Vol.2012-SLP-92 No.12
2012/7/21

帯域に応じた位相差判定閾値に基づく音源分離法 SAFIA による機械雑音下音声認識

徳竹啓佑† 川端豪†

雑音抑圧 SAFIA コムフィルタ 音声認識

SAFIA は音声の倍音構造を取り出し音源分離を行う手法である．しかし，ロボットの厳しい雑音下では SAFIA は音声の重要な成分を除去してしまう可能性がある．本研究では帯域ごとに位相差を扱うことで，音声の倍音成分を保存する．帯域毎に観測位相差と音声の入射角度から算出した位相差の理論値を比較する．提案手法ではこの値から境界周波数を定め，音声成分の保護を行う帯域を選ぶ．また，コムフィルタにより音声の倍音成分を強調した．実験の結果単語誤り率が約 20%改善した．

Speech Recognition under the Mechanical Noise Condition using SAFIA with Band-Sensitive Phase Treatments

KEISUKE TOKUTAKE† TAKESHI KAWABATA†

Noise Reduction SAFIA Comb-Filter Speech Recognition

The sound segregation method SAFIA is a promising technique picking up the speech harmonics. However, for the severe condition with the mechanical body noise of robots, SAFIA sometimes reduces the important harmonics elements of speech recognition. This paper describes the method to improve the SAFIA for protecting the speech harmonics based on the band-sensitive phase treatment. The difference of input channel phases is compared with the theoretical value of the correct incoming direction. The system selects the spectral channel which detects smaller errors than the band-senstive threshold. Also the system emphasize the speech harmonics using the pitch synchronous Comb-Filter. Recognition experiments show that the proposed method improves the WER with about 20%.

1. はじめに

近年のロボット研究の発展に伴い，ロボットと人間が共に生活できることが期待されている．そのためにはロボットと人間のコミュニケーション手段が必要であり，その一つの形態として音声による対話がある．ロボットとの音声対話には，ロボットに人間の対話内容を理解させる必要があり，高精度の音声認識技術が必要不可欠となる．しかし，音声認識を行うにあたりロボット自身が発する機械雑音のため，音声認識の精度を著しく下げてしまうという問題点があり，雑音に頑健なシステムの構築が求められている．

この問題に対し，雑音除去という観点からではマイクロホンの入力信号に対して何らかの処理を施す必要がある．単一マイクロホンの入力に基づく，例としてスペクトルサブトラクション法がある[1]．スペクトルサブトラクションは雑音の入った音声信号から雑音信号を減算することで，目的信号を得るという手法である．この手法は定常性の雑音に対して高い効果を得ることができるが，ロボットの機械雑音のような非定常性雑音では，実際の雑音と推定された雑音の差異により雑音の引きすぎ，引き残しが発生し結果として音声認識精度が向上しない場合がある．

† 関西学院大学
Kwansei Gakuin University

一方で多チャンネル信号処理としてマイクロフォンアレイを用いたアレー信号処理[2]や，音源の独立性に基づく ICA(独立成分分析)[3]などが提案されている．アレー信号処理は指向特性を形成することで音源分離を実現する技術であり，音源の空間的位置の違いによって，非目的音を効率的に抑圧する技術である．しかし，高精度な音声強調を行うには大規模な装置が必要になるという問題点がある．ICA は原信号が統計的独立という課程のもとで，混合信号から原信号を推定する統計的手法である．ICA は音源を事前情報なしで分離することを目的としているため，音源の種類によらず分離を行うことができるという利点がある．一方で演算量が多くリアルタイム処理が困難であるといった欠点がある．

一方，音源位置による複数マイク間のパワー比または位相差を利用して音源分離を実現する青木の SAFIA(sound source Segregation based on estimating incident Angle of each Frequency component of Input signals Acquired by multiple microphones)という手法がある[4]．SAFIA は少数マイクで実現でき，また少ない演算量で音源分離を行える利点がある．本研究では，機械雑音下での音声認識の性能を向上させるために，音源分離法 SAFIA の利用を検討した．しかし，高レベルの雑音状況下では音源分離の課程で音声成分が除去されてしまうという問題がある．この問題に対し，低帯域における位相差判定の緩和，音声の倍音成分の強調によ

り認識精度の向上を図った．

2. 機械雑音下音声認識に対する SAFIA の利用

青木らが提案した SAFIA は音源分離，方向同定の分野において有効な手法であり SN 比の悪い環境においても高い音源分離[4]，方向同定性能[5]が報告されている．本節では音源分離法 SAFIA の音声認識へ利用を検討する．

