人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-B202

ＡＩチャレンジ研究会 (第３６回)

Proceedings of the 36th Meeting of Special Interest Group on AI Challenges

CONTENTS

日　時　2012 年 11 月 15 日　　場　所　慶應義塾大学 日吉キャンパス 来往舎 シンポジウムスペース
Keio University, Kanagawa, Nov. 15, 2012

社団法人 人工知能学会
Japanese Society for Arti cial Intelligence

人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-B202-1

両耳間時間差の脳内処理メカニズムへの心理物理学的アプローチ

Psychophysical approach to understanding interaural-time-difference processing in the brain

古川 茂人 (日本電信電話株式会社　NTT コミュニケーション科学基礎研究所)
Shigeto FURUKAWA (NTT Communication Science Labs, NTT corporation)

furukawa.shigeto@lab.ntt.co.jp

Abstract—This paper reviews three types of psychophysical experiments conducted in the author's laboratory, aiming at understanding interaural-time-difference (ITD) processing in the human auditory system. The first experiment, on the basis of the signal detection theory, indicated that the process for ITD is different between high and low frequency regions: In the low frequency regions, there are partially independent processes for ITD and ILD (interaural level difference), whereas in the high frequency regions, a common mechanism likely processes the two cues. The second experiment was concerned with the hierarchy of ITD information processing and the temporal limit that a mechanism at each level of processing can track changes in ITD. The results indicated that the ITD processor can track changes in ITD up to a rate of 20 Hz and that the loss of tracking ability of the auditory system for a rate slower than that is attributable to a higher level mechanism, which might include a process which integrates ITD and ILD information. In the third experiment, listeners were trained in a pitch discrimination task. The task required the listeners effectively use the information of temporal fine structure of the stimuli, the information being critically important also in ITD processing. The pitch training deteriorated the listeners' performance in the (untrained) ITD discrimination task. The result implies that the pitch and ITD processes compete with each other for limited neural resources.

1. はじめに

耳に届く音の両耳間時間差(interaural time difference, ITD)は両耳間レベル差(interaural level difference, ILD)とならんで、音源定位の主要な手がかりである。競合音が存在する実環境で、目的とする音を聞き取る際にも、この ITD は重要な役割を果たすと考えられている。これまでにメンフクロウ、ネコ、スナネズミなどの動物モデルを用いた生理学実験により、聴覚系における ITD 処理機構の概要がある程度明らかになってきている（例えば[1-3]）。

しかしながら、ヒトはこれらの動物モデルとは、身体の形状、可聴周波数範囲、生態学的な位置づけが異なっている。神経系が身体的、生態学的な条件に合わせて進化・発達しうることを考えると、これら動物モデルから得られた結果をそのままヒトに適用できるとは限らない。ヒトを対象にした実験では、動物に対してのように侵襲的な生理実験を行うことができない。また、既存の脳機能イメージング手法は、空間・時間分解能が十分でなく、ITD 処理が行われる脳幹内の微小な神経核の活動を明らかにすることはできない。

このような状況では、心理物理学的アプローチは、ヒトの ITD 処理メカニズムを探るうえで有効である。本稿では､ITD の脳内処理メカニズムの解明のため、特に脳内処理モジュールの構成や独立性に着目して筆者の研究室で実施された 3 種類の心理物理実験を紹介する。

2. ITD と ILD の処理メカニズムの独立性・周波数間の違い[4]

動物モデルを用いた生理実験により、脳幹にはそれぞれ ITD と ILD の処理に特化した神経核（それぞれ、上オリーブ内側核および外側核；medial superior olive (MSO)および lateral superior olive (LSO)）が存在することが示されている[2]。これは、ITD と ILD が独立した機構で処理されている可能性を意味するものである。

しかし、実際にヒトの聴覚系において、ITD と ILD の処理がどの程度独立かは明らかではない。さらに、ITD および ILD の情報が広い周波数帯域にわたって同じメカニズムで処理されているとは限らない。本研究では､ITD および ILD 処理に関わる仮想的な「チャンネル」を仮定し、信号検出理論[5]の枠組みに従ってチャンネルの重なり・独立性を評価し、周波数帯域間で比較した。

信号検出理論[5]では、各チャンネルの出力は「信号」（手がかりの強度に関連した活動量）とそれに加わる「内部ノイズ」の 2 つの要素によって特徴付けられる。心理物理実験によって計測される、刺激音の検出感度（正答率等から導出される d' 値）は、この信号対ノイズ比によって決定されると考える。ITD または ILD がそれぞれ単独で変化した場合には、それぞれに対する検出感度 d'_{T},および d'_{L} は、対応するチャンネルでの信号対ノイズ比で表されると考える。ITD と ILD が同時に（同方向に）変化するときの感度 d'_{c} は、チャンネルの独立性に依存する。各チャンネルからの出力が、線形かつ最適に組み合わされると仮定すると、d'c は、

$$d'_{\mathrm{C}} = \sqrt{d'^{2}_{\mathrm{T}} + d'^{2}_{\mathrm{I}} + 2\rho d'_{\mathrm{T}} d'_{\mathrm{I}}}$$

で予測される（Fig. 1）[6, 7]。こでρはチャンネルの重なり程度を表す指標（内部ノイズのチャンネル間相関係数）で、値が大きいほどチャンネルの重なりが大きく、0（チャンネル同士が完全に独立）から 1

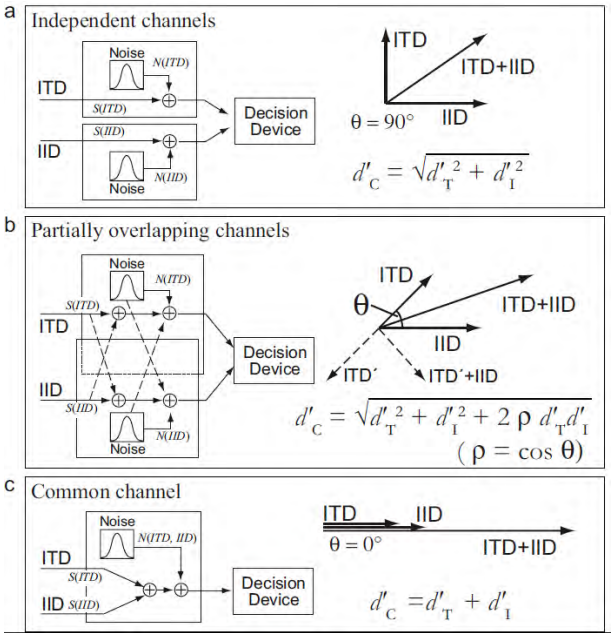


Fig. 1. Illustration of the working model used for evaluating cue interaction [4]. The model assumes channels for ITD and ILD processing, that overlap each other to various degrees. Each channel has internal *noise*, which is added to the *signal* that is related to cue strength. The magnitudes of the internal noise and the signal in the ITD and ILD channels depend on the input ITD and ILD, respectively. The outputs from the channels are added linearly. The decision device was assumed to base its judgments on changes in the weighted linear sum. It was also assumed that the weights on the channel outputs are chosen to optimize the performance. a. An extreme case in which the two channels are completely independent. Each of the channels has independent noise sources. In this case, the dimensions for ITD and ILD are thought of as orthogonal, and the d' value for simultaneous and consonant changes of ITD and ILD would be the root mean square sum of the d' values for individual changes. b. The two channels partially overlap. Fractions of the signal and the noise in one channel are added mutually to the other channel. The degree of channel overlap can be expressed as the angle of the ITD and ILD dimensions, θ (see the vector diagram of solid-line arrows). The cos θ value is equal to the correlation coefficient ρ for the noise at the outputs of the two channels. c. ITD and ILD information is combined linearly in a single channel, in which the common internal noise is added to the signal. Equivalently, ITD and ILD are represented by a single dimension. The d' value for combined ITD and IID changes is the simple sum of the d' values for individual changes.

（ITD と ILD が共通の単一のチャンネル内で処理される）の間の値をとる。d'_T,および d'_L が一定ならば、ρ が大きくなるほど、d'_c も大きくなる。逆に、心理物理実験から得られた d' 値を上式に当てはめ、ρ を推定することにより、チャンネル間の独立性を推測できる。

心理物理実験では、両耳手がかりの変化（ITD のみ、ILD のみ、ITD・ILD 同時）に対するに対する検出感度(それぞれ d'_T, d'_L, d'_c)を測定した。刺激音として次の 3 種を用いた。低周波数刺激として、周波数が 125 Hz および 500 Hz の正弦波音（125-Hz tone, 500-Hz tone）、および高周波数刺激として、搬送周波数が 4 kHz の振幅変調音（4-kHz AM tone）である。4-kHz AM tone の変調波は、125 Hz 正弦波を半波整流し、低域

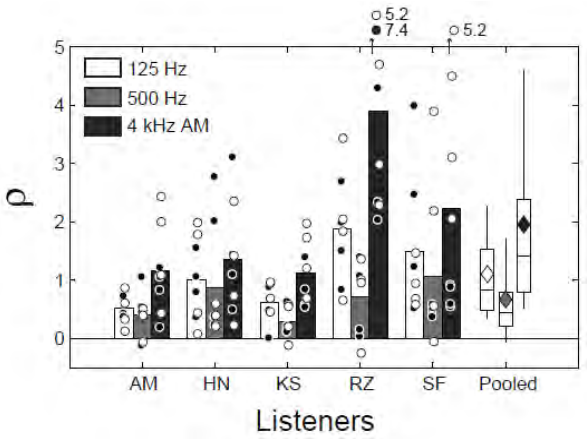


Fig. 2 The ρ values estimated for the three stimulus types. For each stimulus type, individual ρ estimates for 4 ΔITD/ΔIID combinations and 2 ITD/IID change directions are shown by the symbols. A bar indicates the mean of the eight ρ estimates. The 10, 25, 50, 75, and 95th percentiles of the ρ estimates pooled for all listeners and conditions are shown by the box-and-whisker plots.

濾過（<2000 Hz）したものである。この刺激は、4 kHz に同調した聴神経を刺激していながら、神経の発火パターンが、125 Hz の純音に対する 125 Hz に同調した聴神経の発火パターンと実質的に同様になるように設計されている[8]。これらの刺激を、所望の ITD または ILD となるように、ヘッドホンで両耳に呈示した。

検出感度の測定には 2 区間 2 肢強制選択法を用いた。各観察区間は連続した 3 つのトーン・バーストからなった。Reference 区間では、3 つのバーストの ITD, ILD は一定とした(Baseline ITD, ILD = 0 μs, 0 dB)。Signal 区間では、2 番目のバーストについて、Baseline 値よりΔITD またはΔILD だけ外れた ITD または ILD とした。Reference および Signal 区間の順序は、試行ごとにランダムに変え、被験者には、2 つの観察区間を比較して、Signal 区間を示すように指示した。ITD と ILD を同時に変化させた場合、その方向は常に同一であった。つまり、右耳刺激が左耳に対して時間的に先行する ITD 変化には、右耳刺激レベルが相対的に増加する ILD 変化が伴った。

数種類のΔITD とΔILD の組み合わせについて、それぞれ 210 試行の結果を得た。その結果から、検出感度を d' = √2 · z-score(*P*c)によって算出した（*P*c は正答率）[9]。こうして得られた d' 値を上式に当てはめ、最小二乗法により各条件におけるρを推定した。

推定されたρ 値を Fig. 2 に示す。それぞれの点は、ΔITD とΔILD の各組合せに対して求められたρ 値で、それらの平均値が棒グラフで示されている。被験者間での推定値のばらつきは大きいが、刺激種の系統的な効果も顕著である。2 要因分散分析（要因 1：刺激種；要因 2：被験者と変化方向の組合せ）の結果、刺激種の主効果は有意であった($p < 0.001$, $F(2,$

90)=20.08)。刺激種間の多重比較では、4-kHz AM tone におけるρ 値が、他の刺激種よりも有意に大きいことが分かった(Tukey-Kramer test, $p < 0.01$)。低周波刺激同士の差は有意ではなかった($p > 0.05$)。これらの結果は、ITD および ILD の処理過程が、周波数間で異なることを示唆するものである。

4-kHz AM tone におけるρ値の平均は、1.96 であった。これは、本研究で想定したρ 値の範囲(最大値 1)を超える。これは、設定した仮定のいずれかが成立しないことを意味している。実際、観測された d' c 値は、$\rho = 1$（単一チャンネル）を仮定した場合の予測値よりも有意に大きかった(paired *t*-test: $p < 0.001$)。その理由として、ITD と ILD の非線形な加算が可能性として考えられるが、その実態は不明である。

125-Hz tone においては、ρ の平均値は 1.11 であり、共通チャンネル($\rho = 1$)による予測と統計的な差異はなかった。500-Hz tone については、ρ の平均値(0.67)は 0 と 1 の間の値をとった。これは、ITD および ILD チャンネルが部分的に重なる（ITD と ILD を処理する独立した過程がある）ことを示唆するものである。

3. ITD の時間的変化に対する追従特性[10]

両耳機構は、緩慢(sluggish)であるといわれている。つまり、ITD または ILD が変動した場合、比較的ゆっくりとした（< 10Hz）ものであっても、その変化（変調）に追従できないとされている[11]。本研究では、その緩慢性の原因が、聴覚系のどのプロセスにあるかを検討した。

この検討にあたって、両耳機構を単純な 3 段階のプロセスとしてモデル化した。第 1 段階は、ITD および ILD の 2 つの手がかり（cue）を別々に処理する独立の機構から構成される（cue の個別処理レベル）。これら 2 つの機構は、それぞれ脳幹の上オリーブ内側核および外側核に相当すると考えても良い。第 2 段階では、第 1 段階の 2 つの機構からの出力を統合する（cue の統合レベル）。ここで、ITD および ILD の 2 つの cue が、加算され、あるいは互いに打ち消しあう。第 3 段階は、それ以上の高次機構をまとめたもので、判断機構(decision device)もそれに含まれる（高次レベル）。

あるプロセスにおいて「変調に追従できない」状態とは、「変調の方向に関する情報が失われている」状態と考えることができる。本研究では、変調方向に関する情報が、cue の個別処理のレベル（第 1 段階）、または cue の統合レベル（第 2 段階）以降の、いずれの段階で失われるかを検討した。実験では、刺激音の ITD および ILD を同時に正弦波状に変調して、その変調の検出感度を計測した。そして、ITD と ILD 変調の相対位相の効果を調べた。

もし、検出成績が、相対位相に依存するならば、

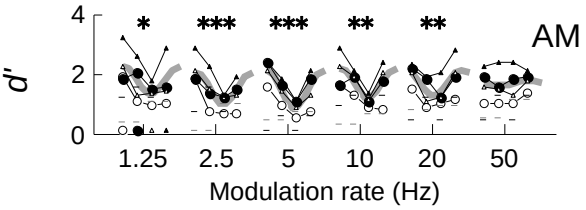


Fig. 3. The *d'* for detecting ITD and/or ILD modulations. The data are sorted according to the modulation rate (horizontal axis) and the relative phase of the ITD and ILD modulations, when the ITD and ILD were modulated simultaneously (connected symbols; relative phases of 0, 90, 180, and 270 degrees from left to right at each rate). Cases with statistically significant effects of relative phase are marked with asterisks above the plots (one-way repeated measures ANOVA; *: $p < 0.05$; **: $p < 0.01$; and ***: $p < 0.001$). The gray curve with each listener and rate is a cycle of sinusoid fitted to the mean data across the combinations of ΔITD and ΔILD.

それは、変調方向の情報が、cue の個別処理のレベルで保持されている証拠ととらえることができる。これは、次のように考えると理解しやすい。ITD と ILD が同時に変調されるとき、cue の個別処理レベルの出力としての ITD と ILD の変調が同位相（たとえば、左耳と比較して、右耳への刺激が時間的に先行するタイミングと音圧レベルが増加するタイミングが同期する）ならば、統合レベルにおいて 2 つの cue は加算され、逆位相ならば同レベルにおいて打ち消しあうであろう。このため、同位相の変調は検出しやすく、逆位相は検出しにくいことが予測される。もし、変調方向の情報が早い段階で失われるならば、このような、位相関係に応じた加算や打ち消しは起こらない。

実験では、中心周波数が 500 Hz の 1/3 オクターブバンド雑音を刺激音として用い、ヘッドホンを通して両耳に呈示した。被験者の検出感度は 2 区間 2 肢強制選択法を用いて計測した。2 つの刺激呈示区間のうち、「信号あり」区間では、刺激音の ITD および ILD を個別または同時に、0 μs および 0 dB を中心とした正弦波状に変調した。変調周波数として 1.25, 2.5, 5, 10, 20, および 50 Hz を用いた。各変調周波数について、ITD 変調の、ILD 変調に対する位相遅延として、0, 90, 180, および 270 度を試した。「信号なし」区間では、変調は行わなかった（つまり diotic な刺激を両耳に呈示した）。各変調周波数について、ITD および ILD 変調の振幅（ΔITD, ΔILD）に関して数首里の組み合わせ 20 種類の条件を試した：ITD と ILD の同時変調について 16 種類（2 種類のΔITD x 2 種類のΔILD x 4 種類の位相遅延）および、ITD または ILD 単独の変調について 4 種類(2 種類のΔITD + 2 種類のΔILD). それぞれの条件において、81~117 試行を行った（被験者および変調周波数ごとに異なる）。得られた正答率より、*d'*値を算出した。

例として、1 名の聴取者から得られた結果を Fig. 1

に示す。各変調周波数（横軸）について、ITD と ILD の同時変調に対する検出感度 *d'*値を示している。各変調周波数について、折線で結ばれた点は各位相遅延に対する結果である（左から、0, 90, 180, および 270 度）。20 Hz までの変調周波数に対して、位相遅延の有意な影響がみられた。この聴取者では、検出感度が最大となるのは、位相遅延が 0°（同位相）の時で、180°（逆位相）で最小となる傾向が見られた。検出感度が最大となる位相遅延は、聴取者によって異なった。しかし、いずれの場合であっても、位相遅延の影響が見られたのは 20 Hz までの変調周波数であった。

前述のモデルから考えると、この結果は、ITD を処理するモジュールは、少なくとも 20 Hz までの変調に追従することができることを意味する。聴覚系全体で考えたときの追従の上限がその周波数よりも低いならば、その要因は、ITD と ILD を統合するモジュール以上のレベルの処理にあるといえる。

4. ピッチの知覚学習の ITD 感度に与える影響[12]

ITD の処理には少なくとも 2 つのメカニズムが関与している。1 つは音の詳細な時間波形（temporal fine structure, TFS）を、聴神経の位相固定（phase locking）によって神経の時間的発火パターンへと変換し、中枢へと伝達するメカニズム（TFS 伝達メカニズム）である。もう 1 つは、前記のメカニズムによって伝えられる神経発火のタイミングを、両耳間で比較するメカニズム（両耳メカニズム）である。ITD の感度は、これら 2 つのメカニズムの精度によってある程度決定されると考えられる。このうち、TFS 伝達メカニズムは、ITD 以外の情報処理にも重要な役割を果たす。例えば、神経発火の時間間隔に基づいて TFS の周期性を計算することによって、ピッチを知覚することができると考えられる。

一般に、ある特定の知覚タスクを繰り返し行うと、その知覚属性について特異的に感度が向上する（知覚学習が起こる）ことがある。これは、このタスクに関わる神経処理モジュールが可塑的であることを意味する。この場合、この神経処理モジュールの出力を利用する他のタスクの成績も向上するであろう。逆に言えば、一見異なる知覚属性に関するタスクであっても、一方に関する知覚学習の結果が、もう一方のタスクの成績に影響するのであれば、これら 2 つの知覚属性は共通のモジュールを利用していると推察することができる。このパラダイムに則り、これまで多くの有効な知見が得られている（例えば[13]）。本研究では、TFS に基づくピッチタスクについての知覚学習が、ITD タスクに影響を与えるかを調べた。

刺激音は、帯域制限された調波複合音（基本周波数 100 Hz ; 7 から 14 倍音成分を含む）である（H 音とする）。この音のすべての成分を、同時に一定の周波数（Δf）だけ高周波側にシフトさせると、振幅包絡（周期 100 Hz）を一定に保ったまま、その詳細な時間波形（TFS）のみが変化する音（S 音とする）を作ることができる。この音は、H 音よりピッチが高い音として知覚される。聴覚末梢においては、その周波数分解能の制約のため、各周波数成分は分解されない。このため、H 音と S 音のピッチの違いは、TFS の違いによるものといえる[14]。実験では、2 種類の系列（HHHH および HSHS）の音をランダムな順序で呈示し、2 区間 2 肢強制選択法において 2 つの系列を弁別するタスクを聴取者に課した。このタスクでは、適応法によってΔf を変化させ、弁別閾値を求めた。聴取者としては、過去に聴覚心理学実験を経験したことのないものを採用した。これを 2 群に分け、一方の群（訓練群）では、このタスクを 12 日間行う訓練を課した。もう一方の群（コントロール群）では、同等な期間に訓練を行わなかった。訓練期間の前後には、同じピッチタスクに加えて、H 音を刺激として用いた ITD タスク、ILD タスク、および強度弁別タスクを行った。これらのタスクで、それぞれ ITD 検出、ILD 検出、および強度弁別閾値を計測した。

Fig. 4 は、訓練群について、訓練前後のピッチ弁別閾値を比較したものである。期待されたとおり、訓練後の閾値は、訓練前のそれから有意に減少した（つまり、成績が向上した）。しかし、ITD タスクについて、ピッチタスク訓練前後の閾値を比較すると、訓練後の閾値は有意に上昇した（成績が悪化した）。他のタスクについては、訓練前後の閾値変化は認められなかった。また、コントロール群については、訓練期間（実際は訓練はしていないが）前後の閾値の変化はいずれのタスクでも認められなかった。

ピッチの知覚学習による ITD 感度の悪化は、期待されていなかった結果である。これは、ピッチ情報と ITD 情報が、一つの神経モジュール内で同時に処

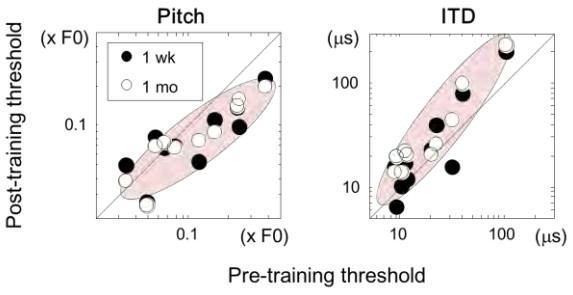


Fig. 4. Comparisons of post-training and pre-training thresholds. Filled and open symbols indicate thresholds obtained 1 week and 1 month after the pitch training, respectively. Each symbol represents one listener.

理されると仮定することで説明できるかもしれない。その神経モジュールの計算リソースは限られており、ピッチ処理と ITD 処理は競合する。ピッチタスクの訓練によって、より多くのリソースがピッチ処理に割かれることにより、その結果、ITD 処理のためのリソースが減少したのかもしれない。

5. おわりに

本稿をとおして、両耳情報の神経処理機構を推定するためのツールとして、心理物理学が有効であることを感じていただければ幸いである。

謝辞

知覚学習に関する研究は、鷲沢史歩氏と柏野牧夫氏との共同研究である。

参考文献

[1] M. Konishi, "Coding of auditory space," *Annu Rev Neurosci,* vol. 26, pp. 31-55, 2003.

[2] T. C. Yin, "Neural mechanisms of encoding binaural localization cues in the auditory brainstem," in *Integrative functions in the mammalian auditory pathway*, D. Oertel*, et al.*, Eds., ed New York: Springer, 2002, pp. 99-159.

[3] B. Grothe, "Sensory systems: New roles for synaptic inhibition in sound localization," *Nat Rev Neurosci,* vol. 4, pp. 540-50, Jul 2003.

[4] S. Furukawa, "Detection of combined changes in interaural time and intensity differences: Segregated mechanisms in cue type and in operating frequency range?," *J Acoust Soc Am,* vol. 123, pp. 1602-17, Mar 2008.

[5] D. M. Green and J. A. Swets, *Signal Detection Theory and Psychophysics*. New York: Krieger, 1974.

[6] D. M. Green, "Detection of multiple component signals in noise," *J. Acoust. Soc. Am.,* vol. 30, pp. 904-911, 1958.

[7] E. R. Hafter*, et al.*, "The combination of interaural time and intensity in the lateralization of high-frequency complex signals," *J Acoust Soc Am,* vol. 87, pp. 1702-8, Apr 1990.

[8] L. R. Bernstein and C. Trahiotis, "Enhancing sensitivity to interaural delays at high frequencies by using "transposed stimuli"," *J Acoust Soc Am,* vol. 112, pp. 1026-36, Sep 2002.

[9] N. A. Macmillan and C. D. Creelman, *Detection Theory: A User's Guide (2nd edition)*. Mahwah, NJ: Lawrence Erlbaum Associates, 2005.

[10] S. Furukawa, "Detection of simultaneous modulation of interaural time and level differences: Effects of modulation rate and relative phase (L)," *J Acoust Soc Am,* vol. 132, pp. 1-4, Jul 2012.

[11] D. W. Grantham, "Spatial hearing and related phenomena," in *Hearing*, B. C. J. Moore, Ed., ed San Diego: Academic Press, 1995, pp. 297-345.

