

# 振幅のみからの相関推定と雑音尖度に基づく 空間サブトラクションアレーの減算係数最適化\*

◎李傑, 宮部滋樹, 小野順貴 (NII/総研大), 山田武志, 牧野昭二 (筑波大)

## 1 はじめに

独立成分分析 (ICA) [1] に代表される、マイクロホンアレーに基づく雑音抑圧処理は、信号の非定常性を許容できる利点がある。しかし、抑圧対象となる雑音が有限個の点音源から発生している場合でなければ、高精度な音声強調を得るためには非常に多くのマイクロホンが求められるという問題がある。また、スペクトル減算 (SS) 法 [2] や最小平均 2 乗誤差短時間振幅スペクトル推定 (MMSE-STSA) [3] に代表される、単一チャンネルの非線形信号処理による音声強調は、音源の位置や形状に依存しない半面、非定常な雑音を抑圧するのが困難であるという問題がある。そこで、ICA と SS 法を効果的に組み合わせることにより、拡散性の非定常雑音を抑圧する空間サブトラクションアレー (SSA) 法が提案されている [4]。SSA 法では非定常雑音の推定のために ICA を用い、これを SS の雑音推定として用いる。しかし、高い音声品質を得るためには、SS 処理の強度を調整する減算係数の最適化が必要で、最適な減算係数は狭帯域ごとに異なるために、これを手動で調整するのは困難である。減算係数の推定としては、ミュージカルノイズフリーという観点から、控え目な雑音抑圧を行う SS 処理の繰り返しによって減算の強度を自動調整する手法が提案されているが [7]、これは狭帯域ごとの減算係数を直接的に最適化するものではない。

本稿では、高品質な出力を得るための減算係数は、音声と雑音の尖度、SNR に強い相関を持っている点に着目し、尖度と SNR の推定を狭帯域ごとに特徴化して回帰分析に施すことにより、帯域ごとの減算係数を直接的に最適化する手法を提案する。更に、ICA の出力から SNR を効果的に推定する手法を提案した。

## 2 従来手法

### 2.1 概要

ここでは音声の音源方位とマイクロホン配置が既知であるという仮定のもとでの SSA 法の処理 [4] について述べる。SSA 法は複数のマイクロホンを用いて非定常雑音を抑圧する。まず ICA と projection back (PB) 処理により、音声と雑音それぞれの多チャンネルの音像を推定し、それぞれを遅延和型ビームフォーマ (DSBF) で単一チャンネル化する。このうちの音声の推定パスは主パスと呼ばれ、ICA では十分な拡散性雑音の抑圧ができないために粗い音声推定が得られる。対して雑音の推定パスは参照パスと呼ばれ、点音源である音声を ICA によって効果的にキャンセルした高精度な雑音推定が得られる。次に、主パスと参照パスの振幅の減算を行う SS 処理によって、主パスの雑音の残留成分を強力に抑圧する。この仕組みにより、ICA による非定常雑音推定性能と SS による拡散性雑音除去性能を両立させることができる。詳細なアル

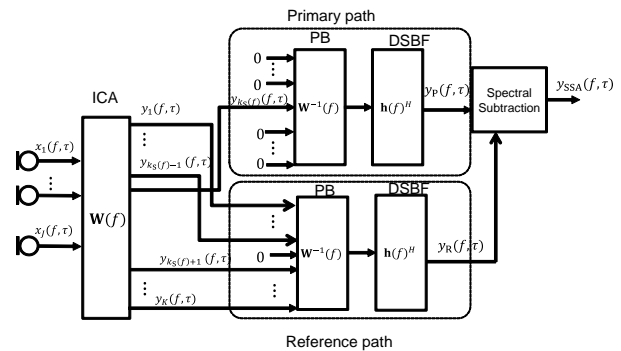


Fig. 1 Block diagram of SSA.

