

# コンピュータで聴く

教授：河原 達也

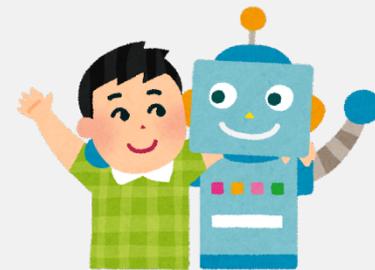
准教授：吉井 和佳

特定助教：井上 昂治

京都大学 大学院情報学研究科  
知能情報学専攻 音声メディア分野

# 音声メディア分野

ロボットとの対話・インタラクション



音声認識

音環境理解

音楽情報処理

パターン認識

音響信号処理

統計的機械学習



# なぜ音声メディア?

---

- 人間の知の創造・伝達の根源
- 幅広い技術をカバー
  - 信号処理
  - パターン認識
  - 知識処理
- 統計的機械学習を本格的にいち早く導入
  - ベイズ学習
  - 深層学習

# 音声認識

# 音声認識ソフトウェア

- オープンソースの大語彙連続音声認識ソフト
- 国内外で幅広く研究利用
- アプリケーション展開
  - ロボット
  - 携帯端末
  - 音声案内
- 無償配布
  - 月4K~5Kダウンロード
  - <http://julius.osdn.jp/>



**Julius** now on **GitHub**

**GitHub**

- GitHub site
- ダウンロード
- Julius最新版
- Julius音声認識パッケージ
- 文法認識キット
- 音素セグメンテーションキット
- 言語モデル・音響モデル
- ドキュメント
- The Juliusbook
- チュートリアル・解説
- インストール
- アプリケーション開発
- マニュアル・ソース資料

### What's Julius?

Julius は、音声認識システムの開発・研究のためのオープンソースの高性能な汎用大語彙連続音声認識エンジンです。数万語彙の連続音声認識を一般のPCやスマートフォン上でほぼ実時間で実行できる軽量さとコンパクトさを持っています。

言語モデルとして単語N-gram、記述文法、ならびに単語辞書を用いることができます。また音響モデルとしてトライフォンのGMM-HMMおよびDNN-HMMを用いたリアルタイム認識を行うことができます。DNN-HMMの出力計算にnumpyを用いた外部モジュールを利用することも可能です。複数のモデルや複数の文法を並列で用いた同時認識も行うことができます。

Juliusの最大の特徴はその可搬性にあります。単語辞書や言語モデル・音響モデルなどの音声認識の各モジュールを組み替えることで、小語彙の音声対話システムからディクテーションまで様々な幅広い用途に応用できます。

Julius はオープンソースソフトウェアです。プログラムはC言語で書かれており、さまざまなプラットフォームへの移植や改造が容易です。ライセンスはオープンライセンスで、商用利用への制限もありません。

Julius の研究・開発に関わっている主な機関は以下の通りです。

Copyright (c) 1991-2019 京都大学 河原研究室  
Copyright (c) 1997-2000 情報処理振興事業協会 (IPA)  
Copyright (c) 2000-2005 奈良先端科学技術大学院大学 鹿野研究室  
Copyright (c) 2005-2019 名古屋工業大学 Julius開発チーム

# 音声認識・対話の実用化と研究

- スマートフォン・スマートスピーカ

- 1ターン = 1発話 = 1文

- 発話時にクリック (push-to-talk)

- マジックワード “Alexa” “OK Google”

- 1問1答

- ユーザが何か話さないと応答しない

- 明確なタスク・検索目標

- ユーザはシステムができることを知っている (天気・乗換)



Hey Siri



課題：人間どうしの自然な話し言葉の音声認識  
人間どうしの会話のようなインタラクション

# 講演の音声認識

- <http://sap.ist.i.kyoto-u.ac.jp/jimaku/>
- 字幕付与
  - Open Course Ware (OCW)
  - 放送大学
- ノートテイク支援
  - 学会講演

音声認識研究者 京都大学 河原 達也 教授  
の協力により日本語字幕配信を始めました。

2012年ノーベル生理学・医学賞 受賞  
山中 伸弥 教授による講演「iPS細胞研究の進展と課題」

(2010年CiRA一般の方対象シンポジウム「iPS細胞研究の最前線」より)

2010年 京都大学 CiRA一般の方対象 > 共有 ▾ 詳細情報

iPS細胞 何ができるか?

iPS細胞

分化誘導

神経細胞 心筋細胞 肝細胞 膵細胞

患者の皮膚細胞

臓器移植治療 (再生医療)

病態モデル、治療薬開発  
毒性、副作用の評価

そういう実験を、病態モデルを作ると言いますが

08:43 / 34:14

YouTube

# 国会審議の音声認識

- <http://sap.ist.i.kyoto-u.ac.jp/diet/>
- 衆議院で運用
  - 会議録作成支援
  - 文字正解率 90%
  - 世界初・唯一