[12] S. Furukawa*, et al.*, "How independent are the pitch and interaural-time-difference mechanisms that rely on temporal fine structure information?," in *16th International Symposium on Hearing (ISH2012)*, Cambridge, UK, 2012.

[13] B. A. Wright and M. B. Fitzgerald, "Different patterns of human discrimination learning for two interaural cues to sound-source location," *Proc Natl Acad Sci U S A,* vol. 98, pp. 12307-12, Oct 9 2001.

[14] G. A. Moore and B. C. Moore, "Perception of the low pitch of frequency-shifted complexes," *J Acoust Soc Am,* vol. 113, pp. 977-85, Feb 2003.

人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-B202-2

ロボットのための実環境ロバストな実時間超解像三次元音源定位

Real-time Noise-robust Super-resolution Sound Source Localization in Three-dimension for Robots

中村圭佑 [1,2]，中臺一博 [1]，奥乃博 [2]

Keisuke NAKAMURA[1,2], Kazuhiro NAKADAI[1], Hiroshi OKUNO[2]

1.（株）ホンダ・リサーチ・インスティチュート・ジャパン，2. 京都大学大学院

1. Honda Research Institute Japan Co., Ltd., 2. Kyoto University

keisuke@jp.honda-ri.com, nakadai@jp.honda-ri.com, okuno@kuis.kyoto-u.ac.jp

Abstract

This paper investigates Three-dimensional Sound Source Localization (3D-SSL) for a robot. 3D-SSL by a robot mainly requires: 1) robustness against high power noise such as robot's ego-noise, 2) sufficiently-high resolution for a 3D space, 3) real-time operation for searching for sound sources in a 3D space. For these, we propose: 1) multiple signal classification based on generalized singular value decomposition (GSVD-MUSIC), 2) transfer function interpolation based on integration of linear interpolation in frequency- and time-domain (FT-DLI), 3) optimal hierarchical sound source localization (OH-SSL). These techniques are integrated into an SSL system using a robot, and the experimenral result showed 3D-SSL in real-time.

1 序論

人とロボットの会話によるインタラクションは重要である．携帯電話などの接話マイクを用いた音声認識に比べ，ロボットに搭載されたマイクを使った音声認識は，発話者からマイクの距離が遠く信号対雑音比が低い，発話者数は単独であると仮定できないという特徴がある．ロボット聴覚[1]では，マイクロホンアレイを用いて空間的に複数の音源を定位・分離することでこれらの問題に対処している．従って，音源定位・分離の性能向上はロボット聴覚システム全体の性能向上に必要不可欠である．

これまで，ロボット聴覚のための音源定位において，解像度が高いという利点のある Multiple Signal Classification (MUSIC[2]) を使用して，主に一次元（方位角）のみの定位が使用されてきた[3; 4; 5]．方位角に対する一次元の音源定位は，複数同時発話[7]に代表されるように，全ての音源がある水平面に近く，高さが近い場合に高い性能を維持する．Figure 1 は，一音源の距離と方位角をそれぞれ 1.0m と 0° に固定し，音源とマイクアレイとの相対高さを変化させたときの MUSIC による方位角推定結果である．図の横軸と縦軸はそれぞれ，方位角と MUSIC スペクトルを示す．図中の黒実線 (0.0m) は音源が水平面上にある時の方位角推定結果，それ以外は水平面から離れている時の方位角推定結果を示す．図のように，音源が高さを持たない場合は 0° に鋭角なピークが見られるが，高さが変化してしまうと解像度が低下し，音源定位性能が劣化してしまう．従って，高さの異なる複数の音源や雑音の定位性能を向上するために三次元の音源定位は必要不可欠である．

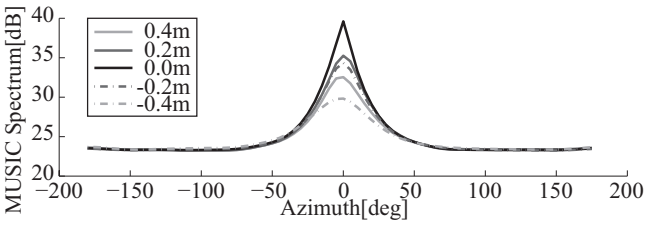


Figure 1: 1D SSL Result with the Variation of Heights

三次元音源定位の必要性に関わらず，ロボットに適用された報告は少なく，これまで，両耳聴音源定位による手法[8]，ビームフォーミングによる手法[9; 10]，MUSIC による手法[11; 12]等が報告されている．両耳聴音源定位による手法[8]は高速に定位できるが，フレーム内に一音源以上存在しないという制約が存在する．ビームフォーミングによる手法[9; 10] は複数音源に対応した高速な音源定位が実現できるが，空間解像度や雑音ロバスト性が MUSIC と比べて十分でない．一方，MUSIC による手法[11; 12]は，計算コストが高く，実時間性が保証されない．

本稿の目的は，高い雑音ロバスト性，高い空間解像度，実時間処理を実現する MUSIC による三次元音源定位の枠組を構築することにある．このため，本稿では次の問題に取り組む．

A1) ロボットの自己雑音等の目的音より大きなパワーを持つ雑音下で定位性能が劣化する．

A2) 三次元空間で高い空間解像度が求められる場合，マイクロホンアレイのキャリブレーションのための伝達関数の計測回数が増大する．

A3) 一次元音源定位に比べ，三次元音源定位では，音源探索にかかる計算コストが増大する．

A1) に対し，一般化特異値展開 (Generalized Singular Value Decomposition, GSVD) を用いた MUltiple SIgnal Classi cation(GSVD-MUSIC) を提案する．この手法は，MUSIC で一般に用いられる標準固有値展開を一般化特異値展開に拡張することにより，マイクロホン間の入力信号の相関行列に加え，自由に設計が可能な相関行列を導入することができる．この相関行列を既知または動的に取得した雑音信号から生成すれば，たとえその雑音が目的音よりも大きなパワーを持っていたとしても，雑音の影響を吸収する効果をこの相関行列が持つため，ロバストに目的音の定位ができる．

A2) に対し，一次元音源定位に対して提案されたハイブリッド伝達関数補間 (Frequency- and Time-Domain Linear Interpolation, FTDLI[6]) にトリリニア補間を導入する．この拡張によって，粗い解像度で計測された少数の三次元空間内伝達関数から所望の解像度の伝達関数を得ることができる．補間された伝達関数を音源定位に応用することで，超解像三次元音源定位を実現する．

A3) に対し，Coarse-to-fine 認識[13]に基づく最適階層的音源定位 (Optimal Hierarchical Sound Source Localization (OH-SSL)) を提案する．本手法により，空間解像度の粗い定位から細かい定位へと階層的に定位を行うことで，定位の精度を制御して精度を維持しつつ，探索にかかる計算コストを削減し，実時間性を向上する．本稿では，探索にかかるコストを最小化する最適な階層数と各階層の粒度について述べる．

2 GSVD-MUSIC による雑音ロバストな三次元音源定位

2.1 MUSIC による三次元音源定位

まずは，MUSIC [2]について概説する．音源の三次元位置を $\psi_{xyz} = [\psi_x, \psi_y, \psi_z]^T$ と表記する．ここで，座標系は円筒座標系，球座標系，直交座標系等，任意のものとする．MUSIC では ψ_{xyz} の音源のインパルス応答の計測もしくは，計算で得られる伝達関数（ステアリングベクトル）$\boldsymbol{A}(\omega, \psi_{xyz})$ を既知情報として用いる．ここで，ω は周波数である．定位では，f フレーム目の入力音響信号を短時間フーリエ変換して得られる $\boldsymbol{X}(\omega, f)$ から，以下のように入力信号の相関行列 $\boldsymbol{R}(\omega, f)$ を計算する．

$$\boldsymbol{R}(\omega, f) = \frac{1}{\mathrm{T_R}} \sum_{\tau=0}^{\mathrm{T_R}-1} \boldsymbol{X}(\omega, f+\tau)\boldsymbol{X}^*(\omega, f+\tau) \quad (1)$$

ここで，$()^*$ は複素共役転置演算子を表す．また，雑音に対するロバスト性向上のため，$\boldsymbol{R}(\omega, f)$ は $\mathrm{T_R}$ フレームで平滑化されている．

MUSIC では $\boldsymbol{R}(\omega, f)$ が張る空間を，以下の標準固有値展開 (SEVD) により，目的音と雑音の部分空間に分解する．

$$\boldsymbol{R}(\omega, f) = \boldsymbol{E}(\omega, f)\boldsymbol{\Lambda}(\omega, f)\boldsymbol{E}^{-1}(\omega, f) \quad (2)$$

以降，提案法と区別するため，式 (2) を用いた MUSIC を SEVD-MUSIC と呼ぶ．空間スペクトルは以下で求められる．

$$P(\omega, \psi_{xyz}, f) = \frac{|\boldsymbol{A}^*(\omega, \psi_{xyz})\boldsymbol{A}(\omega, \psi_{xyz})|}{\sum_{m=L_s+1}^{M} |\boldsymbol{A}^*(\omega, \psi_{xyz})\boldsymbol{e}_m(\omega, f)|} \quad (3)$$

ここで，L_s は音源数を，$\boldsymbol{e}_m(\omega, f)$ は式 (2) の $\boldsymbol{E}(\omega, f)$ の m 番目の固有ベクトルを表す．音源方向を推定するため，$P(\omega, \psi_{xyz}, f)$ を以下のように ω 方向に平均する．

$$\bar{P}(\psi_{xyz}, f) = \frac{1}{j_h - j_l + 1} \sum_{j=j_l}^{j_h} P(\omega_{[j]}, \psi_{xyz}, f) \quad (4)$$

ここで，j_l, j_h は音源定位で用いる最低・最高周波数に相当する周波数ビン番号を表す．音源探索では，$\bar{P}(\psi_{xyz}, f)$ がなす三次元超平面の極大点を探索し，その極大点の集合から極大値が大きいものから L_s 個の ψ_{xyz} を選択し，音源位置を決定する．以降，この L_s 個の ψ_{xyz} を $\psi_{xyz}^{[l]}$ と表記することとする ($1 \le l \le L_s$).

2.2 GSVD-MUSIC への拡張

式 (3) では，$\boldsymbol{A}(\omega, \psi_{xyz})$ は事前情報である為，入力音響信号に関わる項は $\boldsymbol{e}_m(\omega, f)$ のみであり，$\boldsymbol{e}_m(\omega, f)$ の選択が定位の性能に大きく影響する．MUSIC では式 (2) で求まる固有値が音源のパワーと相関があることを利用し，固有値の小さなものが雑音である（雑音のパワーは必ず目的音のパワーより小さい）という仮定のもとで $\boldsymbol{e}_m(\omega, f)$ を選択する．しかし，実環境では，雑音が目的音より大きなパワーを持つ場合が存在し，固有値の大きさに逆転が生じるため性能が劣化することが知られている．

これまで我々はこの問題の解決のため，一般化固有値分解を用いた MUSIC (GEVD-MUSIC) を導入しており，その対雑音ロバスト性を確認した[14]．この手法では，自由に設計が可能な相関行列 $\boldsymbol{V}$ を導入し，$\boldsymbol{V}$ を非発話区間の入力信号から式 (1) を用いて生成することで，非発話区間に存在するロボットの自己雑音等を白色化し，定位のロバスト性を向上させた．しかし，この手法は，計算量が大きい，もしくは，固有ベクトルの直交性が保証できないといった問題をかかえていた．

そこで，こうした問題を解決するため，本稿では一般化特異値分解による MUSIC(GSVD-MUSIC) を導入する．本手法は式 (2) を次のように拡張する．

$$\boldsymbol{V}^{-1}\boldsymbol{R}(\omega,f)=\boldsymbol{E}_l(\omega,f)\boldsymbol{\Lambda}(\omega,f)\boldsymbol{E}_r^*(\omega,f) \tag{5}$$

ここで，$\boldsymbol{E}_l(\omega,f),\boldsymbol{E}_r(\omega,f)$ は，それぞれ左特異ベクトル行列，右特異ベクトル行列を表し，いずれもユニタリ行列で互いの直交性が保証される．音源方向推定において，$\boldsymbol{E}_l(\omega,f)$ を式 (2) の $\boldsymbol{E}(\omega,f)$ として用いることで，MUSIC の対雑音ロバスト性を向上することができる．

3 FTDLI による三次元伝達関数の補間

まず，FTDLI による一次元伝達関数の補間[6]について概説する．本章では，ψ 方向の音源とマイクロホンアレイ間の伝達関数を $\boldsymbol{A}(\omega,\psi)=[A_1(\omega,\psi),\ldots,A_M(\omega,\psi)]^T$ と，M 個のマイクで別に表記する．2 つの $\psi_{\underline{x}}$ 方向と $\psi_{\overline{x}}$ 方向の事前計測伝達関数から，補間によって ψ_x 方向 ($\psi_{\underline{x}}<\psi_x<\psi_{\overline{x}}$) の未知伝達関数 $\boldsymbol{A}(\omega,\psi_x)$ を推定する．

FTDLI は周波数領域での線形補間法 (Frequency Domain Linear Interpolation, FDLI[15]) による位相情報と，時間領域での線形補間法 (Time Domain Linear Interpolation, TDLI[16]) による振幅情報を統合し，高い精度の補間を実現する．二つの線形補間法は以下のように統合される．

B1) FDLI による補間を行う．

$$\hat{A}_{m[\mathrm{F}]}(\omega,\psi_x)=(1-D_x)A_m(\omega,\psi_{\underline{x}})+D_xA_m(\omega,\psi_{\overline{x}}) \tag{6}$$

ここで，$\hat{A}_{m[\mathrm{F}]}(\omega,\psi_x)$ は，$A_m(\omega,\psi_{\underline{x}})$ と $A_m(\omega,\psi_{\overline{x}})$ の事前計測伝達関数を用いて推定された ψ_x 方向の音源と m 番目のマイクロホン間の伝達関数である．また，D_x は補間係数である ($0\le D_x\le 1$).

B2) TDLI による補間を行う．

$$\hat{A}_{m[\mathrm{T}]}(\omega,\psi_x)=A_m^{1-D_x}(\omega,\psi_{\underline{x}})A_m^{D_x}(\omega,\psi_{\overline{x}})\,. \tag{7}$$

B3) B1) と B2) によって補間された伝達関数を以下のように位相と振幅に分解する．

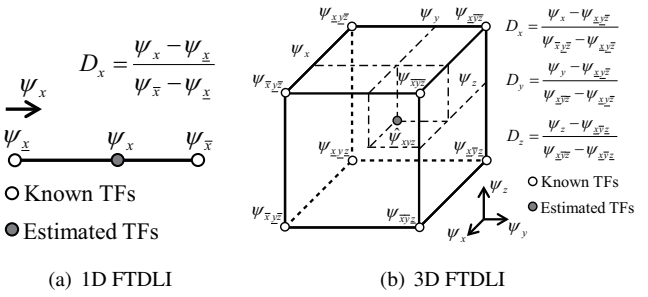
$$\hat{A}_{m[\mathrm{F}]}(\omega,\psi_x)=\lambda_{m[\mathrm{F}]}\exp(-j\omega t_{m[\mathrm{F}]}) \tag{8}$$

$$\hat{A}_{m[\mathrm{T}]}(\omega,\psi_x)=\lambda_{m[\mathrm{T}]}\exp(-j\omega t_{m[\mathrm{T}]}) \tag{9}$$

B4) 補間伝達関数 $\hat{A}_m(\omega,\psi_x)$ は以下で求まる．

$$\hat{A}_m(\omega,\psi_x)=\lambda_{m[\mathrm{T}]}\exp(-j\omega t_{m[\mathrm{F}]})\,. \tag{10}$$

本稿は一次元伝達関数に対する FTDLI を三次元伝達関数 $\boldsymbol{A}(\omega,\psi_{xyz})$ の補間に拡張する．8 つの三次元位置 ($\psi_{\underline{x}\underline{y}\underline{z}}$, $\psi_{\underline{x}\underline{y}\overline{z}}$, $\psi_{\underline{x}\overline{y}\underline{z}}$, $\psi_{\underline{x}\overline{y}\overline{z}}$, $\psi_{\overline{x}\underline{y}\underline{z}}$, $\psi_{\overline{x}\underline{y}\overline{z}}$, $\psi_{\overline{x}\overline{y}\underline{z}}$, $\psi_{\overline{x}\overline{y}\overline{z}}$) の事前計測伝達関数から，補間によって三次元空間の未知伝達関数 $\boldsymbol{A}(\omega,\psi_{xyz})$ を推定する．ただし，$\psi_{\underline{x}}<\psi_x<\psi_{\overline{x}}$, $\psi_{\underline{y}}<\psi_y<\psi_{\overline{y}}$, $\psi_{\underline{z}}<\psi_z<\psi_{\overline{z}}$ とする．



(a) 1D FTDLI (b) 3D FTDLI

Figure 2: Difference of FTDLI between 1D and 3D

FDLI では，式 (6) にトリリニア補間を以下のように導入することで三次元伝達関数の補間に拡張する．

$$\begin{aligned}
\hat{A}_{m[\mathrm{F}]}(\omega,\psi_{xyz})&=\begin{bmatrix}1-D_y & D_y\end{bmatrix}\begin{bmatrix}A_m(\omega,\psi_{\underline{x}yz}) & A_m(\omega,\psi_{\overline{x}yz})\\ A_m(\omega,\psi_{\underline{x}\overline{y}\underline{z}}) & A_m(\omega,\psi_{\overline{x}\overline{y}\underline{z}})\end{bmatrix}\begin{bmatrix}1-D_x\\ D_x\end{bmatrix}\\
\hat{A}_{m[\mathrm{F}]}(\omega,\psi_{xy\overline{z}})&=\begin{bmatrix}1-D_y & D_y\end{bmatrix}\begin{bmatrix}A_m(\omega,\psi_{\underline{x}y\overline{z}}) & A_m(\omega,\psi_{\overline{x}y\overline{z}})\\ A_m(\omega,\psi_{\underline{x}\overline{y}\overline{z}}) & A_m(\omega,\psi_{\overline{x}\overline{y}\overline{z}})\end{bmatrix}\begin{bmatrix}1-D_x\\ D_x\end{bmatrix}\\
\hat{A}_{m[\mathrm{F}]}(\omega,\psi_{xyz})&=(1-D_z)\hat{A}_{m[\mathrm{F}]}(\omega,\psi_{xy\underline{z}})+D_z\hat{A}_{m[\mathrm{F}]}(\omega,\psi_{xy\overline{z}})
\end{aligned} \tag{11}$$

ここで，D_x, D_y, D_z はそれぞれ，補間係数である ($0\le D_x,D_y,D_z\le 1$)．同様に TDLI では，式 (7) にトリリニア補間を以下のように導入する．

$$\begin{aligned}
\hat{A}_{m[\mathrm{T}]}(\omega,\psi_{xyz})=&A_m^{(1-D_x)(1-D_y)(1-D_z)}(\omega,\psi_{\underline{x}\underline{y}\underline{z}})A_m^{(1-D_x)(1-D_y)D_z}(\omega,\psi_{\underline{x}\underline{y}\overline{z}})\\
&A_m^{(1-D_x)D_y(1-D_z)}(\omega,\psi_{\underline{x}\overline{y}\underline{z}})A_m^{(1-D_x)D_yD_z}(\omega,\psi_{\underline{x}\overline{y}\overline{z}})\\
&A_m^{D_x(1-D_y)(1-D_z)}(\omega,\psi_{\overline{x}\underline{y}\underline{z}})A_m^{D_x(1-D_y)D_z}(\omega,\psi_{\overline{x}\underline{y}\overline{z}})\\
&A_m^{D_xD_y(1-D_z)}(\omega,\psi_{\overline{x}\overline{y}\underline{z}})A_m^{D_xD_yD_z}(\omega,\psi_{\overline{x}\overline{y}\overline{z}})
\end{aligned} \tag{12}$$

三次元伝達関数に対する FTDLI は，式 (11) で得られる $\hat{A}_{m[\mathrm{F}]}(\omega,\psi_{xyz})$ と，式 (12) で得られる $\hat{A}_{m[\mathrm{T}]}(\omega,\psi_{xyz})$ とを，B3) と B4) に従って統合し，補間伝達関数 $\hat{A}_m(\omega,\psi_{xyz})$ を得る．最後に，$\hat{A}_m(\omega,\psi_{xyz})$ を式 (3) の $\boldsymbol{A}(\omega,\psi_{xyz})$ として使用することで，事前計測伝達関数よりも細かな空間解像度の音源定位（超解像音源定位）が実現できる．

4 OH-SSL による音源探索コストの削減

4.1 OH-SSL のアルゴリズム

GSVD-MUSIC では，式 (3) の空間スペクトルの算出と音源探索にかかるコストが，$\boldsymbol{A}(\omega,\psi_x)$ の空間解像度に比例した探索数に依存する．OH-SSL では，探索数を削減するために音源探索を階層化する．以下簡略化のため，ψ_x 軸上の一次元音源探索の階層化を扱うが，ψ_y 軸と ψ_z 軸についても同様に階層化が可能である．$\psi_x^{[l]}$, K, $d_{x[k]}$ をそれぞれ，$\psi_{xyz}^{[l]}$ の ψ_x，階層数，k 階層目の ψ_x の空間解像度とする．OH-SSL は，音源探索を以下のように階層化する．

C1) 伝達関数を $d_{x[k]}$ 間隔となるように選ぶ．必要なら FTDLI により伝達関数を補間する．

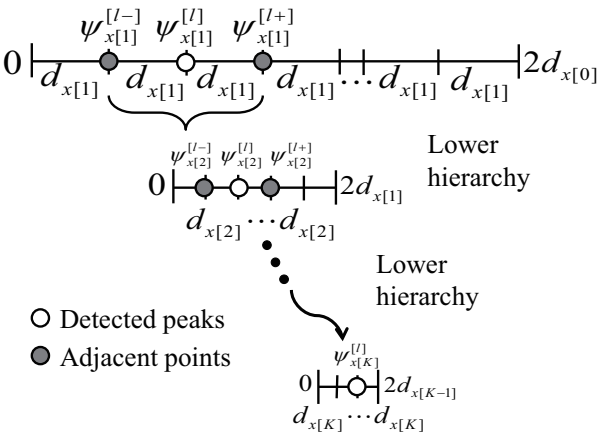


Figure 3: Hierarchical Structure of OH-SSL

C2) $\bar{P}(\psi_{xyz}, f)$ により定位する．ここで，$\psi_{x[k]}^{[l]}$ を k 階層目の $\psi_x^{[l]}$ とする．

C3) $\psi_{x[k]}^{[l]}$ から ψ_x 軸に沿って，その点を囲む二点を選ぶ．その二点を $\psi_{x[k]}^{[l-]}$ と $\psi_{x[k]}^{[l+]}$ とする ($\psi_{x[k]}^{[l-]} < \psi_{x[k]}^{[l]} < \psi_{x[k]}^{[l+]}$)．

C4) $\hat{A}_m(\omega, \psi_{x[k]}^{[l-]})$ と $\hat{A}_m(\omega, \psi_{x[k]}^{[l]})$ を使って，$\psi_{x[k]}^{[l-]}$ と $\psi_{x[k]}^{[l]}$ の間の伝達関数 $\hat{A}_m(\omega, \psi_x)$ を $d_{x[k+1]}$ の空間解像度で式 (10) を使って補間する．

C5) $\hat{A}_m(\omega, \psi_{x[k]}^{[l]})$ と $\hat{A}_m(\omega, \psi_{x[k]}^{[l+]})$ を使って，$\psi_{x[k]}^{[l]}$ と $\psi_{x[k]}^{[l+]}$ の間の伝達関数 $\hat{A}_m(\omega, \psi_x)$ を $d_{x[k+1]}$ の空間解像度で式 (10) を使って補間する．

C6) C4) と C5) で生成された伝達関数を用いて，式 (3) により定位する．

C7) C3)-C6) を k が K になるまで繰り返す．

階層化によって探索の粒度を制御できるため，定位性能を維持しつつ，計算コストを削減できる．

4.2 OH-SSL による探索コストの最小化

Figure 3 に OH-SSL の階層構造を示す．本章では図中の探索点数を最小化する最適な K と $d_{x[k]}$ について述べる．図より，k 階層目の探索点数 $g(k)$ は $g(k) = \frac{2d_{x[k-1]}}{d_{x[k]}}$ と求まる．従って，階層数を K とした時の全探索点数は以下となる．

$$G(K) = \sum_{k=1}^{K} g(k) = \sum_{k=1}^{K} \frac{2d_{x[k-1]}}{d_{x[k]}} \tag{13}$$

$G(K)$ を最小化する粒度を求める．$K=2$ の場合，$d_{x[1]}$ のみが変数となるため，$G(2)$ は $\frac{\partial G(2)}{\partial d_{x[1]}} = 0 \Rightarrow d_{x[1]} = \sqrt{d_{x[0]} d_{x[2]}}$ の時に最小となる．この時，$g(1) = g(2) = \sqrt{d_{x[0]}/d_{x[2]}}$ となるため，各階層の粒度が等しい時に $G(K)$ が最小となる．$K > 2$ も同様に，各階層の粒度が等しい時に $G(K)$ が最小となる．$g(1) = g(2) = \cdots = g(K)$ の条件下で $G(K)$ は以下のように求まる．

$$\tilde{G}(K) = K \left(\frac{d_{x[0]}}{d_{x[K]}} \right)^{\frac{1}{K}} \tag{14}$$

