音声情報処理特論 2003年12月21日

音楽情報処理 ・ 擬音語認識

奥 乃 博

京都大学 大学院情報学研究科
知能情報学専攻
知能メディア講座 音声メディア分野
http://winnie.kuis.kyoto-u.ac.jp/~okuno/
okuno@i.kyoto-u.ac.jp, okuno@nue.org

目次

1. 音楽情報処理とMPEG-7
2. 自動採譜
3. 音高による音色変化に着目した楽器音の音源同定
4. 未知の楽器を考慮した楽器音の音源同定
5. 擬音語認識
6. レポート課題

音楽情報処理とAI

- 作曲
- 演奏
- 音楽理論
  - 音楽心理学
  - 音楽知覚
- ディジタル音響処理
- 信号・記号・変換

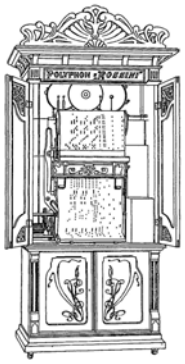


Figure 1. Automatic string instrument with percussion, driven by a perforated paper card (1904).

音楽信号認識

1. ディジタル信号
   - アタックの検出
   - ピッチの決定(あれば)
   - ラウドネスの決定
   - 音源分離(音源分凝)
   - 音源同定
   - ニュアンス(ビブラート・トレモロ・スタッカート)検出
   - 休止検出
2. 音響マップ
   - 調律(チューニング)同定
   - 調同定
   - テンポの揺らぎを追跡
   - 拍子決定
   - 音符と休符の割り当て
   - 音声の分離
   - スラー・タイの決定
3. 音楽マップ

MPEG-7

- Moving Picture Experts Group (MPEG):'88
- Multimedia Content Description Interface:'97
- MPEG-7の構成
  - Audio part, Video part → 音響, 映像だけ規定
  - Multimedia Description Schemes (MDS) part → audio・visual descriptorsを含む記述法を規定
  - Description Definition Language (DDL) → description schemeを表現する言語の標準化
  - System part →実環境での使用に対する糊
  - Reference Software → オープンソースコード

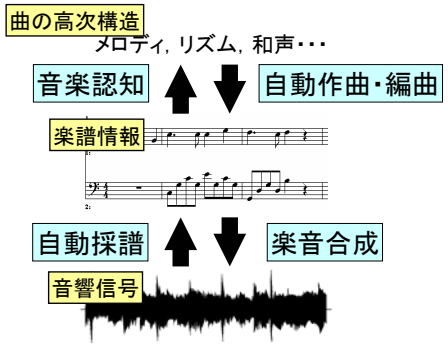
MPEG-7

- MPEG-1, MPEG-2, MPEG-4: audiovisual data に対するツールを提供
- MPEG-7: contents についてのnavigationするための手段を提供
- メタデータの一種: Dublin Core, GDA (Global Data Annotation), UNL (Universal Network Language)
- 設計方針:
  - Wide application base, Relation with content, Wide array of data type, Media independence, Object-base, Format independence, Abstract level, Extensibility

音楽情報処理とは（その1）

音楽情報処理＝音楽を扱う情報処理技術全般

主な研究テーマ①



音楽情報処理とは（その2）

主な研究テーマ②＝**音楽情報検索**

- ディジタル音楽配信の普及
  ⇒入手可能な音楽データの急増
- 音は一覧性のないメディア
  ⇒目的の曲を探すには一曲一曲試聴が必要
- 代表的な3つのアプローチ
  ①鼻歌検索 ②印象語検索 ③類似楽曲検索
- 検索に適した高次アーカイブの構築
  ⇒音楽へのMPEG-7タグ付けが重要な課題
- http://www.ismir.net/

音楽情報処理とは（その3）

主な研究テーマ③＝**演奏の自動表情付け**

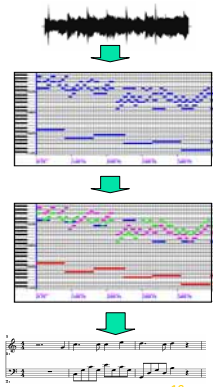
- 入力：楽譜or表情のないMIDIファイル
  出力：表情のついたMIDIファイル
- 2002年より，コンクール(Rencon)開始
- http://shouchan.ei.tuat.ac.jp/~rencon/

