実環境下での音源定位・音源検出の検討

○高橋 徹 (京都大学)　中臺 一博 (東工大/HRI-JP)　石井 Carlos 寿憲 (ATR-IRC)
Jani Even(ATR-IRC)　奥乃 博 (京都大学)

1. はじめに

一般的な会話シーンから，いつ，どこで，誰が発話を行ったかという情報を抽出することは，会議録作成，ライフログ，会話分析研究といった分野で有用である．実際に，speaker diarization，conversational scene analysis といった分野で研究が行われている [1]．また,「よい聞き手ロボット」構築し，認知科学，学習科学の分野に役立て，新しい学術領域の確立を目指す新学術領域「人ロボット共生学」[1]においても，人ロボットインタラクションを通じて学習過程を分析する上で，上述の発話解析は重要な役割を果たす．

このような技術を考える際に，２つの課題を考慮する必要がある．一つは，会話を行う環境における雑音の問題であり，もう一つは実データを扱う際に特有の問題である．前者に関しては，方向性雑音，拡散性雑音，残響，ロボット自己雑音といった実環境での様々雑音を同時に扱う問題として，我々はロボット聴覚分野で研究を行っている．特に，他の話者が会話に割り込むバージインや複数人が同時に発話する同時発話を認識する技術を報告してきた．さらに，我々が培った技術をオープンソースのロボット聴覚ソフトウェア HARK[2] として公開している [2].

一方，後者は，処理のロバスト性を扱う問題であり，想定外の状況にどれだけ対処できるかを扱う問題と捉えることができる．例えば，前者の雑音の課題をある程度解決し，ベンチマークテストで 90%以上の認識率や検出率を達成したとしても，いざ実データを扱うとベンチマークとはかけ離れた低い性能しか出ないことが多い．これは，ベンチマークでは想定されていない複数の要素が，実はトータルな性能に大きな影響を持ち，こうした要素を事前に想定することは難しいことを示している．従って，実データでの性能向上のためには，従来研究で用いられるベンチマークテストだけでは不十分であり，実データを用いた評価が必要である．例えば，AMI プロジェクト[3] では，100 時間にも及ぶ，会議のマルチモーダルデータ AMI Meeing Corpus として提供しており，実データで有効性を検証するために様々な研究に利用されている．

HARK では，実環境を考慮した性能向上の検討が始まっているものの [3, 4, 5]，従来法との比較が行いやすいという観点で，特定のコーパスを用いたベンチマークテストを行うことが多かった．しかし，ベンチマークテストによる評価では，前述のように，２つめの課題に対応することが難しい．本稿では,「人ロボット共生学」で想定している教室内の先生や生徒の自由発話（複数話者発話）の発話分析に適用可能な技術の確立を目指し，音源定位・音源抽出機能に着目して，HARK の実データに対する有効性，実データを扱う際の問題点の抽出を行う．また，マイクロホンアレイには，ATR で開発中の 16 チャネルマイクロホンアレイを用いた．

2. HARK による実データ処理の準備

2·1 実データ収録

データ収録は，図 1 に示す 16 ch のマイクロホンアレイを図 2 に示す部屋に置かれた高さ 0.72 [m] の机の上に設置して行った．部屋の大きさは，$5.3 \times 4.7 \times 2.3$ [m] で残響時間 RT_{20} は約 260 [ms] である．マイクロホンアレイのマイクロホンレイアウトは，図 1a)-c) の３種類を試した．図 2 の座席に男子学生 5 名に座ってもらい，2 分間の自由発話を 16 ch の音響信号として収録した．各座席は，マイクロホンアレイの中心からみて，22.5°, 157.5°, -22.5°, -45°, -157.5° に位置しており，マイクロホンアレイの中心と話者の口元の距離は約 1 [m] であった．話者の口元の高さは，約 1.2 [m]，マイクロホンアレイの高さは 0.8 [m] であった．なお，マイクロホンは, Sony ECM-C10, 多チャンネル A/D は, 東京エレクトロンデバイス社製 TD-BD-16ADUSB を用いた．

