修 士 論 文 の 和 文 要 旨

研究科・専攻	大学院　情報理工学研究科　　情報・ネットワーク工学専攻　博士前期課程		
氏　　名	五反田 聖矢	学籍番号	2031072
論 文 題 目	リモート合唱における発声タイミングずれの評価と自動補正法		

要　旨

新型コロナウイルス感染拡大に伴い従来の合唱が困難な中，複数の個人歌唱録音をミキシングすることで合唱作品を制作する「リモート合唱」が注目を集めている．以前はプロのエンジニアによってリモート合唱のミキシングが行なわれていたが，音源数が増えるにつれ，手動でのミキシングは困難との意見があった．そこで，著者はリモート合唱のミキシングを自動化するべく研究を開始した．

リモート合唱のミキシングにおいて重要なのが，歌唱者の発声タイミングを揃えることである．研究を進める中で，リモート合唱にはシステム由来のずれと歌唱者由来のずれという，2 種類の発声タイミングずれが存在することを発見した．システム由来のずれとは，歌唱録音の時間軸が異なることが原因のずれである．一方で，歌唱者由来のずれとは，歌唱者が周りの声を聞くことができないというリモート合唱特有の歌唱環境が原因で生じるずれである．これらのずれを解消するべく，本研究では 3 点の目的を設定した．第一に，システム由来のずれを自動的に解消する手法を提案すること，第二に，歌唱者由来のずれを定量的に分析すること，第三に，歌唱者由来のずれを自動的に補正する手法を提案することである．

第一の目的のためには，マーク信号と呼ぶ音信号を目印に時間軸を揃える手法を提案し，実験により実用上十分な精度が得られることが確かめられた．第二の目的のためには，FN 法と呼ぶオンセット検出の手法，DP マッチングを用いたオンセット対応付けの手法を提案した．これらの手法を用いて 2 曲のリモート合唱での歌唱者由来のずれを分析した結果，ずれ量の分布は平均−8 ms，標準偏差 51 ms となり，舞台上の合唱よりも大きくずれている可能性が示された．第三の目的のためには，計測したずれ量をもとに時間シフトする補正手法を提案し，実験によりその有効性が確かめられた．

応用として，本研究の成果をリモート合唱用 web プラットフォームに実装した．また，本研究は新聞やテレビなどで取り上げられたほか，提案手法を用いて制作したリモート合唱作品がテレビ・ラジオ番組，音楽イベント等で公開された実績を持つ．

令和 3 年度　修士論文

リモート合唱における
発声タイミングずれの評価と
自動補正法

電 気 通 信 大 学

情 報 理 工 学 研 究 科

情報・ネットワーク工学専攻

電子情報学プログラム

2031072　五反田　聖矢

指導教員　高橋 弘太 准教授

副指導教員　張 熙 教授

提出日 令和4年1月28日

目 次

第 1 章
序論

1.1 研究背景

2019 年末に発生した新型コロナウイルス（以後，新型コロナ）は，やがて世界各地に拡大し，現在もなお大きな社会問題となっている．新型コロナによる影響は，合唱活動にまで及ぶ．合唱では多人数が密集して発声するため，飛沫による感染リスクが高い．2020 年 3 月には，オランダのアムステルダム混声合唱団で，公演後に合唱団 130 人中 102 人が新型コロナに感染し，悲劇的なニュースとして報道された [1]．これを機に，日本においてもアマチュア合唱団の活動や，小・中学校における合唱教育を休止せざるを得ない状況が続いた．日本の合唱人口は 300 万人とも言われることから [2]，新型コロナが合唱に与えた影響の大きさは看過できない．

2020 年 8 月には，福島県の合唱サークルにて，感染対策を行なったにもかかわらず合唱によるクラスター感染が発生してしまった [3]．報道によれば，飛沫による感染防止のため歌唱者にマスクやフェイスシールドを着用させ，歌唱者間の距離を従来よりも広げ，換気も行なっていた．それでも感染を防げなかったことから，安全に合唱を行なうためには，各歌唱者が完全に隔たれた空間で歌唱することが重要だと考えられる．この場合，歌唱を一つにミキシングするための手立てが必要となる．その手立てとしては，遠隔合奏 [4] の技術が有効だと考えられる．

1.1.1 遠隔合奏の従来製品

遠隔合奏とは，遠隔地にいる複数の演奏者で合奏を行なうことである．近年では，遠隔合奏をサポートする製品として，YAMAHA 社のモバイル端末向けアプリケーション「SYNCROOM」[4] が開発された．また，遠隔で合奏を行なうという点で類似したサービスとして，第一興商社が運営するカラオケ「DAM」の会員制サービス「コラボ録音」[5] と「コラボ動画」[6] が存在する．各製品の特徴を，表 1.1 に示す．

表 1.1: 遠隔合奏の従来製品の特徴 [4][5][6]

	SYNCROOM	コラボ録音	コラボ動画
作品コンテンツ	音声・動画	音声	音声・動画
ミキシング方式	リアルタイム	非リアルタイム	非リアルタイム
収録場所	任意	カラオケ店内	カラオケ店内
最大演奏人数[1]	5 人	14 人	6 人

「SYNCROOM」は，合唱に限らず多人数での合奏を遠隔で行なうことを目的としている．YAMAHA 社が独自に開発した遠隔合奏技術「NETDUETTO」[4] が導入され，リアルタイムでのミキシングを実現している．これにより，他人の演奏を聞きながら同時に演奏することが可能である．収録は参加者が所持する端末によって行なうため，任意の場所で収録が可能である．

一方,「コラボ録音」と「コラボ動画」は，カラオケにおけるデュエット歌唱を目的としている．こちらはリアルタイムでのミキシングとは異なり，参加者が伴奏または過去にミキシングされた歌声を聞きながら歌唱し，歌声を 1 トラックずつミキシングしていく方式である．収録はカラオケ店内にある専用の端末を用いて行なう．

こうした製品は，合唱への応用も期待される．しかし，従来製品を用いて遠隔で合唱作品を制作する場合，以下の 4 点が問題となる可能性がある．

第一に，最大演奏人数の制約である．従来の遠隔合奏技術で演奏できる最大人数は，最も多い製品で 14 人である．一方，一般的に合唱曲の歌唱人数はそれ以上である．例として，第 72 回全日本合唱コンクール全国大会（2019 年）大学職場一般部門に出演し

[1] 1 端末あたり 1 人ずつ収録した場合

た合唱団の平均人数は 45.8 人 [7] であった．このことから，従来製品を用いて一般的な規模の合唱作品を制作することは，人数の制約から困難であると言える．なお，1 端末で多人数の同時歌唱を収録すればこの制約は問題とならないが，新型コロナの感染防止という目的上，そのような使用法は考慮しないこととする．

第二に,「SYNCROOM」の場合，高速かつ安定した通信環境が必要となるために対応端末や使用回線が制約される点が挙げられる．たとえ人数の制約が無かったとしても，数十名が参加する合唱ともなれば，この制約がより顕著に影響すると考えられる．

第三に,「コラボ録音」と「コラボ動画」の場合，店頭のカラオケマシンに搭載された楽曲しか歌唱できないため，楽曲が制約される点が挙げられる．特に，一般的にはポピュラーでない合唱曲や新曲を歌唱する場合，カラオケマシンに搭載されていないことが多いため，問題となり得る．

第四に，3 製品全てに当てはまる問題として，舞台上の合唱の録音に近い作品に仕上げるには音量や各種エフェクトの高度な手動調整が必要となる点が挙げられる．歌唱者が多いほど調整も煩雑となるため，合唱作品を制作する上では障壁となる可能性がある．

このように，従来製品を使用して遠隔で合唱作品を制作することは，現状では困難であると考えられる．

1.1.2 リモート合唱の概要

従来製品では困難であった遠隔での合唱作品制作を実現するため，IT 技術により合唱を支援する団体「Harmorearth（ハモラス）」[8] は，リモート合唱の web プラットフォームである「tuttii（トゥッティ）」[9] を構築した．ここで，本研究におけるリモート合唱とは,「スマートフォン等に内蔵された既存の簡易的なアプリで個人歌唱を録音し，それらを非リアルタイムでミキシングして合唱作品を制作する活動」と定義する．

「tuttii」を使用したリモート合唱の概要を，図 1.1 に示す．

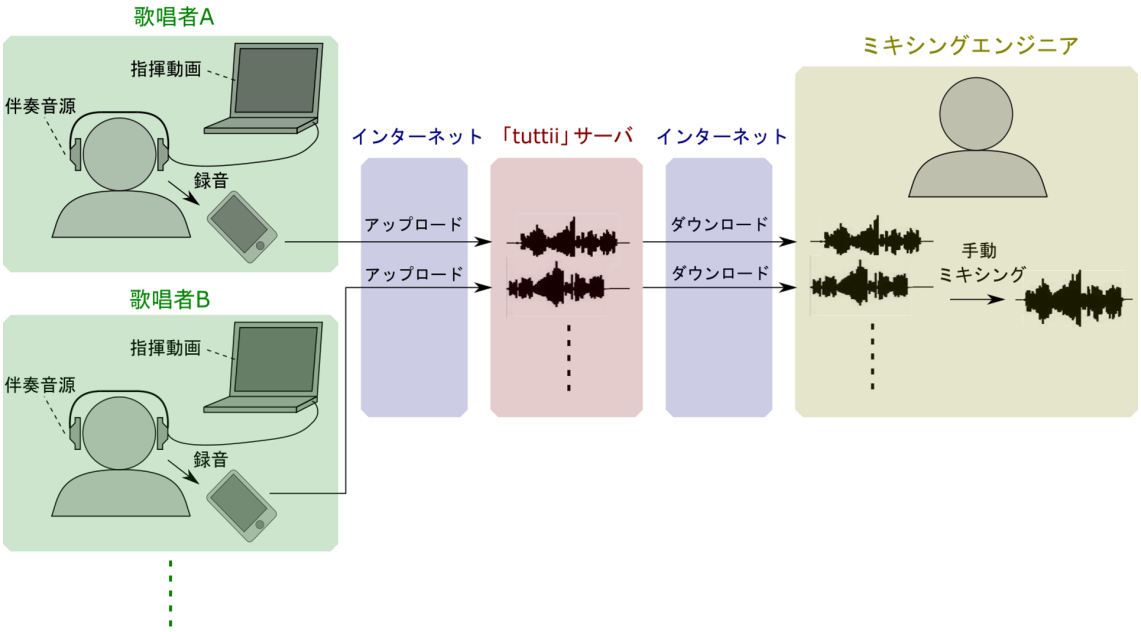

図 1.1: 本研究開始当初のリモート合唱の概要

歌唱者は端末を 2 台用意し，一方で事前収録された伴奏音源と指揮動画を再生，もう一方で個人歌唱を録音する．各歌唱者が録音データを「tuttii」サーバにアップロードし，全員の録音データが揃うと，プロのミキシングエンジニアが手動でミキシングを行なう仕組みである．

リモート合唱では，従来製品で想定された 4 点の問題点が，ユーザの立場においては解決された．その解決法を表 1.2 に示す．

表 1.2: 従来製品で想定された問題点の解決法

問題点	解決法
最大演奏人数の制約	音源数に制約のない専用プラットフォームを構築
対応端末や使用回線の制約	既存の簡易的な録音アプリを使用・非リアルタイムのミキシングを採用
楽曲の制約	ユーザが自由に指揮や伴奏を収録し，公開できる仕組みを構築
ユーザによる高度な手動調整	ミキシングエンジニアに手動ミキシングを依頼

リモート合唱の登場により，ユーザの立場では問題なく遠隔での合唱作品の制作が可能となった．しかし，ミキシングエンジニアからは，歌唱者が 100 人を超えるよう

な大規模なリモート合唱では，手動でミキシングすることが困難であるとの意見が寄せられた．このため，リモート合唱のミキシングを自動化することが望まれていた．

著者はこの問題を早急に解決すべきと考え，2020 年 6 月より「Harmorearth」協力のもと，リモート合唱の自動ミキシングに関する研究を行っている．ここでのミキシングとは，単に信号を加算するだけではない．リモート合唱を作品として成立させるためには，各録音データの時間軸を揃え，音量バランスを適切に調整し，リバーブなどのエフェクトを適切にかけることが重要である．さらに，規模の大きなリモート合唱では，一度に多数の音源に対して処理を施す必要がある．例として，図 1.2 に合唱曲「時代」のリモート合唱音源を示す．ここでは 1 パート 84 人分の音源のみを示すが，全体では 170 人分の音源が存在する．こうした多数の音源に対して，人間が手を加えることなく，ディジタル信号処理により全自動でミキシングを行なうことを目標としている．

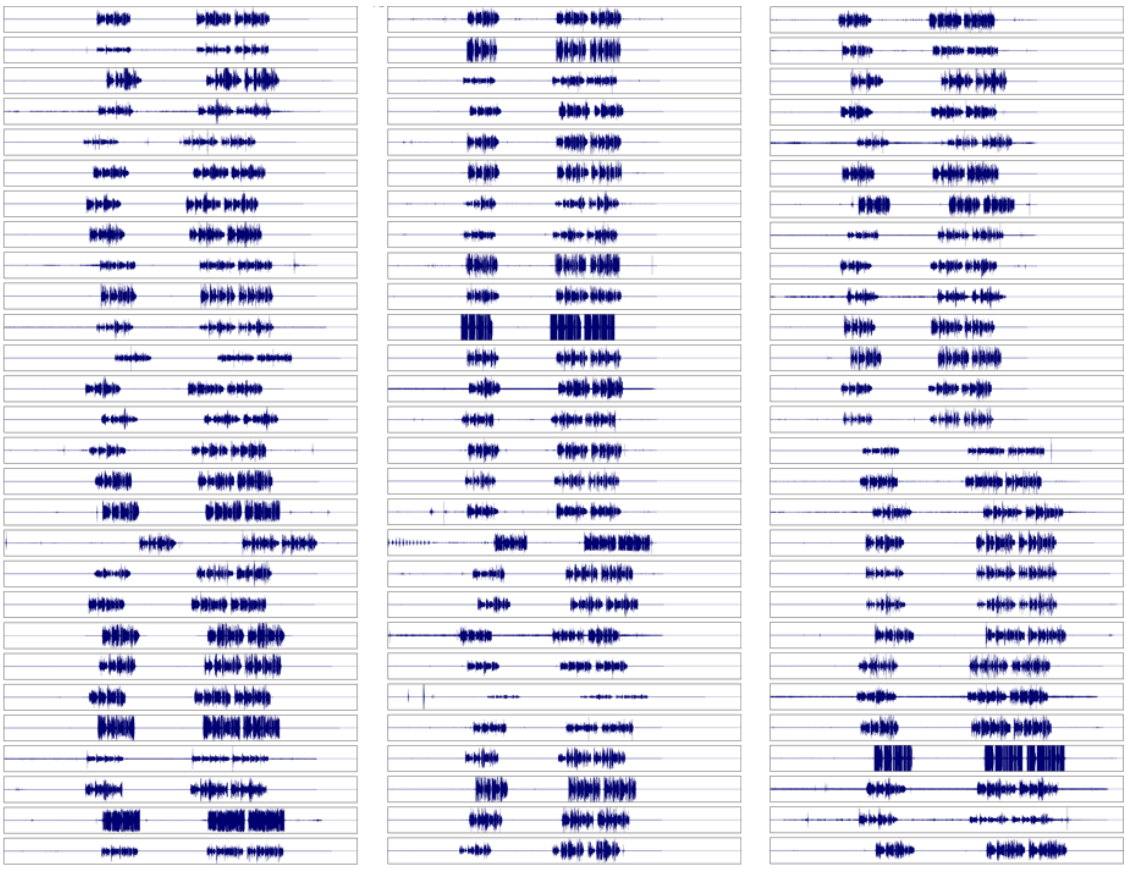

図 1.2: 「時代」のリモート合唱における 1 パート分の音源．各音源の時間長は 244 sec.

1.2 本研究の目的

本研究は，ディジタル信号処理を用いてリモート合唱のミキシングを自動化するための研究である．その中でも特に，各歌唱者の発声タイミングを自動で揃える処理に焦点を当てた．

リモート合唱では，各歌唱者が指揮や伴奏に対して録音を開始するタイミングが揃っていないと，各録音データの時間軸に相違が生じる．本研究では，この相違が原因となる発声タイミングずれを，システム由来の発声タイミングずれ，あるいは単にシステム由来のずれと定義する．録音データの時間軸を揃える仕組みが無い限り，システム由来のずれの発生は避けられない．よって，本研究では，まずリモート合唱における録音データの時間軸を自動的に揃える手法を提案する．

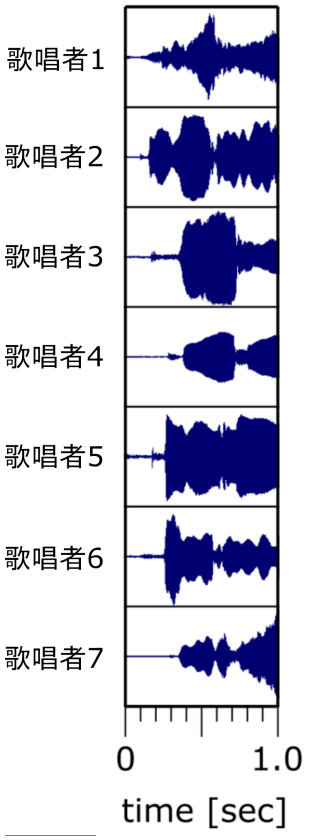

図 1.3:
歌唱者由来のずれの例

研究を進める中で著者は，システム由来のずれを解消しても，なお発声タイミングが大きくずれる場合があることを発見した．例として，合唱曲「心の瞳」のリモート合唱における，ソプラノパート 7 人の歌い出し部分の時間波形を図 1.3 に示す．図 1.3 の歌唱者 7 人の録音データは，時間軸を揃えてシステム由来のずれを解消しているが，歌い出し部分に最大で約 200 ms の時間差が存在する．この 7 人の歌唱を同時に聴取すると，合唱作品としての完成度を損なう程度に歌唱のばらつきを知覚する．このような発声タイミングのずれは，各歌唱者が周りの声を聞けない環境で歌唱するという，リモート合唱特有の歌唱環境が原因で発生すると考えられる．本研究では，このずれを歌唱者由来の発声タイミングずれ，あるいは単に歌唱者由来のずれと定義する．歌唱者由来のずれ量は，舞台上の合唱での発声タイミングずれ量よりも大きいことが予想される．しかし，それを裏付ける研究は過去に行われていない．そこで，本研究では，実験により歌唱者由来のずれ量を計測して定量的な分析を行なう．さらに，リモート合唱を舞台上での合唱に近づけるために，歌唱者由来のずれを自動的に補正する手法を提案する．

以上をまとめると，本研究の目的は 3 点存在する．本研究の目的は，第一に，システム由来のずれを自動的に解消する手法を提案すること．第二に，歌唱者由来のずれを定量的に分析すること．第三に，歌唱者由来のずれを自動的に補正する手法を提案することである．これらの目的を図示すると，図 1.4 のようにまとめられる．

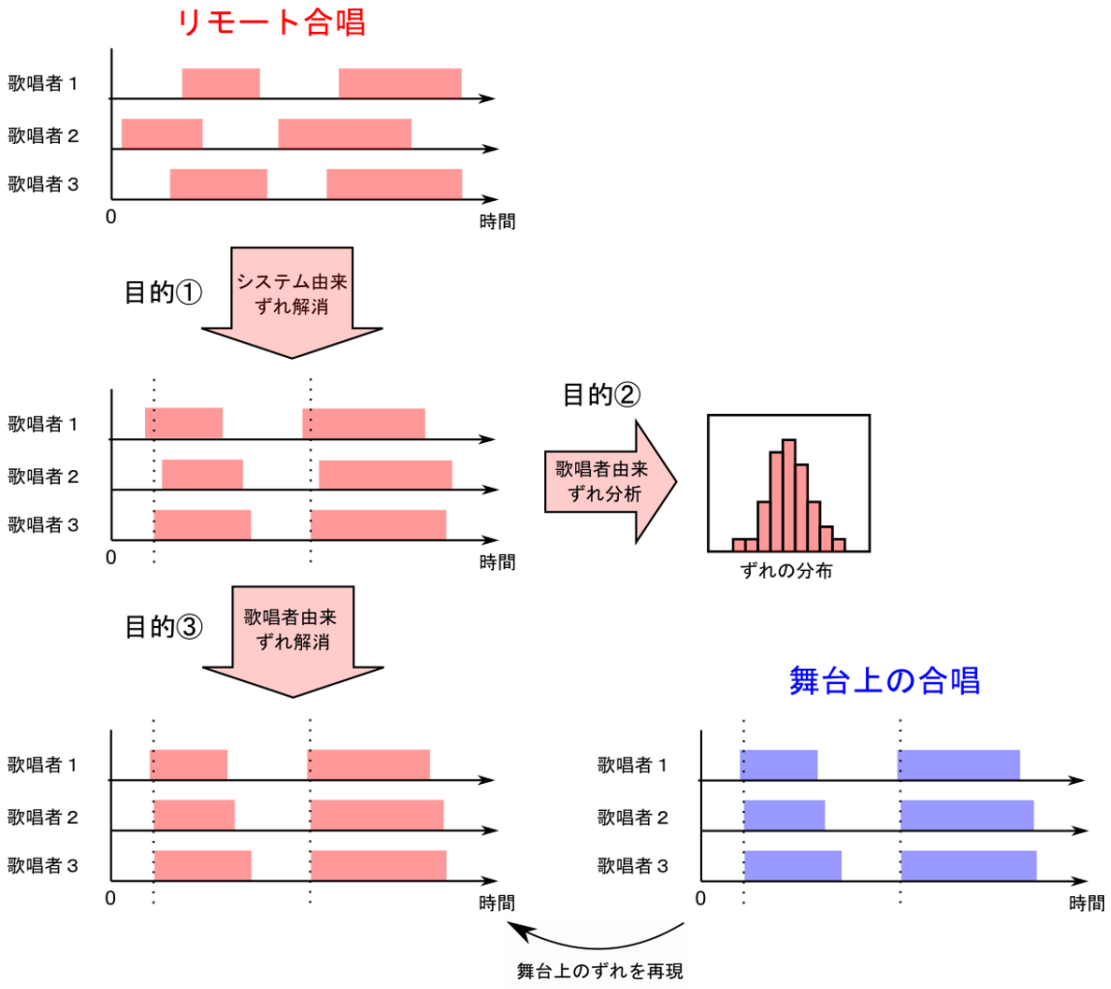

図 1.4: 本研究の目的

1.3 論文の構成

本論文を，全 7 章で構成する．第 2 章では，音楽合奏における同時性に関する先行研究，並びに，オンセット検出に関する先行研究について述べる．第 3 章では，システム由来の発声タイミングずれを自動的に解消する手法，歌唱者由来の発声タイミングずれを定量的に分析する手法，並びに，歌唱者由来の発声タイミングずれを自動的に補

正する手法を提案する．第 4 章では，第 3 章で提案した各手法の性能を評価するための実験，並びに，歌唱者由来のずれを分析する実験について述べる．実験結果の考察についても，この章で述べる．第 5 章では，本研究の応用として構築した自動ミキシングシステムの概要について述べる．また，著者が本研究の一環として行なった，リモート合唱に関する対外活動の実績についてもこの章で述べる．第 6 章では，リモート合唱研究の今後の展望について述べる．最後に，第 7 章では，本研究の結論と今後の課題について述べる．

第 2 章
先行研究

2.1 音楽合奏における同時性

本研究では，リモート合唱における歌唱者ごとの発声タイミングのずれに焦点を当てている．これは，音楽合奏における同時性を扱う研究に位置づけられる．先行研究として，そのメカニズムの解明や，演奏タイミングずれの計測を行なった研究が存在する．

2.1.1 認知メカニズム

音楽合奏では，演奏者間の 1 秒のずれですら失敗とみなされることから，他の集団行動と比較してメンバー間に非常に高いレベルの協調性が必要とされる [10]．演奏者同士が協調するための認知メカニズムとしては，auditory imagery の存在が示唆されている [11]．auditory imagery とは，聴覚に関する内的な感覚のことであり，ピッチや音色，旋律，テンポと言った音楽的要素も含まれる．その中でも，合奏での演奏タイミングの同期については，他の演奏者の先行する音から次の発音タイミングを予測するという，予測的な auditory imagery が関与することが指摘されている [12].

本研究のリモート合唱においては，歌唱者が他の歌唱者の先行音を聞くことができ

ないため，次の発声タイミングを決定するための予測的な auditory imagery を想起させることが，舞台上の合唱と比較して困難になると考えられる．このように，歌唱者由来の発声タイミングずれの問題は，音楽合奏での認知メカニズムの点から説明することが可能である．

2.1.2 演奏タイミングのずれ

音楽合奏において演奏者がずれたと認識し始める演奏タイミングのずれ量は，100 ms 以上であるとの研究結果がある [11]．言い換えれば，タイミングの合った合奏では 100 ms 以下のずれ量で演奏されているはずである．その具体的なずれ量を計測した研究について，器楽合奏と声楽合奏に分けて述べる．

まず，器楽合奏での演奏タイミングずれについて述べる．先行研究では，プロ演奏者によるリコーダーのトリオ，弦楽器のトリオ，ピアノデュオの各合奏におけるずれ量（絶対値）が計測されており，結果はいずれも 30~50 ms であった [11]．このほか，12 人のアマチュア奏者からなるヴァイオリンセクションでの演奏タイミングずれの分布を分析した研究が存在する．この研究によれば，分析結果は図 2.1 に示すような平均 40.35 ms，標準偏差 0.00 ms の分布であった [13]．

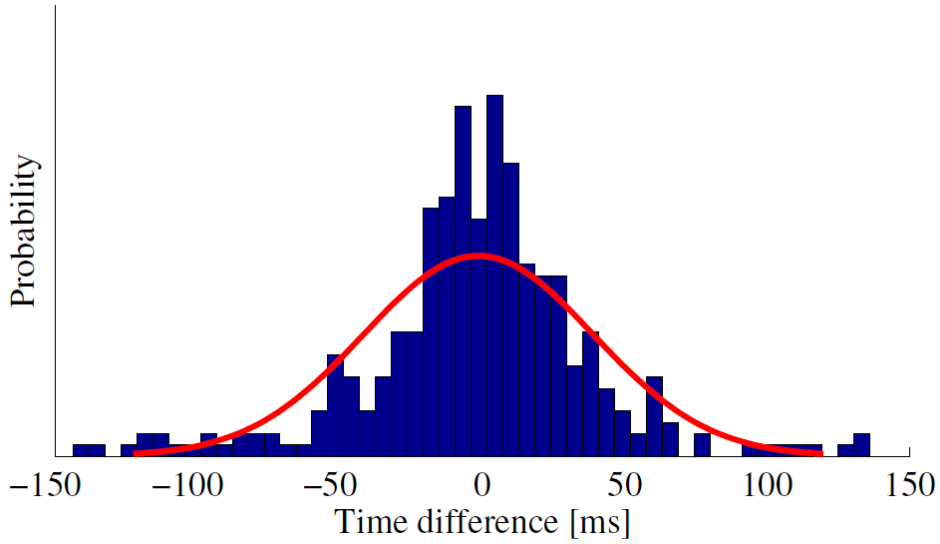

図 2.1: ヴァイオリンセクションにおける演奏タイミングずれ量のヒストグラム．実線は比較のためのガウス分布を示す．[13] の Figure 3 より引用．

次に，声楽合奏での演奏タイミングずれについて述べる．著者の調査した限り，合唱のように多数の歌唱者で行なう声楽合奏の演奏タイミングずれを計測した研究は存在しない．その理由は，歌唱者が多数の場合，歌唱者 1 人に 1 本ずつマイクロフォンを設置したとしても他の歌唱者からのクロストークが避けられず，正確な分析が困難であるからだと考えられる．ヴァイオリンセクションを対象とした研究 [13] では，各楽器の駒にコンタクトマイクを設置することでクロストークの問題を解決していた．しかし，人の歌唱を対象とする場合，電気声門図（EGG）などを利用しなければクロストークは避けられず，そうした研究は過去に行われていない．

舞台上の合唱における発声タイミングずれの研究例は無いが，ソロ歌唱者が伴奏を聴取しながら歌唱した場合に，発声タイミングが譜面上の拍点からどれだけずれるかを計測した研究は存在する [14]．この研究では，歌唱者のグルーヴ感が発声タイミングのずれに及ぼす影響を調べるため，プロ歌唱者が「通常通り」歌唱した場合と，「棒歌い」で歌唱した場合の 2 条件で計測が行なわれた．計測結果を図 2.2 に示す．この結果からは，母音オンセットは拍点を中心におよそ ±50 ms の範囲に分布し，子音オンセットは拍点よりも前に，母音オンセットよりも広い範囲に分布することが示されている．

