スペシャルセッション 音声B/音声A/聴覚
[ここまで来た声質変換技術 ー実用可能性の視点からの現状認識と将来展望ー]

歌声インタフェース:
歌声を対象とした信号処理と それに基づくインタフェース構築

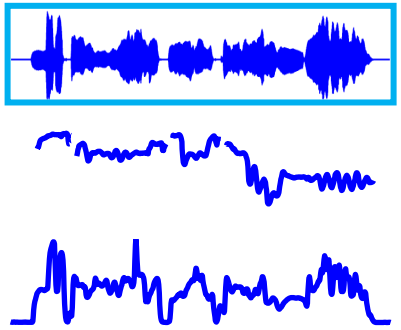
中野 倫靖，後藤 真孝
（産業技術総合研究所）

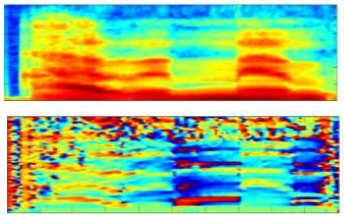
2013年9月27日
日本音響学会 2013年秋期研究発表会（講演番号3-7-3）

ここまで来た声質変換技術 ー実用可能性の視点からの現状認識と将来展望ー

歌声インタフェースとは

- 歌声信号処理に基づくインタフェース構築やインタラクションによって人々の音楽生活をより豊かにする研究アプローチ





実現

必要性

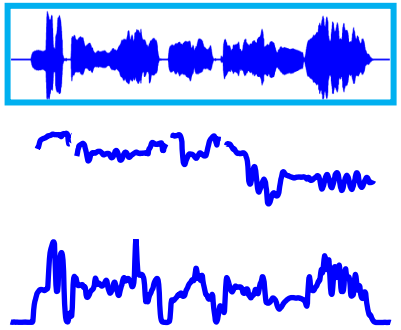
歌声信号処理

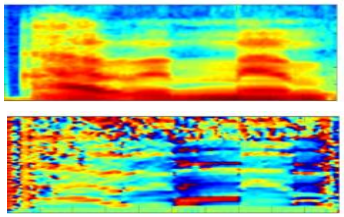
インタフェース構築（インタラクション）

歌声インタフェースとは

- 歌声信号処理に基づくインタフェース構築やインタラクションによって人々の音楽生活をより豊かにする研究

有機的な融合





実現

必要性

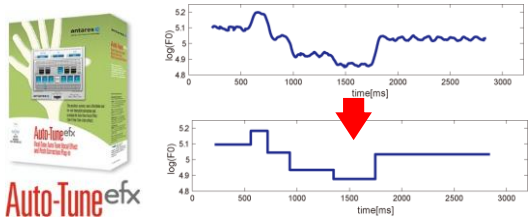
歌声信号処理

インタフェース構築（インタラクション）

ここまで来た声質変換技術 ―実用可能性の視点からの現状認識と将来展望―

インタフェース構築・インタラクションデザイン

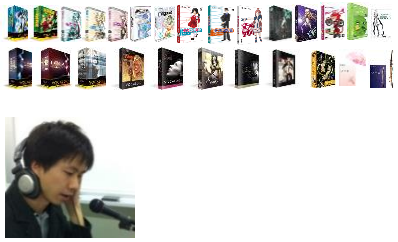
- 様々なユーザに技術を使ってもらうため
  - 対象ユーザの特性に合わせたインタフェース構築
    - ユーザ視点での問題発見（インタラクションデザイン）
- 歌声インタフェースの現状
  - プロ向けのインタフェースは既に多く存在
    - 音高補正ツールAuto-tune 等



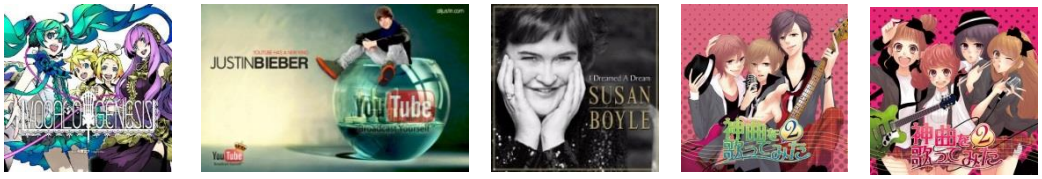
  - エンドユーザ向けのインタフェースは少ない
    - 一方で、近年はエンドユーザが気軽に歌を発表したり、それを聴いて楽しんだりする文化が広がっている

エンドユーザによる歌声の公開

- エンドユーザによるWeb上の歌声
  - 歌声合成（VOCALOID）
  - 歌ってみた / Me Singing
- 多くの視聴者もいる



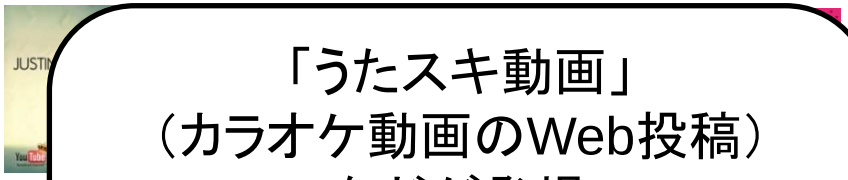
- エビデンス
  - 音楽CDの販売
  - ニコニコ動画
    - 27万曲+ (VOCALOID楽曲）64万曲+（歌ってみた)
    - 10万回+ 再生 (それぞれ3200曲+, 4,650曲+）
    - 100万回+ 再生 (それぞれ170曲+, 200曲+)



エンドユーザによる歌声の公開

- エンドユーザによるWeb上の歌声
  - 歌声合成（VOCALOID）
  - 歌ってみた / Me Singing
- 多くの視聴者もいる

- エビデンス
  - 音楽CDの販売
  - ニコニコ動画
    - 27万曲+ (VOCALOID楽曲) 64万
    - 10万回+ 再生 (それぞれ3200曲+
    - 100万回+ 再生 (それぞれ170曲+



「うたスキ動画」
（カラオケ動画のWeb投稿）
などが登場
http://utadou.jp/

カラオケユーザへ**新たな体験を提供する「場」**としての機能

歌声信号処理の全体像

歌声生成（人間）

歌声生成（機械）

歌声変換（ピッチシフトなど）・声質変換

歌声合成

歌声合成（歌声入力による合成）

自動作曲

表情付け

歌声分析

三要素の推定（音高、音量、音色）

歌手の特性（声質、性別、声種等）

歌い方の特性（声区、音高軌跡等）

楽曲の特性（歌詞、ジャンル等）

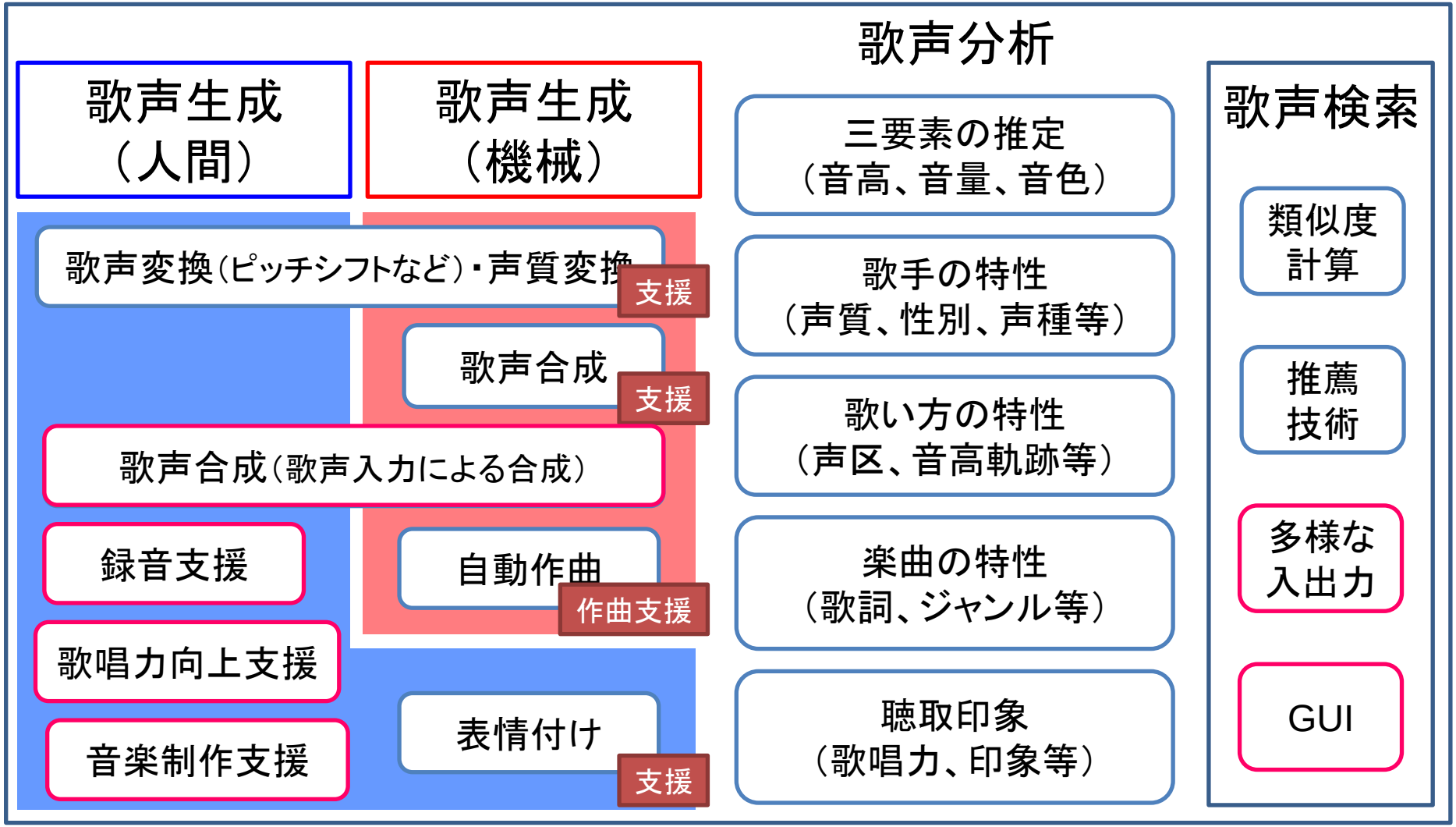
聴取印象（歌唱力、印象等）

歌声検索

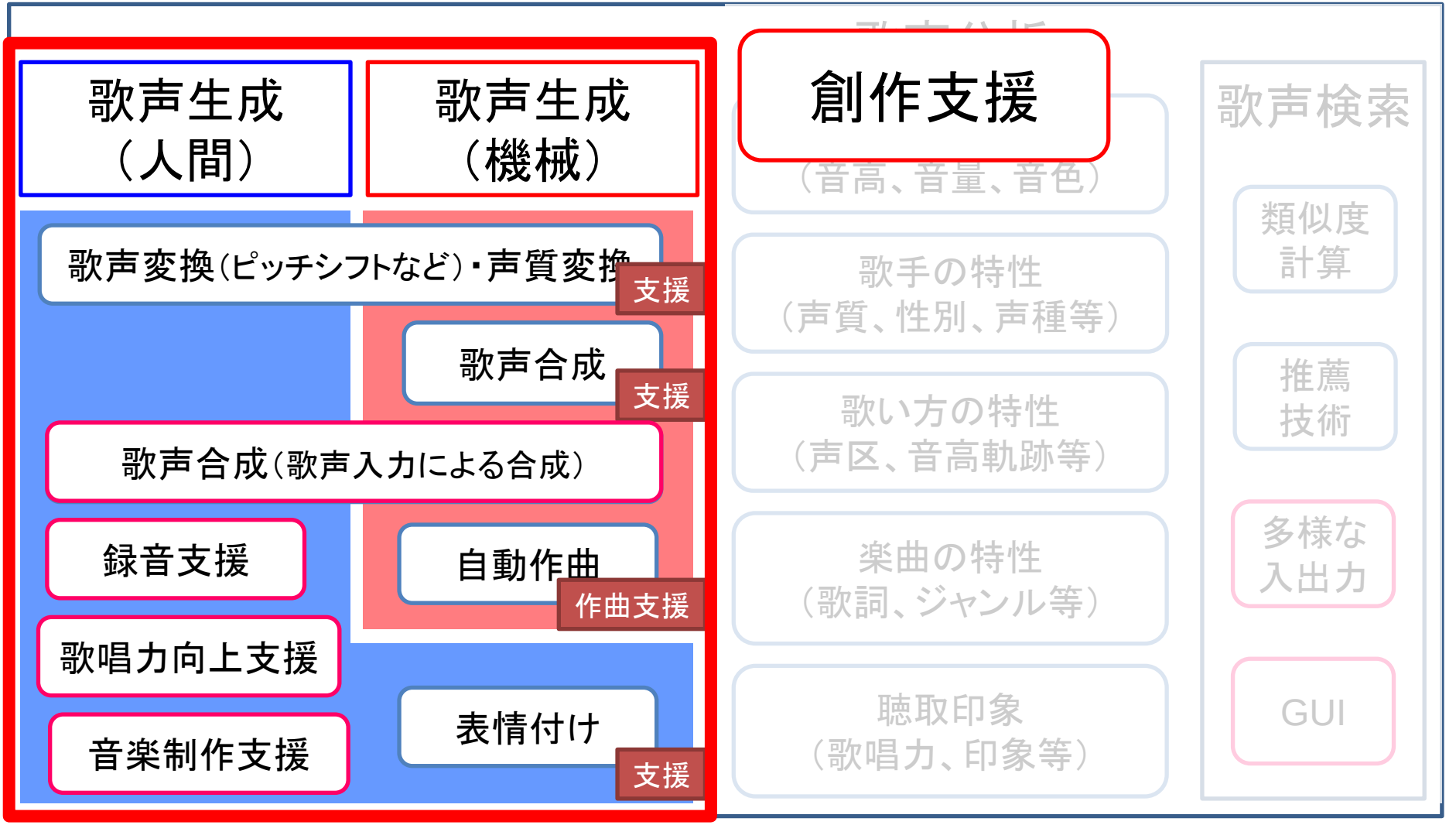
類似度計算

推薦技術

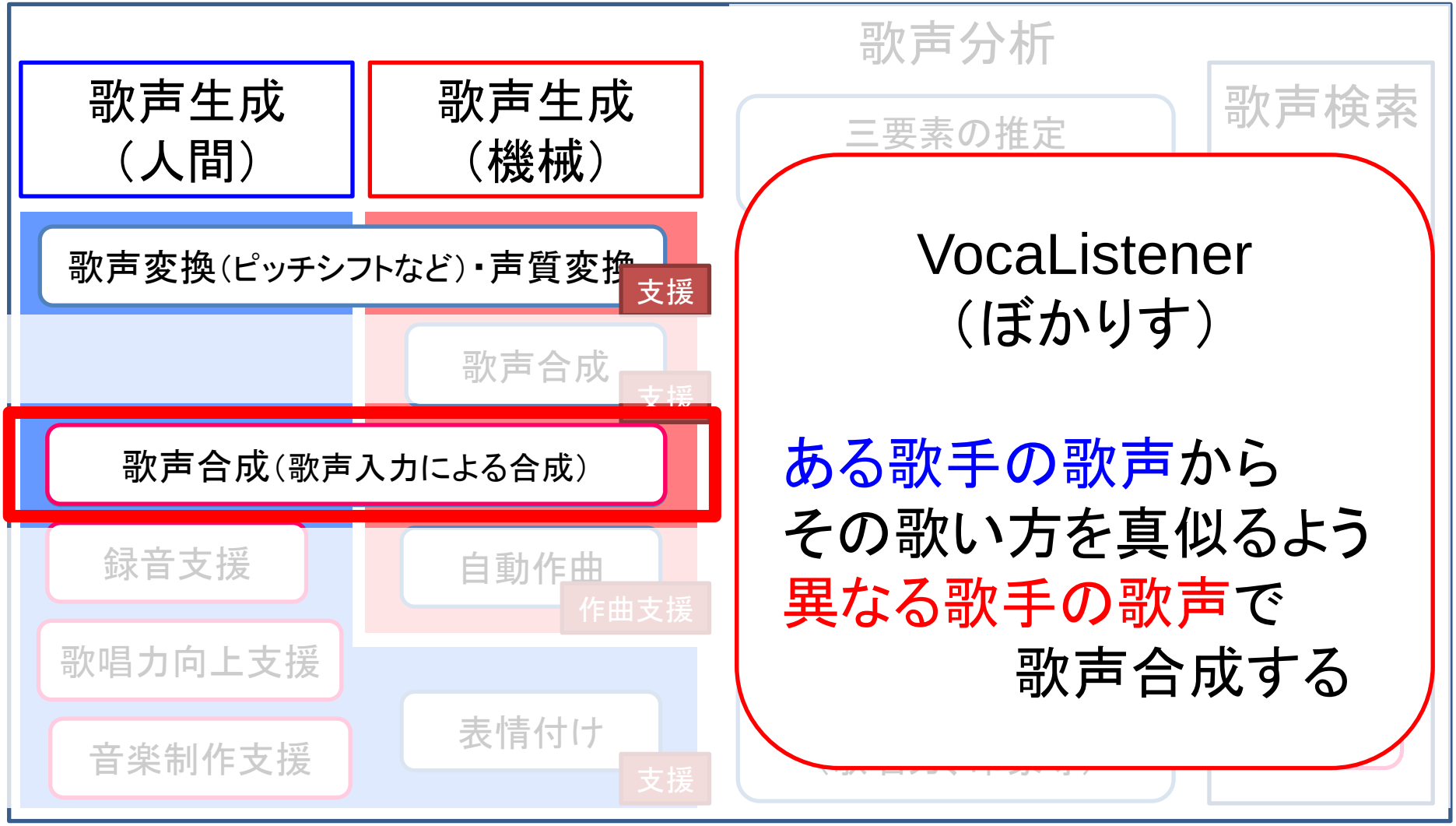
歌声インタフェースの全体像



歌声インタフェースの全体像



歌声インタフェースの全体像



VocaListener（ぼかりす）による合成結果

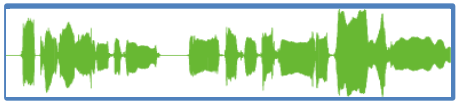
- 【初音ミク】PROLOGUE【ぼかりす】
  - 2008/04/28 [sm3128145] ～
  - *361,055 views*