2·2 HARK の設定

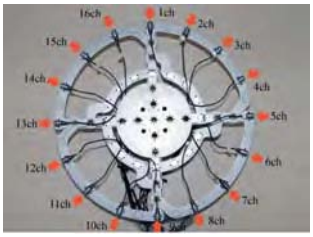
HARK の GUI プログラミング環境で図 3 に示されるネットワークを作成して，収録データの解析を行った．このネットワークは，録音ファイルを読み込んで (AudioStreamFromWave)，フーリエ変換による周波数解析を行い (MultiFFT)，*MUltiple SIgnal Classification (MUSIC)* 法を用いてフレーム[4]ごとに音源定位を行う (LocalizeMUSIC)．その後，音源追跡 (SourceTracker)，発話の検出遅れを防ぐため音源開始時刻を検出結果より遡って早くする処理 (SourceIntervalExtender) を行い，結果をファイルに保存 (SaveLocalization) する．

MUSIC 法は，M チャネル音響信号（本稿では $M = 16$）から，最大 $M-1$ 個の音源の方向を推定できる手法で，遅延和ビームフォーミングなどと比較し，雑音にロバストであることが知られている．MUSIC では，式 2 から求まる MUSIC 空間-周波数スペクトルの周波数方向の和として，式 1 から得られる MUSIC 空間スペクトル $\bar{P}(\theta)$ を算出して，閾値以上のピークの方向 θ を音源方向とする手法である．

$$\bar{P}(\theta) = \sum_{\omega=\omega_s}^{\omega_e} \sqrt{\lambda_1} P(\theta, \omega) \tag{1}$$

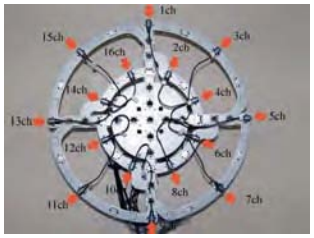
[1] http://www.irc.atr.jp/human-robot-symbiosis/
[2] http://winnie.kuis.kyoto-u.ac.jp/HARK/
[3] http://corpus.amiproject.org/

[4] 1 フレームの長さは 512 サンプル (32 [ms])，フレームのシフト長は 160 サンプル (10 [ms]) である．サンプリングレートは 16[kHz] を用いた．



a) 円状レイアウト（大）

b) 円状レイアウト（小）



c) 円状レイアウト（Mix）

図 1 使用したマイクロホンアレイ:外周直径 40cm，内周直径 20cm，マイクロホンは円周上に等間隔配置，いずれも 16 本利用

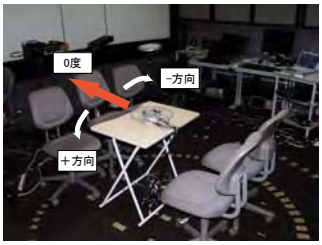


図 2 録音した部屋 (0 度は，奥手前と奥の真ん中の椅子の間方向)

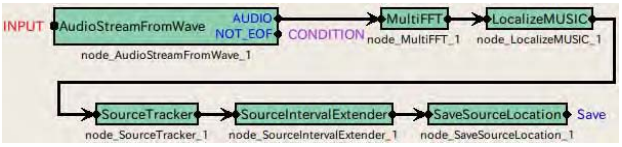


図 3 解析で用いた音源定位用ネットワーク

$$P(\theta,\omega) = \frac{|H^H(\theta,\omega)H(\theta,\omega)|}{\sum_{i=N_s+1}^{M}|H^H(\theta,\omega)e_i(\omega)|} \qquad (2)$$

$e_i(\omega)$ は，入力の相関行列を固有値分解した際の固有ベクトル，λ_1 は，その時の第一 (最大) 固有値，N_s は入力に含まれる音源数，$H(\theta,\omega)$ は，事前計測によって得られる音源とマイクロホンアレイ間のインパルス応答の直接音成分から計算される伝達関数である．本稿では，インパルス応答は，スピーカ (GENELEC 1029A) の高さを 1.2 [m] とし，マイクロホンアレイ中心から 1 [m] の距離から，5° おきに 72 点計測した．従って，定位の角度分解能は 5° である．

