2008－SLP－74（4）
2008 / 12 / 9

ベイジアンネットワークを用いた単一チャネル信号による背景音楽抑圧

伊藤 弘章† 西野 隆典† 北岡 教英† 武田 一哉†

† 名古屋大学大学院 情報科学研究科 〒464-8603 名古屋市千種区不老町 1
E-mail: †hiroaki@sp.m.is.nagoya-u.ac.jp, ††nishino@media.nagoya-u.ac.jp, †††kitaoka@is.nagoya-u.ac.jp, ††††{kazuya,takeda}@nagoya-u.jp

あらまし 本研究では背景音楽混じり音声の背景音楽抑圧のための，単一チャネル信号によるノンパラメトリックな確率的手法を提案する．提案手法では，ベイズ識別器が混合信号の時間-周波数成分の局所依存性をモデル化するように拡張し，そのモデルに基づいたバイナリマスキングを行う．時間-周波数成分の依存性は，ベイズ識別器を拡張したグラフとして表現する．また提案手法は音源分離のためのベクトル量子化に基づく射影法の拡張であることも示す．最適な音響分析条件やグラフ構造は実験的に決定する．提案手法の有効性を確認するために，6 種類の音声とポップ音楽を計算機上で加算した混合信号に対して背景音楽抑圧実験を行い，音質の改善率を評価した．実験結果より，従来法よりも改善率が 4%上回った．

キーワード 単一チャネル音源分離，背景音楽抑圧，バイナリマスキング，ベイジアンネットワーク

Music suppression method for single channel speech mixed with BGM using Bayesian networks

Hiroaki ITOU†, Takanori NISHINO†, Norihide KITAOKA†, and Kazuya TAKEDA†

† Graduate School of Information Science, Nagoya University Furo-cho 1, Chikusa-ku, Nagoya, 464-8603 Japan
E-mail: †hiroaki@sp.m.is.nagoya-u.ac.jp, ††nishino@media.nagoya-u.ac.jp, †††kitaoka@is.nagoya-u.ac.jp, ††††{kazuya,takeda}@nagoya-u.jp

Abstract A non-parametric stochastic method of the single-channel speech extraction from a mixture of speech and music is proposed. In the proposed method, conventional binary masking based on Bayesian classifier is extended so that the local dependency among time-frequency components of the mixture signal is modeled. A graphical representation of the dependency is introduced as a natural extension of the Bayesian classifier. We show that the proposed approach also extends the VQ based projection method for signal separation. Training scenario and optimal graph topology of the proposed method is exploit through experimental discussions. Finally, the performance of the proposed method is evaluated using six mixtures of speech and pop music. Through the experiments, the effectiveness of the method is clarified by overperforming the conventional method by 4 % relative improvement of sound quality.

Key words single-channel sound source separation, background music suppression, binary masking, Bayesian networks

1. はじめに

音楽混じりの音声から音声のみを抽出するという研究は多く試みられている．これらの研究は，放送番組から得られる音響信号に対する音声認識技術への応用が期待される．

白色雑音，ピンク雑音，車内雑音，バブル雑音のような，従来の研究対象とされてきた定常雑音とは異なり，音楽信号は非定常雑音なので，スペクトル減算法 [1], [2] やウィナーフィルタリング [3], [4] のような従来の雑音抑圧手法で必要とされる，局所 SNR の推定を困難にする．近年では，空間的に疎に分布した複数の音源を複数のマイクで観測した混合信号に適用されるブラインド音源分離技術が提案されている [5]〜[8]．さらに，そ

の拡張として単一チャネル信号に対するブラインド音源分離も研究されている [9]. 多くの単一チャネル音源分離アルゴリズムは分離問題に対して，音声の調波性のような信号のスパース性を制約として利用している．しかし，音楽は一般的に多重ピッチ構造の信号であるため，音声同士の混合信号のために設計されたアルゴリズムの多くは音声信号と音楽信号の分離には不適当である．

