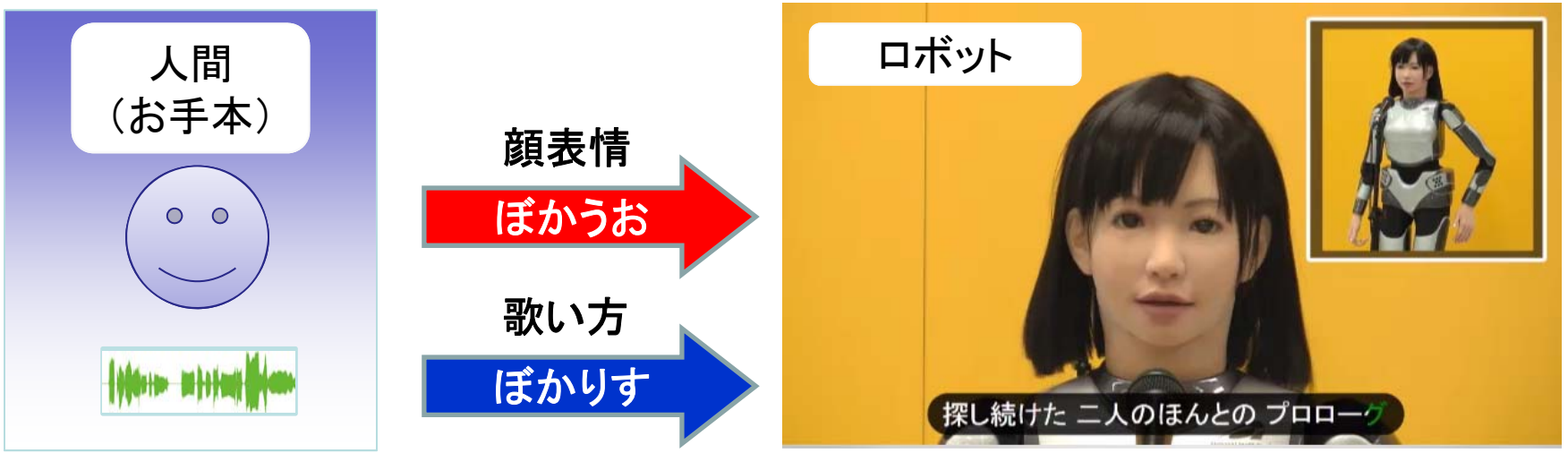

VocaWatcher (ぼかうお)
人間の歌唱時の表情を真似る
ヒューマノイドロボットの顔動作生成システム
中野 倫靖, 後藤 真孝, 梶田 秀司,
松坂 要佐, 中岡 慎一郎, 横井 一仁
(産業技術総合研究所)
SIGMUS94 & SIGSLP90 スペシャルセッション「歌声情報処理最前線!!」第2部 (2012/2/3

ヒューマノイドロボット

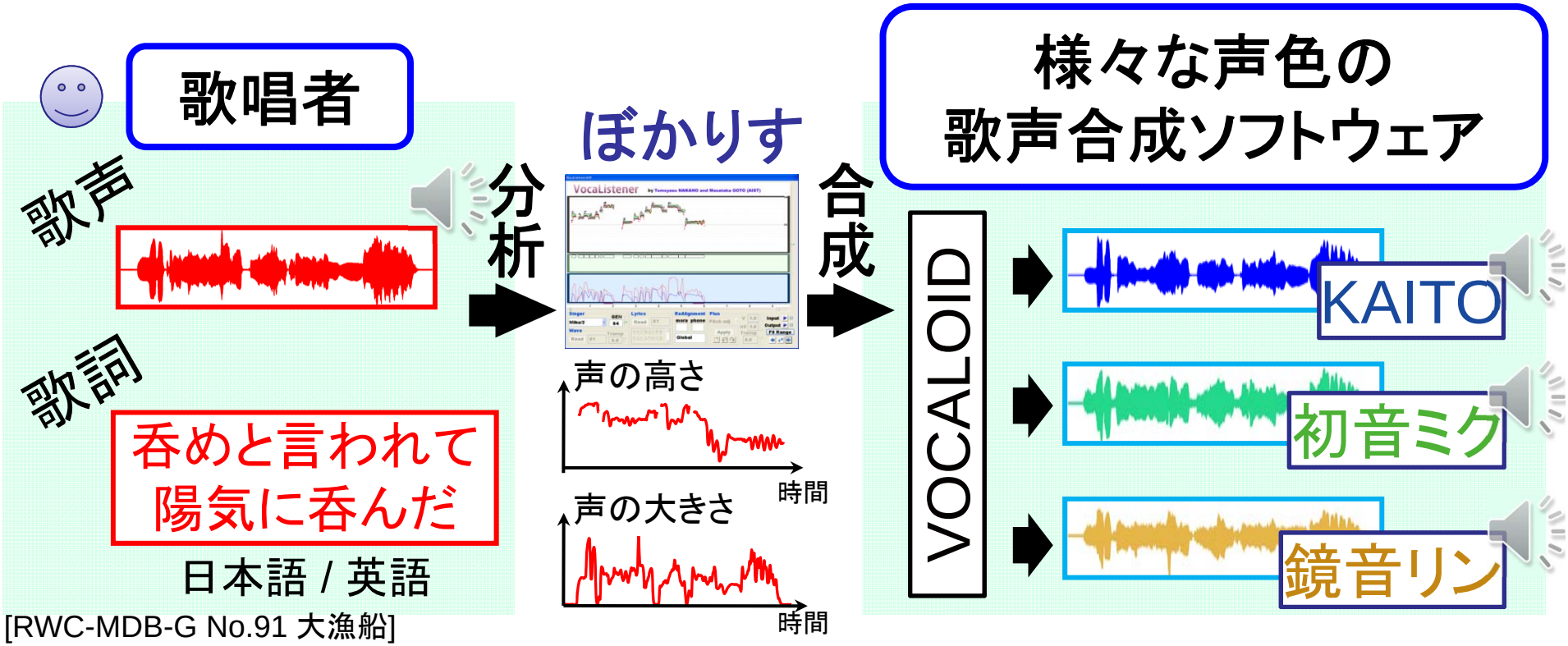
HRP-4C 未夢が自然な歌声と表情で歌唱

- お手本歌唱を真似る二つの技術により実現
  - **VocaListener（ぼかりす）**
    - 「歌い方」を真似て歌声合成する技術
  - **VocaWatcher（ぼかうお）**
    - 「顔表情」を真似てロボットの顔動作を生成する技術



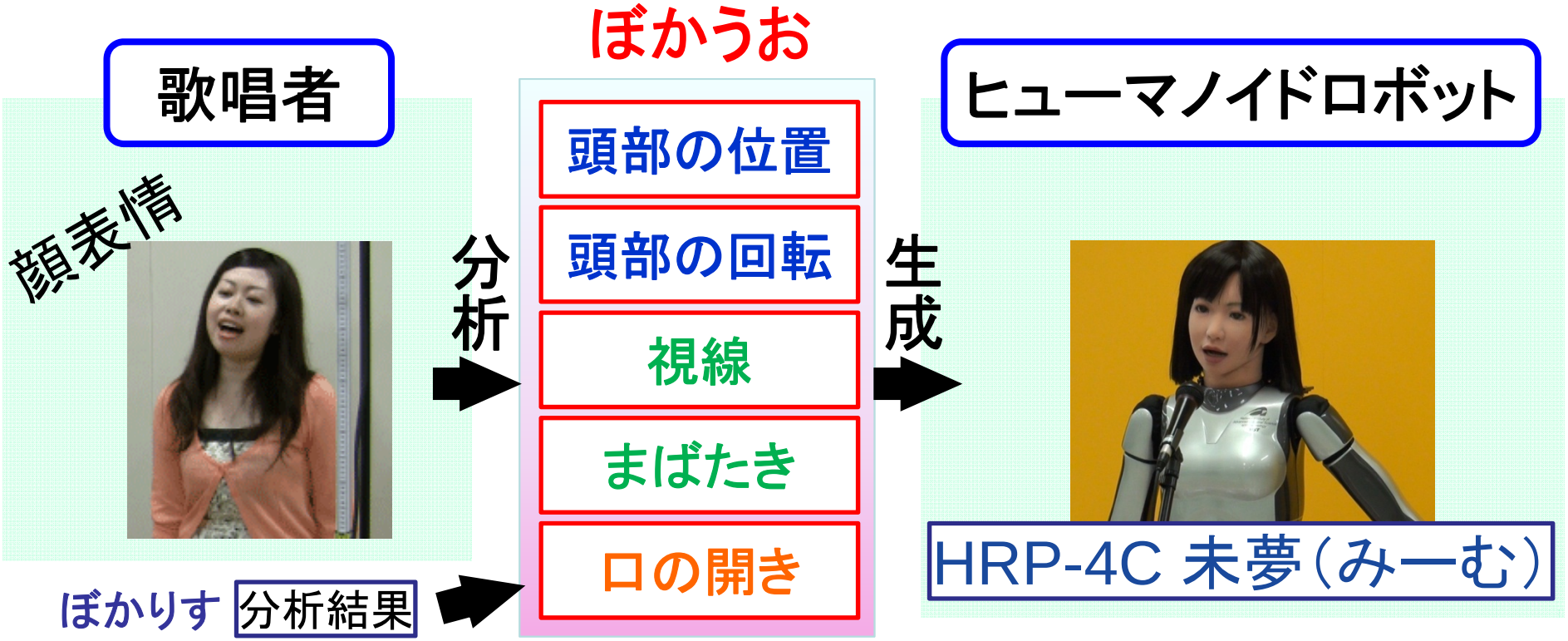
VocaListener（ぼかりす） [中野, 後藤 (2008～)]

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



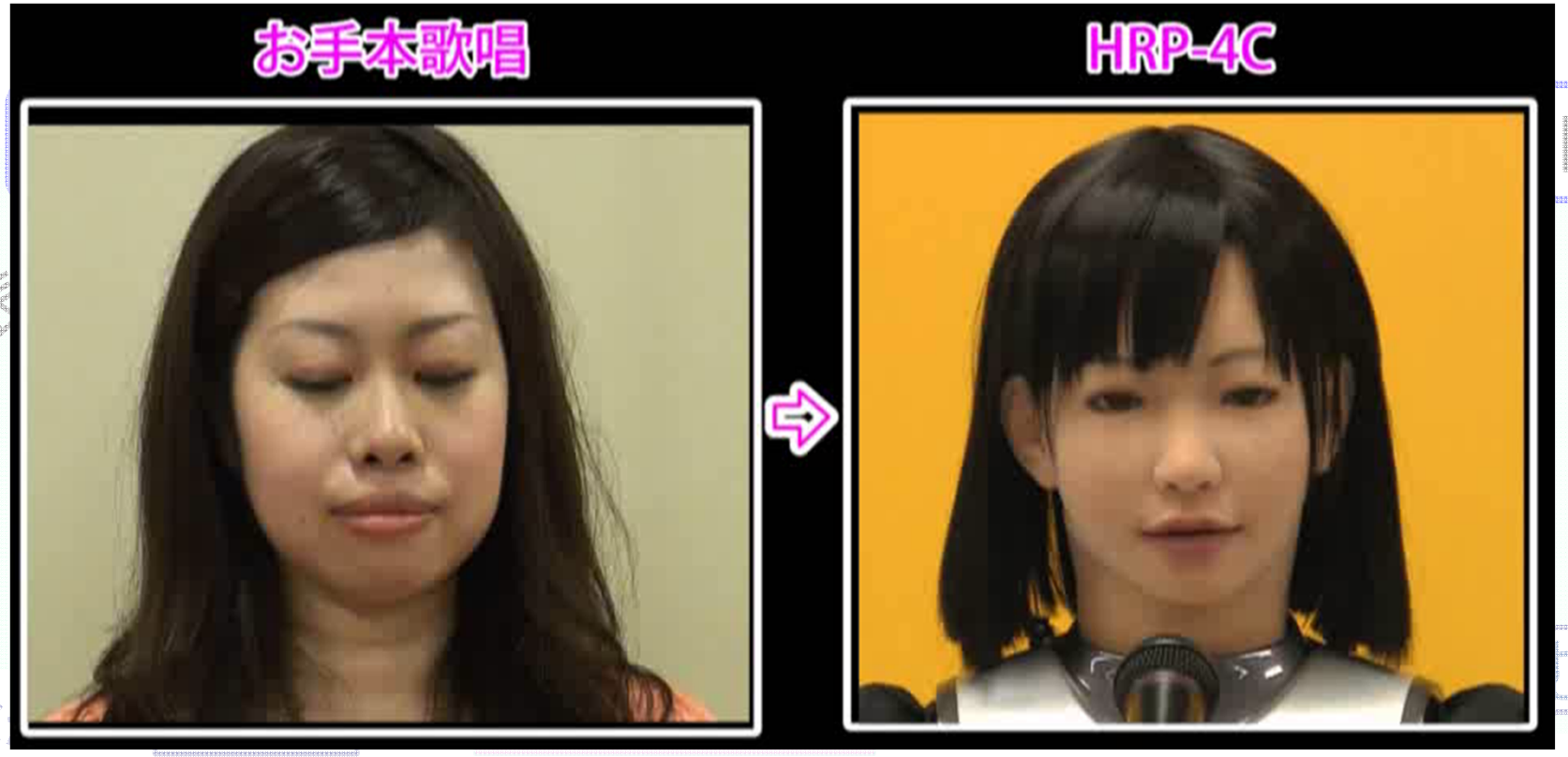
VocaWatcher（ぼかうお）

- 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



VocaWatcher（ぼかうお）

- 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

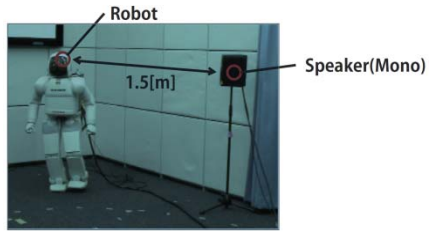


研究背景:「歌うロボット」の意義

- **ヒューマノイドロボット**の有望な応用事例
  - **人々の関心を惹き付けやすい**
    - 我々の展示経験(CEATEC2010、産総研オープンラボなど)
    - アイドルロボット構想(経済産業省「技術戦略マップ2010」)
  - **歌声合成とロボットの高い親和性**
    - ロボット技術のエンターテインメント分野への展開
- 人間との自然なコミュニケーションの実現
  - まずは**人間を「真似る」ことで「自然さ」を追求**
  - 未来社会において不可欠な基礎技術
    - 声の音響信号処理、顔の画像処理、ロボット制御の技術

