Computational Musical Instrument Recognition and Its Application to Content-based Music Information Retrieval

（計算機による楽器音の認識および内容に基づく音楽情報検索への応用）

北原 鉄朗

京都大学大学院情報学研究科知能情報学専攻
奥乃研究室

15 Feb. 2007

導入

研究の背景 (1/2)

現状の計算機技術の不足点

実世界の認識

特に**聴覚メディア**の認識は，
単一話者の音声認識以外発展途上

聴覚メディア認識の難しさ

複数の音響イベントの同時発生

従来の典型的な音声認識研究では対象外

⬍

計算機による聴覚的情景分析（CASA）

音声に限定されない
混合音のグラウンディング

研究の背景 (2/2)

音楽情景分析 [Kashino'94] etc.

音楽音響信号を対象としたCASA研究

- 音楽はCASA研究の絶好の題材
  - 複数楽器の同時発音，楽譜による記号表現
  - どれがsignalで，どれがnoiseか一意的でない
- 応用サイドからの要請
  - 電子的音楽配信・大容量音楽プレイヤーの普及
    →音楽情報検索（MIR）技術の必要性
  - 音楽の「中身」に基づく検索には不可欠

研究の目的

本研究の目的

多重奏に対する楽器認識

入力：音楽音響信号 ➡ 出力：楽器に関する記述

楽器名 or 他の表現

- 音楽情景分析における重要性
  - 複数楽器を「聞き分け」る最も基本的な手がかり
- 音楽情報検索における重要性
  - 「何の楽器か」は楽曲を特徴づける重要な要素
    e.g. ピアノソロ，弦楽四重奏

国内外の研究状況 (1/2)

音楽音響信号処理

自動採譜・F0推定がメイン

- 単一楽器による和音演奏の採譜 [Katayose ’89] etc.
- 複数楽器による多重奏の認識 [Kashino ’99] etc.
- 市販CDに対するメロディとベースのF0推定 [Goto ’99]
- 拘束付GMM [Kameoka’04,’05], NMF [Smaragdis’03] etc.

多重奏に対する楽器認識は発展途上

- 単一音対象は近年増加（10～30楽器に対して70～80%）
- 多重奏対象は数件のみ（3～5楽器による二～三重奏）

国内外の研究状況 (2/2)

音楽情報検索（MIR）

音楽類似度の研究がさかん

[Aucouturier ’02]
[Pampalk’04] etc.

- MFCC，クロマなどの low-level の特徴量がメイン
  - 楽器の音色，ハーモニー，録音時の音質を区別せずに表現
  - ハーモニー（和音）に関連する特徴量

  ⇒ 楽器，和音進行，リズムなどの各音楽要素に対応する higher-level の特徴量 必要

- 主な応用例：
  類似楽曲検索（Query-by-Example），楽曲コレクションの可視化[Rauber’02]，プレイリスト生成[Aucouturier’02]etc.

単一音に対する楽器認識

音高による音色変化

- 音響特徴量のF0による変化を関数近似する「**F0依存多次元正規分布**」の提案
- 19楽器6,247音の単一音DBで認識率約80%

未知楽器

- 未知楽器の**カテゴリーレベルでの認識**
- カテゴリーレベルの認識に適した**楽器カテゴリーを音響類似から**自動的に獲得
- 自然楽器音のみを学習したシステム上で未知の電子楽器音の約77%を正しく認識

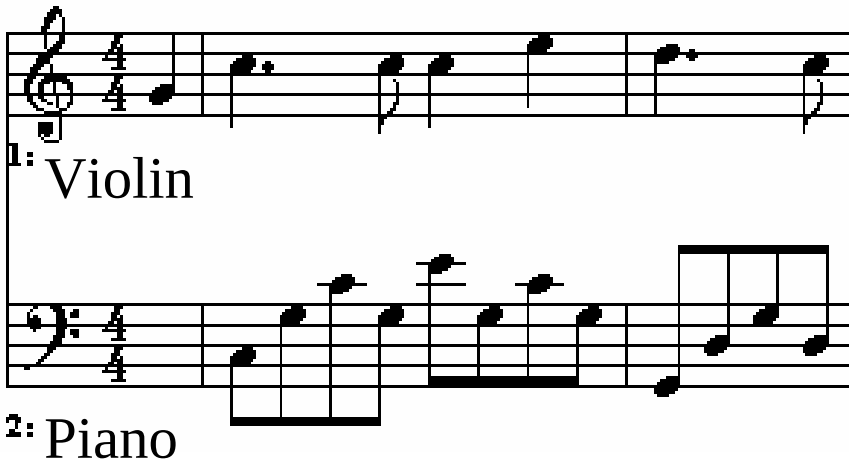
多重奏に対する楽器認識

多重奏に対する楽器認識 (1/2)

基本的な問題設定

- 入力信号には複数の単音$n_1, \ldots, n_K$を含み，同時に鳴る場合もある
  - $n_1, \ldots, n_K$：1音符の音
- 同時に鳴る楽器は，同種の場合も異種の場合も
- 各単音に対して，それを演奏する楽器名を出力

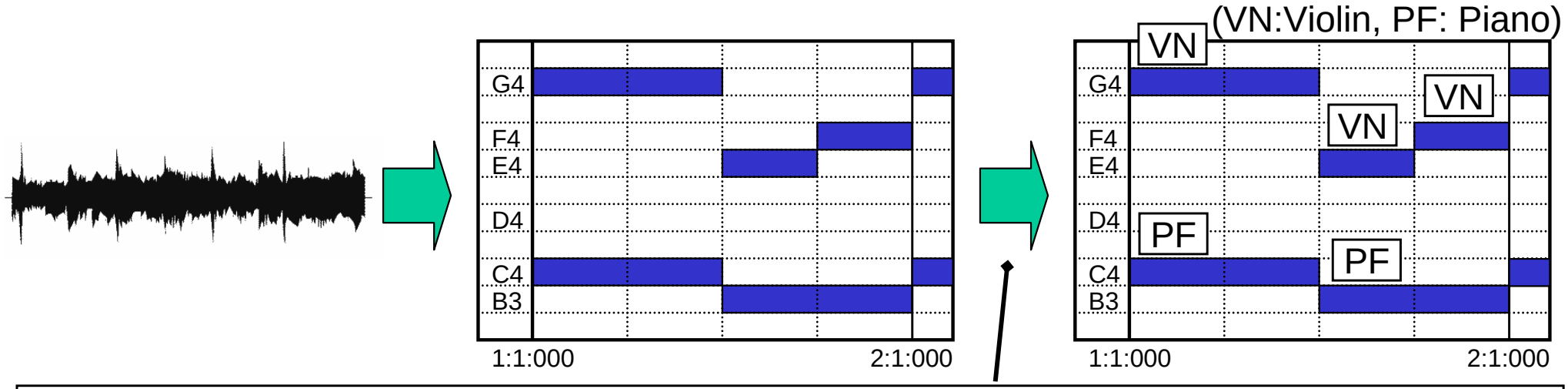
入力楽曲例



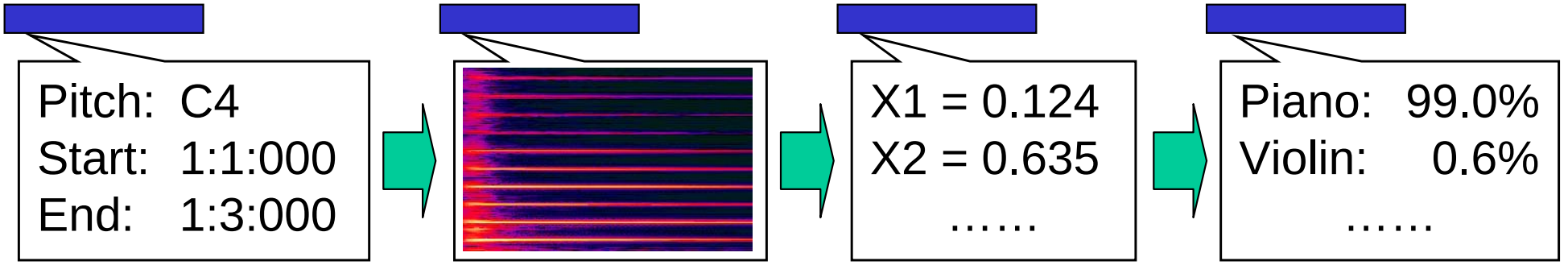
多重奏に対する楽器認識 (2/2)