# 技術トレンド

- End-to-End/Seq-to-Seq学習に基づく音声認識
  - 従来：HMMベース
    - 単語列  $Z$  → 音声信号  $X$  の生成モデルを推定
      - 言語モデル  $p(Z)$ ：単語列から学習
      - 音響モデル  $p(X|Z)$ ：ペアデータで学習
      - $p(Z|X) \propto p(X|Z)p(Z)$ ：両モデルを利用
  - 現在：RNNベース
    - 音声信号  $X$  → テキスト  $Z$  の写像  $p(Z|X)$  を直接学習
      - 注意 (attention) 機構付きEncoder-Decoderモデル
      - Connectionist Temporal Classification (CTC)モデル

# 音声対話・インタラクション

# 自律型アンドロイド ERICA

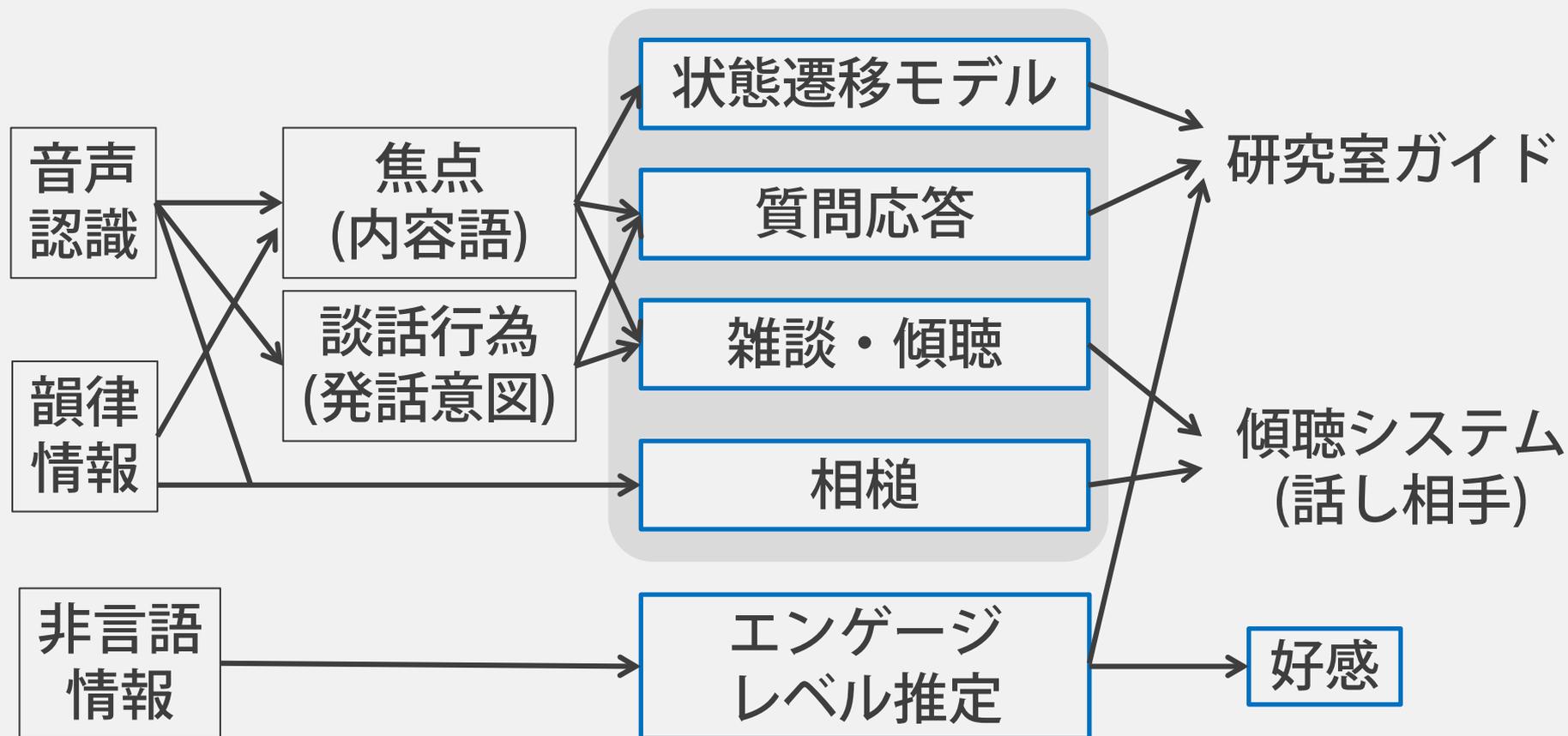
- <http://sap.ist.i.kyoto-u.ac.jp/erato/>
  - 目標：人間レベルの音声対話 **“Total Turing Test”**
    - 音声対話
    - 相槌・フィラー
    - 視線・表情
- ↓
- 受付・ガイド
  - 傾聴