先行研究：（二足歩行）ロボットと音楽

- Sony QRIO (4体) [Kuroki *et al.*, 2003]
  - ビブラート付アカペラコーラス



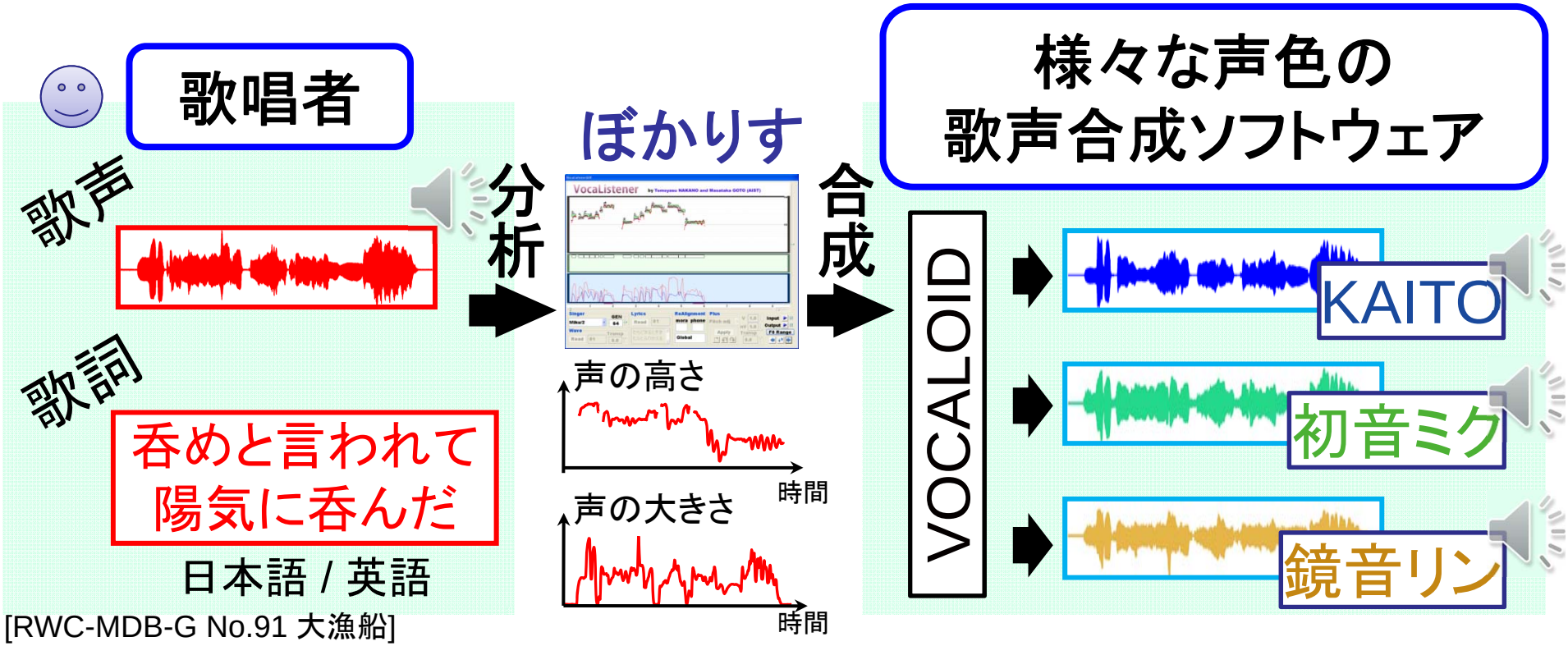
- Honda ASIMO [Murata *et al.*, 2008]
  - ビートトラッキングとオンボードマイクを用いた歌唱

- 産総研 HRP-4C [Tachibana *et al.*, 2010]
  - VOCALOIDと組み合わせた表情制御と歌唱
  - CEATEC2009で展示

人間を真似て、歌声分析と組み合わせた例はなかった

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（声色）を手軽に切り替えて合成できる



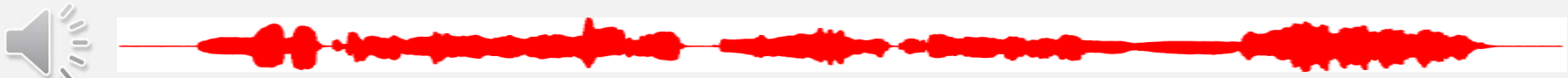
[RWC-MDB-G No.91 大漁船]

VocaListener（ぼかりす） [中野, 後藤 (2008～)]

[RWC-MDB-G-No.31 大漁船]

VocaListener（ぼかりす） [中野, 後藤 (2008～)]

【入力】 人間の歌声信号と歌詞

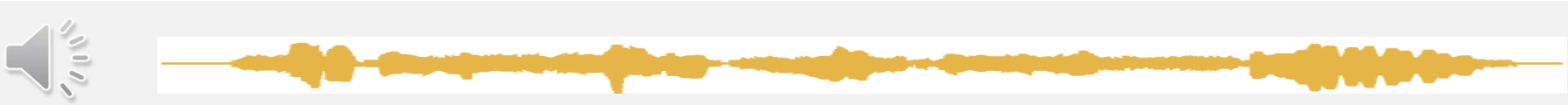


歌詞 「呑めといわれて陽気に呑んだ」

1．歌声用の音響モデル（HMM)を用いて
歌詞の発声タイミングを得る

（モーラ） N no me to i wa re te yo u ki ni no N da

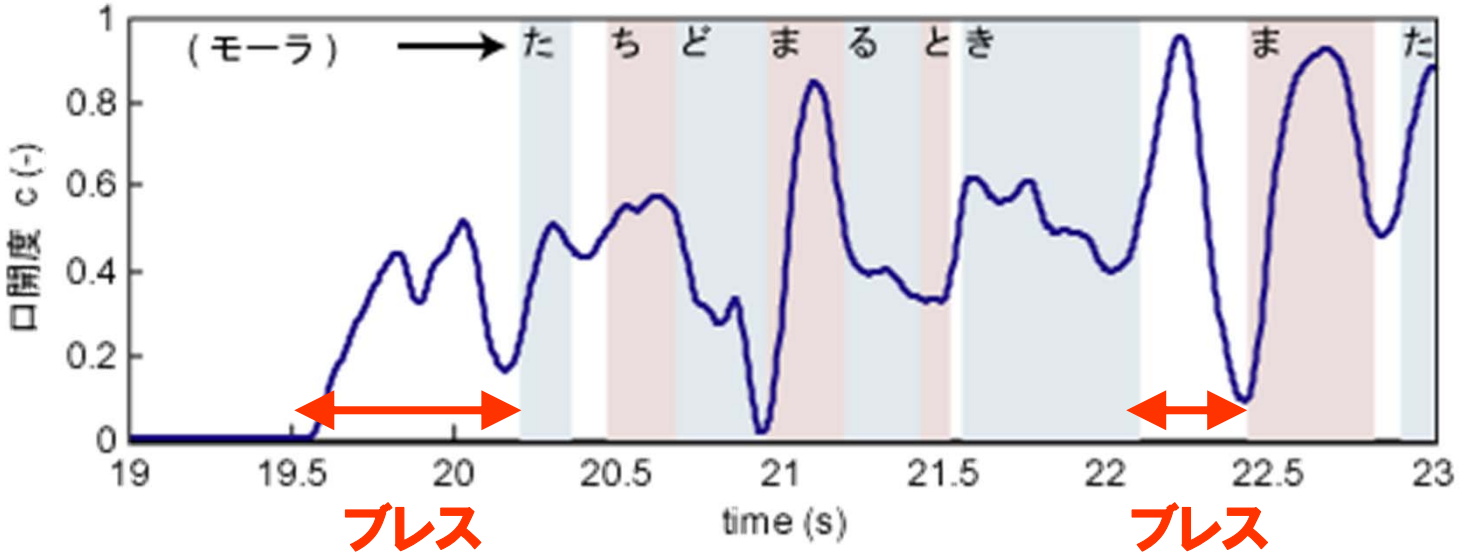
2．歌声合成パラメータ（音高と音量）を
反復更新して出力を目標に近づける



ブレス検出・合成を新規開発

- 人間の歌手は**歌唱中にブレス（吸気）**する
  - 顔動作を真似ると口を開けてしまう
  - **問題点: 口が開くのみで音がないと不自然**

    ⇒ブレス音を真似て歌声合成

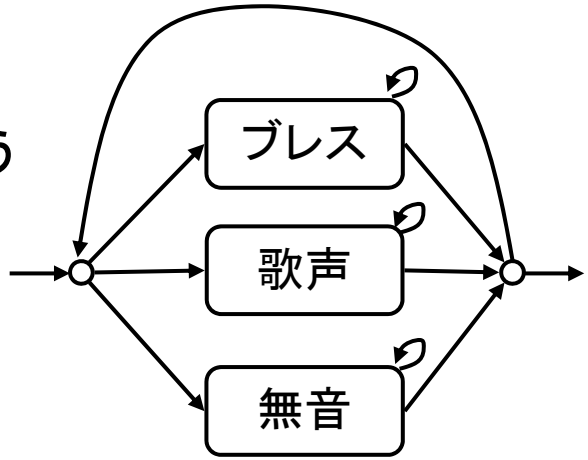


ブレス音の合成なし／合成した場合

HMMによるブレス検出 [中野, 緒方, 後藤, 平賀 (2008～)]

ブレス/歌声/無音の3種のHMM

- 「歌声」区間は、子音・母音は区別せずに扱う
- 3状態16混合



特徴量

- MFCC (12), ⊿MFCC (12), ⊿Power (1)
- 分析条件
  - 16kHzサンプリング、ハミング窓（窓幅 25ms、シフト 10ms）

学習用歌声データ

- **1488個のブレス、438秒**
  - **16人が歌った27曲**（RWC研究用音楽データベース）とAISTハミングデータベースから女性2人分
    - » 男性 15 曲、女性 14 曲
    - » 日本語歌詞 20 曲、英語歌詞 9 曲

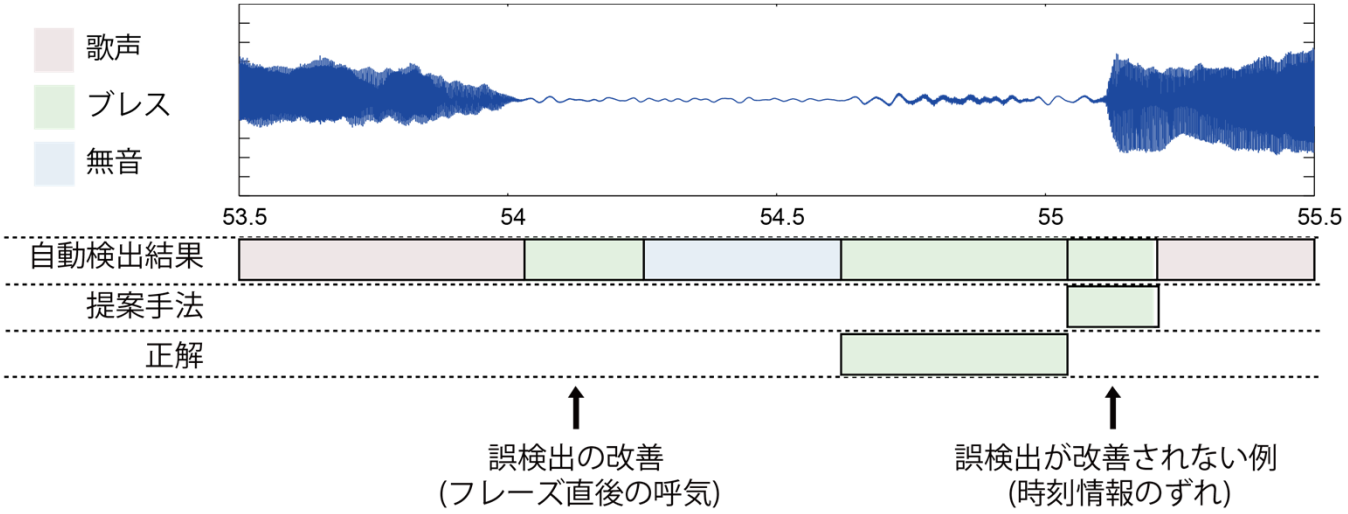
ブレス検出結果

- 先行研究の知見[中野 他, 2008] に基づく検出精度の向上
  - HMMによる検出のみでは、再現率は高いが呼気や子音で誤検出する
    - ⇒ 歌唱フレーズの直前以外の結果を削除
    - ⇒ 継続時間長が50ms～1225ms以外の結果を削除

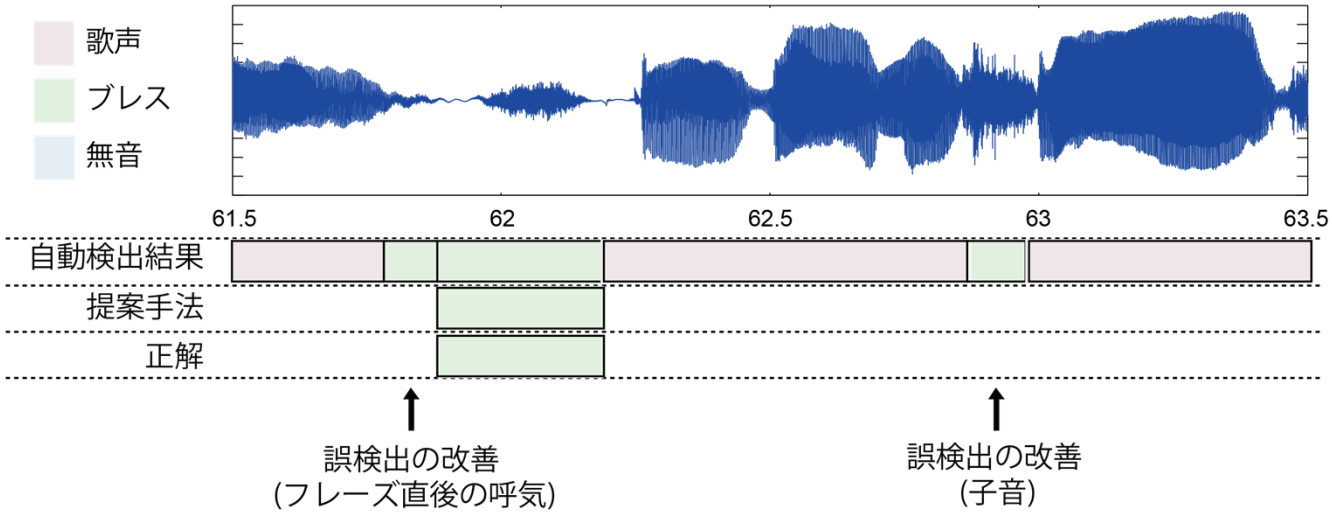
- 実験条件
  - 楽曲：RWC-MDB-P No.7 PROLOGUE
  - 歌唱者：日本人女性1名