典型的なアプローチ

1. 単音ごとに発音時刻・F0を推定して調波構造抽出
2. 抽出された調波構造から特徴抽出し，楽器同定



For each note…



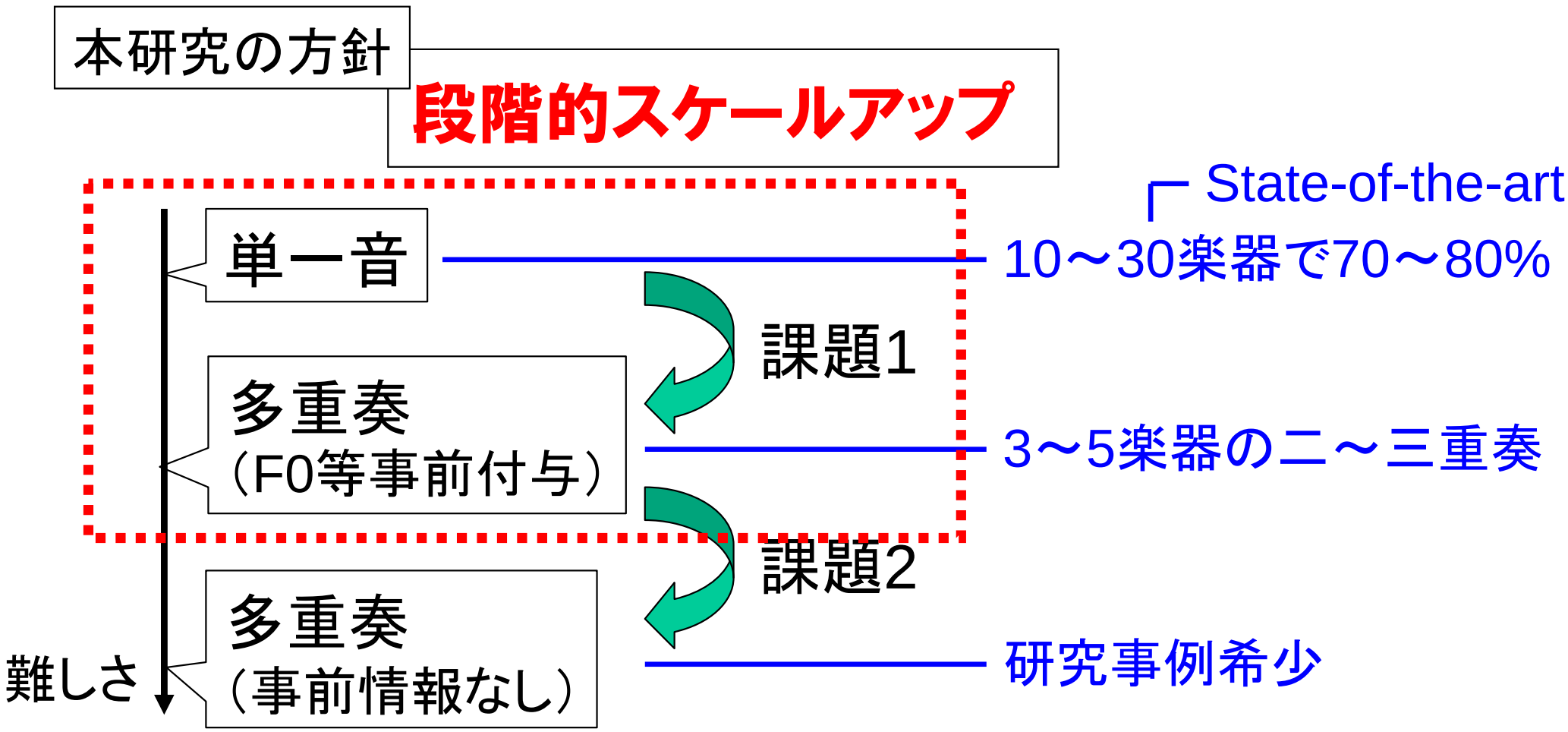
多重奏に対する楽器認識の課題

課題1　**音の重なり**による特徴量の変化

課題2　**発音時刻・F0推定のエラー**の影響

本研究の方針

段階的スケールアップ



課題１　音の重なりによる特徴量の変化

課題1 音の重なりによる特徴量の変化

複数の楽器が同時に発音

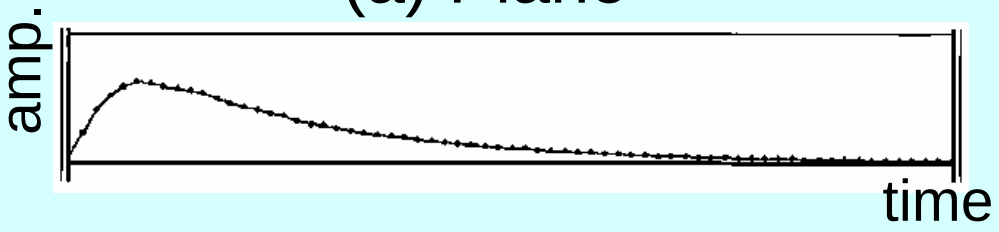
周波数が共通の
周波数成分が重複・干渉

そこから抽出した特徴量が
単一音の場合と変化

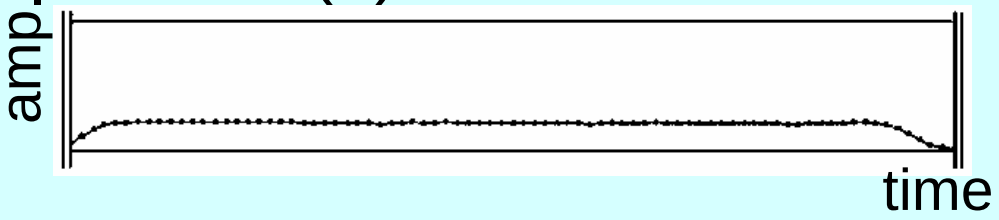
周波数成分重なりの例

ある周波数におけるパワー包絡

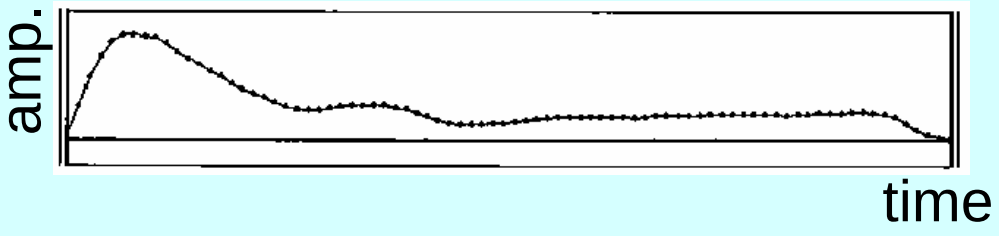
(a) Piano



(b) Clarinet



(c) Piano+Clarinet



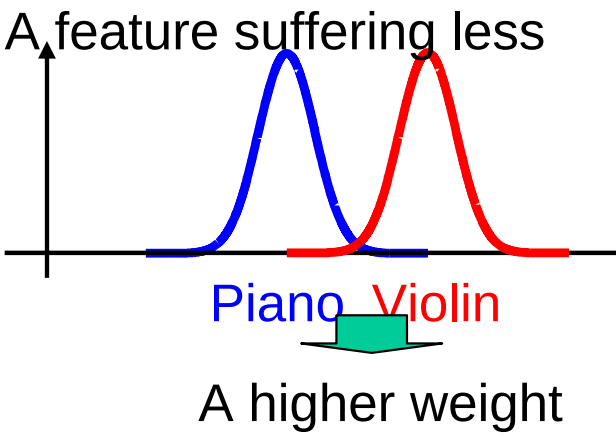
音の重なりに頑健な特徴量への重み付け

ねらい

音の重なりによる影響が
- 小さな特徴量に大きな重みを
- 大きな特徴量に小さな重みを

課題

音の重なりの影響の度合いの定量化

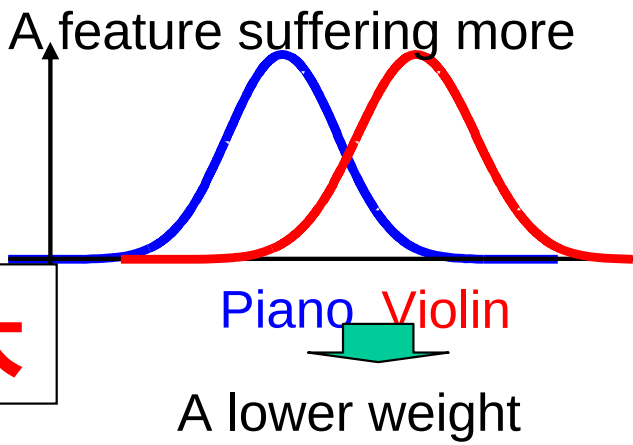


着眼点

特徴抽出を混合音から行った場合

干渉音が様々に変われば,
特徴量の変化の仕方も様々なはず

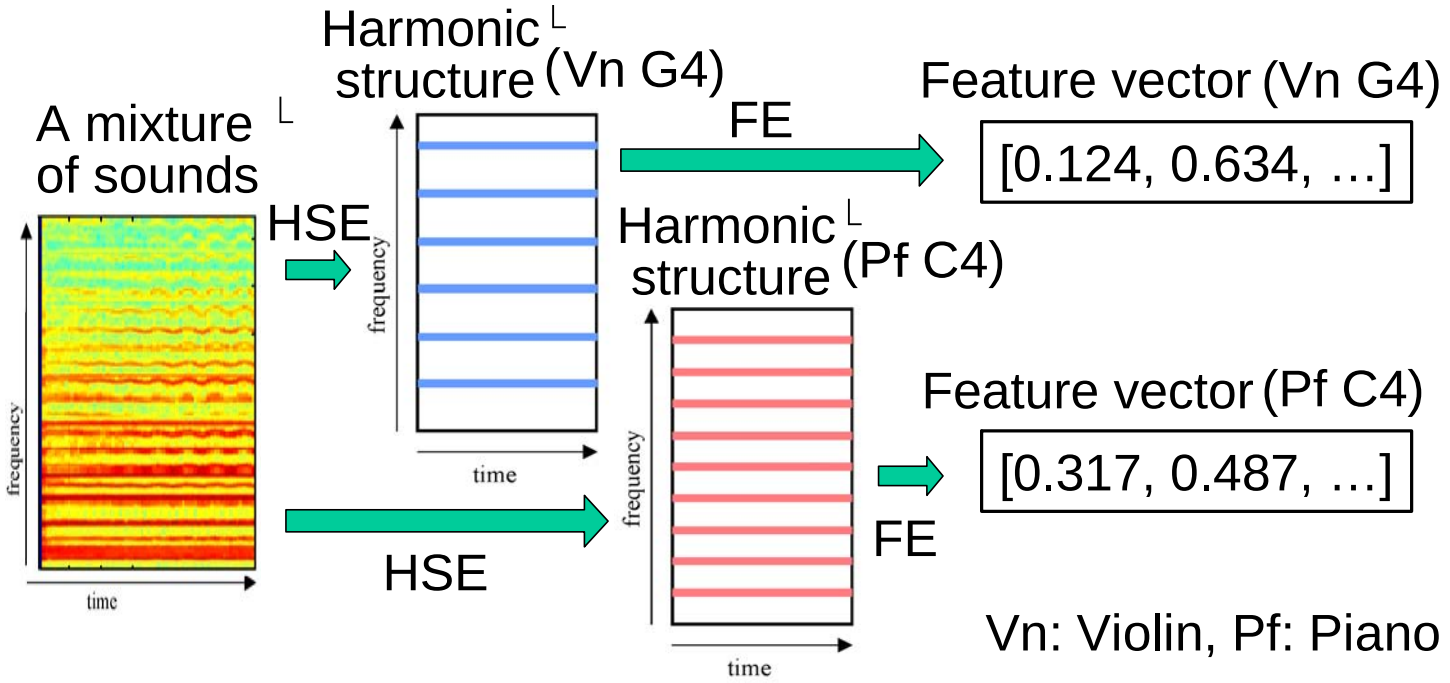
⇒ 影響を受けやすい特徴量は**分散が大**



音の重なりに頑健な特徴量への重み付け

解決策

- 学習データを混合音から作成（**混合音テンプレート**）
- 混合音から得た学習データに対する**クラス内分散・クラス間分散比**を音の重なりの影響度と定義
  ⇨これを最小化する部分空間を計算（**線形判別分析**）



