社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2008－MUS－74 (25)
2008－SLP－70 (25)
2008／2／9

サブバンド信号振幅変化に着目した音源分離

荒井 佑真 † 東山 三樹夫 †† 白井 克彦 †

† 早稲田大学大学院理工学研究科情報・ネットワーク専攻
†† 早稲田大学国際情報通信研究センター

概要 本稿では，サブバンド信号振幅変化に着目し，複数の楽器音が混在したモノラル音響信号を対象とする音源分離手法を提案する．サブバンド信号振幅に関する制約条件を用いることで，分離対象を楽音と限定することなく非楽音も同様に分離することが目的である．我々の手法においては，復元性誤差関数，類似性誤差関数，時間的連続性誤差関数からなるモデル化誤差関数を最小化するようにそれぞれの楽器音をモデル化することで，尤もらしい分離結果を得る．評価実験により，NMF（Nonnegative Matrix Factorization）を用いた手法よりも高い精度で分離することができた．

Sound Separation Based on the Grouping Cues Focused on the Amplitude of Subband Signals

Yuma ARAI † Mikio TOYAMA †† Katuhiko SHIRAI †

† Graduate School of Science and Engineering, Waseda University
†† Global Information and Telecommunication Institute, Waseda University

Abstract We propose in this paper a new approach for the separation of sound sources in one channel music signals. The algorithm is based on the grouping cues focused on the amplitude of subband signals, and can separate not only piched musical instrument sounds but also drum sounds. In our method, each sound source is modeled by minimizing the error function composed reconstruction error function, similarity error function and temporal continuity error function. The performance of the proposed method was compared with the metod based on NMF(Nonnegative Matrix Factorization). According to these simulations, the proposed method enables a better separation than the method based on NMF.

1 はじめに

音響信号を計算機上で扱うようになって久しいが，未だに複数の音源情報を含む音響信号を扱うのは困難である．音源分離とは，様々な音源からの音響信号が混在した信号中から，特定の音響信号，もしくはその音響信号に関する情報を取り出す作業を指す．曲からの自動情報抽出，音楽の個人的なカスタマイズ，雑音除去，コンピュータの耳等，様々な応用が考えられる研究課題である．

音源分離問題は一般的に以下のように定式化される．

$$x(t) = \sum_{k=1}^{K} g_k s_k(t) \qquad (1)$$

観測信号 $x(t)$ から，信号中に含まれる音源 k の信号 $s_k(t)$ とその重み g_k を求める問題であるが，これだけでは一意に解くことができないので，対象とする分離問題に合わせて含まれる各音源信号 $s(t)$ になにかしらの制約を与える必要がある．本研究では，楽器音からなるモノラル混合音響信号を入力とし，その中に混在している各音響信号に関する知識をほとんど持たない状況下での音源分離問題を対象とする．

人間の聴覚を計算機上で実現しようとする計算論的聴覚情景分析（CASA: Computational Auditory Scene Analysis）においては，Bregman が挙げた規則 [1] を音源分離問題を解くための制約とし，それらの制約を数学的に解釈したモデルが提案されている [2][3]．しかし，これらの手法は調波構造による制約が不可欠であるため，楽音[1]を分離するには適しており，音高推定とも相性が良いが，打楽器等の非楽音[2]を分離対象とはしていない．分離対象を限

[1] 明確な音の高さを感じることができ，基音とその整数倍の倍音から構成される音．
[2] 振動に一定の規則性が認められず，音の高さが不明瞭な音．

定しない手法として，入力信号を多次元の特徴量に変換することによりモノラル信号に対しても ICA を適用させた独立部分空間分析（ISA: Independent Subspace Analysis）を用いた音源分離手法も提案されている [4][5]．ISA を利用することにより単純なモデルでの分離が可能となるが，分離された各信号と混合音中の音源の関係性は不明であり，グルーピングの問題も残っている．また，ISA と同様のモデルを用いる手法に NMF（Nonnegative Matrix Factorization）がある．ただし，NMF においては各信号の独立性は用いずに，各信号が非負であることを制約条件として信号の分離を行う [6]．分離対象を振幅スペクトル，もしくはパワースペクトルに設定することで音源分離への適用も可能であり [7]，NMF をベースにその他の制約を加えた音源分離手法も提案されているが [8]，NMF も ISA と同様の問題を抱えている．

