

平成26年度国立情報学研究所市民講座第7回

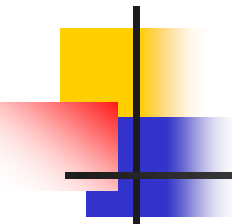
機械が音を聞き分ける！？ ～暮らしを変える音源分離技術～

小野順貴

<http://www.onn.nii.ac.jp/>

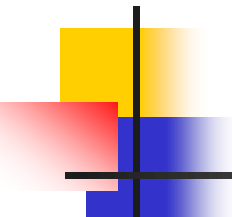
onono@nii.ac.jp

国立情報学研究所 情報学プリンシプル研究系／
総合研究大学院大学 情報学専攻



講演の流れ

- 音源分離への研究動機
- 音の性質
- 音源分離の原理1: 引き算アプローチ
 - マイクロフォンアレイによる音源分離
- 音源分離の原理2: 掛け算アプローチ
 - 調波音打楽器音分離
 - ボーカル抽出／ボーカルキャンセル
- まとめと今後の展望

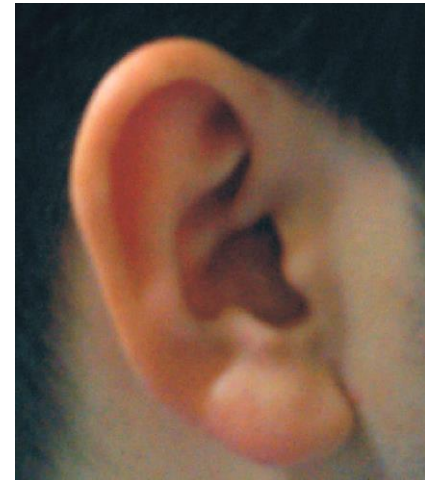


講演の流れ

- 音源分離への研究動機
- 音の性質
- 音源分離の原理1: 引き算アプローチ
 - マイクロフォンアレイによる音源分離
- 音源分離の原理2: 掛け算アプローチ
 - 調波音打楽器音分離
 - ボーカル抽出／ボーカルキャンセル
- まとめと今後の展望

音による環境理解

- 人間は音により外界の様々な情報を得ている
 - 音声によるコミュニケーション
 - 言語による意思の疎通、感情の伝達
 - 危険や異常の検知
 - 物体の接近
 - 聞きなれない音への違和感
 - 様々な環境把握
 - ドアの開閉、機械の動作、天候、物体の衝突、...



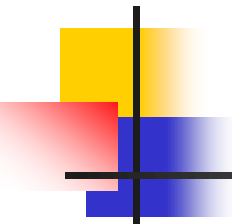
機械が音を理解するには？

- 問題：ファーストフード店で機械が音声対話するとしたらどのくらい難しいか？
- 複数の話し声
- 足音、衝突音
- 機械音
- 報知音(ピーピーピー・・・)



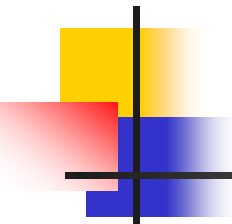
 日本建築学会編, 建築と環境のサウンドライブラリ(SMILE2004)より

実世界は混合音に満ちている！



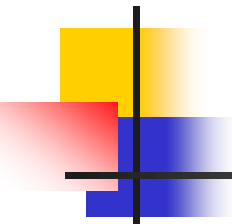
聞きたい音を取り出す技術の必要性

- 携帯電話の音声認識は実用レベル
 - Siri, しゃべってコンシェル→きれいな音声は認識できる
- 遠隔発話の認識では残響や干渉音の重畳が問題
- 実環境で特定の音だけを取得する技術
 - 接話マイク(音源近くにマイクを配置。可能ならこれがベスト)
 - ガンマイク(大きなサイズが必要、音源方向にむける必要)
 - アレイマイク(複数マイクの信号を用いて指向特性を制御)



音源分離の応用例

- 音源分離：混ざった音を個々の音に分ける技術
- 応用
 - **音声認識**：雑音中の音声を認識する
 - **遠隔会議**：会話のオーバーラップや雑音の中で高い品質の音声で通話する
 - **補聴器**：必要な音だけを大きくして聞きやすくする
 - **音楽や動画の加工**：楽器音や歌声を個別に編集したり、好みのバランスで音楽を楽しむ
 - **音による環境のモニタリング**

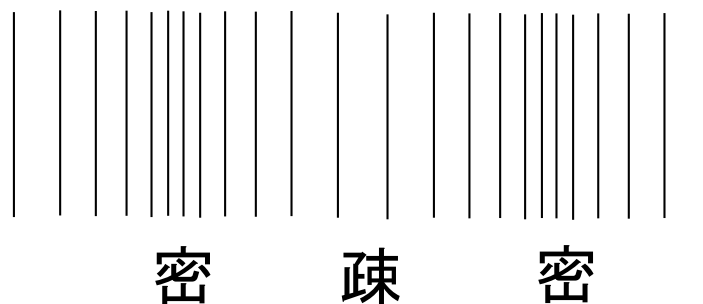


講演の流れ

- 音源分離への研究動機
- 音の性質
- 音源分離の原理1: 引き算アプローチ
 - マイクロフォンアレイによる音源分離
- 音源分離の原理2: 掛け算アプローチ
 - 調波音打楽器音分離
 - ボーカル抽出／ボーカルキャンセル
- まとめと今後の展望

音とは

- 音: (空気の) 圧力の粗密波
 - 波としての性質をもつ
 - 干渉: 複数の波は重なり合う(足し算)
 - 反射: 面にぶつかり進行方向が変化する
 - 回折: 物の後ろに回りこむ



音の伝播

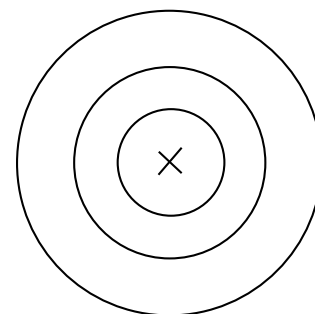
■ 速度

- 空気中: 340m/s
- 水中: 1500m/s
- 鉄中: 6000m/s
(光の速度: 3.0×10^8 m/s)
- 空気中では、17cm(人間の頭と同サイズ)を進むのに0.5ms:
聴覚系で検出可能→方向知覚の手がかり

結構遅い！

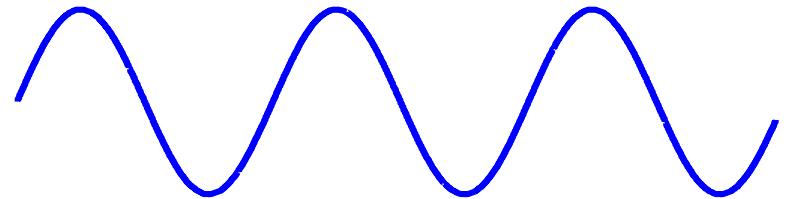
■ 距離減衰

- 点音源からの伝播: 球面波
音圧は振幅に反比例して減衰
- 例) 点音源から10cm、1mでの距離で音を比較すると?
振幅で10倍減衰 → 20dB減衰



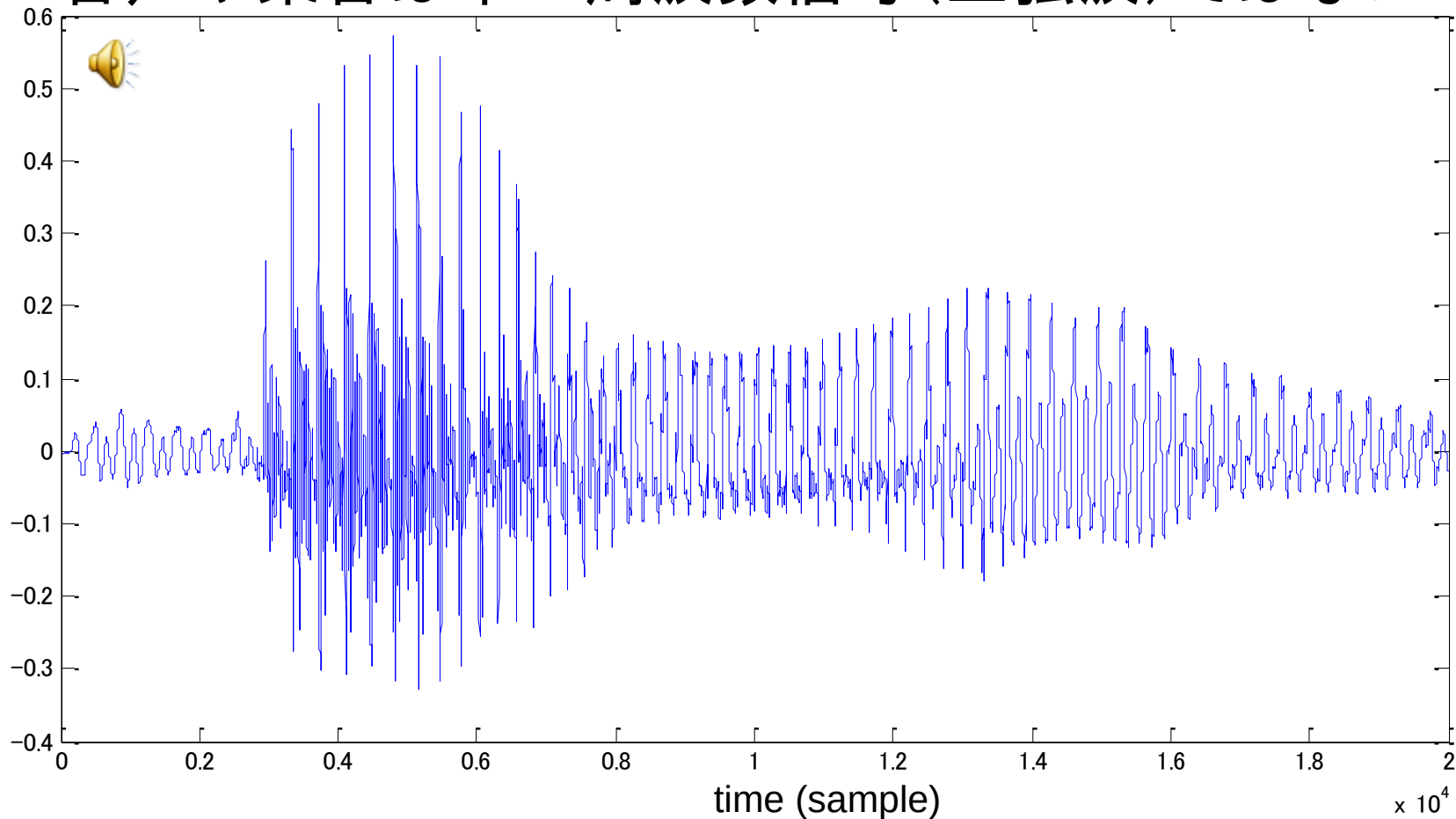
音の高さ(周波数)

- 音の周波数: 1秒あたりの圧力の周期的変化の回数
 - 可聴音で20Hz～20000Hz
 - 20kHz以上は超音波
 - 20Hz以下は超低周波音などとも呼ばれる
 - 非常に広帯域 (1000倍≒10オクターブ！)
 - 豊富な情報量
- 波長(=速度／周波数)
 - 3.4cm@10kHz
 - 34cm@1kHz
 - 3.4m@100Hz
 - (周波数に依存するが) 長波長