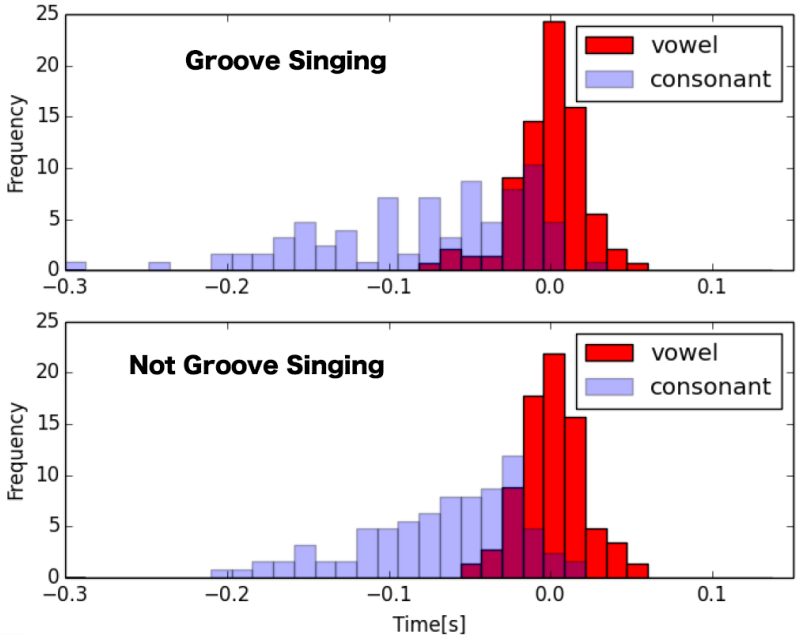

図 2.2: ソロ歌唱における，譜面上の拍点からのオンセットずれ量のヒストグラム．[14] の図 3 より引用．

本研究で行なう歌唱者由来の発声タイミングずれの分析は，歌唱者が伴奏のみを聴取しながら歌唱した場合のオンセット時刻を計測するという点で，ソロ歌唱を対象と

した研究 [14] と共通する．しかし，本研究では，複数の歌唱者間でオンセット時刻がどのように異なるのかを分析する点で新規性を持つと言える．

2.2 オンセット検出

発声タイミングずれを分析するためには，歌唱のオンセットを検出する必要がある．オンセット検出に関しては，先行研究で複数の手法が提案されている [15]．

2.2.1 オンセットの定義

はじめに，オンセットの定義を明確にするために，楽音を構成する時間区間について述べる．一般的に，楽音は音響的特徴によって分類された 4 つの時間区間から構成される [15][16]．各時間区間の定義を表 2.1 に示す．

表 2.1: 楽音を構成する 4 つの時間区間 [15][16]

名称	定義
アタック（attack）	振幅包絡が増加する区間
ディケイ（decay）	楽器本体の共振周波数での振動を残して減衰する区間
サスティン（sustain）	楽器本体の共振周波数での振動が持続する区間
リリース（release）	楽器本体の共振周波数での振動が減衰する区間

各時間区間とオンセットの関係を表した模式図を，図 2.3 に示す．

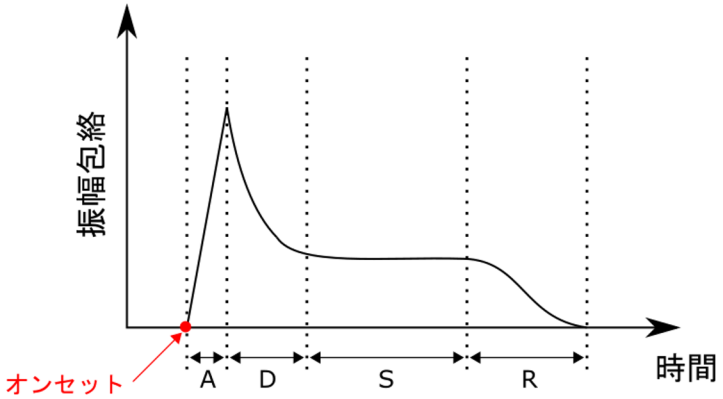

図 2.3: 4 つの時間区間とオンセットの関係 [15][16]．A:アタック，D:ディケイ，S:サスティン，R:リリースに対応する．

図 2.3 に示したように，一般的にオンセットはアタックの開始時刻のことを指す．本研究でもこの定義を用いる．

2.2.2 オンセット検出関数

オンセット検出アルゴリズムでは，原信号から局所的な構造を反映した中間信号へと変換し，そのピークを検出することでオンセット時刻を求める．原信号を中間信号に変換する関数は，オンセット検出関数と呼ばれる．ここでは，代表的な 3 点のオンセット検出関数を用いた手法について述べる．

- 手法 1：HFC（high frequency content）法 [15]

HFC 法 [15] では，信号のスペクトル特徴量を用いる．まず，信号 $x[i]$ の短時間フーリエ変換（STFT）を式 (2.1) に示す．

$$X[i,k] = \sum_{n=-\frac{N}{2}}^{\frac{N}{2}-1} x[ih+n]\ \omega[n]\ e^{-\frac{2j\pi nk}{N}} \tag{2.1}$$

ここで，i はサンプル番号，k は周波数ビン番号，$\omega[n]$ は N 点のウィンドウであり，h はウィンドウのシフト量を表す．信号のパワーは一般に低周波数域に集中する一方，トランジェントによる変動は高周波数域に現れる．これを強調するために，式 (2.2) のように重み付けを行なったパワー尺度を検出関数とする．

$$E_{HFC}[i] = \frac{1}{N} \sum_{k=-\frac{N}{2}}^{\frac{N}{2}-1} |k|\ |X[i,k]|^2 \tag{2.2}$$

HFC 法では，アタック区間で鋭いピークが得られるため，打楽器的なオンセットを検出するのに有効である．一方で，ピッチの低い音や，低周波数域に変化が集中する音に対するロバスト性は低い．

- 手法 2：スペクトル差（spectral difference）法 [15]

スペクトル差法 [15] では，各時刻におけるスペクトルを N 次元空間内の点と考え，それらの距離を検出関数として定式化する．スペクトル間の距離として L_1 ノルムを用いた場合の検出関数を式 (2.3) に示す．

$$E_{SD}[i] = \sum_{k=-\frac{N}{2}}^{\frac{N}{2}-1} H[|X[i,k]| - |X[i-1,k]|]^2 \tag{2.3}$$

ここで，$H[x] = \frac{x+|x|}{2}$ である．この操作により，パワーが増加する周波数ビンのみをカウントしている．

スペクトル差法では，HFC 法とは異なり広い周波数域におけるオンセットを検出できる．一方で，振幅情報のみに依存するため，調波音間の遷移や弦楽器音の立ち上がりといったパワー変化の小さいオンセットは検出が困難とされる．

- 手法 3：位相偏差（phase deviation）法 [15]

位相偏差法 [15] では，位相スペクトルの特徴量を用いる．まず，STFT 表現 $X[i,k]$ のアンラップされた位相を $\theta[i,k]$ とする．定常な正弦波音では，ある時刻の位相 $\theta[i,k]$ と 1 つ前のウィンドウの位相 $\theta[i-1,k]$ から，式 (2.4) によって瞬時周波数を推定することができる．

$$f[i,k] = \frac{\theta[i,k] - \theta[i-1,k]}{2\pi h} F_s \tag{2.4}$$

ここで，F_s はサンプリング周波数である．このとき，正弦波が定常であれば，ウィンドウ間の位相変化量は式 (2.5) のようにほぼ一定である．

$$\theta[i,k] - \theta[i-1,k] \simeq \theta[i-1,k] - \theta[i-2,k] \tag{2.5}$$

このことから，式 (2.6) のように位相偏差を位相の 2 階微分として定義できる．

$$\Delta\theta[i,k] = \theta[i,k] - 2\theta[i-1,k] + \theta[i-2,k] \simeq 0 \tag{2.6}$$

楽音の定常部分では位相偏差が 0 となるが，トランジェントでは瞬時周波数が一定とならないため，位相偏差は大きくなるはずである．これを利用し，式 (2.7) のように位相偏差の絶対値の平均をとることで検出関数とする．

$$E_{PD}[i] = \frac{1}{N} \sum_{k=1}^{N} |\Delta\theta[i,k]| \tag{2.7}$$

位相偏差法では，スペクトル差法や HFC 法とは異なり，パワー変化の小さいオンセットも検出可能である．ただし，定常的なピッチを持たない楽音のオンセットは検出が困難である．

2.2.3 オンセット検出関数の性能比較

前述の HFC 法，スペクトル差法，位相偏差法について，実験により性能比較を行なった研究が存在する [15]．実験では,「ピッチあり非打楽器性音」としてヴァイオリンソロ,「ピッチあり打楽器性音」としてピアノソロ,「ピッチなし打楽器性音」としてドラムを含むポップスについて，3 手法でオンセット検出が行われた．またその結果から，各手法の真陽性率（TP 率）と偽陽性率（FP 率）が計算された．実験結果を表 2.2, 表 2.3，表 2.4 に示す．この結果からは，ドラムのような打楽器性の強い楽音には HFC 法が，明確なピッチを持つ楽音にはオンセット差法や位相偏差法が比較的適していると結論付けられている．

表 2.2: ピッチあり非打楽器性音のオンセット検出結果 [15]．オンセット数：93

手法	TP 率 [%]	FP 率 [%]
HFC 法	81.7	14.7
オンセット差法	87.1	8.6
位相偏差法	95.7	4.3

表 2.3: ピッチあり打楽器性音のオンセット検出結果 [15]．オンセット数：489

手法	TP 率 [%]	FP 率 [%]
HFC 法	94.1	5.4
オンセット差法	94.9	1.6
位相偏差法	95.5	0.3

表 2.4: ピッチなし打楽器性音のオンセット検出結果 [15]．オンセット数：212

手法	TP 率 [%]	FP 率 [%]
HFC 法	96.7	0.0
オンセット差法	81.6	5.5
位相偏差法	80.7	5.5

第 3 章
提案手法

本研究の目的は，システム由来のずれを自動的に解消すること，歌唱者由来のずれを定量的に分析すること，歌唱者由来のずれを自動的に補正することの 3 点であった．これらを達成するための提案手法を，目的ごとに述べる．

3.1 システム由来のずれの解消

システム由来のずれの原因は，各歌唱者が指揮や伴奏に対して録音を開始するタイミングが異なるために，各録音データの時間軸に相違があることであった．そこで，リモート合唱における録音データの時間軸を揃える手法を提案する．

システム由来のずれを解消する最も単純な手法は，歌唱者全員が伴奏音源の再生開始と同時に録音を開始することである．しかし，この手法では歌唱者が録音ボタンを押下するタイミングの精度がそのまま反映されてしまう．さらに，伴奏音源が含まれる指揮者動画は，権利の関係からストリーミング再生とせざるを得ない場合が多い．その場合，動画のダウンロード時のバッファが原因で，再生ボタ ンの押下と再生開始時間が一致しないことがある．このような理由から，録音を開始するタイミングによらず，時間軸を自動的に揃える手法が必要となる．

手法を検討するにあたり，前提として，リモート合唱の各録音データには 1 人の歌唱

者による歌唱のみが含まれ，伴奏音は含まれないものとする．第一に考えられる手法は，オンセット検出 [15] や F0 推定 [17] などから得られた歌声特徴量の分布から，歌唱者間の相対的な時間関係を推定することである．ただし，この手法は同一パート内では有効な可能性があるが，リズムや音高の異なる別パートの間では使用できない．よって，この手法は不適当と判断する．第二に考えられる手法は，得られた歌声特徴量を楽譜情報と照らし合わせ，各歌唱の絶対的な歌唱位置を求めることである．しかし，あらゆる合唱曲の楽譜情報を網羅することは困難なため，この手法も適さないと判断する．第三に考えられる手法は，伴奏音源に時間軸の基準となる信号を挿入し，それを歌唱の録音に含めることで時間軸の目印とすることである．挿入する信号を適切に選択すれば，楽曲によらず汎用的に使用できると期待されるため，本研究ではこの手法を選択する．以後，時間軸の基準となる信号をマーク信号と呼ぶ．

3.1.1 相互相関関数

提案手法を数学的に定義する．まず，マーク信号を $x_m[i]$（$i = 1, 2, \cdots, L_m$）とし，これをヘッドフォンから再生しマイクロフォンを近づけて録音した信号を $x_o[i]$（$i = 1, 2, \cdots, L$）とする．ここで，i はサンプル番号であり，$L_m < L$ とする．録音データ $x_o[i]$ の中からマーク信号 $x_m[i]$ の位置を特定するには，これらの信号の相互相関を計算すれば良い．一般的な相互相関関数 [18] を式 (3.1) に示す．

$$\phi_o[i] = \frac{1}{L_m} \sum_{l=0}^{L_m-1} x_m[l]\ x_o[i+l] \tag{3.1}$$

$x_m[i]$ を適切に選択すれば，$\phi_o[i]$ はパルス状の関数となり，マーク信号の位置を特定できるはずである．しかし，この手法をリモート合唱に適用させるためには，式 (3.1) に 2 点の変更を加える必要がある．第一の変更点は，$x_o[i]$ のエネルギーで規格化することである．$x_o[i]$ にはマーク信号のほかに歌唱が録音されるが，多くの場合，歌唱はヘッドフォンから再生された $x_m[i]$ よりも大きなエネルギーで録音される．その場合，相互相関関数のピークが偶発的に歌唱音声中に現れてしまうことが考えられる．これを防ぐため，$x_o[i]$ のエネルギーで規格化を行なう必要がある．第二の変更点は，相互相関を絶対値で評価することである．これは，ヘッドフォンから再生された音をマイクロフォンで受信する場合，再生系によっては $x_o[i]$ に録音された $x_m[i]$ の位相が反転している場合があるためである．以上の変更を加えた相互相関関数を，式 (3.2) に示す．

$$\phi[i] = \frac{|\sum_{l=0}^{L_m-1} x_m[l]\ x_o[i+l]|}{\sqrt{\sum_{l=0}^{L_m-1} x_o[i+l]^2}} \tag{3.2}$$

ここで，$\phi[i]$ を最大化する i を i_{max} とすると，式 (3.3) に示す時間シフトによって任意の $x_o[i]$ を時間軸が一元化された歌唱信号 $x[i]$ に変換することが可能となる．

$$x[i] = x_o[i + i_{max} + L_m] \tag{3.3}$$

3.1.2 マーク信号

提案手法に用いる具体的なマーク信号について述べる．マーク信号に求められる条件は，次の 3 点が考えられる．第一に，高い時間分解能を有すること．第二に，部分的に録音が欠けても機能すること．第三に，人間が聴取して不快ではないことである．特に，第三の条件は本研究特有のものであり，リモート合唱に実用化する上で重要となる．これらの条件を満たす音信号として，8 個のチャープ信号から構成される，長さ 3.62 sec のマーク信号「mark8」を提案する．「mark8」の時間周波数平面と時間波形を，図 3.1 に示す．

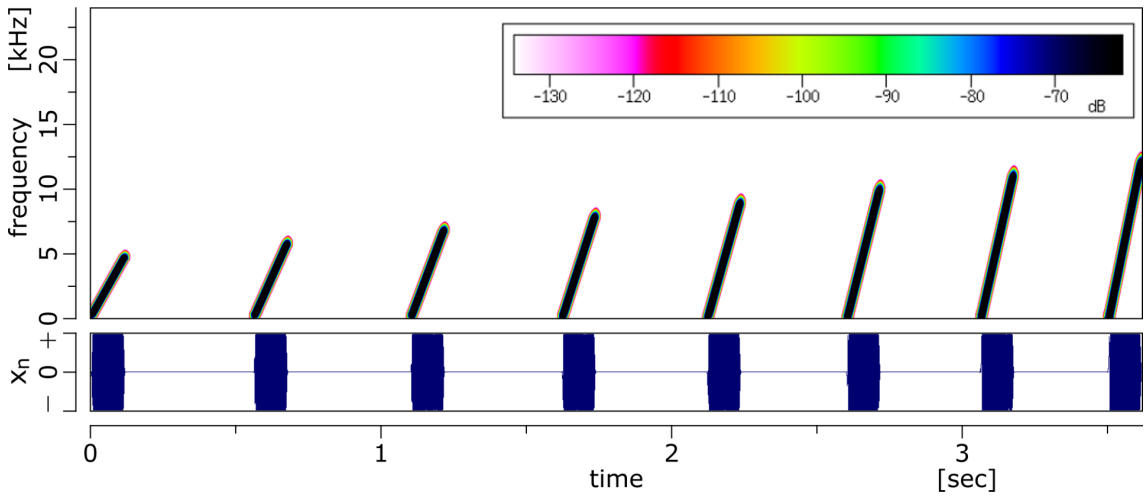

図 3.1: 「mark8」の時間周波数平面（上）と時間波形（下）

8 個のチャープ信号はそれぞれ周波数変化率が異なるものとし，各チャープの間隔は不均等としてある．これにより，8 個あるチャープ信号のうち少なくとも 1 つが録音されていれば，理論上 $\phi[i]$ はパルス状の関数となる．よって，マーク信号の第一，第二

条件の両方を満足する．一方で，第三条件を満たすため，聴取した際に耳障りとならないように 12 kHz 以上の高周波数域は使用していない．また，チャープ信号の小気味良さも相まって，全体的には小鳥のさえずりのような聞こえの良い音となるように設計してある．

3.2 歌唱者由来のずれの分析

歌唱者由来のずれを分析するために，同一パート，すなわち同じ音高，リズム，歌詞を歌唱する集団における，歌唱者間での各オンセットのずれ量の計測を行なう．類似の計測を行なう先行研究 [13][14] では，1. オンセット検出，2. オンセットマッチング，3. オンセットずれ量の計測という 3 段階の手順で計測が行なわれた．本研究においてもこの手順に則る．ただし，図 3.2 に示すように，歌唱者由来のずれ量の計測は分析のためだけでなく，リモート合唱の自動ミキシングシステムの一部ともなる．よって，自動ミキシングにも対応した手法を選択する必要がある．

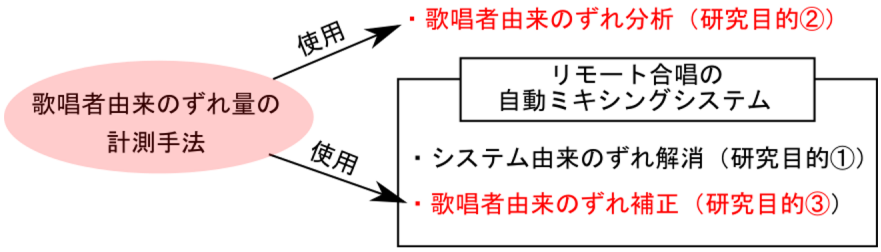

図 3.2: 歌唱者由来のずれ量計測手法の位置付け

手順の第 1 段階であるオンセット検出に関して，ヴァイオリンセクションを対象とした計測 [13] では，スペクトル差法 [15] が用いられた．一方で，ソロ歌唱を対象とした計測 [14] では，波形の目視と聴取による手動での検出が行なわれた．歌唱のオンセット検出に既存手法が用いられなかった理由として，母音と子音のオンセットを検出するにあたり，既存の手法では精度が 100 %に及ばないためとされた．しかし，合唱を対象とする本研究において，全員のオンセットを手動で検出することは，その作業量の膨大さから困難である．さらに，リモート合唱の自動ミキシングシステムに組み込むことから，この手順も自動化する必要がある．そこで，本研究では歌唱のオンセットを自動で検出する手法を提案する．

第2段階のオンセットマッチングは，リファレンスとなるオンセットと検出されたオンセットを対応付ける作業である．リファレンスとして通常用いられるのは，楽譜上の拍点である [13][14]．しかし，第 3.1 節で述べたように，本研究では楽譜情報を一切使用せずに処理を行なうことを目指す．そこで，歌唱の録音データのみからリファレンスを作成する手法を提案する．このリファレンスとなるオンセットを，本研究では標準オンセットと呼ぶ．先行研究におけるオンセットマッチングの作業は，リファレンスからの誤差時間が閾値以内にあるオンセットを対応付けるという手法 [13] が用いられた．しかし，著者が行なった先行実験では，この単純な手法では誤った対応付けが行なわれる場合があると判明した．そこで，本研究では，より合理的なオンセットマッチングの手法を提案する．

第3段階のオンセットずれ量の計測は，検出されたオンセットに対して，それと対応付けられたリファレンスのオンセットとの時間差を記録する作業である．本研究では，検出されたオンセットに対して，それとマッチングした標準オンセットとの時間差を記録することに相当する．

以上の手順をまとめたブロック図を，図 3.3 に示す．なお，表記の関係上，オンセットずれ量の計測はオンセットマッチングの処理ブロックに含めるものとする．

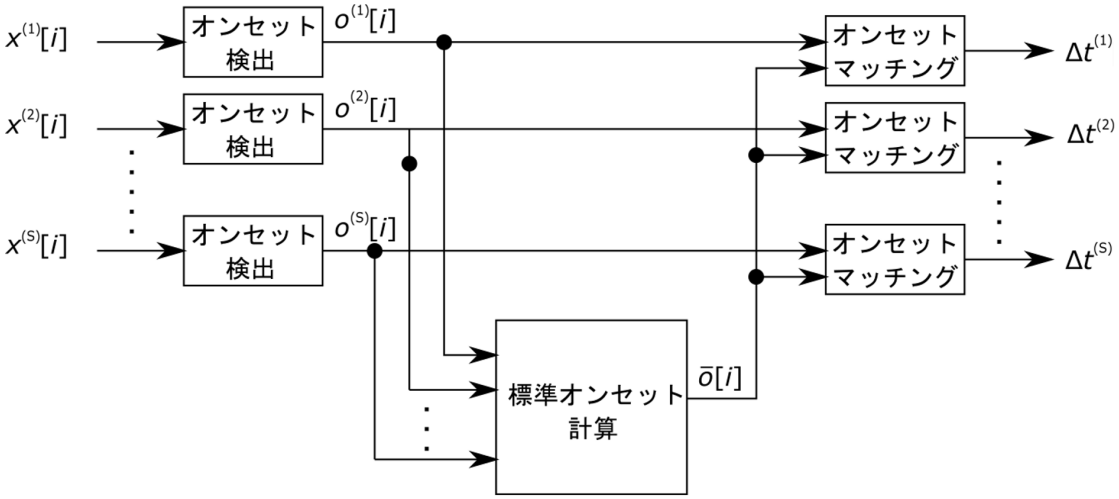

図 3.3: 歌唱者由来のずれを計測するための提案手法

ここで，入力音 $x^{(s)}[i]$（$s = 1, 2, \cdots, S$）は s 番目の歌唱者の歌唱信号であり，式 (3.3) の時間シフトにより時間軸を一元化したものである．$o^{(s)}[i]$ は，各歌唱者について検出されたオンセットが存在する時刻で 1，それ以外の時刻で 0 をとる二値の信号である．

これをオンセット信号と呼ぶ．一方，$\bar{o}[i]$ は標準オンセットが存在する時刻で 1，それ以外の時刻で 0 をとる二値の信号である．これを標準オンセット信号と呼ぶ．最後に，$\Delta t^{(s)}$ は標準オンセットと各歌唱者のオンセットとの時間差であり，歌唱者由来のずれ量に相当する．

3.2.1 オンセット検出手法（手法の検討）

オンセット検出の手法は，その検出精度により本研究の価値を左右すると言っても過言ではないため，特に慎重に検討する必要がある．第 2.2.3 項で述べたように，既存の代表的なオンセット検出関数は，HFC 法のようにパルス的な鋭いオンセットに対して感度が高いものと，スペクトル差法や位相偏差法のように周期的な音のオンセットに対して感度が高いものに二分される．よって，検出対象とする楽音の性質によって使い分ける必要がある．

歌声の音響的性質の特徴は，調波構造を持つ有声音と，白色的なスペクトル構造を持つ無声音に分類されることである．日本語の音韻における有声音と無声音の分類を，表 3.1 に示す．

表 3.1: 日本語の音韻における有声音と無声音の分類 [19]

	有声音	無声音
母音	/a/ /i/ /ɯ/ /e/ /o/	-
子音	/g/ /ŋ/ /z/ /dz/ /ʑ/ /dʑ/ /d/ /n/ /ɲ/ /b/ /m/ /j/	/k/ /s/ /ɕ/ /t/ /tɕ/ /ts/ /h/ /ç/ /ɸ/ /p/ /ɾ/ /ɯ̹/ /ɴ/

母音は全て有声音であり，子音は有声音と無声音が約半数ずつである．ここで，例として歌詞「ち」「う」「も」を歌唱する女声の時間周波数平面と時間波形を，図 3.4，図 3.5，図 3.6 に示す．

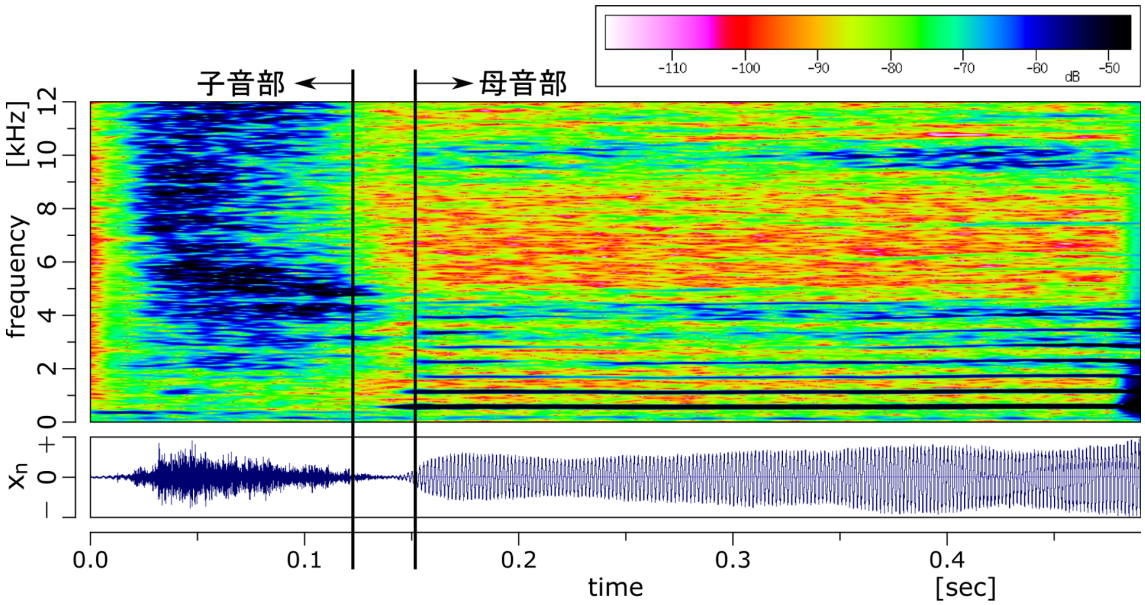

図 3.4: 歌詞「ち」を歌唱する女声の時間周波数平面（上）と時間波形（下）

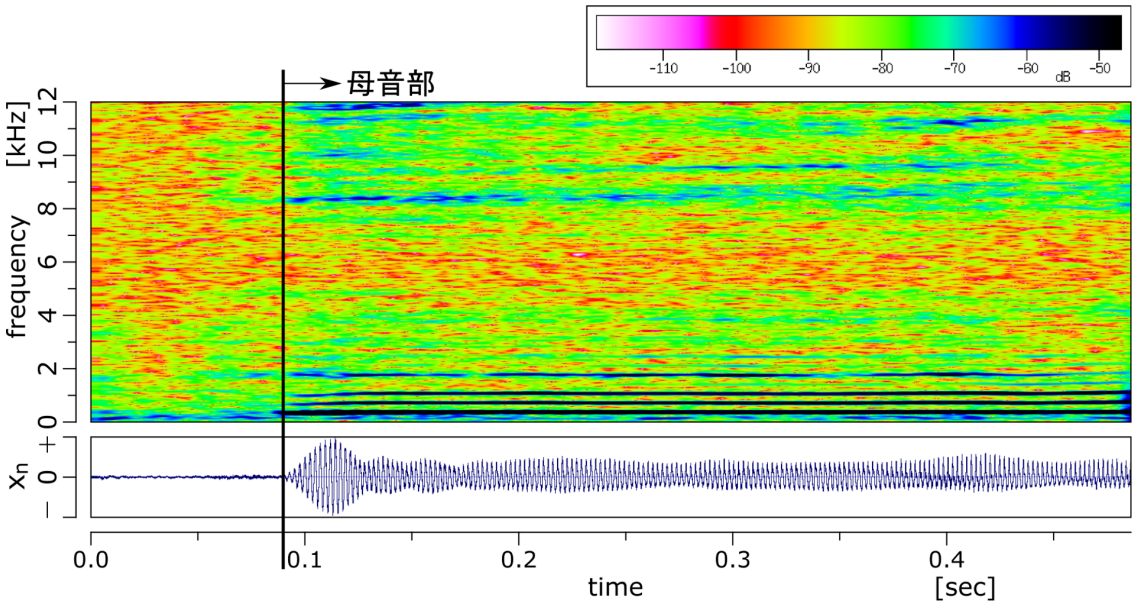

図 3.5: 歌詞「う」を歌唱する女声の時間周波数平面（上）と時間波形（下）

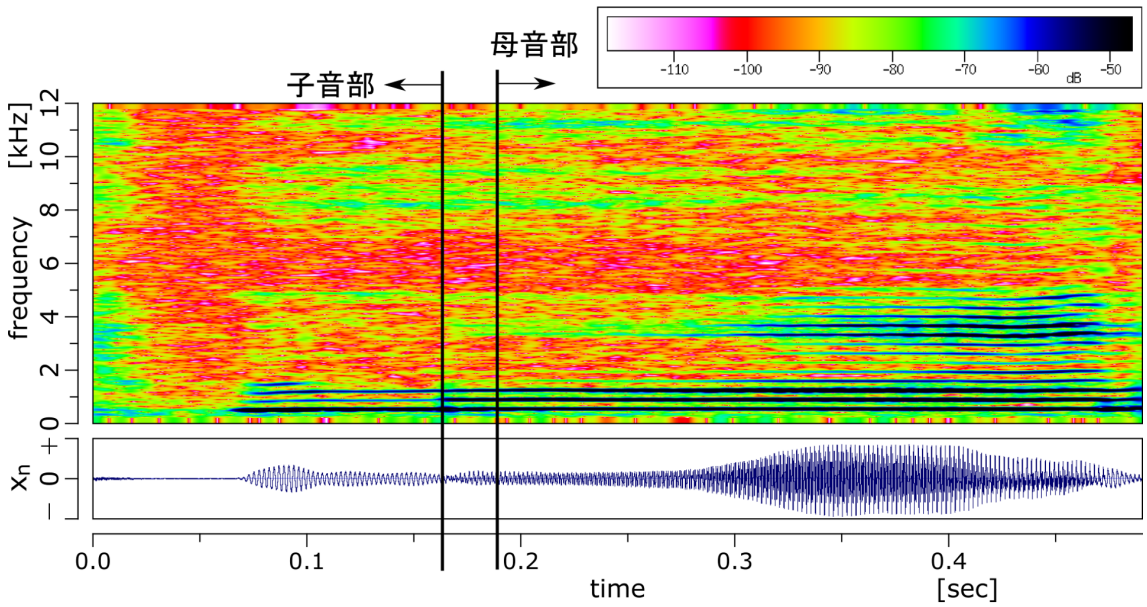

図 3.6: 歌詞「も」を歌唱する女声の時間周波数平面（上）と時間波形（下）

「ち」の子音/tɕ/は無声音に分類され，子音部からフォルマント遷移部を経て母音部へと移行する [20]．時間周波数平面を見れば，子音部と母音部のスペクトル構造の違いが明らかである．「う」/u/は母音のみのため，全区間で調波構造を持つ．「も」の子音/m/は有声音に分類され，子音部から母音部まで連続して調波構造を持つ．ただし，「う」とは異なり，子音部から母音部へ遷移する部分で高調波が増加している．