本研究においては式 (1) 中の $s(t)$ を楽器毎の信号として捉える．つまり，音源分離問題を

1. 観測信号を楽器毎の信号に分離
2. 楽器毎の信号を音高別の信号に分離

と階層化し，第一段階である楽器毎の分離を目的とする．楽器毎の分離を目的とすることで，CASA のアプローチをとっても調波構造に関する制約を考慮する必要がなくなり，対象を楽音に限定することなく同一の方法論で非楽音も分離することができると考えた．そこで，調波構造に関する制約の代わりにサブバンド信号振幅に関する制約条件を用いることにより各音源のサブバンド信号振幅の分離を目指す．

2 提案手法

2.1 概要

観測された混合音のサブバンド信号振幅を行列 X により表現する．

$$X = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,T} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,T} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N,1} & x_{N,2} & \cdots & x_{N,T} \end{bmatrix} \quad (2)$$

$x_{n,t}$ は帯域 n，時間 t における観測振幅値である．分離問題を以下のように定式化する．

$$X \simeq \sum_{k=1}^{K} W_k M_k \quad (3)$$

M_k は楽器音モデル k を記述した行列であり，W_k はその重みである．それぞれ以下の通りである．

$$M_k = \begin{bmatrix} f_{k,1,1} & f_{k,1,2} & \cdots & f_{k,1,T} \\ f_{k,2,1} & f_{k,2,2} & \cdots & f_{k,2,T} \\ \vdots & \vdots & \ddots & \vdots \\ f_{k,N,1} & f_{k,N,2} & \cdots & f_{k,N,T} \end{bmatrix} \quad (4)$$

$$W_k = \begin{bmatrix} w_{k,1} & 0 & \cdots & 0 \\ 0 & w_{k,2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & w_{k,N} \end{bmatrix} \quad (5)$$

ここで $f_{k,n,t}$ は楽器音モデル k，帯域 n，時間 t における振幅値である．ただし，以下のようにサブバンド毎のパワーが 1 になるように正規化したものを用いる．

$$\sum_{t=1}^{T} f_{k,n,t}^2 = 1 \quad (6)$$

また，各重み行列 W は最小二乗法に基づいた以下の式を解くことで一意に定まるものとする．

$$\frac{\partial}{\partial w_{k,n}} \sum_{t=1}^{T} (x_{n,t} - \sum_{k=1}^{K} w_{k,n} f_{k,n,t})^2 = 0 \quad (7)$$

ただし，重みは非負であるので，$w_{k,n} < 0$ の時 $w_{k,n} = 0$ とする．

以上の分離問題を踏まえ，提案手法の手順を示す．

1. フィルタバンク分析により X を算出する．
2. 振幅に対する制約を用いて各 M を算出する．
3. 式（7）により各 W を算出する．
4. 各 WM を分離結果とする．

ここで，まず考えるべきは各楽器音モデル行列 M を一意に定めるための制約条件である．提案手法で用いた制約を 2.2 で述べる．また，制約の数学的記述，それらの条件下でのモデル算出アルゴリズムは 2.5 で示す．

2.2 制約

2.2.1 振幅の加法性

式（3）に示した分離問題は，サブバンド信号振幅の加法性を前提としているため，算出した振幅の加法性次第で分離精度が大きく左右されてしまう．フィルタバンク分析における振幅はヒルベルト変換により算出した解析信号の絶対値を取るのが一般的であるが，そのように算出した振幅の加法性を認めるのは困難である．そこで，2.4 において，近似的に加法性が成り立つような振幅の算出

方法の定式化と実装方法を示す．そして，近似的に加法性が成り立つとすると，X と $\sum_{k=1}^{K} W_k M_k$ の誤差に関する制約を与えることができる．具体的な実装方法は 2.5 で示す．

2.2.2 隣接する帯域間の振幅類似性

Bregman は聴覚上において音を一つにまとめる要因として調波成分の共通振幅変化を挙げた [1]．これは調波成分のみに限定されず，定 Q フィルタバンクによる楽器音のサブバンド信号振幅においても同様の傾向が見られる．その傾向は帯域同士が近い場合に特に強く，良く類似した振幅変化になることが多い．そこで，隣接する帯域間の振幅変化が類似するような制約を用いる．具体的な実装方法は 2.5 で示す．