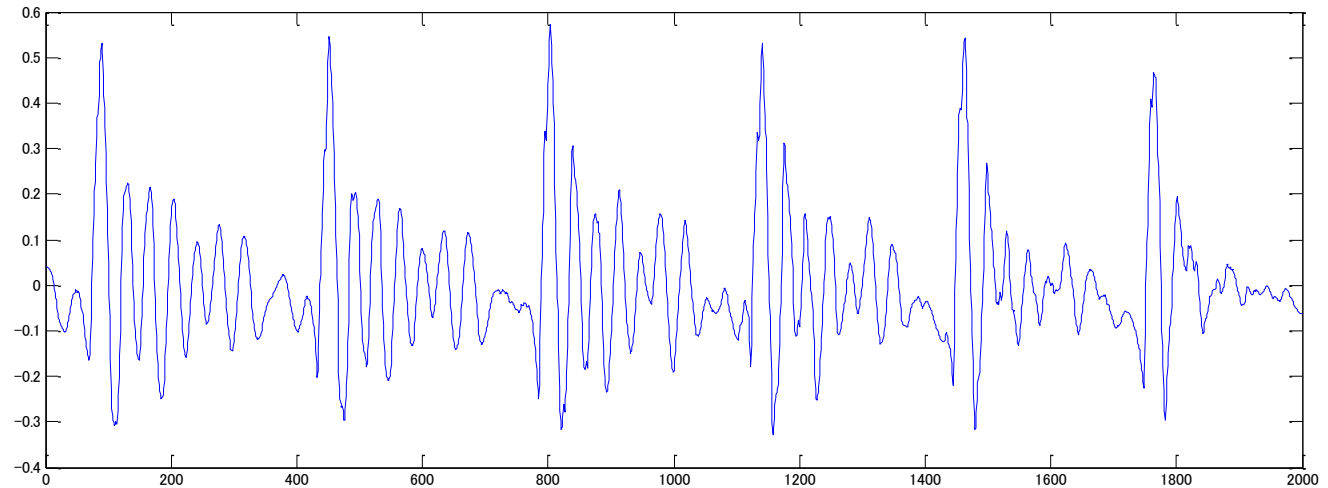
音声信号の例

- 音声や楽音は単一周波数信号(正弦波)ではない

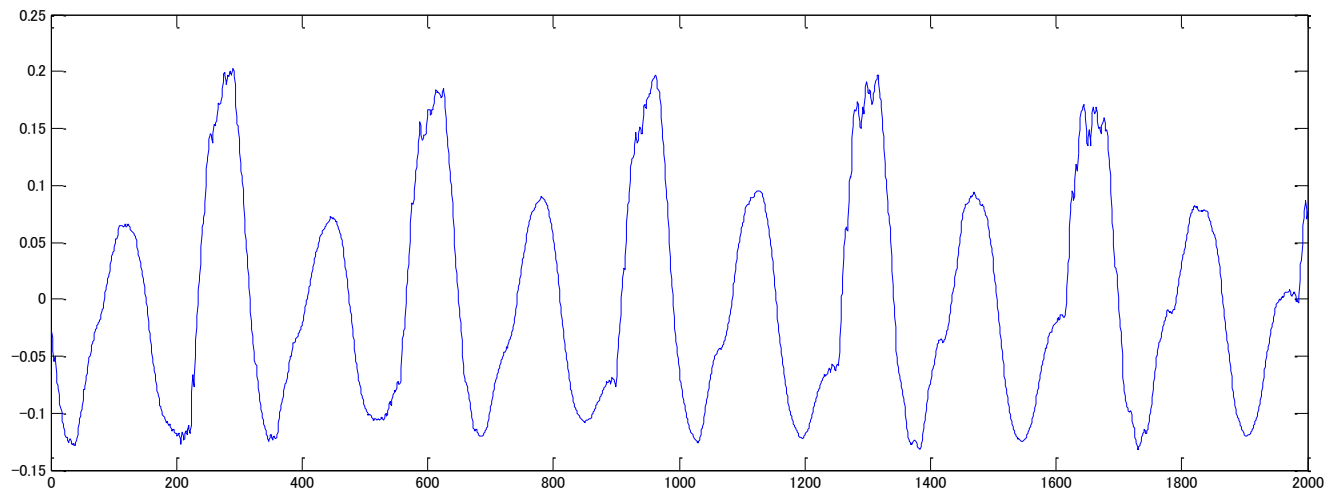


音声信号の例(拡大図)

/a/の付近

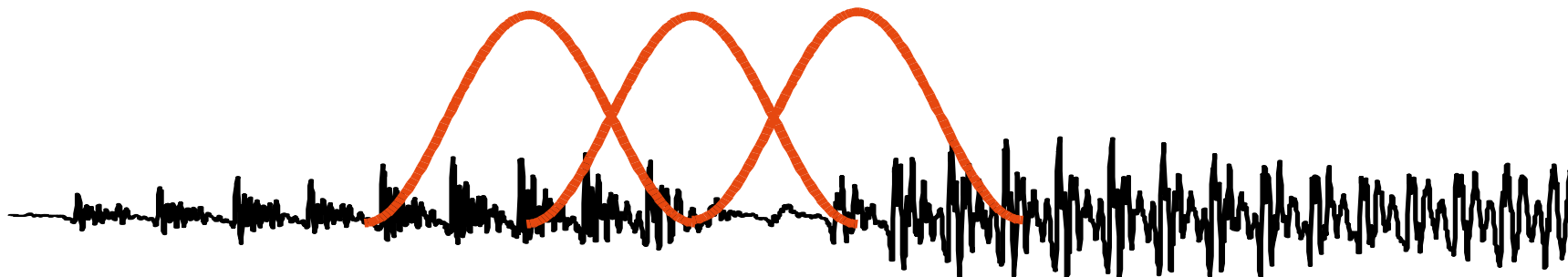


/i/の付近

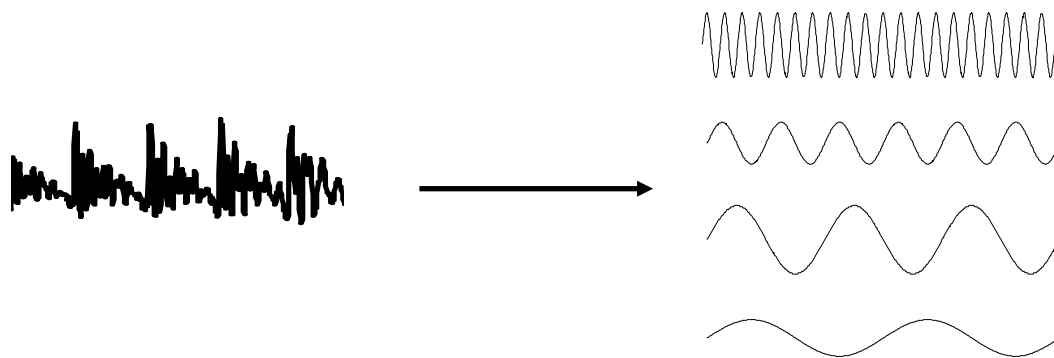


信号のスペクトログラム解析

- 信号を短い時間区間に分解

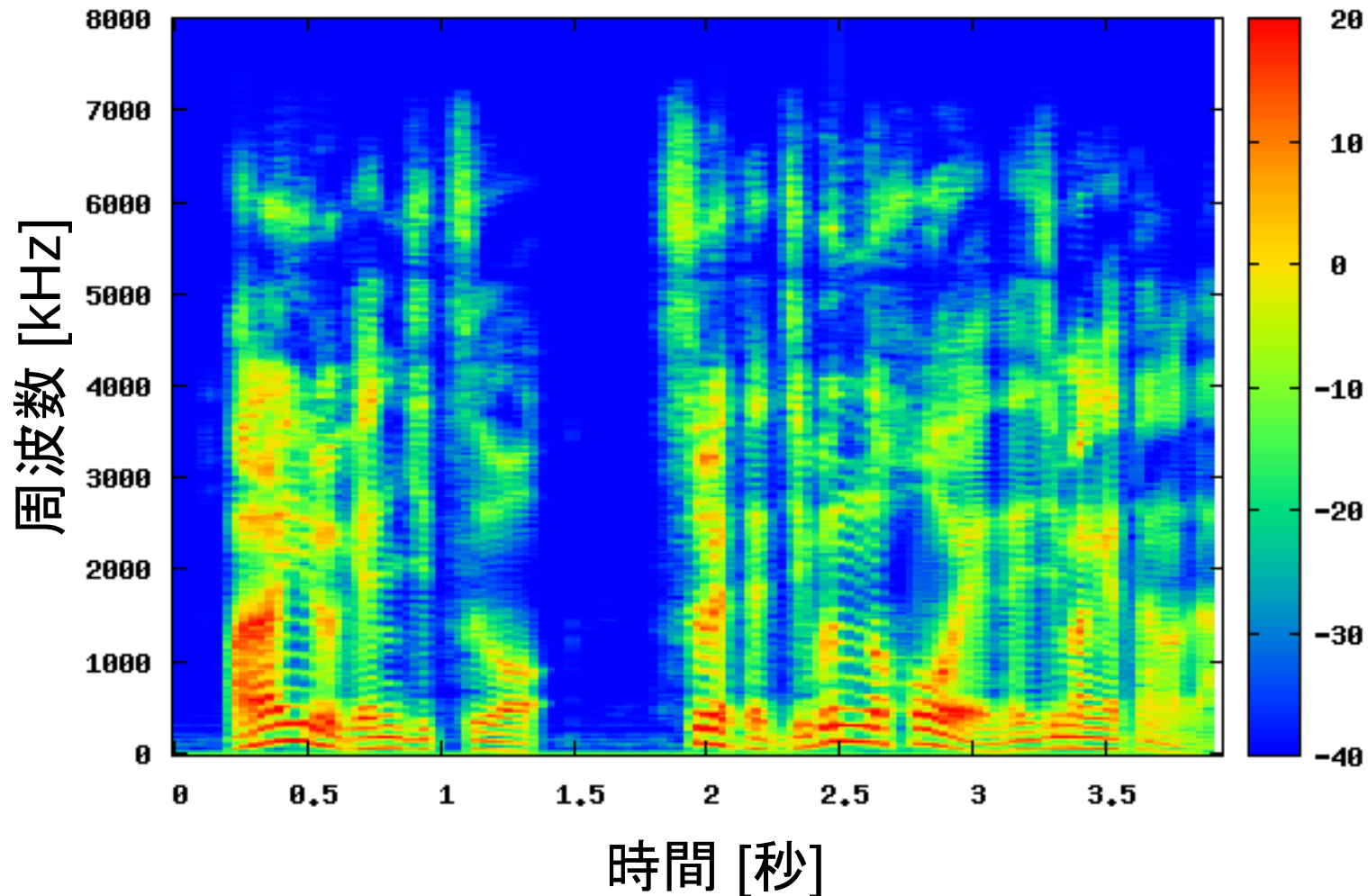


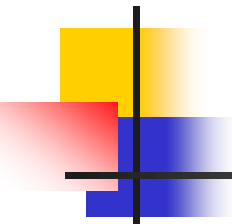
- 信号を単純な波(正弦波)の重ねあわせに分解



周波数:
1秒間の
振動の回数

音声信号の例(スペクトログラム表現)



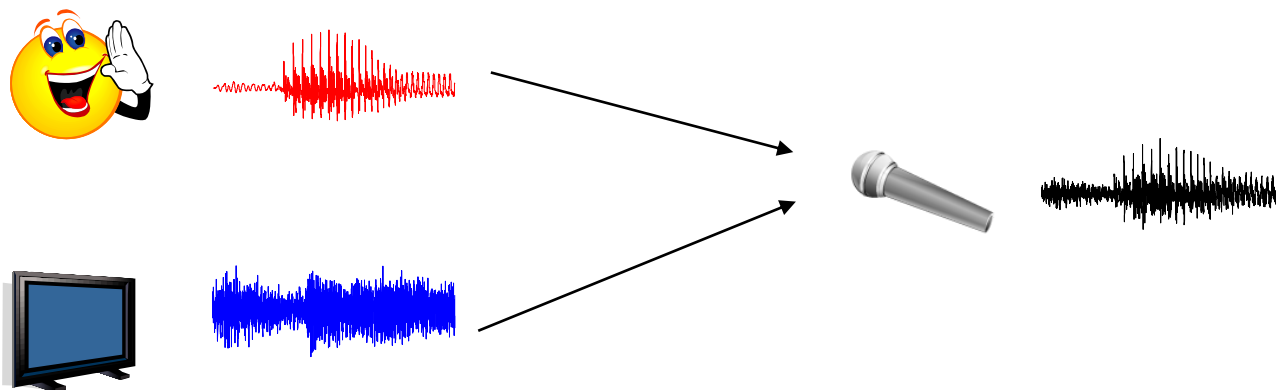


講演の流れ

- 音源分離への研究動機
- 音の性質
- 音源分離の原理1: 引き算アプローチ
 - マイクロフォンアレイによる音源分離
- 音源分離の原理2: 掛け算アプローチ
 - 調波音打楽器音分離
 - ボーカル抽出／ボーカルキャンセル
- まとめと今後の展望