混合音テンプレート

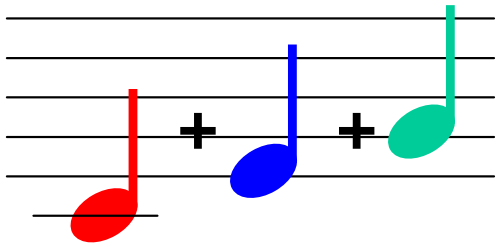
「混合音テンプレート」とは

学習用に用意された混合音から同定と同じ手順で調波構造を抽出し，特徴抽出したものの集合

問題点

音の混合の組み合わせが無限

⇨網羅的な混合音収集は不可能

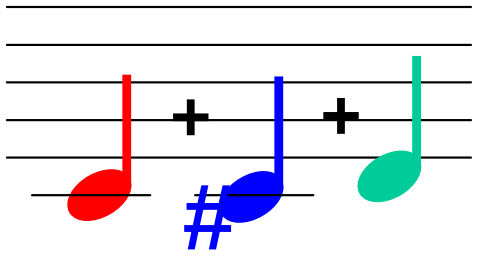


frequently used

着眼点

実際に使われる組み合わせは多くない

⇨よく使われるものを重点的に収集



rarely used

実楽曲の楽譜に基づいて混合音を作成

混合音テンプレートを用いた特徴量の重み付け

まとめると・・・

混合音から学習データを作成
（混合音テンプレート）

音の重なりによる特徴量の変化の度合いが，
学習データの分布の分散に現れてくる

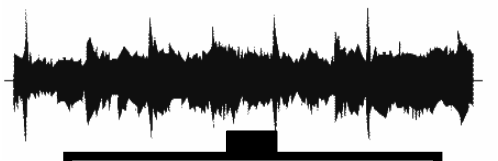
線形判別分析でクラス内分散・クラス間分散比最小化

音の混合パターンの組み合わせ爆発

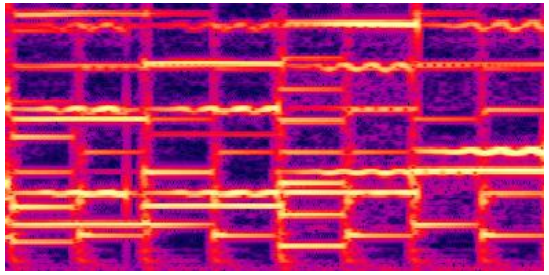
音楽で使われる混合パターンはごく一部

実際の楽曲の楽譜から混合音作成

実装システム

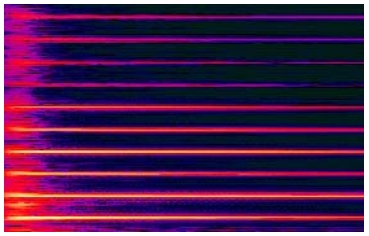


STFT



単音毎に

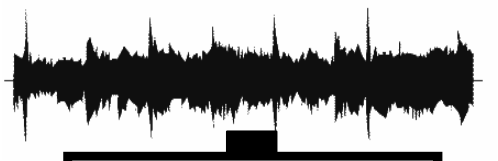
調波構造抽出



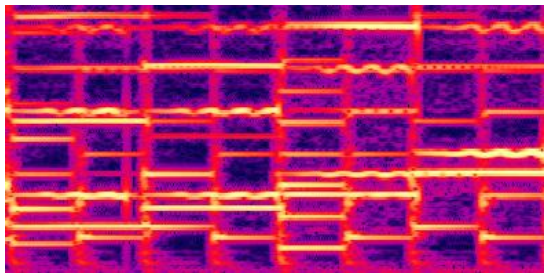
各単音の
発音時刻・F0等

└ 正解を付与

実装システム

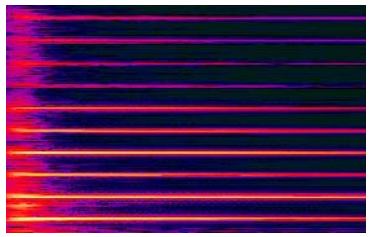


STFT



単音毎に

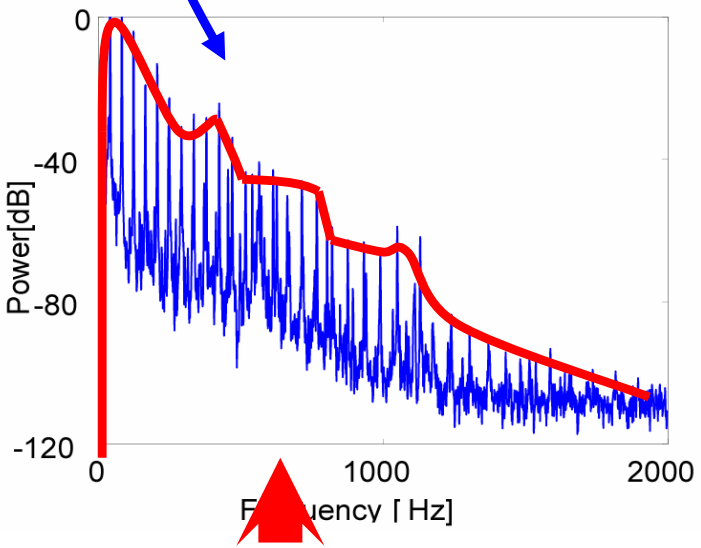
調波構造抽出

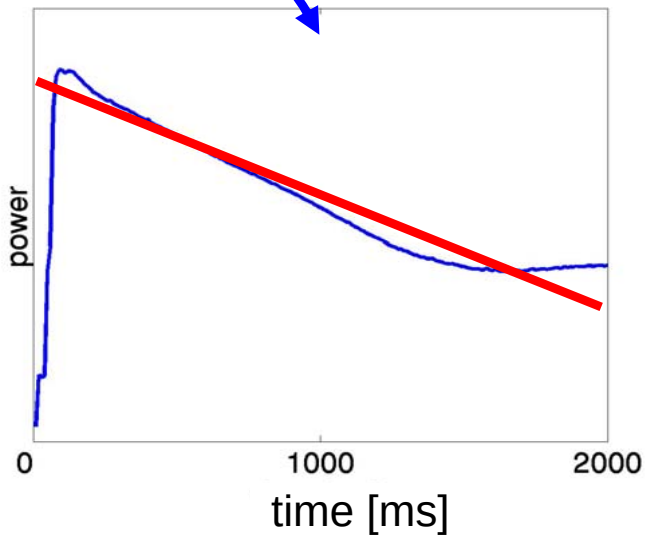


特徴抽出

次元圧縮

43個の特徴量を抽出
e.g. 周波数重心（時間軸方向の中央値）
パワー包絡の近似直線の傾き

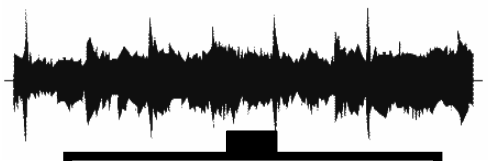




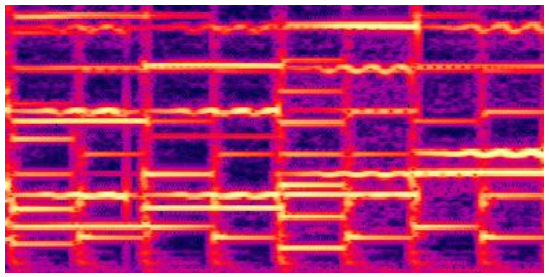
主成分分析（PCA）&線形判別分析（LDA）で
43→4次元に圧縮

変換行列を**混合音テンプレート**から得る
ことで，**音の重なりの影響度**を最小化

実装システム

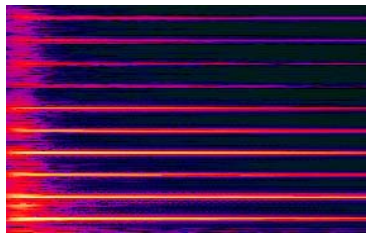


STFT



単音毎に

調波構造抽出



特徴抽出

次元圧縮

事後確率計算

$p(\omega_i \mid \mathbf{x}_k)$

事後確率再計算

$p(\omega_i \mid \mathbf{x}_k)$

楽器同定

Violin

楽器毎に $p(\omega_i \mid \boldsymbol{x}_k)$ を計算（特徴ベクトル$\boldsymbol{x}_k$が楽器ω_iのものである確率）

