7D－7

マルチエージェントシステムによる音響ストリーム分離
― ストリーム分離の排他性の向上 ―

中谷 智広　奥乃 博　川端 豪

NTT 基礎研究所

1 はじめに

一般環境での音を媒介とした状況理解を，**音環境理解** (Auditory Scene Analysis) と呼ぶ [1, 2]. 音環境理解が機械的に実現できれば，ロボットの耳など，音響処理の応用分野は飛躍的に拡大する．通常の環境では，複数の音が同時に存在するのが普通であり，混合音から環境内の各事象を理解するためには，混合音中の各音を個別に聞き分ける必要がある．混合音から分離された音を，**音響ストリーム**と呼ぶ．我々は，音響ストリーム分離の枠組をマルチエージェントシステムを用いて実現した (本稿では旧システムと呼ぶ) [3]．各音響ストリームの管理を一つの追跡エージェントが担当し，エージェントどうしが相互作用しながらストリームの分離を行なう．このシステムでは，エージェントの追加により，容易に機能が拡張できる [4]．また，動的な追跡エージェントの生成・消滅機構により，ストリームの数を自動的に制御できる．しかし，ストリームの排他的分離の誤差が多い場合，ストリーム数のコントロールが必ずしも適切には機能しなくなる．本稿では，この排他的分離機構を適正化して，ストリーム数が増えた場合にも安定してストリーム分離を行なえるようにする．４つのストリームが存在する場合のストリーム分離について，実験結果を示す．

2 マルチエージェントの音響ストリーム分離

2.1 システムの概要

調波構造ストリームを分離するエージェント (調波構造追跡エージェント) 群の構成を図1に示す．(本稿では，調波構造追跡エージェントに限定して話を進める．) 一つの生成エージェント (Generator) と，動的に生成・消滅する調波構造追跡エージェント (Harmonic Tracer) からなる．生成エージェントは，各時刻の入力中に未知音を発見すると，調波構造追跡エージェントを生成する．調波構造追跡エージェントは，以後，その音を追跡し，逐次的にストリームを分離する．調波構造追跡エージェントは，調波構造の強さ $E(\omega)$ [1] を極大にする ω を追跡して，各時刻の基本周波数を求める．次に，各 $H_k(\omega)$ [1] の値から，各倍音の強さ・位相を求める．

2.2 旧システムの排他的分離法

調波構造追跡エージェントは，追跡時に相互のストリームの影響を低減させるために，三つの原理に基づいて相互作用する．(1) まず，各時刻において，各調波構造追跡エージェントは，自分が追跡しているストリームの次フレームの波形を予測して，他のエージェントの入力から減算する．(2) 次に，エージェントの共有検知域変数 (周波数帯ごとに値を持つベクトル変数) を，調波構造追跡エージェントは動的に変更して，他のエージェントの周波数帯ごとの感度を自動調整する．各エージェントは，検知域値以下の強さの倍音は無視しながらストリームを追跡する．すべての倍音が検知域値以下になった調波構造追跡エージェントは消滅する．検知域変更は，(1) で生じる予測誤差の影響を取り除く役割を果たしている．(3) さらに，本システムでは，追跡誤差が原因となって複数の追跡エージェントが同じストリームを追跡することが起きる．これに対して，追跡エージェントどうしは相互にストリームを比較して，同じストリームを追跡していると判断した場合，冗長な追跡エージェントは消滅させる．本システムでは，これまで，実験により，二つの声や合成音からなる混合音が分離できることを示した．

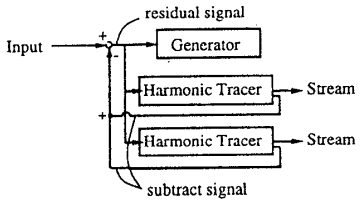


図 1: 調波構造に基づくストリーム分離

2.3 旧システム問題点

旧システムの排他的分離には，次の問題がある．

- エージェントの数が増加すると，検知域値が高くなりシステムの感度が悪くなるため，ストリームが発見されないことがある．
- エージェント数が少ない時，システムの感度が良くなり，予測誤差に反応した不要な追跡エージェントを生成する．