音の混合

- 音は波なので加法的に(足し算)で混合する
 - 画像の場合: 複数のオブジェクトがあったら足し算というよりは隠す(occlusion)
 - 匂いの場合: 加法的ともいえるが引き算はない



音源分離の原理1:引き算 (1/2)

- 音は波であり、加法的に(足し算で)混ざっている
→ だったら引き算すれば元に戻る！
- 連立方程式の解法と同じ

$$2X + Y = 4$$

$$X + 2Y = 5$$

→
Yの大きさを
あわせる

$$4X + 2Y = 8$$

$$X + 2Y = 5$$

↓ 引き算によりYを消去

$$3X = 3$$



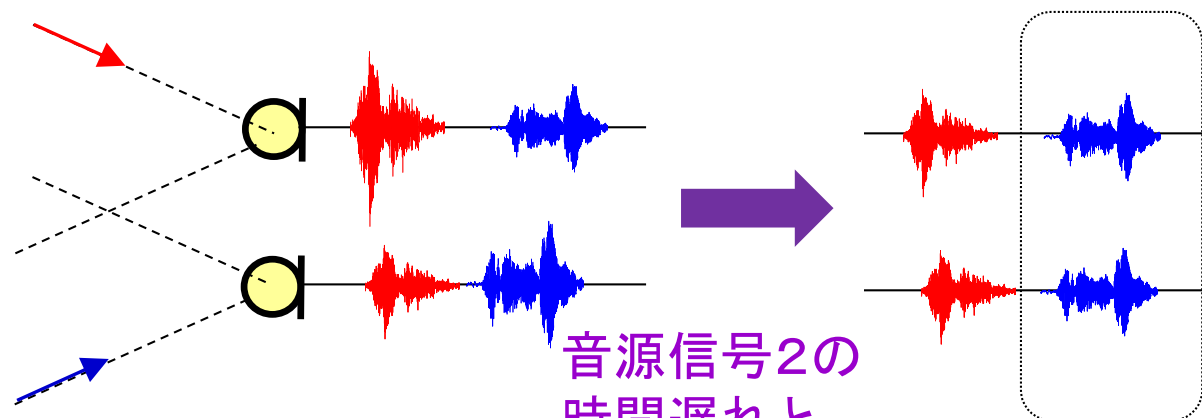
$$X = 1$$

Xの大きさを
1倍に戻す

音源分離の原理1:引き算 (2/2)

■ 以下は原理説明。実際は周波数成分毎に処理

音源1



音源2

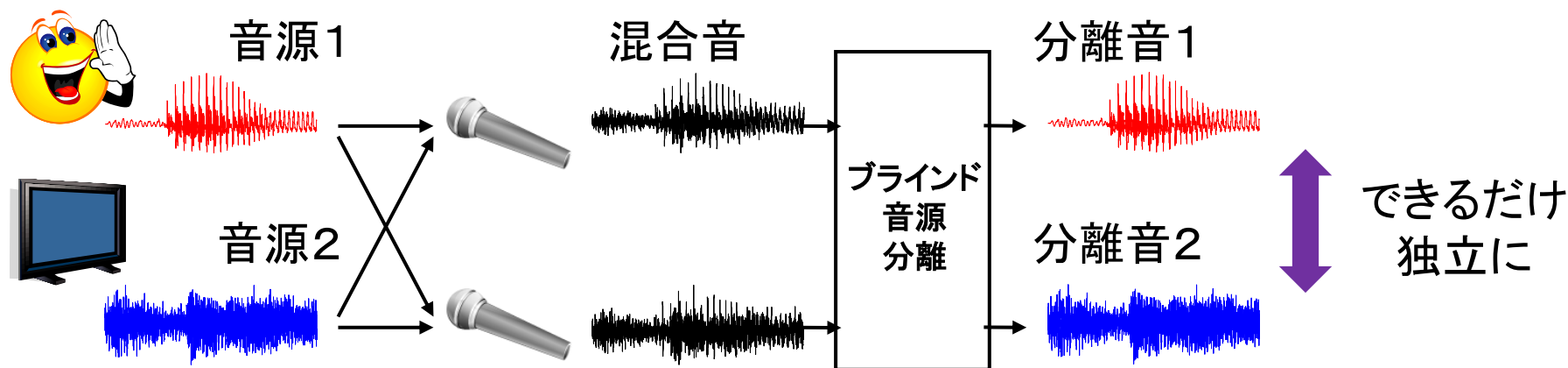
どうやって
あわせるかは
別の問題

引き算により
音源信号2を消去



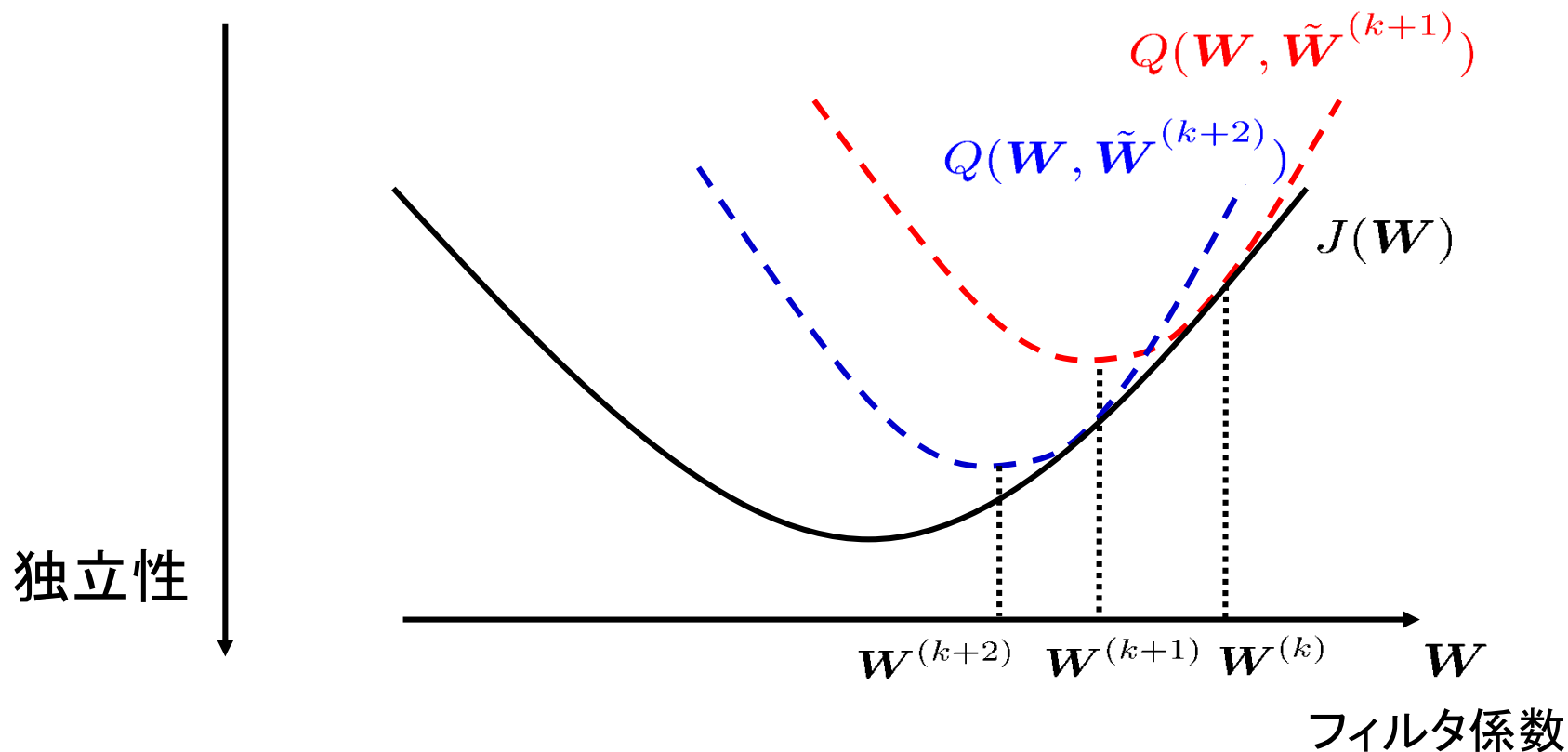
マイクロフォンアレイによる ブラインド音源分離

- 複数マイクを用い、音源位置が未知(**ブラインド**)な状態で、混合音を個々の音に分離する技術
- 信号の統計的な性質(独立性)を用いる
独立成分分析(独立ベクトル分析)という手法が発展
(ただし、一般には音源数以上のマイクロフォンが必要)



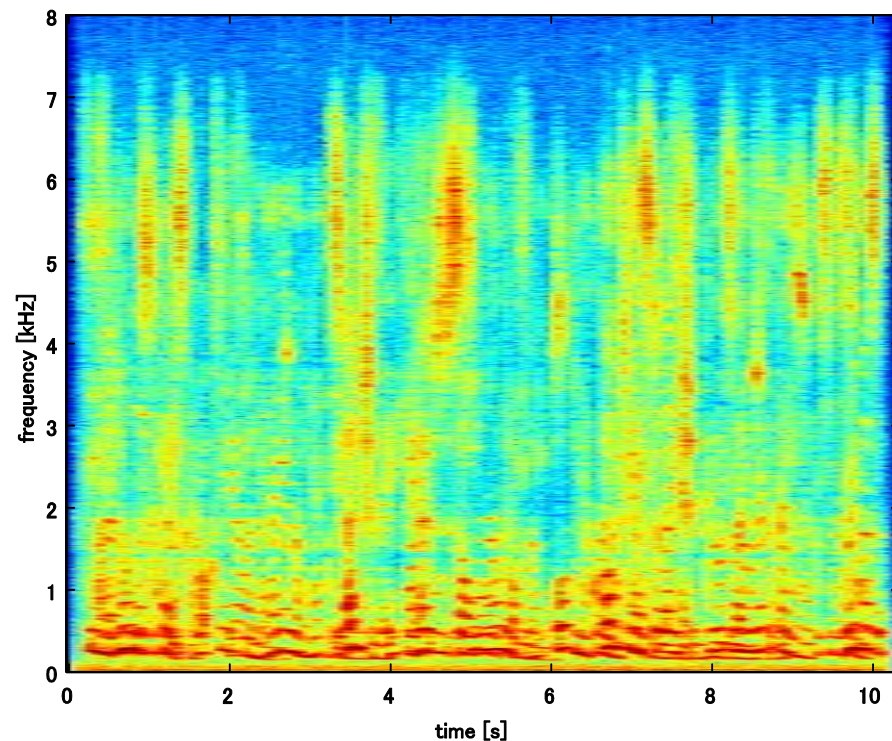
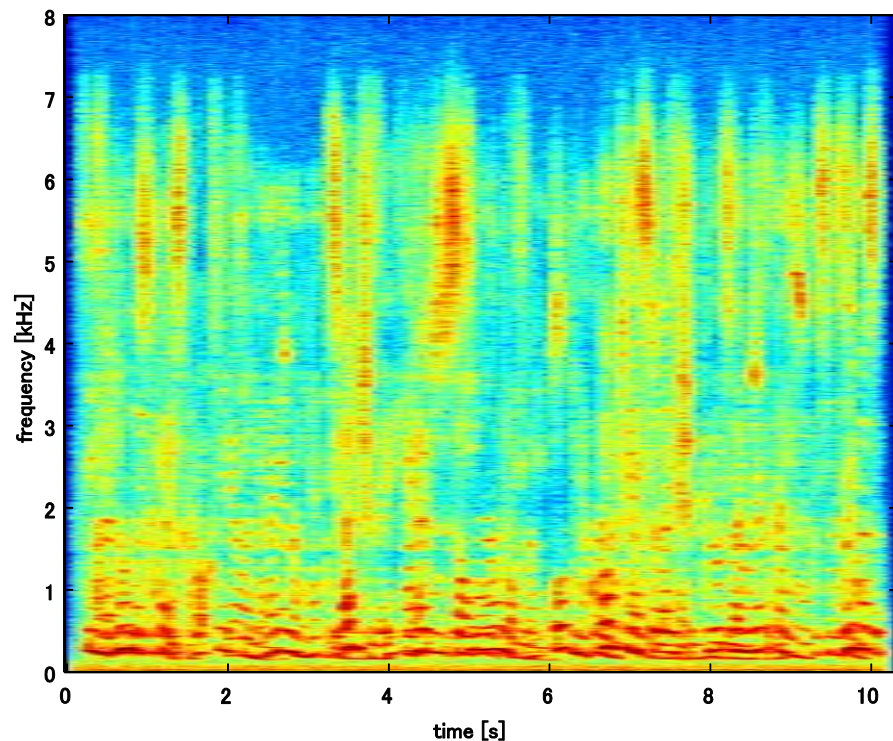
反復計算による最適化