Alexey ら [10] は音源を 1 つの混合ガウスモデル（GMM）で表現し，学習した音源 GMM を単一チャネル信号ブラインド音源分離問題を解く際の制約とした．ここで GMM を構成する各ガウス関数は，音源特有のスペクトル概形に対応する．図 1 に示すように，彼らの手法では音源分離はウィナーフィルタリングとして実装される．最小 2 乗誤差基準により各音源 GMM から最適なガウス関数を選択し，それらの線形結合として伝達関数を推定する．同じような考え方で，ベクトル量子化して得られるプロトタイプによって張られる空間において，混合信号から音声信号の部分空間へと射影することによって音声と音楽を分離するという研究もある [11].

これらの手法には，十分一般的で幅広い種類の音響信号に対応できるという利点がある．しかし，効果的なモデルや代表ベクトルは混合信号中に含まれる音源の統計的性質を正確に表現できなければならない．このような場合，GMM では多数のガウス関数を，ベクトル量子化では多数の代表ベクトルを必要とし，様々な問題が生じる．例えば音楽特有の全てのスペクトル概形を GMM でモデル化するのは特に非現実的である．文献 [10] では，混合信号の発話区間を検出し，音楽 GMM を適応することで問題を解決している．

本研究では以下のような，ある時間-周波数成分 (n,k) について局所的なスペクトル情報 $S(i,j)$ が与えられた下で音声信号によって支配される確率を直接学習する手法を提案する．

$$\Pr\Big\{ Z(n,k) = \mathcal{S} \;\Big|\; \{S(i,j)\}, (i,j) \in \Omega(n,k) \Big\}$$

ここで，n は時間インデックス，k は周波数インデックスを表す．Z は 2 値 $(\mathcal{S}, \mathcal{M})$ の確率変数であり，時間-周波数成分 (n,k) が音声に支配されるか否かを表す．Ω は (n,k) を中心とした局所的な時間-周波数領域を表し，$\Omega(n,k) = \big\{ (i,j) \;\big|\; \|(i,j) - (n,k)\| < \theta \big\}$ である．

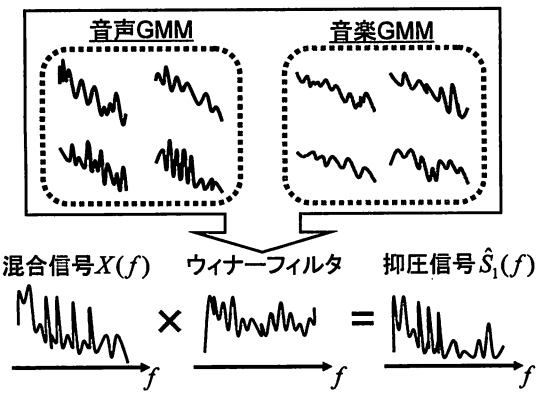

図 1　GMM を用いた音源分離 [10]
Fig. 1　Source separation using GMM [10]

前述した従来手法と提案手法との大きな違いが 2 点ある．1 つ目は学習データに混合信号を用いる点である．2 つ目は局所的なスペクトル情報 $\Omega(n,k)$ を用いる点である．$\Omega(n,k)$ は目標とする時間-周波数成分 (n,k) を中心とする時間，周波数方向の近傍成分の集合を表すが，従来法では時間方向の制約を利用していない．この考えを実装するため，$Z(i,j)$ は学習時には既知である仮定の下で，$Z(i,j)$ を隠れ変数，$S(i,j)$ を観測変数としたグラフ表現を用いる．

以下，2 章では本手法のアルゴリズムについて詳しく述べる．3 章で背景音楽抑圧実験と評価実験の結果を示し，この結果を踏まえ 4 章で考察し，5 章でまとめる．

2. 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 [12]