その他の研究テーマ

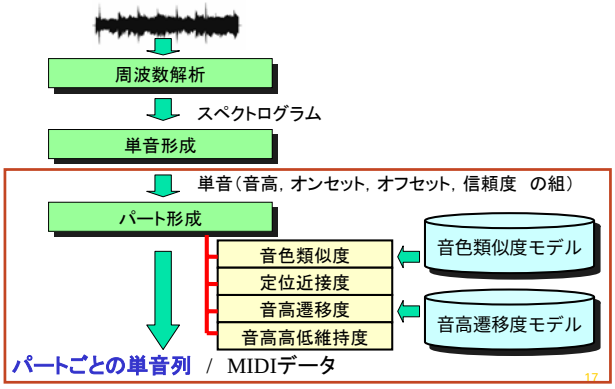
- ジャムセッションシステム
- 音楽表現のための新インターフェース
- 感性情報処理，メディアアート，etc.

自動採譜とは

- 自動採譜＝音響信号から楽譜へ
- 自動採譜の中心となる処理
  1. 音楽音響信号から個々の単音（音高，オンセット，オフセット）を推定 ⇒ 単音形成
  2. 単音（音符）をパートごとに分類 ⇒ パート形成
  3. 音価推定，テンポ推定，調認識，・・・

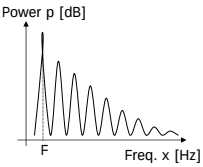


標準的なシステムの概念図



単音形成（混合音からの音高推定）

- 基本周波数Fの調波構造p(x|F)をモデル化（音モデル）
- 混合音のスペクトル p(x) を，あらゆる基本周波数Fの音モデルの加重混合とみなす

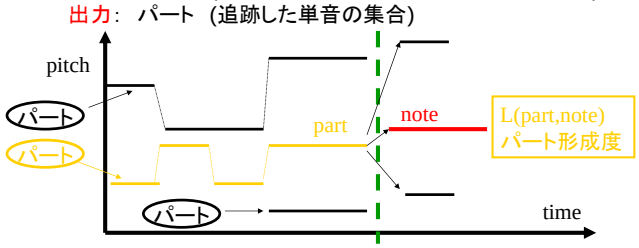


$$p(x) = \int_{Fl}^{Fh} w(F)\, p(x \mid F)\, dF$$

- 重みw(F)をEMアルゴリズムで推定
- w(F)の値が大きい *iff* Fに基本周波数成分がある
- 特長①音源数仮定せず②Missing fundamental対応

パート形成

入力: 単音列 (音高, オンセット, オフセット, 信頼度の組)
出力: パート (追跡した単音の集合)



- パート形成度 L(part,note) パートpartと単音候補noteとの結びつきの強さ
- パート形成度が最大となる単音候補とパートを形成

パート形成

- パート形成L(part, note)をどのように算出するか
  ⇒以下の4つの手がかりを利用

1. 一つの旋律は類似した音色の系列を持つ
   ⇒ 音色類似度
2. 一つの旋律は近接した定位の系列を持つ
   ⇒ 定位近接度
3. 一つの旋律に出現する音高の遷移には傾向がある
   ⇒ 音高遷移度
4. 旋律同士の音高の高低関係は維持する傾向にある
   ⇒ 音高高低維持度

音色類似度 $L_t(part, note)$

- 音色類似度 L(part, note) を2単音間の音色類似度の平均で求める

$$L_t(part, note) = \frac{1}{c} \sum_{note_j \in part}^{c} L_t(note_j, note)$$

- 二単音が同じ/異なる楽器である群 $\prod_0 / \prod_1$
- 二単音$note_j$, $note_k$の音色特徴ベクトルの差 x_{jk}

$$L_t(note_j, note_k) = p(\Pi_0 \mid x_{jk}) = \frac{p(x_{jk} \mid \Pi_0)}{p(x_{jk} \mid \Pi_0) + p(x_{jk} \mid \Pi_1)}$$

$$p(x \mid \Pi_i) = \frac{1}{(2\pi)^d |\Sigma|} \exp\left(-\frac{D_i^2(x;\mu_i)}{2}\right)$$