- フィルタ係数(大きさをどうあわせるか、時間遅れをどうあわせるか)を、分離信号の独立性が最も大きくなるように決める
- 一度では解けないので反復計算で解く



音声分離例

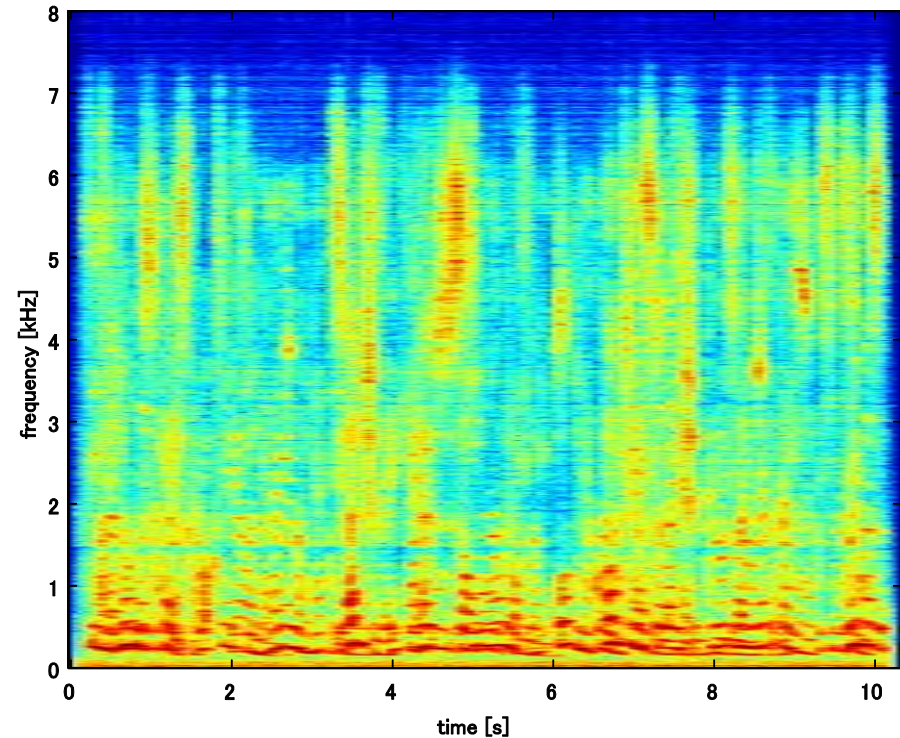
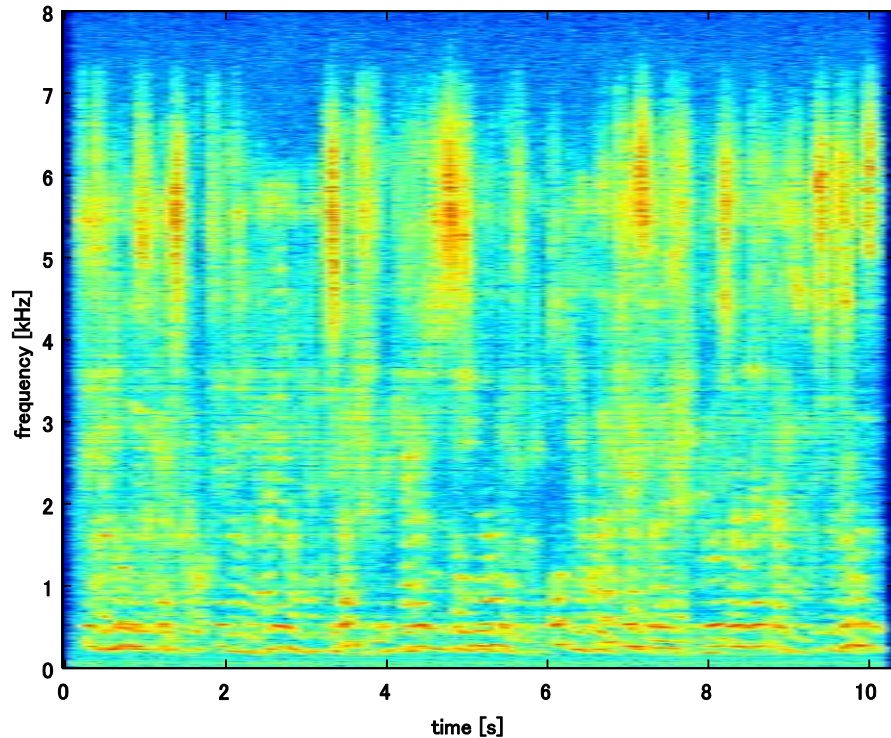
(残響時間 250ms、ライブ録音、マイク間隔5cm)



計算前

音声分離例

(残響時間 250ms、ライブ録音、マイク間隔5cm)

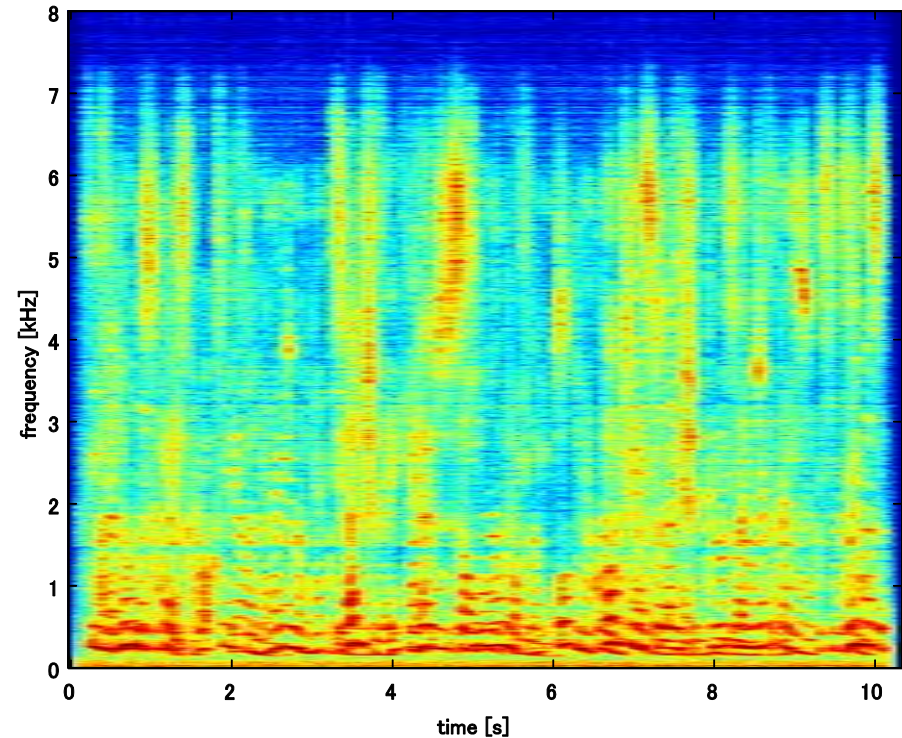
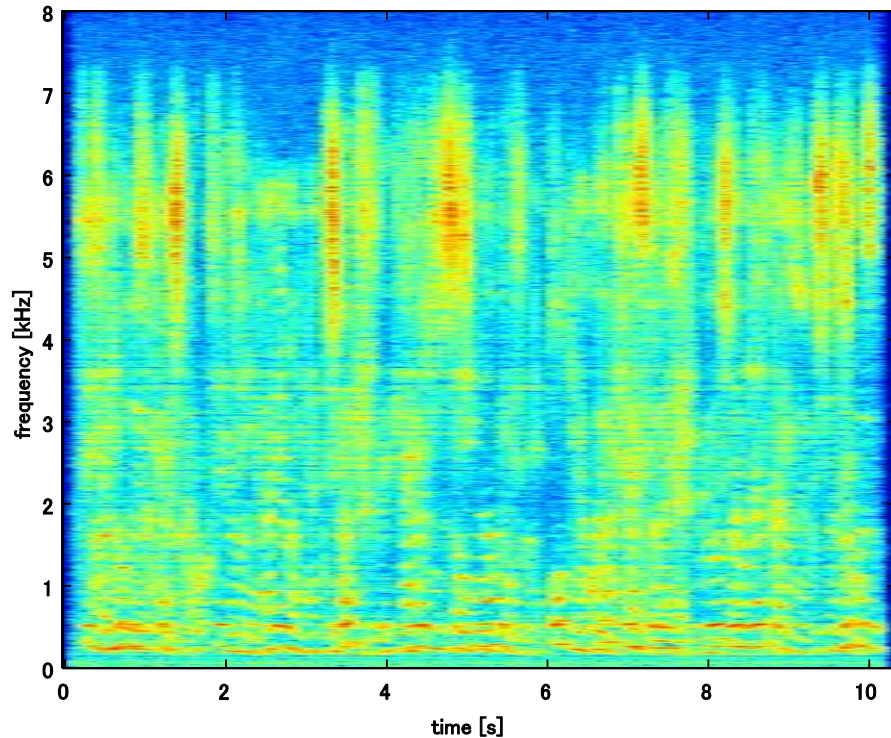