つまり，予測誤差の影響を除くために大きな検知域値を用いると，たくさんのストリームが扱えないことになり，逆に，検知域値を下げると誤差の影響を受けてしまう．

3 排他的分離の適正化

本稿では，システムが扱えるストリームの数の範囲を広げるために，(a) 減算信号の誤差を減少し，(b) 不要な追跡エージェントの生成を抑制し，(c) 検知域値の変更方法を適正化する．

(a) 減算信号の精度をあげるために，各追跡エージェントは，現在の状態から次フレームの入力信号を予測するだけではなく，次フレームの波形を使って基本周波数を同定する．この方法では，各入力フレームに対し基本周波数

Multi-agent Based Sound Stream Segregation — exclusive allocation of sound stream.

Tomohiro Nakatani, Hiroshi G. Okuno and Takeshi Kawabata
NTT Basic Research Laboratories
3-1, Morinosato-Wakamiya, Atsugi, Kanagawa 243-01, Japan

[1] $E(\omega) = \sum_{k=1}^{n} \|H_k(\omega)\|^2$, $H_k(\omega) = \sum_t x(t) \cdot \exp(-jk\omega t)$. ($\omega$: 基本周波数, $x(t)$: 入力, n: 倍音の数, t: 時間, を表す.)

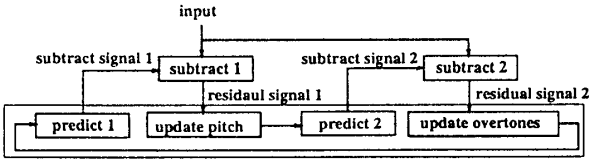


図 2: 各時刻の追跡エージェントの処理のながれ

同定を行なうための減算処理と，ストリーム同定のための減算処理の 2 回の処理を行なう．これにより，2 回目の減算処理では，基本周波数予測誤差に基づく減算誤差が減少する．各追跡エージェントは各時刻において，(1) 現在のストリームの状態に応じて次フレームの予測信号を作る．(2) 減算処理を行う．(3) 一回目の残差信号を受け取る．(4) 基本周波数の同定を行う．(5) 基本周波数と各倍音の位相を更新して 2 回目の減算処理を行なう．(6) 二回目の残差信号を受け取る．(7) 各倍音の強さと位相を同定する．(図 2) 旧システムとの主な違いは，(5) と (6) の処理が加わったことである．

(b) 生成エージェントは，まず，一回目の残差信号を用いて，旧システムと同じ方法で追跡エージェントを生成する．この追跡エージェントは，他のストリームの予測誤差に反応して生成された可能性がある．これを検査するために，二回目の残差信号を用いる．追跡エージェントは，二回目の残差信号 $e_2(t)$ と，追跡エージェント自身の減算信号 $s(t)$ との相関 $r = (s(t), e_2(t))/(s(t), s(t))$ を求める．$r < c\ (= -0.65)$ の時，この追跡エージェントは，他のストリームの周波数予測誤差に反応して生成されたものと判断して直ちに消滅する．

(c) (a), (b) で示した方法を用いることにより，ストリームの基本周波数予測誤差が原因となる不要な追跡エージェントの発生が減少するので，各追跡エージェントが検知域値を上昇させる周波数帯域幅をより小さくすることができる．実験で用いたシステムでは，追跡エージェントは，各倍音 (強さ: A, 周波数, ω) ごとに，以下の式に基づいて，検知域変数 $\theta(\omega')$ (ω': 周波数帯の代表周波数) を上昇させる．

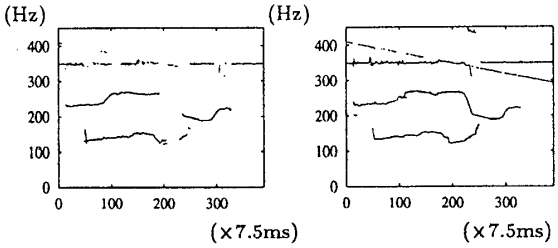
$$\theta(\omega') = \max(\theta(\omega'), 0.25 \cdot T(0.5 \cdot (\omega + \omega')).$$