ここで，歌声のオンセットを検出するための検出関数について議論する．前述のように，歌声は有声音と無声音という性質の異なる要素を持つため，それぞれに対する適性を考慮する必要がある．第 2.2.2 項で述べた既存の代表的な 3 種類の検出関数の適性を，表 3.2 に示す．

表 3.2: 3 種類の検出関数の適性．適性の高さを ◯ > △ > × の順に表す．

	有声音	無声音
HFC 法	×	△
スペクトル差法	△	×
位相偏差法	◯	×

有声音は，調波構造を持ち比較的持続時間が長いことから，ピッチあり非打楽器性音に分類される．よって，有声音に対する適性は，表 2.2 に示したピッチあり非打楽器性音の検出率をもとに評価した．表 2.2 では，真陽性率，偽陽性率ともに位相偏差法が最も良い結果となっている．したがって，同手法は有声音の検出に最も高い適性があると言える．一方で，無声音は白色的なスペクトル構造を持ち比較的持続時間が短いことから，ピッチなし打楽器性音に分類される．よって，無声音に対する適性は，表 2.4 に示したピッチなし打楽器性音の検出率をもとに評価した．ただし，無声音は一般的な打楽器よりも立ち上がりが鈍いため，鋭い立ち上がりに敏感な HFC 法の検出率は，表 2.4 の結果よりも悪化すると考えられる．このことから．無声音の検出には相対的に HFC 法が適しているものの，有声音に対する位相偏差法ほどの適性があるとは言えない．なお，オンセット差法については有声音にのみやや適性があるものの，位相偏差法には及ばない．よって，この時点で候補からは除外する．

続いて，位相偏差法と HFC 法について，リモート合唱に対する適性を考える．リモート合唱では，SN 比の悪い音源や突発的なノイズが混入した音源を処理することもある．その例を図 3.7 に示す．

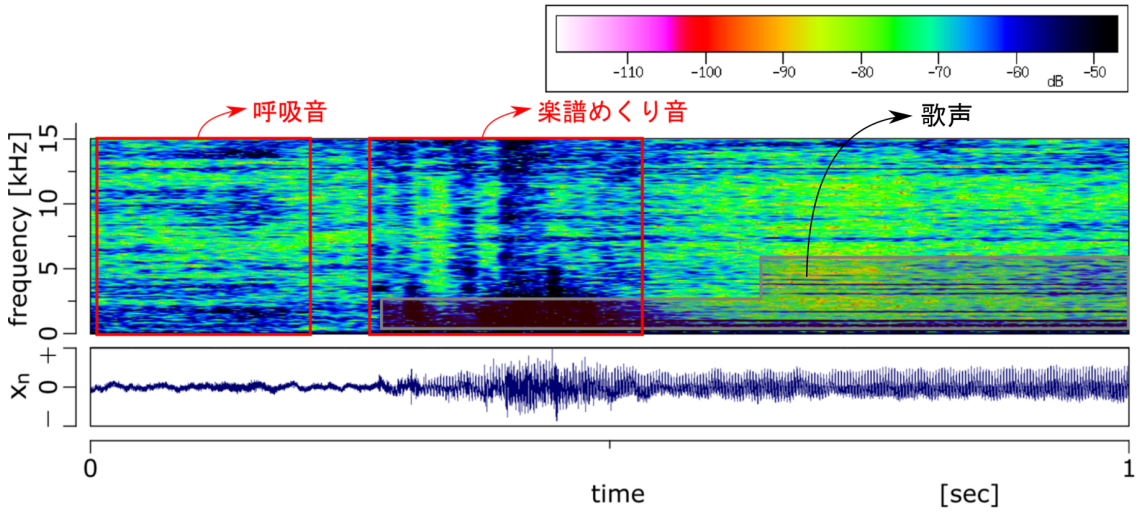

図 3.7: リモート合唱の歌唱録音に混入するノイズの例．時間周波数平面（上）と時間波形（下）を示す．

図 3.7 は，実際にリモート合唱の歌唱録音として投稿された音源の一部である．この例では呼吸音や楽譜をめくる音がノイズとして混入しているのに加え，歌唱に比べて背景雑音が大きい．こうした場合に，フェイルセーフとして偽陰性が増加することはあまり問題とならないが，偽陽性が増えることは問題となる．有声音検出の候補であ

る位相偏差法は，周期性を持つ音を選択的に陽性とするため，白色的なノイズで偽陽性を生じる可能性は低い．周期性を持つノイズに対しては偽陽性を生じるが，そうしたノイズとしては電源ノイズが多いため，事前にローカット処理を行なうことで偽陽性を抑制できる．一方で，無声音検出の候補である HFC 法は，単にエネルギーが増加した場合を陽性とするため，ノイズの種類によらず偽陽性を生じる可能性がある．前述のように HFC 法は検出精度の点でも問題があることから，リモート合唱のオンセット検出に用いることは不適当と判断する．したがって，無声音の検出は行なわず，有声音のオンセットのみ検出を行なうこととする．

以上のことから，リモート合唱のオンセット検出手法として既存手法の中で最も適性が高いのは，位相偏差法であると言える．しかし，位相偏差法についても 2 点の問題が考えられる．

第一に，計算量の問題である．位相偏差法をリモート合唱の自動ミキシングシステムに組み込んだ場合，1 回のミキシングで STFT を歌唱者の人数分行なうことになる．例えば，5 分間の楽曲を 50 人で歌唱するリモート合唱の場合，$F_s = 48000$ Hz の各歌唱録音をシフト量 1 ms で STFT するには，計 1.5×10^7 回の FFT を行なう必要がある．この計算量は，処理速度の点でボトルネックとなる可能性がある．本研究で取り組むリモート合唱のミキシング自動化は，実用化されてこそ意義があるため，処理の高速化も重要な要素である．位相偏差法の検出精度を保ちつつ，STFT を必要としない検出手法を提案するため，位相偏差法の検出関数である式 (2.7) を再度考察する．

$$E_{PD}[i] = \frac{1}{N} \sum_{k=1}^{N} |\Delta\theta[i,k]| \tag{2.7}$$

式 (2.7) は，正弦波が含まれる周波数ビンのほうが，正弦波が含まれない周波数ビンよりも位相偏差 $\Delta\theta[i,k]$ の値が小さくなるという性質を利用した検出関数である，つまり，位相偏差は入力波形と正弦波形の類似度を表す指標と言える．これと同等の性質を持つ指標を検出関数に導入できれば，位相偏差法と同様に有声音のオンセット検出に適した手法を構築することが可能であると考えられる．よって，STFT のボトルネック問題を解決するためには，位相偏差と同等の性質を持ち，なおかつ STFT を用いずに計算できる指標を導入した検出関数を提案することが必要である．

第二に，歌声のオンセットの定義に関する問題である．式 (2.7) では，全ての周波数ビンの位相偏差 $\Delta\theta[i,k]$ を平均化している．このため，検出関数を閾値処理して計算さ

れるオンセット時刻には，複数の調波成分の生起時刻が反映される．全ての調波成分が同時に生起する場合は問題ないが，歌声のオンセットを対象とした研究 [21] によれば，各高調波の生起時刻には最大で 50 ms のずれがあることが指摘されている．この場合，どの調波成分の生起をオンセットとするかを明確に定義し，その成分に絞ってオンセットを検出することが望ましい．本研究では，有声音の種類に限らず比較的高い強度で存在する，基本波の生起をオンセットと定義する．よって，基本波に絞ってオンセットを検出する手法が必要となる．

以上の 2 点の問題を解決する可能性があるのが，F0 推定の手法 [17] を応用することである．この手法では，まず前処理として基本波が含まれる帯域を抽出する．次に，各帯域で入力波形と正弦波形の類似度を表す指標である「基本波らしさ（fundamentalness）」の計算を行なう．この「基本波らしさ」は，位相偏差法における位相偏差 $\Delta\theta[i,k]$ と類似した性質を持つ指標である．よって，この指標をオンセット検出関数に導入することで，位相偏差法と同質のオンセット検出を行える可能性がある．また,「基本波らしさ」は STFT を用いず時間領域のみで計算できるため，第一に挙げた計算量の問題を解決できる可能性がある．さらに，基本波に絞った解析を行なうので，第二のオンセットの定義に関する問題も解決できる．

以上のことから，本研究では「基本波らしさ」を指標として導入したオンセット検出手法を提案する．この提案手法をFN（fundamentalness）法と呼ぶ．

3.2.2 オンセット検出手法（手法の提案）

FN 法の概要を示した処理ブロック図を，図 3.8 に示す．

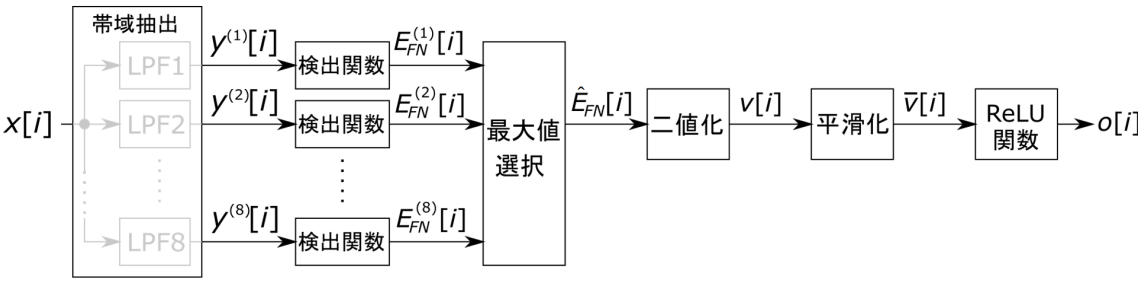

図 3.8: FN 法の処理ブロック図

各処理ブロックについて，個別に詳細を述べる．なお，入力として用いる歌唱信号 $x[i]$ は，サンプリングレート $F_s = 48000$ Hz で標本化されたディジタル信号と仮定する．

- 処理ブロック 1：帯域抽出

この処理ブロックは,「基本波らしさ」を計算するための前処理に相当する．その目的は，ある歌唱者の歌唱信号 $x[i]$ から複数の帯域を抽出し，いずれかの帯域に有声音の基本波成分のみを含ませることである．時間領域で特定の帯域のみを抽出するには，バンドパスフィルタ（BPF）により帯域を細かく分割する手法が考えられる．しかし，本研究では単音の調波構造から基本波成分のみを分離できれば良い．このような場合には，カットオフ周波数の異なる複数のローパスフィルタ（LPF）を使用する手法 [17] も有効である．フィルタ特性の急峻さが同じ場合，BPF よりも LPF のほうがフィルタ次数が小さいため計算量は少なくて済む．このことから，LPF による手法を選択する．FIR フィルタと IIR フィルタの選択についても，同様に計算量の少ない IIR フィルタを選択する．IIR フィルタの場合には群遅延が存在するため，周波数によってフィルタの遅延量が異なるという問題がある．オンセット検出における群遅延の悪影響として，周波数が異なる複数の音のオンセットを検出する場合，周波数によって検出時刻の時間軸が異なってしまうという問題がある．しかし，歌唱者由来のずれ量を計測する際は，同一パートのほぼ同じ周波数の音同士でオンセット時刻を比較することになる，よって，本研究においては群遅延の影響は小さいと考えられる．

続いて，LPF のカットオフ周波数を選択する．合唱における基本波の周波数は，およそ 80 Hz∼900 Hz[22][23] であると言われている．よって，この周波数範囲で帯域の抽出を行えば良い．カットオフ周波数の異なる複数の LPF を用いて単音の基本波を抽出する場合，LPF のカットオフ周波数が 1 オクターブあたりに 2 個存在すれば十分精度が良いことが知られている [17]．これを満たすためには，下限周波数の $(\sqrt{2})^n$（n は自然数）倍をカットオフ周波数とすれば良いが，ここでは便宜的に次のようにカットオフ周波数を設定する．まず，周波数の下限である 80 Hz の 1 オクターブ上の 160 Hz にカットオフを設け，これらを二等分する 120 Hz にもカットオフを設ける．次は，160 Hz の 1 オクターブ上の 320 Hz にカットオフを設け，これらを二等分する 240 Hz にもカットオフを設ける．このようにして，上限の 900 Hz を超えるまでカットオフ周波数を設定する．ここで，カットオフ周波数の低い帯域からに順に帯域番号 $k = 1, 2, \cdots$ を定義すると，カットオフ周波数 $F_c^{(k)}$ との対応は表 3.3 のようになる．

表 3.3: 帯域番号 k とカットオフ周波数 $F_c^{(k)}$ の対応

帯域番号 k	カットオフ周波数 $F_c^{(k)}$ [Hz]
1	120
2	160
3	240
4	320
5	480
6	640
7	960
8	1280

歌唱信号 $x[i]$ を帯域番号 k の LPF に入力し，その出力を $y^{(k)}[i]$ と表す．すると，有声音が存在する時刻にて，いずれか 1 つの $y^{(k)}[i]$ には $F_c^{(k-1)}$ [Hz]（$k = 1$ の場合は 80 Hz）$\sim F_c^{(k)}$ [Hz] の基本波のみが含まれる．これにより，基本波の抽出が可能である．

ここで，IIR フィルタの設計について述べる．IIR フィルタは 6 次のチェビシェフフィルタで，フィルタ設計ツール [24] を使用した計算機設計である．通過域リップルは 0.5 dB とする．例として，カットオフ周波数 1280 Hz の場合の振幅特性と位相特性を，図 3.9 に示す．

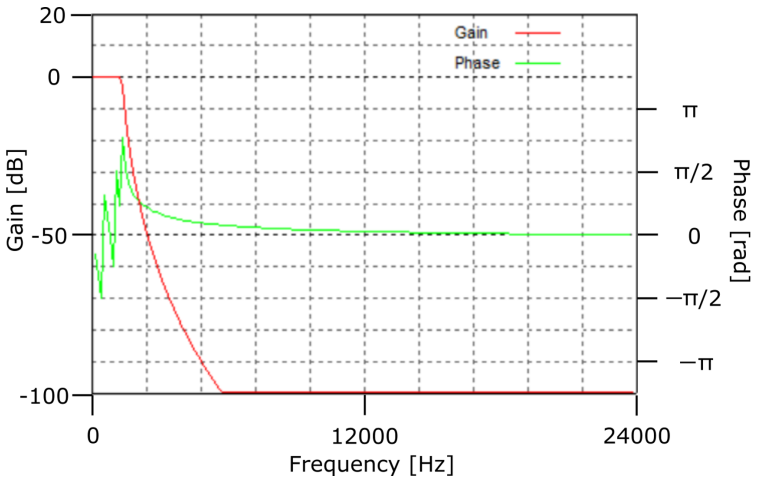

図 3.9: $F_c^{(8)} = 1280$ Hz の LPF の振幅特性と位相特性

次に，計算量について述べる．位相偏差法では，前処理に用いる STFT が計算量の点でボトルネックとなる問題があったが，FN 法においても前処理が最も計算量を要する部分である．IIR フィルタと FFT の計算量を比較すると，IIR フィルタではフィルタ次数 n に対して $O(n^2)$，FFT では FFT 点数 N に対して $O(n \log n)$ と表される．ここで，FN 法における IIR フィルタの次数は $n = 6$ であり，位相偏差法の FFT 点数は一般的に使用される $N = 2048$ を仮定する．その場合，$n^2 \ll \log n$ となるため，FN 法は位相偏差法よりも十分高速な手法であると言える．

- 処理ブロック 2：検出関数

FN 法における検出関数の目的は，各時刻において基本波が存在するかどうかを評価することである．まず，検出関数を計算するために，$y^{(k)}[i]$ を長さ t_a [ms] の分析フレームで分割する．このフレーム長は，オンセット検出の時間精度を直接決定づけるものであるから，重要なパラメータである．ここでは $t_a = 1$ ms とする．

次に，各フレームの始端サンプル $i = nt_aF_s$（$n = 0, 1, \cdots$）にて，検出関数を計算する．FN 法では，検出関数として「基本波らしさ」[17] の指標を用いる．この指標は信号波形と正弦波の類似性を示し，式 (3.4)[17] のように計算される．

$$E_{FN}^{(k)}[i] = \exp(-\frac{f_{std}^{(k)}[i]}{f_{ave}^{(k)}[i]}) \tag{3.4}$$

ここで，$f_{std}^{(k)}[i]$ は，図 3.10 のように定義する 4ヶ所の周期の標準偏差の逆数，$f_{ave}^{(k)}[i]$ は同じく平均である．

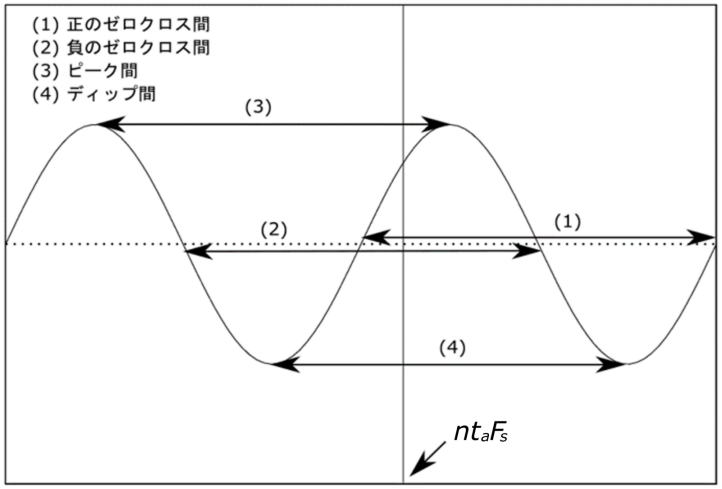

図 3.10: 波形の 4 カ所の周期．[17] の Fig. 3 を参考に作成．

$E_{FN}^{(k)}[i]$ は 0 から 1 の値をとり，信号波形と正弦波の類似度が大きいほど値は 1 に近づく．4 箇所の周期のうち 1ヶ所でも検出できなかった場合には，$E_{FN}^{(k)}[i] = 0$ とする．また，検出された周期から計算された周波数が，その帯域のカットオフ周波数の 1.5 倍以上だった場合にも $E_{FN}^{(k)}[i] = 0$ とする．なお，$E_{FN}^{(k)}[i]$ は，分析フレーム内で一定の値を出力する．すなわち，$nt_aF_s \leq i < (n+1)t_aF_s$ において $E_{FN}^{(k)}[i] = E_{FN}^{(k)}[nt_aF_s]$ となる．よって，$E_{FN}^{(k)}[i]$ は 1 ms ごとに値が変化するステップ状の関数となる．

図 3.10 に示した 4ヶ所の周期は，サンプル番号 nt_aF_s を中心に $\pm\frac{t_T}{2}$ [ms] の範囲で計測される．この計測範囲は，オンセットを検出する音の周波数の下限を決定づける．合唱における基本波の最低周波数は 80 Hz であるから，$t_T = \frac{1000}{80} = 12.5$ ms 以上の範囲とすれば良い．ここでは $t_T = 20$ ms と設定する．

- 処理ブロック 3：最大値選択

$k = 1, \cdots, 8$ の 8 個の帯域のうち，どの帯域に基本波成分が含まれているのかはまだ判明していない．それを推定するのが，この処理ブロックの目的である．

有声音が存在する時刻において，基本波が含まれる帯域のみ検出関数 $E_{FN}^{(k)}[i]$ は 1 に近い値をとる．それと比較して，基本波が含まれない帯域，あるいは基本波に加え高調波を含む帯域では，検出関数の値が小さくなる．また，有声音が存在しない時刻では，いずれの帯域も検出関数は 0 に近い値をとる．このことから，各時刻で検出関数の最大値をトレースしていけば，基本波の生起を検出することが可能となる．つまり，式 (3.5) に示す処理を行なう．

$$\hat{E}_{FN}[i] = \max\{E_{FN}^{(1)}[i], E_{FN}^{(2)}[i], \cdots, E_{FN}^{(8)}[i]\} \tag{3.5}$$

ここで，$\hat{E}_{FN}[i]$ は，基本波を含むと推定される帯域の検出関数である．

リモート合唱の歌唱信号での $\hat{E}_{FN}[i]$ 計算結果を，図 3.11 に示す．例として提示するのは，合唱曲「ほたるこい」のソプラノパート 1 人の歌唱録音で,「やまみちこい」と歌唱する部分である．

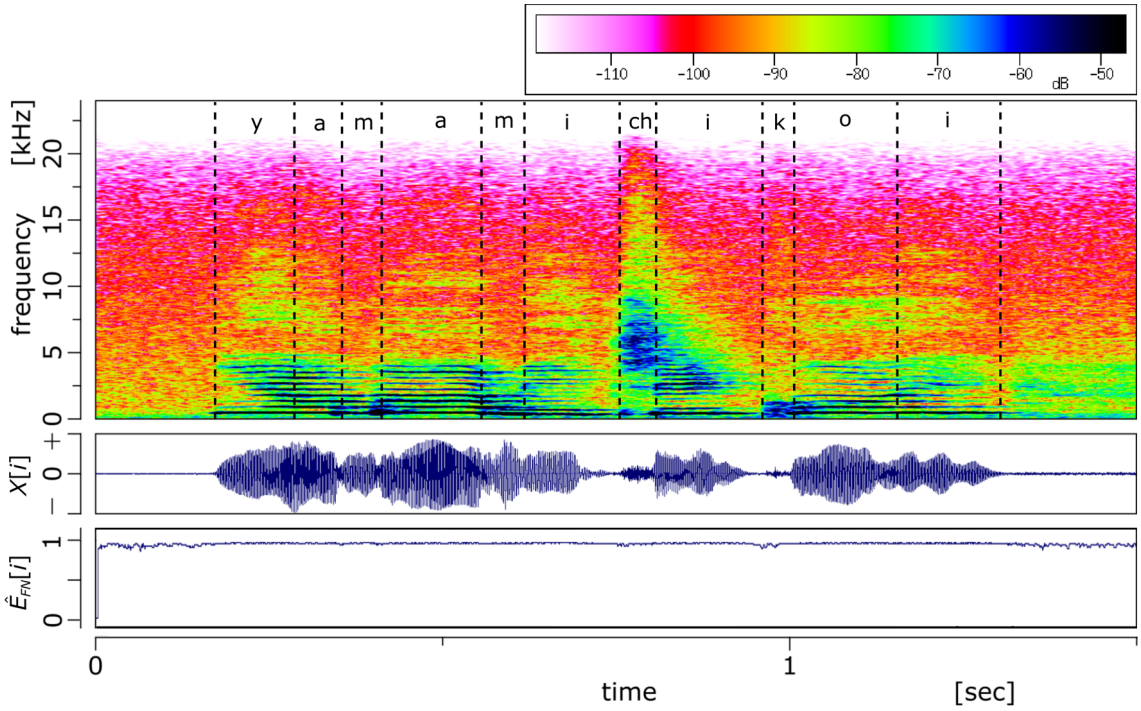

図 3.11: リモート合唱の歌唱信号での $\hat{E}_{FN}[i]$ 計算結果．時間周波数平面（上），時間波形（中），$\hat{E}_{FN}[i]$ 計算結果（下）を示す．

周波数平面に併記した破線は，著者が聴取して推定した音韻の開始時刻である．音源の時刻 0 sec 付近は過去方向に周期の計測ができないため，$\hat{E}_{FN}[i]$ の値は 0 を示す．その後は約 0.8〜1.0 の間で小刻みに変動しているが，有声音が存在する部分では，存在しない部分よりもわずかに大きな値を示していることが分かる．

- 処理ブロック 4：二値化

この処理ブロックの目的は，検出関数の値に閾値を設け，基本波が存在するかどうかの判定を行うことである．はじめに，その閾値の設定について議論する．式 (3.4) において，指数部の $\frac{f_{std}^{(k)}[nt_aF_s]}{f_{ave}^{(k)}[nt_aF_s]}$ は，4ヶ所で観測された瞬時周波数の標準偏差をその平均周波数で規格化したものである．これは，周波数によらず信号波形と正弦波の相違度を表す変動係数に相当する．ここで，この変動係数を $C^{(k)}[i] = \frac{f_{std}^{(k)}[nt_aF_s]}{f_{ave}^{(k)}[nt_aF_s]}$ と改めて定義する．$C^{(k)}[i]$ は $E_{FN}^{(k)}[i]$ とは正反対の意味を持つ指標であることと，値の範囲には上限が存在しないことに注意されたい．$C^{(k)}[i]$ と $E_{FN}^{(k)}[i]$ の値の対応を，図 3.12 に示す．

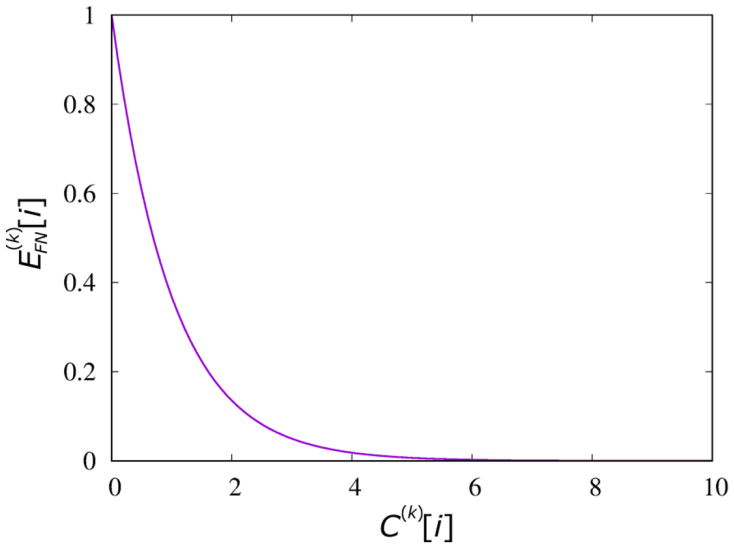

図 3.12: $C^{(k)}[i]$ と $E_{FN}^{(k)}[i]$ の値の対応

信号波形が定常な純音であれば，その平均周波数によらず $C^{(k)}[i]$ の値は 0 になる．一方，狭帯域信号 $y^{(k)}[i]$ のうち基本波のみを含む帯域であっても，歌唱ピッチの変動やノイズの混入により，$C^{(k)}[i]$ が 0 になることはほぼない．そこで，本研究では，$C^{(k)}[i]$ の許容誤差を 2%と設定する．$C^{(k)}[i] = 0.02$ に対応する検出関数 $E_{FN}^{(k)}[i]$ の値は $e^{-0.02} = 0.98$ であるため，これを閾値 E_{th} とする．

$\hat{E}_{FN}[i]$ を閾値処理した結果を $v[i]$ とする．$v[i]$ は，$\hat{E}_{FN}[i]$ が E_{th} より大きい場合に 1，小さい場合に 0 をとる二値の信号である．数式では，式 (3.6) のように定義する．

$$v[i] = \begin{cases} 0 & (\hat{E}_{FN}[i] < E_{th}) \\ 1 & (\hat{E}_{FN}[i] \geq E_{th}) \end{cases} \tag{3.6}$$

$v[i]$ は，有声音が存在する時刻に 1，無声音または無音である時刻で 0 をとる．よって，$v[i]$ は有声区間を検出する信号であると言える．

図 3.11 に示した音源を用いて $v[i]$ を計算した結果を，図 3.13 に示す．

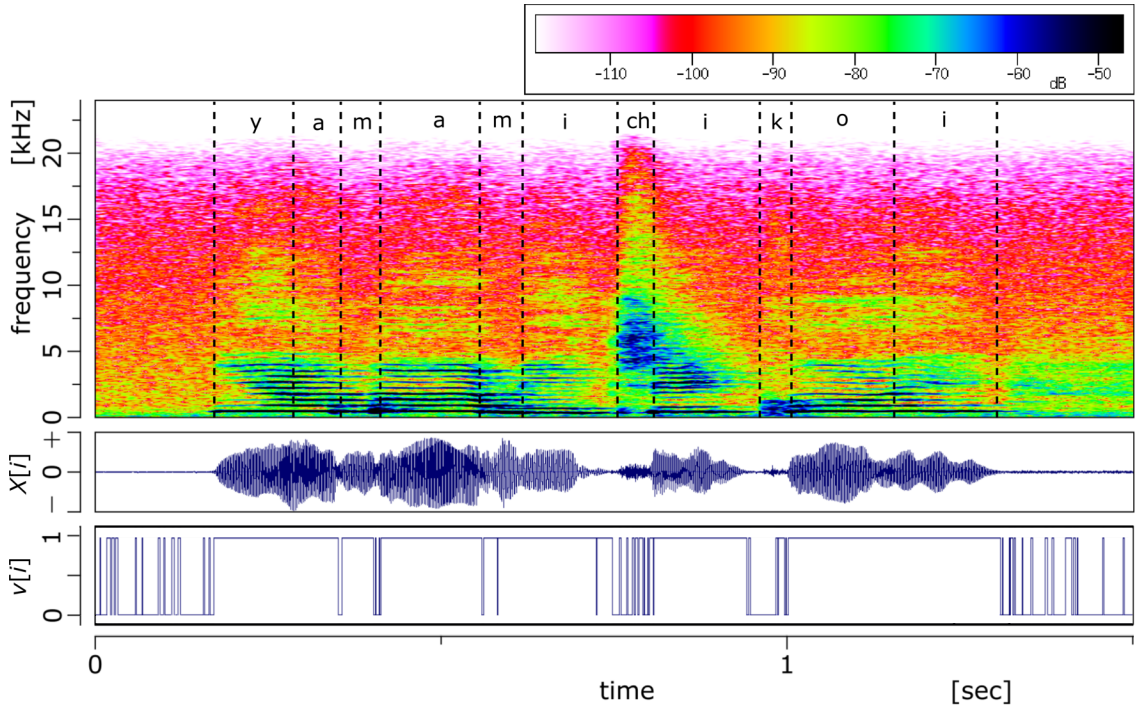

図 3.13: リモート合唱の歌唱信号での $v[i]$ 計算結果．時間周波数平面（上），時間波形（中），$v[i]$ 計算結果（下）を示す．

有声音が存在する部分では安定して $v[i] = 1$ を維持しているが，歌声が存在しない部分や無声音の部分では頻繁に値が 0 と 1 を行き来している．これは，位相偏差法でも指摘されていたように，検出関数がエネルギーの大小にかかわらず反応することを示している．