反復計算1回



音声分離例

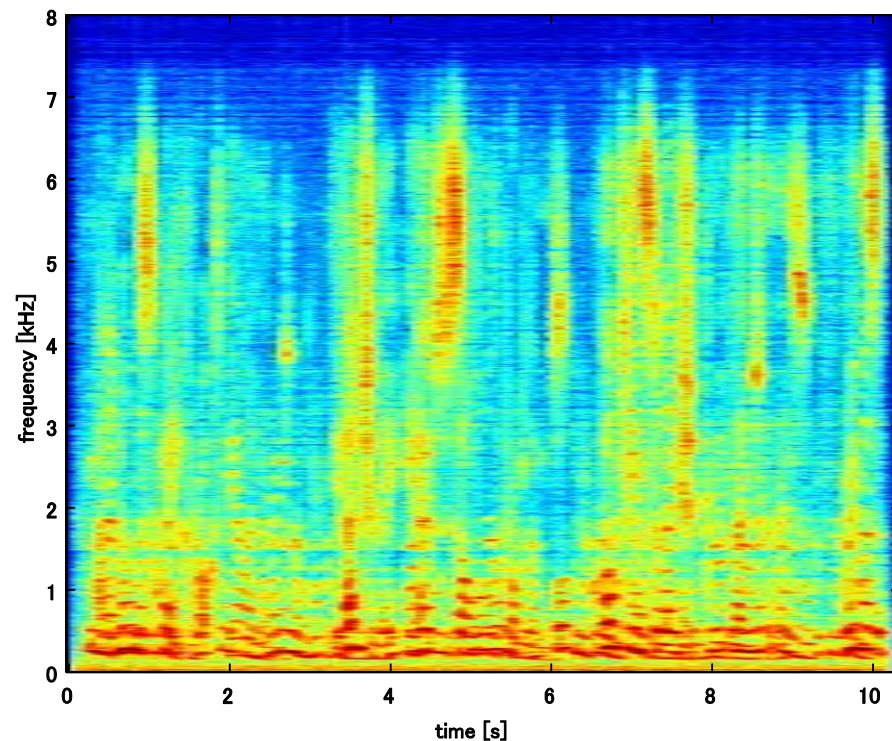
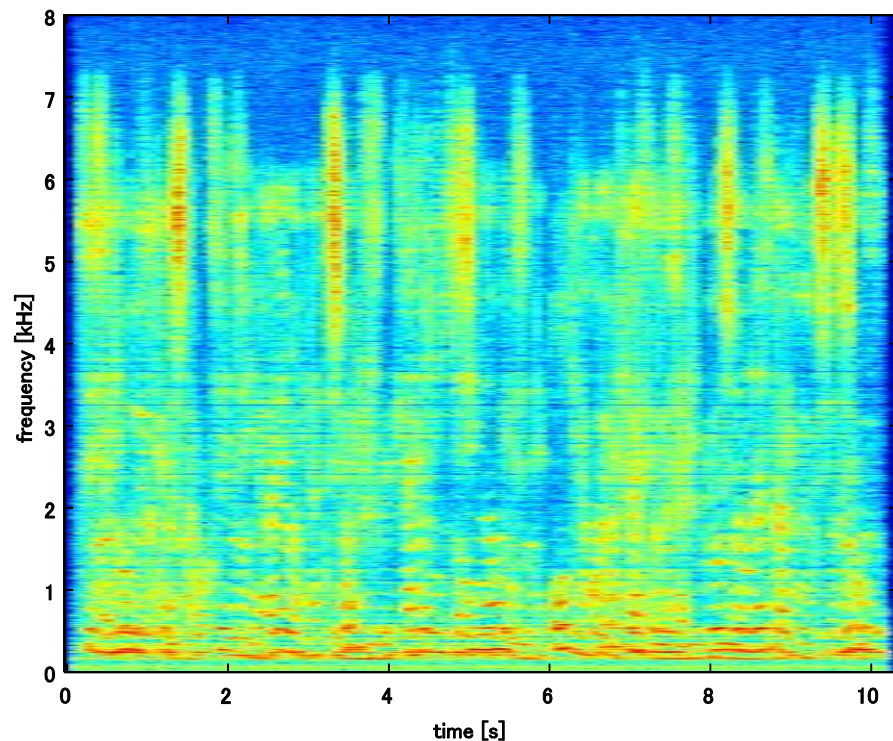
(残響時間 250ms、ライブ録音、マイク間隔5cm)



反復計算2回

音声分離例

(残響時間 250ms、ライブ録音、マイク間隔5cm)

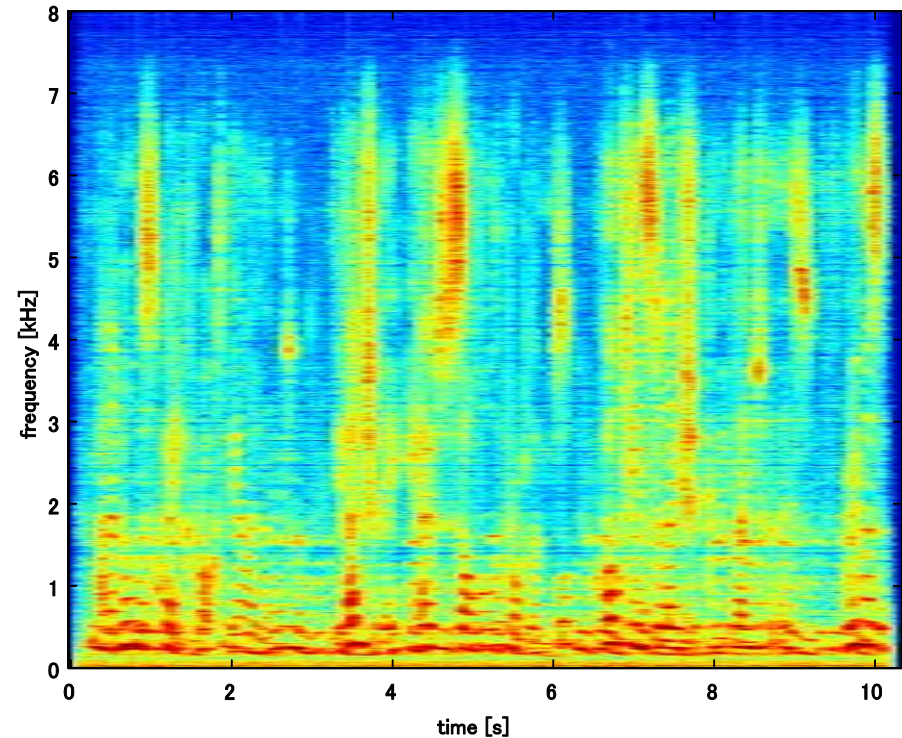
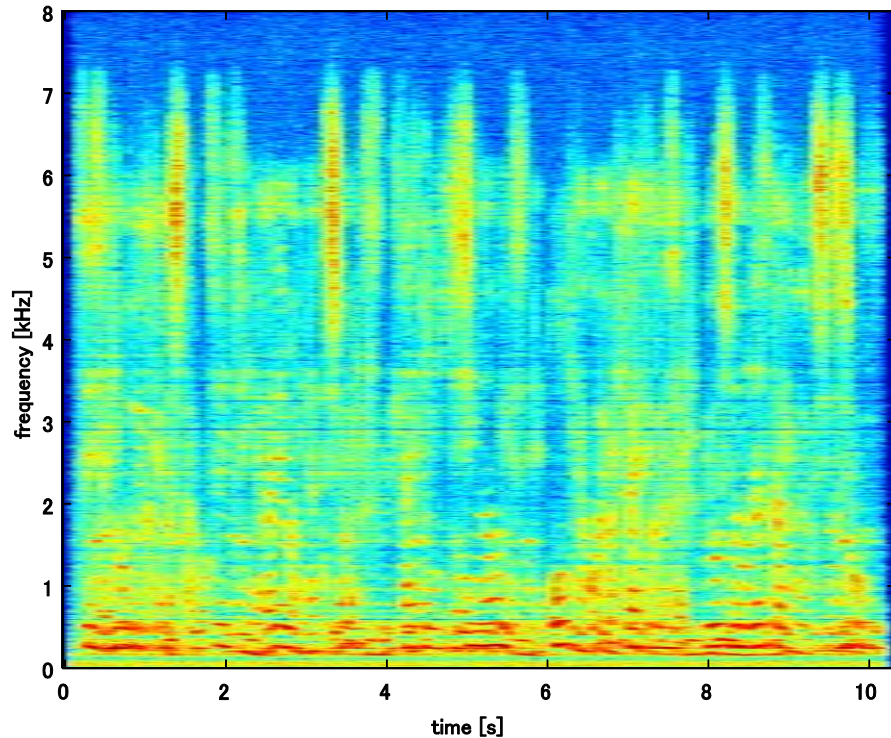


反復計算5回



音声分離例

(残響時間 250ms、ライブ録音、マイク間隔5cm)

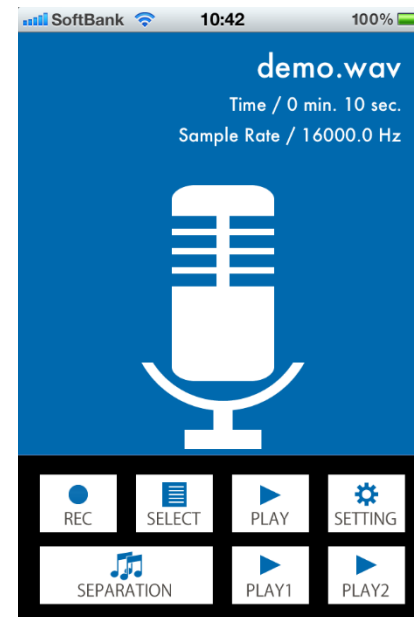


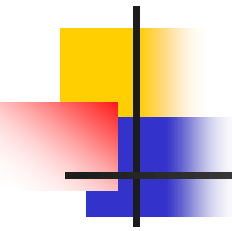
反復計算10回



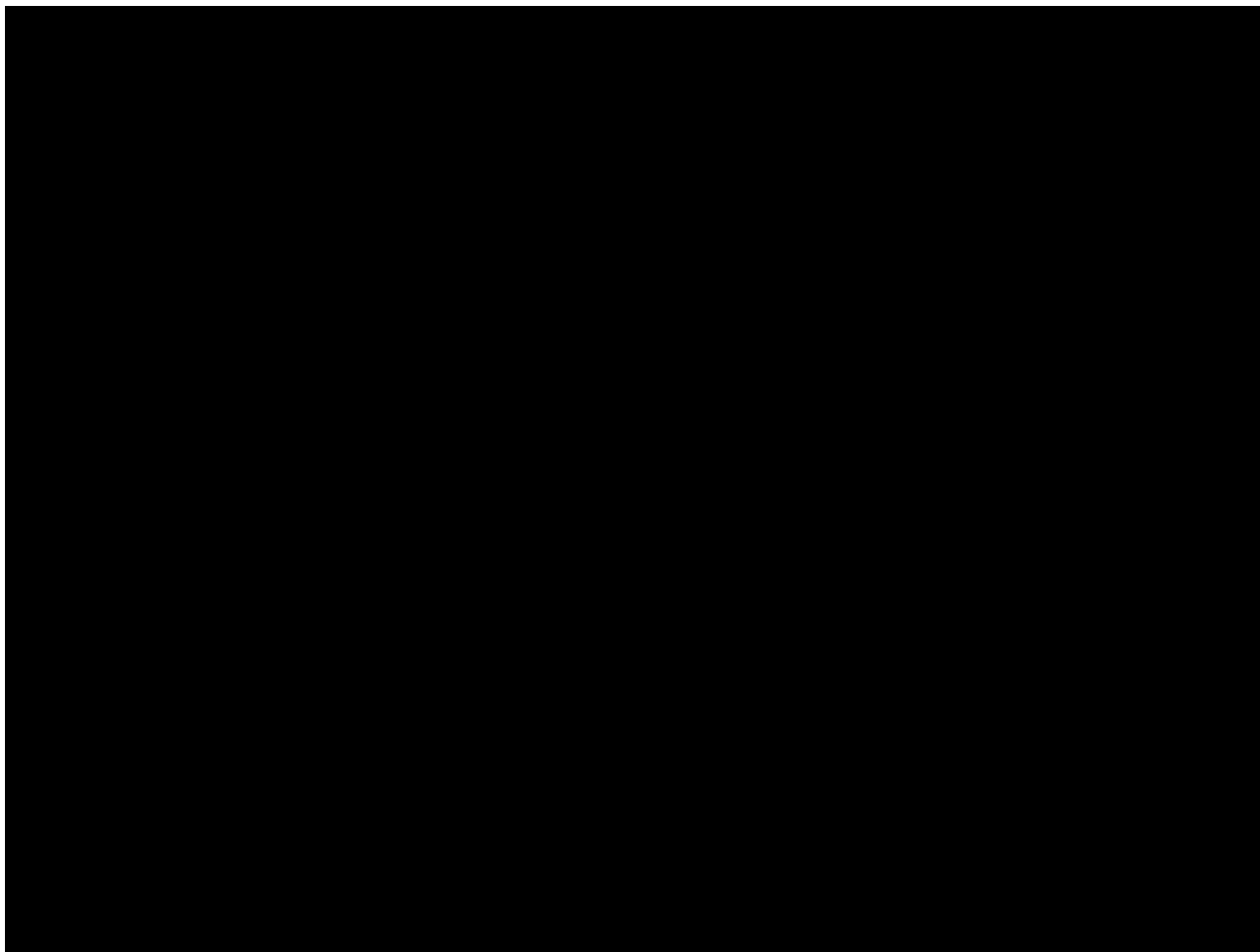
高速ブラインド音源分離のiPhone実装

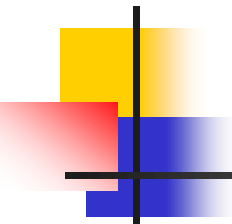
- 高速なブラインド音源分離アルゴリズムを開発 [Ono2011]
- iPhoneアプリとして実装 [Ono2012]
- ステレオマイクを装着したiPhoneで2つの音を高速に分離（10秒の音を2秒で分離）