[original]

- 【メグッポイド】Packaged【ぼかりす】
  - 2010/10/4 [sm12320140] ～
  - *31,404 views*

歌声合成とは

- コンピュータ上で歌声を作ること
  - 入力
    - 楽譜
    - 歌詞
    - 歌声合成パラメータ(声の高さ、音量)
  - 出力
    - 歌声の音響信号



音符を並べて歌詞入力

＋ 細かなパラメータ調整

これまでの歌声合成の問題点

- 音符を並べて入力

問題

音符を並べただけでは
自然性が低い

歌 詞
「ためらいもなく飲み込んで
次々と消化してゆくの」

従来

VocaListener

これまでの歌声合成の問題点

- 音符を並べて入力

問題

音符を並べただけでは
自然性が低い

歌 詞
「ためらいもなく飲み込んで
次々と消化してゆくの」

従来

VocaListener

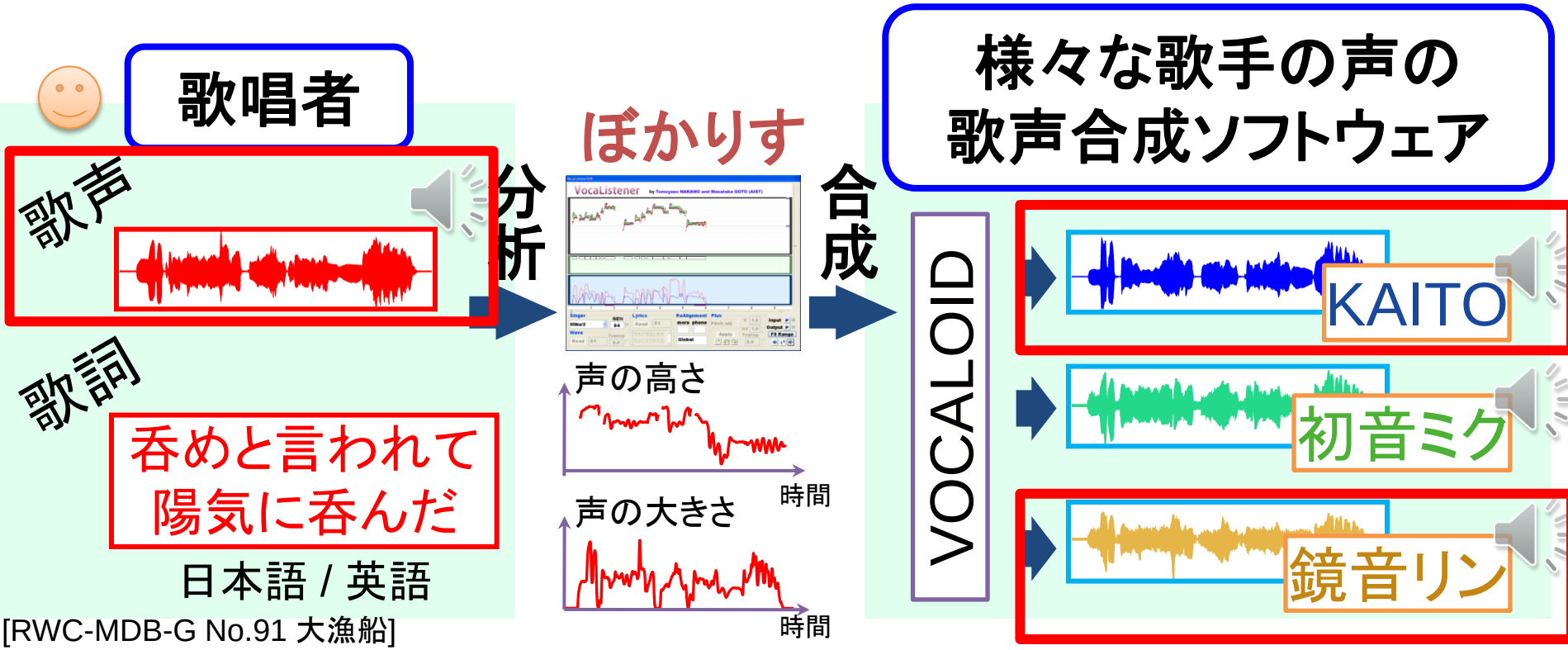
問題

品質を高くしようとすると
パラメータ調整に数時間
（数十時間）かかる

VocaListener（ぼかりす）

[中野, 後藤, 2008-]

- 歌うだけで楽譜と自然なパラメータを自動生成
- 音源（歌手）を手軽に切り替えて合成できる



これまでの歌声合成の問題点

- 音符を並べて入力

問題

音符を並べただけでは
自然性が低い

歌 詞

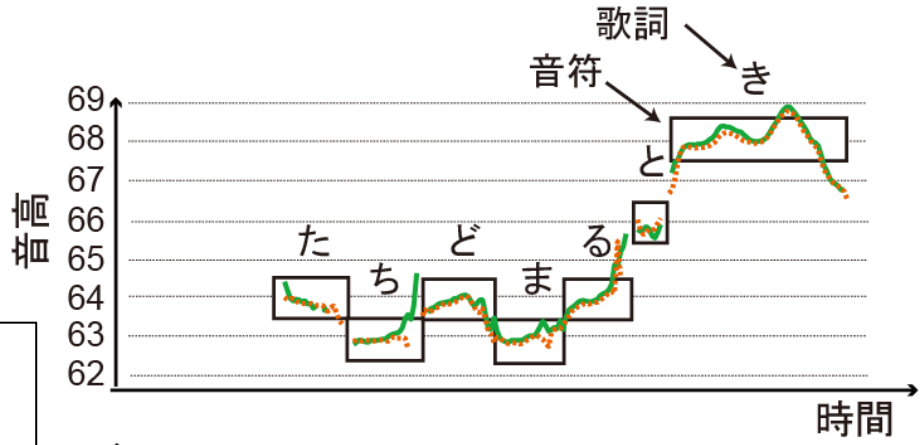
「呑めと言われて
陽気に呑んだ」

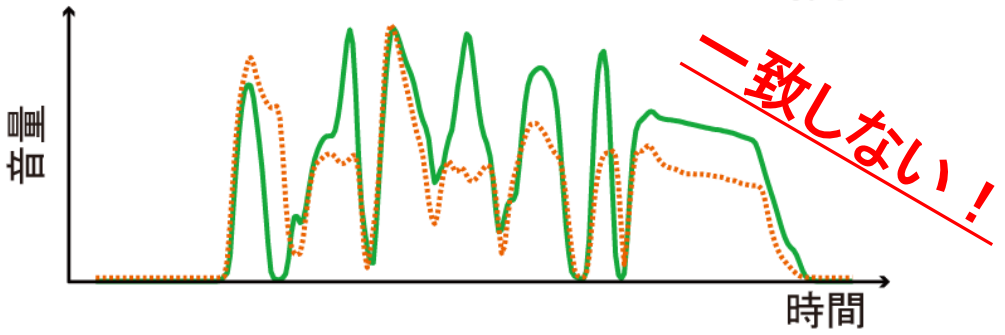
ぼかりすの実現課題(1)

問題

与えたパラメータ通りに合成されない

- 声の高さと声の大きさが歌声合成ソフトに依存





同じパラメータを与えて
二つの歌声合成ソフト
AとB
で合成した結果

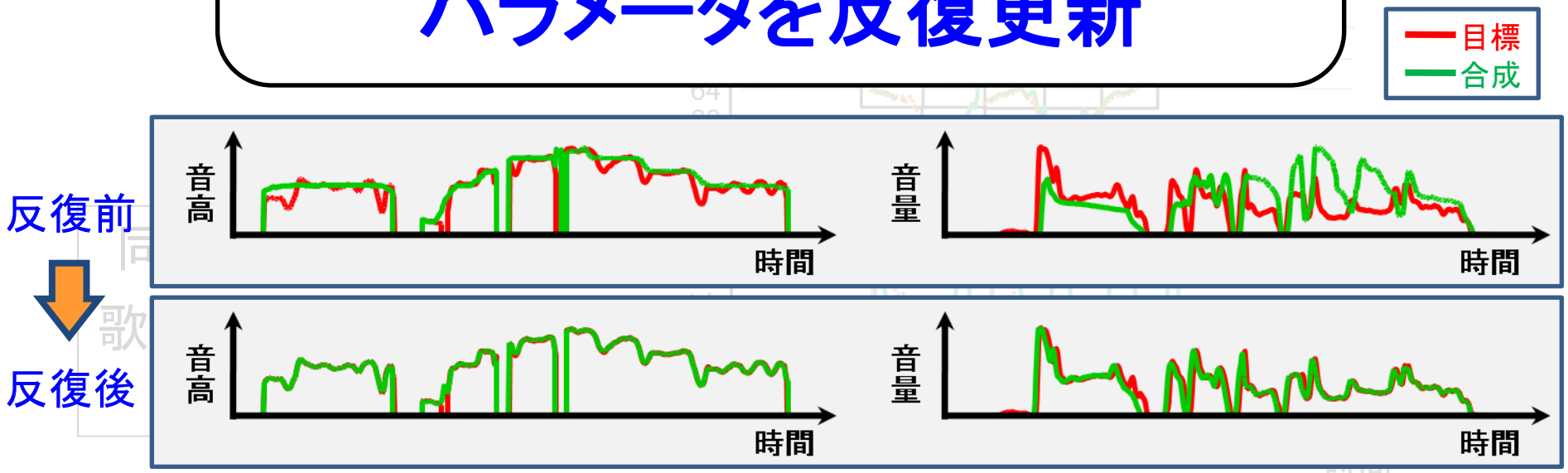
ぼかりすの実現課題(1)

問題

与えたパラメータ通りに合成されない

解決方策

合成結果を分析
して、目標の値に近くなるように
パラメータを反復更新

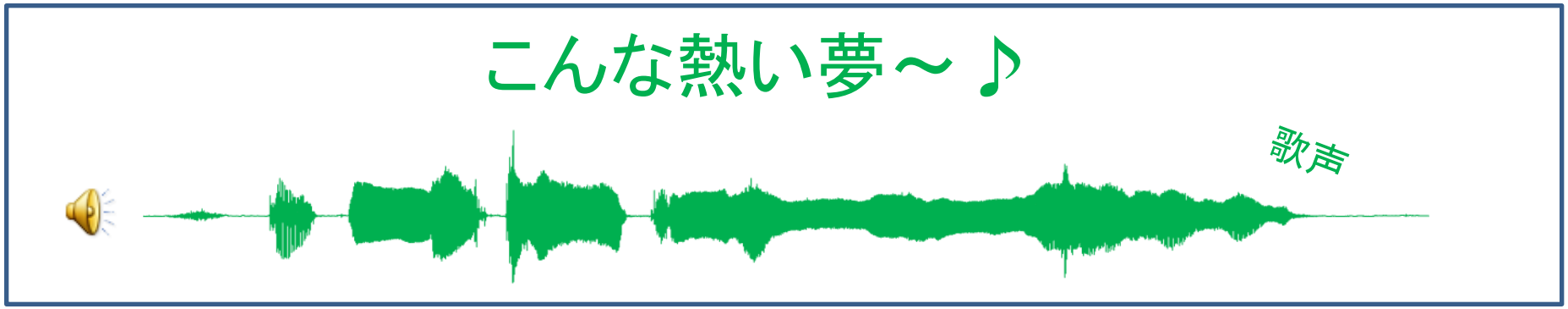


ぼかりすの実現課題(2)

問題

歌詞のどこを、いつ歌っているか分からない

- 歌声の信号と歌詞のテキストだけがある



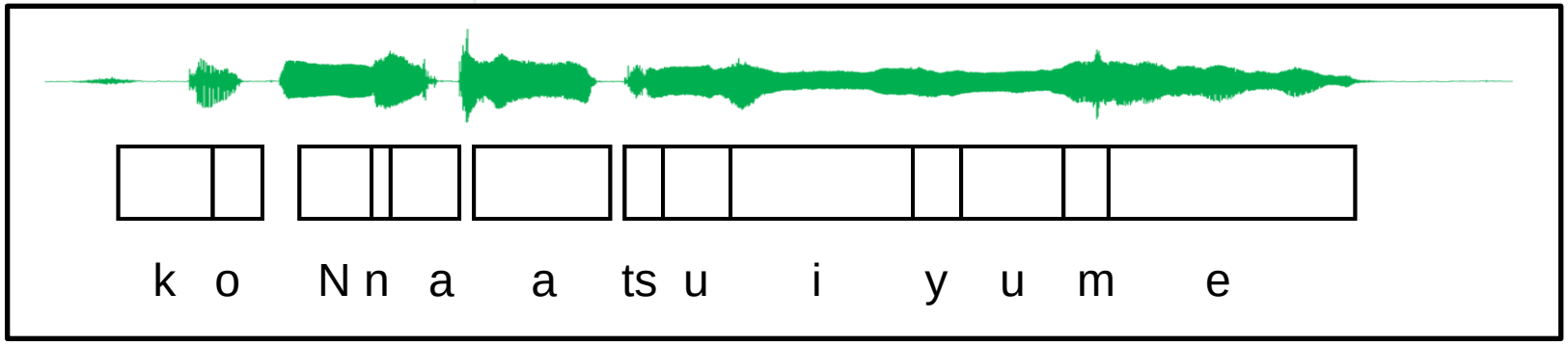
ぼかりすの実現課題(2)

問題

歌詞のどこを、いつ歌っているか分からない

解決方策

歌声専用の音響モデル
（HMM: 声の響き方を学習したモデル）
を構築して高精度に対応付け



ぼかりすの実現課題(2)

問題

歌詞のどこを、いつ歌っているか分からない

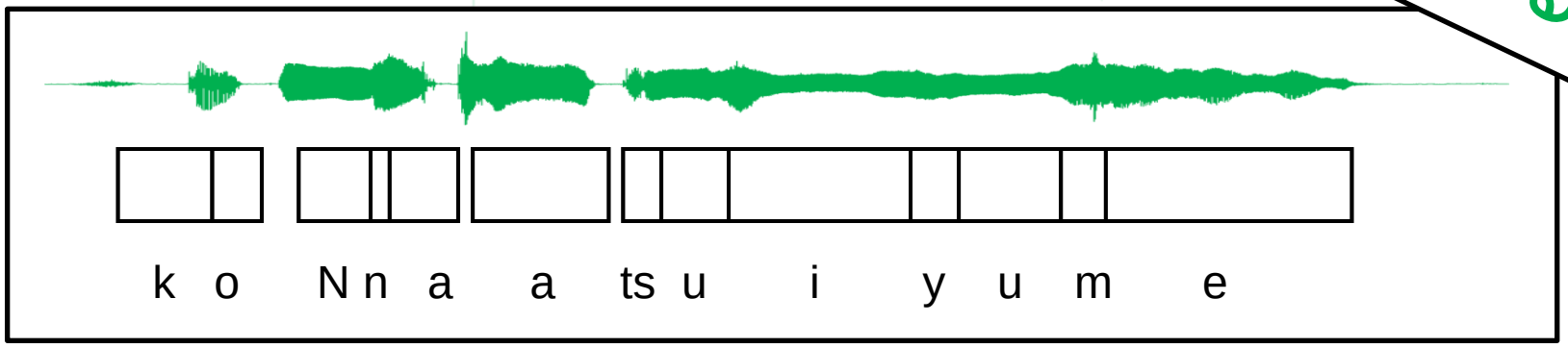
解決方策

歌声専用の音[illegible]デル

（HMM: 声の響き方を学[illegible]）

を構築して高精度に対応[illegible]

しかしまだ問題がある

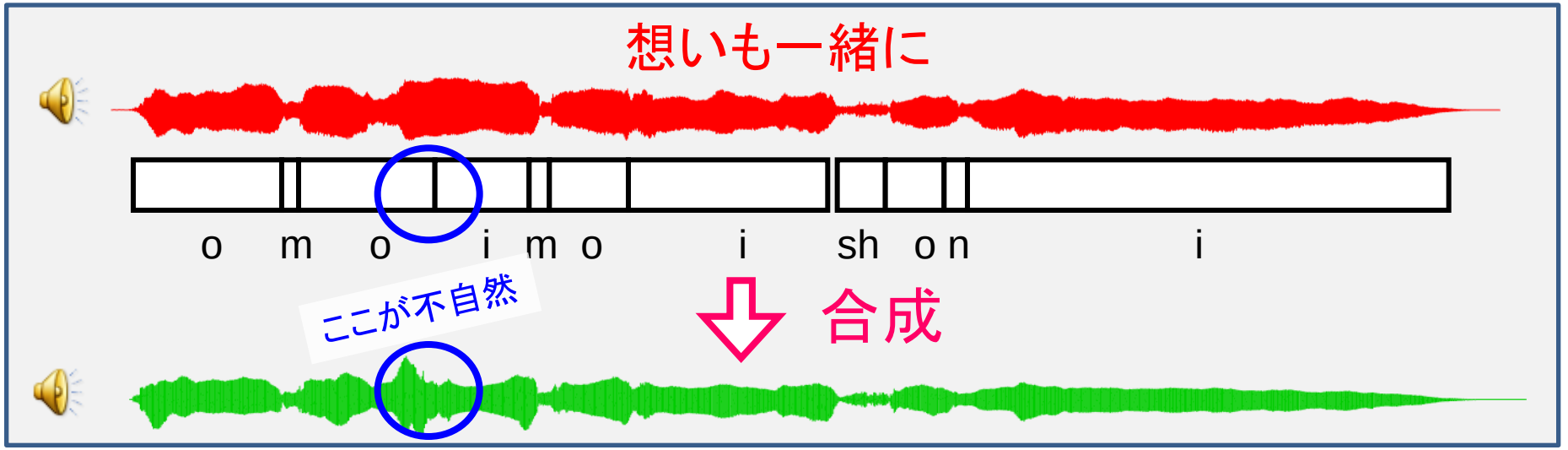


ぼかりすの実現課題(3)

問題

合成のためには100%に近い精度が必要

- 高い精度であるが、数%の誤りは不可避



ぼかりすの実現課題(3)