- 検出結果
  - 従来：再現率 100%, 精度 66.25%
  - **改良：再現率 100%, 精度 100%** ※ブレス位置のみの結果であり検出時刻のずれを無視した場合

ブレス検出結果



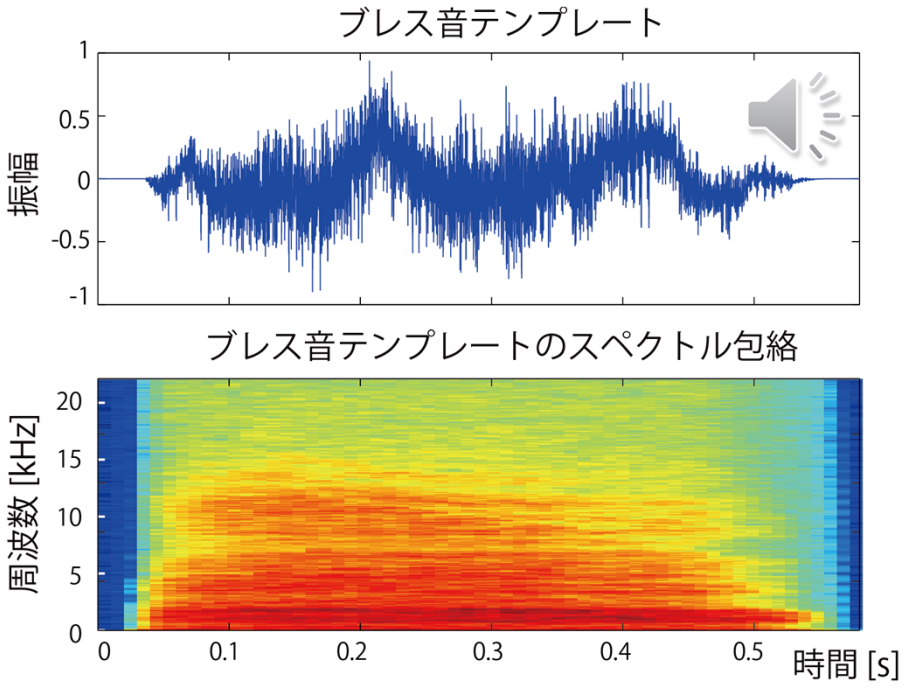
※歌唱者によっては
時刻誤りが多かった



※同一歌唱者の
別の曲でも
良く動作した

ブレス合成手法

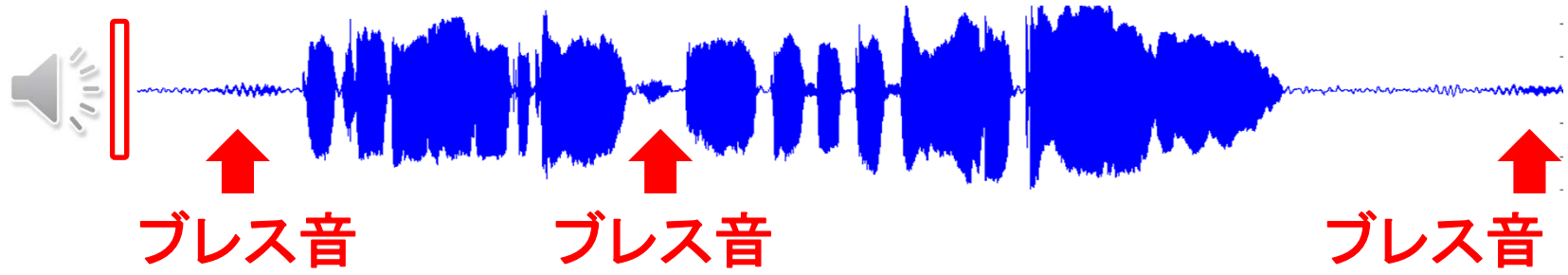
- ソース・フィルタ分析に基づく合成
  1. ブレス音テンプレートを合成
     - 歌声合成ソフトウェアから手作業で合成
  2. スペクトル包絡を推定
  3. ブレス検出時刻で合成
     1. テンプレートの包絡
        を時間伸縮
     2. 音量を反映
     3. インパルス応答と
        ガウス雑音を畳み込む



ブレス合成結果

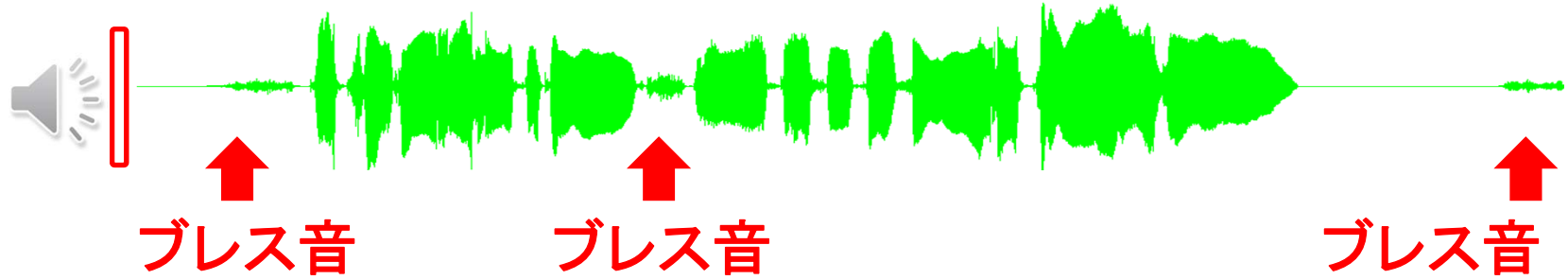
- お手本歌唱

[曲：RWC-MDB-P No.7 PROLOGUE] [歌唱: サリヤ人 様]



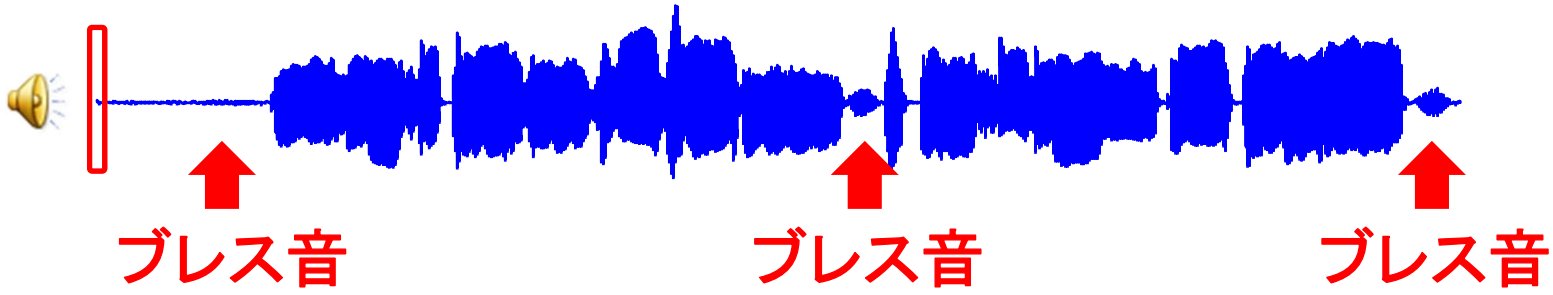
合成歌唱

[歌声合成ソフトウェア：Vocaloid2 初音ミク]

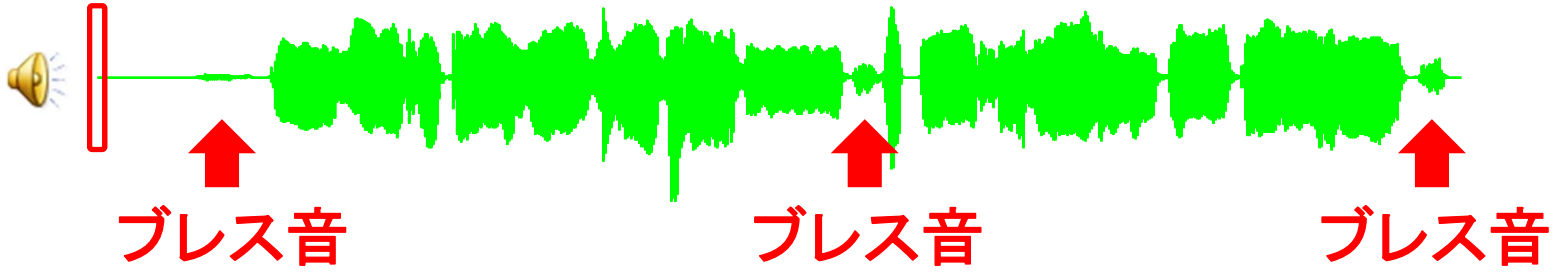


ブレス合成結果（Mix）

- お手本歌唱 [曲：Packaged by kz 様] [歌唱: サリヤ人 様]

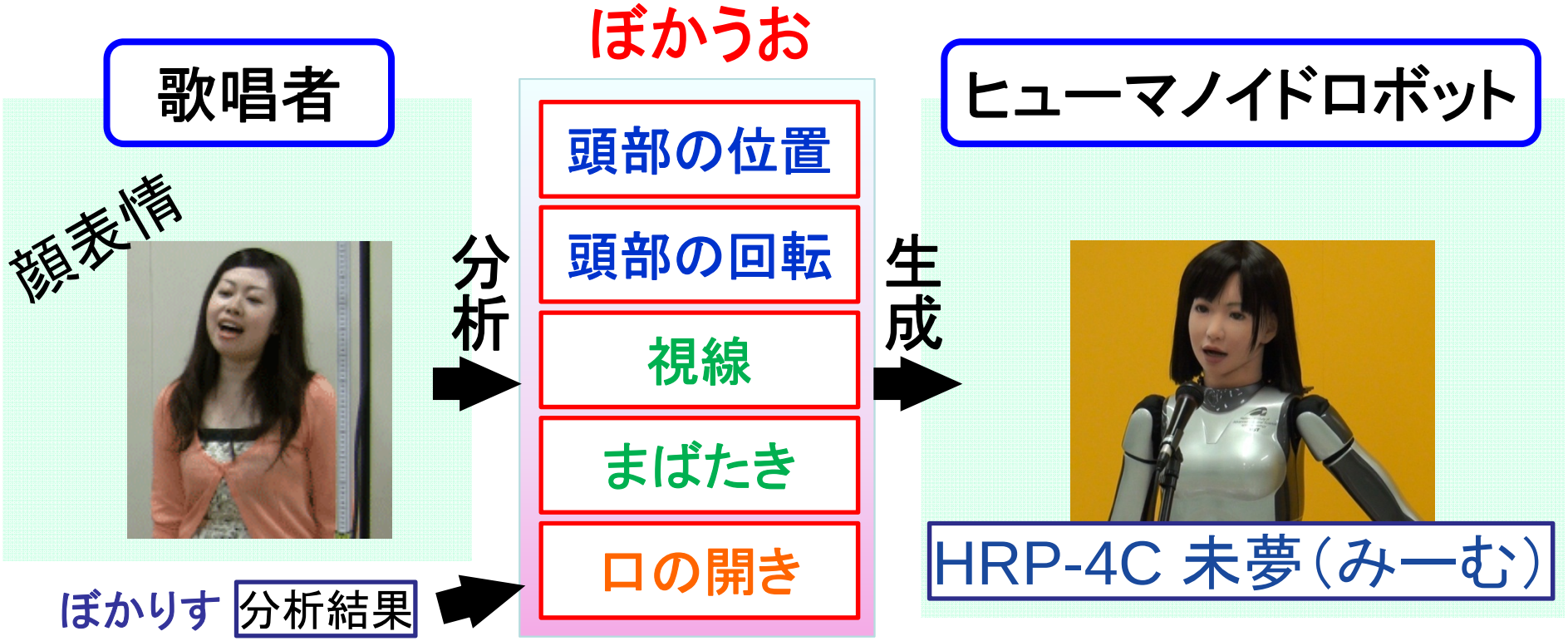


合成歌唱 [歌声合成ソフトウェア：Vocaloid2 Megpoid]