学習データを混合音から得る（**混合音テンプレート**）

データの分布が**F0依存多次元正規分布**に従うと仮定

└音高による音色変化を考慮して多次元正規分布を拡張

音楽的文脈（前後関係）を考慮

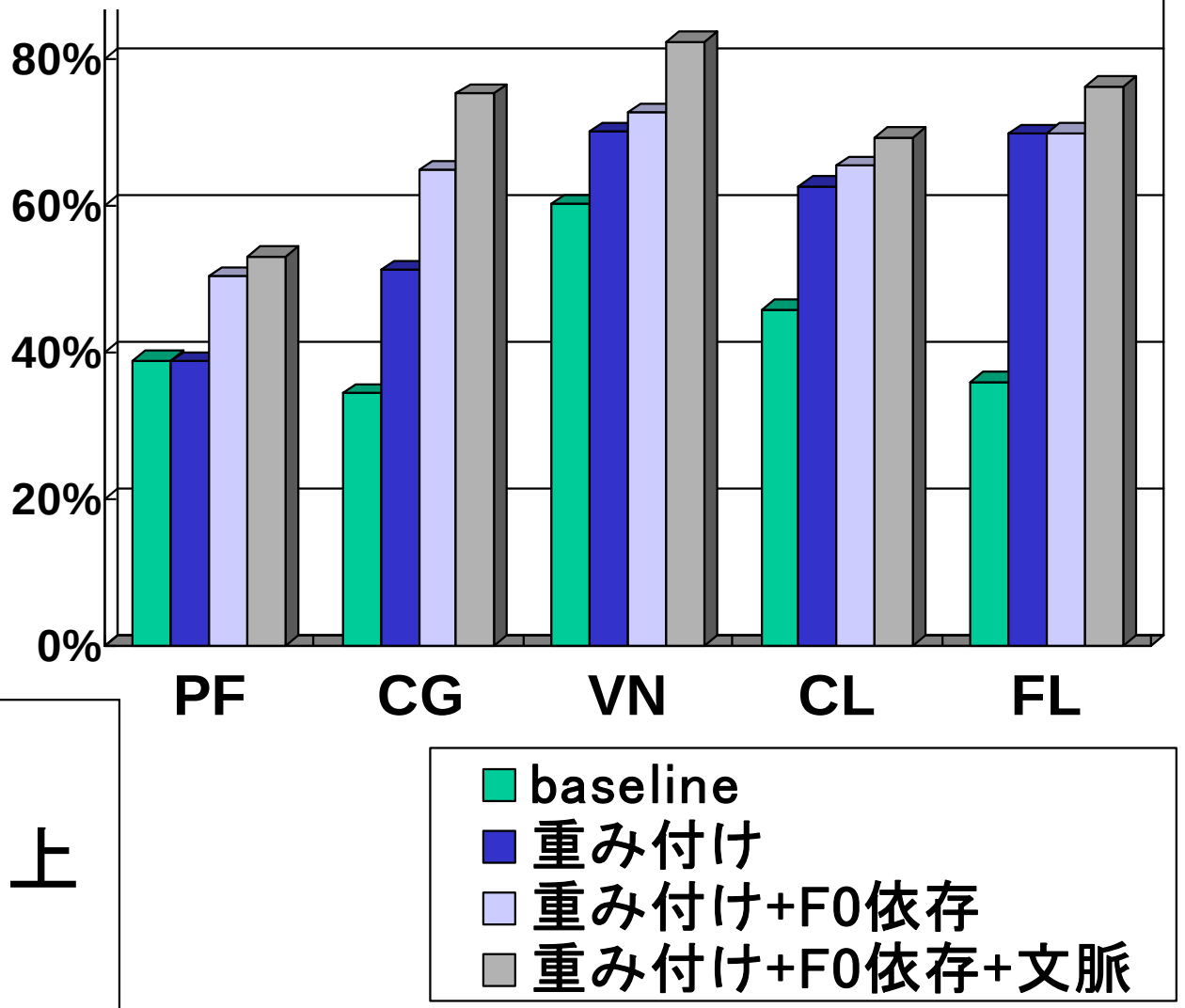
（詳細略）

$p(\omega_i \mid \boldsymbol{x}_k)$ が最大の楽器を出力

評価実験 (1/2)

実験条件

- 認識対象:四重奏
- 学習データ:
  単旋律＋二重奏
- 対象楽器:5種類



実験結果

- 重み付け:
  PF以外10%以上向上
- F0依存:
  PF, CG10%以上向上

全体的に認識率低め
調波構造だけではCGと区別困難

評価実験 (2/2)

実験目的

混合音テンプレートを用いるが
線形判別分析を用いない場合
(PCAだけで次元圧縮)との比較

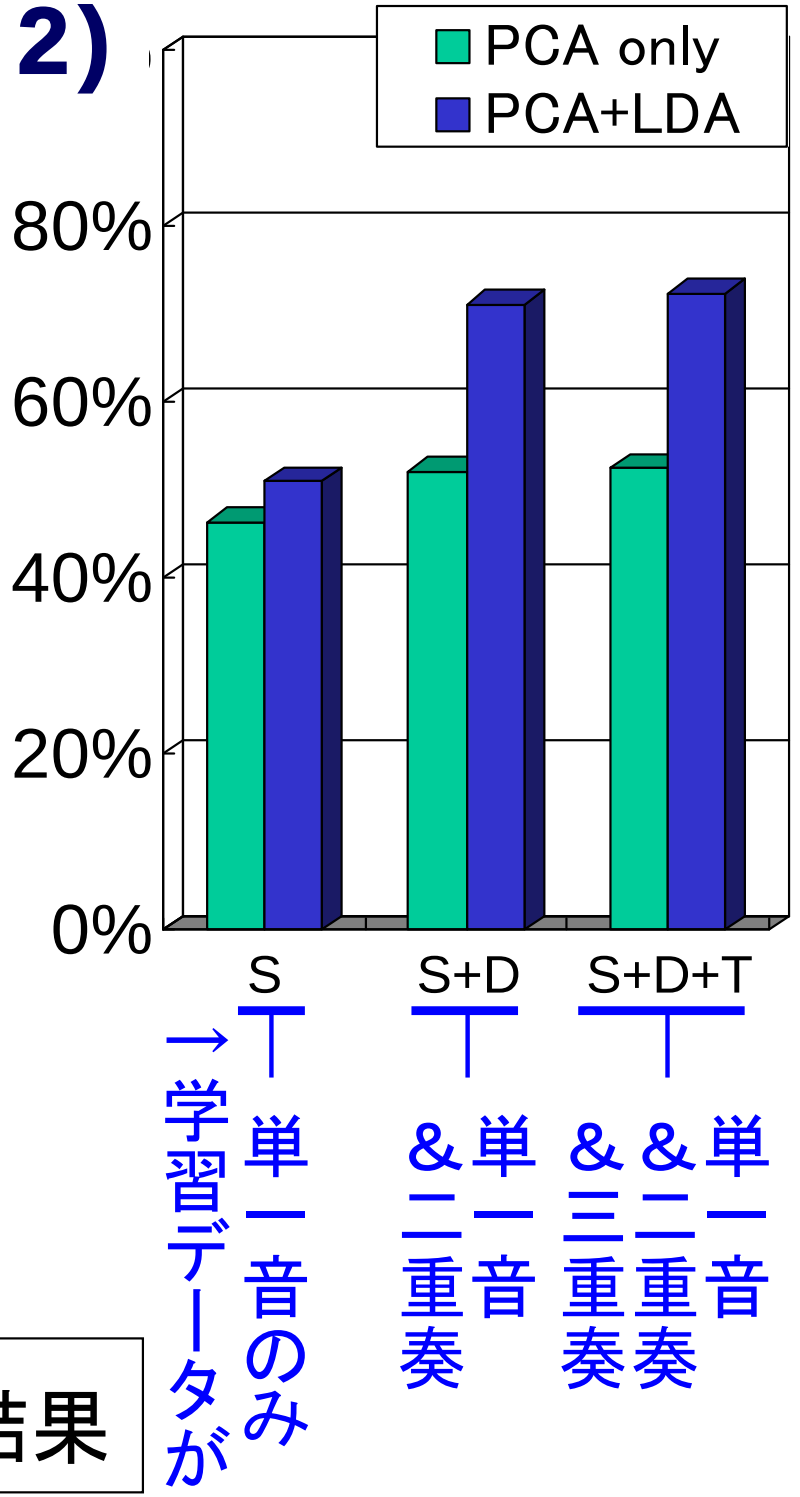
実験結果

線形判別分析使用時のほうが
混合音テンプレートの効果大

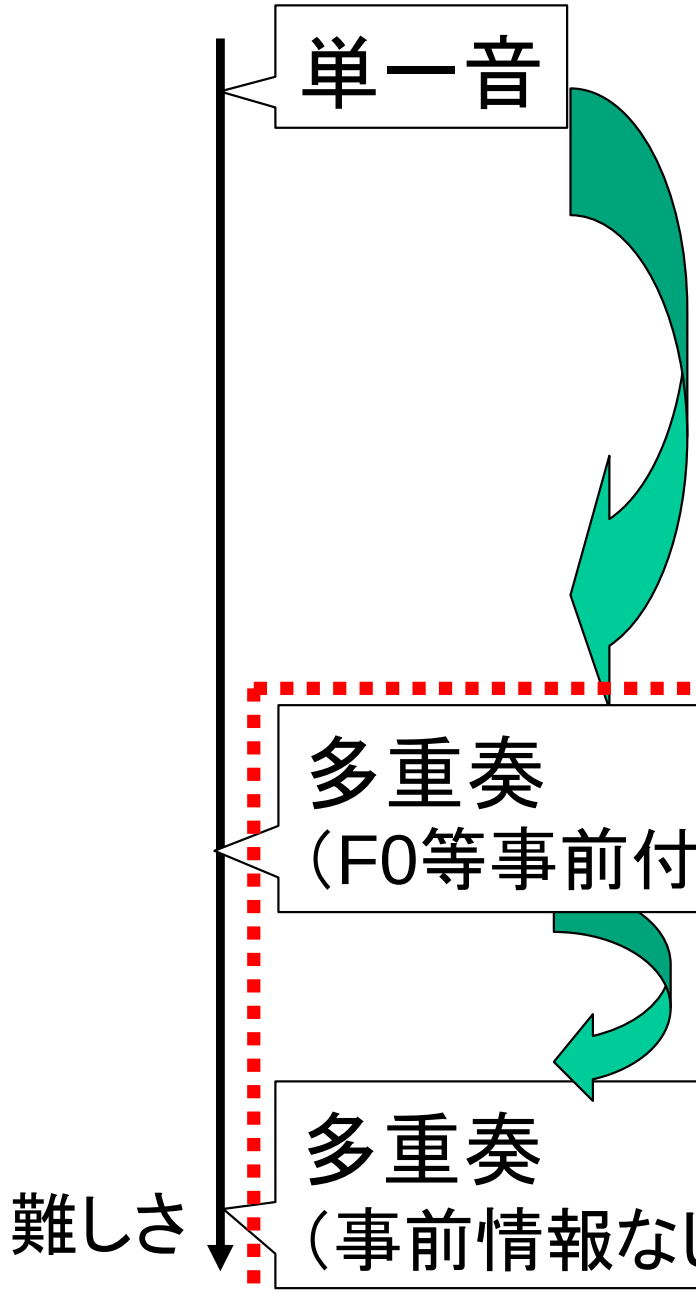
考察

パワーに関する特徴量の重みが
相対的に小さくなる傾向

二重奏・三重奏についても同様の結果



ここまでのまとめ



課題1　**音の重なりによる特徴量変化**

アイディア

音の重なりに頑健な**特徴量への重み付け**

- 学習データを混合音から作成
  （混合音テンプレート）
- 音の重なりの影響度を「クラス内分散・クラス間分散比」として定量化
- 少ないデータからのテンプレート作成