インタラクション

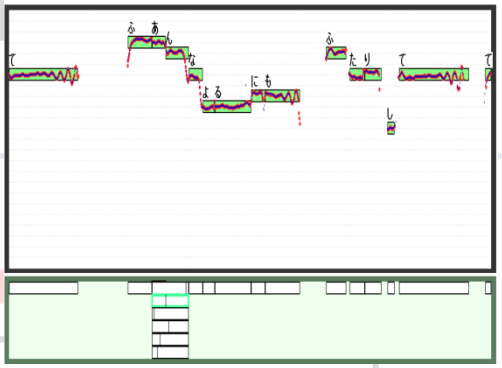
問題

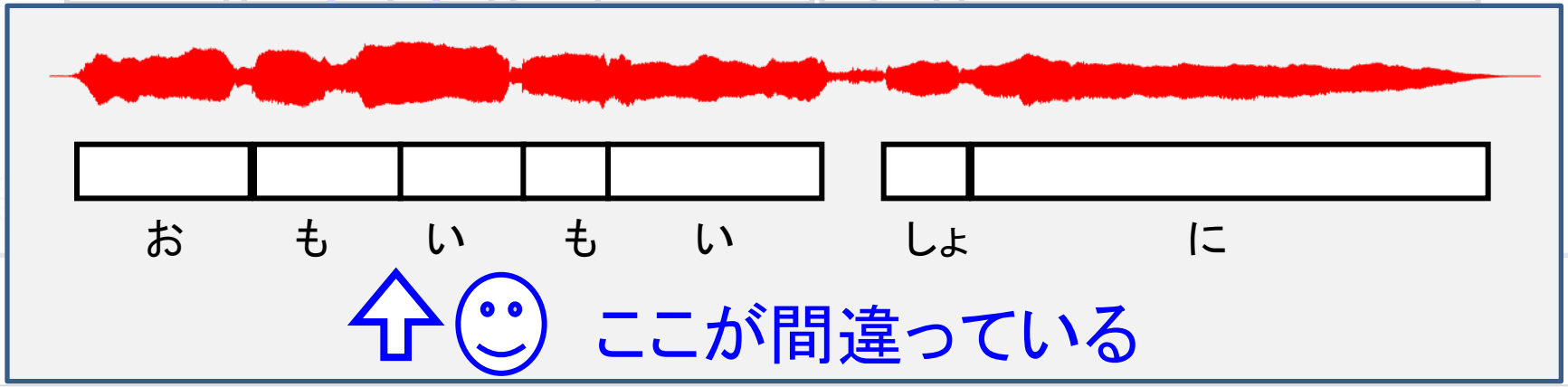
合成のためには100%に近い精度が必要

解決方策

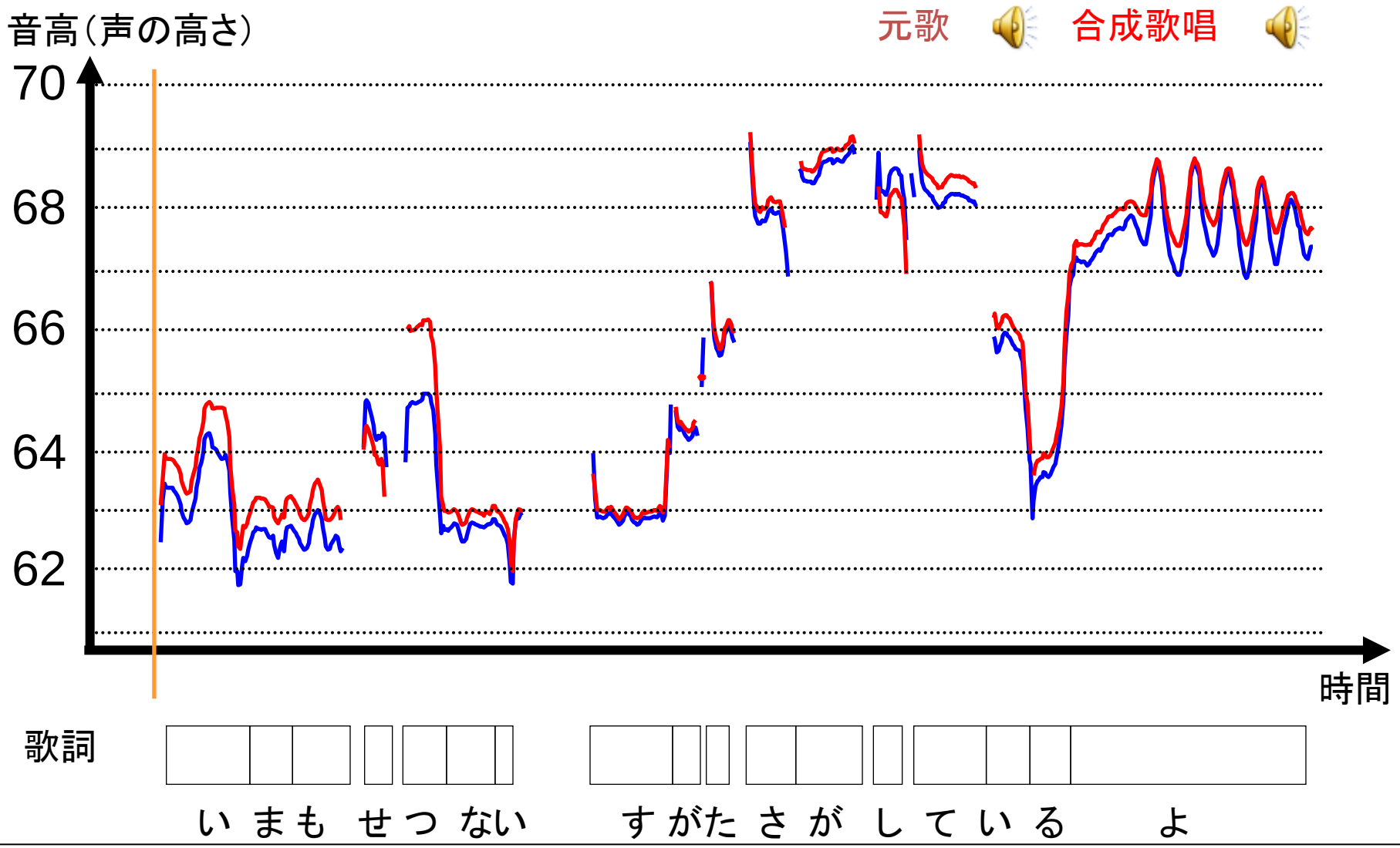
ダメ出しインタフェース

（聴いて指摘するだけで訂正）





歌唱力補正も可能



VocaListener VOCALOID3 Job Plugin

- 2012/10/19 ヤマハ株式会社より製品化

VocaListener
VOCALOID™3 Job Plugin

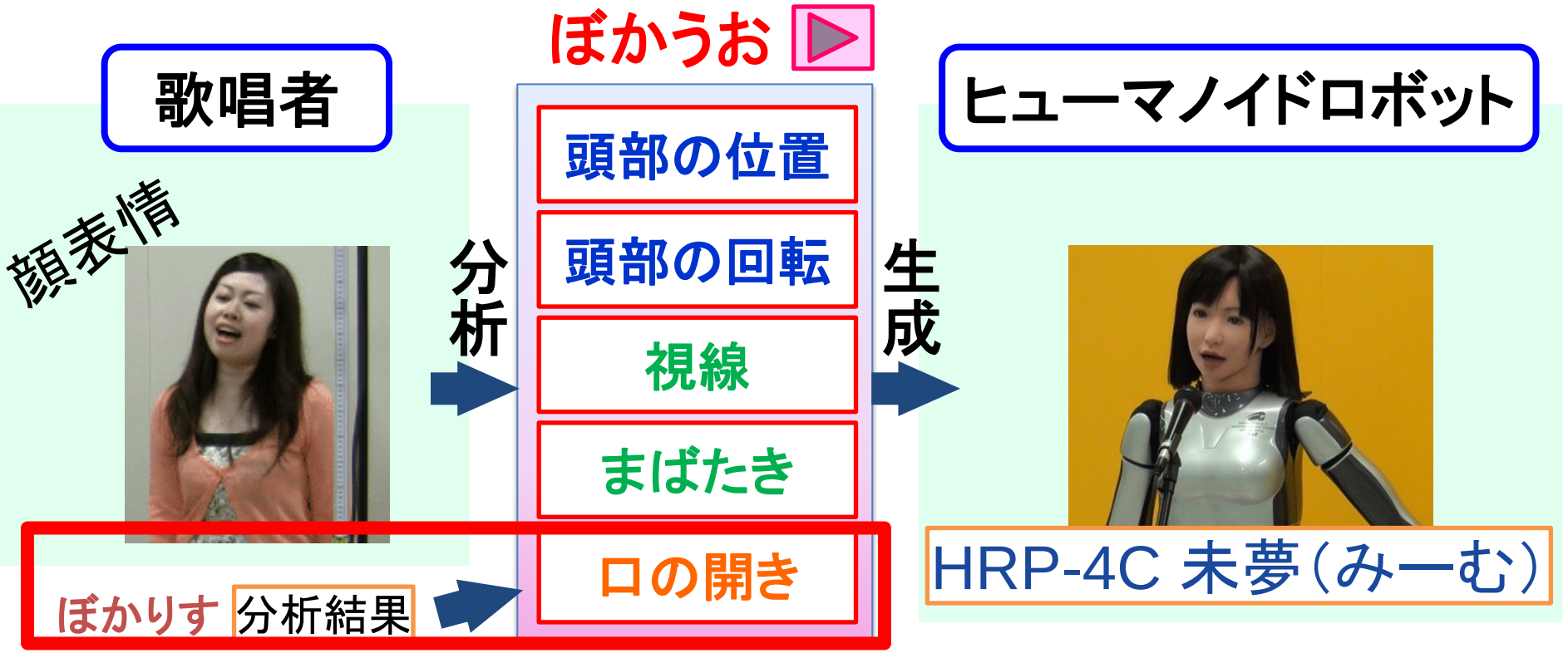
- 【ぼかりす公式デモ】迷子ライフ【歌：がくっぽいどPower】
  - 2012/10/22 [sm19181691] ～
  - *8,489 views*

- 塾講師なので本気出して卒業ソングをつくってみたよ【初音ミク など】
  - 2013/3/22 [sm20403611] ～
  - *111,977 views*

VocaWatcher（ぼかうお）

[Kajita, Nakano, Goto+ 2010-]

- 歌うだけでヒューマノイドロボットの
  自然な顔動作パラメータを自動生成



複数の声色で歌声合成が可能

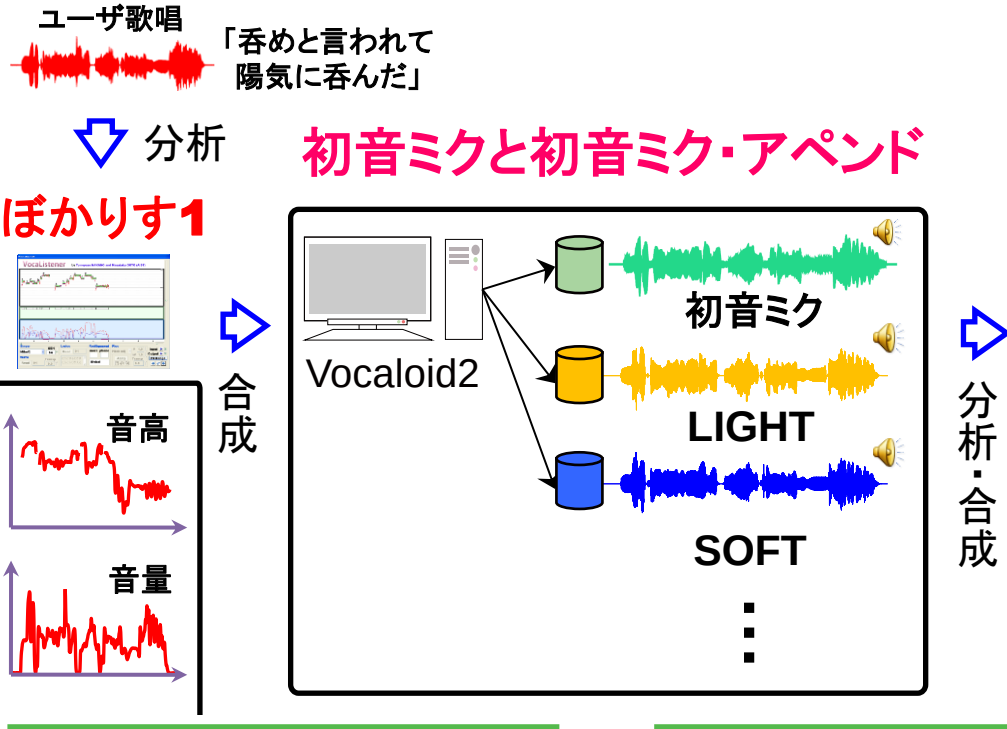
- 中間の声色を作ることはできなかった
- ユーザの声色変化を真似ることはできなかった

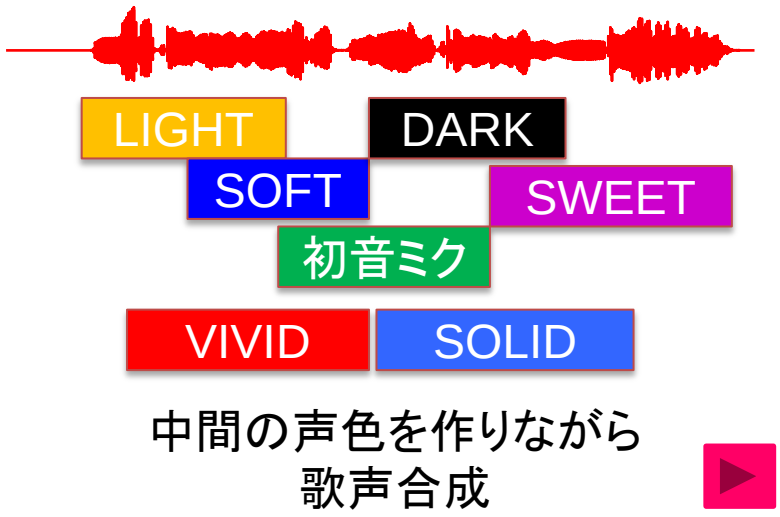
[RWC-MDB-G No.91 大漁船]

VocaListener2（ぼかりす2）

[中野, 後藤, 2010-]

