2008－SLP－72 (10)
2008／7／19

ベイジアンネットワークを用いたバイナリマスキングに基づく音源分離

伊藤 弘章† 大石 康智† 宮島 千代美† 北岡 教英† 武田 一哉†

†名古屋大学大学院情報科学研究科

†{hiroaki,ohishi}[at]sp.m.is.nagoya-u.ac.jp, miyajima[at]is.nagoya-u.ac.jp, {kitaoka,kazuya.takeda}[at]nagoya-u.jp

あらまし 本研究では，音楽混じりの音声に対する単一チャネル音源分離手法を提案する．バイナリマスキングの原理に基づくと，混合信号の各周波数成分のパワーは，個々の音源のうち，その周波数成分のパワーが最も大きい音源に由来するものであると考えることができる．したがって，混合信号の時間-周波数成分の中から，個々の音源が支配する成分を選択的に残し，他の成分をマスクするためのマスクパターンを決定すれば，個々の音源を分離することが可能となる．しかし，混合信号に含まれる各音源は未知であるため，個々の音源に対してこのマスクパターンを最適に推定する必要がある．そこで，時間-周波数成分の周囲の依存関係を仮定し，ベイジアンネットワークを用いて確率的に成分を選択する手法を提案する．提案手法の有効性を確認するために，6 種類の SNR で非定常な音楽を重畳した混合信号に対して音源分離実験を行い，目的の音源成分を選択するマスクパターンの正解率と音質の評価を行った．実験結果より，マスクの正解率と音質評価ともに従来のベイズ識別器を用いる手法よりも良い結果が得られることが確認された．

Sound Source separation based on binary masking using bayesian network

Hiroaki ITOU† Yasunori OHISHI† Chiyomi MIYAJIMA† Norihide KITAOKA† Kazuya TAKEDA†

†Graduate School of Information Science, Nagoya University

Abstract In our study, we propose a method of single-channel sound source separation for a mixture of speech and music. Based on the principle of binary masking, we can assume that each time-frequency bin is dominated by a certain source whose power is highest of all original sources at the bin. Therefore, if we decide the mask pattern to selectively retain components dominated by the target signal and mask out the other signal, we can segregate the target signal from the mixed one. However, since original sources are unknown, we need to optimally estimate this mask pattern for each original source. So we assume the dependency among neighboring time-frequency components, and propose a probabilistic mask estimation method using bayesian networks. To prove the effectiveness of proposed method, we performed an experiment of source separation of mixture of speech and nonstationary musics with six defferent levels of SNRs and evaluate the accuracy of estimated mask pattern to select the target components and obtained sound quality. As a result, both accuracy and sound quality were better in comparison with conventional method which used bayesian classifier.

1 はじめに

本研究では，音楽混じりの音声に対する単一チャネル音源分離手法を提案する．従来の単一チャネル信号の音源分離問題に対する戦略としては，各音源信号に含まれる典型的なスペクトル概形パターンを事前知識として獲得し，最大事後確率 (MAP) 基準や最小二乗誤差 (MMSE) 基準によって混合信号を各パターンの重み付き和で表現することが一般的であった [1]．例えば，Benaroya らは GMM を用いて音源に含まれる局所パワースペクトル密度をモデル化し，事前知識として用いた [2]．また Jang らは ICA によって得られる音源固有の基底関数を事前知識として用いる手法を提案した [3]．これらの手法では，事前知識の学習に用いた音源データと実際の混合信号に含まれる音源信号が一致しないという問題があり，事前知識として蓄えたパターンでは混合信号に含まれる音源信号を表現することができず，分離性能が低下する場合がある．

そこで本研究では，簡素なバイナリマスキングに基づく音源分離手法を考える．この手法は，混合信号の時間-周波数成分の中から，特定の音源が支配する成分を選択的に残し，他の成分をマスクするためのマスクパターンを決定することでその音源を分離する．しかし，混合信号に含まれる各音源は未知であるため，個々の音源に対して最適なマスクパターンを推定する必要がある．そこで，ある時間，ある周波数の混合信号の成分が与えられた下で，その成分が 2 クラス (音声 or 音楽) に属する確率をそれぞれ求め，その確率の大小を比較することでクラスを決定し，個々の音源が支配する周波数成分を決定する．先行研究としてベイズ識別器を用いるものがある [4]．これは周波数帯域ごとにクラスを決定するため，周波数分解能が低下し，分離性能も低下する．そこで提案手法では，高い周波数分解能で分析した周波数ビンごとにクラスを決定する．また注目する周波数ビンのクラスは周辺の時間-周波数成分に依存すると仮定し，この依存関係をベイジアンネットワークでモデル化する．これによって，目的の音源に対応する周波数ビンを選択することができ，分離性能の向上が期待される．またさらなる応用として，音楽が重畳した音声から音楽を抑圧することによって，音声認識の認識率向上も期待される．