課題2　**発音時刻・F0推定エラー**の影響

課題2　発音時刻・F0推定エラーの影響

課題2 発音時刻・F0推定エラーの影響

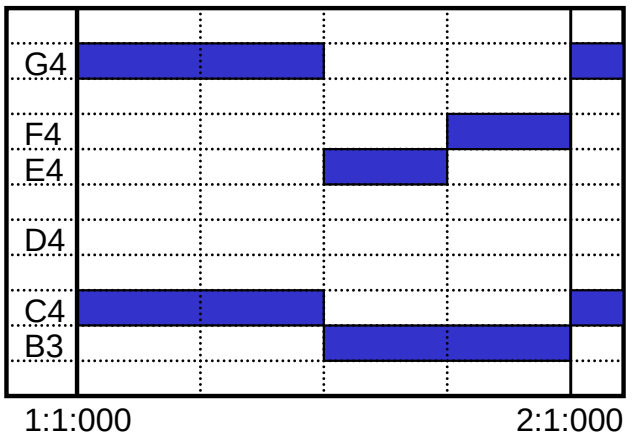
従来の一般的な楽器音認識では

単音(1音符の音)ごとに楽器を認識

⇨各単音の発音時刻・F0の推定が必要

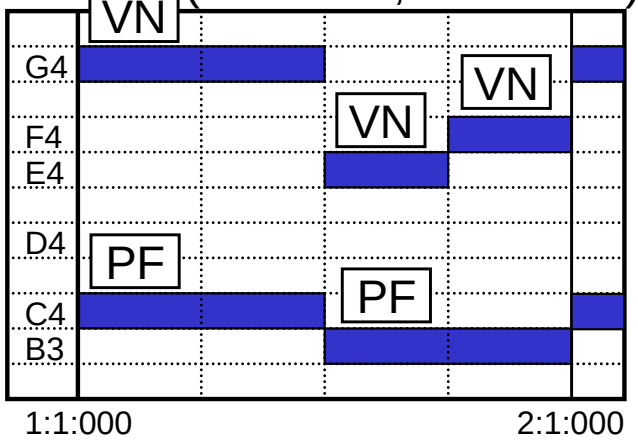
特に混合音に対しては
いまだ困難な課題

Audio signal



Estimate onset time, pitch and duration of each note

(VN:Violin, PF: Piano)



Identify instrument name of each note

問題解決のために (1/2)

問題提起

①「発音時刻・F0推定→楽器同定」という
　流れは本当に正しいのか

②単音ごとに楽器を認識する必要はあるのか

我々の考え

①楽譜を頭に思い浮かべ（発音時刻・F0推定に
　相当）なくても，どんな楽器かはわかる

②自動採譜以外の応用では
　必ずしも単音単位の必要はない

問題解決のために (2/2)

方針

①F0推定などを前処理として行うことなく,

②単音ごとではなく楽曲全体をみていって,

(1) いつ ―――― 従来の発音時刻検出に相当

(2) どの高さで ―― 従来のF0推定に相当

(3) どの楽器が ― 従来の楽器同定に相当

音を鳴らすのか，大まかにとらえる枠組み

言い換えると

上3つの処理を**同時並行的**に行う**統一的枠組み**

Instrogram

楽器音認識の新たな考え方

楽器音認識を

「**楽器存在確率** $p(\omega_i; t, f)$ を全時刻・全F0にわたって網羅的に計算する問題」

ととらえる

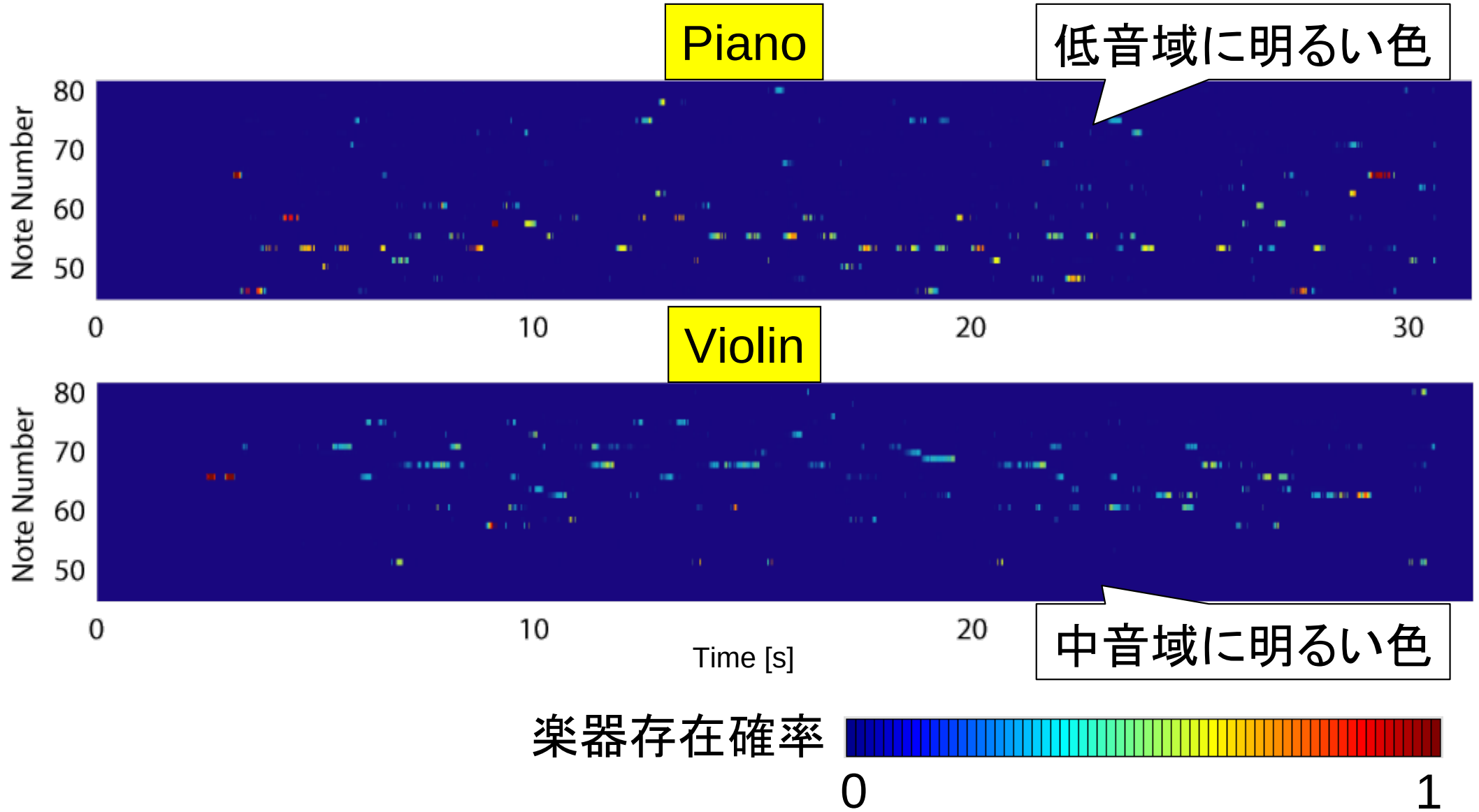
時刻 t において周波数 f をF0とする楽器 ω_i の音が存在する確率

Instrogramとは

楽器存在確率をスペクトログラムのように時間・周波数平面に可視化したもの

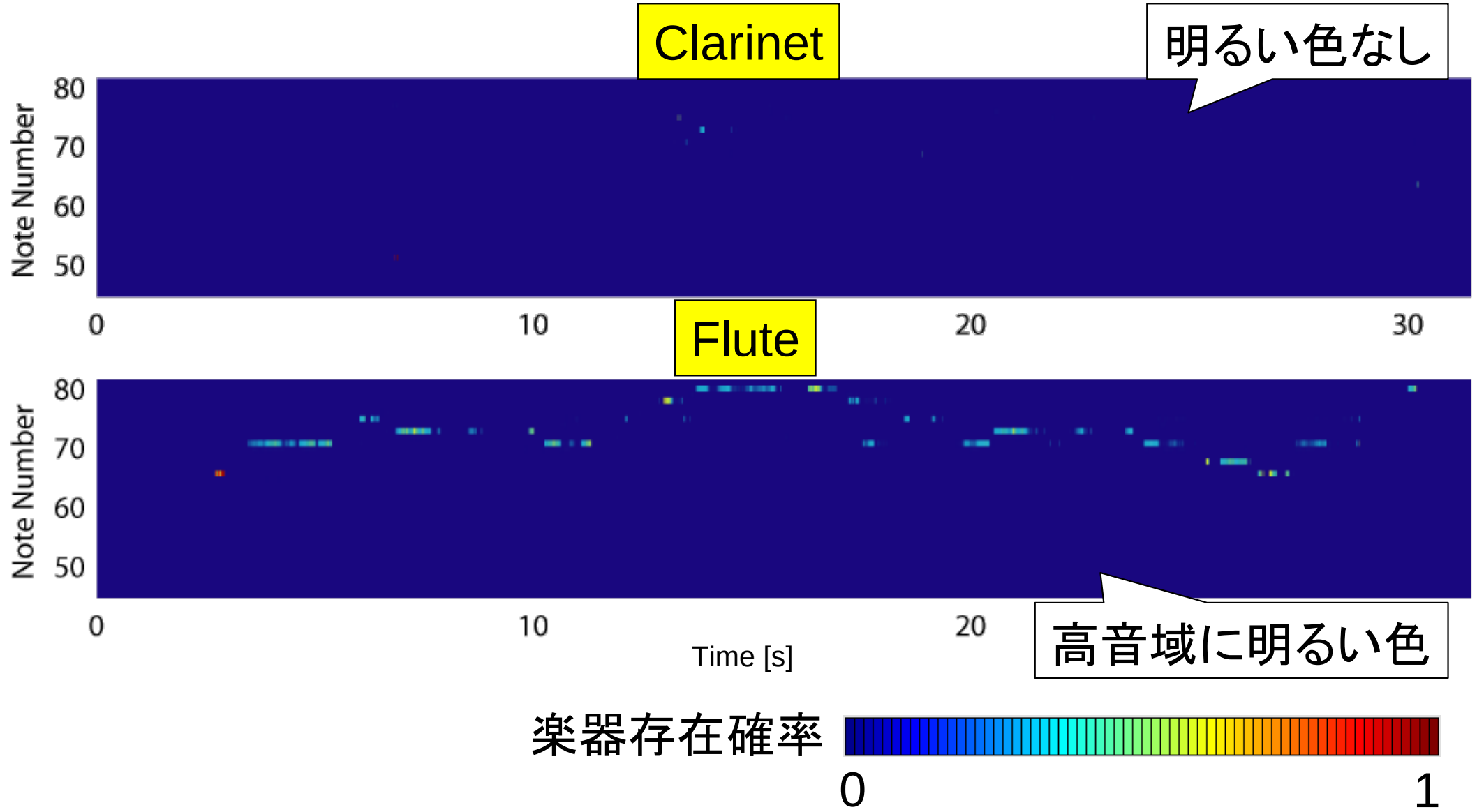
Instrogram (1/2)