2.2.3 振幅変化のなめらかさ

単独楽器音によるサブバンド信号振幅の時間変化は，微細な振動を除けばなめらかであり急激な変化は見られない．CASA に関する多くの従来研究においても，調波成分に関してではあるが，振幅変化のなめらかさを制約として利用している [2][3]．さらには NMF にこの制約を加えることで分離精度を上げている手法も提案されている [8]．提案手法においても，事前実験で効果が確認できたため制約とした．具体的な実装方法は 2.5 で示す．

2.3 定 Q フィルタバンク

聴覚神経の特性，音高変動への頑健さ，さらには隣接する帯域間の同一楽器音による振幅類似性も考慮して定 Q フィルタバンクを用いてサブバンド信号を算出する．具体的には，ガンマトーンフィルタを基本 wavelet とする wavelet 変換により，60Hz から 8000Hz まで 1/4 オクターブ毎に 29 個のフィルタバンクを配置した．中心周波数 f_cHz のガンマトーンフィルタのインパルス応答は次式で与えられる．

$$gt(t) = at^{n-1}\exp(-2\pi bt)\cos(2\pi f_c t + \varphi) \quad (8)$$

n はフィルタの次元であり Q 値に関係し，b はインパルス応答の長さ，つまりフィルタのバンド幅に関係するパラメータである．今回は [9] を参考に $n = 4$，$b = 1.019ERB(f_c)$ とした．ただし，

$$ERB(f) = 24.7(\frac{4.37f}{1000} + 1) \quad (9)$$

である．また，$f_c = 1024$ とした．

2.4 振幅算出

観測信号 $x(t)$ を瞬時振幅 $A(t)$，瞬時位相 $\theta(t)$ を用いて，

$$x(t) = A(t)\cos\{\theta(t)\} \quad (10)$$

と表す時，$x(t)$ から加法性を持つような $A(t)$ を一意に定めたい．

$x^{(T)}(t)$ を時間フレーム T 内における観測信号とし，正弦波の和で近似すると，

$$x^{(T)}(t) \simeq \sum_{m=1}^{M} a_m^{(T)}\sin(\omega_m^{(T)}t + \phi_m^{(T)}) \quad (11)$$

$a^{(T)}$，$\omega^{(T)}$，$\phi^{(T)}$ はそれぞれ近似に用いた各正弦波のパラメータである．ここで $A^{(T)} = \sum_{m=1}^{M} a_m^{(T)}$ とおくと，

$$x^{(T)}(t) \simeq A^{(T)} \sum_{m=1}^{M} \frac{a_m^{(T)}}{A^{(T)}}\sin(\omega_m^{(T)}t + \phi_m^{(T)}) \quad (12)$$

$-1 \leq \sum_{m=1}^{M} \frac{a_m^{(T)}}{A^{(T)}}\sin(\omega_m^{(T)}t + \phi_m^{(T)}) \leq 1$ であるから，

$$x^{(T)}(t) \simeq A^{(T)}\cos\{\theta^{(T)}(t)\} \quad (13)$$

信号を加算する時，加算前の各信号を近似した正弦波の中に互いに逆位相のものが含まれない限り，近似に必要な正弦波は加算されるはずである．よって，近似に用いた各正弦波の和を取ることにより算出した振幅は近似的に加法性を示すことができる．

以上に示した振幅算出法を実装するのに最も向いているであろうアルゴリズムは一般調和解析であるが，フレーム毎の最適化問題の繰り返しに膨大な計算量が必要となるため，帯域が広いサブバンド信号に対して適用するのは現実的でない．そこでフレーム内において $x^{(T)}(t)$ の近似に用いた正弦波のほとんどの位相が $\pm\frac{\pi}{2}$ に近づく時間 t が存在すると仮定し，$|x^{(T)}(t)|$ の最大値を $A^{(T)}$ とする．提案手法においてはこちらの手法で実装を行った．また，フレーム長を 10ms，オーバーラップはなしとし，スプライン補間により内挿を行っている．

2.5 楽器音モデル

モデル誤差関数を定義し，各モデルがそれを最小化するように最適化する．誤差関数は復元性誤差関数，類似性誤差関数，時間的連続性誤差関数の重み付け和からなるものとする．M_k の誤差関数 $e(M_k)$ を以下のように定義する．

$$e(M_k) = e_r(M_k) + \alpha e_s(M_k) + \beta e_t(M_k) \quad (14)$$

$e_r(M_k)$ は復元性誤差関数，$e_s(M_k)$ は類似性誤差関数，$e_t(M_k)$ は時間的連続性誤差関数であり，α，β はそれらの重みである．それぞれの誤差関数を以下に示す．