ただし，$T(\omega') = A \cdot \| \sum_t \sin(\omega t) \exp(-j\omega' t) \|$．旧システムとの主な違いは，300 Hz 以上の帯域では，変更帯域幅が狭くなったことと，検知域値の変更量を最大値で代表させることによって，エージェント数が増えた場合も検知域があまり大きくならなくなったことである．

4 実験による評価

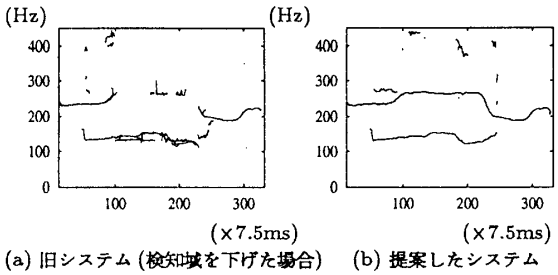
実験では，男性の声 ("あいうえお")，女性の声 ("あいうえお")，合成音 1 (基本周波数固定 350 Hz)，合成音 2 (基本周波数変動 410 Hz - 290 Hz) の四つの音を用いる．合成音は指数減衰する倍音を，6 kHz までもっている．男性の声を基準としたとき，各音の SN 比は，女性の声 -1.6 dB，合成音 1 10 dB，合成音 2 10 dB である．

実験 1 旧システムと，提案したシステムの両方に，四つすべての音の混合音を入力した時に，分離されたすべてのストリームの基本周波数の流れを図 3(a), (b) に示す．図より，旧システムでは，分離ストリームに多くの欠落部分があるのに比べて，提案した方法では，ストリームの欠落が減少していることがわかる．



(a) 旧システム　(b) 提案したシステム

図 3: 四つの音の混合音入力時に分離されたストリーム



(a) 旧システム (検知域を下げた場合)　(b) 提案したシステム

図 4: 二つの声の混合音入力時に分離されたストリーム

実験 2 旧システムで検知域変更手続きだけを本稿で提案した方法にしたときの性能を，提案したシステムと比較する．これらのシステムに，男性の声，女性の声の混合音を入力したときに分離されたストリームの基本周波数の流れを，図 4に示す．旧システムでは，検知域を下げると不要な追跡エージェントが多数発生して安定したストリーム分離が行えない (図 4(a))．これに対し，本システムでは，ストリーム数が少ない場合でも安定したストリーム分離が行える (図 4(b))．

5 まとめ

マルチエージェントによるストリーム分離を報告した．このシステムでは，非明示的にストリーム数のコントロールが行なえる．しかし，ストリームの予測誤差が大きいと，システムの感度の自動調整機能がうまく働かない．本稿では，減算信号の予測誤差を減少させ，追跡エージェント生成ルーチンを適正化して，予測誤差が排他的分離に与える影響を減少させることによって，旧システムに比べて安定で感度がよいシステムを実現した．実験では，混合音中に含まれる音の数が，それぞれ，二つ，四つのどちらの場合に対しても，提案したシステムが効率良くストリーム分離できることを示した．今後の課題としては，音声に特徴的な性質を利用してストリーム分離を行なう新しいエージェントの設計が挙げられる．

参考文献

[1] Bregman, A.S.: *Auditory Scene Analysis - the perceptual organization of sound,* MIT Press.

[2] Brown, G.: Computational auditory scene analysis: A representational approach. *PhD thesis, Dept of Computer Science, University of Sheffield,*.

[3] Nakatani, T., Okuno, H.G. and Kawabata, T.: Auditory Stream Segregation in Auditory Scene Analysis with a Multi-Agent System, *Proc. of AAAI94,* pp.100-107, Aug.

[4] 中谷, 奥乃, 川端: マルチエージェントシステムによる音響ストリーム分離のダイナミクス, 人工知能学会全国大会, 1994.