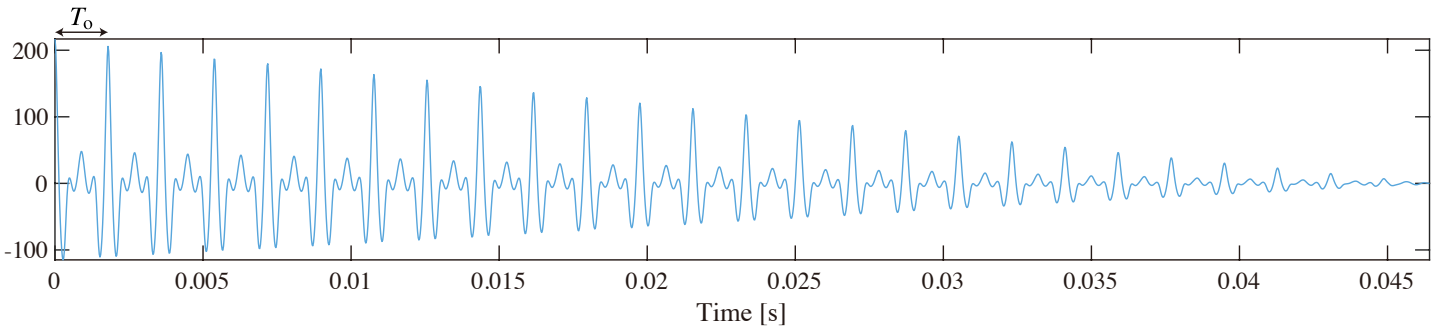


図 2.4　自己相関関数の例．時刻 0 で最大値を持つ周期的な形状を持つ．図では T_o は 0.0018 s で，基本周波数 f_o は 5.6×10^2 Hz と推定される．

2.3.1　ゼロ交差法

ゼロ交差法は，信号の振幅が 0 となるゼロ交差点の時刻をもとに基本周波数を決定する手法である．基本周波数を超える周波数の成分の信号に大きく影響されるため，$2f_o$ 以下の低域のみを通過させるフィルタを適用する必要があり，このフィルタリングは基本波フィルタリングとよばれる．

2.3.2　ケプストラム法

基本周波数は，スペクトル上では櫛形で周期が f_o の周期的な形状となって現れる．ケプストラムにおいては，図 2.2 に示すように，高いケフレンシーに音源波に対応する成分が現れる．したがって，ケプストラム上で特定のケフレンシーにあるピークを抽出することで，基本周波数を推定できる．ピークのあるケフレンシーが周期 T_o に相当する．

2.3.3　自己相関法

自己相関関数は，音響特徴量の 1 つであり，与えられた信号とその信号を時間的にシフトした信号の類似度を与える関数である．すなわち，ある信号と，その信号を τ だけシフトした信号を考え，その信号の類似度を $r(\tau)$ とし，それを信号にわたって計算する．信号 $y(t)$ の自己相関関数 $r(\tau)$ は次式で定義される．

$$r(\tau) = \int_{-\infty}^{\infty} y(t)y(t+\tau)dt \tag{2.9}$$

有声音などの周期的な信号の自己相関は，図 2.4 のように，時刻 0 を最大値とし，周期 T_o の周期的な形状を持つ．そのため，自己相関関数のピークを検出することで，周期 T_o を推定できる．

2.3.4　発展的な基本周波数推定法

YIN は，自己相関法に類似する基本周波数推定法である [49]．自己相関法では，T_o 未満の時刻のピークを誤検出する可能性がある．YIN では，自身の信号との相互相関関数を利用することで，この問題を抑制する．

DIO は，ゼロ交差法にもとづく基本周波数推定法である [50, 51]．ゼロ交差法においては，低域通過フィルタの遮断周波数を適切に決定する必要がある．DIO は，複数の低域通過フィルタを用いて基本周波数を推定し，そのうちもっともらしい基本周波数を選択する手法である．

さらに発展したゼロ交差法にもとづく基本周波数推定法として，Harvest が提案されている [52]．DIO と同様に複数のフィルタを用いて基本周波数の候補を抽出するが，雑音などによってゼロ交差法で基本周波数の候補が得られない可能性を考慮して，基本周波数の候補を時間的にぼかすことで基本周波数の候補を増加させる．さらに，得られた基本周波数の候補の信頼度を瞬時周波数を用いて計算し，種々の方法で時間的な連続性を考慮することで，非常に自然な基本周波数系列を推定する．Harvest はとくに高い品質の音声合成を実現できる基本周波数推定法として知られている．

深層学習による基本周波数推定法も提案されている．Crepe（convolutional representation for pitch estimation）は，畳み込みニューラルネットワーク（convolutional neural network; CNN）によって信号を分類することで，基本周波数を推定する手法である [53]．16 kHz で標本化された 1024 サンプルの信号を入力とし，32.7 Hz から 1975.5 Hz までの対数領域で等間隔な 360 の候補に信号を分類することで，基本周波数を推定する．PYIN（probabilistic YIN）[54] などの既存手法と比較して，精度が高く雑音に対する頑健性が強いことが示されている．各時刻で独立に基本周波数の確率分布を推定するが，ビタビ探索によって基本周波数系列を平滑化できる．

第３章

歌唱者ダイアライゼーション

3.1　はじめに

歌唱者ダイアライゼーションは，図 3.1 のように複数人が歌唱する楽曲から「誰がいつ歌唱しているか」を推定する技術である．歌唱者ダイアライゼーションによって，複数人が歌唱する楽曲の自動アノテーションや，それを利用したアプリケーションの構築などが可能になる．

歌唱者ダイアライゼーションは，複数人の会話から「誰がいつ発話しているか」を推定する話者ダイアライゼーション技術に着想を得たものである．話者ダイアライゼーションは 1990 年代ごろから研究されており，多様な会話音声の自動アノテーションなどに応用可能である [55]．話者ダイアライゼーションのチャレンジには CHiME Challenges [56] や DIHARD Challenges [57] などが開催されており，ダイアライゼーションそのものの性能や，話者ごとの音声認識の性能などによって評価される．映像などを利用したマルチモーダルな話者ダイアライゼーション手法についても研究がなされている [58, 59].

歌唱者ダイアライゼーションは，アフリカの民族音楽に対して行われた研究が報告されている [37]．この研究では，古典的な話者ダイアライゼーションの手法を適用することで歌唱者ダイアライゼーションを実現している．しかし，いくつかの理由で，話者ダイアライゼーションの手法を適用するだけでは高い精度の歌唱者ダイアライゼーションを行うことはできないと考えられる．まず，古典的な話者ダイアライゼーションの手法では，発話のオーバーラップ，すなわち複数人の同時発声に対応できない．オーバーラップを考慮した話者ダイアライゼーション手法がいくつか提案されており [60, 61, 62]，それらの歌唱者ダイアライゼーションにおける適用可能性について議論する必要がある．また，この研究では伴奏音による影響が示唆されている．音源分離技術や音声強調技術などを導入することで，ダイアライゼーション性能の向上が期待できる．さらに，話声と歌声では音響的な違いがあり，これによって歌唱者ダイアライゼーションが困難になると考えられる．たとえば，表 3.1 に示すように，歌声は基本周波数の幅が広く，より音素の継続長が長い．基本周波数の幅が広いことで，音の高さによってスペクトル包絡に影響が出たり，音高が高い箇所でスペクトル包絡の抽出精度が低下したりすることが懸念される．また，音素の継続長が長いことで，Δ 特徴量や $\Delta\Delta$ 特徴量が有効に活用できないことが懸念される．同一の歌詞の発話と歌声を用いた研究では，これらの違いが音響特徴量に影響を及ぼすことが指摘されている [63]．さらに，歌声と会話音声では，同時に発話している時間や，同時に発話するときの人数が大きく異なる．図 3.2 に，CHiME-6 Challenge で用いられた評価セットと，本研究の評価に用いたデータセットにおける，同時に発声している人数の分布を示す．歌声においては，多数の歌唱者が同時に発声している区間がより長い．これらの理由から，高品質な歌唱者ダイアライゼーションを実現するためには，既存の話者ダイアライゼーション手法の有効性について議論する必要がある．

本章では，話者ダイアライゼーション手法におけるこれらの問題を解決する歌唱者ダイアライゼーション手法を提案する．提案法では，これまでの話者ダイアライゼーション手法では後処理として用いられることが多かったオーバーラップ検出について再考する．また，オーバーラップ検出を高い精度で行うため，自己相関関数を活用した新たな音響特徴量を導入する．さらに，ブラインド音源分離（blind source separation; BSS）法を利用することで，伴奏音のある楽曲に対しても適用可能な手法を構築する．加えて，歌唱者の情報を短時間の音声から抽出する

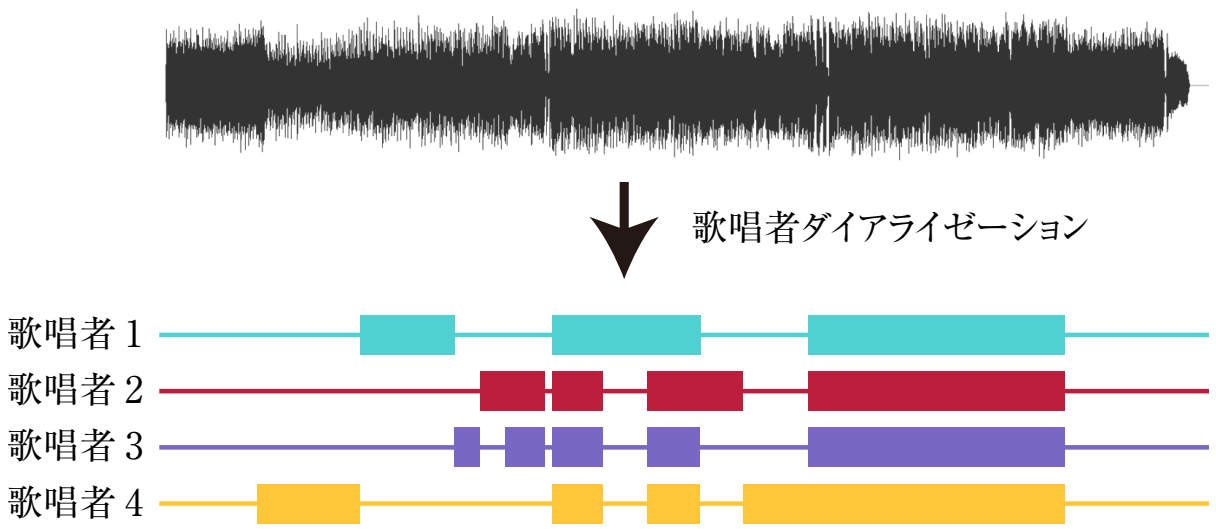
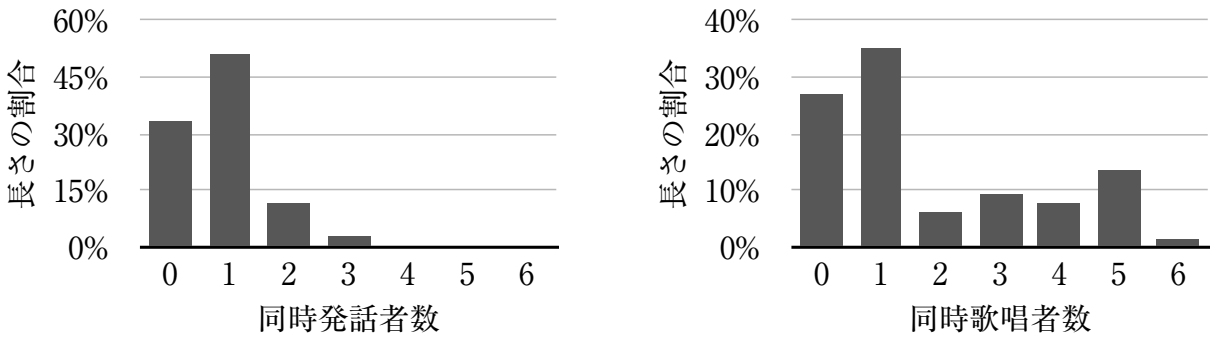


図 3.1　歌唱者ダイアライゼーションの模式図．入力された音楽音響信号から，誰がいつ歌唱しているかを推定する．

表 3.1　発話と歌声における音響特徴量の統計的な違い．発話のデータセットには，ATR 日本語音声データベースの話者 FTK が発話した日本語音素バランス文を用いた [64]．歌声のデータセットには，東北きりたん歌声データベース [65] を利用した．± は標準偏差を表す．

カテゴリ	$\log f_0$ [log Hz]	音素継続長 [ms]
発話	5.44 ± 0.24	37 ± 82
歌声	5.87 ± 0.31	151 ± 191



(a) CHiME-6 Challenge の評価データセット [56, 62]

(b) 本論文の実験の評価に用いた歌声のデータセット

図 3.2　会話音声と歌声における同時発話者数のヒストグラム．会話音声のデータセットには 4 人の話者が会話する音声のみが含まれているため，同時に 5 人以上が発話している区間は存在しない．

ために，ArcFace とよばれる埋め込み表現抽出法を導入する．これらの技術により，人数や歌唱者の情報が未知であっても適用可能な歌唱者ダイアライゼーション手法が実現される．本章では，市販の CD に収録されている楽曲を用い，提案法の実験的評価を行う．

複数人の歌唱における認識においては，特定の歌唱者がその楽曲に含まれているかを認識する target-singer detection，各時刻で特定の歌唱者が歌唱しているかを検出する target-singer tracking などの技術が提案されている [35,36]．本章で議論する歌唱者ダイアライゼーションは，事前知識のない target-singer tracking と解釈できる．

本章では，まず話者ダイアライゼーションにおいて複数人の発声を効果的に扱うことのできる手法を関連研究として解説する．次に，話者ダイアライゼーションに用いられる既存手法を 2 つ述べる．これらは本論文で提案する歌唱者ダイアライゼーション手法の基礎となる技術である．そして，提案法である歌唱者ダイアライゼーション手法について詳細に述べる．さらに，提案法を評価する実験について述べる．実験は，提案した音響特徴量についてのみ評価する実験と，提案法全体の性能について評価する実験を行う．

3.2 関連研究

話者ダイアライゼーション研究においては，複数人の同時発声を効果的に扱うことのできる手法がいくつか提案されている．

Target-speaker VAD（TS-VAD）は，各話者の話者表現と音響特徴量系列を入力として，各話者の発声状態を推定する機構であり，これを用いた話者ダイアライゼーション手法が提案されている [62]．TS-VAD を用いた話者ダイアライゼーション手法は，本章におけるベースライン手法であり，3.3 節に詳細に解説する．

End-to-end neural diarization（EEND）は，ニューラルネットワークによって，直接ダイアライゼーション結果を推定する手法である [66]．ニューラルネットワークを用いて直接結果を推定する場合，正解ラベルと推定ラベルの対応付け，すなわちパーミュテーションを考慮する必要がある．EEND では permutation-invariant training（PIT）損失関数を用いることでこの問題を解決する．EEND の性能を向上するため，自己注意（self-attention）機構を導入した self-attentive EEND（SA-EEND）も提案されている [61]．EEND は，TS-VAD と同様，認識できる最大の話者数がニューラルネットワークのアーキテクチャによって事前に定まっている．任意話者数の音声の認識に対応した EEND を構築するため，encoder-decoder based attractor calculation（EDA）とよばれる機構が導入されている [67].

オーバーラップを事前に検出し，同時発声人数に応じて異なるアプローチで認識を行う手法も提案されている [68,69]．こうした手法は，本章で提案する歌唱者ダイアライゼーション手法のアプローチと類似した手法である．

音源分離を事前に行い，分離された音源に対してそれぞれ認識を行うことで、オーバーラップのある会話音声のダイアライゼーションを実現する手法も提案されている [70]．こうした手法も，事前にオーバーラップを検出する手法の 1 つといえる．ことに歌唱者ダイアライゼーションにおいては，重なり合う歌声が調和・同期していたり，

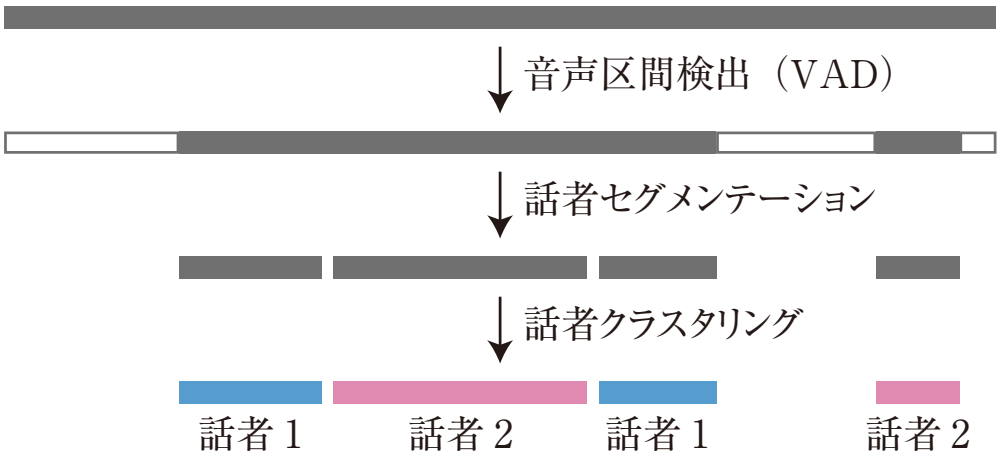


図 3.3　クラスタリング法のダイアグラム．まず発話区間を検出し，発話の区間に対して話者に応じたセグメンテーションおよびクラスタリングを行うことで，ダイアライゼーションを実現する．

ほぼ同一の音高であったりするなど，音源分離技術を適用することは難しく，こうしたアプローチを採用することはできない．

3.3　話者ダイアライゼーションを目的としたベースライン手法

本節では，本論文で提案する歌唱者ダイアライゼーション手法の基礎となる，話者ダイアライゼーション手法について述べる．本節で扱う技術は，会議音声などの会話音声を対象としたダイアライゼーション技術である．

3.3.1　クラスタリング法

もっとも古典的なダイアライゼーション手法は，クラスタリング法である [55, 71]．この手法はおもに，音声区間検出（voice activity detection; VAD），セグメンテーション，クラスタリングの 3 手順からなる．図 3.3 にこの手法のダイアグラムを示す．この手法は，オーバーラップを考慮せず，各時刻で最大でも 1 人の話者が発話していると仮定する．

音声区間検出

VAD は，音声の各時刻で少なくともいずれかの話者が発話しているかを検出する手順である．これは，誰も発話していない区間に対して後段の処理を行わないようにするための手順である．発話していない区間によって，各話者の音響モデルなどに影響が現れるのを避けるために行う．また，ダイアライゼーションの評価指標では VAD の誤りがそのまま評価指標に現れるため，最終的な評価指標の改善に重要な手順である．

VAD は，各時刻の MFCC などの音響特徴量を音声区間と無音区間の 2 クラスに分類することで実現される．サポートベクトルマシン（support vector machine; SVM）や線形判別分析（linear discriminant analysis; LDA）などの識別モデルを用いた検出法が提案されている [72, 73, 74]．隠れマルコフモデル（hidden Markov model; HMM）や，回帰型ニューラルネットワーク（recurrent neural network; RNN）や長短期記憶（long-short term memory; LSTM）

ネットワークなどの記憶を持つニューラルネットワークなど，時系列情報を考慮した識別モデルも導入されている [75, 76, 77].

ことに音楽情報処理においては，類似の問題として歌声区間検出がある．歌声の認識においては伴奏音が含まれた混合音から歌声の区間を検出する必要があることが多く，それを目的とした手法がいくつか提案されている．まず，新たな音響特徴量が導入されている．MFCC のほかに，フラクトグラムとよばれる歌声のピッチのゆらぎを検出できる音響特徴量や，スペクトルの平坦さを評価する音響特徴量，声道特徴の時間変化を検出できる vocal variances とよばれる音響特徴量を導入することで，誤った歌声区間の検出を抑制する手法が提案されている [78]. 事前に調波音・打楽器音分離（harmonic/percussive sound separation; HPSS）を行い，それぞれから抽出した音響特徴量を入力することで，高い精度で歌声区間検出を行う手法も提案されている [79]．また，統計的手法を用いる場合，パワーを利用して認識を行ってしまいパワーの違いに頑健な手法を構築するのが困難な問題が指摘されており，CNN のフィルタ係数に制約を与えることでパワーの違いに頑健な歌声区間検出法も提案されている [80].

話者セグメンテーション

話者セグメンテーションの手順では，各セグメントが単一の話者によって発話されているように音声を分割する．話者交代検出などともよばれる．音声全体を 1 つのセグメントからなるとする仮説か 2 つのセグメントからなるとする仮説かをなんらかの規準で選び，その分割する過程を繰り返すことでセグメンテーションを行う，トップダウンの方式を採用することが多い．選択の規準には，ベイズ情報量規準（Bayesian information criterion; BIC）を用いる手法が古典的である [81]．各セグメントを混合ガウスモデル（Gaussian mixture model; GMM）などでモデル化し，より BIC の小さい仮定を選択する．BIC は調節が困難なため，BIC における尤度比とパラメータ数に対する重みの比率を可変にした修正ベイズ情報量規準（modified Bayesian information criterion; mBIC）が話者ダイアライゼーションでは広く用いられる．

話者クラスタリング

話者クラスタリングの手順では，セグメントを話者に応じてクラスタリングすることで，最終的なダイアライゼーションを実現する．これは話者ダイアライゼーションにおけるもっとも本質的な手順である．クラスタリングでは，主にトップダウン型（分割型）の手法とボトムアップ型（凝集型）の手法に分類される．この模式図を図 3.4 に示す．トップダウン型のクラスタリングでは，すべてのセグメントが同一話者（クラスタ）に属するとまず仮定し，繰り返しクラスタを分割する．一方，ボトムアップ型のクラスタリングでは，すべてのセグメントが異なる話者（クラスタ）に属するとまず仮定し，繰り返し類似のクラスタを凝集する．話者セグメンテーションと同様に mBIC を規準として用いる手法のほか，Kullback–Leibler（KL）ダイバージェンス規準でのクラスタリングなどが提案されている [81, 82].

スペクトルクラスタリングも話者ダイアライゼーションにおけるクラスタリングの手法として利用されている [83]．スペクトルクラスタリングは，類似度行列から構成したグラフのラプラシアン行列を利用して埋め込み表

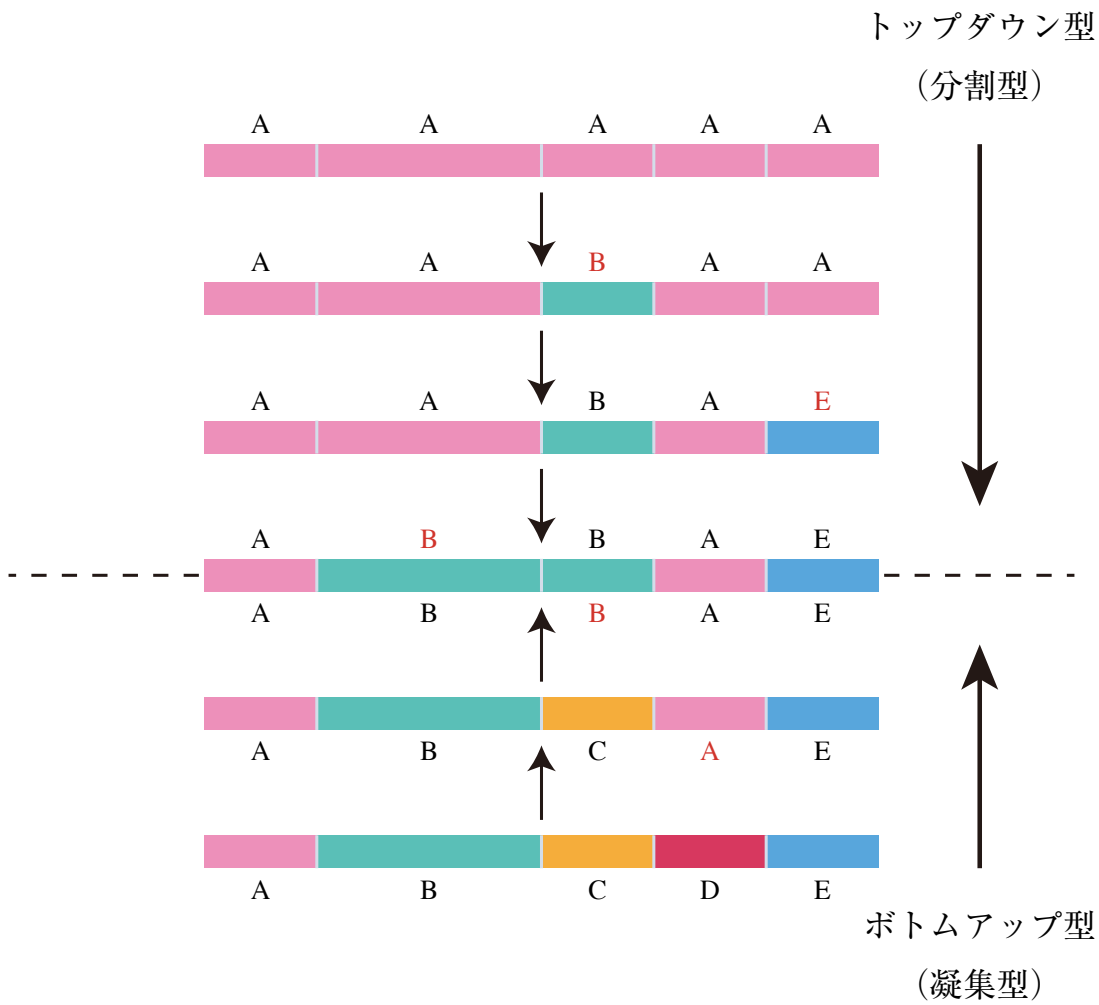


図 3.4　トップダウン型（分割型）クラスタリングとボトムアップ型（凝集型）クラスタリング．トップダウン型ではすべてのセグメントが同一話者による発話であると仮定して分割を行う一方，ボトムアップ型ではすべてのセグメントが異なる話者による発話であると仮定して凝集を行う．理想的にはともに同一の結果に至る．

現を得て，これを k-means アルゴリズムなどによってクラスタリングするものである．単純な k-means アルゴリズムなどの位置関係をもとにクラスタリングする手法と比較して，グラフのノードの集まりをもとにクラスタリングするため，より複雑な境界面を持つ特徴量のクラスタリングが可能であるという特徴がある．また，類似度行列の固有値を昇順に並べたときの差分の列である eigengap に着目し，これを正規化した normalized maximum eigengap（NME）とよばれる値を利用してクラスタ数などのパラメータを決定する手法が提案されている [84].

話者表現の活用

古典的な話者クラスタリングでは，各セグメントを GMM でモデル化して評価する．一方，話者認識においては，話者表現とよばれる可変長の音響信号から抽出可能な話者の情報を表す特徴量が用いられる場合があり，代表的な話者表現として i-vector が挙げられる [85]．話者ダイアライゼーションにおいては，こうした話者表現を話者クラスタリングに活用する手法が提案されている [86, 87]．話者表現がごく短時間の音声から抽出できる場合，音声を 500 ms 程度で等間隔にセグメンテーションし，各セグメントから抽出された話者表現をクラスタリングする

ことで，話者ダイアライゼーションを実現できる [88]．こうした手法は，話者交代に応じた緻密な話者セグメンテーションを行う必要がないという利点がある．

3.3.2　Target-speaker VAD を用いた話者ダイアライゼーション手法

クラスタリング法は，各時刻で最大でも 1 人の話者が発話していると仮定するため，複数人の同時発声を認識することができない．これを解決するため，ダイアライゼーションの後処理として同時発声を認識する手法がいくつか提案されている [60]．本論文では，TS-VAD を用いた話者ダイアライゼーション手法 [62] をベースライン手法として引用する．TS-VAD は，各時刻の音響特徴量と，各話者の話者表現にもとづき，各話者がその時刻で発話しているかをそれぞれ認識するものである．当該研究では，4 人の話者が会話するディナーパーティを対象としてシステムを構築している．この手法は次の 4 手順からなる．

1. **初期ダイアライゼーション**．クラスタリング法により，各時刻で最大でも 1 人が発話していると想定して話者ダイアライゼーションを行う．当該研究では x-vector [89] を短時間の音声から抽出しクラスタリングすることで話者ダイアライゼーションを実現する．
2. **話者ごとの話者表現の抽出**．話者ダイアライゼーションの結果をもとに，各話者の i-vector [85] を推定する．
3. **TS-VAD**．各時刻で，各話者が発話しているかどうかを推定する．各話者に対する VAD を行っていることに相当するため target-speaker VAD とよばれる．TS-VAD のモデル $f_{\mathrm{TS-VAD}}$ は次式のように TS-VAD を実現する．

$$\left[s_{1,[1:T]}, s_{2,[1:T]}, s_{3,[1:T]}, s_{4,[1:T]}\right] = f_{\mathrm{TS-VAD}}\left(\boldsymbol{x}_{[1:T]}, \boldsymbol{y}_1, \boldsymbol{y}_2, \boldsymbol{y}_3, \boldsymbol{y}_4\right), \tag{3.1}$$

ここで，$\boldsymbol{y}_i$ は i 番目の話者の話者表現，$\boldsymbol{x}_{[1:T]}$ は音響特徴量の系列，$s_{i,[1:T]}$ は i 番目の話者の発話状態の系列を表す．TS-VAD では事前に話者数を既知として認識を行う．当該研究では 4 人の話者表現を用いて 4 人の発話状態をそれぞれ推定する．

4. **後処理**．不適切な結果を抑制するため，後処理を適用する．メディアンフィルタ，短時間の発話区間の除去，スコアの閾値処理，HMM を用いた平滑化が後処理として提案されている．

話者ごとの話者表現は，オーバーラップ区間を含めた発話区間全体から推定するため，オーバーラップ区間により品質が低下する可能性がある．TS-VAD を用いた手法では，手順 2 と 3 を繰り返すことでこの問題を抑制する．TS-VAD の結果を用いて i-vector を再度抽出することで，TS-VAD の結果が次第に改善し収束する．

TS-VAD では，あらかじめ認識される話者の人数がニューラルネットワークのアーキテクチャにより決定されている．そのため，話者数が未知の場合に TS-VAD を適用することができない．そこで，TS-VAD により任意の話者数の会話音声を認識する手法が提案されている [90]．この手法では，アーキテクチャにより事前に決定している話者数よりも認識する話者数が少ない場合には学習に用いた話者の話者表現を無作為に入力し，多い場合にはもっ

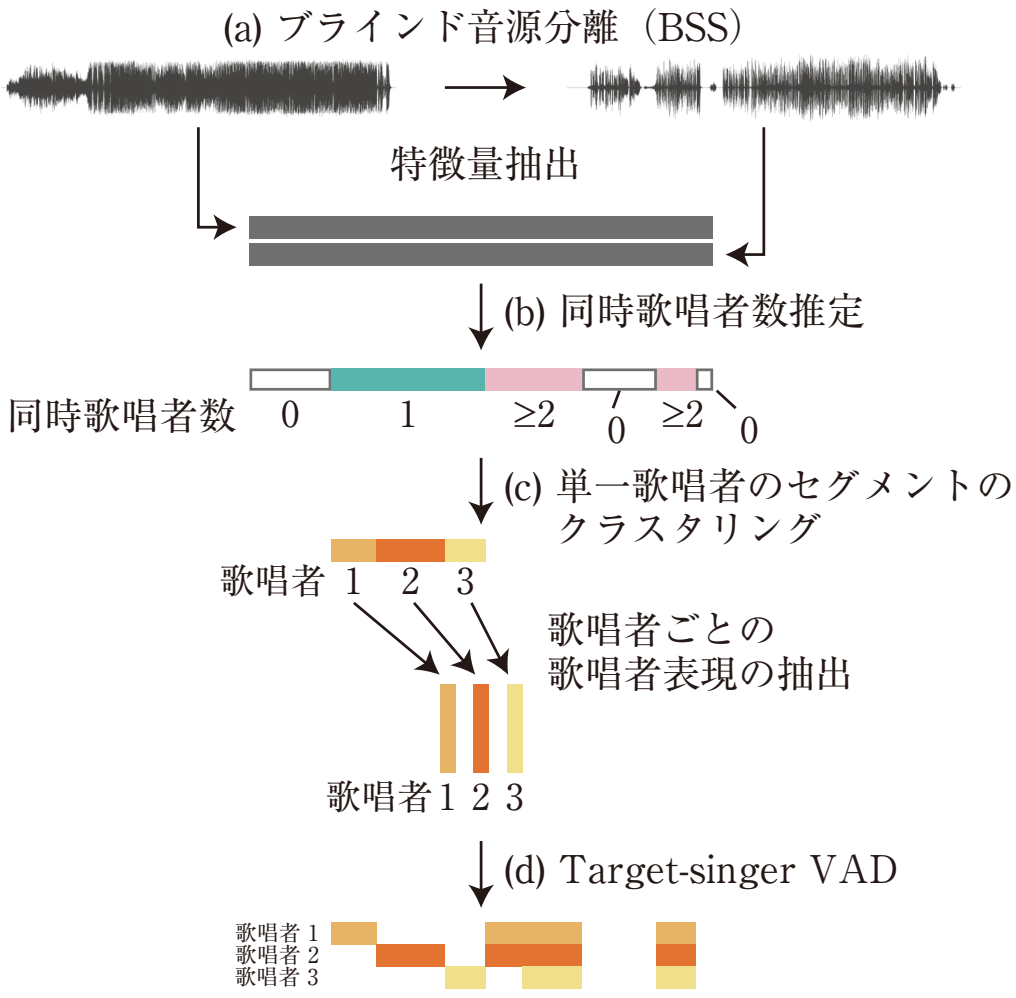


図 3.5　提案する歌唱者ダイアライゼーション法の概観図.

とも発話時間の少ない話者を無視する.

3.4　歌唱者ダイアライゼーションを目的とした提案法

3.4.1　手法の概観

本論文では，歌唱者ダイアライゼーションを目的とした，新たなダイアライゼーション手法を提案する．この概観を図 3.5 に示す．提案法は TS-VAD を用いた話者ダイアライゼーション法を基礎としているが，複数人歌唱の認識をクラスタリングの前の手順で行うことで，人数の推定誤りを抑制する．また，BSS を導入することで，伴奏音のある歌声の認識を実現可能にする．本節では，提案法の詳細について，手順に従って述べる．

3.4.2　前処理

BSS 手法により，伴奏音との混合音から歌声のみを抽出する．図 3.5 の (a) にあたる．本論文では，BSS 手法として Spleeter [14] を採用する．Spleeter は，音楽音響信号を対象としたオープンソースの BSS ライブラリである．Spleeter の学習済みモデルでは，12 層の U-Net [91] を用い，各時刻・周波数ビンがどのクラスに属するかを示す

スペクトログラムのソフトマスクを推定する．理想的には，BSS を行い分離した歌声のみを用いれば歌唱者ダイアライゼーションを実現できるが，Spleeter を用いた BSS は歌声の歪みが大きく分離した歌声のみでは高い品質の認識を実現できない．そこで，提案法では，分離前の混合音および分離後の歌声からそれぞれ音響特徴量を抽出し，これを結合して利用する．

3.4.3　同時歌唱者数推定

複数人の同時発声を扱えるダイアライゼーション手法においては，多くの場合後処理によって同時発声を認識する．すなわち，まずすべての区間が最大でも単一の話者によって発話されていると仮定してダイアライゼーションを行い．その事後処理として複数人同時発声を扱う．しかし，歌声においては会話音声と比較して複数人が同時に発声している区間が長いため，こうしたアプローチでは複数人同時歌唱が影響して最終的なダイアライゼーション結果が劣化しやすい．そこで提案法では，ダイアライゼーションを行う前にあらかじめ同時に歌唱している人数を推定しておき，この情報を用いて同時歌唱者数に応じたダイアライゼーションを行う．

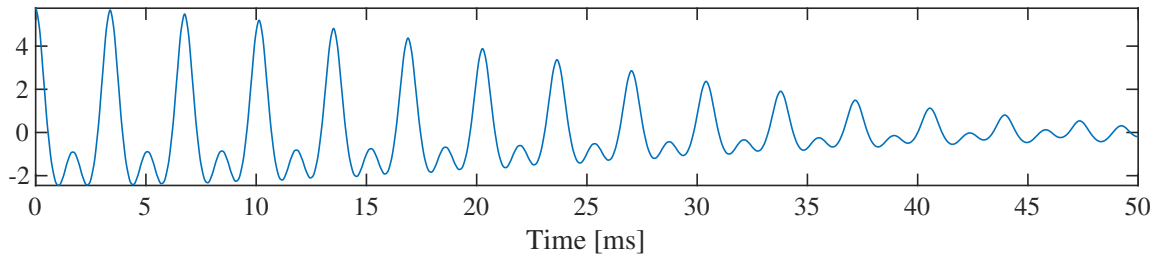
同時歌唱者数推定の手順では，各時刻の音響特徴量を同時に歌唱している人数が 0 人（非歌唱区間），1 人，2 人以上の 3 クラスに分類する．したがって，このステップは voice activity と overlap の検出を同時に行っていることに当たり，voice activity and overlap detection と表現できる．この手順は図 3.5 の (b) にあたる．提案法では分類に双方向長短期記憶（bidirectional LSTM; BiLSTM）を用い，MFCC，パワー，Cosacorr スコアの 3 種類の音響特徴量を入力して分類を行う．

Cosacorr スコアは，より高い精度で同時歌唱者数推定を行うために導入する，本論文で提案する音響特徴量である．複数人が同時に歌唱しているとき，たとえ複数人の歌唱がユニゾンであったとしても，その基本周波数はわずかに異なる．音響信号の自己相関関数は基本周波数に対応する周期を持つ周期的な波形を描くが，複数人が同時に歌唱している場合この基本周波数の違いによって自己相関関数が歪む．図 3.6 に自己相関関数の例を示す．複数人が同時に歌唱していることにより，自己相関関数の周期的な波形に歪みが生じていることが確認できる．Cosacorr スコアは，このような自己相関関数の歪みをコサイン距離を用いて計測し，信号の複数人らしさを評価する音響特徴量である．$X = [x_1, x_2, \ldots, x_N]$ を音響信号の自己相関関数とする．まず，ピーク検出アルゴリズムを用いて，局所最大値 $x_{p_1}, x_{p_2}, \ldots, x_{p_P}$ を得る．P は検出するピークの数で，あらかじめ決定する．自己相関関数は $n = 1$ でもっとも高い値となるため，p_1 は常に 1 である．ここで，k 次の Cosacorr スコア $\mathrm{Cosacorr}_k$ は次式で計算される．

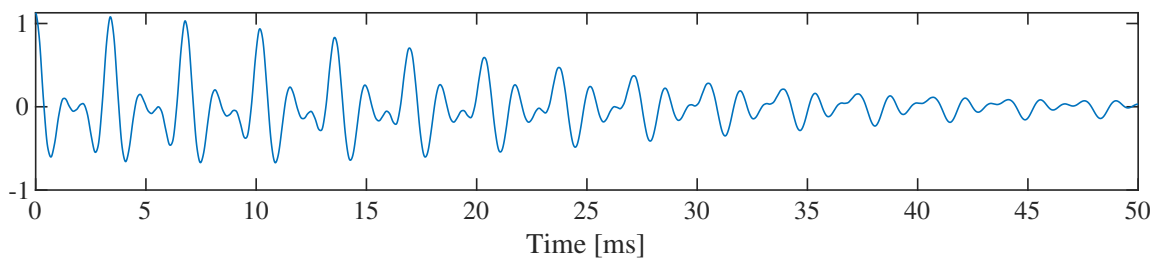
$$\mathrm{Cosacorr}_k = \frac{P_{k+1}}{P_1}\left(1 - \frac{\sum_{i=1}^{p_2-1} x_i y_{k,i}}{\sqrt{\sum_{i=1}^{p_2-1} x_i^2}\sqrt{\sum_{i=1}^{p_2-1} y_{k,i}^2}}\right) \tag{3.2}$$

$$P_m = \frac{1}{p_{m+1} - p_m}\sum_{i=p_m}^{p_{m+1}-1} x_i^2 \tag{3.3}$$

ここで $y_{k,1}, y_{k,2}, \ldots, y_{k,p_2-1}$ は，自己相関関数の部分 $x_{p_{k+1}}, x_{p_{k+1}+1}, \ldots, x_{p_{k+2}-1}$ を線形補間によりリサンプリングすることで得られる系列である．この補間は，第 1 周期 $x_1, x_2, \ldots, x_{p_2-1}$ とのコサイン距離を計算するために行



(a) 単一歌唱者が歌唱している歌声の信号の自己相関関数．8 次 Cosacorr スコアの和は 0.0029 である．



(b) 2 歌唱者が歌唱している歌声の信号の自己相関関数．8 次 Cosacorr スコアの和は 0.3048 である．

図 3.6　歌声の音響信号の自己相関関数の例．2 つの自己相関関数は同一楽曲 BM11『フタリの記憶』の同一箇所「ずっとずっと 空で見守っているよ」の自己相関関数を示している．

う．理想的な自己相関関数では第 1 周期と第 k 周期の長さは一致するが，信号が標本化されており，また信号が歪んでいるため，現実的にはこれらの周期が一致しない．そこで，第 k 周期の長さをリサンプリングによって第 1 周期の長さに一致させる必要がある．また，パワーによる重み付けを行うため $\frac{P_{k+1}}{P_1}$ を乗じている．Cosacorr スコアは第 1 周期と第 k 周期の違いを測る音響特徴量のため，より高い Cosacorr スコアはより複数人らしいことを示す．図 3.7 にこの Cosacorr スコアの導出方法の模式図を示す．人の発声できる f_0 と比較して適切なピークが検出できない場合，Cosacorr スコアの計算は失敗する．この場合，本論文では計算できなかった値を 0 として扱い音響特徴量として用いる．また，Cosacorr スコアは伴奏音のない歌声から抽出することを想定して設計されている．伴奏音が含まれる場合，計算を行うことはできるものの，伴奏音によって自己相関関数が影響を受けるため，適切な値が得られない場合がある．しかし，本論文では伴奏音が含まれる信号から抽出した Cosacorr スコアも認識において効果があると考えるため，分離前後の両者の信号から抽出した Cosacorr スコアを音響特徴量として用いる．

3.4.4　歌唱者表現の抽出

固定長の短時間セグメントから，歌唱者表現を抽出する．ダイアライゼーションにおいて，話者表現の品質は最終的な性能を左右する重要な要素である．古典的な話者ダイアライゼーション手法では，話者表現として i-vector [85] や x-vector [89] が採用されることが多い．しかし，3.1 節でも述べたように，歌声は発話に比べて f_0 の範囲が広く音素継続長も長いなどの特徴があり，短時間の歌声から識別性の高い歌唱者表現を得ることは難しい．本論文では，より識別性の高い歌唱者表現を得るために，ArcFace を用いて歌唱者表現を抽出する．

ArcFace（additive angular margin loss）は，より識別性の高い埋め込み表現を得ることのできるネットワーク

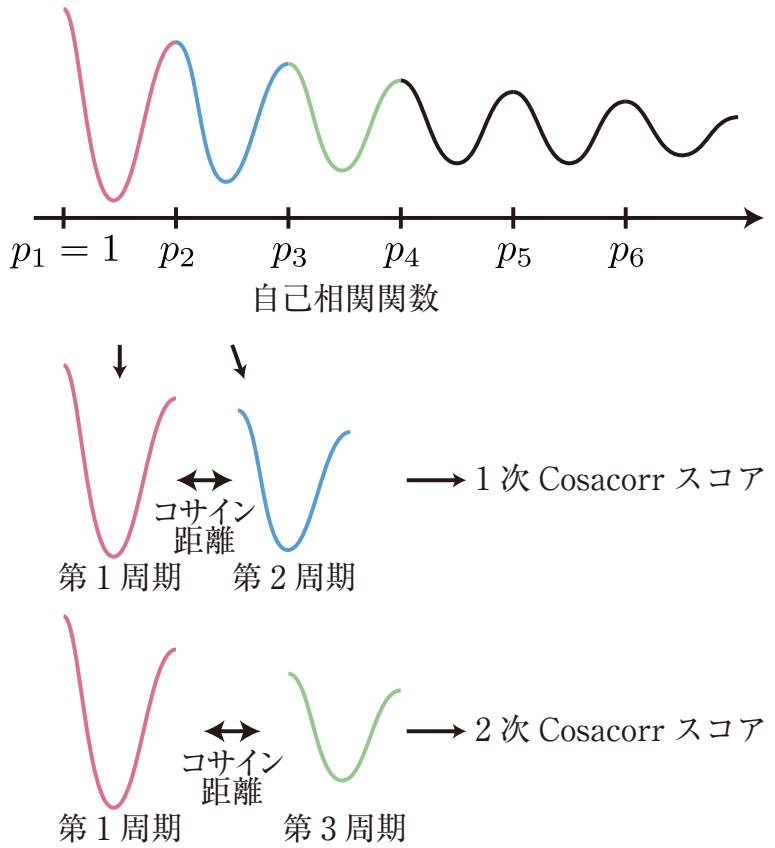


図 3.7　Cosacorr スコアの計算手順の模式図．第 k 次の Cosacorr スコアは，自己相関関数の第 1 周期と第 $k+1$ 周期の形状の違いをコサイン距離によって表現したものである．

アーキテクチャであり，顔認識を目的として導入された [92]．従来の埋め込み表現は，最終層の活性化関数を softmax とし，交差エントロピー損失にもとづいて学習された分類器の，ボトルネック層（多くの場合最終層の直前層）を利用することが多い．このような損失関数は，次式で表現できる．

$$L = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{W_{y_i}^{\top}x_i}}{\sum_{j=1}^{N}e^{W_j^{\top}x_i}} \tag{3.4}$$