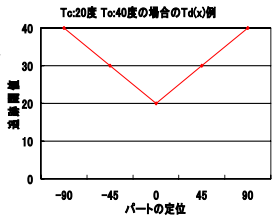
d:正規分布の次元数, Σ:共分散行列
$D_i^2(x;\mu_i)$: xと群Π_iの平均μ_iとのマハラノビス距離
各群の事前確率は等しいと仮定

定位近接度 $L_d(part, note)$

- T_c:0度(中央)おける定位閾値
- T_o:左右90度(真横)における定位閾値
- D:定位(パートの定位はIPDから求める)

$$L_d(part, note) = 1 - \frac{|D(part) - D(note)|}{T_d(D(part))}$$

$$T_d(x) = T_c + (T_o - T_c) \cdot \frac{|x|}{90}$$



音高推移度 $L_p(part, note)$

- 音高のトライグラムモデルを利用
- 学習データ
  - RWC研究用音楽データベース:クラシック(50曲)
  - 付属MIDIデータの単旋律のトラックを利用
  - 総音符数(167179個)
  - 長調と短調それぞれ統計をとる
  - 調性で正規化(調は楽譜から得る)

音高高低維持度 $L_r(part, note)$

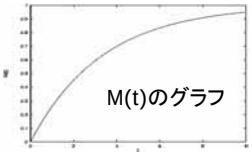
- パートpartが単音noteとパート形成する
  ⇒ 他のパートより音高が高くなる

$$part_i (1 \le i \le N)$$

$$L_r(part, note) = \frac{1}{N} \sum_{i}^{N} \left(\frac{1}{2} + (q_i - \frac{1}{2}) \cdot M(time_i) \right)$$

$$M(t) = 1 - \exp(-C \cdot t)$$

q_i: $part$ の音高が $part_i$ より高かった割合 , C:定数
$time_i$: $part$ と $part_i$の同時発音時間 (1小節1として正規化)



$time_i$が十分小さい: $L_r(part, note) = \frac{1}{2}$

$time_i$が十分大きい: $L_r(part, note) = \frac{1}{N} \sum_{i}^{N} q_i$

パート形成　実験結果

音色類似度	定位近接度	音高遷移度	音高高低維持度	精度(%)「カノン」	精度(%)「蛍の光」
○	ー	ー	ー	63.32	79.87
ー	○	ー	ー	77.39	84.90
ー	ー	○	ー	66.50	75.17
ー	ー	ー	○	57.03	63.76
○	○	○	ー	85.01	90.27
○	○	ー	○	77.39	84.90
○	ー	○	○	66.50	79.19
ー	○	○	○	84.60	91.61
○	○	○	○	84.86	90.27

楽器音の音源同定とは

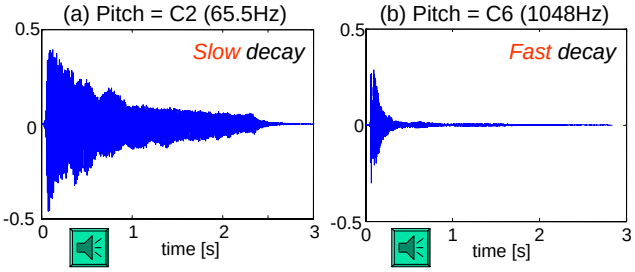
楽器音の音源同定
（音からの楽器名の同定）
‖
自動採譜・音楽アーカイブ構築
などにおいて重要な課題

<inst begin=00:00:00
end=02:10:00>
piano
</inst>
……

関連研究：
•音楽認識関連の多くは，音高推定を指向（音源同定は少数）
•現状の性能は，
単独音：70-80% / 10-30クラス
混合音：60-70% / 3-5クラス

特徴量抽出での問題点の所在

同一楽器でも音高によって音色が変化する
（∵楽器音は，他の音に比べて音域広）
例 Piano



音高による音色変化（音高依存性）への対応

[課題] 音高による音色変化の確率分布としての
モデル化（∵Bayes-basedの様々な識別器利用可）

[こんな方法は?]
半音毎に異なる多次元正規分布を学習し，
同定時に適切な多次元正規分布を選択する.