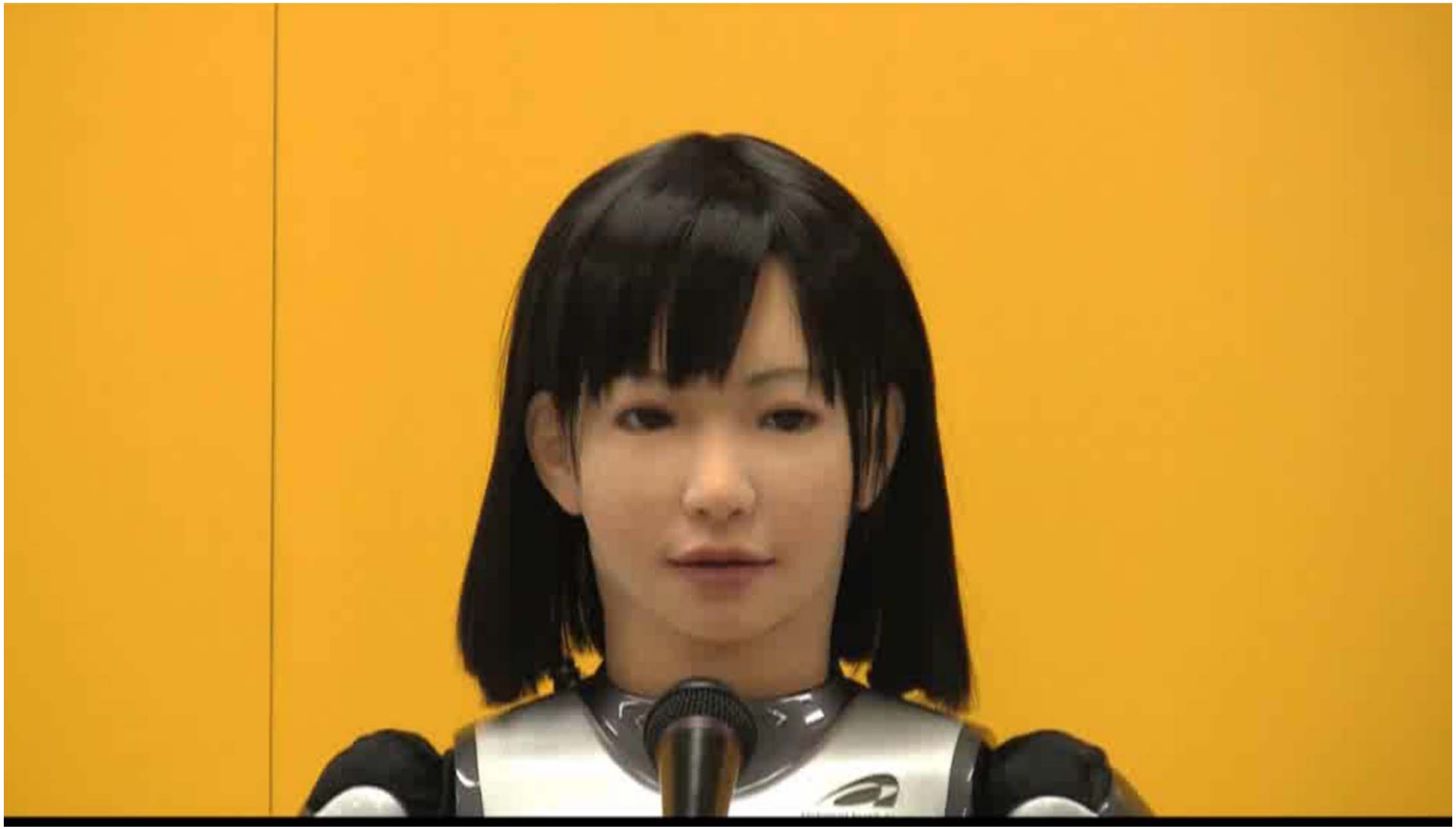


VocaWatcher（ぼかうお）

- 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

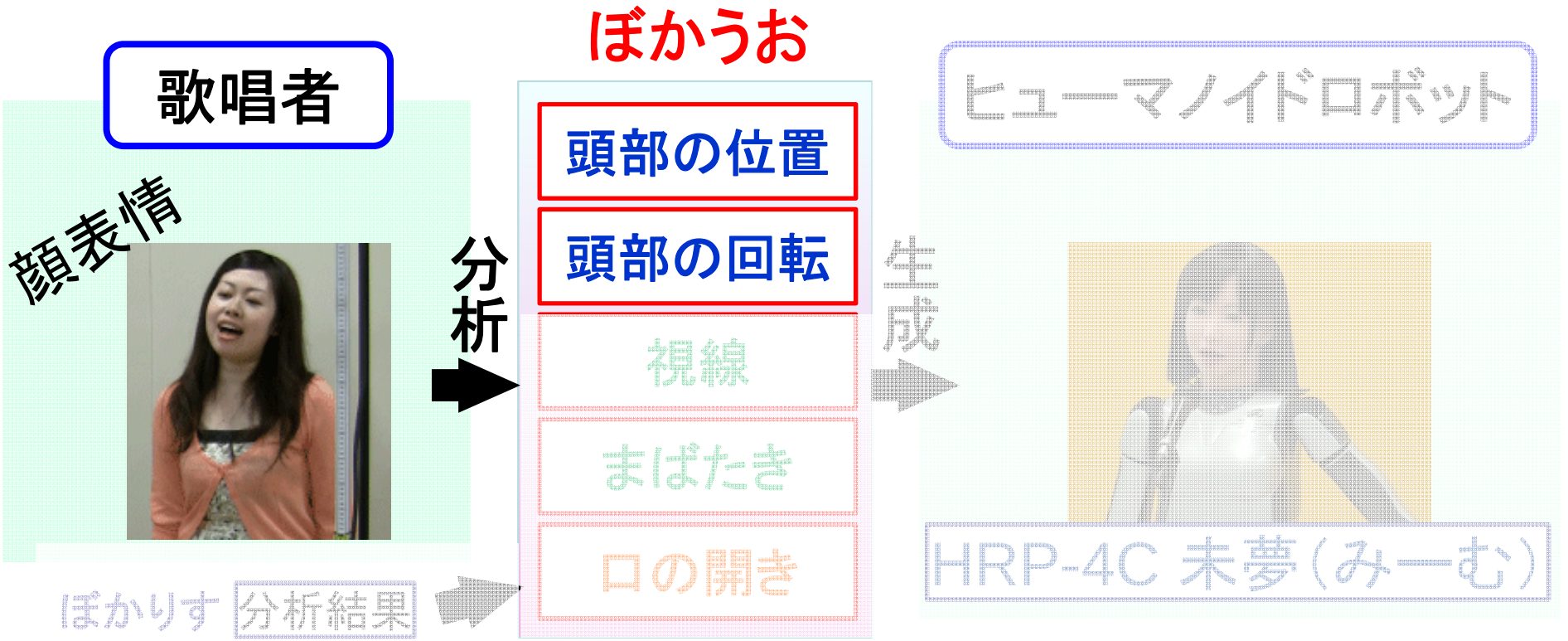


VocaWatcher（ぼかうお）



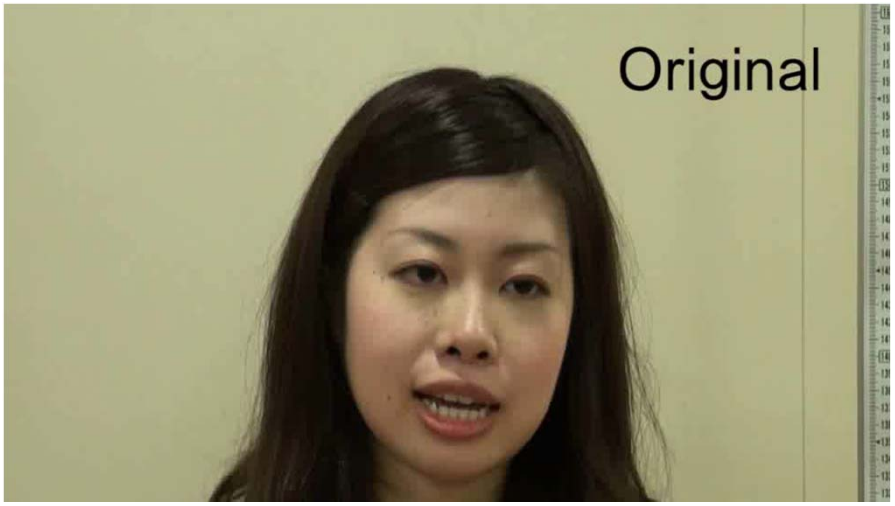
VocaWatcher（ぼかうお）

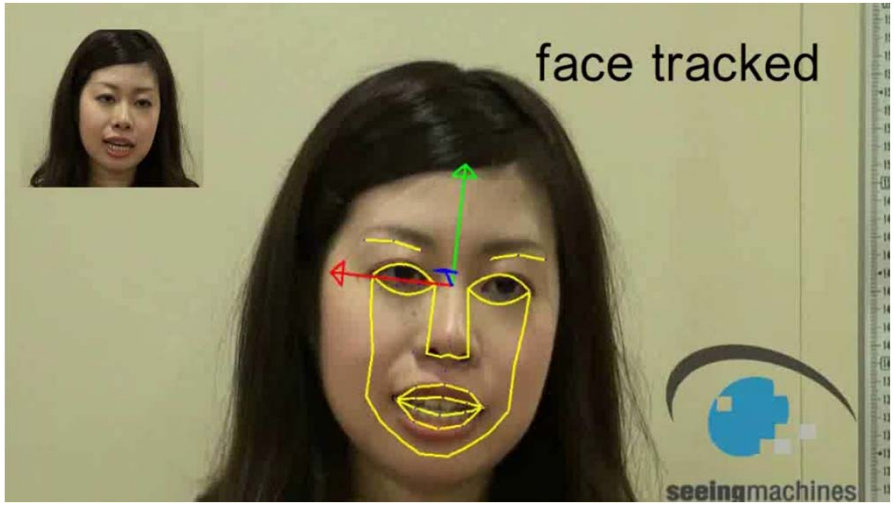
ビデオ映像からの画像処理



ビデオ映像からの顔追跡

- 頭部の位置と回転を検出
  - 市販ソフトウェアを使用
    - faceAPI ver.3.2.6
      - Seeing Machines社, http://www.seeingmachines.com

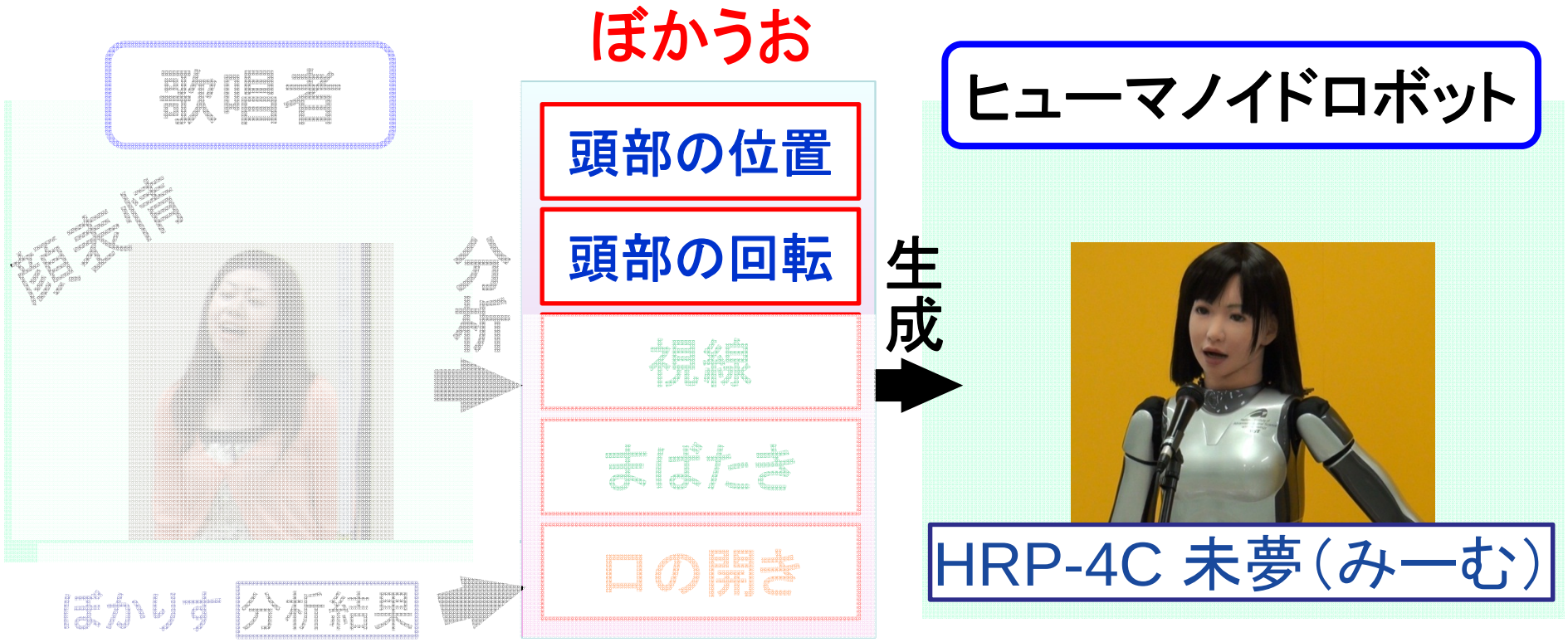




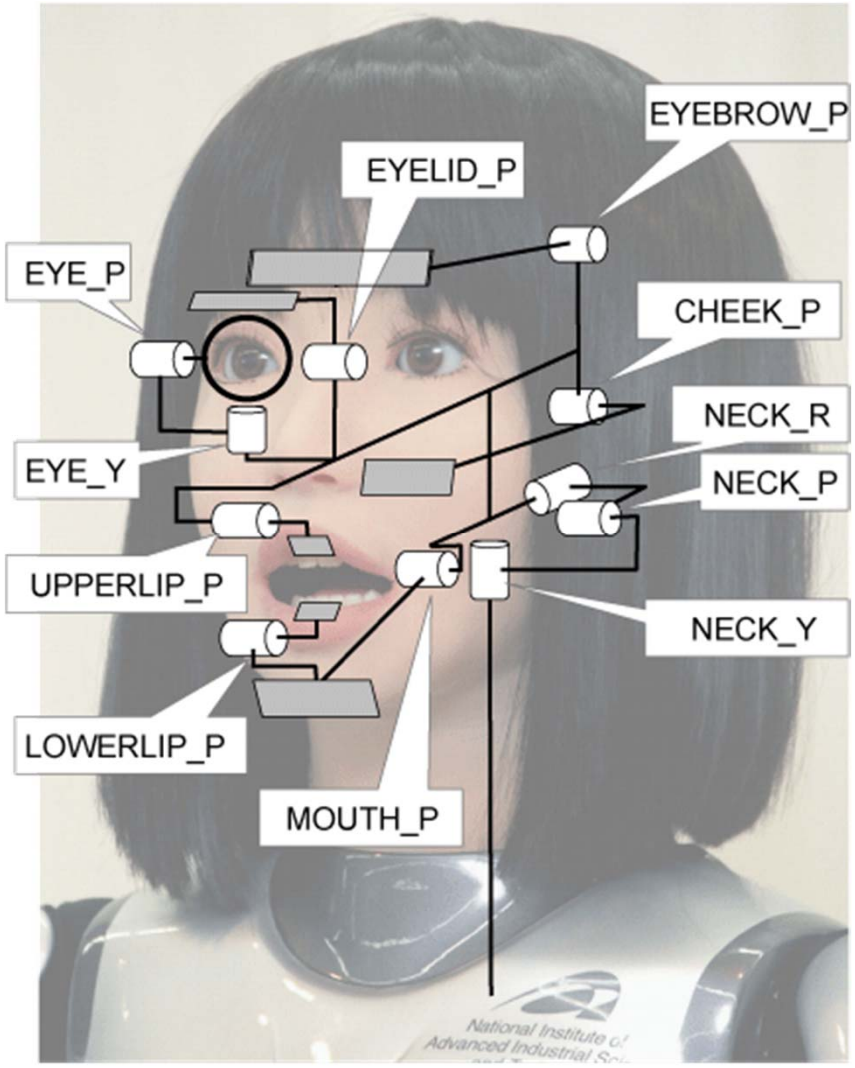
[RWC-MDB-P-2001 No.7 PROLOGUE]