ここで，N はバッチサイズ，y_i は i 番目の埋め込み表現 x_i が属するクラスの番号，W_j はニューラルネットワークの最終層の重み行列の j 番目の列を表す．また最終層に適用されるバイアスベクトルは省略されている．しかし，この手法では，同一クラスの特徴量がより近く，逆に異なるクラスの特徴量がより遠くなるような，クラスタリングに適した識別性の高い埋め込み表現を得ることは難しい．そこで，損失関数にマージンを与えることで，より識別性の高い埋め込み表現を得る手法が提案された [93, 94]．ArcFace はそうしたマージンベースの損失関数を用いたネットワークアーキテクチャの一種であり，softmax を次式で置き換えるものである．

$$L = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{s(\cos(\theta_{y_i}+m))}}{e^{s(\cos(\theta_{y_i}+m))}+\sum_{j=1,j\neq y_i}^{n}e^{s\cos\theta_j}} \tag{3.5}$$

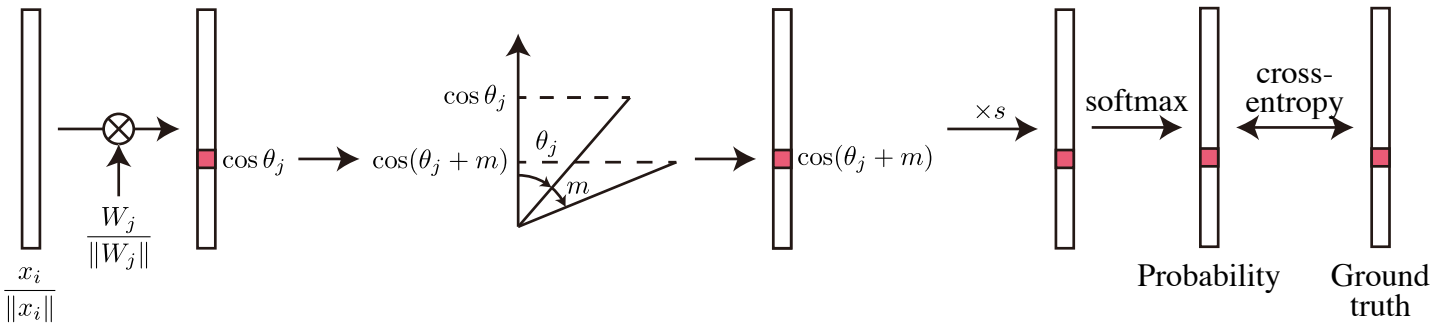


図 3.8　ArcFace における損失関数の計算の模式図．x_i は埋め込み表現，W_j は重み行列の第 j 列，m はマージン，s はスケーリングパラメータを表す．埋め込み表現ベクトルを正規化し，それに重み行列を乗じて得られたベクトルに対して，直接 softmax を計算するのではなく，正解ラベルの値にのみマージンを加えた上で softmax を計算する．これにより，より大きなマージンをもって正解ラベルを推測するように学習することができ，より識別的な埋め込み表現が得られる．

ここで，m はマージン，s はスケーリングパラメータである．また，θ_j は次式で定義される．

$$\theta_j = \arccos \frac{W_j^\top x_i}{\|W_j\|\|x_i\|} \tag{3.6}$$

この損失関数の計算の模式図を図 3.8 に示す．ArcFace は，従来の類似手法に比べて実装が容易であり，顔認識において高い性能を示す．このような利点から，提案法では ArcFace を歌唱者表現の抽出に用いる．

3.4.5　クラスタリング

得られた歌唱者表現を利用して，単一歌唱者が歌唱する区間に対してクラスタリングを行う．この手順は図 3.5 の (c) にあたり，3.3.1 節で述べたクラスタリング法によるダイアライゼーションに相当する．提案法では，NME を用いたスペクトルクラスタリング法 [84] を用いてクラスタリングを行う．

この手順では，クラスタリングに時間情報を利用しないため，歌唱者の頻繁な切り替わりやごく短時間のセグメントなどの不適切な結果を生じる可能性がある．このような結果を抑制するため，後処理として HMM を用いて平滑化を行う．用いる HMM を図 3.9 に示す．遷移確率と出力確率はハイパーパラメータとして事前に決定する必要がある．

最後に各クラスタの歌唱者表現を平均し，歌唱者ごとの歌唱者表現を得る．

3.4.6　Target-singer VAD

複数人が歌唱している区間において，各歌唱者が歌唱しているかを推定する．この手順は図 3.5 の (d) にあたり，TS-VAD を用いたベースライン手法においては TS-VAD のステップに相当する．提案法では，歌唱者がいない区間や単独の歌唱者が歌唱している区間を含めた音声全体に対して BiLSTM ネットワークを用いて識別する．同時歌唱者数が 1 人以下の区間については，この手順で推定した情報は用いず，これまでの手順で推定した結果を用い

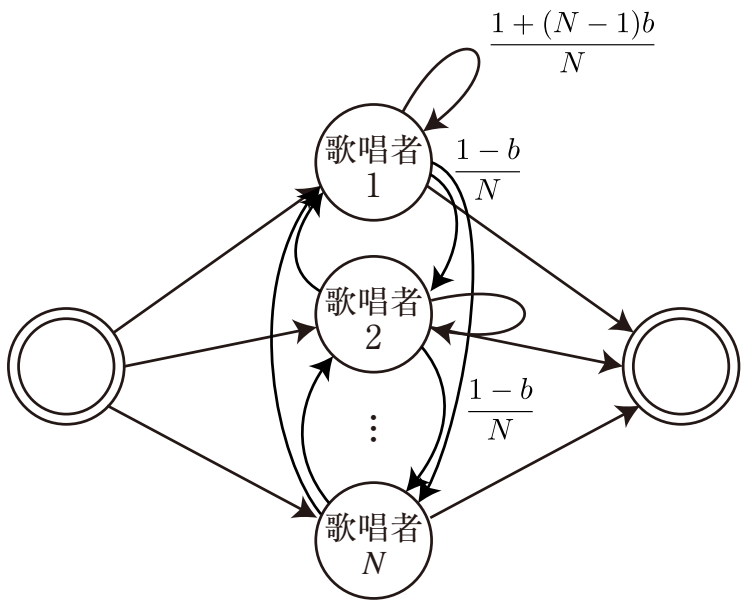


図 3.9　クラスタリングの後処理の平滑化に用いる HMM．N は歌唱者の数を表し，b は $0 < b < 1$ なるハイパーパラメータである．初期状態と最終状態を除く各状態は，各歌唱者の発声状態に対応する．

て最終的な結果を構成する．

TS-VAD は，よく現れる発話の重なり合いや話者交代のパターンを BiLSTM によって基礎知識として学習していると解釈できる．歌声の重なり合いや歌唱者の交代は，会話音声のそれとは異なり，様々なパターンが楽曲によって存在し，そこに創作性が現れる．したがって，事前に重なり合いや交代のパターンを学習する TS-VAD では，多様な楽曲の認識に適用できない可能性が高い．そこで，TS-VAD に類似した Personal VAD [95] を用いて，各歌唱者ごとにそれぞれ独立に歌唱状態の推定を行う．Personal VAD では，各時刻で各歌唱者について次の 3 クラスに分類する．

1. どの歌唱者も歌唱していない
2. 当該歌唱者が歌唱している
3. 当該歌唱者以外の歌唱者のみが歌唱している

推定時には 1 と 3 の区別は行わず，2 の場合のみその歌唱者が歌唱しているものとする．各時刻の音響特徴量，当該歌唱者の歌唱者表現，またその歌唱者表現とその時刻における歌唱者表現のコサイン類似度を入力として，歌唱状態を推定する．このアーキテクチャは，Personal VAD の提案論文 [95] では score and embedding conditional training と示されているものである．図 3.10 にこのアーキテクチャの概要を示す．

3.5　Cosacorr スコアの有効性の評価

本節では，Cosacorr スコアの有効性を評価するための実験について述べる．本実験では，少量の歌声を用い，1 人の歌唱者が歌唱した歌声か 2 人の歌唱者が歌唱した歌声かを 1 秒間の音声セグメントから認識する．本実験では，Cosacorr スコアのほかに i-vector を利用して人数を推定するシステムを構築する．Cosacorr スコアを用いた

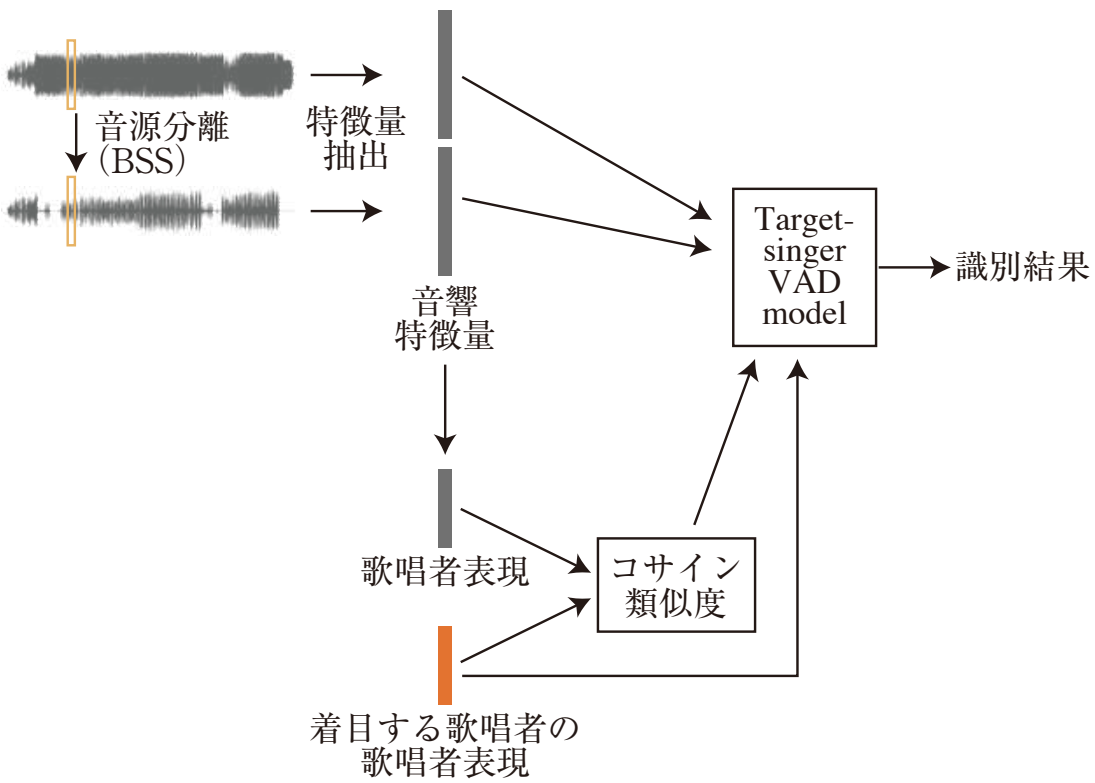


図 3.10　Target-singer VAD のアーキテクチャ．図中では，入力に用いる音響特徴量と歌唱者表現抽出に用いる音響特徴量を同一のものと示しているが，実験では異なる条件で抽出した音響特徴量を用いる．

システムおいても，i-vector を用いたシステムおいても，ともに SVM を用いて 2 クラス分類することにより人数を推定する．誰も歌唱していないセグメントや 3 人以上が歌唱しているセグメントは認識対象に含まれない．

3.5.1　実験条件

データセットとして，市販の CD から抽出した 12 楽曲を用いた．データセットには，複数の歌唱者がそれぞれ個別に歌唱した歌声が，伴奏音とミックスされた状態で収録されている．データセットの詳細は A.1 に記す．楽曲の長さは楽曲 AM12 を除いて約 2 分であり，楽曲 AM12 のみ 3 分 49 秒である．サンプリング周波数は 44.1 kHz である．歌唱者の数は 18 名であり，AM12 を除いた 11 楽曲は 9 人から 12 人が歌唱し，AM12 のみ 4 人が歌唱している．CD に収録されている伴奏音を用いて歌声を抽出し，本実験ではこれを利用した．音楽音響信号から伴奏音の差分を得る際には，歌声りっぷ[*1] を利用した．同一楽曲の歌声から 2 つ無作為に選んでユニゾン音声を作成する操作を楽曲総数と同数回行うことで，ソロの歌声と同数の 2 歌唱者によるユニゾンの歌声を作成した．学習には 12 曲中 AM1–AM10 の 10 曲を用い，評価には AM11 と AM12 の 2 曲を用いた．AM11 の歌唱者は学習データ中に含まれおり，AM12 の歌唱者は学習データ中に含まれていない．すなわち，楽曲 AM11 は closed-singer 条件，楽曲 AM12 は open-singer 条件である．すべての音楽音響信号はステレオで収録されているが，伴奏音との差分を計算後，チャンネル間の平均値を算出し，モノラルの歌声を利用して学習および認識を行った．

[*1] https://www.vector.co.jp/soft/win95/art/se127635.html

表 3.2　ソロ・ユニゾンの識別の正解率．ソロ・ユニゾンのラベルは正解ラベルを表す．

音響特徴量	楽曲	ソロ	ユニゾン	全体
Cosacorr スコア	AM11（closed-singer）	78.4%	63.5%	70.9%
	AM12（open-singer）	66.6%	87.9%	77.3%
i-vector	AM11（closed-singer）	81.8%	82.9%	82.4%
	AM12（open-singer）	81.9%	85.1%	83.5%
Cosacorr スコア，i-vector	AM11（closed-singer）	82.6%	82.9%	82.8%
	AM12（open-singer）	81.7%	95.3%	88.5%

フレーム周期は 10 ms とした．各フレームから 8 次の Cosacorr スコアを抽出し，その合計をそのフレームの Cosacorr スコアとした．連続する 100 フレームから Cosacorr スコアを計算し，その平均と分散を音響特徴量として用いた．すなわち，1 秒の歌声から 2 次元の音響特徴量を抽出した．

I-vector の抽出には MSR Identity Toolbox [96] を用いた．音響特徴量には，16 次の MFCC 及びその Δ，ΔΔ 特徴量を用いた．Universal background model（UBM）には 2048 混合の GMM を用いた．I-vector の次元数は 100 次元とした．Cosacorr スコアと同様に，1 秒の歌声から 100 次元の i-vector を抽出した．

本実験では，Cosacorr スコアと i-vector をそれぞれ単独で使用したシステムに加えて，Cosacorr スコアと i-vector の両者を音響特徴量として用いて認識を行うシステムを構築した．このシステムでは，入力される特徴量は 102 次元である．

識別モデルには SVM を用い，カーネル関数にはガウシアンカーネルを用いた．5 分割交差検証により SVM を学習した．

3.5.2　実験結果

実験結果を表 3.2 に示す．Cosacorr スコアを用いたシステムでは，少なくとも 70% の正解率での認識を達成した．したがって，Cosacorr スコアがユニゾンの検出に利用できることが示唆された．Open-singer 条件である楽曲 AM12 においても同等以上の認識が実現されており，歌唱者が既知か未知かに関わらず認識できることが確認された．I-vector を用いたシステムにおいても，80% 以上の正解率での識別を達成した．また，Cosacorr スコアと同様に，i-vector を用いても open-singer の楽曲 AM12 において同等の認識精度が確認された．さらに，Cosacorr スコアと i-vector を同時に用いたシステムでは，それぞれの音響特徴量を単独で用いた場合と比較して，高い正解率での識別を達成した．したがって，Cosacorr スコアと MFCC に由来する i-vector，ともにユニゾン検出に効果的であることが示された．

3.6　ダイアライゼーションシステム全体の評価

本節では，ダイアライゼーションシステム全体での提案法の有効性を評価するための実験を行う．本実験では，CD に収録されているパート割りのある 25 楽曲をそのまま評価に用いた．データセット中では複数人が歌唱するとき全歌唱者がユニゾンで歌唱する．

3.6.1　データセットの準備

学習および評価に用いるため，3 種類のデータセットを作成した．このデータセットは，学習セット，同時歌唱者数推定用学習セット，評価セットの 3 つに分かれる．

学習セット

3.5.1 節で述べたデータセットと同様の，各歌唱者がそれぞれ個別に歌唱した歌声を利用して構築した．すべての歌声は，伴奏音とミックスされた状態で収録されている．楽曲数は 55 で，歌唱者は各楽曲およそ 9 人であり，合計ののべトラック数は 500 である．楽曲の時間長の合計はおよそ 32 時間で，歌唱時間の合計はおよそ 24 時間である．楽曲および歌唱者の詳細は A.2.1 に示す．

本実験では学習のためパート割りのある楽曲を次の手順で生成した．まず，カラオケトラックを利用して 500 トラックから歌声のみを抽出した．同一楽曲の歌声を無作為にミックスすることで，パート割りのあるトラックを，各楽曲最大 10 トラックとして，526 トラック生成した．パート割りは 3 人の歌唱者によって行われるものとした．すなわち，これらのトラックには，誰も歌唱していない区間，1 人のみが歌唱している区間，2 人が歌唱している区間，3 人が歌唱している区間が入れ替わりながら存在している．53 楽曲のうち 51 楽曲を学習に，2 楽曲（19 トラック）を開発セットとして利用した．

カラオケトラックは，カラオケトラックが存在しない，あるいはボーカルとミキシングされているトラックとマスタリングが異なる，などの理由で利用できない場合がある．この場合，複数人がそれぞれ歌唱しているトラックから共通部分を抽出することで伴奏音を生成した．具体的には，まず同一楽曲の全トラックを相互相関関数を用いてサンプル単位で時間同期させる．次に，各トラックから振幅スペクトログラム $s_{n,f,t}$ を抽出する．ここで，n は歌唱者のインデックス，f および t はそれぞれ周波数・時間ビンのインデックスである．そして，カラオケトラックの振幅スペクトログラム $\hat{s}_{f,t}$ を次のように推定する．

$$\hat{s}_{f,t} = \min_{n} s_{n,f,t} \tag{3.7}$$

これによって得られたスペクトログラム $\hat{s}_{f,t}$ から Griffin–Lim 法 [45] によって位相復元することで，カラオケトラックを得る．位相復元の際には，もとの歌唱者が含まれる信号から抽出した位相スペクトログラムを初期値として利用することで，理想に近い位相をもつカラオケトラックを得た．

同時歌唱推定のための学習セット

同時歌唱者数推定部の学習にのみ用いるデータセットを構築した．このデータセットは，CD から抽出したパート割りのある楽曲と，その人数ラベルからなる．楽曲の時間長の合計はおよそ 113 分である．楽曲および歌唱者の詳細は A.2.2 に示す．

楽曲数は 25 曲で，このうち 19 楽曲を学習セットに，6 楽曲を開発セットに用いた．開発セットのうち 2 楽曲（CM20，CM21）の歌唱者は 19 楽曲に含まれ，4 楽曲（CM22–25）の歌唱者は 19 楽曲に含まれない．すなわち，前者は closed-singer 条件で，後者は open-singer 条件である．

人数のラベルは筆者の聴取により作成した．

評価セット

最終的なダイアライゼーションの性能を評価するためのデータセットを構築した．このデータセットは CD から抽出したパート割りのある楽曲と，各歌唱者の歌唱状態のラベルからなる．楽曲の時間長の合計はおよそ 112 分である．楽曲および歌唱者の詳細は A.2.3 に示す．

楽曲数は 25 曲であり，正解ラベルは筆者の聴取によって与えた．歌唱者数は楽曲によって 2 人から 6 人と異なる．歌唱者数は 57 名で，どの歌唱者も前節で述べた 2 データセットに含まれない．楽曲 DM12 および DM15 にはユニゾンではない区間が合計約 32 秒存在するが，全体の曲長と比較して短く，最終的な性能の議論には影響しない．

3.6.2　実験条件

音源分離

伴奏音とミックスされた音楽音響信号から歌声を抽出するために Spleeter [14] を用いた．事前学習され提供されている `2stems-16kHz` モデルを利用した．このモデルは，ミックスされた音楽音響信号を歌声と伴奏音の 2 つに分離するモデルである．このモデルの制約上，分離された音響信号はおよそ 16 kHz 以下に帯域制限されている．前述したように，MFCC や Cosacorr スコアなどの音響特徴量は，分離前後の音響信号からそれぞれ抽出し，結合して用いる．これは，分離後の歌声のみから抽出した音響特徴量のみでは十分な性能が得られないためである．

同時歌唱者数推定（voice activity and overlap detection）

同時歌唱者数推定には，BiLSTM ネットワークを用いた．このネットワーク構造を表 3.3 に示す．入力特徴量は，0 次項を含む 24 次 MFCC，パワー，8 次元の Cosacorr スコアである．フレーム周期は 100 ms で，21 フレームの音響特徴量を結合したものを入力とした．したがって，入力の次元数は総じて $(25+1+8)\times 2\times 21 = 1428$ 次元である．前述のように，このネットワークは，誰も歌唱していない，1 人が歌唱している，2 人以上が歌唱して

表 3.3　同時歌唱者数推定および target-singer VAD に用いた BiLSTM ネットワークのアーキテクチャ.

レイヤーの種類	出力次元数
入力層	1428 もしくは 849
Batch normalization	
BiLSTM	1024
Fully connected	1024
Batch normalization	
Fully connected	1024
Batch normalization	
BiLSTM	1024
Fully connected	1024
Batch normalization	
Fully connected	1024
Batch normalization	
Fully connected	1024
Batch normalization	
Fully connected	1024
Batch normalization	
Fully connected	3
Softmax	3

いる，の 3 クラス分類を行う．ネットワークは Adam を用いて学習し，学習率は 0.001 とした．バッチサイズは 19 とし，200 エポック学習した．本実験では開発セットを検証セットとして利用した.

Target-singer VAD

Target-singer VAD には，同時歌唱者数推定と同様に BiLSTM を用いた．このネットワークの構造を表 3.3 に示す．入力および出力の構造は図 3.10 に示すとおりである．入力特徴量は，0 次項を含む 24 次 MFCC，パワー，8 次元の Cosacorr スコア，対象歌唱者の歌唱者表現，また当該フレームにおける歌唱者表現と対象歌唱者の歌唱者表現のコサイン類似度である．フレーム周期は 100 ms で，11 フレームの音響特徴量を結合したものを入力とした．したがって，入力の次元数は総じて $(25 + 1 + 8) \times 2 \times 11 + 100 + 1 = 849$ 次元である．前述のように，このネットワークは，誰も歌唱していない，当該歌唱者のみ歌唱している，当該歌唱者ではない別の歌唱者が歌唱している，の 3 クラス分類を行う．ネットワークは Adam を用いて学習し，学習率は 0.001 とした．バッチサイズは 8 とし，20 エポック学習した．本実験では開発セットを検証セットとして利用した.

表 3.4　歌唱者表現に用いた ArcFace のアーキテクチャ.

レイヤーの種類	出力次元数
入力層	4050
Fully connected	2048
Fully connected	2048
Batch normalization	
Fully connected	2048
Batch normalization	
Fully connected	2048
Fully connected	100
Normalization	100
Margin addition（学習時）	100
Softmax	100

歌唱者表現

歌唱者表現の抽出には，ArcFace [92] をもとにしたネットワークアーキテクチャを用いた．このネットワーク構造を表 3.4 に示す．入力特徴量は，0 次項を含む 79 次の MFCC，パワー，8 次元の Cosacorr スコアである．フレーム周期は 20 ms で，25 フレームの音響特徴量を結合したものを入力とした．したがって，入力の次元数は総じて $(80+1)\times 2\times 25 = 4050$ 次元である．出力される歌唱者表現は 100 次元とした．また，ArcFace のマージンは 0.05 とした．式 (3.5) 中のスケーリングパラメータ s は学習可能なパラメータとした．ネットワークは Adam を用いて学習し，学習率は 0.0001 とした．バッチサイズは 32768 として，1572 エポック目のパラメータを採用した．学習データのうち 1% を検証セットとして用い，開発セットは用いなかった．

クラスタリング

クラスタリング法を用いた単一歌唱者の区間のクラスタリングには，自動チューニング可能なスペクトルクラスタリング法 [84] を採用し，オープンソースの実装[*2]を利用した．NME-SC と当該論文中で示されている，正規化最大 eigengap を利用した自動パラメータチューニング法を用い，自動でハイパーパラメータをチューニングした．クスタリングの後，HMM を用いた平滑化により，不適切な短時間での歌唱者の切り替わりや極端に短い歌唱区間などを排除した．この HMM の構造は図 3.9 に示す．図中で示されているハイパーパラメータの b は 0.9999 とした．また，出力確率は次のように構成した．

- 推定された歌唱者と状態に対応する歌唱者が等しい場合，$0.8 + 0.2/N$
- そうでない場合 $0.2/N$

[*2] https://github.com/tango4j/Auto-Tuning-Spectral-Clustering

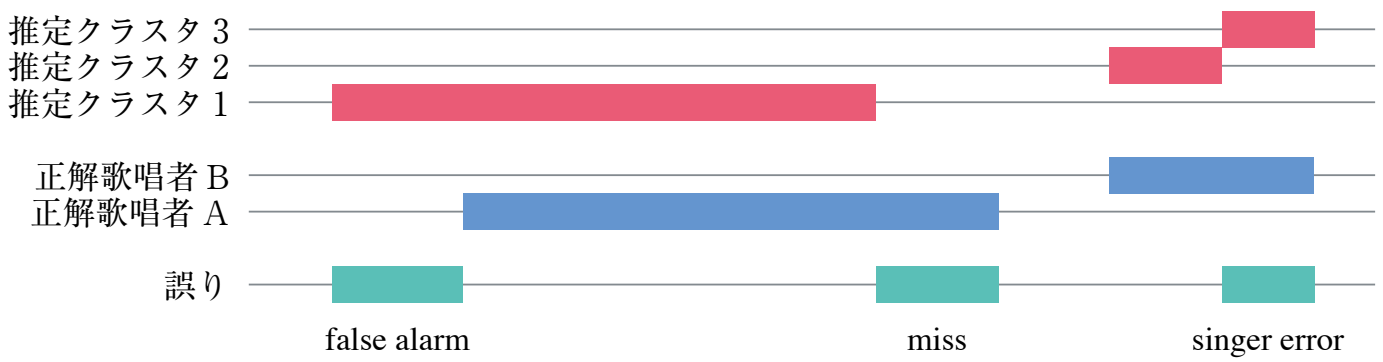


図 3.11　DER を計算する上での 3 種類の誤りの模式図．DER を計算する上では推定話者と正解話者との一対一対応を推定するが，図中では推定歌唱者 1 と正解歌唱者 A，推定歌唱者 2 と正解歌唱者 B がそれぞれ対応づけられている．推定歌唱者 3 に対応する正解歌唱者はいない．DER を計算する上での分母は，正解の歌唱区間の合計長，すなわち青い線の長さの合計にあたる．

これらのハイパーパラメータは開発セットを利用して決定した．

客観評価指標

ダイアライゼーション結果の客観評価指標には diarization error rate（DER）[97] を用いた．DER は，正解ラベルの全発話時間に対する，次に示す 3 種類の誤りの時間の割合で定義される．

- Singer error．異なる歌唱者としてラベル付けされた．
- False alarm．本来歌唱している人数よりも多い人数が歌唱しているとラベル付けされた．
- Miss．本来歌唱している人数よりも少ない人数が歌唱しているとラベル付けされた．

図 3.11 にこの 3 種類の誤りの模式図を示す．DER は次式で定義される．

$$\mathrm{DER} = \frac{\sum_{s=1}^{S} \tau_s \left(\max\left(N_s^{(\mathrm{ref})}, N_s^{(\mathrm{hyp})} \right) - N_s^{(\mathrm{correct})} \right)}{\sum_{s=1}^{S} \tau_s N_s^{(\mathrm{ref})}} \tag{3.8}$$

ここで，S はセグメントの数，τ_s は s 番目のセグメントの長さ，$N_s^{(\mathrm{ref})}$ および $N_s^{(\mathrm{hyp})}$ は s 番目の正解および推定セグメント中で歌唱している歌唱者の数，$N_s^{(\mathrm{correct})}$ はそのうち合致した歌唱者の数である．歌唱者ダイアライゼーションは正解歌唱者の情報を与えないため歌唱者認識を行わない．そのため，DER を計算する上は，あらかじめ推定歌唱者と正解歌唱者の一対一対応を得る必要がある．これにはハンガリアンアルゴリズムが用いられることが多いが，この対応付けには任意性があり，ハンガリアンアルゴリズムの影響で DER が高くなることを防ぐため，本論文ではもっとも DER が低くなるような一対一対応を用いて DER を計算する．DER は定義上 100% を超えることがある．DER の計算の際には，正解ラベルの歌唱者の切り替わり周辺の時刻の結果を無視することで，切り替わり時刻の違いによる誤りの検出を抑制する手法がとられる場合がある．本論文では，提案法の緻密な評価を行うため，このような手法は採用しない．

表 3.5　同時歌唱者数推定の精度の比較．同時歌唱者数推定用のデータセット中の開発セットに対する正解率を示している．

モデル	音響特徴量	正解率	
		CM20，CM21（Closed-singer）	CM22–25（Open-singer）
LSTM	MFCC，パワー	81.1%	68.4%
LSTM	MFCC，パワー，Cosacorr スコア	88.1%	76.6%
BiLSTM	MFCC，パワー	86.6%	75.1%
BiLSTM	MFCC，パワー，Cosacorr スコア	**89.9%**	**79.7%**
	ベースライン手法	52.5%	60.4%

3.6.3　同時歌唱者数推定

同時歌唱者数推定の性能を評価するため，複数のシステムを構築し，開発セットに対しての精度を比較した．本節では，BiLSTM 層を LSTM 層に置き換えたシステム，および Cosacorr スコアを用いないシステムを構築した．また，加えてベースライン手法における性能についても評価した．ベースライン手法は，TS-VAD を用いたダイアライゼーション手法と同様に，同時歌唱者数推定を明示的に行わず，target-singer VAD によって全歌唱区間の認識を行う．このダイアグラムを図 3.12b に示す．ベースライン手法においては，開発セットに対してすべてのダイアライゼーション手順を行い，その上で同時歌唱者数推定の観点での精度を得た．

実験の結果を表 3.5 に示す．BiLSTM ネットワークおよび Cosacorr スコアを導入することで，closed-singer の楽曲で 89.9%，open-singer の楽曲で 79.7% の正解率での認識を実現した．Cosacorr スコアを用いないシステムと比較して，Cosacorr スコアを用いたシステムは高い正解率での認識を実現しており，Cosacorr スコアの有効性が示された．また，LSTM ネットワークと BiLSTM ネットワークを比較すると後者が高い正解率を示した．本論文で扱う歌唱者ダイアライゼーションにおいては，オンラインで認識する必要がないため，BiLSTM ネットワークを用いたシステムを採用した．

また，ベースライン手法の認識精度は，同時歌唱者数推定を独立で行う場合と比較して大きく劣っていた．これは，target-singer VAD による認識精度の低さを示唆している．同時歌唱者数推定を独立して行うことで，miss や false alarm の誤りを抑制できると推察される．

3.6.4　歌唱者表現

ArcFace により抽出された歌唱者表現の有効性を確認するため，ArcFace に加えて x-vector [89] を用いて歌唱者表現を抽出し，これを利用してシステムを構築してその性能を比較した．X-vector の抽出には，ArcFace と同様に音源分離前後の音声から抽出した 0 次項を含む 79 次 MFCC およびパワーを用いた．フレーム周期は 10 ms とし，連続する 51 フレームから歌唱者表現を抽出した．

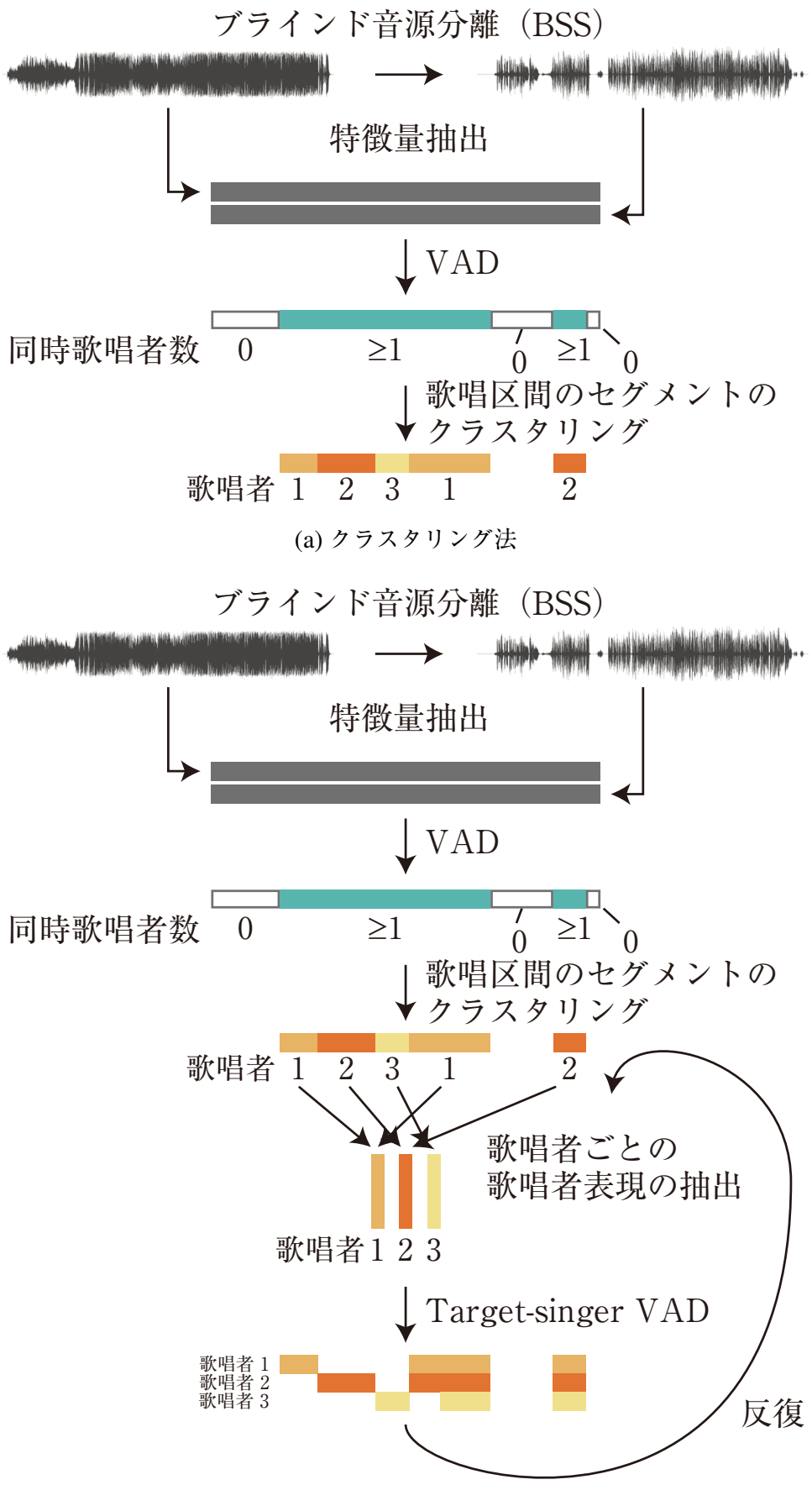


図 3.12　実験において比較したクラスタリング法およびベースライン手法のダイアグラム．図 3.5 に示した提案法のダイアグラムと比較せよ．

表 3.6　異なる歌唱者表現によるダイアライゼーション性能の比較.

歌唱者表現	歌唱者の人数情報の有無	
	あり	なし
ArcFace	**36.7%**	**48.7%**
x-vector	44.8%	59.3%

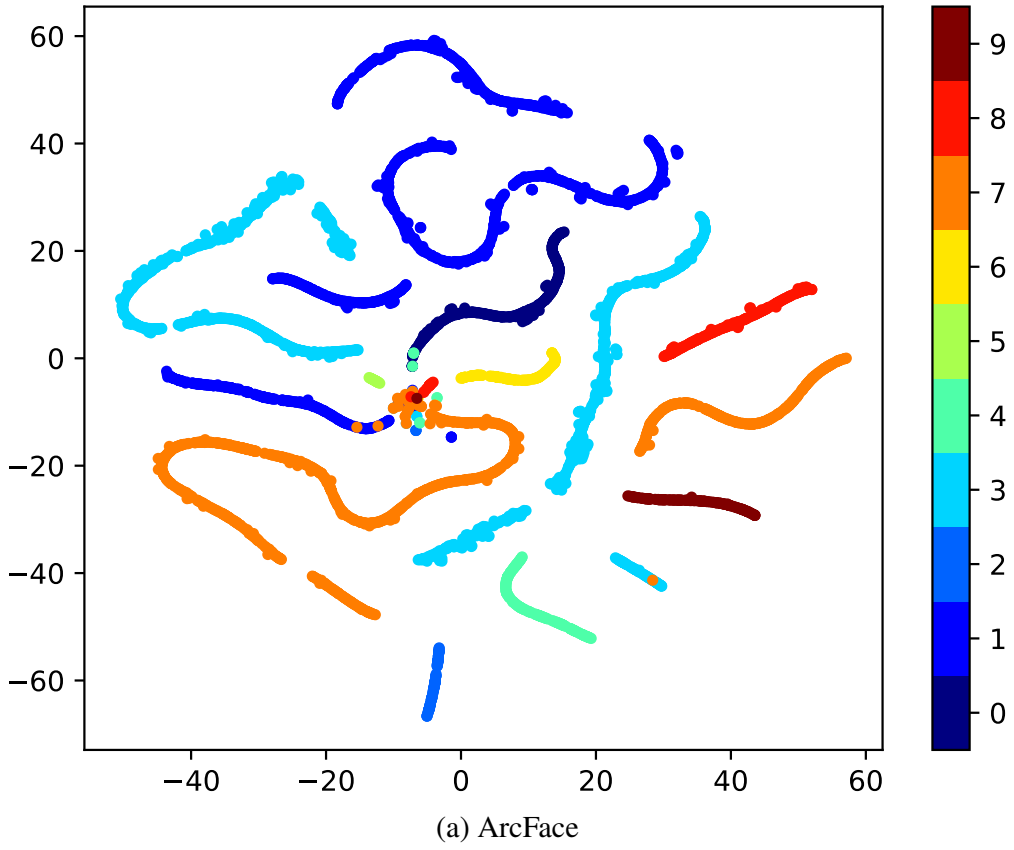
まず，抽出した歌唱者表現を *t*-SNE [98] により可視化した．これを図 3.13 に示す．ArcFace を用いて抽出した歌唱者表現は，x-vector と比較してクラスタを形成していることが観察される．一方，x-vector による歌唱者表現は散逸している．

次に，ダイアライゼーションにおける性能を比較した．ここでは，開発セットにおける単一歌唱者の区間に対してクラスタリングによりダイアライゼーションを行い，その DER を比較した．結果を表 3.6 に示す．人数の情報の有無によらず，ArcFace を用いて抽出した歌唱者表現を利用することで，DER が 10 ポイント程度低減された．この結果は ArcFace を用いて抽出された歌唱者表現がより分離性能が高く，ダイアライゼーションにおいても有効であることを示唆している．

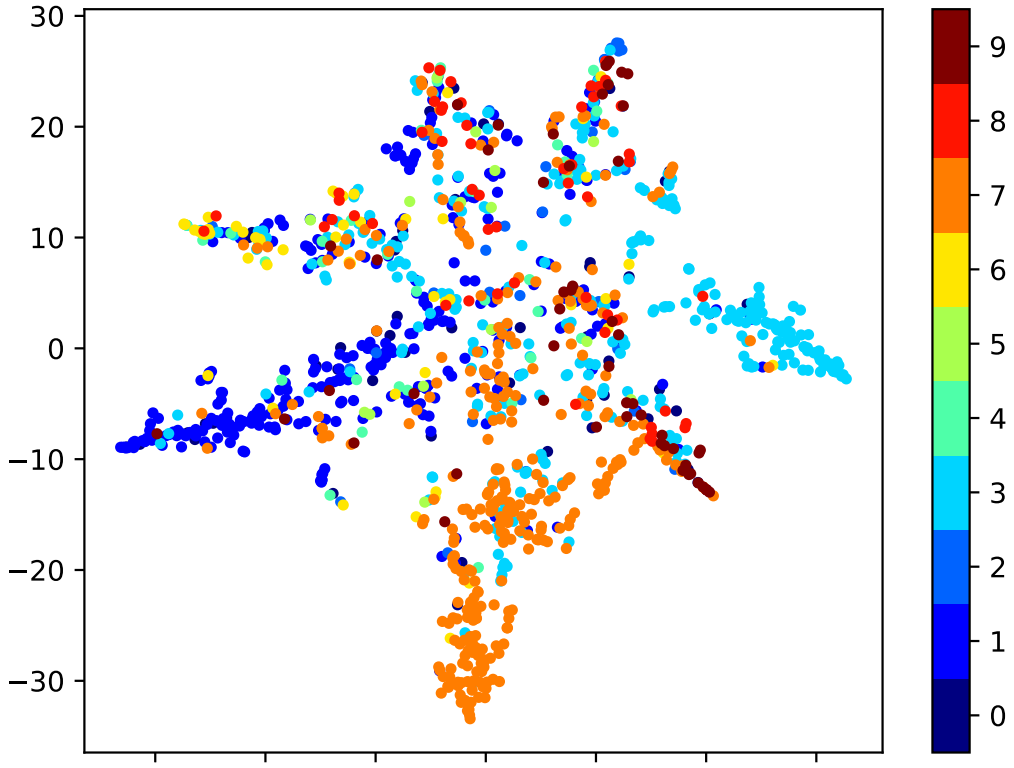
3.6.5　Target-singer VAD

Target-singer VAD の性能を評価するため，クラスタリングの正解を与えた状態での性能を評価した．すなわち，同時歌唱者数推定およびクラスタリングの結果として正解を用い，その結果をもとに抽出した歌唱者表現を利用して，複数人が歌唱する区間に対して歌唱者ダイアライゼーションを行った．開発セットに対する DER は 24.4% で，評価セットに対する DER は 14.3% であった．なお，クラスタリングの結果に対して正解を与えているため，DER を計算する際の一対一対応はあらかじめ正解が与えられている．

また，target-singer VAD と比較して，TS-VAD の評価についても行った．ここでは，任意人数の認識を行うことのできる手法 [90] を採用し，TS-VAD の認識可能な最大人数を，評価セットに現れる最大の同時歌唱人数である 6 人とした．TS-VAD の学習においては，学習セットをそのまま用いる場合，歌唱区間・非歌唱区間の割合が大きく非歌唱区間に偏るため，適切に学習を行うことができない．そこで，歌唱区間と非歌唱区間の割合がおよそ均等になるようなデータセットを合成して，これを用いて学習を行った．ネットワークのアーキテクチャは TS-VAD の提案論文 [62] と同様のものとした．開発セットに対する DER は 36.3% で，target-singer VAD と比較して高い値であった．これは，学習に用いるデータセットが無作為にミックスされたものであり，実データにみられるパート割りとは異なるパターンであったためであると考えられる．TS-VAD は同時に複数の歌唱者の歌唱状態を推定するため，歌唱者間の混同を抑制できるモデルではあるものの，パート割りの事前知識を BiLSTM によって学習するため，本実験のデータセット内では高い性能を示さなかったと考えられる．



(a) ArcFace



(b) x-vector

図 3.13　*t*-SNE [98] による歌唱者表現の可視化．色は歌唱者を表す．

3.6.6　ダイアライゼーション精度の評価

提案法の性能を評価するため，次の 3 つのシステムを比較した．

- 複数人歌唱を考慮しないクラスタリング法
- ベースライン手法
- 提案法

図 3.12 にクラスタリング法およびベースライン手法のダイアグラムを，図 3.5 に提案法のダイアグラムを示す．クラスタリング法においては，同時歌唱者数推定により 1 人以上と認識された区間に対して，歌唱者表現をクラスタリングすることによりダイアライゼーションを行った．ベースライン手法では，TS-VAD を target-singer VAD に置換する点を除いて，TS-VAD を用いたダイアライゼーション手法 [62] と同一の手順によってダイアライゼーションを行った．すなわち，同時歌唱者数推定により 1 人以上と認識された区間に対してクラスタリング法によりダイアライゼーションを行い，その結果をもとに歌唱者表現を抽出し，それを用いて target-singer VAD を行った．歌唱者表現の抽出と target-singer VAD は 3 回繰り返した．音響特徴量や歌唱者表現などの条件は同一とした．歌唱者の人数を既知とした場合と未知とした場合それぞれにおいて性能を評価した．

結果を表 3.7 に示す．提案法は，他の 2 手法と比較して低い DER を示した．したがって，あらかじめ同時歌唱者数推定を行う提案法の有効性が示唆された．一方，クラスタリング法の DER は非常に高い DER を示した．これは，同時歌唱が認識されないため，同時歌唱の区間で多くの miss が発生したためであると推察される．また，歌唱者の人数の情報により大きな DER の差がみられた．これは，歌唱者の人数が既知の場合，false alarm や miss が発生しにくくなるためであると考えられる．

3.7　考察

3.6.6 節で述べた結果より，提案法はベースライン手法と比較して低い DER での認識を実現した．表 3.8 に評価に用いたすべての楽曲の結果を示す．25 曲中 3 曲ではベースライン手法が優れていたが，それ以外の 22 楽曲では提案法がより高いダイアライゼーション性能を示した．

典型的なダイアライゼーション結果の例として，図 3.14 に楽曲 DM4（Tomorrow Program）のダイアライゼーションの結果を示す．ベースライン手法では，同時歌唱者数推定を明示的に行わないため，単一歌唱者の区間での false alarm がとくに多くみられる．表 3.5 にも示したように，ベースライン手法では同時歌唱者数の観点での誤りが提案法と比較して多く，これにより false alarm や miss が多く発生し，DER が高まったと考えられる．一方，提案法は同時歌唱者数推定を明示的に行うため，人数の誤りに起因する誤りを抑制できている．したがって，この結果は同時歌唱者数推定を明示的に行う提案法の有効性を示唆している．また，とくに同時歌唱者数推定では Cosacorr スコアが有効であり，Cosacorr スコアの導入が全体の DER の低下に貢献していることが確認された．

提案法と比較してベースライン手法が高い精度でダイアライゼーションを行うことができた例として，図 3.15

表 3.7　3 手法によりダイアライゼーションを行った結果の DER．歌唱者の人数を既知とした場合および未知とした場合それぞれについて評価した．

(a) 歌唱者の人数を既知とした場合

手法	開発セット				評価セット			
	DER	singer error	false alarm	miss	DER	singer error	false alarm	miss
クラスタリング法	149.0%	12.6%	1.2%	135.2%	191.2%	31.7%	2.7%	156.7%
ベースライン手法	77.7%	3.1%	15.9%	58.6%	72.9%	8.0%	20.3%	44.6%
提案法	**55.9%**	7.9%	4.0%	44.0%	**52.9%**	10.6%	26.5%	15.8%

(b) 歌唱者の人数を未知とした場合

手法	開発セット				評価セット			
	DER	singer error	false alarm	miss	DER	singer error	false alarm	miss
クラスタリング法	153.5%	17.2%	1.2%	135.2%	191.9%	32.3%	2.7%	156.9%
ベースライン手法	92.0%	4.6%	22.5%	64.9%	129.5%	12.1%	25.1%	92.3%
提案法	**69.3%**	10.2%	3.3%	55.8%	**79.3%**	13.2%	24.8%	41.2%

に楽曲 DM18（囚われの TeaTime）のダイアライゼーション結果を示す．提案法では複数人が同時に歌唱している区間での miss が多くみられる．提案法での同時歌唱者数推定は 2 人以上の推定はせず，target-singer VAD によって解決している．そのため，歌唱者表現の生成すなわち初期クラスタリングの性能が不十分なことにより，target-singer VAD での誤りが多くみられたと考えられる．一方ベースライン手法では，楽曲全体から歌唱者表現を抽出するため，複数人が歌唱している際に全員が歌唱していると判断しやすくなると推察され，これにより全員が歌唱している区間での miss を抑制できたと考察される．

提案法は，オーバーラップ区間においてもベースライン手法と比較して高い精度での推定を実現した．表 3.9 に，推定された歌唱人数に関する混同行列を示す．提案法は，歌唱者の人数について，複数人歌唱区間においてより少ない誤りでの推定を実現していることが示された．表 3.7 の結果においても，提案法の結果はベースライン手法と比較して miss が少ない．これらの結果は，抽出される歌唱者表現の品質の差に起因するものと考えられる．提案法は，最初に同時歌唱者数推定を行うことで，1 人が歌唱している区間のみから歌唱者表現を抽出するため，すべての歌唱区間から歌唱者表現を抽出するベースライン手法と比較して，高い品質の歌唱者表現を得られると考えられる．ベースライン手法では，歌唱者表現の抽出と target-singer VAD を反復することで歌唱者表現の品質を高めるが，そのようなアプローチと比較して提案法のアプローチが効果的であることが示された．

最終的な DER は，歌唱者数を与えた場合でも平均で 50% 以上であり，まだ十分な性能が得られているとは言えない．この原因の 1 つとして，クラスタリング法による単一歌唱者の区間でのクラスタリングの性能が十分得られていない点がある．クラスタリングの結果は，各歌唱者の歌唱者表現を得るためにも用いられるため，この手順は最終的なダイアライゼーションの品質を大きく左右する．もし仮にこのクラスタリングに誤りがなければ，

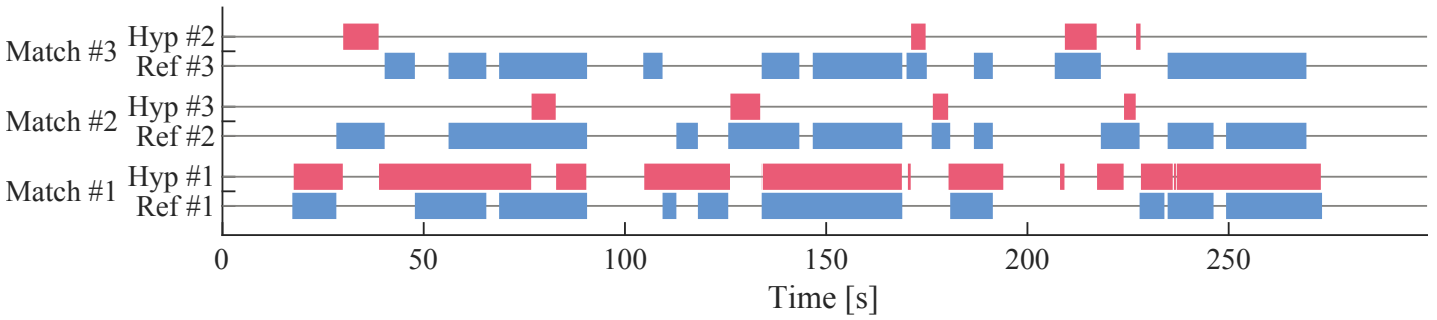
表 3.8　評価に用いた全楽曲に対するダイアライゼーション結果の DER.