↓

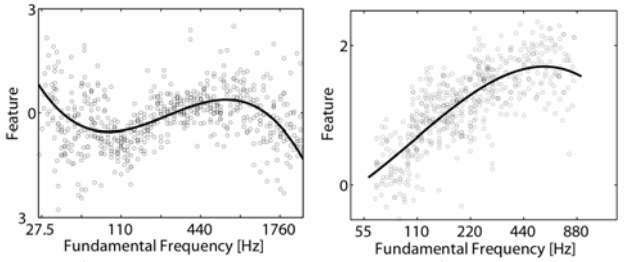
十分な学習データを用意するのは非現実的
（多次元正規分布学習には大量の学習データ）

↓

少ない学習データのためには，**確率分布そのものに音高による音色変化を表現する機能**が必要

音高依存性の表現法

- 音色を表す各特徴量の音高依存性を
基本周波数(F0)の関数として近似
- **このF0の関数をパラメータとする確率分布**を使用

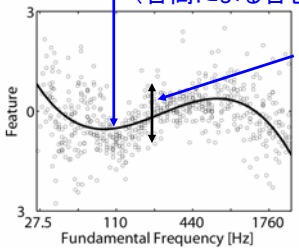


左はピアノの第4軸，右はチェロの第1軸（いずれも次元圧縮後）

F0依存多次元正規分布

多次元正規分布の拡張
＝平均ベクトルが基本周波数の関数

代表値関数
各特徴量のF0依存性を近似するF0の関数
（音高による音色変化を表現）



F0正規化共分散行列
各特徴量を代表値関数で
正規化したうえで算出
（音高以外の要因による
音色変化を表現）

音源同定アルゴリズム

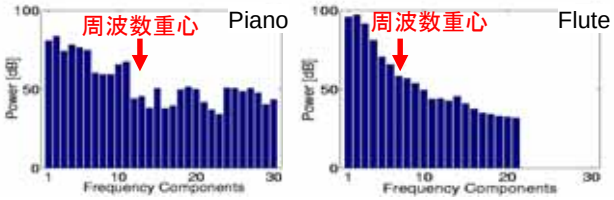
1. 特徴抽出（129個）
2. 主成分分析で次元圧縮
   （累積寄与率99%で79次元に圧縮）
3. 線形判別分析でさらに次元圧縮
   （19楽器なので18次元に圧縮）
4. F0依存多次元正規分布のパラメータ推定
5. ベイズ決定規則に基づいて楽器名を同定
   （事後確率が最大になる楽器名を見つける）

$$g_i(x; f) = \log\ p(x \mid \omega_i; f) + \log\ p(\omega_i; f)$$

使用する特徴空間

音源同定に用いるものと同じ特徴空間を使用
⇒[北原，音情研2002]で用いたものを使用

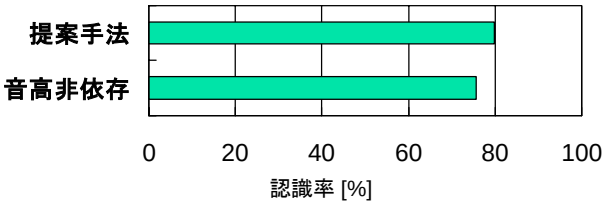
- 「周波数重心」，「パワー包絡線の近似直線の傾き」など，129個の特徴量を抽出
- 主成分分析で79次元（累積寄与率：99%）に圧縮し，さらに線形判別分析で18次元に圧縮



実 験 方 法

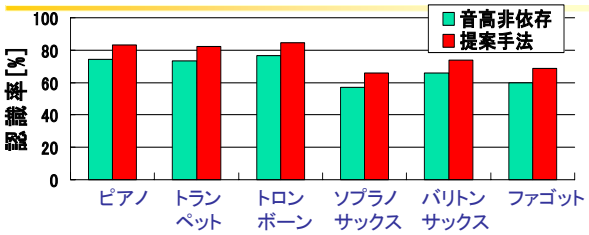
- 使用データベース：RWC-MDB-I-2001
  - 実楽器の単独発音を半音ごとに収録
  - 今回は19種類の楽器を使用
  - 各楽器に，3楽器個体，3種類の音の強さ
  - 今回は，通常の奏法のみ使用
  - 使用したデータ総数：6,247個
- 上記のデータを無作為に10等分し，クロスバリデーション
- 音高は既知

実 験 結 果



提案手法：79.73%の認識率を実現．
音高非依存に比べ，4.00%認識率向上．
誤り削減率は，16.48%．

実験結果：認識率が7%以上改善された楽器



- ピアノ：最も性能改善
  （認識率9.06%改善，誤り削減35.13%）
  ∵音域が広く音高による音色変化が顕著
- PF, TR, TBで約33～35%の認識誤りを削減
- SS, BS, FGでも20%以上の認識誤りを削減