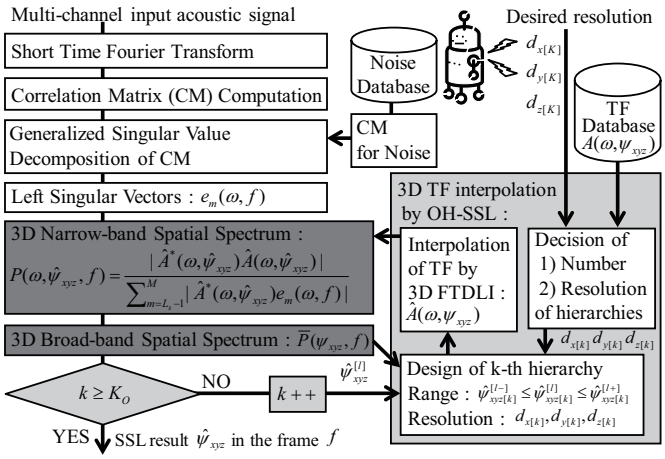


Figure 4: Block Diagram of Noise-robust Real-time 3D Super-resolution SSL

この時，$G(K)$ は最小となる．また，$\tilde{G}(K)$ を最小化する K は，$\frac{\partial \tilde{G}(K)}{\partial K} = \left(\frac{d_{x[0]}}{d_{x[K]}} \right)^{\frac{1}{K}} \left(1 - \frac{1}{K} \log \left(\frac{d_{x[0]}}{d_{x[K]}} \right) \right) = 0$ で以下のように求まる．

$$K = \log \left(\frac{d_{x[0]}}{d_{x[K]}} \right) \tag{15}$$

最後に，各階層の粒度 $d_{x[k]}$ は以下となる．

$$d_{x[k]} = d_{x[0]}^{\frac{K-k}{K}} d_{x[K]}^{\frac{k}{K}} \tag{16}$$

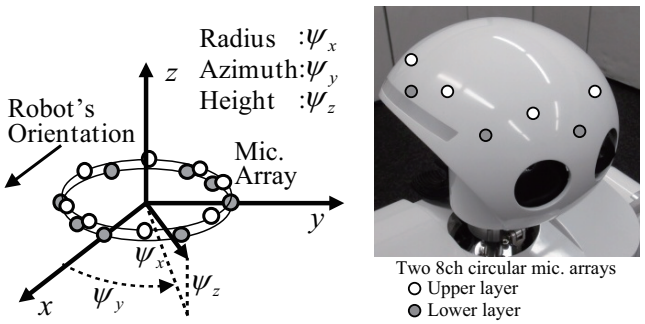
OH-SSL では式 (15)-(16) の K と $d_{x[k]}$ を音源探索の階層化に使用する．以降，式 (15) の K を K_O と表記する．

5 システム構成

Figure 4 に GSVD-MUSIC, FTDLI, OH-SSL を適用した三次元音源定位システムの処理フローを示す．評価ではこのシステムをオープンソースのロボット聴覚ソフトウェア HARK[19] 上に実装し，2.0GHz の Intel Core i7 の CPU を持つ計算機で実時間動作することを確認した．本稿では，マイクロホンアレイを搭載したロボットを残響時間が 0.2 秒 (RT_{20}) の 7m×4m の部屋の中央に，x 軸方向を向くように設置した．マイクロホンアレイは Figure 5(b) のように 8 チャネルの円状アレイが二つ上下に重なった形状の 16 チャネルのものを用いた．$\boldsymbol{A}(\omega, \psi_{xyz})$ は円筒座標上に事前に計測した（距離 1m，方位角 5° 毎，高さ -0.4m〜0.4m を 0.2m 毎に計測）．Figure 5(a) のように ψ_x, ψ_y, ψ_z の各軸はそれぞれ，円筒座標系の半径，方位角，高さとした．入力音響信号は 16kHz, 16 ビットでサンプリングした．音響信号処理のフレーム長とシフト長はそれぞれ，512, 160 サンプルとした．

6 評価実験

本章では，以下の実験を行い，各機能の評価を行う．



(a) Coordinates for SSL (b) Microphone Array

Figure 5: Conditions for Experiment

D1) GSVD-MUSIC の雑音ロバスト性評価

D2) FTDLI の伝達関数の補間性能評価

D3) OH-SSL の音源探索コスト評価

D4) ロボットへの応用と音源定位性能評価

D1) では，SEVD-MUSIC[2], GEVD-MUSIC[14], GSVD-MUSIC の 3 手法を比較し，提案手法の雑音ロバスト性を評価する．D2) では，FDLI, TDLI, FTDLI の 3 手法の補間誤差を比較し，統合の有効性を確認する．D3) では，OH-SSL といくつかの既存手法の音源探索回数を比較し，提案手法の有効性を検証する．最後に，D4) において，Figure 4 の音源定位システムを実際のロボットに適用する．評価の簡略化のため，D2) と D4) の評価では二次元空間（方位角と仰角）を，D1) と D3) の評価では一次元空間（方位角）を対象とした．

6.1 GSVD-MUSIC の雑音ロバスト性評価

実験には，60° に目的音（白色雑音）を，180° に雑音（定常状態でのロボットファン雑音）を配置した．ロボットと各音源との距離は 1[m] とした．また，事前情報のステアリングベクトル $\boldsymbol{A}(\omega, \psi_{xyz})$ の方位角は 5° ごとに計測し，使用した．式 (1) の相関行列の平均化のためのフレーム数はそれぞれ，$T_R = 25$ とした．

評価指標には，信号対雑音比 (Signal-to-Noise Ratio, SNR) 及び，定位正解率 (SSL Correct Rate) を用いた．SNR は以下のように定義した．

E1) M 個のマイクロホンの平均入力音響信号のスペクトル $X_{s_a}(\omega)$ の，パワースペクトル密度 (PSD) $P_{s_a}(\omega)$ を求める（k_{wl} は窓長）．$P_{s_a}(\omega) = \frac{1}{k_{wl}} X_{s_a}(\omega) X^*_{s_a}(\omega)$.

E2) 雑音についても同様に PSD を求める $P_{n_a}(\omega)$.

E3) 音源定位で使用する周波数帯に相当する周波数ビンを用いて平均 SNR を求める．

$$\mathrm{SNR} = 10\log_{10}\left(\frac{1}{j_h - j_l + 1}\sum_{j=j_l}^{j_h}\frac{P_{s_a}(\omega_{[j]})}{P_{n_a}(\omega_{[j]})}\right) \qquad (17)$$

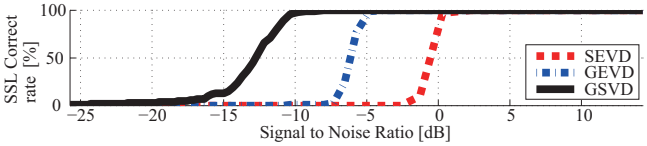


Figure 6: SSL correct rate of SEVD-, GEVD-, and GSVD-MUSIC

定位で用いた周波数帯は 500[Hz]〜2800[Hz] とした．

また定位正解率は，各フレームにおいて，空間スペクトル（式 (4)）が最大となる方向を音源数個取得し，その方向が目的音の方向から 10° 以内に入っているかを判定し，100 フレーム中で正解と判定されたフレーム数の割合として定義した．

Figure 6 に定位性能の比較結果を示す．図の横軸は SNR を，縦軸は定位正解率を表す．図中の実線は GSVD-MUSIC を，鎖線は GEVD-MUSIC を，点線は SEVD-MUSIC の結果を表す．図より，GSVD-MUSIC が GEVD-MUSIC や SEVD-MUSIC に比べ，低い SNR で高い定位精度を維持（SEVD-MUSIC に比べて約 10dB，GEVD-MUSIC に比べて約 5dB の向上）していることから，GSVD-MUSIC の高い雑音ロバスト性が確認できた．

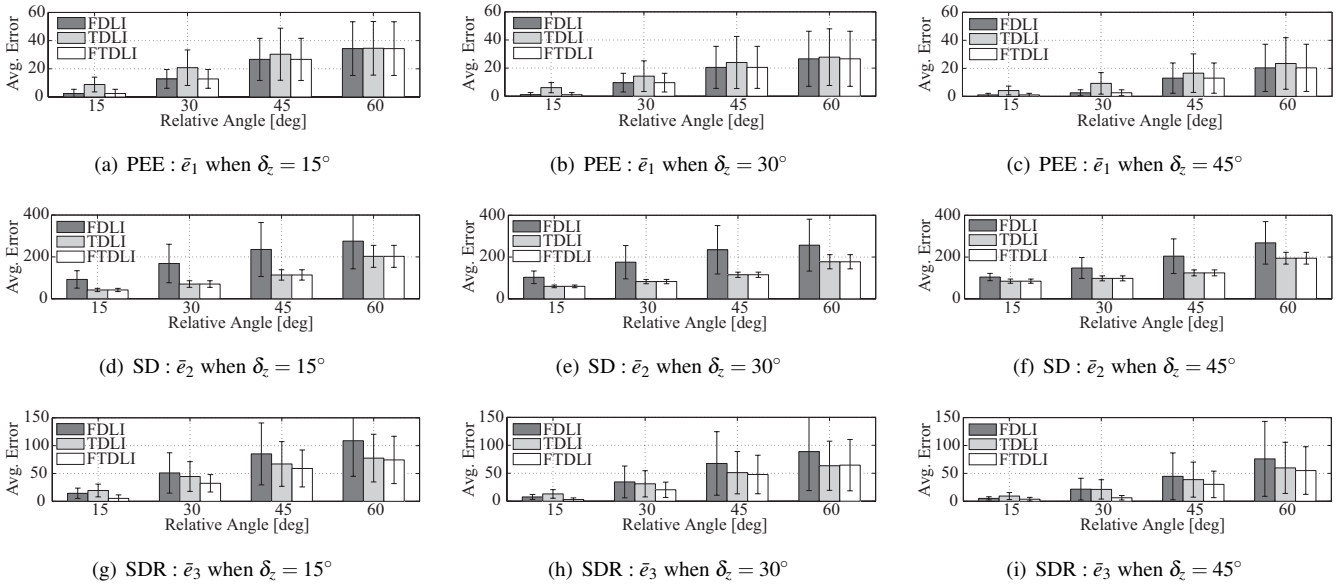
6.2 FTDLI の伝達関数の補間性能評価

FDLI, TDLI, FTDLI による三次元伝達関数の補間誤差を評価し，統合の有効性を確認した．本稿では，FTDLI の汎用性を確認するため，板倉らの伝達関数データベース [18]でまず評価し，次にロボットの事前計測伝達関数を用いて評価を行う．

板倉らの伝達関数データベースは両耳音響信号処理のための頭部伝達関数として知られており，球状の計測器によって録音されている．補間性能の評価のため，三次元座標を球座標（ψ_x, ψ_y, ψ_z をそれぞれ，半径，方位角，仰角とする．）で定義する．実験では，ψ_x=1.2m に固定し，ψ_y と ψ_z を変化させたときの $\hat{\boldsymbol{A}}(\omega, \psi_{xyz})$ と $\boldsymbol{A}(\omega, \psi_{xyz})$ の誤差を評価した．補間に使用する 4 つの事前計測伝達関数の位置 ($\psi_{\underline{xyz}}$, $\psi_{\underline{xy}\bar{z}}$, $\psi_{\underline{x}\bar{y}\underline{z}}$, $\psi_{\underline{x}\bar{y}\bar{z}}$) は，それらがなす中点が $[\psi_x, \psi_y, \psi_z] = [1.2\mathrm{m}, 0°, 0°]$ となるように取った．すなわち，4 点は，$[\psi_x, \psi_y, \psi_z] = [1.2\mathrm{m}, \pm\delta_y, \pm\delta_z]$ と表せる．ただし，δ_y と δ_z はそれぞれ，事前計測伝達関数の方位角と仰角である．評価では，$\delta_y = \{15°, 30°, 45°, 60°\}$ と，$\delta_z = \{15°, 30°, 45°\}$ を用いた．$\hat{\boldsymbol{A}}(\omega, \psi_{xyz})$ を ψ_y と ψ_z について 5° ごとに推定し，計測して得た 5° 毎の伝達関数との誤差を次式で表される加算平均により求めた．

$$\bar{e} = \frac{1}{i_\psi}\sum_{i=1}^{i_\psi}\frac{1}{j_h - j_l + 1}\sum_{j=j_l}^{j_h} f(\omega_{[j]}, \psi_{xyz[i]}) \qquad (18)$$

ここで，$f(\omega_{[j]}, \psi_{xyz[i]})$ は，補間点 $\psi_{xyz[i]}$ での $\omega_{[j]}$ に相当する周波数ビンの補間誤差である．i_ψ は評価に用いた補間

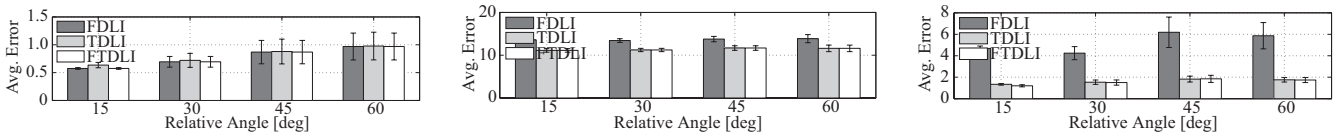


(a) PEE : $\bar{e}_1$ when $\delta_z = 15°$ (b) PEE : $\bar{e}_1$ when $\delta_z = 30°$ (c) PEE : $\bar{e}_1$ when $\delta_z = 45°$

(d) SD : $\bar{e}_2$ when $\delta_z = 15°$ (e) SD : $\bar{e}_2$ when $\delta_z = 30°$ (f) SD : $\bar{e}_2$ when $\delta_z = 45°$

(g) SDR : $\bar{e}_3$ when $\delta_z = 15°$ (h) SDR : $\bar{e}_3$ when $\delta_z = 30°$ (i) SDR : $\bar{e}_3$ when $\delta_z = 45°$

Figure 7: Interpolation error of PEE, SD, SDR using Itakura's TF Database[18]



(a) PEE : $\bar{e}_1$ when δ_z = 0.2m (b) SD : $\bar{e}_2$ when δ_z = 0.2m (c) SDR : $\bar{e}_3$ when δ_z = 0.2m

Figure 8: Interpolation error of PEE, SD, SDR using TFs of a robot-embedded microphone array

点個数であり，$-\delta_y < \psi_y < \delta_y$ と $-\delta_z < \psi_z < \delta_z$ の範囲内の事前計測伝達関数 $\boldsymbol{A}(\omega, \psi_{xyz})$ の個数で決まる．

式 (18) の $f(\omega_{[j]}, \psi_{xyz[i]})$ の誤差指標として，位相推定誤差 (PEE)，スペクトル歪み (SD)，信号対歪み比 (SDR) を用いた．PEE は以下で表される位相誤差指標である．

$$f_1(\omega, \psi_{xyz}) = \sum_{m=1}^{M} \left| \frac{A_m(\omega, \psi_{xyz}) \cdot \hat{A}_m(\omega, \psi_{xyz})}{|A_m(\omega, \psi_{xyz})||\hat{A}_m(\omega, \psi_{xyz})|} - 1 \right| \quad (19)$$

SD は以下で表され，振幅誤差を示す．

$$f_2(\omega, \psi_{xyz}) = \sum_{m=1}^{M} \left| 20 \log \frac{|\hat{A}_m(\omega, \psi_{xyz})|}{|A_m(\omega, \psi_{xyz})|} \right| \quad (20)$$

SDR は以下で表され，伝達関数自体の誤差を示す．

$$f_3(\omega, \psi_{xyz}) = \sum_{m=1}^{M} \frac{|A_m(\omega, \psi_{xyz}) - \hat{A}_m(\omega, \psi_{xyz})|^2}{|A_m(\omega, \psi_{xyz})|^2} \quad (21)$$

以下，$\bar{e}_1, \bar{e}_2, \bar{e}_3$ をそれぞれ，PEE, SD, SDR を用いた時の $\bar{e}$ とする．

Figure 7 に δ_z を変化させた時の $\bar{e}_1, \bar{e}_2, \bar{e}_3$ を示す．図の横軸は δ_y を表す．図より，FDLI が振幅誤差にあたる SD において，TDLI が位相誤差にあたる PEE において，それぞれ補間性能が劣化していることがわかる．FTDLI は FDLI と TDLI を統合することにより，$\bar{e}_1, \bar{e}_2, \bar{e}_3$ の全てにおいて最小の誤差となった．

Table 1: Comparison of computational cost

Condition		$G(K)$				
$d_{x[0]}$	$d_{x[K]}$	H1	H2	H3	SG	OS
360	10.0	36	11	9	11	8
360	1.0	360	36	21	17	12
360	0.1	3600	120	45	25	16
360	0.01	36000	1200	213	37	26

同様にロボットの伝達関数についても評価を行った．板倉らのデータベースと異なり，事前計測伝達関数を円筒座標系で計測したため，ψ_z を音源高さとして定義した．Figure 8 に，δ_z=0.2m に固定した時の，$\bar{e}_1, \bar{e}_2, \bar{e}_3$ の比較を示す．図より，FTDLI が全ての誤差指標において最小の誤差となり，統合の有効性を確認できた．

6.3 OH-SSL の音源探索コスト評価

計算コスト評価のため，OH-SSL と既存手法による総探索数 $G(K)$ を比較した．2 章より，$d_{x[0]}$ と $d_{x[K]}$ を変化させた時の $G(K)$ を計算した．評価対象として，以下の 5 つの手法を比較した．

H1) 階層化を行わない手法

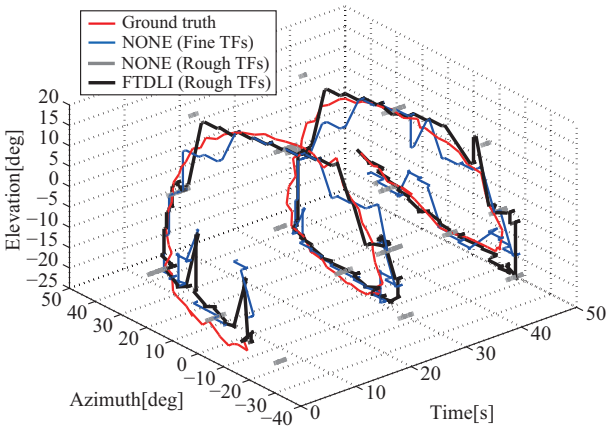


Figure 9: Trajectory of the 3D Localization

H2) 2 階層（各階層の粒度は式 (16) で決定）

H3) 3 階層（各階層の粒度は式 (16) で決定）

SG) Spherical grid 法[10]

OS) OH-SSL（K と $d_{x[k]}$ を式 (15)-(16) で決定）

Table 1 に求まった $G(K)$ を示す．表より，全ての所望の解像度に対して OS が $G(K)$ を最小化していることがわかり，有効性を確認できた．

実際の音源定位計算コスト評価のため，1000 フレーム分の処理を行い，音源定位の計算にかかる平均処理時間を算出した．所望の空間解像度を 1° に設定し，OH-SSL の有無 (OS と H1) による平均処理時間を比較したところ，それぞれ，5.2[ms] と 23.4[ms] となり，78%の計算コストを削減できた．また，フレーム周期である 10ms 以下で処理が実現できていることから，フレーム毎実時間処理を達成できた．

6.4 ロボットへの音源定位の応用と定位性能評価

FTDLI の有無による動的音源に対する音源定位の方向推定誤差を評価する．本稿では，1 つの白色雑音源を動かして動的音源とした．評価対象として，以下の 3 条件を比較した．

F1) 5° ごとの方位角，0.2m ごとの高さの事前計測伝達関数を用いた音源定位（高い空間解像度）

F2) 30° ごとの方位角，0.4m ごとの高さの事前計測伝達関数を用いた音源定位（低い空間解像度）

F3) F2) を用いて，5° ごとの方位角，0.01m ごとの高さの伝達関数を FTDLI を用いて推定した時の音源定位

また，リファレンスデータを得るため，超音波タグ位置測位システムを別途用いた（計測精度：20〜80mm）．Figure 9 に比較結果を示す．図の x 軸，y 軸，z 軸はそれぞれ，時間，推定された音源方位角 ψ_y，推定された音源仰角 ψ_z を示す．図中の赤線，青線，灰色線，黒線はそれぞれ，リファレンスデータ，F1 の結果，F2 の結果，F3 の結果を示す．F2(灰色線) の結果より，事前計測伝達関数の空間解像度が粗い場合，音源軌道が切断されているのが確認できる．また，F2 はリファレンスデータから離れた位置の音源として推定されており，推定結果とリファレンスデータとの平均方向誤差 $\bar{e}_\psi$ は $\bar{e}_\psi = 10.5°$ となった．F1(青線) は連続した軌道が確認できるものの，軌道にばらつきがあり，$\bar{e}_\psi = 7.2°$ となった．これは事前計測伝達関数の ψ_z の解像度が十分でなかったためと考えられる．F3(黒線) は，FTDLI により ψ_z 軸上に高い解像度の伝達関数を生成できるため，$\bar{e}_\psi = 6.5°$ となり，F1 に比べてばらつきの少ない軌道を実現できた．このことから，FTDLI の三次元音源定位での有効性を実環境で確認することができた．

また，OH-SSL の有効性を検証するため，OH-SSL の有無による音源定位の平均処理時間を比較した．所望の解像度を ψ_y 軸上で 5°，ψ_z 軸上で 0.01m に設定し，1000 フレーム分の定位を行い，平均処理時間を算出した．結果，OH-SSL の有無による平均処理時間はそれぞれ，0.028 秒，1.073 秒となった．これは，およそ 97%の処理時間の削減に相当し，OH-SSL の有効性を確認できた．

7 結論

本稿では，ロボットによる三次元音源定位について以下の問題点を扱った．

1) 実環境下のロボットが持つ自己雑音等の大きなパワーを持つ雑音に対するロバスト性

2) 音源定位の三次元空間での十分な空間解像度

3) 三次元空間内の音源探索の実時間性

1) に対して GSVD-MUSIC による音源定位における雑音の白色化を，2) に対してトリリニア補間を用いた FTDLI による三次元伝達関数の空間解像度の向上を，3) に対して OH-SSL による音源探索コストの軽減を提案した．評価では，1) GSVD-MUSIC が，SEVD-MUSIC に比べて約 10dB，GEVD-MUSIC に比べて約 5dB の SNR を向上すること，2) FTDLI が既存の補間手法に比べて誤差の小さな推定を実現すること，3) OH-SSL が音源探索数を約 78%軽減することを確認できた．最後に，これらの手法を実環境下のロボットに適用し，FTDLI による超解像音源定位と OH-SSL による実時間音源定位を確認でき，提案法の有効性を確認した．

今後の課題として，音源定位から得られる三次元位置情報を，画像から得られる定位情報と統合することによる定位のロバスト性の向上などが考えられる．

参考文献

[1] K. Nakadai *et al.*, “Active Audition for Humanoid”, in *Proc. of AAAI-2000*, pp. 832–839, 2000.

[2] R. Schmidt, “Multiple emitter location and signal parameter estimation”, *IEEE Trans. Ant. Prop.*, vol. 34, no. 3, pp. 276–280, 1986.

[3] K. Nakadai *et al.*, “Robust Tracking of Multiple Sound Sources by Spatial Integration of Room and Robot Microphone Arrays”, in *Proc. of IEEE ICASSP*, vol. IV, pp. 929–932, 2006.

[4] F. Asano *et al.*, “Real-time sound source localization and separation system and its application to automatic speech recognition”, in *Proc. of EUROSPEECH-2001*, pp.1013–1016.

[5] S. Argentieri and P. Danés, “Broadband variations of the MUSIC high-resolution method for sound source localization in Robotics”, in *Proc. of IEEE/RSJ IROS*, pp. 2009–2014, 2007.

[6] K. Nakamura *et al.*, “Real-time Super-resolution Sound Source Localization for Robots,” in *Proc. of 2012 IEEE/RSJ IROS*, accepted.

[7] K. Nakadai *et al.*, “A robot referee for rock-paper-scissors sound games”, in *Proc. of IEEE ICRA*, pp. 3469–3474, 2008.

[8] H. Nakashima *et al.*, “A Localization Method for Multiple Sound Sources by Using Coherence Function”, in *Proc. of 18th EUSIPCO*, pp. 130–134, 2010.

[9] B. Rudzyn *et al.*, “Real time robot audition system incorporating both 3D sound source localization and voice characterization”, in *Proc. of IEEE ICRA*, pp. 4733–4738, 2007.

[10] J.-M. Valin *et al.*, “Localization of simultaneous moving sound sources for mobile robot using a frequency- domain steered beamformer approach”, in *Proc. of IEEE ICRA*, vol. 1, pp. 1033–1038, 2004.

[11] J.-S. Hu *et al.*, “Simultaneous localization of mobile robot and multiple sound sources using microphone array”, in *Proc. of IEEE ICRA*, pp. 29–34, 2009.

[12] C. T. Ishi *et al.*, “Evaluation of a MUSIC-based real-time sound localization of multiple sound sources in real noisy environments”, in *Proc. of IEEE/RSJ IROS*, pp. 2027–2032, 2009.

[13] D. Ringach, “Look at the big picture (details will follow)“, *Nature Neuroscience*, vol. 6, no. 1, pp. 7–8, 2003.