- **VocaListenerで音高と音量**を真似て全て合成
- 声色変化を真似るようにそれらを**混ぜ合わせる**

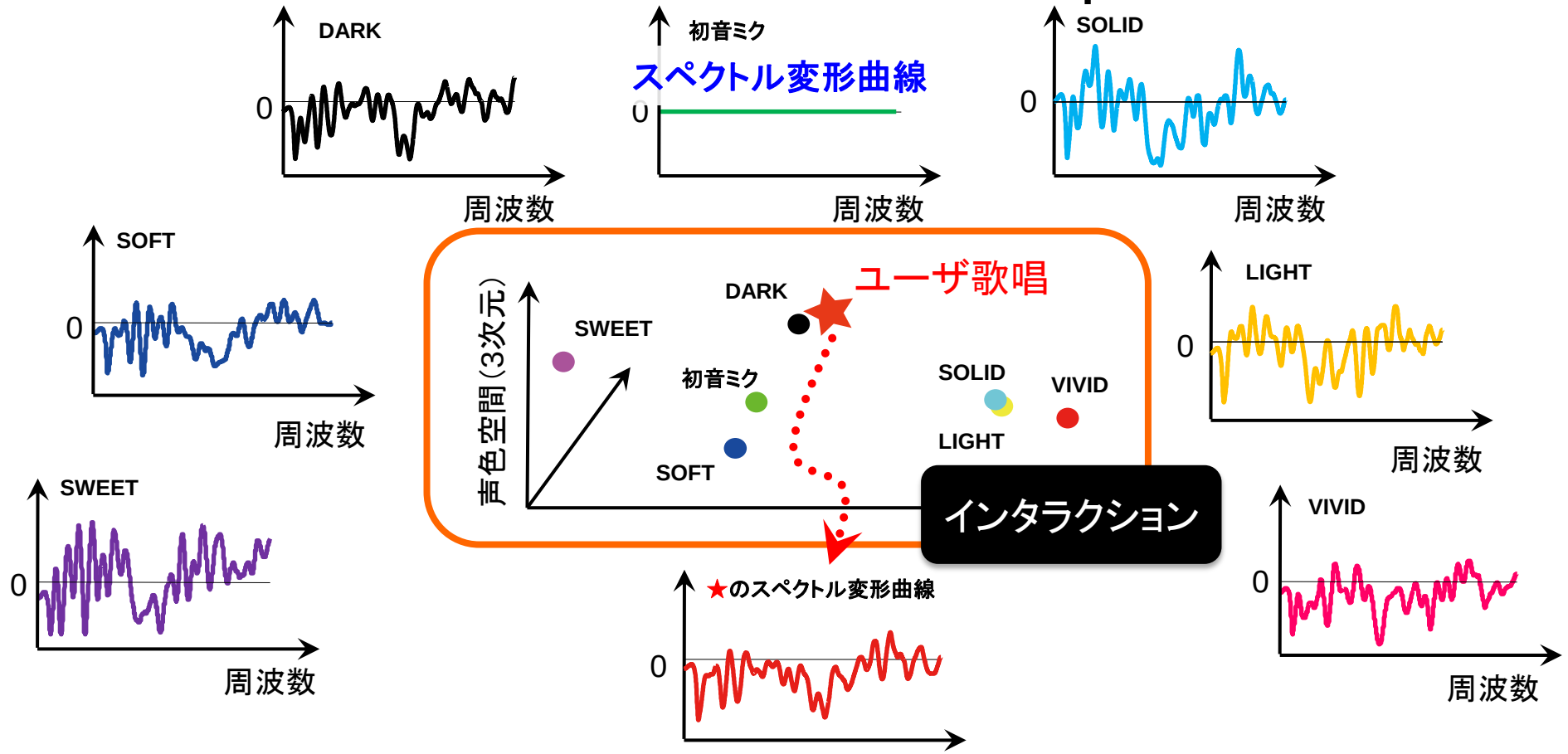




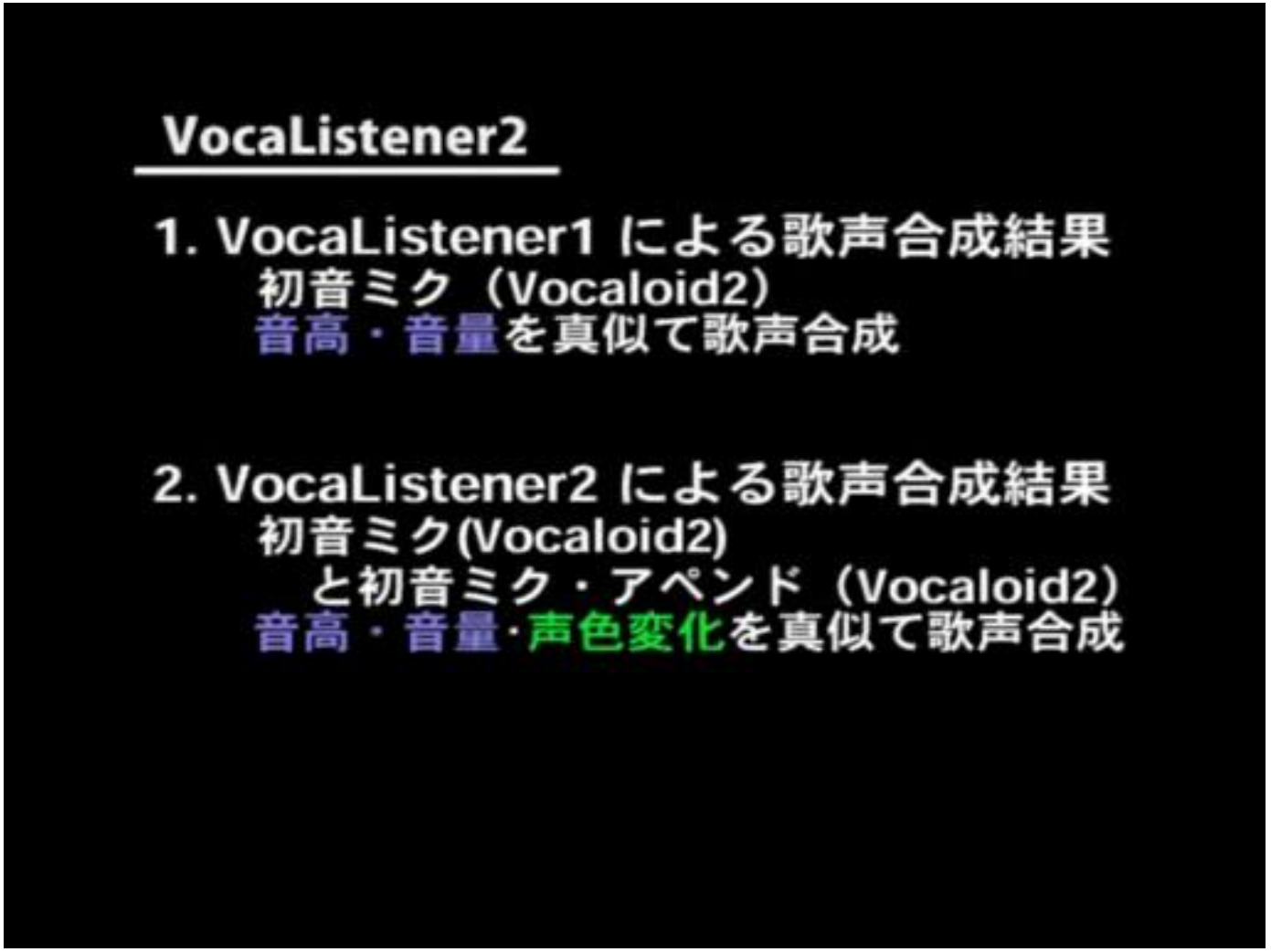
声色空間に基づくスペクトル変形曲線の推定

- スペクトル包絡の主成分分析により3次元空間へ射影
- **Radial Basis Functionを用いたVariational Interpolation**

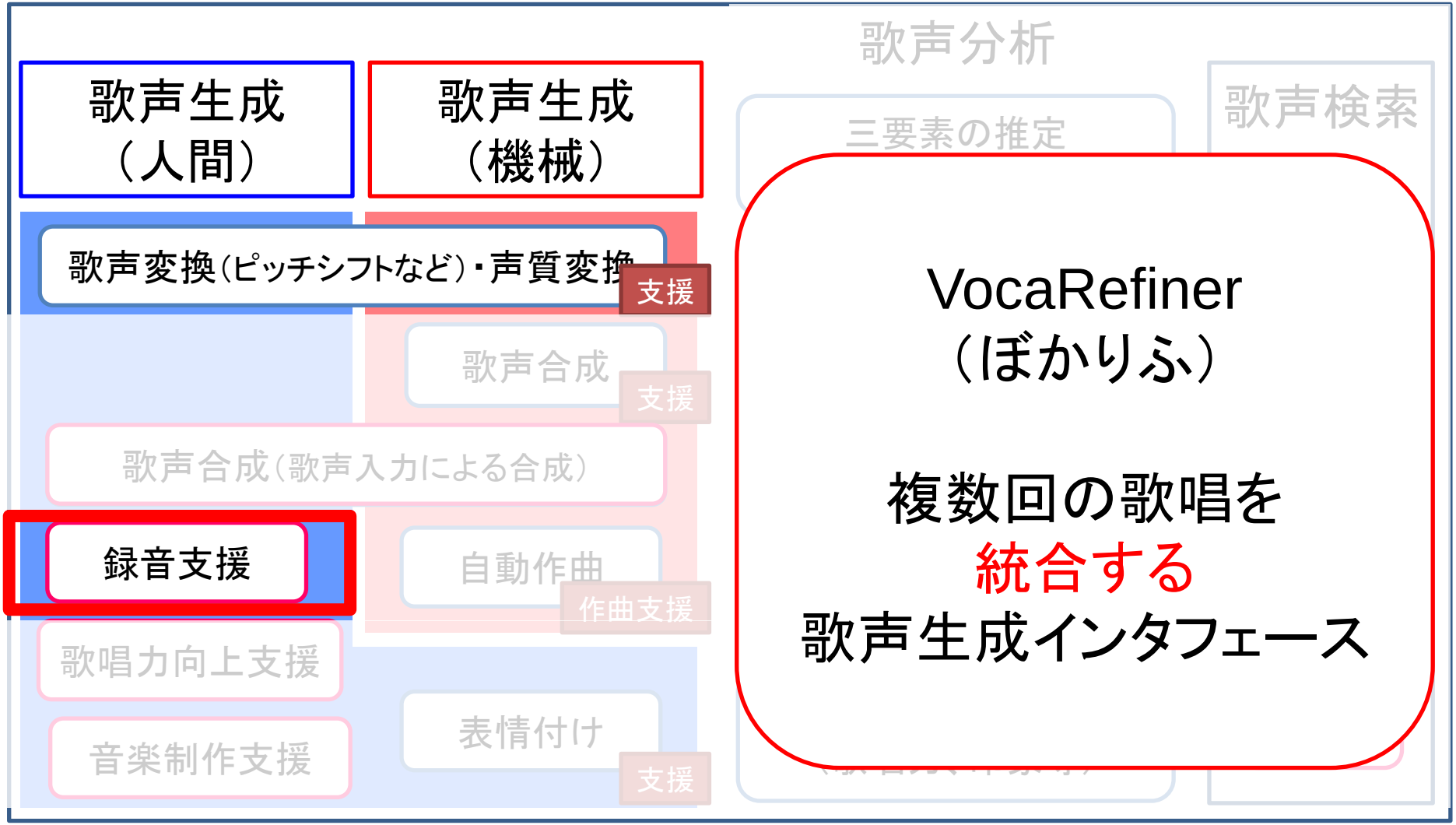
[Turk et al., 2008]



ぼかりす2による歌声合成デモ



歌声インタフェースの全体像



歌をうまく録音したい

- 誰でも気軽に自分の歌を公開して共有する文化

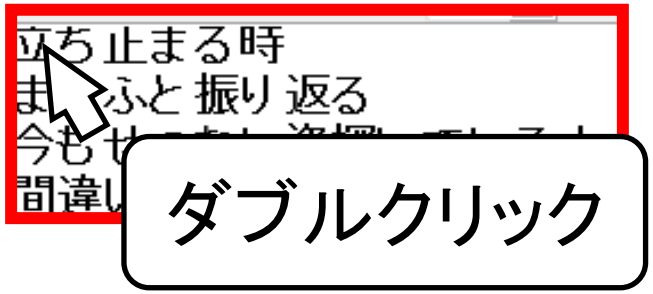
- 歌ってみた（ニコニコ動画）
  - 投稿動画数：64万件以上

歌を一度で完璧に録音するのは困難

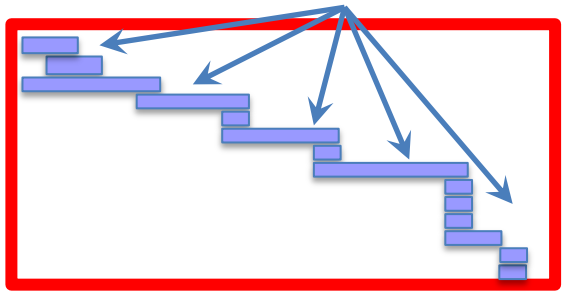
F0適応多重フレーム統合分析法とF0適応F0推定法に基づく[中野, 後藤, 2012]

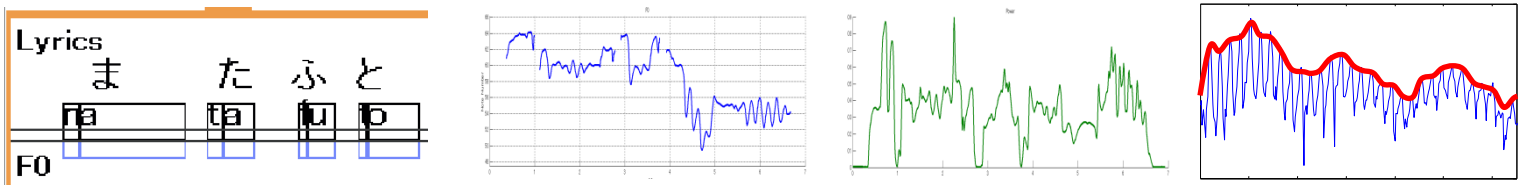
歌声生成インタフェースVocaRefiner

- 歌詞を使って手軽に録音する
  - 効率的に録音できる
  - 日本語と英語に対応



- 歌を歌い直して（複数回録音）統合する
  - 音素単位で統合できる
  - 音の三要素で統合できる
  - 過去に歌った歌を活用する



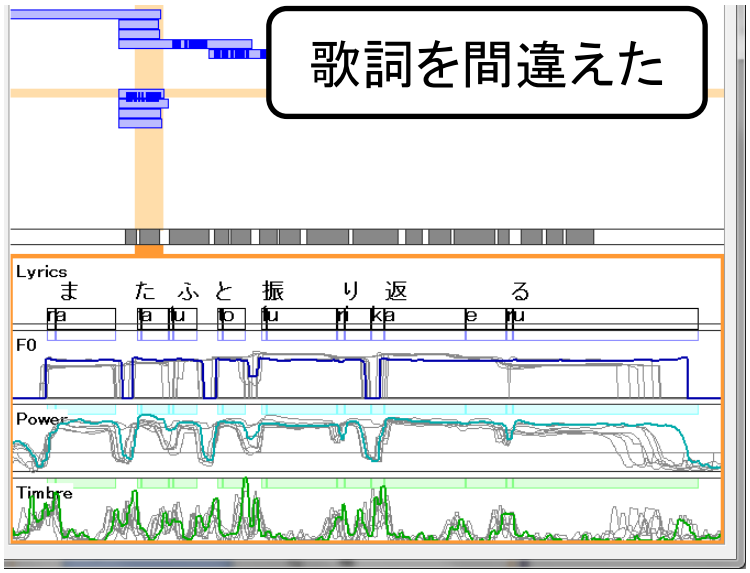


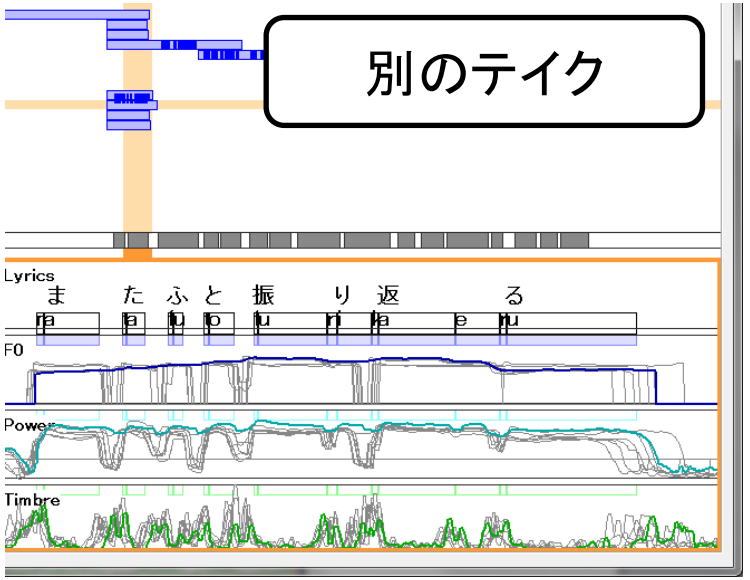
統合機能

※歌詞を間違えたと想定して
ハミング（ラララ歌唱）で歌った

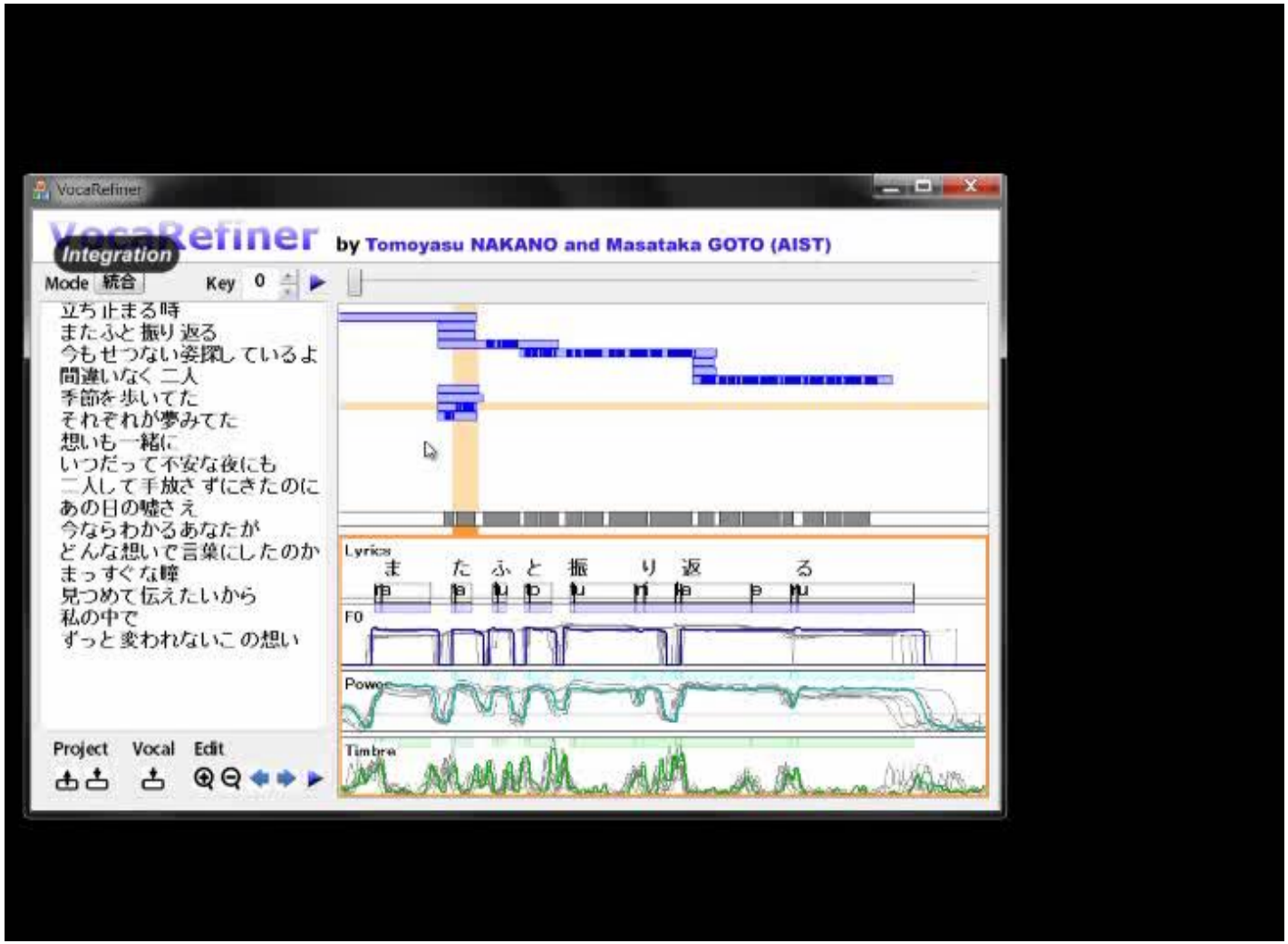
- 三要素を組み替えながら統合
  - 「感情は完璧だったが歌詞を間違えた」場合
    - 歌い方（音高・音量）はその録音を使用