楽曲	楽曲名	クラスタリング法	ベースライン手法	提案法
DM1	アルストロメリア	119.5%	76.0%	**53.8%**
DM2	ハピリリ	128.7%	53.0%	**42.2%**
DM3	ZETTAI×BREAK!! トゥインクルリズム	144.6%	67.5%	**64.4%**
DM4	Tomorrow Program	114.3%	35.9%	**23.2%**
DM5	Melty Fantasia	138.3%	36.9%	**30.3%**
DM6	I.D 〜EScape from Utopia〜	144.6%	74.1%	**50.1%**
DM7	花ざかり Weekend ✤	181.1%	56.9%	**35.5%**
DM8	RED ZONE	212.4%	53.7%	**33.5%**
DM9	咲くは浮世の君花火	279.7%	29.1%	**24.1%**
DM10	BORN ON DREAM! 〜HANABI ☆ NIGHT〜	290.2%	30.2%	**29.1%**
DM11	だってあなたはプリンセス	100.0%	61.3%	**37.9%**
DM12	ミラージュ・ミラー	74.4%	137.4%	**27.1%**
DM13	月曜日のクリームソーダ	161.7%	183.8%	**58.0%**
DM14	I did+I will	162.0%	**50.0%**	62.4%
DM15	ピコピコ IIKO! インベーダー	217.0%	28.7%	**26.7%**
DM16	Get lol! Get lol! SONG	158.7%	56.8%	**52.1%**
DM17	dans l'obscurité	210.0%	90.0%	**83.2%**
DM18	囚われの TeaTime	230.5%	**46.2%**	68.3%
DM19	Tulip	236.1%	**51.6%**	81.7%
DM20	ハイファイ☆デイズ	299.7%	152.4%	**147.3%**
DM21	イリュージョニスタ!	243.1%	68.8%	**52.7%**
DM22	Yes! Party Time!!	258.7%	125.9%	**46.4%**
DM23	shabon song	176.4%	122.5%	**75.2%**
DM24	ガールズ・イン・ザ・フロンティア	224.1%	70.2%	**70.0%**
DM25	夢で夜空を照らしたい	273.4%	62.7%	**46.3%**
	平均	191.2%	72.9%	**52.9%**

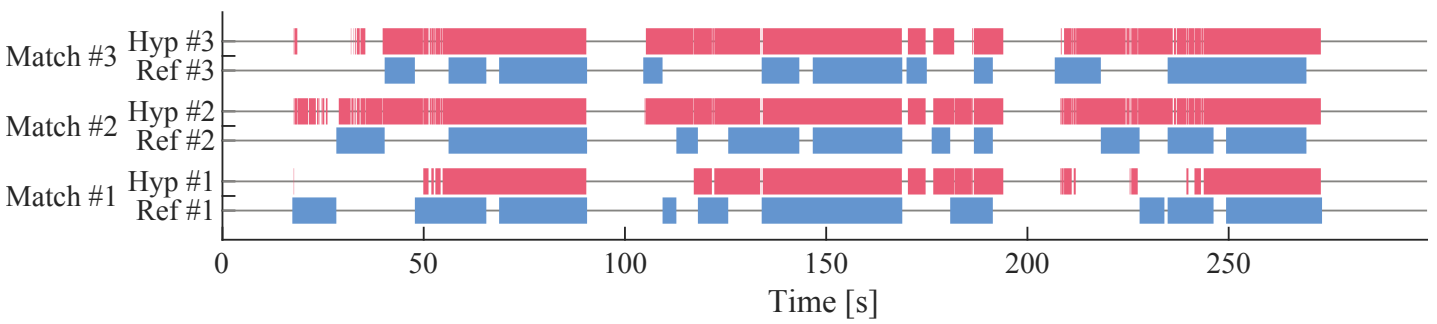
3.6.5 節に示したように DER を大きく減少させることができると考えられる．したがって，歌唱者表現およびクラスタリング法の改善が，さらに高品質なダイアライゼーションの実現には不可欠である．

3.8　本章のまとめ

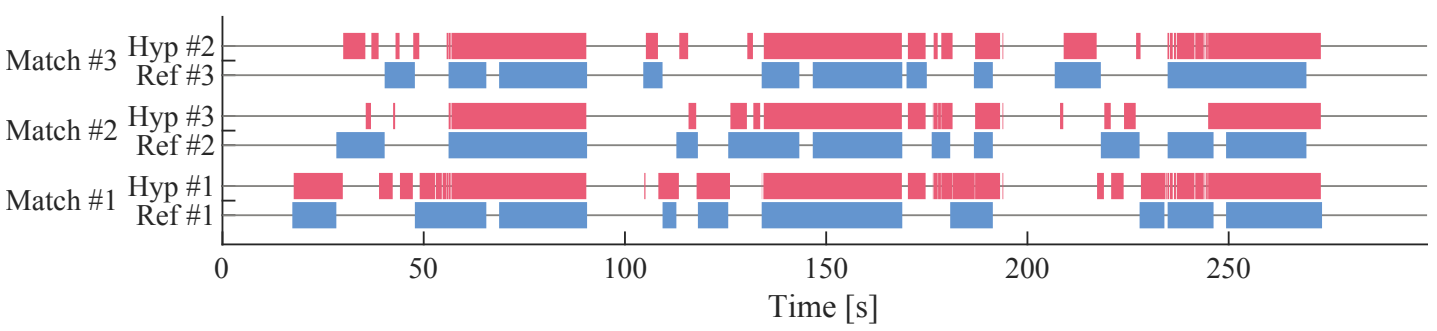
本章では，複数人が歌唱している楽曲のうち，とくにパート割りのある楽曲から，誰がいつ歌唱しているかを推定する歌唱者ダイアライゼーション技術について扱った．会話音声における同様の技術である話者ダイアライゼーションと比較して，音響的な特徴や同時に歌唱している人数の違いがあり，歌唱者ダイアライゼーションに対して話者ダイアライゼーション手法をそのまま適用することは困難であると考えられる．そこで本章では，同時歌



(a) クラスタリングのみを利用したベースライン手法（DER=114.3%）



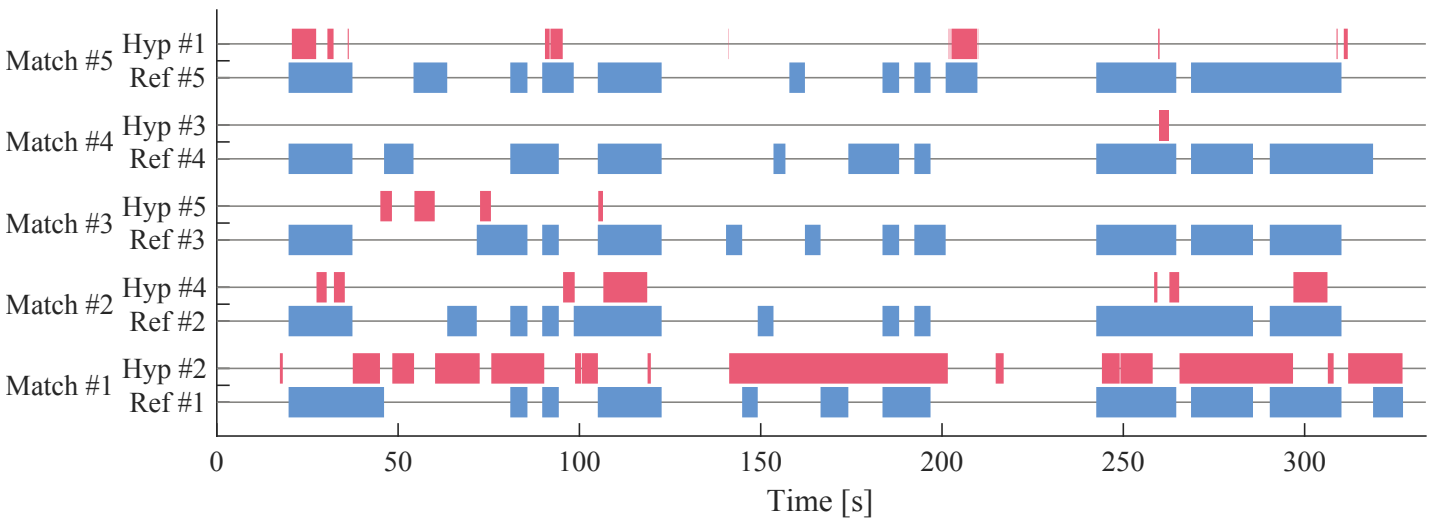
(b) ベースライン手法（DER=35.9%）



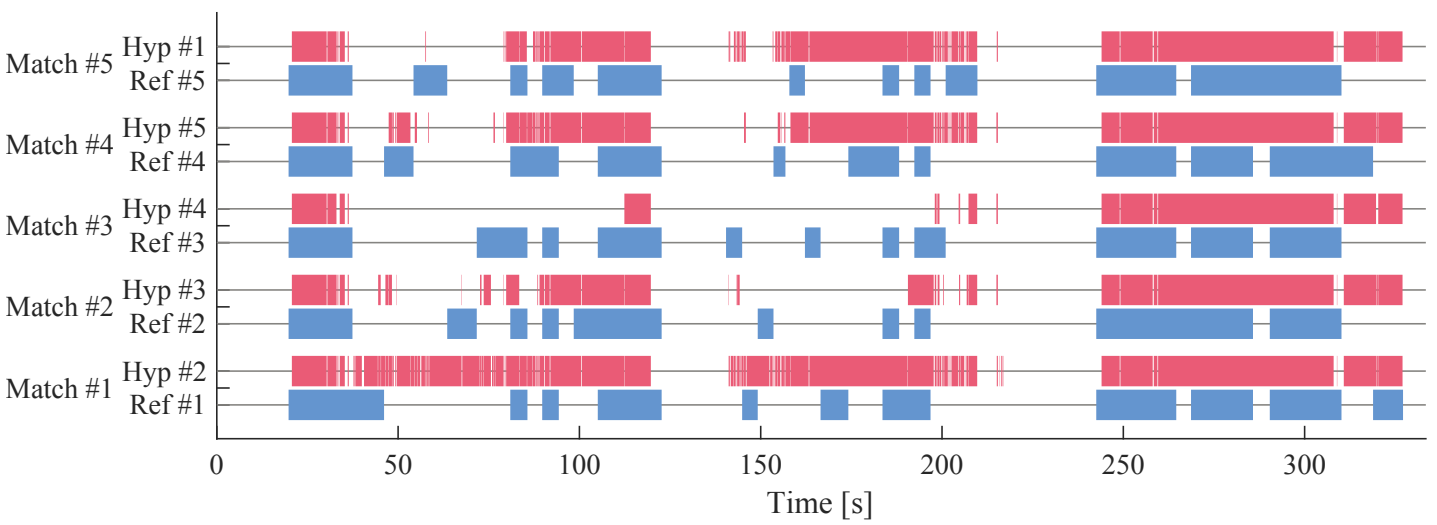
(c) 提案法（DER=23.2%）

図 3.14　楽曲 DM4『Tomorrow Program』のダイアライゼーション結果．青線（Ref）は正解を，赤線（Hyp）は推定結果を表す．DER を計算する際に得られた一対一対応を Match として示している．どちらの結果も歌唱者数の情報を与えてダイアライゼーションを行ったものである．

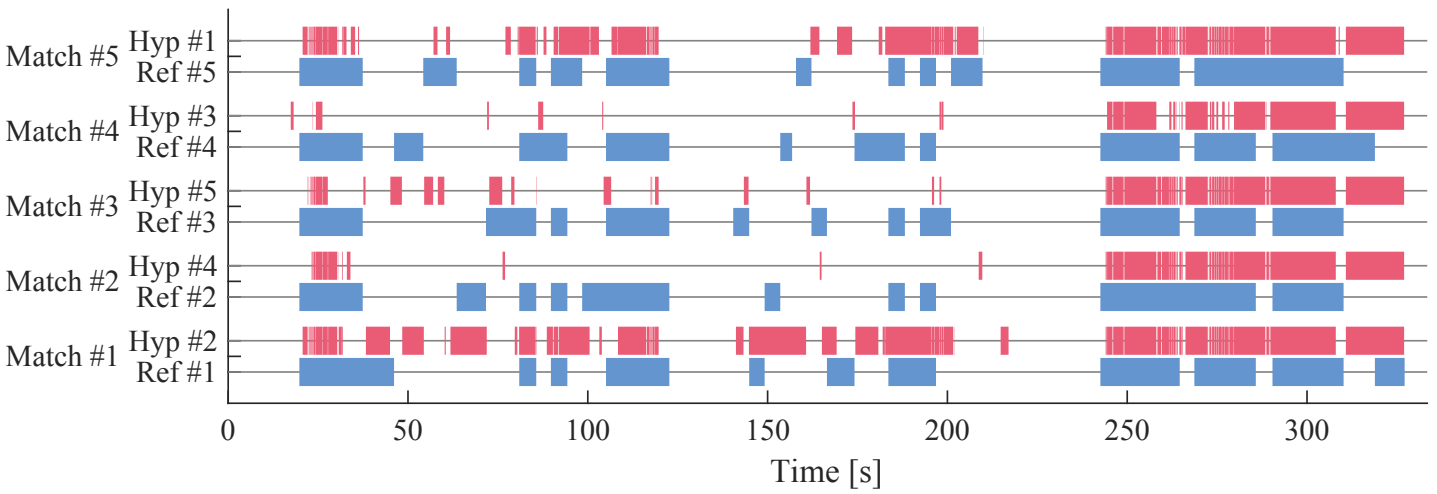
唱者数推定を明示的に行う手法を提案し，これを前処理として用いることで，ダイアライゼーションにおける誤りを抑制するような手法を提案した．高い精度での同時歌唱者数推定のため，Cosacorr スコアとよばれる音響信号の自己相関関数の形状を評価する音響特徴量を導入した．伴奏音がミックスされた歌声を扱うため，本論文では Spleeter [14] を用いた音源分離を行い，抽出された歌声を活用して認識を行った．また，歌唱者の埋め込み表現，すなわち歌唱者表現を短時間の歌声から抽出するため，ArcFace [92] とよばれる顔認識に用いられる埋め込み表現抽出法を採用し，ダイアライゼーション精度の向上を実現した．最終的に，既存の話者ダイアライゼーション手法をもとにしたベースライン手法と比較して，低い誤り率での認識を実現した．



(a) クラスタリングのみを利用したベースライン手法（DER=230.5%）



(b) ベースライン手法（DER=46.2%）



(c) 提案法（DER=68.3%）

図 3.15　楽曲 DM18『囚われの TeaTime』のダイアライゼーション結果．青線（Ref）は正解を，赤線（Hyp）は推定結果を表す．DER を計算する際に得られた一対一対応を Match として示している．どちらの結果も歌唱者数の情報を与えてダイアライゼーションを行ったものである．

表 3.9　人数推定における混同行列．単位は秒である．

(a) ベースライン手法（正解率 49.0%）

True \ Predicted	0	1	2	3	4	5	6
0	1712	54	27	11	6	12	
1	645	551	549	254	166	198	
2	51	150	126	21	40	10	
3	41	45	277	238	22	8	
4	19	58	36	167	228		
5	87	41	164	75	105	447	
6	6			4	61	29	

True number of singers

Predicted number of singers

(b) 提案法（正解率 68.5%）

True \ Predicted	0	1	2	3	4	5	6
0	1699	61	15	11	23	12	
1	182	1558	191	103	166	156	7
2	33	84	188	9	71	11	2
3	35	62	190	304	11	21	8
4	30	17	28	73	360		
5	103	55	152	50	92	467	
6	12		2	2	8	34	42

True number of singers

Predicted number of singers

第 4 章

非負値行列因子分解を用いたノンパラレル声質変換

4.1　はじめに

声質変換は，音声の言語情報を保持したまま，それ以外の特定の情報を変換する技術である [99]．たとえば，音声をあたかも別の話者が発話したかのように変換する話者変換や，音声にこめられた感情を変換する感情変換などが，声質変換技術に含まれる．本論文では，話者変換を指して声質変換とよぶ．声質変換においては，入力される音声の話者を入力話者，出力される目標となる声質の話者を出力話者という．声質変換技術は，エンターテインメント分野での活用やテキスト音声合成（text-to-speech; TTS）への適用など種々の応用可能性があり，第 5 章では歌声の声質変換を利用したアプリケーションを提案する．

声質変換システムは，学習と変換の 2 つの手順に分けられる．学習時には，入出力話者が発話した音声から音響特徴量を抽出し，音響特徴量間の変換関数をなんらかの方法で学習する．そして変換時には，入力話者の任意の発話から音響特徴量を抽出し，学習した変換関数を用いて変換し，変換された音響特徴量を用いて合成することで，変換後の音声を得る．古典的な声質変換システムでは，入力話者と出力話者がそれぞれ同一の言語内容を発声したパラレルデータを用いて変換関数を学習する．パラレルデータを用いた声質変換システムの模式図を図 4.1 に示す．学習時には，まずパラレルデータからメルケプストラム係数などの音響特徴量を抽出する．次に，動的時間伸縮（dynamic time warping; DTW）などの手法を用いて時間的なアラインメント（対応付け）を得る．これは，たとえパラレルデータであっても，わずかな発話スタイルの違いにより厳密に発話長が一致することはないためである．学習には音響特徴量間の一対一対応を得る必要があるため，時間的なアラインメントが必要である．そして，一対一対応が得られたパラレルデータを用いて変換モデルを学習する．変換モデルには，GMM [99, 100, 101]，制限付きボルツマンマシン（restricted Boltzmann machine; RBM）[102, 103]，ニューラルネットワーク [104, 105]，RNN [106]，非負値行列因子分解（non-negative matrix factorization; NMF）[107, 108] などが用いられる．古典的な手法では，学習にパラレルデータが必要なため，パラレルデータを収録できない状況や，多数の文を入出力話者に発話させられない状況に用いることができない．また，DTW によってパラレルデータのアラインメントを得る際の誤りによって，最終的な変換モデルの品質が低下することが指摘されている [109][*1]．そこで，パラレルコーパスを用いずに変換モデルを学習するノンパラレル声質変換法が多数提案されている．

ノンパラレル声質変換法は，外部データの有無によって大きく 2 つに大別できる．第 1 の手法は，外部データを用いて言語情報と話者情報を分離し，話者情報のみを変換する手法である．GMM の適応を利用する手法では，あらかじめなんらかのパラレルデータを利用して GMM を学習し，それを目標の入出力話者に適応することでノンパラレル声質変換を実現する [111]．固有声（eigenvoice）を用いる手法では，GMM の平均ベクトルを結合したスーパーベクトルを話者の特徴と見なし，スーパーベクトルに対して主成分分析（principal component analysis; PCA）を行うことで，話者を表現する低次元のベクトルを得る [112, 113]．目標の出力話者に対してこの話者を表現するベクトルを推定することで，ノンパラレルな条件でも GMM による声質変換モデルを構築できる．GMM

[*1] なお，sequence-to-sequence モデルを用いることで，DTW を注意機構（attention mechanism）によって置き換える手法も提案されている [110].

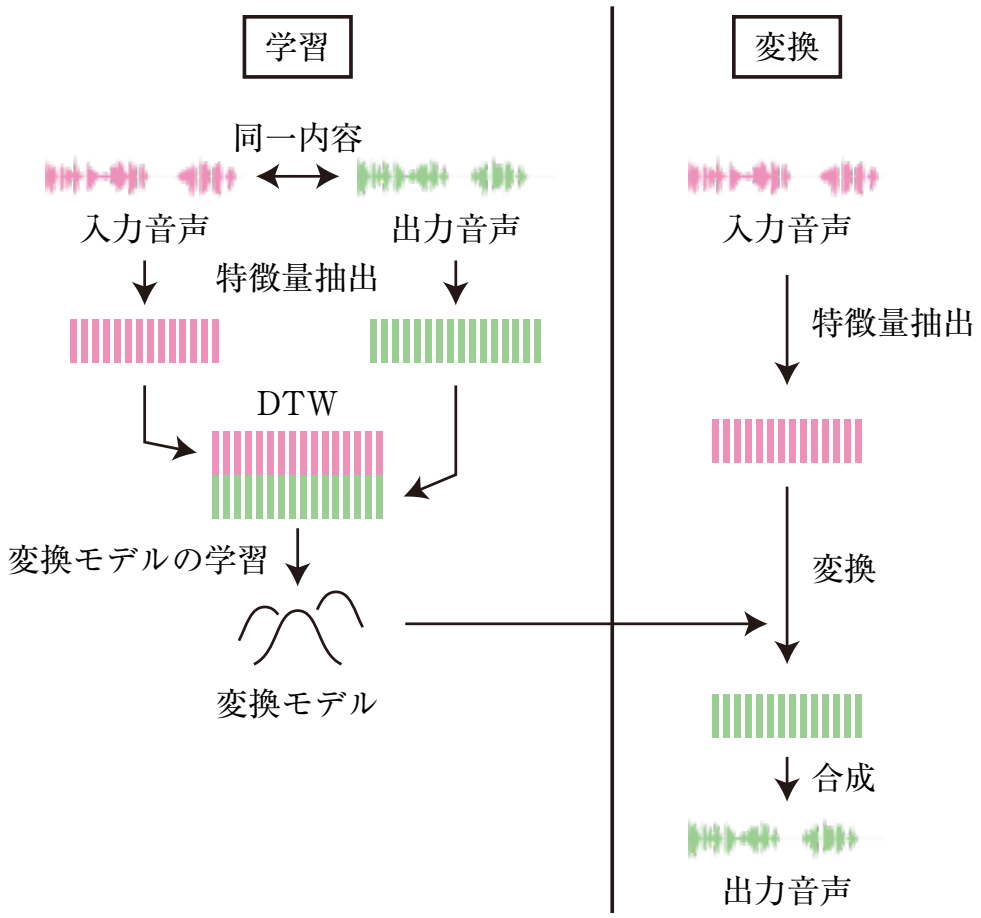


図 4.1　パラレルデータを用いる古典的な声質変換手法の模式図．大きく学習と変換の 2 手順に分かれる．

を用いた手法では，ともに GMM のパラメータが話者表現，GMM の事後確率が言語情報であると捉えることができる．話者認識で用いられる i-vector [85] も，声質変換における話者表現として利用されている [114]．この手法では，入力された音声が目標の話者表現を保持するように修正する．一方，言語情報を明示的に利用して，話者情報と言語情報を分離する手法も提案されている．Sequence-to-sequence ネットワークを用いる手法は，テキスト情報を学習時に用いることで話者情報と言語情報を分離し，話者情報を置き換えることを可能にしている [115]．自動音声認識（automatic speech recognition; ASR）のような仕組みを用いて，各時刻にどの音素が発音されているかを確率的に表した音素事後確率（phonetic posteriogram）を抽出し，これを用いて TTS のようなアプローチで声質変換を実現するシステムも提案されている [116]．ASR 部分を話者非依存にしたり，TTS に話者表現を入力できるようにしたりすることで，複数話者声質変換も実現可能である．書き起こしを用いる手法は，入出力話者以外の話者を必要としない手法も構築可能であるが，一方言語情報が外部データとして用いられるため第 1 の手法に分類する．これらの外部データを用いる手法は，入出力話者の音声に関してデータ量の制約が少ない．一方，多人数の音声コーパスや書き起こしなどの外部データを必要とするため，外部データの用意が難しく，また外部データの品質や量によって最終的な変換品質が決まるなど，システム全体を高品質に構築することが難しい．

第 2 の手法は，外部データを用いずに入出力間の変換関数を学習する手法である．CycleGAN-VC は，入力話者から出力話者への変換だけでなく，出力話者から入力話者への変換を同時に学習する手法である [117, 118, 119]．

変換された発話が入出力話者の発話らしいかを評価する adversarial loss，両変換を適用したときに元の発話に戻るかを評価する cycle-consistency loss，言語情報が一貫しているかを評価する identity-mapping loss の 3 種の損失関数を組み合わせた損失関数を利用して学習される．変分自己符号化器（variational autoencoder; VAE）を利用した手法では，話者を表す one-hot ベクトルで VAE を条件付けすることで，言語情報を保持する潜在表現を得て，声質変換を実現する [120]．CycleGAN-VC や VAE を用いた手法は複数話者声質変換にも拡張されているが，それぞれの手法の考え方は 2 話者間変換の場合と同様である [121, 122]．INCA アルゴリズムは，ノンパラレルな入出力話者の発話からパラレルデータを生成する手法である [123]．INCA アルゴリズムはパラレルデータの生成アルゴリズムであるため，別途古典的なパラレルデータを利用した声質変換法を利用することで声質変換システムを実現する．これらの手法は，入出力話者の発話以外の学習音声を必要としないが，一方で第 1 の手法と比較してより多くの入出力話者の発話が必要である．

第 2 の手法が多くの学習音声を必要とする要因には，入出力話者を同等に扱っている点が挙げられる．たとえば CycleGAN-VC は，入力話者から出力話者への変換と，出力話者から入力話者への変換を同時に同等に学習する．INCA アルゴリズムでも同様に，入力話者の発話から出力話者の発話への対応と，その逆の対応を同時に得る．VAE を用いた手法では，入出力話者の発話を同一のアーキテクチャを用いて潜在表現を得る．そもそも声質変換システムは，入出力話者間での変換と出力話者の生成器を同時に構築しさえすればよい [124]．すなわち，入力話者に関しては，言語的整合性を保って出力話者の生成器を駆動できるだけの十分な音響モデルがありさえすればよい．第 1 の手法では外部データを利用して言語的整合性を保っているが，第 2 の手法ではそれらが実現できていない．したがって，言語的整合性と高品質な出力話者の生成器をそれぞれ独立に得る手法を構築することで，入力話者に関してはより少ない量の発話でノンパラレルな声質変換システムを構築できると考えられる．ただし出力話者に関しては，言語情報から音声を合成する都合上，入力話者と比較してより多くの発話が必要である．

より少ない入力話者の発話で学習できる手法として，本論文では新たに Soft INCA アルゴリズムとよばれる手法を提案する．この手法は，NMF の時不変な特徴（基底）と時変な特徴（生起状態）に分解する性質を利用したものである．音声においては，時不変な特徴は話者情報に，時変な特徴は言語情報にあたると解釈できるため，NMF は言語情報と話者情報を音声のみから分解することが可能である．これを利用して，音声の話者情報のみを置換する手法を構築する．また，NMF は基底から分解対象の特徴量へのアラインメントを得る手法であると解釈できる．このアラインメントは一対一ではなく，注意機構のように非負の連続的なアラインメントである．このアラインメントは，ノンパラレルな音響特徴量間のアラインメントを得る INCA アルゴリズムと同様に行うことができる．すなわち，Soft INCA アルゴリズムは，NMF はアラインメントであるという解釈を利用して INCA アルゴリズムを拡張した手法である．Soft INCA アルゴリズムはこれらの NMF の特徴を利用することで，出力話者の生成器を高品質で構築しつつ，言語的な整合性の保たれた変換器を得ることができ，少量の発話でも自然性の高い変換を実現できる．これまで述べたノンパラレル声質変換法の中での本手法の立ち位置を図 4.2 に示す．本手法は，背景知識を必要とせず，また少量の入力話者の発話のみを必要とする手法である．

本章では，まず提案法の基礎である NMF およびそれを利用したパラレル声質変換法，また INCA アルゴリズム

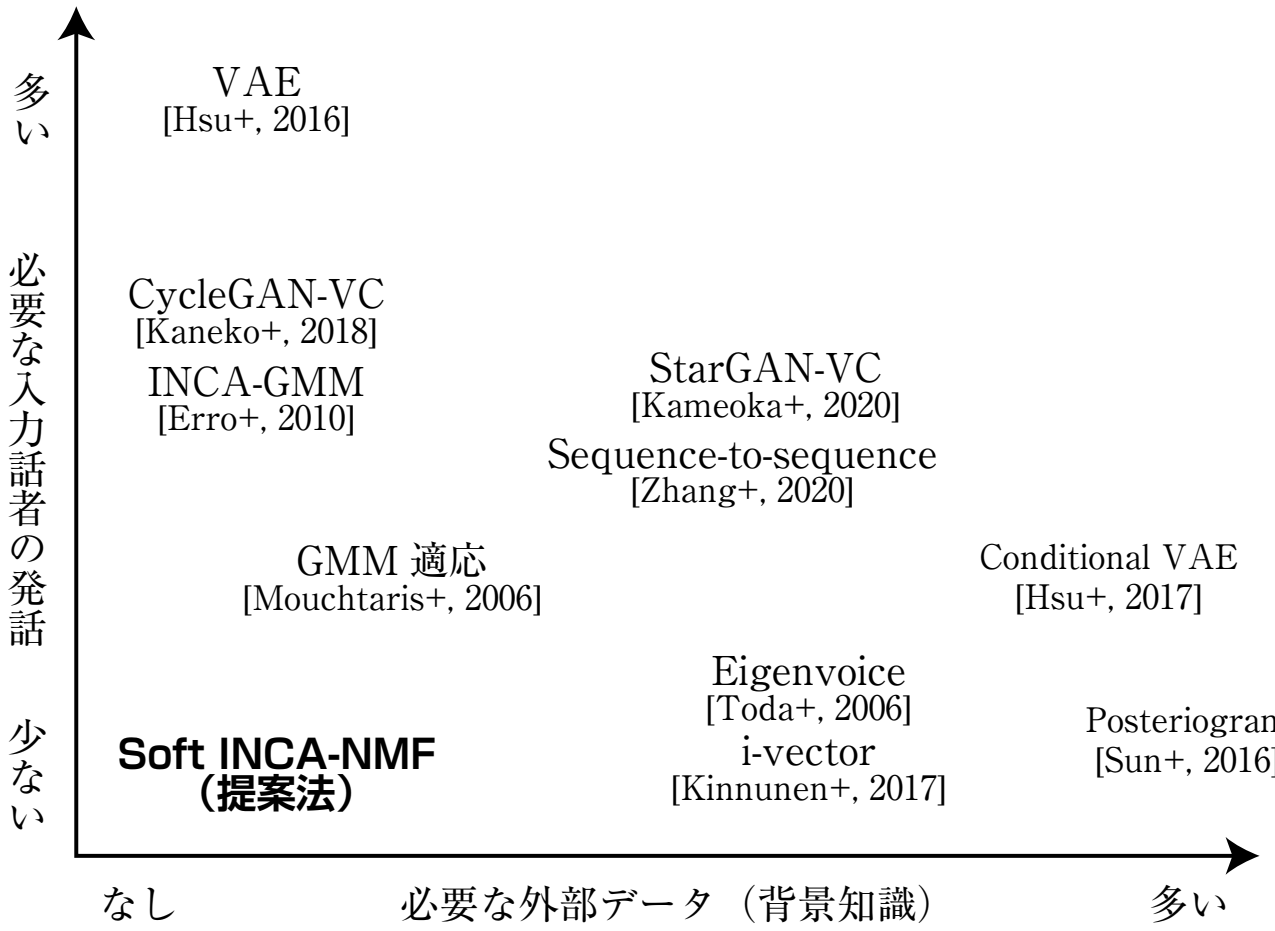


図 4.2　ノンパラレル声質変換法の分布図．本論文で引用した種々のノンパラレル声質変換法を，手法の前提とする必要な背景知識および入力話者の発話の量によって分類した．各手法の位置はあくまで目安である．また，手法の変換品質は学習データ量に大きく依存するため考慮していない．

について解説する．次に，提案法である Soft INCA アルゴリズムについて詳細に説明する．さらに，提案法の評価実験について説明し，その結果について考察する．

4.2　非負値行列因子分解を利用したパラレル声質変換法

4.2.1　非負値行列因子分解

NMF は，非負値の行列を 2 つの非負値行列の積に分解する操作である [125]．$\boldsymbol{Y} \in \mathbb{R}^{\geq 0, K \times T}$ を分解対象の行列とすると，NMF は次式のように行列 $\boldsymbol{H} \in \mathbb{R}^{\geq 0, K \times N}$ および $\boldsymbol{U} \in \mathbb{R}^{\geq 0, N \times T}$ を近似するものである．

$$\boldsymbol{Y} \approx \boldsymbol{H}\boldsymbol{U} \tag{4.1}$$

ここで，$\boldsymbol{H}$ および $\boldsymbol{U}$ はそれぞれ基底および生起状態とよばれる．N は基底の数を表す．

$\boldsymbol{Y} = [\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_T]$ をスペクトログラムのような時系列特徴量であると解釈すると，式 (4.1) の近似は次式のよ

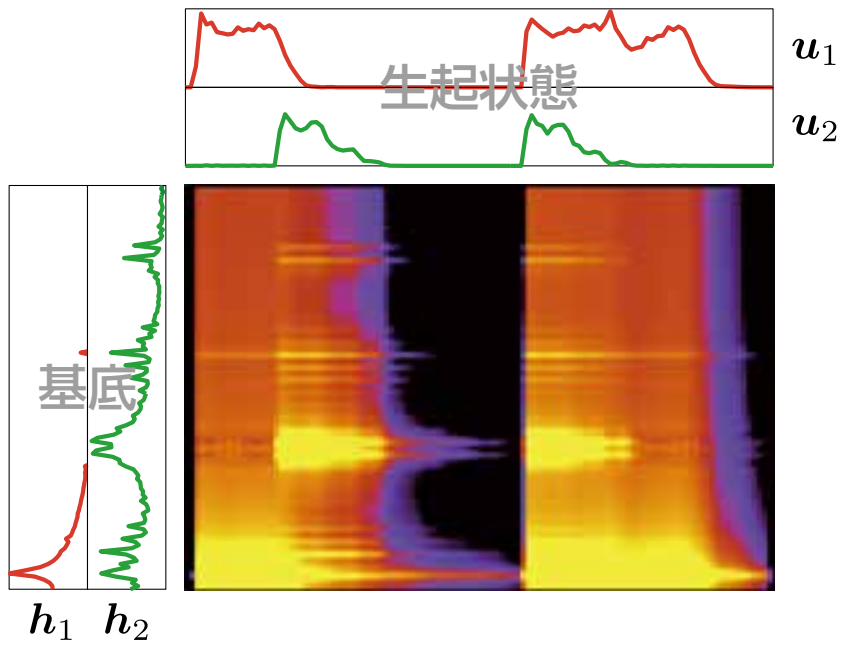


図 4.3　楽音のスペクトログラムに対して NMF を適用した例．スペクトログラムが，基底 $\boldsymbol{h}_1$ および $\boldsymbol{h}_2$ と，その生起状態に分解されている．

うに表現できる．

$$\boldsymbol{y}_t \approx \sum_{n=1}^{N} \boldsymbol{h}_n u_{n,t} \tag{4.2}$$

ただし，$\boldsymbol{h}_n$ および $u_{n,t}$ は次式で定義される．

$$\boldsymbol{H} = [\boldsymbol{h}_1, \boldsymbol{h}_2, \ldots, \boldsymbol{h}_N] \tag{4.3}$$

$$\boldsymbol{U} = [\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_N]^\top \tag{4.4}$$

$$\boldsymbol{u}_n = \left[u_{n,1}, u_{n,2}, \ldots, u_{n,T}\right]^\top \tag{4.5}$$

ここで，$^\top$ は行列の転置を表す．$\boldsymbol{Y}$ がスペクトログラムであれば，K および T は周波数ビンおよび時間フレームの数にあたる．式 (4.2) によれば，時刻 t での観測 $\boldsymbol{y}_t$ は，時不変の基底 $\boldsymbol{h}_1, \boldsymbol{h}_2, \ldots, \boldsymbol{h}_N$ およびその生起状態 $u_{1,t}, u_{2,t}, \ldots, u_{N,t}$ で表されていると解釈できる．$\boldsymbol{Y}$ がスペクトログラムであれば，各基底 $\boldsymbol{h}_n$ はスペクトルのテンプレートを示し，各生起状態 $u_{n,t}$ はその強さを示していると解釈できる．この例を図 4.3 に図示する．各基底がスペクトルのテンプレートを表すことから，$\boldsymbol{H}$ は辞書ともよばれる．この NMF の性質により，自動採譜や雑音抑制，帯域拡張など，NMF は様々な信号処理分野での技術に利用されている [126, 127, 128].

NMF は，観測をほとんど含むような部分空間の基底を得る手法と解釈できる．図 4.4 にこの概念図を示す．部分空間の次元数が基底数にあたり，その部分空間を成すベクトルが基底にあたる．基底は観測と同一の空間にあって，ともに非負値であるため，観測 $\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_T$ と同様の物理量として解釈できる[*2].

[*2] ただし，基底は部分空間を成すベクトルであるため，その大きさについては定めることができない．この点で観測の物理量とまったく同一に解釈できるわけではない．

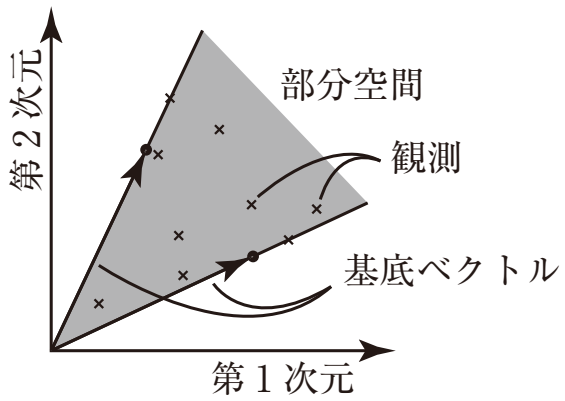


図 4.4　NMF が部分空間の基底を得るアルゴリズムであるという観点での，NMF の概念図．NMF はほとんどの観測を含むような部分空間を成すベクトルを得るアルゴリズムである．それぞれの基底ベクトルが，角錐の辺のベクトルにあたる．

非負値行列 $\boldsymbol{H}$ および $\boldsymbol{U}$ は，ダイバージェンス $\mathcal{D}(\boldsymbol{Y} \mid \boldsymbol{HU})$ を最小化することで得られる．ダイバージェンスには，ユークリッド距離，一般化 KL ダイバージェンス（I ダイバージェンス），板倉斎藤擬距離などが用いられる．この問題は解析的には解けず，補助関数法を用いて反復的に最小化するアルゴリズムが提案されている [129]．補助変数を用いたダイバージェンスの上限関数を設定し，上限関数の最小化と補助変数の更新を繰り返すことで，目的関数であるダイバージェンスを最小化する．たとえば $\mathcal{D}$ が一般化 KL ダイバージェンスの場合，$\mathcal{D}$ は次式で定義される．

$$\mathcal{D}_{\mathrm{KL}}(\boldsymbol{Y} \mid \boldsymbol{X}) = \sum_{k,t} \left(y_{k,t} \log \frac{y_{k,t}}{x_{k,t}} - y_{k,t} + x_{k,t} \right) \tag{4.6}$$

ここで，$y_{k,t}$ および $x_{k,t}$ は次式で定義される．

$$\boldsymbol{y}_t = \left[y_{1,t}, y_{2,t}, \ldots, y_{K,t} \right]^\top \tag{4.7}$$

$$\boldsymbol{X} = \boldsymbol{HU} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_T] \tag{4.8}$$

$$\boldsymbol{x}_t = \left[x_{1,t}, x_{2,t}, \ldots, x_{K,t} \right]^\top \tag{4.9}$$

ここで $-\log x$ は凸関数であるから，イェンセンの不等式より $-\log x_{k,t}$ について次式が成立する．

$$-\log x_{k,t} = \log \sum_n \lambda_{k,n,t} \frac{h_{k,n} u_{n,t}}{\lambda_{k,n,t}} \leq -\sum_n \lambda_{k,n,t} \log \frac{h_{k,n} u_{n,t}}{\lambda_{k,n,t}} \tag{4.10}$$

ただし，$\boldsymbol{h}_n = \left[h_{1,n}, h_{2,n}, \ldots, h_{K,n} \right]^\top$ である．なお，等号成立条件は次式である．

$$\frac{h_{k,1} u_{1,t}}{\lambda_{k,1,t}} = \frac{h_{k,2} u_{2,t}}{\lambda_{k,2,t}} = \cdots = \frac{h_{k,N} u_{N,t}}{\lambda_{k,N,t}} \tag{4.11}$$

$$\Leftrightarrow \lambda_{k,n,t} = \frac{h_{k,n} u_{n,t}}{x_{k,n}} \tag{4.12}$$

したがって，次式で定義される関数 G_{KL} は，$\mathcal{D}_{\mathrm{KL}}$ の上限関数であり補助関数としての要件を満たす．

$$G_{\mathrm{KL}} = \sum_{k,t}\left[y_{k,t}\log y_{k,t} - \sum_{n} y_{k,t}\lambda_{k,n,t}\log\frac{h_{k,n}u_{n,t}}{\lambda_{k,n,t}} - y_{k,t} + \sum_{n} h_{k,n}u_{n,t}\right] \tag{4.13}$$

補助関数法によって，次の 3 パラメータの更新を繰り返すことで，目的関数 $\mathcal{D}_{\mathrm{KL}}$ は単調に減少する．

$$\boldsymbol{\lambda} \leftarrow \arg\min_{\boldsymbol{\lambda}} G_{\mathrm{KL}} \tag{4.14}$$

$$\boldsymbol{H} \leftarrow \arg\min_{\boldsymbol{H}} G_{\mathrm{KL}} \tag{4.15}$$

$$\boldsymbol{U} \leftarrow \arg\min_{\boldsymbol{U}} G_{\mathrm{KL}} \tag{4.16}$$

式 (4.14) の更新は式 (4.12) に示すとおり行えばよい．式 (4.15) の更新は，G_{KL} を $h_{k,n}$ について偏微分を行うことで導かれる．すなわち，次式で $h_{k,n}$ を更新する．

$$h_{k,n} \leftarrow \frac{\sum_t y_{k,t}\lambda_{k,n,t}}{\sum_t u_{n,t}} = h_{k,n}\frac{\sum_t \frac{y_{k,t}}{x_{k,t}}u_{n,t}}{\sum_t u_{n,t}} \tag{4.17}$$

これを $u_{n,t}$ についても同様に計算すると，次の更新式が得られる．

$$u_{n,t} \leftarrow u_{n,t}\frac{\sum_k \frac{y_{k,t}}{x_{k,t}}h_{k,n}}{\sum_k h_{k,n}} \tag{4.18}$$

NMF はダイバージェンスに対応した生成モデルにおける最尤推定問題と等価である．ダイバージェンスが一般化 KL ダイバージェンスの場合，$\boldsymbol{Y}$ は $\boldsymbol{HU}$ にポワソン分布に従うノイズを加えて生成されたものと仮定される [125].

