

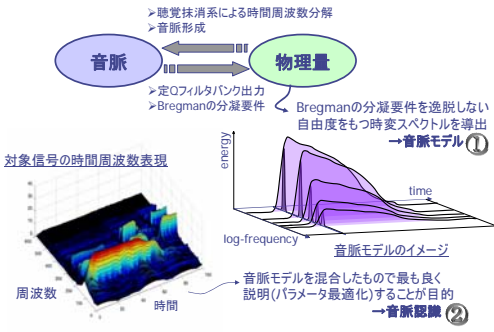
調波時間構造化クラスタリングによるCASAへのアプローチ

亀岡弘和 ルルー・ジョナトン 小野順貴 嵯峨山茂樹



概要

- ◆人間の聴覚機能をヒントにした多重音解析法の確立
- ◆調波時間構造化クラスタリング(HTC)



◆本発表で扱う問題:

- ◆混合音から目的音の情報を分離推定
 - ◆ロボット聴覚、音声認識、音楽の自動採譜への応用
- ⇒ 数理的には大変難しい!
- 一にもかわらず、我々人間は容易に混合音から目的音だけを選択的に聴き取ることができる
- 一もしかしたら人間の聴覚機能に解法のヒントがあるかも?

◆聴覚情景分析 (Auditory Scene Analysis; ASA)

- ◆A. Bregmanによる示唆:
 - 『音を通じて環境を把握する聴覚の働き』
- ⇒ 聴覚の低次の分離能力:
 - ◆音響信号はスペクトログラムに似た要素に「分解」される
 - ◆同じ音源に由来する要素は「群化」されて音脈を形成する
 - ◆群化のされやすさ(Bregmanの分離要件)は、
 - 周期性(調波構造)
 - 調波成分の共通の立ち上がり
 - 調波成分の振幅および周波数の共通の変化
 - 成分の連続性
 - 共通の音源位置

◆計算論的聴覚情景分析 (CASA)

- ◆人間の優れた聴覚機能を計算機で実現する試み
- ◆目的: Bregmanの分離要件に基づく混合音分離
 - ◆音脈の認識に有用な特徴量(ピッチ周波数など)の抽出
 - ◆目的音に対応する音脈の分離再構成
- ◆従来法:
 - ◆規則記述的な人工知能的アプローチ (Cooke'93, Brown'94, Ellis'94, Fishbach'94, 中谷ら'96)
 - ◆制約つき最適化問題を解く数理的アプローチ (西ら'98, 鷗木ら'99, 安部ら'00, Wuら'03)
- ⇒ 多段処理的アプローチに基づく:

(step1) 周波数方向の群化(調波構造に相当する瞬時特徴量抽出)

(step2) 時間方向の群化(瞬時特徴量の時間方向のスムージング)

本当にこれが群化プロセスの最適な実践方法か? 個々の音源の時間周波数全域に渡ったスペクトル構造を一挙に推定できる方法論が不可欠ではないか?

という問題意識のもと、定式化した手法こそが調波時間構造化クラスタリング(HTC)である

① 定QフィルタバンクとBregmanの分離要件に基づく混合音脈モデルの導出

■「周期性」「調波成分の共通の周波数変化」

◆擬似周期信号の解析信号表現

$$f_k(t) = \sum_{n=1}^N \tilde{w}_{k,n}(t) e^{j(\theta_{k,n}(t) + \phi_{k,n})}$$

初期位相
瞬間振幅
瞬間位相
k: 音源インデックス
n: 倍音インデックス

◆定Qフィルタバンク出力の導出

◆ウェーブレット基底関数の定義: $\psi_{a,b}(t) \equiv \frac{1}{\sqrt{2\pi a}} \psi\left(\frac{t-b}{a}\right)$

◆連続ウェーブレット変換: $W_k(\log \frac{1}{a}, b) \equiv \langle f_k(t), \psi_{a,b}(t) \rangle_{t \in \mathbb{R}}$

$$= \int_{-\infty}^{\infty} \sum_{n=1}^N \tilde{w}_{k,n}(t) e^{j(\theta_{k,n}(t) + \phi_{k,n})} \psi_{a,b}(t) dt$$

中心周波数1のアライズング
ウェーブレット

◆瞬間位相と瞬間振幅の時刻 b 周辺で0次および1次近似:

$$\tilde{w}_{k,n}(t) \approx \tilde{w}_{k,n}(b), \quad \theta_{k,n}(t) \approx \theta_{k,n}(b) + \theta'_k(b)(t-b)$$