2.1 音源分離法 SAFIA

青木らは，複数マイクを用いて特定の音源信号を取り出す手法として，SAFIA(sound source Segregation based on estimating incident Angle of each Frequency component of Input signals Acquired by multiple microphones)を提案した．SAFIA における信号の流れを図 1 に示す．単一の目的音源 S_1 と単一の雑音源 S_2，及び 2 つのマイクロホン 1 とマイクロホン 2 が配置されている場合を考える．目的音源はマイクロホン 1 に近くに配置されているとする．ここで，目的音と雑音(不要音)は，音声のように調波構造を持った信号であると仮定する．2 チャンネル入力された信号 $x_1(n)$ と $x_2(n)$ に対して離散フーリエ変換により周波数分析する．各周波数 ω における周波数成分を $X_1(\omega)$ と $X_2(\omega)$ とする．到達位相差・到達レベル差の算出部において，式に定義されるチャネル間到達位相差 $\Delta\phi(\omega)$ 及び到達レベル差 $\Delta A(\omega)$ を算出する．

$$\Delta\phi(\omega)=\arg(X_1(\omega))-\arg(X_2(\omega)) \tag{1}$$

$$\Delta A(\omega)=20\log_{10}\left(\frac{|X_1(\omega)|}{|X_2(\omega)|}\right) \tag{2}$$

目的音および雑音の調波構造がスパースであれば，多くの周波数成分において目的音と雑音のいずれかが支配的な成分となる．このため，各周波数では支配的な成分となる単独の音源に基づくマイク間の位相差及びレベル差が観測される．これらの値を基に，各周波数成分が目的音，あるいは雑音のどちらに属するかを判定することができる．図 1 のような音源配置においては，$X_1(\omega)$ に含まれる目的音成分のレベルは，$X_2(\omega)$ に含まれるものより大きく，位相も変化する．

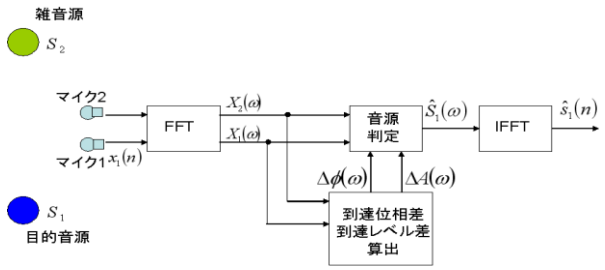


図 1 SAFIA の動作

そのため判定部で，$\Delta A(\omega)$ 及び $\Delta\phi(\omega)$ が正である帯域は目的音が支配的である成分であると判定できる．逆に $\Delta A(\omega)$ 及び $\Delta\phi(\omega)$ が負である帯域は雑音が支配的周波数成分であると判定できる．$\Delta A(\omega)$ による判定式を式(3)に，$\Delta\phi(\omega)$ による判定式を式(4)に示す．

$$\left.\begin{aligned}&\hat{S}_1(\omega)=X_1(\omega),\quad \hat{S}_2(\omega)=0,\quad (\Delta A(\omega)\geq 0)\\&\hat{S}_1(\omega)=0,\quad \hat{S}_2(\omega)=X_2(\omega),\quad (\Delta A(\omega)\leq 0)\end{aligned}\right\} \tag{3}$$

$$\left.\begin{aligned}&\hat{S}_1(\omega)=X_1(\omega),\quad \hat{S}_2(\omega)=0,\quad (\Delta\phi(\omega)\geq 0)\\&\hat{S}_1(\omega)=0,\quad \hat{S}_2(\omega)=X_2(\omega),\quad (\Delta\phi(\omega)\leq 0)\end{aligned}\right\} \tag{4}$$

式(3)，式(4)によって得られた目的音成分の推定値の各周波数成分 $\hat{S}_1(\omega)$ に対し逆フーリエ変換を施し，時間領域の目的信号 $\hat{s}_1(n)$ を復元する．このように SAFIA は音源の位置に基づく特徴量を用いて，特定領域内にある音源のみを抽出する．

本研究では，無指向性マイクを用い判定には到達位相差を用いる．事前検討の結果，無指向性マイク用いた場合，2 チャネル間の到達レベル差の分布は真の分布より大きく外れてしまった．一方で到達位相差は，真の分布との差異が少ないため，到達位相差を用いた．