4.2.2　非負値行列因子分解を利用したパラレル声質変換法

NMF の時系列特徴量を時変成分と時不変成分に分解する特徴を利用して，NMF を利用した声質変換法が提案されている [107]．この手法の概略図を図 4.5 に示す．

$\boldsymbol{Y}^{(s)} = \left[\boldsymbol{y}_1^{(s)}, \boldsymbol{y}_2^{(s)}, \ldots, \boldsymbol{y}_{T_s}^{(s)}\right]$ および $\boldsymbol{Y}^{(t)} = \left[\boldsymbol{y}_1^{(t)}, \boldsymbol{y}_2^{(t)}, \ldots, \boldsymbol{y}_{T_t}^{(t)}\right]$ を，同一内容を入力話者および出力話者がそれぞれ発話した音声のスペクトログラムとする．スペクトログラムとして，声道特徴であるスペクトル包絡の振幅もしくはパワースペクトログラムを利用することが多い．DTW などにより時間的なアラインメントを行い，アラインメントされた入出力話者の音声のスペクトログラム $\boldsymbol{Y'}^{(s)} = \left[\boldsymbol{y'}_1^{(s)}, \boldsymbol{y'}_2^{(s)}, \ldots, \boldsymbol{y'}_T^{(s)}\right]$ および $\boldsymbol{Y'}^{(t)} = \left[\boldsymbol{y'}_1^{(t)}, \boldsymbol{y'}_2^{(t)}, \ldots, \boldsymbol{y'}_T^{(t)}\right]$ を得る．NMF を用いた声質変換法では，次のようにこれらのスペクトログラムを話者依存の基底と話者非依存の生起状態に分解する．

$$\boldsymbol{Y'}^{(s)} \approx \boldsymbol{H}^{(s)}\boldsymbol{U} \tag{4.19}$$

$$\boldsymbol{Y'}^{(t)} \approx \boldsymbol{H}^{(t)}\boldsymbol{U} \tag{4.20}$$

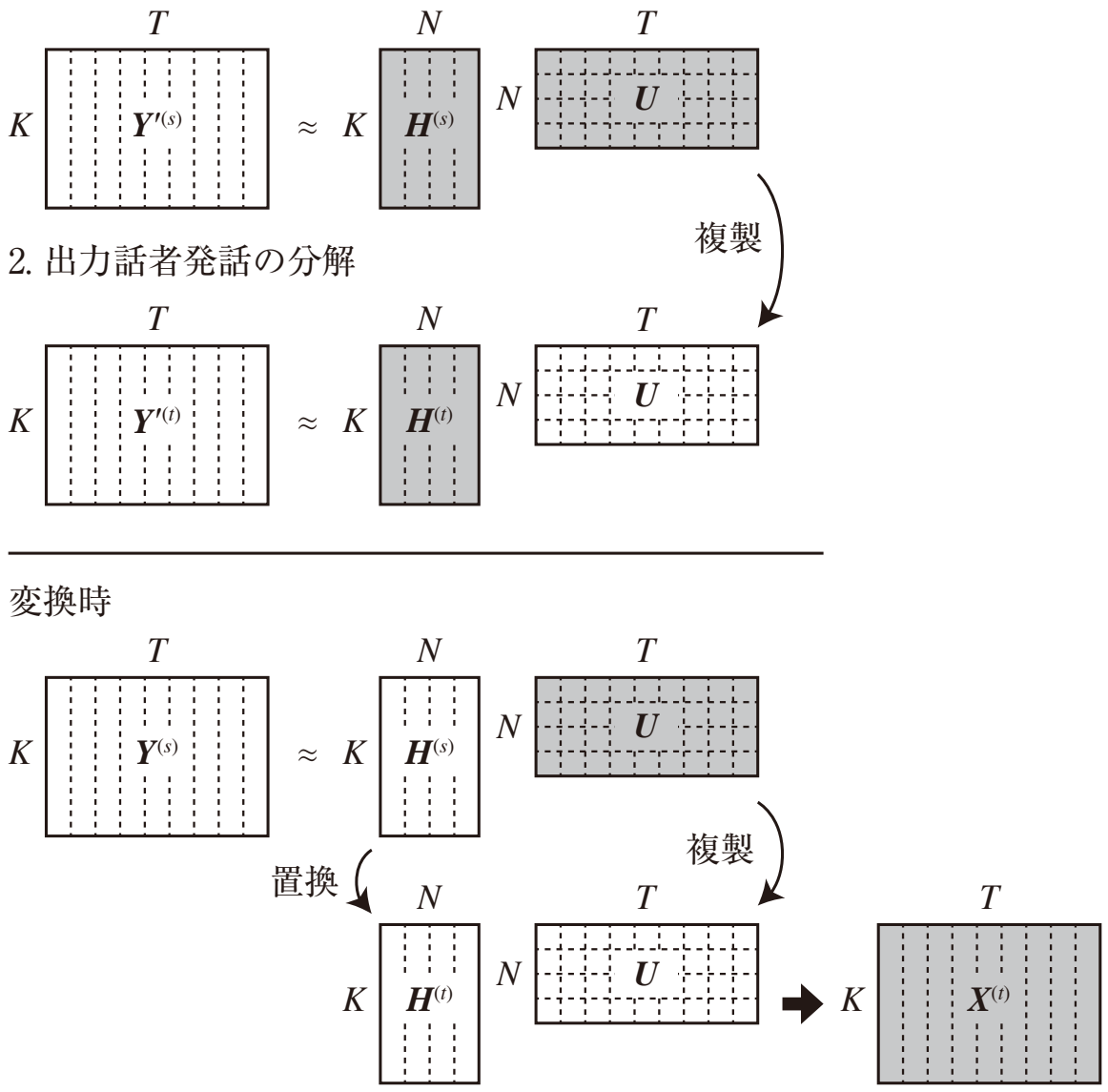


図 4.5　パラレルデータを利用した NMF による声質変換法の概略図 [107]．各手順で得る行列を灰色で示している．

各辞書 $\boldsymbol{H}$ は各話者のスペクトルテンプレートを保持しており，$\boldsymbol{U}$ はそれらの基底の生起状態を表す．$\boldsymbol{U}$ は話者非依存のため，各基底インデックス n について $\boldsymbol{h}_n^{(s)}$ と $\boldsymbol{h}_n^{(t)}$ は音響的に対応付けされている．最終的に，辞書のペア $\boldsymbol{H}^{(s)}$ および $\boldsymbol{H}^{(t)}$ を変換モデルとして保持する．GMM を利用した声質変換法では入出力話者の音声の音響特徴量を結合して学習する [99] が，NMF を用いた声質変換法ではそれらを個別に分解する．これは，結合したスペクトログラムを分解した場合，アラインメントを得る際の誤りなどにより自然性が大きく低下することが実験的に知られているためである．

変換時には入力音声のスペクトログラム $\boldsymbol{Y}^{(s)}$ を入力話者辞書 $\boldsymbol{H}^{(s)}$ を用いて分解し $\boldsymbol{U}$ を得て，これを出力話者辞書 $\boldsymbol{H}^{(t)}$ と乗算することにより変換後スペクトログラム $\boldsymbol{X}^{(t)} = \boldsymbol{H}^{(t)}\boldsymbol{U}$ を得る．

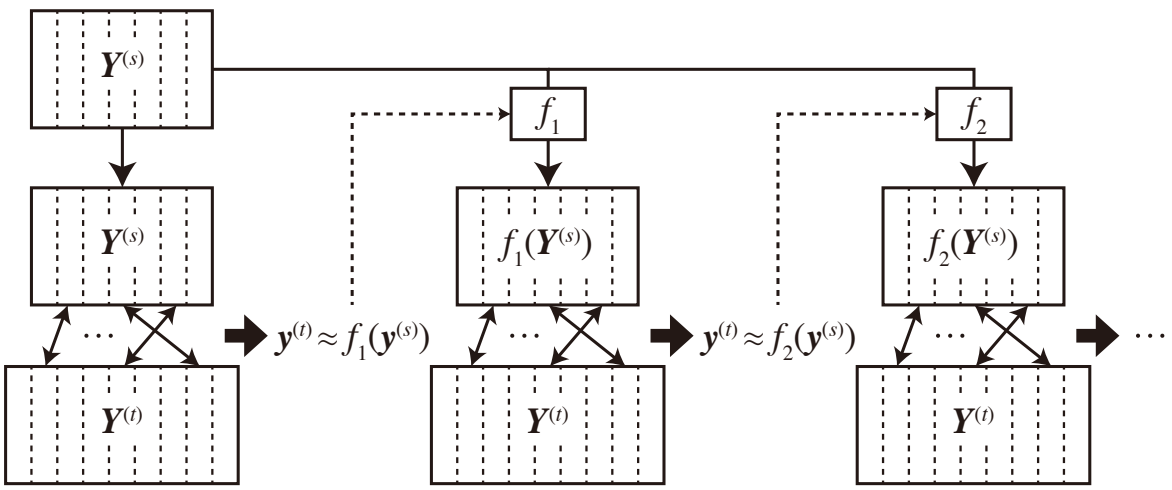


図 4.6　INCA アルゴリズムの手順の概要 [123]．反復することで，$f_i(\boldsymbol{Y}^{(s)})$ がより出力話者による発話らしくなり，アラインメントが収束する．

NMF による声質変換法は，NMF によって，スペクトログラムが話者情報 $\boldsymbol{H}$ と言語情報 $\boldsymbol{U}$ に教師なしに分解されることを利用している．これは，$\boldsymbol{H}$ が時不変かつ $\boldsymbol{U}$ が時変であり，話者情報は時不変かつ言語情報は時変であると期待されるためである．しかし，NMF による分解は，いかなる明示的な制約のない分解であるため，この分解は理想的なものとはならない．すなわち，生起状態 $\boldsymbol{U}$ に話者の情報が含まれている．生起状態に話者情報が含まれる場合，最初に分解される発話の話者情報が生起状態に含まれるため，次に分解される発話の話者情報とミスマッチが生じる．すなわち，話者情報と言語情報の分解の不完全さによって，最終的な変換品質に影響が及ぶ．

NMF による声質変換の本質は，出力話者辞書 $\boldsymbol{H}^{(t)}$ と，発話に対応する生起状態 $\boldsymbol{U}$ のペアを得ることである．ここで，入力話者辞書 $\boldsymbol{H}^{(s)}$ は，$\boldsymbol{U}$ を得るための道具にすぎない．もし仮に出力話者基底 $\boldsymbol{H}^{(t)}$ と入力話者音声 $\boldsymbol{Y}^{(s)}$ から生起状態 $\boldsymbol{U}$ が得られれれば，変換後音声 $\boldsymbol{X}^{(t)}$ を得ることができる．これが，本論文で提案するノンパラレル声質変換法の基礎である．

4.3　INCA アルゴリズム

INCA は，an iterative combination of a nearest neighbor search step and a conversion step alignment method の略であり，ノンパラレルな入出力話者の発話から，フレーム毎のアラインメントを得て，パラレルデータを作成するアルゴリズムである [123]．INCA アルゴリズムはノンパラレルな発話の組からパラレルデータを得る手法であって，任意のパラレル声質変換法と組み合わせることでノンパラレル声質変換を実現する．図 4.6 に INCA アルゴリズムの手順の概要図を示す．

$\boldsymbol{Y}^{(s)} = \left[\boldsymbol{y}_1^{(s)}, \boldsymbol{y}_2^{(s)}, \ldots, \boldsymbol{y}_N^{(s)}\right]$ および $\boldsymbol{Y}^{(t)} = \left[\boldsymbol{y}_1^{(t)}, \boldsymbol{y}_2^{(t)}, \ldots, \boldsymbol{y}_M^{(t)}\right]$ を，それぞれ入出力話者による発話の音響特徴量系列とする．ここで，これらの発話は同一内容である必要はない．INCA アルゴリズムは，入力話者発話の音響特徴量の変換，最近傍法によるアラインメント，変換モデルの学習，の 3 手順を反復するものである．

1. **入力話者発話の音響特徴量の変換**．入力話者の発話を，1 つ前の反復で学習された変換モデルによって次式

のように変換する．

$$\boldsymbol{y}_{i,n}^{(s)} = f_{i-1}\left(\boldsymbol{y}_{n}^{(s)}\right) \tag{4.21}$$

ここで，i は反復回数のインデックス，f_{i-1} は $i-1$ 回目の反復で学習された変換モデルである．最初の反復では，f_0 は恒等変換，すなわち $\boldsymbol{y}_{1,n}^{(s)} = \boldsymbol{y}_{n}^{(s)}$ とする．

2. **最近傍法によるアラインメント**．変換された入力話者発話の音響特徴量 $\boldsymbol{y}_{i,n}^{(s)} = f_{i-1}\left(\boldsymbol{Y}_{i}^{(s)}\right)$ と，出力話者発話の音響特徴量 $\boldsymbol{Y}^{(t)}$ に対して，次式のように最近傍法によってアラインメントを得る．

$$p_i(n) = \arg\min_{m} d\left(\boldsymbol{y}_{i,n}^{(s)}, \boldsymbol{y}_{m}^{(t)}\right) \tag{4.22}$$

$$= \arg\min_{m} d\left(f_{i-1}\left(\boldsymbol{y}_{n}^{(s)}\right), \boldsymbol{y}_{m}^{(t)}\right) \tag{4.23}$$

$$q_i(m) = \arg\min_{n} d\left(\boldsymbol{y}_{i,n}^{(s)}, \boldsymbol{y}_{m}^{(t)}\right) \tag{4.24}$$

$$= \arg\min_{n} d\left(f_{i-1}\left(\boldsymbol{y}_{n}^{(s)}\right), \boldsymbol{y}_{m}^{(t)}\right) \tag{4.25}$$

ここで，d はユークリッド距離などの距離関数，p_i および q_i は得られたアラインメントを表す．この手順によって，入力話者発話の音響特徴量系列と，出力話者発話の音響特徴量系列のアラインメントが得られる．すなわち，パラレルデータを生成することが可能になる．

3. **変換モデルの学習**．変換モデル f_i を，パラレルデータ化された音響特徴量系列 $\left[\boldsymbol{y}_{n}^{(s)\top}, \boldsymbol{y}_{p_i(n)}^{(t)\top}\right]^\top$ および $\left[\boldsymbol{y}_{q_i(m)}^{(s)\top}, \boldsymbol{y}_{m}^{(t)\top}\right]^\top$ を用いて学習する．この変換は，パラレルな声質変換と同様のモデルを用いることができるが，最近傍法による探索によって実効的なサンプル数が少なくなり過学習しやすいため，少ない混合数の GMM などを用いる．

反復を繰り返すことで，アラインメントが次第に収束する．最終的に，得られたパラレルデータ $\left[\boldsymbol{y}_{n}^{(s)\top}, \boldsymbol{y}_{p_i(n)}^{(t)\top}\right]^\top$ および $\left[\boldsymbol{y}_{q_i(m)}^{(s)\top}, \boldsymbol{y}_{m}^{(t)\top}\right]^\top$ を利用して，変換モデルを学習する．

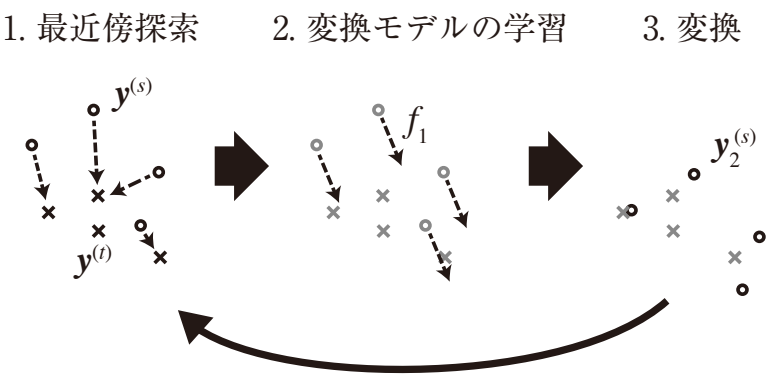
INCA アルゴリズムの収束は，次式の二乗誤差によって判断できる．

$$d_i = \frac{1}{N+M}\left(\sum_{n=1}^{T_s}\left\|\boldsymbol{y}_{i,n}^{(s)} - \boldsymbol{y}_{p_i(n)}^{(t)}\right\|^2 + \sum_{m=1}^{T_t}\left\|\boldsymbol{y}_{i,q_i(m)}^{(s)} - \boldsymbol{y}_{m}^{(t)}\right\|^2\right) \tag{4.26}$$

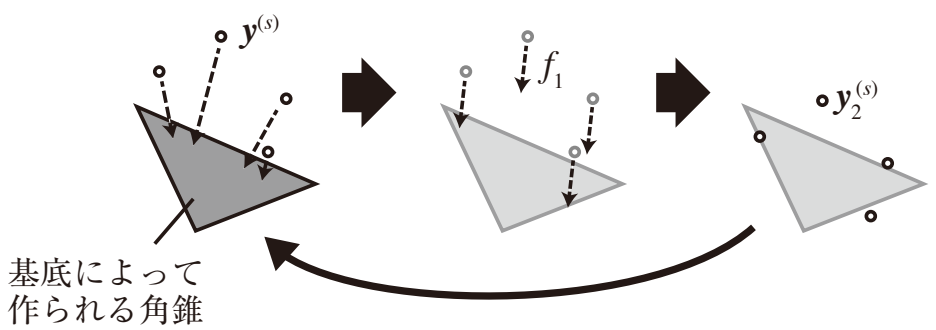
この二乗誤差は単調に減少することが数学的に示されている [130].

4.4 Soft INCA アルゴリズム

INCA アルゴリズムは，入力話者発話の音響特徴量に対して，出力話者発話の音響特徴量からもっとも近いものを選択する手順によって実現される．図 4.7a にこの概念図を示す．INCA アルゴリズムは，各音響特徴量に対して対応する音響特徴量を最近傍法によって探索する．学習に用いる発話の量が少ない場合，INCA アルゴリズムでは適切な対応する音響特徴量が発見できず，アラインメント後の音響特徴量系列に影響する．とくに，観測されて



(a) INCA アルゴリズム [123]



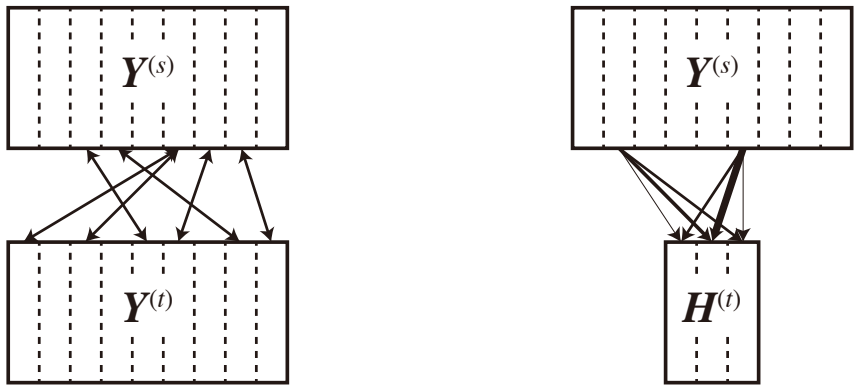
(b) Soft INCA アルゴリズム

図 4.7　INCA アルゴリズムおよび Soft INCA アルゴリズムの概念図．どちらの手法も，アラインメント，変換モデルの学習，変換の 3 手順によって，入力話者発話の音響特徴量 $\boldsymbol{y}^{(s)}$ を出力話者発話の音響特徴量に近づける手法である．INCA アルゴリズムでは音響特徴量を観測された出力話者発話の音響特徴量に近づけるが，Soft INCA アルゴリズムの場合には出力話者発話の音響特徴量系列を NMF によって分解された部分空間に近づける．

いない音素などがある場合，理想的なアラインメントを得ることができない．また，INCA アルゴリズムでは，類似の音響特徴量に対してまったく異なる音響特徴量が割り当たる場合がある．このため，INCA アルゴリズムによって得られるパラレルデータは，不自然で非連続的である．

このような INCA アルゴリズムの問題点を補うため，本論文では Soft INCA アルゴリズムを提案する．Soft INCA アルゴリズムは，NMF を利用して，入力話者発話の音響特徴量を出力話者発話の音響特徴量の空間に分解する．NMF によるアラインメントは一対一ではなく，生起状態は連続的な値であるため，連続的に出力話者のもつ空間に分解できる．これによって，INCA アルゴリズムと比較して自然で滑らかなパラレルデータを生成できると考えられる．図 4.7b に Soft INCA アルゴリズムの概念図を示す．NMF は生起状態が疎になりやすく，部分空間が小さくなりやすいという性質を持つ [125]．この性質によって出力話者の空間が小さくなるため，十分に入力話者発話の音響特徴量を出力話者の空間に近づけることができる．

Soft INCA アルゴリズムは．INCA アルゴリズムの音響特徴量から音響特徴量への対応付けを連続的な対応付けに置き換えたものと解釈できる．NMF を用いた声質変換では，スペクトログラムを，辞書と生起状態に分解する．この生起状態は，辞書の各基底がどの程度の大きさで用いられているかを表すため，スペクトログラムと基底間の



(a) INCA = 離散的なアラインメント　(b) NMF = 連続的なアラインメント
入力話者特徴量 ↔ 出力話者特徴量　入力話者特徴量 ↠ 出力話者基底

図 4.8　NMF の生起状態が連続的なアラインメントであるという概念の可視化．INCA アルゴリズムでは入出力話者発話の音響特徴量間の離散的な対応付けを得るが，Soft INCA アルゴリズムでは入力話者発話の音響特徴量から出力話者の基底への連続的な対応付けを得る．

アラインメントと解釈できる．この概念を図 4.8 に可視化する．Soft INCA アルゴリズムは，INCA アルゴリズムのアラインメントを soft にしたものであることから，Soft INCA アルゴリズムとよぶ．生起状態が非負であることは Soft INCA アルゴリズムにおいて重要な要素であり，様々な行列分解法のなかで NMF は最適な分解法である．

連続的なアラインメントを INCA アルゴリズムと同様の手続きで得ることで，パラレルデータや大量の背景知識を利用することなく，NMF を用いた声質変換システムを構築できる．観測された音響特徴量を補間してアラインメントを得るため，入力話者の発話に関しては少量であっても自然な変換を実現できることが期待される．

Soft INCA アルゴリズムは，出力話者基底の学習，入力話者発話に対応する生起状態の推定，入力話者基底の推定，の 3 手順からなる．図 4.9 に Soft INCA アルゴリズムの概略図を示す．$\boldsymbol{Y}^{(s)} = \left[\boldsymbol{y}_1^{(s)}, \boldsymbol{y}_2^{(s)}, \ldots, \boldsymbol{y}_N^{(s)}\right]$ および $\boldsymbol{Y}^{(t)} = \left[\boldsymbol{y}_1^{(t)}, \boldsymbol{y}_2^{(t)}, \ldots, \boldsymbol{y}_M^{(t)}\right]$ をそれぞれ入力話者および出力話者による発話の音響特徴量系列とする．

まず，$\boldsymbol{Y}^{(t)}$ を NMF により分解することで出力話者基底 $\boldsymbol{H}^{(t)}$ を得る．この手順は出力話者のモデル化にあたる．この手順では制約なく分解を行うため，可能な限り緻密なモデル化を行うことができる．

次に，生起状態 $\boldsymbol{U}$ を，入力話者の発話 $\boldsymbol{Y}^{(s)}$ から次の手順で推定する．

1. **入力話者発話の音響特徴量の変換**．入力話者発話の音響特徴量を次式のように変換し，$\boldsymbol{Y}_i^{(s)}$ を得る．

$$\boldsymbol{y}_{i,n}^{(s)} = f_{i-1}\left(\boldsymbol{y}_n^{(s)}\right) \tag{4.27}$$

   ここで，i は反復回数のインデックス，f_{i-1} は $i-1$ 回目の反復で学習された変換モデルである．最初の反復では，f_0 は恒等変換，すなわち $\boldsymbol{y}_{1,n} = \boldsymbol{y}_n^{(s)}$ とする．

2. **NMF による音響特徴量の分解**．変換された音響特徴量系列 $\boldsymbol{Y}_i^{(s)}$ を，出力話者の辞書 $\boldsymbol{H}^{(t)}$ を用いて分解し，生起状態 $\boldsymbol{U}_i$ を得る．この手順は，INCA アルゴリズムにおけるアラインメントの手順にあたり，$\boldsymbol{U}_i$ は音響特徴量系列 $\boldsymbol{Y}^{(s)}$ と出力話者の辞書 $\boldsymbol{H}^{(t)}$ の間のアラインメントを表す．INCA アルゴリズムとは異なり，こ

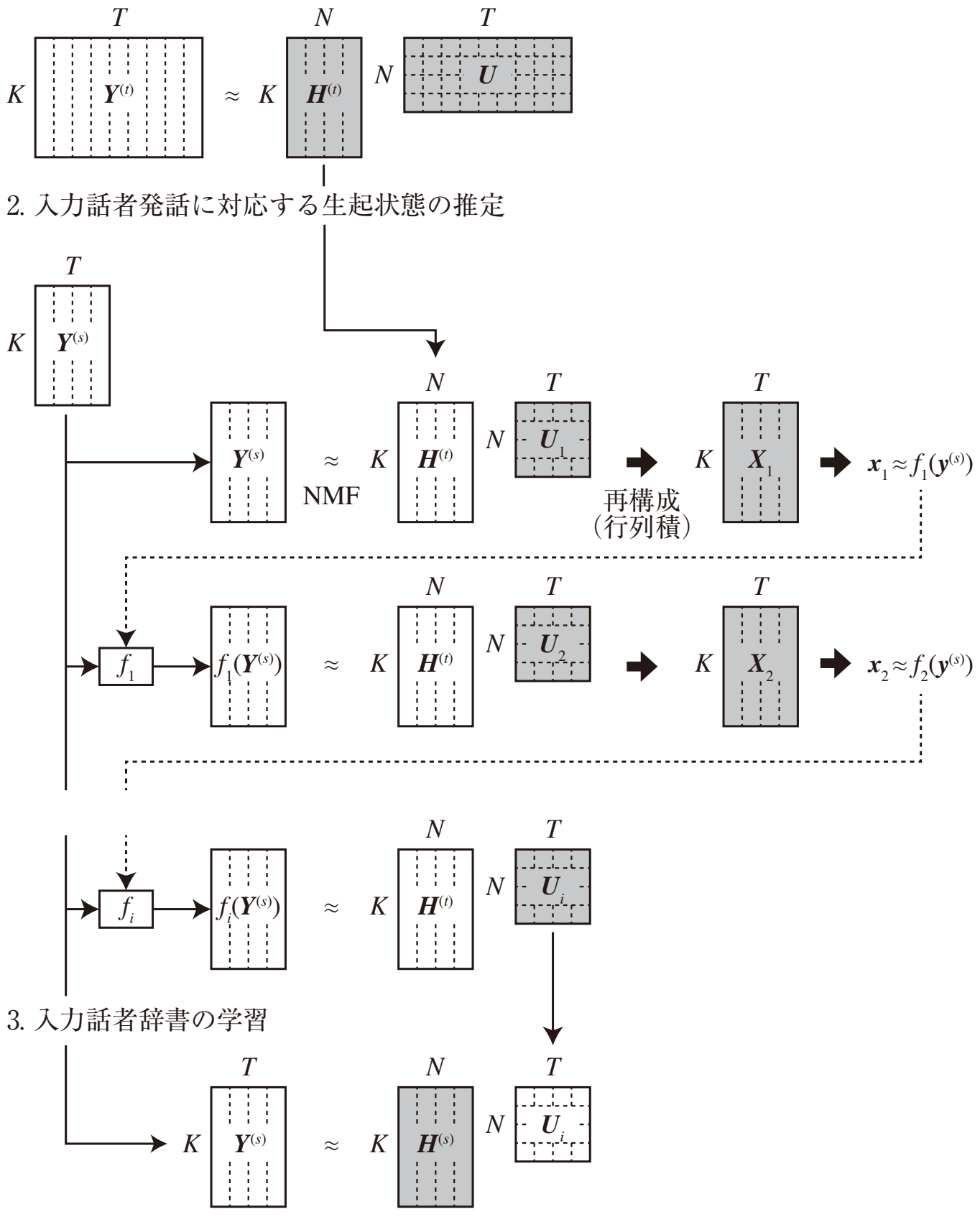


図 4.9　Soft INCA アルゴリズムの手順の概略図．各手順で得る行列を灰色で示している．

のアラインメントは連続的である．

3. **再構成によるパラレルデータの生成**．変換された音響特徴量系列 $\boldsymbol{X}_i$ を，次式のように行列積によって再構成することで得る．

$$\boldsymbol{X}_i = \boldsymbol{H}^{(t)}\boldsymbol{U}_i \tag{4.28}$$

$\boldsymbol{X}_i$ は $\boldsymbol{Y}^{(s)}$ を声質変換したものであり，この手順がパラレルデータの生成にあたる．

4. **変換モデルの学習**．変換モデル f_i を，$\boldsymbol{Y}^{(s)}$ および $\boldsymbol{X}_i$ を用いて学習する．INCA アルゴリズムと同様に，この変換はパラレル声質変換法と同様の変換モデルを用いることが可能であるが，過学習を避けるために少ない混合数の GMM などを用いる．

収束の様子は NMF のダイバージェンス $\mathcal{D}\left(f_i(\boldsymbol{Y}^{(s)}) \,\middle|\, \boldsymbol{X}_i\right)$ により観察できる．変換モデルの学習と NMF の分解で異なる距離規準を用いるため，このダイバージェンスは単調に収束しないが，単調に収束すると仮定しても問題ないことを後述の実験で確認する．

最後に，入力話者発話の音響特徴量 $\boldsymbol{Y}^{(s)}$ を，推定された生起状態 $\boldsymbol{U}$ を用いて NMF により分解し，入力話者基底 $\boldsymbol{H}^{(s)}$ を得る．

変換時には，辞書のペアを利用して，NMF を用いたパラレル声質変換法と同様の手順で声質変換を実現できる．

Soft INCA アルゴリズムは，INCA アルゴリズムと同様に，パラレルデータの生成アルゴリズムでもあるため，任意のパラレルデータを用いた声質変換法を利用することも可能である．この場合は，$\boldsymbol{Y}^{(s)}$ および $\boldsymbol{X}_i$ をパラレルデータと見なして学習すればよい．本論文では，出力話者基底 $\boldsymbol{H}^{(t)}$ を有効に活用するため，NMF を用いて声質変換を行う．

4.5　実験条件

データセットには，日英・日中バイリンガル独話音声データベース[*3]のうち，日英バイリンガル話者の発話を用いた．データセットに含まれる話者のうち，筆者の聴感上発声の上手な 8 人を選んだ．このうち話者 EJF101，EJF102，EJM101 はプロフェッショナル話者，その他の 5 人は非プロフェッショナル話者である．表 4.1 に，8 人の話者についての詳細を示す．学習には音素バランス文[*4]を用い，変換には合文法無意味文[*5]を利用した．音素バランス文は各およそ 5 秒，合文法無意味文は各およそ 3.5 秒である．NMF の分解をより安定させるため，音声はすべて 24 kHz にダウンサンプリングした．音声の分析合成には WORLD [131]（D4C edition [132]）を用いた．合成時には WORLD のバリエーションである Requiem を用い，自然性を向上させるためゼロ位相フィルタによる音

[*3] https://alaginrc.nict.go.jp/slc-outline.html
[*4] その言語に現れる代表的な音素がすべて用いられ，かつ音素が利用されるバランスが全体として自然な文章に近くなるように設計された，文の集合のこと．日本語の代表的な音素バランス文には，ATR 音素バランス文 [64]，声優統計コーパスなどがある．
[*5] 「裏目の重さがまじめに汗ばむ．」といった，文法的には正しいが意味の成さない文のこと．

表 4.1　実験に用いた話者の詳細な情報．プロフェッショナル話者の母語はデータセット中に示されていない．

話者	性別	母語	プロフェッショナルか否か
EJF04	女性	日本語，英語	非プロフェッショナル
EJF08	女性	日本語	非プロフェッショナル
EJM09	男性	日本語，英語	非プロフェッショナル
EJM13	男性	日本語	非プロフェッショナル
EJF101	女性		プロフェッショナル
EJF102	女性		プロフェッショナル
EJM11	男性	日本語	非プロフェッショナル
EJM101	男性		プロフェッショナル

表 4.2　INCA アルゴリズムおよび Soft INCA アルゴリズムで用いる変換モデルのスケジュール．

反復	変換モデル	パラメータ数
1–10	声道長変換	1
11–20	GMM 声質変換（$M = 1$）	500
21–30	GMM 声質変換（$M = 2$）	1,001
31–40	GMM 声質変換（$M = 4$）	2,003
41–50	GMM 声質変換（$M = 8$）	4,007
51–60	GMM 声質変換（$M = 16$）	8,015

声合成を行った．フレーム周期は 1 ms とした．基本周波数は次式のように平均と分散をもとに線形に変換した．

$$\hat{\phi}_t^{(t)} = \frac{\sigma^{(t)}}{\sigma^{(s)}}\left(\phi_t^{(s)} - \mu^{(s)}\right) + \mu^{(t)} \tag{4.29}$$

ここで $\phi_t^{(s)}$ および $\hat{\phi}_t^{(t)}$ はそれぞれ第 t フレームでの変換前後の対数基本周波数であり，μ および σ は対数基本周波数の平均および分散である．非周期性指標については変換を行わなかった．NMF を用いた手法では，256 次のメル振幅スペクトログラムを用い，分解規準には一般化 KL ダイバージェンスを用いた．基底数は 128 とした．より識別性の高い基底を得るため，出力話者の基底を学習する際には，GMM によりメルケプストラム係数をクラスタリングしたのち，この平均ベクトルから基底ベクトルを生成し，これを初期値として NMF を行った．

INCA アルゴリズムおよび Soft INCA アルゴリズムでは，100 次のメルケプストラム係数に対して変換を学習した．収束を速め学習の安定性を向上させるため，反復に応じて徐々に変換モデルを複雑にした[*6]．表 4.2 に変換モデルのスケジュールを示す．声道長変換は，z 変換領域で，次式のように定義される．

$$\hat{z}^{-1} = \frac{z^{-1} - \alpha}{1 - \alpha z^{-1}},\ z = e^{j\omega},\ \hat{z} = e^{j\hat{\omega}} \tag{4.30}$$

[*6] 変換モデルを徐々に複雑にして学習を行う手法は，MSR Identity Toolbox や Hidden Markov Model Toolkit（HTK）などにも用いられている．とくに GMM の混合数を徐々に増やす手法は mix up とよばれている．

ここで，α は $-1 < \alpha < 1$ なるウォーピングパラメータ，ω および $\hat{\omega}$ は変換前後の正規化角周波数である [133]．この変換は，ケプストラム領域で再帰的に計算できる [134]．GMM 声質変換では，入力話者発話の音響特徴量系列 $\boldsymbol{x} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_T]$ および出力話者発話の音響特徴量系列 $\boldsymbol{y} = [\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_T]$ を用いて，GMM を利用して変換モデルを構築する [99]．GMM 声質変換では，結合した音響特徴量 $\boldsymbol{z} = [\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_T]$（$\boldsymbol{z}_t = \left[\boldsymbol{x}_t^\top, \boldsymbol{y}_t^\top\right]^\top$）を次式のように GMM によってモデル化する．

$$p(\boldsymbol{z}) = \sum_{m=1}^{M} w_m \mathcal{N}(\boldsymbol{z}; \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m) \tag{4.31}$$

$$\boldsymbol{\mu}_m = \begin{bmatrix} \boldsymbol{\mu}_m^{(x)} \\ \boldsymbol{\mu}_m^{(y)} \end{bmatrix} \tag{4.32}$$

$$\boldsymbol{\Sigma}_m = \begin{bmatrix} \boldsymbol{\Sigma}_m^{(xx)} & \boldsymbol{\Sigma}_m^{(xy)} \\ \boldsymbol{\Sigma}_m^{(xy)} & \boldsymbol{\Sigma}_m^{(yy)} \end{bmatrix} \tag{4.33}$$

ここで，M は GMM の混合数，$\boldsymbol{\mu}_m$ および $\boldsymbol{\Sigma}_m$ は第 m 混合の平均ベクトルおよび分散共分散行列である．過学習を防ぐため，分散共分散行列の各ブロック要素 $\boldsymbol{\Sigma}^{(xx)}$，$\boldsymbol{\Sigma}^{(yy)}$，$\boldsymbol{\Sigma}^{(xy)}$ は対角行列と仮定する．GMM 声質変換における変換関数は，次式のように最尤推定によって導かれる．

$$\hat{\boldsymbol{y}} = \arg\max_{\boldsymbol{y}} p(\boldsymbol{y} \mid \boldsymbol{x}) \tag{4.34}$$

ここで，$\boldsymbol{x}$ および $\hat{\boldsymbol{y}}$ はそれぞれ入力音響特徴量および変換後音響特徴量である [101]．式 (4.34) において，$p(\boldsymbol{y} \mid \boldsymbol{x})$ は次式で与えられる．

$$p(\boldsymbol{y} \mid \boldsymbol{x}) = \sum_{m=1}^{M} p(m \mid \boldsymbol{x}) p(\boldsymbol{y} \mid \boldsymbol{x}, m) \tag{4.35}$$

ここで，$p(m \mid \boldsymbol{x})$ および $p(\boldsymbol{y} \mid \boldsymbol{x}, m)$ は次のように計算できる．

$$p(m \mid \boldsymbol{x}) = \frac{w_m \mathcal{N}\left(\boldsymbol{x}; \boldsymbol{\mu}_m^{(x)}, \boldsymbol{\Sigma}_m^{(xx)}\right)}{\sum_{m'=1}^{M} w_{m'} \mathcal{N}\left(\boldsymbol{x}; \boldsymbol{\mu}_{m'}^{(x)}, \boldsymbol{\Sigma}_{m'}^{(xx)}\right)} \tag{4.36}$$

$$p(\boldsymbol{y} \mid \boldsymbol{x}, m) = \mathcal{N}(\boldsymbol{y}; \boldsymbol{E}_m, \boldsymbol{D}_m) \tag{4.37}$$

式 (4.37) において，$\boldsymbol{E}_m$ および $\boldsymbol{D}_m$ はそれぞれ次式で与えられる．

$$\boldsymbol{E}_m = \boldsymbol{\mu}_m^{(y)} + \boldsymbol{\Sigma}_m^{(xy)} \boldsymbol{\Sigma}_m^{(xx)^{-1}} \left(\boldsymbol{x} - \boldsymbol{\mu}_m^{(x)}\right) \tag{4.38}$$

$$\boldsymbol{D}_m = \boldsymbol{\Sigma}_m^{(yy)} - \boldsymbol{\Sigma}_m^{(xy)} \boldsymbol{\Sigma}_m^{(xx)^{-1}} \boldsymbol{\Sigma}_m^{(xy)} \tag{4.39}$$

INCA アルゴリズムおよび Soft INCA アルゴリズムにおいては，変換モデルにはどのようなパラレル声質変換法を用いることもできるが，徐々にモデルを複雑にできること，また少ないパラメータで変換モデルを表現できることから，本論文では GMM 声質変換を採用した．

4.6　手法の収束の客観的評価

4.6.1　実験条件

入力話者には EJF04 および EJM09 を，出力話者には EJF08 および EJM13 を選んだ．学習文数は，出力話者については 30 文，入力話者については 1 文とした．学習および評価に用いた発話はすべて日本語である．

4.6.2　非負値行列因子分解のダイバージェンスによる収束の評価

Soft INCA アルゴリズムの収束を確認するため，NMF のダイバージェンス $\mathcal{D}\left(f_i\left(\boldsymbol{Y}^{(s)}\right) \,\middle|\, \boldsymbol{H}^{(t)}\boldsymbol{U}\right)$ の推移を調査した．このダイバージェンスは，変換された音響特徴量 $f_i\left(\boldsymbol{Y}^{(s)}\right)$ がどの程度 NMF によって分解できたか，すなわちどの程度出力話者の空間にあるかを示唆するものであり，4.4 節で述べたように，このダイバージェンスは手法の収束を評価する手掛かりとなる．

図 4.10 にダイバージェンスの推移を示す．すべての話者ペアで，NMF のダイバージェンスにおいて収束が確認された．この収束は 4.4 節で述べたとおり単調に収束することが数学的には示されず，実験的にも単調な収束はみられなかったが，実験的に大域的な収束を仮定できることが確認された．また，図 4.11 に，常に 16 混合の GMM で学習した場合の NMF のダイバージェンスの推移を示す．モデルを徐々に複雑にしたシステムと比較して，常に 16 混合の GMM を用いた場合は収束が遅い．したがって，徐々にモデルを複雑にすることで高速な収束を実現できることが実験的に確かめられた．

4.6.3　変換品質の評価

INCA アルゴリズムや Soft INCA アルゴリズムで用いられる変換 f_i の品質を評価するため，このモデルで変換された音響特徴量と，パラレルデータから抽出した出力話者発話の音響特徴量（目標音響特徴量）のメルケプストラム歪み（mel-cepstral distortion; MCD）を計算した．MCD は次式のように定義される[*7]．

$$\mathrm{MCD[dB]} = \frac{10}{\log 10}\sqrt{2\sum_{d=1}^{24}\left(mc_d^{(y)} - \hat{mc}_d^{(y)}\right)^2} \tag{4.40}$$

ここで，$mc_d^{(y)}$ および $\hat{mc}_d^{(y)}$ はそれぞれ第 d 次元の目標音響特徴量と変換後音響特徴量である．MCD によって変換された音声の話者性が目標の話者性からどの程度離れているかを客観的に評価できる．したがって，MCD が低いほど変換性能が高いことを示唆する．本実験では，INCA アルゴリズムと Soft INCA アルゴリズムを組み合わせたコンビネーション法を導入し，その性能についても評価した．コンビネーション法では，INCA アルゴリズムを 25 反復適用し，その後 Soft INCA アルゴリズムを適用する．

[*7] 本論文では，統一して 24 次メルケプストラム係数に対して MCD を計算している．高次のメルケプストラム係数はごく小さいため，評価には大きな影響を及ぼさない．

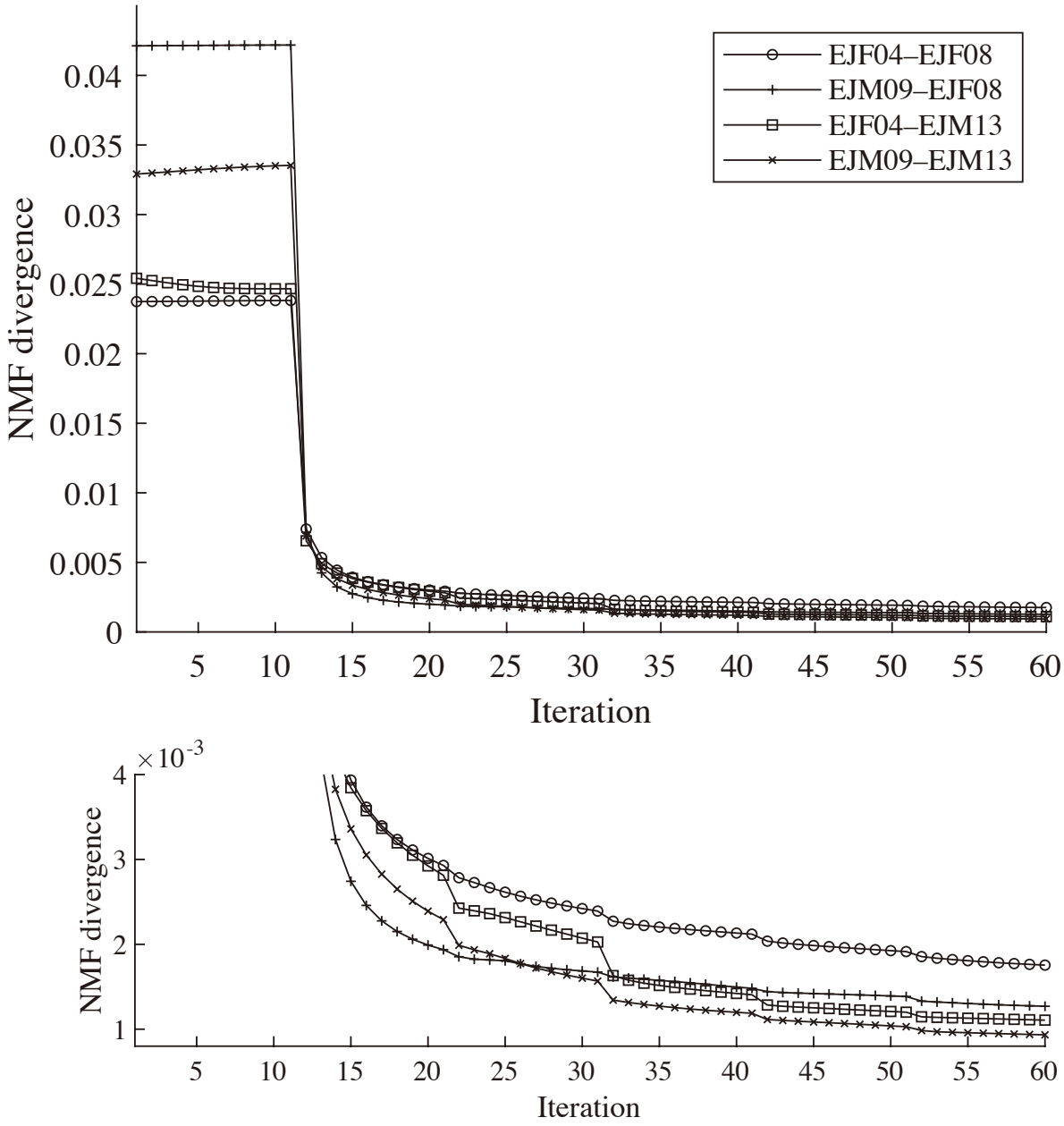


図 4.10　NMF のダイバージェンス $\mathcal{D}\left(f_i\left(\boldsymbol{Y}^{(s)}\right) \,\middle|\, \boldsymbol{H}^{(t)}\boldsymbol{U}\right)$ の推移．下段は上段を拡大した図である．