未知楽器の問題

- 学習データにない楽器（未知楽器）が入力されたときに，それをどう扱うかという問題
- 実際のタグ付けでは不可避な問題

> 実際の音楽では，オーケストラ用楽器，民族楽器，シンセサイザーによる合成音など**多種多様な楽音**が使用され，これらの学習データを**網羅的に収集するのは困難**

- 従来研究では指摘されてこなかった

未知楽器の問題

- 本研究における解決策：
  既知楽器⇒**楽器名レベル**で認識
  未知楽器⇒**カテゴリーレベル**で認識

- 具体的なアルゴリズム
  (1) **楽器名レベル**で認識
  (2) (1)の結果が本当に正しいかどうか判定
  （「既知」か「未知」かの判定に相当）
  (3) (2)でFalse（＝未知）と判定されたら，
  カテゴリーレベルで再認識

楽器カテゴリーの設計

楽器の発音機構に基づく階層表現

大分類	中分類	小分類	属する楽器
弦楽器	——	打弦楽器	PF
		撥弦楽器	CG, UK, AG
		擦弦楽器	VN, VL, VC
管楽器	木管楽器	無簧楽器	PC, FL, RC
		単簧楽器	SS, AS, TS, BS, CL
		複簧楽器	OB, FG
	金管楽器	——	TR, TB
打楽器	（省略）	（省略）	（省略）

楽器カテゴリーの設計

- 音源同定に適した楽器カテゴリーとは？
  ⇒楽器の音響的類似性を総合的にとらえた楽器カテゴリー
  ⇒楽器の音響的類似性に基づく階層表現を自動獲得し，そこから楽器カテゴリーを作成

従来からある楽器の発音機構に基づく階層表現が使えるのでは？

No! 楽器の発音機構に基づく階層表現は，必ずしも**音響的類似性をとらえていない**．

e.g. バイオリンとギターはともに弦楽器だが音響的には大きく異なる

楽器階層の獲得における課題と解決策

課題1 使用する特徴空間によって結果が変化
音源同定で用いるものと同じ特徴空間を使用
⇒任意の音源同定システムに対して，適切な階層表現を自動的に獲得

課題2 音高などにより特徴空間上の位置が変化
各楽器多数の音響信号を用意し，各楽器の分布に対して階層的クラスタリング
⇒各楽器1音のみに比べ，各楽器の特徴空間上の位置関係を適切に把握可能

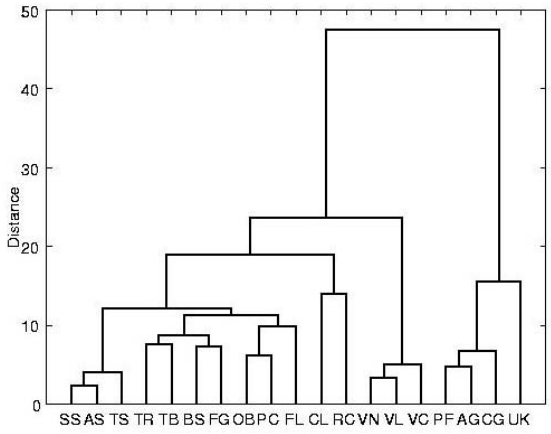
使用する楽器音の音響信号

1楽器あたり130～700個，計6,247個を使用

- 19種類のオーケストラ楽器の実楽器音データを「RWC-MDB-I-2001」から抜粋
- 半音ごとに全音域収録
- 各楽器，3楽器個体，3種類の音の強さ
- 通常の奏法のみ使用

⬇

以上のデータから得られる各楽器の特徴空間上の分布を多次元正規分布で近似し，各楽器間のマハラノビス汎距離を使って階層的クラスタリング

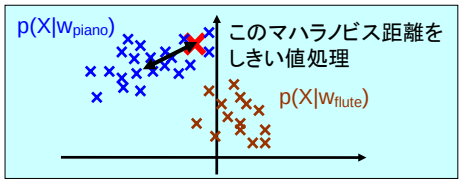


楽器カテゴリー設計結果

大分類	中分類	小分類	属する楽器
減衰系楽器	——	ウクレレ以外	PF, CG, AG
		ウクレレ	UK
持続系楽器	弦楽器	——	VN, VL, VC
	管楽器	サックス	SS, AS, TS
		クラリネット	CL
		リコーダー	RC
		低音系＋α	TR, TB, BS, FG
		高音系	OB, PC, FL