      歌詞が正しい過去録音の音色を使用

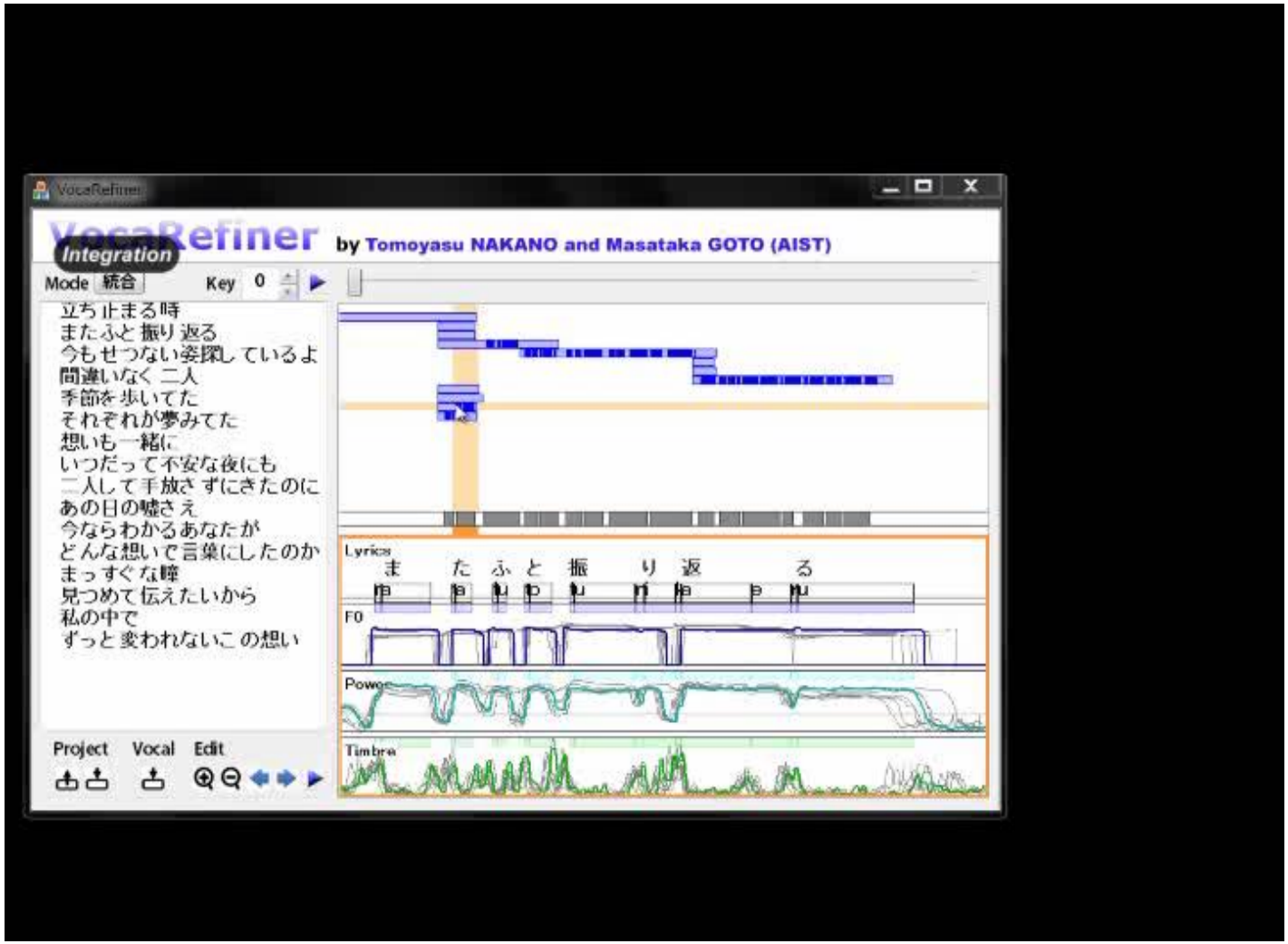




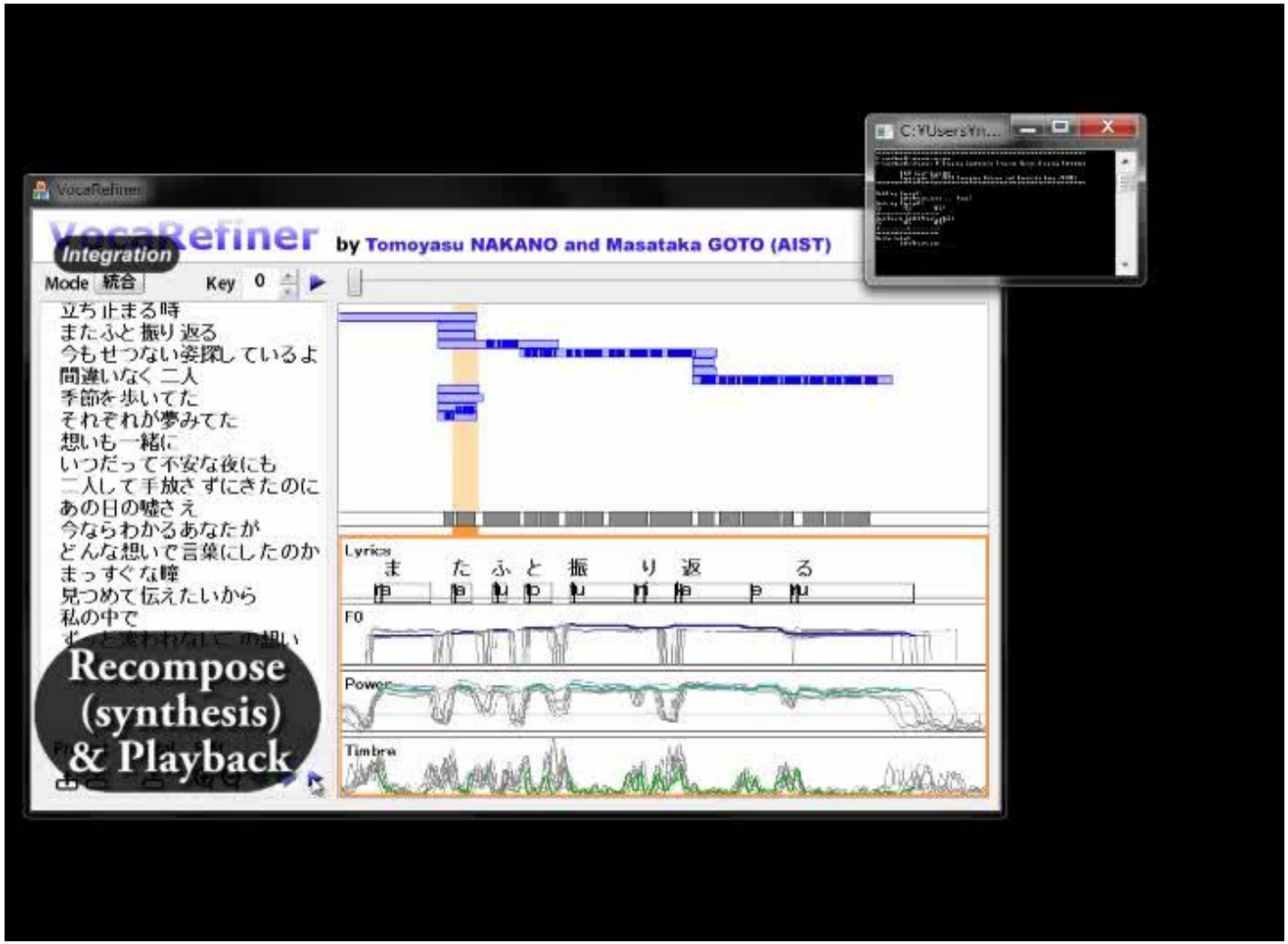
テイク1:音高は正しいが歌詞を間違えた



テイク2：歌詞は正しいが音高が誤った



統合（合成）：正しい歌詞と正しい音高

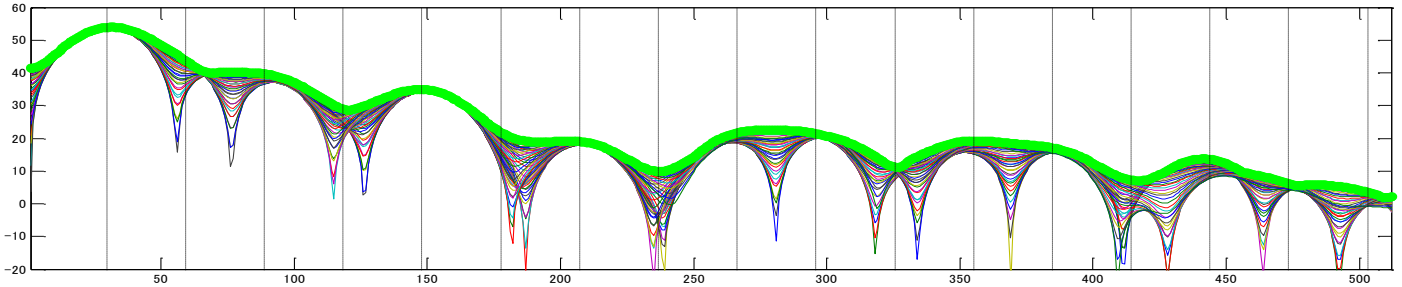


話声と歌声の分析合成のための F0適応多重フレーム統合分析法

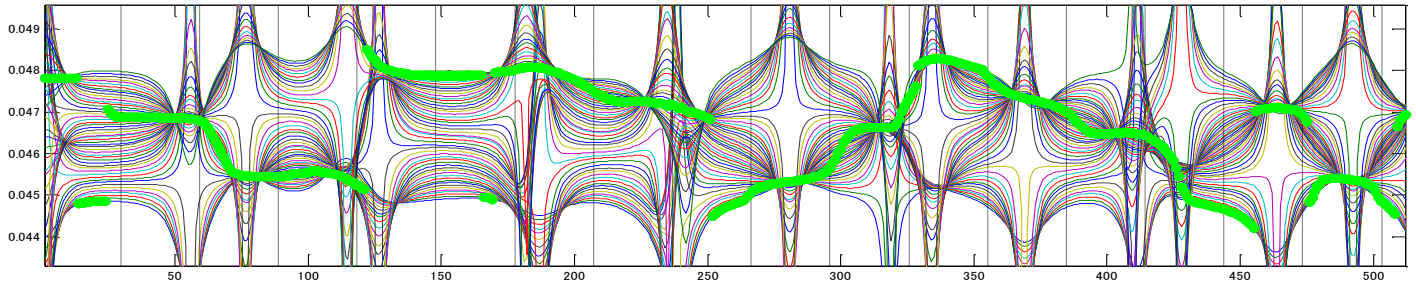
[中野, 後藤, 2012]