以下，2 章ではバイナリマスキングに基づく音源分離手法について述べる．3 章では先行研究としてベイズ識別器を用いたマスキング関数決定法について述べ，4 章では提案するベイジアンネットワークを用いたマスキング関数決定法を述べる．5 章で音源分離実験結果と音質評価結果を示す．この結果を踏まえ，6 章で考察し，7 章でまとめる．

2 バイナリマスキングに基づく音源分離

2 つの音源信号とそれらの混合信号を短時間フーリエ変換して得られる複素スペクトル $S_1(n,k), S_2(n,k), X(n,k)$ の間には，式 (1) に示す加法性が成り立つ．ここで n は時間，k は周波数を表す．

$$X(n,k) = S_1(n,k) + S_2(n,k) \tag{1}$$

バイナリマスキングに基づく音源分離では，混合信号の短時間スペクトル $X(n,k)$ にマスキング関数 $M_c(n,k), (c = 1,2)$ を乗算することで各音源の時間-周波数成分を抽出する．例えば音源 1 は，式 (2) にしたがって分離される．すなわち，マスキング関数は時刻 n, 周波数ビン k ごとに 0 か 1 の値をとり，分離したい音源を構成する周波数ビンを選択する関数である．ここで各音源が未知であるため，観測される混合信号からマスキング関数を推定する必要がある．

$$\hat{S}_1(n,k) = M_1(n,k)X(n,k) \tag{2}$$

3 ベイズ識別器を用いたマスキング関数の決定 [4]

Seltzer らは，ベイズ識別器によってマスキング関数を決定し，非定常な雑音環境下における音声認識特徴量を抽出する手法を提案した [4]．スペクトログラムから，雑音の影響を受けていない音声のみの周波数帯域を識別し，認識特徴量に利用することによって認識性能を向上させた．

スペクトルをフィルタバンク分析して各帯域ごとに特徴ベクトル $\boldsymbol{x}_{n,k}$ を作成し，音声と雑音の事後確率 $P(c \mid \boldsymbol{x}_{n,k})$ (ただし $c = \{$ 音声, 雑音 $\}$) を求め，その大小を比較して，以下のようにマスキング関数を決定する．ここで $M_{\text{音声}}(n,k)$ は音声の帯域を選択するマスキング関数である．また事後確率は，特徴ベクトルを用いて学習された混合正規分布 (GMM) を用いて得られた尤度をもとに，ベイズ則を用いて求める．

$$M_{\text{音声}}(n,k) = \begin{cases} 1 & \left(P(c = \text{音声} \mid \boldsymbol{x}_{n,k}) > P(c = \text{雑音} \mid \boldsymbol{x}_{n,k}) \text{ のとき}\right) \\ 0 & \left(P(c = \text{雑音} \mid \boldsymbol{x}_{n,k}) > P(c = \text{音声} \mid \boldsymbol{x}_{n,k}) \text{ のとき}\right) \end{cases}$$

この手法ではフィルタバンク内の帯域は同じ識別結果であり，すなわちマスキング関数の周波数分解能が悪くなってしまう．決定したマスキング関数を音源分離手法に用いるには，周波数帯域ごとではなく，周波数ビンごとに選択する必要がある．

4　ベイジアンネットワークを用いたマスキング関数の決定

ベイジアンネットワークとはグラフ構造を持つ確率モデルの1つであり，複数の確率変数間の依存関係をグラフ構造によって表し，依存関係の強さを条件付き確率で表す．ベイジアンネットワークにおいて，各ノードは確率変数を表す．提案手法では，ある時間，ある周波数における音源の種類を1つの状態と考え，1つのノードに対応させる．確率変数は混合信号のパワースペクトルの値 $Y_{n,k} = |X(n,k)|^2$ と音源の種類を決定するクラス $Z_{n,k} = \{0,1\}$ である．ここでは2つの音源が混合した信号を想定する．ここで n は時間(フレーム)，k は周波数を表す．本研究の問題を定式化すると，時間 n_0，ある周波数 k_0 における近傍 Ω_ϵ のパワースペクトルの値が分かった下での Z_{n_0,k_0} の条件付き確率 $P(Z_{n_0,k_0} \mid Y_{i,j}, \{i,j\} \in \Omega_\epsilon)$ を求め，その確率をもとにマスクを決定する問題である．以下の手順に従って，ベイジアンネットワークを用いてマスクを決定する．