- 処理ブロック 5：平滑化

この処理ブロックの目的は，図 3.13 に見られたような $v[i]$ の小刻みな変動を抑制し，オンセット検出における偽陽性を減らすことである．具体的には，次に示す 2 点の処理を行なう．

第一に，$v[i]$ に対して長さ $2Nt_aF_s + 1(N$ は自然数) サンプルのメジアンフィルタをかける．この処理は，突発的なノイズ等により途切れてしまった有声区間を連結させる目的がある．$v[i]$ は t_a [sec] 間の分析フレーム内では一定値を取るため，各フレーム内の 1 点のみを取り出してメジアンを求めれば良い．よって，フィルタをかけた信号

を $\bar{v}'[i]$ とすると，この処理は式 (3.7) のように表せる．

$$\bar{v}'[i] = \begin{cases} 1 & (\frac{\sum_{n=-N}^{N} v[i+nt_aFs]}{2N+1} > \frac{1}{2}) \\ 0 & (\frac{\sum_{n=-N}^{N} v[i+nt_aFs]}{2N+1} < \frac{1}{2}) \end{cases} \tag{3.7}$$

フィルタ長は，有声音や無声音の持続時間よりも十分短ければ良い．本研究では，$N = 2$，すなわちフィルタ長を約 4 ms とした．

第二に，$\bar{v}'[i]$ の各有声区間の開始時刻から t_v [sec] 前までを，$\bar{v}[i] = 0$ とする処理を行なう．この処理は，合唱曲においてオンセットが t_v [sec] 以内に近接することはないという仮定のもとで，信頼性の低い有声区間を削除する目的がある．信頼性の低い有声区間とは，図 3.13 に見られるように，主に無声子音やフォルマント遷移の部分に多く発生する短い有声区間である．よって，閾値となる t_v は，子音部とフォルマント遷移部を合わせた時間長よりも長く，合唱における一般的な音価よりも短ければ良い．ここでは $t_v = 150$ ms と設定した．

図 3.11 に示した音源を用いて $\bar{v}[i]$ を計算した結果を，図 3.14 に示す．

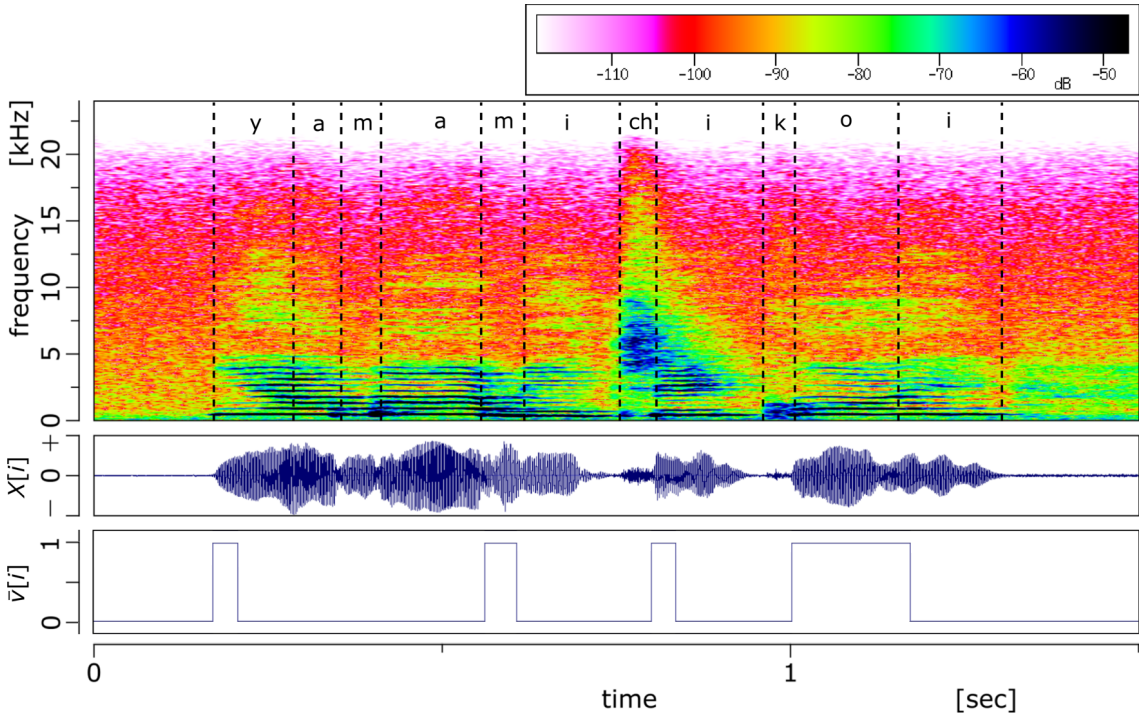

図 3.14: リモート合唱の歌唱信号での $\bar{v}[i]$ 計算結果．時間周波数平面（上），時間波形（中），$\bar{v}[i]$ 計算結果（下）を示す．

歌詞の冒頭「やま」と末尾の「こい」は同じ音高でなおかつ有声音が連続するので，

基本波成分はほぼ連続して存在する．このような場合には，平滑化処理により最初のオンセットのみが残される．図 3.13 に示した $v[i]$ は大まかに有声区間を検出する信号であったが，$\bar{v}[i]$ では有声区間の開始部分をのみを検出する信号となっていることが分かる，

- 処理ブロック 6：ReLU 関数

最後の処理の目的は，有声区間の開始部分に反応する信号 $\bar{v}[i]$ を，オンセット信号 $o[i]$ に変換することである．そのためには，$\bar{v}[i]$ の立ち上がり部分を取り出せば良い．ここでは式 (3.8) に示す ReLU 関数を用いる．

$$o[i] = \begin{cases} \bar{v}[i] - \bar{v}[i-1] & (\bar{v}[i] - \bar{v}[i-1] > 0) \\ 0 & (\text{otherwise}) \end{cases} \tag{3.8}$$

図 3.11 に示した音源を用いて $o[i]$ を計算した結果を，図 3.15 に示す．

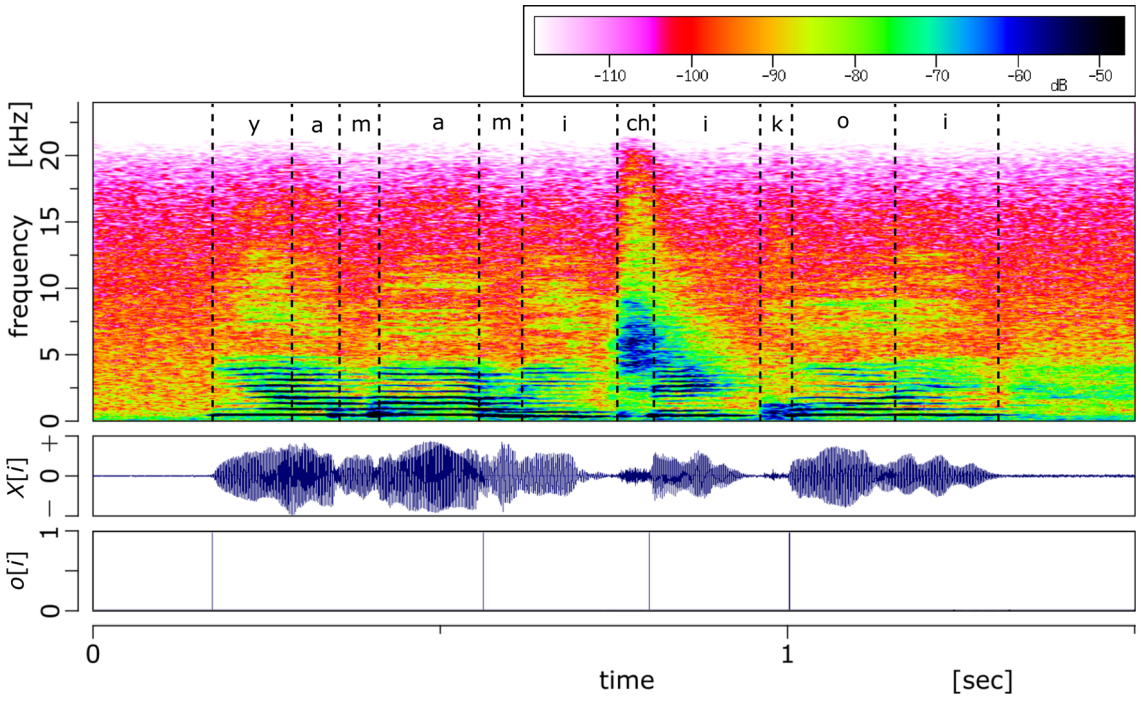

図 3.15: リモート合唱の歌唱信号での $o[i]$ 計算結果．時間周波数平面（上），時間波形（中），$o[i]$ 計算結果（下）を示す．

最終的に検出されたオンセットは，波形の聴取と時間周波数平面の観察により手動で検出したものとほぼ一致していることが分かる．

3.2.3 標準オンセット計算

ここでは，各歌唱者のオンセットずれ量を求める際にリファレンスとして必要な，標準オンセット時刻を設定する．従来であれば楽譜上の音符が始まる拍点をリファレンスとする [13][14] が，本研究では楽譜情報を用いずに処理を行なうため，各歌唱者のオンセット時刻をもとに設定する必要がある．楽譜上のある一つの音符が始まる時刻は，大数の法則により，その音を歌唱する複数の歌唱者のオンセット時刻の平均値から求められると考えられる．しかし，リモート合唱の場合には他の歌唱者の声が聞けないため,「いつも遅れ気味に発声する」というような癖に気づかないまま歌唱する場合や，リズムを間違えて歌唱することも考えられる．そのような場合，オンセット時刻の分布が偏ったり，外れ値が存在する可能性がある．よって，平均値よりも中央値の方が標準オンセット時刻として相応しいと考えられる．そこで本研究では，各歌唱者のオンセット信号 $o^{(s)}[i]$ から，標準オンセット信号 $\bar{o}[i]$ として各オンセット時刻の中央値を求める手法を提案する．

まず，時間軸上の点で表される各歌唱者のオンセットを拡張し，一定の時間幅 t_{on} を持つ線として表す．線状に拡張されたオンセット信号を便宜的に $o'^{(s)}[i]$ と表記すると，この操作は式 (3.9) のように定義できる．

$$o'^{(s)}[i] = \begin{cases} 1 & (i_{on}^{(s)} \leq i < i_{on}^{(s)} + t_{on}F_s) \\ 0 & (\text{otherwise}) \end{cases} \tag{3.9}$$

ここで，$i_{on}^{(s)}$ は，$o^{(s)}[i] = 1$ となるときのサンプル番号である．この操作を説明した概略図を，図 3.16 に示す．

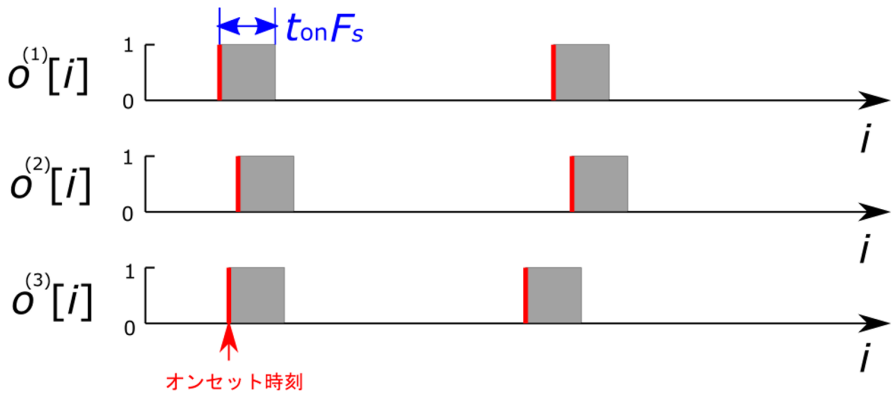

図 3.16: オンセット拡張の概略図

ここで，一つの音符に対するオンセット時刻の中央値から，$\pm t_{on}$ [sec] の間に全員のオンセットが存在すると仮定する．このとき，図 3.17 に示すように過半数の歌唱者が $o'^{(s)}[i] = 1$ となる時刻がオンセット時刻の中央値，すなわち標準オンセット時刻であると言える．

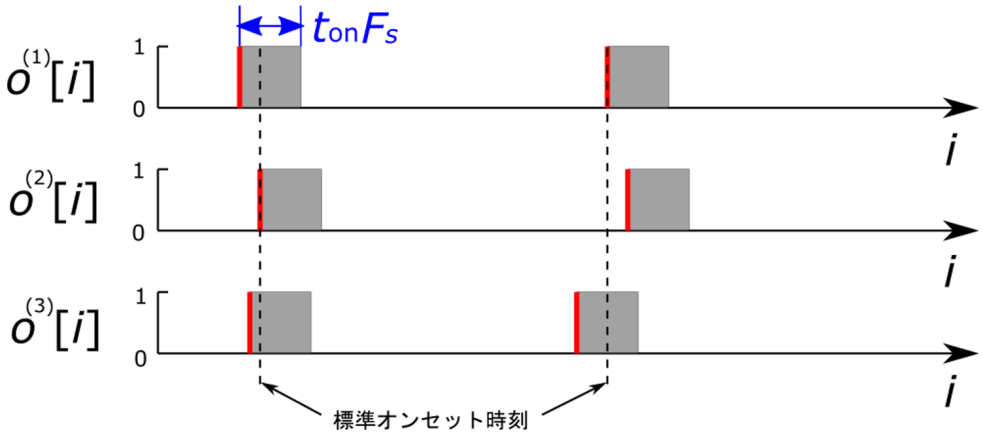

図 3.17: 標準オンセット時刻の設定の概略図

よって，標準オンセット時刻で 1，それ以外の時刻で 0 をとる標準オンセット信号 $\bar{o}[i]$ は，式 (3.10) のように定義できる．

$$\bar{o}[i] = \begin{cases} 1 & (\sum_{s=1}^{S} o'^{(s)}[i] \geq \frac{S}{2}) \\ 0 & (\text{otherwise}) \end{cases} \tag{3.10}$$

なお，時間幅 t_{on} は，オンセットずれ量の最大値を仮定するパラメータであると言える．時間幅が短すぎると式 (3.10) の陽性となる条件を満たせず，標準オンセットの個数が減ってしまう．反対に長過ぎると，次の音符のオンセットを現在の音符の標準オンセットとしてしまう可能性がある．特に後者は，オンセットずれ量の計測結果の信頼性を損ねてしまうため，t_{on} は 1 音の音価よりも短い時間幅となるよう注意しなければならない．本研究においては，先行研究 [13] に則り $t_{on} = 150$ ms と設定する．

3.2.4 オンセットマッチング

オンセットマッチングでは，各歌唱者のオンセット信号 $o^{(s)}[i]$ と標準オンセット信号 $\bar{o}[i]$ をもとに，検出されたオンセットがどの標準オンセットに対応するかを決定し，そ

の時間差を計測する．ある歌唱者から検出されたオンセットと標準オンセットが完全に 1 対 1 で対応しているならば，オンセット信号の先頭から順番に対応付ければ良い．しかし，実際には偽陽性や偽陰性の存在により．検出されたオンセットと標準オンセットが同数であることは少ない．そこで，先行研究 [13] では．リファレンスから ±150 ms 以内あるオンセットのみが対応付けられ，範囲内に存在しない場合，その音符は演奏されなかったものとして扱われた．本研究でも，標準オンセットを決定するにあたっては，オンセットずれ量の最大値を ±150 ms と仮定した．しかし，歌唱者由来のずれを補正することを考慮すると，これよりも大きいずれこそ補正を行なうべきである．単純には，リファレンスからの対応付けの範囲を拡大すれば良いが，それに伴い誤った対応付けが増加してしまう．そこで，単に時間関係で対応付けるのではなく，歌声の音響特徴量を用いた弾性マッチングを行なうことにより，この問題の解決を図る．

まず，提案手法の概略を述べる．オンセットマッチングを行なうことは，ともに 0 と 1 の二値をとる時系列データ $o^{(s)}[i]$ と $\bar{o}[i]$ の弾性マッチングを行なう問題に置き換えられる．しかし，合唱曲のオンセットは 1 秒間に高々5 個程度である．このため，$F_s = 48000$ Hz で標本化されたオンセット信号では，1 秒間に 48000 個近い「0」の中に最大でもわずか 5 個の「1」が点在するだけである．よって，$o^{(s)}[i]$ と $\bar{o}[i]$ を直接弾性マッチングさせる場合，両者の局所距離はほぼ常に 0 となり，マッチング結果は単に「1」同士を時系列順に対応させただけのものに過ぎない．そこで，各歌唱信号から計測される音響特徴量の時系列データと，リファレンスとなる音響特徴量の時系列データを弾性マッチングさせることで，歌唱部分の大局的な同定を行なう．その結果を用いて，各歌唱者のオンセットと標準オンセットの局所的なマッチングを行なう．弾性マッチングの手法としては，動的計画法を用いた DP マッチングが広く知られており [25]，本研究でもこれを使用する．

使用する音響特徴量に求められる条件は，歌唱部分の同定を行なうために，音韻の種類により異なる値を示すことである．そのような特徴量として，スペクトルフラットネス [26] を導入する．スペクトルフラットネスの計算結果を時系列に並べた信号を，フラットネス信号と呼ぶこととする．フラットネス信号 $F_{la}[i]$ は，式 (3.11) により計算される．

$$F_{la}[i] = \frac{\sqrt[K]{\prod_{k=1}^{K} P^{(k)}[i]}}{\frac{1}{K}\sum_{k=1}^{K} P^{(k)}[i]} \tag{3.11}$$

ここで，$P^{(k)}[i]$ は帯域番号 k における信号のパワーである．STFT を使用する場合は，それぞれパワースペクトルと周波数ビンに対応する．しかし，本研究では STFT を用いない処理を行いたいため，オンセット検出での帯域抽出と同じく IIR フィルタを用いて帯域分割を行なう．信号から取り出す帯域は，歌声成分が強く存在する帯域に限定する．そこで，表 3.4 に示すようなカットオフ周波数 $F_{low}^{(k)}$，$F_{high}^{(k)}$ を持った BPF により，帯域分割を行なう．

表 3.4: 帯域番号 k とカットオフ周波数 $F_{low}^{(k)}$，$F_{high}^{(k)}$ の対応

帯域番号 k	低域側カットオフ $F_{low}^{(k)}$ [Hz]	高域側カットオフ $F_{high}^{(k)}$ [Hz]
1	100	1100
2	1100	2100
3	2100	3100
4	3100	4100
5	4100	5100
6	5100	6100
7	6100	7100
8	7100	8100

$F_{low}^{(1)} = 100$ Hz としているのは，低周波のノイズを除くためである．分割された各帯域の信号を $y^{(k)}$ とすると，各帯域でのパワー $P^{(k)}[i]$ は式 (3.12) で計算される．

$$P^{(k)}[i] = \frac{1}{t_p} \sum_{n=-\frac{t_p}{2}F_s}^{\frac{t_p}{2}F_s} (y^{(k)}[i+n])^2 \tag{3.12}$$

ここで，t_p はパワー算出に用いる分析フレーム長であり，$t_p = 1$ ms と設定した．

式 (3.11) により計算されるスペクトルフラットネスは，信号のスペクトルの平坦さを表す指標であり，0~1 の値をとる．白色雑音や無声音ではスペクトルが一様な分布となるため，$F_{la}[i]$ の値は 1 に近づく．一方，有声音では調波構造によりスペクトルの凹凸が激しくなるため，$F_{la}[i]$ の値は 0 に近づく．図 3.11 に示したリモート合唱の歌唱信号での $F_{la}[i]$ 計算結果を，図 3.18 に示す．

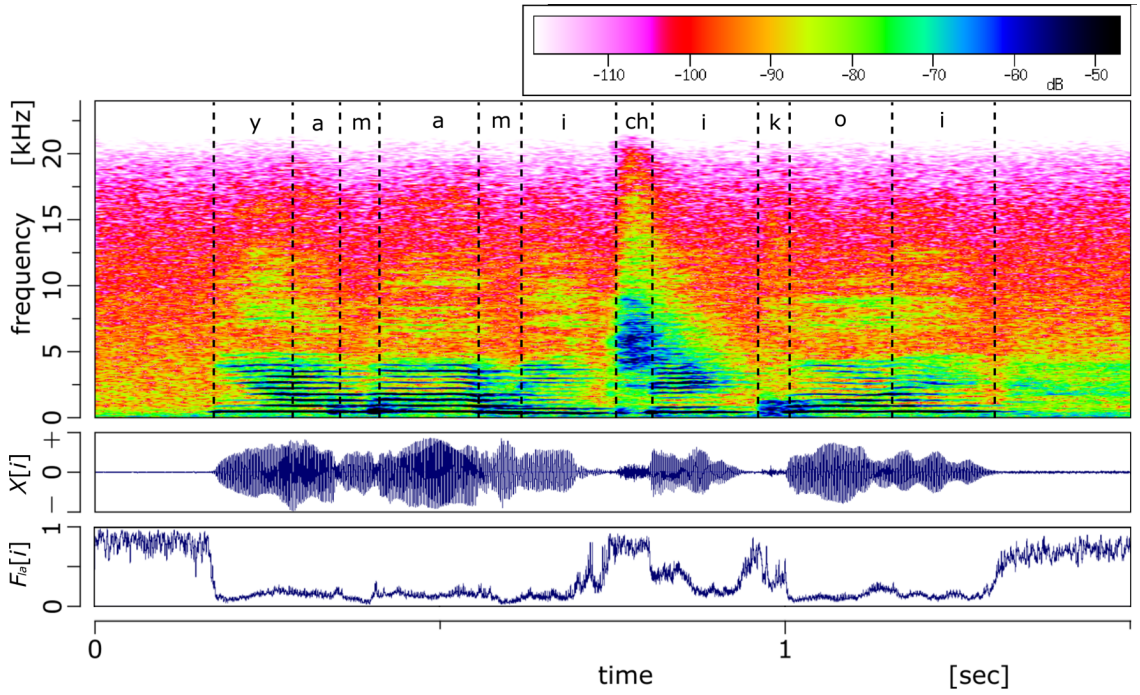

図 3.18: リモート合唱の歌唱信号での $F_{la}[i]$ 計算結果．時間周波数平面（上），時間波形（中），$F_{la}[i]$ 計算結果（下）を示す．

「k」,「ch」のような無声の子音や無音の部分では値が山型に変化し，母音や有声の子音では谷型に変化することが分かる．よって，リファレンスとなるフラットネス信号と比較すれば，大まかな歌唱位置を同定することができると考えられる．なお，$F_{la}[i]$ は各歌唱者の歌唱信号で計算されることから，以後 s 番目の歌唱者のフラットネス信号を $F_{la}^{(s)}[i]$ と表す．

次に，$F_{la}^{(s)}[i]$ の各時刻における歌唱部分を同定するために，リファレンスとなるフラットネス信号を作成する．ここでは，各歌唱者のフラットネス信号の，各時刻での平均値をリファレンスとする．これを標準フラットネス信号と呼び，$\bar{F}_{la}[i]$ と表す．標準フラットネス信号の導出式を，式 (3.13) に示す．

$$\bar{F}_{la}[i] = \frac{1}{S}\sum_{s=1}^{S} F_{la}^{(s)}[i] \tag{3.13}$$

$F_{la}^{(s)}[i]$ と $\bar{F}_{la}[i]$ を弾性マッチングすることにより，各歌唱信号の歌唱部分を同定する．このとき，考慮しなければならない事項が 2 点存在する．第一に，空間計算量の問題点を解決する必要がある．DP マッチングを行なうためには，対象とする 2 個の時系列データの長さを I，J とした場合，$I \times J$ の 2 次元配列を確保する必要がある．本研究で

は $F_s = 48000$ Hz で標本化したデータを用いるため，例えば時間長 5 min の信号同士を DP マッチングさせる場合，$I \times J = (300 \times F_s)^2 = 2.0736 \times 10^{14}$ 個のデータを記憶する必要がある．1 個のデータ量を 32 bit とすれば，必要となるメモリ容量は 0.7 PB 以上となるため，これは現実的ではない．第二に，フラットネス信号のマッチング結果を，オンセットマッチングと紐付けるための指針を設定しなければならない．スペクトルフラットネスではオンセット位置を正確に知ることはできず，あくまでも音韻の位置を大局的に同定することしかできない．そこで，各歌唱信号に含まれる音韻を同定した後，さらにオンセットを局所的に対応付けるための指針が必要となる．これら 2 点の事項を考慮したオンセットマッチングの提案手法を，図 3.19 に示す．

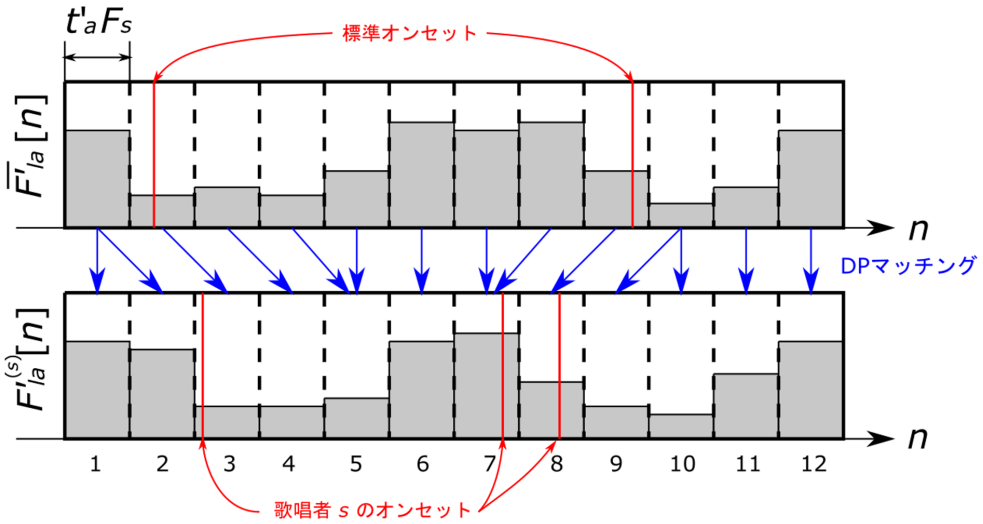

図 3.19: DP マッチングを用いたオンセットマッチングの概略図

まず，$F^{(s)}_{la}[i]$ と $\bar{F}_{la}[i]$ を，長さ t'_a [sec] の分析フレームで分割する．さらに，各フレーム内で $F^{(s)}_{la}[i]$ と $\bar{F}_{la}[i]$ それぞれの平均値を求め，その値をフレーム番号 $n = 1, 2, \cdots, N$ の順に並べる．こうして圧縮されたフラットネス信号と標準フラットネス信号を，$F'^{(s)}_{la}[n]$, $\bar{F}'_{la}[n]$ と表現する．このとき，$\bar{F}'_{la}[n]$ は式 (3.14) のように定義できる．$F'^{(s)}_{la}[n]$ も同様である．

$$\bar{F}'_{la}[n] = \frac{1}{t'_a F s} \sum_{m=0}^{t'_a F_s - 1} \bar{F}_{la}[n t'_a F s + m] \tag{3.14}$$

重要なのは，分析フレーム幅 t'_a は，オンセット検出の際の分析フレーム t_a に対して $t'_a > t_a$ とすることである．本研究では，$t_a = 1$ ms に対して $t'_a = 100$ ms と設定する．ここで，毎秒 48000 個のデータを持つ（標準）フラットネス信号 $F^{(s)}_{la}[i]$，$\bar{F}_{la}[i]$ の代わ

りに，毎秒 10 個のデータに圧縮された（標準）フラットネス信号 $F'^{(s)}_{la}[n]$，$\bar{F}'_{la}[n]$ で DP マッチングを行なう．こうすることで，メモリ使用量を $\frac{10^2}{48000^2}$ 倍に抑えることができる．時間長 5 min の信号であれば必要となるメモリ容量は約 34 MB となるため，実装にあたって現実的な数値であると言える．なお，DP マッチングに用いる要素間のコスト (局所距離)$c[i]$ は，$c[i] = |F'^{(s)}_{la}[n] - \bar{F}'_{la}[n]|$ と定義する．

次に，DP マッチングの結果を用いてオンセットマッチングを行なうための指針を示す．DP マッチングにより，標準フラットネス信号 $\bar{F}_{la}[i]$ の第 p フレームと，フラットネス信号 $F^{(s)}_{la}[i]$ の第 q フレームが，1 対 1 でマッチングしたと仮定する．このとき，標準オンセット信号 $\bar{o}[i]$ のうち第 p フレームに相当する時間範囲と，オンセット信号 $o^{(s)}[i]$ のうち第 q フレームに相当する時間範囲を参照する．範囲内に $\bar{o}[i] = 1$ となる時刻と $o^{(s)}[i] = 1$ となる時刻が存在した場合，それらの時刻のオンセットをマッチングさせる．図 3.19 の例では，$\bar{F}_{la}[i]$ の第 2 フレームと $F^{(s)}_{la}[i]$ の第 3 フレームがマッチングすることから，第 2 フレームに存在する標準オンセットと，第 3 フレームに存在する歌唱者 s のオンセットをマッチングさせることに相当する．一方，歌唱者 s のオンセットのうち第 7 フレームに存在するオンセットは，マッチングする標準オンセットが存在しないため，このような場合はマッチング不成立とする．ここで，オンセット検出における平滑化の処理ブロックで，各オンセット間の最小間隔を $t_v = 150$ ms と設定していた．そのため，$t'_a = 100$ ms の分析フレームで区切った場合，各フレームに存在するオンセットの個数は最大で 1 個であることが保証される．よって，1 対 1 でマッチングするフレーム内では，必ずオンセットも 1 対 1 でマッチングする．しかし，弾性マッチングは非線形な対応付けを行なうため，1 個のフレームが複数のフレームとマッチングする場合もある．それにより，オンセットが 1 対 1 でマッチングしない場合には，誤った対応付けを避けるためにマッチング不成立とする．