3. 実データ解析結果

収録した音響信号を，前節で述べた HARK のネットワークを用いて解析した結果を図 4 に示す．縦軸は音源方向，横軸は時間を表し，フレームごとの $\bar{P}(\theta)$ をグレースケールマップとして表している．青線は，HARK によって得られた発話区間・方向の推定結果，赤線は，リファレンスデータの発話区間・方向である．リファレンスデータは，聴取して作成した．実際には複数名が同時に発話したり，他の雑音源があるため，正確に発話区間を抽出することは聴取でも難しい場合がある．本稿では，HARK に備わっている音源分離機能を用いて，発話者の座席方向（22.5°, 157.5°, -157.5°, -45°, -22.5°）の音源を分離抽出し，その分離音を聴取することで，こうした問題を解決した．このように，正解データは簡単に得られず，また得られる正解データも真値ではない．実際に本稿で用いているリファレンスデータも発話方向については，単に座席とマイクアレイの位置関係から計算した値であるため，比較的大きな誤差がある．また，解析結果を見てわかるように話者数が動的に変化しているといった状況を扱わなければいけないことからも，ベンチマークテストにはない難しさがあることがわかる．N_s を 3 としているが，話者が 5 名いても，フレーム単位でみた場合に同時に発話してる人は高々3 名程度であろうという仮定をおき設定した．実際に N_s を変化させて処理を行ったが，$N_s = 3$ の時が最も良い結果を示した．音源検出時は，$\bar{P}(\theta)$ の閾値以上のピークを抽出することで行う．閾値を小さくとれば，雑音源を拾いやすく，ゴースト音源が抽出されやすい．逆に大きくとると，目的音源が抽出されないことになる．実際には最適な閾値はフレームごとに異なるため，適応的に設定することが好ましいが (閾値の適応制御は [4] で発表)，本稿では，以下で定義する定位性能を評価するための発話ベースの指標，発話正解率 C と発話正解精度 A のバランスが全体として最も良くなる値を用いた．

表 1 マイクロホンレイアウトによる性能差

Array	N	D	I	S	C (%)	A (%)
円状（大）	89	17	52	1	79.7	21.3
円状（小）	108	33	79	0	69.4	-3.7
円状（mix）	91	23	58	0	74.7	10.9

$$C = (N - D - S)/N \times 100, \qquad (3)$$
$$A = (N - D - S - I)/N \times 100, \qquad (4)$$

これらは音声認識の単語ベースの評価指標である単語正解率，単語正解精度を参考に定義したものであり，N, D, I, S はそれぞれ，正解発話数, 削除誤り発話数, 挿入誤り発話数, 置換誤り発話数を示している．なお，リファレンスから ±20° 以内に定位される発話を正解[5]，±20°〜±30° 以内に定位される発話を置換誤り，それ以上離れて定位された発話は挿入誤りとした．

3·1 マイクロホンレイアウトの違い

3 種類のマイクロホンレイアウトに関して，エラーを分析した結果を表 1 に示す．発話正解率，挿入誤りまで考慮した発話正解精度ともに，円状（大）が突出してよいことがわかる．また，全体的に挿入誤りが多く，誤りの大部分を占めていることがわかる．

この現象を考察するため，図 5 に，3 種類のレイアウトに関して，-45° に音源がある場合の指向特性を示す．円状（大）が一番鋭い指向特性を示していることがわかる．この結果は，一般的なビームフォーミングにおけるアレイ形状が大きいほどメインローブの解像度がよいという知見とも合致する．この例では，どのレイアウトでも-160° 付近に別のピークが見られることがわかる．これは部屋の反射の影響と考えられるが，目的

[5]正解の許容誤差は，定位の角度分解能 5° と比較し，大きい値であるが，顔を横に向けると口の位置が 8 [cm]，つまり，約 5° 変化する．また体が 34cm 動けば 20° 程度のずれとなるため，±20° の許容誤差は大きな値ではないと考える．

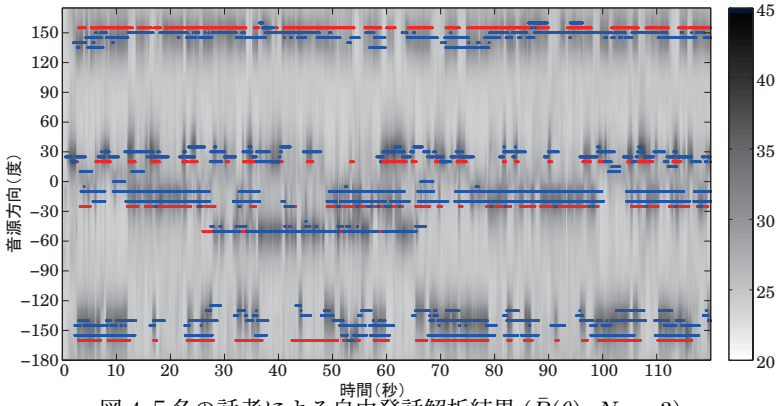


図 4 5名の話者による自由発話解析結果 ($\bar{P}(\theta)$, $N_s = 3$)