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．

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

ここでは音楽混じりの音声を音声と音楽に分離する手法を提案する．分離手法にはバイナリマスキングに基づく音源分離手法を用いる．

2.1　バイナリマスキングに基づく音源分離

バイナリマスキングとは，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しているとする考え方である．この原理に基づくと，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ようなマスキング関数を決定することで，個々の音源を分離することが可能である．この手法では図 2 に示すように，混合信号の短時間スペクトル $X(n,k)$ にマスキング関数 $\mathcal{M}_c$, $(c = \{ \text{音声}, \text{音楽} \})$ を乗算することで各音源の時間-周波数成分を抽出する．ここで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時間-周波数成分を選択する関数である．図 2 のマスキング関数において，白い部分が “1” の値をとり，選択する部分に対応する．また黒い部分が “0” の値をとり，マスクする部分に対応する．ただし，各音源が未知の場合には，観測された混合信号からマスキング関数を推定する必要がある．本論文では，確率的にマスキング関数を推定する手法を提案する．

2.2　ベイジアンネットワークを用いたマスキング関数の推定

ベイジアンネットワークとはグラフ構造の確率モデルの 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確率変数と考え，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．ただし $Z(n,k)$ について "0" が音楽に対応し，"1" が音声に対応する．また n は時間 (フレーム)，k は周波数を表す．マスキング関数の推定問題を定式化すると，ある時間 n_0，ある周波数 k_0 における近傍 $\Omega(n_0,k_0)$ の混合信号のパワースペクトル値が与えられた下での $Z(n_0,k_0)$ の条件付き確率 $\Pr\big(Z(n_0,k_0)=1 \mid \{Y(i,j)\},(i,j)\in\Omega(n_0,k_0)\big)$ を求め，その確率の大小を比較することでクラスを決定する問題である．

まずネットワークのグラフ構造について述べる．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．各グリッドには音源の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．またグリッド間は時間方向に N フレーム，周波数方向に K ビンで依存関係があると仮定し，図3のようなグラフ構造を仮定した．これは推定したいマスキング関数における局所的な変化パターンを考慮しており，周波数振幅の時間変化が少ないという特徴や，注目する成分が周囲の成分に依存するという特徴がグラフ上のリンクで表されている．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．従ってノード数を削減する必要があるが，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，各サブバンドごとに確率モデルを構築するとともに，サブバンド間は独立と仮定した．．提案するグラフ構造の全体像を図4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と同時に，1つ1つのモデルの構造が簡単であるため，計算コストの削減が可能である．

次にモデルの学習について述べる．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，混合信号のパワースペクトル値をとるノード $Y(n,k)$ は連続値をとる．学習データには，クリーンな音声と音楽信号を計算機上でSNRが5dBとなるように加算した混合信号を用いる．$Y(n,k)$ には混合信号のパワースペクトル $|X(n,k)|^2$ を与え，$Z(n,k)$ には混合前の各音源信号（音声と音楽）から求められる理想的なマスキング関数を与える．ここで理想的なマスキング関数とは，各音源信号の振幅スペクトルの大小を比較することで得られる "1","0" のパターンである．モデルの学習において，離散ノード $Z(n,k)$ に対しては条件付き確率をデータの出現頻度から計算し，連続ノード $Y(n,k)$ に対しては条件付き確率分布を最尤推定する．例えば図4において $(K,N)=(16,3)$ とすると，ノード $Z(n,k)$ については図5のような条件つき確率 $p\big(Z(n,k)\mid \mathbf{pa}_{Z(n,k)}\big)$ が学習され，ノード $Y(n,k)$ については図6のような条件つき確率分布 $p\big(Y(n,k)\mid \mathbf{pa}_{Y(n,k)}\big)$ が学習される．また全てのノードに関する同時確率は式 (2) で表される．

$$p(\mathbf{Z},\mathbf{Y})=\prod_{n,k}p\big(Z(n,k)\big|\mathbf{pa}_{Z(n,k)}\big)p\big(Y(n,k)\big|\mathbf{pa}_{Y(n,k)}\big) \quad (2)$$