# 人間らしい見た目・動き



# ERICAの音声対話システム



# 人間らしい音声の反応

---

- 相槌の生成

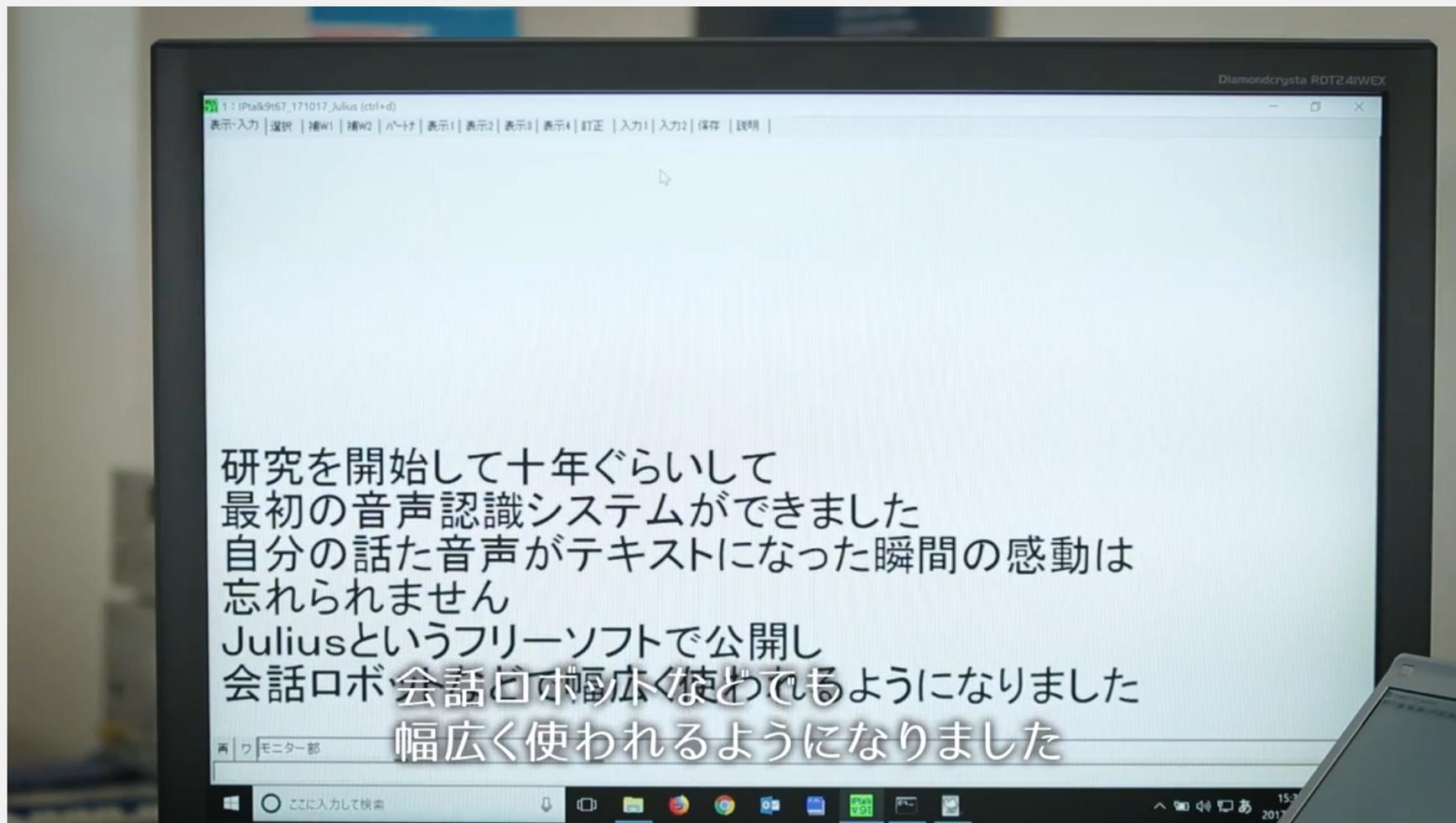
- 同一のパターン 
- ランダムに4種類 
- 文脈に応じて選択 

- フィラーの生成 

- 同調的笑いの生成



# 音声認識・対話・インタラクション



# 音楽情報処理

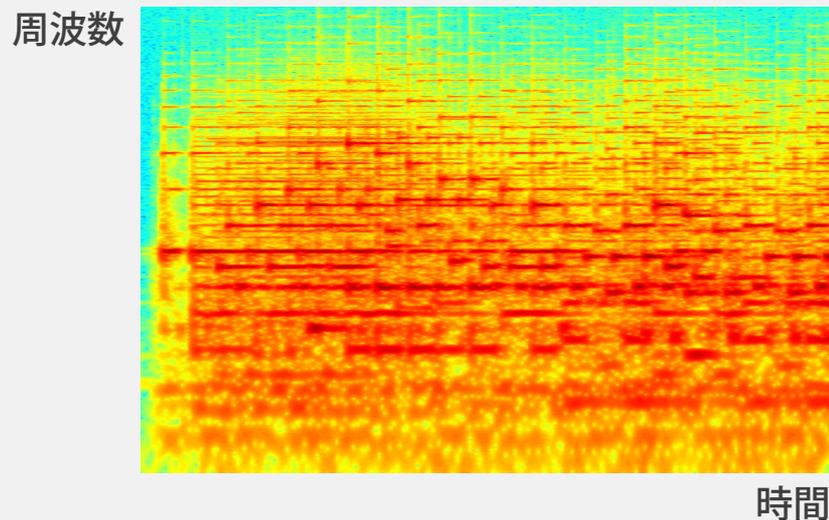
# OngaACCELプロジェクト

- 音楽を深く理解できるコンピュータを作りたい
  - Web上でみんなが使えるようにする (<http://songle.jp>)