2.2 低帯域における到達位相差のばらつき

SAFIA は各 ω に対し音声成分と雑音成分のどちらが主成分になっているかを位相差に基づき判定し音声を復元する手法である．そのため周波数 ω の変化に対し位相差 $\Delta\phi(\omega)$ が一定の傾向を持って変化していれば，判定が容易になり精度の良い主成分判定が行える．しかし，現実のデータではもっと複雑な問題が観測される．図 2 に 2 チャネルマイクに対して角度 30 度，距離 1m に音源を配置して観測した信号の位相差を示す．図 2 の位相差では，楕円で囲んだ低帯域において位相差のばらつきがみられることが分かる．図のような位相差では低帯域において誤判定がおき，雑音成分と判断される帯域が多くなる．

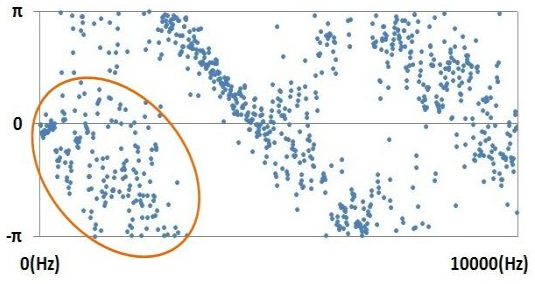
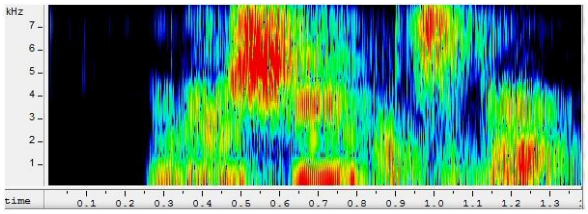
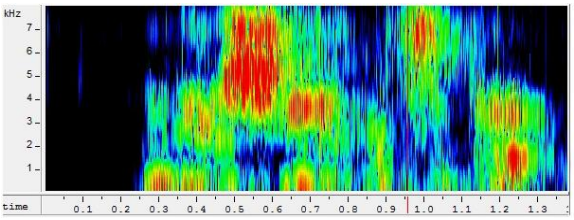


図 2 音声の各周波数成分に対する 2 チャネル間位相差



(a)音声「うれしいはずが」



(b)SAFIA 処理後の音声「うれしいはずが」

図 3　スペクトログラムの比較

図 3 に SAFIA 処理によるスペクトログラムの比較を示す．(a)に音声のスペクトログラム，(b)に SAFIA 処理後のスペクトログラムを示す．図の 0.7 秒付近,0～2kHz の帯域に注目すると音声のパワーが失われていることが分かる．そのため，音声認識精度に重大な支障をきたす可能性がある．

2.3 改良のための着眼点

ここでは，SAFIA 処理の過程で失われる音声成分の保護について述べる．先にも述べたが，低帯域の位相差のばらつきにより，音声成分の除去が生じる．一方でそれ以外の帯域では規則性を持った位相差構造を持つため，誤判定が起こりづらい．そのため，低帯域と信頼できない高帯域を分け SAFIA を用いることで音声認識率が向上する可能性がある．そこで，以下の 2 点の検討を行った．

(1)　低帯域において位相差判定閾値の緩く設定の緩和
(2)　コムフィルタによる音声の倍音成分の強調

以上の 2 点により音声成分の保護・強調を行い，音声認識精度の向上を試みた．

3.　帯域に応じた位相差判定閾値に基づく音源分離法 SAFIA による機械雑音下音声認識

本節では，SAFIA 処理において音声認識に有効な帯域成分が過度に除去されることがないように改良した手法について説明する．

3.1 帯域制限による閾値設定

SAFIA においては入力 x_1 及び x_2 の各周波数 ω に対する位相差 $\Delta\phi(\omega)$ に閾値を設定し，その ω に対し音声と雑音のどちらが支配的な成分であるか決定するが，前述の通り低周波数においては ω と $\Delta\phi(\omega)$ の分布が大きく広がり，適切な判定が困難になる．多数のデータについて周波数 ω と位相差 $\Delta\phi(\omega)$ の分布を観察したところ，前記の位相差が大きくばらつく領域は，比較的低周波数に集中することが分かった．そこで入力データに対してある閾値を定め，この周波数より低い領域では SAFIA において音声と雑音の判定を行う閾値の幅を 2 倍に拡大することを考えた．境界周波数は以下の様に定める．　ある周波数 ω_i に対し帯域幅 n 点の範囲で位相差の理論値と観測値の平均二乗誤差を算出する．

$$g_i = \frac{1}{n}\sum_{j=-\frac{n}{2}}^{j=\frac{n}{2}} (\Delta\phi(\omega_{i+j}) - \Delta\phi_0(\omega_{i+j}))^2 \qquad (5)$$