制約条件として振幅の加法性を挙げた．正しく分離されるほど，各楽器音の振幅和は観測信号の振幅に近付くはずである．そこで，以下の復元性誤差関数を定義する．a は正規化するための定数であり，$a = \sqrt{\sum_{n=1}^{N}\sum_{t=1}^{T} x_{n,t}^2}$ である．

$$e_r(M_k) = \frac{1}{a}\sum_{n=1}^{N}\sum_{t=1}^{T}(\sum_{k=1}^{K} w_{k,n}f_{k,n,t} - x_{n,t})^2 \quad (15)$$

制約条件として隣接する帯域間の振幅類似性を挙げた．隣接するサブバンド間において二乗誤差をとり，類似性誤差関数と定義する．

$$e_s(M_k) = \frac{1}{N}\sum_{n=1}^{N-1}\sum_{t=1}^{T}(f_{k,n+1,t} - f_{k,n,t})^2 \quad (16)$$

制約条件として，単独楽器音によるサブバンド信号振幅時間変化のなめらかさを挙げた．実装に用いた振幅算出方法においては，微細な時間振動は観測されないので，隣接する時間のみに着目して以下の時間的連続性誤差関数を定義する．

$$e_t(M_k) = \frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T-1}(f_{k,n,t+1} - f_{k,n,t})^2 \quad (17)$$

式（15），式（16），式（17）により各誤差関数の勾配は以下のように算出されるので，

$$[\nabla e_r(M_k)]_{n,t} = \frac{2}{a}(\sum_{k=1}^{K} w_{k,n}f_{k,n,t} - w_{k,n}x_{n,t}) \quad (18)$$

$$[\nabla e_s(M_k)]_{n,t} = \frac{2}{N}(2f_{k,n,t} - f_{k,n+1,t} - f_{k,n-1,t}) \quad (19)$$

$$[\nabla e_t(M_k)]_{n,t} = \frac{2}{N}(2f_{k,n,t} - f_{k,n,t+1} - f_{k,n,t-1}) \quad (20)$$

誤差関数の勾配 $\nabla e(M_k)$ を求め，

$$\nabla e(M_k) = \nabla e_r(M_k) + \alpha\nabla e_s(M_k) + \beta\nabla e_t(M_k) \quad (21)$$

最急降下法による以下の更新則を得る．

$$M_k \leftarrow M_k - \gamma\nabla e(M_k) \quad (22)$$

以上を踏まえ，モデルの最適化アルゴリズムを，

step1 各 M を初期化する．

step2 式（7）を解いて各 W を算出する．

step3 更新則（22）により各 M を更新する．

step4 step2，step3 を収束するまで繰り返す．

とする．各楽器音モデルの初期化は，モデル毎に帯域を一つ選び，選んだ帯域において観測されたサブバンド信号振幅を正規化して全帯域に用いることで行う．

3 評価実験

3.1 実験概要

提案手法の評価のために実験を行った．2.5 で述べたパラメータは事前実験により $\alpha = 0.1$，$\beta = 0.1$ とした．また，楽器音モデル初期化に利用する帯域の最適な選択方法は未だ検討中であるため，事前知識を用いることで初期化を行った．具体的には，他の楽器音との振幅の重なりが最も少ない帯域を混合前の楽器音から事前に調べた上で各モデルの初期化に利用した．実験に用いた混合音は 2 楽器音からなるものとし，混合比は 0dB，それぞれ同時に 1 秒間発音した．混合音に含まれる各楽器音はソフトウェア音源によって作成した．各楽器音は 16kHz でサンプリング，16bit で量子化したものである．実験に用いた楽器音はピアノ，バイオリン，バスドラム，スネアの 4 種類であり，その中から 2 種類を楽器音 A，楽器音 B としてそれぞれ選び出して混合音グループとした．1 つの混合音グループは 8 種類の混合音からなり，楽器音 A は同一の楽器音，楽器音 B は一オクターブ分（C4，D4，E4，F4，G4，A4，B4，C5）の異なる音高を用いている．実験に用いた混合音グループを表 1 に示す．