4.1　確率モデルのグラフ構造の決定

横軸に時間フレーム，縦軸に周波数ビンをとった平面における格子点をグリッドと呼ぶ．各グリッドにはクラスを決定するノード $Z_{n,k}$ とパワースペクトルの値をとるノード $Y_{n,k}$ が存在し，$Y_{n,k}$ は $Z_{n,k}$ に依存するという関係があるとする．

また，グラフ構造を決定するグリッド間の依存関係を仮定して，本手法では時間方向に N フレーム，周波数方向に K ビンに依存関係があるとする．周波数方向について，全周波数ビンを用いて1つのグラフ構造を決定すると計算量が増える．そこで，周波数方向を粗くとり，帯域ごとにノードを対応させてノード数を削減する方法も考えられるが，周波数分解能が低下し，分離信号の音質が悪化してしまう．そこで全周波数帯域を U 個のサブバンドに線形分割し，各サブバンド内では精細な周波数ビン単位で依存関係があるとしてサブバンドごとに確率モデルを構築した．すなわち，FFT ポイント数を $2F$ とすると，式 (3) の関係が成り立つ．

$$F = U \times K \tag{3}$$

ただしサブバンド間は独立と仮定する．本手法で用いるグラフ構造を図1に示す．このような確率モデルを構築することで周波数分解能の低下を防ぐことが可能である．

4.2　確率モデルの学習

本節ではベイジアンネットワークの学習方法を述べる．ベイジアンネットワークにおいて，音源の種類を決定するノード $Z_{n,k}$ は離散値をとり，パワースペクトルの値をとるノード $Y_{n,k}$ は連続値をとる．学習データ

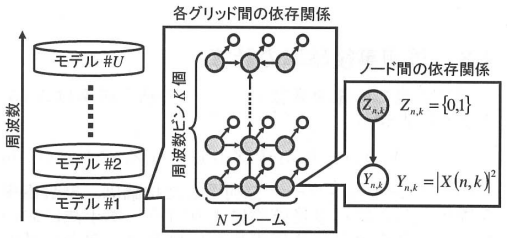
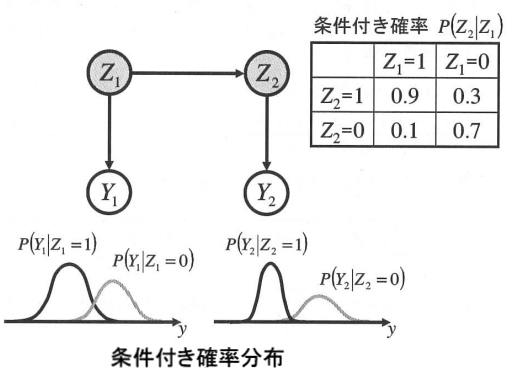


図1　提案するグラフ構造



	Z_1=1	Z_1=0
Z_2=1	0.9	0.3
Z_2=0	0.1	0.7

図2　ベイジアンネットワークの例

として雑音の混ざっていない音声に音楽を重畳した混合信号を用意した．$Y_{n,k}$ には，混合信号のパワースペクトル $|X(n,k)|^2$ を与える．$Z_{n,k}$ には，混合信号に含まれる各音源 (音声と音楽) より求められる理想的なマスクを与える．理想的なマスクの求め方を述べる．まず式 (4) のように，音声と音楽の各信号の複素スペクトル $S(n,k), M(n,k)$ の全ての時間-周波数成分について局所 SNR を求める．

$$\mathrm{SNR}_{\mathrm{local}}(n,k) = 20\log_{10}\frac{|S(n,k)|}{|M(n,k)|} \tag{4}$$