VocaWatcher（ぼかうお）

分析結果からの軌道生成



HRP-4Cの頭部アクチュエータ



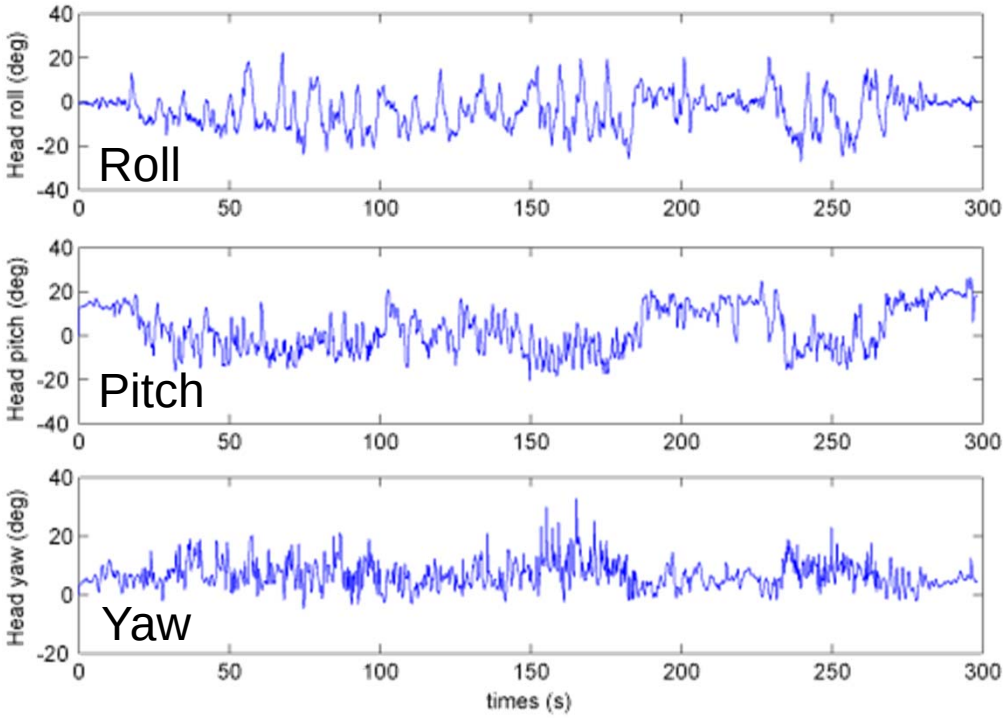
首
- NECK_Y
- NECK_P
- NECK_R

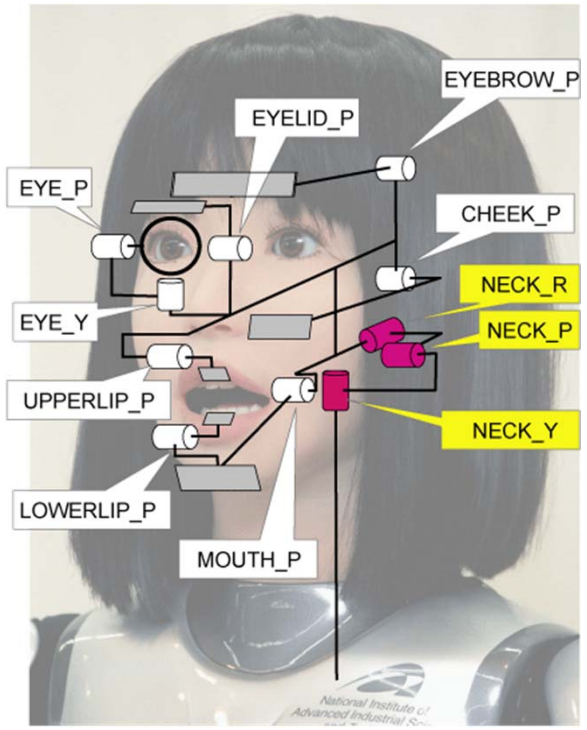
目元
- EYEBROW_P
- EYELID_P
- EYE_P
- EYE_Y

口元
- MOUTH_P
- UPPERLIP_P
- LOWERLIP_P
- CHEEK_P

faceAPI ver.3.2.6 の特性

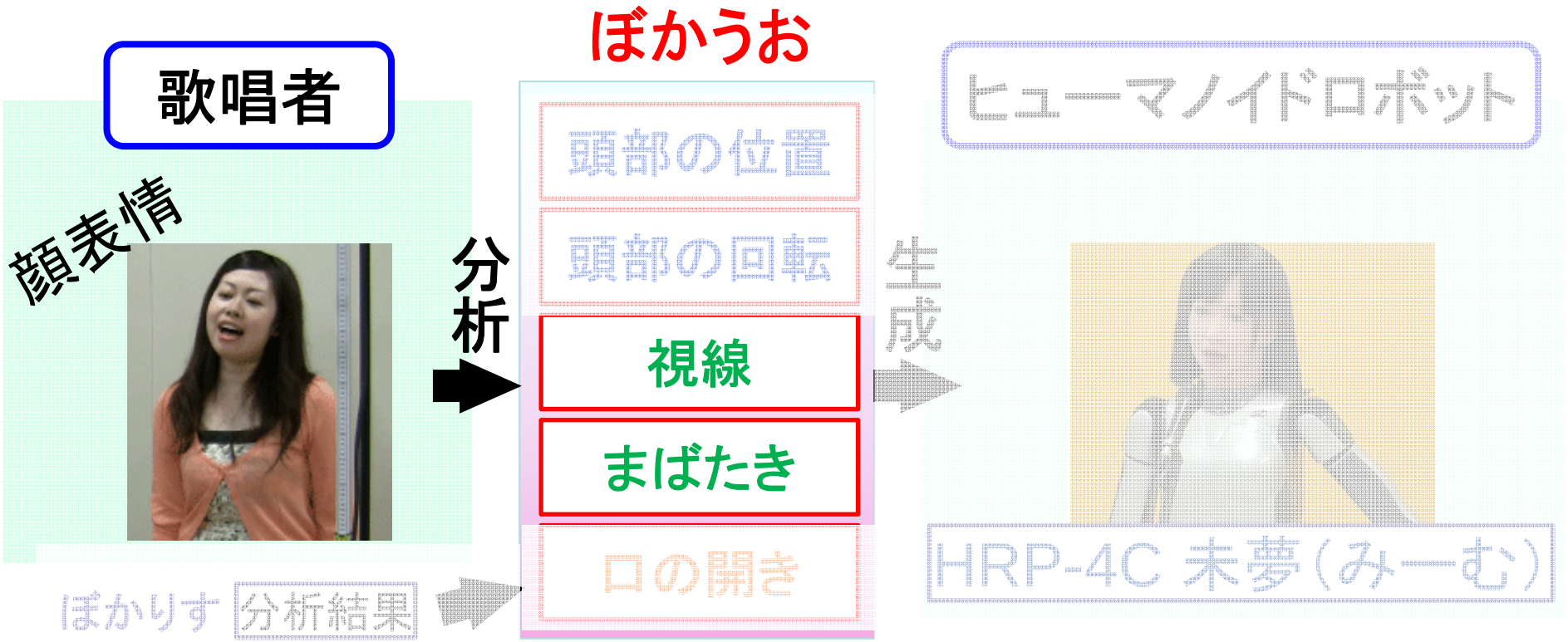
- 顔の姿勢(Roll, Pitch, Yaw)は良好に検出
- フィルタ処理を行い、首3軸の動作角として利用