以上の手順により，音響特徴量に基づいたオンセットマッチングが可能となる．マッチング完了後には，歌唱者のオンセット時刻と標準オンセット時刻の差 $\Delta t^{(s)}$ を出力する．ここで，$\bar{o}[i] = 1$ となるサンプル番号を i_{sd} とし，これとマッチングした歌唱者 s のオンセットのサンプル番号を $i^{(s)}_{ind}$ とすると，$\Delta t^{(s)}$ は式 (3.15) のように定義される．

$$\Delta t^{(s)} = \frac{i^{(s)}_{ind} - i_{sd}}{F_s} \tag{3.15}$$

この $\Delta t^{(s)}$ の値は，マッチングが成立したオンセットの組の個数だけ出力を行なう．

3.3 歌唱者由来のずれの補正

歌唱者由来のずれを補正する提案手法の処理ブロック図を，図 3.20 に示す．

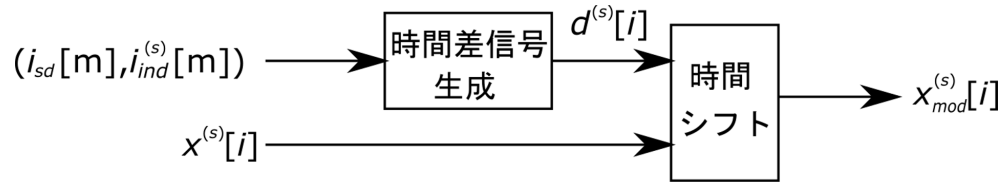

図 3.20: 歌唱者由来のずれ補正法の処理ブロック図

各処理ブロックについて，個別に詳細を述べる．

- 処理ブロック 1：時間差信号生成

この処理ブロックに入力されるのは，オンセットマッチングにより対応付けられた，標準オンセット時刻と各歌唱者のオンセット時刻の組 $(i_{sd}[m], i^{(s)}_{ind}[m])$（$m = 1, 2, \cdots, M$）である．$m$ は標準オンセットに対して時系列順に付した番号で，最大値 M は標準オンセットの個数を表す．ここで，m 番目の標準オンセットにマッチングする歌唱者 s のオンセットが存在しない場合．$i^{(s)}_{ind}[m]$ は定義されないことに注意されたい．なお，このように $(i_{sd}[m], i^{(s)}_{ind}[m])$ によって表されるオンセット時刻の組を，マッチングテーブルと呼ぶこととする．

この処理ブロックの出力は，歌唱信号 $x^{(s)}[i]$ の各時刻における標準的な歌唱からのずれ量 $d^{(s)}[i]$ である．これを時間差信号と呼ぶこととする．時間差信号 $d^{(s)}[i]$ は，マッチングテーブルを用いて式 (3.16) のように計算される．

$$d^{(s)}[i] = \begin{cases} 0 & (i < i_{sd}[1]) \\ i^{(s)}_{ind}[m] - i_{sd}[m] & (i_{sd}[m] \leq i < i_{sd}[m+1]) \end{cases} \tag{3.16}$$

時間差信号 $d^{(s)}[i]$ は，時間の経過とともに標準オンセット時刻 $i_{sd}[m]$ になるたびに値が変化する信号であり，標準オンセット時刻以外では値が変化しない．これは，歌唱者が各オンセットの時刻でのみ歌唱のずれ量が変化し，それ以外では一定のずれ量を保ち続けるという仮定を表している．ここで，オンセット時刻と $d^{(s)}[i]$ の関係を表した模式図を，図 3.21 に示す．

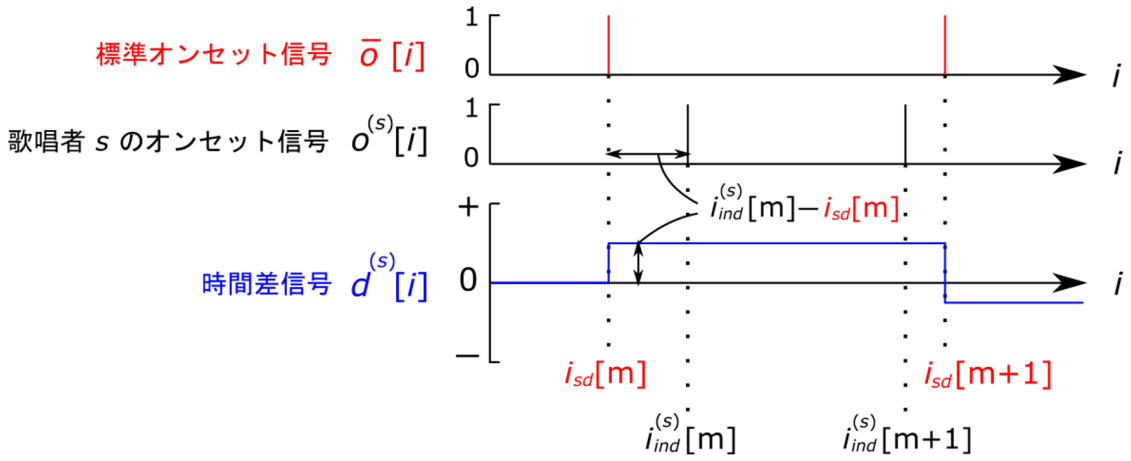

図 3.21: オンセット時刻と $d^{(s)}[i]$ の関係

図 3.21 に示したように，$d^{(s)}[i] < 0$ のときは標準的な歌唱よりも早く，$d^{(s)}[i] > 0$ のときは遅れて歌唱していることを表す．

- 処理ブロック 2：時間シフト

この処理ブロックの入力は，各歌唱者の歌唱信号 $x^{(s)}[i]$ と，時間差信号 $d^{(s)}[i]$ である．一方で出力は，時間差信号をもとに歌唱者由来のずれを補正した歌唱信号 $x^{(s)}_{mod}[i]$ である．ここでは，各歌唱者のオンセット時刻 $i^{(s)}_{ind}[m]$ が標準オンセット時刻 $i_{sd}[m]$ と完全に一致するように補正するものとする．この場合，式 (3.17) のように変換を行なうことで，目的を達成できる．

$$x^{(s)}_{mod}[i] = \begin{cases} x^{(s)}[i] & (i < i_{sd}[1]) \\ x^{(s)}[i + i^{(s)}_{ind}[m] - i_{sd}[m]] & (i_{sd}[m] \leq i < i_{sd}[m+1]) \end{cases} \tag{3.17}$$

時間シフトにより，標準オンセット時刻 $i_{sd}[m]$ では歌唱信号が不連続となってしまう．これを避けるため，$d^{(s)}[i_{sd}[m+1]] > d^{(s)}[i_{sd}[m]]$ のとき，すなわち次のオンセットが遅れているため時間軸を縮める場合は，クロスフェードを行なう．一方，$d^{(s)}[i_{sd}[m] > d^{(s)}[i_{sd}[m+1]]$ のとき，すなわち次のオンセットが早いため時間軸を伸ばす場合は，フェードアウト，フェードインをした上で，間に無音を挿入する．この操作の模式図を，図 3.22 に示す．

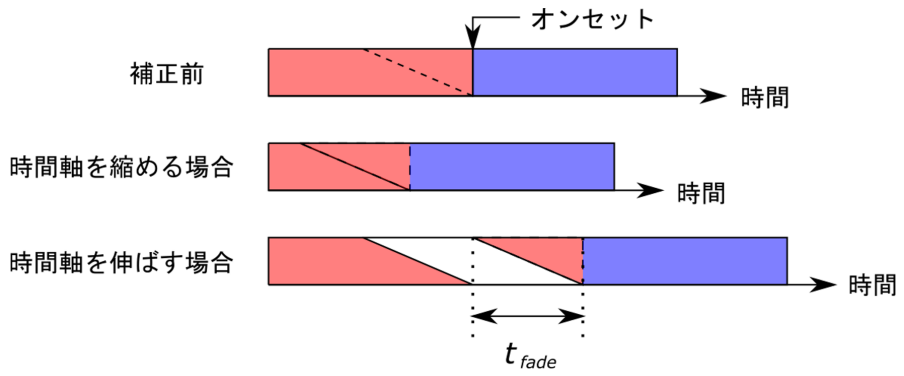

図 3.22: フェード処理の模式図

図 3.22 に示したように，フェード処理はオンセット時刻の直前で行なう．これは，音のアタック部分ではオリジナルの波形が残るようにするためである．ここで，フェード時間 t_{fade} は 100 ms と設定する．

以上の処理により，歌唱者由来のずれを補正した歌唱信号 $x^{(s)}_{mod}[i]$ が得られる．

3.4 提案手法における入力信号の格納法

本研究では，提案手法の実装を全て C 言語プログラムにて行なう．実装において，提案手法をリモート合唱に実用化するために重要となるのが，入力信号の格納法である．

例として，2 入力の単純加算ミキサーをプログラムで実装する場合を考える．この場合，あらかじめ 2 個の配列を用意しておき，入力信号を各配列に記憶して加算すれば良い．しかし，本研究では入力信号を $x^{(s)}[i]$（$i = 1, 2, \cdots, S$）と表現したように，リモート合唱のミキシングシステムに入力される信号の個数は，毎回同じとは限らない．しかし，個数が変わるたびに別のプログラムを使用することは，実用性に乏しい．そこで，入力信号の個数にかかわらず同一のプログラムで処理できるようにする工夫が必要である．

本研究では，この解決策として，マルチチャネル信号を一つの時系列データとして扱う規格を設け，この規格に沿って格納した信号を各処理ブロックへの入力とする．この格納法の概要を，図 3.23 に示す．

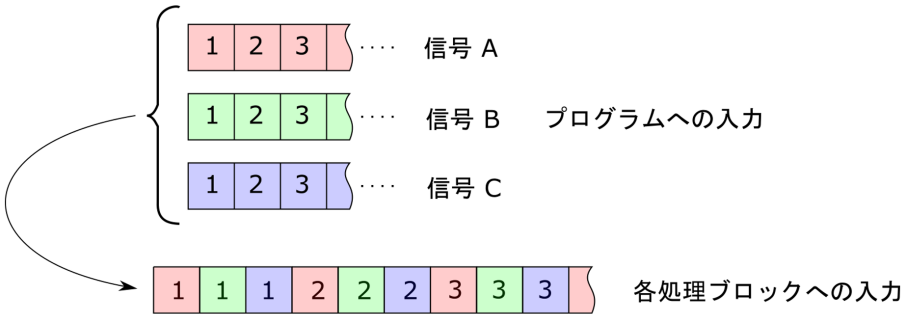

図 3.23: マルチチャネル信号の格納法

図 3.23 ではチャネル数が 3 の場合を表したが，他の場合も同様で，入力チャネル数にかかわらず同一のプログラムで処理が可能となる．提案手法のうち，マルチチャネル信号を扱う必要がある歌唱者由来のずれの分析，および補正のプログラムには，この手法を導入する．

第 4 章
実験

本研究では，5 項目の実験を行なった．第一に，システム由来のずれを解消する手法の精度評価，第二に，オンセット検出のために提案した FN 法の精度評価，第三に，オンセットマッチング手法の精度評価，第四に，歌唱者由来のずれ量の分析，第五に，歌唱者由来のずれ補正の検証である．実験の概要と結果の考察を，項目ごとに述べる．

4.1 システム由来のずれを解消する手法の精度評価

4.1.1 概要

システム由来の発声タイミングずれを解消する手法の時間精度を評価する実験を行なった．本研究をリモート合唱の自動ミキシングシステムに実用化する場合を考慮し，本番環境にならった次の実験方法をとった．まず，合唱曲の伴奏音源を含む指揮動画の冒頭，末尾の 2 カ所にマーク信号を挿入したものを用意した．被験者は，この動画を使用して次のような作業を行った．1. 録音端末を用意し，任意の時間に録音を開始，2. ヘッドフォンを接続した別の端末を用意し，任意のタイミングで指揮動画のストリーミング再生を開始，3. 直ちにヘッドフォンを録音端末に近づけて冒頭のマーク信号を録音，4. 録音状態のままヘッドフォンを装着し，指揮と伴奏に合わせて歌唱，5. 歌唱

後，録音状態のままヘッドフォンを録音端末に近づけ，末尾のマーク信号を録音，6. 録音を停止．手順 3 と手順 5 で行なうマーク信号の録音の概要を，図 4.1 に示す．

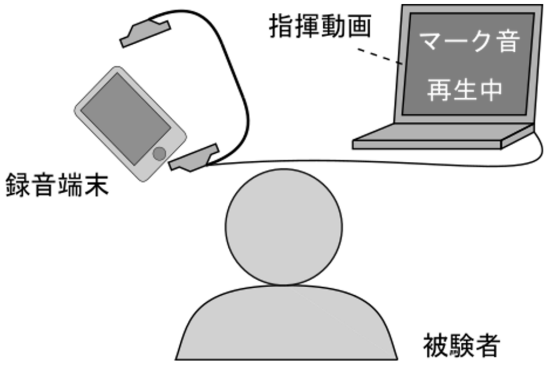

図 4.1: マーク信号の録音方法

次に，録音された全ての歌唱信号に対し式 (3.2) を適用し，2 カ所のマーク信号間の時間を推定した．ここで，t_m は指揮動画に挿入された 2 カ所のマーク信号間の時間，$\hat{t}_m$ は歌唱信号から推定された 2 カ所のマーク信号間の時間とする．指揮動画における t_m と歌唱録音における $\hat{t}_m$ の定義を，図 4.2 に示す．

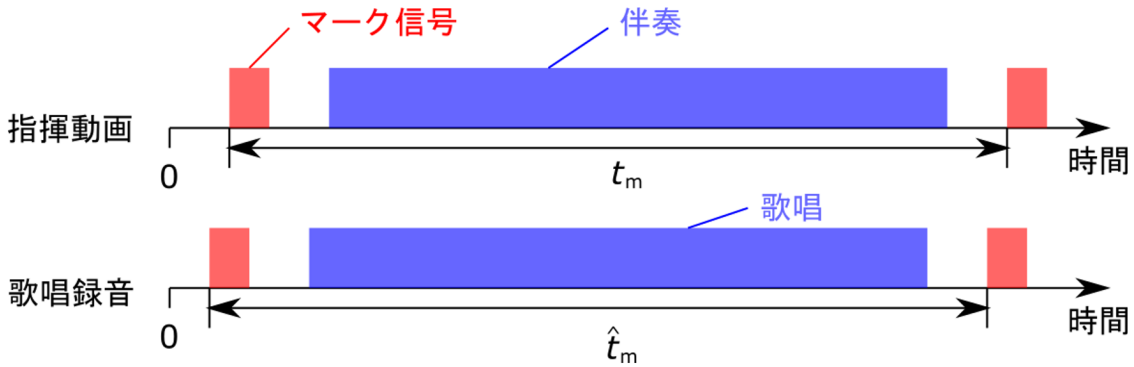

図 4.2: t_m と $\hat{t}_m$ の定義

最後に，式 (4.1) で表される値 r を用いて，提案手法の時間精度を評価した．

$$r = \frac{|\hat{t}_m - t_m|}{t_m} \times 100\ \% \tag{4.1}$$

ここで，r は本来のマーク信号間の時間長に対する，録音によって生じた時間ずれの割合に相当する．

エンジニアでない一般人が，近年の端末やアプリを任意に選択し録音した結果，どのような音源が集まるのかは興味深い．そこで，被験者らが使用する端末，ソフトウ

エア，並びに録音を行う場所は完全に任意とした．ただし，録音形式はサンプリングレート $F_s = 48000$ Hz，モノラルの，非可逆圧縮でない形式を推奨した．録音データはアマチュア合唱団員，合唱経験を問わない非合唱団員から募集した．なお，被験者には録音データを研究目的で使用することへの許諾を得た．また，新型コロナ感染防止の観点から，著者は被験者の歌唱録音への立ち会いを行なっていない．

本実験で使用した楽曲，並びに各楽曲における被験者情報を，表 4.1 に示す．楽曲は，いずれもピアノ伴奏が存在する合唱曲である．

表 4.1: 楽曲，被験者情報

曲名	人数 [人]	被験者所属
雨	15	アマチュア合唱団員
群青	17	アマチュア合唱団員
旅立ちの日に	29	非合唱団員

4.1.2 結果と考察

まず，音源の収集結果から述べる．本実験のために収集した全 61 音源のうち，非合唱団員の 6 人から提供されたものは，非可逆圧縮の形式であった．内訳は，AAC 形式が 3 個，MP3 形式が 3 個であった．サンプリングレートについては，全体の約 90 %にあたる 56 音源が $F_s = 48000$ Hz で録音され，残りの 5 音源は $F_s = 44100$ Hz であった．

収集した音源は全て $F_s = 48000$ Hz，ビット深度 24 bit，モノラルの WAVE 形式に変換した．変換にあたっては，フリーソフトウェアの「FFmpeg」[27] を使用した．なお，非合唱団員のうち 4 名の歌唱録音では末尾のマーク信号が未収録であったため，評価対象からは除外した．

次に，各歌唱録音に対する $\hat{t}_m$ の計算結果を，楽曲ごとに示す.「雨」,「群青」,「旅立ちの日に」における計算結果を，それぞれ図 4.3，図 4.4，図 4.5 に示す．

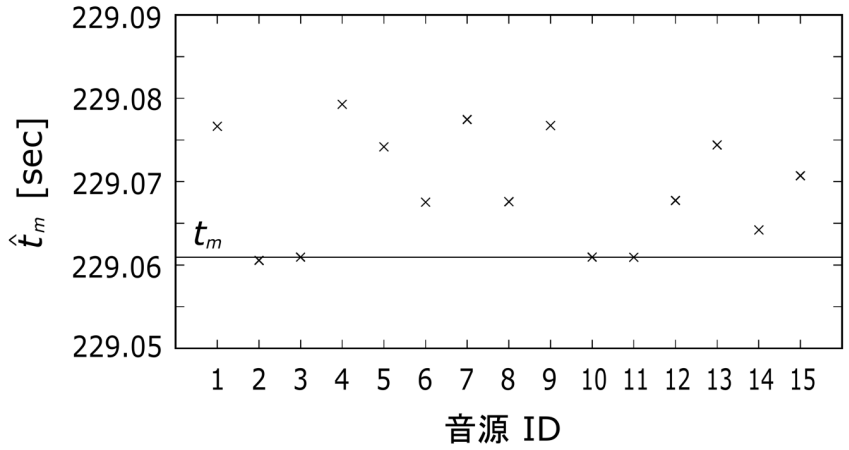

図 4.3: 「雨」における $\hat{t}_m$ の計算結果

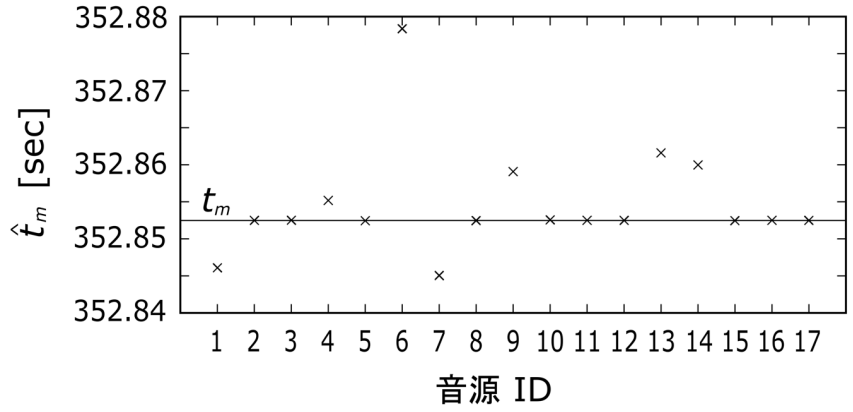

図 4.4: 「群青」における $\hat{t}_m$ の計算結果

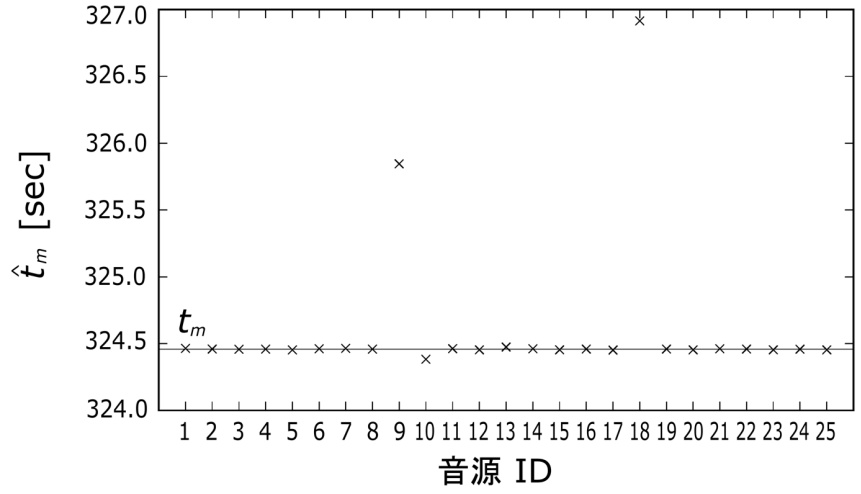

図 4.5: 「旅立ちの日に」における $\hat{t}_m$ の計算結果．他 2 曲とは時間スケールが異なることに注意．

$\hat{t}_m$ 計算結果のグラフには，指揮動画のマーク信号間時間長 t_m を併記している．3 曲の結果に共通して，録音により本来の時間軸が拡大される傾向が見られる．この原因としては，録音端末のクロック周波数のばらつき，指揮動画のストリーミング再生時の遅延，Bluetooth イヤホンを使用していた場合にはその遅延量の変動も考えられる．音源によっては時間軸が縮小されるものも存在するが，縮小の場合にはクロック周波数のずれである可能性が高い．「旅立ちの日に」では，9 番，18 番の音源で 1 sec 以上の誤差があり，他とは明らかに異なるオーダーでずれが生じている．クロック周波数の許容偏差は最大でも ±100 ppm 程度 [28] とされており，5 min 間で 1 sec 以上ずれることは考えにくい．Bluetooth イヤホンの遅延変動も，その変動量は一般的に最大 ±50ms[29] とされていることから，原因としては考えにくい．よって，この 2 音源の大きなずれは，ストリーミング再生時の遅延によるものと考えるのが妥当である．

続いて，本来の時間軸からのずれ量の割合 r を計算した結果を，表 4.2 に示す．ここでは，各楽曲についての平均値 $\bar{r}$ を示す．

表 4.2: $\bar{r}$ 計算結果

曲名	$\bar{r}$ [%]
雨	0.0037
群青	0.0011
旅立ちの日に	0.0494 (0.0021)[1]

表 4.2 を見ると，時間軸ずれの割合は，2 個の外れ値を除いて楽曲時間長のおよそ 0.001 %から 0.004 %である．0.004 %の場合，5 min 間の楽曲で最大 12 ms の時間ずれが生じることを意味する．音楽合奏において，演奏者がずれたと認識し始めるオンセットのずれ量は 100 ms 以上である [11] が，それよりも十分に短い．よって，2 個の外れ値の原因と考えられる動画再生時の遅延さえ無ければ，実用上無視できる程度の時間ずれであると言える．

なお,「雨」,「群青」の被験者のうち 8 人は，両楽曲の実験に参加したことが確認された．しかし，同じ被験者でも r の値は 2 曲で異なる結果となった．さらに,「雨」では $r = 0$ となった被験者は 0 人であったが,「群青」では 5 人が $r = 0$ となった．これら

[1] 括弧内は 2 個の外れ値を除いた計算結果

の 2 点から，録音による時間軸のずれは偶発的な要因が影響しており，条件が良い場合には時間軸のずれが全く生じないことが示された．

以上の考察から，提案手法は，動画のストリーミング再生中に遅延が生じない環境であれば，ほとんどの場合でシステム由来のずれの解消に有効であると言える．

4.2 FN 法の精度評価

4.2.1 概要

歌声のオンセットを検出するための提案手法である，FN 法の精度を評価する実験を行なった．実験に用いたのは，表 4.1 に示した音源の中から混声四部合唱「雨」のソプラノパート歌唱と，同バスパート歌唱の各 1 音源ずつである．各パートには歌唱者の異なる音源が 3 個ずつ存在するが，その中から無作為に 1 個ずつを選択した．また，オンセット検出の前処理として，図 3.9 と同等の遮断特性を有する遮断周波数 100 Hz のハイパスフィルタで，低周波ノイズを減衰させた．各音源の特徴を，表 4.3 に示す．

表 4.3: 音源情報

	時間長 [sec]	テンポ [BPM]	最低周波数 [Hz]	最高周波数 [Hz]
ソプラノ歌唱	180	116	262 (C4)	698 (F5)
バス歌唱			104 (A♭2)	233 (B♭3)

表 4.3 に記した時間長は，前奏と後奏を除いた歌唱部分だけの長さである．また，各音源の最低周波数，最高周波数はともに 80 Hz∼900 Hz の範囲内であることから，第 3.2.2 項で仮定した基本波周波数の条件を満たしている．

これらの音源に FN 法を適用し，オンセット検出を行なった．オンセット検出結果は，オンセット信号 $o[i]$ をグラフに出力することで可視化した．このグラフと時間周波数平面の比較，あるいは音源を聴取してオンセット時刻を比較することで，オンセット検出の精度を評価した．評価にあたっては，表 4.4 に示す真陽性率（TP 率）と偽陽性率（FP 率）の計算を行なった．

表 4.4: オンセット検出精度の評価項目

評価項目	定義
真陽性率（TP 率）	実際のオンセット数に対する検出されたオンセット数の割合
偽陽性率（FP 率）	検出されたオンセット数に対する誤検出数の割合

4.2.2 結果と考察

表 4.3 に示した各音源に FN 法を適用し，オンセット検出を行なった結果を図 4.6，図 4.7 に示す.