楽器音同定アルゴリズム

(1) **楽器名レベル**で認識
(2) (1)の結果が本当に正しいか判定
（Falseなら「未知楽器である」とみなす）
認識対象音から学習データ（分布）までのマハラノビス距離がしきい値以内ならTrue
(3) (2)の結果がFalseなら**カテゴリーレベル**で再認識



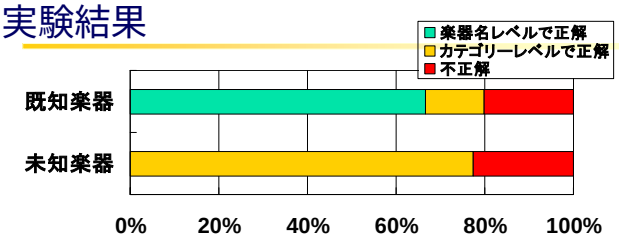
評価実験条件

- 既知楽器なら**楽器名レベル**で，
未知楽器なら**カテゴリーレベル**で認識
- 学習データ，評価用データともに
単音を1つ1つ個別に収録したものを使用
- 認識（楽器名・カテゴリーともに）では，
129次元の特徴空間を**PCA**で**79**次元に，
LDAでさらに**18**次元に圧縮したものを使用
- **既知／未知の判定**では，129次元の特徴空間を**PCA**で**23**次元に圧縮したものを使用
- **既知／未知の判定**で用いるしきい値は**40**

使用データベースの詳細

- **学習データ: 自然楽器音**
（RWC-MDB-I-2001から抜粋した
19楽器6,247音のうち，ランダムに半分を選択）
- **既知楽器の評価データ: 自然楽器音**
（上記の残り半分）
- **未知楽器の評価データ: 電子楽器音**
（ヤマハ製MU2000に収録されている
- エレクトリックピアノ**(ElecPf)**,
- シンセストリングス**(SynStr)**,
- シンセブラス**(SynBrs)**.
※各々2バリエーションずつ使用）

実験結果



- 誤り率は，既知楽器で約20%，未知楽器で約23%.
- このような楽器音理解は，情報統合においても有用
e.g. 音から「楽器名はわからないが弦楽器」と同定
画像から「ある民族楽器」
⇒弦楽器に属する新たな楽器として再学習

発達論的コミュニケーション

- 脳モデル化による対話能力自己獲得仮説原理
- 社会的対話能力獲得原理、ソーシャルインタラクション
- 対話能力を獲得する上での視覚認識機能
- 未発達レベルでの、曖昧語の発話・音声認識機能
- 音素認識、擬音語認識、真似発話（鸚鵡返し）
- 対話を通した運動と知覚の双方向発達スキーム、アクティブパーセプション
- 対話による視覚、聴覚の発達的融合機能
- その他感覚モダリティの融合

環境音の擬音語認識とは

- 発達論的コミュニケーションの立場から、
  聴覚機構の工学的実現の第一歩として、
  「環境音を擬音語として認識」

- 環境音とは？
  音声・楽音以外の音全般
  - 自然音（風の音、動物の声など）
  - 人工音（機械の駆動音など）
  - 楽器音（楽器の単発音）

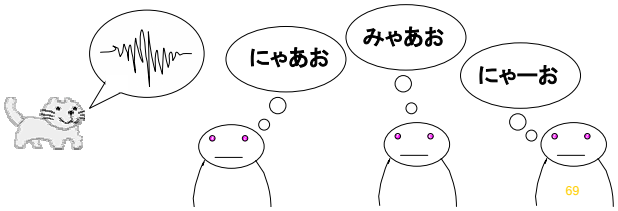
環境音の擬音語認識の背景

- 擬音語使用の利点
  - 自然で人間らしい表現を実現
    日本語では日常生活で擬音語が多用される
  - 音の詳細な表現が可能　[Wake01]