- 高い精度と時間分解能で推定可能
  - スペクトル包絡：音素、声質、声色
  - 群遅延（位相の周波数微分）：波形の形、声の質感

F0適応スペクトル
と
最大包絡

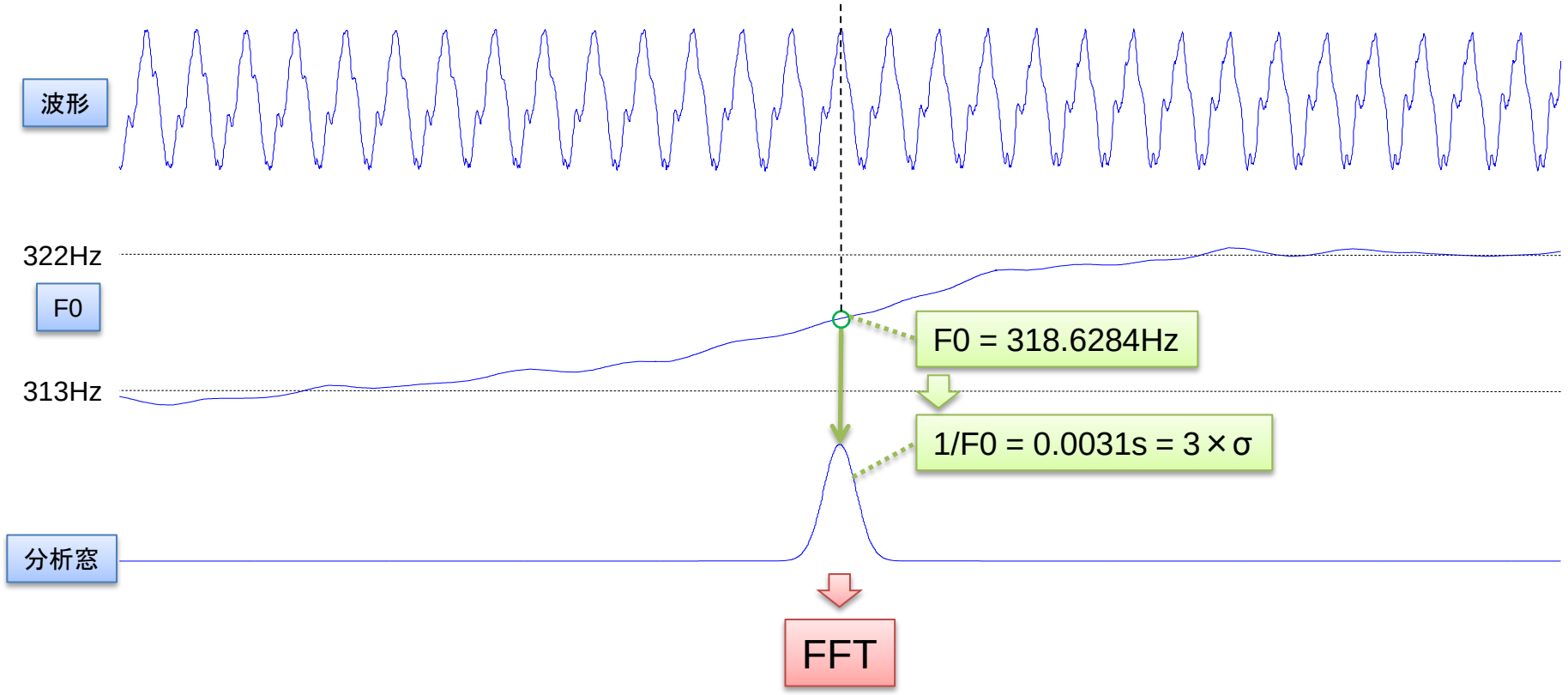


最大包絡
に対応する
群遅延



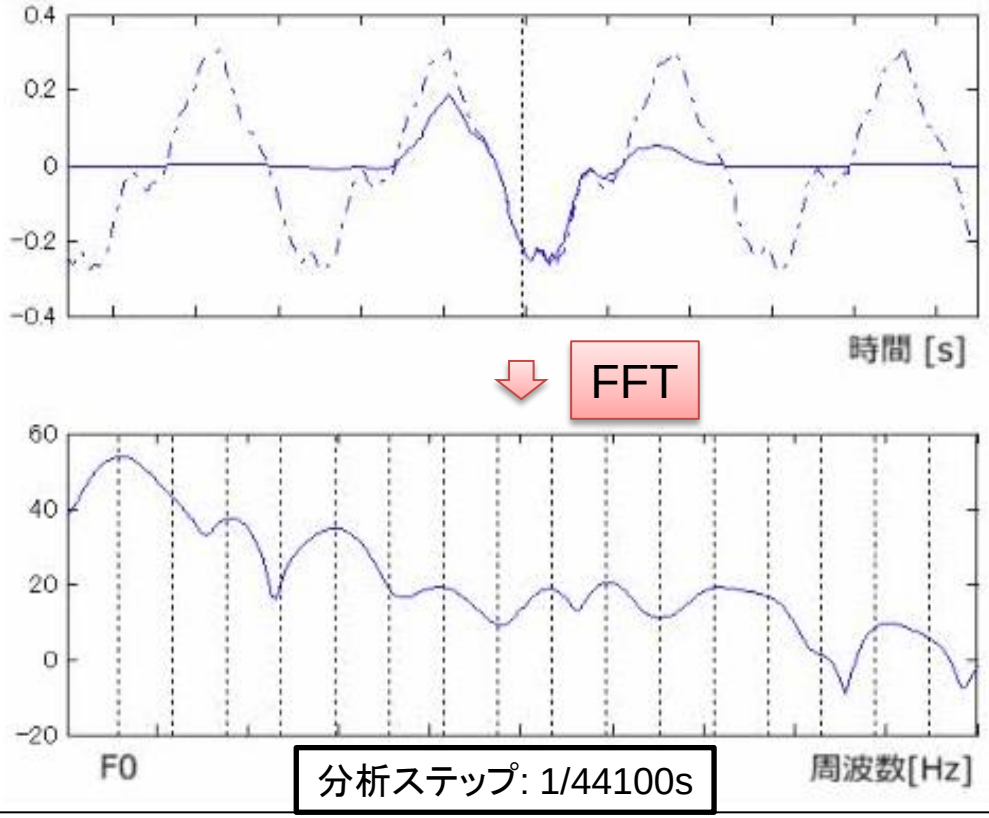
F0適応分析:F0に適応した長さの短い窓

- F0に適応した長さの短い窓で
  全時間(全サンプリング点)に対してFFT



F0適応分析：全時間に対してFFT

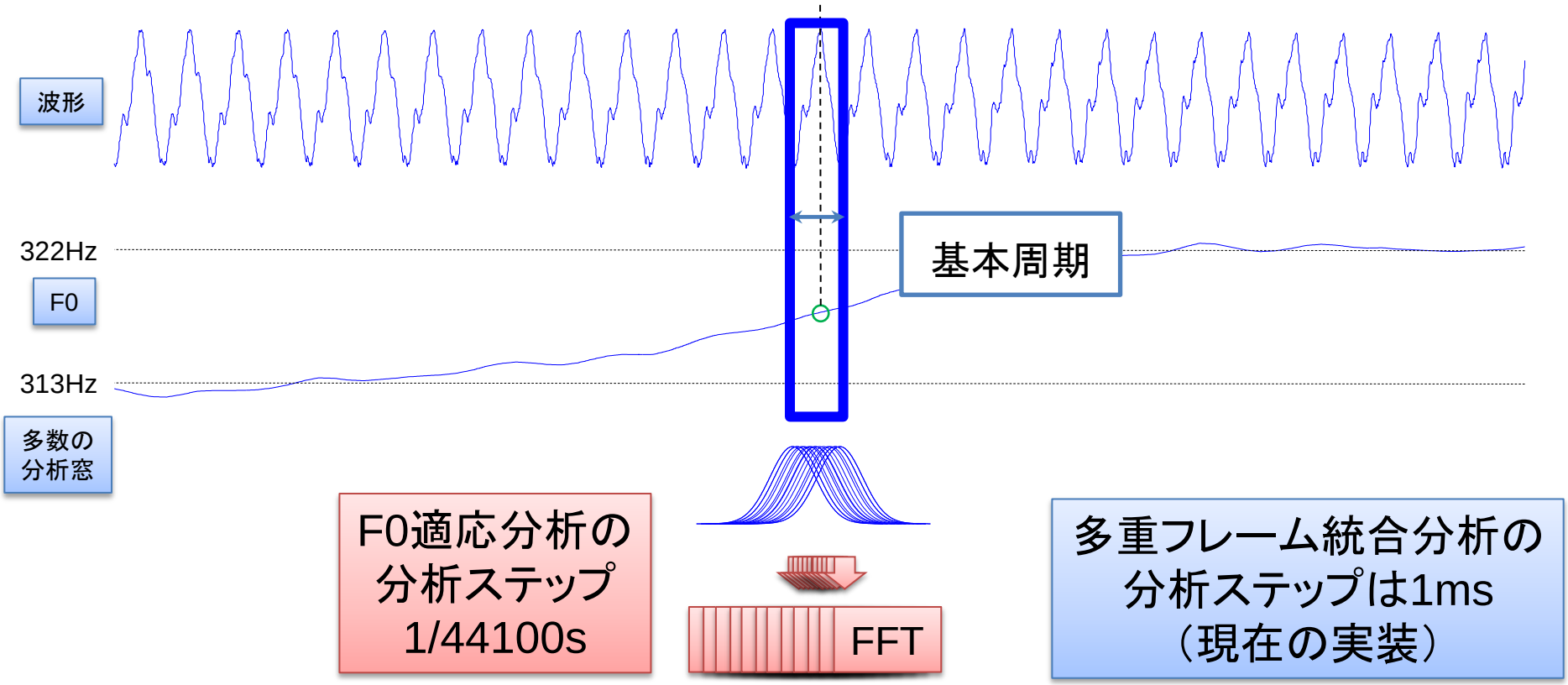
- F0に適応した長さの短い窓で
  全時間（全サンプリング点）に対してFFT



多重フレーム統合分析:基本周期の範囲で統合

- 基本周期の範囲で

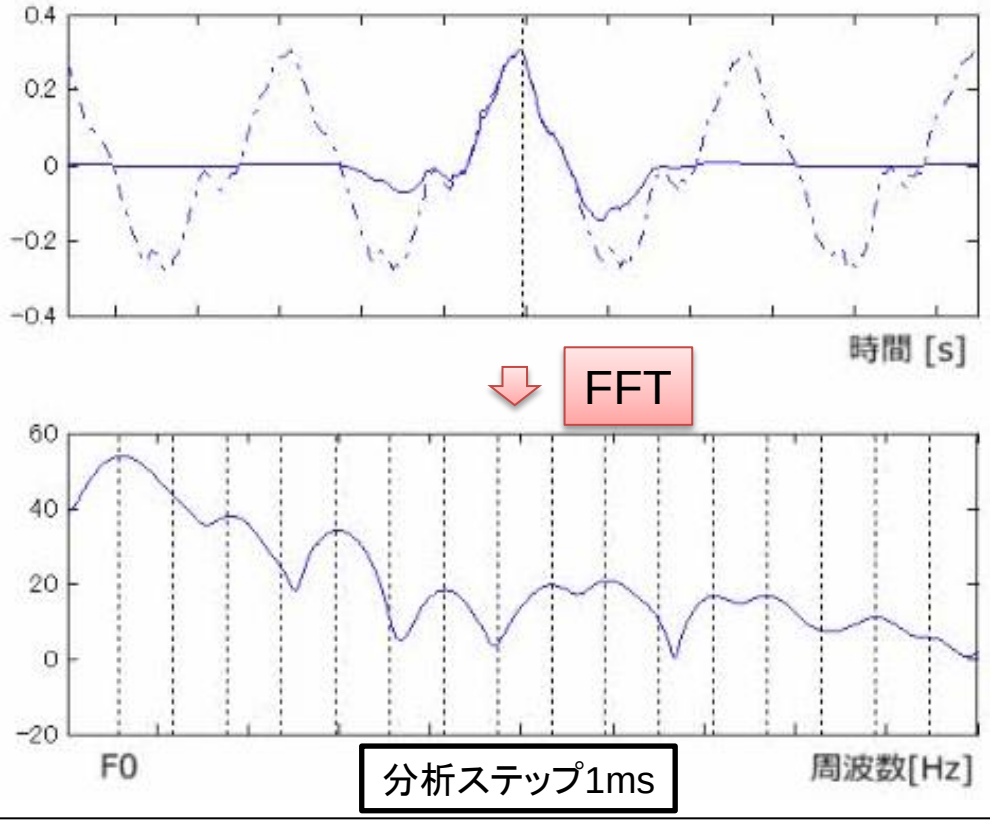
  F0適応スペクトルを統合して変動を消去



多重フレーム統合分析:基本周期の範囲で統合

- 基本周期の範囲で

  F0適応スペクトルを統合して変動を消去



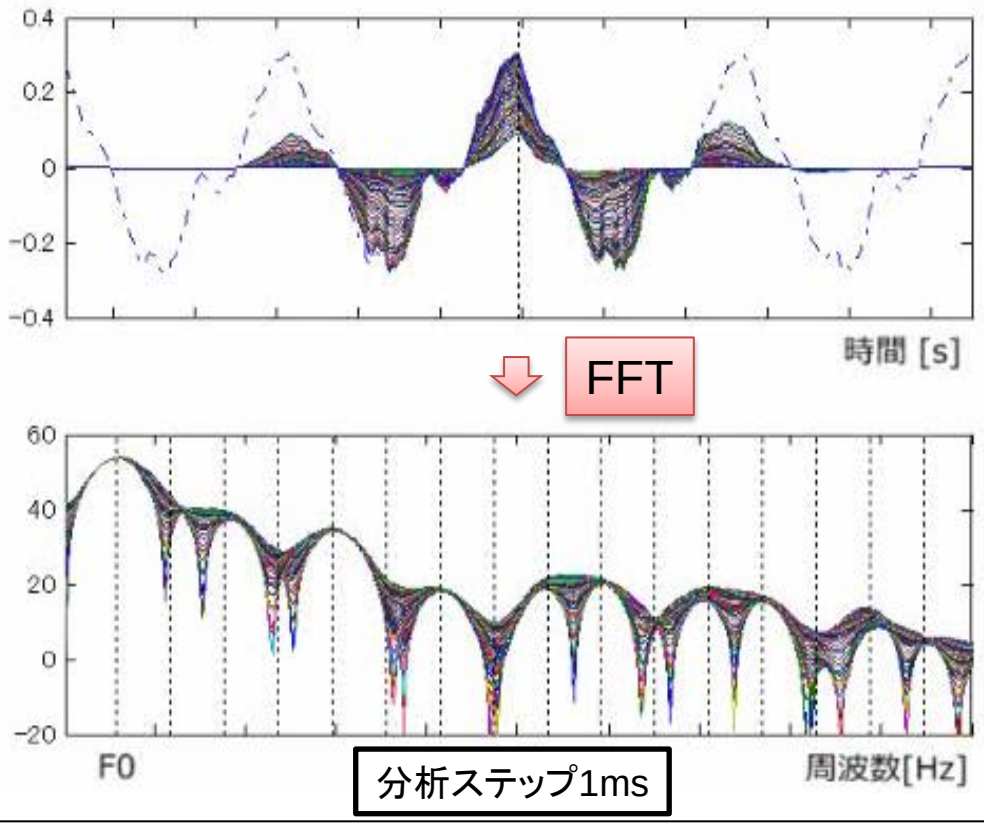
統合を
行わない場合

谷が発生

多重フレーム統合分析:基本周期の範囲で統合

- 基本周期の範囲で

  F0適応スペクトルを統合して変動を消去



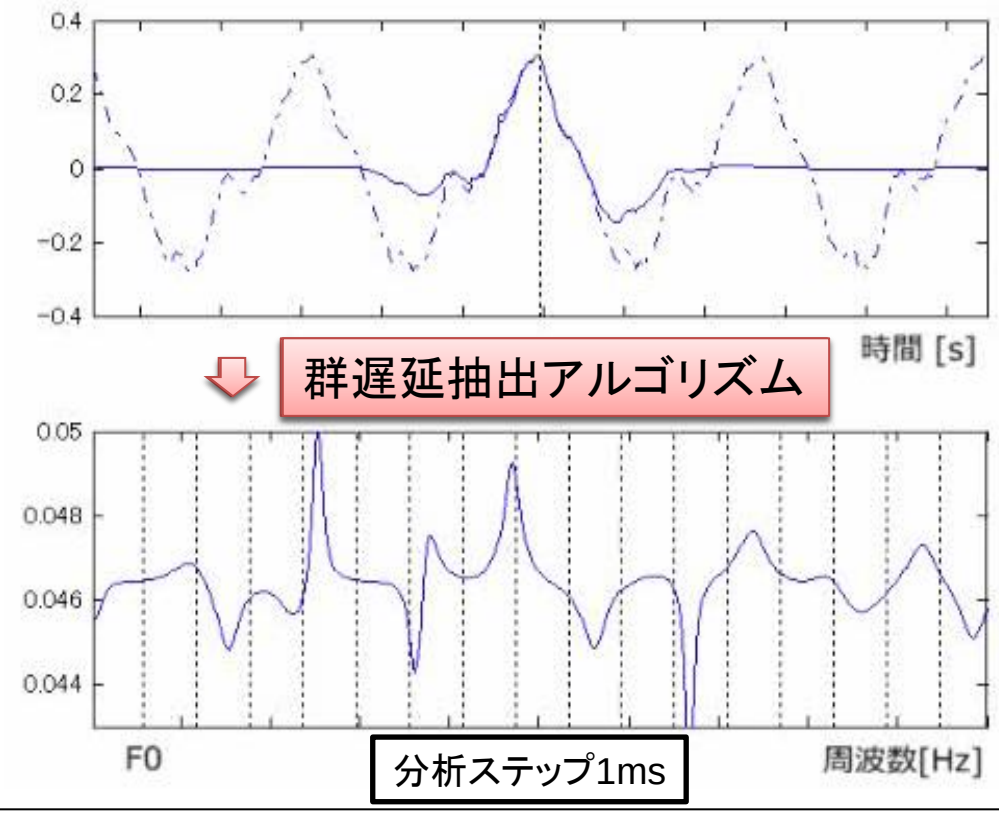
統合(重畳)により

どこで分析しても谷が埋まる

多重フレーム統合分析:基本周期の範囲で統合

- 基本周期の範囲で

  F0適応群遅延を統合して変動を消去



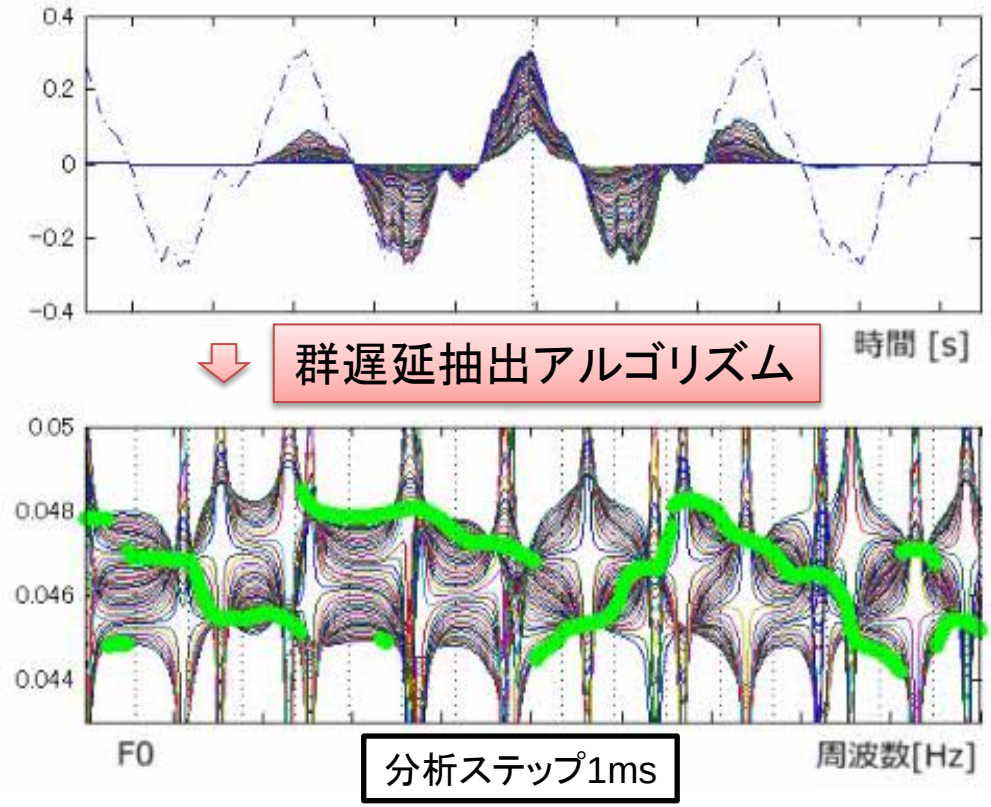
統合を
行わない場合

ピークが発生