ただし $\mathbf{Z}=\big\{Z(1,1),\ldots,Z(n,k)\big\},\mathbf{Y}=\big\{Y(1,1),\ldots,Y(n,k)\big\}$ である．また $\mathbf{pa}_{Z(n,k)}$ は $Z(n,k)$ の親ノードの集合を表し，$(K,N)=(16,3)$ の場合，図5のような3パターンが考えられる．同様に $\mathbf{pa}_{Y(n,k)}$ は $Y(n,k)$ の親ノード集合を表し，図6より $\mathbf{pa}_{Y(n,k)}=Z(n,k)$ である．$Z(n,k)$ に着目し，"0","1" のパターンがいくつ表現できるか図5を参考にして求めてみる．図5の (1) は2通り，(2) は $2^3=8$ 通り，(3) は $2^4=16$ 通りの2値パターンが表現される．また (1) に対応するノードが32個，(2) に対応するノードが1個，(3) に対応するノードが15個あるので，全体のノードでは，

$$(2\times 32)\times(2^3\times 1)\times(2^4\times 15)=122,880 \quad (\text{通り})$$

の2値パターンが表現できる．これは全てのノードにリンクがない場合（2^{48} 通り）に比べれば非常に小さく，一種のベクトル量子化が行われていると考えられる．

最後にマスキング関数の推定について述べる．マスキング関

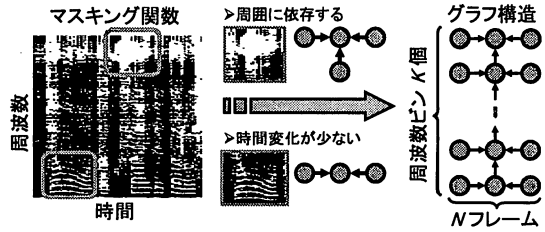

図3　グリッド間のグラフ構造の決定
Fig. 3　A decision of graph structure among grids

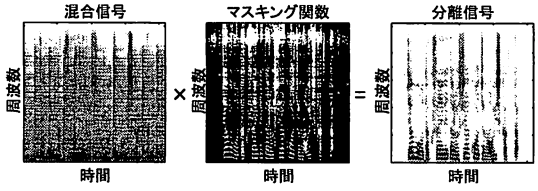

図2　バイナリマスキングに基づく音源分離
Fig. 2　Source separation based on binary masking

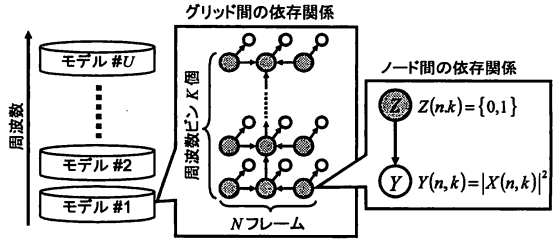

図4　提案するグラフ構造
Fig. 4　Proposed graph structure

数を決定するために求めたい確率は，混合信号のパワースペクトル値が観測された下で，ある時間-周波数ビンの音源の種類が音声である確率であり，$p\big(Z(i,j)=1 \mid \mathbf{Y}\big)$ で表される．これはモデル学習の際に求めた式 (2) の同時確率を周辺化することで求められる．

$$p(Z(i,j)=1 \mid \mathbf{Y}) = \frac{p(Z(i,j)=1,\ \mathbf{Y})}{p(\mathbf{Y})} = \frac{\sum_{\mathbf{Z}_{\setminus(i,j)}} p(Z(i,j)=1,\ \mathbf{Z}_{\setminus(i,j)},\ \mathbf{Y})}{\sum_{\mathbf{Z}} p(\mathbf{Z},\mathbf{Y})}$$

ここで $\mathbf{Z}_{\setminus(i,j)}$ は $Z(i,j)$ を除いた $\mathbf{Z}$ の全ての確率変数の集合を表す．求めた事後確率の閾値 ϵ を超えた部分を音声と判定することでマスキング関数が決定できる．