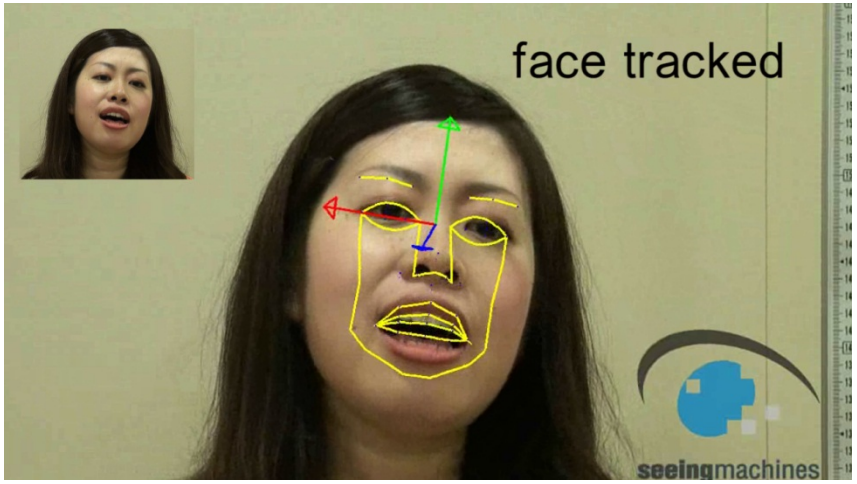
VocaWatcher（ぼかうお）

ビデオ映像からの画像処理



faceAPI ver.3.2.6 の問題点

- まばたきと視線を検出できない
- 唇のトラッキングにしばしば失敗する



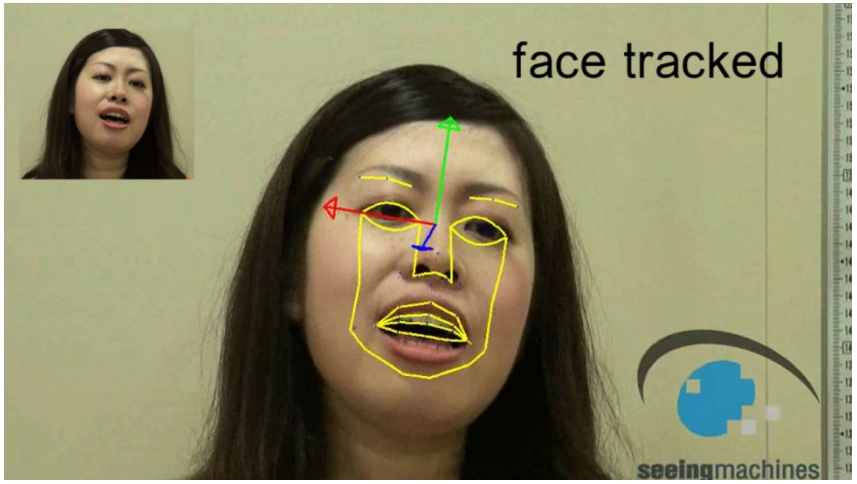
faceAPIで得られた情報を元に、新たな画像処理を行う

右目の領域　⇒　まばたきと視線の検出

顔中心線　⇒　口開度の検出

faceAPI ver.3.2.6 の問題点

- まばたきと視線を検出できない
- 唇のトラッキングにしばしば失敗する



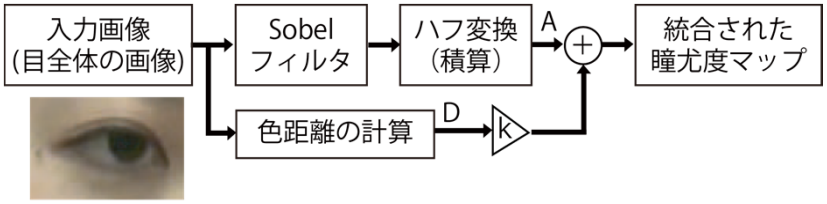
faceAPIで得られた情報を元に、新たな画像処理を行う

右目の領域　⇒　まばたきと視線の検出

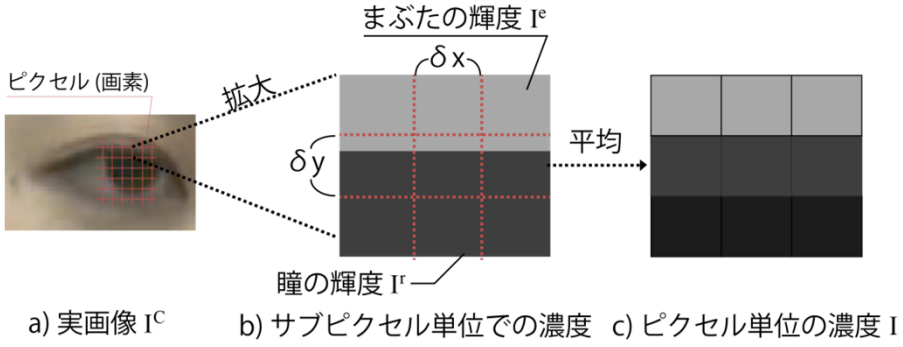
顔中心線　⇒　口開度の検出

まばたきと視線の検出

- faceAPIで検出された右目の領域を画像処理
  - 円形ハフ変換と色距離に基づく**視線検出**
    - 瞳の存在確率を計算して決定

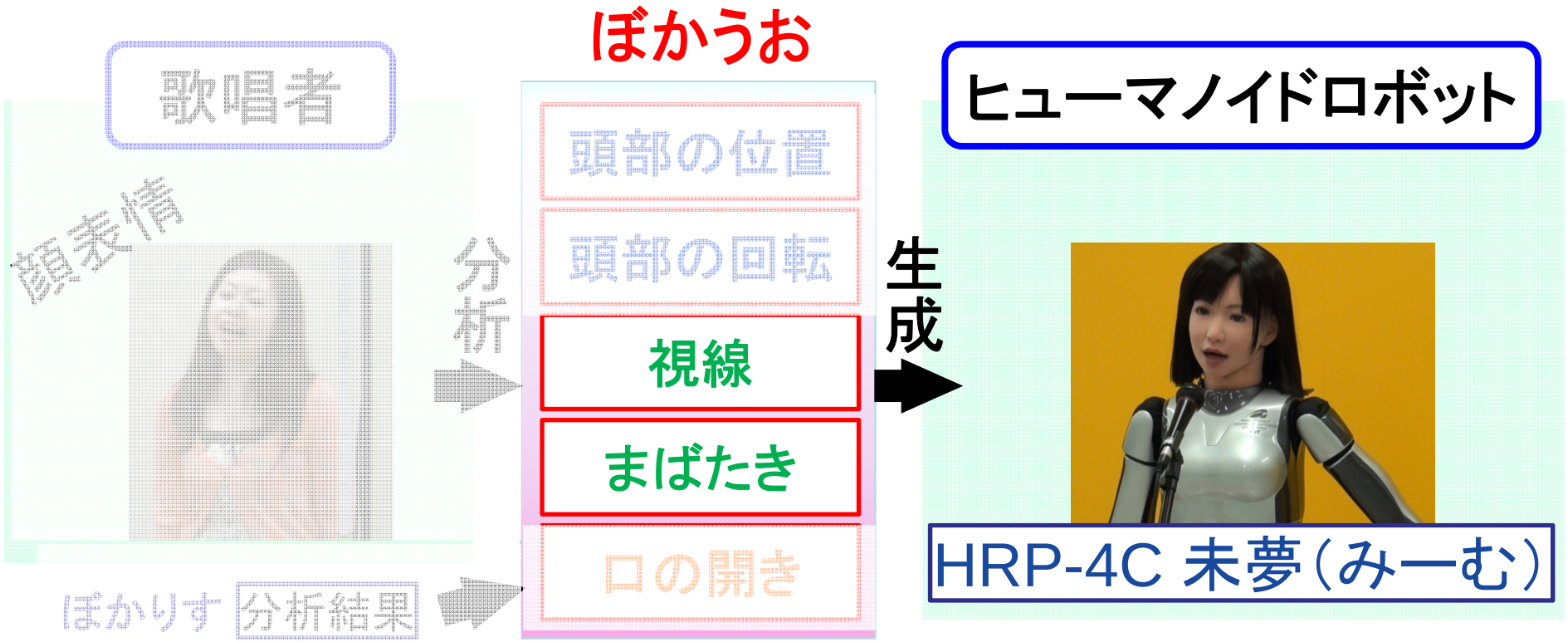


  - **まぶた開度の推定（まばたき検出）**
    - 瞳とまぶたの輝度から、サブピクセル精度で推定



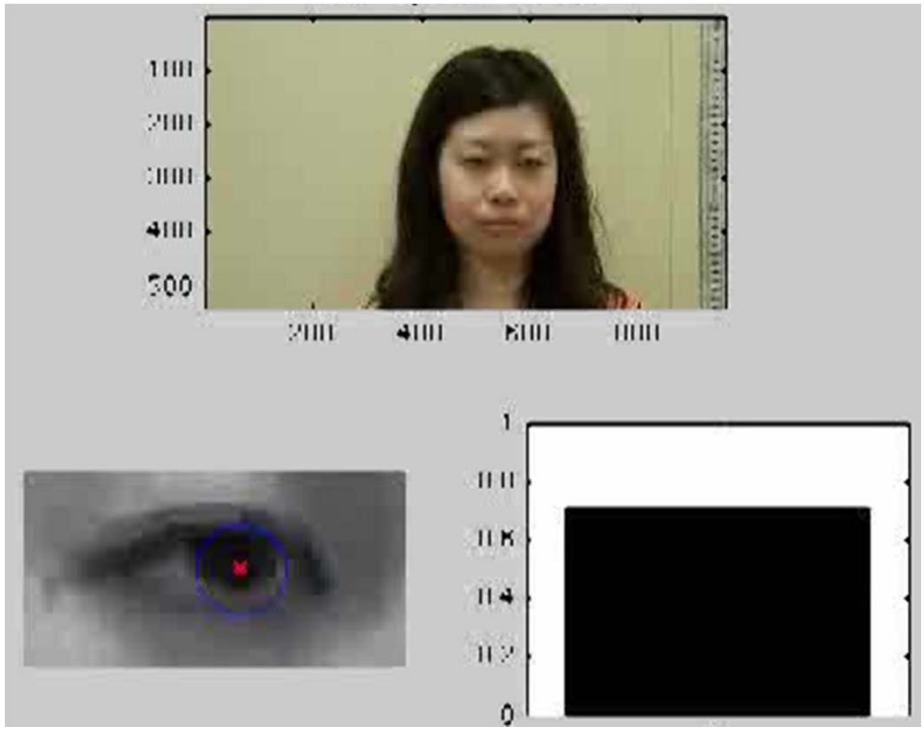
VocaWatcher（ぼかうお）

分析結果からの軌道生成



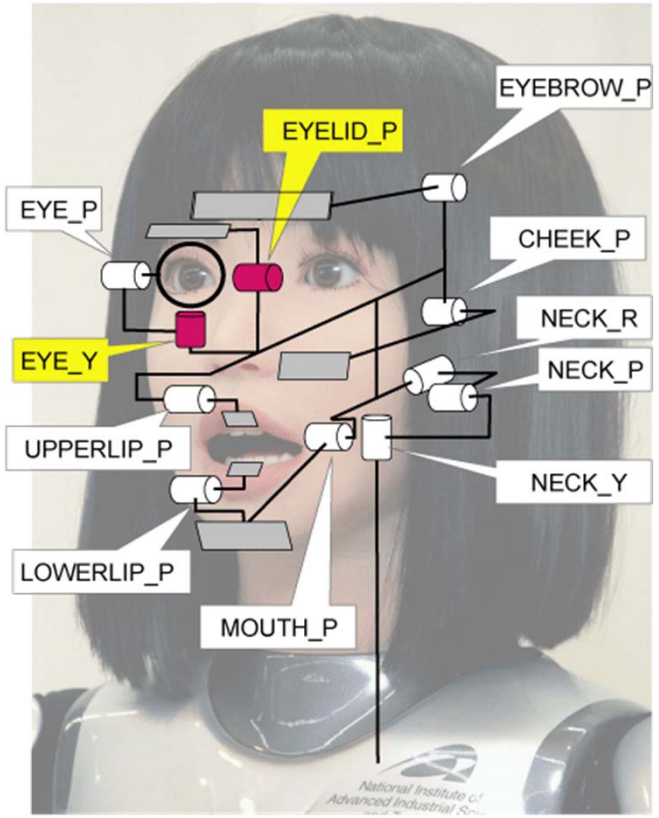
まばたきと視線動作の制御