多重フレーム統合分析:基本周期の範囲で統合

- 基本周期の範囲で

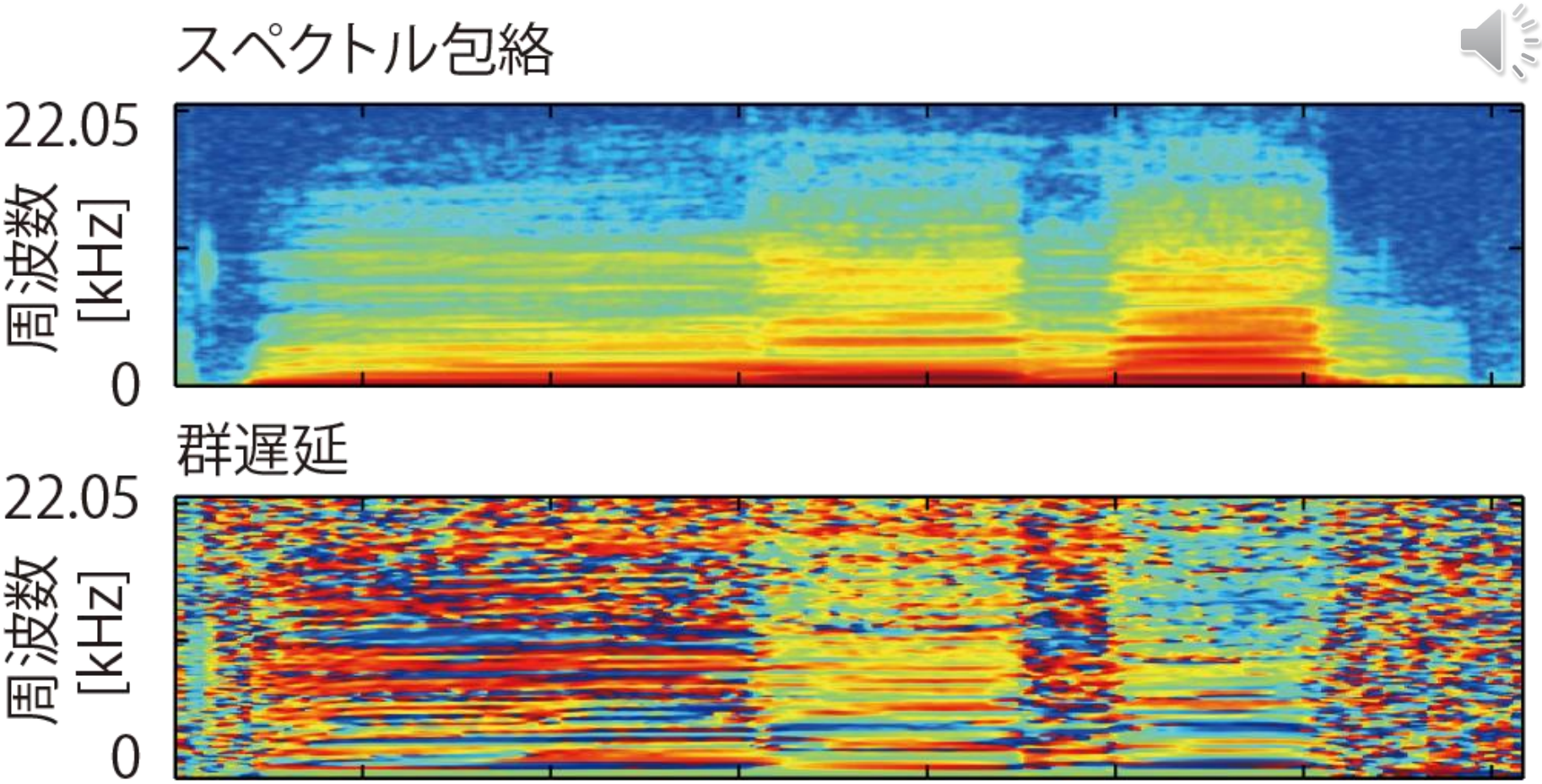
F0適応群遅延を統合して変動を消去



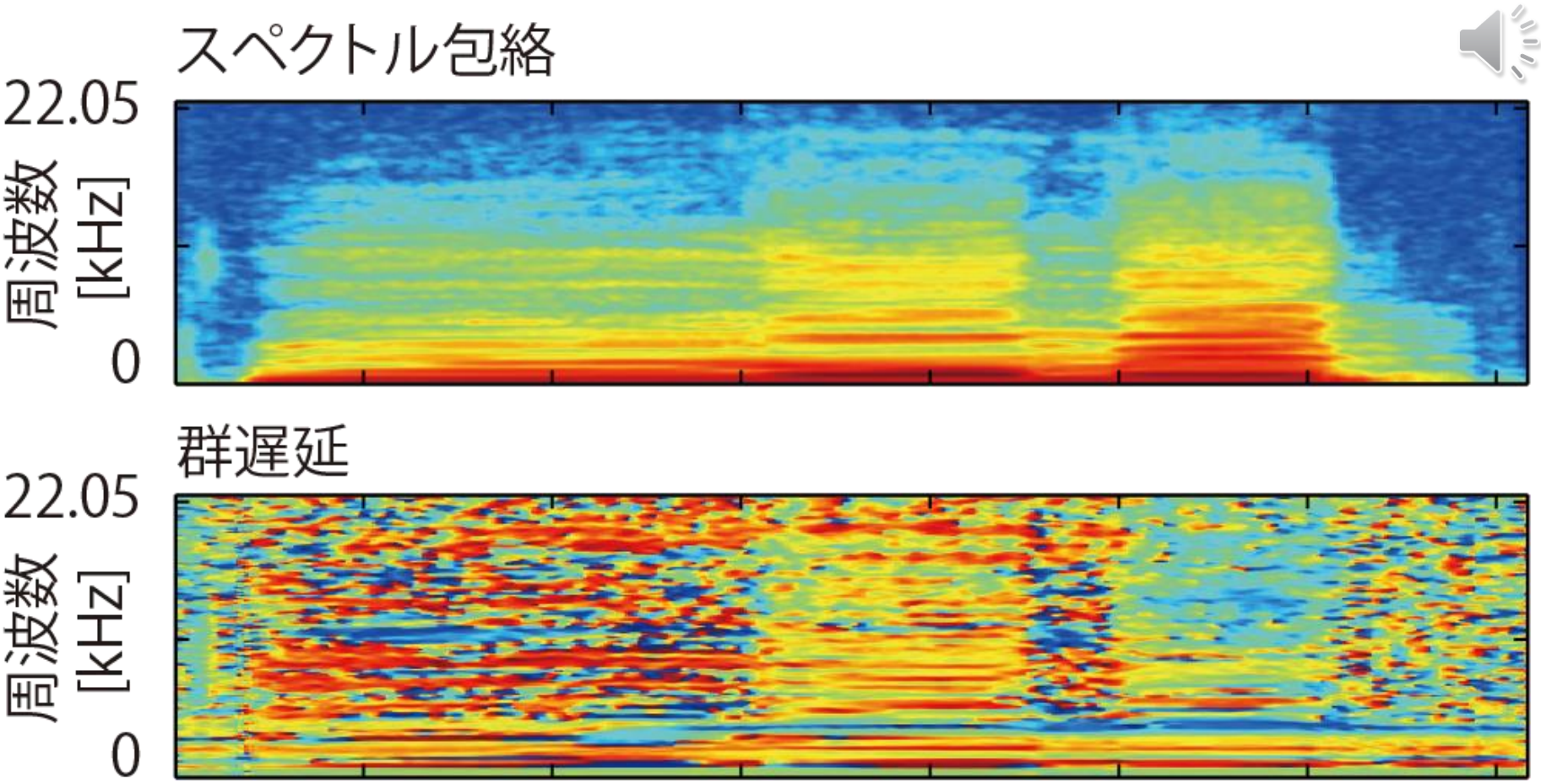
統合(重畳)により

どこで分析しても相対的に同じ形状

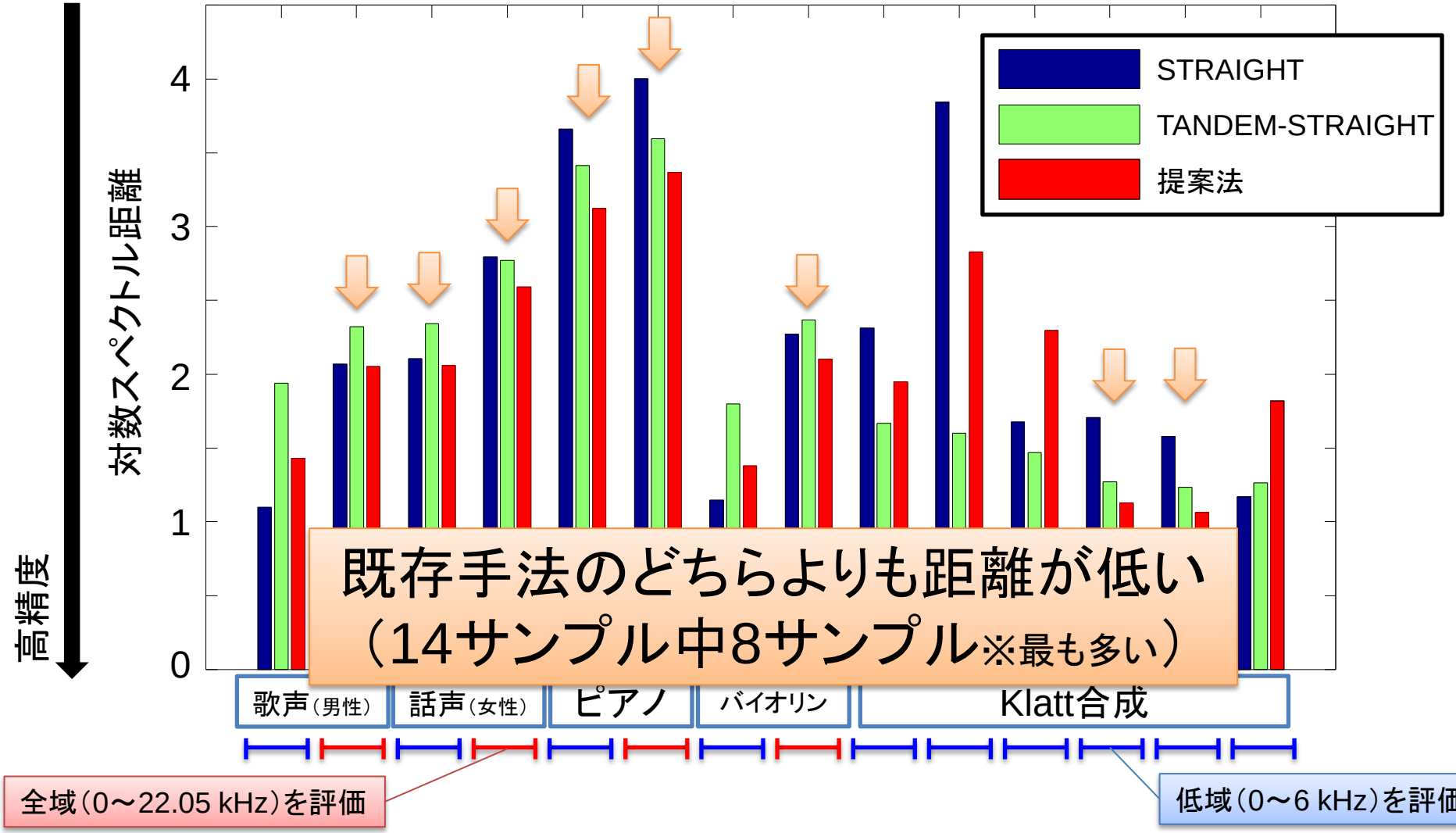
歌声の例（分析結果）



歌声の例（再合成結果）：音高・時間軸も変更可能



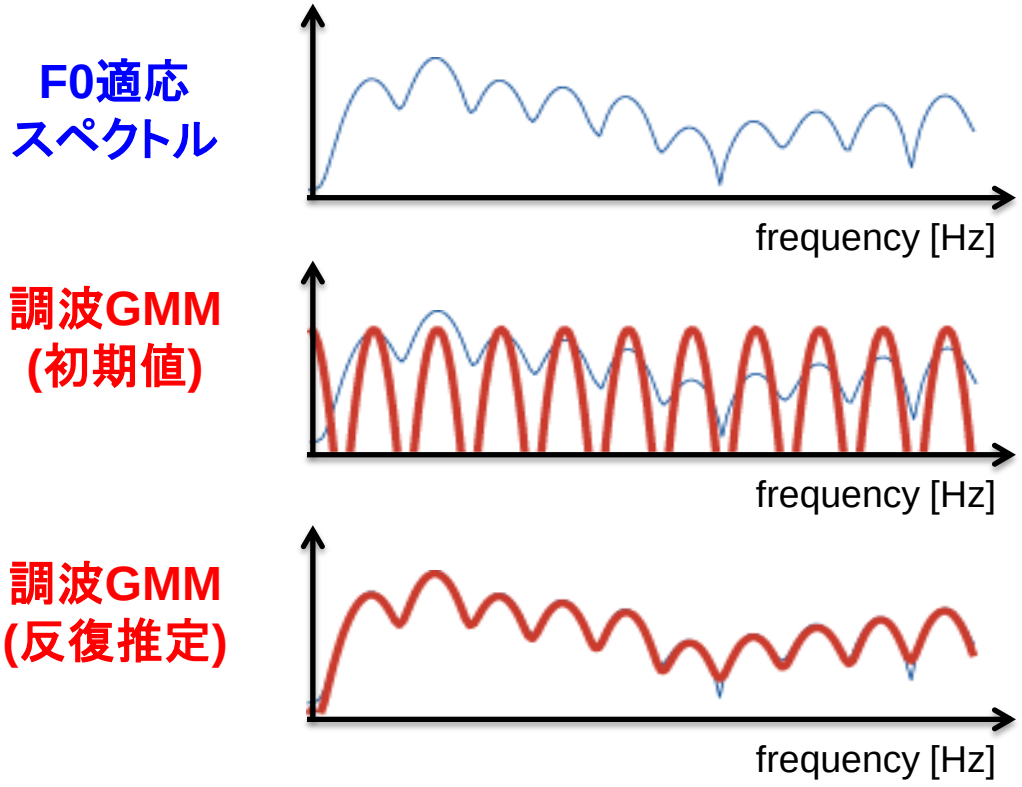
スペクトル包絡の推定精度



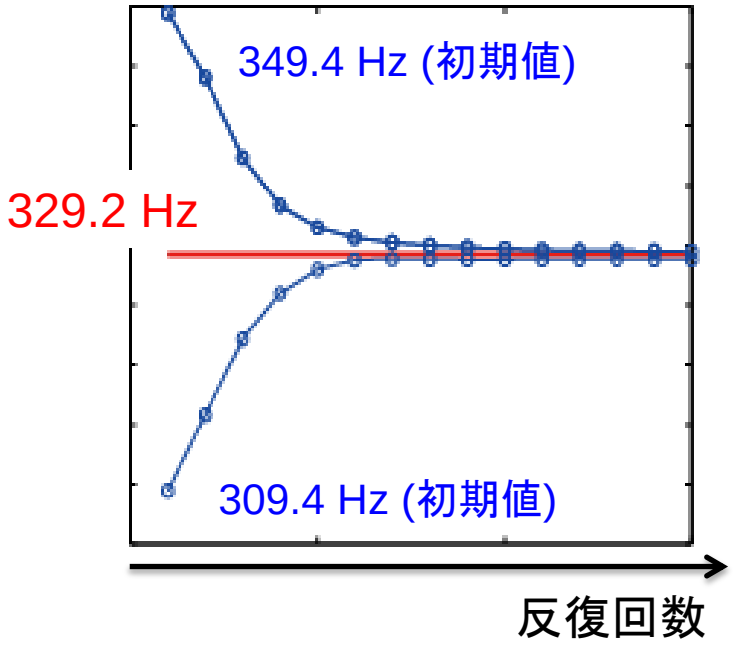
歌声の音高(F0)推定手法

[Nakano and Goto, 2013]

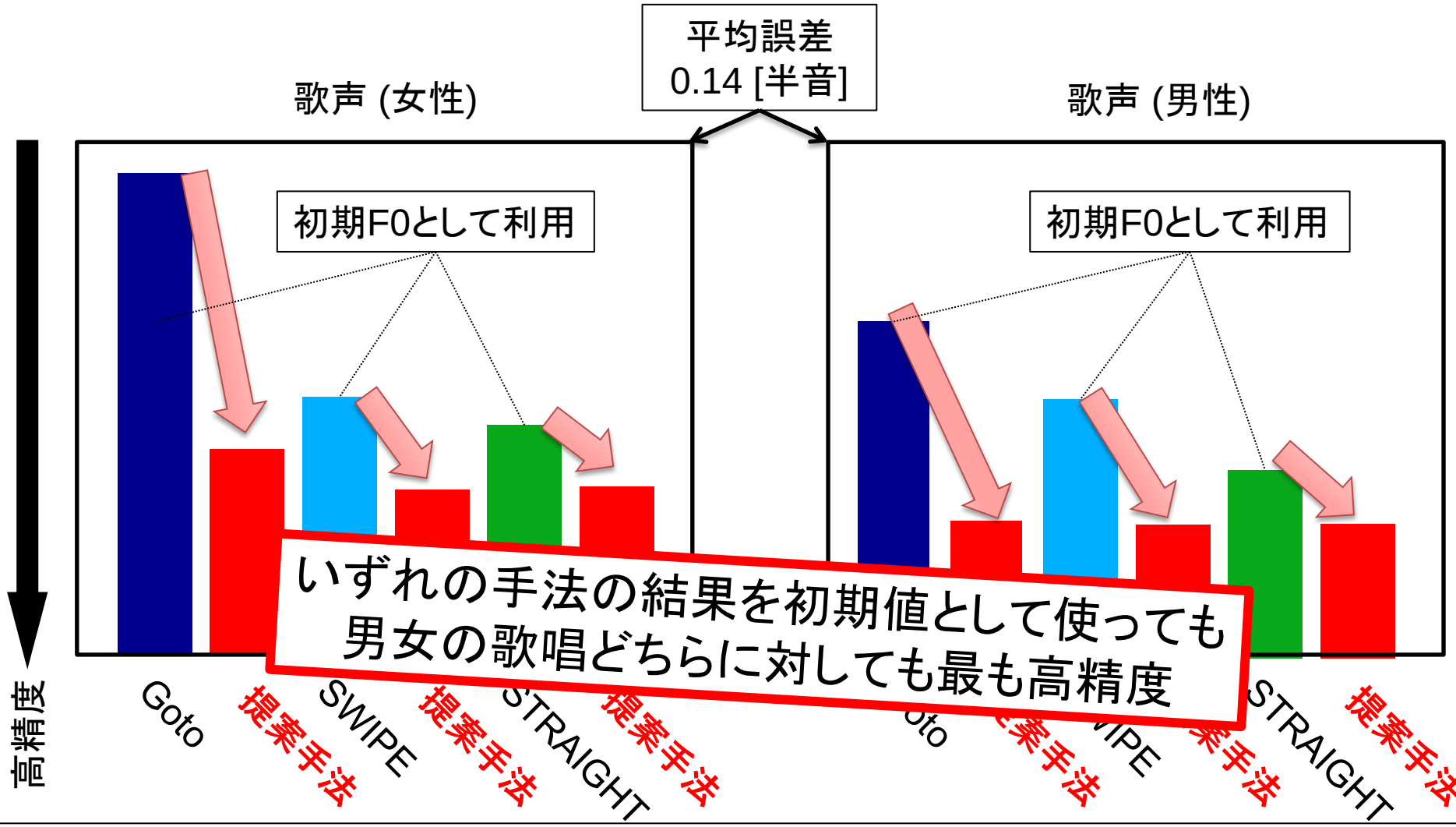
- F0情報からF0適応スペクトルを算出し、調波GMMを用いて**高い精度と時間分解能でF0を再推定**する



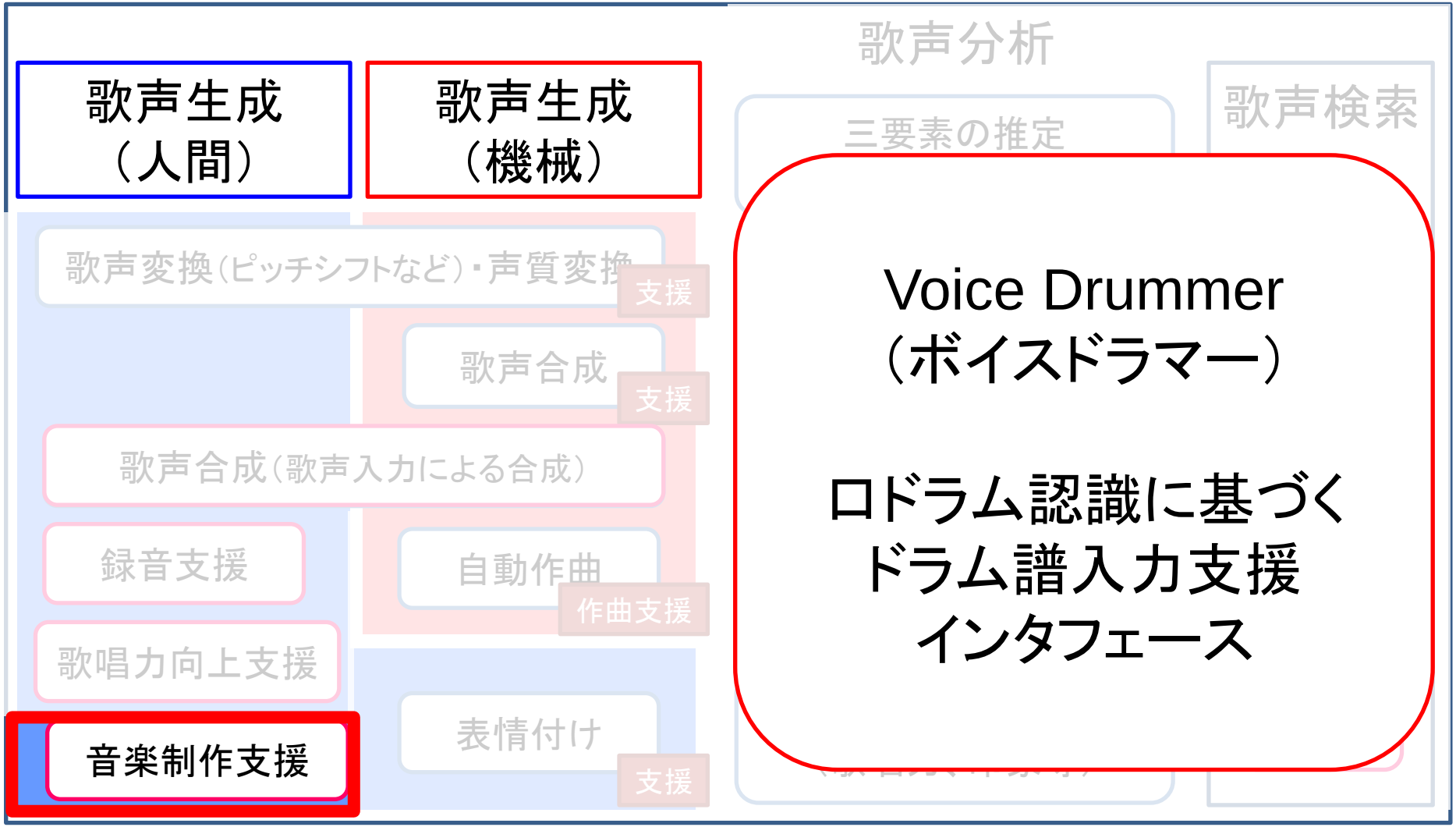
初期値の誤りにもロバスト



既存手法よりも推定誤差を低くできる

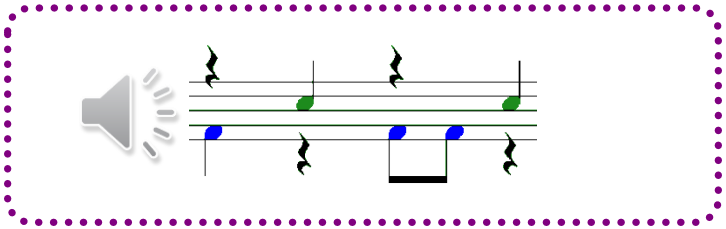


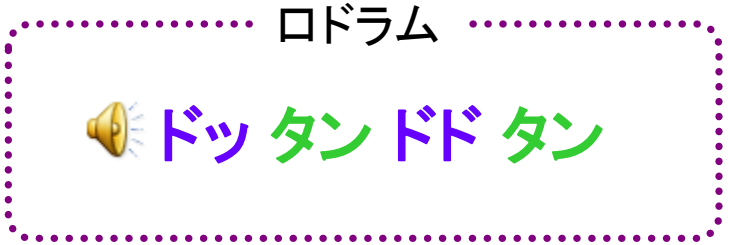
歌声インタフェースの全体像



ロドラムとは

- ドラム音（楽器音）の擬音語表現（リズミカルな歌声）





- ロドラム認識の重要性
  - 楽譜入力の新しい方式の実現
    - ロドラムによる直感的で手軽な楽譜入力
  - 音楽情報検索へ応用可能
    - 「ラララー」等のハミングでの検索はこれまであったが、ドラムパートを検索する目的では使えなかった