3. 背景音楽抑圧実験と評価実験

音楽が重畳した音声に対して提案法を用いた背景音楽抑圧実験を行い，音質評価を行う．提案法において，最適なグラフ構造と音響分析条件は実験的に求める．すなわち，グラフ構造において時間方向情報である N と，音響分析条件においてフレーム長を変化させて抑圧実験を行う．グラフ構造については図 7 に示すような 3 パターンを用いる．フレーム長については表 1 に示すような 4 パターンを用いる．また，音質評価ではケプストラム距離を用いて従来法 [10] と提案法を比較する．ただし，これは文献 [10] で用いられた評価尺度でないため，文献 [10] で用いられたデータを使用して，従来法の評価も行う．また，提案法で用いたデータは文献 [10] で用いられたデータとは異なる．

3.1 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 SNR が 5dB となるように加算することで作成した．JNAS データベース [13] より男性 4 名，女性 4 名の発話を用意し，RWC 研究用音楽データベース [14] よりポップス音楽 7 曲，クラシック音楽 1 曲を用意した．なお音楽には歌声も含まれている．音声 6 種類，ポップス音楽 6 種類を評価データとし，その他のデータはモデル学習の際に用いた．

モデル学習データについて，次の 3 つの場合に分類し，さらに性別依存のモデルを作成した．

（1） クローズ条件：楽曲を特定できた場合

（2） セミ・クローズ条件：楽曲のジャンルが特定できた場合

（3） オープン条件：楽曲に関して全く未知の場合

すなわち，クローズ条件では評価データでない音声と評価データと同じ音楽を加算した信号で学習し，セミ・クローズ条件では評価データでない音声と評価データにないポップス音楽を加算した信号で学習し，オープン条件では評価データでない音声とクラシック音楽を加算した信号で学習した．その他の実験条件を表 1 に示す．

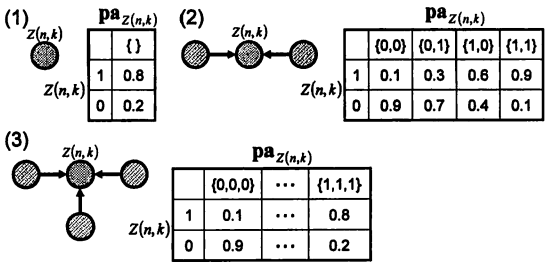

図 5 ノード $Z(n,k)$ において学習される条件付き確率表の例

Fig. 5 An example of conditional probability table of a node $Z(n,k)$

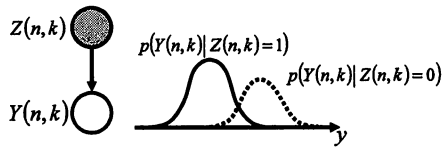

図 6 ノード $Y(n,k)$ において学習される条件付き確率分布の例

Fig. 6 An example of conditional probability distribution of a node $Y(n,k)$

3.2 背景音楽抑圧結果

各モデル学習条件での音源分離結果の例を図 8 に示す．図 8 より，クローズ条件のほうが音楽のみの部分での抑圧効果がみられ，モデル学習条件が悪くなるほど抑圧効果が低くなるのことが分かる．

3.3 音質の評価

抑圧信号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 (3) で定義され，評価値にはケプストラム距離のフレーム平均を用いた．

$$\mathrm{CD} = \frac{1}{T}\sum_{t=1}^{T}\sqrt{\sum_{k=1}^{F}\Big(c_x(t,k) - c_{ref}(t,k)\Big)^2} \qquad (3)$$

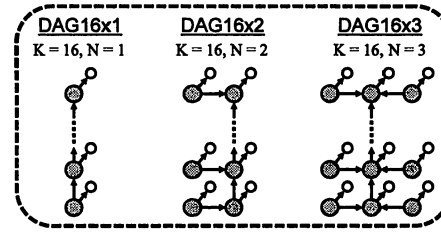

図 7 グラフ構造のパターン

Fig. 7 Patterns of graph structure

表 1 実験条件

Table 1 Experimental condition

サンプリング周波数	16 kHz
フレーム長	16, 32, 64, 128 ms
フレームシフト幅	フレーム長の 1/4
分析窓	ハニング窓
FFT ポイント数	2048
サブバンド分割数 U	64
事後確率の閾値 ϵ	0.8