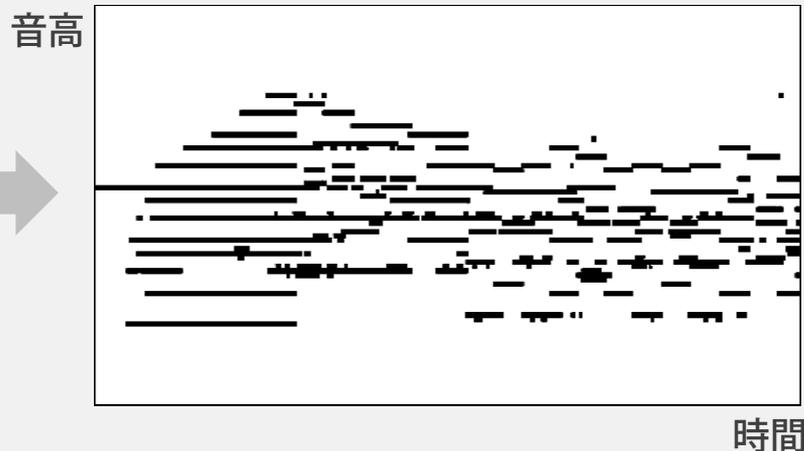


# 音楽音響信号に対する自動採譜

- 音楽音響信号から楽譜 (ピアノロール) を推定したい
  - 音響的なスペクトル構造に着目するだけでは不十分

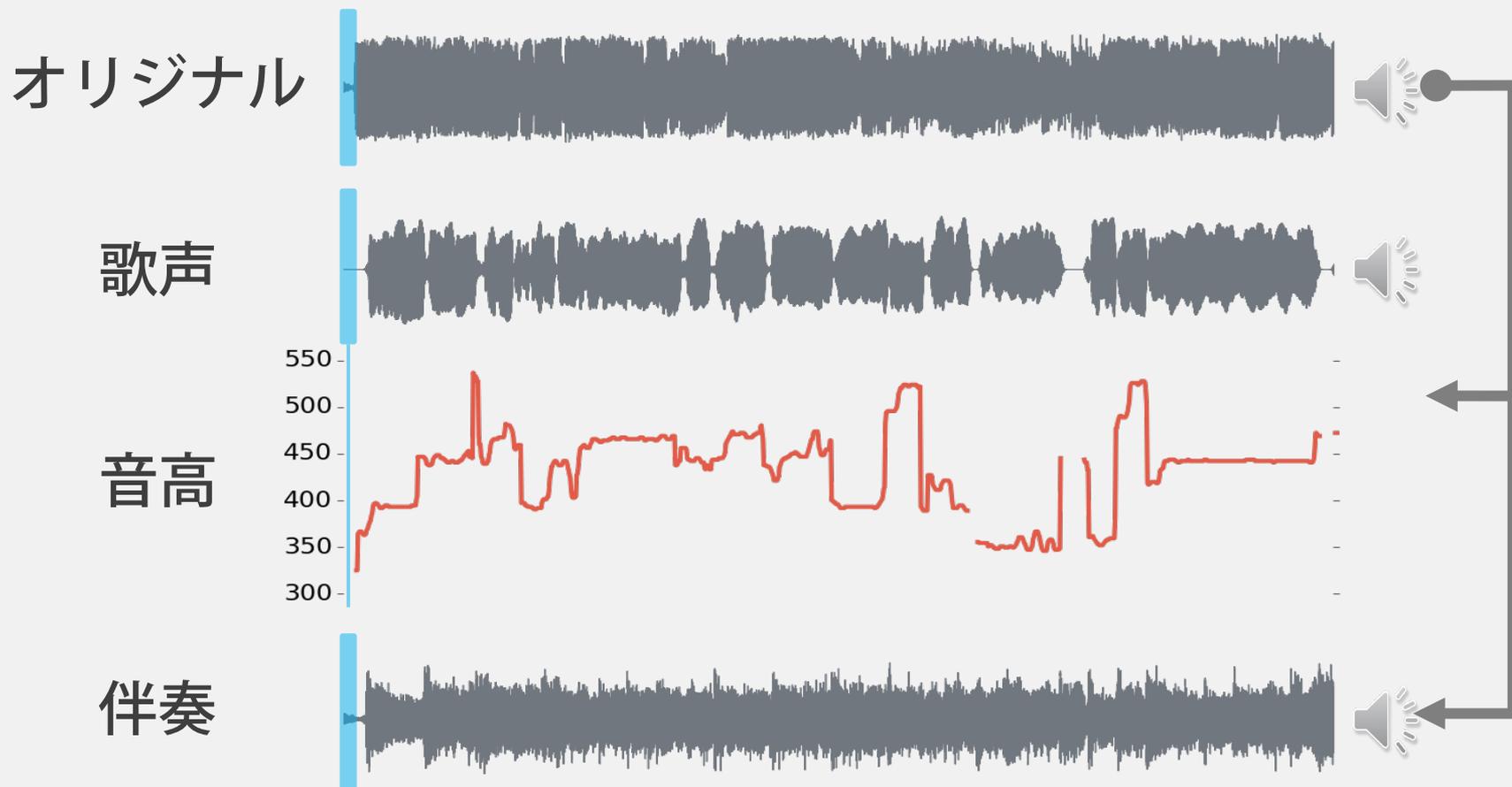


音楽的に不自然な音符の配置が多発



音楽的に「正しい」音符配置となるよう誘導するベイズモデルを考案