- 擬音語認識の期待される効用
  - マンマシンインターフェースの高度化
  - 聴覚障碍者の補助（擬音語を用いた字幕）
  - 音声認識の未定義語問題を処理

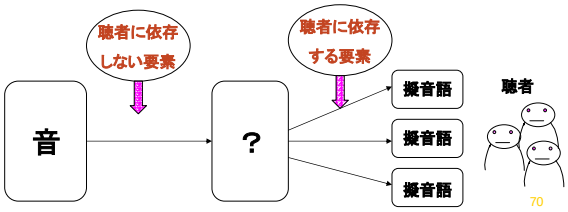
環境音の擬音語表現の問題

- **擬音語変換の問題点**
  環境音を表す擬音語は聴者により異なる
  「にゃあお」「みゃあお」「にゃーお」・・・
  ⇒『正しい』認識を定義することが難しい



擬音語表現の曖昧性の解決策

- **解決策**
  ・共通項⇒聴者に依存しない要素
  ・その他⇒聴者に依存する要素
  これらを分けて段階的に処理を行う
  ＊音節レベル/モーラレベル/音素レベル



音節とモーラ

- 音節　　例：漢字（音読み）、など
- モーラ　例：かな文字、など
  - 日本語はモーラ言語
- 音素　　例：アルファベット、など

	新聞紙	こけこっこー
音節	しん ぶん し	こ け こっ こー
モーラ	し ん ぶ ん し	こ け こっ こ ー
音素	sh i N b u N sh i	k o k e k o Q k o R

各レベルの聴者非依存性

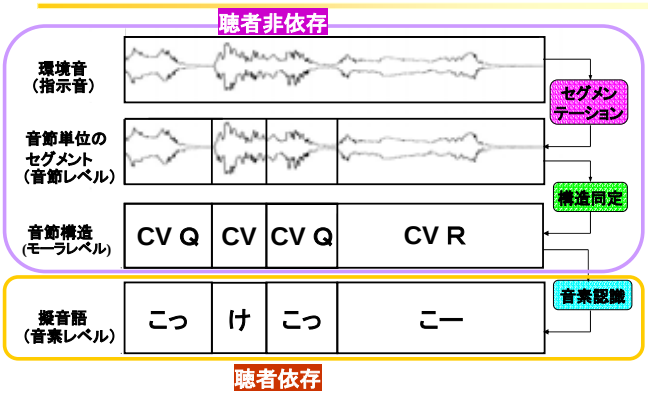
- 音節レベル　（の聴者非依存性）
  音節数やアラインメントが聴者に依らず一致

- モーラレベル（の聴者非依存性）
  右表の4種類のモーラ記号に
  よる表現が聴者に依らず一致

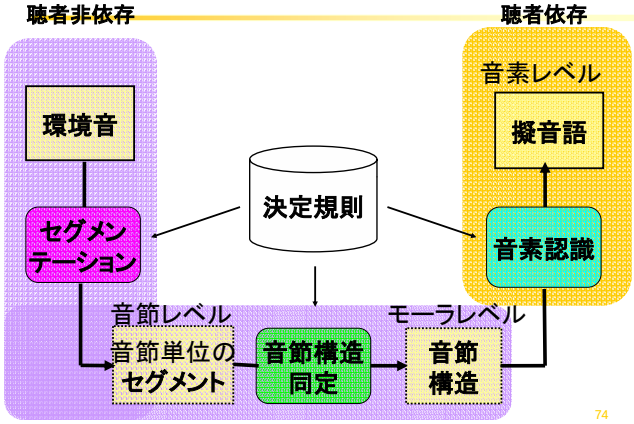
モーラ記号
CV:かな　R:長音
Q :促音　N:撥音

- 音素レベル　（の聴者非依存性）
  モーラレベルが一致し、対応する各CVの
  音素もまた一致

擬音語認識の階層的処理



3レベルによる変換



各レベルにおける聴者非依存性

- 音節レベル （聴者非依存）
  Sonority Theory（聞こえ度理論）
  「音節＝**Sonority**の一山」
- モーラレベル（聴者非依存）
  日本人の感覚では一般に非依存 [比屋根ら]
  ⇒ 聴取実験で妥当性を検証
- 音素レベル（聴者依存）
  聴者依存であるが何らかの傾向はある