そして求めた局所 SNR が -5dB より大きいグリッドに"1"を割り当て，それ以外のグリッドに"0" を割り当てる．ここで"1" は音声に対応し，"0" は音楽に対応する．こうして得られた"1","0" のパターンを理想的なマスクと呼ぶ．

ベイジアンネットワークの統計的学習では，条件付き確率の学習を行う．離散値をとるノードに対しては条件付き確率を学習データの出現頻度から計算し，連続値をとるノードに対しては条件付き確率分布を EM アルゴリズムにより推定する．図2に示すように，学習によって離散値をとるノード Z_1, Z_2 には条件付き確率 $P(Z_2|Z_1)$ の表を与え，連続値をとるノード Y_1, Y_2 には条件付き確率密度分布として1混合のガウス分布を仮定して推定しておく．

4.3 周辺事後確率の計算

図 2 の全ての確率変数 (ノード) の結合確率は式 (5) のように表せる.

$$P(Z_1, Z_1, Y_1, Y_2) = P(Z_1)P(Z_2|Z_1)P(Y_1|Z_1)P(Y_2|Z_2) \quad (5)$$

ベイジアンネットワークを利用した推論は，結合確率を周辺化した周辺事後確率により行う．本手法において，マスク決定のために求めたい確率は，混合信号のパワースペクトルの値が観測された下で，音源の種類が音声である確率である．Z_2 に着目すると，求めたい確率は $P(Z_2 = 1|Y_1 = y_1, Y_2 = y_2)$ であり，式 (6) のように計算できる.

$$\begin{aligned}
&P(Z_2 = 1|Y_1 = y_1, Y_2 = y_2) \\
&= \frac{P(Z_2 = 1, Y_1 = y_1, Y_2 = y_2)}{P(Y_1 = y_1, Y_2 = y_2)} \\
&= \frac{\sum_{Z_1} P(Z_1)P(Z_2 = 1|Z_1)P(Y_1 = y_1|Z_1)P(Y_2 = y_2|Z_2 = 1)}{\sum_{Z_1}\sum_{Z_2} P(Z_1)P(Z_2|Z_1)P(Y_1 = y_1|Z_1)P(Y_2 = y_2|Z_2)} \\
&= \alpha P(Y_2 = y_2|Z_2 = 1)\sum_{Z_1} P(Z_1)P(Z_2 = 1|Z_1)P(Y_1 = y_1|Z_1)
\end{aligned} \quad (6)$$

ここで α は $\sum_{Z_2} P(Z_2|Y_1 = y_1, Y_2 = y_2) = 1$ となるための正規化定数である．求めた周辺事後確率がある閾値 ϵ を超えれば，音声と判定する.

5 音源分離実験と評価実験

音楽が重畳した音声に対して従来法 [4] と提案法を用いた音源分離実験を行い，推定したマスクの評価と分離音声の音質評価を行う.

5.1 実験条件

混合信号は雑音の混ざっていない音声と音楽を加算することで作成した．JNAS データベース [5] より男性 5 名，女性 5 名の計 10 発話の音声を用意し，RWC 研究用音楽データベース [6] より 10 種類の音楽ジャンルの楽曲を用意した．なお音楽は歌声の混ざっていない部分を使用した．表 1 に示すように音声 5 種類，音楽 5 種類，SNR6 種類の 150 サンプルのデータを 2 セット作成し，1 セットで確率モデルを学習し，もう一方を評価データとした．SNR は式 (7) で定義する．ただし $s(n)$ は音声，$m(n)$ は音楽，L は全信号長を表す.

$$\mathrm{SNR} = 10\log_{10}\frac{\sum_{n=0}^{L-1} s(n)^2}{\sum_{n=0}^{L-1} m(n)^2} \quad [\mathrm{dB}] \quad (7)$$

グラフ構造のパラメータ U, K, N を含めた実験条件を表 2 に示す．ここで N, K はそれぞれネットワークの時間方向，周波数方向のグリッド数を表し，U はサブバンド数を表す.