Voice Drummer

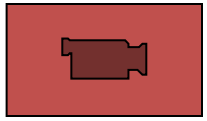
[Nakano, Goto, Ogata, Hiraga, 2005-]

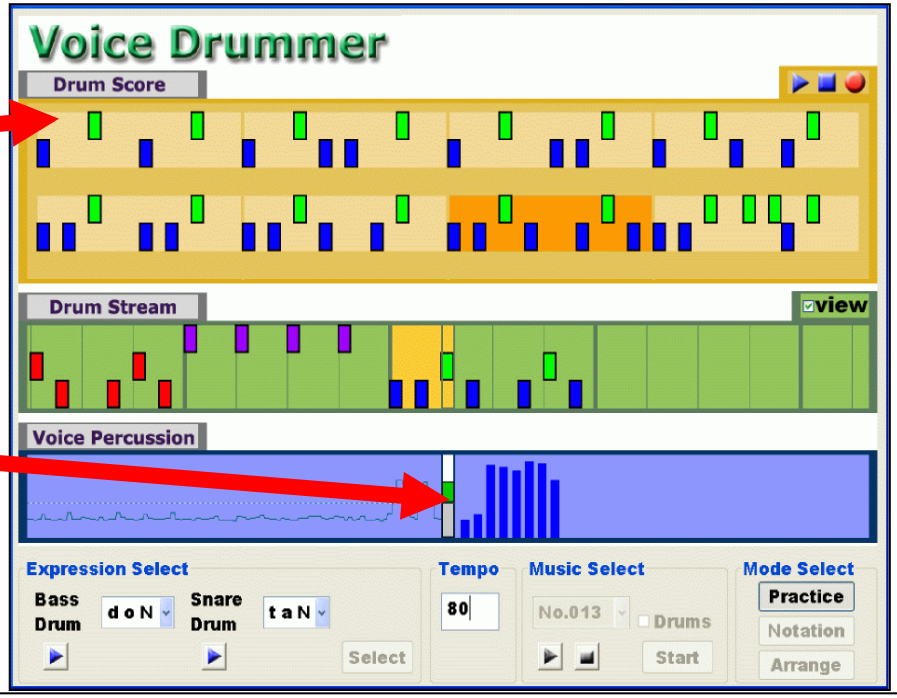
- 口(くち)ドラムによるドラム譜入力システム
  - 口ドラム(ボイスパーカッション)によってドラムパターンを検索
    - ドラム音を真似た「ドンタンドドタン」のような発声
  - 練習することで認識モデルを歌唱者へ適応
  - 既存の楽曲のドラムパートだけを差し替えて編曲

インタラクション

認識結果

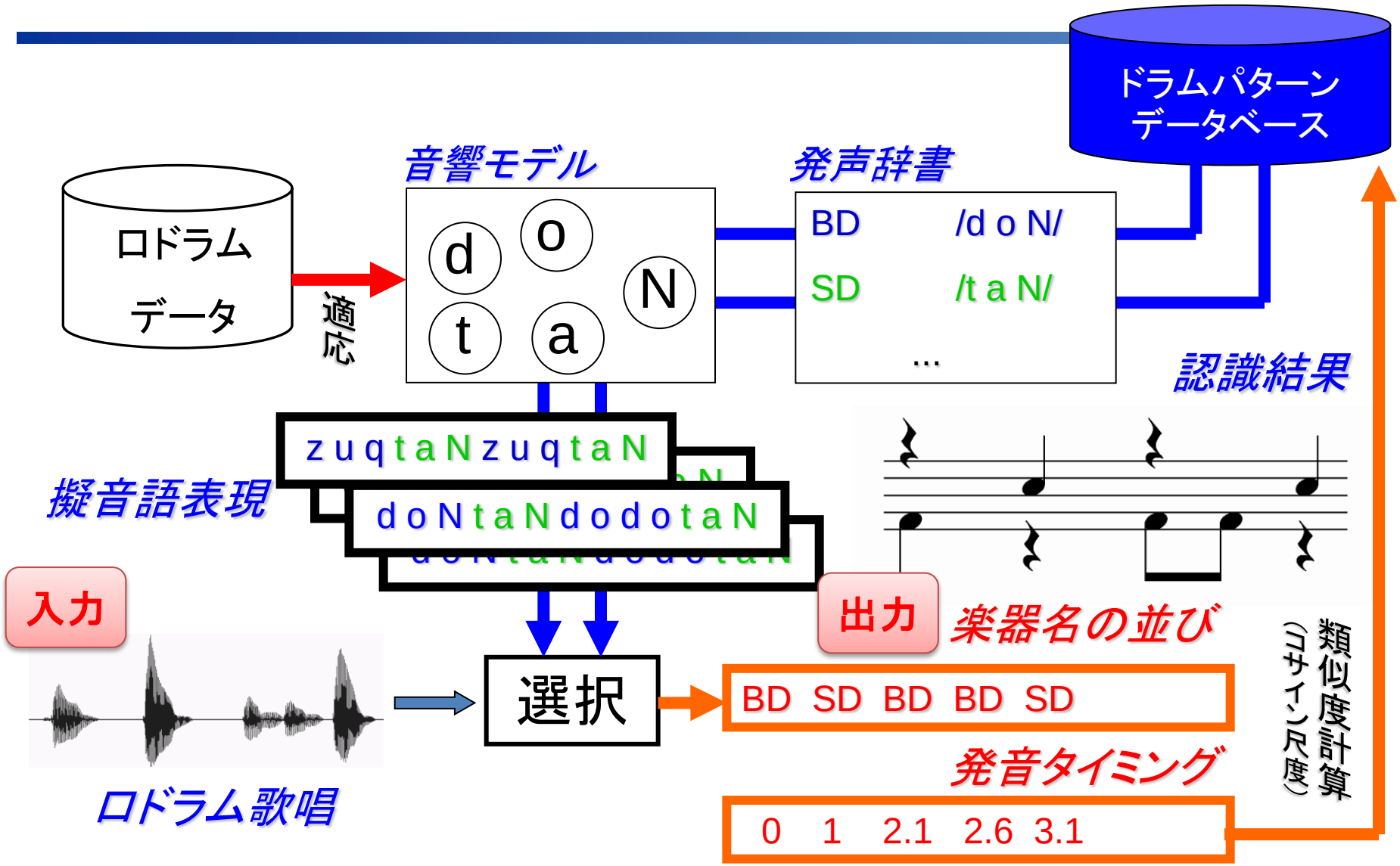
口ドラム入力
「ドンタンドドタン」





口(くち)ドラム認識手法

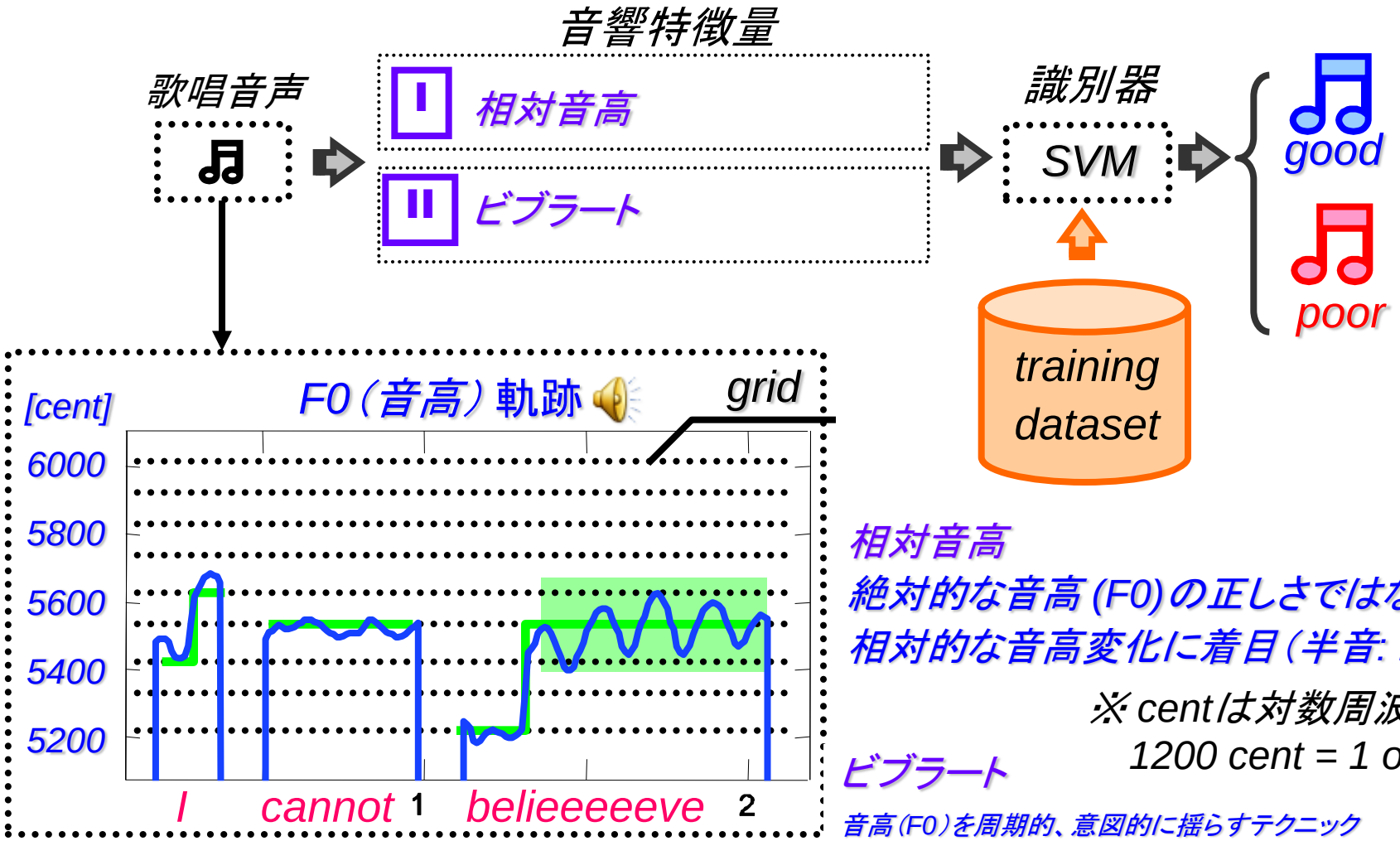
[Nakano, Goto, Ogata, Hiraga, 2004-]



歌声インタフェースの全体像

歌唱力自動評価（楽譜なし）

[Nakano, Goto, Hiraga, 2006-]



$$\text{Grid Frequency}_n = n \times 100 + F \quad (0 \le F < 100)$$

相対音高

絶対的な音高（F0）の正しさではなく、
相対的な音高変化に着目（半音: 100cent）

※centは対数周波数
1200 cent = 1 octave

ビブラート

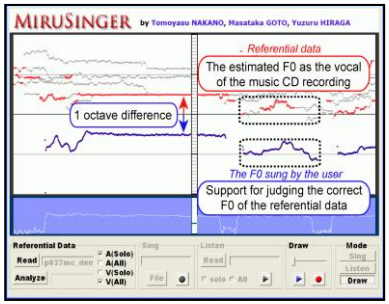
音高（F0）を周期的、意図的に揺らすテクニック
この区間を自動的に推定する

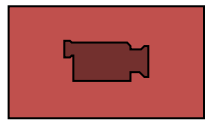
MiruSinger

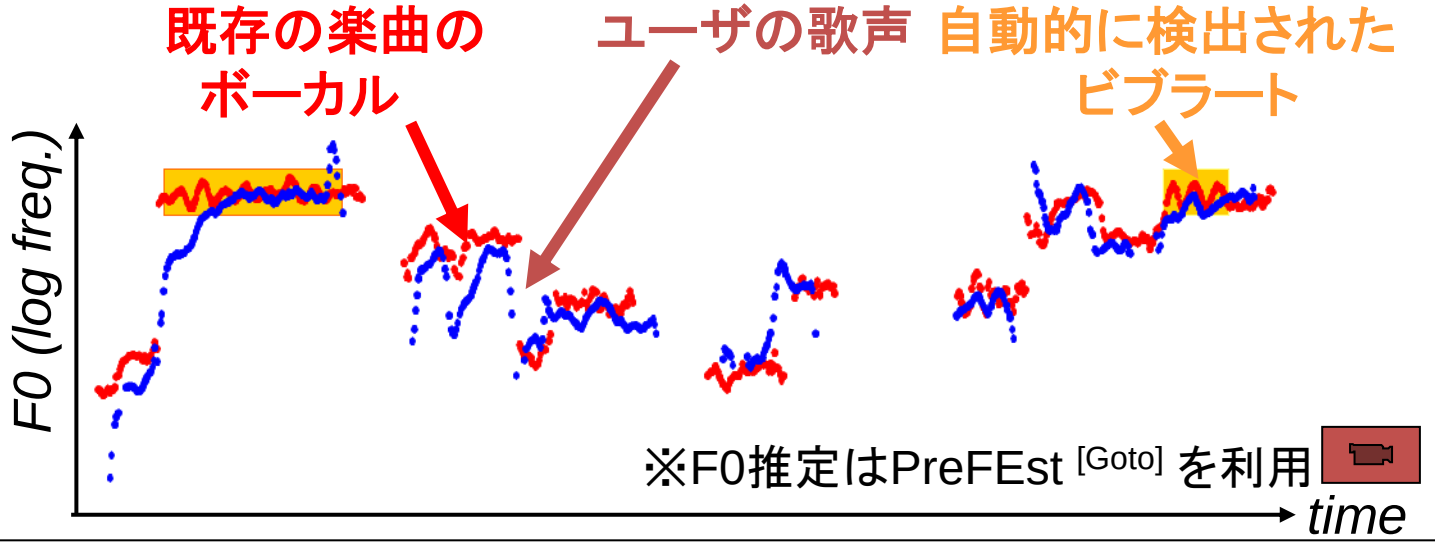
[Nakano, Goto, Hiraga, 2007-]

- 歌唱力向上支援インタフェース
  - 既存曲のボーカルの歌い方に忠実に歌いたい！
  - 混合音中のボーカルを分析して可視化して**修正可能**
  - それに合わせてユーザの歌声も比較表示
  - リアルタイムに音高（F0）が可視化され、ビブラート区間も表示

インタラクション







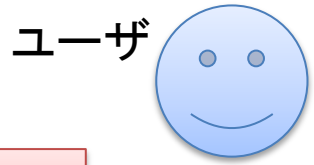
歌声インタフェースの全体像

まとめ

- 「歌声インタフェース」と名付けた研究アプローチ
  - 歌声信号処理に基づくインタフェース構築やインタラクションによって人々の音楽生活をより豊かに
- 技術が世の中で広く活用されるために
  - 対象ユーザの特性に合わせたインタフェース構築
    - ユーザ視点での問題発見（インタラクションデザイン）
      - VocaRefiner：1度で完璧に歌えないユーザ
      - VocaListener, Voice Drummer:楽譜に不慣れなユーザ
      - MiruSinger:自分の好きな歌手の歌で練習したいユーザ

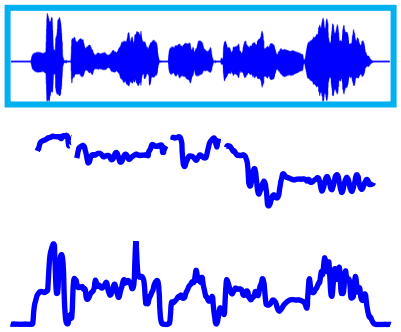
歌声インタフェースの未来

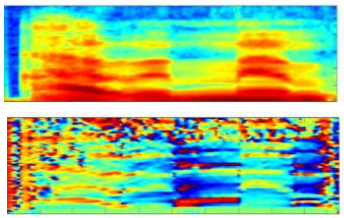
- 人間と技術・インタラクションが
  相互に成長できる未来



ユーザ自身の歌唱力や表現力等が向上

必要な機能の実現：歌声インタフェース自体の機能を豊かで高度に





実現

必要性

歌声信号処理

インタフェース構築（インタラクション）

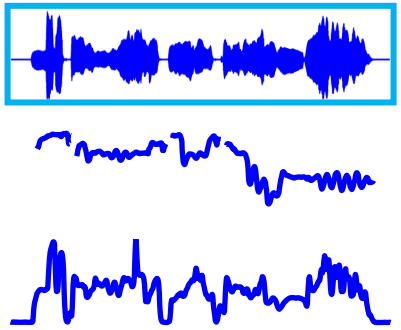
歌声インタフェースの未来

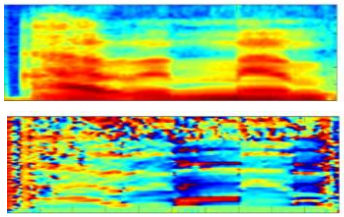
新たな歌唱表現を追求したり
音楽に関する理解が深まったり
できるようになる

単に技術のみが
高度化していく
のではない未来
が切り拓ける

ユーザ自身の歌唱力
や表現力等が向上

必要な機能の実現：歌声インタフェース自体
の機能を豊かで高度に





実現

必要性

歌声信号処理

インタフェース構築（インタラクション）

スペシャルセッション 音声B/音声A/聴覚
[ここまで来た声質変換技術 ー実用可能性の視点からの現状認識と将来展望ー]

歌声インタフェース:
歌声を対象とした信号処理と それに基づくインタフェース構築

中野 倫靖，後藤 真孝
（産業技術総合研究所）

2013年9月27日
日本音響学会 2013年秋期研究発表会（講演番号3-7-3）