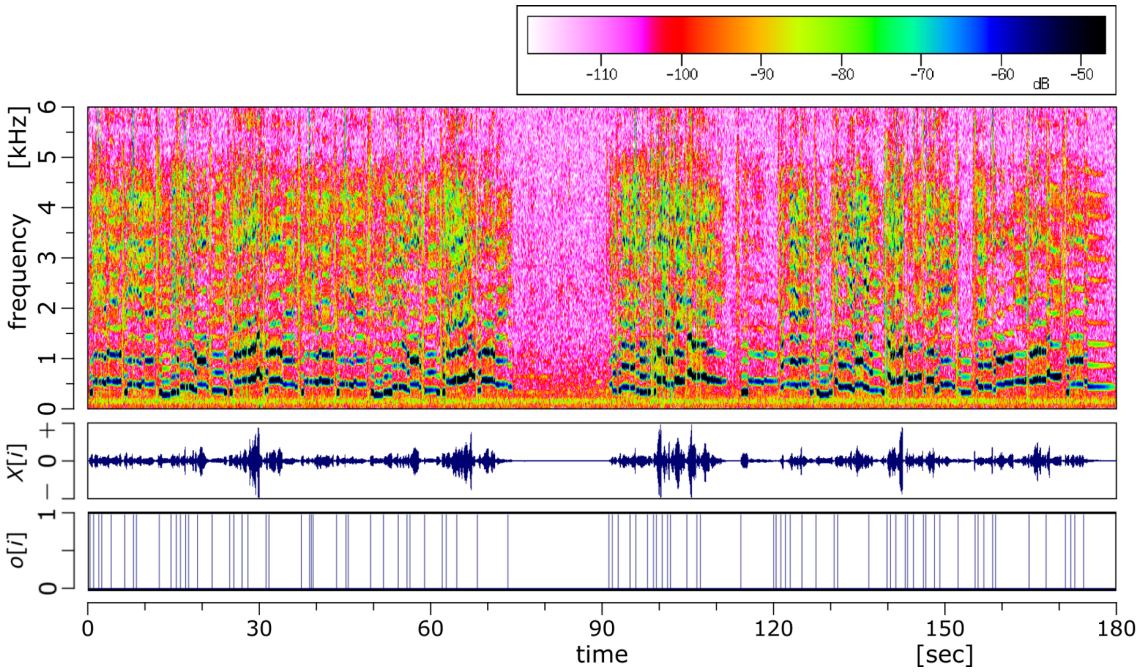

図 4.6: 「雨」ソプラノ歌唱におけるオンセット検出結果

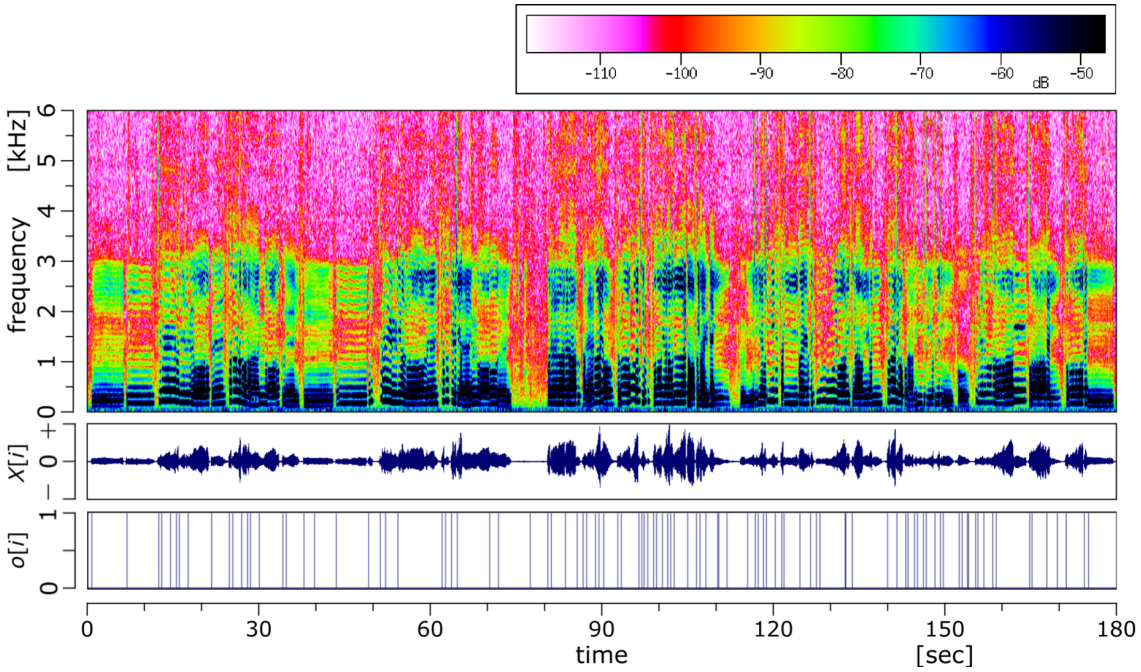

図 4.7: 「雨」バス歌唱におけるオンセット検出結果

続いて，各音源に実在するオンセットの個数と，検出されたオンセットの個数を表 4.5 に示す．ここで，各音源に実在するオンセットの個数は，音源を聴取することでカウントを行なった．オンセットとする条件は，歌詞における平仮名 1 文字に相当する音の立ち上がり，または，平仮名 1 文字の中でもリズムに合わせて音程が変化する場合は，その音程変化もオンセットとカウントした．

表 4.5: オンセット個数の比較

	実在するオンセット個数	検出されたオンセット個数
ソプラノ歌唱	185	89
バス歌唱	169	101

最後に，検出されたオンセットの TP 率，FP 率を計算した．TP 率に関しては歌唱のフレーズ始端部分と中間部分で大きな差が存在したため，それぞれ分けて記載する．ここで，フレーズの始端部分とは無音状態から発声を開始する部分のことを指し，休符の後や息継ぎの後の音のオンセットに相当する．一方でフレーズの中間部分とは，音と音が途切れずに歌唱される場合に，その移り変わり部分のことを指す．聴取による

カウントの結果，フレーズの始端部分に実在するオンセットの個数は，ソプラノ音源で 27 個，バス音源で 29 個であった．これ以外のオンセットはすべて，フレーズの中間部分のオンセットである．各音源のオンセット検出結果から計算された TP 率，FP 率を，表 4.6 に示す．

表 4.6: TP 率と FP 率の計算結果

	TP 率 [%]		FP 率 [%]
	フレーズ始端	フレーズ途中	
ソプラノ歌唱	100.0	38.2	2.2
バス歌唱	96.6	45.7	9.9

実験結果をもとに，FN 法のオンセット検出精度について考察する．まず，表 4.6 の TP 率を見ると，フレーズ始端では音源によらずほぼ確実にオンセットを検出できることが分かる．音楽合奏において，フレーズ始端のオンセットずれは特に演奏のずれを知覚させやすいと言われている [11]．よって，この手法を歌唱者由来のずれの補正に適用する場合，フレーズ始端のずれをほぼ確実に補正できると言えるので，この結果は好ましいことである．一方，フレーズ途中の検出率は 4 割程度であることから，フレーズ内の発声タイミングずれ量を計測または補正することは，フレーズ始端と比較して困難であると言える．FP 率はどちらの音源も 1 割未満であるが，これが発声タイミングずれ量の計測または補正にどの程度影響するかは，次の実験で検証するオンセットマッチングの精度次第である．オンセットマッチングで FP のオンセットと対応付けられる割合が高い場合，FP 率をさらに低下させることが求められる．

ここで，FP の原因について考察する．はじめに，ソプラノ歌唱で検出された FP の個数は 2 個であり，どちらも図 4.6 の 120 秒付近に存在する．この部分を拡大した図を，図 4.8 に示す．

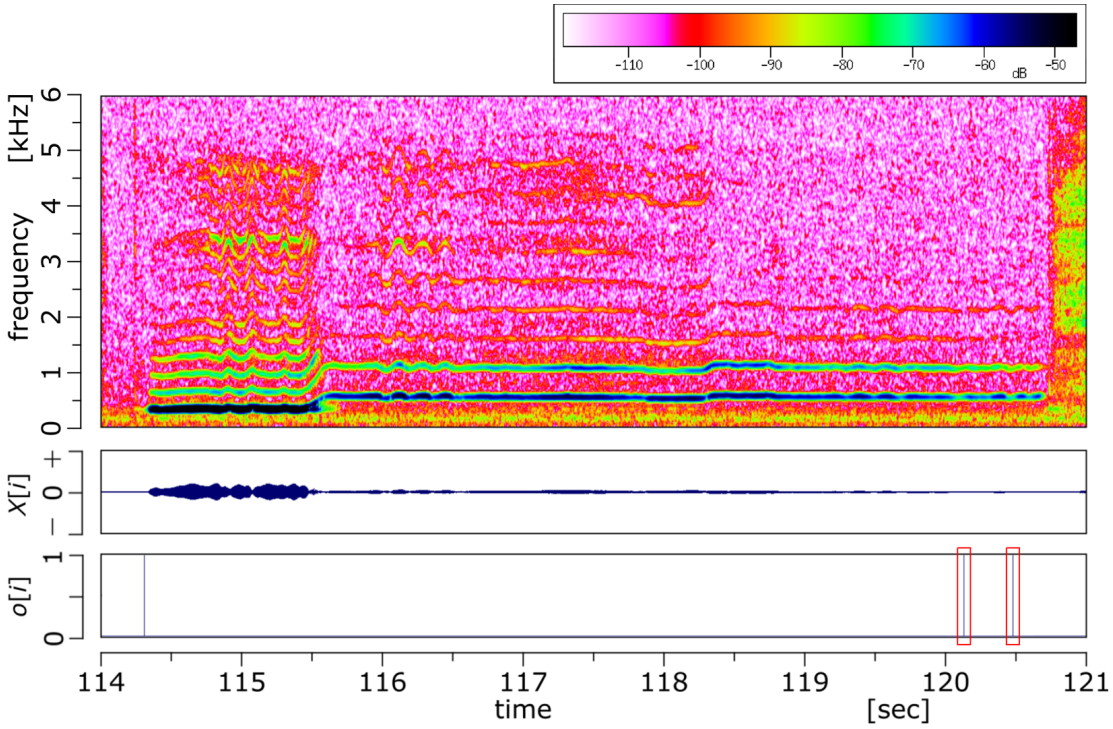

図 4.8: 「雨」ソプラノ歌唱において検出された FP．赤枠部分が FP に相当する．

114.3 sec にあるオンセットは正しく検出されたものであり，フレーズの最後に母音を長く伸ばす部分で FP が検出された．時間周波数平面上では，FP が検出される付近で基本波成分のパワーが局所的に弱まっている．そのため，有音区間が一度途切れたと判定され，誤ったオンセットが検出されてしまったと考えられる．このような誤検出は，式 (3.7) で表される平滑化のフィルタ長係数 N を大きく取ることで改善される可能性がある．一方，ソプラノ歌唱と比較してバス歌唱における FP 率が高い原因は，低周波ノイズによるものと考えられる．図 4.6 と図 4.7 の時間周波数平面を見ると，ソプラノ歌唱にも 100 Hz 付近に周期的なノイズが確認されるが，これは歌声の成分と比較し十分にパワーが弱いため，誤検出にはつながりにくい．一方で，バス歌唱では歌声の基本周波数よりも下の帯域において，歌声成分に匹敵する程度の強さのノイズが確認される．このような場合，歌声の基本波成分の波形に影響を与えたり，歌声が存在しない部分で偶然位相が周期的になってしまうことで誤検出が発生する．これを防ぐためには，前処理として，実験で用いたものよりもさらに遮断特性が急峻なサブソニックフィルタを用いて，低周波ノイズを除去することが必要だと考えられる．

最後に，FN 法によって検出できるオンセットの性質について考察する．2 音源から

検出されたオンセットを，有声音なのか無声音なのか，直前の音はどのような音であったのかによって分類した．その結果を表 4.7 に示す．表には，1 個以上オンセットが検出された場合には ◯ を，1 個も検出されなかった場合には × を記した．

表 4.7: 検出されたオンセットの分類

			オンセットが検出された音	
			有声音	無声音
直前の音	有声音	周波数変化あり	◯	×
		周波数変化なし	×	
	無声音または無音		◯	×

まず，FN 法では無声音の検出が不可能であることが分かる．これは，FN 法では信号の周期が安定ている場合に音が存在すると判定しているため，周期が不安定な無声音では音が存在しないと判定された結果である．次に，有声音であれば基本的には検出できるが，有声音から有声音に周波数変化を伴わずに変化する場合は検出できないことが分かる．これは，同じ周波数で有声音が連続する場合，基本波成分が途切れることが無く継続するためである．このケースとして多いのは，有声子音を持つ音における母音のオンセットである．さらに，この検出の特性は，歌声のオンセット検出における重要な制約を与える．その制約とは，FN 法で歌声のオンセット検出を行なった場合，歌詞の平仮名 1 文字に相当する音に対しては 1 つだけオンセットが検出されることである．すなわち，平仮名 1 文字に対して，子音の立ち上がりと母音の立ち上がりを両方検出することは無い．歌唱者由来のずれを補正する場合，1 音に対して子音部と母音部両方のオンセットを補正してしまうと，歌詞の発音の明瞭度が低下するといった弊害を及ぼす恐れがある．よってこの制約は，過剰な補正を防止するという点で，本研究においては好ましい影響を与えると言える．

4.3 オンセットマッチング手法の精度評価

4.3.1 概要

本研究で提案する，オンセットマッチング手法の精度を評価する実験を行なった．先行研究 [13] では，リファレンスとなるオンセット時刻から ±150 ms 以内に存在するオンセットを，一つの音に対するオンセットとみなしていた．この範囲外にあるオンセットを対応付けるためには時間範囲を広げれば良いが，その場合，実際には異なる音のオンセットを対応付けてしまう可能性がある．こうした場合に，時間関係に縛られること無く適切な範囲で対応付けを決定することが，提案手法のねらいである．よって，これまでに収集したリモート合唱音源の中でも特に大きな発声タイミングのずれを知覚するものを選定し，適切な対応付けが可能であるかどうかを評価した．

実験に用いた音源は，混声四部合唱「心の瞳」のソプラノパート歌唱である．歌唱人数は 7 人で，第 4.1.1 項に示したリモート合唱の収録方法に則り，各音源には 1 人ずつの歌唱が録音されている．また，実験の前処理として，第 3.1 節で提案した手法を用いてシステム由来のずれを解消した．実験には，この楽曲の末尾 10 sec 間の「きみをみつめれば」と歌唱する部分を抜粋して使用した．各音源の時間波形を，図 4.9 に示す．

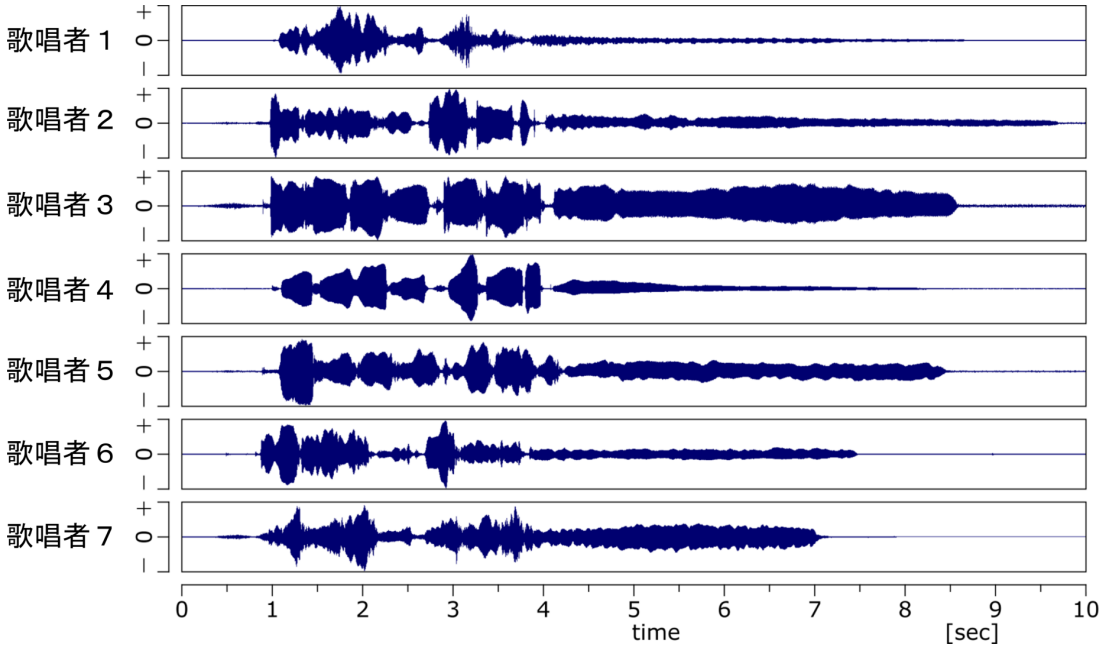

図 4.9: 「心の瞳」ソプラノパート 7 人の歌唱信号の時間波形

抜粋した区間では，曲のテンポがだんだん遅くなるリタルダンドの表現が用いられている．リモート合唱ではまわりの歌唱者の声が聞こえないため，歌唱者ごとのテンポの遅らせ方の違いが顕著に現れる．よって，歌唱者の間には最大250 msもの発声タイミングずれが生じている．

実験手順を示す．まず，各音源に対しFN法を用いてオンセット検出を行った．次に，第3.2.3項の手法を用いて標準オンセットを計算した．さらに，第3.2.4項の手法を用いてオンセットマッチングを行なった．最後に，各音源のオンセット信号$o[i]$をグラフ出力し，マッチングテーブルを参照して標準オンセットとの対応付けを確認することにより，マッチング結果を評価した．評価にあたっては，検出された各オンセットを表4.8に示すマッチング結果の項目に当てはめ，さらに式(4.2)と式(4.3)に示す2種類のマッチングエラー率E_1，E_2を計算した．

表 4.8: マッチング結果の定義

項目	定義
真陽性（TP）	検出されたオンセットのうち，それ自身と同じ音の標準オンセットとマッチングしたもの
偽陽性（FP）	検出されたオンセットのうち，それ自身とは異なる音の標準オンセットとマッチングしたもの
真陰性（TN）	検出されたオンセットのうち，それ自身と同じ音の標準オンセットが存在しないため，マッチングしなかったもの
偽陰性（FN）	検出されたオンセットのうち，それ自身と同じ音の標準オンセットが存在するが，マッチングしなかったもの

$$E_1 = \frac{N_{FP}}{N_{TP} + N_{FP} + N_{FN}} \tag{4.2}$$

$$E_2 = \frac{N_{FN}}{N_{TP} + N_{FN}} \tag{4.3}$$

ここで，N_{TP}，N_{FP}，N_{TN}，N_{FN}はそれぞれTN，FP，TN，FNの検出個数である．E_1は全体の検出個数に占めるFPの割合であり，この値が小さいほど歌唱者由来のず

れの分析結果の信頼性が高いことを表す．一方 E_2 は，誤り無く検出されたオンセットのうちマッチングできなかったオンセットの割合である．この値は，オンセット検出の不正確性を排除し，オンセットマッチング手法単体に対するエラーレートに相当する．

4.3.2 結果と考察

図 4.9 に示した 7 音源に対してオンセットマッチングを行なった結果を，図 4.10 に示す．図 4.10 は，上から第三段以降に各音源から計算されたオンセット信号を示し，参考用として最上段に歌唱者 1 の音源の時間周波数平面と歌詞，第二段に同じく時間波形を示す．

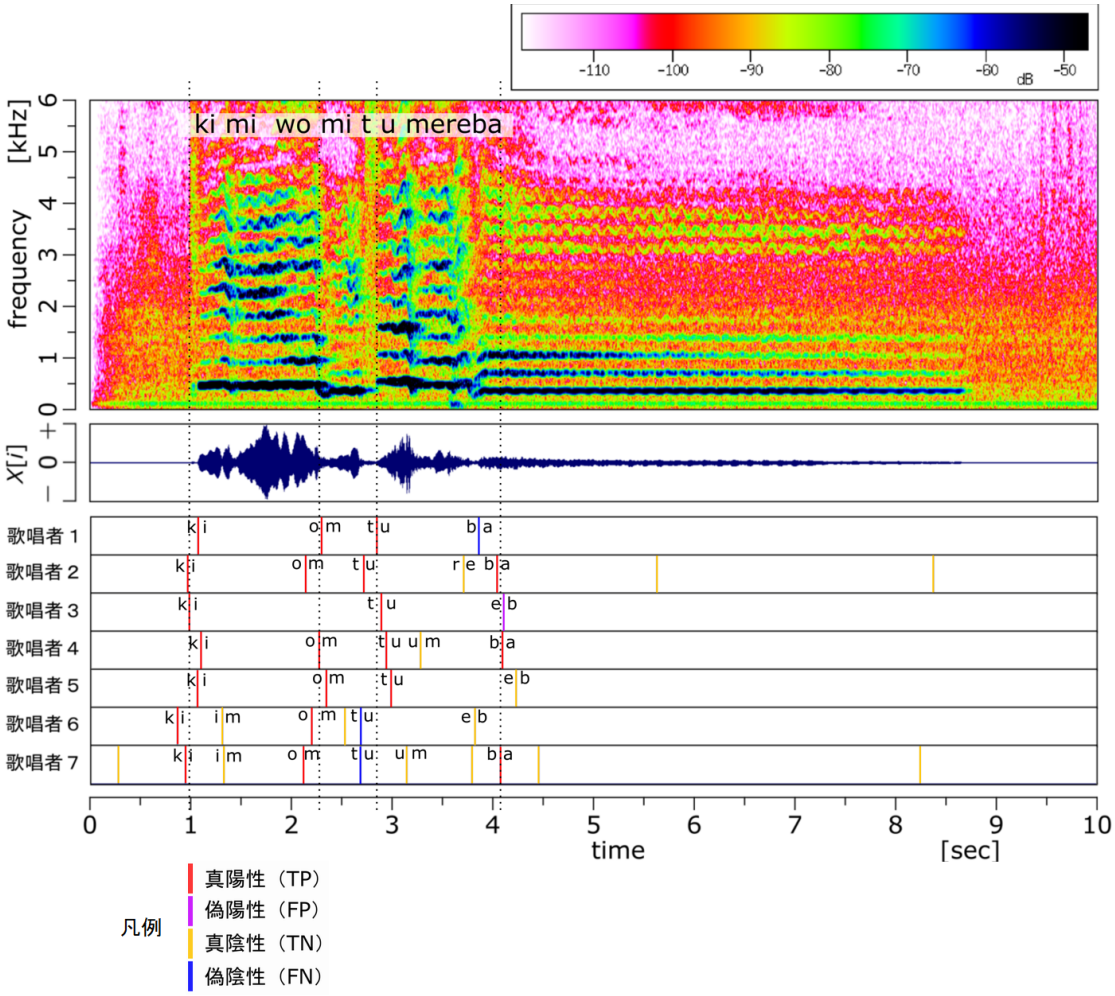

図 4.10: オンセットマッチング結果

図中の点線は標準オンセットの位置を表し，これに対するマッチング結果として，凡

例の色の通りにオンセット信号を表記した．検出された各オンセットには，それぞれどの音韻からどの音韻に変化する部分であるのかをローマ字で併記した．併記していないものは，検出された時刻にはオンセットが存在しないことを表す．

マッチング結果は，$N_{TP} = 21$，$N_{FP} = 1$，$N_{TN} = 14$，$N_{FN} = 3$ であった．これをもとに E_1 と E_2 を計算すると，$E_1 = 2.6$ %，$E_2 = 12.5$ %となる．E_1 の値が大きい場合，このマッチング結果から歌唱者由来のずれ量を計測し分析したとしても，同一のオンセット同士での比較でない可能性が高く，分析結果の信頼性が低下する．本実験では，歌唱者間でのずれが特に大きく，マッチング難易度の高い音源を使用したにもかかわらず，$E_1 = 2.6\%$ という低いエラーレートを示している．よって，本手法を使用して歌唱者由来のずれを分析した場合，十分な信頼性が得られると考えられる．

本実験では唯一，歌唱者 3 の歌詞「ば」において FP が発生した．ここでは，標準オンセットは母音の立ち上がりであるのに対し，検出されたオンセットは子音の立ち上がりである．よって，標準オンセットとマッチングするのは誤りである．この誤りの原因としては,「ば」の子音/b の持続時間が 25 ms 程度と短いため，100 ms の分析フレームで行なう DP マッチングでは，子音と母音の特徴量が平均化されて同一フレームに含まれてしまったことが考えられる．そもそも本手法では，同一音から検出されたオンセットにおいて，歌唱者によって検出部分が子音部と母音部で異なることは想定されていない．このケースでは，図 4.11 に示すように，歌唱者によって有声音である/b/が無声音である/p/に近い発音となり，その違いによって検出部分が異なったことも間接的な原因と言える．

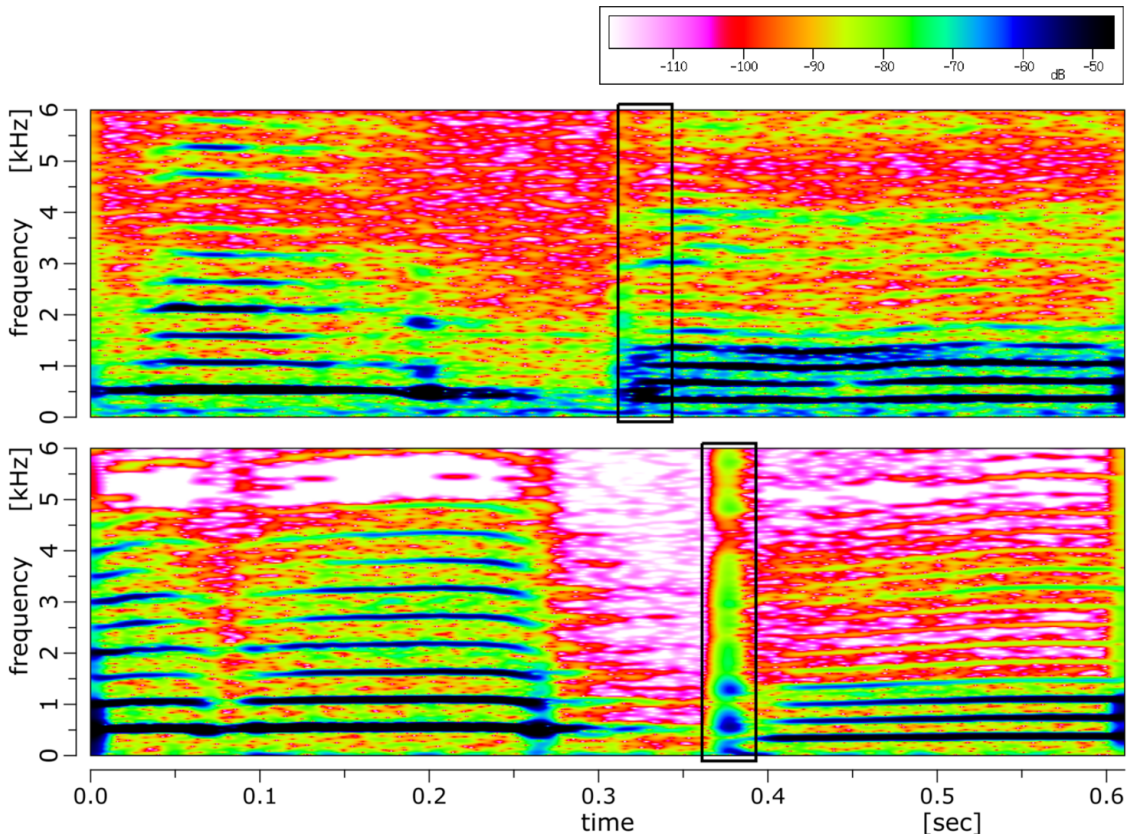

図 4.11: 歌唱者によって異なる「ば」の発音．/b/に近い発音の時間周波数平面（上）と，/p/に近い発音の時間周波数平面（下）．黒枠部が子音部分を表す．

このマッチングエラーは有声子音を検出する場合に起こり得るが，表 4.7 に示したように，FN 法で有声子音のオンセットが検出可能となるのは，無音の直後と周波数変化を伴う場合に限られる．よって，このようなエラーが全体のオンセット検出数に占める割合は小さく，歌唱者由来のずれの分析への影響は小さいと考えられる．

E_2 の値は 12.5 %と計算されたが，これはオンセットマッチングの手法を改善することにより，ずれを補正することのできるオンセット数が 12.5 %増加する余地があるということである．単純には DP マッチングのフレーム長を増大させることで，より多くのマッチングを行なうことができる．しかし，その場合には誤った対応付けも増加するため，最適なフレーム長を検討することが今後の課題である．一方，DP マッチングを行なわず， 時間長 100 ms の分析フレームを完全に線形に対応付けた場合のマッチングと比較すると，本手法の優位性が示される．DP マッチングを行なわなかった場合，図 4.12 の影を付けた部分に存在する 10 個のオンセットのみがマッチングする．こ

のときのエラーレートは $E_2 = 58.3$ %と計算される．すなわち，DP マッチングを導入することで，マッチングするオンセットの個数が約 46 %増加したと言える．

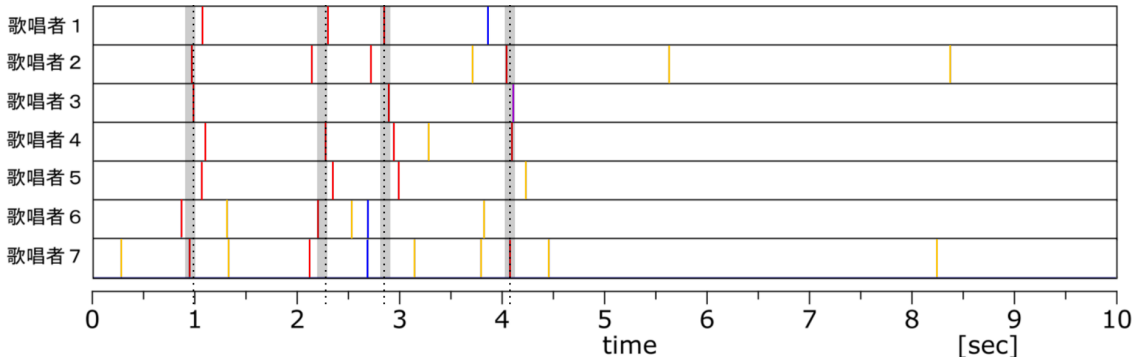

図 4.12: 線形対応付けの場合のマッチング範囲

4.4 歌唱者由来のずれ量の分析

4.4.1 概要

歌唱者由来の発声タイミングずれ量を分析する実験を行なった．ずれ量の計測には，図 3.3 に示した提案手法を使用した．分析に用いた音源は，第 4.1 節で述べたリモート合唱音源のうち,「群青」と「旅立ちの日に」の 2 曲の音源である．音源に関する情報を，表 4.9 に示す．

表 4.9: 音源情報

	人数 [人]	歌唱者所属
群青	ソプラノ:3，アルト:3，テノール:5，バス:3	アマチュア合唱団員
旅立ちの日に	ソプラノ:9，メゾソプラノ:7，アルト:5	非合唱団員

	時間長 [sec]	伴奏テンポ [BPM]
群青	275	72~79
旅立ちの日に	225	84

音源には複数の楽曲とパートが含まれるが，ずれ量の計測は各楽曲の各パートごとに分けて行なった．また，音源の総数が表 4.1 に示したものと一致しないが，これは次

の 2 点の理由による．第一に，第 4.1 節で行った実験の結果，システム由来のずれ量が他の音源に比べて明らかに大きいことが判明した「旅立ちの日に」の 2 音源を除外したためである．第二に，標準オンセット時刻として各歌唱者のオンセット時刻の中央値を設定するため，音源数が偶数の場合，計測結果に偏りが生じる恐れがある．そこで，各パートの音源数が奇数となるように一部の音源を除外したためである．

図 3.3 に示した提案手法では，各標準オンセットに対する各歌唱者のオンセットのずれ量が出力される．この出力結果には，オンセット時刻がメジアンとなる歌唱者のオンセット時刻と，標準オンセット時刻の差も含まれる．これは歌唱者間の発声タイミングずれ量には相当しないため，結果の分析対象からは除外した．

4.4.2 結果と考察

歌唱者由来のずれ量の分析結果について述べる．なお，ずれ量の計測結果をもとに各音源を第 3.3 節に示した手法で時間シフトし，パートごとの音源を加算して聴取した．その結果，大きな発声タイミングずれは知覚しなかったため，オンセットマッチングに大きな誤りは発生していないことが期待される．分析にあたっては，計測された各歌唱者の発声タイミングずれ量 Δt のヒストグラムを作成し，その分布を推定した．まず，計測された Δt の個数を表 4.10 に示す．

表 4.10: 計測された Δt の個数

楽曲	パート	パートごとの個数	楽曲ごとの個数	総数
群青	ソプラノ	277	1403	3903
	アルト	302		
	テノール	512		
	バス	312		
旅立ちの日に	ソプラノ	1174	2500	
	メゾソプラノ	769		
	アルト	557		

次に，2 曲合わせた Δt のヒストグラムとその分布に関する統計量を，図 4.13 と表

4.11 にそれぞれ示す.