ブラインド音源分離デモ



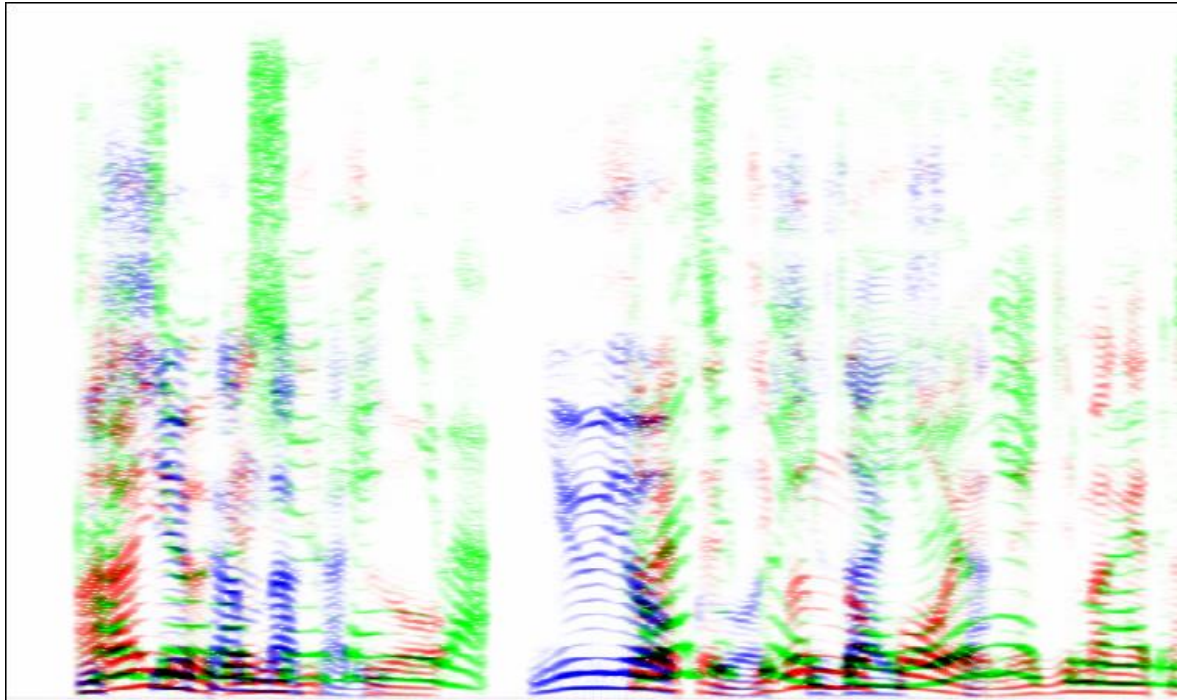


講演の流れ

- 音源分離への研究動機
- 音の性質
- 音源分離の原理1: 引き算アプローチ
 - マイクロフォンアレイによる音源分離
- 音源分離の原理2: 掛け算アプローチ
 - 調波音打楽器音分離
 - ボーカル抽出／ボーカルキャンセル
- まとめと今後の展望

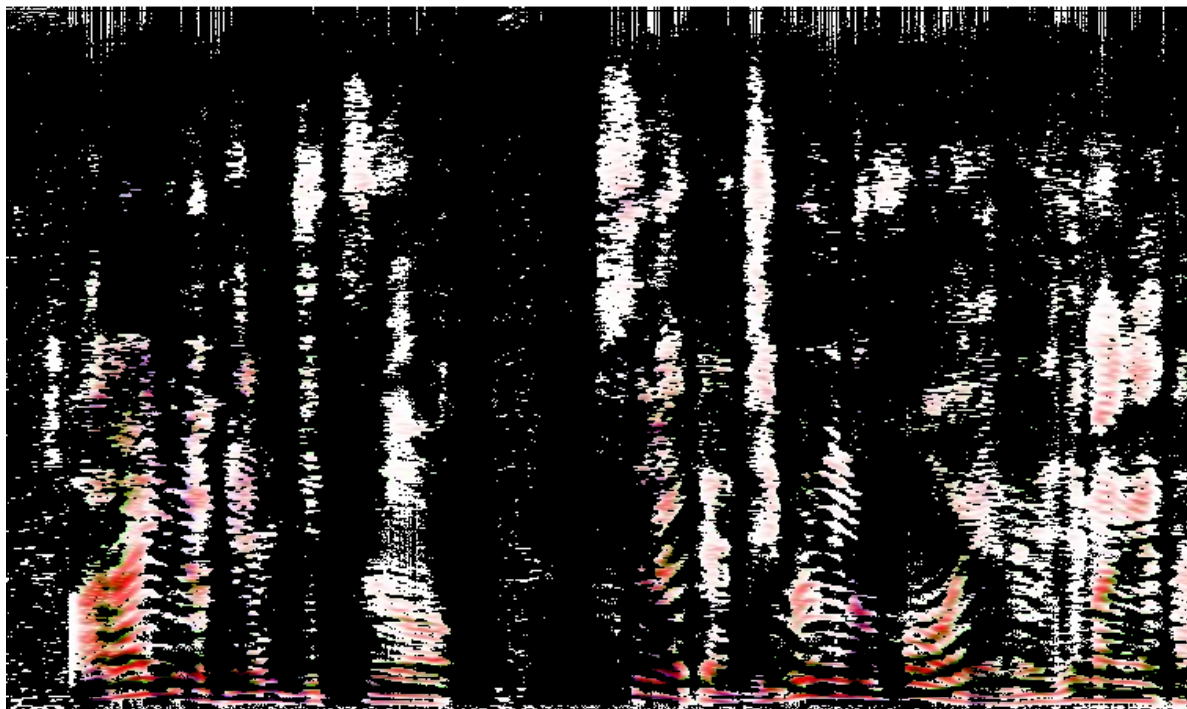
音源分離の原理2: 掛け算 (1/2)

- 戦略: できるだけばらばらにして、できるだけ目的のものを集めてくる
- 例: 3つの音声信号の重ね合わせ

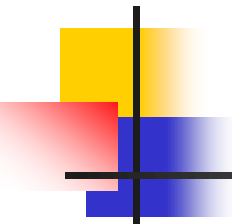


音源分離の原理2: 掛け算 (2/2)

- 時間波形としては重なっていても、
スペクトログラム上では割りと重なっていない(スパース性)
- 赤い成分が強かったら通す(1を掛ける)、
他の成分が強かったら通さない(0を掛ける)フィルタをつくと？



どうやって
つくるかは
別の問題



講演の流れ

- 音源分離への研究動機
- 音の性質
- 音源分離の原理1: 引き算アプローチ
 - マイクロフォンアレイによる音源分離
- 音源分離の原理2: 掛け算アプローチ
 - 調波音打楽器音分離
 - ボーカル抽出／ボーカルキャンセル
- まとめと今後の展望

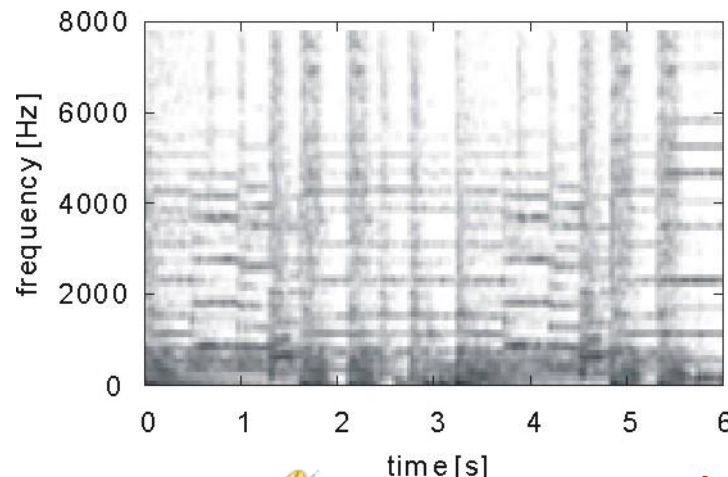
調波音／打楽器音分離

■ 動機

- 多くの楽曲(特にポップスやジャズなど)は、2つの性質の異なる音:
調波音と打楽器音を含んでいる
- 調波音(ピアノ、ギター、トランペット……): 主にメロディやコードに対応
- 打楽器音(バスドラム、ハイハット……): 主にリズムに対応

■ これらの分離は機械による音楽の認識や 能動的音楽鑑賞に有用

ポピュラー音楽の例
(RWC-MDB-P034)