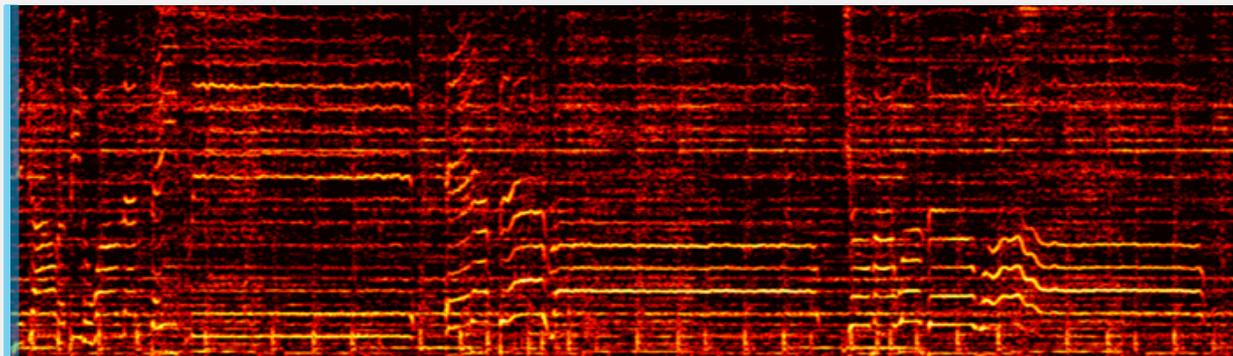
# 歌声と伴奏音の分離



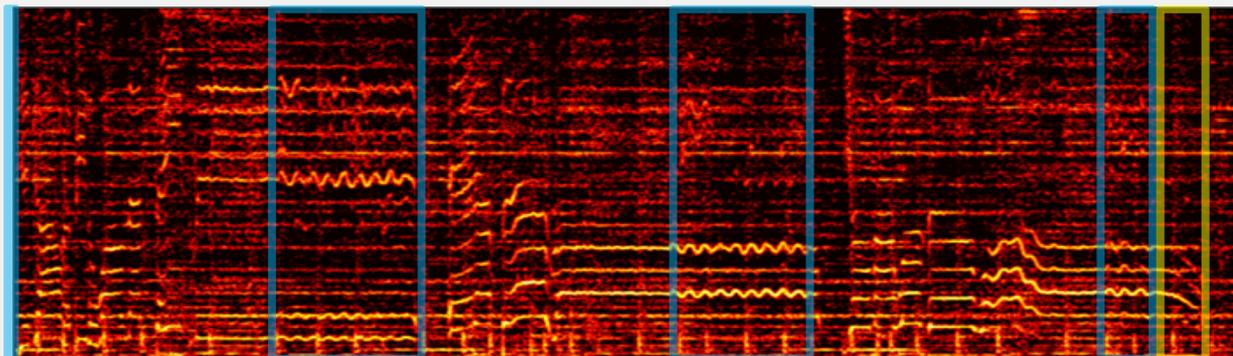
# ビブラート・グリッサンドを付与

- 混合音中の歌声の音高軌跡を自由に編集できる

編集前



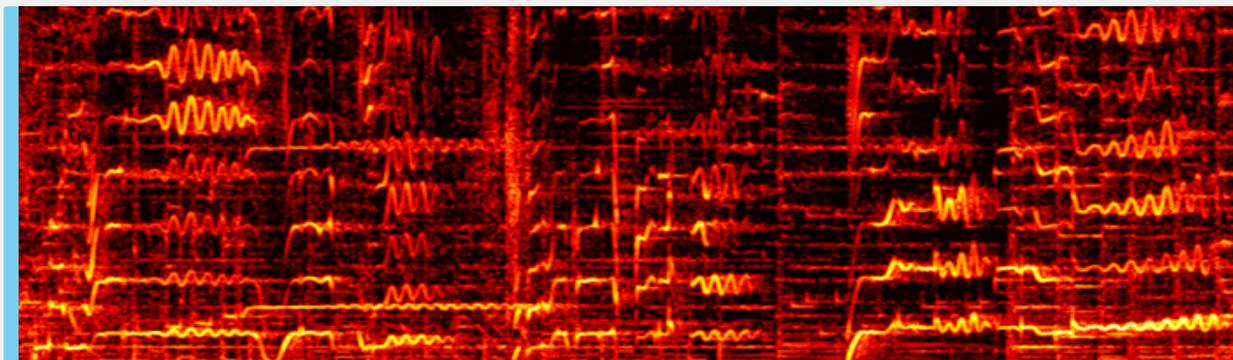
編集後



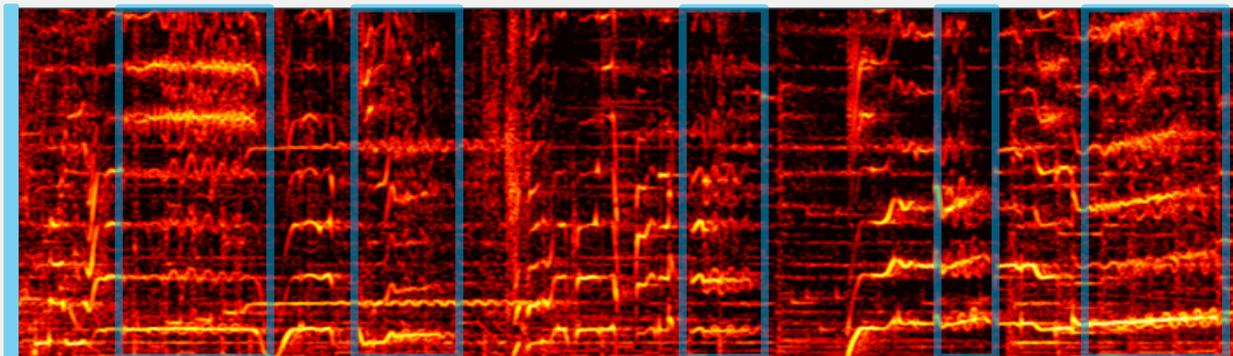