[14] K. Nakamura *et al.*, “Intelligent Sound Source Localization for Dynamic Environments”, in *Proc. of IEEE/RSJ IROS*, pp. 664–669, 2009.

[15] T. Nishino *et al.*, “Interpolating Head Related Transfer Functions in the median plane”, in *Proc. of IEEE WASPAA*, pp. 167–170, 1999.

[16] M. Matsumoto *et al.*, “A method of interpolating binaural impulse responses for moving sound images”, in *Acoust. Sci. & Tech.*, vol. 24, no. 5, pp. 284–292, 2003.

[17] W. G. Gardner and K. D. Martin, “HRTF measurements of a KEMAR”, *J. Acoust. Soc. Am.*, vol. 97, pp. 3907–3908, 1995.

[18] T. Nishino *et al.*, ”Interpolating head related transfer functions,” in *Proc. of 7th WESTPRAC VII*, 1A-1-3, pp.293-296, 2000.

[19] K. Nakadai *et al.*, “Design and Implementation of Robot Audition System HARK”, *Advanced Robotics*, vol. 24, pp. 739–761, 2009.

人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-B202-3

両耳間レベル差を用いた音源定位における耳孔位置の最適化

Optimization of the ear canal position for sound localization using interaural level difference

木元大輔，公文誠
Daisuke KIMOTO and Makoto KUMON
熊本大学
Kumamoto University
d.kimoto@ick.mech.kumamoto-u.ac.jp

Abstract

耳介のあるバイノーラル聴覚システムでは，両耳間レベル差を用いた音源定位によって正中面内の音源の上下方向の定位を行うことが可能とされている．この場合，限られたマイクロホン数の制限と，マイクロホンの配置の影響で，両耳間レベル差によっては定位の困難な，または定位精度が著しく低下してしまう方向が生じる問題がある．このような方向はマイクロホン位置や頭部形状などの影響を受けるため，マイクロホンの配置を適切に選ぶことで，定位性能を改善できる可能性がある．

本研究では，マイクロホン周辺に取り付けた耳介の音響特性を考慮して，耳孔位置を最適化することで両耳間レベル差を用いた音源定位性能の向上を図る．実験を通して特に最適化しない耳介を用いた場合と比較し，最適化した耳介を用いた場合，音源定位性能が向上することが確かめられた．

1　はじめに

ロボットが自動で柔軟に適切な対応を行うためには，周辺環境の認識がとても重要である．この周辺環境の認識を行う方法の一つして，音信号を利用する方法がある．これは，ロボット聴覚として盛んに研究されている[奥乃, 2010]．その中でも，音の到来方向や音源方向を正確に推定することは，将来，ロボットが人と共生する上で必要な能力であり，重要な基礎機能と言える．

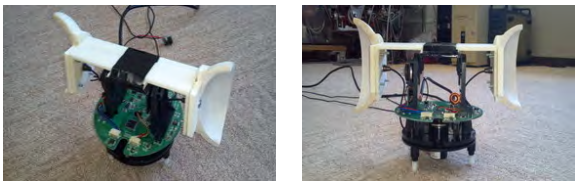
人間や動物は 2 つの耳のみで現実的な音源定位を実現しており，音源定位のための必要最小限の構成だと考えられる．ロボットでも 2 つのマイクロホンだけで音源定位を実現することは，聴覚システムの簡素化や音源定位の原理の解明など，興味深い課題を含んでいると言える．

また，人間や動物では，耳に耳介と呼ばれる音の反射・集音を果たす器官が存在している．耳介形状は一般的には複雑なため，耳介の音響特性が音の到来方向に応じて異なることが知られている[E.A.Lopez, 1976]．特に，耳介のゲイン特性が顕著に抑制されている帯域を耳介ノッチ[E.A.G.Shaw, 1968]と呼ぶが，耳介ノッチの周波数が音源方向の関数になっているので，耳介ノッチの検出を行えば音源方向を推定することが可能である．実際に，章ら[章, 2008]は，音源方向の推定のための特徴量として両耳間レベル差 (Interaural Level Difference，ILD) を用いて，事前に学習した特徴量との相関を求めることで仰伏角および方位角を推定している．

本研究では，耳介に埋蔵するマイクロホン位置を適切に選ぶことでこの方法の改善を試みる．このため，音源から直接マイクロホンに収録される直接波，および耳介上で反射し収録される信号（2 回反射まで）を考慮した耳介の周波数応答モデルを考え，数値的に最適な耳介を設計することとした．最適化には特に ILD が音源方向毎に十分異なるような指標を考え，マイクロホン位置を最適化パラメータとした．なお，本研究では定位において ILD が重要とされる仰伏角方向に着目するため，モデルでの音源は正中面内に位置するものを考える．

2　バイノーラル聴覚ロボット

本研究では Figure1 に示すバイノーラル聴覚ロボットを用いる．2 つのマイクロホンには同じ形状の耳介に取り付けられているが，その取り付け位置は後述するように異なることがある．本研究では, 左右のマイクロホンで受聴される信号が音源方向やマイクロホン位置によって異なる周波数伝達特性の影響を受けると考え，それを利用することを想定している．



(a) ロボット頭部　　(b) ロボット正面

Figure 1: バイノーラル聴覚ロボット

ロボットはFigure2に示すような方位，仰伏角の2方向に動作可能である．使用しているモータが超音波モータのため, ロボットが動作中であっても可聴域でのエゴノイズはほとんど生じない．ロボットを制御する計算機にはPCを使用する．PCは姿勢角および角速度の規範値を指定し，シリアル通信によりモータを制御する．また，PCのマイク端子よりマイクロホンが受信した音信号を44.1[kHz]で左右の2chともサンプリングする．

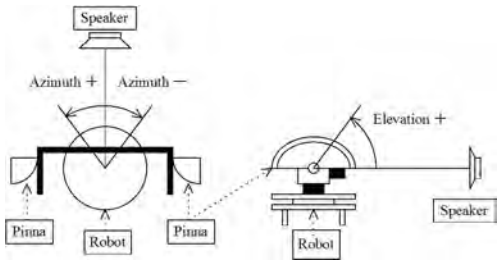


Figure 2: ロボット回転方向

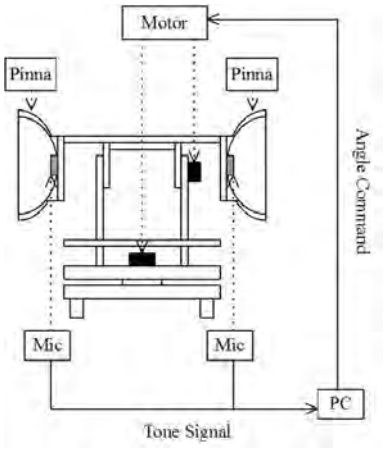


Figure 3: システム

3　両耳間レベル差

環境やロボット身体の影響によりロボットの受聴する音信号は原信号とは異なったものとなる．今，ロボットを基準とした音源までの距離，方位角，仰伏角をそれぞれr，θ，ϕとすると左右のマイクロホンへの伝達関数のうちロボット身体によるものは$H_l(\theta,\phi;\omega)$，$H_r(\theta,\phi;\omega)$，環境によるものは$H_{le}(r,\theta,\phi;\omega)$，$H_{re}(r,\theta,\phi;\omega)$と表すことができる．また，原信号$s_O(\omega)$に対してロボットの左右のマイクロホンで受聴する音信号をそれぞれ$s_l(r,\theta,\phi;\omega)$,$s_r(r,\theta,\phi;\omega)$とすればFigure4に示すように

$$\begin{aligned} s_l(r,\theta,\phi;\omega) &= H_l(\theta,\phi;\omega)H_{le}(r,\theta,\phi;\omega)s_O(\omega) \\ s_r(r,\theta,\phi;\omega) &= H_r(\theta,\phi;\omega)H_{re}(r,\theta,\phi;\omega)s_O(\omega) \end{aligned}$$

の関係がある．

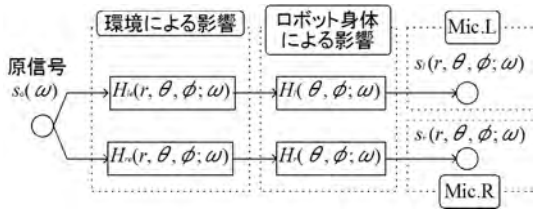


Figure 4: 受聴される音信号

もし環境の影響が$H_{le} \fallingdotseq H_{re}$であれば，両耳間レベル差 (Internal Level Difference，ILD)Sは

$$\begin{aligned} S &\equiv 20\log|s_l(r,\theta,\phi;\omega)| - 20\log|s_r(r,\theta,\phi;\omega)| \\ &= 20\log|H_l(\theta,\phi;\omega)H_{le}(r,\theta,\phi;\omega)s_O(\omega)| \\ &\quad -20\log|H_r(\theta,\phi;\omega)H_{re}(r,\theta,\phi;\omega)s_O(\omega)| \\ &\fallingdotseq 20\log|H_l(\theta,\phi;\omega)| - 20\log|H_r(\theta,\phi;\omega)| \end{aligned}$$

と近似でき，ILDSが音源方向θ,ϕの関数として$S(\theta,\phi;\omega)$となり，ロボット身体の影響のみで特徴づけられることになる．特に耳介はマイクロホン近傍にあるため，伝達特性に大きく影響していると考えられ，音源定位の情報を与えると期待される．

4　両耳間レベル差を用いた音源定位法

ILD$S(\theta,\phi;\omega)$が音源方向によって特徴づけられたものであるので, 本研究では対象とする周波数帯域のILDを特徴量ベクトルの1つとする．なお，原信号s_Oに含まれていない (あるいは非常に小さい) 周波数成分ついてはSが正しく求まらない可能性があるため, 除外して考える必要がある．このため，適当な正定数εに対して

$$f(x,a,b) = \begin{cases} 0 & (\text{if } a<\varepsilon \text{ or } b<\varepsilon) \\ x & (\text{otherwise}) \end{cases}$$

となる関数fを用いて，特徴量ベクトルS_{ILD}を

$$\begin{aligned} S_{\mathrm{ILD}} &= [f(\Delta S(\omega_1),|s_l(\omega_1)|,|s_r(\omega_1)|), \\ &\quad \cdots, f(\Delta S(\omega_N),|s_l(\omega_N)|,|s_r(\omega_N)|)]^T \end{aligned}$$

と定める．ここで$\omega_1,\cdots,\omega_N$は対象とする周波数成分である．

本手法では音源方向 θ，ϕ から周波数成分を十分に含んでいる音信号を与え，ILD の特徴ベクトルを事前に計測し，学習ベクトル $S_d(\theta,\phi)$ として保存する．このような推定対象となる方向を $\theta_1,\cdots,\theta_{N_\theta},\phi_1,\cdots,\phi_{N_\phi}$ とすると，$N_\theta \times N_\phi$ 点の特徴ベクトルを学習ベクトルとしてデータベースに保存することになる．

次に方向を推定したい音信号が与えられ，これの特徴ベクトル S が得られたとする．この時 S と同じ方向から得られた学習ベクトルは，S との間で高い相関を示すと考えられる．相関は以下のように内積として表わされる．

$$X_{\mathrm{ILD}}(S,S_d) = \frac{< S,S_d >}{|S||S_d|} \tag{1}$$

ここで，X_{ILD} は特徴ベクトル S が与えられたとき θ,ϕ の関数として求まるので $X_{\mathrm{ILD}}(S,S_d) = X_{\mathrm{ILD}}(S,\theta,\phi)$ となり，音源方向の推定値として X_{ILD} を最大とするものを用いる．

本研究では，そのまま学習ベクトルを使用するのではなく，学習ベクトルデータベースの擬似逆行列を使用する．これにより，単純に学習ベクトルを使用するのに比べ方向推定性能が改善されると期待できる[公文, 2011]．

5　耳介の周波数応答モデル

ここでは，マイクロホン取り付け位置が ILD に与える影響を調べるため，耳介の音響特性を与える周波数応答モデルを考える．

5.1　耳介の周波数応答モデル

$H(\boldsymbol{P}_s,\omega)$ を $\boldsymbol{P}_s$ に位置する音源からの角周波数 ω の伝達特性 (ゲイン) とし，正弦波でのモデル (音源が発している音信号を $\sin\omega t$ と仮定) を考える．

音源から直接マイクロホンに収録される信号 (直接波) は，音信号の振幅が距離により減衰するので，マイクロホン位置では以下のように表すことができる．

$$r(\boldsymbol{P}_s,\omega,t) = \frac{1}{l}\sin(\omega(t-\frac{l}{V_s})) \tag{2}$$

ここで，l は音源からマイクロホンまでの距離，V_s は音速を表している．

次に，耳介上で反射し，その後マイクロホンに収録される音信号を考える．反射点までの音信号は直接波と同様に考えると，耳介の反射点での信号は

$$\frac{1}{l_{s1}}\sin(\omega(t-\frac{l_s}{V_s})) \tag{3}$$

となる．ここで l_{s1} は音源位置から反射点までの距離を表す．反射後の信号は反射点を音源位置とする (3) 式の音信号と考えられ，すると，反射した音信号はマイクロホン位置において

$$r_{r1}(\boldsymbol{P}_s,\omega,t) = \frac{\lambda(\omega)}{l_{s1}l_{m1}}\sin(\omega(t-\frac{l_s+l_m}{V_s})) \tag{4}$$

と考えられる．ここで $\lambda(\omega)$ は耳介での反射率，l_{m1} は反射点からマイクロホンまでの距離を表す．

同様に，耳介や耳介の周りで反射したのち，再び耳介で反射しその後マイクロホンに収録される 2 回反射を考える．これは 1 回反射の時と同様に考え

$$r_{r2}(\boldsymbol{P}_s,\omega,t) = \frac{\lambda^2(\omega)}{l_{s2}l_{e2}l_{m2}}\sin(\omega(t-\frac{l_{s2}+l_{e2}+l_{m2}}{V_s})) \tag{5}$$

のように表せる．ここで l_{s2} は音源位置から反射点 1 までの距離，l_{e2} は反射点 1 から反射点 2 までの距離，l_{m2} は反射点 2 からマイクロホンまでの距離を表す．

これらより，正弦波での応答モデルは以下のように示すことができる．

$$\begin{aligned} h(\boldsymbol{P}_s,\omega,t) = r(\boldsymbol{P}_s,\omega,t) &+ \iint_S r_{r1}(\boldsymbol{P}_s,\omega,t)dS \\ &+ \iint_S r_{r2}(\boldsymbol{P}_s,\omega,t)dS \end{aligned} \tag{6}$$

本研究では ILD に着目しているため，(6) 式のゲイン特性のみを考慮すればよい．したがって，(6) 式のゲイン特性を $H(\boldsymbol{P}_s,\omega)$ として用いる．

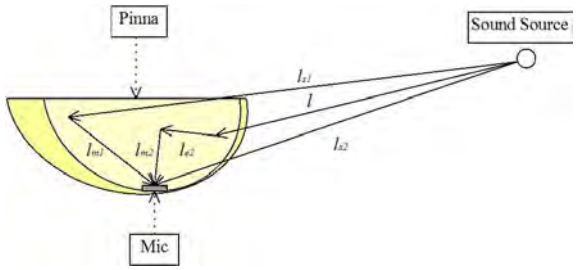


Figure 5: 反射モデル

5.2　モデルの検証

作成した耳介応答モデルの検証を行う．検証に用いた耳介形状は $x=40$[mm]，$y=30$[mm]，$z=20$[mm] であり，マイクロホン位置は原点の位置とした．

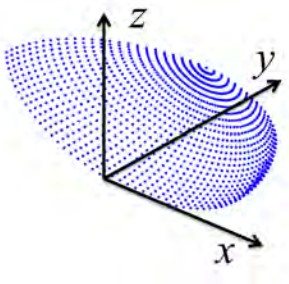


Figure 6: 耳介形状

また，音の収録は 9.4[m] × 3.45[m] × 2.8[m] の Figure7 に示すような部屋で行った．

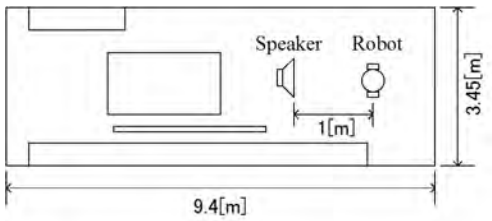
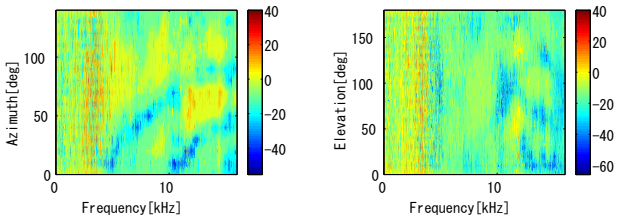


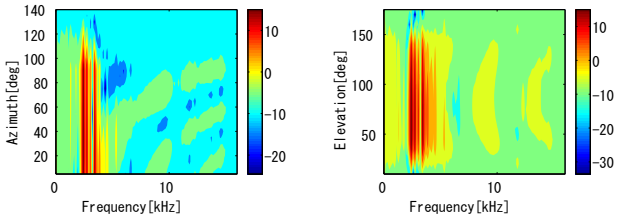
Figure 7: 収録環境

まず，収録を行い計測により求めた ILD 結果を Figure8 に示す．また，ILD をモデルより推定した結果を Figure9 に示す．なお，反射率 $\lambda(\omega)$ は実測した周波数応答に良く合うよう適宜調整を行った．



(a) 方位角方向　　(b) 仰伏角方向

Figure 8: 収録結果より求めた ILD



(a) 方位角方向　　(b) 仰伏角方向

Figure 9: 応答モデルより求めた ILD

計測により求めた ILD とモデルより推定した ILD を比較すると，ノッチの応答に酷似した傾向があることがわかる．

6　耳孔位置の最適化

6.1　耳孔位置の最適化法

左側の耳介の応答を H_l，右側の耳介の応答を H_r とし，5 節で導入したモデルを用いて ILD を求めることができる．

今，ILD を用いた音源定位では音源方向毎に S が大きく異なっていることが望ましい．このことから，音源 s_i，s_j の音源位置を $\boldsymbol{P}_{si}$，$\boldsymbol{P}_{sj}$ とすると，それらの音源方向の ILD 間の内積

$$X(\boldsymbol{P}_{si},\boldsymbol{P}_{sj}) = \frac{< S(\boldsymbol{P}_{si},\omega), S(\boldsymbol{P}_{sj},\omega) >}{|S(\boldsymbol{P}_{si},\omega)||S(\boldsymbol{P}_{sj},\omega)|} \tag{7}$$

について，すべての音源位置 $\boldsymbol{P}_{si}$，$\boldsymbol{P}_{sj}$ の組み合わせについて考える．ここで，ある値 $\epsilon(<1)$ よりも小さい $X(\boldsymbol{P}_{si},\boldsymbol{P}_{sj})$ が多い場合に，音源方向によって特性により違いがあると言えるので，本研究では $X(\boldsymbol{P}_{si},\boldsymbol{P}_{sj}) < \epsilon$ となる $X(\boldsymbol{P}_{si},\boldsymbol{P}_{sj})$ の集合を M_ϵ，また，M_ϵ の要素数を m_ϵ とし m_ϵ の多いものを適当なマイクロホン位置と考える．

$$M_\epsilon = \{X(\boldsymbol{P}_{si},\boldsymbol{P}_{sj}) | X(\boldsymbol{P}_{si},\boldsymbol{P}_{sj}) < \epsilon\} \tag{8}$$

$$m_\epsilon = |M_\epsilon| \tag{9}$$

なお，本研究では ILD が重要となる仰伏角方向に着目し，モデルでの音源位置を正中面内に限定し，音源の上下方向の変化に対応した ILD を対象に最適化を行った．

6.2　耳孔位置の最適化

応答モデルを用いて耳孔（マイクロホン）位置の最適化を行う．今回は，計算量の観点から最適化を行うマイクロホン位置を Figure11 に示すように 5[mm] 四方のメッシュ上に限定した．また，マイクロホンは耳介表面上にあるとして最適化を行った．

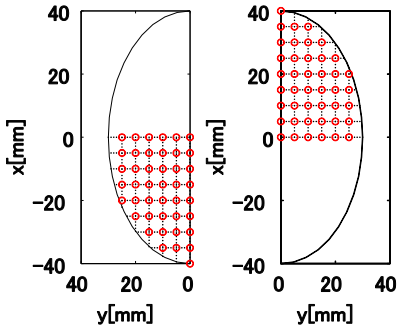


Figure 10: マイクロホン位置

$\epsilon = 0.985$ とし最適化計算を行った結果を Figure11 に示す．Figure11 は m_ϵ の値が大きいものの上位 6 つを示しており，色毎に左右のマイクロホン位置が対応している．

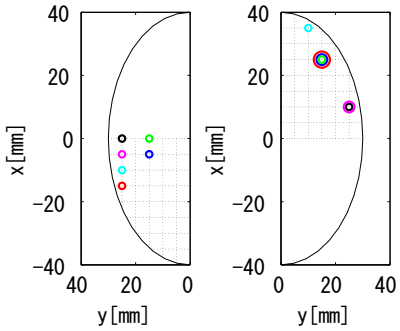


Figure 11: マイクロホン位置推定結果

Figure11 に示す結果から，マイクロホン位置が $x_l = 25[\mathrm{mm}], y_l = 15[\mathrm{mm}], x_r = -15[\mathrm{mm}], y_r = 20[\mathrm{mm}]$ 近傍にあることが望ましいということがわかる．

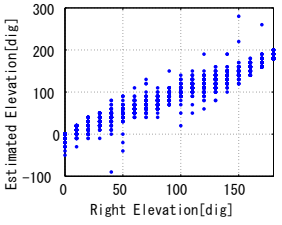
6.3 最適化結果の検証

6.2 節で求めた最適化結果の検証を行う．比較のために，

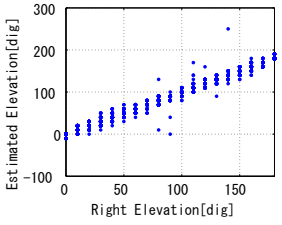
- Symmetric pinnae - マイクロホン位置が左右とも $x = 0[\mathrm{mm}], y = 0[\mathrm{mm}]$ のもの
- Non-optimal pinnae - マイクロホン位置が $x_l = 15[\mathrm{mm}], y_l = 10[\mathrm{mm}], x_r = -15[\mathrm{mm}], y_l = 10[\mathrm{mm}]$ のもの
- Optimal pinnae - 最適化結果 $x_l = 25[\mathrm{mm}], y_l = 15[\mathrm{mm}], x_r = -15[\mathrm{mm}], y_r = 20[\mathrm{mm}]$ のもの

とする．

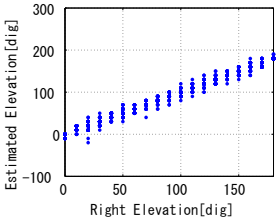
実際に，以上の位置にマイクロホンを有する耳介を作成し，音源定位を行った結果を Figure12 に示す．さらに，音源定位性能を比較した結果を Figure13 に示す．図中の縦軸は定位誤差の RMS を表す．



(a) Symmetric pinnae



(b) Non-optimal pinnae



(c) Optimal pinnae

Figure 12: 音源定位結果

Figure13 より，まずマイクロホン位置を単純に変化させた場合でも音源定位の性能に大きな影響があることがわかる．さらに，単純にマイクロホン位置を変化させた Non-optimal pinnae に比べ，最適化結果を用いた Optimal pinnae の方が誤差の値が小さいことから，最適化結果が比較したマイクロホン位置の中でもっとも音源定位性能が良く，Symmentric pinnae と比較すると 55.9%，Non-optimal pinnae と比較すると 21.5%の誤差を削減することができた．このことから，今回導入した最適化法の有効性が言える．

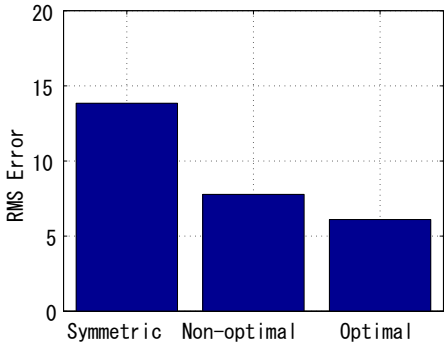


Figure 13: 最適化結果の評価 (仰伏角方向)

7 結言

本研究では，ILD を用いた音源定位において，耳介形状，特に耳孔位置の変更により定位性能の向上を試みた．その結果，耳孔位置の最適化を行うことにより，音源定位性能の向上を確認することができた．

今後は，今回導入した最適化法を耳介形状に適用させ，耳介形状の最適化を行いたい．

参考文献

[奥乃, 2010] 奥乃　博 : ロボット聴覚の現状と展望，日本ロボット学会誌 Vol.28 No.1，pp.2-5，2010.

[章, 2008] 章　忠，井 和章，三宅　哲夫，今村　孝，堀畑　聡 : バイノーラルモデルを用いた音源方向定位，日本機械学会論文集 C 編，Vol.74-739,pp.642-649,2008.

[C.Knapp, 1976] C.Knapp，G.Carter : The generalized Correlation Method for Estimation of Time Delay, IEEE Trans.on Acoustic Speech and Signal Process.，vol.4,pp.320-327,1976.