harmonic component

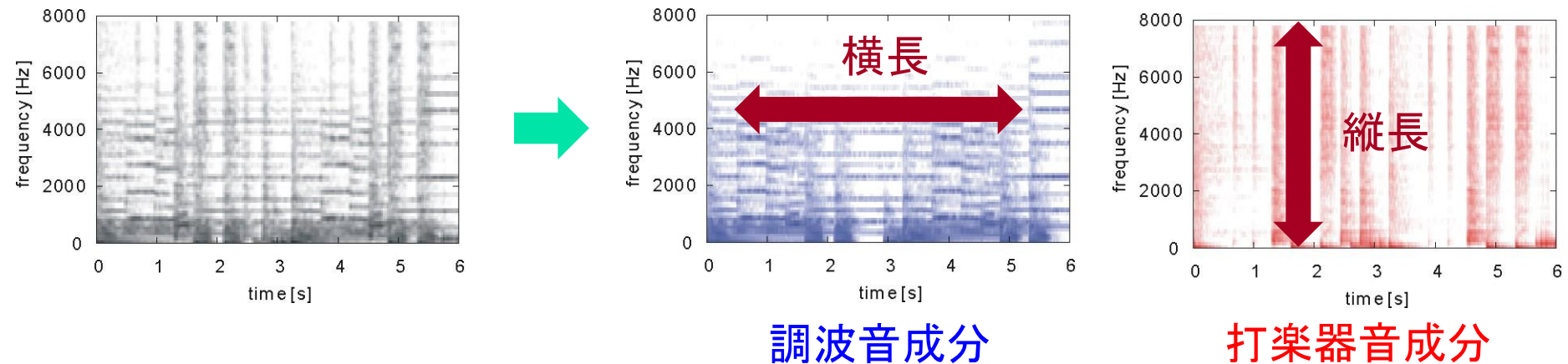


percussive component



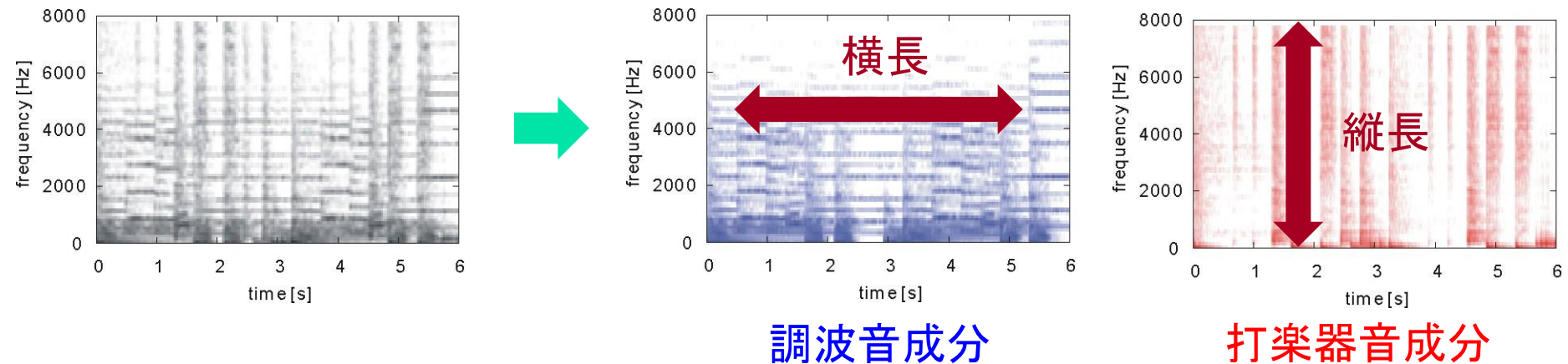
スペクトログラムの形の違い

- スペクトログラム上での形の違いに着目
 - 調波音: 横方向(時間方向)に滑らか
 - 打楽器音: 縦方向(周波数方向)に滑らか
- 入力信号をスペクトログラム上で2成分に分離する問題として定式化
- HPSS (Harmonic/Percussive Sound Separation)



HPSSの原理

- どうやって分離するか→方程式を解く
- $W(\text{混合音}) = H(\text{調波音}) + P(\text{打楽器音})$
 - これだけでは、無数の解がありえる
 - 例: $10 = 1 + 9, 2 + 8, 3 + 7, \dots$
 - Hはできるだけ横長に、Pはできるだけ縦長の成分になる
答えをみつける



実際に行っている計算

- 以下の計算を繰り返し行う（反復法）

$$H_{\omega,\tau} \leftarrow \left(\frac{b_{H_{\omega,\tau}} + \sqrt{b_{H_{\omega,\tau}}^2 + 4a_{H_{\omega,\tau}}c_{H_{\omega,\tau}}}}{2a_{H_{\omega,\tau}}} \right)^2$$

$$P_{\omega,\tau} \leftarrow \left(\frac{b_{P_{\omega,\tau}} + \sqrt{b_{P_{\omega,\tau}}^2 + 4a_{P_{\omega,\tau}}c_{P_{\omega,\tau}}}}{2a_{P_{\omega,\tau}}} \right)^2$$

$$m_{P_{\omega,\tau}} = \frac{P_{\omega,\tau}}{H_{\omega,\tau} + P_{\omega,\tau}}$$

$$a_{P_{\omega,\tau}} = \frac{2}{\sigma_P^2} + 2$$

$$b_{P_{\omega,\tau}} = \frac{(\sqrt{P_{\omega-1,\tau}} + \sqrt{P_{\omega+1,\tau}})}{\sigma_P^2}$$

$$c_{P_{\omega,\tau}} = 2m_{P_{\omega,\tau}}W_{\omega,\tau}$$

$$m_{H_{\omega,\tau}} = \frac{H_{\omega,\tau}}{H_{\omega,\tau} + P_{\omega,\tau}}$$

$$a_{H_{\omega,\tau}} = \frac{2}{\sigma_H^2} + 2$$

$$b_{H_{\omega,\tau}} = \frac{(\sqrt{H_{\omega,\tau-1}} + \sqrt{H_{\omega,\tau+1}})}{\sigma_H^2}$$

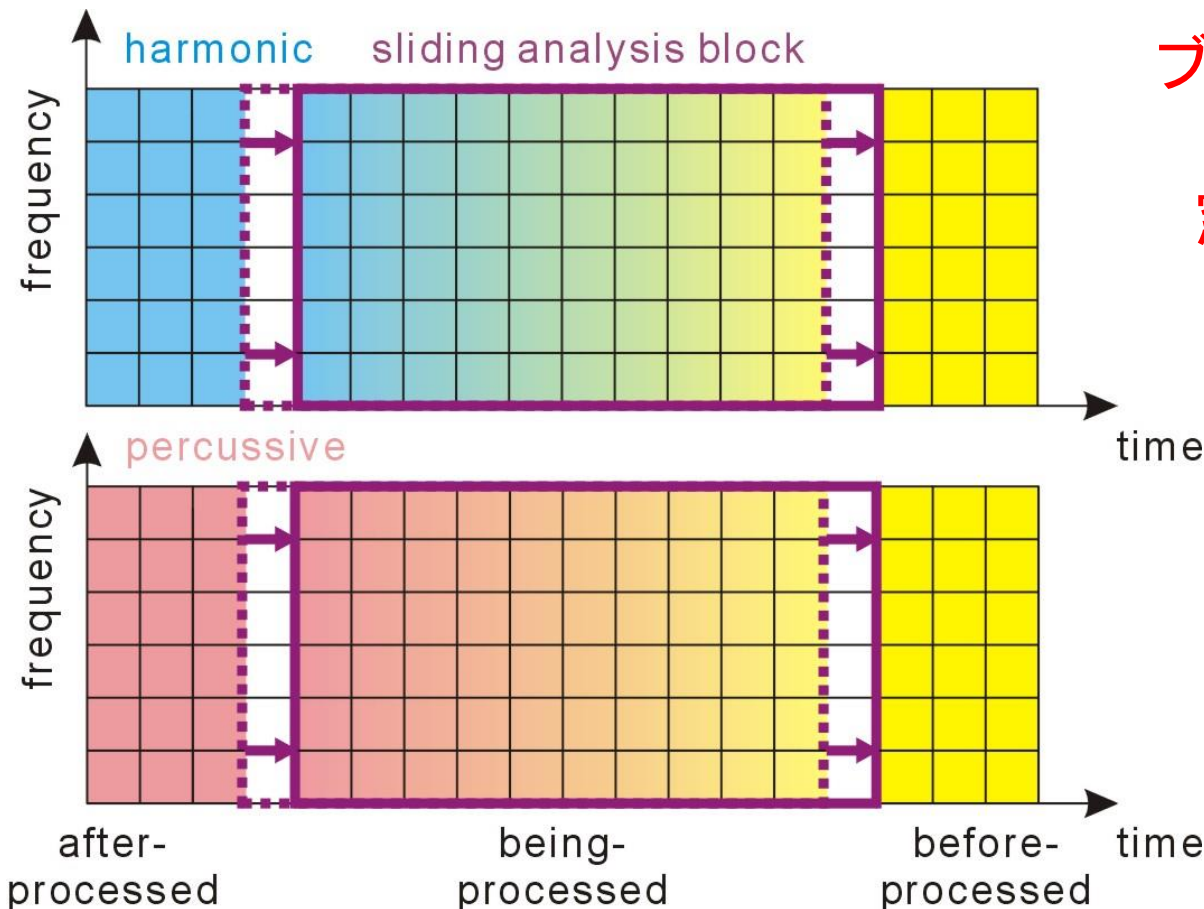
$$c_{H_{\omega,\tau}} = 2m_{H_{\omega,\tau}}W_{\omega,\tau}$$

音楽信号分離例

楽曲名	元信号	H	P
RWC-MDB-P-7 “PROLOGUE ”			
RWC-MDB-P-12 “KAGE-ROU ”			
RWC-MDB-P-18 “True Heart”			
RWC-MDB-P-25 “tell me”			
RWC-MDB-J-16 “Jive ”			

リアルタイムのための実装

■ スライディング窓分析

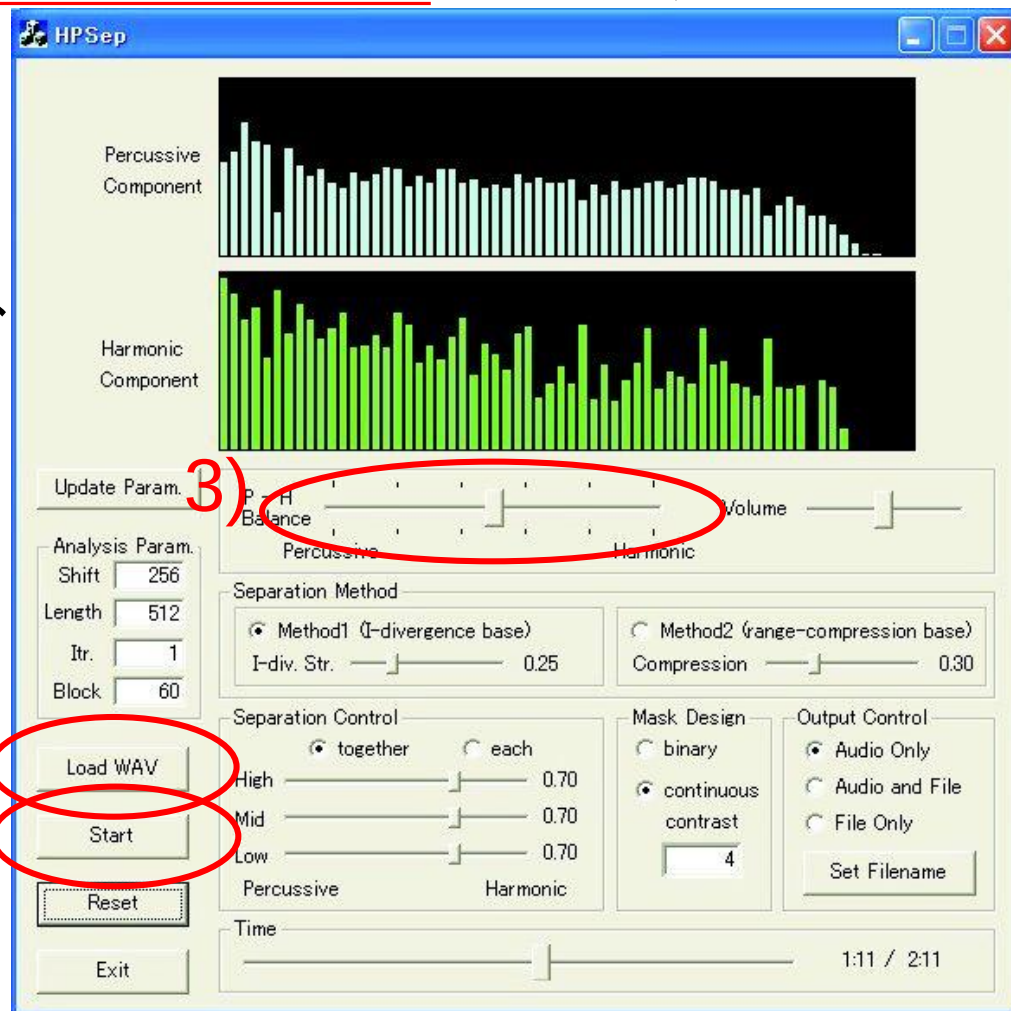


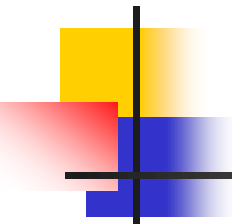
ブロック窓の中でだけ
反復計算を実行し
窓を少しずつシフト

オープンソフトウェア:実時間H/Pイコライザ

<http://www.hil.t.u-tokyo.ac.jp/software/HPSS/> からダウンロード可

- 音響信号のH/P成分を実時間で制御
- 操作方法:
 - 1) “Load WAV”ボタンをクリックし、WAVフォーマットの音響信号を読み込む
 - 2) “Start”ボタンをクリックすると、音響信号が再生
 - 3) H/P バランスバーをスライドさせることにより、H成分、P成分のバランスを自由に制御