表 1 学習データと評価データ

	学習データ	評価データ
音声	5 名 (男性 2 名, 女性 3 名)	5 名 (男性 3 名, 女性 2 名)
音楽	5 曲 (カントリー, ボサノバ, ジャズ, フラメンコ, バラード)	5 曲 (ロマン派, ゴスペル, インド, クラシック, ハウス)
SNR	0,5,10,15,20,25 [dB]	0,5,10,15,20,25 [dB]

表 2 実験条件 (U はサブバンド数,K は周波数方向のグリッド数,N は時間方向のグリッド数)

	従来法	提案法
特徴量	有声区間 7 次元 無声区間 5 次元	混合信号の対数パワー 1 次元
サンプリング周波数	16 kHz	16 kHz
フレーム長	25 ms	32 ms
フレームシフト幅	10 ms	16 ms
分析窓	ハミング窓	ハミング窓
FFT ポイント数	2048 ポイント	2048 ポイント
(U, K, N)	(20,1,1)	(64,16,3)

5.2 マスクの評価

推定したマスクのうち理想的なマスクと合致した割合を正解率として算出し，推定したマスクを評価する．理想的なマスクは，4.2 節で述べた手順に従って作成した.

推定したマスクを $E_{n,k}$，理想的なマスクを $I_{n,k}$ とし，以下の集合を考える.

$$\begin{aligned}
C_1 &= \{(n,k) \mid I_{n,k} = 1 \wedge E_{n,k} = 1\} \\
C_2 &= \{(n,k) \mid I_{n,k} = 1 \wedge E_{n,k} = 0\} \\
C_3 &= \{(n,k) \mid I_{n,k} = 0 \wedge E_{n,k} = 1\} \\
C_4 &= \{(n,k) \mid I_{n,k} = 0 \wedge E_{n,k} = 0\}
\end{aligned}$$

ただし C_1 は理想的なマスクにおいて"1" かつ推定したマスクにおいて"1" とラベル付けされた状態の集合である．また正解率を式 (8) のように定義する.

$$(\text{正解率}) = \frac{|\,C_1\,| + |\,C_4\,|}{T \times K \times U} \quad (8)$$

ここで $|\,C_1\,|$ は集合 C_1 の要素数を表し，T は全フレーム数を表す．従来法と提案法で推定したマスクの例を図 3 の 1,2 列目に示し，評価結果を図 4 の左側に示す．結果から，提案法の方が悪い結果となった．これは周波数方向のグリッド数の違いによると考える．そこで提案法で推定したマスクの周波数方向のグリッド数を，従来法の周波数方向のグリッド数に合わせる．すなわち提案法で推定したマスクについて，サブバンドごとにクラスの多数決をとり，サブバンドごとにクラスを決定する．こうして得られたサブバンド単位のマスクの例を図 3 の 3 列目に示す.

サブバンド単位のマスクの評価結果を図 4 の右側に示す．この結果，提案法の方が良い結果が得られた.

5.3 音質の評価

分離音声の音質を評価するためにケプストラム距離 (Cepstrum Distance : CD) を用いた．ケプストラム距離は式 (9) で定義され，評価値にはケプストラム距

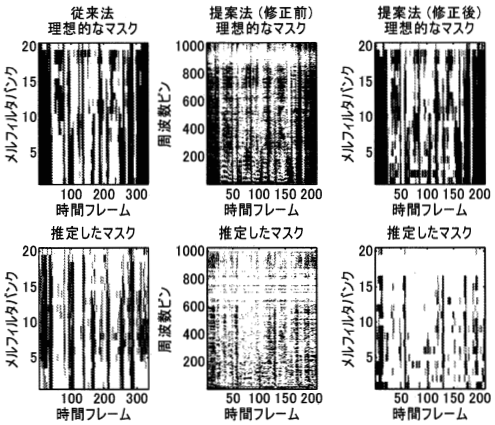


図 3　推定したマスク

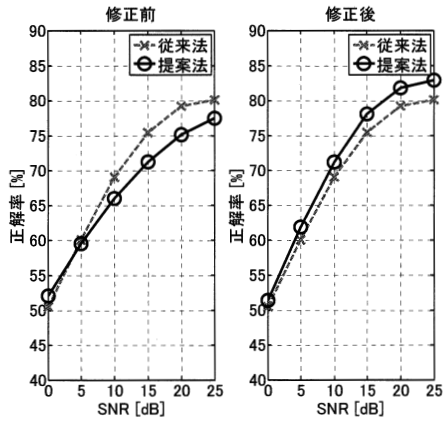


図 4　マスク正解率

離のフレーム平均を用いた．この値が小さいほど音質が良いといえる．ここで $c_x(t,k), c_{ref}(t,k)$ は，時刻 t フレーム目の分離音声と元の音声の k 次ケプストラム係数を表す．F は分析ポイント数を示し，FFT ポイント数と同じとした．T はフレーム数を表す．またフレーム長は 32ms とした．

$$\mathrm{CD} = \frac{1}{T}\sum_{t=1}^{T}\sqrt{\sum_{k=1}^{F}\left(c_x(t,k) - c_{ref}(t,k)\right)^2} \qquad (9)$$