以後、 $\mu_k(t) \equiv \theta'_k(t)$

$$\Psi_{a,b}^*(t) = \sum_{n=1}^N \tilde{w}_{k,n}(b) e^{j(\theta_{k,n}(b) + \phi_{k,n})} \int_{-\infty}^{\infty} e^{j\mu_k(t)(t-b)} \psi_{a,b}(t) dt$$

$$\Psi^*(am_k(b)) = (\mathcal{F}[\Psi(t)]) = \Psi(\omega)$$

◆混合音信号の定Qフィルタバンク出力

$$W(\log \frac{1}{a}, b) = \sum_{k=1}^K W_k(\log \frac{1}{a}, b)$$

K: 音源数

◆変数変換: $x = \log \frac{1}{a} (a = e^{-x})$

$$\Omega_k(t) = \log \mu_k(t)$$

$$\therefore W(x, b) = \sum_{k=1}^K \sum_{n=1}^N \tilde{w}_{k,n}(b) \Psi^*(ne^{-x \log \Omega_k(b)} e^{j(\theta_{k,n}(b) + \phi_{k,n})})$$

◆アナライズングウェーブレットの具体形:

$$\Psi(\omega) = \begin{cases} \exp\left(-\frac{(\log \omega)^2}{4\sigma^2}\right) & (\omega > 0) \\ 0 & (\omega \leq 0) \end{cases}$$

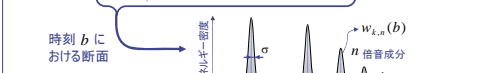
$$W(x, b) = \sum_{k=1}^K \sum_{n=1}^N \tilde{w}_{k,n}(b) \exp\left(-\frac{(x - \Omega_k(b) - \log n)^2}{4\sigma^2}\right) e^{j(\theta_{k,n}(b) + \phi_{k,n})}$$

◆スパース性によるパワースペクトルの加法性の近似仮定:

交差項が自乗項に比べて十分小さい

$$\|W(x, b)\|^2 \approx \sum_{k=1}^K \sum_{n=1}^N \tilde{w}_{k,n}^2(b) \exp\left(-\frac{(x - \Omega_k(b) - \log n)^2}{4\sigma^2}\right) e^{j(\theta_{k,n}(b) + \phi_{k,n})}$$

$$= \sum_{k=1}^K \sum_{n=1}^N \frac{\tilde{w}_{k,n}^2(b)}{w_{k,n}(b)} \exp\left(-\frac{(x - \Omega_k(b) - \log n)^2}{2\sigma^2}\right)$$



■「調波成分の共通の立ち上がり/振幅変化」「成分の連続性」

◆調波成分が共通のパワーエンベロープをなす:

$$w_{k,n}(t) = v_{k,n} u_k(t)$$

n倍音の
パワーエンベロープ

共通パワーエンベロープ

非負
連続的に変化

$$u_k(t) = \sum_{y=0}^{Y-1} \frac{u_{k,y}}{\sqrt{2\pi\phi_k}} \exp\left(-\frac{(t - \tau_k - y\phi_k)^2}{2\phi_k}\right)$$

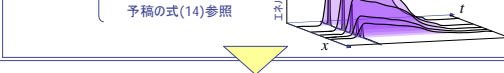
拘束つき混合正規分布モデル

$$\|W(x, t)\|^2 = \sum_{k,n,y} \frac{v_{k,n}^2 u_{k,y}^2}{\sqrt{2\pi\phi_k}} e^{-\frac{(x - \Omega_k(t) - \log n)^2}{2\sigma^2} - \frac{(t - \tau_k - y\phi_k)^2}{2\phi_k}}$$

(混合音脈モデル)

◆瞬時ピッチ周波数:

$$\Omega_k(t) \equiv \begin{cases} \Omega_{k,0} + \Omega_{k,1} + \dots \\ \text{3次スプライン関数:} \\ \text{予稿の式(14)参照} \end{cases} \times K$$



◆評価実験

- ◆目的:
 - ◆方法論の工学的な実用可能性を探る
 - ◆Bregmanの分離要件がいかに物理的事実と結びついているかを示す
- ◆評価方法:
 - ◆ピッチ周波数推定精度(音脈認識精度の目安)
- ◆対象: 単一話者音声信号、混合音声信号、音楽信号

◆実験結果:

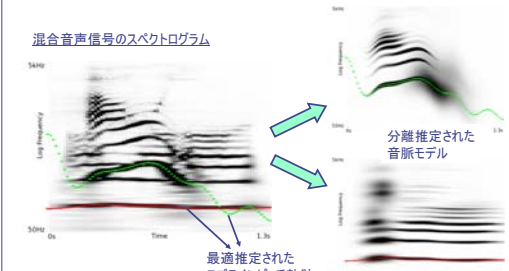
- ◆単一話者音声のピッチ周波数推定精度
 - ◆実験データ: 女性と男性話者による計100個の英文読み上げ音声のLaryngograph信号
 - ◆瞬時ピッチ周波数モデル: 3次スプライン関数
 - ◆許容する正解ピッチからの誤差: ±20%
 - ◆比較対象: 最先端のピッチ抽出法「VIN」を含む他多数