ゴリズムについては以下で説明する。SSA 法の処理の流れは図 1 に示す。

### 2.2 SSA による音声強調

まずマイクロホンアレーで観測される  $J$  チャンネル信号の短時間フーリエ領域表現  $\mathbf{x}(f, \tau)$  が得られているものとするものとして、これを ICA に施すことによって音声の粗い推定と音声を取り除いた雑音の推定に分離する。ICA による観測信号の処理は以下のようなになる。

$$\begin{aligned} \mathbf{o}(f, \tau) &= [o_1(f, \tau), \dots, o_K(f, \tau)]^T \\ &= \mathbf{W}(f)\mathbf{x}(f, \tau) \end{aligned} \quad (1)$$

ここで  $K$  は出力チャンネル数 (ただしここでは  $K = J$  として処理を行う)、 $o_k(f, \tau)$  は第  $k$  番目の分離信号 ( $k = 1, \dots, K$ )、 $\mathbf{W}(f)$  は  $K \times J$  次元の分離行列を表している。ICA による出力同士を独立にする分離行列  $\mathbf{W}(f)$  の最適化には様々な方法が提案されているが、たとえば周波数領域の自然勾配法は以下の更新の繰り返しにより最適化する [1]。

$$\begin{aligned} \mathbf{W}(f) &\leftarrow \mu \left( \mathbf{I} - \left\langle \phi(\mathbf{o}(f, \tau)) \mathbf{o}(f, \tau)^H \right\rangle_{\tau} \right) \mathbf{W}(f) \\ &+ \mathbf{W}(f) \end{aligned} \quad (2)$$

ここで  $\mu$  は更新係数、 $\mathbf{I}$  は  $K$  次元単位行列、 $\langle \cdot \rangle_{\tau}$  は  $\tau$  についての平均、 $\phi(\cdot)$  は非線形関数ベクトルを表す。

ICA の出力から、音声と雑音それぞれの音像  $\mathbf{c}_S(f, \tau)$ 、 $\mathbf{c}_N(f, \tau)$  を分離信号の PB 処理により推定する。

$$\mathbf{c}_S(f, \tau) = \mathbf{W}(f)^{-1} \mathbf{U}(f) \mathbf{o}(f, \tau) \quad (3)$$

$$\mathbf{c}_N(f, \tau) = \mathbf{W}(f)^{-1} \bar{\mathbf{U}}(f) \mathbf{o}(f, \tau) \quad (4)$$

ここで  $\mathbf{U}(f)$  および  $\bar{\mathbf{U}}(f)$  は分離出力  $\mathbf{o}(f, \tau)$  から推定対象成分以外を取り除く  $K$  次元対角行列であり、

\*Optimizing subtraction coefficients of spatial subtraction array based on correlation estimation from amplitude and noise kurtosis, by Jie LI, Shigeki MIYABE (University of Tsukuba), Nobutaka ONO (NII/SOKENDAI), Takeshi YAMADA, Shoji MAKINO (University of Tsukuba)

音声推定に相当するただ一つの出力のインデックスを  $k_s(f)$  として、それぞれ以下のように定義される。

$$\mathbf{U}(f) = [\delta_{ij} \delta_{jk_s(f)}]_{ij} \quad (5)$$

$$\bar{\mathbf{U}}(f) = [\delta_{ij} \bar{\delta}_{jk_s(f)}]_{ij} \quad (6)$$

ここで  $[x]_{ij}$  は第  $i$  行  $j$  列要素が  $x$  となる行列、 $\delta_{ij}$  はクロネッカーのデルタ、 $\bar{\delta}_{ij}$  はクロネッカーのデルタの 0 と 1 を反転させたものを表す。また、 $k_s(f)$  の同定には  $\mathbf{W}(f)$  の行とステアリングベクトル  $\mathbf{h}(f)$  のコサインから音声成分を出力するかどうかを評価するなどによって可能であるが、本稿の実験では真のインデックスを手動で求めて評価した。このようにして音声成分を強調した音声画像推定  $\mathbf{c}_S(f, \tau)$  および音声成分を除去した雑音画像推定  $\mathbf{c}_N(f, \tau)$  を DSBF で処理することにより、最終的な主パスの音声推定  $y_P(f, \tau)$  および参照パスの雑音推定  $y_R(f, \tau)$  を得る。