例：蛍の光（Flute-Violin-Piano）



Instrogram (2/2)

例：蛍の光（Flute-Violin-Piano）



楽器存在確率の定式化

楽器存在確率を2種類の確率の積に分解する

$$p(\omega_i; t, f) = p(\mathrm{X}; t, f)\ \ p(\omega_i|\mathrm{X}; t, f)$$

不特定楽器存在確率

何らかの楽器音が
存在する確率

従来の**発音時刻推定・F0推定**に相当

条件付楽器存在確率

何らかの楽器音が
存在するとすると，
その楽器がω_iである確率

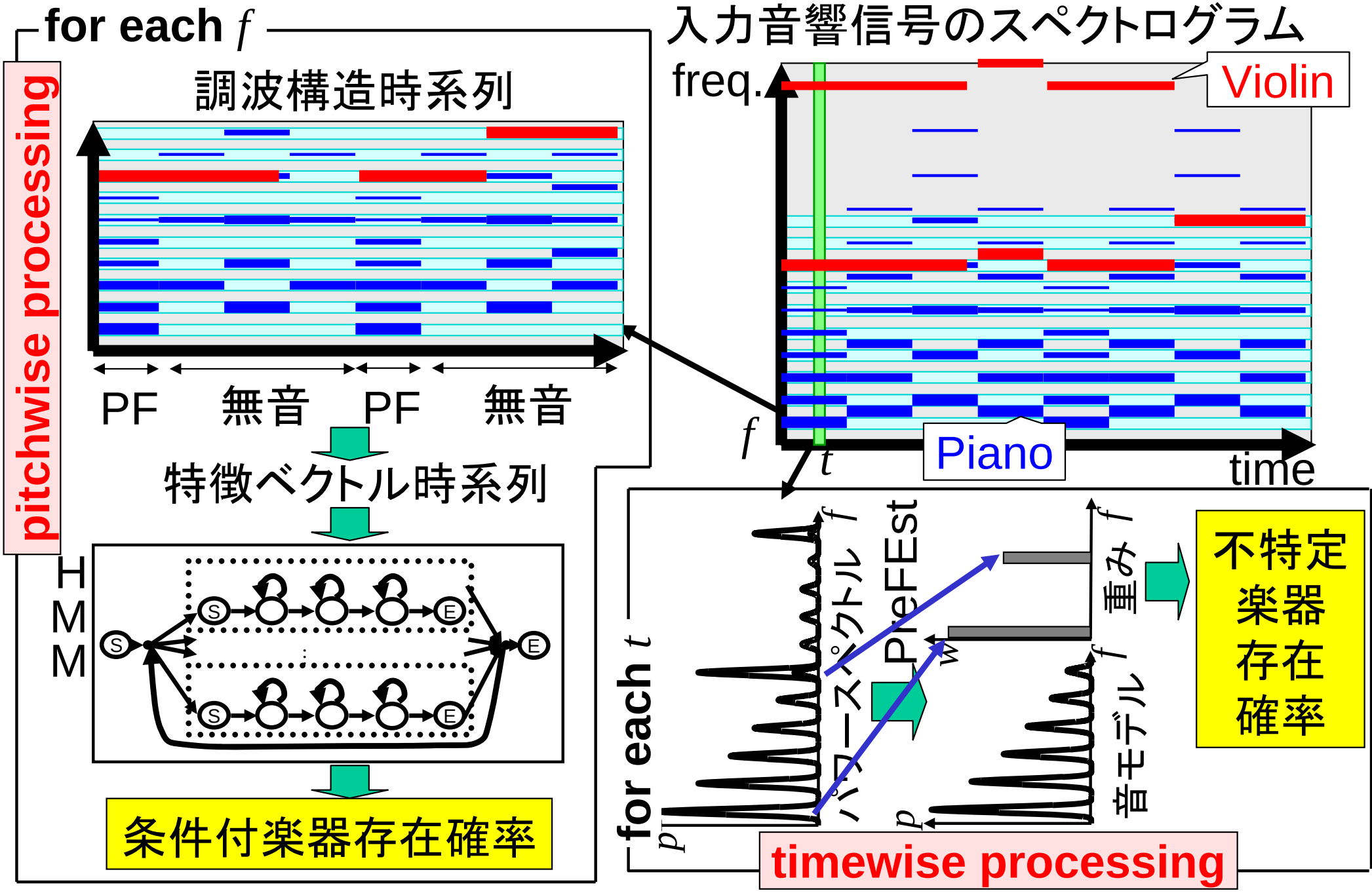
従来の**楽器同定**に相当

補足

同時刻・同F0に複数音はないと仮定

$\mathrm{X} = \omega_1 \cup \cdots \cup \omega_m$ とおき，$\omega_i \cap \mathrm{X} = \omega_i$ を利用