- 「目の泳ぐ様子」や「目を細めて歌う様子」等が抽出できた
  - フィルタ処理後、制御パラメータとする



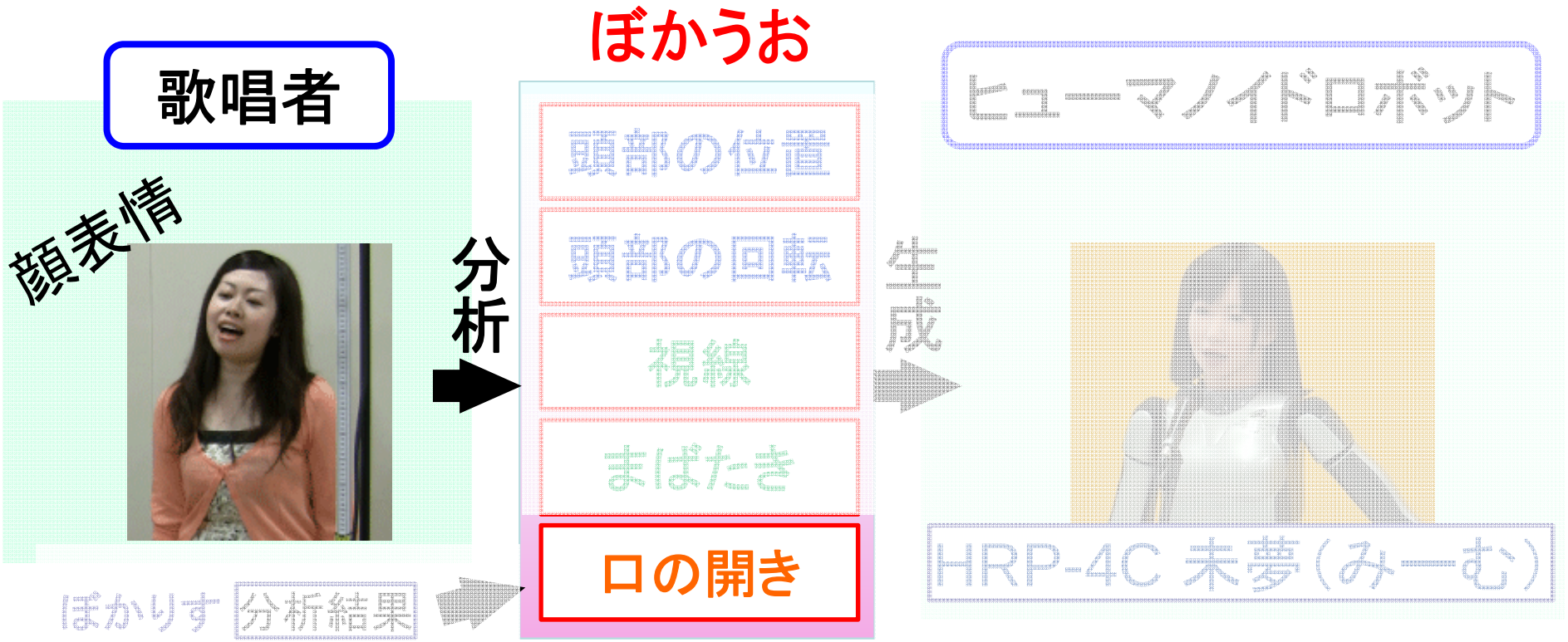
右目領域
（検出された瞳）

目の開度
（0～1）



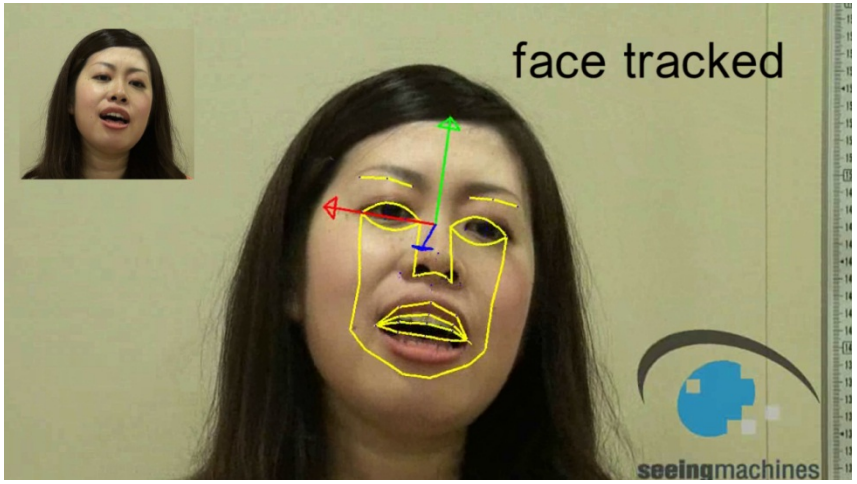
VocaWatcher（ぼかうお）

ビデオ映像からの画像処理



faceAPI ver.3.2.6 の問題点

- まばたきと視線を検出できない
- 唇のトラッキングにしばしば失敗する



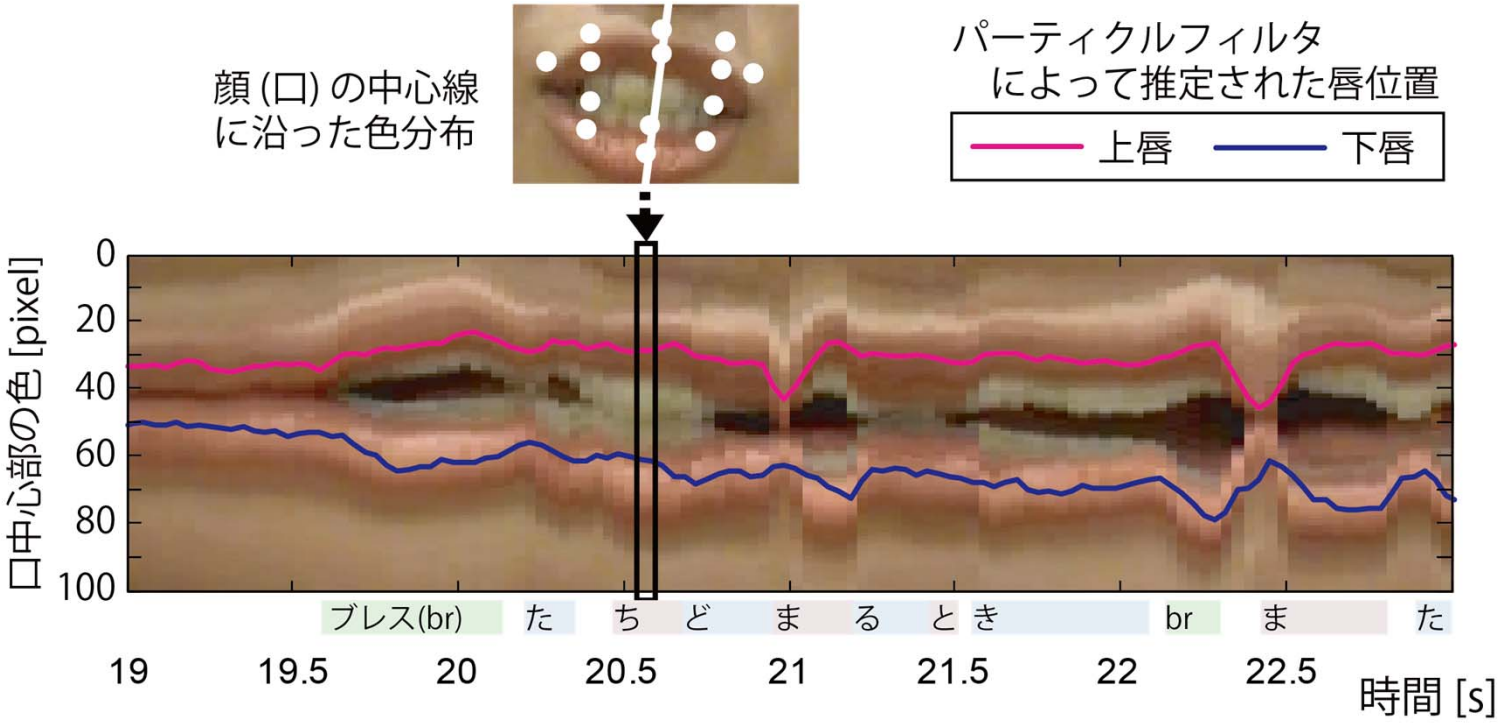
faceAPIで得られた情報を元に、新たな画像処理を行う

右目の領域　⇒　まばたきと視線の検出

顔中心線　⇒　口開度の検出

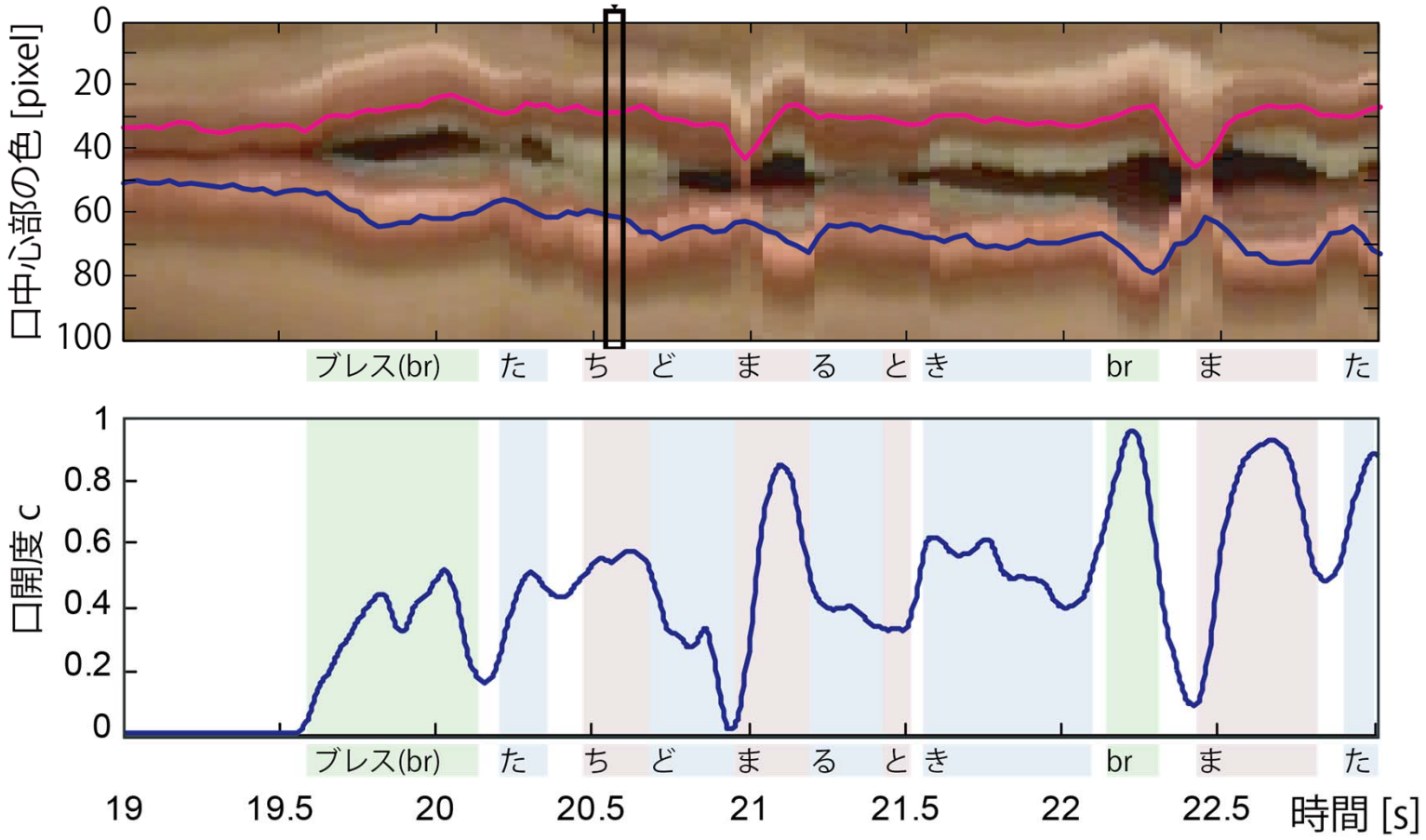
口開度の検出

1. 顔中心線の色分布を各時刻で抽出
2. RGBの色距離を用いたパーティクルフィルタで上下の唇の中心線を推定



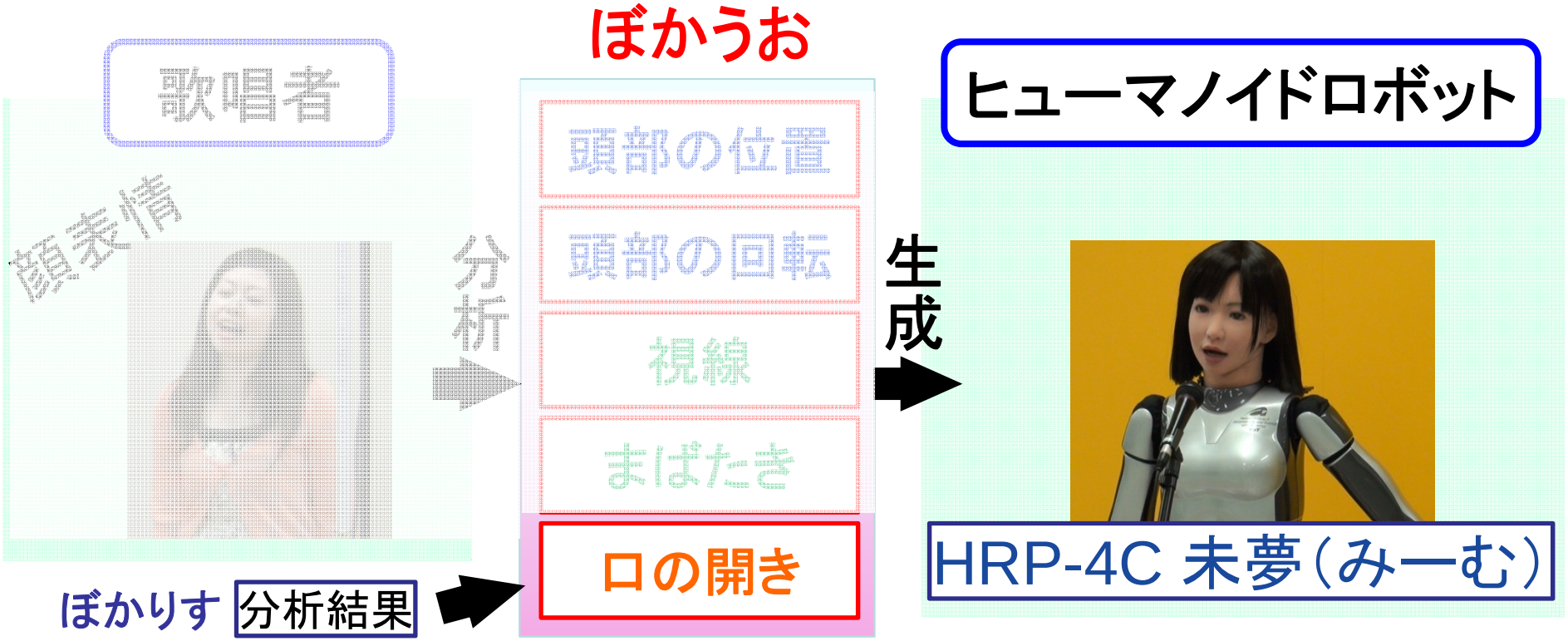
口開度の検出

3. 上下の唇の距離を[0 1]で正規化



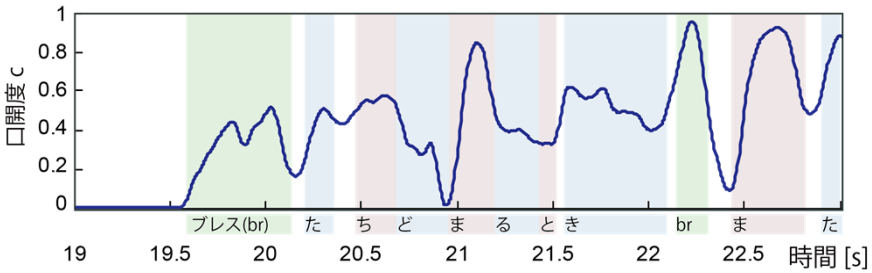
VocaWatcher（ぼかうお）

分析結果からの軌道生成



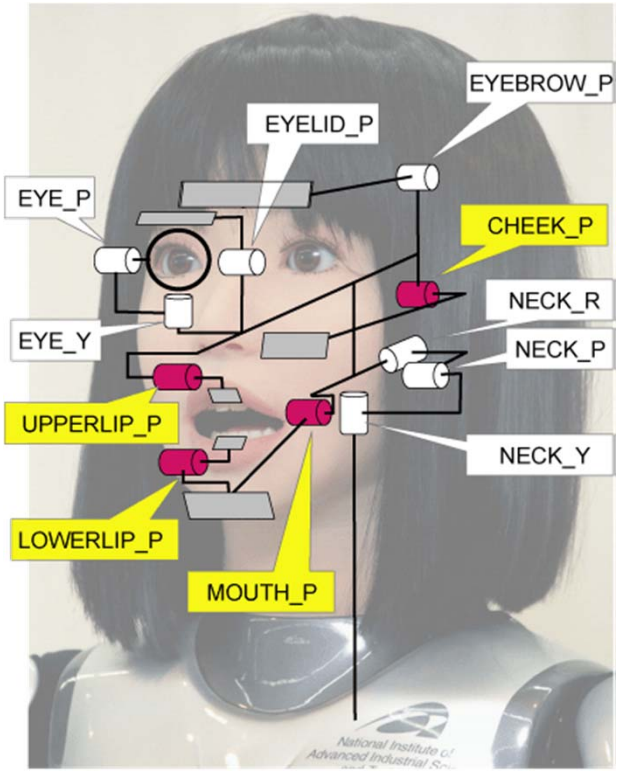
口の開き(唇形状)制御の問題点

- 推定された口開度が一次元データ
- 一方、唇形状はアクチュエータ4つで制御
  - 単に口の開閉だけでは不適切



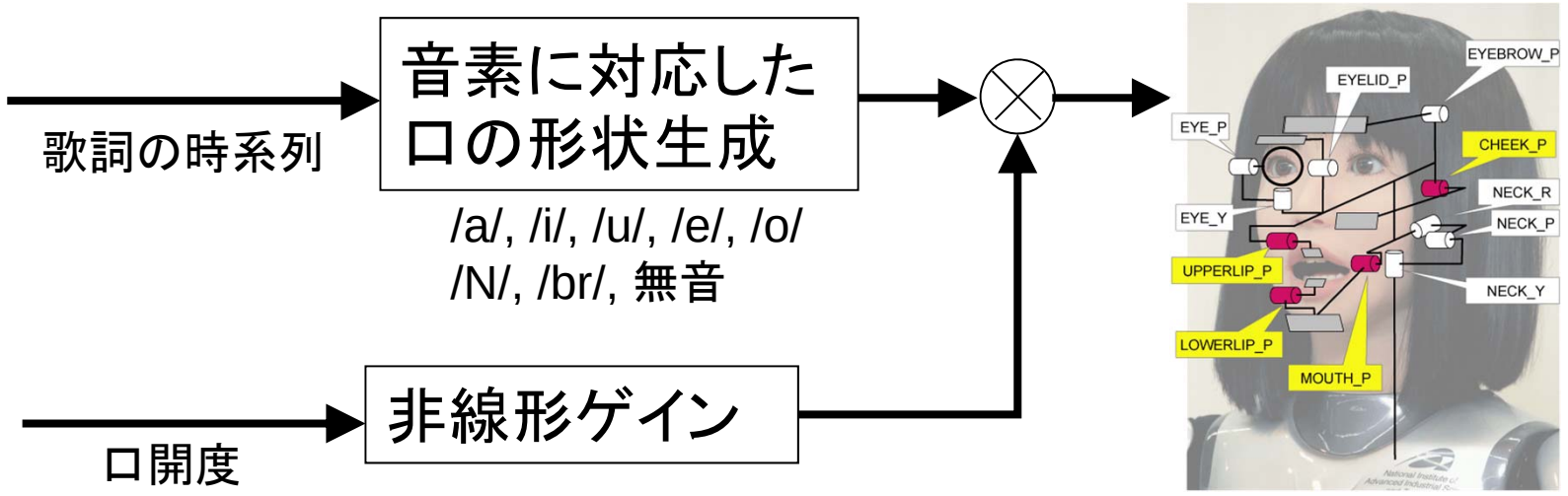
検出された口開度

?