Method	HTC	WWB	HTC	WWB	HTC	
Gross error (%)	19.0	16.8	15.8	9.2	8.8	12.6
	1.9	1.7	3.8	3.2	1.4	3.5

◆混合音声のピッチ周波数推定精度

- ◆実験データ: 2話者(男性と女性、女性同士、男性同士)の音声を平均パワーが等しくなるように混合して作った計150個の混合信号
- ◆瞬時ピッチ周波数モデル: 3次スプライン関数
- ◆許容する正解ピッチからの誤差: ±10%, ±20%
- ◆比較対象: WuらによるCASAの最先端手法 (WWB法)

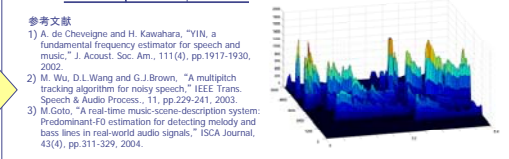
許容誤差	20%		10%	
	HTC	WWB	HTC	WWB
男性と女性	93.3	81.8	86.8	81.5
男性同士	96.1	83.4	87.9	69.0
女性同士	98.9	95.8	95.6	90.8
total	96.1	87.0	90.2	83.5



◆音楽信号のピッチ周波数推定精度

- ◆実験データ: クラシックとジャズの8楽曲(RWC研究用音楽DBより)
- ◆瞬時ピッチ周波数モデル: 0次多項式
- ◆許容する正解ピッチからの誤差: ±3%
- ◆比較対象: 最先端の音楽音高推定法 (PreEst)

手法名	HTC	PreEst
data1	81.2	74.2
data2	77.9	71.8
data3	64.2	55.9
data4	75.2	76.2
data5	62.2	62.3
data6	63.8	48.8
data7	63.2	53.6
data8	70.9	57.6
total	70.4	62.4



◆まとめ

- ◆人間の聴覚機能をヒントにした多重音解析法の確立
- ◆個々の音源の時間周波数全域に渡ったスペクトル構造を一挙に推定できる手法、調波時間構造化クラスタリング(HTC)を提案
 - ◆Bregmanの分離要件から逸脱しない範囲の自由度をもつ時変スペクトル構造を直接的にモデル化→音脈モデル
 - ◆音脈モデルのパラメータを反復アルゴリズム(EMアルゴリズムと形式上は等価)により効果的に推定
- ◆単一話者音声、2話者混合音声、音楽を対象としたピッチ周波数推定精度の評価実験を通し、各分野の最先端従来法を上回る結果を得た

② 音脈認識のアルゴリズム(音脈モデルのパラメータの最適化)

- ◆対象信号を最も良く説明するモデルパラメータ: $\Theta = \{\Omega_{k,i}, \{v_{k,n}\}_{n=1}^N, \{u_{k,y}\}_{y=0}^{Y-1}, \tau_k, \phi_k\}_{k=1}^K$ を求めることが目的
- ◆対象信号の定Qフィルタバンク出力のパワースペクトルが $\|Y(x, t)\|^2$ のとき、 $\|Y(x, t)\|^2 \approx \|W(x, t)\|^2$ となる Θ を求める問題
- ◆非負関数間の近さを表す尺度の導入

ダイバージェンス: 目的関数

$$I(\Theta) \equiv \iint \left(\|Y(x, t)\|^2 \log \frac{\|Y(x, t)\|^2}{\|W(x, t)\|^2} - \|Y(x, t)\|^2 + \|W(x, t)\|^2 \right) dx dt + P(\Theta)$$

ペナルティ関数項

◆目的関数の最小化: 解析的には難しいが、補助関数を使った効果的な反復推定アルゴリズムが導ける

◆準備: 凸不等式による補助関数の設計

$$I(\Theta) \leq I^*(\Theta, m) \equiv \iint \sum_{k,n,y} \left(m_{k,n,y}(x, t) \|Y(x, t)\|^2 \log \frac{m_{k,n,y}(x, t) \|Y(x, t)\|^2}{w_{k,n,y}(x, t)} - m_{k,n,y}(x, t) \|Y(x, t)\|^2 + w_{k,n,y}(x, t) \right) dx dt + P(\Theta)$$

ただし、 $\sum_{k,n,y} m_{k,n,y}(x, t) = 1, m_{k,n,y}(x, t) \in (0, 1)$

$$w_{k,n,y}(x, t) \equiv \frac{v_{k,n}^2 u_{k,y}^2}{\sqrt{2\pi\phi_k}} e^{-\frac{(x - \Omega_k(t) - \log n)^2}{2\sigma^2} - \frac{(t - \tau_k - y\phi_k)^2}{2\phi_k}}$$