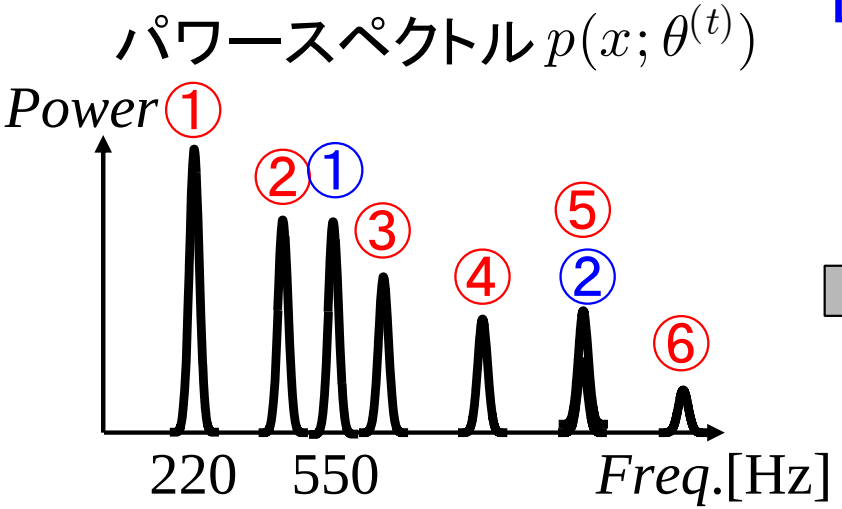
確率計算アルゴリズムの概要

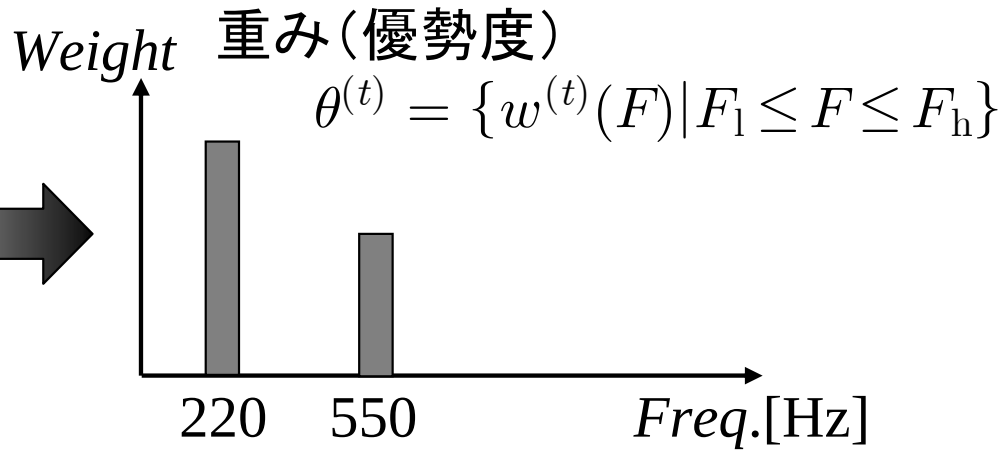


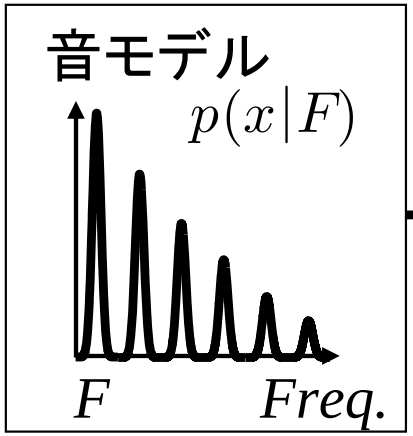
不特定楽器存在確率の計算

フレーム毎にPreFEst [Goto'99]を適用

調波構造をパラメトリックにモデル化し，
混合音中の各調波構造の優勢度を
EMアルゴリズムで推定







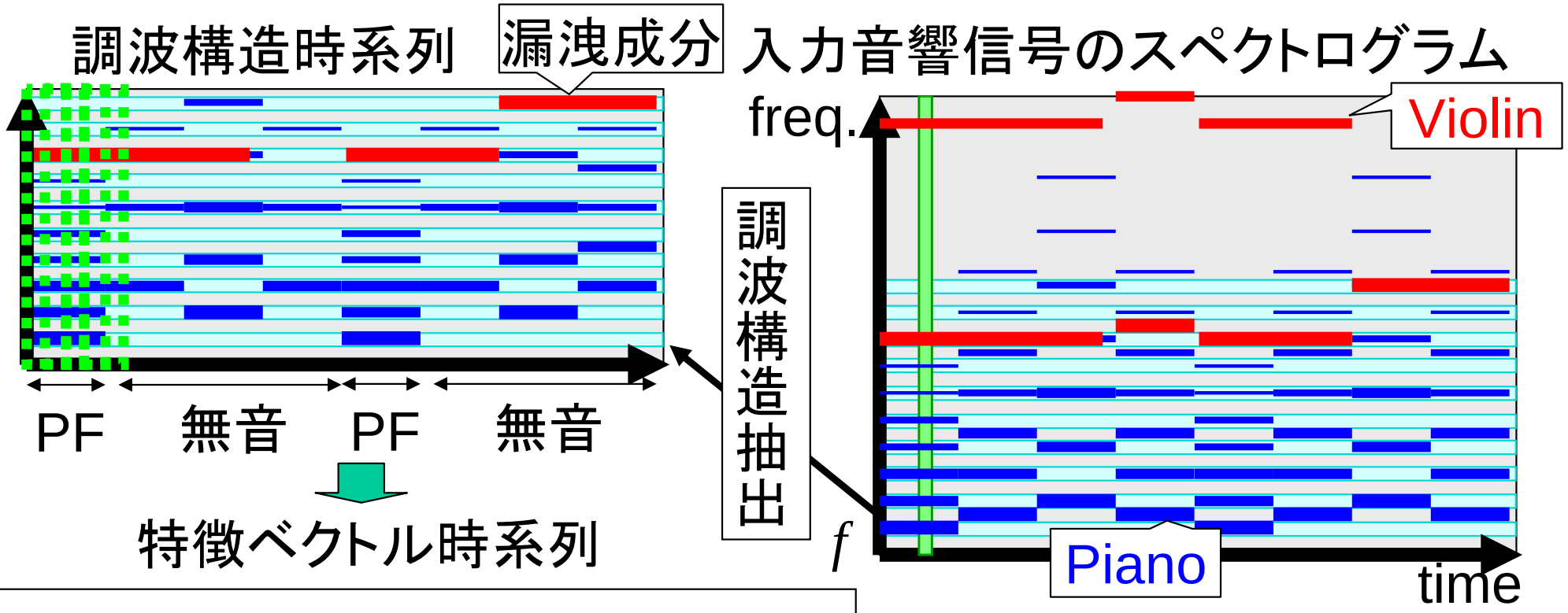
$$\underline{p(x;\theta^{(t)})} = \int_{F_l}^{F_h} \underline{w^{(t)}(F)}\,\underline{p(x|F)}\,dF$$

観測スペクトル　重み　音モデル

不特定楽器存在確率

x：対数周波数

条件付楽器存在確率の計算

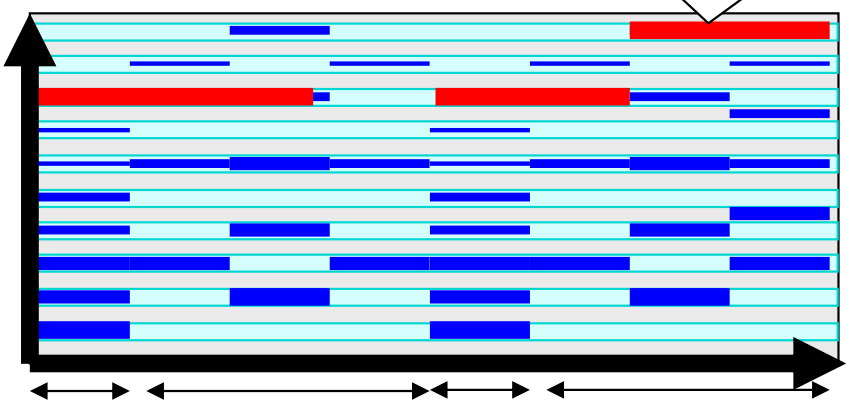


1. 調波構造から短い断片を切り出し
2. 断片から28個の特徴量を抽出
3. 主成分分析で12次元に圧縮
4. 以上の処理を10ms分ずらしながら繰り返す

- 周波数重心
- パワー包絡の近似直線の傾き
- AM, FMの振幅と振動数etc.

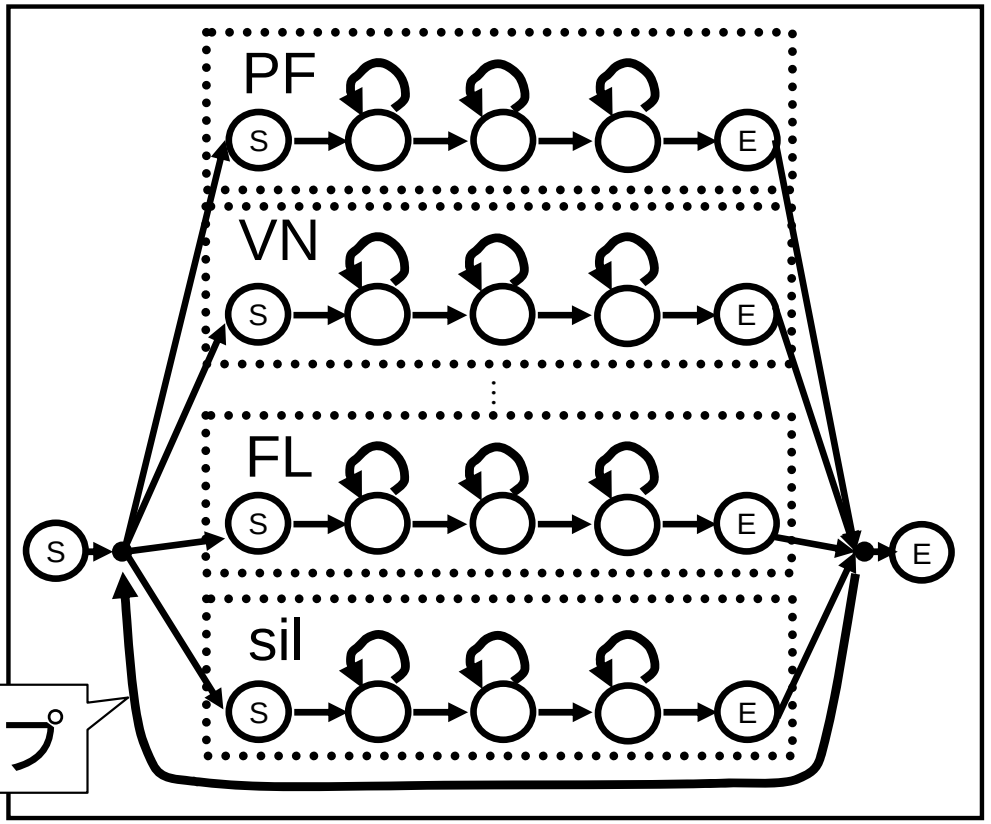
条件付楽器存在確率の計算

調波構造時系列



特徴ベクトル時系列

隠れマルコフモデル（HMM）



ループ

各HMMが各楽器or無音に対応

毎に各状態に停留する尤度を計算

条件付楽器存在確率

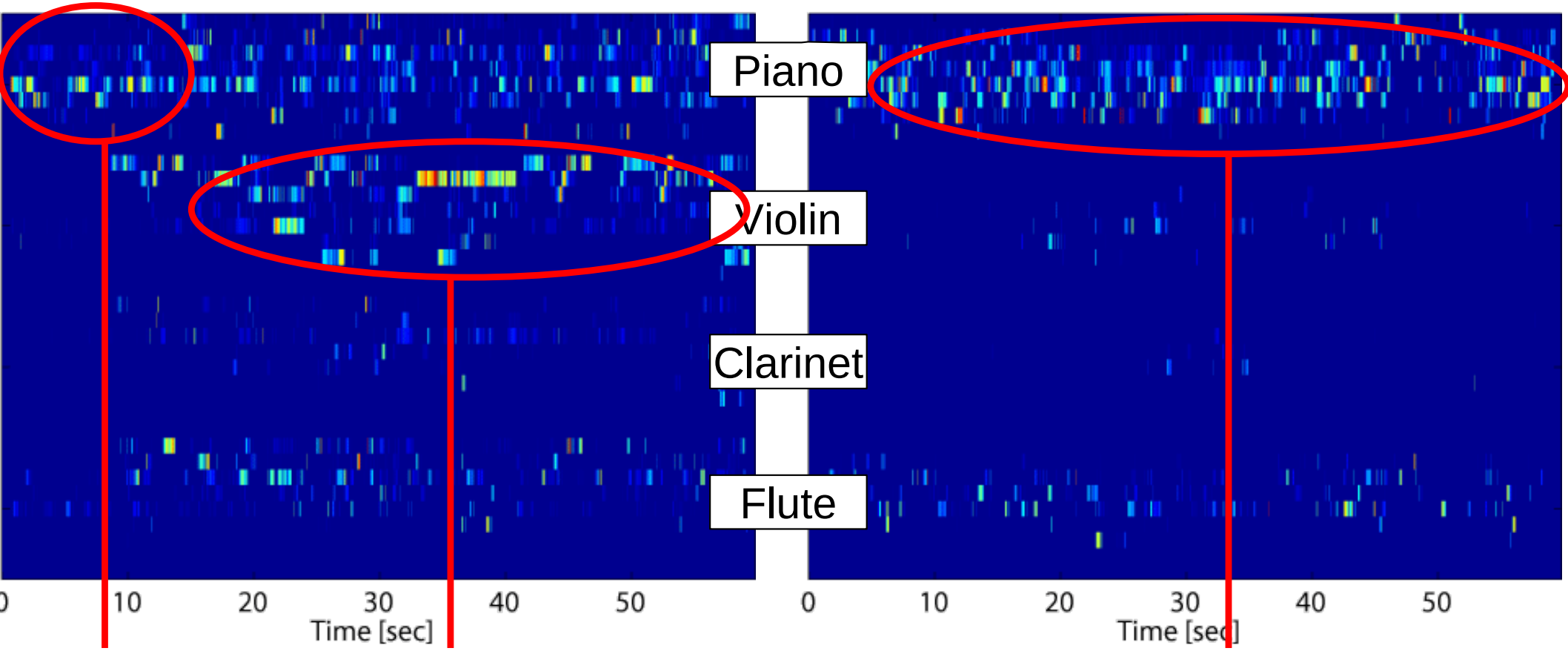
HMM学習時

- 単音毎のラベル付きデータを用いて単音毎にHMM学習
- 音の重なりへの頑健化のため学習データを混合音から作成（混合音テンプレート）

Instrogram作成実験（抜粋）

RM-C No.40 (Pf.+Str.)