# ビブラートを除去

- 混合音中の歌声の音高軌跡を自由に編集できる

編集前



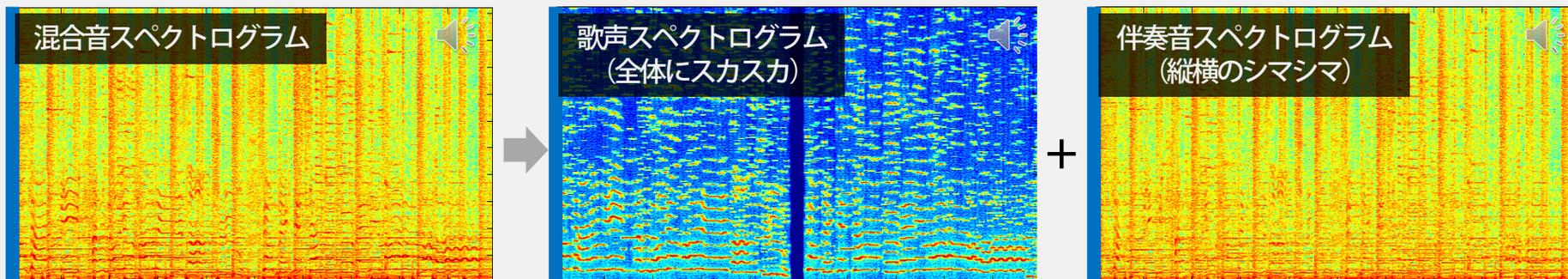
編集後



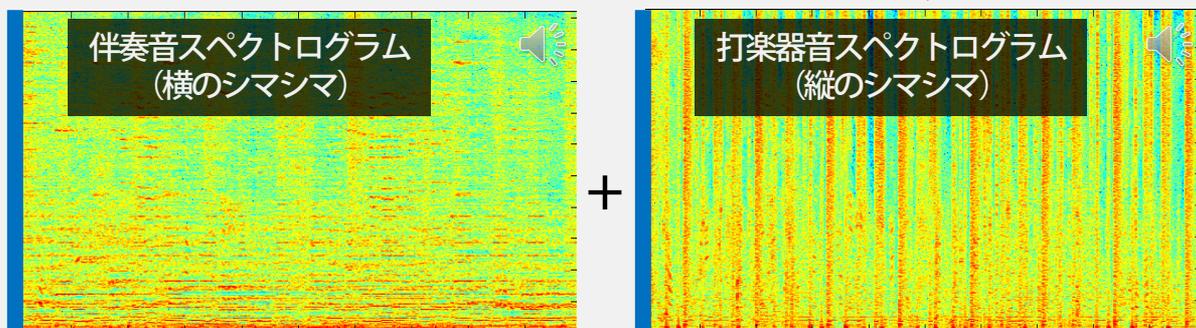
# 歌声・伴奏音・打楽器音分離

- それぞれの音の特徴に着目すると分離可能
  - 分離の良さを評価する関数を最大化する最適化問題を解く

ロバスト主成分分析を用いた分離 [Ikemiya 2014]