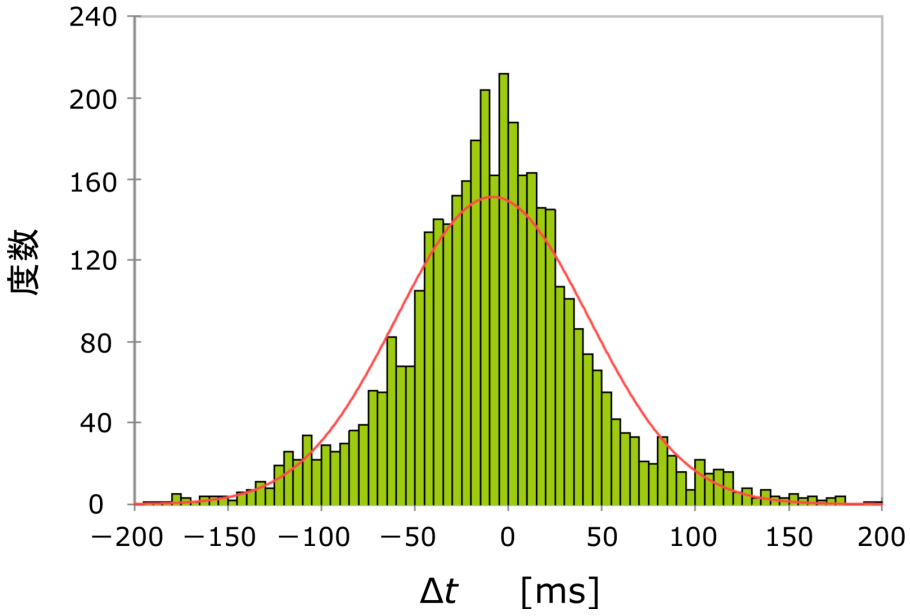

図 4.13: 2 曲合わせた Δt のヒストグラム．赤線は参考用のガウス分布を示す.

表 4.11: 2 曲合わせた Δt 分布の統計量

最小値 [ms]	最大値 [ms]	平均 [ms]	標準偏差 [ms]
−194	198	−8	51

以上に示したように，2 曲合わせた Δt の分布は，平均 −8 ms，標準偏差 51 ms となった．平均が 0 とならないのは，歌唱者の中央値となる発声タイミングからのずれ量を計測したためである．ヒストグラムを見ると，ガウス分布よりは平均値付近に分布が集中しており，スーパーガウシアン分布であると言える．舞台上のヴァイオリンセクションにおける演奏タイミングずれ量は，標準偏差 40 ms の分布となる [13] ことが知られている．それと比較すると，標準偏差が 51 ms となるリモート合唱の発声タイミングずれ量は，舞台上のヴァイオリンよりも大きくばらけていると言える．この差が器楽と声楽による違いなのか，あるいは舞台上とリモートの違いなのかは定かではない．しかし，一般に合奏のずれを知覚すると言われるずれ量は ±100 ms[11] とされるが，分析の結果，リモート合唱では ±100 ms 以上のずれが約 1 割存在すると判明した．これは，舞台上の合唱では修正されるべきずれ量である．よって，リモート合

唱の完成度を高めるためには，こうした大きなずれを補正する必要があると言える．

次に，曲ごとに分類した Δt の分析結果を示す．「群青」と「旅立ちの日に」で計測された Δt のヒストグラムを，それぞれ図 4.14 と図 4.15 に示す．また，各曲の Δt の分布に関する統計量を，表 4.12 に示す．

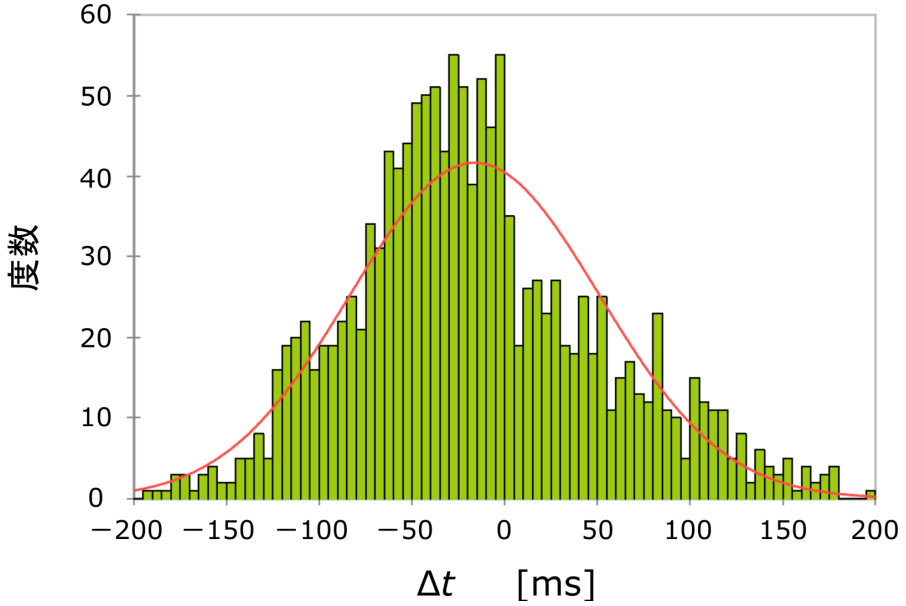

図 4.14: 「群青」における Δt のヒストグラム．赤線は参考用のガウス分布を示す．

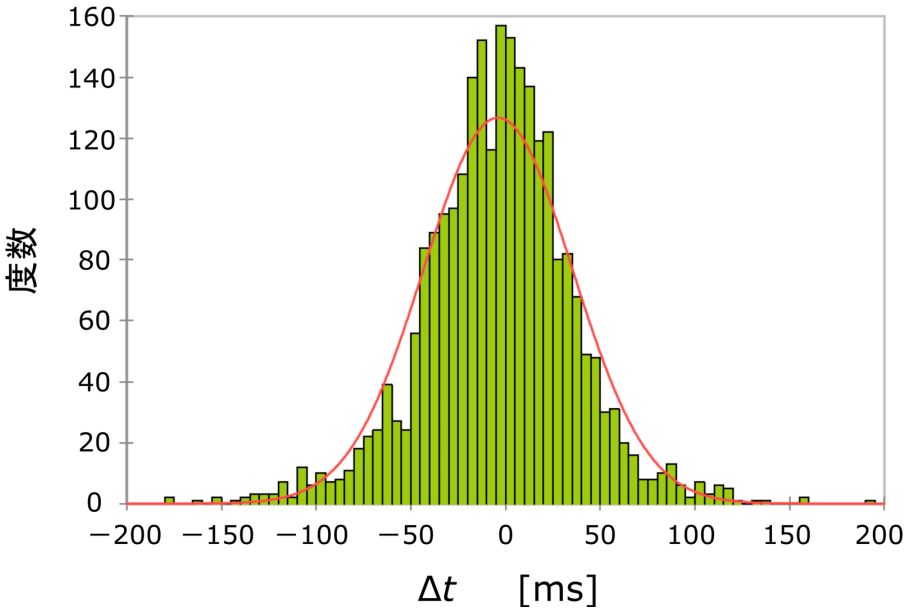

図 4.15: 「旅立ちの日に」における Δt のヒストグラム．赤線は参考用のガウス分布を示す．

表 4.12: 「群青」,「旅立ちの日に」における Δt 分布の統計量

	最小値 [ms]	最大値 [ms]	平均 [ms]	標準偏差 [ms]
群青	−194	198	−16	67
旅立ちの日に	−177	192	−4	39

以上に示したように，Δt の分布は曲ごとに違いが見られる.「群青」は「旅立ちの日に」と比較して，標準偏差の値が約 1.7 倍となっている．それだけでなく，ヒストグラムを見ると「旅立ちの日に」では分布がスーパーガウシアン分布となっているのに対し,「群青」では平均値よりも負の方向に分布が集中していることが分かる．この違いが生じる要因としては，曲の特性の違いが考えられる．表 4.9 を見ると,「旅立ちの日に」はテンポが終始 84 BPM で一定なのに対し,「群青」はテンポが曲中で 72 BPM から 79 BPM の範囲で変動している.「群青」に関しては，曲の冒頭は静かな曲想で始まり，後半に進むに連れて迫力を増していくような曲の構成になっている．また，それに合わせて伴奏のテンポも自然に増加していく．このテンポ変動は楽譜に規定のあるものではなく，演奏者の心理状態からもたらされるものと推測される．このため，歌唱者によりテンポ変動のペースが異なり，その結果テンポ一定で演奏される「旅立ちの日に」と比較して，発声タイミングずれが大きくなったことが考えられる．

最後に，歌唱者ごとに分類した Δt の分析結果を示す．ここでは,「群青」のソプラノパート 3 人で計測された Δt のヒストグラムを，それぞれ図 4.16，図 4.17，図 4.18 に示す．また，各歌唱者の Δt の分布に関する統計量を，表 4.13 に示す.

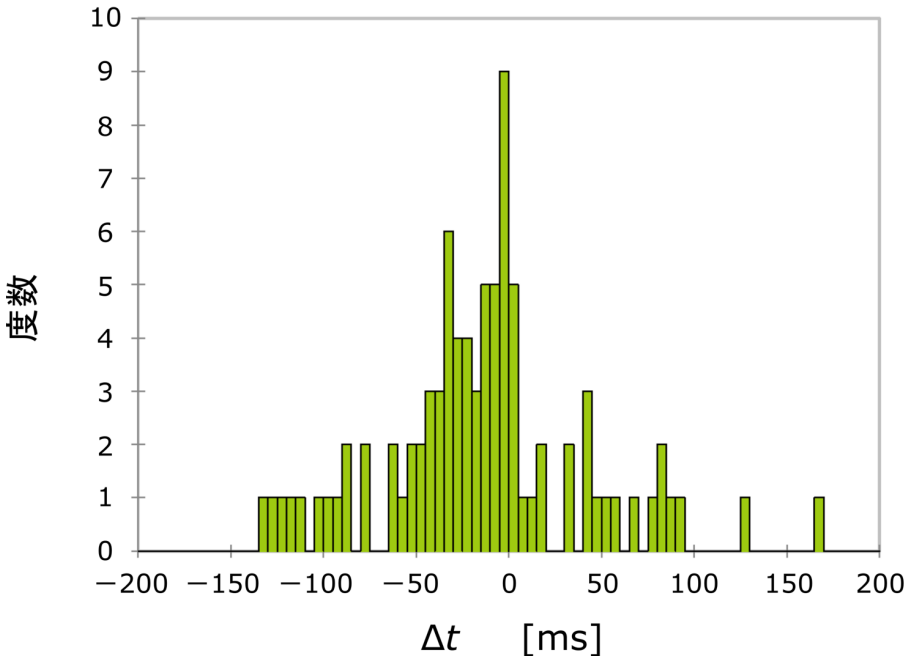

図 4.16: 「群青」のソプラノ歌唱者 1 における Δt 分布のヒストグラム

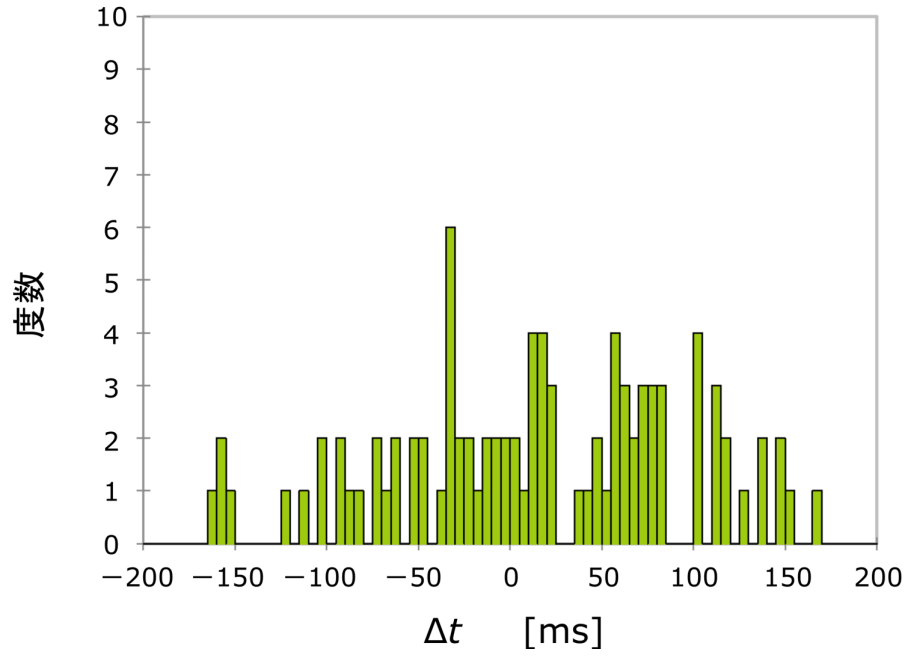

図 4.17: 「群青」のソプラノ歌唱者 2 における Δt のヒストグラム

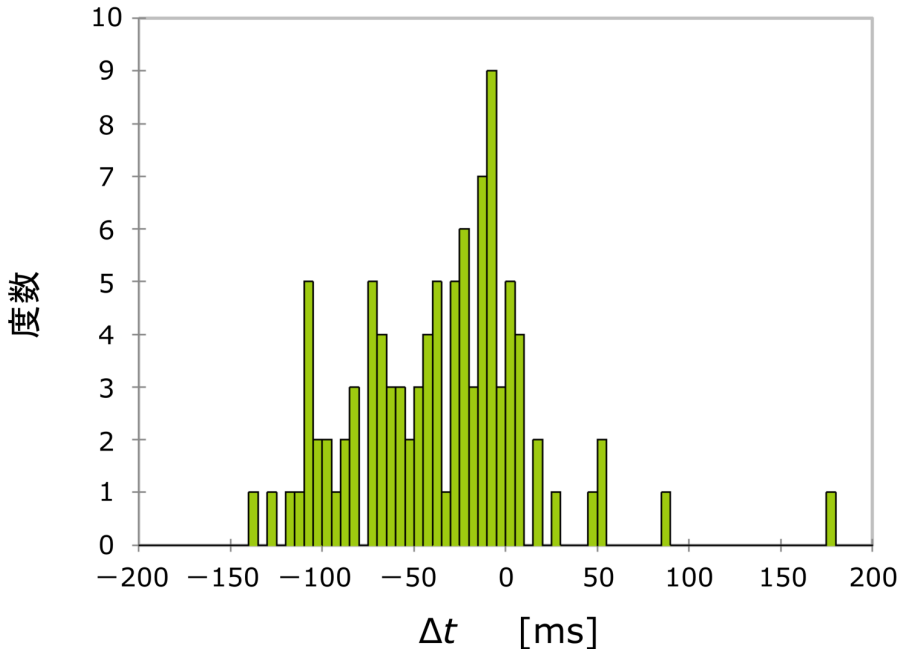

図 4.18: 「群青」のソプラノ歌唱者 3 における Δt のヒストグラム

表 4.13: 「群青」の 3 人のソプラノ歌唱者における Δt の統計量

	最小値 [ms]	最大値 [ms]	平均 [ms]	標準偏差 [ms]
歌唱者 1	−140	175	−36	48
歌唱者 2	−165	167	14	79
歌唱者 3	−133	168	−15	55
(「群青」全体)	(−194)	(198)	(−16)	(67)

以上に示したように，Δt の分布は曲ごとに違いが見られる．歌唱者 1 については，分布の標準偏差は「群青」全体よりも小さいが，標準的な歌唱よりは早めに歌う傾向があると言える．歌唱者 2 については，ガウス分布のような山型の分布ではなく，比較的平坦な分布となっている．すなわち，曲の中で早く歌ってしまう部分と遅く歌ってしまう部分が同程度に混在している．前述のように,「群青」は曲中で徐々にテンポが上昇していくため，そのテンポ変動のペースが他の歌唱者と合っていない可能性も考えられる．歌唱者 3 については，この 3 人の中で平均と標準偏差が「群青」全体の分布と最も合致している．つまり，この中では最も標準的な歌い方をしていると言える．

ここまで述べたように，3 分類の発声タイミングずれ量について分析した結果，曲と歌唱者の特性に応じて，ずれ量の分布に差違が生じることが明らかとなった．本実験では,「群青」はアマチュア合唱団員によって歌唱され,「旅立ちの日に」は非合唱団員によって歌唱された．こうした歌唱者の属性によって分布に差違が生じるかどうかは,同一の曲を属性の異なる歌唱者が歌った場合の結果を比較して判断する必要があると言える．

4.5 歌唱者由来のずれ補正の検証

4.5.1 概要

第 3.3 節で述べた提案手法を用いて，歌唱者由来のずれ補正を検証する．実験に用いた音源は，第 4.3 節で使用した「心の瞳」のソプラノ歌唱者 7 人の歌唱音源である．検証にあたっては，まず前処理として 7 音源のラウドネスを揃えた．次に，第 3.3 節で述べた手法で各音源に時間シフトを施した後，全 7 音源をミキシングした．結果の評価は，時間シフトを施す前にミキシングした音源（以降，単純加算音源）と，時間シフトを施した後にミキシングした音源（以降，補正済み音源）の聴き比べにより行なった．また，単純加算音源と補正済み音源の波形，時間周波数平面を比較することによっても評価を行なった．

4.5.2 結果と考察

単純加算音源と補正済み音源の波形と時間周波数平面を，図 4.19 に示す．

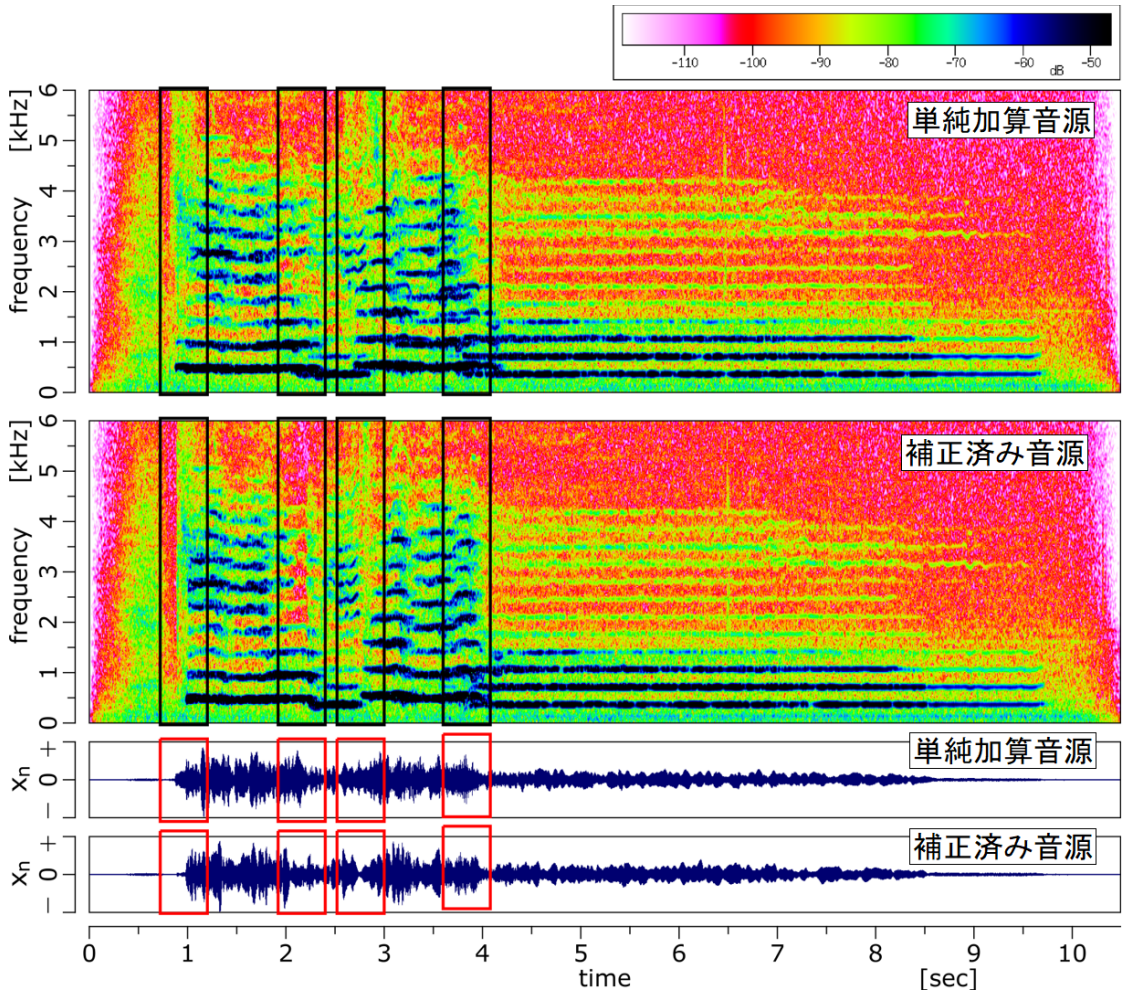

図 4.19: 上から順に，単純加算音源の時間周波数平面，補正済み音源の時間周波数平面，単純加算音源の時間波形，補正済み音源の時間波形

単純加算音源は，第 4.3 節でも述べたように，歌唱者間に最大 250 ms の発声タイミングずれが存在する．そのため，この音源を聴取すると明らかな歌唱のばらけを知覚した．一方で，補正済み音源を聴取すると，合唱作品として許容できる範囲まで発声タイミングが揃ったように感じられた．図 4.19 では，時間シフト量の変化点付近を枠で囲い示してある．この部分では，特にずれが解消されたように感じられた．時間周波数平面を比較すると，最初の変化点部分を除いて単純加算音源では周波数の異なる調波成分が重なっているのに対し，補正済み音源ではその重なりがほとんど見られなくなっている．これは，各歌唱者のオンセットが揃ったことを示している．一方で，時間波形を観察すると，補正済み音源の一部のずれ量変化点においてパワーが局所的に小さくなっている部分が存在する．これは，ずれ補正の手法に含まれるフェードアウ

ト，フェードインの処理が，複数の音源で同時に行なわれたことが原因である．この部分を聴取すると，音量感が減少し，歌唱人数が一時的に減ったように知覚される．これは，リモート合唱を作品として仕上げる場合には，好ましくない音量変化と言える．この問題を解決するためには，時間シフト量変化点に挟まれた各区間で話速変換を行い，各区間をなめらかに接続させることが有効だと考えられる．また，本手法では歌唱者間でオンセットのずれ量が 0 ms となるように補正したが，舞台上の合唱を再現する場合，ある程度のずれを残したほうがより実感的に聞こえる可能性がある．その場合，オンセット時刻の分布を，実際の合唱における分布に従わせることが理想的である．

第 5 章
応用

本研究で提案した手法を応用し，実用化する取り組みが既に行なわれている．その一環として著者は，自動ミキシングシステムの開発，並びに，リモート合唱研究に関連する対外的な活動を行なってきた．それぞれの概要と実績について述べる．

5.1 自動ミキシングシステムの開発

第 1.1.2 項で述べたように，本研究は IT エンジニアグループ「Harmorearth」[8] と共に，リモート合唱のミキシングを自動化することを目的としている．そこで，著者は本研究の一部を応用の上，リモート合唱のための自動ミキシングプログラムを作成した．このプログラムは,「Harmorearth」が運営するリモート合唱の web プラットフォームである「tuttii」[9] のサーバ上に実装され，2021 年 5 月より稼働中である（2022 年 1 月現在，システム移行に伴い休止中.）．ここで，自動ミキシングシステムの構成を，図 5.1 に示す.

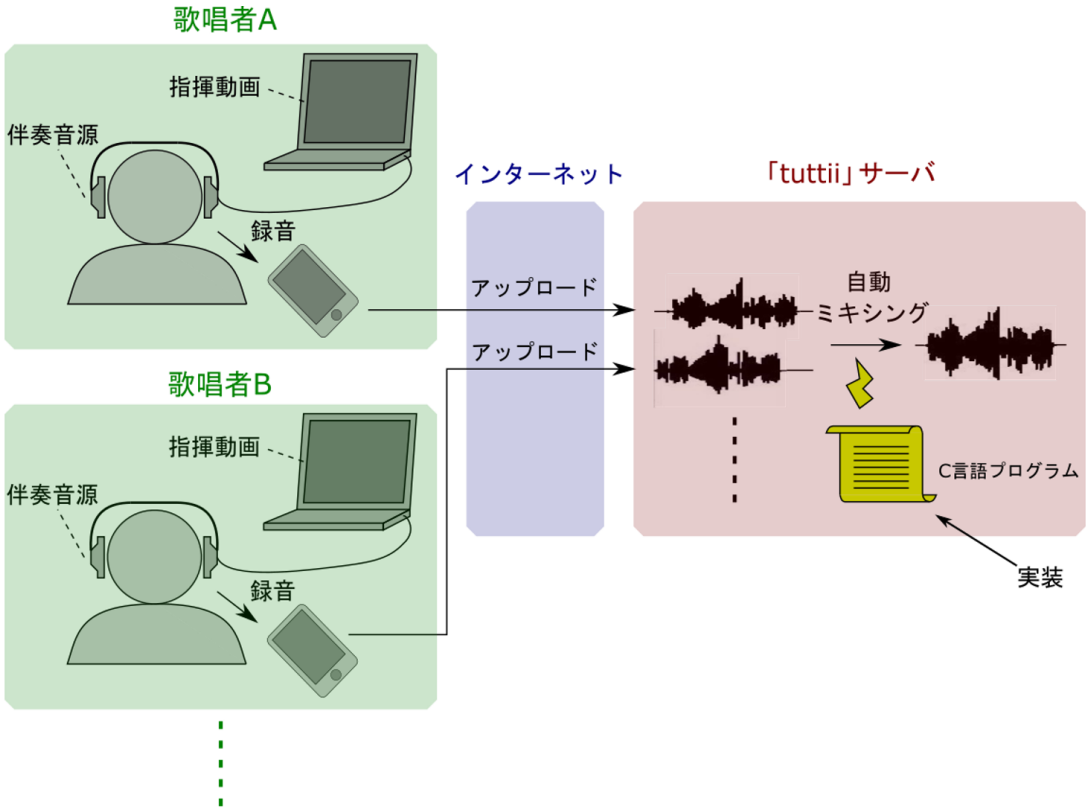

図 5.1: リモート合唱の自動ミキシングシステムの構成

2021 年 5 月以前は図 1.1 に示したように，プロのミキシングエンジニアが手動でミキシングを行なっていた．それを自動化することで，ユーザは随時音源を投稿してミキシング結果を受け取ることが可能となった．

自動ミキシングプログラムには，本研究内容の一部の他，著者の研究成果である複数の機能を組み込んだ．その機能の内容を，サーバ上には未実装のものも含めて表 5.1 に示す．

表 5.1: 自動ミキシングプログラムの機能

	サーバ上への実装	備考
システム由来のずれ補正	○	−
歌唱者由来のずれ補正	×	−
サブソニックフィルタ	○	−
パート内音量調整	○	−
パート間音量調整	○	2022 年 1 月現在，改良中
リバーブ	○	著者が所属する研究室のメンバーが作成
パン配置	○	−

本研究以外の研究成果をもとに作成した機能について，以下に述べる．

- 機能 1：サブソニックフィルタ

低域に集中するノイズは聴取時に不快感を与えるだけでなく，再生機器に無駄な負荷を与えることになるため，ミキシング時に除去しておくことが望ましい．ここでは入力信号を FFT し，約 70 Hz 以下に相当する周波数ビンのエネルギーを 0 とすることで実現している．

- 機能 2：パート内音量調整

サーバに投稿される音源は歌唱者により音量が異なるため，そのまま加算した場合には特定の歌唱者の声が目立ったり，別の歌唱者の声がマスキングされて聞こえにくいといった事態が生じる．それを防ぐため，ミキシング前に各歌唱者のラウドネスを統一する．ここでは，電波産業界（ARIB）により定められたラウドネス運用規定 [30] に則り，K 特性の聴覚補正フィルタを通した平均ラウドネスを指標として用いる．

- 機能 3：パート間音量調整

パート間音量調整は，混声合唱のように多声部で構成される合唱において，パート間の音量バランスを調整する処理である．例として，混声四部合唱であれば主旋律の多いアルトパートはソプラノパートよりも 3 dB 弱め，テノール，バスパートはアルトパー

トよりもさらに 3 dB 弱めるといったように規範を設けることもできる．しかし，本ミキシングシステムでは 20 声部までに対応するために，どのようなパート編成であってもバランスよく聞こえるような規範を定める必要がある．自動ミキシングシステムには既に仮の手法が実装されているが，2022 年 1 月現在，より適した音量バランスを実現するために新手法を構築中である．その新手法について，簡潔に述べる．

図 5.2 に，あるプロのミキシングエンジニアが混声四部合唱「彼方のノック」と混声三部合唱「心の瞳」をミキシングした際の，各パートの平均周波数とラウドネスの関係を示す．これを見ると，平均周波数が 2 倍になるごとに平均ラウドネスが 7~8 LKFS ほど上昇していることが分かる．これは，基本周波数が 2 倍の関係にある 2 音をミックスした場合，低音の高調波 2 個分のエネルギーと，高音の高調波 1 個分のエネルギーがおよそ等しいときに心地よいハーモニーが得られるという仮説を与える．この仮説がより多くのミキシング例に当てはまることが示されれば，パート間の音量バランスを決定する規範として使用できると考えられる．