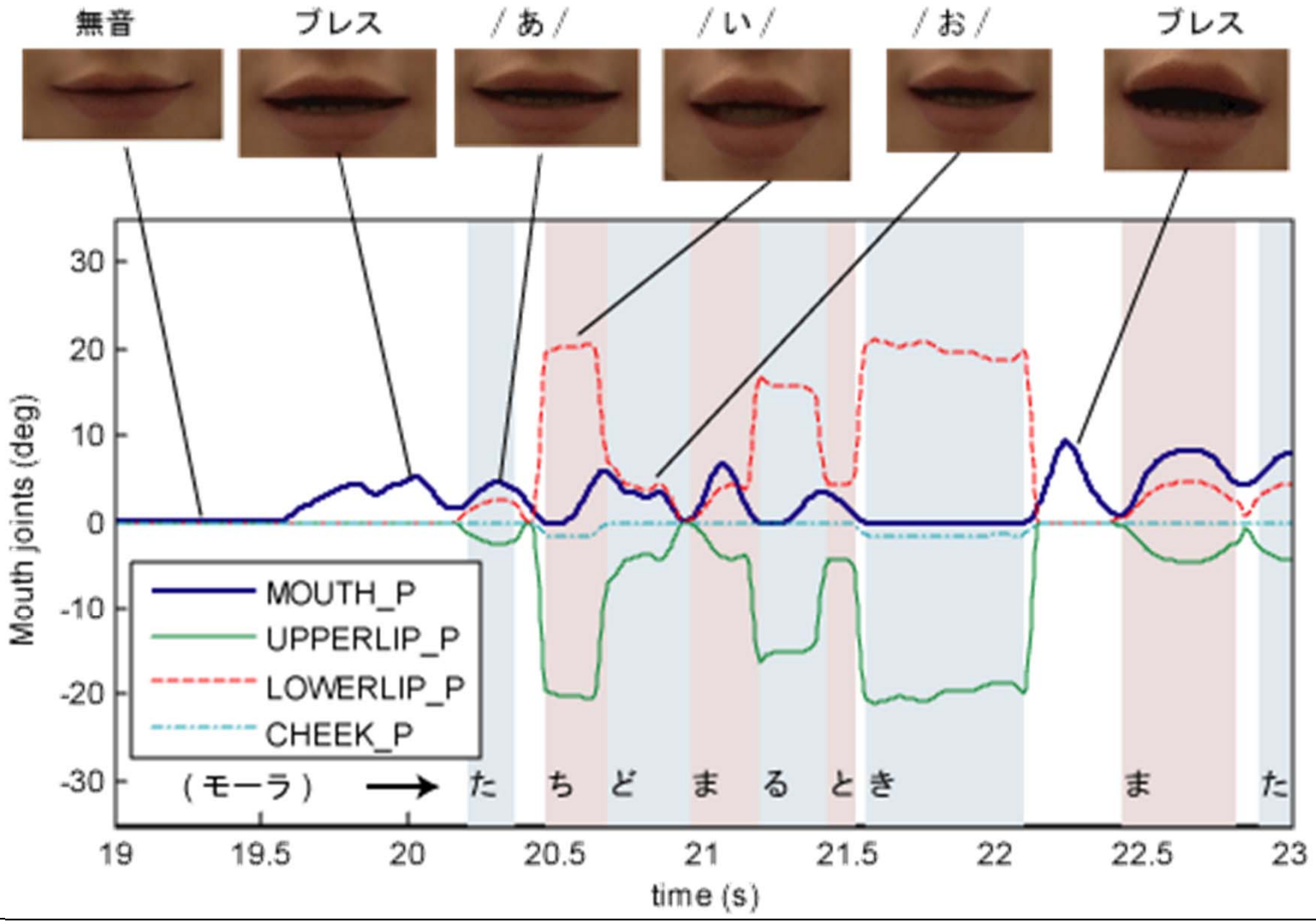


歌声合成時の推定結果を活用

1. VocaListenerで得られた音素情報に対応した唇形状（テンプレート）を生成
   - 5母音（/a/, /i/, /u/, /e/, /o/）, 撥音 /N/, ブレス, 無音
2. 口開度に基づいて唇形状を非線形変調



口開度と発音情報をもとにした口動作生成



まとめ

本研究の三つの新規性

- 人の歌唱をお手本に**自然な歌声**を
  自動生成する技術**VocaListener**を
  ロボットで初めて使用

- 人の歌唱をお手本に**ロボットの顔動作**を
  自動生成する技術**VocaWatcher**を
  新規開発

- **VocaListener**の追加機能として
  ブレス音の自動検出＆合成技術を新規開発

考察

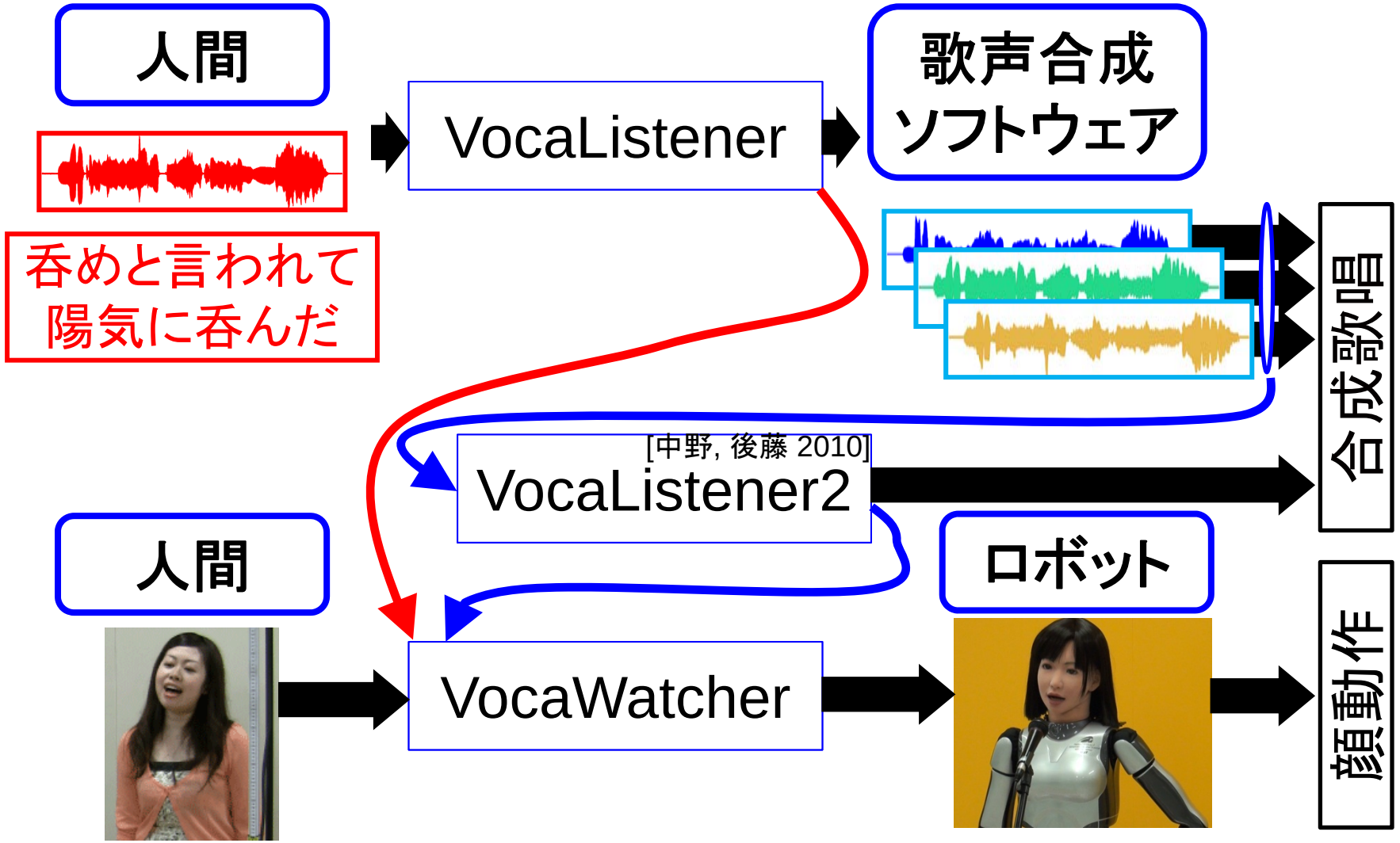
- 「真似る」アプローチによる人間らしい歌唱の生成
  - 少し視聴しただけでは人間と区別がつかないぐらいの印象を受けた人がいた
  - 一方、歌声合成、ロボットの機構の限界もあった
- 「不気味の谷」現象
  - 人間から遠くデフォルメされていると親近感を得やすくても，人間に近づく過程で違和感を感じる現象
    - 顔の動作や声の質、皮膚や顔形状などに不自然さが残るため
  - 自然な表現に関する技術が漸次的に進歩すれば、どこかの段階で「不気味の谷」にさしかかる
    - 辿り着き、その上で技術の力で乗り越えることが重要

「真似る」研究アプローチの意義

- **調整時間の減少**によって、**表現へ注力**できる
  - 歌うだけで、自然な歌声合成・顔動作生成のためのパラメータが調整される
- **ツールの汎用性・多様性の向上**、分野の発展
  - 調整に関する知識を持たないユーザが使用可能
    - **楽譜やピアノロールでの編集が難しいユーザ**など
  - **音楽家にとって**手で弾いた演奏をリアルタイムに
    MIDIで（楽譜）入力するインタフェースは直感的で不可欠
- 人間の**知覚・生成機構の解明**に繋がる
  - 高品質な歌声合成と顔動作生成技術の実現

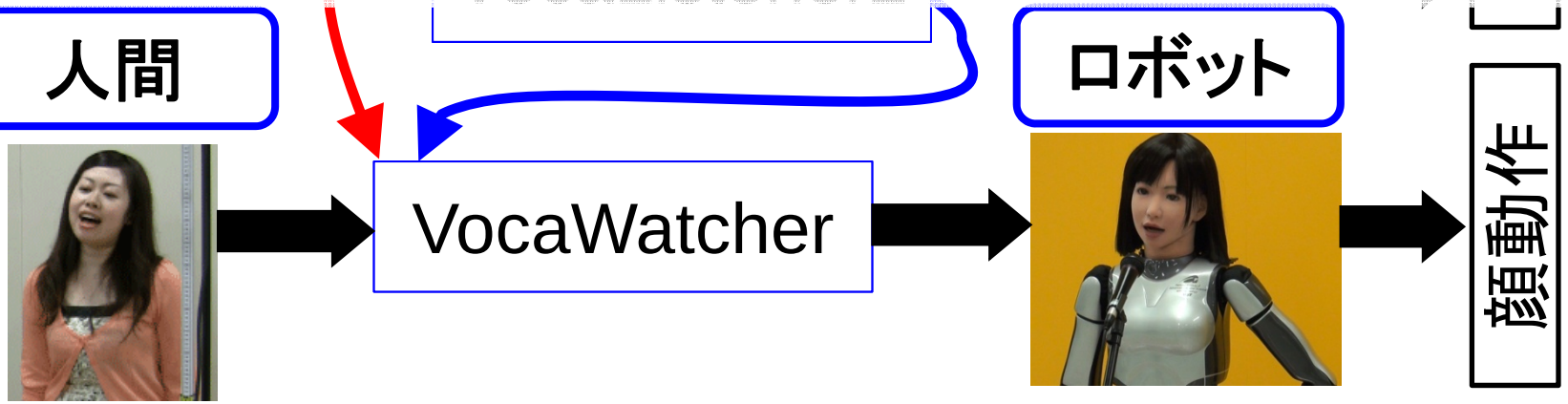
未来

我々の一連の「真似る」研究アプローチ

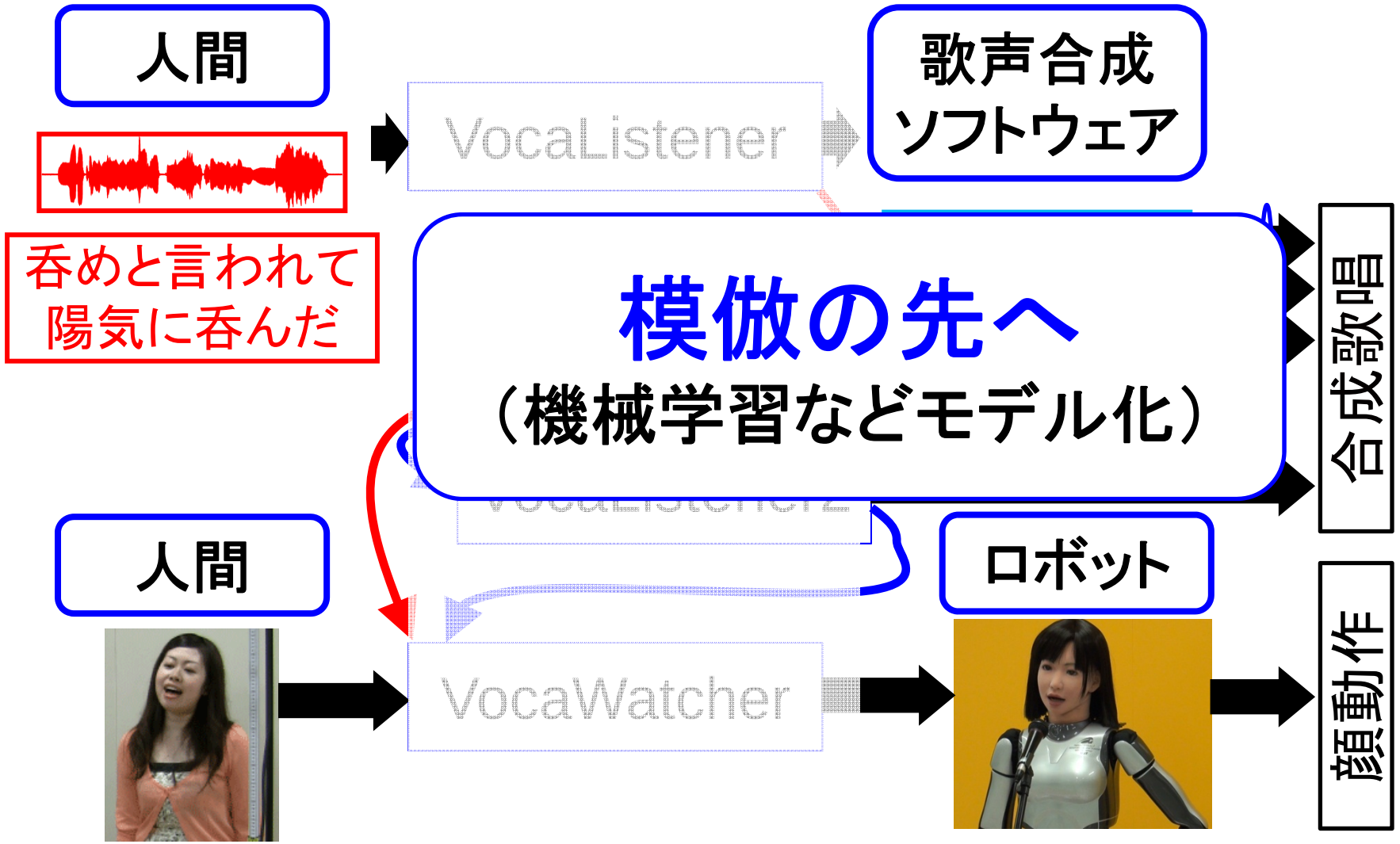


さらなる発展可能性

多様な種類の
ヒューマノイドロボットへの対応
（ぼかりすから反復推定の考え方）



さらなる発展可能性


人間
歌声合成
ソフトウェア
VocaListener
呑めと言われて
陽気に呑んだ
模倣の先へ
（機械学習などモデル化）
合成歌唱
人間
ロボット
VocaWatcher
顔動作

お手本を真似て「歌うロボット」の未来

- 歌声合成、ヒューマノイドロボットの特長を有効活用
  - 人間の機能を**人工的に再現**できる
  - 人間の**限界を超える**ことができる
    - 高い歌や早い歌を歌う
    - 人間らしくない声質で歌う
    - 同じ動きで歌う、等
  - クリエータが**自分単独ではできない表現**
    - 複数人の歌声・振付でコンテンツ作成

- 歌声合成、音楽ロボットの普及
  - 楽音合成（シンセサイザ）が普及した歴史

お手本を真似て「歌うロボット」の未来

- クリエータの立場から
  - **気兼ねなく、柔軟な自由な発想**を表現できる
  - **多数のクリエータが同じ声・同じロボットで**
    コンテンツ制作することで、表現が多様になる

- リスナーの立場から
  - **複数の声・複数のロボット・複数のクリエータ**
    から、**好きな組み合わせ**を楽しむことができる

お手本を真似て「歌うロボット」の未来

- **ロボットによる歌唱を**
  個人が自分好みに制御して楽しむ未来
  - コンピュータグラフィックスの分野では一般的
    - 例）MikuMikuDance
  - 実際にロボットを音楽に合わせて踊らせる人もいる

  ⇒（VocaWatcherのような）

  顔の動きで入力できるインタフェースが、手作業で個々の動作を指定するインタフェースより**便利な可能性**がある

- **新しい感動へ**
  - **ロボットが歌うからこそ意味のある歌詞**