3.2 評価

NMF と比較することで提案手法を評価する．NMF は分離対象を楽音に限定せずに振幅スペクトルの分離を行うことができるので比較対象とした．[6] においてはユークリッド距離を最小化するアルゴリズムとダイバージェンスを最小化するアルゴリズムが提案されているが，ここではユークリッド距離を最小化するアルゴリズムを用いた．グルーピングは [4] で提案されている手法を採用した．提案手法で採用している定 Q フィルタバンクにより算出したサブバンド信号振幅に対する NMF の適用では精度が良くなかったので，短時間フーリエ変換（窓長 40ms，ハニング窓，オーバーラップなし）により観測スペクトログラムを算出した．

両手法における分離精度の評価には SNR を利用した．SNR は原音と誤差のエネルギー比であり，以下のように算出される．

$$SNR[dB] = 10\log_{10}\frac{\Sigma_{n,t}A_{n,t}^2}{\Sigma_{n,t}(\hat{A}_{n,t} - A_{n,t})^2} \quad (23)$$

表 1: 実験に用いた混合音グループ

混合音グループ	楽器音 A	楽器音 B
mix1	ピアノ（C4）	バイオリン
mix2	バイオリン（C4）	ピアノ
mix3	バスドラム	ピアノ
mix4	バスドラム	バイオリン
mix5	スネア	ピアノ
mix6	スネア	バイオリン

表 2: 提案手法と NMF における各混合音グループ毎の平均 SNR(dB)

混合音グループ	提案手法	NMF
mix1	7.98	5.15
mix2	6.11	4.78
mix3	8.09	4.55
mix4	18.95	8.40
mix5	8.27	4.83
mix6	12.95	5.74

表 3: 混合音グループ mix1 における SNR(dB)

楽器音 B 音高	楽器音 A SNR	楽器音 B SNR	平均
C4	0.44	2.74	1.59
D4	3.37	5.56	4.47
E4	3.46	4.42	3.94
F4	11.17	12.33	11.75
G4	10.39	16.20	13.29
A4	10.83	12.64	11.74
B4	5.82	8.95	7.38
C5	7.69	11.65	9.67

$A_{n,t}$ は混合前の各楽器音から算出したサブバンド信号振幅であり，$\hat{A}_{n,t}$ は分離後のサブバンド信号振幅である．

実験結果として，提案手法と NMF における各混合音グループ毎の平均 SNR を表 2 に示す．また，提案手法における傾向を確認するために，混合音グループ mix1，mix5 における各 SNR をそれぞれ表 3，表 4 に示す．さらに，ピアノ音（C4）とスネア音からなる混合音の分離を例として図示しておく．提案手法における分離結果を図 1，図 2 に，混合前の原音を図 3，図 4 に示す．

NMF と比較しても，全体的にある程度の分離精度を示すことができた．特にバイオリンとバスドラム，もしくは音高がある程度異なるピアノとバイオリンのように，振幅変化が異なっており，なおかつ帯域が大きく重ならない楽器音同士の分離は高い精度で実現できた．逆に，音高が同じピアノとバイオリンのように各楽器音の主要な帯域が重なった場合の分離精度は著しく低下してしまった．また，図 1 にも見られるように，ある帯域の

表 4: 混合音グループ mix5 における SNR(dB)

楽器音 B 音高	楽器音 A SNR	楽器音 B SNR	平均
C4	8.86	10.03	9.44
D4	7.36	9.22	8.29
E4	7.44	10.30	8.87
F4	9.07	7.75	8.41
G4	8.74	8.59	8.67
A4	9.06	10.73	9.89
B4	7.28	6.86	7.07
C5	4.89	6.22	5.55

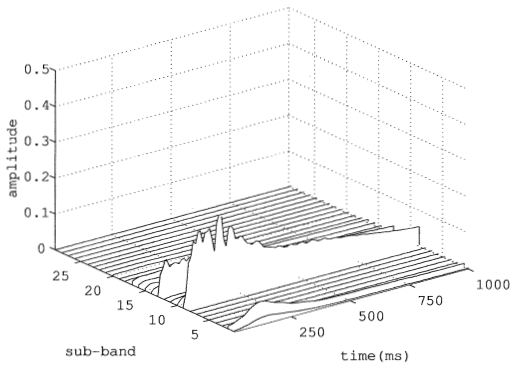

図 1: 分離されたピアノ音サブバンド信号振幅

振幅が部分的に抜け落ちてしまう現象も少なくなかった．この現象は重みの計算時に周波数的連続性の制約を用いる等の対策により改善する必要があると考えられる．

4 まとめ

分離対象を楽音に限定しない音源分離手法を提案した．サブバンド信号振幅変化に着目することで，分離対象によっては高い分離精度を得ることができた．今後は特定の帯域の振幅が抜け落ちてしまう問題，楽器音モデルの初期化の問題等に取り組んでいく予定である．