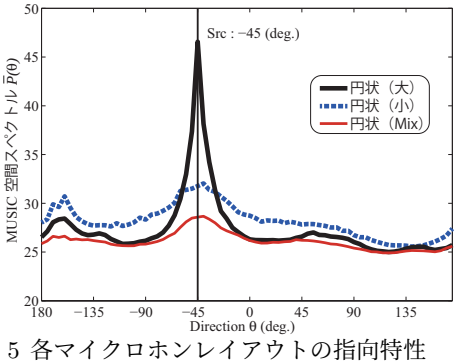


図 5 各マイクロホンレイアウトの指向特性
(インパルスに対する $P(\theta, \omega)$)

表 2 誤りの詳細解析 (誤りの発話数分類)

誤り	ID	誤り原因	大	小	Mix
削除	*D1*	ピーク不明瞭，不検出	17	33	23
挿入	*I1*	ゴースト検出	36	66	42
	I2	発話断片化	16	13	16
置換	*S1*	20°～30° の音源定位ミス	1	0	0
その他	*M1*	発話の頭尾切れ	18	38	23
	M2	発話の頭尾伸長	50	37	44
	M3	音源定位の揺れ	14	5	9

方向 (-45°) の指向特性が低いと-160° 付近のピークの影響が出やすくなる．実際に円状（小）や円状（mix)では，目的方向のピークは低く，閾値処理で区別することは難しい．つまり，目的方向のピークを抽出すると，副作用として-160° 付近のピークも抽出する可能性が高くなる．これが挿入誤りの増加につながっていると考えられる．なお，紙面の都合で割愛するが，目的方向が-45° 以外でも同様の傾向が見られた．

4. 音源定位・検出誤りの詳細解析

前節では，音声認識の評価法に倣い，削除誤り，挿入誤り，置換誤りという分類を行い，分析を行ったが，これらの誤りの原因を探るため，各誤りについて，より詳細な解析を行う．図 4 に対し，誤りの分析を行った結果を図 6 に示す．また，前節の削除誤り，挿入誤り，置換誤りとの関係を表 2 にまとめた．これまでの誤り分類では，解析しきれない現象が起きていることがわかり，これらをその他の誤りとしてまとめた．

実際には，定位誤差が 30° 以上ある音源定位ミスは，削除誤りと挿入誤りが同時に起きた場合として，処理すべきであるが，今回のデータではそうした事例は見られなかったので，表 2 には含めていない．

削除誤り： *D1* は,「ああ」「ええ」「へえ～」などフィラーや相槌など音量が小さい発話によく見られ，検出すべき発話が検出できない現象である．典型的な箇所の MUSIC 空間-周波数スペクトル ($P(\theta, \omega)$) を図 7b), c) に示す．図 7b) は，図 6 の 17.5[s] の $P(\theta, \omega)$ であり，リファレンスでは，-22.5°，22.5°，157° に音源が存在している．最初の 2 音源は，ピークが現れているが，157° の音源はピークのパワーが弱く検出しにくくなっていることがわかる．実際，これは笑い声であり，他の発話と比べて，パワーが小さく空間スペクトルを求めた際に閾値を超えないため，検出されなかった．図 7c) は，図 6 の 28[s] の $P(\theta, \omega)$ であり，リファレンスでは，±157°, ±22.5°, -45° に音源が存在している．図 6 の該当時刻の MUSIC スペクトルを見ると，-157° は，ほぼ発話終了となっており，空間スペクトルのパワーがほとんどないことがわかる．157° は，フレーム単位でみると，息継ぎなどで無音となっているが，音源は前後のフレームでピークが検出されているため，音源検出に問題は起きていない．一方，-22.5° の音源はピークの確認ができない．この発話は「リスニング」という音量が小さい発話であり，雑音に紛れてしまい検出ができなかった．