[E.A.Lopez, 1976] E.A.Lopez，Ray Meddis : A physical model of sound diffraction and reflections in the human concha，J.Acoust. oc.Am.Volume100，Issue 5，pp.3248-3259，1996.

[公文, 2011] 公文　誠，木元　大輔 : 耳介を持つバイノーラル聴覚ロボットの音源方向推定の検討，人工知能学会研究資料 SIG-Challenge-B102-11，2011.

[森下, 2009] 森下 巖, 小畑 秀文 : 信号処理，コロナ社，2009.

[E.A.G.Shaw, 1968] E.A.G.Shaw，R.Teranishi : Sound Pressure Generated in an External-Ear Replica and Real Human Ears by a Nearby Point Source，J.of Acoust.Soc.Am.44 1)，pp.240-249，1968.

人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-B202-4

複数ロボットによる音源定位結果を統合し発話者を特定するシステム

Speaker Identification System
Based on Sound Source Localization Results from Multiple Robots

中島大一, 駒谷和範, 佐藤理史
Taichi Nakashima, Kazunori Komatani, Satoshi Sato
名古屋大学大学院 工学研究科
Graduate School of Engineering, Nagoya University
{taichi_n, komatani, ssato}@nuee.nagoya-u.ac.jp

Abstract

Humanoid robots need to head toward human participants when answering to their questions in multi-party dialogues. Some positions of participants are difficult to localize from robots in multi-party situations, when the robots can only use their own sensors. We present a method of identifying who is speaking more accurately by integrating the multiple sound source localization results obtained from two robots. This method employs two robots and places them so as to compensate for each other's localization capabilities and then integrate their two results. Our experimental evaluation revealed that using two robots improved speaker identification compared with when only one robot was used. We furthermore implemented our method using humanoid robots and constructed a demo system.

1 はじめに

ヒューマノイドロボットを用いた複数人会話システムの開発を行っている．複数人会話システムとは，2 人以上のユーザと会話するシステムである[12].

今日までに開発されてきた複数人会話システムには 2 つの問題がある．1 つ目の問題は，多くのシステムが特殊なデバイスを利用しており，特定の環境でしか動作しないことである．複数人会話システムは，ユーザが何を発話したのかだけでなく，そのユーザがどこにいるのか，そして誰に対して発話したのかを特定する必要がある．このようなユーザの行動を検出，追跡するために，超音波センサ[7]や，高解像度 (HD) カメラ[4]，広角カメラ[2]といったデバイスが利用されている．もしくは，カメラやマイクロフォンが多く設置されたスマートルーム[5]が前提とされてきた．2 つ目の問題は，ユーザがシステムに対して何を発話すればよいのかが分からず，会話が中断してしまうことである．この問題は，単一のユーザを相手にする対話システムでも発生する[3]ため，複数人会話システムでも起こりうる問題である．

本研究では以下のアプローチでこれらの問題の解決を図る．

1. ロボットに搭載されたマイクとカメラのみを利用する．つまり，特殊なデバイスが準備された特別な環境を仮定しない．
2. 2 体のロボットを利用する．ユーザが発話を止めてしまったときに，ロボット同士で会話を行うことで会話を持続させる．さらに，ロボットに搭載されたカメラやマイクロフォンの能力は十分ではないため，2 体を用いることでそれを補う．

複数人会話を行う状況として，図 1 に示すような複数ユーザが机を囲み，その机の上に配置されたロボットと行う会話を想定する．この状況設定は，複数人会話システムにおけるユーザの位置の決定を簡単化するものである．つまり，複数のユーザが机を囲んだ状況では，ユーザの位置を机の周辺に限定できる．

本論文では，複数人会話システムの実現の第一歩として，発話したユーザがどこにいるかを特定する「発話者の特定」を扱う．発話者を特定できれば，ロボットに発話したユーザに対して顔を向け，応答を行わせることができる．この挙動は，ユーザに自分が聞き手であることを自覚させ[8]，かつ会話に参加していると感じさせる[1]ことが期待できるため，複数人会話システムにとって重要な要素である．このようなロボットの挙動に加え，複数人会話ではシステムが個々のユーザに質問したり，話題を提供したりするのも重要である．例えば，まだ発話をしていないユーザに対して発話を促せるのが望ましい．このために

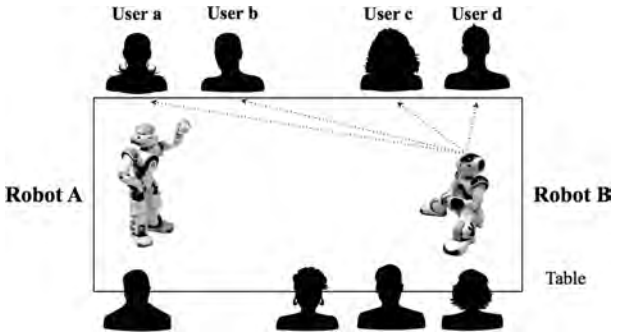


Figure 1: 想定する会話状況．複数ユーザが机を囲み，机の上に配置したロボットと会話を行う．

は，システムは個々のユーザの位置を個別に特定する必要がある．

本研究では音源定位を用いて発話者の特定を行う．音源定位とは，音源がどの方向から到来したかを推定する手法である．ここで，図 1 に示す本稿で想定する会話状況では，以下の 2 つの問題がある．

1. ロボットに搭載されたマイクのみを用いる場合，定位が困難な位置が存在する．
2. 雑音の誤検出が避けられない．つまり，音源定位結果が常に発話者の位置を示しているとは限らない．

本稿で想定する会話状況において，ロボットから遠く離れるほど，2 人のユーザの位置を示す角度差は小さくなる．例えば，図 1 において，Robot B から遠い位置に存在する User a と User b の間の角度差は小さい．このため，2 人のユーザを個別に定位するのは難しい．

本研究では，2 体のロボットを図 1 のように対面的に机に配置し，それらから得られる音源定位結果を統合することで，これら 2 つの問題の解決を図る．この配置により，2 体のロボットのマイクロフォンによる音源定位能力を互いに補うことができる．例えば，Robot B の位置からは，User a と User b の間に十分な角度差はないが，Robot A の位置からは十分な角度差があり，それらを個別に定位することは容易である．この 2 体のロボットから得られる音源定位結果に対して，パワー (音圧) で重みを付け，統合を行う．この音源定位結果の正しさを示す指標としてパワーを用いる．

さらに，本統合手法をヒューマノイドロボット NAO[1]を用いて実装し，デモシステムを構築した．本システムは発話したユーザを特定し，音声認識に基づきそのユーザに対して顔を向けて応答を返す．統合された定位結果のパワーが低い場合は，ロボットに搭載されたカメラを用いて発話者の存在を確認する．

[1]http://www.aldebaran-robotics.com/en/

2 関連研究

複数人会話において発話者の特定を行う最も単純な手法は，ユーザそれぞれにマイクロフォンを持たせることである[6]．マイクロフォンとユーザの位置からシステムは発話者がどこにいて，いつ発話したのかを特定できる．この手法は会話を開始する前に，各ユーザにマイクを準備する必要がある．他にも，多くのマイクロフォンやカメラを設置したスマートルームで会話を行うことで，発話者を特定する手法も考えられる[11]．この手法では，室内に設置された複数のセンサを利用して，ユーザの行動を追跡し，発話者を特定する．この手法は，このような特定の環境でのみ有効である．

これに対して，本研究ではロボットに搭載されたマイクロフォンやカメラのみを用いて発話者の特定を行う．この場合，本稿の想定する会話状況では以下の 2 つが前提となる．

1. ユーザがカメラの視野角内に常に存在するとは限らない．これはロボットに搭載されたカメラの視野角は狭く，解像度も低いためである．
2. 狭い視野角を補うために，常に周りを見回し続けるのは不適当である．これはロボットが発話の当事者であり，このような挙動は会話として不自然になるためである．

Faih らは，唇領域を検出することで，発話者を特定する手法を提案した[4]．Bohus らのシステムは，画像情報に基づいてユーザを追跡し，ジェスチャなどのユーザ行動を認識することで発話者の特定を行う[2] ．これらの手法はユーザが常にカメラの視野角内に存在する場合には有効である．Bennewitz らは，ユーザがロボットのカメラの視野角外に存在する場合であっても，顔検出に基づき参加者の存在について確率的な信頼度を維持する手法を提案した[1]．この手法では，常にユーザをカメラの視野角内に捉えておく必要はないが，一度もカメラの視野角に入らない発話者の位置を特定するのは困難である．

本研究では，主に音源定位結果に基づき発話者の特定を行う．音源定位結果を用いれば，ロボットのカメラの狭い視野角内にユーザが存在することが仮定できなくても，発話者の特定ができる．さらに，一度も視野角内に入らない発話者の特定も期待できる．

3 音源定位結果の統合

本研究では 2 体のロボットから得られる音源定位結果を統合する．本章ではその統合方法について述べる．

2 体のロボットのマイクロフォンを通して，音源の到来方向を角度で示す音源定位結果とそのパワー (音圧) を得る．音源定位結果は，ロボットの正面方向を 0 度とし，反時計回りを正方向として得られる (例えば，ロボットの左

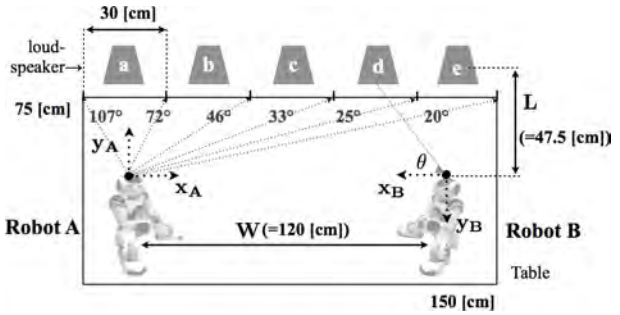


Figure 2: ロボットとスピーカの配置．スピーカをユーザが存在すると考えられる位置に配置する．

手方向は 90 度である)．この設定で，以下の 3 つのステップにより 2 体のロボットから得られる音源定位結果を統合する．

1. 2 体のロボットで異なる座標系を一致させる．
2. 音源定位結果に対して，パワーで重みを付ける．
3. 2 体のロボットから得られた重み付きの音源定位結果を足し合わせる．

まず，図 2 に示す 2 体のロボットの座標系を一致させる．図 2 における各変数を以下のように定義する．

- (x_R, y_R): ロボットごとの座標．ここで $R \in \{A, B\}$ とし，A と B は図 2 の Robot A，Robot B と対応する．座標の原点はロボットの頭部である．x_R 軸の正方向はロボットの正面方向とし，y_R 軸の正方向は正面から見て左方向である．本研究ではロボットは同じの高さの平面上に存在すると仮定し，垂直方向については考慮しない．
- W: 2 体のロボット間の水平距離．
- L: ロボットとユーザが存在すると考えられる位置の最短の水平距離．

ここでは，Robot B の座標系を Robot A の座標系に一致させる．Robot B が音源定位結果 θ を得たとき，Robot B の座標系における音源の座標 (x_B, y_B) は式 1 で表せる．

$$
\begin{aligned}
(x_B, y_B) &= (\frac{l}{\tan\theta}, l) \\
l &= \begin{cases} L & , \quad if \quad \theta > 0 \\ -L & , \quad if \quad \theta < 0 \end{cases}
\end{aligned} \tag{1}
$$

Robot A の座標系における (x_B, y_B) は式 2 で表せる．ここで，α は 2 体のロボットの座標系間の回転角度差であり，(u, v) は，Robot A の座標系における Robot B の原点座標を示す．

$$
\begin{pmatrix} x_A \\ y_A \end{pmatrix} = \begin{pmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{pmatrix} \begin{pmatrix} x_B \\ y_B \end{pmatrix} + \begin{pmatrix} u \\ v \end{pmatrix} \tag{2}
$$

これにより，ロボット間で異なる座標系の一致をとる．実際に，ロボットを図 2 のように配置したとき，各変数の値はそれぞれ，$L = 48.5$ [cm]，$\alpha = 180$ [度]，$(u, v) = (W, 0)$，そして $W = 120$ [cm] であった．現在，L や α，W は人手で計測している．

次に，音源定位結果に対して，パワーで重みを付ける．雑音による音源定位結果のパワーは小さいと仮定し，パワーを音源定位結果の正しさを示す指標として利用する．音源定位結果 θ_r，パワー p_r が得られた場合，音源定位結果の曖昧さは正規分布に従うと仮定し，確率密度関数 $f_r(\theta)$ を定義する (式 3)．ここで，r は ID を示す，例えば，Robot A の音源定位結果は θ_A と表現する．式 3 において，σ_r^2 は分散であり，音源定位結果がどれだけ不確かであるかを示す．

$$
f_r(\theta) = \frac{1}{\sqrt{2\pi\sigma_r^2}} \exp(-\frac{(\theta - \theta_r)^2}{2\sigma_r^2}) \tag{3}
$$

確率密度関数の最大値 $f_r(\theta_r)$ は音源定位結果 θ_r のパワー p_r に比例すると仮定する (式 4)．この仮定は，パワー p_r が大きいほど，音源定位結果が θ_r である確率が大きくなることを示す．ここで，式 4 の C は定数であり実験的に決定する．

$$
f_r(\theta_r) = \frac{1}{\sqrt{2\pi\sigma_r^2}} = \frac{1}{C} p_r \tag{4}
$$

式 4 より σ_r を定める (式 5)．式 5 より，σ_r はパワー p_r に反比例する．つまり，パワーが大きいほど音源定位結果は散らばりが小さいとしている．

$$
\sigma_r = \frac{C}{\sqrt{2\pi}} \frac{1}{p_r} \tag{5}
$$

確率密度関数 $f_r(\theta)$ の例を図 3 に示す．グラフの横軸は，音源定位結果を示し，縦軸はその確率を示す．

最後に，上記のステップの後に得られた音源定位結果を足し合わせる．Robot A，Robot B から音源定位結果 θ_A，θ_B とそのパワー p_A，p_B が得られたとき，それぞれ確率密度関数 $f_A(\theta)$，$f_B(\theta)$ を定義する．$f_A(\theta)$ と $f_B(\theta)$ を式 6，7 に適用し，統合による音源定位結果 θ_{mix} とそのパワー p_{mix} を得る．

$$
\theta_{mix} = \arg\max_{\theta}(f_A(\theta) + f_B(\theta)) \tag{6}
$$

$$
p_{mix} = C(f_A(\theta_{mix}) + f_B(\theta_{mix})) \tag{7}
$$

さらに，統合によって得られるパワーに閾値を設定する．そして，閾値より小さいパワーを持つ音源定位結果を削除する．これにより，ノイズ等に基づく誤った音源定位結果の悪影響を減らすことを期待できる．この閾値は実験により最も精度の高い値を設定する．

4 評価実験

2 体のロボットによる音源定位結果の統合により，1 体のロボットのみを用いる場合と比較して，発話者の特定性能が向上することを確認する．本実験では，スピーカから音声を再生し，そのスピーカの位置を特定した．

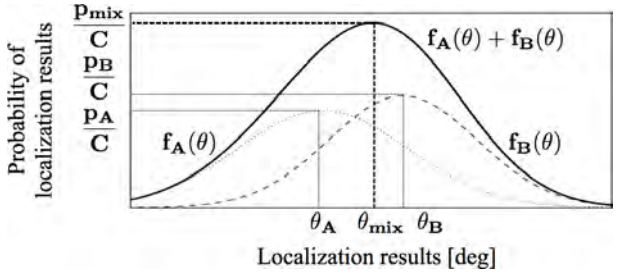


Figure 3: 確率密度関数の足し合わせの例

4.1 実験設定

図 2 に示すように，机 (150cm × 75cm) を準備し，ユーザが存在すると考えられる位置にスピーカを配置した．スピーカはそれぞれ 30cm 間隔で配置し，その中心から ±15cm をそのスピーカの領域とした．図 2 には Robot A から見たスピーカの領域を示す．

本研究では，スピーカから音声を再生しているときに，音源定位結果がそのスピーカの領域内にあれば，その定位結果を正解とみなす．例えば，図 2 において，b のスピーカから音声を再生し，Robot A で音源定位を行う場合を考える．このとき，Robot A からみて b のスピーカは 46 度から 72 度の間に存在するため，定位結果がその間にあれば，正解とする．

音源定位には，ロボット聴覚システム HARK [9]を用いた．HARK は MUltiple SIgnal Classification (MUSIC) 法[10]に基づき，1 フレーム (0.01 秒) ごとに音源定位結果とそのパワーを出力する．MUSIC 法は，音源と入力に用いるマイクロフォン間のインパルス応答 (伝達関数) に基づき，音源を定位する．マイクロフォンには，ヒューマノイドロボット NAO の頭部の前後左右に搭載された 4 つのマイクロフォンを用いた．伝達関数を計算するためのインパルス応答は，マイクロフォンの中心より 1m から，10 度間隔で 36 点計測した．したがって，音源定位結果の角度分解能は 10 度である．

4.2 データと評価指標

スピーカから音声ファイルを再生し，それをロボットのマイクロフォンで録音したデータに対して音源定位を行った．各音声ファイルには，1 名のユーザによる発話が録音されていた．発話は全部で 5 種類で，平均の発話長は 1.0 秒であった．それらを男女 4 名により録音し，図 2 に示す a から e の全 5 カ所から再生した．評価は発話ごと，つまり，全 100 発話について行った．

評価指標には *Precision* (P)，*Recall* (R)，そして F 値 (F) を用いた．本研究では，それらを以下のように定義する．

$$P = \frac{\text{発話中にスピーカの領域内を定位したフレーム数}}{\text{全検出フレーム数}}$$

$$R = \frac{\text{発話中にスピーカの領域内を定位したフレーム数}}{\text{全発話フレーム数}}$$

$$F = (\frac{1}{P} + \frac{1}{R})^{-1}$$

4.3 実験結果

5 カ所 (a から e) に配置したスピーカの特定結果を表 1 に示す．表は左から Robot A のみ，Robot B のみ，そして統合による特定結果である．パワーの閾値は，Robot A のみ，Robot B のみ，統合 ($C = 800$) の各条件でそれぞれ，24, 25.5, 25 のとき，ALL の F 値が最大となった．

1 体のロボットのみを用いた場合では，ロボットから遠い位置のスピーカの特定が難しい．例えば，Robot A から遠い位置にある，スピーカ c，d，e の特定性能は低い．これは，正解とするスピーカの領域が，ロボットとスピーカが離れるほど狭くなり，その領域の定位が困難であるためである．さらに，ロボット間で特定性能に差がある．これは，ロボットのマイクロフォンの性能が異なるためである．

統合により，1 体のロボットのみでは困難であった位置の特定が可能になっている．特に，c のスピーカはどちらのロボットも 1 体のみではほとんど特定ができていないが，統合により他の位置のスピーカと同等の F 値が得られている．c 以外のスピーカの F 値は若干の精度の低下がみられる．これは，どちらのロボットも常に正しい音源定位結果を出力するわけではなく，誤った音源定位結果が統合に悪影響を与えることがあるためである．

4.4 パワーごとの音源定位結果

本研究では，統合によるパワー (p_{mix}) を音源定位結果の正しさを示す指標として用いる．パワーごとのスピーカの特定結果を調べることにより，パワーが音源定位結果の正しさを示す指標として利用できることを確認する．図 4 に，パワーごとの平均誤り率とその平均角度誤差を示す．グラフの横軸は，パワーで 8dB ごとに集計している．左の縦軸は誤り率であり，右の縦軸は平均角度誤差を示す．誤り率は，音源定位結果の出力数のうち，誤りであったものの割合である．ここで，正解であるスピーカの領域以外で検出された音源定位結果を，誤りとした．平均角度誤差は，誤りであった音源定位結果と，スピーカの領域の中心との誤差の平均として計算した．

図 4 の平均角度誤差をみると，パワーが大きいほど平均角度誤差は小さい．つまり，パワーが大きいほど誤りであってもスピーカの領域に近い位置を定位できている．パワーが小さいときは，誤り率は高く，平均角度誤差も大きい．これより，統合後のパワーを用いることで，音源定位結果の正誤の区別が期待できる．例えば，パワーが小さい音源定位結果が得られたときは，音源定位結果は誤り

Table 1: スピーカの特定結果

	Robot A			Robot B			統合		
スピーカ	*Presicion*	*Recall*	*F* 値	*Presicion*	*Recall*	*F* 値	*Presicion*	*Recall*	*F* 値
a	0.56	0.89	0.69	0.00	0.00	-	0.57	0.85	0.68
b	0.49	0.65	0.56	0.00	0.00	-	0.40	0.50	0.45
c	0.00	0.00	-	0.13	0.13	**0.13**	0.38	0.49	**0.43**
d	0.06	0.03	**0.04**	0.63	0.83	0.72	0.48	0.67	0.56
e	0.09	0.03	**0.05**	0.50	0.69	0.58	0.39	0.61	0.48
ALL	0.33	0.32	0.33	0.39	0.33	0.36	0.45	0.62	0.52

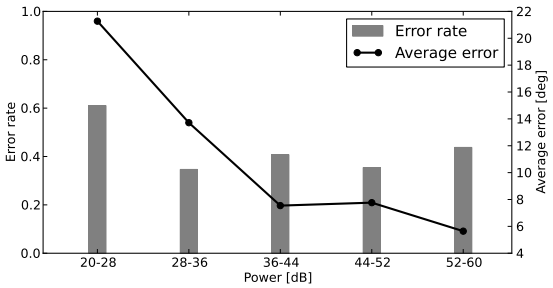


Figure 4: パワーごとの誤り率と平均角度誤差

である可能性が高い．そのため，システムは顔検出などを行うことで，発話者の存在を確認するのが望ましい．

5 デモシステム

本統合手法をヒューマノイドロボットを用いて実装し，デモシステムを構築した．図 5 に，構築したシステムと複数ユーザによるインタラクションの様子を示す．システムのタスクは我々の研究室の紹介で，ユーザはロボットに研究室に関する質問ができる (例えば,「研究室の生活について教えて」)．2 体のロボットにはそれぞれ役割を設定した．役割は主にユーザの質問に答える説明役と，ユーザと共に質問を行う質問役である．構築したシステムには以下の 4 つの特徴がある．

- 複数ユーザの中から発話者を特定し，そのユーザに顔を向けて応答を行う (図 5 の上の写真)．
- 統合によるパワー (p_{mix}) を，音源定位結果の正しさを示す指標として用いる．パワーの小さい音源定位結果が得られたとき，ロボットはその方向に向けて自身に搭載されたカメラを用いて顔検出を行い，発話者の存在を確認する．パワーが大きいときは，確認は行わず，そのまま応答を返す．
- 2 体のロボットで同時に音声認識を行い，音源定位結果を用いて発話者に近いロボットが得た音声認識結果を採用する．
- 一定時間ユーザの沈黙を検出したとき，質問役のロボットが説明役のロボットに質問を行い，会話の間を繋ぐ．

Figure 5: 複数ユーザとシステムのインタラクションの様子．上の写真では，システムが発話者を特定し，そのユーザに顔を向け応答を返す様子を示す．下の写真は，パワーの低い音源定位結果が得られた際に顔検出を行い，発話者の存在を確認する様子を示す．

構築したシステムと複数ユーザのインタラクションのデモ動画はオンラインで視聴できる[2]．

6 まとめと今後の課題

本論文では，2 体のロボットから得られる音源定位結果を統合し，発話者の特定行う手法について述べた．評価実験により，1 体のロボットのみを利用する場合と比較して，統合により発話者の特定性能が相対的に向上することを示した．しかし，絶対的な発話者の特定精度は F 値にして 0.52 程度 (表 1) であり，複数人会話を行うには性能向上が必要である．本研究の問題設定において，発話者の特定精度が低い理由は以下の 2 点である．まず第一の理由は，正解条件，つまりスピーカの正解領域を厳しく設定しているためである．例えば，図 2 において，Robot A の位置からみた d, e のスピーカの間の角度差は，他の位置のスピーカと比べて狭い．このような正解条件を用いるのは，複数人会話においてシステムが個々のユーザを個別に定位できることが重要であるためである．第二の理由は，マイクロフォンとスピーカが離れているためである．

[2]http://sslab.nuee.nagoya-u.ac.jp/en/?page_id=112

このような設定では，接話型のマイクロフォンを利用する場合と比較して，部屋の残響や環境雑音の影響が避けられない．

発話者の特定精度を向上させるための今後の課題は，発話者の存在を示す別の情報源も同時に用いることである．別の情報源には，例えば，ロボットのカメラから得られる画像情報や，会話開始からある時点までに得られた音源定位結果が考えられる．これらの情報を現在の手法と同時に用いることで発話者の特定精度のさらなる向上が期待できる．

個別のユーザの定位に基づいたインタラクションの実現も今後の課題である．例えば，まだ発話していないユーザに対して，発話を促すといった挙動を生成する．

謝辞

Nao と HARK を接続するプログラムは，京都大学の水本武志氏と協力して作成した．本研究の一部は，JST 戦略的創造研究推進事業さきがけの支援を受けた．

参考文献

[1] Maren Bennewitz, Felix Faber, Dominik Joho, Michael Schreiber, and Sven Behnke. Integrating vision and speech for conversations with multiple persons. In *Proceedings of IEEE/RSJ the International Conference on Intelligent Robots and Systems (IROS)*, pages 2523–2528, 2005.