マスキング関数作成のための周辺事後確率の閾値を 0.5 とした場合の音楽抑圧結果を図 5 に示す．また閾値を 0.5 とした場合の音質の客観評価結果を図 7 に示す．図 7 より，提案法の方が高い評価が得られたが，実際に分離された音声を聞いてみると，提案法はうまく分離できておらず，音楽が残ったままであった．

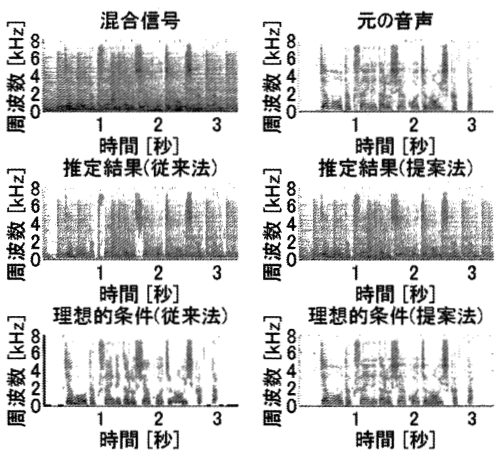


図 5　分離音声の結果:女性発話とハウス音楽を SNR0dB で混合 (理想的条件とは，理想的なマスクを用いた場合)

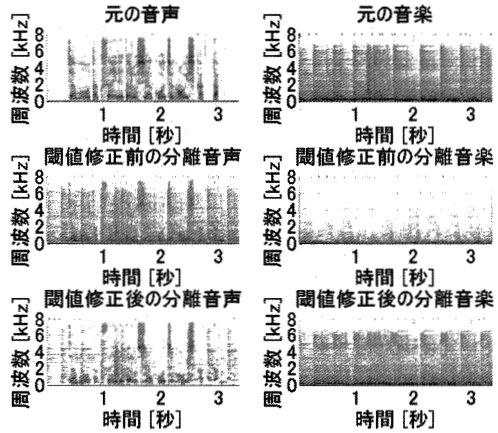


図 6　最適な閾値を用いた分離結果

6　考察

周辺事後確率からバイナリ値のマスクを決定するための閾値について考察する．低 SNR の混合信号に対して，閾値を 0.5 とすると提案法では分離音声に音楽が残ったままであった．これは理想的には音楽に対応するビンが音楽と識別されていないことが原因である．マスク評価結果の低 SNR において，正解率が低いことからもこのことが分かる．閾値を調整すれば，より音声であると信頼できる部分のみを選択することが可能であると考えられる．そこで音楽が一番抑圧される閾値を各混合音声ごとに決定する．閾値を決定するための尺度として，Noise Reduction Ratio(以下 NRR) を用いる．これは雑音がどれだけ抑圧されたかを測る尺度であり，式 (10) のように分離処理後の SNR から処

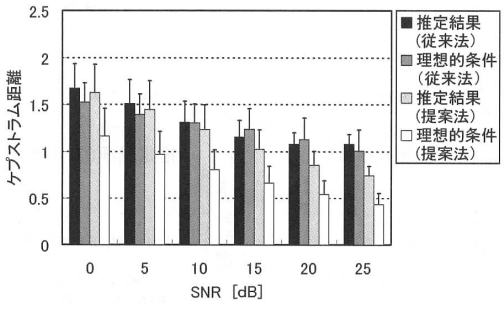


図 7　音質の客観評価結果:各 SNR におけるケプストラム距離の平均を示す.