挿入誤り： *I1* は，実際に音源がない方向に音源が検出される現象で，反射の影響などが考えられる（関連：図 5）．例えば，図 6 の 3.5[s] では，±140° 近辺にゴーストが確認できる．図 7a) は，図 6 の 3.5[s] の $P(\theta, \omega)$ であり，±157° のピークが周波数に沿ってみると揺らぎがあることがわかる．これが空間スペクトルを算出した際に複数のピークとなって現れる．このような揺らぎが起きる原因としては，反射が考えられる．*I2* は，1 発話が, 2 つ以上に分割して検出される現象である．図 6 の例では,「君さあ、あ～、旅行好きなん?」といった息継ぎや言い澱みがある場合や，発話が長時間に及ぶ場合に, 途中でトラッキングが失敗し, 断片化が起きる傾向がある．

置換誤り： *S1* は，*D2* や *I3* と同様音源定位ミスによる誤りであるが，ミスの程度が 20°～30° に収まっている場合を指す．この誤りは，ほとんど見られなかったので，音源定位自体の性能は高いといえる．

その他の誤り： 時間方向の誤りと音源方向の誤りに関するもので，*M1* と *M2* が前者，*M3* が後者に対応する．開始時刻が遅れる場合は，発話の冒頭部の音量が小さいためであり，前述した MUSIC の閾値に関わる問題である．終了時刻が早くなってしまう場合は，語尾の音量が小さい場合と考えられる．HARK では LocalizeMUSIC のパラメータとして発話継続時間長を設定でき，これによってある程度問題を緩和できる．これら 2 つの問題は発話区間を実際より短く見積もってしまうことになり，音声認識では致命的な問題である．HARK では，前者の問題は，図 3 の SourceIntervalExtender を用いて，後者は，LocalizeMUSIC の発話継続時間長パラメータを適切に設定することにより，これらの問題をある程度緩和することができる．より抜本的には，MUSIC の閾値を適応的に決める方法 [4] や雑音抑圧機能付きの MUSIC (GSVD-MUSIC) [5] を導入してさらなる改善を図る予定である．また，開始時刻が実際よりも早

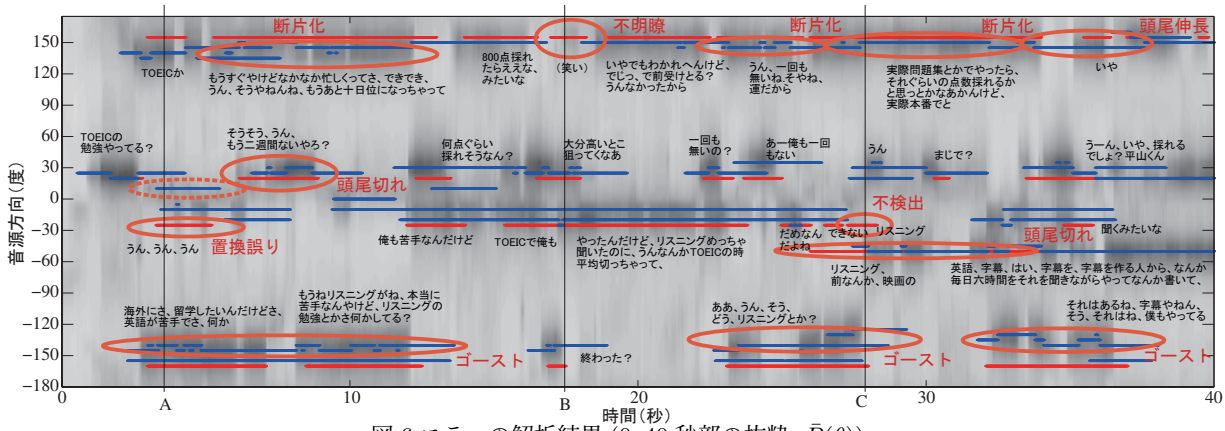
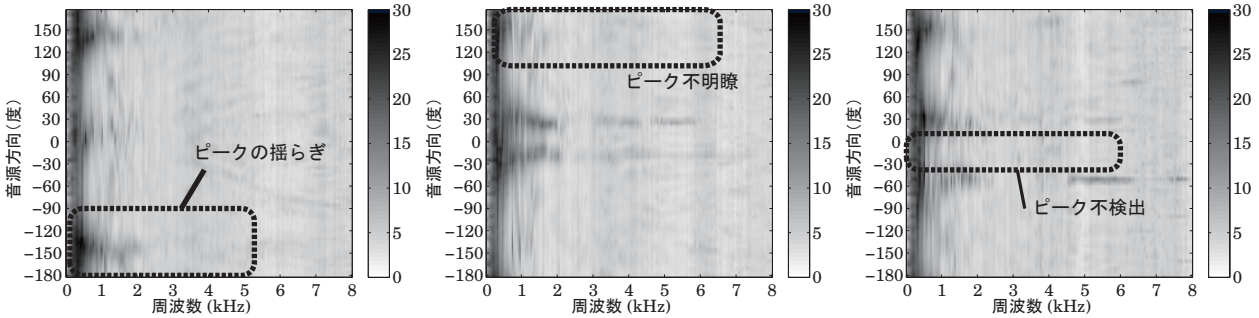