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 フレーム目の評価信号と参照信号の 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，512 点とした．T はフレーム数を表す．またフレーム長は 32ms とした．

モデルの学習条件を固定した場合のグラフ構造による評価結果の比較を図 9 に示す．どの学習条件の場合も，最適なフレーム長を選べば DAG16x3 のグラフ構造が一番良い結果を示すことが分かる．またグラフ構造のパターンを固定した場合のモデル学習条件による評価結果の比較を図 10 に示す．どのグラフ構造の場合も，モデル学習条件が悪くなると評価結果が悪くなるという傾向が見られる．

図 9 より，各モデル学習条件，各グラフ構造において最適なフレーム長を決定したものを表 2 に示す．最適なフレーム長を用いた場合の従来法と提案法の評価結果の比較を図 11 に示す．ここで，従来法で用いた評価データと提案法で用いた評価データが異なるため，抑圧前の信号のケプストラム距離も異なる．従って，式 (4) の改善率を評価することで，どれだけ音質が改善されたかという点を評価する．

$$(\text{改善率}) = \frac{\text{CD}_{\text{抑圧前}} - \text{CD}_{\text{抑圧後}}}{\text{CD}_{\text{抑圧前}}} \times 100 \quad [\%] \qquad (4)$$

ただし $\text{CD}_{\text{抑圧前}}$ は式 (3) における c_x に混合信号を用いた場合のケプストラム距離であり，$\text{CD}_{\text{抑圧後}}$ は式 (3) における c_x に抑圧信号を用いた場合のケプストラム距離である．図 11 において，クローズ条件かつ DAG16x3 の場合で従来法より提案法が 4%良い結果が得られた．

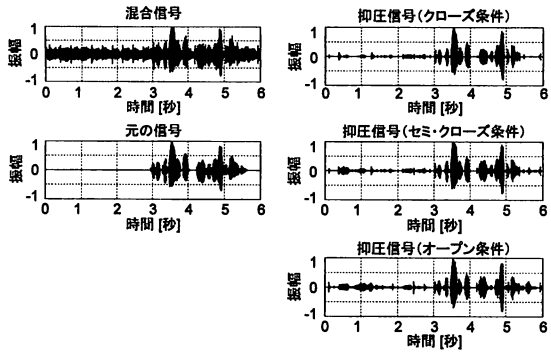

図 8 背景音楽抑圧結果の例（DAG16x3，フレーム長 16ms の場合）
Fig. 8 Waveforms of music suppression (DAG16x3, and frame length is 16ms)

表 2 各モデル学習条件，各グラフ構造における最適なフレーム長
Table 2 Optimal frame lengths for each training conditions and graph structures

	クローズ	セミ・クローズ	オープン
DAG16x1	128 ms	128 ms	128 ms
DAG16x2	128 ms	128 ms	64 ms
DAG16x3	16 ms	16 ms	16 ms

4. 考　　察

まずモデル学習条件について考察する．図 8 の抑圧結果の例や図 10 からも分かるように，学習データと混合信号の性質の違いが大きくなればなるほど，抑圧性能は低下してしまう．特に学習データと混合信号に含まれる音楽の種類が違うと，抑圧性能は格段に低下する．この問題を解決するためには，従来法 [10] でも行われているようなモデルの適応が必要であると考える．したがって，混合信号中の音楽のみの区間を発話区間検出によって見つけ出し，その部分のみを用いてモデルを適応するといった枠組みを考えることが今後の課題である．