図 4.12 に結果を示す．Soft INCA アルゴリズムは緩やかに MCD を減少させるが，その値は INCA アルゴリズムと比較して高い．一方，コンビネーション法は INCA アルゴリズムと同等の MCD を示している．したがって，コンビネーション法により，INCA アルゴリズムと同等の話者性の品質を持つより自然な変換を実現できることが確認された．また，INCA アルゴリズムは，とくに異性間変換において，混合数が多い場合に過学習がみられた．一方で，Soft INCA アルゴリズムやコンビネーション法では過学習が起きにくく，より安定して緩やかに学習できることが示された．

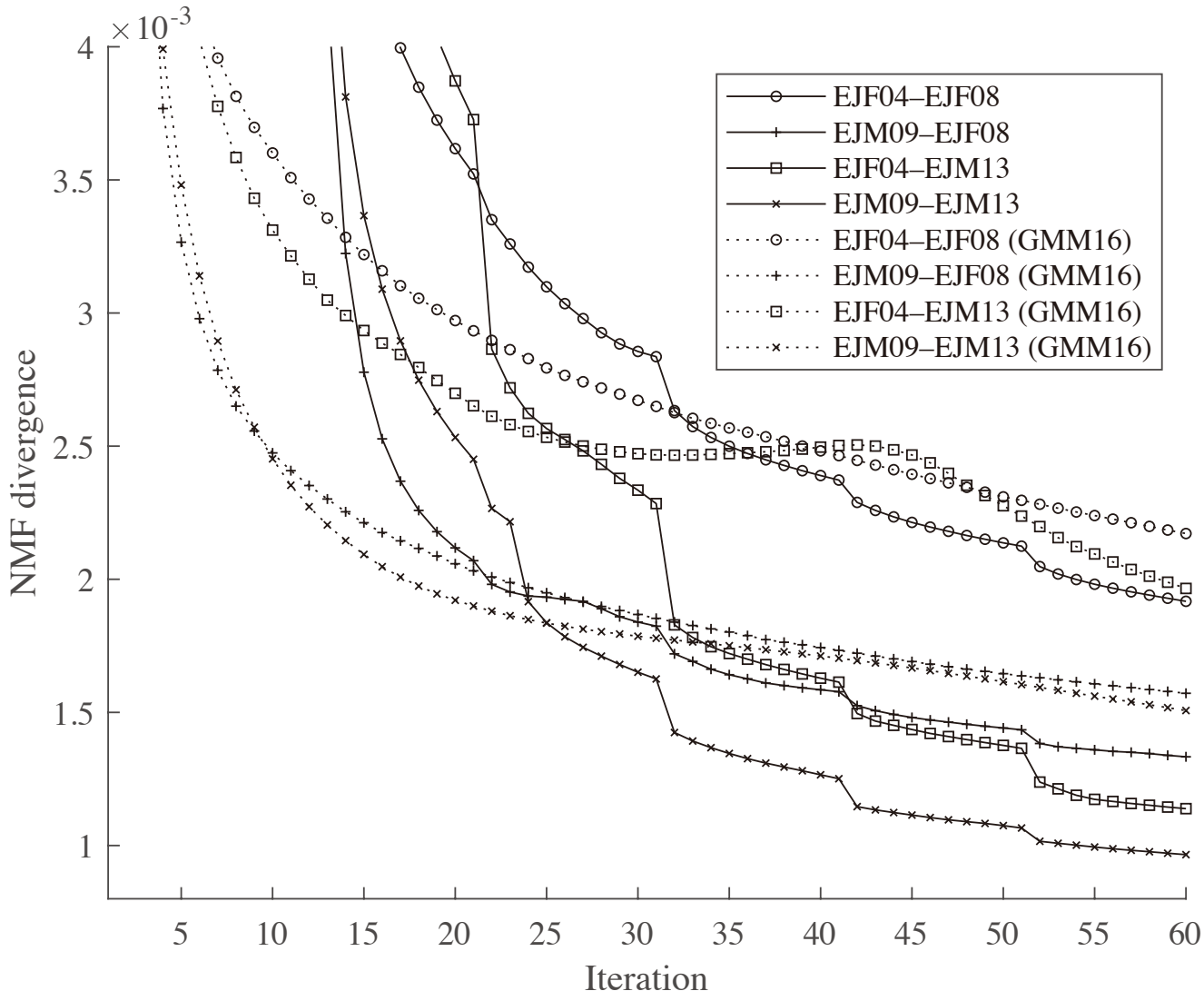


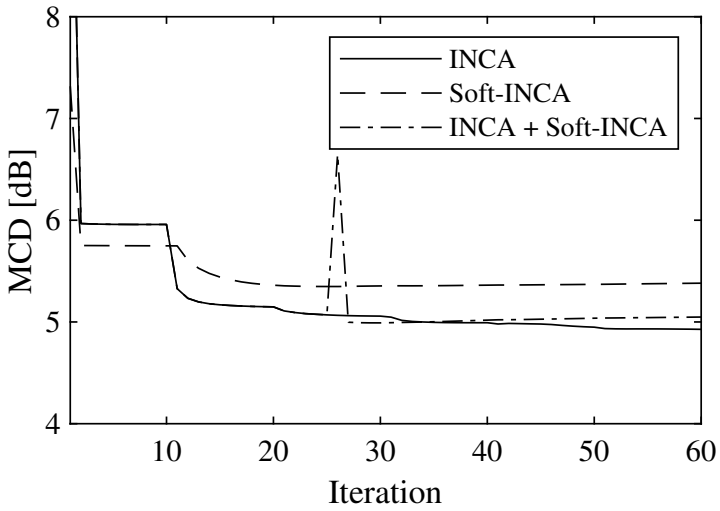
図 4.11　NMF のダイバージェンス $\mathcal{D}\left(f_i\left(\boldsymbol{Y}^{(s)}\right) \middle| \boldsymbol{H}^{(t)}\boldsymbol{U}\right)$ の推移．常に 16 混合の GMM で学習したシステムは，GMM16 と示されている．

4.7　変換品質の評価

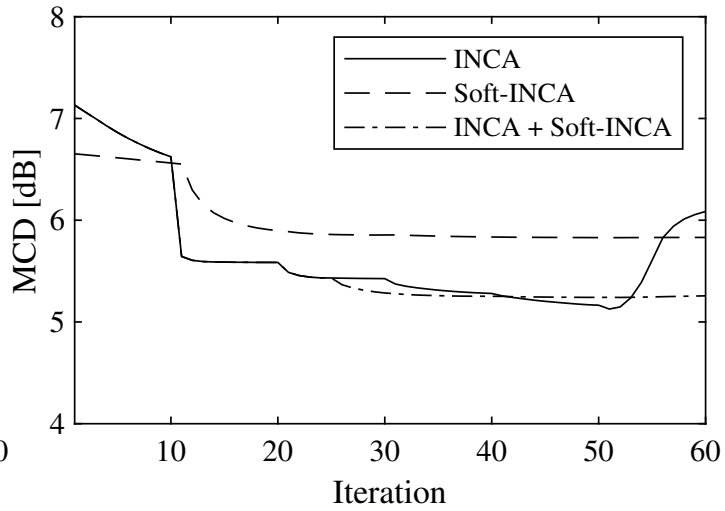
4.7.1　実験条件

入力話者には EJF101 および EJM11 を，出力話者には EJF102 および EJM101 を選んだ．

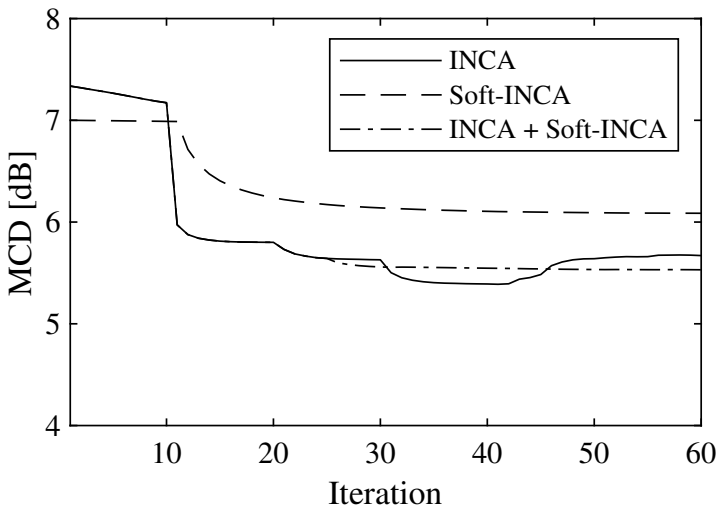
主観評価として，自然性については AB 選好試験，話者性については ABX 試験を行った．AB 選好試験は，2 つの音声 A および音声 B を提示し，「音声 A のほうが自然である」「音声 B のほうが自然である」の 2 選択肢から選択させるものである．ABX 試験は，2 つの音声 A および音声 B と目標音声（出力話者音声）X を提示し，「音声 A のほうが音声 X の話者に近い」「音声 B のほうが音声 X の話者に近い」の 2 選択肢から選択させるものである．すべての評価ペアにおいて，被験者は少なくとも 25 人であり，すべての評価者は各ペアについて 2 問答えた．そのため，すべての評価ペアにおいて，標本数は 50 以上である．評価に用いたすべての音声は日本語で，被験者はすべて母語が日本語の日本人である．主観評価はクラウドソーシングサービスを通じて行い，各被験者は回答した



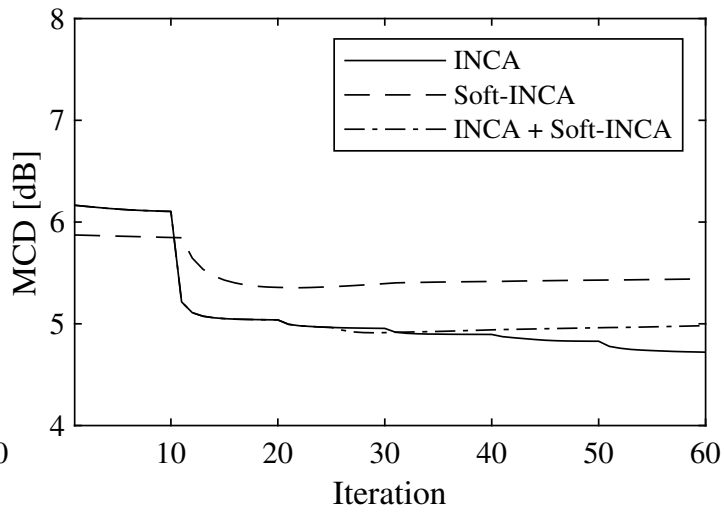
(a) EJF04–EJF08



(b) EJF04–EJM13



(c) EJM09–EJF08



(d) EJM09–EJM13

図 4.12　INCA アルゴリズムおよび Soft INCA アルゴリズムにおける，変換後の音響特徴量と目標の音響特徴量間の MCD の推移.

問題数に応じて報酬を支払われた.

4.6 節に述べた実験より，INCA アルゴリズムおよび Soft INCA アルゴリズムの反復数は，それぞれ 40 および 50 とした.

4.7.2　同一言語内変換

入出力話者ともに日本語の音素バランス文を用いて学習を行い，日本語の合文法無意味文を変換し，その品質を主観評価により評価した.

評価されたシステム

本節では，次の 5 システムについて評価を行った.

- **Soft**: Soft INCA アルゴリズムによる NMF を用いたノンパラレル声質変換法．出力話者については 30 発話，入力話者については 10 発話を用いて学習した．
- **Combi**: INCA アルゴリズムと Soft INCA アルゴリズムのコンビネーション法による NMF を用いたノンパラレル声質変換法．4.6 節で述べたシステムと同様に，INCA アルゴリズムを 25 反復適用した後，Soft INCA アルゴリズムを 25 反復適用した．学習に用いた文は Soft と同様である．
- **INCA**: INCA アルゴリズムによる NMF を用いたノンパラレル声質変換法．学習に用いた文は Soft，Combi と同様である．
- **CycleGAN**: CycleGAN-VC によるノンパラレル声質変換法 [117]．オープンソースの実装[*8]を利用した．学習に用いた文は Soft，Combi，INCA と同様である．エポック数は 10000 とし，自然性を向上させるため 0 次のメルケプストラム係数については変換を行わなかった．
- **Para**: NMF を用いたパラレル声質変換法 [107]．入出力話者が発話したパラレルの 30 発話を学習に用いた．内容は，Soft で出力話者の学習に用いた 30 文と同一である．時間的アラインメントには，Affine-DTW [135] を用いた．Affine-DTW の反復数は 5 回とした．出力話者の基底を同一にするため，学習においてはまず出力話者発話の音響特徴量を分解し，得られた生起状態を利用して入力話者の基底を学習した．

NMF を利用する Soft，Combi，INCA，Para の 4 システムで用いた出力話者の基底はすべて同一である．4.2.2 節では，NMF を用いたパラレル声質変換法においては，まず入力話者の発話を分解し，次に出力話者の発話を分解すると述べたが，本実験では出力話者の基底を共通にするため，出力話者発話の次に入力話者発話を分解した．

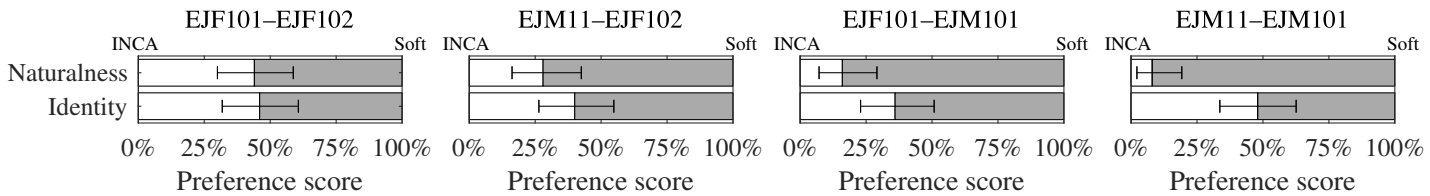
Soft，INCA，Combi の 3 手法間の比較

Soft，INCA，Combi の 3 手法の性能を評価した．図 4.13 にこの結果を示す．INCA と Soft の比較では，自然性において Soft が優れており，話者性においては有意な差はみられなかった．一方，INCA と Combi の比較では，自然性および話者性の両観点で Combi が優れていた．Combi は，高い自然性を保持しつつ，INCA や Soft と比較して高い話者類似性を持った変換を実現できることが確認された．4.6.3 節で述べた客観評価を考慮すれば，話者性においては INCA と Combi は同等の性能を持つと考えられるが，主観評価においては Combi が優れていた．この結果は，Combi が高い自然性を持つため，聴感上話者性においてもより優れて感じられたと考えられる．

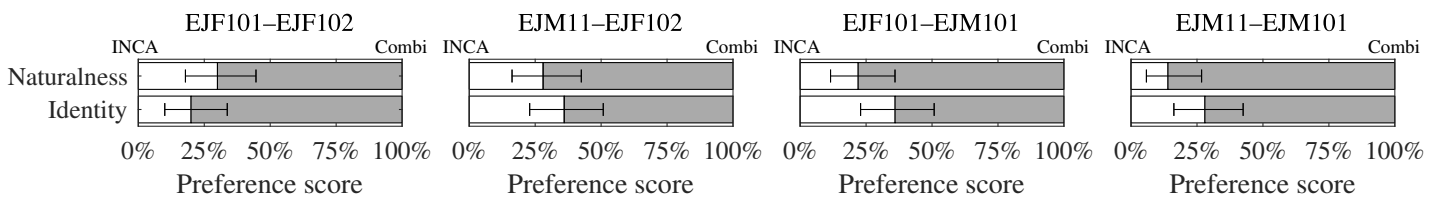
学習に用いる入力話者発話が 1 文の場合の Soft，INCA，Combi の 3 手法間の比較

学習に用いる入力話者の発話が 1 文の場合の，Soft，INCA，Combi の 3 システムの比較を行った．図 4.14 にこの結果を示す．前節で述べた学習に用いた発話が 10 文の場合と異なり，Combi だけでなく Soft についても INCA と比較して高い話者類似性を保持した変換が実現された．この結果は，学習に用いる発話数が少ない場合，INCA アルゴリズムはサンプル数の不足による影響を大きく受けるためであると考えられる．また，Soft と Combi の比

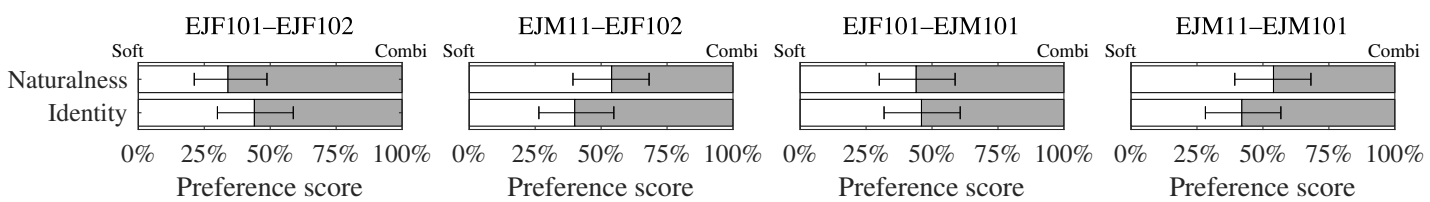
[*8] https://github.com/leimao/Voice_Converter_CycleGAN



(a) INCA と Soft の比較



(b) INCA と Combi の比較



(c) Soft と Combi の比較

図 4.13　Soft，INCA，Combi の 3 システムの主観評価による比較．エラーバーは 95% 信頼区間を示す．

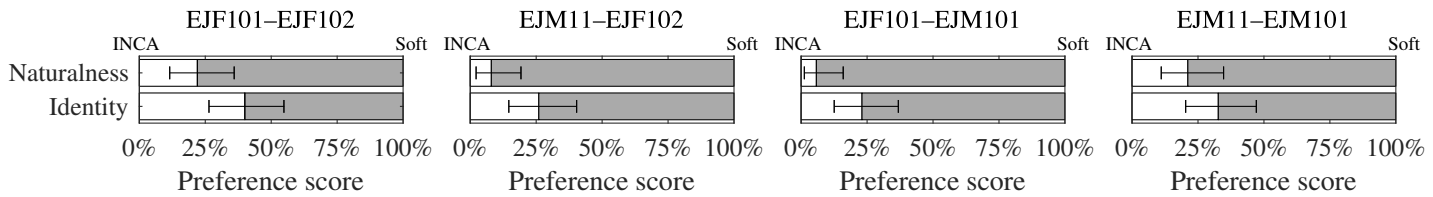
較においては，EJF101–EJF102 のペアを除いては，特に自然性において Soft が優れていた．これは，Combi における INCA アルゴリズムによる初期値がサンプル数の不足により劣化したためであると考えられる．

学習に用いる入力話者の発話数の違いによる影響

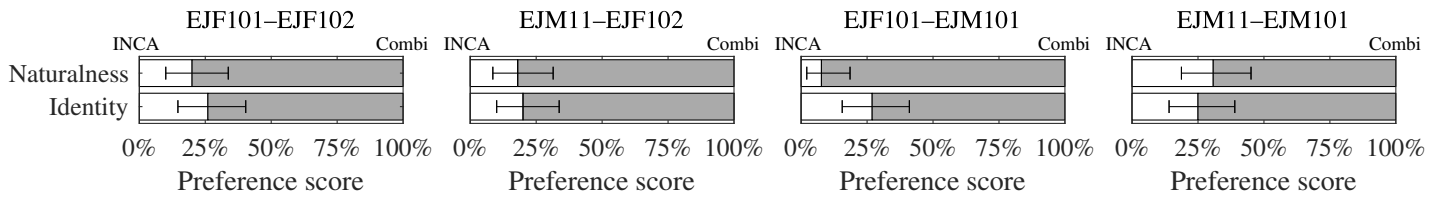
Combi において，学習に用いる入力話者の発話数が 1 文の場合と 10 文の場合を比較した．この結果を図 4.15 に示す．学習に用いた発話数が少ない場合，話者性もしくは自然性に影響が現れることが確認された．これは，入力話者の基底の品質が劣化した，すなわち言語的整合性が劣化したためであると推測される．

学習に用いる出力話者の発話数の違いによる影響

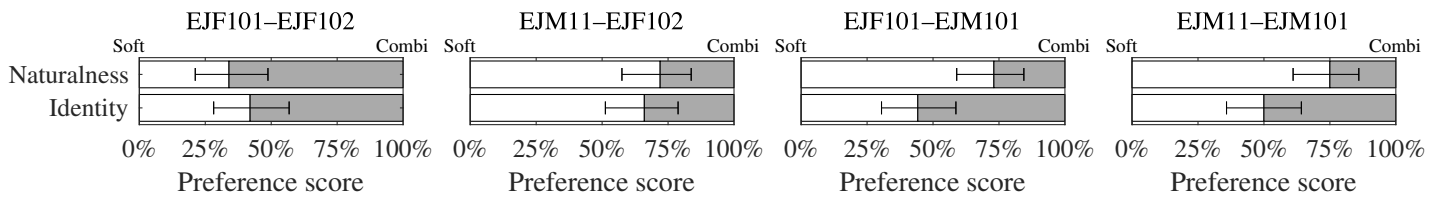
Combi において，学習に用いる出力話者の発話数を 5 文，10 文，30 文の間で変化させたときの変換品質を比較した．この結果を図 4.16 に示す．出力話者の学習に用いる発話数が多いほど，より高い変換品質が得られることが示された．10 文の場合と 30 文の場合の比較では，品質に有意差がみられる話者ペアとみられない話者ペアがあり，10 文の場合により品質が高いと評価された話者ペアも存在した．一方，5 文の場合と 30 文の場合の比較では，EJM11–EJF102 のペアを除き，5 文の場合の品質の劣化が顕著にみられた．これらの結果は，高い品質の出力話者の基底を得るためには十分な発話数が必要であり，この基底の品質が最終的な変換品質に影響することを示唆している．



(a) INCA と Soft の比較



(b) INCA と Combi の比較



(c) Soft と Combi の比較

図 4.14　学習に用いる入力話者の発話が 1 文の場合の，Soft，INCA，Combi の 3 システムの主観評価による比較．エラーバーは 95% 信頼区間を示す．

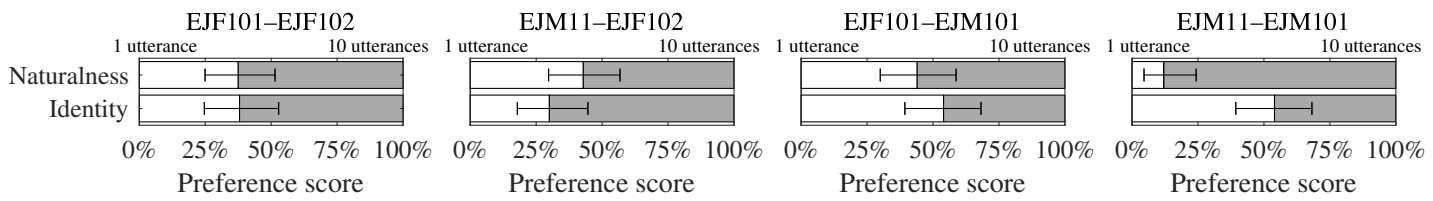
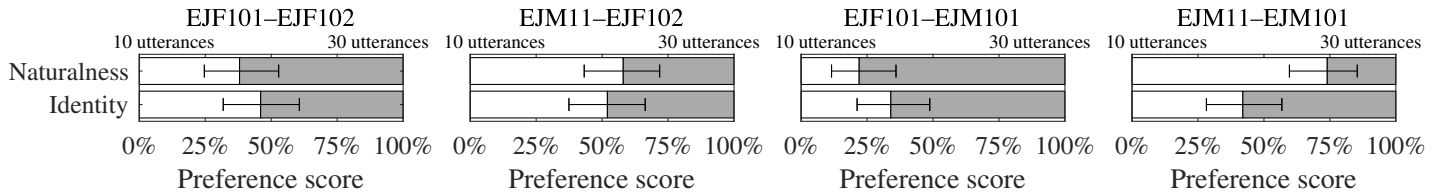
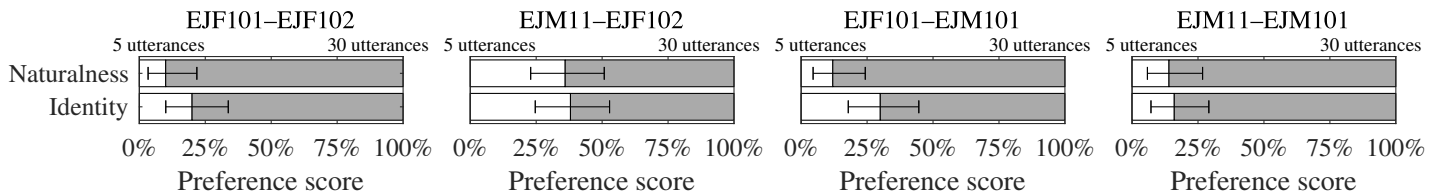


図 4.15　Combi における，学習に用いる入力話者の発話数の違いによる影響の主観評価．エラーバーは 95% 信頼区間を示す．



(a) 10 文の場合と 30 文の場合の比較



(b) 5 文の場合と 30 文の場合の比較

図 4.16　Combi における，学習に用いる出力話者の発話数の違いによる影響の主観評価．エラーバーは 95% 信頼区間を示す．

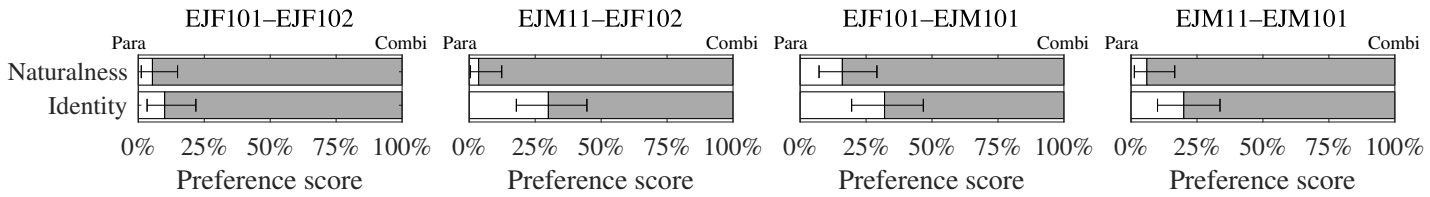


図 4.17　Combi と Para の主観評価による比較．エラーバーは 95% 信頼区間を示す．

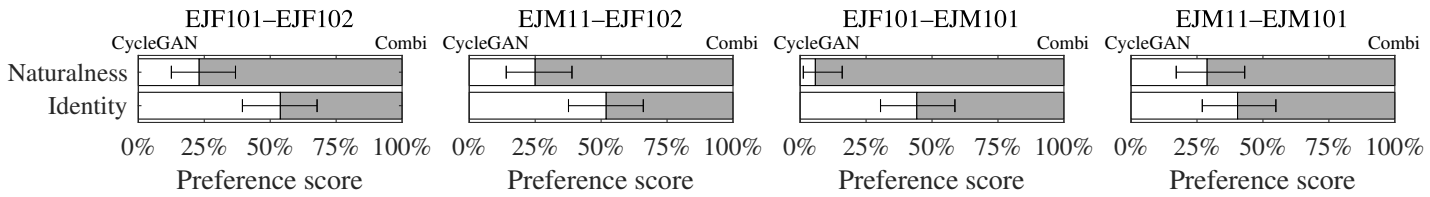


図 4.18　Combi と CycleGAN の主観評価による比較．エラーバーは 95% 信頼区間を示す．

Combi と Para との比較

ノンパラレルシステムである Combi と，パラレルシステムである Para の品質を比較した．この結果を図 4.17 に示す．いずれの話者ペアに対しても，Combi が有意に高い品質を示した．Para は，Affine-DTW を利用しているものの，DTW によるアラインメントの不整合が影響して品質が低下したと考えられる．

Combi と CycleGAN との比較

NMF を用いたシステムである Combi と，まったく異なるシステムである CycleGAN を比較した．この結果を図 4.18 に示す．自然性において，CycleGAN と比較して Combi は有意に高い品質を示した．一方，話者性においては有意差はみられなかった．本実験のような少量の発話のみ利用できる環境では，自然性の観点で Soft INCA アルゴリズムが効果的であることが示された．

4.7.3　異言語間変換

異言語間変換における Soft INCA アルゴリズムの有効性を調査するため，異言語間の声質変換システムを構築し，その性能を同一言語間変換の場合と比較した．異言語間変換の場合には，学習に用いる出力話者の発話を英語の音素バランス文とした．これにより，英語の話者である出力話者に日本語を発話させるシステムが構築される．英語と日本語に基本周波数のわずかな違いがみられたため，基本周波数の変換モデルは同一言語間変換で用いたものをそのまま利用した．

結果を図 4.19 に示す．入力話者が EJF101 の場合では，話者性もしくは自然性において，同一言語間変換の場合に有意に高い変換品質を示した．一方，入力話者が EJM11，出力話者が EJM101 の場合では，異言語間変換において有意に高い自然性が認められた．異言語間変換の場合には，出力話者の基底の学習に用いる発話の言語が異なるため，その基底によって入力話者の日本語の発話が分解できるかどうかが品質に大きな影響を及ぼすと考えられる．したがって，EJF101 の発話は英語の基底によって分解されにくく，一方で EJM11 の発話は英語の基底を

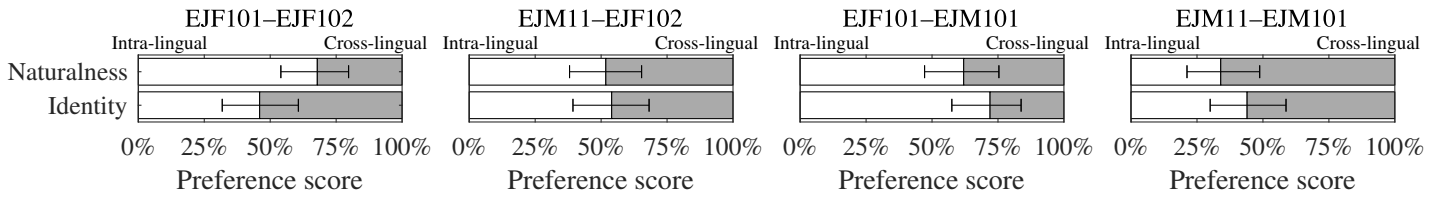


図 4.19　システム Combi における，同一言語内変換と異言語間変換での主観評価による品質の比較．エラーバーは 95% 信頼区間を示す．

用いても分解できる，などの入出力話者に依存した結果が導かれたと推測される．

4.7.4　One-shot システム

本節では，one-shot システムにおける提案法の有効性について確認する．ここでいう one-shot システムとは，入力話者の発話を学習に用いることなく，直接それを変換することをいう．すなわち，入力話者に関しては学習の必要のないシステムをいう．

INCA アルゴリズムや Soft INCA アルゴリズムでは，パラレルデータの生成を反復中に行うため，このパラレルデータをそのまま利用することで one-shot の声質変換システムを構築できる．本論文では，INCA アルゴリズムの場合には，変換後の音響特徴量 $\boldsymbol{y}_{p_i(n)}^{(t)}$ を用いて合成する．これは，最近傍法によって得られたパラレルデータは不連続で不自然なためである．一方，Soft INCA アルゴリズムの場合には，NMF の再構成によって得られた $\boldsymbol{X}_i$ を用いて合成する．図 4.20 に示すように，これらの変換後の音響特徴量は過平滑化による影響が無視できないため，global variance [101][*9]を補償するようにポストフィルタリングを適用した．

INCA，Soft，Combi の 3 手法を比較した結果を図 4.21 に示す．INCA と Combi の比較においては，話者性については有意差がみられなかったが，自然性においてはとくに異性間変換において Combi が有意に高い性能を示した．また，Combi と Soft においても，自然性において Soft が有意に高い変換性能を示した．これらの結果を総合して，話者性においては大きな違いがみられないものの，自然性においては Soft，Combi，INCA の順に高いことが示された．Soft において Combi と比較して高い自然性の変換が実現されており，この結果は 4.7.2 節で示した声質変換システムの評価の結果とは異なっている．これらの結果は，INCA アルゴリズムと比較して Soft INCA アルゴリズムがより自然な中間特徴量を生成すること，また Combi は INCA による不自然な初期値によって最終的な品質が劣化しやすいことを示唆している．

本論文では，one-shot システムに加えて，one-utterance システムの性能についても評価した．ここでいう one-utterance システムとは，one-shot システムと同様のデータの条件において，声質変換システム全体を構築する手法をいう．すなわち，入力話者の発話 1 文を用いて学習し，同じその発話を変換するシステムをいう．したがって，入力話者の基底を学習および利用する操作により合成音声の品質が変化することが期待される．INCA，Soft，

[*9] 合成音声の品質を示す値の 1 つで，発話内での音響特徴量の分散のこと．この値が自然な発話の値と比較して極度に小さい場合，ぼやけた不自然な音声であることを示唆する．

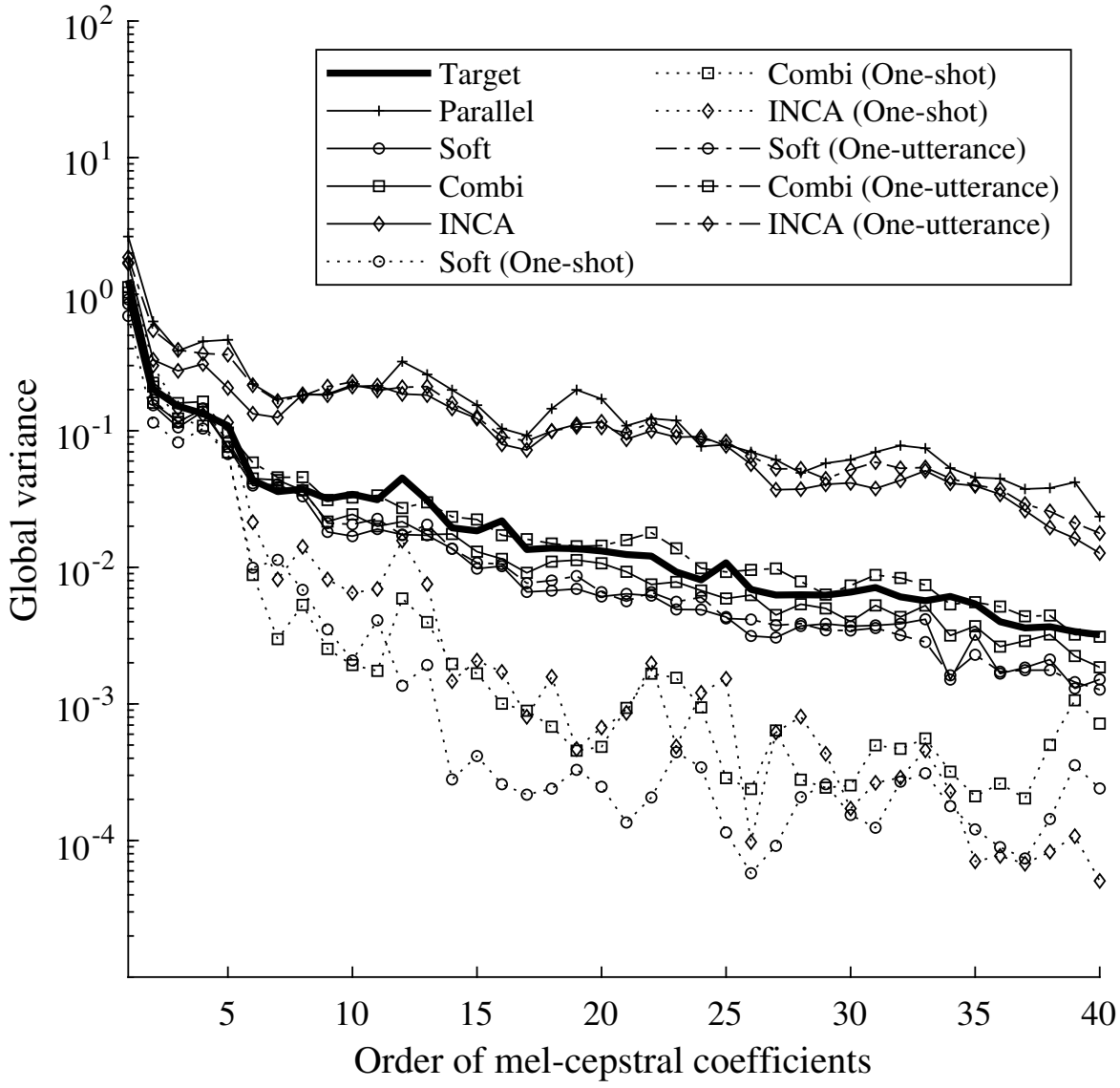


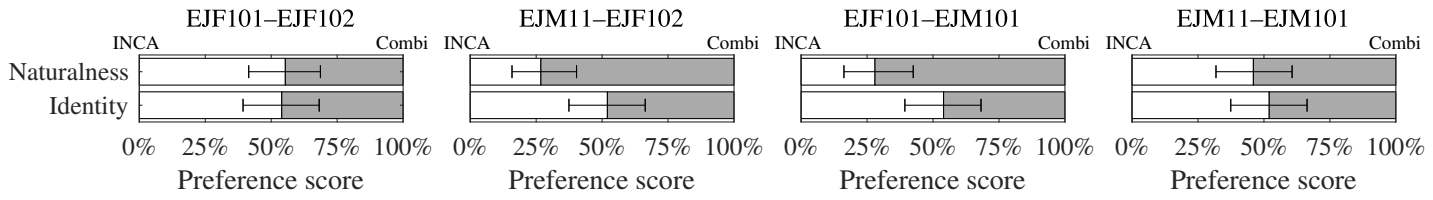
図 4.20　変換された音声の global variance．入力話者は EJM11，出力話者は EJF102 である．変換されたすべての発話に対して global variance を計算し，その平均を示す．統計モデルを利用して生成された音響特徴量は，global variance の小さい過平滑化されたものになる場合があり，これを考慮することで聴感上品質の高い音響特徴量を生成できる．

Combi の各手法において，one-shot システムと one-utterance システムを比較した結果を図 4.22 に示す．INCA については，one-utterance システムと比較して one-shot システムにおいて高い自然性および話者性を示した．一方，Soft については，one-utterance システムにおいて高い自然性および話者性を示した．また，Combi については，自然性については one-utterance システムが，話者性においては one-shot システムが高い性能を示した．

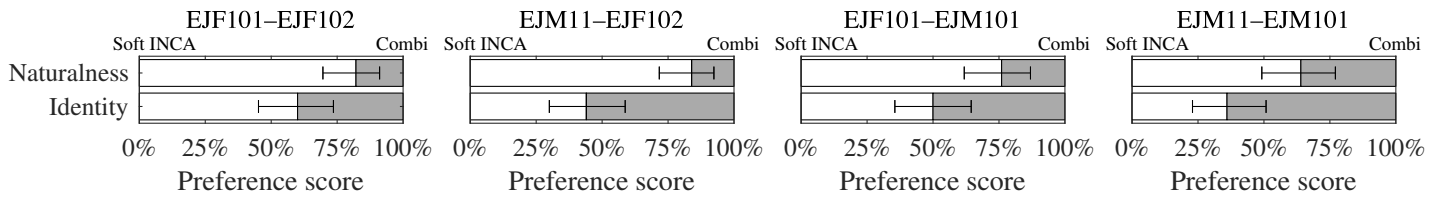
4.7.5　メルケプストラム歪み

すべてのシステムを客観的に評価するため，すべてのシステムに対して MCD の値を計算した．ここでは，合文法無意味文のうちはじめの 50 文を変換し，各文の MCD の値の平均を計算した．

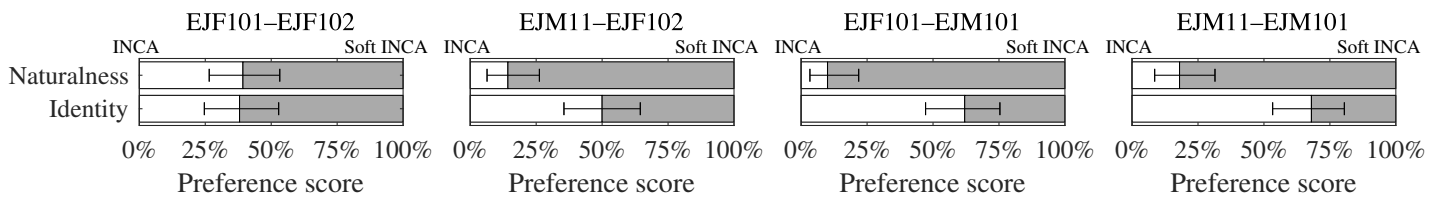
表 4.4 にこの結果を示す．Para および INCA は，他の手法と比較して高い MCD を示した．また，ノンパラレ



(a) INCA と Combi の比較



(b) Soft と Combi の比較



(c) INCA と Soft の比較

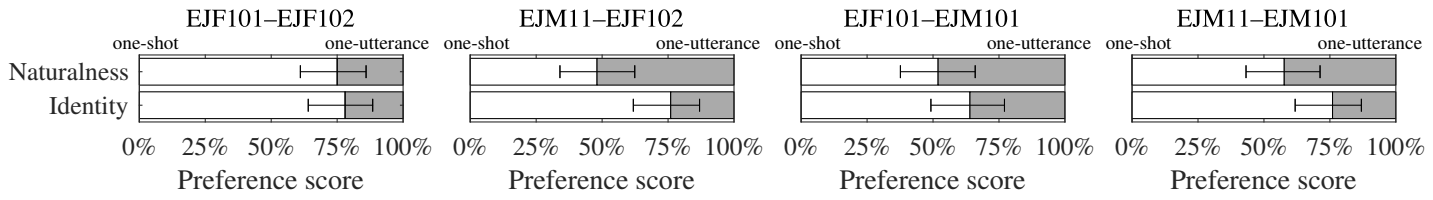
図 4.21　Soft，INCA，Combi による one-shot システムの主観評価による品質の比較．エラーバーは 95% 信頼区間を示す．

ルかつ入力話者の発話数が 10 文の条件では，CycleGAN がもっとも低い MCD を示した．一方，one-shot の条件では，Soft がわずかに高い MCD を示したものの，手法間に大きな差はみられなかった．One-utterance システムにおいては，Soft がもっとも MCD が低く，通常のノンパラレル声質変換と同等の品質で変換できることが示された．

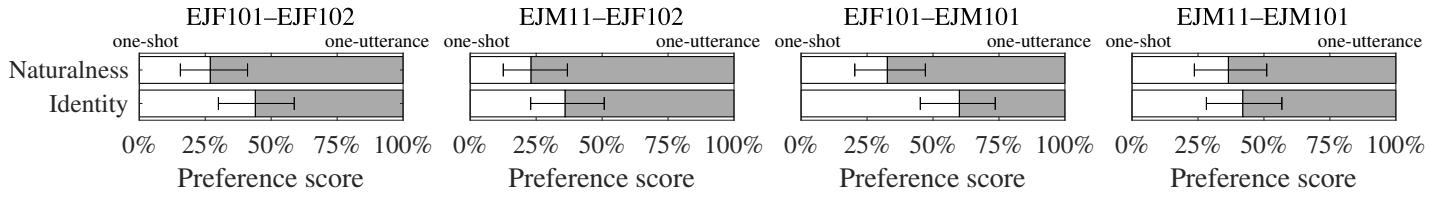
4.7.6　平均オピニオン評点による評価

すべてのシステムを主観的に評価するため，平均オピニオン評点（mean opinion score; MOS）による評価を行った．自然性については，1（まったく自然でない），2（やや自然でない），3（どちらでもない），4（やや自然である），5（とても自然である）の 5 段階から，話者性については，1（完全に異なる話者である），2（どちらかというと異なる話者だと思う），3（どちらかというと同じ話者だと思う），4（完全に同じ話者である）の 4 段階から評価させた．AB 選好試験，ABX 試験の場合と同様に，クラウドソーシングサービスを通じて少なくとも 25 人の被験者が各システムに対して 2 問答えた．