参考文献

・“Sound Retrieval with Intuitive Verbal Descriptions”
和氣早苗 旭敏之 IEICE TRANS.INF.&SYST.vol.E84-D,No.11, 2001
・「犬は「びよ」と鳴いていた －日本語は擬音語・擬態語が面白い－」
山口仲美 光文社 2002
・「オノマトペ －形態と意味－」 田守育啓 くろしお出版 1999
・“Industrial Sound Design” 前田修(http://member.nifty.ne.jp/~ISD/)
・異音の表現における擬音語の検討
田中基八郎 日本機械学会論文集C編 Vol.61 No.592, 1995
・単発音のスペクトル構造とその擬音語表現に関する検討
比屋根一雄 他, 信学技法 SP97-125, 1998
・**“Sound Retrieval with Intuitive Verbal Expressions”**
Sanae Wake, Toshiyuki Asahi, Proc. Of ICAD '98 (International Conference on Auditory Display), The proceedings have been published as an eWiC, 1998

レポート問題（奥乃分）

- 下記2問を各々10ptで A4 4ページ以上でまとめる。

1. 音声に限らず一般の音（楽音・環境音・混合音）に対する、他のメディアとの情報統合のモデル化について述べよ。ただし、実現可能な機能やロバストとなる機能を具体的に取り上げて考察すること。
2. 発達論的コミュニケーションの具体的なモデル化・処理方法について述べよ。

- あるいは、MPEG-7論文（50ページ）を10pt・A4 10ページ以上でまとめる（論文は奥乃まで）
- 津崎先生・奥乃分の締切りは
  1月23日（金）午後5時・10号館事務室

レポート問題（津崎先生分）

[1] A群に示した日常体験として聴覚的な特徴に対して、それぞれ最も関わりが深いと思われる事項をB群から選択し、アラビア数字とローマ数字の組み合わせで答えよ。
[2] 設問[1]での選択の根拠が具体的に分かるように、聴覚的な処理の流れを概説せよ。
（字数制限：全角で1600字程度（ワープロ使用の場合は、12ポイントのフォントでシングル・スペースでA4用紙１枚程度））

1. 振幅の等しい純音の大きさは音の高さが違うと違って聞こえる。
2. 音として聞こえる空気振動の振幅のダイナミック・レンジは非常に大きい。
3. ピアノの調律師はひとつのピアノの鍵盤を叩いた音の中にさらに部分音を聞き出すことができる。
4. ピアノ、ギター、トランペットなど管弦楽器の音や音声の母音は、物理的には複数の周波数成分の加算として表されるにもかかわらず、通常はひとつの高さを持った一つの音として知覚されている。
5. 男性と女性が同じ旋律を歌うと実際には１オクターブのずれがあることが多いが、それに気がつくことは少ない。

i. 非線形応答性
ii. 中耳によるインピーダンス整合
iii. 内耳の基底膜の振動
iv. 位相固定した聴神経発火
v. 両耳間時間差
vi. 臨界帯域
vii. 外耳・中耳の周波数応答特性
viii. 鼓膜の振動
ix. 同期性、調波構造の制約
x. 音脈分凝

[1] A群に示した日常体験として聴覚的な特徴に対して、それぞれ最も関わりが深いと思われる事項をB群から選択し、アラビア数字とローマ数字の組み合わせで答えよ。
[2] 設問[1]での選択の根拠が具体的に分かるように、聴覚的な処理の流れを概説せよ。（字数制限：全角で1600字程度｛ワープロ使用の場合は、12ポイントのフォントでシングル・スペースでA4用紙１枚程度｝）

1. 振幅の等しい純音の大きさは音の高さが違うと違って聞こえる。
2. 音として聞こえる空気振動の振幅のダイナミック・レンジは非常に大きい。
3. ピアノの調律師はひとつのピアノの鍵盤を叩いた音の中にさらに部分音を聞き出すことができる。
4. ピアノ、ギター、トランペットなど管弦楽器の音や音声の母音は、物理的には複数の周波数成分の加算として表されるにもかかわらず、通常はひとつの高さを持った一つの音として知覚されている。
5. 男性と女性が同じ旋律を歌うと実際には１オクターブのずれがあることが多いが、それに気がつくことは少ない。

i. 非線形応答性
ii. 中耳によるインピーダンス整合
iii. 内耳の基底膜の振動
iv. 位相固定した聴神経発火
v. 両耳間時間差
vi. 臨界帯域
vii. 外耳・中耳の周波数応答特性
viii. 鼓膜の振動
ix. 同期性、調波構造の制約
x. 音脈分凝