$$y_P(f, \tau) = \mathbf{h}(f, \tau)^H \mathbf{c}_S(f, \tau) \quad (7)$$

$$y_R(f, \tau) = \mathbf{h}(f, \tau)^H \mathbf{c}_N(f, \tau) \quad (8)$$

ここで  $\mathbf{h}(f)$  は、既知である音声の音源方位とマイクロホン配置から求めた音声のステアリングベクトルである。

次に、参照パスの雑音推定を用いた主パスの音声推定に含まれる残留雑音の抑圧を、SS 法により行う。出力  $y_{SSA}(f, \tau)$  は以下のように表される。

$$y_{SSA}(f, \tau) = \begin{cases} (|y_P(f, \tau)| - \beta(f) |y_R(f, \tau)|) \text{sign}[y_P(f, \tau)] \\ \quad \text{(if } |y_P(f, \tau)| - \beta(f) |y_R(f, \tau)| > 0) \\ \eta y_P(f, \tau) \quad \text{(otherwise)} \end{cases} \quad (9)$$

ここで  $\text{sign}[\cdot]$  は符号を表し、 $\beta(f)$  は減算係数を、 $\eta$  はスペクトル減算の結果が負の値を取る場合に観測信号に掛けるフロアリング係数を表す。なお、本稿の実験ではフロアリング係数を 0 としている。

### 2.3 従来のスペクトル減算の最適化

SSA を含んだスペクトル減算処理では減算係数  $\beta(f)$  の最適化が問題となる。 $\beta(f)$  は大きいほど雑音除去の効果が大きい、しかし除去された成分と残留成分の起伏が激しくなるために、ミュージカルノイズと呼ばれる不快な人工雑音が残留して音声品質を低下させてしまう。最適な減算係数は狭帯域ごとに大きく異なるため、雑音除去と音声品質のトレードオフを  $\beta(f)$  の手動調整により解決するのは難しい。

Uemura *et al.* [5] は、雑音除去による雑音の尖度の変化がミュージカルノイズの知覚と相関があることを見出し、この知見を用いた様々な手法が同研究グループにより提案されている。たとえば Miyazaki *et al.* は、雑音の尖度を変えないスペクトル減算を繰り返すことによりミュージカルノイズの発生しないスペクトル減算 [6] を提案し、これを SSA に適用している [7]。これらの手法では、まず減算係数  $\beta(f)$  を全帯域で共通の小さい値に手動で設定し、与えた  $\beta(f)$  と雑音のモデルから、ミュージカルノイズを発生させないためのフロアリング係数  $\eta$  を推定して SS 処理を行う。小さい減算係数とフロアリングによって雑音抑

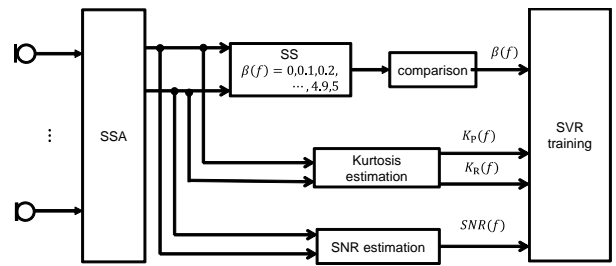


Fig. 2 Block diagram of SVR training.