[2] Dan Bohus and Eric Horvitz. Models for multiparty engagement in open-world dialog. In *Proceedings of the SIGDIAL 2009 Conference*, pages 225–234, 2009.

[3] Alexander Gruenstein and Stephanie Seneff. Releasing a multimodal dialogue system into the wild: User support mechanisms. In *Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue*, pages 111–119, 2007.

[4] Fasih Haider and Samer Al Moubayed. Towards speaker detection using lips movements for human-machine multiparty dialogue. In *FONETIK 2012*, 2012.

[5] Natasa Jovanovic, Rieks op den Akker, and Anton Nijholt. Addressee identification in face-to-face meetings. In *Proceedings of the 11th Conference of the EACL*, 2006.

[6] Yoichi Matsuyama, Hikaru Taniyama, Shinya Fujie, and Tetsunori Kobayashi. Framework of communication activation robot participating in multiparty conversation. In *Proceedings of AAAI Fall Symposium, Dialog with Robots*, pages 68–73, 2010.

[7] Samer Al Moubayed, Jonas Beskow, Mats Blomberg, Bjrn Granstrm, Joakim Gustafson, Nicole Mirnig, and Gabriel Skantze. Talking with furhat - multi-party interaction with a back-projected robot head. In *FONETIK 2012*, 2012.

[8] Bilge Mutlu, Toshiyuki Shiwaand Takayuki Kanda, Hiroshi Ishiguro, and Norihiro Hagita. Footing in human-robot conversations: how robots might shape participant roles using gaze cues. In *Proceedings of the 4th ACM/IEEE International Conference on Human Robot Interaction*, pages 61–68, 2009.

[9] Kazuhiro Nakadai, Toru Takahashi, Hiroshi G. Okuno, Hirofumi Nakajima, Yuji Hasegawa, and Hiroshi Tsujino. Design and implementation of robot audition system 'HARK' - open source software for listening to three simultaneous speakers. *Advanced Robotics*, 5:739–761, 2010.

[10] Ralph O. Schmidt. Multiple emitter location and signal parameter estimation. *IEEE Transactions on Antennas and Propagation*, 34:276 – 280, 1986.

[11] Rainer Stiefelhage, Jie Yang, and Alex Waibel. Modeling focus of attention for meeting indexing based on multiple cues. *IEEE Transactions on Neural Networks*, 13:928–938, 2002.

[12] David Traum and Jeff Rickel. Embodied agents for multi-party dialogue in immersive virtual worlds. In *Proceedings of the First International Joint Conference on Autonomous Agents and Multiagent Systems*, pages 766 – 773, 2002.

多チャンネルマイクロホンアレイを用いた音声区間検出および音源定位の精度の向上の検討

On Improving the Accuracy of Voice Activity Detection and Sound Source Localization by Microphone Array

黄 楊暘 †, 大塚 琢馬 †, 中臺 一博 ‡, 奥乃 博 †

Yangyang Huang†, Takuma Otsuka†, Kazuhiro Nakadai‡, Hiroshi G. Okuno†

† 京都大学大学院情報学研究科, ‡(株) ホンダ・リサーチ・インスティチュート・ジャパン

†**Graduate school of Informatics, Kyoto University,** ‡**HONDA Research Institute Japan, Co., Ltd.**

†{yangyang, ohtsuka, okuno}@kyoto-u.ac.jp, ‡nakadai@jp.honda-ri.com

Abstract

In Real-World Auditory Scene Analysis concerning human-robot interaction, three types of information are essential and need to be extracted from the observation data – **WHO** speaks **WHEN** and **WHERE**. This paper presents such a system that is used to accomplish the resolution of these objects. To evaluate such a system, we formulate the use of evaluation indicators which are precision rate, recall rate, localization error and speaker ID error rate. Multiple Signal Classification (MUSIC) is a powerful method used for analysing WHEN and WHERE, more specifically, voice activity detection (VAD) and direction of arrival estimation (DOA). In this paper, we describe our system and compare its performance in VAD and DOA with MUSIC method.

1 はじめに

人とロボットが共生するためには, ロボットの聴覚機能の開発は不可欠である. 特に重要な聴覚機能としては，ロボットが人間と会話する場面を考えると，様々な人が発話する観測音の中から，いつ，どこで，誰が，何を話したかを認識する機能が挙げられる (図 2). これらの機能は, 音声区間検出, 音源定位, 音源同定や, 音源分離問題として, 様々な手法が開発されている. [Nakadai *et al.*, 2010; Tranter and Reynolds, 2006; Nakamura *et al.*, 2011].

本稿では, 上記のいつ，どこで，誰が話しているかを推定する話者ダイアライゼーション問題を取り扱う. 本問題は，マイクロホンアレイで収録された複数話者同士の自由発話に対して，各話者の音声区間検出や音声到来方向, および話者の推定を行う複合的な問題である. この問題には次の 2 点が重要である.

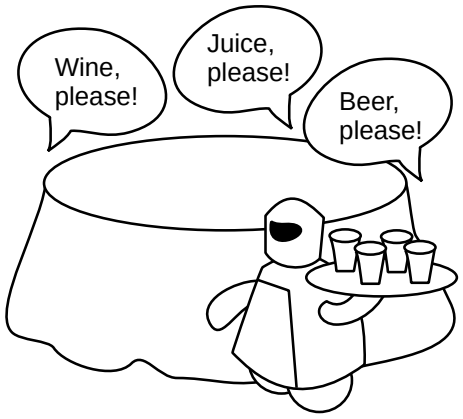


Figure 2: 例えば, 図示のカクテルパーティで接客するロボットの場合を考えて, いつ, どの方向から誰が注文を理解するのが重要である.

- 各部分問題に対してどのような要素技術を選択すれば全体の性能向上に寄与するかを明らかにすること,
- 複数の要素技術を直列につないで処理を行う場合，前段の処理の結果が後段の処理に影響するため，前段の処理は様々な観測音に対して頑健な手法が望ましい.

例えば，ロボット聴覚システム HARK[Nakadai *et al.*, 2010] では，全体の処理を multiple signal classification (MUSIC) 法による音源定位を行い，その音源方向推定結果に基づいて音源分離など，各音源に関する処理を行う. 本話者ダイアライゼーション問題についてもまず各話者の方向を推定し，その結果を用いて話者同定などを行う枠組

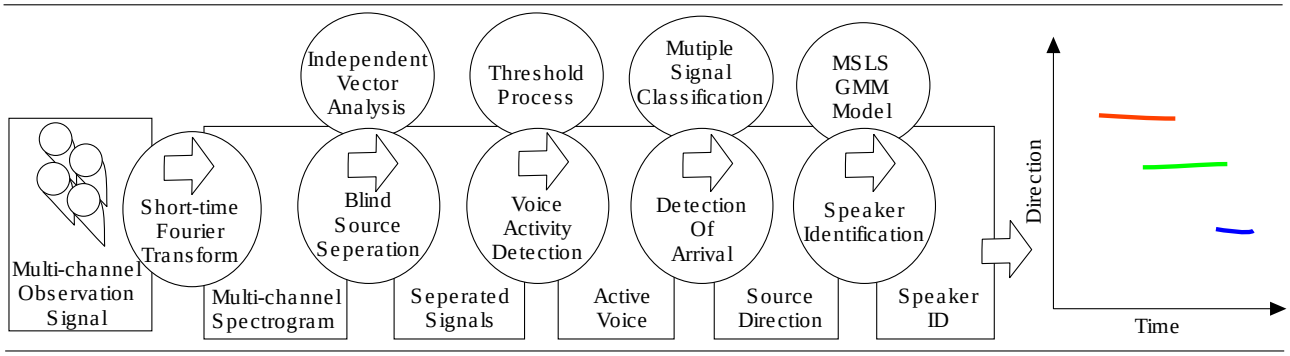


Figure 1: 処理の流れおよび出力結果の図示.
ひとつひとつの発話を線分で時間ー方向座標系で示すように, 色は音源 ID の違いを指している.

みが考えられる．しかし，表 1 でも示すように，MUSIC 法には入力音に依存した音源数や閾値などのパラメータにより，出力が大きく変わるという問題がある．従って，システム全体を最適化するには，注意深く MUSIC 法で用いるパラメータを選択する必要があるという問題があった．

上記に示して 2 点に関する本稿の貢献は次の通りである．

- 効率的な性能評価のため，収録した発話に対して正解データを付与し，話者ダイアライゼーション問題に対して性能評価指標を定義し，
- 前処理に頑健性の高い音源分離手法である independent vector analysis (IVA) を用いることで全体の性能を改善した．

正解データは，各話者に接話型マイクを使用して音声区間を決定した他，話者の位置を計測する MAC 3 D システム [角康之 *et al.*, 2008] を利用して音源方向の正解データを作成した．また，話者ダイアライゼーションシステムの評価指標としては，各話者の音声区間に対しては適合率，再現率を用いて F 値を定義し，音源定位誤差も導入した．

本稿は次のように構成されている．2 節では問題設定と提案手法の処理の流れ，および各要素技術を示す．3 節では評価用データの収録環境と正解データ作成法を説明し，および評価指標を定義し，4 節では評価実験結果を報告する．5 節で本稿をまとめる．

2　問題設定とシステム構成

本節では，話者ダイアライゼーションシステムの問題設定を示した後，提案手法の枠組みを示し，利用するそれぞれの要素技術を概説する．本稿で扱う問題設定は次の通りである．

以下に本稿で扱う問題設定を示す:

- 入力：多チャンネルの音声信号
- 出力：音声区間, 音源の到来方向および話者 ID
- 条件 1：各話者の事前学習データが入手可能
- 条件 2：マイクロホンアレイの伝達関数が既知

条件 1 に関して, 音声区間と話者についての正解ラベルが与えられた音声データを用いて, 各話者クラス構築のための事前学習を行う. 条件 2 に関して, MUSIC による音源定位では, マイクロホンアレイの伝達関数が必要である. 伝達関数は各方向からの音の伝達特性を表す.

提案手法は図 1 に従って処理する. 入力である多チャンネル音声信号を短時間フーリエ変換の後に，音源分離手法 IVA を適用する．得られた各話者の分離音声に対してパワーの閾値処理による音声区間検出を行う．また，各分離音声に対して MUSIC 法を用いて各話者の方向推定と，mel-scale log spectrum (MSLS) 音声特徴量を用いた話者同定を行う．話者同定には，混合ガウスモデル (GMM) による判別を行う．

2.1　IVA を用いたブラインド音源分離

IVA は多チャンネルの時間周波数領域における音源分離法であり，独立成分分析 (independent component analysis; ICA) の拡張手法である．本節ではまず ICA について概観し，IVA への拡張を簡潔に説明する．

ICA は時間周波数領域における多チャンネル観測信号 $\mathbf{Z}_{t,f} = [z^1_{t,f}, ..., z^M_{t,f}]^T$ が次式の観測モデルで表す.

$$\mathbf{Z}_{t,f} = \mathbf{A}_f \mathbf{Y}_{t,f}$$

ただし，$\mathbf{Y}_{t,f} = [Y^1_{t,f}, ..., Y^M_{t,f}]^T$ は時間フレーム t，周波数ビン f における各音源の信号で，$\mathbf{A}_f$ は混合行列である．このとき，ICA は観測信号 $\{\mathbf{Z}_{t,f}\}_{t=1}^T$ から，

$$\hat{\mathbf{Y}}_{t,f} = \mathbf{W}_f \mathbf{Z}_{t,f}$$

に従って計算される $\hat{\mathbf{Y}}_{t,f}$ の各成分が統計的に独立になるよう分離行列 $\mathbf{W}_f$ を求める．これは，元音源である $\mathbf{Y}_{t,f}$

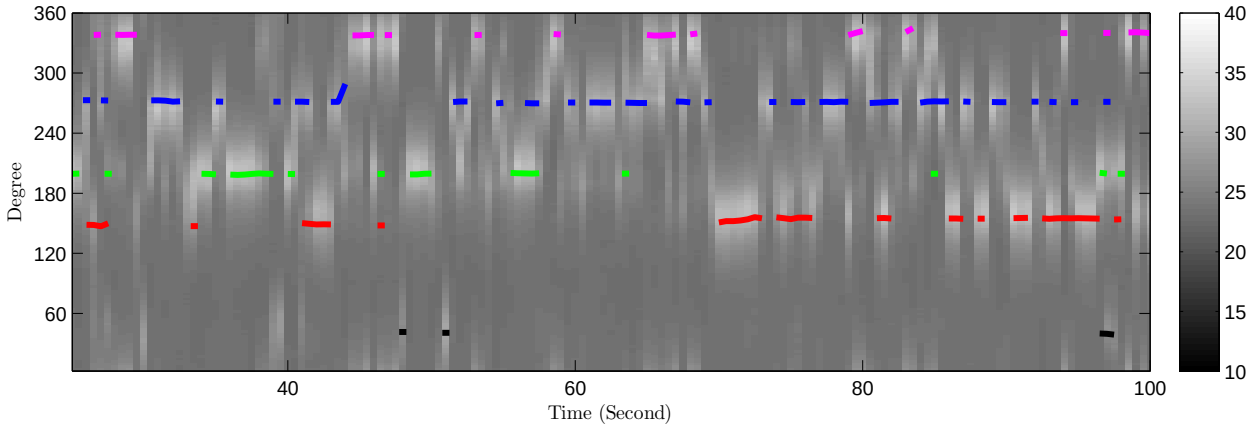


Figure 3: 作成した正解データと MUSIC スペクトルを重ね合わせて描いた図. MUSIC スペクトルのピークが音声区間の対応関係を確認できる.

の各成分が統計的に独立であるという仮定に従って音源分離に適用されている.

ICA における問題点は，式に従って各周波数ビン f ごとに計算された $\hat{\mathbf{Y}}_{t,f}$ の各成分は，必ずしも元の $\mathbf{Y}_{t,f}$ と同じ順になっていないというパーミュテーション問題である．そのため，元の音声信号を復元する際には，各周波数ビンごとに同一音源に属する成分を正しく選ぶ必要があった．それに対して IVA は，$\{\mathbf{Y}_{t,f}\}_{f=1}^{F}$ の各成分を F 次元のベクトルとみなして，全周波数ビンに関して $\{\mathbf{W}_f\}_{f=1}^{F}$ を最適化することで，パーミュテーション問題を回避している [Lee *et al.*, 2007; Ono, 2011].

2.2 音量閾値処理による音声区間検出

入力の多チャンネルスペクトログラムを波形信号 $y_{t,d}$ に変換して, 時間領域の信号 $Y_{t,d}$ に対して, 一定長 Δt の区間中において, 絶対値が閾値 T_v 以上の波形のサンプル数が T_s を超える場合に, 音声区間と見なす. 各分離音声の音声区間に含まれた部分をこれ以後の処理を続きます. 多チャンネルのスペクトログラムを算出した音声区間で切り出して出力する.

2.3 MUSIC 法による音源定位

MUSIC 法は音声信号の部分区間と雑音信号の部分区間が直交することを利用して, 高い精度の音源定位ができている. MUSIC スペクトルが得られたら, 事前に閾値を設定する. 閾値より以上の値が出た場合に, 音源定位と音声区間検出の同時推定ができる. 本手法では, MUSIC 法を音源定位に使う. MUSIC 法は，観測信号に対して MUSIC スペクトル $P_{b,\theta}$ と呼ばれる，各ブロック b，方向 θ に対応するエネルギーを計算し，一定以上の $P_{b,\theta}$ を持つ方向に音源が存在するという閾値処理を行うことで音源定位を行う．その算出の手順が次のようになる. 入力スペクトログラム $z_{t,f}$ の自己相関形式

$$R_{b,f} = \sum_{t=(b-1)*\Delta T}^{b\Delta T} \mathbf{z}_{t,f}\mathbf{z}_{t,f}^{H}$$

を取って, 安定の定位結果を得るために, フレーム ΔT 分の自己相関行列を足し合わせる, 一つのブロックと見なす. 各時間ブロック b と周波数ビン f の $R_{b,f}$ に対して固有値分解を行なって, チャンネル数と同じ M 個の固有値と固有ベクトルが得られる $\{\lambda_{b,f,m}, \mathbf{e}_{b,f,m}\}$. 固有値の大きい順から, 固有値と固有ベクトルを並べる. その時間ビンと周波数ビンの MUSIC エネルギーは算出された固有ベクトル $\mathbf{v}_{b,f,m}$ と事前に測定した伝達関数 $a_{f,\theta}$ を利用する. 算出式は次のようになる.

$$P_{b,f,\theta} = \frac{\|a_{f,\theta}a_{f,\theta}^{H}\|}{\sum_{m=N+1}^{M}|a_{f,\theta}\mathbf{e}_{b,f,m}^{H}\|}$$

計算式では, $N+1$ 番目の固有ベクトルから, $N-m$ 個の固有ベクトルを利用する. 周波数ビンの統合は周波数ビン $1,\cdots,F$ に対して, 最大の固有値 $\lambda_{t,f,1}$ の平方根による重み付け和によって行う.

$$P_{b,\theta} = \sum_{f=1}^{F}\sqrt{\lambda_{t,f,1}}P_{b,f,\theta}$$

MUSIC 法の詳細は[Schmidt, 1986]を参照する.

2.4 MSLS 特徴量の計算

本稿では, 話者同定の音声特徴量として MSLS 特徴量を利用する. MSLS 特徴量は，人間の聴覚機能を反映した対数

周波数軸上のパワーに基づく特徴量である．MSLS 特徴量は音源分離時に生じた漏れノイズに対する頑健性が期待でき，たとえば分離音声の音声認識などに利用されている[Yamamoto *et al.*, 2007].

MSLS 特徴量の計算の手順は次のようになる. メル周波数窓を使って,257 次元の線形周波数軸の分離音声の絶対値 $|V_{f,i}|(f = 0...256)$ を 13 次元の特徴ベクトル **r** に変換する.

1. メル周波数と周波数の関係の計算式は次のようになる.

$$m = 1127\log(1.0 + \frac{f}{700.0})$$

2. 周波数領域で等間隔各成分の窓をかけて, 得られた各成分に対して, 対数値を取って, **h** が得られる.
3. 13 次元のベクトル $h(i)$ を以下のように $r(i)$ 正規化する. $i = 1,\cdots,13$.

$$r(i) = \frac{1}{13}\sum_{p=0}^{12}\left\{\sum_{r=1}^{1} 3\left\{h(r)\cos(\frac{\pi p(r-0.5)}{13})\right\}\cos(\frac{\pi p(i-0.5)}{13})\right\}$$

2.5 GMM のパラメータ学習

GMM による識別は，IVA で分離した事前にラベルを付けた各話者からの 20s 程度の分離音声を学習データとして. ラベル付けた音声特徴量データを EM アルゴリズムで混合ガウス分布の各混合の重み, 平均と分散 g^l, μ^l, Σ^l を学習する. $l(= 1,\cdots,3)$ は各混合のインデックスを表す, 本稿では混合数を 3 にした. c をクラスの番号として, クラスの決定は次の式で行う. $\mathcal{N}$ はガウス分布の確率密度関数で, **r** は音声特徴量ベクトルを指す.

$$Class = \underset{c}{\operatorname{argmax}} \sum_{l=1}^{3} g_c^l N(\mathbf{v}|\mu_c^l, \Sigma_c^l)$$

3 実験データ収録環境

本節では, 実録音対話データからの正解データ作成手順と, 音声区間検出, 音源定位, および話者同定に関する評価指標の設計を説明する.

マイクロホンアレイの入力音声信号を, 長さが 0.5 秒のブロックに分割して, 方向ごとに, 音声区間であるかどうかおよび音源の ID を目標として, 結果の形式は, ブロック数 × 方向数の二次元アレイのデータ構造として扱う. $x_{b,\theta}$ は b 番のブロックにおいて, θ 方向の推定結果の値を表す. $x_{b,\theta}$ は 0 以上の整数である, 0 の場合は無音区間, 0 より大きい場合はその音源 ID の音声区間であることを示す.

3.1 正解データの作成

実環境の音源は, 今回収録したデータを含めて, 一般に移動する. 複数音源が時々刻々位置を変化させながら音を発したり黙ったりするデータに対して, 音源位置や音声区間の評価用フィファレンスデータが必要であるため, 今回は次の手順で正解データを作成した.

1. 今回の複数話者による発話データは図 4-a のように収録した. 机の上に 16 チャンネルのマイクロホンアレイ (図 4-b) を設置し, 机の周りに, 五人の話者が座った. 各話者が着席した状態でマイクロホンアレイに向いて発話を行った. 話者の首の動きなどによる音源移動はあるが, 席替えなどの音源方向の大きな変動は今回のデータには含まれない.
2. 音声区間のリファレンスデータは, 各話者の襟元につけた接話型マイクロホンによる録音データと収録時に同時に録画されたビデオを元に手動で作成した.
3. 各話者の位置の正解データはリアルタイム光学式モーションキャプチャシステム (MAC3D システム) を利用して取得した. このシステムは, 図 4-c のように各話者の肩と頭部に付けられたマーカーとカメラアレイによって各話者の位置を追跡する. 本システムにより得られた, 各話者を天井から見下ろした場合の, マイクロホンアレイを減点とする x-y 座標をプロットすると, 図 3 のようになる．話者の x-y 座標から，マイクロホンアレイからみたその話者の方向も容易に計算が可能である．マイクロホンアレイからの話者方向の範囲で話者 ID を定め，線を色分けした.
4. 2. で作成した音声区間は,3. で付与した音声 ID と対応付けることで, $x_{b,\theta}$ を作成した.

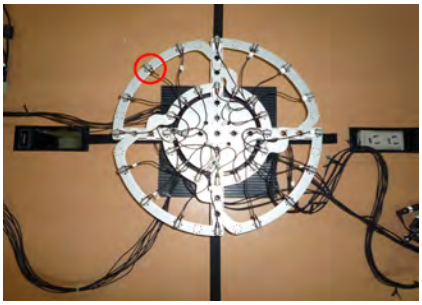
3.2 評価指標

音声区間検出, 音源定位, 音源同定の結果について, 以下の評価指標を設計する.

3.2.1 音源 ID を考慮しない場合

音声区間検には, 挿入エラーと削除エラーを考えて. それらは適合率と再現率で定量的に評価する．挿入エラーは，正解データでは無音区間となっている区間に対して, 音声を検出する誤認識のことである．それに対し削除エラーは，正解データでは音声区間であることを示しているのに，アルゴリズムが発話を検出しないという誤認識である．挿入・削除エラーの計算には音源方向にある程度の誤差を許容する．たとえば，正解データでは 30° 方向に音声が存在するのに，35° 方向に音声を検出した場合を考えた時，30° 方向の音に対する削除エラーに加えて 35° 方向への挿入エラーが生じたとみなすのではなく，定位誤差はあるものの, 挿入・削除は生じなかったとみなす．具体的には，許容誤差が θ_p で正解データではブロック b にて θ 方向に音源があるとき，$[x_{b,\theta-\theta_p}, x_{b,\theta+\theta_p}]$ の範囲内に

(a) 録音風景



(b) マイクロホン配置, 今回は外側の 16 個のマイクロホンが収録したデータを利用した, 赤い丸で囲んだのはその一つである.

(c) MAC3D システムマーカー, 帽子と肩にある白い円状物がマーカーである.

Figure 4: 実験風景

Table 1: MUSIC スペクトルに基づくベースライン手法の適合値率・再現率評価. 行: MUSIC スペクトルで音が存在すると判定する閾値. 列：MUSIC スペクトル計算時に仮定する音源数. パラメータの変化に伴って, 精度が大きく変わることがみられる.

	1		2		3	
	P	R	P	R	P	R
25	0.541	0.679	0.268	0.770	0.155	0.719
27	0.641	0.621	0.323	0.766	0.155	0.719
29	0.766	0.539	0.457	0.742	0.156	0.719
31	0.827	0.317	0.600	0.667	0.179	0.711

存在する $x_{b,\theta}$ の値が 0 より大きい場合は，音声区間検出については正解とみなす．ただ, 一つの音源方向の許容範囲に複数の推定結果が含まれる場合は, 挿入エラーとなる. 音源 ID を考慮しない場合には, マイクロホンアレイ処理によって検出された音声区間，すなわち $x_{b,\theta} > 0$ の数. その内の推定結果が正しい (挿入エラーでない) 数を S_c とする. また，正解データ中の音声区間 $x_{b,\theta} > 0$ の数を S_d とする． 音源方向について，正解データと推定結果の誤差の絶対値の和を Δ_{dir} とする．これらを用いて, 音声区間検出における評価指標は次のように定義される.