メディアンフィルタを用いた分離 [FitzGerald 2010]

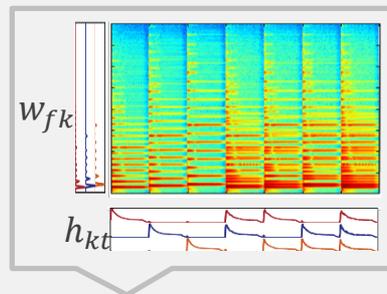
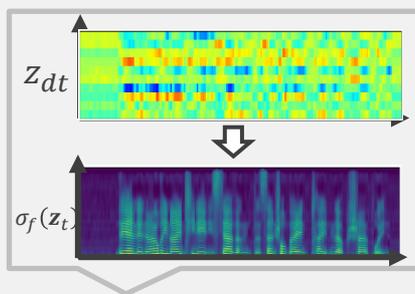


# 音環境理解

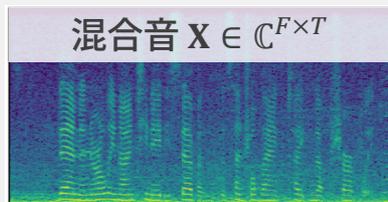
# 深層生成モデルに基づくモノラル音声強調

- 実環境下での音声認識には音声強調が必要
  - スパースモデル → 音声スペクトルにはフィットしない
- 深層生成モデルを用いた音声事前分布を利用
  - DNNを用いて音声のスペクトル構造の分布を事前学習