圧効果は弱められるが、減算を繰り返すにより残留雑音が低減されていく。

この仕組みは、ミュージカルノイズフリーという観点からスペクトル減算の残留雑音抑圧と品質のトレードオフを効果的に解決するものであるが、フロアリングによって抑制された減算の繰り返しは帯域ごとの減算係数を間接的に最適化するのと同様の効果を与えていて、何らかのより直接的なアプローチの方が、副作用が小さくより効果的になる可能性がある。そこで、ミュージカルノイズフリーなアプローチとの比較は今後の課題として残っているが、帯域ごとの減算係数を直接的に最適化する手法を提案する。

## 3 提案手法

### 3.1 提案手法の枠組み

本節では Miyazaki *et al.* のスペクトル減算とは異なった、回帰分析によって帯域ごとの減算係数を直接的に最適化するアプローチを試みる。Uemura *et al.* が示したように、最適なスペクトル減算は尖度などの分布形状との関係が深いといえる。また、雑音と音声特定の分布によってモデル化できる場合には、スペクトル減算の減算係数のような雑音抑圧の強度は SNR の推定によって最適化できる (例えば [3])。SNR と雑音形状の両方に依存するのが減算係数最適化の困難な点であるが、本稿では分布形状と SNR を表す特徴量を帯域ごとに抽出し、この単純に非線形回帰によって音声品質尺度を最大化する係数を求めることにより、帯域ごとの抑圧強度の直接的な最適化が可能になると考えられる。

提案手法は減算係数推定器の教師有り事前学習と、最適減算係数推定を用いた雑音抑圧の 2 つのステージからなる。まず学習の枠組を図 2 に示す。非線形な回帰分析にはサポートベクトル回帰 (SVR) [8] を用いる。帯域ごとの減算係数の最適性の基準は、狭帯域の音声品質を評価できるあらゆる尺度を用いることが可能であるが、ここでは signal-to-distortion ratio (SDR) [9] を帯域ごとに評価して用いる。まず様々な雑音環境の、クリーンな音像を有したコーパスを用意し、離散化した減算係数  $\beta(f)$  の候補から、SSA 処理の SDR が最も大きくなる減算係数を帯域ごとに選択する。ここでは減算係数  $\beta(f)$  は 0, 0.1, ..., 5.0 の合計 51 個を候補とした。次に、これらのコーパスの後述する分布形状特徴 ( $K_P(f)$ ,  $K_R(f)$ ) と SNR の狭帯域特徴  $SNR(f)$  の分析を行う。最後に特徴量から最適な減算係数を推定する SVR を学習する。

抑圧のステージでは、主パスと参照パスの処理を行い、その出力の特徴量を求めて SVR に入力し、SVR から求められた推定最適減算係数  $\beta(f)$  を用いてスペ

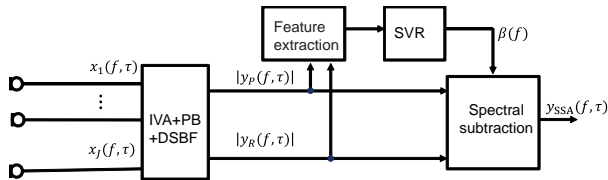


Fig. 3 Block diagram of noise suppression.

クトル減算を行う。雑音抑圧のブロック図を Fig. 3 に示す。

### 3.2 分布形状の特徴分析

まず狭帯域信号の分布形状を特徴化する。これは、何らかの形状パラメタを持つ分布の、最尤推定による形状パラメタによって特徴化する。ここでは分析対象となる信号の狭帯域のパワースペクトログラム  $X(f, \tau)$  が、密度関数が以下の  $f(X(f, \tau); \kappa(f), \theta(f))$  のように与えられるガンマ分布に従うと仮定する。

$$f(X(f, \tau); \kappa(f), \theta(f)) = \frac{X(f, \tau)^{\kappa(f)-1} \exp\left(-\frac{X(f, \tau)}{\theta(f)}\right)}{\Gamma(\kappa(f)) \theta(f)^{\kappa(f)}} \quad (10)$$

ここで  $\kappa(f) (> 0)$  は形状パラメタ、 $\theta(f) (> 0)$  は尺度パラメタを表す。帯域  $f$  ごとの最尤推定 [10] によって  $\kappa(f)$  を求めたのちに、これを分布形状の特徴とする。複素信号のパワーがガンマ分布に従うと仮定すると、複素信号の尖度は形状パラメタに反比例するため、この特徴化は尖度に準ずる特徴抽出であるといえる。このような特徴分析を主パスの出力パワー  $|y_P(f, \tau)|^2$  と参照パスの出力パワー  $|y_R(f, \tau)|^2$  のそれぞれを  $X(f, \tau)$  に代入した  $\kappa(f)$  の最尤推定を行い、それぞれ  $K_P(f)$ ,  $K_R(f)$  として特徴量として用いる。