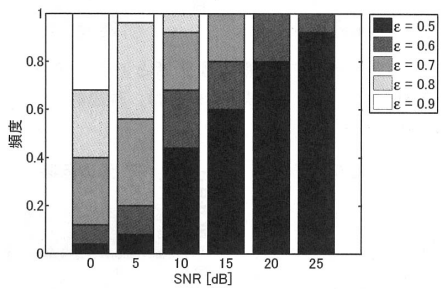


図 8　SNR 毎の最適な閾値 ϵ の内訳

理前の SNR を減算することで得られる．分離処理前後の SNR は式 (11),(12) のように算出する．ここで，$s(n)$ は元の音声，$m(n)$ は元の音楽，$\hat{s}(n)$ は分離した音声，L は全信号長を表す.

$$\mathrm{NRR} = \mathrm{SNR}_{\mathrm{out}} - \mathrm{SNR}_{\mathrm{in}} \tag{10}$$

$$\mathrm{SNR}_{\mathrm{in}} = 10\log_{10}\left(\frac{\sum_{n=0}^{L-1} s^2(n)}{\sum_{n=0}^{L-1} m^2(n)}\right) \tag{11}$$

$$\mathrm{SNR}_{\mathrm{out}} = 10\log_{10}\left(\frac{\sum_{n=0}^{L-1} \hat{s}^2(n)}{\sum_{n=0}^{L-1} \left(\hat{s}(n) - s(n)\right)^2}\right) \tag{12}$$

閾値を 0.5 から 0.9 まで 0.1 ずつ変化させて分離信号を作成し，NRR を算出した．そして NRR が最大となる閾値を最適値とした．SNR ごとに最適な閾値の割合を図 8 に示す．図 8 より，低 SNR の混合音声について音楽を抑圧するためには 0.5 より大きい閾値が必要であることが分かる．最適な閾値を用いた場合の音源分離結果を図 6 に示し，マスクの評価結果を図 9 に示す．最適な閾値を用いることにより，より音声と信頼できる部分のみを選択することができ，分離性能が向上したと言える.

7　まとめと今後の展開

バイナリマスキングに基づく音源分離手法のためのマスクをベイジアンネットワークを用いて推定する手法を提案した．音声 10 種類と音楽 10 種類を用いて，

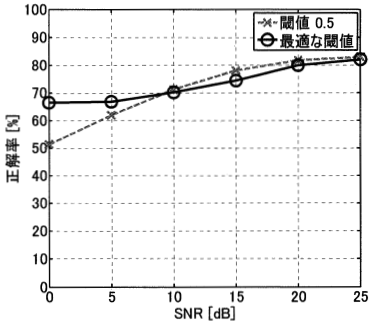


図 9　最適な閾値を用いたマスク評価結果

従来法と提案法との性能を評価するための実験を行った．推定したマスクが理想的なマスクとどれくらい合致しているか，分離信号の音質は良いかという点で評価した．結果としてマスクの評価，音質の客観評価ともに提案法の方が良い結果が得られたようにみえたが，実際に聞いたところ提案法で分離した音声には音楽が混ざったままであった．そこで混合信号に含まれる各音源を用いて最も音楽を抑圧するような閾値を決定したところ，低 SNR の混合音声に対して音楽が抑圧され，うまく分離された．今後は混合信号のみから最適な閾値を決定する手法の検討が必要である．また音質の主観評価実験と音声認識実験も行う予定である.

参考文献

[1] Blouet, R., Raraport, G., Cohen, I. and Fevotte, C.: Evaluation of several strategies for single sensor speech/music separation, *Proc. ICASSP2008*, pp. 37–40 (2008).

[2] Benaroya, L., Bimbot, F. and Gribonval, R.: Audio source separation with a single sensor, *IEEE Trans. Audio, Speech and Language Processing*, Vol. 14, No. 1, pp. 191–199 (2006).

[3] Jang, G.-J. and Lee, T.-W.: A probabilistic approach to single channel source separation, *Proc.NIPS'02*, pp. 1173–1180 (2002).

[4] Seltzer, M. L., Raj, B. and Stern, R. M.: A Bayesian classifier spectrographic mask estimation for missing feature speech recognition, *Speech Communication*, Vol. 43, pp. 379–393 (2004).

[5] Itou, K., Yamamoto, M., Takeda, K., Takezawa, T., Matuoka, T., Kobayashi, T., Shikano, K. and Itahashi, S.: JNAS:Japanease speech corpus for large vocabulary continuous speech recognition research, *J. Acoust. Soc. Jpn E*, Vol. 20, No. 3, pp. 199–206 (1999).

[6] 後藤真孝，橋口博樹，西村拓一，岡　隆一：RWC 研究用音楽データベース:音楽ジャンルデータベースと楽器音データベース，情処研報音楽情報科学，Vol. 2002, No. 40, pp. 19–26 (2002).