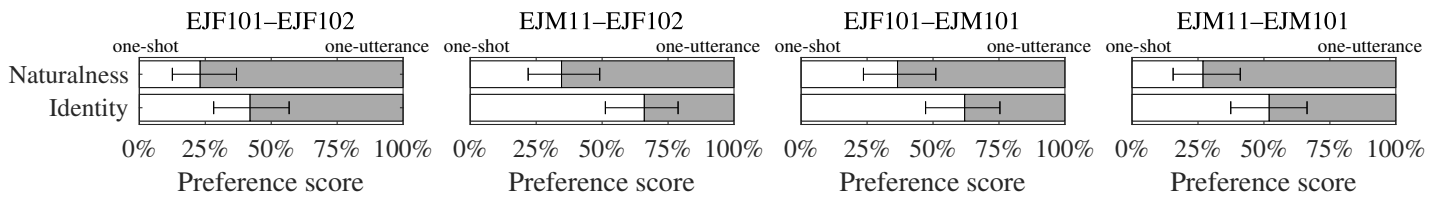
表 4.3 にこの結果を示す．自然性においては，INCA や Para の MOS 値が低い一方，Soft や Combi については 3.3 以上の評価を得た．また，one-utterance における Soft についても，3.29 と通常のノンパラレル条件と同等の自然性で変換できることが示された．一方，話者性においては，いずれの手法も 2 を下回っており，十分な変換品



(a) INCA



(b) Soft



(c) Combi

図 4.22　INCA，Soft，Combi それぞれのシステムにおける，one-shot システムと one-utterance システムの主観評価による品質の比較．エラーバーは 95% 信頼区間を示す．

質が得られたとは言い難い．

4.8　考察

4.8.1　中間特徴量の品質

4.7.4 節で述べた one-shot システムの評価の結果は，中間特徴量からそのまま合成した音声を用いており，中間特徴量の品質そのものを評価していると解釈できる．実験結果によれば，Soft INCA アルゴリズムは INCA アルゴリズムと比較して自然な中間特徴量を生成することが示された．INCA アルゴリズムは，フレーム毎の離散的なアラインメントを得るため，不適切なアラインメントが得られやすいことが可能性として考えられる．この問題は，系列情報を利用した INCA アルゴリズムの拡張によって緩和できる可能性がある [136, 137]．一方，Soft INCA アルゴリズムは連続的なアラインメントを得るため，不自然なアラインメントを抑制できる．また，INCA アルゴリズムと Soft INCA アルゴリズムのコンビネーション法においては，Soft INCA アルゴリズムのみの場合と比較して自然性の低い中間特徴量が得られた．これは，コンビネーション法では INCA アルゴリズムによって得られた不自然な中間特徴量を初期値として利用しているためであると考えられる．話者性においては，客観評価においては Soft INCA アルゴリズムが低い性能を示しているものの，主観評価では 3 手法間に大きな違いはみられなかった．

表 4.3　すべての変換条件における MOS 評価値．自然性については 1 から 5 の 5 段階評価，話者性については 1 から 4 の 4 段階評価による．「同性間」の列は同性ペア EJF101–EJF102 および EJM11–EJM101 の平均を，「異性間」の列は異性ペア EJM11–EJF102 および EJF101–EJM101 の平均を，「平均」の列は全 4 ペアの平均を示している．

条件	手法	学習に用いた発話数		自然性			話者性		
		入力話者	出力話者	同性間	異性間	平均	同性間	異性間	平均
	入力話者（自然音声）			—	—	4.82	1.36	1.04	1.20
	出力話者（自然音声）			—	—	4.95	—	—	3.61
パラレル		30	30	1.84	1.52	1.68	1.97	1.86	1.92
ノンパラレル	INCA	10	30	2.92	2.32	2.62	2.00	1.40	1.70
	Soft	10	30	3.60	3.11	3.35	1.92	1.52	1.72
	Combi	10	30	3.63	3.13	3.38	1.96	1.39	1.68
	CycleGAN	10	30	2.54	1.78	2.16	2.11	1.58	1.84
	Combi	1	30	3.25	2.85	3.05	1.88	1.42	1.65
	Combi	10	10	3.78	2.94	3.36	1.87	1.40	1.64
	Combi	10	5	2.72	2.43	2.57	1.59	1.44	1.51
	Combi（異言語間）	10	30	3.90	2.95	3.43	1.93	1.37	1.65
One-shot	INCA	1	30	2.21	1.36	1.78	2.15	1.71	1.93
	Soft	1	30	2.50	2.31	2.41	1.91	1.53	1.72
	Combi	1	30	2.14	1.51	1.82	1.97	1.67	1.82
One-utterance	INCA	1	30	1.90	1.40	1.65	2.01	1.53	1.77
	Soft	1	30	3.59	2.99	3.29	1.83	1.40	1.62
	Combi	1	30	3.03	2.42	2.73	2.12	1.52	1.82

この理由として，INCA アルゴリズムやコンビネーション法で得られた音響特徴量が不自然で歪んでいるため，主観的な話者性が劣化したことが考えられる．

4.8.2　声質変換システム全体を用いた場合の品質

声質変換システム全体を利用した場合の結果は，one-shot システムの場合とは異なるものであった．Soft INCA アルゴリズムは 4.6.3 節や 4.7.5 節で示したように，中間特徴量の話者性が低い．この問題がコンビネーション法によって改善されたことで，聴感上の話者性においてもコンビネーション法は優れた変換性能を示したと考えられる．また，コンビネーション法は，INCA アルゴリズムと比較して自然性，話者性ともに高い性能を示した．これは，INCA アルゴリズムではアラインメントが不自然で，入力話者の基底が劣化したためであると考えられる．また，コンビネーション法は Soft INCA アルゴリズムと同等の自然性での変換が可能であることが示された．One-shot の条件ではコンビネーション法の自然性が低かったものの，入力話者基底を介すことでこの問題が緩和

されたと考えられる．全体として，コンビネーション法によって，Soft INCA アルゴリズムと同等の自然性をもち INCA アルゴリズムと同等の話者性をもつ変換が実現できることが示された．

4.8.3　アラインメントの誤りによる影響

主観評価や客観評価の結果から，INCA アルゴリズムを用いた場合やパラレル声質変換では，他の手法と比較して自然性が低く MCD の高い音声が合成されたことが確認された．これらの結果は，NMF を用いた声質変換法がアラインメントの誤りに影響を受けやすいことを示唆している，One-shot システムと one-utterance システムの比較においても，この問題が示唆されている．Soft INCA アルゴリズムやコンビネーション法では，連続的なアラインメントを利用するため，この問題を抑制できると考えられる．したがって，連続的なアラインメントを得るという提案法のアプローチが，アラインメントの誤りに影響を受けやすい NMF 声質変換において有効であることが示唆された．

4.8.4　出力話者音響モデルと言語的整合性の品質

4.7.2 節で示した結果から，入力話者の発話数について 10 文で学習した場合と 1 文で学習した場合では，1 文で学習した場合に自然性や話者性への影響がみられた．同一の出力話者の基底を利用していることから，1 文の場合は言語的整合性が劣化していると解釈できる．したがって，言語的整合性の低下が話者性や自然性へ影響を及ぼすことが確認された．提案法は少量の発話を有効に活用して変換モデルを構築できるものの，高い品質の変換には依然として十分な量の発話が必要であることが示唆された．同様に，学習に用いる出力話者の発話数が少ない場合についても自然性や話者性への影響がみられ，出力話者の基底を高い品質で得るためには十分な学習データ量が必要であった．また，異言語間変換では，出力話者の音響モデルすなわち出力話者基底の言語が異なることによって，自然性や話者性が低下する場合がみられた．したがって，出力話者の基底や一時変換のモデルの品質，すなわち出力話者の音響モデルや言語的整合性の品質によって，話者性や自然性に影響が及ぶことが示唆された．一方で，異言語間変換においては，話者のペアによっては異言語間の場合に品質がより高い場合があり，また同言語間と異言語間で MCD に大きな差がみられないことから，入出力話者間の言語の違いが必ずしも変換品質の低下には導かれないことも確認された．これは，異言語の音声を利用しても出力話者の基底を十分な品質で構築できる場合があることを示している．

4.8.5　One-shot システムと one-utterance システムの比較

One-shot システムと one-utterance システムの比較により，INCA アルゴリズムや Soft INCA アルゴリズムの NMF 声質変換における有効性を判断できる．INCA アルゴリズムにおいては，one-shot システムは one-utterance システムと比較して高い自然性および話者性の変換を実現した．これは，上述したアラインメントの誤りの影響によるものと考えられる．NMF 声質変換はアラインメントの誤りによって品質が劣化しやすいため，NMF 声質変

換を介した one-utterance システムは性能が低くなったと考えられる．一方で，Soft INCA アルゴリズムでは逆に，one-utterance システムが高い性能を示した．変換時の生起状態の Wiener エントロピー[*10]の平均を比較すると，one-utterance システムでは 0.1331，one-shot システムでは 0.5530 であり，one-utterance システムがより疎な生起状態を利用していた．これによって，出力話者の基底がよりスペクトルテンプレートのように利用されたため，より自然な音声が合成できたと考えられる．これは，NMF 声質変換を組み込んだことによる効果である．コンビネーション法では，INCA アルゴリズムと Soft INCA アルゴリズムの結果がともに発生したため，話者性と自然性で異なる結果が得られたと考えられる．すなわち，NMF 声質変換を利用したことでより自然な合成が実現された一方，話者性においては非連続的で不自然なアラインメントの影響を受けたと考えられる．総括すると，one-shot の条件においては，Soft INCA アルゴリズムを利用した one-utterance システムが，もっとも品質の高い音声を合成できると結論付けられる．

4.8.6　CycleGAN-VC との比較

CycleGAN-VC は，MCD においては高い性能を示した一方，主観評価においてはコンビネーション法と比較して話者性に有意差がなく自然性が低いことが示された．CycleGAN-VC は，メルケプストラム係数を規準としてより変換後の話者に近くなるよう学習を行うため，メルケプストラム係数の観点ではより目標話者に近づけることができる．一方，過平滑化によって音声としての自然性が低下する．NMF 声質変換は過平滑化が起こりにくく，提案法の枠組みでは自然性が低下しにくいため，主観的には CycleGAN-VC と比較して高い自然性の音声が合成できると考えられる．主観評価実験によって，話者性においては提案法と CycleGAN-VC は同等の変換性能を示したため，CycleGAN-VC と比較して提案法の有効性が確かめられた．

4.9　本章のまとめ

本章では，短時間の歌声から声質変換モデルを学習することを目的として，Soft INCA アルゴリズムとよばれる新たなノンパラレル声質変換法を提案した．この手法は，INCA アルゴリズムに着想を得たもので，NMF の分解すなわちアラインメントと変換モデルの学習を交互に反復して行うものである．INCA アルゴリズムと異なり，出力話者の音響モデルを利用した音響特徴量の補間によって入力話者発話の音響特徴量を表現し，そのアラインメントが連続的であるため，より自然性の高い変換を実現できる点が特徴である．主観評価により，Soft INCA アルゴリズムは，既存の INCA アルゴリズムや CycleGAN-VC と比較して，同程度の話者性で高い自然性の音声が合成できることが示された．また，INCA アルゴリズムと Soft INCA アルゴリズムのコンビネーション法により，さらなる品質の向上が可能であることを示した．本章では one-shot 声質変換，すなわち入力話者を学習せずに行う声質変換についても評価を行い，提案法により優れた品質で one-shot 変換を行うことが可能であることを示した．

[*10] 特徴量，特にスペクトルなどの平坦さを示す指標．幾何平均を算術平均で除算したもの．

表 4.4　すべての変換条件における MCD [dB].

条件	手法	学習に用いた発話数		話者ペア			
		入力話者	出力話者	EJF101–EJF102	EJM11–EJF102	EJF101–EJM101	EJM11–EJM101
入力話者（自然音声）				6.77	7.29	7.20	6.49
パラレル		30	30	10.74	11.11	9.52	10.25
ノンパラレル	INCA	10	30	7.82	10.04	8.71	8.36
	Soft	10	30	6.00	6.32	6.72	6.06
	Combi	10	30	6.00	6.46	6.69	6.28
	CycleGAN	10	30	5.35	5.93	6.27	5.58
	Combi	1	30	6.40	6.61	7.18	6.58
	Combi	10	10	6.23	6.65	6.56	5.93
	Combi	10	5	7.14	7.57	6.72	6.34
	Combi（異言語間）	10	30	5.97	6.69	6.83	5.98
One-shot	INCA	1	30	6.24	6.42	6.54	6.20
	Soft	1	30	6.57	6.89	7.10	6.37
	Combi	1	30	6.14	6.31	6.78	6.30
One-utterance	INCA	1	30	10.37	10.61	10.55	9.87
	Soft	1	30	6.07	6.44	6.72	6.21
	Combi	1	30	6.68	6.84	7.29	6.45

第 5 章

VocalRemixer: 歌唱者ダイアライゼーションと声質変換を利用した音楽鑑賞アプリケーション

5.1　はじめに

音楽鑑賞技術は，1877 年にエジソンによって発明された蓄音機をはじめとして，人々の音楽鑑賞を豊かにしてきた．古典的な音楽鑑賞は，音楽を選んで再生し，それを聴取するという，いわば受動的な音楽鑑賞である．一方で，音楽情報処理技術，とくに音楽理解技術を利用して，アプリケーションをインタラクティブに操作できる能動的音楽鑑賞技術が広く検討されている [138]．能動的音楽鑑賞技術の方向性は，おもに音楽再生，音楽加工（タッチアップ），音楽検索（ブラウジング）の 3 つに分類される．

音楽再生の能動的音楽鑑賞技術には，音楽の再生位置の変更（シーク）や音楽のスキップといったナビゲーション機能，音楽の可視化，歌詞の表示などの技術が挙げられる．音楽のナビゲーションによって，好きな音楽の好きな箇所を再生したり，気に入らない楽曲をスキップしたりなどの操作が可能になる．音楽の可視化は，鑑賞している音楽により没入できる効果がある．歌詞の表示においても，カラオケ画面のように今再生されている箇所をハイライトなどすることで，歌詞を目で追うことなく容易に歌詞を把握および解釈し，音楽をさらに楽しむことができる．これらの技術は種々のアプリケーションにおいて既に導入されているが，その場で演奏される音楽の鑑賞では成されなかった音楽情報処理による成果である．こうした技術と音楽理解技術を組み合わせたアプリケーションとして，SmartMusicKIOSK [139]，Cindy [11]，Songle [19]，TextAlive [21] などが提案されている．SmartMusicKIOSK は，楽曲のサビの自動検出技術である RefraiD [140] を利用することでサビ区間を検出し，楽曲のサビのみを選択的に試聴できるインタフェースである [139]．知らない楽曲に対して，楽曲のサビを手作業で探すことなく確認することができ，好きな音楽を探す作業を効率化できる．Cindy は，音楽のビートを自動で推定するビートトラッキング技術を利用し，ビートに同期して 3D のダンサーが踊る様子を表示するインタフェースである [11]．音楽鑑賞システム Songle では，SmartMusicKIOSK のようなサビの頭出し機能のほかに，図 5.1 に示すようにメロディー，コード進行，ビートを可視化するインタフェースを備えている [19]．TextAlive は，VOCALOID 楽曲のミュージックビデオなどにみられる歌詞をダイナミックに表現する kinetic typography を自動で生成し表示するインタフェースである [21]．従来モーショングラフィックスソフトウェアなどを利用しなければ制作できなかった歌詞を印象的に見せるミュージックビデオを，楽曲と歌詞の自動アラインメント技術を利用して自動で生成することが可能である．

音楽加工の能動的音楽鑑賞技術には，楽曲を自分好みに修正したり改変したりする技術が含まれる．たとえば，グラフィックイコライザなどの信号処理を利用したエフェクタによる音楽加工は，音楽プレーヤなどに広く導入されている．また，音源分離技術などを活用することで，音楽的な情報を活用した高次の音楽加工技術も考えられる．INTER:D は，楽曲中のドラム音を能動的に変更できるインタフェースである [141]．ドラム音の発音時刻を自動で推定し，バスドラムとスネアドラムの音量や音色を変化させることができる．Drumix とよばれるドラムパターンの修正を可能にするインタフェースも提案されている [142]．図 5.2 のようにドラムパターンを各小節ごとに詳細に編集することが可能である．

音楽検索の能動的音楽鑑賞技術には，好みの音楽や新しい音楽を効率的，効果的に発見する技術が含まれる．従

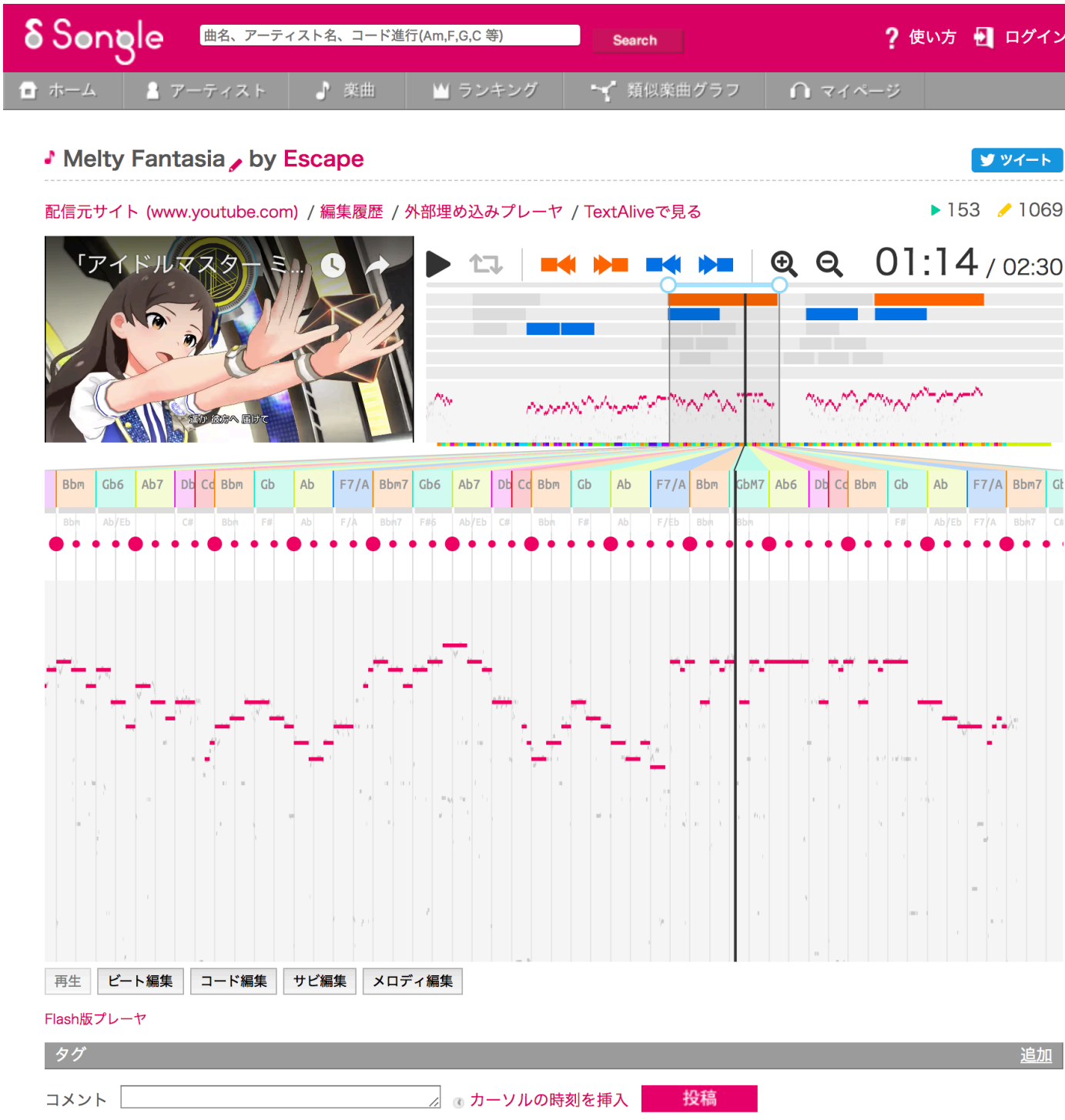


図 5.1　能動的音楽鑑賞サービス Songle [19] のスクリーンショット．左上に YouTube に投稿された動画が再生され，右上に楽曲構造，中段にコードとビート，下段に旋律が可視化されている．ビート，コード，旋律，楽曲構造の情報はすべて自動解析によって推定されるが，最下部に表示されている編集ボタンをクリックすることで編集することができる．画像中の楽曲では各情報は手作業で修正されている．

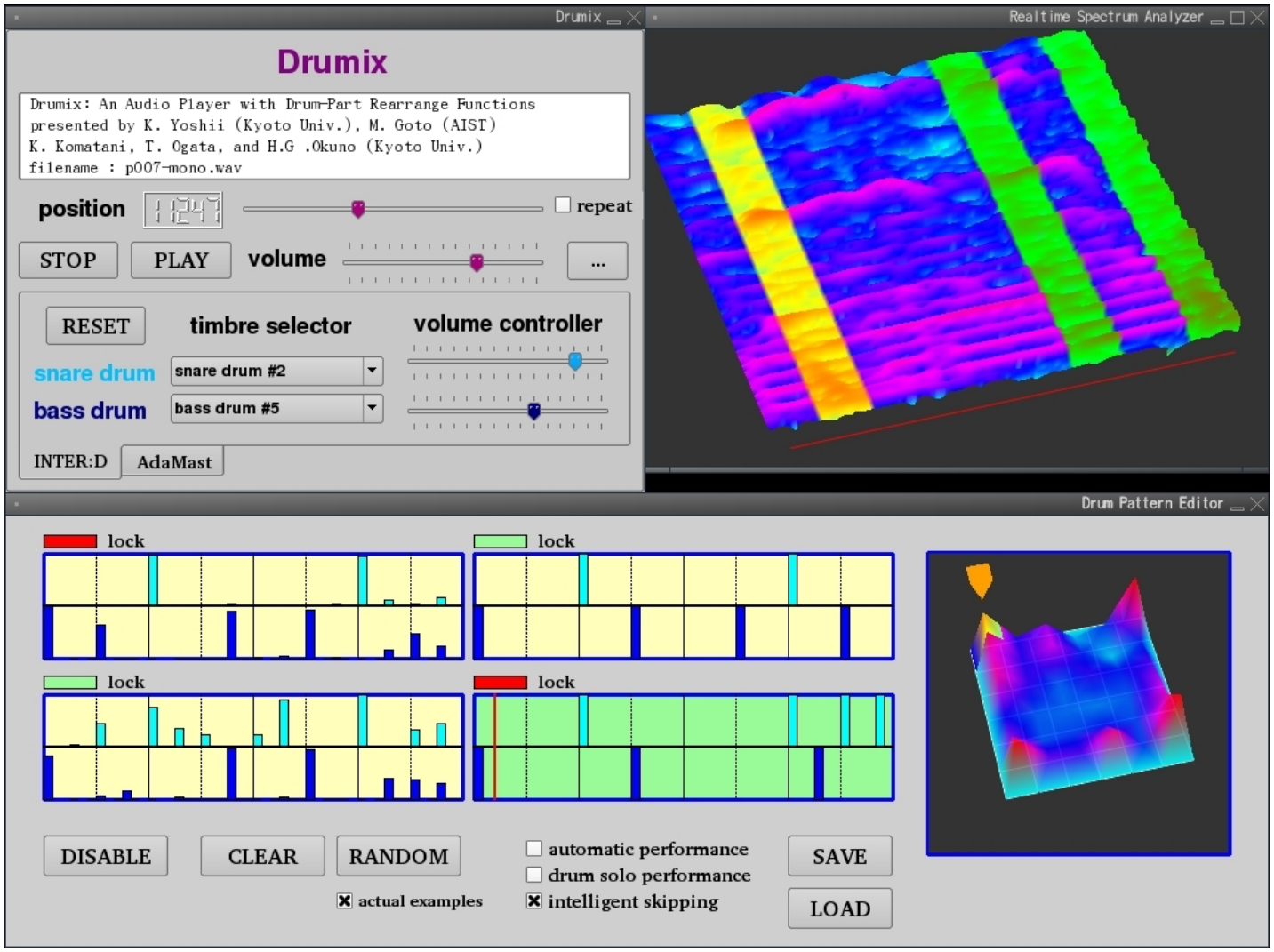


図 5.2　ドラムパートのリアルタイム編集機能を持つオーディオプレーヤ Drumix [142] のスクリーンショット[*1]．左上の画面では，全体の音量やナビゲーションなどの操作ができるほか，スネアドラムとバスドラムの音を選択できる．右上の画面では，リアルタイムにスペクトログラムが表示され，編集を行う前と後の小節がそれぞれ緑色および黄色で示されている．下の画面では，ドラムパターンを詳細に編集することができる．手動でドラム音を追加する機能や，小節ごとに無作為にドラムパターンを切り替える自動リアレンジ機能が利用できる．ドラムパートのみをソロで再生したり，お気に入りのパターンを保存したり読み込んだりすることも可能である．

来の音楽検索技術は，曲名やアーティスト名などの人手で与えたメタデータを利用していた．Apple Music などの音楽サブスクリプションサービスでは，大規模なユーザの嗜好情報を活用して，音楽を推薦する機能が備えられている．音楽理解技術を活用して，自動認識されたジャンルやムードなどを起点に検索する手法も考えられる．myMoodplay は，ムードを基準に楽曲の位置を可視化する Web インタフェースである [143]．Songrium は，ニコニコ動画などに投稿された VOCALOID を利用して制作された楽曲などの繋がりを可視化することで，投稿された楽曲を発見できる Web インタフェースである [144]．こうしたインタフェースは，2 次元や 3 次元の星図状に楽曲の位置を可視化している．Musicream は，新たな楽曲を発見できるインタフェースであり，楽曲を流し聴きしたり，楽曲の雰囲気にもとづいて類似楽曲を検索したりする機能を提供している [145].

[*1] http://sap.ist.i.kyoto-u.ac.jp/members/yoshii/drumix/

本章では，能動的音楽鑑賞技術のテーマのうち，音楽再生と音楽加工の 2 つに着目して，複数人が歌唱するパート割りのある楽曲を能動的に鑑賞できる Web インタフェースを提案し，これを VocalRemixer とよぶ．歌声を加工する技術としては歌声の声質変換技術が考えられるが，本インタフェースは同一楽曲内で複数人の歌唱者が歌唱していることに着目して，歌唱者内での声質変換を行う．これによって，オリジナルの楽曲にはないパート割りを合成および再現できる．

本章では，まず VocalRemixer の概要とその実装について説明する．次に，VocalRemixer の主観評価実験について述べる．また，VocalRemixer をさらによい音楽鑑賞インタフェースとして発展させるための構想について述べる．

5.2 VocalRemixer の概要

複数人が歌唱する楽曲には，複数人がそれぞれ順番にソロで歌いサビを全員で歌うなどといったパート割りが用いられることがあり，アイドルソングやアニメの主題歌などのような楽曲にみられる．このようなパート割りは，あらかじめ楽曲の制作者が意図して決定するもので，基本的に楽曲の聴き手はそれを変更することはできない．しかし，本来のパート割りとは異なるパターンのパート割りを聴くことができたり，あるいはある歌唱者が単独で歌唱しているかのような歌声を聴くことができるようになれば，1 つの楽曲をさらに楽しむことができると考えられる．メディアミックスコンテンツである『アイドルマスターシャイニーカラーズ』や『ラブライブ!』などでは，本来複数人で歌唱されていた楽曲を各歌唱者が単独で歌唱したトラックを収録した CD が発売されており[*2]，複数人が歌唱する楽曲に対して異なる聴き方ができる魅力が示唆されている．このような CD はすべての複数人歌唱楽曲で提供されているわけではなく，好みの楽曲でこうした楽しみ方をすることは制作者から提供されている音楽のみを利用している限り難しい．そこで本論文では，声質変換を利用して，このような複数人歌唱楽曲に対して異なる聴き方を与えるインタフェースを提案する．

図 5.3 に VocalRemixer の外観を示す．VocalRemixer は，楽曲のナビゲーション機能，パート割りの表示機能，パート割りの手動編集機能，パート割りの一括編集機能を持つ．

- **楽曲のナビゲーション**．再生，一時停止といった基本的な再生機能に加えて，パート割り表示部をクリックすることによる特定の時刻へのジャンプ機能を持つ．
- **パート割り表示**．VocalRemixer は楽曲のパート割りを可視化する．各歌唱者について，歌唱している区間を青色で，歌唱していない区間を灰色で表示する．楽曲を開いた初期状態として，オリジナルのパート割りが表示される．
- **パート割り手動編集**．パート割り表示部の，青色および灰色のセルをクリックすることで，各歌唱者について各区間の歌唱非歌唱を切り替えることができる．全体を表示した状態では編集しづらいため，

[*2] https://shinycolors.lantis.jp/releaseinfo/lacz-10090_1/ や https://www.lantis.jp/release-item/LACA-39331.html など．

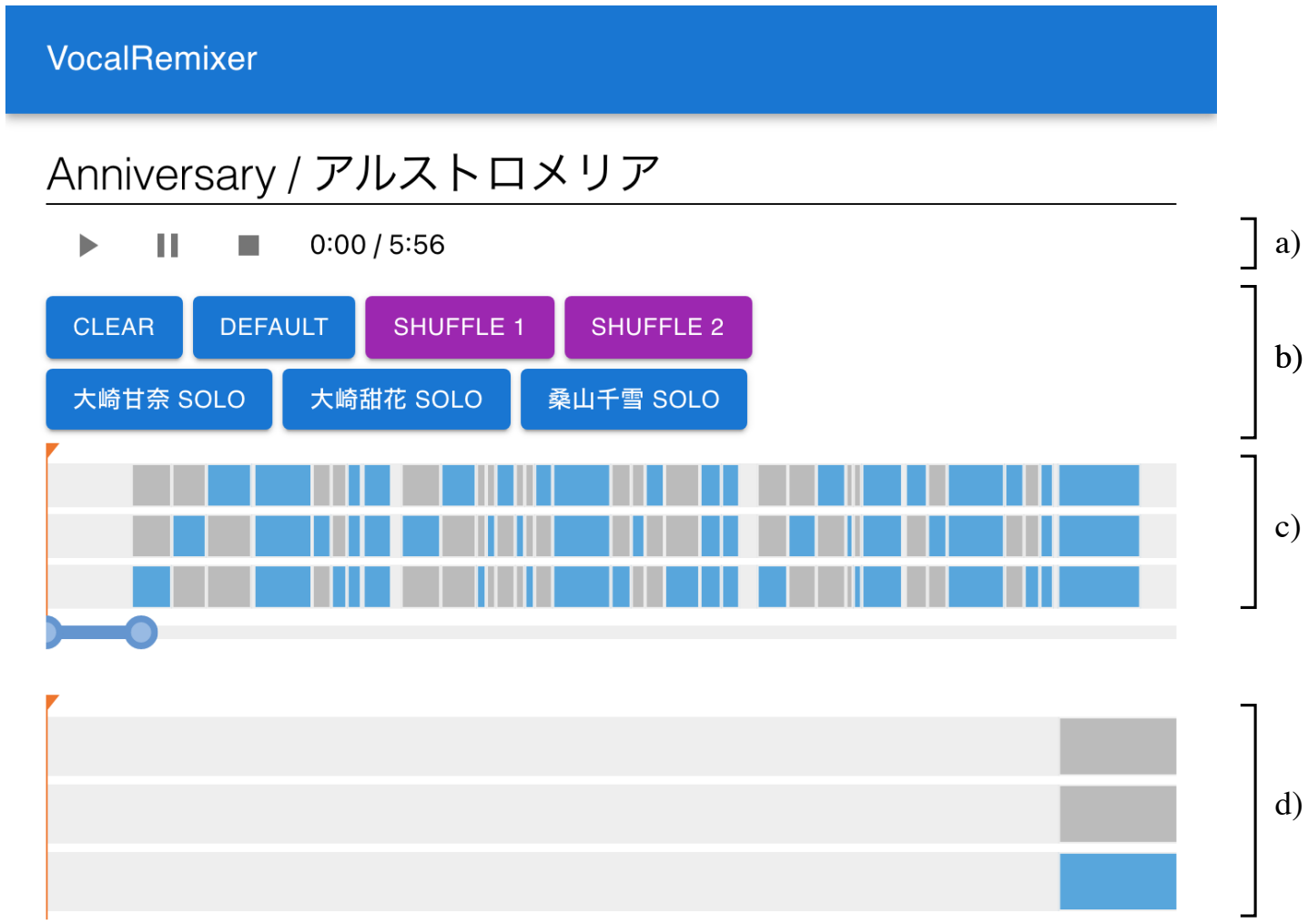


図 5.3　複数人歌唱楽曲鑑賞インタフェース VocalRemixer の画面．a) ナビゲーションボタン，b) パート割り一括変更ボタン，c) パート割り表示，編集画面，d) c) の拡大表示．

VocalRemixer では拡大表示を採用している．拡大表示部分は楽曲の進行に自動で追従する．

- **パート割り一括編集**．すべての歌唱者をオフにして伴奏音のみを再生する機能，初期状態であるオリジナルのパート割りに復元する機能，特定の歌唱者に全区間を単独で歌唱させる機能，すべての区間の歌唱非歌唱をシャッフルする機能を持つ．シャッフルには 2 つのアルゴリズムを実装している．第 1 のアルゴリズムは，歌唱者同士を入れ替える方式である．たとえば歌唱者 A，B，C の歌唱する楽曲を仮定し，歌唱者 B が本来歌っているパートを歌唱者 A に，歌唱者 C が本来歌っているパートを歌唱者 B に，歌唱者 A が本来歌っているパートを歌唱者 C に，といった要領で無作為に入れ替える．第 2 のアルゴリズムは，各区間の歌唱者の人数のみを固定し，歌唱者を無作為に選択する方式である．同一の歌唱者が 2 つの区間を連続して歌唱しにくくなるよう，1 次マルコフモデルを採用して確率分布を決定する．サビを全員で歌うなどといった本来の演出を維持しつつ，まったく異なるパート割りを演出することが可能である．

5.3 VocalRemixer の実装

5.3.1 インタフェースの実装

Web インタフェースは，スクリプト言語 JavaScript，Web フロントエンドライブラリ React[*3]，CSS フレームワーク Material UI[*4]を利用して実装した．インタフェースでは，楽曲の曲名やアーティスト名を記したメタデータ，パート割りの区間やオリジナルの歌唱者を記したラベル，各歌唱者がソロ歌唱した歌声，楽曲の伴奏音を利用して構築した．すべての歌声と伴奏音は時間的に同期されている．

5.3.2 歌声と伴奏音の分離

このインタフェースでは，歌声と伴奏音を分離した音響信号が必要である．CD などに収録されているミックスされた楽曲からこれらを分離することが理想的であるが，分離の性能が低く，声質変換時の分析合成で大きく歌声の品質が劣化する．そのため本論文では，A.2.1 に示した伴奏音分離法で分離された伴奏音や，別収録された伴奏音を用いて歌声の分離を行う．

5.3.3 各歌唱者がソロ歌唱した歌声が利用できる場合のデータ作成

各歌唱者がソロ歌唱した歌声が利用できる場合，それらをそのまま用いればよく，声質変換を行う必要がない．この場合，A.2.1 に示した分離法を用いることで，伴奏音のない歌声と伴奏音を用意する．

5.3.4 各歌唱者がソロ歌唱した歌声が利用できない場合のデータ作成

パート割りのあるミックスされた歌声のみが利用できる場合，声質変換によって変換および合成することで，各歌唱者があたかもソロで歌唱したかのような歌声を用意できる．VocalRemixer では，第 4 章で提案した声質変換法を用い，次の手順でこの歌声を用意する．この手順を図 5.4 に示す．

1. **歌唱者ダイアライゼーション**．第 3 章に提案した手法などを用いて歌唱者ダイアライゼーションを行う．本論文での実装にあたっては，自動推定を行わず，あらかじめ聴取によって手作業で生成したパート割り情報を用いた．
2. **歌唱者それぞれの音響モデルの学習**．各歌唱者がソロで歌唱している区間を利用して，各歌唱者の音響モデルを学習する．
3. **変換モデルの学習**．歌唱者の組み合わせを入力，各歌唱者を出力として，変換モデルを学習する．入力には，たとえば A と B の歌唱者が同時に歌唱した場合など，現れる歌唱者の組み合わせすべてについて変換

[*3] https://reactjs.org/
[*4] https://v4.mui.com/

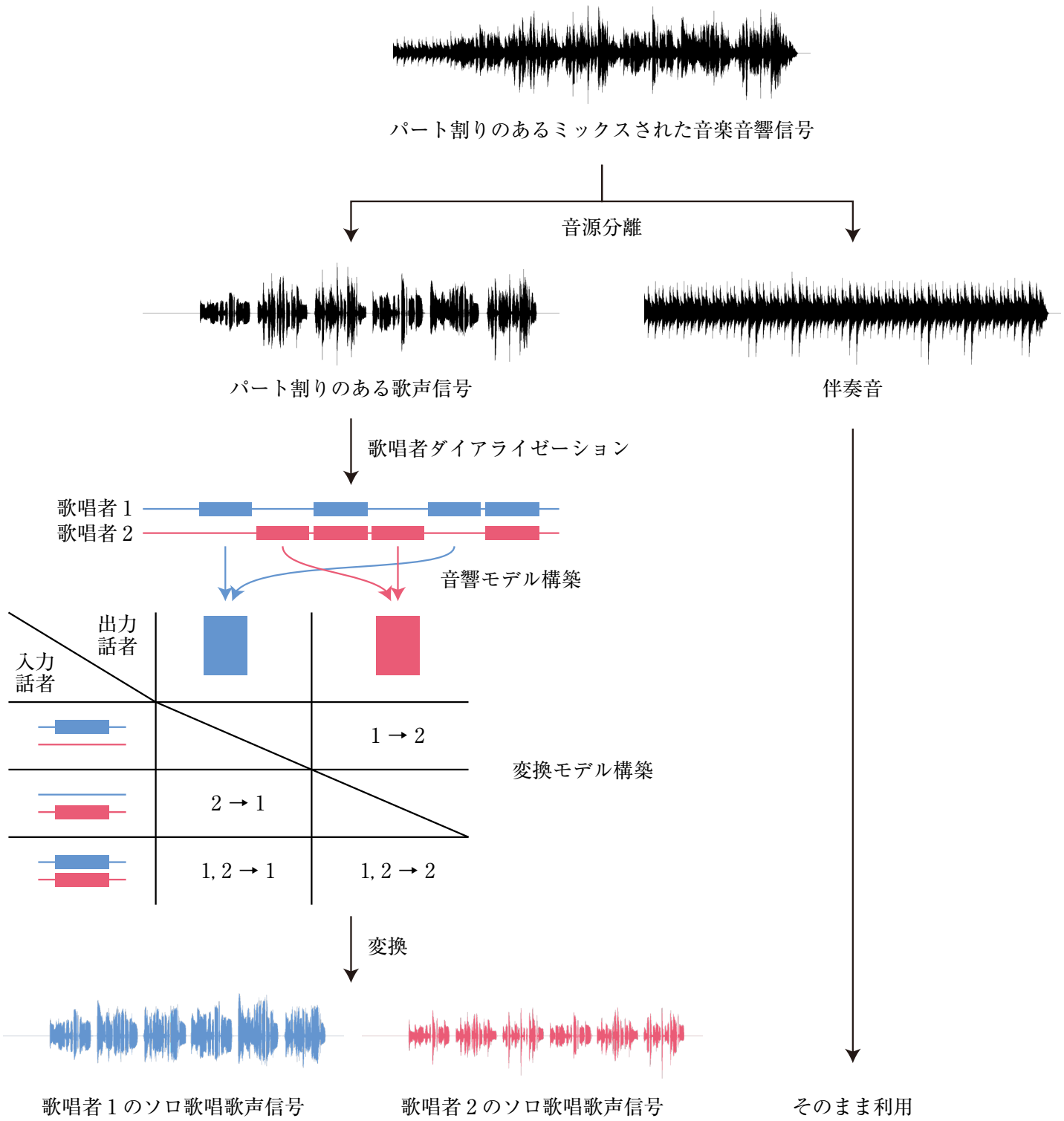


図 5.4　パート割りのある楽曲を入力した場合の，VocalRemixer で利用するソロ歌唱信号と伴奏音信号を作成する手順．2 人の歌唱者がパート割りのある楽曲を歌唱した場合の例を示している．実験では，音源分離や歌唱者ダイアライゼーションについて，ブラインド音源分離や自動的な歌唱者ダイアライゼーションではなく，手作業で与えた理想的なデータを用いる．

モデルを学習する．仮に歌唱者が A と B であった場合，A から B，A と B の同時歌唱から B，B から A，A と B の同時歌唱から A，の 4 通りについて変換モデルを学習する．

4. **変換**．歌声全体を，学習した変換モデルを用いて変換し，それぞれの歌唱者があたかもソロで歌唱したかのような音響信号を合成する．

この手順では，歌声のパラレルデータを用意できないため，ノンパラレル声質変換法が必要である．またこの手順では，複数の歌唱者が同時に歌唱している歌声を，分離せずにそのまま分析，変換，合成する．本章ではユニゾンの歌声を仮定しているため本手順が適用できるが，異なるピッチのパート割りが存在する場合には，主旋律のみを選択的に利用して変換および合成する必要がある．

本手法では，合成時にすべての歌唱者において同一の基本周波数系列を用いる必要がある．仮にすべての歌唱者の基本周波数系列が同一の場合，同時に歌唱させるとあたかも単一の歌唱者が歌唱しているかのような歌声に聴こえる．これを解決するため，VocalRemixer では基本周波数系列に各歌唱者によって位相の異なる正弦波によるゆらぎを与える．このゆらぎは，次式で与えられる．

$$\hat{f}_0 = (1+\alpha)^{\sin\left(\frac{2t}{T}+\frac{2n}{N}\right)\pi} f_0 \tag{5.1}$$

ここで，f_0 および $\hat{f}_0$ はゆらぎ前後の基本周波数，α はゆらぎの幅，t は時刻，T はゆらぎの周期，N は歌唱者数，n は歌唱者のインデックス（$0 \leq n \leq N-1$）である．

5.4　主観評価

本節では，童謡『かたつむり』を用いて行った評価実験について説明する．

5.4.1　実験条件

本実験では，歌声コーパス JVS-MuSiC [146] に収録されている，童謡『かたつむり』を歌唱した歌声を用いた．歌唱者には，同一の性別で，同一の調で歌唱した歌声が収録されている JVS004 および JVS010 を選んだ．楽曲『かたつむり』に対して 2 歌唱者で歌唱するパート割りを表 5.1 のように作成し，これをオリジナルのパート割りとして利用した．本実験では，次の 2 つの場合について評価を行った．

- 声質変換を用いない場合．**VC なし**と表現する．JVS-MuSiC に収録されている歌声と，筆者が作成した伴奏音を用いてインタフェースを構成した．
- 声質変換を用いる場合．**VC あり**と表現する．JVS-MuSiC に収録されている歌声を用いてパート割りのある『かたつむり』の歌声部分をミキシングによって作成し，これを利用して声質変換により各歌唱者のソロの歌声を合成したものを利用した．すなわち，オリジナルのパート割りとして歌唱している区間のみが与えられたものと仮定して声質変換を用いてソロの歌声を作成した．

表 5.1　実験で用いたパート割りのある『かたつむり』のパート割り.

歌唱者	歌詞
JVS004	でんでんむしむし かたつむり
JVS010	おまえのあたまは どこにある
JVS004，JVS010	つのだせやりだせ あたまだせ
JVS010	でんでんむしむし かたつむり
JVS004	おまえのめだまは どこにある
JVS004，JVS010	つのだせやりだせ めだまだせ

実験に用いた VocalRemixer のインタフェースは
https://www.gavo.t.u-tokyo.ac.jp/~hitoshi/vocalremixer/ で体験できる.

『かたつむり』の歌声には JVS-MuSiC [146] に収録されているものをそのまま用いた．サンプリング周波数は 24 kHz である．歌唱者ダイアライゼーションの結果として，ミックス時に利用した正解をそのまま用いた．伴奏音には，筆者が楽譜作成ソフトウェア Finale[*5]を利用して作成したピアノによる伴奏音を利用した.

主観評価にはクラウドソーシングサービスを利用し，アンケート形式で調査を行った．各被験者は報酬を支払われ，被験者数は 50 人であった.

5.4.2　VC ありの場合の実験条件

歌声の分析合成には WORLD [131]（D4C edition [132]）を用いた．合成時には WORLD のバリエーションである Requiem を用い，自然性を向上させるためゼロ位相フィルタによる音声合成を行った．フレーム周期は 1 ms とした．基本周波数には 5.3.4 節に示した方法でゆらぎを与えた．ゆらぎの周期 T は 0.4 s，ゆらぎの幅 α は 0.01 とした．このゆらぎの幅はおよそ 17 cents に相当する．合唱経験者による合唱の基本周波数の平均帯域幅が ±25 cents 程度である [44] ことから，この値に近い $\alpha = 0.01$ を選んだ．非周期性指標については変換を行わなかった.