### 3.3 SNR に準ずる特徴抽出

ここで推定する必要があるのは  $y_P(f, \tau)$  の SNR であり、これをクリーンな音声を与えられていない条件で行わなければならない。音声と雑音は無相関であると仮定できるため、仮に  $y_P(f, \tau)$  に含まれる雑音成分と推定雑音  $y_R(f, \tau)$  の相関係数が 1 であった場合には、 $y_P(f, \tau)$  と  $y_R(f, \tau)$  の相関係数から SNR を推定することができる。しかし ICA は出力を無相関化するため、これらの信号間の相関係数はおよそ 0 となり、このような SNR 推定は成り立たない。そのため無相関化の影響を受けない相関係数の推定を考える必要がある。

SSA の後段処理のスペクトル減算が有効であるのは、ICA によって  $y_P(f, \tau)$  と  $y_R(f, \tau)$  が無相関化されているにも拘わらず、両チャンネルに混入する雑音成分の振幅増減が十分類似しているからであるといえる。そこで、一旦相関を推定する 2 つの信号の位相を捨ててしまい、位相の失われた絶対値振幅  $|y_P(f, \tau)|$  と  $|y_R(f, \tau)|$  から、元の複素信号  $y_P(f, \tau)$  と  $y_R(f, \tau)$  の相関係数を推定することを考える。このような位相が失われた条件での複素信号の相関係数推定として、我々は元の信号が共に複素正規分布に従うと仮定し、位相を隠れた変数とした最尤推定を提案している [11]。これを用いた相関係数推定は、ICA による無相関化が位相のみに影響を与えるわけではないために、残留雑音に対応するものと一致する推定とはならない。

しかし、この最尤推定の手がかりとなるのは振幅の増減の類似性のみであり、同じ振幅の増減の類似性を利用した SSA に適した相関係数推定になると期待できる。

この最尤推定は EM アルゴリズムにより達成される。最初に  $y_P(f, \tau)$  と  $y_R(f, \tau)$  の分散  $\sigma_P(f)^2$ ,  $\sigma_R(f)^2$  を求め、更に  $y_P(f, \tau)$  と  $y_R(f, \tau)$  の相関係数  $|\rho(f)|$  に 0 から 1 の範囲内の初期値を与える。次に、E ステップの更新

$$\lambda(f, \tau) \leftarrow \frac{I_1\left(\frac{2|\rho(f)||y_P(f, \tau)||y_R(f, \tau)|}{\sigma_P(f)^2 \sigma_R(f)^2 (1-|\rho(f)|^2)}\right)}{I_0\left(\frac{2|\rho(f)||y_P(f, \tau)||y_R(f, \tau)|}{\sigma_P(f)^2 \sigma_R(f)^2 (1-|\rho(f)|^2)}\right)} \quad (11)$$

と M ステップの更新

$$|\rho(f)| \leftarrow \frac{\langle |y_P(f, \tau)| |y_R(f, \tau)| \lambda(f, \tau) \rangle_\tau}{\sigma_P(f) \sigma_R(f)} \quad (12)$$

を繰り返す。ここで  $I_\nu(\cdot)$  は次数  $\nu$  の第 1 種修正ベッセル関数である。このようにして得られる  $|\rho(f)|$  は、 $y_P(f, \tau)$  に含まれる雑音成分と  $y_R(f, \tau)$  の相関係数の推定であり、両雑音成分の相関係数が 1 であることを改めて仮定した  $y_P(f, \tau)$  の SNR [dB] の推定という形で特徴量  $SNR(f)$  を以下のように得る。

$$SNR(f) = 10 \log_{10} \frac{1 - |\rho(f)|^2}{|\rho(f)|^2} \quad (13)$$

この SNR の推定には複数の仮定が入っているため、SNR を高精度に推定するものではないが、汎化性能の高い SVR に入力する SNR に準ずる特徴抽出としては、十分効果的なものであると期待される。