適合率:$R_p = \frac{S_c}{S_a}$　　再現率:$R_r = \frac{S_c}{S_d}$

音源定位誤差: $E_{dir} = \frac{\Delta_{dir}}{S_c}$　　F 値:$F = \frac{2R_pR_r}{R_p+R_r}$

3.2.2　音源 ID を考慮する場合

音源 ID を考慮する場合では, 音声区間と音源定位の推定結果が正しいにも関わらず, 音源 ID の付与が間違った場合がある. [高橋徹 *et al.*, 2009] ここで, 推定結果と正解データの同じ音源 ID である部分を取り出して, 各音源に対して, 前節の指標で評価することができる. この評価方法は音源 ID が正しい推定されたことを仮定して, 評価を行う. 音源 ID を考慮した音声区間検出・音源定位精度の評価指標としては，推定された音源 ID の正解データについて前節の評価指標を適用することが考えられる．この手法は容易に評価計算を行えるが，音源 ID の誤推定が評価スコアを著しく低下させる要因となる．したがって, 音源 ID の誤推定を定量的に評価するのが望ましい. ここで, 音源ごとに評価する時, 推定結果が正しいと考えられる数をすべて足しあわせて, その総数を S_e とする. 音源 ID の誤推定率 E_{ID} を次のように定義する. $E_{ID} = \frac{S_c - S_e}{S_c}$

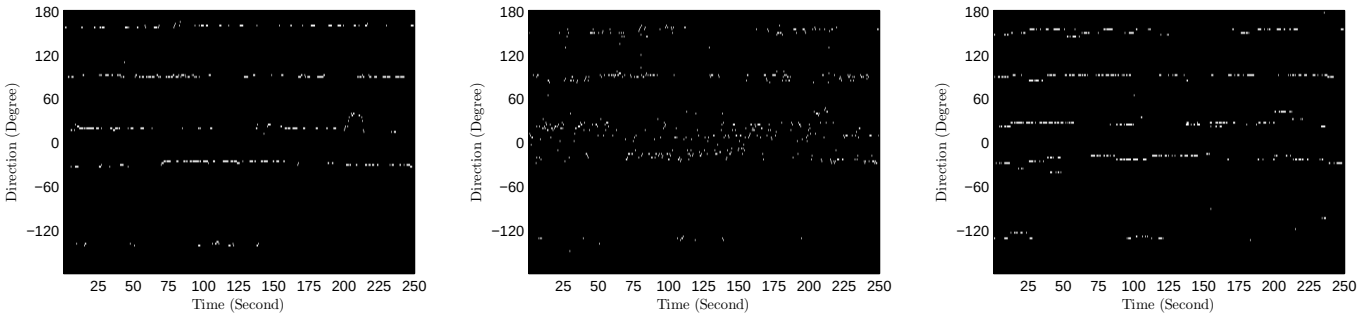
4　実験結果

4.1　ベースライン手法の評価

MUSIC スペクトルに対して, 以下の処理を順に行って, 音声区間検出, 音源定位を行う. MUSIC スペクトルでは, 閾値以下の範囲である部分を無音区間と見なす．一つのブロックにおいて, 連続の方向区間 $\Delta_\theta (= 15°)$ 内に連続で閾値より大きい場合, そのなかの最大値が位置する $x_{b,\theta}$ を音源の方向にして, 区間内の他の $x_{b,\theta}$ を無音区間と見なす. 以上の手順で計算された MUSIC 法による音源定位結果を表 1 にまとめる.

4.2　提案手法の評価

IVA 音源分離処理では, 音源数をその場にいた話者数 5 に設定している. 音声区間検出の閾値処理の部分では, T_v を 0.01 に設定して, T_s を 8000 サンプル中の 100 に設定している. 音声区間検出と音源定位の推定結果について, 図 5 で示したように, 評価実験の結果と MUSIC 法による結果の比較を行った. リファレンスデータに対して, 提案手法がより精度の高い結果が得られることがわかった. 数値的な評価に関しては, 図 6 で示したように, 精度の定量的な向上が確認できた. 図 6 の左辺では, MUSIC 法のデータ点が多い理由は閾値と音源数を変えて結果 MUSIC 法を評価



(a) リファレンスデータの音声区間と音源方向　(b) MUSIC 法による推定結果　(c) 提案手法による推定結果

Figure 5: 提案手法と MUSIC 法に基づいたベースライン手法の比較, 提案手法のほうの推定結果が MUSIC 法だけを利用した手法より精度が高いことがわかる.

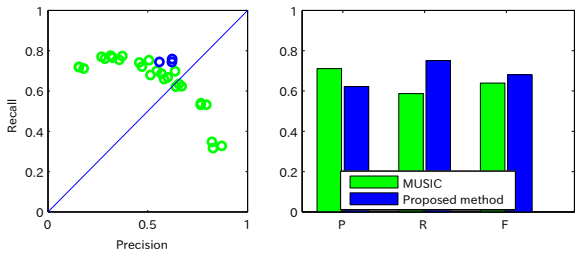


Figure 6: 比較結果の図である. 左辺は適合率ー再現率分布で. 右辺は適合率 (P), 再現率 (R), F 値 (F) 評価の比較である.

した結果をプロットしている. 右辺はそれぞれの手法の適合率, 再現率と F 値評価を示している. 三つの 4 分程度の実録音データに対して, MUSIC 法の評価結果については, 各録音データに対して, F 値の高いほうを選んで各指標の平均を取っている. また, 提案手法の音源 ID の誤推定率 E_{ID} は 0.23 である. 提案手法と MUSIC 法による結果の音源定位誤差が同じく 7.5 度ぐらいとなる.

4.3 考察

実験を通じて, 提案手法はより高い再現率と F 値を示した. しかし, 本手法には次の制約が存在する. (1)IVA 音源分離は音源が動かない前提で分離行列を推定しているので, 本手法の移動音源への対応が必要となる. (2) 音声区間検出の閾値処理だけでは, 環境雑音に対して頑健性が足りないと予想している.

5 まとめ

本稿では, いつ, どこで, 誰が話しているかを推定する話者ダイアライゼーションシステムの構成を述べた. 話者ダイアライゼーション問題は複合的な問題なので, 様々な処理を直列につないで対処するが, 本手法は様々な観測音に対して頑健な IVA を前処理とすることで, 全体のパフォーマンスの改善に寄与している. 評価実験では, MUSIC 法をベースとした手法により音声区間検出と音源定位精度の向上を確認した.

謝辞: 本研究の一部は科研費基盤 (S) の支援を受けた.

参考文献

[Lee *et al.*, 2007] I. Lee, T. Kim, and T.W. Lee. Fast fixed-point independent vector analysis algorithms for convolutive blind source separation. *Signal Processing*, 87(8):1859–1871, 2007.

[Nakadai *et al.*, 2010] K. Nakadai, T. Takahashi, H.G. Okuno, H. Nakajima, Y. Hasegawa, and H. Tsujino. Design and implementation of robot audition system'hark'—open source software for listening to three simultaneous speakers. *Advanced Robotics*, 24(5-6):739–761, 2010.

[Nakamura *et al.*, 2011] K. Nakamura, K. Nakadai, F. Asano, and G. Ince. Intelligent sound source localization and its application to multimodal human tracking. In *In Proceedings of the IEEE/RSJ International Conference on IROS*, pages 143–148. IEEE, 2011.

[Ono, 2011] N. Ono. Stable and fast update rules for independent vector analysis based on auxiliary function technique. In *2011 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 189–192. IEEE, 2011.

[Schmidt, 1986] R. Schmidt. Multiple emitter location and signal parameter estimation. *IEEE Transactions on Antennas and Propagation*, 34(3):276–280, 1986.

[Tranter and Reynolds, 2006] S.E. Tranter and D.A. Reynolds. An overview of automatic speaker diarization systems. *In Proceedings of the IEEE Transactions on Audio, Speech, and Language Processing*, 14(5):1557–1565, 2006.

[Yamamoto *et al.*, 2007] S. Yamamoto, J. Valin, K. Nakadai, M. Nakano, H. Tsujino, K. Komatani, T. Ogata, and HG Okuno. Simultaneous speech recognition based on automatic missing feature mask generation by integrating sound source separation. *Journal of the Robotics Society of Japan*, 25(1):92, 2007.

[角康之 *et al.*, 2008] 角康之, 西田豊明, 坊農真弓, and 來嶋宏幸. Imade: 会話の構造理解とコンテンツ化のための実世界インタラクション研究基盤. **情報処理**, 49(8):945–949, 2008.

[高橋徹 *et al.*, 2009] 高橋徹, 中臺一博, 石井 Carlos 寿憲, Jani Even, and 奥乃博. 実環境したでの音源定位・音源検出の検討. **第 29 回日本ロボット学会学術講演会**, 29(1F3-3), 2009.

社団法人 人工知能学会
Japanese Society for
Artificial Intelligence

人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-B202-6

ノンパラメトリックベイズによるメディア処理

Media processing via Bayesian nonparametrics

中野允裕

Masahiro Nakano

NTT コミュニケーション科学基礎研究所

NTT communication science laboratories

nakano.masahiro@lab.ntt.co.jp

Abstract

本稿では音, 画像, 動画, 自然言語など多くのメディア処理における近年のベイズ（特にノンパラメトリックベイズ）的な手法の発展を分野横断的に紹介する. 混合モデル, 隠れマルコフモデル, 確率文脈自由文法, n-gram, 非負値行列因子分解, 独立成分分析など多くの確率モデルに対するベイズ的な取扱いと, それらの無限モデルを構成する際に登場するディリクレ過程やレヴィ過程の特別な場合について紹介する. さらにモデルの拡張するための基本的な方法として階層化, 相関の導入, 入れ子構造の利用, 木構造化などを示し, それらが各種メディア処理のどのような場面で活用されているかを紹介する.

1 はじめに

音, 画像, 動画, 自然言語などのメディア処理全般において, 近年確率的な生成モデルを用いて課題を解決しようとする研究が少なくない. 特にここ最近 10 年ほどはノンパラメトリックベイズモデルの普及, 推論アルゴリズムの発達, 計算機能力の向上などと相まって, メディア処理における各種課題への有力な選択肢の一つとなってきている.

ベイズ的な手法を用いたメディア処理の基本的な戦略は, 対象の確率的な生成モデルを描き, それを実際の観測データにフィッティングさせ, 得られたモデルを介して課題を解くことにある. 対象に対していかに適切なモデルを設計するかが重要になってくるため、対象とするメディアや扱う課題ごとに全く別の方針を考えなければならないように思われるが, 実際のモデルの設計には共通した部分が非常に多く, また, あるメディアにおける標準的な戦略を別のメディアに輸入し成功した例も多い. 本稿では各種メディアに対するベイズ的な手法を分野横断的に整理することでそれらの発展を概観したい (Fig. 1).

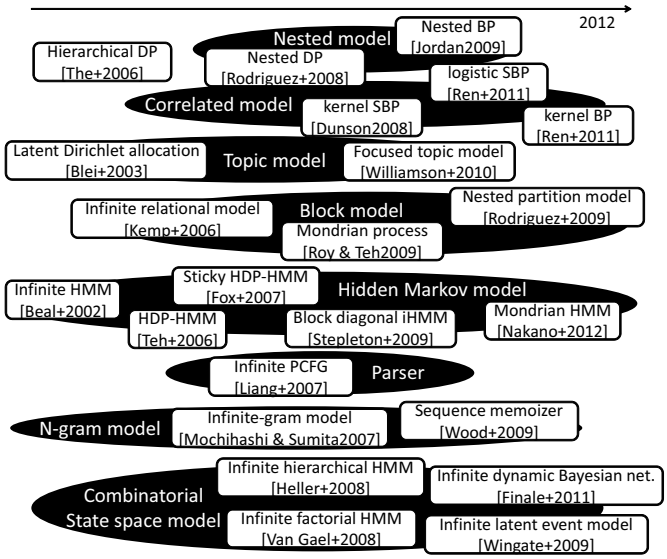


Figure 1: 本稿で扱うベイズモデルの一部. メディア処理におけるベイズモデルとして重要なもの全てを網羅することは出来ないが, その一部を対象とするメディアに関して分野横断的に概観する.

2 確率的生成モデルを用いたメディア処理

2.1 モデルの設計と推論の基本方針

ベイズ的な戦略の基本は, 観測データを説明する確率的な生成モデルを設計することにある. 観測データ（例えば 1 次元の実数）x が平均 μ, 分散 σ^2 の正規分布 $\mathrm{Normal}(\mu, \sigma^2)$ から生成されたとみるのが妥当であれば, $x \sim \mathrm{Normal}(\mu, \sigma^2)$ のように観測データの生成過程を記述する. もし 100 個の観測データ $x_1, x_2, \ldots, x_{100}$ がある一つの正規分布から独立に生成されたと見なすことが出来るのであれば $x_n \sim \mathrm{Normal}(\mu, \sigma^2)$ $(n = 1, 2, \ldots, 100)$ のように表せばよい. ここで, 正規分布の平均は 0 付近にあるはずだと分かっていれば, μ を確率変数と見なし, $\mu \sim \mathrm{Normal}(0, \tau^2)$ のように μ の生成過程を設計すればよい.

モデルが設計出来た後, 我々の興味は観測データが与えられたときにそれらのパラメータがどうなっているか, またはパラメータの分布がどうなっているかにある. 先の例であれば, $x_1, x_2, \ldots, x_{100}$ が観測されたときのパラメータ μ, σ の分布 $p(\mu, \sigma \mid x_1, x_2, \ldots, x_{100})$ が我々の興味である. このようなパラメータの分布を知るための推論には大きく二つの方針が用いられることが多く, 一つはマルコフ連鎖モンテカルロ法と呼ばれるもので, もう一つは変分ベイズ法と呼ばれるものである. それぞれがどのようなものであるかについては膨大な教科書があるのでそれらを参照して頂きたい.

2.2 モデルの複雑度の設定

対象の確率的な生成モデルを設計する上で, データを説明するパラメータの個数をどの程度にすべきかは重要な問題である. パラメータを増やせば観測データをよく説明出来るようになるが, 一方で過学習を起こし未知のデータを説明する能力に乏しくなる恐れがある. パラメータが少なすぎれば観測データを説明するための能力が足りなくなる恐れがある. このようなに, データを説明する際のパラメータ数（モデルの複雑度）をどのように設定すべきかは極めて重要な問題で, 汎用の設計指針が求められてきた.

モデルの複雑度の設定に対する近年の標準的な取扱い方がノンパラメトリックベイズと呼ばれる枠組みで, これは無限のパラメータを用いた生成モデルを描こうとする考え方である. そもそもデータを説明するだけのパラメータの数などというものはそれ自身不確かなもので, 本来観測データに説明させるべきものである. そこで, 十分なパラメータを用意しておき, 観測データにフィッティングさせた際に必要な分だけ説明に寄与するようなモデルの設計の仕方が考えられてきた. このとき, 観測データは有限なため, 無限のパラメータを使った生成モデルを考えることが出来ればデータの説明に寄与する分は観測データそれ自身の要請に従って必要な分だけ表出してくるように働くはずである. では無限のパラメータを使った生成モデルを描くにはどうすれば良いだろうか.

2.3 モデル化のための確率変数の設計

ノンパラメトリックベイズの基本的な戦略は無限のパラメータを用いた確率的な生成モデルを描くことにあったが, ではどのように無限のパラメータを取り扱うことが出来るだろうか. 直観的には無限のパラメータを作った際にそれらの出現に対する確率が与えられているようなモデルと捉えることが出来る. つまり確率分布 $f(\theta)$ から確率変数 x が生成ことを表す $x \sim f(\theta)$ において, x として例えば無限実数列のようなものが作れればよい訳である. ノンパラメトリックベイズモデルは多くの場合確率過程を用いてこのような無限のパラメータから成るモデルを構成する. このことを以降の議論で直感的に了解するために, 確率空間, 確率変数, 分布とはどんなものであったかを簡単に整理し, 先の無限数列のような色々なものが確率変数として扱うことが出来ることを確認しておく.

集合 Ω の要素 ω を標本点と呼び, σ 加法族 $\mathcal{B}$, すなわち, 1) $\phi \in \mathcal{B}$, 2) $A_n \in \mathcal{B}$ $(n = 1, 2, \ldots) \Rightarrow \cup_n A_n \in \mathcal{B}$, 3) $A \in \mathcal{B} \Rightarrow A^c \in \mathcal{B}$, によって決まる $(\Omega, \mathcal{B})$ を可測空間と呼ぶ. $\mathcal{B}$ を定義域とする関数 P が, 1) $A \in \mathcal{B} \Rightarrow 0 \leq P(A) \leq 1$, 2) $P(\omega) = 1$, 3) $\{A_1, \ldots, A_N\}$ が排反ならば $P(\cup_n A_n) = \sum_n P(A_n)$, のとき確率測度と呼び, $P(A)$ を事象 A の確率, $(\Omega, \mathcal{B}, P)$ を確率空間と呼ぶ. 今 Ω 上に定義され実数値をとる関数 $X(\omega)$ で $B(r) = \{\omega; X(\omega) \geq r\}$ に対して $P(B(r))$ が定まるものを考えたとき, これを可測関数といい確率変数とみなすことが出来る. つまり確率変数とはとる値に対して必ずその確率を計算できるものであると捉えればよい. 直観的な把握には色々な例があると思うが, 例えばサイコロのようなものを連想しそのサイコロが相当に柔軟なものと考えてしまうのが簡単である. サイコロは偶然現象を表しており, サイコロを振ったときに一つの面が出るのが標本点 ω に対応している. 偶然現象は「あの面 ω_i が出る」「この面 ω_i もしくはあの面 ω_j が出る」などの事象として表すことができ, これらの集合が σ 加法族 $\mathcal{B}$ に対応している. 各事象はサイコロの歪み方など P によって確率を計算することが出来る. いまサイコロの各面にはある実数が割り当てられていて, サイコロを振ってその目が出たとき, そこに描かれた実数を吐き出されるように見なすことで確率変数と捉えることが出来る.

確率変数としての可測関数は必ずしも実数値をとる必要はなく $X(\omega)$ として多次元の実数や, 無限次元の数列, 関数（このような時に確率過程と呼ばれる）を考えることも出来る. 確率変数として確率過程のようなものを考える場合においても, あくまでサイコロのような偶然現象を表すものがあり, 出た面に割り当てられた出力が確率変数だという捉えることが直観的な理解を助けてくれる.

前述の通りノンパラメトリックベイズモデルにおいては無限のパラメータを扱う性質から, 確率過程を導入することが多い. 具体的なモデル化において登場するディリクレ過程, レヴィ過程（の一部であるベータ過程やガンマ過程）, などのそれぞれについては各論で触れていきたいが, それらと密接な関係があり以降の議論でも登場するポアソン過程を例としてここで簡単に紹介しておく [13].

いま時間 t にともなって変化する偶然現象 $X(t, \omega)$ を考えることにする. 偶然を司る ω が一つに決まれば (サイコロをようなものを振ったとして ω が決まれば)ω を省略して $X(t)$, すなわち時間 t に伴って変化する関数が生成されたと捉えることが出来る. パラメータ $\lambda > 0$ を持つポアソン過程 $X(t)$ とは次の性質を持つものを指す:

- X(t) は非負整数値をとり, X(0)=0.
- $0 \leq t_1 \leq t_2 \rightarrow X(t_1) \leq X(t_2)$.

- $0 \leq t_1 < t_2 < \cdots < t_N$ について, $X(t_1) - X(0), X(t_2) - X(t_1), \ldots, X(t_N) - X(t_{N-1})$ は独立.
- $0 \leq t_1 \leq t_2$, $h > 0$ のとき, $X(t_2) - X(t_1)$ と $X(t_2 + h) - X(t_1 + h)$ の分布が同一.
- $P(X(h) = 1) = \lambda h + o(h)$, $P(X(h) \geq 2) = o(h)$ $(h \downarrow 0)$. ただし, $o(h)$ は $P(X(h) \geq 2)/h \to 0$ $(h \downarrow 0)$.

$X(t)$ は時間 t に伴って発生する何らかのイベントの発生回数を数えているもので, 発生のしやすさの程度が時間の経過に対して一律で λ に支配されていると考えると直感的に理解しやすい. ポアソン過程には様々な良い性質があるがその一部だけ紹介すると:

- 時刻 t_1 から時刻 t_2 までに発生するイベントの回数は $\mathrm{Poisson}(\lambda(\mathrm{t}_2 - \mathrm{t}_1))$ に従う.
- あるイベントから次のイベントの発生までの時間は $\mathrm{Exp}(\lambda)$ に従う.
- $(0, t]$ において一回だけイベントが発生したとすれば, その発生時刻は $\mathrm{Uniform}(0, t)$ に従う.

ポアソン過程を実際に構成したい場合（サイコロを振って一つの具体的なポアソン過程 $X(t)$ を生成したい場合）には上記の 2 番目の性質を使って, 指数分布に従うイベント発生時間間隔を次々に生成していけばよい. またイベント発生の程度を表すパラメータ λ は時間の経過に対して一律である必要はなく, 時刻に依存するように $\lambda(t)$ とすることも出来る. 例えば上記一つ目の性質は次のようになる:

- 時刻 t_1 から時刻 t_2 までに発生するイベントの回数は $\mathrm{Poisson}(\lambda \int_{\mathrm{t}_1}^{\mathrm{t}_2} \lambda(\mathrm{t}))$ に従う.

1 次元のポアソン過程の説明として, 簡単のため t を時刻のように捉えてきたが, これは必要に応じて所望の対象を置き換えてよく, また多次元として考えることも出来る. ノンパラメトリックベイズモデルに頻出のディリクレ過程, ベータ過程, ガンマ過程, ベルヌーイ過程などはこのポアソン過程と密接な関係にある.

3 混合モデルと因子モデル

多くの確率的生成モデルは混合モデル（隠れマルコフモデル, 確率文脈自由文法, n-gram）と因子モデル（非負値行列因子分解, 独立成分分析）の二つに大別して捉えることが出来る. まずこの 2 つのモデルの基本的な考え方とそれぞれの無限モデル化について整理する.

3.1 混合モデル

データの分類やクラスタリングは様々なメディアに現れる頻出の問題である. 例えば画像に「花」や「人」のラベルを付ける問題や, 音楽において各楽曲を「ジャズ」, 「ポップス」のように分類する問題がこの典型である. ベイズ的な方法によってこのような問題を扱う場合, 混合モデルという考え方が重要な役割を果たす.

N 個の観測データ $y_1, y_2, \ldots, y_N$ それぞれに対して, K 種類のラベルの中の一つを割り当てることを考える. いま n 番目のデータ y_n に割り当てられるラベルのインデックス（$1, 2, \ldots, K$ の中の一つ）を z_n と表すことにする. 典型的な分類やクラスタリングは $y_1, y_2, \ldots, y_N$ と分類したいクラスタの数 K が与えられたときに $z_1, z_2, \ldots, z_n$ を推定する問題だと捉えることが出来る (K が未知の場合の取扱い方は後述する). ベイズ的なアプローチにおける基本的な戦略は, 観測データの確率的な生成過程を描くことにある. 混合モデルは次のような生成過程によって観測データが確率的に生成されたと考えるものである.

1. 各ラベル k ごとにデータ生成用のパラメータ θ_k をそれらの事前分布 F から生成する: $\theta_k \sim F$ $(k = 1, 2, \ldots, K)$.
2. 各データに K 種のラベルの中の一つを確率的に割り当てるための離散分布 (和が 1 となる K 次元ベクトル) の重みを生成する: $\pi = (\pi_1, \ldots, \pi_K) \sim \mathrm{Dirichlet}(\alpha/K, \ldots, \alpha/K)$.
3. 各データ y_n に離散分布 π から一つのラベル z_n が割り当てられる: $z_n \sim \mathrm{Discrete}(\pi)$ $(n = 1, \ldots, N)$.
4. 各データ y_n が z_n 番目のラベルの持つパラメータ θ_{z_n} からある分布 $f(\theta_{z_n})$ から生成される: $y_n \sim f(\theta_{z_n})$ $(n = 1, \ldots, N)$.

3.1.1 ディリクレ過程と無限混合モデル

一般にベイズ的なモデリングにおいてモデルの複雑度をどのように設定するかは極めて重要な問題であり, 近年の標準的な取扱いとして, 無限混合モデルがよく用いられる (Rasmussen, 2000). 最もよく用いられるものの一つとして, ディリクレ過程 (Ferguson, 1973) を使ったモデルを紹介する.

ディリクレ過程は確率測度に対する確率分布であり, 集中度と呼ばれる正の実数 γ と可測空間 $(\Theta, \mathcal{B})$ 上の基底測度 F が与えられたとき, ディリクレ過程から生成された確率測度 $G \sim \mathrm{DP}(\gamma, F)$ は $G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k}$, $\theta_k \sim F$, の形で表されることが知られている. 重み $\boldsymbol{\beta}$ を構成する方法としては棒折り過程 (Sethuraman, 1994) が有名である: $\beta_k = \beta'_k \prod_{l=1}^{k-1}(1 - \beta'_l)$, $\beta'_k \sim \mathrm{Beta}(1, \gamma)$. (この分布は簡単に $\boldsymbol{\beta} \sim \mathrm{GEM}(\gamma)$ と表されることが多い)

ディリクレ過程の性質は前述の有限混合モデルと関連付けて捉えると分かりやすい. 有限混合モデルでは, 各コンポーネントの重みをディリクレ分布から生成し, k 番目のコンポーネントに紐付いたラベルのもつパラメータ θ_k を事前分布 F から生成していた. ディリクレ過程から生成された確率測度は各コンポーネントの重み π_k と, そのラベルに紐付いたパラメータ θ_k を同時に表現していると捉えると, ディリクレ過程は潜在的に無限のコンポーネン

トを持った混合モデルの構成に用いることが出来る:

$$G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k} \sim \mathrm{DP}(\gamma, F), \quad z_n \sim \pi, \quad y_n \sim f(\theta_{z_n}). \quad (1)$$

実際, 前述の有限混合モデル (ディリクレ分布-離散分布) において $K \to \infty$ とすればディリクレ過程混合モデルとなることが知られている（Ishwaran & Zarepour, 2002）. 具体的な構成法や推論 (例えば Kenichi et al., 2007, Paisley et al., 2011, Walker, 2007, Papaspiliopoulos, 2008) については紙面の都合上省略させて頂く.