RM-J No.1 (Piano)



最初は
ピアノだけ

途中から弦楽器も

終始ピアノだけ

実演奏に対してInstrogramを作成
シンボル化（詳細略）の結果，適合率約70%

ここまでのまとめ

課題1　**音の重なりによる特徴量変化**

↳ 音の重なりに頑健な**特徴量への重み付け**

課題2
発音時刻・F0推定エラーの影響

アイディア

発音時刻・F0推定を**同時並行的**に行う

- 楽器認識を**Instrogram**（楽器存在確率の時間・周波数表現）の計算ととらえる
- **楽器存在確率**を**2つの確率の積**と定義
- 各々の確率を**PreFEst**と**HMM**で計算

類似楽曲検索への応用

類似楽曲検索とは？

ユーザが指定した楽曲に似た曲を提示

- 別名 : Query-by-Example
- Content-based MIRの一形態
- 誰でも簡単に利用可能

課題

何をもって似ていると判断するか

従来はlow-levelの特徴量を利用 [Aucouturier ’02] [Pampalk’04] etc.

Low-level 特徴量の限界

類似尺度のユーザ依存性に対応できない

類似楽曲検索における課題

類似尺度のユーザ依存性

ユーザがどの音楽要素を重視して類似を判断するか分からない
e.g. メロディ，リズム，ベースライン，楽器構成

⇒ 音楽要素ごとに類似尺度を定義し，user-adaptiveに重みを付けて統合

本研究では第1段階として

楽器構成のみに着目した音楽類似度

∵ 楽器構成は聴いたときの印象に影響大
（e.g. 同じ楽曲を異なる楽器で演奏した時）

楽器構成に着目した音楽類似度

要求条件

- 楽器構成の移り変わりを考慮したい
- 長さの異なる楽曲同士も比較できるようにしたい

キーアイディア

楽器存在確率を特徴量としてDTW

Dynamic Time Warping

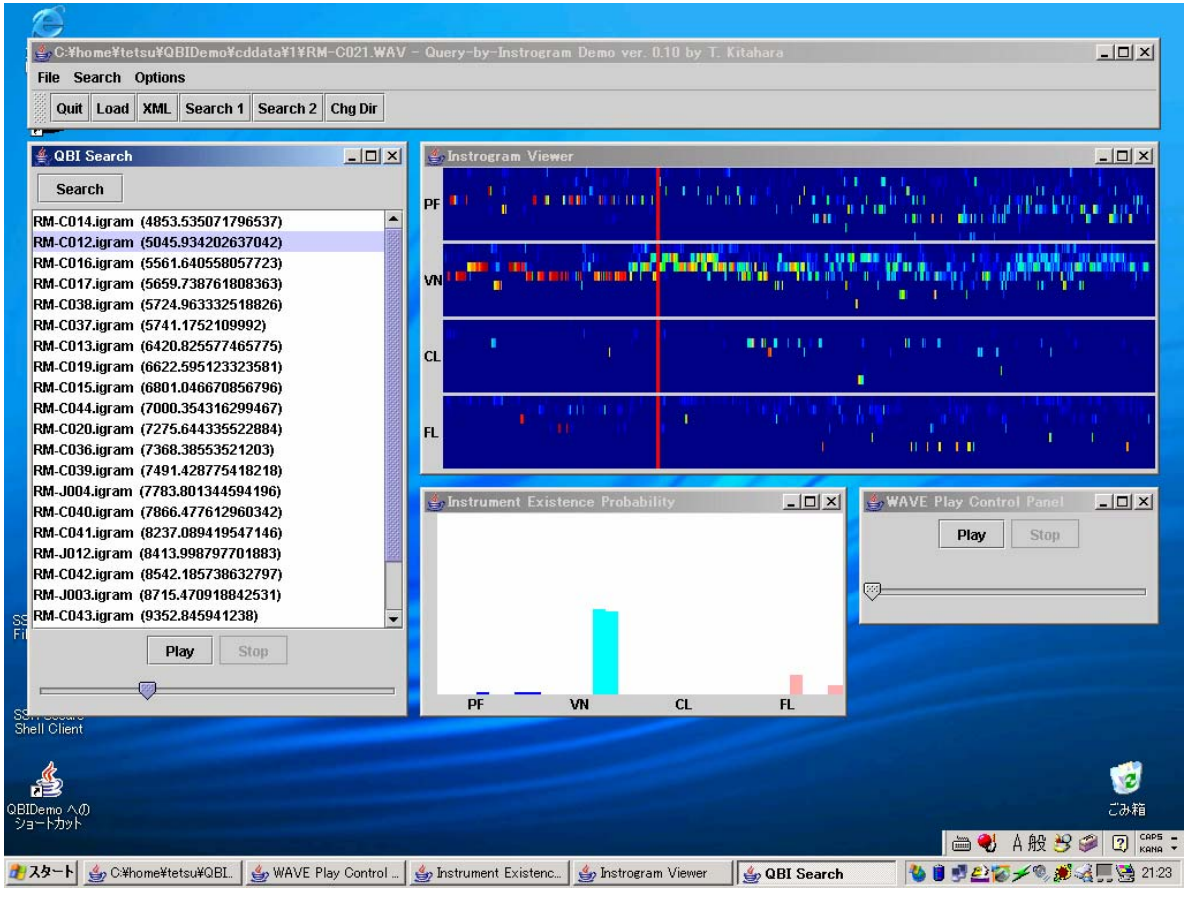
結果

- 楽器構成が同じ楽曲間の非類似度： <7000
- 弦楽曲-ピアノ曲間の非類似度： >9000
- 弦楽曲の類似楽曲ベスト3はすべて弦楽曲
  非弦楽曲の類似楽曲ベスト3はすべて非弦楽曲

MFCCを用いた場合には両者の混同しばしば

Demo

- ユーザが指定した楽曲に楽器構成が似ている楽曲を検索
- 再生の際に楽器存在確率を可視化



議論

到達点

単一音 ——— 19楽器に対して約80%

課題1　**音の重なりによる特徴量変化**

→ **混合音テンプレート**に基づく
音の重なりに頑健な**特徴量への重み付け**

多重奏
（F0等事前付与） ——— 5楽器による二～四重奏に対して80%前後

課題2　**発音時刻・F0推定エラー**の影響

→ **Instrogram**に基づく
発音時刻・F0推定との**確率的統合**

多重奏
（事前情報なし） ——— 実演奏にも適用可能に

難しさ ↓

類似楽曲検索への応用

本研究の主な貢献 (1/3)

CASAの観点から

2種類のグルーピング問題に対する新たな解法

不特定楽器存在確率 ⇒ Simultaneous grouping

条件付楽器存在確率 ⇒ Sequential grouping

両者を確率計算として定式化し，確率的に統合

「分離→認識」のcascade frameworkからの脱却

完全な音源分離を前提としない混合音認識手法

① 音の重なりに頑健な特徴量の重み付け

② Instrogramに基づく複数楽器音のパラレルな認識

本研究の主な貢献 (2/3)

Content-based MIRの観点から

Instrogram: Middle-level descriptor

- 高次の音楽要素(楽器構成)を表現し，自動抽出可能
- 音楽要素ごとの特徴を考慮した
  「きめ細かな音楽類似度」実現の第1歩

詳細(高分解能)で質にムラのないアノテーション

- 楽器存在確率を任意の分解能でアノテーション可能
- マニュアルアノテーションでは困難
  cf. 大人数のプロミュージシャンによる人海戦術

本研究の主な貢献 (3/3)

音楽可視化の観点から

楽器構成のグラフィカルな視覚表現の確立

- 一覧性のない音楽に一覧性を付加する重要なテーマ
  - ⇨ 音楽の中身を一目で確認⇒音楽の選択を効率化

従来研究

- MIDI信号（演奏情報）の可視化 [Hiraga ’02]
- 倍音抑制による pianoroll 的可視化 [Sagayama’04]
- 楽曲構造の解析と可視化 [Goto ’06]

本研究

- 楽器構成の可視化：
  - 楽器構成の移り変わりを見ながら聴く⇒教育効果
  - 「フルートが吹き始めるところから聴く」のも容易

残された課題・今後の展開 (1/2)

楽器音認識の観点から

音楽知識・トップダウン情報の導入

- 本研究の成果は主にボトムアップ処理のモデル化
- トップダウン情報を事前確率として導入

歌声・ドラム音への対応

- 当研究室の他の学生による成果との統合

計算量の削減

- Instrogram表現の圧縮
- 並列処理化による全体のスループット向上

残された課題・今後の展開 (2/2)

Content-based MIRの観点から

他の音楽要素(e.g. 和音進行・リズム)との統合

要素ごとの特徴量と類似度計算により,
user-adaptive な類似度計算を実現

音楽推薦やPersonalRadioのような応用への展開

目標は「自分専用音楽DJ」

音楽理解のメカニズムの解明を目指して

楽譜表現に基づかない音楽理解技術の確立

「音楽理解≠採譜」に基づき各種音楽要素を認識

論文の構成

第1章

Introduction

第2章

Literature Review

第3章

F0-dependent Timbre Modeling

関連発表

- 情報処理学会論文誌, vol.44, no.10, 2003.
- Applied Intelligence, vol.23, no.6, 2005.
- IEEE-ICASSP 2003, IEA/AIE-2003

第4章

Category-level Recognition of Non-registered Instruments

関連発表

- 情報処理学会論文誌, vol.45, no.3, 2004.
- IEEE-ICASSP 2004

ここまでは単一音

論文の構成

第5章

Feature Weighting based on Mixed-sound Template

関連発表

- 電子情報通信学会論文誌, vol.J89-D, no.12, 2006.
- EURASIP JASP, vol.2007, no.51979, 2007.
- ISMIR 2005.

本発表における課題1

第6章

Note-estimation-free Instrument Recognition

関連発表

- IPSJ Journal（英文）, vol.48, no.1, 2007.
- IEEE-ICASSP 2006, IEEE-ISM 2006.

本発表における課題2

第7章

Application

関連発表は「第6章」に含む

MIRへの応用

第8章

Discussion

第9章

Conclusion

まとめ

研究の目的

多重奏に対する楽器認識

・F0依存
・未知楽器

混合音認識の基礎技術

単一音 ——— 19楽器に対して約80%

課題1　**音の重なりによる特徴量変化**

混合音テンプレートに基づく**特徴量の重み付け**

多重奏（F0等事前付与） ——— 5楽器による二～四重奏に対して80%前後

課題2　**発音時刻・F0推定エラー**の影響

Instrogramに基づく**確率的枠組み**

多重奏（事前情報なし） ——— 実演奏にも適用可能に

難しさ

応用

楽器構成に基づく類似楽曲検索