次にグラフ構造について考察する．図 9 より，DAG16x3 が一番良い結果を示したことで，時間方向の情報を使用することの有効性が確認できた．これは音声信号の周波数振幅は，時間方向にゆっくり変化するといった制約もモデルに含ませること

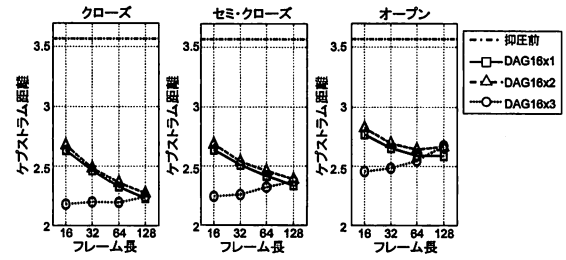

図 9 グラフ構造における比較
Fig. 9 Comparison among graph structures

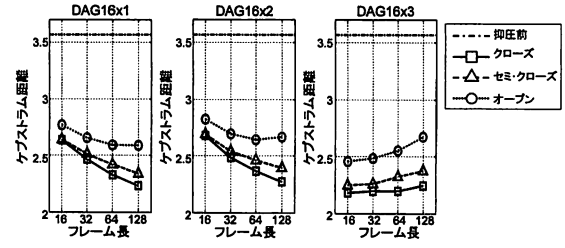

図 10 モデル学習条件における比較
Fig. 10 Comparison among experimental conditions

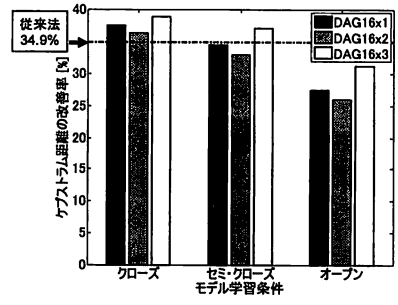

図 11 従来法と提案法の評価結果の比較
Fig. 11 Comparison between conventional method and proposed method

ができたためではないかと考える．しかし，学習されたモデルの中身（条件付き確率分布など）の考察が不十分であるため，今回のグラフ構造が最適であるかは更なる考察が必要である．

最後にバイナリマスキングの理論上の限界について議論する．バイナリマスキングにおける理想的なマスキング関数は，混合前の音声信号と音楽信号の各時間-周波数成分のパワーを比較することで求められる．マスキング関数は2値のフィルタであるため，抑圧信号のスペクトルが歪んでしまうことが予想される．従って理想的なマスキング関数が推定されたとしても，局所SNRからゲインを推定するウィナーフィルタのような連続値のフィルタより性能が悪いことが考えられる．図12に理論上の限界も含めた評価結果を示す．図12において，BMは理想的なバイナリマスキングを，WFは理想的なウィナーフィルタリングを表す．理想的なウィナーフィルタリングとは，混合前の音声信号の振幅スペクトルと混合信号の位相スペクトルを用いて再合成した場合である．結果から，提案法は従来法よりは性能が良かったが，理想的なバイナリマスキングには及ばないことが分かる．また理想的なバイナリマスキングが実現されたとしても，理想的なウィナーフィルタリングには敵わないといえる．これを踏まえて，今後は隠れ変数 $Z(n,k)$ を2値変数としてではなく連続変数としてモデル化し，ウィナーフィルタを推定する枠組みを考えていく予定である．

5. まとめと今後の展開

音楽混じり音声の背景音楽抑圧のためのノンパラメトリックな確率的手法を提案した．提案手法では混合信号の時間-周波数成分の局所依存性をベイジアンネットワークを用いてモデル化し，そのモデルに基づきバイナリマスキングを行う．音声6種類と音楽6種類を用いて，従来法と提案法の性能を評価するための実験を行った．ネットワークのグラフ構造や音響分析条件の中のフレーム長をパラメータとして，実験的に最適なパラメータを決定したところ，グラフ構造はDAG16x3，フレーム長は16msであった．ケプストラム距離を用いて，抑圧信号の音質の改善率で評価を行った．結果として，従来法より提案法が4%上回った．しかし理想的なバイナリマスキングを行った場合の性能にはまだ達しておらず，グラフ構造に関する考察などが必要であると考える．今後はモデル学習条件がオープンの場合での性能を向上させるために，学習したモデルを混合信号の性質に適応する枠組みを考える予定である．また，バイナリマスキングはウィナーフィルタリングに劣るため，連続値のフィルタゲインを推定するモデルを構築する予定である．