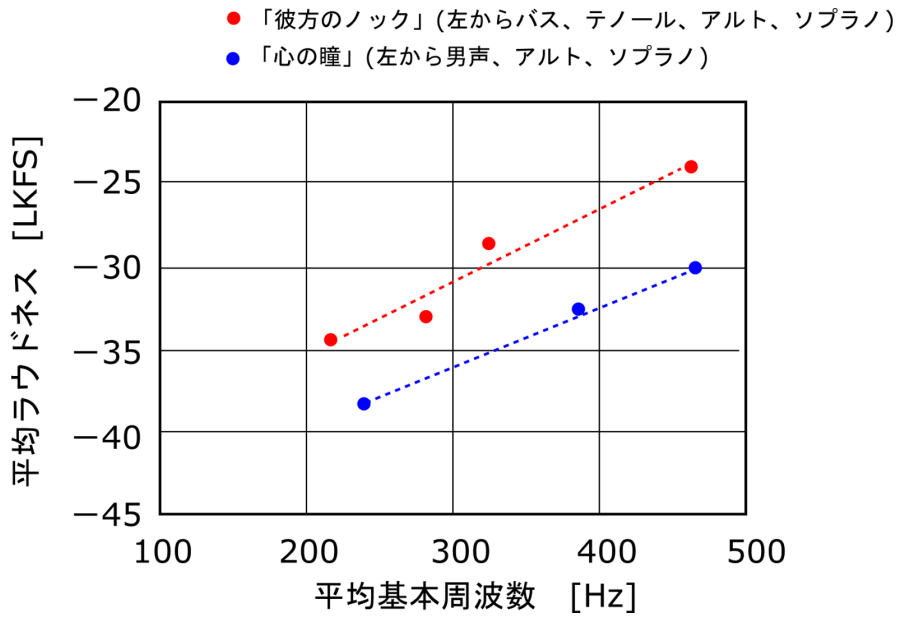

図 5.2: 各パートの平均周波数と音量バランスの関係．破線は近似直線を表す．

- 機能 4：リバーブ付加

リバーブの機能のみ，著者が所属する研究室のメンバーである中内氏の研究成果 [31] によるものである．リモート合唱は一般的な住宅内で録音されることが多く，歌唱者によって残響の付き方に差がある．そこで，ミキシング後にリバーブを付加すること

によって各歌唱者の響きの差を低減すると共に，ホールでの合唱のような空間の広がりを与えることができる．ここでは，並列コムフィルタによるリバーブエフェクタを使用する．

- 機能 5：パン配置

リモート合唱に用いる個人歌唱音源はモノラルを前提としているが，ミキシング後の出力をステレオとするため，各歌唱者にパンを振り分ける処理が必要である．一般的には強度差ステレオが用いられるが，リモート合唱では位相差ステレオを導入している．位相差ステレオは，人間の頭部を円で近似し，両耳までの音波の到達時間の差を再現するものである．そのモデル化された音波の到達経路を，図 5.3 に示す．

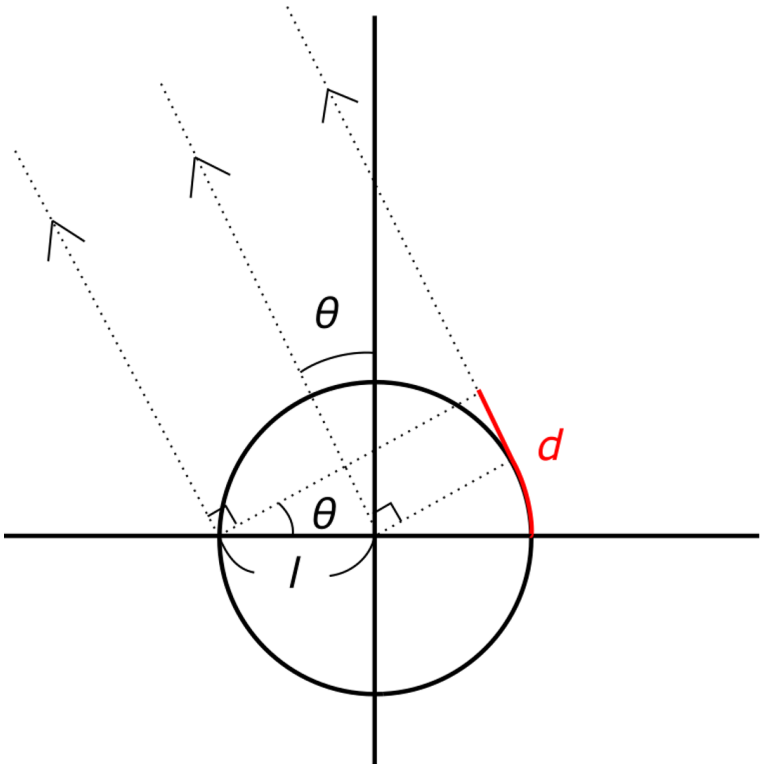

図 5.3: 位相差ステレオにおける音波の到達経路モデル

図は人間の頭部を真上から見た状態を表す．位相差ステレオにおける音波の到来方向を θ [rad] とすると，式 (5.1) と式 (5.2) より，LR チャンネル間に与えるべきサンプル差 Δs を求めることができる．

$$d = l \sin\theta + l\theta \tag{5.1}$$

$$\frac{d}{v_s} = \frac{\Delta s}{F_s} \tag{5.2}$$

ここで，v_s は音速を表す．例として，$l = 10$ cm，$F_s = 48000$ Hz，$\theta = \frac{\pi}{2}$ rad とすると，$\Delta s \simeq 36$ サンプルと求まる．

自動ミキシングプログラムでは，定位の広がりを補完するために位相差ステレオと強度差ステレオを併用している．ただし，強度差ステレオは聴取者に過剰な分離感を与えないために，強度差を従来の 30 %に抑えて組み合わせている．

5.2 リモート合唱研究の対外的な活動

著者はこれまで，新型コロナ流行下において通常の合唱が困難な団体から依頼を受け，研究成果を生かしたリモート合唱のミキシングを手がけてきた．依頼を受けた団体は多岐にわたり，アマチュア合唱団や学校，さらにはラジオ局主催の音楽イベント等にもリモート合唱作品を納めてきた．このほか，リモート合唱の研究は，新聞やテレビ等のメディアにも取り上げられた経歴を持つ．そこで，これまでの対外的な活動の実績を，表 5.2 に年表形式で示す．

表 5.2: リモート合唱研究の主な対外活動実績

日付	活動	楽曲	人数 [人]
2020/06/30	朝日新聞にて紹介	—	—
2020/07/26	公明新聞にて紹介	—	—
2020/09/09	テレビ東京「WBS」にて紹介	「恋音と雨空」	11
2020/09/20	「Harmorearth」「機械 vs 人間」プロジェクト	「心の瞳」	18
2020/09/24	エイベックス「合唱のアソビバ」	「はじまりのとき」	6
2021/01/02	FM 福岡 ジングルミキシング	（ジングル）	16
2021/03/29	「Harmorearth」初演プロジェクト	「音楽である私」	41
		「音楽がぼくに囁いた」	38
2021/03/31	YouTube チャンネル「あされん」コラボ	「旅立ちの日に」	33
2021/03/31	FM 福岡「50 周年記念特番」	「時代」	170
2021/04/20	中学生によるリモート吹奏楽	「ルパン三世」	21
		「美女と野獣」	26
		「宝島」	26
2021/06/15	「デジタル TEPIA」出展	「大地讃頌」	5
2021/08/17	読売新聞にて紹介	—	—
2021/10/31	FM 福岡「九州ゴスペルフェスティバル 2021 in 博多」	「OH HAPPY DAY」	64

各活動の内容を簡潔に述べる．2020 年 6 月 30 日には，朝日新聞にて「コロナウイルスが分断した合唱のハーモニー よみがえれ IT の力で」[32] と題して，同年 7 月 26 日には，公明新聞にて「コロナ禍でも合唱を楽しむ」と題してリモート合唱が紹介された．同年 9 月 9 日には，テレビ東京系列の番組「ワールドビジネスサテライト（WBS）」の

コーナー「トレンドたまご」にて，リモート合唱が取り上げられた他，アナウンサーらが歌うリモート合唱「恋音と雨空」のミキシングを担当し，放送された．この際，スタジオで録音した音源を著者の自宅の計算機に転送し，収録終了までにミキシングして送信するという緊迫した作業を担当した．同月 20 日には，自動ミキシングとプロのミキシングを判別できるかというアンケートが,「Harmorearth」の Twitter 上で行なわれた．著者はこれに際し,「心の瞳」の自動ミキシングを担当した．結果として，回答者の 56 %が，自動ミキシングによる作品をプロのミキシングによる作品と回答した．さらに同月 24 日には．エイベックス・エンタテインメント株式会社主催の音楽イベント,「合唱のアソビバ」に,「はじまりのとき」のミキシング作品を提供した．

2021 年 1 月 2 日に放送された FM 福岡のラジオ番組コーナー「おもろい家族」では，リモートで行なうジングルのミキシングを担当した．同年 3 月 29 日には,「Harmorearth」主催のプロジェクトとして，リモート合唱用に作曲された「音楽である私」,「音楽がぼくに囁いた」の初演ミキシングを担当した．同月 31 日に公開された，合唱系 YouTuber「あされん」とのコラボ動画では,「旅立ちの日に」のミキシングを担当した．さらに同日放送された FM 福岡の 50 周年記念特番では,「時代」[33] のミキシングを担当した．ラジオリスナーやアナウンサーなど，総勢 170 名の歌声をミキシングし，過去最大規模のリモート合唱となった．同年 4 月 20 日には，合唱と同じく活動が困難な吹奏楽部の中学生らを支援するため，初の器楽曲となる「ルパン三世」,「美女と野獣」,「宝島」の吹奏楽をミキシングした．同年 6 月 15 日には，先端技術のバーチャル・ミュージアムである「デジタル TEPIA」[34] に出展し，技術紹介を行なうと共に,「大地讃頌」のミキシング作品を提供した．2021 年 8 月 17 日には，読売新聞にて「コロナ禍 合唱の喜び絶やさず」[35] と題してリモート合唱が紹介された．同年 10 月 31 日に開催された音楽イベント「九州ゴスペルフェスティバル 2021 in 博多」にて，ゴスペル曲「OH HAPPY DAY」[36] のミキシングを担当した．

この他にも，アマチュア合唱団から依頼された「Ave Maria」,「雨」,「群青」,「ほたるこい」のミキシングを担当した．以上を合計して，これまでに 17 作品，539 音源に自動ミキシングを適用してきた．また，著者は，高橋弘大研究室のホームページ上 [37] に研究用音楽データベース「SIS-DB」を作成中である．2022 年 1 月現在は未公開であるが,「九州ゴスペルフェスティバル 2021 in 博多」におけるミキシングの際に提供された個人歌唱音源について，歌唱者の了承を得てデータベースに掲載する予定である．

第 6 章
リモート合唱研究の展望

リモート合唱研究は，本研究を発端として今後さらなる発展を遂げることが期待される．その展開としては，主に 2 種類に大別される．第一に，舞台上の合唱をより良く再現すること．第二に，舞台上では成し得ない合唱「超合唱」の実現に向けた取り組みである．この章では，それぞれの研究の方向性について述べる．

6.1 舞台上の合唱の再現

リモート合唱を舞台上の合唱に近づける上で，自動ミキシングのさらなる改良の余地が存在する．そこで，本研究の主題とは異なる内容について，今後期待される研究内容をまとめる．

- 研究内容 1：リバーブ除去

現在の自動ミキシングでは，歌唱者ごとに異なる残響の付き方に対して，同一のリバーブを付加することで違和感の低減を行なっている．多くの場合はこれによって良好な結果が得られるが，極端に残響の多い場所で録音した音源や，既にリバーブエフェクタがかかってしまっている音源に対しては，残響が付きすぎてしまう．そこで，既に付加されている残響を除去するか，あるいは音源に応じて適した残響を付加する，適応型リバーブエフェクタの研究が必要とされる．

- 研究内容 2：AGC 補正

オートゲインコントロール（AGC）とは，録音レベルに応じてマイク感度を自動で調整する，録音端末の機能である．これを使用してリモート合唱を録音した場合，休符の後の出だしが突出して大音量で録音されるため，ミキシング後に違和感が生じてしまう，これを補正するためには，細かい分析フレームで区切ってラウドネスを計測し，各区間で適性な音量となるよう経時的にゲインを変化させる手法が有効だと考えられる．

- 研究内容 3：歌詞誤りの検出

リモート合唱で投稿される音源の中には，歌詞を間違えて歌唱しているものや,「あー！間違えた！」などといった歌詞とは関係のない声が録音されてしまっている場合がある．このような歌詞誤りを検知する手法として，隠れマルコフモデル（HMM）を用いた手法 [38] が提案されている．歌詞誤りを検出した場合には，問題のある部分をミュートするか，その音源自体をミキシング対称から除外するといった措置を取ることが望ましい．

- 研究内容 4：語尾を揃える処理

本研究では歌詞の語頭を合わせることに着目したが，歌詞の語尾が揃いにくいこともリモート合唱ならではの問題である．日本語のように必ず母音で終わる言語であれば，各歌唱者のエネルギーを揃えることで語尾が揃う．しかし，語尾が子音で終わることもある外国語では，その子音を合わせるための手立てが必要である．これについては後述する時間周波数平面での処理が有効である可能性がある．

6.2 超合唱の実現

これまでは舞台上の合唱を再現することをリモート合唱の目標としていたが，その目標は達成されつつある．今後リモート合唱を発展させる上では，現実では不可能な合唱「超合唱」の実現に向けた取り組みが期待される．ここでは，そのための研究アイデアを提案する．

- 研究内容 1：時間周波数平面上の処理

これまでは，時間領域での処理が主体であったが，時間周波数平面上での信号操作を導入すれば，さらに自由度が高まる．例えば，全ての歌唱者のピッチをコントロールし，最も心地よいハーモニーとなるよう周波数シフトを施すことが挙げられる．あるいは，時間周波数平面を二次元画像として捉え，各歌唱者の語頭や語尾，ピッチまでもを局所的に一致させることができれば，これまで聞いたことのない理想的な合唱が実現できる可能性がある．

- 研究内容 2：コーラスエフェクタ

現状では，歌唱録音を行なった人数が，そのままリモート合唱の規模として反映されている．そこで，実際の歌唱人数以上の規模に聞こえるような処理が可能になれば，リモート合唱の価値がさらに高まると言える．通常，再現度の高いコーラスエフェクタを作成する場合，合奏において各演奏者からクロストークのない演奏音を録音し，解析の上でモデル化する [13]．舞台上の合唱においてクロストークのない歌唱録音を得るのは困難であったが，リモート合唱であればセパレート音源が揃っているので，実現可能性は十分にあると考えられる．

第 7 章
結論と課題

本研究の結論と，今後の課題について述べる．

7.1 結論

本研究は，リモート合唱のミキシングを自動化するための研究に位置付けられる．本研究の最初の成果は，リモート合唱においてはシステム由来のずれと歌唱者由来のずれという，2 種類の発声タイミングずれを定義したことである．その上で，ミキシング自動化のために 3 点の目的を設定した．第一に，システム由来のずれを自動的に解消する手法を提案すること．第二に，歌唱者由来のずれを定量的に分析すること．第三に，歌唱者由来のずれを自動的に補正する手法を提案することである．これらの目的を達成するために，各手法の提案と評価を行なった．

第 1 章では，リモート合唱に関する背景を述べるとともに，本研究の必要性について論じた．第 2 章では，本研究と密接に関連する内容である音楽における同時性，またそれを分析するために必要となる，オンセット検出手法の先行研究について述べた．第 3 章では，本研究の 3 点の目的を達成するために必要となる手法を提案した．このうち，システム由来のずれを解消する手法として，マーク信号を用いた手法を提案した，一方で，歌唱者由来のずれを分析，補正するための手法として，オンセット検出

の FN 法，標準オンセットの設定法，オンセットマッチングの手法を提案した．特に，オンセットマッチングには DP マッチングを導入し，従来の時間閾値でマッチングを行なう手法よりも精度の向上を図った．第 4 章では，第 3 章で提案した各手法の精度を検証する実験と，その結果の考察について述べた．実験は 5 項目を実施した．第一に，システム由来のずれを解消する手法の時間精度を評価し，その精度は実用上十分であることが示された．第二に，FN 法によるオンセット検出の精度を評価し，本手法は有声音の検出に有効であることが示された．第三に，オンセットマッチングの精度を評価し，歌唱者由来のずれを計測する際に問題となるマッチングエラーの割合は 2.6 %と低く，計測結果の信頼性が示された．第四に，歌唱者由来のずれ量を計測し，2 曲全体，曲ごと，歌唱者ごとの 3 分類について分布を分析した．その結果，全体として歌唱者由来のずれ量は平均 −8 ms，標準偏差 51 ms のスーパーガウシアン分布となった．これは舞台上での合唱よりもずれ量が大きい可能性があり，補正の必要性が示された．また，ずれ量の分布は曲ごと，歌唱者ごとに差違があることが判明した．第五に，歌唱者由来のずれを補正する手法を適用し，適用前に知覚された大きなずれが解消したことが示された．しかし，処理を加える部分で音量感が変化してしまう問題が生じることが明らかとなった．第 5 章では，リモート合唱研究の応用として，自動ミキシングシステムの概要と対外的な活動実績について述べた．第 6 章では，今後のリモート合唱研究の方向性や解決すべき問題を取り上げた．

以上をまとめると，本研究の成果は 4 点存在する．第一に，リモート合唱におけるシステム由来のずれと歌唱者由来のずれを定義したこと．第二に，システム由来のずれを解消する手法を提案し，その有効性が示せたこと．第三に，歌唱者由来のずれの計測手法を提案し，実験にて計測した結果，ずれを補正する必要性を示せたこと．第四に，歌唱者由来のずれを補正する手法を提案し，その有効性を示せたことである．

7.2 今後の課題

今後の課題は，歌唱者由来のずれを補正する手法に話速変換を導入し，ずれ量変化点において信号をなめらかに接続させることである．既存の話速変換アルゴリズムには，位相ボコーダ [39] や WSOLA[40] といった代表的な手法が存在する．しかし，こうした典型的なアルゴリズムを使用した場合，位相が不連続になることによるアーティ

ファクトの発生が避けられず，聴感上の違和感を生じる恐れがある．それを防ぐ手法として，ハーモニック・パーカッシブ（HP）分離を用いた手法 [41] が提案されているが，歌声の場合には音響的性質の複雑さが原因で，その効果は限定的であるとの指摘がある．この問題を解決できる手法として，歌声の調波成分を擬似正弦波に置き換え，位相が不連続とならないように調整しながら話速変換を行なうことが考えられる．著者の所属する研究室では，こうした擬似正弦波での信号表現として「跡」[42] が提案されている．今後は,「跡」を用いて聴感上自然な話速変換を行なう手法を構築し，歌唱者由来のずれ補正に導入することが課題となる．

参考文献

[1] DPG Media Group,“ Die ene Passion die wel doorging, met rampzalige gevolgen, ”https://www.trouw.nl/verdieping/die-ene-passion-die-wel-doorging-met-rampzalige-gevolgen~b4ced33e/, Trouw, 2022/01/06 閲覧.

[2] 戸ノ下達也・横山琢哉,“ 日本の合唱史, ”青弓社, 2011.

[3] 朝日新聞社,“「合唱クラスター」で中学生ら5人感染　福島・郡山, ”https://www.asahi.com/articles/ASN8N760WN8NUGTB00T.html, 朝日新聞 DIGITAL, 2022/01/06 閲覧.

[4] ヤマハ株式会社,“ SYNCROOM, ”https://syncroom.yamaha.com/, YAMAHA, 2021/07/21 閲覧.

[5] 株式会社第一興商,“ DAM ともヘルプコラボ録音, ”https://www.clubdam.com/app/damtomo/common/support.do?type=damtomo&source=recording_collaboration&subType=support, DAM とも, 2021/06/06 閲覧.

[6] 株式会社第一興商,“ DAM ともヘルプコラボ動画, ”https://www.clubdam.com/app/damtomo/common/support.do?type=damtomo&source=movie_collaboration&subType=support, DAM とも, 2021/06/06 閲覧.

[7] 全日本合唱連盟,“ 第72回全日本合唱コンクール全国大会 大学職場一般部門 出演スケジュール（11/23）, ”https://jcanet.or.jp/event/concour/72kyoto_schedule.pdf, 全日本合唱連盟 Japan Choral Association, 2021/01/06 閲覧.

[8] Harmorearth,“ Harmorearth TOP, ”https://www.harmorearth.com/, Harmorearth, 2022/01/06 閲覧.

[9] Harmorearth,“ tuttii リモート 合唱 プラットフォーム, ”http://tuttii.harmorearth.com/, tuttii, 2021/07/21 閲覧.

[10] Vivienne M. Young, Andrew M. Colman,“ Some psychological processes in string quartets, ”Psychology of Music Psychology of Music, vol.7, pp.12-18, 1979.

[11] 河瀬 諭,“ 合奏における演奏者間コミュニケーション, ”心理学評論, vol.57, pp.495-510, 2014.

[12] Peter E. Keller, Mirjam Appel,“ Individual differences, auditory imagery, and the coordination of body movements and sounds in musical ensembles, ”Music Perception, vol.28, pp.27-46, 2010.

[13] J. Patynen, S. Tervo and T. Lokki,“ Simulation of the violin ection ound ba ed on the analn i of orche tra performance, ” IEEE WAS-PAA, pp.173-176, 2011.

[14] 的場達也, 馬場隆, 成山隆一, 松本秀一, 森瀬将雅, 片寄晴弘,“ 歌唱のグルーブ感の構成要因の分析, ”情報処理学会研究報告, Vol.2014-MUS-102, No.12, 2014.

[15] Juan Pablo Bello, Laurent Daudet, Samer Abdallah, Chris Duxbury, Mike Davies, and Mark B Sandler,“ A tutorial on onset detection in music signals. ”IEEE Transactions on speech and audio processing, Vol.13, No.5, pp.1035-1047, 2005.

[16] 島村楽器,“ アナログシンセ超入門〜その 2：EG（エンベロープジェネレーター、ADSR）編, ”https://info.shimamura.co.jp/digital/guide/2018/02/122103, Digiland, 2022/01/27 閲覧.

[17] 森勢将雅, 河原英紀, 西浦敬信,“ 基本波検出に基づく高 SNR の音声を対象とした高速な F0 推定法 , IEICE Trans., D. 情報・システム, Vol.93(2), pp.109-117, 2010.

[18] 中村尚五,“ デジタルフーリエ変換 ”, 東京電機大学出版局, 1989.

[19] 姫野伴子, 小森和子, 柳澤絵美,“ 日本語教育学入門 ”, 研究社, 2015.

[20] 中尾睦彦, 岸本昌也, 濱田憲治,“ 子音/ｔ/をもつ日本語音声の合成, 明石工業高等専門学校研究紀要, Vol.48, pp.11-18, 2005.

[21] 東栄伸,“ 歌唱音源における周波数ごとのオンセット時間の差異の傾向と歌唱への影響,“ 電気通信大学中間発表予稿, 2021.

[22] 楽譜作成ドットコム,“ 混声合唱のパートと音域, ”https://gakufu-ya.com/oniki, 2022/01/17, 楽譜作成.com, 2022/01/17 閲覧.

[23] 小林音楽教室,“ 6 種類の声域と特徴について，”https://www.kobayashi-music.com/tips/seiiki/, 小林音楽教室，2022/01/17 閲覧.

[24] Toshio Iwata,“ デジタルフィルタアナライザ, ”http://digitalfilter.com/products/dfalz/jpdfalz.html, DIGITALFILTER.COM, 2022/01/17 閲覧.

[25] 内田誠一,“ DP マッチング概説 〜基本と様々な拡張〜, ”信学技報, PRMU2006-166, 2006.

[26] 山田真司, 三浦雅展, “ 音楽情報処理で用いられる音響パラメータによる音楽理解の可能性, ”日本音響学会誌, vol.70(8), pp.440-445, 2014.

[27] Fabrice Bellard,“ FFmpeg, ”https://www.ffmpeg.org/, FFmpeg, 2022/01/22 閲覧.

[28] 日本パルスモーター株式会社,“ LSI 製品の基準クロック精度について ”,https://www.pulsemotor.com/, NPM Japan, 2022/01/22 閲覧.

[29] 株式会社ディーアンドエムホールディングス,“ Bluetooth, その音質と遅延について ”,DENON Official Blog, https://www.denon.jp/ja-jp/blog/3853/index.html, 2022/01/22 閲覧.

[30] 電波産業会,“ デジタルテレビ放送番組におけるラウドネス運用規定, ”ARIB TR-B32, 2011.

[31] 中内優,“ 音源に適応した残響レベルの自動調整法, ”電気通信大学卒業論文,2020.

[32] 朝日新聞社,“ コロナウイルスが分断した合唱のハーモニー よみがえれ IT の力で, ”https://www.asahi.com/articles/ASN6Z5SHRN6YULZU00P.html, 朝日新聞 DIGITAL, 2022/01/27 閲覧.

[33] FM 福岡,“ 開局 50 周年ファイナル“ しみうた ”特番 最後を飾ったリモート合唱の感動がふたたび！「時代」オリジナル動画が完成 !! , ”https://fmfukuoka.co.jp/topics/530.html,FM FUKUOKA, 2022/01/22 閲覧.

[34] 一般財団法人 高度技術社会推進協会,“ TEPIA TOP, ”https://digital-tepia.com/, TEPIA 先端技術館 Digital, 2022/01/22 閲覧.

[35] 読売新聞社,“ [関心アリ！] コロナ禍　合唱の喜び絶やさず…動画で口の形指導　リモート技術活用, ”https://www.yomiuri.co.jp/life/20210816-OYT8T50121/, 読売新聞オンライン, 2022/01/22 閲覧.

[36] 九州ゴスペルフェスティバル in 博多 実行委員会,“ 九州ゴスペルフェスティバル in 博多, ”https://fmfukuoka.co.jp/gospel/, 九州ゴスペルフェスティバル in 博多, 2022/01/22 閲覧.

[37] 高橋弘太研究室,“ トップページ, ”http://www.it.cei.uec.ac.jp/, 電気通信大学 情報･ネットワーク工学専攻/II類 電子情報学プログラム 高橋弘太研究室, 2022/01/22 閲覧.

[38] W. Tsai, V. Tran and S. Kung,“ Automatic Detection of Mispronounced Lyrics in Singing, ”2019 International Conference on Machine Learning and Cybernetics (ICMLC), pp.1-5, 2019.

[39] J. L. Flanagan and R. M. Golden,“ Phase vocode, ”Bell Syst. Tech. l J., vol.45, pp.1493-1509, 1966.

[40] W. Verhelst and M. Roelands,“ An overlap-add technique based on waveform similarity (WSOLA) for high quality time-scale modification of speech, ”in Proc. IEEE International Conf. Acoustics, Speech, and Signal Processing (ICASSP), 1993.

[41] J. Driedger, M. Muller and S. Ewert,“ Improving Time-Scale Modification of Music Signals Using Harmonic-Percussive Separation, ”in IEEE Signal Processing Letters, Vol.21, No.1, pp.105-109, 2014.

[42] 山田諒太郎,“ 擬似複素正弦波による信号表現を用いた新しい音混合法, ”電気通信大学卒業論文, 2019.

謝辞

本研究を進める上で，たくさんの方々にお世話になりました．

まず，指導教員の高橋弘太先生には，常日頃から大変お世話になりました．リモート合唱研究については，私には思いもよらないようなアイデアや的確なアドバイスを頂戴し，本研究を進める上で大きな助けとなりました．また，研究以外の日常生活についても気にかけて下さり，リモート時代にあっても安心して研究に打ち込める環境を作って下さいました．ありがとうございました．

続いて，研究室の学生の方々に感謝申し上げます．今井秀氏は，自身の音楽経験を活かしてハウリングの研究に専念する姿や，研究室独自のソフトウェアである tf の mac 移植に際して粘り強く問題解決にあたる姿が印象的で，感銘を受けました．中内優氏は，ソフトからハードまで幅広く勉強を進めようという貪欲さが，研究室メンバー全体のモチベーション向上につながっているように思われました．また，リモート合唱の自動ミキシングプログラムに，自作の素晴らしいリバーブプログラムを提供して下さり，大変助かりました．西谷悠人氏は，豊富な知識と経験で計算機管理やハード関連の問題を解決に導いて下さり，研究を進める上で大きな助けとなりました．村嵜啓介氏は，音楽に対する情熱や研究に対する独創的なアイデアを持っておられ，私にとっては常日頃から良い刺激を頂ける存在でした．太田晴紀氏は，豊富な知識と確かなプログラミング技術で，西谷氏と共に計算機間系のトラブルを迅速に解決に導いて下さいました．東栄伸氏は，素早い習得力で短期間のうちに確かな研究成果を獲得され，同じくリモート合唱を研究する身として，モチベーションを大いに高めて下さいました．水谷達也氏は，難しい研究テーマでありながらも試行錯誤して研究を進めていく姿が，私自身も研究を進めていく上での模範となりました．

研究室以外でも，多くの方々のお力添えを頂きました．

大澤暁様，音森一輝様を始めとする「Harmorearth」の方々には，感謝してもしきれません．本研究を開始するきっかけを与えて下さり，さらには多数の貴重なデータの提供，並びに広報活動にも取り組んで下さいました．「Harmorearth」の方々がいらっしゃらなければ，この研究は存在しなかったと言って間違いありません．心より，感謝申し上げます，

FM 福岡の皆様，八千代市立村上中学校吹奏楽部の皆様，慶應義塾大学混声合唱団楽友会の皆様，九州産業大学グリークラブの皆様，電気通信大学グリークラブの皆様，中村学園大学クリスタル・ハーモニーの皆様，九大混声合唱団の皆様，合唱団瑠衣の皆様，女声合唱団さはほたるの皆様，混声合唱団歌バカの皆様，福岡シンフォニック合唱団の皆様はじめ，私が作成した自動ミキシングシステムを利用して下さった全ての方々に感謝申し上げます．リモート合唱という新技術にご賛同いただき，実践的な研究活動の機会を与えて頂いたことで，本研究を進める大きな力となりました．

最後に，家族に心から感謝します．学域時代から 6 年間に渡って大学生活を支えて下さり，生活には何一つ不自由なく過ごすことができました．家族の支えなしには，この研究も成し得ませんでした．

ご支援頂いた皆様に，重ねて感謝申し上げます．ありがとうございました．

発表実績

(1) 五反田聖矢, 大澤暁, 音森一輝, 高橋弘太,“ リモート合唱の音混合における音声時間軸の自動調整と発声タイミングずれの評価, ”音楽音響研究会 2021 年 8 月研究会, 2021.