ここで $\Delta\phi$ は観測された位相差，$\Delta\phi_0$ は音声の到達角度から計算される位相差の理論値である．今回はサンプリング周波数 48kHz，分析窓 4096 点で分析した音声のスペクトルに対して，帯域幅 50 点の範囲で分析を行った．g_i は位相差のばらつきが少ないほど小さな値をとる．音声に対し雑音 5dB が重畳したデータの g_i を下の図 4 に示す．図 4 のような低域から高域に向けて，徐々に値が小さくなり，ばらつきが減っていく．そこで実線で示す閾値を定め，境界周波数を決定した．この例では破線で示す 2.2Hz が境界周波数となる．また，境界周波数の最大値は 4kHz とした．

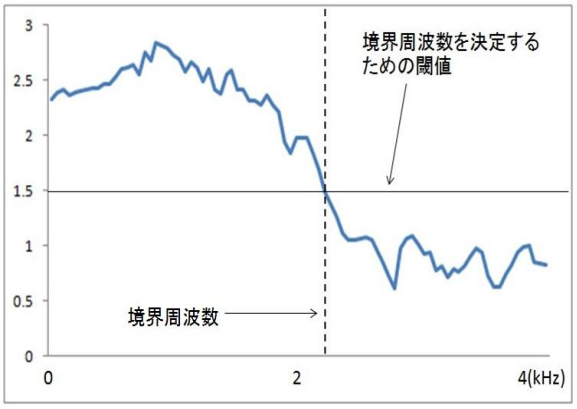


図 4 音声「うれしいはずだ」の時刻 0.7 秒における到達位相差と理論値との平均二乗誤差

3.2 コムフィルタによる音声の調波構造の強調

前項で保護された音声成分をさらに強調することを検討する．基本周波数の整数倍の高調波すなわち倍音を強調することで音声認識率が上がる可能性がある[6]．そこで，図 5 に示すようなコムフィルタ(くし型フィルタ)をかける事で音声の強調を行う．式(6)にコムフィルタの式を示す．通常の SAFIA では音声の低域成分が失われるため，倍音の強調は難しいが，先に述べた手法により音声の低域成分が保護されているため，コムフィルタが有効だと考えられる．

$$F(\omega,\omega_0)=\sqrt{(1+\alpha^2)+2\alpha\cos(2\pi\omega/\omega_0)} \qquad (6)$$

コムフィルタは基本周波数 ω_0 を推定しその整数倍の成分を強調する．以下の式により基本周波数の推定を行った．

$$\hat{\omega}_0=\arg\max_{\omega_0}\sum_{\omega}X(\omega)F(\omega,\omega_0) \qquad (7)$$

ここで $X(\omega)$ は入力信号，$F(\omega,\omega_0)$ はコムフィルタを表す．

つまり様々な基本周波数を持つコムフィルタをかけその中で最もスペクトル構造を保存できるコムフィルタの基本周波数を推定値として利用する．

4. 連続音声認識による提案手法の有効性評価

本章では帯域に応じた位相差判定閾値に基づき処理した SAFIA のデータを使い，連続音声認識に対する有効性の評価を行う．

4.1 実験条件

学習データとしては日本音響学会研究用連続音声データベース(ASJ)を用いた．声データは ATR 音素バランス文(503 文)を 64(男 30 名，女 34 名)の話者が発声した約 9600 文の音声を用いた．音声は 48kHz，16 ビットでデジタル化されたデータを用いた．実験には市販のトイロボットを利用した．