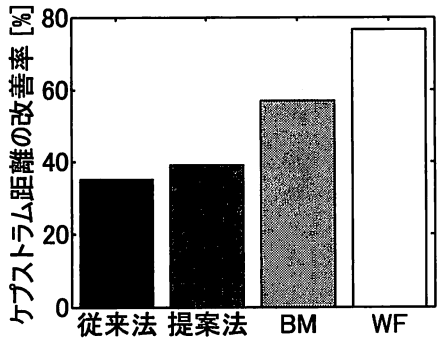

図12　理想的な場合も含めた評価結果

Fig. 12　Result of sound quality evaluation including theoretical limit

謝辞

本研究の一部は総務省戦略的情報通信研究開発推進制度（SCOPE）によるものである．

文　献

[1] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Audio, Speech and Language Processing*, vol. ASSP-27, no. 2, pp.113-120, 1979.

[2] J. S. Lim, and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proc. IEEE*, vol. 67, no. 12, pp.1586-1604, 1979.

[3] R. J. McAulay, and M. L. Malpass, "Speech enhancement using a soft decision noise suppression filter," *IEEE Trans. Audio, Speech and Language Processing*, vol. ASSP-28, no. 2, pp.137-145, 1980.

[4] Y. Ephraim, and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Audio, Speech and Language Processing*, vol. ASSP-32, no. 6, pp.1109-1121, 1984.

[5] M. Aoki, M. Okamoto, S. Aoki, H. Matsui, T. Sakurai, and Y. Kaneda, "Sound source segregation based on estimating incident angle of each frequency component of input signals acquired by multiple microphones," *Acoust. Sci. Tech.*, vol. 22, no. 2, pp.149-157, 2001.

[6] O. Yilmaz, and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. Signal Proc.*, vol. 52, no. 7, pp. 1830-1847, 2004.

[7] 猿渡 洋, "音声・音響信号を対象としたブラインド音源分離," 電子情報通信学会 DSP 研究会 (DSP2001-194), pp.59-66,2002.

[8] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," *Neurocomputing*, vol. 22, no. 1-3, pp. 21-34, 1998.

[9] R. Blouet, G. Raraport, I. Cohen, and C. Fevotte, "Evaluation of several strategies for single sensor speech/music separaton," *ICASSP2008*, pp.37-40, 2008.

[10] A. Ozerov, P. Philippe, F. Bimbot, and R. Gribonval, "Adaptation of bayesian models for single-channel source separation and its applicatio to voice/music separation in popular songs," *IEEE Trans. Audio, Speech and Language Processing*, vol. 15, no. 5, pp.1564-1578, 2007.

[11] D. P. W. Ellis, and R. J. Weiss, "Model-based monaural source separation using a vector-quantized phase-vocoder representation," *ICASSP2006*, pp.957-960, 2006.

[12] 伊藤弘章, 大石康智, 宮島千代美, 北岡教英, 武田一哉, "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," 情報処理学会研究報告, SLP72-10, pp.51-56, 2008.

[13] K. Itou, M. Yamamoto, K. Takeda, T. Takezawa, T. Matuoka, T. Kobayashi, K. Shikano, and S. Itahashi, "JNAS:Japanese speech corpus for large vocabulary continuous speech recognition research," *J. Acoust. Sci. Jpn E*, vol.20, no.3, pp.199-206, 1999.

[14] 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," 情報処理学会 音楽情報科学研究会 研究報告 2002-MUS-45-4, vol.2002, no.40, pp.19-26, 2002.