謝辞

本研究の一部は，早稲田大学理工学研究所の研究課題「自発的コミュニュケーション機構を有するマルチモーダルヒューマンインタフェースの研究」，平成 19 年度科学研究費基盤研究（B）課題

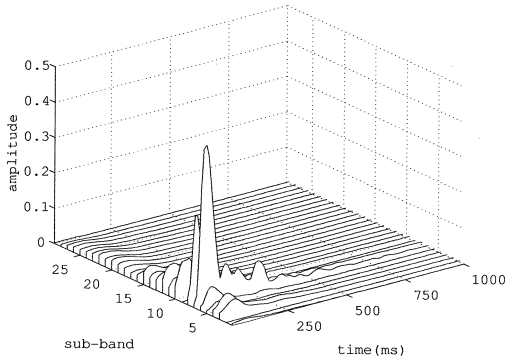

図 2: 分離されたスネア音サブバンド信号振幅

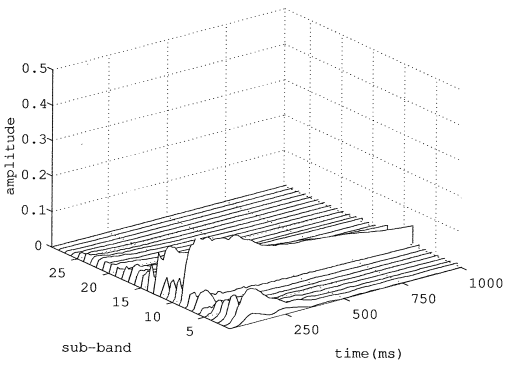

図 3: ピアノ原音のサブバンド信号振幅

番号 17300066「対話状況に応じた自発的コミュニケーション機構の研究」によるものである.

参考文献

[1] A.S. Bregman, "Auditory Scene Analysis," MIT Press, Cambridge, 1990.

[2] 鵜木裕史, 赤木正人, "聴覚の情景解析に基づいた雑音下の調波複合音の一抽出法," 電子情報通信学会論文誌, vol.J82-A, no.10, pp.1497-1507, Oct. 1999.

[3] 亀岡弘和, ルルー・ジョナトン, 小野順貴, 嵯峨山茂樹, "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," 日本音響学会聴覚研究会, H-2006-103, vol.36, no.7, pp.575-580, 2006.

[4] M.A. Casey, and A. Westner, "Separation of mixed audio sources by independent subspace analysis," Proc. International Computer Music Conference, Berlin, Germany, Aug. 2000.

[5] S. Dubnov, "Extracting sound objects by independent subspace analysis," Proc. 22nd International Conference on Virtual, Synthetic, and Entertainment Audio, Espoo, Finland, May. 2002.

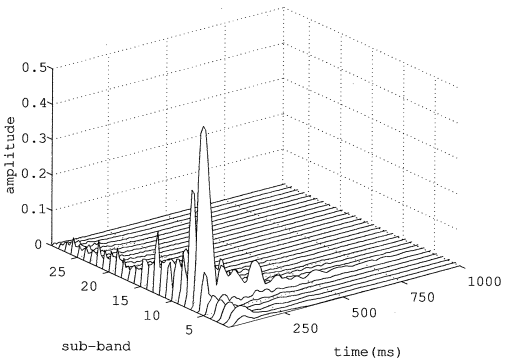

図 4: スネア原音のサブバンド信号振幅

[6] D.D. Lee, and H.S. Seung, "Algorithms for nonnegative matrix factorization," Advances in Neural Infomation Processing Systems, vo.13, pp.556-562, 2001.

[7] P. Smaragdis, and J.C. Brown, "Non-negative matrix factorization for polyphonic music transcription," 2003 IEEE Workshop on Applications of Signal Processing to Audio Acoustics, pp.177-180, Oct. 2003.

[8] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," IEEE Transactions on Audio, Speech, and Language Processing, vol.15, no.3, pp.1066-1073, Mar. 2007.

[9] 赤木正人, "聴覚フィルタとそのモデル," 電子情報通信学会誌, vol.77, no.9, pp.948-956, Sep. 1994.