講演の流れ

- 音源分離への研究動機
- 音の性質
- 音源分離の原理1: 引き算アプローチ
 - マイクロフォンアレイによる音源分離
- 音源分離の原理2: 掛け算アプローチ
 - 調波音打楽器音分離
 - ボーカル抽出／ボーカルキャンセル
- まとめと今後の展望

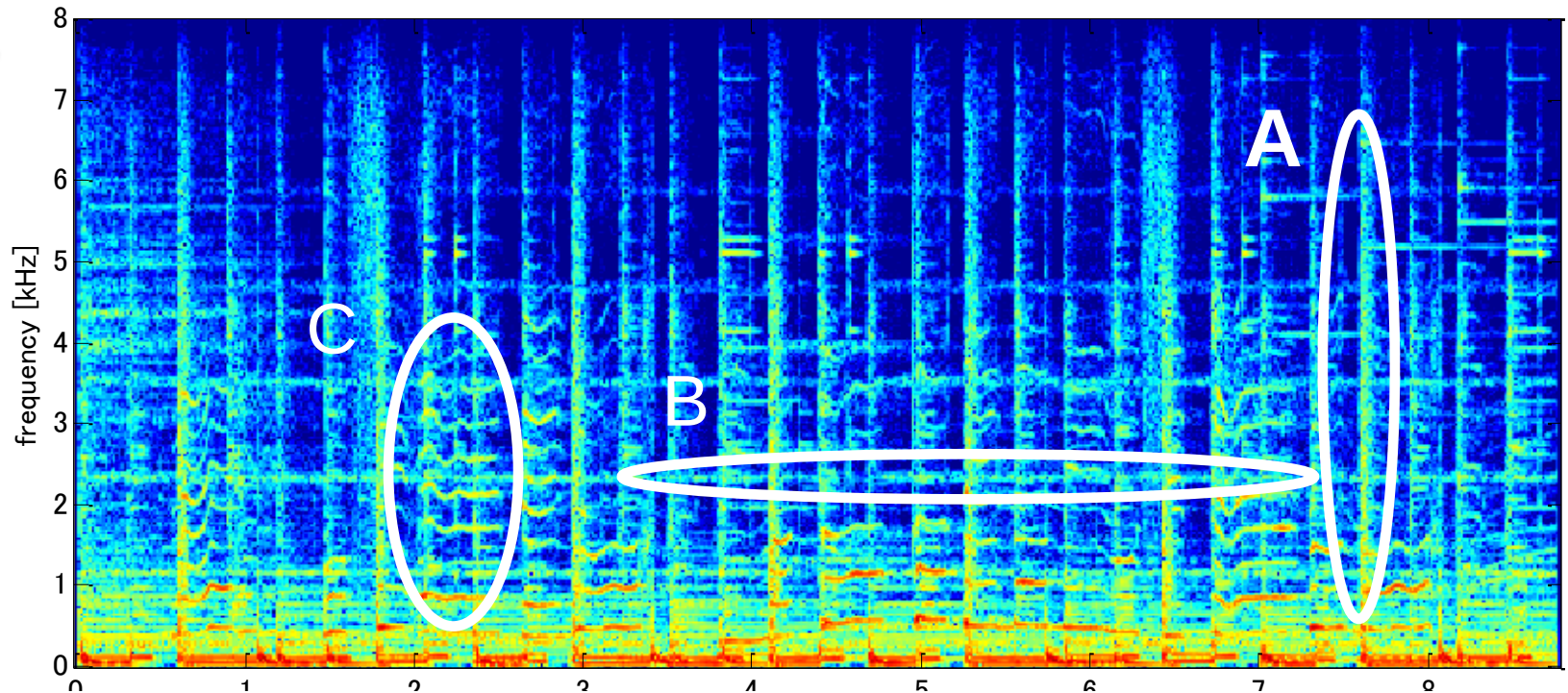
ボーカル分離への応用

- 目的
音楽音響信号のボーカルと伴奏への分離
- 応用分野の例
 - カラオケの自動生成
 - 歌唱データベース作成→音楽検索への応用



歌声を含むスペクトログラム

RWC-MDB-P-25 “tell me”

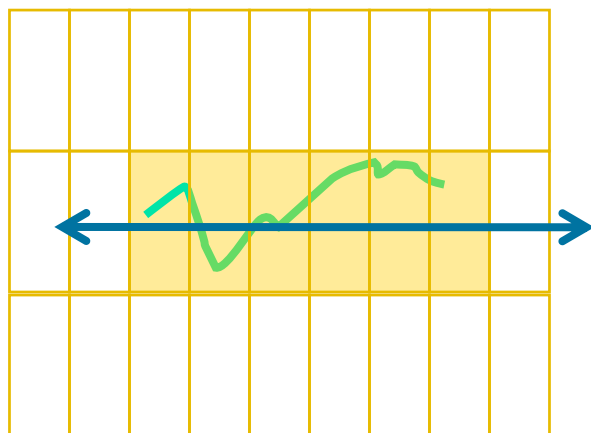
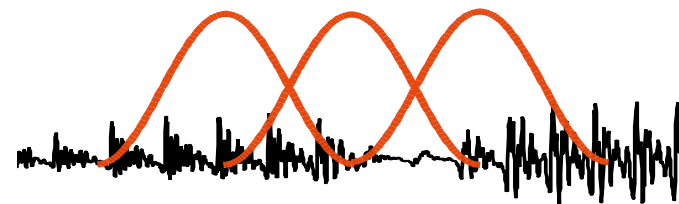


- A. 縦成分: 打楽器音
- B. 横成分: 調波音
- C. ゆらぎのある成分: 歌声

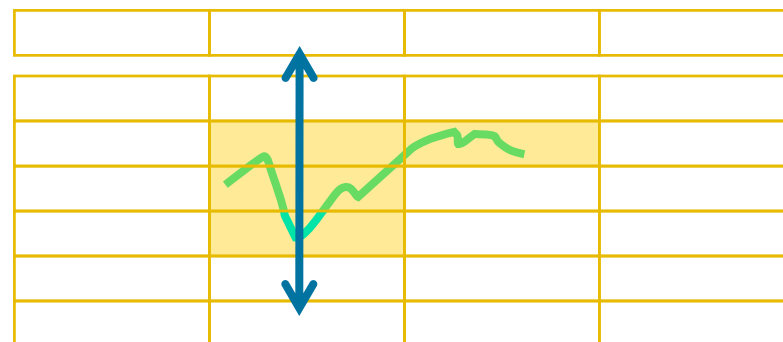
歌声のスペクトログラムは縦長？横長？

スペクトログラム分析時の分析フレーム長に依存

- 短い窓では横長に見える
- 長い窓では縦長に見える



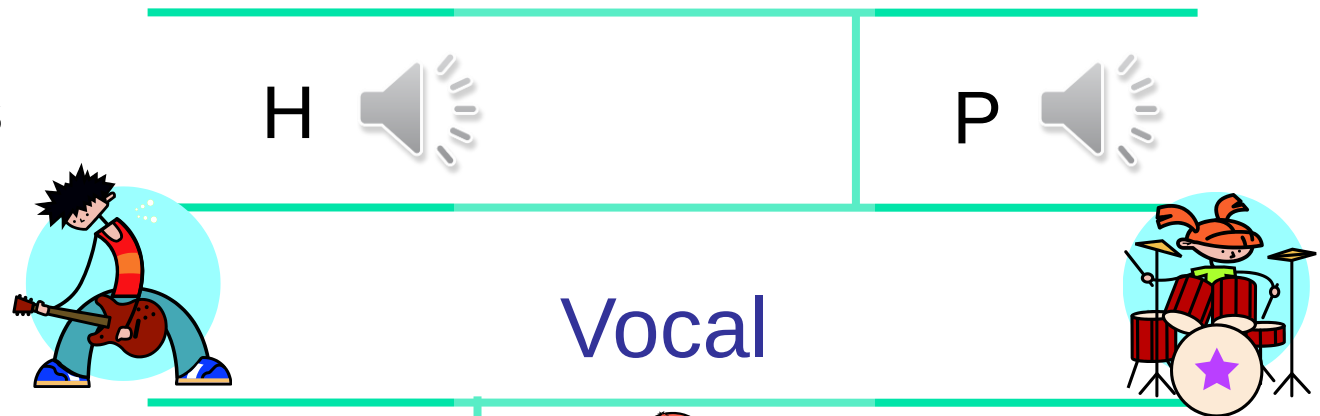
調波的



打樂器的

HPSSのフレーム長による違い

フレーム長: 16ms



フレーム長: 512ms



2段階HPSS

元信号

短時間フレームHPSS

打楽器音

長時間フレームHPSS

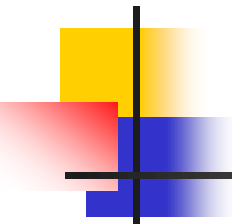
調波音

歌声



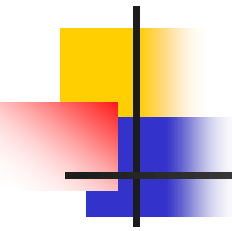
歌声と伴奏の分離例

タイトル	入力	歌声	伴奏	ジャンル
tell me				F, R&B
PROLOGUE				F, J-POP
恋に落ちる時間に関する考察				M, 渋谷系
枯れ葉色のTwilight				F, J-POP
Weekend				F, ユーロビート
Dance Together				M, Jazz
1999				M, メタル
七つの子				F, 童謡
歌劇「リゴレット」より, 「女心の歌」				M, クラシック



講演の流れ

- 音源分離への研究動機
- 音の性質
- 音源分離の原理1: 引き算アプローチ
 - マイクロフォンアレイによる音源分離
- 音源分離の原理2: 掛け算アプローチ
 - 調波音打楽器音分離
 - ボーカル抽出／ボーカルキャンセル
- まとめと今後の展望



まとめ

- 実世界では様々な音が存在し、混合して聞こえてくる
→機械による音の認識に音源分離が有用
- 音源分離の原理1:引き算アプローチ(線形フィルタ)
 - 引き算で干渉音を消去(連立方程式と同じ仕組み)
 - 音の方向の違いを利用
 - 複数マイクが必要
- 音源分離の原理2:掛け算アプローチ(時間周波数マスキング)
 - スペクトログラム上で掛け算により目的音を強調(選り分け)
 - 音のスペクトルの違いを利用
 - 単一マイクでも可能

今後の展望

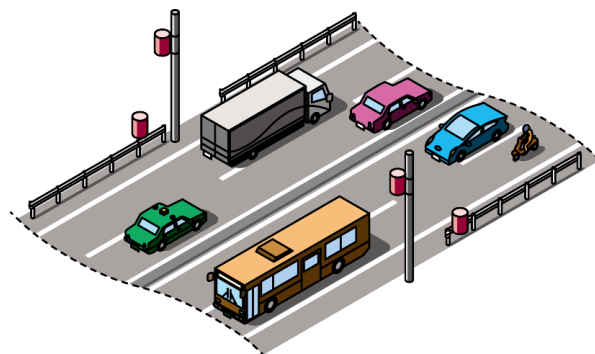
- スマートフォンの普及
→音の入出力デバイスがいたるところに
- 機械学習・信号処理理論の発展
→音声、音楽からあらゆる音の認識へ
- 暮らしを変える「機械の耳」の実現を目指して



どこでも音声認識



撮影した動画の音を
自由に編集



音による交通量の
モニタリング



危険な音の認識・検知