## 4 実験

### 4.1 実験条件

提案手法による減算係数最適化の有効性を実験により評価する。まず、減算係数推定モデルを作るために、大量の混合音声を準備してシミュレーション実験を行った。鏡像法により複数の残響時間の 2 チャンルの室内インパルス応答を作成し、複数の SNR と種類の拡散性雑音を重畳した。音声は ATR コーパスから選択し、学習データの雑音のコーパスは SiSEC 2011 [12] の Two-channel mixtures of speech and real-world background noise タスクの開発データから選択した Cafeteria, Subway, Square noise を使用し、テストデータの雑音はゲームセンターと地下鉄で新しく収録した実環境雑音を用いた。テストはオープンテストである。ICA の学習則は補助関数法に基づく独立ベクトル分析 [13] を用いた。観測信号はどれもサンプリング周波数 16 kHz の 3 秒の信号で、フレーム分析は窓幅 4096 サンプル、フレームシフト 1024 サンプルのハミング窓を用いた。SVR は、RBF カーネルを用いた  $\nu = 0.5$  の  $\nu$ -SVR を用いた [14]。学習とテストの各条件を表 1 を示す。

比較手法として、SSA の減算係数を複数の方法で最適化したものを評価した。まず未処理の観測信号 (Unprocessed)、IVA と DSFB のみによる音声推定 (Fixed IVA only)、 $\beta(f) = 3$  に固定した SSA 処理 (Fixed coefficient  $\beta(f) = 3$ )、SNR 推定  $SNR(f)$  のみを持つ

Table 1 Experimental conditions

	Training	Test
Room size	5 × 4 × 6 m	7 × 5 × 4 m
Reverberation time $T_{60}$	0.4, 0.8, 1.2, 1.6, 2.0 s	0.3, 0.5, 0.7 s
Number of speech utterances	4	2
Microphone interval	0.086 m	0.05 m
SNR	0, 10, 20, 30 dB	-10, 0, 10 dB
Speaker-microphone distance	1 m	1 m
Speaker directions	7 directions -90°, -60°, ..., 90°	3 directions 10°, 40°, 70°

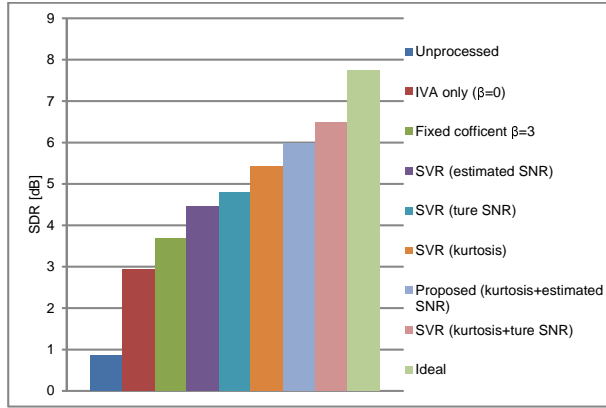


Fig. 4 Experiment result

微量とした SVR による減算係数最適化 (SVR (estimated SNR))、 $y_P(f, \tau)$  の真の SNR を手動で与えて特徴量とした SVR (SVR (true SNR))、尖度特徴量  $K_P(f)$ ,  $K_R(f)$  を特徴量とした SVR (SVR (kurtosis))、尖度と推定 SNR を特徴量とした提案手法 (Proposed (kurtosis+estimated SNR))、尖度と真の SNR を特徴量とした SVR (SVR (kurtosis+true SNR))、狭帯域 SDR を最大化する理想的な減算係数による SSA (Ideal) を評価した。評価尺度には目的音対歪比 (SDR) [9] を用いた。