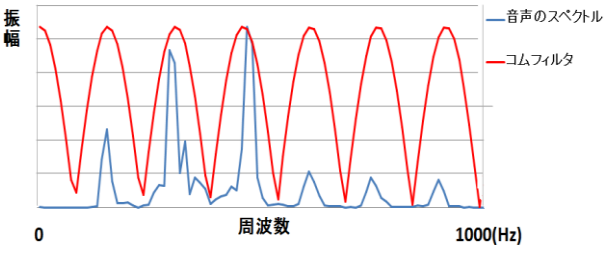


図 5　コムフィルタ

このロボットの左肩と右肩に無指向性マイクをそれぞれ装着する．図 6 に示すように，ロボットに対して距離 1m，角度-90°，-60°，-30°，0°，30°，60°，90°の方向から音声を入射した．

2 チャンネルで録音したデータに対して雑音データを、SN 比 10dB, 5dB, 0dB となるように付加し SAFIA 処理を行う．このデータを使い HMM で音響モデルを作成し，音声認識の評価を行う．音響モデルとしてはトライフォンを使用した．音声デコーダは Julius を用いる．その他の実験条件に関しては表 1 を示す．

評価データとしては学習データと同じように 2 チャンネルで録音したデータに対して録音したデータに雑音を付加させ，学習データと同条件の SAFIA 処理をしたデータで行う．評価方法に交差確認法を用いる．交差確認法とは図 7 のように全データをいくつかに分割してそのうちの 1 つを評価データ，その他を学習データとして用いる．今回の実験では ASJ データを 9 分割(A～I セットに分割)して，それぞれを評価用・学習用に分けて評価を行う．

表 1　実験条件

特徴量	MFCC，デルタ項，各対数パワー 計 26 次元
分析窓	Hamming 窓
分析窓長	25ms
フレーム周期	10ms

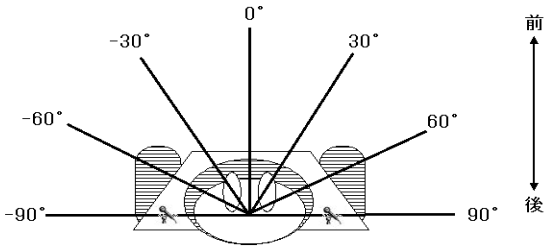


図 6　実験における音声の入射角度

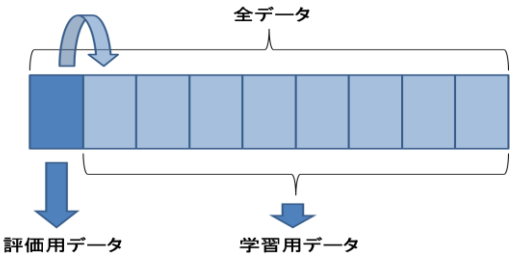


図 7　交差確認法

4.2 認識結果と考察

認識結果を表 2 に示す．結果は単語正解率で評価を行った．横方向は SN 比を示しており雑音の少ないほうから 10dB, 5dB, 0dB となっている．「SAFIA」は通常の SAFIA，「帯域制限 SAFIA」は 3.1 の項で示した手法を用いて低域成分を保護した SAFIA,「帯域制限 SAFIA＋コムフィルタ」は帯域制限 SAFIA に対してコムフィルタをかけたデータである．

SAFIA の位相判定閾値制限による効果を考察する．表 2(c)に示した，角度 30°の認識結果に注目すると SN 比 10dB で「SAFIA」では「処理なし」と比べ，11.44%認識率が向上した．対して「帯域制限 SAFIA」において認識率は「処理なし」と比べ 15.74%の改善がみられた．帯域に応じて位相差判定閾値を定めることで音声成分を保護でき，認識率が向上したと考えられる．更に「帯域制限 SAFIA+コムフィルタ」では「処理なし」と比べ，18.35%の認識率の改善がみられ，低域の音声成分を保護し強調することで認識率を更に向上できることが分かった．SN 比を 5dB に下げると,「SAFIA」では認識率の向上が 7.1%にとどまった．一方で「帯域制限 SAFIA」では 17.35%,「帯域制限 SAFIA＋コムフィルタ」では 19.75%の認識精度の向上がみられ，雑音が多い場合でも安定して認識率の向上がみられた．更に SN 比が 0dB という環境下の中でも「SAFIA」の認識率 6.94%に対して「帯域制限 SAFIA」では 11.67%の認識率の改善を実現できた．

同様に，いずれの角度の場合であっても「SAFIA」と比べ認識精度が向上していることが分かる．これにより低帯域の閾値を制限することにより,「SAFIA」では除去してしまっていた低域の音声成分を保護することができ，音声認識に有効であることが判った．