精緻な  
音声モデル  
→ 事前学習可能



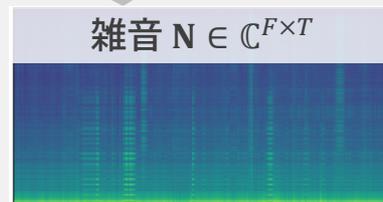
低ランクな  
雑音モデル  
→ 事前学習不要



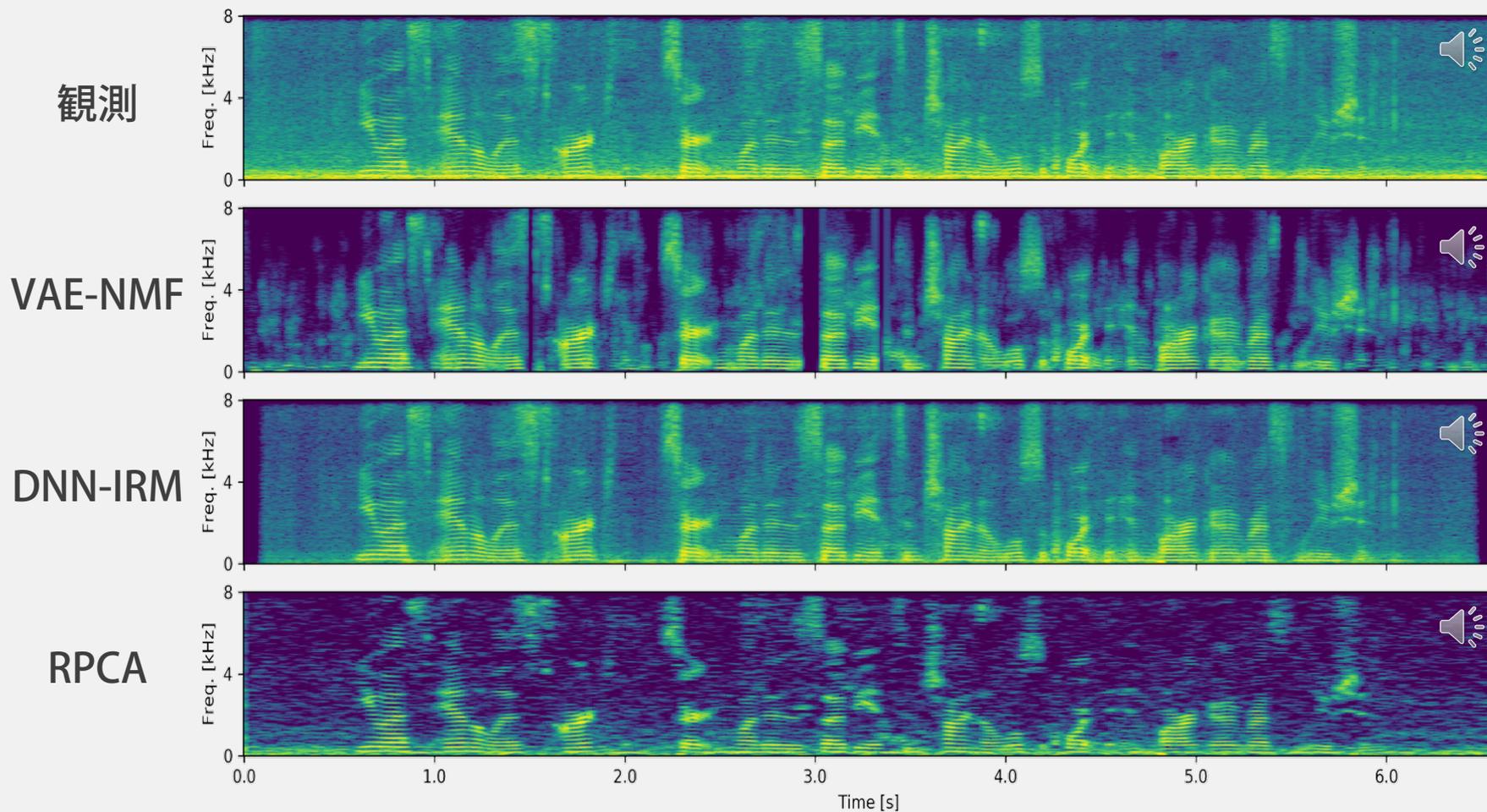
=



+

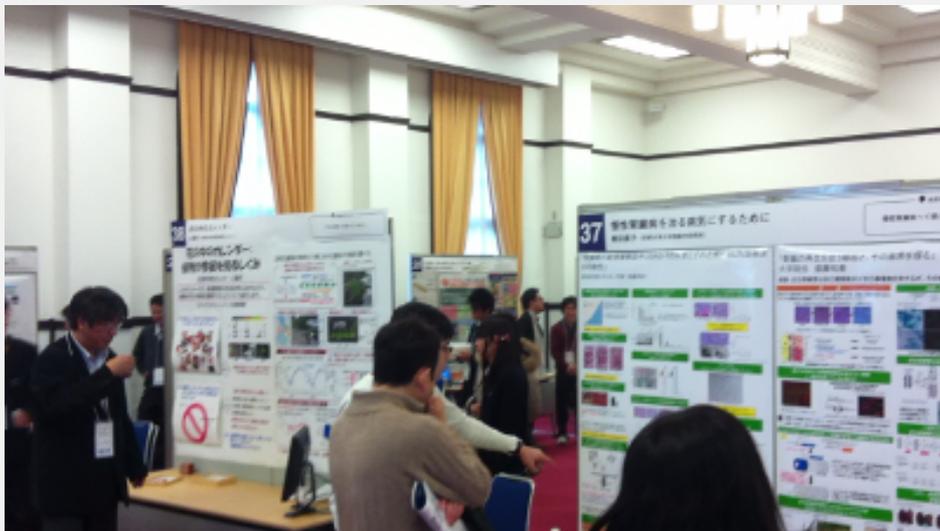


# モノラル音声強調例



# 生成モデルに基づくマルチチャンネル音声強調

時計台国際交流ホール



観測した混合音



マイクロホンアレイ



分離音



発表者の発話



聴衆の発話



背景雑音



# マルチモーダルな会話のセンシング・解析

- <http://sap.ist.i.kyoto-u.ac.jp/crest/>
- Smart Posterboard
  - 大型電子掲示板
  - マイク・カメラ群搭載
  - プレゼン支援
- **環境認識機能**
  - 音声・映像収録
  - 発話・視線検出
  - 興味箇所推定



# 音声メディア分野の講習内容

- 音声強調・認識・対話の理論的基盤・最新動向解説
- アプリケーション作成からモデルの深層学習演習

最先端の研究+核となる基礎理論  
(座学)

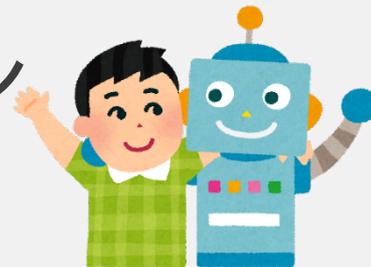
実装体験  
(演習)

	10:00	12:00	13:00	14:30	16:00	17:30	19:00
Day 1 木	音声技術の展望		音声認識の概要	音声認識の基盤技術	音声認識演習 (Julius)		懇親会
Day 2 金	音声対話の概要		音声対話の基盤技術	音声対話演習 (MMDAgent)	音声対話演習 (Dialogflow)		Q&A
Day 3 木	系列写像学習		seq2seqモデル学習演習 (Speech Commands)				Q&A
Day 4 金	音声強調		音声強調演習	音源分離	音源分離演習		Q&A

# 音声メディア分野の講習内容

Day 2

ロボットとの**音声対話**・インタラクション



Day 1

**音声認識**

**音環境理解**

**音楽情報処理**



**パターン認識**

**音響信号処理**



Day 3

**統計的機械学習**

Day 4