#### 4.2 実験結果

各条件における雑音抑圧の結果を、全ての条件の平均を図 4 に、SNR ごとの結果を図 5 に示す。まず、図 4 について議論する。SVR を用いた全ての手法が固定の減算係数による処理よりも品質が高いため、帯域ごとの減算係数の推定が重要であることがわかる。また、SNR と尖度の特徴の組み合わせは、SNR または尖度の特徴のどちらか片方のみを用いた場合よりも精度が高く、これらの特徴量の組み合わせが効果的であることがわかる。最尤推定による推定 SNR を用いた SVR 学習は、やはり真の SNR を与えた場合よりわずかに性能が劣るものの、減算係数推定のための特徴量として有効に機能しており、また尖度のみを特徴とした推定を大きく改善させていることがわかる。同様の関係は、図 5 の全ての SNR について保たれているため、様々な雑音・残響環境のデータを学習することによって、雑音抑圧を行う環境に適した減算係数推定が SVR の学習により効果的に行えていることがわかる。以上より、提案する SNR 推定と尖度推定を特徴量とした SVR による減算係数推定の有効性が確認された。

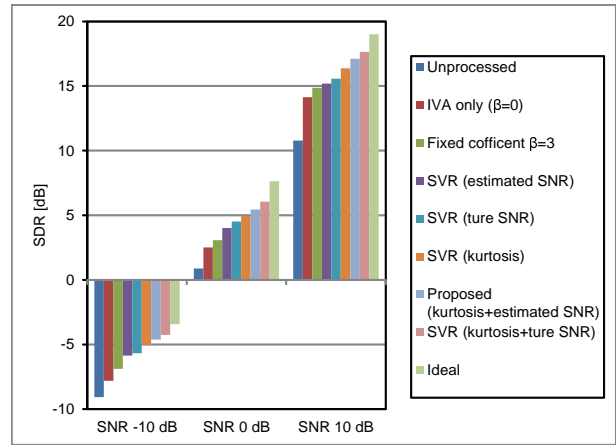


Fig. 5 Experimental result for each SNR.

## 5 おわりに

本稿では、SSA 法の狭帯域ごとの最適な減算係数推定をするために、SNR と尖度の推定を特徴量とした SVR を用いる手法を提案した。また、ICA の無相関化のために、参照パスの雑音推定から主パスに含まれる雑音の SNR を推定するのは困難であるが、位相を捨てた振幅のみからの最尤推定により、SNR に準ずる特徴量の抽出を提案した。評価実験により、尖度と SNR の推定が特徴抽出として有効であり、効果的な減算係数の回帰が行えることを確認した。今後の課題はミュージカルフリーな SSA 法との性能比較である。

## 参考文献

- [1] Makino *et al.*, *Blind Speech Separation*, 2007.
- [2] Boll, *IEEE Trans. ASSP*, 27(2), 113–120, 1979.
- [3] Ephraim, Malah, *IEEE Trans. ASSP*, 32(6), 1109–1121, 1984.
- [4] Takahashi *et al.*, *IEEE Trans. Audio Speech Lang. Process.*, 17(4), 650–664, 2009..
- [5] Uemura *et al.*, *Proc. IWAENC*, 2008.
- [6] Miyazaki *et al.*, *IEEE Trans. Audio Speech Lang. Process.*, 20(7), 2080–2094, 2012.
- [7] Miyazaki *et al.*, *Signal Process.*, 102, 226–239, 2014.
- [8] Smola *et al.*, *Statist.& Comput.*, 14(3), 199–222, 2004.
- [9] Vincent *et al.*, *IEEE Trans. Audio Speech Lang. Process.*, 14(4), 1462–1469, 2006.
- [10] Minka, *Microsoft Tech. Rep.*, 2002.
- [11] Miyabe *et al.*, *Proc. LVA/ICA*, 2015.
- [12] Araki *et al.*, *Proc. LVA/ICA*, 414–422, 2012.
- [13] Ono, *Proc. WASPAA*, 189–192, 2011.
- [14] Chang, Lin, *ACM Trans. Intell. Syst. Technol.*, 2(3), 27:1–27:27, 2011.