5. 結論

機械雑音下での音声認識の性能を向上させるために，音源分離手法 SAFIA の利用を検討した. 高レベルの機械雑音が混入した音声に SAFIA を適用すると, 特に低周波数帯域において，雑音が支配的になる周波数チャネルが増え音声認識に必要な音声成分が過度に除去されるという問題が起きるため以下の 2 点の検討を行った．

(1)低帯域における位相差判定の閾値の緩和
(2)コムフィルタによる音声の倍音成分の強調

連続音声認識実験によって評価を行った結果，SN 比 10dB の少ない雑音下では SAFIA を特に手当てなく利用しても効果があるが，SN 比が劣化すると向上が減少することが分かった．これに対し低周波数の帯域に応じた位相差の閾値設定とコムフィルタによって音声成分を保護するという考えに基づいた提案法では認識率の減少を抑え安定した性能が得られた．

参考文献

[1]J.チェン, K.K.パリワル, 松井知子: 長時間パワースペクトル減算による雑音下音声認識,信学技報, SP2000-77(2000)
[2]岡本拓磨, 岩谷幸雄, 鈴木陽: 包囲型マイクロホンアレイを用いた音源放射指向特性抽出に関する基礎的検討, 信学技法報, EA109(166), p31-36(2009)
[3]高橋 祐, 高谷 智哉, 猿渡 洋, 鹿野 清宏: 独立成分分析に基づく空間的サブトラクション，EA106(125), p13-18(2006)
[4］Mariko Aoki, Manabu Okamoto, Shigeaki Aoki, Hiroyuki Matsui, Tetsuma Sakurai, Yutaka Kaneda: Sound source segregation based on estimating incident angle of each frequency component of input signals acquired by multiple microphones, Acounst,& Tech, 22, 2, pp. 149-157(2001)
[5]川野恵右, 春木智貴, 川端豪: 音源分離法 SAFIA を用いたロボット動作雑音中の話者方向判定，SLP2010-082, pp.1-6(2010)
[6]Jae S.Lim, Aln V. Oppenheim, Louis D.Braida: Evaluation of Adaptive Comb Filtering Method for Enhancing Speech Degraded by White Noise Addition, Transaction On Acoustics, Speech, And Signal Processing, ASSP-26, No4(1978)

表 2　連続音声認識による評価(単語正解率)

手法＼SN 比	10dB	5dB	0dB
処理なし	70.7	55.27	43.38
SAFIA	69.9	52.68	42.19
帯域制限 SAFIA	72.54	57.55	48.34
帯域制限 SAFIA＋コムフィルタ	73.38	62.3	54.23

(a)　角度 90°

手法＼SN 比	10dB	5dB	0dB
処理なし	61.44	46.07	43.23
SAFIA	66.95	51.07	44.6
帯域制限 SAFIA	67.44	53.8	48.31
帯域制限 SAFIA＋コムフィルタ	72.62	60.35	54.45

角度-90°

手法＼SN 比	10dB	5dB	0dB
処理なし	69.01	59.48	46.57
SAFIA	74.51	63.85	49.11
帯域制限 SAFIA	80.58	70.71	55.24
帯域制限 SAFIA＋コムフィルタ	82.31	72.28	60.07

(b) 角度 60°

手法＼SN 比	10dB	5dB	0dB
処理なし	71.45	57.33	41.42
SAFIA	77.57	64.5	52.68
帯域制限 SAFIA	82.01	71.47	58.34
帯域制限 SAFIA＋コムフィルタ	84.4	75.15	61.2

(e) 角度-60°

手法＼SN 比	10dB	5dB	0dB
処理なし	67.23	56.45	43.36
SAFIA	78.67	63.55	47.3
帯域制限 SAFIA	82.97	73.8	52.1
帯域制限 SAFIA＋コムフィルタ	85.58	76.2	58.1

(c) 角度 30°

手法＼SN 比	10dB	5dB	0dB
処理なし	68.24	56.28	46.23
SAFIA	75.65	67.24	49.76
帯域制限 SAFIA	79.53	72.1	56.21
帯域制限 SAFIA＋コムフィルタ	83.47	75.48	63.46

(f) 角度-30°

手法＼SN 比	10dB	5dB	0dB
処理なし	70.45	57.2	40.09
SAFIA	75.56	64.45	44.51
帯域制限 SAFIA	79.12	70.32	50.87
帯域制限 SAFIA＋コムフィルタ	82.74	74.56	55.49

(d) 角度 0°