NMF には 256 次のメル絶対値スペクトログラムを用い，分解規準には一般化 KL ダイバージェンスを用いた．基底数は 128 とした．より識別性の高い基底を得るため，出力話者の基底を学習する際には，GMM によりメルケプストラム係数をクラスタリングしたのち，この平均ベクトルから基底ベクトルを生成し，これを初期値として NMF を行った．短時間の歌声から各歌唱者の音響モデルを緻密に構築するため，NMF には最小体積 NMF [147] を 1000 反復行ったのち，通常の一般化 KL ダイバージェンス規準の NMF を 10 反復行った．最小体積 NMF では，最小化するダイバージェンスを次式とする.

$$\mathcal{D}(\boldsymbol{Y} \mid \boldsymbol{X}) = \mathcal{D}_{\mathrm{KL}}(\boldsymbol{Y} \mid \boldsymbol{X}) + \lambda \log\det\left(\boldsymbol{H}^{\top}\boldsymbol{H} + \delta \boldsymbol{I}\right) \tag{5.2}$$

[*5] https://www.finalemusic.com/

表 5.2　VocalRemixer のソロ歌唱信号を生成する際の，Soft INCA アルゴリズムで用いる変換モデルのスケジュール．

反復	変換モデル	パラメータ数
1–10	声道長変換	1
11–20	GMM 声質変換（$M = 1$）	500
21–30	GMM 声質変換（$M = 2$）	1,001
31–40	GMM 声質変換（$M = 4$）	2,003
41–50	GMM 声質変換（$M = 8$）	4,007

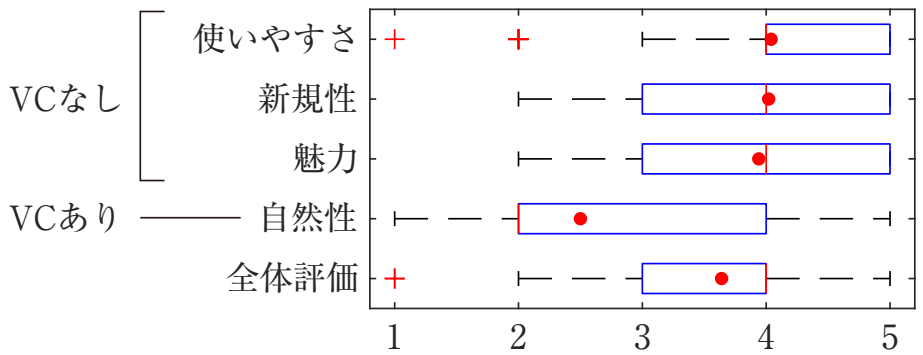


図 5.5　VocalRemixer に関する主観評価の結果の箱ひげ図．赤丸は平均を，赤＋は外れ値を表す．

ここで，λ は体積によるペナルティ項の重み，δ は行列式の計算を安定化させるための小さな値である．本実験では λ に 0.1，δ に 1 を選んだ．また，反復中で用いるパラメータである γ の初期値には 0.5 を選んだ．

変換には，第 4 章で提案した Soft INCA アルゴリズムを用いた．反復数は 50 とし，表 5.2 に示すスケジュールで一時変換モデルを徐々に複雑化した．

5.4.3　VC なしについての評価

使いやすさ，新規性，魅力について，5 段階で評価した．使いやすさについては「このアプリは使いやすいと思いましたか」，新規性については「このアプリには，目新しさ，斬新さはあると思いますか」，魅力については「もし好きな楽曲を入力して使えるようになったとしたら，このアプリはどの程度魅力的だと思いますか」と質問した．回答は，「5—そう思う」「4—ややそう思う」「3—どちらともいえない」「2—あまりそう思わない」「1—思わない」といった 5 段階の選択肢から選択させた．図 5.5 に結果を示す．MOS は，3.9 から 4.0 であった．多くの人が使いやすいと回答し，また新規性や魅力についても認められた．

また，「このアプリの使いやすい・使いにくい点や，アプリに関する感想などがありましたら自由に教えてください。」と質問し，自由記述による感想を収集した．表 5.3 にすべての感想を示す．「操作が直感的である」「簡単で使いやすい」といった意見や，「自由に組み合わせを変えられるのがおもしろい」といった意見がみられた．一方で，「用途が多いとは思えない」「あまり魅力を感じない」といった意見もみられた．また，インタフェースのデ

表 5.3　VC なしについてのインタフェースに関して収集された感想．複数のカテゴリにわたる感想が書かれていた場合には分割した．

カテゴリ	感想
わかりやすい	使い方が簡単。
	感覚的にパートを弄れるので分かりやすい
	直感的に使えて便利。子供でも高齢者でも使えると思う。
	操作が簡単でとても使い勝手は良いと思います。
	シンプルで使いやすいです。
	見た目がわかりやすくて、クリックのみで操作できるので使いやすいです。
	簡単で使いやすいと思います
	最初は見たときはどう操作すればよいかわからなかったが、2 分ほどでだいたい理解できた。再生位置を変えられるのはよい。
	操作が簡単で使いやすいと思う
	直観的で操作し易かったです
	使い方が感覚的にわかりやすく、すぐに思ったように扱えるようになりました。
	単純、シンプルで使いやすいと思います。
	直感的に使えてわかりやすい。
	説明文を読んだ後とは言え、音楽の素人の私でも直感的に自由自在に使うことができたのでこのアプリは使いやすいと思いました。
	直感的に操作することができますので、操作性はいいと思いました。
おもしろい，おもしろそう	追加機能しだいで面白くなりそう。
	ランダムで声が切り替わる点が面白い。
	歌っている最中にも歌い分けを操作できるのは良い。
	簡単なボタン操作でカラオケ・シングル・混声が出来るのが面白い。
	自分の好きな曲で使ってみたら楽しいだろうなと思いました
	操作が簡単でわかりやすく使いやすかったです。クリックすると歌唱・非歌唱が切り替わったり、伴奏のみを聞けたり面白いなと思いました。
	自由に歌うパートを指定できたり面白いアプリだと思いました。
	歌い分けの区間で歌っている最中でも、クリックするとその場で切り替わって音声が変わるので、楽しめると思う。
	初めての感じなので新鮮な気分です
	多声の楽曲で自由に組み合わせを変えられるのは面白いと思います.
	このアプリを使用して好きな曲を色々聞いてみたいです。
	ボタンを押すだけで色々なバージョンに出来るのが良いと思いました。
	複数の声を合成したり切り替えたり、自分で好きにリミックスできてすごく面白いと思いました。自分が好きな歌手や、好きな曲でこれをやれたら、とても楽しいと思います。
用途がわからない，わかりづらい	使い方は簡単だと思いますが、使用用途が歌声を消してカラオケに使用するくらいしか思いつきませんでした。
	最初、説明を読んで用途がわからなかった
	魅力についてはそれほど用途が多いとは思えません。
	アプリそのものは使いやすいとは思うが現時点では良くも悪くもそれだけであまり魅力を感じません。
	切り替えも早いしスムーズなので使いやすいと思いますが、個人的にはただ好きな曲を聴くだけなのでこうやって切り替えて遊ぶことはあまりしないと思います。
	これを音楽鑑賞用として、どういうシチュエーションで使うのかということを考えてはみたのですが、なかなかいい答えが出せないでいます。カラオケですとか、好きなアーティストさんとのコラボなどに使えそうだとは思うのですが。
UI に関する問題点	子供向けの場合は英語表記や現状の説明では難しかなと思いました。
	CLEAR など、英語より日本語で書いてあるとわかりやすいと思いました。
	幅広い年代で使えるよう日本語表記してもらえると良いと思う。
	まだまだ UI がわかりにくい。説明のページを見なくても、なんとなく操作がわかるような作りが必要。ポインターをバーに置くと自動的に拡大したバーが下段に現れるが、非表示や自動的に拡大しないようにするなどした方が便利。
	CLEAR ボタンに関してですが、押すとボーカルトラックが消える、までは非常に自然なのですが、さらにもう一度押すことでボーカルトラックが復活する、という設定にした方が良いのでは、と感じました。操作性が向上する気がします
	再生中に好きな部分の飛べるともっとよかったと思います。(そういうことができたのかも知れませんが、色々とやってみてもできませんでした)[*6]
	今まであまり無かったアプリだと思うが、その分使い方が難しい所がある。
	開発中だと思いますが、UI デザインが付けばもっと良くなると思います
	もう少し色やデザインが入ると良くなりそうですね。
	シンプルに視覚的に使うには面倒に感じてしまいます。
	アーティストごとに色分けする
とくにズームに関する意見	ズームの調整を上手く使いこなせませんでした。
	拡大表示の必要性はあまり感じない。
	上下段で幅の違いがほぼ無く、拡大表示の意味がまるで無い
	箇所の拡大バー？ が操作感と違和感がある。
	区間の幅が大きいためズームの恩恵がない。そもそもズーム機能は必要ないのでは
とくにシャッフルに関する意見	「SHUFFLE」がもう少したくさんあるとよいと思います。
	シャッフルはいらない
	シャッフルボタンを連打してみたのですが、途中でシャッフルがならなくなります。
	シャッフルについては、歌唱が斉唱になる部分が必ずあるようなので意味が無い。全区間でランダム化するべき
	シャッフル機能はあまり必要性を感じませんでした。
その他	この「かたつむり」のように、歌い手の区別がはっきりとついている場合には、歌い分けの聴き比べをする楽しさが加わる明確なメリットがあると思います。
	音データーの分析と合成とがどれくらいうまくゆくかが、このアプリの品質を左右すると思います。
	音程も自由に変えられると面白いと思った
	この 2 種類の声に自分の声も合わせられるともっと楽しく使えるのではないかと思った。
	アーティストごとにボリューム設定を加える・アーティストごとにパン（音が聞こえる位置）の設定を加える
	伴奏も消せるとよい

ザインについても「歌唱者ごとに色を変えてほしい」などの問題点が指摘され，とくにズームやシャッフル機能の必要性についての意見が多くみられた．ズーム機能については，VocalRemixer は本来 4 分程度あるポップスの楽曲を対象としてデザインされており，『かたつむり』が短くパート割りが単純なためこのような意見が多かったと考えられる．

5.4.4　VC ありについての評価

声質変換を用いた場合の自然性について，「声質変換技術を用いた合成音声は自然でしたか」と質問し，5.4.3 節と同様に 5 段階で評価させた．図 5.5 に結果を示す．MOS は 2.5 で，不自然であるとする回答が多くみられた．『かたつむり』の場合，入出力話者の発声時間がそれぞれ 10 秒程度であるため，高い品質でのノンパラレル声質変換が行えず，不自然であったと考えられる．

5.4.5　インタフェース全体に関する評価

インタフェース全体に関する評価として，「もし好きな楽曲を入力でき，自然に合成できたとして，このような音楽鑑賞アプリを使ってみたいと思いますか」と質問し，5.4.3 節と同様に 5 段階で評価させた．図 5.5 に結果を示す．MOS は 3.6 で，やや使ってみたいとする意見が多くみられた．

また，「アプリを使ってみた感想や，こういう機能がほしいなどの意見がありましたらコメントしてください」と質問し，自由記述による感想を収集した．表 5.4 にすべての感想を示す．「面白そうと感じた」といった意見や「簡単で使いやすい」といった意見がみられた一方，「有用な活用方法が思いつかない」といった意見がみられた．追加機能については，「歌詞が書いてあると便利」といった意見や，「声質や楽器音を自由に変えられるとよい」や「テンポやキーが変えられる機能がほしい」といった楽曲をアレンジできる機能が提案された．「自分の声を入れて一緒に歌えるとよい」などの，自分の声を入れたいという意見もいくつかみられた．また，特定のユーザのアレンジを共有できる機能などを提案する意見もみられた．声質変換の自然性が低い問題はあるものの，肯定的な意見が多く得られた．

5.5　さらなる拡張の構想

本節では，VocalRemixer をさらに魅力的なインタフェースにするための様々な拡張の構想について述べる．

5.5.1　任意の楽曲を扱える機能

YouTube やニコニコ動画などの動画投稿サイトに投稿された楽曲や，ユーザの持っている楽曲を利用できる機能が考えられる．これを実現するためには，自動的な歌唱者ダイアライゼーションをはじめとした分析技術の構築

[*6] VocalRemixer には，パート割り表示部のすぐ上をクリックすることでシークできる機能が備わっているが，容易に発見できなかったと思われる．

表 5.4　インタフェースの全体に関して収集された感想．複数のカテゴリにわたる感想が書かれていた場合には分割した．

カテゴリ	感想
おもしろかった，使ってみたい	個人的には初めての感じのアプリだったので結構面白かったです。 自分の好きな曲をアレンジできるのは面白そうと感じました。 ランダムで声が切り替わる点が面白かった。 本格的に開発が進めば、面白いアプリになると思う。 今まで使ったことがないアプリなので新鮮です。好きな楽曲などこのアプリで色々と変えて楽しんでみたいと思いました。ただ音楽を聴くだけではなくて、自分であれこれ変えることで今までにない楽しみ方が出来そうで気になります。 最近のアニソンは声優に歌わせるタイプが多く、パートが分かれてたり、同じ曲でも複数のキャラがそれぞれ歌ってたりするので、好きなキャラだけをチョイスできそうなこのアプリは面白いと思います。 好きな曲でそれぞれのパートを単独で聴いてみたり等、好きなように聞けるのはとても魅力的な機能なので色々試してみたいです。 私も癒されたいときにかわいい声で童謡を使って、このアプリを楽しんでみたいと思いました。 面白いアイデアで、既存の曲でも使えるなら試してみたい。
使わないと思う	面白いアイデアなので使ってみたい気がしますが、積極的じゃありません。 話のネタに使うとしてもその後の有用な活用方法が思いつかないので使わないかと思う。 自由に音源を入力して歌い分けが楽しめるのであれば面白いアプリかと思うが、違和感が出るレベルでまで歌い分けを聞きたいとは思わない
使いやすい	構造が簡単で使いやすい アプリは操作が比較的簡単で使いやすいので良いと思います。 とにかく操作が簡単で、素人でも使いやすい親切なアプリだと思いました。
機能がほしい	とにかく自由にカスタマイズや編曲できる機能や素材があるとよい。 声のキーを変えられる機能が欲しいです。 歌い分けをもっと細かい語句で行いたい。現在は「でんでんむしむしかたつむり」で歌い分けられているが、「でんでん」「むしむし」「かたつむり」で歌い分けたい。 もっとパターンがあると面白いです テンポを変えたり変調が出来ると面白いと思います。 ドラッグ＆ドロップでシャッフルできたらいい 任意でそれぞれの音量を調節できれば面白いかなと感じました 自分の好きな音声を設定できたら良いと思いました。 歌い手に名前をつけられるとよいと思います。 白い部分のスペースに音楽に関係した画像が表示されるようになると、より楽しめるのではないかと思いました。 キャラクターが歌ってる様子を見たいです。
自分の声を入れたい	自分の声も録音できて同じように歌うパートを変えたり、合唱したり、輪唱したりといろんな楽しみ方ができると良いのではと思った。 楽曲に自分の声を入れて一緒に歌えたり歌い分けできたりしたらいいなと思いました。 自分の音声を認識して録音できて、アプリ上で合成できたらいいと思った。 面倒くさがり屋なのでわざわざ他の声に変換させて音楽を楽しむことは考えられませんが、自分の声を録音して映像作品にのせれば面白いかもしれません。
UI に関する意見	もう少し直感的に使えたらいいと思う。 もっとわかりやすく編集できるような機能。 再生・停止・一時停止のボタンはもう少し大きくデザインされていた方が使いやすいかと思いました。 5 秒や 10 秒など、少しずつ早送りや巻き戻しができる機能があると良いです。 再生中に再生バーのドラッグなどにて好きな部分に飛べるようにしたいです。 見落としていたら申し訳ないのですが、複数人の音声を全てオンにするボタンが無かったのでそういうボタンもあると尚良いと思いました。 歌詞が書いてあると、どの箇所か分かりやすくなるなと感じました。 歌詞の表示もあればパート分けがしやすいかなと思います
声質変換に関する意見	変換技術の質が低すぎる 「VC あり」「VC なし」なしではなく音声の変化量、を自分で決められる機能が欲しいと思いました。 声質変換機能に関しては、さまざまなシンガーやジャンル風の声に変えられると楽しいと思う。 例えば、この技術だと楽器なども音を変えられて面白そうですね。あと、男が女の声なんかも出来るのでしょうか？ 音色をいくつかの選択肢から例えば高いや低いなどから選んで好きな音色を合成できたらいいなと思います。 男性ボイスと女性ボイスを選択できるようにしてほしいです。 声の速度を変えたり、声のトーンを高くしたり低くしたり出来ると面白いなと思いました。
その他	ボーカルトラック数に関してですが、20〜30 人くらいのセパレートのボーカルトラックを編集出来ると面白いかもしれないな、と感じました。既存曲のサビのハモリパートなどを自分の好みで組み合わせて、色々なバリエーションで聴けたら面白そうです。 無料ならこのアプリを使ってみることもあると思いました。やはり、個々の歌い手の区別と分離がしっかりとできるかどうかがポイントであると思います。 用途があれば，皆使うのでは。プロが必要なくなるかも。 マウスなどで操作するのはわかりやすいが、面倒なのでやりたいことを自動化できると楽でよいと感じた。好きな曲を自然に合成できるなら結構楽しめると思った。 面白いと思いますが、私はアイドルやアニメの曲はあまり聴かないので、こういう遊び方も知りませんでした。しかし、幼い子供たちが遊んだり、シニアや高齢者の方、特にデイケアや老人ホームの方々が、童謡や懐メロなどをこのアプリで楽しむには、とても有用ではないかと思いました。なので、高齢者の方用に別でアプリを作って、すべて日本語表記にするなど、改善なさってはと思いました。 このアプリは社会に役立ちそうな感じがしました。特に学校の音楽の授業やカラオケ教室で使うと先生の指導も楽になりそうです。「こういう機能がほしい」で思ったのですが、たとえば「歌手の MISIA さんが AKB48 の曲を歌ったら」のようなことが可能になれば面白いなと思いました。MISIA さんの音声を抜き出して、MISIA さんの声だけでなく歌い方の特徴も完全コピーして別の歌手の曲を歌わせれば（MISIA と AKB の合唱でも構いません）今後の音楽の遊び方も変わってくるような気がします。 好きなアーティスの楽曲を集中して聴くというアプリではなくて、曲をアレンジして楽しむというコアファン用のエンタメ性重視のアプリだと感じました。ですので、もっとユーザーがアレンジできる機能が増えると楽しめると思います。例えば、あるユーザーがアレンジ（編集？）した「かたつむり」が聴けるとか。このアレンジが一番人気ですのようにリコメンドが出るとかですね。 2 人だけじゃなくて、もっと多数の人間の声を切り替えたり、混ぜたり出来ると面白いだろうなと思いました。

が必須である．一方で，Songle [19] のように分析結果をユーザが修正できる機能を備えることも可能で，完全な分析が難しい場合はこうしたインタフェースを構築することも有効である．

また，VocalRemixer ではユニゾンを前提として変換および合成を行っている．さらに広範な楽曲を扱うためには，合唱や重唱などのように音程やリズムの異なる歌声の重なりに対して，分離などを行う技術の構築が必要である．

5.5.2　歌詞との連携

VocalRemixer では，歌詞の情報を利用していない．歌詞を表示することで，特定の歌詞にシークしたり，歌詞を見ながら鑑賞できるようになると考えられる．また，歌詞に応じてさらに細かい単位のパート割りを設定できる機能なども考えられる．さらに，歌声合成技術を導入することで，楽曲の歌詞を好みに編集する技術も考えられる．これまでは楽器音に対する編集技術が多く検討されてきたが，歌詞に対する編集技術が実現されれば，好きなメロディに自分の作詞した歌詞をのせることができるようになる．

5.5.3　ボリューム・パン・エフェクト操作

VocalRemixer では，各ボーカルトラックに対する操作として，ミュートもしくはミュート解除のみが利用できる．一方，デジタル・オーディオ・ワークステーション（digital audio workstation; DAW）ソフトウェアでは，各トラックのボリュームやパン，またエフェクトなどを操作できる．こうした機能を導入することで，より編集の幅が広がると考えられる．とくに，エフェクトとして声質変換エフェクトを導入し，特定の好きな歌手の声に変換したり，男声と女声を切り替えたり，年齢を操作したりなどが可能になれば，好みの声質で楽曲を鑑賞できるようになると考えられる．

5.5.4　楽器音の操作

VocalRemixer では，DAW ソフトウェアでいえば各ボーカルのトラックを編集できる機能にあたる．これを拡張して，各楽器音に対してもトラック化し，ボリュームやパンなどを操作できる機能が考えられる．この場合，1人が歌唱している楽曲にも適用可能な汎用的なインタフェースとなる．さらに，自分で新たなトラックを追加できるようになれば，自分の声や自分で演奏したドラムパートなどを追加することができ，鑑賞の幅が広がるだけでなく，表現する楽しみも生まれると考えられる．

5.5.5　ソーシャル機能

提案する VocalRemixer では，各ユーザが個別に編集し鑑賞するのみで，それを共有する機能などが存在しない．動画投稿サイトなどのように楽曲を編集した結果をソーシャル・ネットワーキング・サービス（social networking service; SNS）上などで他人と共有できるようになれば，あるユーザが加工した楽曲を楽しんだり，それを起点に

したさらなる加工が可能になり，インタフェースがより活発に利用されると考えられる．

5.6　本章のまとめ

本章では，新たな音楽体験を与える音楽鑑賞インタフェースとして，VocalRemixer を提案した．VocalRemixer は，パート割りのある楽曲に対し，各歌唱者がソロで歌唱した歌声信号を利用することで，好みのパート割りに楽曲を編集できるインタフェースである．第 3 章で提案した歌唱者ダイアライゼーション手法や，第 4 章で提案したノンパラレル声質変換法を組み合わせることで，任意の楽曲に対してこのインタフェースを適用することが可能である．VocalRemixer は Web インタフェースとして実装されており，ユーザが気軽に楽しむことができる．主観評価によって，インタフェースの使いやすさ，新規性，魅力が確かめられた．自由記述による感想においても，自由に組み合わせを変えられるのがおもしろい，といった意見がみられた．一方で，声質変換を組み合わせる場合，学習に用いることができる歌声の区間が非常に短く変換品質が低いことが確認され，ノンパラレル声質変換法の性能向上が求められることが明らかになった．

第 6 章

結論

6.1　本論文のまとめ

音楽情報処理は，音楽にまつわる情報処理全般を指し，音楽から様々な情報を抽出したり，その情報をインタフェースなどで活用したりすることで，人々の音楽体験を拡張する技術を含む．古典的な歌声を扱う音楽情報処理では，歌唱者が 1 人であることを仮定した技術や歌唱者の人数によらない技術が提案されており，複数人歌唱にとくに着目した技術が検討されることは少ない．本論文では，複数人が歌唱する楽曲に対する音楽情報処理として，そのような楽曲の分析や加工技術について論じた．

第 3 章では，複数人が歌唱する楽曲を分析する技術として，パート割りのある楽曲から「誰がいつ歌唱しているか」を推定する歌唱者ダイアライゼーションを実現する手法を提案した．歌唱者ダイアライゼーションは，パート割りのある楽曲に対してそのパート割りを分析できるほか，その楽曲が何人の歌唱者によるものかを推定できる．すなわち，歌唱者ダイアライゼーションによって，パート割りの認識だけでなく，各歌唱者の声質を推定したり，楽曲構造の解析に役立てたりすることが可能である．歌唱者ダイアライゼーションは，会話音声中から「誰がいつ話しているか」を推定する話者ダイアライゼーション技術に類似しているが，伴奏音の影響があることや，複数人が同時に歌唱する区間が会話音声と比べて長いことなどが原因で，既存の話者ダイアライゼーション手法をそのまま適用しても高い精度での推定が行えない．そこで本論文では，話者ダイアライゼーション手法を拡張し，既存の話者ダイアライゼーション手法と比較して高い精度でダイアライゼーションを行うことができる歌唱者ダイアライゼーション手法を提案した．提案法は，特定の話者表現を持つ話者が各フレームで発話しているかを推定する TS-VAD とよばれる機構を持つ話者ダイアライゼーション手法を基礎とし，各歌唱者の歌唱状態を個別に推定する target-singer VAD とよばれる機構を導入した．提案法では，音楽音響信号を対象とした音源分離ソフトウェアである Spleeter を導入し，伴奏音が含まれる楽曲に対する分析を可能にした．また，同時歌唱者数の推定を行いこの結果を利用することで，target-singer VAD の推定誤りによる最終的な推定ラベルの劣化を抑制した．この同時歌唱者数推定を高精度に行うために，Cosacorr スコアとよばれる新たな音響特徴量を導入した．さらに，提案法の性能を向上させるため，顔認識に用いられる埋め込み表現抽出法である ArcFace を利用して歌唱者表現を抽出した．実環境下での実験を行うため，市販の CD に収録されているパート割りのある楽曲を利用して評価した．実験により，既存の同時歌唱者数推定を明示的に行わないダイアライゼーション手法と比較して，提案法を用いることで高い精度でのダイアライゼーションが実現できることが確かめられた．

第 4 章では，歌声を変換する技術として，Soft INCA アルゴリズムとよばれる短時間のノンパラレルな音声から声質変換モデルを学習できるノンパラレル声質変換における変換モデルの学習法を提案した．声質変換モデルの学習には，従来，入力話者と出力話者が同一の内容を発話したパラレルデータが必要であった．しかし，同一の内容を発話した音声が利用できない場合，これらの手法を利用することができない．そこで，同一でない発話を学習に利用できるノンパラレル声質変換法が多数提案されている．一方，ノンパラレルな声質変換法は，入出力話者以外が発話した音声から構築される膨大な背景知識を必要としたり，入出力話者による発話を大量に必要としたりするため，小規模な量の発話のみからモデルを学習することは困難であった．この原因として，音声の言語情報と話

者情報を分離することが既存の手法では難しいことが挙げられる．そこで，提案法では NMF の時変成分と時不変成分に分解する特徴を利用して，この問題を解決せんとした．提案法では，まず出力話者の音響モデル，すなわち NMF による基底を緻密に構成する．次に，入力話者による発話と出力話者の音響モデルのアラインメントを得る．このアラインメントは NMF における生起状態にあたり，この生起状態を得ることで入力話者による発話の言語情報を推定することができ，この生起状態を利用して入力話者の音響モデルを得ることができる．アラインメントには，INCA アルゴリズムとよばれるノンパラレルデータのアラインメントを得る手法から着想を得た手法を利用した．INCA アルゴリズムは離散的なアラインメントを得る手法であるため不適切なアラインメントが最終的なモデルの品質を劣化させるが，提案法では連続的なアラインメントを得るためこの劣化を回避できる．話声を用いた実験では，提案法は既存手法である INCA アルゴリズムや CycleGAN-VC と比較して短時間の音声を有効に活用できることが示された．とくに，アラインメントの誤りに脆弱な NMF 声質変換の枠組みでも提案法は高い変換品質を示し，連続的なアラインメントを得る有効性が示唆された．また，入力音声をそのまま変換する one-shot 変換においても提案法が有効であることが確認された．

第 5 章では，複数人が歌唱する楽曲を加工し鑑賞できるインタフェースである VocalRemixer を提案した，音楽を加工できるインタフェースには，信号処理的なエフェクトを追加したり，特定の楽器音を編集したりできるインタフェースなどが提案されており，音楽情報処理による音楽鑑賞の拡張技術として研究がなされている．VocalRemixer では，複数人が歌唱しているパート割りのある楽曲に対して，各歌唱者を相互に変換することにより，各歌唱者があたかもソロで歌唱した歌声を合成する．これにより，本来 CD などに収録されていたパート割りとは異なるパート割りで楽曲を鑑賞できる．実験では，あらかじめ各歌唱者がソロで歌唱した音声を利用した場合と，パート割りのある歌声のみから変換して合成した場合の，2 つの条件における主観評価を行った．主観評価により，VocalRemixer の使いやすさ，新規性，魅力が確認され，異なるパート割りを鑑賞できる楽しみについても示された．一方で，声質変換を用いた場合には，学習に用いることのできる歌声がごく短時間であるため，高い自然性での変換を行うことはできなかった．

6.2　今後の展望

歌唱者ダイアライゼーションにおいては，十分な精度でダイアライゼーションが実現されているとは言い難い．とくに，単一歌唱者が歌唱している区間のダイアライゼーションについて検討が必要である．この結果に応じて各歌唱者の歌唱者表現が決定されるため，この手順の精度が手法全体の性能に大きく影響する．本論文では ArcFace によって得られた歌唱者表現に対してスペクトルクラスタリングを適用することによってダイアライゼーションを行っているが，深層学習を用いたクラスタリング法などを導入することで改善する可能性がある．また，target-singer VAD についてもさらなる改善の余地がある．現在は，各歌唱者に対して，その歌唱者の表現と各フレームの音響特徴量から歌唱状態を推定している．しかし，似た声の歌唱者がいる場合，ある歌唱者の歌唱状態が別歌唱者の歌唱状態の推定に大きく影響してしまう．そこで，注意機構などの可変長の入力を受け付ける機構を導

入することで，似た声の歌唱者がいる場合でも高い精度で推定できると考えられる．さらに，同時歌唱者数推定においても改善の余地がある．提案した同時歌唱者数推定では，歌唱者数が 0 人，1 人，2 人以上の 3 クラス分類のみを行っていた．この推定において，2 人，3 人，4 人などとさらに多数の推定をも可能になれば，それを考慮したダイアライゼーションが実現できると考えられる．この場合，分類問題ではなく回帰問題として扱う必要があるほか，10 人を超えるような多人数になった場合正確に推定できないことが考えられるため，とくに注意して問題を吟味する必要がある．

Soft INCA アルゴリズムにおいては，話者性の変換が十分でない点が課題である．提案した手法においても，既に CycleGAN-VC と同程度の話者性が再現できているものの，聴感上目標話者に限りなく近いとは言い難い．この問題は，出力話者の音響モデルである辞書の大きさに依存すると考えられる．出力話者の辞書が表現できる空間が狭ければ狭いほど，話者性をより近づけることができると期待される．これを解決する手段として，最小体積 NMF [147] やスパース正則化 [148, 149] などの手法が考えられる．たとえば VocalRemixer の実験では少量の歌声から効果的に学習するために最小体積 NMF を導入した．一方で，これらの手段を利用すると，音響モデルの構築時に制約が生じるため，最終的な自然性が低下することが見込まれる．自然性の低下を抑制しつつ，出力話者の辞書の表現できる空間を小さくする技術が求められる．また，NMF をニューラルネットワークに置き換えた non-negative autoencoder（NAE）が提案されている [150, 151]．NMF と比較して音源分離において高い性能を示しており，Soft INCA アルゴリズムにおいても有効に利用できる可能性がある．本論文では適切な評価のため音声の合成に WORLD を用いたが，深層学習による音声合成手法 [46, 152, 153] や WaveCycleGAN [154] などを導入することで，音声の話者性や自然性をさらに向上できることが期待される．声質変換法は多数提案されており，学習に利用できる音声によって適切な手法が異なる．様々なノンパラレル声質変換法と Soft INCA アルゴリズムを比較することで，提案法の立ち位置や効果的に利用できる条件を見出す必要がある．

VocalRemixer については，現在提案しているインタフェースはプロトタイプ的であり，まだ十分な機能を備えているとは言い難い．任意の楽曲を利用できるなど，5.5 節に示したような様々な機能が考えられ，自由記述による感想においても多数の意見が寄せられた．DAW のような楽曲を編集したり再構成したりできる技術や機能の開発が求められる．既存の音楽鑑賞インタフェースとの統合も課題として挙げられる．とくに Songle [19] や TextAlive [21] などの音楽鑑賞インタフェースと同期できれば，1 つの楽曲を起点とした様々な音楽鑑賞体験が実現できると考えられる．

複数人の歌声を扱う音楽情報処理は，これまで歌声という 1 つのトラックとして扱われてきた要素を，複数のトラックに分解する技術である．とくに歌声については，複数の歌声間での相互なインタラクションが可能であり，VocalRemixer で実現したような「本来歌っていない区間を歌わせる」といった機能が実現できる．一方で，本論文で検討した技術のみでは，複数人の歌声を扱う音楽情報処理としては未熟である．複数人が同時に歌唱することによる相互作用を考慮した歌声合成や，複数人の歌唱者の魅力を引き出せるような自動作曲技術や作曲支援技術など，複数人歌唱にまつわる音楽情報処理技術は様々な方向性の展開が可能である．「みんなで歌う」という当たり前を解釈し実現する技術が音楽情報処理には求められる．

謝辞

指導教員である本学大学院工学系研究科の齋藤大輔准教授には，普段から熱心なご指導をいただき，また多くのご助言を賜りました．これまで筆者が所属した峯松・齋藤研究室および齋藤研究室において 6 年もの長きにわたるご指導をいただきました．筆者の研究に対するあり方や哲学を構成するもっとも大きな存在といっても過言ではありません．また，研究に対する指導だけでなく，研究室のサーバ管理をはじめとして，研究環境の面においても支えていただきました．心より感謝を申し上げます．

本学大学院工学系研究科の峯松信明教授には，直接の指導教員であった 1 年間に加えて，5 年もの間第二の指導教員として研究に関わっていただき，多くのご助言を賜りました．研究内容に対する深い洞察や的を射たご指摘は，筆者と筆者の研究の成長においてとても重要でした．深く感謝いたします．

本研究のうち，NMF を用いたノンパラレル声質変換の研究においては，本学大学院情報理工学系研究科の高道慎之介助教に様々なご助言をいただきました．深く感謝いたします．

本研究，とくに歌唱者ダイアライゼーションについて，その出発点は産業技術総合研究所でのインターンシップにありました．産業技術総合研究所の後藤真孝博士，深山覚博士，中野倫靖博士には，インターンシップ中の本研究のテーマ提案を含め，研究に関する多くのご助言をいただきました．またインターンシップ終了後も継続して本研究に関わっていただき，ディスカッションや論文執筆を通じて様々なご意見をいただきました．この出会いがなければ，本研究がこのように形を成すことはありませんでした．深く感謝を申し上げます．また，産業技術総合研究所におけるインターンシップは，濱崎雅弘博士をはじめとした産業技術総合研究所の方々，また同時期にインターンシップを行っていた 5 人のみなさんに支えられておりました．インターンシップを有意義なものにしてくださったみなさまに感謝いたします．

齋藤研究室および峯松研究室のみなさまには，日頃の研究生活をともに送ってくださり支えてくださいました．とくに，齋藤研究室の同期である小谷岳氏には，ディスカッションなどを通じて研究に貢献いただきました．さらに，技術専門員である高橋登氏，事務補佐員である押田美智子さんおよび池上恵さんには，齋藤研究室および峯松研究室のメンバーとして研究室を支えていただきました．ありがとうございました．

今日まで筆者は，家族と数多くの友人に支えられておりました．筆者を支えてくださったすべての方々に感謝の意を表します．ありがとうございました．

最後に，本論文での評価には，『アイドルマスター』シリーズおよび『ラブライブ!』シリーズの楽曲が使われております．こうした楽曲は，もっとも基本的なアイドルソングであると同時に，非常に多くの人々に楽しまれている素晴らしい楽曲です．本研究の課題はこうした楽曲をきっかけとするものであり，こうした楽曲によって本論文において緻密な評価を行うことができました．こうした楽曲を作ってくださった作詞家，作曲家，編曲家のみなさま，キャラクターに息を吹きこんでくださる声優のみなさま，こうしたコンテンツの制作に携わってくださるみな

さま，こうしたコンテンツを支えてくださるファンのみなさま，そして主役である『アイドルマスター』のアイドルや『ラブライブ!』のメンバーのみなさまに，心から感謝を申し上げます.

2021 年 12 月 1 日

須田仁志

参考文献

[1] Julian Dodd. What 4′33″ is. *Australasian Journal of Philosophy*, Vol. 96, No. 4, pp. 629–641, October 2018.

[2] Steven Brown, Björn Merker, and Nils L. Wallin, editors. *The origins of music*. A Bradford Book, November 1999.

[3] 柴田 南雄. 音楽史と音楽論. 岩波書店, April 2014.

[4] Robert Fink. *Neanderthal Flute: Oldest musical instrument*. Greenwich, 1997.

[5] Philip G. Chase and April Nowell. Taphonomy of a suggested middle paleolithic bone flute from Slovenia. *Current Anthropology*, Vol. 39, No. 4, pp. 549–553, 1998.

[6] Earliest music instruments found. *BBC News*, May 2012.

[7] 谷口 高士. 現代の音楽生活と感情. 佐藤 香（編）, 感情現象の諸相, pp. 27–42. ナカニシヤ出版, December 2005.

[8] 北原 鉄朗, 永野 秀尚, 亀岡 弘和, 東条 敏, 齋藤 大輔, 深山 覚, 後藤 真孝, 吉井 和佳, 帆足 啓一郎, 竹川 佳成, 伊藤 貴之, 濱崎 雅弘, 馬場 哲晃, 水本武志, 寺島 裕貴. 特集 音楽を軸に拡がる情報科学. 情報処理学会論文誌, Vol. 57, No. 6, pp. 504–543, May 2016.

[9] Anne Steele. Apple Music reveals how much it pays when you stream a song. *Wall Street Journal*, April 2021.

[10] Johan Pauwels, Ken O'Hanlon, Emilia Gómez, and Mark B. Sandler. 20 years of automatic chord recognition from audio. In *Proc. the 20th Conference of the International Society for Music Information Retrieval*, November 2019.

[11] Masataka Goto. An audio-based real-time beat tracking system for music with or without drum-sounds. *Journal of New Music Research*, Vol. 30, No. 2, pp. 159–171, June 2001.

[12] Zhouyu Fu, Guojun Lu, Kai Ming Ting, and Dengsheng Zhang. A survey of audio-based music classification and annotation. *IEEE Transactions on Multimedia*, Vol. 13, No. 2, pp. 303–319, April 2011.

[13] Fabian-Robert Stöter, Stefan Uhlich, Antoine Liutkus, and Yuki Mitsufuji. Open-unmix - a reference implementation for music source separation. *Journal of Open Source Software*, Vol. 4, No. 41, p. 1667, September 2019.

[14] Romain Hennequin, Anis Khlif, Felix Voituret, and Manuel Moussallam. Spleeter: A fast and efficient music source separation tool with pre-trained models. *Journal of Open Source Software*, Vol. 5, No. 50, pp. 2154–2157, June 2020.

[15] A. Ozerov, P. Philippe, R. Gribonval, and F. Bimbot. One microphone singing voice separation using source-adapted models. In *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp.

90–93, October 2005.

[16] Martin Piszczalski and Bernard A. Galler. Automatic music transcription. *Computer Music Journal*, Vol. 1, No. 4, pp. 24–31, November 1977.

[17] Emmanouil Benetos, Simon Dixon, Zhiyao Duan, and Sebastian Ewert. Automatic music transcription: An overview. *IEEE Signal Processing Magazine*, Vol. 36, No. 1, pp. 20–30, January 2019.

[18] Annamaria Mesaros and Tuomas Virtanen. Automatic recognition of lyrics in singing. *EURASIP Journal on Audio, Speech, and Music Processing*, Vol. 2010, No. 1, pp. 1–11, December 2010.

[19] Masataka Goto, Kazuyoshi Yoshii, Hiromasa Fujihara, Matthias Mauch, and Tomoyasu Nakano. Songle: A web service for active music listening improved by user contributions. In *Proc. the 12th International Society for Music Information Retrieval Conference*, pp. 311–316, October 2011.

[20] Masahiro Hamasaki, Masataka Goto, and Tomoyasu Nakano. Songrium: A music browsing assistance service with interactive visualization and exploration of protect a web of music. In *Proc. the 23rd International Conference on World Wide Web*, pp. 523–528, April 2014.

[21] Jun Kato, Tomoyasu Nakano, and Masataka Goto. TextAlive: Integrated design environment for kinetic typography. In *Proc. the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pp. 3403–3412, April 2015.

[22] 藤原 弘将, 後藤 真孝. VocalFinder: 声質の類似度に基づく楽曲検索システム. 情報処理学会研究報告音楽情報科学（MUS）, Vol. 2007, No. 81 (2007-MUS-071), pp. 27–32, August 2007.

[23] Alexios Kotsifakos, Panagiotis Papapetrou, Jaakko Hollmén, Dimitrios Gunopulos, and Vassilis Athitsos. A survey of query-by-humming similarity methods. In *Proc. the 5th International Conference on PErvasive Technologies Related to Assistive Environments*, pp. 1–4, June 2012.

[24] Òscar Celma. *Music recommendation and discovery: The long tail, long fail, and long play in the digital music space*. Springer, 2010.

[25] Jose David Fernández and Francisco Vico. AI methods in algorithmic composition: A comprehensive survey. *Journal of Artificial Intelligence Research*, Vol. 48, pp. 513–582, November 2013.

[26] 渡邉 研斗, 松林 優一郎, 深山 覚, 中野 倫靖, 後藤 真孝, 乾 健太郎. メロディと歌詞の相関に基づく自動歌詞生成. 研究報告自然言語処理（NL）, Vol. 2017-NL-231, No. 16, pp. 1–12, May 2017.

[27] Satoru Fukayama, Kei Nakatsuma, Shinji Sako, Yuichiro Yonebayashi, Tae Hun Kim, Si Wei Qin, Takuho Nakano, Takuya Nishimoto, and Shigeki Sagayama. Orpheus: Automatic composition system considering prosody of Japanese lyrics. In *Proc. the Entertainment Computing 2009*, Lecture Notes in Computer Science, pp. 309–310, Berlin, Heidelberg, September 2009. Springer.

[28] Cheng-Zhi Anna Huang, David Duvenaud, and Krzysztof Z. Gajos. ChordRipple: Recommending chords to help novice composers go beyond the ordinary. In *Proc. the 21st International Conference on Intelligent User*

Interfaces, pp. 241–250, March 2016.

[29] 阿部 ちひろ, 伊藤 彰則. patissier-アマチュア作詞家のための作詞補助システム-. 研究報告音声言語情報処理（SLP）, Vol. 2012-SLP-90, No. 17, pp. 1–6, January 2012.

[30] Perry R. Cook. Singing voice synthesis: History, current work, and future directions. *Computer Music Journal*, Vol. 20, No. 3, pp. 38–46, 1996.

[31] 大浦 圭一郎, 間瀬 絢美, 山田 知彦, 徳田 恵一, 後藤 真孝. Sinsy:「あの人に歌ってほしい」をかなえる HMM 歌声合成システム. 研究報告音楽情報科学（MUS）, Vol. 2010-MUS-86, No. 1, pp. 1–8, July 2010.

[32] Ajay Kapur. A history of robotic musical instruments. In *Proc. International Computer Music Conference*. Michigan Publishing, September 2005.

[33] Darius Petermann, Pritish Chandna, Helena Cuesta, Jordi Bonada, and Emilia Gomez. Deep learning based source separation applied to choir ensembles. In *Proc. the 21st International Society for Music Information Retrieval Conference*, October 2020.

[34] Fabian-Robert Stöter, Antoine Liutkus, Roland Badeau, Bernd Edler, and Paul Magron. Common fate model for unison source separation. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 126–130, March 2016.

[35] Wei-Ho Tsai and Hsin-Min Wang. Automatic detection and tracking of target singer in multi-singer music recordings. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, Vol. 4, pp. 221–224, May 2004.

[36] Wei-Ho Tsai and Hsin-Min Wang. Automatic singer recognition of popular music recordings via estimation and modeling of solo vocal signals. *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 14, No. 1, pp. 330–341, January 2006.

[37] Marwa Thlithi, Claude Barras, Julien Pinquier, and Thomas Pellegrini. Singer diarization: Application to ethnomusicological recordings. In *Proc. the 5th International Workshop on Folk Music Anaysis*, pp. 124–125, June 2015.

[38] Sten Ternstrom, Anders Friberg, and Johan Sundberg. Synthesizing choir singing. *Journal of Voice*, Vol. 1, No. 4, pp. 332–335, January 1988.

[39] 桑原 彰宏. 合唱音声の合成における基本周波数制御に関する基礎研究. 修士論文, 北陸先端科学技術大学院大学, March 2010.

[40] 野田 雄也. 合唱における基本周波数の同期現象に関する基礎研究. 修士論文, 北陸先端科学技術大学院大学, March 2008.

[41] 勝瑞 雄介, 齋藤 大輔, 峯松 信明. 自然な斉唱音声合成のための複数歌唱者の基本周波数パターン制御に関する検討. 研究報告音楽情報科学（MUS）, Vol. 2021-MUS-131, No. 11, pp. 1–7, June 2021.

[42] 菊地 晏南, 齋藤 大輔, 峯松 信明. 固有声変換法を用いた重唱における調和度制御に関する検討. 研究報告音声

言語情報処理（SLP）, Vol. 2021-SLP-137, No. 27, pp. 1–6, June 2021.

[43] 山内 孔貴, 須田 仁志, 齋藤 大輔, 峯松 信明. ソースフィルタ分解に基づく複数歌唱者の調和制御に関する検討. 研究報告音声言語情報処理（SLP）, Vol. 2020-SLP-132, No. 35, pp. 1–6, May 2020.

[44] Johan Sundberg. 歌声の科学. 東京電機大学出版局, March 2007.

[45] D. Griffin and Jae Lim. Signal estimation from modified short-time Fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. 32, No. 2, pp. 236–243, April 1984.

[46] Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. WaveNet: A generative model for raw audio. *arXiv:1609.03499 [cs]*, September 2016.

[47] 徳田 恵一, 小林 隆夫, 今井 聖. メル一般化ケプストラムの再帰的計算法. 電子情報通信学会論文誌 A, Vol. 71, No. 1, pp. 128–131, January 1988.

[48] Toshiaki Fukada, Keiichi Tokuda, Takao Kobayashi, and Satoshi Imai. An adaptive algorithm for mel-cepstral analysis of speech. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, Vol. 1, pp. 137–140, March 1992.

[49] Alain de Cheveigné and Hideki Kawahara. YIN, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America*, Vol. 111, No. 4, pp. 1917–1930, April 2002.

[50] Masanori Morise, Hideki Kawahara, and Haruhiro Katayose. Fast and reliable f0 estimation method based on the period extraction of vocal fold vibration of singing voice and speech. In *Proc. the Audio Engineering Society 35th International Conference: Audio for Games*. Audio Engineering Society, February 2009.

[51] Masanori Morise, Hideki Kawahara, and Takanobu Nishiura. Rapid f0 estimation for high-SNR speech based on fundamental component extraction. *IEICE Transactions on Information and Systems (Japanese Edition)*, Vol. J93-D, No. 2, pp. 109–117, February 2010.

[52] Masanori Morise. Harvest: A high-performance fundamental frequency estimator from speech signals. In *Proc. INTERSPEECH*, pp. 2321–2325, August 2017.

[53] Jong Wook Kim, Justin Salamon, Peter Li, and Juan Pablo Bello. Crepe: A convolutional representation for pitch estimation. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 161–165, April 2018.

[54] Matthias Mauch and Simon Dixon. PYIN: A fundamental frequency estimator using probabilistic threshold distributions. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 659–663, May 2014.

[55] Xavier Anguera, Simon Bozonnet, Nicholas Evans, Corinne Fredouille, Gerald Friedland, and Oriol Vinyals. Speaker diarization: A review of recent research. *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 20, No. 2, pp. 356–370, February 2012.