図 6 エラーの解析結果 (0–40 秒部の抜粋, $\bar{P}(\theta)$)



a) ゴースト部 (図 6 中の A)　b) ピーク不明瞭部 (図 6 中の B)　c) ピーク不検出部 (図 6 中の C)

図 7 MUSIC 空間-周波数スペクトル ($P(\theta, \omega)$)

い時刻として検出される場合は，雑音の影響もしくは，一つ前の発話とつながってしまった場合と考えられる．また，終了時刻が，遅れる場合も，雑音の影響や次の発話とつながった場合が考えられる．これらについても，[4] や [5] を用いて改善可能だと考える．一方，*M3* は，頭部の動作や，実際の音源方向が測定した伝達関数の方向とずれている場合が原因として考えられる．この問題については，音源定位の解像度を測定した伝達関数の解像度以上に向上させる相関行列内挿法 [6] を用いることで，滑らかな音源の追跡が実現できると考える．

5. まとめ

本稿では，一般的な会話シーンから，いつ，どこで，誰が発話を行ったかという情報を抽出する技術の確立を目指し，5 名の 2 分間に及ぶ自由発話の分析を行った．これまで研究開発を行ってきたオープンソースのロボット聴覚ソフトウェア HARK を用いて 3 種類の 16 ch マイクロホンアレイを試したところ，以下の知見を得た．

1. 音声認識でよく用いられる削除・挿入・置換誤りを用い，ドレードオフ関係にある挿入誤りと削除誤りのバランスを考えて閾値を設定すると，挿入誤りの占める割合が大きくなることが分かった．
2. 上記の誤りの主な原因は 7 つに大別されることが分かった．これらには，削除・挿入・置換誤りとは別に，発話時刻に関するエラーや，音源定位の揺れに関する 3 つの要素も含まれる．これらに対処するには，動的な音源数の変化や雑音レベルの変化に対応した適応的な閾値処理や定位用の雑音抑圧処理が必要であろう．実際に別稿 [4, 5] でこうした処理の有効性が示されている．
3. 3 種類のマイクロホンレイアウトでは，発話正解率，発話正解精度アレイのサイズが大きいと良好な性能を示すことが分かった．これは，アレイサイズが大きい方が鋭く解像度の高い指向特性が得られることによるものであると考えられる．

今後は，これらの誤りを解消する手法を適用し，「人ロボット共生学」で想定している教室内の先生や生徒の自由発話の発話分析にも適用可能な技術の確立を目指したい．

謝辞　本研究の一部は，新学術領域研究 No.22118502, 京大 GCOE, 科研若手 (B) No.21700195 の支援を受けた．

参考文献

[1] Otsuka et. al, “A realtime multimodal system for analyzing group meetings by combining face pose tracking and speaker diarization”, ICMI 2008, pp.257–264.
[2] K. Nakadai et. al, “Design and implementation of robot audition system HARK”, Advanced Robotics, vol.24, pp.739–761, 2010.
[3] T. Mizumoto *et al.*: Design and Implementation of Selectable Sound Separation on a Texai Telepresence System Using HARK. IEEE-RAS ICRA 2011, pp.2130-2137.
[4] 大塚ら, “MUSIC 法を用いた音源定位のベイズ拡張”, ロボット学会学術講演会, 2011
[5] 中村ら, “ロボットを対象にした複数同時発話にロバストな音源定位の検討”, ロボット学会学術講演会, 2011
[6] Nakamura *et al.*, “Correlation Matrix Interpolation in Sound Source Localization for a Robot”, IEEE ICASSP 2011, pp.4324-4327.