[56] Shinji Watanabe, Michael Mandel, Jon Barker, Emmanuel Vincent, Ashish Arora, Xuankai Chang, Sanjeev Khudanpur, Vimal Manohar, Daniel Povey, Desh Raj, David Snyder, Aswin Shanmugam Subramanian, Jan Trmal, Bar Ben Yair, Christoph Boeddeker, Zhaoheng Ni, Yusuke Fujita, Shota Horiguchi, Naoyuki Kanda, Takuya Yoshioka, and Neville Ryant. CHiME-6 challenge: Tackling multispeaker speech recognition for unsegmented recordings. In *Proc. the 6th International Workshop on Speech Processing in Everyday Environments (CHiME 2020)*, pp. 1–7, May 2020.

[57] Neville Ryant, Prachi Singh, Venkat Krishnamohan, Rajat Varma, Kenneth Church, Christopher Cieri, Jun Du, Sriram Ganapathy, and Mark Liberman. The third DIHARD diarization challenge. *arXiv:2012.01477 [cs, eess]*, April 2021.

[58] Kazuhiro Otsuka, Shoko Araki, Kentaro Ishizuka, Masakiyo Fujimoto, Martin Heinrich, and Junji Yamato. A realtime multimodal system for analyzing group meetings by combining face pose tracking and speaker diarization. In *Proc. the 10th International Conference on Multimodal Interfaces*, ICMI '08, pp. 257–264, New York, NY, USA, October 2008. Association for Computing Machinery.

[59] Athanasios Noulas, Gwenn Englebienne, and Ben J.A. Krose. Multimodal speaker diarization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 34, No. 1, pp. 79–93, January 2012.

[60] Scott Otterson and Mari Ostendorf. Efficient use of overlap information in speaker diarization. In *Proc. IEEE Workshop on Automatic Speech Recognition Understanding*, pp. 683–686, December 2007.

[61] Yusuke Fujita, Naoyuki Kanda, Shota Horiguchi, Yawen Xue, Kenji Nagamatsu, and Shinji Watanabe. End-to-end neural speaker diarization with self-attention. In *Proc. IEEE Automatic Speech Recognition and Understanding Workshop*, pp. 296–303, December 2019.

[62] Ivan Medennikov, Maxim Korenevsky, Tatiana Prisyach, Yuri Khokhlov, Mariya Korenevskaya, Ivan Sorokin, Tatiana Timofeeva, Anton Mitrofanov, Andrei Andrusenko, Ivan Podluzhny, Aleksandr Laptev, and Aleksei Romanenko. Target-speaker voice activity detection: A novel approach for multi-speaker diarization in a dinner party scenario. In *Proc. INTERSPEECH*, pp. 274–278, October 2020.

[63] Zhiyan Duan, Haotian Fang, Bo Li, Khe Chai Sim, and Ye Wang. The NUS sung and spoken lyrics corpus: A quantitative comparison of singing and speech. In *Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, pp. 1–9, October 2013.

[64] Akira Kurematsu, Kazuya Takeda, Yoshinori Sagisaka, Shigeru Katagiri, Hisao Kuwabara, and Kiyohiro Shikano. ATR Japanese speech database as a tool of speech recognition and synthesis. *Speech Communication*, Vol. 9, No. 4, pp. 357–363, August 1990.

[65] Itsuki Ogawa and Masanori 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, May 2021.

[66] Yusuke Fujita, Naoyuki Kanda, Shota Horiguchi, Kenji Nagamatsu, and Shinji Watanabe. End-to-end neural speaker diarization with permutation-free objectives. In *Proc. INTERSPEECH*, pp. 4300–4304, 2019.

[67] Shota Horiguchi, Yusuke Fujita, Shinji Watanabe, Yawen Xue, and Kenji Nagamatsu. End-to-end speaker diarization for an unknown number of speakers with encoder-decoder based attractors. In *Proc. INTERSPEECH*, pp. 269–273, 2020.

[68] Hee-Soo Heo, Jee-weon Jung, Youngki Kwon, You Jin Kim, Jaesung Huh, Joon Son Chung, and Bong-Jin Lee. NAVER CLOVA submission to the third DIHARD challenge. Technical report, 2021.

[69] Jee-weon Jung, Hee-Soo Heo, Youngki Kwon, Joon Son Chung, and Bong-Jin Lee. Three-class overlapped speech detection using a convolutional recurrent neural network. In *Proc. INTERSPEECH*, pp. 3086–3090, 2021.

[70] Xiong Xiao, Naoyuki Kanda, Zhuo Chen, Tianyan Zhou, Takuya Yoshioka, Sanyuan Chen, Yong Zhao, Gang Liu, Yu Wu, Jian Wu, Shujie Liu, Jinyu Li, and Yifan Gong. Microsoft speaker diarization system for the VoxCeleb speaker recognition challenge 2020. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 5824–5828, June 2021.

[71] Sue E. Tranter and Douglas A. Reynolds. An overview of automatic speaker diarization systems. *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 14, No. 5, pp. 1557–1565, September 2006.

[72] Tomi Kinnunen, Evgenia Chernenko, Marko Tuononen, Pasi Fränti, and Haizhou Li. Voice activity detection using MFCC features and support vector machine. In *Proc. International Conference on Speech and Computer*, Vol. 2, pp. 556–561, October 2007.

[73] Andrey Temko, Dusan Macho, and Climent Nadeu. Enhanced SVM training for robust speech activity detection. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, Vol. 4, pp. 1025–1028, April 2007.

[74] Elias Rentzeperis, Andreas Stergiou, Christos Boukis, Aristodemos Pnevmatikakis, and Lazaros C. Polymenakos. The 2006 Athens Information Technology speech activity detection and speaker diarization systems. In *Proc. Machine Learning for Multimodal Interaction*, pp. 385–395, May 2006.

[75] Jongseo Sohn, Nam Soo Kim, and Wonyong Sung. A statistical model-based voice activity detection. *IEEE Signal Processing Letters*, Vol. 6, No. 1, pp. 1–3, January 1999.

[76] Thad Hughes and Keir Mierle. Recurrent neural networks for voice activity detection. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 7378–7382, May 2013.

[77] Florian Eyben, Felix Weninger, Stefano Squartini, and Björn Schuller. Real-life voice activity detection with LSTM recurrent neural networks and an application to Hollywood movies. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 483–487, May 2013.

[78] Bernhard Lehner, Gerhard Widmer, and Reinhard Sonnleitner. On the reduction of false positives in singing

voice detection. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 7480–7484, May 2014.

[79] Simon Leglaive, Romain Hennequin, and Roland Badeau. Singing voice detection with deep recurrent neural networks. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 121–125, April 2015.

[80] Jan Schlüter and Bernhard Lehner. Zero-mean convolutions for level-invariant singing voice detection. In *Proc. the 19th International Society for Music Information Retrieval Conference*, pp. 321–326. ISMIR, September 2018.

[81] Scott Shaobing Chen and P. S. Gopalakrishnan. Speaker, environment and channel change detection and clustering via the Bayesian information criterion. In *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, pp. 127–132, February 1998.

[82] J. E. Rougui, M. Rziza, D. Aboutajdine, M. Gelgon, and J. Martinez. Fast incremental clustering of Gaussian mixture speaker models for scaling up retrieval in on-line broadcast. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, Vol. 5, pp. 521–524, May 2006.

[83] Huazhong Ning, Ming Liu, Hao Tang, and Thomas S. Huang. A spectral clustering approach to speaker diarization. In *Proc. INTERSPEECH*, pp. 2178–2181, September 2006.

[84] Tae Jin Park, Kyu J. Han, Manoj Kumar, and Shrikanth Narayanan. Auto-tuning spectral clustering for speaker diarization using normalized maximum eigengap. *IEEE Signal Processing Letters*, Vol. 27, pp. 381–385, December 2019.

[85] Najim Dehak, Patrick J. Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet. Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 19, No. 4, pp. 788–798, May 2011.

[86] Jan Prazak and Jan Silovsky. Speaker diarization using PLDA-based speaker clustering. In *Proc. the 6th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems*, Vol. 1, pp. 347–350, September 2011.

[87] Gregory Sell and Daniel Garcia-Romero. Speaker diarization with PLDA i-vector scoring and unsupervised calibration. In *Proc. IEEE Spoken Language Technology Workshop*, pp. 413–417, December 2014.

[88] Daniel Garcia-Romero, David Snyder, Gregory Sell, Daniel Povey, and Alan McCree. Speaker diarization using deep neural network embeddings. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4930–4934, March 2017.

[89] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. X-vectors: Robust DNN embeddings for speaker recognition. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 5329–5333, April 2018.

[90] Maokui He, Desh Raj, Zili Huang, Jun Du, Zhuo Chen, and Shinji Watanabe. Target-speaker voice activity detection with improved i-vector estimation for unknown number of speaker. In *Proc. INTERSPEECH*, pp. 3555–3559, 2021.

[91] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Proc. Medical Image Computing and Computer-Assisted Intervention*, pp. 234–241, October 2015.

[92] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. ArcFace: Additive angular margin loss for deep face recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4690–4699, June 2019.

[93] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. SphereFace: Deep hypersphere embedding for face recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, July 2017.

[94] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. CosFace: Large margin cosine loss for deep face recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, June 2018.

[95] Shaojin Ding, Quan Wang, Shuo-Yiin Chang, Li Wan, and Ignacio Lopez Moreno. Personal VAD: Speaker-conditioned voice activity detection. In *Proc. Odyssey the Speaker and Language Recognition Workshop*, pp. 433–439, November 2020.

[96] Seyed Omid Sadjadi, Malcolm Slaney, and Larry Heck. MSR identity toolbox v1.0: A MATLAB toolbox for speaker-recognition research. *Speech and Language Processing Technical Committee Newsletter*, Vol. 1, No. 4, pp. 1–32, September 2013.

[97] Jonathan G. Fiscus, Jerome Ajot, Martial Michel, and John S. Garofolo. The rich transcription 2006 spring meeting recognition evaluation. In *Proc. Machine Learning for Multimodal Interaction*, pp. 309–322, 2006.

[98] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, Vol. 9, No. 86, pp. 2579–2605, November 2008.

[99] Alexander Kain and Michael W. Macon. Spectral voice conversion for text-to-speech synthesis. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 285–288, May 1998.

[100] Yannis Stylianou, Olivier Cappé, and Eric Moulines. Statistical methods for voice quality transformation. In *Proc. EUROSPEECH*, pp. 447–450, September 1995.

[101] Tomoki Toda, Alan W. Black, and Keiichi Tokuda. Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory. *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 15, No. 8, pp. 2222–2235, November 2007.

[102] Ling-Hui Chen, Zhen-Hua Ling, Yan Song, and Li-Rong Dai. Joint spectral distribution modeling using restricted

Boltzmann machines for voice conversion. In *Proc. INTERSPEECH*, pp. 3052–3056, August 2013.

[103] Toru Nakashika, Tetsuya Takiguchi, and Yasuo Ariki. Sparse nonlinear representation for voice conversion. In *Proc. IEEE International Conference on Multimedia and Expo*, pp. 1–6, June 2015.

[104] B. Makki, S. A. Seyedsalehi, N. Sadati, and M. N. Hosseini. Voice conversion using nonlinear principal component analysis. In *Proc. IEEE Symposium on Computational Intelligence in Image and Signal Processing*, pp. 336–339, April 2007.

[105] Srinivas Desai, Alan W. Black, B. Yegnanarayana, and Kishore Prahallad. Spectral mapping using artificial neural networks for voice conversion. *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 18, No. 5, pp. 954–964, July 2010.

[106] Lifa Sun, Shiyin Kang, Kun Li, and Helen Meng. Voice conversion using deep bidirectional long short-term memory based recurrent neural networks. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4869–4873, April 2015.

[107] Ryoichi Takashima, Tetsuya Takiguchi, and Yasuo Ariki. Exemplar-based voice conversion in noisy environment. In *Proc. IEEE Spoken Language Technology Workshop*, pp. 313–317, December 2012.

[108] Zhizheng Wu, Tuomas Virtanen, Eng Siong Chng, and Haizhou Li. Exemplar-based sparse representation with residual compensation for voice conversion. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 22, No. 10, pp. 1506–1521, October 2014.

[109] Elina Helander, Jan Schwarz, Jani Nurminen, Hanna Silen, and Moncef Gabbouj. On the impact of alignment on voice conversion performance. In *Proc. INTERSPEECH*, pp. 1453–1456, September 2008.

[110] Jing-Xuan Zhang, Zhen-Hua Ling, Li-Juan Liu, Yuan Jiang, and Li-Rong Dai. Sequence-to-sequence acoustic modeling for voice conversion. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 27, No. 3, pp. 631–644, March 2019.

[111] Athanasios Mouchtaris, Jan Van der Spiegel, and Paul Mueller. Nonparallel training for voice conversion based on a parameter adaptation approach. *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 14, No. 3, pp. 952–963, May 2006.

[112] Tomoki Toda, Yamato Ohtani, and Kiyohiro Shikano. Eigenvoice conversion based on Gaussian mixture model. In *Proc. INTERSPEECH*, pp. 2446–2449, September 2006.

[113] Daisuke Saito, Keisuke Yamamoto, Nobuaki Minematsu, and Keikichi Hirose. One-to-many voice conversion based on tensor representation of speaker space. In *Proc. INTERSPEECH*, pp. 653–656, August 2011.

[114] Tomi Kinnunen, Lauri Juvela, Paavo Alku, and Junichi Yamagishi. Non-parallel voice conversion using i-vector PLDA: Towards unifying speaker verification and transformation. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 5535–5539, March 2017.

[115] Jing-Xuan Zhang, Zhen-Hua Ling, and Li-Rong Dai. Non-parallel sequence-to-sequence voice conversion with

disentangled linguistic and speaker representations. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 28, pp. 540–552, December 2019.

[116] Lifa Sun, Kun Li, Hao Wang, Shiyin Kang, and Helen Meng. Phonetic posteriorgrams for many-to-one voice conversion without parallel data training. In *Proc. IEEE International Conference on Multimedia and Expo*, pp. 1–6, July 2016.

[117] Takuhiro Kaneko and Hirokazu Kameoka. CycleGAN-VC: Non-parallel voice conversion using cycle-consistent adversarial networks. In *Proc. the 26th European Signal Processing Conference*, pp. 2100–2104, September 2018.

[118] Takuhiro Kaneko, Hirokazu Kameoka, Kou Tanaka, and Nobukatsu Hojo. CycleGAN-VC2: Improved CycleGAN-based non-parallel voice conversion. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 6820–6824, May 2019.

[119] Takuhiro Kaneko, Hirokazu Kameoka, Kou Tanaka, and Nobukatsu Hojo. CycleGAN-VC3: Examining and improving CycleGAN-VCs for mel-spectrogram conversion. In *Proc. INTERSPEECH*, pp. 2017–2021, October 2020.

[120] Chin-Cheng Hsu, Hsin-Te Hwang, Yi-Chiao Wu, Yu Tsao, and Hsin-Min Wang. Voice conversion from non-parallel corpora using variational auto-encoder. In *Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, pp. 1–6, December 2016.

[121] Takuhiro Kaneko, Hirokazu Kameoka, Kou Tanaka, and Nobukatsu Hojo. StarGAN-VC2: Rethinking conditional methods for StarGAN-Based voice conversion. In *Proc. INTERSPEECH*, pp. 679–683, September 2019.

[122] Hirokazu Kameoka, Takuhiro Kaneko, Kou Tanaka, and Nobukatsu Hojo. ACVAE-VC: Non-parallel voice conversion with auxiliary classifier variational autoencoder. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 27, No. 9, pp. 1432–1443, September 2019.

[123] Daniel Erro, Asunción Moreno, and Antonio Bonafonte. INCA algorithm for training voice conversion systems from nonparallel corpora. *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 18, No. 5, pp. 944–953, July 2010.

[124] Daisuke Saito, Shinji Watanabe, Atsushi Nakamura, and Nobuaki Minematsu. Statistical voice conversion based on noisy channel model. *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 20, No. 6, pp. 1784–1794, August 2012.

[125] Daniel D. Lee and H. Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, Vol. 401, No. 6755, pp. 788–791, October 1999.

[126] Paris Smaragdis and Judith C. Brown. Non-negative matrix factorization for polyphonic music transcription. In *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 177–180, October 2003.

[127] Mikkel N. Schmidt, Jan Larsen, and Fu-Tien Hsiao. Wind noise reduction using non-negative sparse coding. In *Proc. IEEE Workshop on Machine Learning for Signal Processing*, pp. 431–436, August 2007.

[128] Paris Smaragdis and Bhiksha Raj. Example-driven bandwidth expansion. In *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 135–138, October 2007.

[129] Daniel D. Lee and H. Sebastian Seung. Algorithms for non-negative matrix factorization. In *Proc. the Advances in Neural Information Processing Systems 13*, pp. 556–562, December 2001.

[130] Nirmesh J. Shah and Hemant A. Patil. On the convergence of INCA algorithm. In *Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, pp. 559–562, December 2017.

[131] Masanori Morise, Fumiya Yokomori, and Kenji Ozawa. WORLD: A vocoder-based high-quality speech synthesis system for real-time applications. *IEICE Transactions on Information and Systems*, Vol. E99-D, No. 7, pp. 1877–1884, July 2016.

[132] Masanori Morise. D4C, a band-aperiodicity estimator for high-quality speech synthesis. *Speech Communication*, Vol. 84, pp. 57–65, November 2016.

[133] Michael Pitz and Hermann Ney. Vocal tract normalization equals linear transformation in cepstral space. *IEEE Transactions on Speech and Audio Processing*, Vol. 13, No. 5, pp. 930–944, September 2005.

[134] Keiichi Tokuda, Takao Kobayashi, and Satoshi Imai. Recursion formula for calculation of mel generalized cepstrum coefficients. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences (Japanese Edition)*, Vol. 71, No. 1, pp. 128–131, January 1988.

[135] Gaku Kotani, Hitoshi Suda, Daisuke Saito, and Nobuaki Minematsu. Experimental investigation on the efficacy of Affine-DTW in the quality of voice conversion. In *Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, pp. 119–124, November 2019.

[136] H. Benisty, D. Malah, and K. Crammer. Non-parallel voice conversion using joint optimization of alignment by temporal context and spectral distortion. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 7909–7913, May 2014.

[137] Nirmesh Shah and Hemant Patil. Effectiveness of dynamic features in INCA and temporal context-INCA. In *Proc. INTERSPEECH*, pp. 711–715, September 2018.

[138] Masataka Goto. Active music listening interfaces based on signal processing. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, Vol. 4, pp. 1441–1444, April 2007.

[139] Masataka Goto. SmartMusicKIOSK: Music listening station with chorus-search function. In *Proc. the 16th Annual ACM Symposium on User Interface Software and Technology*, pp. 31–40, November 2003.

[140] Masataka Goto. A chorus section detection method for musical audio signals and its application to a music listening station. *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 14, No. 5, pp. 1783–1794, September 2006.

[141] Kazuyoshi Yoshii, Masataka Goto, and Hiroshi G. Okuno. INTER:D: A drum sound equalizer for controlling volume and timbre of drums. In *Proc. the 2nd European Workshop on the Integration of Knowledge, Semantic and Digital Media Technologies*, pp. 205–212, January 2005.

[142] Kazuyoshi Yoshii, Masataka Goto, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G. Okuno. Drumix: An audio player with real-time drum-part rearrangement functions for active music listening. *Information and Media Technologies*, Vol. 2, No. 2, pp. 601–611, 2007.

[143] Alo Allik, György Fazekas, Mathieu Barthet, and Mark Swire. myMoodplay: An interactive mood-based music discovery app. In *Proc. Web Audio Conference*, April 2016.

[144] Masahiro Hamasaki and Masataka Goto. Songrium: A music browsing assistance service based on visualization of massive open collaboration within music content creation community. In *Proc. the 9th International Symposium on Open Collaboration*, pp. 1–10, August 2013.

[145] Masataka Goto and Takayuki Goto. Musicream: Integrated music-listening interface for active, flexible, and unexpected encounters with musical pieces. *Journal of Information Processing*, Vol. 17, pp. 292–305, 2009.

[146] Hiroki Tamaru, Shinnosuke Takamichi, Naoko Tanji, and Hiroshi Saruwatari. JVS-MuSiC: Japanese multi-speaker singing-voice corpus. *arXiv:2001.07044 [cs, eess]*, January 2020.

[147] Valentin Leplat, Nicolas Gillis, and Andersen M.S. Ang. Blind audio source separation with minimum-volume beta-divergence NMF. *IEEE Transactions on Signal Processing*, Vol. 68, pp. 3400–3410, May 2020.

[148] Patrik O. Hoyer. Non-negative matrix factorization with sparseness constraints. *Journal of Machine Learning Research*, Vol. 5, pp. 1457–1469, December 2004.

[149] Andrzej Cichocki, Rafal Zdunek, and Shun-ichi Amari. New algorithms for non-negative matrix factorization in applications to blind source separation. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, Vol. 5, pp. 621–624, May 2006.

[150] Paris Smaragdis and Shrikant Venkataramani. A neural network alternative to non-negative audio models. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 86–90, March 2017.

[151] Shrikant Venkataramani, Cem Subakan, and Paris Smaragdis. Neural network alternatives to convolutive audio models for source separation. In *Proc. the IEEE 27th International Workshop on Machine Learning for Signal Processing*, pp. 1–6, September 2017.

[152] Nal Kalchbrenner, Erich Elsen, Karen Simonyan, Seb Noury, Norman Casagrande, Edward Lockhart, Florian Stimberg, Aaron van den Oord, Sander Dieleman, and Koray Kavukcuoglu. Efficient neural audio synthesis. *arXiv:1802.08435 [cs, eess]*, June 2018.

[153] Ryan Prenger, Rafael Valle, and Bryan Catanzaro. WaveGlow: A flow-based generative network for speech synthesis. *arXiv:1811.00002 [cs, eess, stat]*, October 2018.

[154] Kou Tanaka, Takuhiro Kaneko, Nobukatsu Hojo, and Hirokazu Kameoka. WaveCycleGAN: Synthetic-to-

natural speech waveform conversion using cycle-consistent adversarial networks. *arXiv:1809.10288 [cs, eess, stat]*, September 2018.

発表文献

ジャーナル論文

[1] **Hitoshi Suda**, Gaku Kotani, and Daisuke Saito. INmfCA algorithm for training of nonparallel voice conversion systems based on non-negative matrix factorization. IEICE Transactions on Information and Systems, Vol. E105-D, No. 6, June 2022.（採録決定済み）

[2] **Hitoshi Suda**, Daisuke Saito, Satoru Fukayama, Tomoyasu Nakano, and Masataka Goto. Singer diarization for polyphonic music with unison singing. IEEE/ACM Transactions on Audio, Speech, and Language Processing.（条件付き採録）

国際会議

[1] **Hitoshi Suda**, Gaku Kotani, Shinnosuke Takamichi, and Daisuke Saito. A revisit to feature handling for high-quality voice conversion based on Gaussian mixture model. In *Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, November 2018.

[2] **Hitoshi Suda**, Daisuke Saito, and Nobuaki Minematsu. Voice conversion without explicit separation of source and filter components based on non-negative matrix factorization. In *Proc. the Speech Synthesis Workshop 10*, September 2019.

[3] Shunsuke Goto, Yuma Shirahata, Gaku Kotani, **Hitoshi Suda**, Daisuke Saito, and Nobuaki Minematsu. The UTokyo speech synthesis system for Blizzard Challenge 2019. In *Proc. the Blizzard Challenge 2019 workshop*, September 2019.

[4] Gaku Kotani, **Hitoshi Suda**, Daisuke Saito, and Nobuaki Minematsu. Experimental investigation on the efficacy of Affine-DTW in the quality of voice conversion. In *Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, November 2019.

[5] **Hitoshi Suda**, Gaku Kotani, and Daisuke Saito. Nonparallel training of exemplar-based voice conversion system using INCA-based alignment technique. In *Proc. INTERSPEECH*, October 2020.

全国大会

[1] **須田 仁志**，齋藤 大輔，峯松 信明．ソースフィルタ非負値行列因子分解によるボコーダを用いない声質変換の実験的検討．日本音響学会 2017 年秋季研究発表会，2017 年 7 月．

[2] **須田 仁志**，小谷 岳，高道 慎之介，齋藤 大輔．高品質声質変換のための特徴量分析再訪．日本音響学会 2018

年春季研究発表会，2018 年 3 月．

[3] 小谷岳，**須田 仁志**，齋藤 大輔，峯松 信明．パラレルデータ声質変換の品質改善に向けた Affine-DTW の実験的評価．日本音響学会 2019 年秋季研究発表会，2019 年 9 月．

[4] **須田 仁志**，小谷 岳，齋藤 大輔．非負値行列因子分解による声質変換における INCA アルゴリズムをもとにした基底のノンパラレル学習法．日本音響学会 2020 年春季研究発表会，2020 年 3 月．

[5] 今村 奏海，増田 尚建，**須田 仁志**，齋藤 大輔，峯松 信明．自然音声の人工感を連続的に制御する技術の検討と評価．日本音響学会 2022 年春季研究発表会，2022 年 3 月．（投稿済み）

研究会

[1] **須田 仁志**，深山 覚，中野 倫靖，齋藤 大輔，後藤 真孝．グループアイドルソングを対象とした歌唱者ダイアライゼーション手法の基礎的検討．情報処理学会音楽情報科学研究会第 121 回研究発表会，2018 年 11 月．

[2] **須田 仁志**，小谷 岳，高道 慎之介，齋藤 大輔．A revisit to feature handling for high-quality voice conversion based on Gaussian mixture model．第 126 回情報処理学会音声言語情報処理研究発表会（既発表セッション），2019 年 2 月．

[3] 山内 孔貴，**須田 仁志**，齋藤 大輔，峯松 信明．ソースフィルタ分解に基づく複数歌唱者の調和制御に関する検討．第 127 回音楽情報科学・第 132 回音声言語情報処理合同研究発表会（音学シンポジウム 2020），2020 年 6 月．

[4] 五来 丈瑠，**須田 仁志**，齋藤 大輔，峯松 信明．テキスト音声合成における劣化音声を活用したデータ拡充に関する検討．第 127 回音楽情報科学・第 132 回音声言語情報処理合同研究発表会（音学シンポジウム 2020），2020 年 6 月．

学位論文

[1] **須田 仁志**．ソースフィルタ非負値行列因子分解を用いたボコーダフリー声質変換に関する研究．東京大学工学部卒業論文．2017 年 3 月．

[2] **須田 仁志**．歌声分析によるグループアイドルソングのパート割り構造認識に関する基礎的検討．東京大学大学院工学系研究科修士論文．2019 年 3 月．

受賞

[1] 情報処理学会音楽情報科学研究会第 121 回研究発表会学生奨励賞受賞（2018 年 11 月）

付録 A

第3章の実験で用いたデータセットの詳細

すべての楽曲は，アニメやゲームなど，キャラクターの歌唱するキャラクターソングとして制作されたものである．そのため，本章では歌唱者名を「キャラクター名（声優名）」と表記する．データセット内では，同一の声優が複数の異なるキャラクターを演じることはないが，同一のキャラクターが異なる声優によって演じられている場合がある．すべての楽曲は市販の CD から抽出されたものであり，44.1 kHz で標本化，16 bit で量子化された，ステレオの音楽音響信号である．

A.1　データセット A: 3.5 節で用いたデータセット

3.5 節で述べた，Cosacorr スコアの評価についての実験において用いたデータセットの詳細を記す．このデータセットをデータセット A とよぶ．各音楽音響信号には，各歌唱者がそれぞれ独立に歌唱した音楽音響信号が，伴奏音とともにミキシングおよびマスタリングされた上で収録されている．また，各楽曲にはボーカルのミキシングされていない伴奏音のみのカラオケトラックが存在し，実験ではこれを用いている．表 A.1 に歌唱者の一覧を，表 A.2 に楽曲の一覧を示す．

A.2　3.6.1 節で用いたデータセット

3.6.1 節で述べた，ダイアライゼーションシステム全体の評価実験において用いたデータセットの詳細を記す．データセットは，学習用データセット，同時歌唱者数推定のための学習・評価用データセット，評価用データセットに分かれ，それぞれデータセット B，C，D とよぶ．

A.2.1　データセット B: 学習用データセット

データセット A と同様に，各音楽音響信号には，各歌唱者がそれぞれ独立に歌唱した音楽音響信号が，伴奏音とともにミキシングおよびマスタリングされた上で収録されている．またこのうち表中で○で示した 16 曲には，ボーカルのミキシングされていない伴奏音のみのカラオケトラックが存在し，実験ではこれを用いて差分を計算し，歌声のみの音響信号を得ている．一方その他の 37 曲は伴奏音のみの音響信号が用意されていない，あるいはミックスされたトラックとカラオケトラックのマスタリング環境が異なるなどの理由により，単純には伴奏音との差分を得ることができない．そこで本論文では，同一楽曲を複数人がそれぞれ個別に歌唱した音楽音響信号に対して，時間的にサンプル単位で同期させ，スペクトログラム $m_{s,t,f}$ を得て，このスペクトログラムの各周波数・時刻ビンにおける最小値 $\hat{m}_{t,f} = \arg\min_s \{m_{s,t,f} \mid 1 \leq s \leq S\}$ を計算し，$\hat{m}_{t,f}$ から Griffin–Lim 法 [45] により位相推定することで，伴奏音の音響信号を得る．ここで S は楽曲を歌唱する歌唱者数，s は歌唱者のインデックス，t および f は時間および周波数ビンのインデックスである．このような手法で構成した分離信号は，必ずしも高い品質の分離信号とはなっていない場合があるが，実験中ではこの信号をさらに伴奏音とミックスした上で学習に用いるため，実験には影響しない．

表 A.3 に歌唱者の一覧を，表 A.4 に楽曲の一覧を示す．

A.2.2　データセット C: 同時歌唱者数推定の学習・評価に用いたデータセット

本データセットは，楽曲を編集することなくそのまま用いた．人数の正解ラベルについては筆者の聴取により作成した．表 A.5 に歌唱者の一覧を，表 A.6 に楽曲の一覧を示す．

A.2.3　データセット D: システムの評価に用いたデータセット

本データセットは，楽曲を編集することなくそのまま用いた．各時刻の歌唱者の発声状態のラベルは筆者の聴取により作成した．表 A.7 に歌唱者の一覧を，表 A.8 に楽曲の一覧を示す．表 A.7 に示されているデータセット D の歌唱者には，表 A.3 および表 A.5 で示されたデータセット B，C の歌唱者が含まれないことに留意する．

[*1] 2 人のキャラクターを区別せず収録されている．

表 A.1　データセット A に含まれる楽曲を歌唱する歌唱者.

番号	歌唱者名
AS1	天海春香（中村繪里子）
AS2	如月千早（今井麻美）
AS3	星井美希（長谷川明子）
AS4	萩原雪歩（長谷優里奈）
AS5	萩原雪歩（浅倉杏美）
AS6	高槻やよい（仁後真耶子）
AS7	菊地真（平田宏美）
AS8	水瀬伊織（釘宮理恵）
AS9	四条貴音（原由実）
AS10	秋月律子（若林直美）
AS11	三浦あずさ（たかはし智秋）
AS12	双海亜美／真美[※1]（下田麻美）
AS13	我那覇響（沼倉愛美）
AS14	佐々木千枝（今井麻夏）
AS15	櫻井桃華（照井春佳）
AS16	市原仁奈（久野美咲）
AS17	龍崎薫（春瀬なつみ）
AS18	赤城みりあ（黒沢ともよ）

表 A.2　データセット A に含まれる楽曲.

目的	番号	楽曲名	曲長	歌唱者
学習	AM1	9:02 pm	2:03	AS1–AS4, AS6–AS8, AS10–AS12
	AM2	エージェント夜を往く	2:04	AS1–AS4, AS6, AS8, AS10–AS12
	AM3	蒼い鳥	2:06	AS1–AS4, AS6–AS8, AS10–AS12
	AM4	おはよう!! 朝ご飯	2:01	AS1–AS4, AS6–AS8, AS10–AS12
	AM5	きゅんっ! ヴァンパイアガール	2:00	AS1–AS3, AS5–AS13
	AM6	キラメキラリ	2:02	AS1–AS4, AS6–AS13
	AM7	Little Match Girl	2:08	AS1–AS3, AS5–AS13
	AM8	迷走 Mind	2:04	AS1–AS4, AS6–AS13
	AM9	relations	2:05	AS1–AS4, AS6–AS8, AS10–AS12
	AM10	隣に…	2:04	AS1–AS4, AS6–AS13
評価（closed-singer）	AM11	Kosmos, Cosmos	2:01	AS1–AS4, AS6–AS13
評価（open-singer）	AM12	ハイファイ☆デイズ	3:49	AS14–AS18

表 A.3　データセット B に含まれる楽曲を歌唱する歌唱者.

番号	歌唱者名
BS1	天海春香（中村繪里子）
BS2	如月千早（今井麻美）
BS3	星井美希（長谷川明子）
BS4	萩原雪歩（長谷優里奈）
BS5	萩原雪歩（浅倉杏美）
BS6	高槻やよい（仁後真耶子）
BS7	菊地真（平田宏美）
BS8	水瀬伊織（釘宮理恵）
BS9	四条貴音（原由実）
BS10	秋月律子（若林直美）
BS11	三浦あずさ（たかはし智秋）
BS12	双海亜美／真美（下田麻美）
BS13	我那覇響（沼倉愛美）
BS14	高坂穂乃果（新田恵海）
BS15	絢瀬絵里（南條愛乃）
BS16	南ことり（内田彩）
BS17	園田海未（三森すずこ）
BS18	星空凛（飯田里穂）
BS19	西木野真姫（Pile）
BS20	東條希（楠田亜衣奈）
BS21	小泉花陽（久保ユリカ）
BS22	矢澤にこ（徳井青空）

表 A.4　データセット B に含まれる楽曲.

目的	番号	カラオケトラックあり	楽曲名	曲長	歌唱者
学習	BM1	○	エージェント夜を往く	2:04	BS1–4, BS6, BS8, BS10–12
	BM2	○	蒼い鳥	2:06	BS1–4, BS6–8, BS10–12
	BM3	○	おはよう!! 朝ご飯	2:01	BS1–4, BS6–8, BS10–12
	BM4	○	きゅんっ! ヴァンパイアガール	2:00	BS1–3, BS5–13
	BM5	○	キラメキラリ	2:02	BS1–4, BS6–13
	BM6	○	Kosmos, Cosmos	2:01	BS1–4, BS6–13
	BM7	○	Little Match Girl	2:08	BS1–3, BS5–13
	BM8	○	迷走 Mind	2:04	BS1–4, BS6–13
	BM9	○	relations	2:05	BS1–4, BS6–8, BS10–12
	BM10	○	隣に…	2:04	BS1–4, BS6–13
	BM11	○	フタリの記憶	2:03	BS3, BS6, BS4, BS11, BS13
	BM12	○	Honey Heartbeat	2:02	BS1–3, BS5–13
	BM13	○	MEGARE!	2:03	BS1–3, BS5–13
	BM14	○	スタ→トスタ→	2:00	BS1–4, BS6–13
	BM15	○	THE IDOLM@STER	2:03	BS1–4, BS6–8, BS10–12
	BM16		だってだって噫無情	5:31	BS14–22
	BM17		愛してるばんざーい!	4:51	BS14–22
	BM18		Angelic Angel	4:55	BS14–22
	BM19		嵐のなかの恋だから	5:12	BS14–22
	BM20		僕たちはひとつの光	4:52	BS14–22
	BM21		僕らは今のなかで	4:34	BS14–22
	BM22		COLORFUL VOICE	3:39	BS14–22
	BM23		Dancing stars on me!	4:30	BS14–22
	BM24		どんなときもずっと	4:35	BS14–22
	BM25		Happy maker!	4:54	BS14–22
	BM26		HEART to HEART!	4:37	BS14–22
	BM27		輝夜の城で踊りたい	4:31	BS14–22
	BM28		KiRa-KiRa Sensation!	4:55	BS14–22
	BM29		これからの Someday	4:25	BS14, BS16–18, BS21, BS22
	BM30		LOVELESS WORLD	5:13	BS14–22
	BM31		Oh, Love & Peace!	4:58	BS14–22
	BM32		Love wing bell	4:45	BS15, BS18–22
	BM33		ミは μ'sic のミ	4:35	BS14–22
	BM34		もぎゅっと “love” で接近中!	5:43	BS14–22
	BM35		MOMENT RING	6:09	BS14–22
	BM36		Music S.T.A.R.T!!	4:54	BS14–22
	BM37		No brand girls	4:04	BS14–22
	BM38		Paradise Live	5:08	BS14–22
	BM39		るてしキスキしてる	5:29	BS14–22
	BM40		さようならへさよなら!	5:06	BS14–22
	BM41		きっと青春が聞こえる	4:07	BS14–22
	BM42		SENTIMENTAL StepS	3:57	BS14–22
	BM43		Shangri-La Shower	5:09	BS14–22
	BM44		それは僕たちの奇跡	4:15	BS14–22
	BM45		START:DASH!!	4:18	BS14–22
	BM46		SUNNY DAY SONG	4:45	BS14–22
	BM47		Super LOVE=Super LIVE!	5:38	BS14–22
	BM48		タカラモノズ	3:44	BS14–22
	BM49		WILD STARS	4:19	BS14–22
	BM50		Wonderful Rush	4:42	BS14–22
	BM51		ユメノトビラ	4:48	BS14–22
開発	BM52	○	9:02 pm	2:03	BS1–4, BS6–8, BS10–12
	BM53		Wonder zone	3:57	BS14–22

表 A.5　データセット C に含まれる楽曲を歌唱する歌唱者.

目的	番号	歌唱者名
学習・評価	CS1	天海春香（中村繪里子）
	CS2	如月千早（今井麻美）
	CS3	星井美希（長谷川明子）
	CS4	萩原雪歩（長谷優里奈）
	CS5	萩原雪歩（浅倉杏美）
	CS6	高槻やよい（仁後真耶子）
	CS7	菊地真（平田宏美）
	CS8	水瀬伊織（釘宮理恵）
	CS9	四条貴音（原由実）
	CS10	秋月律子（若林直美）
	CS11	三浦あずさ（たかはし智秋）
	CS12	双海亜美／真美（下田麻美）
	CS13	我那覇響（沼倉愛美）
	CS14	水谷絵理（花澤香菜）
	CS15	高坂穂乃果（新田恵海）
	CS16	絢瀬絵里（南條愛乃）
	CS17	南ことり（内田彩）
	CS18	園田海未（三森すずこ）
	CS19	星空凛（飯田里穂）
	CS20	西木野真姫（Pile）
	CS21	東條希（楠田亜衣奈）
	CS22	小泉花陽（久保ユリカ）
	CS23	矢澤にこ（徳井青空）
評価のみ	CS24	月岡恋鐘（礒部花凛）
	CS25	田中摩美々（菅沼千紗）
	CS26	白瀬咲耶（八巻アンナ）
	CS27	三峰結華（成海瑠奈）
	CS28	幽谷霧子（結名美月）
	CS29	関裕美（会沢紗弥）
	CS30	白菊ほたる（天野聡美）
	CS31	森久保乃々（高橋花林）
	CS32	鷹富士茄子（森下来奈）

表 A.6　データセット C に含まれる楽曲.

目的	番号	楽曲名	曲長	歌唱者
学習	CM1	GO MY WAY!!（M@STER VERSION）	4:50	CS1–4, CS6–8, CS10–12
	CM2	LOST	5:52	CS2, CS8, CS11, CS12, CS14
	CM3	My Best Friend（M@STER VERSION）	4:21	CS4, CS6, CS9, CS10, CS12
	CM4	いっしょ	4:51	CS1, CS3, CS6, CS11, CS12
	CM5	またね（M@STER VERSION）	5:02	CS1, CS4, CS7, CS8, CS10, CS13
	CM6	エージェント夜を往く（M@STER VERSION）	4:11	CS7, CS9, CS12, CS13
	CM7	キミはメロディ（M@STER VERSION）	4:01	CS2, CS3, CS6, CS9, CS11, CS12
	CM8	after school NAVIGATORS	4:48	CS19, CS22, CS23
	CM9	Cutie Panther	4:30	CS16, CS20, CS23
	CM10	Listen to my heart!!	4:28	CS19, CS22, CS23
	CM11	Love marginal	4:35	CS15, CS17, CS22
	CM12	Mermaid festa vol.2 〜Passionate〜	5:25	CS15, CS19
	CM13	Pure girls project	4:35	CS15, CS17, CS22
	CM14	soldier game	3:42	CS16, CS18, CS20
	CM15	START:DASH!!	4:17	CS14–23
	CM16	sweet & sweet holiday	4:17	CS14, CS17, CS22
	CM17	UNBALANCED LOVE	4:26	CS14, CS16, CS22
	CM18	これからの Someday	4:26	CS14, CS16–20, CS22, CS23
	CM19	微熱から Mystery	3:19	CS18, CS19, CS21
開発セット（closed-singer）	CM20	思い出をありがとう（M@STER VERSION）	4:29	CS2, CS3, CS7, CS10, CS13
	CM21	告白日和、です!	3:38	CS16, CS22
開発セット（open-singer）	CM22	NEO THEORY FANTASY	5:50	CS24–28
	CM23	ラビリンス・レジスタンス	3:37	CS24–28
	CM24	ステップ&スキップ（M@STER VERSION）	4:16	CS29–31
	CM25	幸せの法則 〜ルール〜（M@STER VERSION）	4:46	CS30, CS32

表 A.7　データセット D に含まれる楽曲を歌唱する歌唱者.

番号	歌唱者名
DS1	大崎甘奈（黒木ほの香）
DS2	大崎甜花（前川涼子）
DS3	桑山千雪（芝崎典子）
DS4	中谷育（原嶋あかり）
DS5	七尾百合子（伊藤美来）
DS6	松田亜利沙（村川梨衣）
DS7	真壁瑞希（阿部里果）
DS8	白石紬（南早紀）
DS9	北沢志保（雨宮天）
DS10	桜守歌織（香里有佐）
DS11	豊川風花（末柄里恵）
DS12	北上麗花（平山笑美）
DS13	馬場このみ（髙橋ミナミ）
DS14	高山紗代子（駒形友梨）
DS15	横山奈緒（渡部優衣）
DS16	高坂海美（上田麗奈）
DS17	佐竹美奈子（大関英里）
DS18	福田のり子（浜崎奈々）
DS19	徳川まつり（諏訪彩花）
DS20	エミリー スチュアート（郁原ゆう）
DS21	ロコ（中村温姫）
DS22	舞浜歩（戸田めぐみ）
DS23	永吉昴（斉藤佑圭）
DS24	周防桃子（渡部恵子）
DS25	木下ひなた（田村奈央）
DS26	箱崎星梨花（麻倉もも）
DS27	大神環（稲川英里）
DS28	望月杏奈（夏川椎菜）
DS29	速水奏（飯田友子）
DS30	塩見周子（ルゥ・ティン）
DS31	城ヶ崎美嘉（佳村はるか）
DS32	宮本フレデリカ（高野麻美）
DS33	一ノ瀬志希（藍原ことみ）
DS34	佐々木千枝（今井麻夏）
DS35	櫻井桃華（照井春佳）
DS36	市原仁奈（久野美咲）
DS37	龍崎薫（春瀬なつみ）
DS38	赤城みりあ（黒沢ともよ）
DS39	本田未央（原紗友里）
DS40	佐久間まゆ（牧野由依）
DS41	鷺沢文香（M・A・O）
DS42	輿水幸子（竹達彩奈）
DS43	新田美波（洲崎綾）
DS44	島村卯月（大橋彩香）
DS45	渋谷凛（福原綾香）
DS46	安部菜々（三宅麻理恵）
DS47	小日向美穂（津田美波）
DS48	相葉夕美（木村珠莉）
DS49	上条春菜（長島光那）
DS50	早坂美玲（朝井彩加）
DS51	木村夏樹（安野希世乃）
DS52	高海千歌（伊波杏樹）
DS53	桜内梨子（逢田梨香子）
DS54	渡辺曜（斉藤朱夏）
DS55	津島善子（小林愛香）
DS56	国木田花丸（高槻かなこ）
DS57	黒澤ルビィ（降幡愛）

表 A.8　データセット D に含まれる楽曲.

番号	楽曲名	曲長	歌唱者
DM1	アルストロメリア	4:47	DS1–3
DM2	ハピリリ	4:23	DS1–3
DM3	ZETTAI×BREAK!! トゥインクルリズム	4:27	DS4–6
DM4	Tomorrow Program	4:59	DS4–6
DM5	Melty Fantasia	4:23	DS7–9
DM6	I.D ～EScape from Utopia～	3:48	DS7–9
DM7	花ざかり Weekend ✤	4:12	DS10–13
DM8	RED ZONE	4:11	DS10–13
DM9	咲くは浮世の君花火	4:21	DS14–18
DM10	BORN ON DREAM! ～HANABI ☆ NIGHT～	4:06	DS14–18
DM11	だってあなたはプリンセス	4:42	DS19, DS20
DM12	ミラージュ・ミラー	4:31	DS19, DS20
DM13	月曜日のクリームソーダ	4:23	DS21–24
DM14	I did+I will	4:38	DS21–24
DM15	ピコピコ IIKO! インベーダー	4:18	DS25–28
DM16	Get lol! Get lol! SONG	4:18	DS25–28
DM17	dans l’obscurité	5:24	DS5, DS7, DS21, DS23, DS28
DM18	囚われの TeaTime	5:33	DS5, DS7, DS21, DS23, DS28
DM19	Tulip（M@STER VERSION）	3:53	DS29–33
DM20	ハイファイ☆デイズ（M@STER VERSION）	3:49	DS34–38
DM21	イリュージョニスタ!（M@STER VERSION）	4:17	DS39–DS43
DM22	Yes! Party Time!!（M@STER VERSION）	4:21	DS38, DS39, DS44–46
DM23	shabon song	4:40	DS31, DS38, DS47–49
DM24	ガールズ・イン・ザ・フロンティア（M@STER VERSION）	4:14	DS30, DS45, DS47, DS50, DS51
DM25	夢で夜空を照らしたい	5:34	DS52–57