3.1.2 ディリクレ過程混合モデルを用いる際の注意: コンポーネント数は推定出来るのか？

ディリクレ過程はコンポーネントの数が未知の場合の混合モデルの構成に広く用いられるようになっているが,「コンポーネントの数を推定する」という文脈で使う場合には注意が必要であることが近年の研究で指摘されている. ディリクレ過程混合モデルは, 混合分布の密度関数や混合の重みへのフィッティングの意味においては適しているが, コンポーネントの個数の推定には不向きであることが報告されている (Miller & Harrison [5]).

詳細は Miller & Harrison [5]を参照して頂きたいが, ここでは概略だけ紹介したい. ディリクレ過程は非常に少数のデータが割り当てられているようなコンポーネントが大量に発生することに寛容な性質を持っており, データ生成の尤度部分がこのディリクレ過程の事前分布における性質を上回ることが出来ずに無駄なコンポーネントの発生を抑えられない場合がある. コンポーネントの数自体を推定したい場合の代替案も併せて議論されている. .

3.1.3 混合モデルの例: 隠れマルコフモデル

紙面の都合上無限隠れマルコフモデル (Beal et al., 2002) を例にとり, その生成モデルが混合モデルとして記述できることを確認していく. 隠れマルコフモデルは観測系列 $\boldsymbol{y} = (y_1, y_2, \ldots, y_T)$ の背後に隠れ状態の系列 $\boldsymbol{z} = (z_1, z_2, \ldots, z_T)$ を考えることにより系列データのモデル化を行う. z_t は時刻 t における状態のインデックスを表している. 通常の隠れマルコフモデルでは 1 次のマルコフ性が仮定され, 状態の遷移 $i \to j$ は状態遷移確率 $\pi_{i,j} = p(z_t = j \mid z_i = i)$ によって決まるものとして取り扱う. 現在の状態 z_t は前の状態 z_{t-1} から $z_t \sim \boldsymbol{\pi}_{z_{t-1}}$ と表現することが出来る. 各状態 k はパラメータ θ_k を持ち, 観測系列は出力分布 $y_t \sim f(\theta_{z_t})$ から生成されたものと見なす. では無限モデルである無限隠れマルコフモデルはどのように構成できるだろうか. 状態遷移確率 π_i $(i = 1, 2, \ldots)$ は重みの総和が 1 となる多項分布であると考えることが出来ることから, それぞれの π_i をディリクレ過程に従う確率測度の重みと見なすことで無限次元多項分布のように振る舞うことが期待される. つまり, $\pi_i \sim \mathrm{GEM}(\gamma)$ $(i = 1, 2, \ldots)$ とすることで無限状態を持つ隠れマルコフモデルが構成出来るかに思われる. しかしここで各重みに割り当てらたパラメータに注意を払わなければならない. 例えば π_3, π_5 が独立なディリクレ過程の確率測度の重みだと考えたとき, $\pi_{3,2}, \pi_{5,2}$ ともに “状態 2” のアトム θ_2 に対応した重みである必要がある. しかし前述のような確率測度の重みだけの議論ではこのようなアトムの共有が陽に保証されていない. そこで次のような階層ディリクレ過程 (Teh et al., 2006) が用いられる.

階層ディリクレ過程とは, あるディリクレ過程から生成された測度を基底測度として共有するディリクレ過程の集合のことを指す. すなわち, 確率測度の集合 $G_i \sim DP(\alpha, G_0)$ $(i = 1, 2, \ldots)$ で, 基底測度が $G_0 \sim \mathrm{DP}(\gamma, H)$ となるものを指す. ディリクレ過程の性質から, 上の階層のディリクレ過程に従う G_0 はアトミックになり, 下の階層のディリクレ過程はアトミックになった基底測度を用いることで G_0 においてアクティブなアトムにだけ重みを持つようになる. 各確率測度は $G_i = \sum_{k=1}^{\infty} \pi_{i,k} \delta_{\theta_k}$ $(i = 1, 2, \ldots)$. と表せるようになり, 各重みの受け持つアトム (パラメータ) を陽に共有化することが出来る. そこで, $\pi_{i,j}$ を状態 i から状態 j への遷移確率, θ_k を状態 k のもつ出力分布パラメータだと見なすと, 階層ディリクレ過程に基づく隠れマルコフモデルは次のように表すことが出来る:

$$\begin{aligned} \boldsymbol{\beta} \sim \mathrm{GEM}(\gamma),\ \pi_i \sim \mathrm{DP}(\alpha, \boldsymbol{\beta}),\ z_t \mid z_{t-1} \sim \pi_{z_{t-1}}, \\ \theta_k \sim H,\ y_t \mid z_t, \theta \sim f(\theta_{z_t}) \quad . \end{aligned} \quad (2)$$

3.2 因子モデル

混合モデルは各データが一つのラベルに割り当てられたパラメータから生成されるモデルになっていた. もう一つの典型的なモデルとして, 各データがいくつかの特徴量の組み合わせから生成されたと見なすものが因子モデルである. 各データが特徴量からの線形モデルとなる場合は非負値行列因子分解や独立成分分析などのメディア処理に頻出なモデルと捉えることが出来る. ノンパラメトリックベイズにおいては無限の因子とその結合の係数に対する事前分布を適切に設定することが重要になる.

3.2.1 ベータ過程, ガンマ過程, ベルヌーイ過程と無限因子モデル

潜在的に無限個の特徴量を持つような因子モデルを構成する方法として, 特別な場合のレヴィ過程がよく用いられる. レビィ過程 $X(\omega)$ は可測空間 $(\Phi, \mathcal{F})$ 上に独立な跳ねを持つ確率過程で, 無限因子モデルにはおいては離散的でそれらの跳ねが正となるようなベータ過程, ガンマ過程, ベルヌーイ過程がよく用いられる. 上記の特別な場合のレヴィ過程は $R^+ \times \Omega$ の直積に対してレヴィ測度と呼ばれる ν を持つポアソン点過程と見なすことが知られている（Sato, 1999, Wang & Carin, 2012）.

まずレヴィ過程の特別な場合の例として, ベータ過程を紹介する. 集中度を $c > 0$, 基底測度を μ とするベータ過程 $B \sim \mathrm{BP}(c, \nu)$ とは $B(d\omega) \sim \mathrm{Beta}(c\mu(d\omega), c(1-\mu(\omega)))$ となるようなレヴィ過程のことをいう. このとき, B はレヴィ測度を $\nu(d\pi, d\omega) = c\pi^{-1}(1-\pi)^{c-1}d\pi\mu(d\omega)$ とする $R^+ \times \Omega$ へのポアソン点過程とも見なすことが出来る. ここで $c\pi^{-1}(1-\pi)^{c-1} = \mathrm{Beta}(0, c)$ であり, $(0, 1)$ において積分したときに無限になっていることに注意するとポアソン点過程の性質から, B には無限の点が発生していると見なすことができ, $B = \sum_{i=1}^{\infty} \pi_i \delta_{\omega_i}$ のようにアトム ω_i において跳ね π_i を持つような形で表すことが出来ることが知られている. アトムを各特徴量を表すパラメータ, 跳ねを各特徴量のデータへの貢献を表す重みとして用いることで無限因子モデルを構成することが出来る.

無限因子モデルの構成には必ずしもベータ過程を使う必要はなく, レヴィ測度はモデル化や推論の都合に合わせて適切なものを設定すればよい. 以下に述べるガンマ過程も実用的によく用いられる. レヴィ測度を $\nu(d\phi, d\theta) = c\theta^{-1}e^{-c\theta}d\theta F(d\phi)$, のように選ぶ. ただし, $c > 0$ は集中度を表している. $c\theta^{-1}e^{-c\theta}$ は $\mathrm{Gamma}(0, 1/c)$ となっており, $(0, \infty)$ で積分した際に無限になるため, ガンマ過程 $G \sim \Gamma\mathrm{P}(c, F)$ は無限のアトムをもち, 次のように表すことが出来る: $G = \sum_{d=1}^{\infty} \theta_d \delta_{\phi_d}$. ベータ過程の跳ねは $(0, 1)$ だったのに対し, ガンマ過程の跳ねは $\mathbb{R}^+$ になっている.

無限因子モデルを利用する際に, 無限の因子の各データに対する寄与の有無, すなわち 0 または 1 のバイナリな情報だけ用いたい場合も少なくない. そのような場合にはベータ過程-ベルヌーイ過程の階層モデルを用いるのが一般的である（Te et al., 2007, Thibaux & Jordan, 2007）. 無限因子モデルの構成に用いられるインド料理過程はこの階層モデルの特別な場合に対応している（Thibaux & Jordan, 2007）.

3.2.2 因子モデルの例: 非負値行列因子分解

因子モデルの代表的のものとして非負値行列因子分解 (の確率的生成モデル) による音楽信号を分離する問題を例に, それの構成法を紹介する. 振幅スペクトログラムもしくはパワースペクトログラム $Y = (Y_{\omega,t})_{\Omega \times T} \in R^{\geq 0, \Omega \times T}$ (ただし, $\omega = 1, \cdots, \Omega$ は周波数インデックス, $t = 1, \cdots, T$ は時間インデックスを表す) が基底 $H = (H_{\omega,d})_{\Omega \times D} \in R^{\geq 0, \Omega \times D}$ とアクティベーション $U = (U_{d,t})_{D \times T} \in R^{\geq 0, D \times T}$ の積で表現できるという仮定に基づいている. これはすなわち, $Y_{\omega,t} \approx \sum_d H_{\omega,d} U_{d,t}$, のように観測スペクトログラム Y を D 個の頻出の基底スペクトル $h_d = [H_{1,d}, \cdots, H_{\Omega,d}]$ とそれぞれの音量変化を表すアクティベーションで近似しようとしていることに相当する. 音響信号によく合った尤度関数の選び方はそれ自身大きな研究対象の一つではあるが, ここでは標準的なものの一つとして, 指数分布を用いると次のようになる: $Y_{\omega,t} \sim \mathrm{Exp}(1/\sum_d H_{\omega,d} U_{d,t})$. これは板倉斎藤距離規準の非負値行列因子分解に対応している（Févotte et al., 2009）. $\boldsymbol{H}$ と $\boldsymbol{U}$ に対する事前分布には例えばガンマ分布を用いるのが標準的である: $H_{\omega,d} \sim \mathrm{Gamma}(a_H, b_H)$, $U_{d,t} \sim \mathrm{Gamma}(a_U, b_U)$. 一般にデータを説明するのに必要な因子の適切な数 D を事前に決めるのは困難であり, 無限非負値行列因子分解が有用になってくる. 先のように尤度関数として指数分布を選んだときには推論の都合上ガンマ過程を用いるのが簡単である（Hoffman et al., 2010）. 各コンポーネントごとのゲインの総量を表す θ_d $(d = 1, 2, \dots)$ を導入し, $Y_{\omega,t} \sim \mathrm{Exp}(1/\sum_d \theta_d H_{\omega,d} U_{d,t})$, $G = \sum_{d=1}^{\infty} \theta_d \delta_{\phi_d} \sim \Gamma\mathrm{P}(c, F)$ のように, θ_d $(d = 1, 2, \dots)$ をガンマ過程の跳ねだと考える方法が提案されている. ϕ_d は d 番目のコンポーネントにおけるパラメータ, すなわち $(H_{1,d}, \dots, H_{\Omega,d}), (U_{d,1}, \dots, U_{d,T})$ を表しており, F はそれらに対する事前分布（ガンマ分布など）を考えればよい.

4 メディア処理のためのベイズモデル

各種メディアに対し数多のノンパラメトリックベイズモデルが提案されており, それらを網羅することは到底不可能であるが, 紙面の許す範囲で一部だけでも紹介したい.

4.1 bag-of-words を表現したモデル

自然言語処理において文書中に登場する単語の生成過程を描く頻出のモデルとしてトピックモデルがある（Blei et al., 2003）. その汎用性の高さからコンピュータヴィジョンや音響信号処理分野でも盛んに応用されている. 最も素朴なモデルは階層ディリクレ過程の例題としてもよく用いられる. 各単語はある隠れたトピックから生成されたと仮定し, 各文書における各トピックの出やすさと各単語における隠れたトピックの割り当てを同時に考えたモデルになっている. J 個の文書が観測されたとき, j 番目の文書における n 番目の単語の生成過程は次のように描かれる:

1. 潜在的に無限のトピックの出やすさを表す確率測度を生成する: $G_0 = \sum_{t=1}^{\infty} \lambda_t \delta_{\boldsymbol{\phi}_t} \sim \mathrm{DP}(\gamma, F)$. ϕ_t は t 番目のトピックにおける各単語の出現頻度を表す離散分布であり, それらへの事前分布となるディリクレ過程の基底測度 F には例えばディリクレ分布を使うのが一般的である.
2. j 番目の文書中での各トピックの出現しやすさを表す確率測度を生成する: $G_j = \sum_{t=1}^{\infty} \theta_{j,t} \delta_{\boldsymbol{\phi}_t} \sim \mathrm{DP}(\alpha, G_0)$. 各トピックがどの程度現れやすいかを表す G_0 の傾向が階層ディリクレ過程によって G_j $(j = 1, 2, \dots)$ に反映されている.
3. j 番目の文書の n 番目の単語に対するトピックの割り当て $z_{j,n}$ を生成する: $z_{j,n} \sim \boldsymbol{\theta}_j$.
4. j 番目の文書の n 番目の単語に対して, $z_{j,n}$ 番目の

トピックにおける単語の出現頻度を表した離散分布 $\boldsymbol{\phi}_{z_{j,n}}$ から単語 $x_{j,n}$ を生成する: $x_{j,n} \sim \boldsymbol{\phi}_{z_{j,n}}$.

トピックモデルの改良として, 混合モデルと因子モデルの両方を組み込んだ focused topic model と呼ばれるモデルが提案されている（Williamson et al., 2010）. 先のトピックモデルは階層ディリクレ過程のおかげで, 各文書においてあらゆるトピックの出現を 0 でない確率で考えることが出来た. 一方, 各文書中に現れるトピックは非常にスパース, つまり同一文書には数少ないいくつかのトピックしか出現しないものと仮定したモデルｍｏ考えられうる. このような性質を陽を組み込んだ拡張として, 先のトピックモデルにおける G_j $(j = 1, 2, \dots)$ を次のように修正することが考えられる:

$$B_{j,t} \sim \text{Bernoulli}(a_t), \quad a_t \sim \text{Beta}(a_0/T, b_0(T-1)/T), \tag{3}$$

$$\boldsymbol{\theta}_j \sim \text{DP}(\alpha, \boldsymbol{B}_j \odot \boldsymbol{\lambda}) \quad . \tag{4}$$

$(a_1, \dots, a_T)$ はベータ過程（簡単のため有限打ち切りで表記）の重み, $(B_{j,1}, \dots, B_{j,T})$ はそれらを上の階層にしたベルヌーイ過程のバイナリな重みである. $\boldsymbol{B}$ はインド料理過程から生成されたと見なしても構わない. $(B_{j,1}, \dots, B_{j,T})$ は j 番目の文書における各トピックの出現有無をバイナリに表現しており, $\boldsymbol{B}_j \odot \boldsymbol{\lambda}$(ベクトルの要素ごとの積) によってトピックの候補がスパースになるように働く.

4.2 状態空間を用いた系列データのモデル

隠れマルコフモデルのように観測系列をスパースな状態によって表現しようとするモデルはメディア処理の多くの場面で非常に役に立つ. 観測データのそれぞれに一つの状態で表現されその状態遷移が一次のマルコフ性を持つという意味で隠れマルコフモデルを最も基本的な系列データのモデルと見なすことが出来るが, 対象とするメディアや課題に合わせ様々なモデルが用いられている. いくつか主要な拡張として次のようなものを挙げることが出来る.

- 状態遷移行列を所望の形へ誘導する: stick HDP-HMM（Fox et al., 2007）, block diagonal infinite HMM（Stepleton et al., 2009）
- 系列の長い依存関係を表現すべく高次のマルコフ性を考える: n-gram（Mochihashi& Sumita, 2007）, sequence memoizer（Wood et al., 2009）
- 系列の構文を解析すべく木構造の状態割り当てを考える: 確率文脈自由文法（Liang et al., 2007）
- 状態空間を因子の組み合わせによって表現する: infinite factorial HMM（Van Gael et al., 2008）, infinite latent event model（Wingate et al., 2009）, infinite dynamic Bayesian net.（Doshi-Velez et al., 2011）

無限隠れマルコフモデルにおける状態遷移行列を所望の形に誘導しようとする研究として sticky HDP-HMM や block diagonal infinite HMM などが知られている. Fox et al. (2008) はスティッキー階層ディリクレ過程隠れマルコフモデルと呼ばれるモデルを提案した. これは, 系列をゆっくりと変化する隠れ状態の遷移によって表現しやすくなるように改良を加えたものである. 何らかの系列を隠れマルコフモデルで解析する状況において, 隠れ状態がゆっくりと変化してほしい (すなわち自己遷移が起こりやすくなってほしい) 場合は多々ある. このような性質を無限隠れマルコフモデルに導入する方法として, 式 (2) において遷移確率を次のように変更する拡張が提案された: $\boldsymbol{\pi}_i \sim \text{DP}(\alpha + \kappa, \frac{\alpha\boldsymbol{\beta} + \kappa\delta_i}{\alpha+\kappa})$. $\kappa > 0$ によって自己遷移への遷移確率にバイアスがかかることでゆっくりと変化するダイナミクスを捉えやすくなる. このように状態遷移行列を所望の形に誘導する拡張としてブロック対角の形 (Stepleton et al., 2009) がある.

系列の長い依存関係を表現するには高次のマルコフ性を考えた n-gram が有用である. n-gram は $(n-1)$ 次のマルコフ性をもった系列のモデルであり, $(n-1)$ 個の系列に依存して次が候補の中から確率的に選ばれるものと考えることで混合モデルの一種と見なすことが出来る. 一般に依存関係の長さは系列の中でも可変として考えられる方が便利であり, 無限モデルとしての可変長 n-gram が提案され, その構成には階層ピットマンヨー過程（Pitman & Yor, 1997）が用いられている. ピットマンヨー過程もディリクレ過程同様それ自身が確率測度と見なすことが出来るため無限混合モデルの構成に用いることが出来る. また同種の系列データのために sequence memoizer と呼ばれるモデルも提案されている. これらは自然言語処理における単語系列のモデルとして使われることが多いが, 近年は音楽における旋律のモデル化などにも利用されている. さらに, 単語レベルでの ∞-gram に, 文字レベルの ∞-gram を入れ子にしたモデルを用いることで教師なしの単語分割に用いる研究も行われている.

自然言語処理における構文解析への応用に代表されるように確率文脈自由文法もメディア処理で度々登場するモデルで, これは一種の混合モデルとして構成することが出来る. 簡単のためチョムスキー標準形（親ノードが二つの子ノードに分岐）を考えるとすると, ある状態から二つの状態のペアへの確率的な分岐規則を司る確率測度が作ることが出来ればよく, これは隠れマルコフモデルの時と同様に階層ディリクレ過程によって表現することが出来る. 確率文脈自由文法は自然言語処理における構文解析のみならず, 近年は音楽の多重音を 2 次元の確率文脈自由文法と因子モデルの組み合わせで表現しようという試みもある（Nakano et al., 2011, Kameoka et al., 2012）.

系列データを因子モデルによって表現しようとする研究もここ数年盛んに行われている. 無限階乗隠れマルコフモデル (Van Gael et al., 2009) や無限階層隠れマルコフモデ

ル (Heller et al., 2009) では, 観測系列を表現するために無限の (バイナリもしくは整数値をとる) 因子の組み合わせが導入された. また, 無限イベントモデル (Wingate et al., 2009) や無限ダイナミックベイジアンネットワーク (Doshi-Velez et al., 2011) では因子の組み合わせおよびそれらの間のネットワークを同時に推定する手法が提案されてきた.

4.3 入れ子構造を用いたモデル

モデルの中に入れ子構造を導入することでしばしば有用な拡張を行うことが出来る（Jordan, 2009）. 例えば画像のクラスタリングを階層的に行う問題, すなわち階層クラスタリングを考えてみる. バラの花が映ったある画像は「花」という大きな単位で分類され, その「花」の中でさらに細かい単位として「バラ」のように分類したい. いま n 番目のデータ y_n に割り当てられる「花」のような大きなレベルのラベルのインデックス（$1, 2, \ldots, K$ の中の一つ）を z_n と表すことにすると混合モデルとして $z_n \sim \boldsymbol{\pi}$, $G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k} \sim \mathrm{DP}(\gamma, G_0)$ と表せる. ここで k 番目のコンポーネントが持つ ϕ_k およびその事前分布 G_0 をどのように扱えばよいだろうか. 「花」のような大きな単位でのラベル z_n が割り当てられた後, データ y_n はさらに細かい単位のラベルとして「バラ」のような割り当てが行われることが期待されている. そこで細かい単位でのラベルのインデックスを z'_n とすると, これも混合モデルとみなすことが出来ることから

$$z'_n \sim \boldsymbol{\pi}'_{z_n},\ G'_{z_n} = \sum_{m=1}^{\infty} \pi'_{z_n,m} \delta_{\phi_{z_n,m}} \sim \mathrm{DP}(\alpha, F), \qquad (5)$$

$y_n \sim f(\phi_{z_n, z'_n})$ のように表現したい. このように考えると $G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k}$ のアトム θ_k には細かいレベルでのラベル割り当てを行うための確率測度（すなわちディリクレ過程）G'_k が対応していることが分かる. これはディリクレ過程が入れ子になっている様子を陽に表した形として

$$G = \sum_{k=1}^{\infty} \pi_k \delta_{G'_k} \sim \mathrm{DP}(\gamma, \mathrm{DP}(\alpha, F)) \qquad (6)$$

のようにも表される. $\phi_{k,m}$ はデータを生成するためのパラメータが格納されており F がそれらの事前分布に対応していると考えればよい.

ここでは2段の入れ子構造を持ったディリクレ過程（Rodriguez et al., 2008）のみを紹介したが, 多段や無限の深さを考えることも出来る. また同種の考え方は因子モデルにおいても用いられており, nested beta process や nested gamma process などが考えられている（Jordan, 2009）.

4.4 共変量を介して依存関係を陽に表現したモデル

観測データのそれぞれには何らかの共変量があり, それらが似ているものについてはデータの生成過程も似た振る舞いになることが期待される場面がしばしばある. 例えば, 画像の領域分割はピクセルまたはスーパーピクセルのようにあらかじめ細分化された各領域に対してラベル付け（同一ラベルは同一領域）を行う問題と捉えることが出来るが, もともと画像上で隣り合った領域はそもそも同一ラベルがつきやすいはずである. したがって画像上での「位置」は細分化された領域に対するある種の共変量とみなすことが出来る. 別の例として, 音楽のような繰り返し構造を持った信号を各楽器音に分解する問題を考えたとき, これは信号が楽器音の線形結合で表せるとみなした因子モデルを適用するのが標準的な扱いであるが, 音楽のようにAメロ, Bメロ, サビのような構造を持った信号の場合1回目のサビと2回目のサビでは似たような因子がアクティブになることが期待される. したがって, 音楽の（隠れた）繰り返し構造は各時刻の信号を分解する上での共変量となりうる. このようにデータ間の依存関係を導入しようとした研究は混合モデル, 因子モデルともに数多く報告されている（Williamson et al., 2010, Ren et al., 2011）.

データ間の依存関係を導入した因子モデルの一つに kernel Beta process [42]を用いたものが挙げられる. これは各データと因子に共変量を導入しそれらが似ているものは共起しやすくなるようにベータ過程の跳ねを修正することで, 共変量を介してデータ間の依存関係を表現しようとしたものである. データ Y_n に対応する共変量を $\boldsymbol{x}_n \in \mathcal{X}$, factor ω_i に対応する共変量を $\boldsymbol{x}^*_i \in \mathcal{X}$ と表すことにする. 各因子は共変量空間において, それぞれの局所的な性質を表すためのパラメータ $\psi^* \in \Psi$ を用いると, 共変量空間においてデータと因子の間の類似度 $K(\boldsymbol{x}, \boldsymbol{x}^*; \psi^*)$（例えば $K(\boldsymbol{x}, \boldsymbol{x}^*; \psi^*) = \exp(-\psi^* ||\boldsymbol{x} - \boldsymbol{x}^*||_2)$）を考えることが出来る. ここで, 共変量 $\boldsymbol{x}$ の影響を受けた $B_{\boldsymbol{x}}$:

$$B_x = \sum_{i=1}^{\infty} \pi_i K(\boldsymbol{x}, \boldsymbol{x}^*; \psi^*) \delta_{\omega_i} \qquad (7)$$

を考えると, これはベータ過程の重みが共変量の近さによって修正されたものと見なすことが出来る. つまり, 共変量を介してデータ間の依存関係を陽に導入することが出来る. $\boldsymbol{x}^*, \psi^*$ がそれぞれ確率測度 S, Q から生成されたとすると, $K(\boldsymbol{x}_n, \boldsymbol{x}^*_i; \phi^*_i) \in (0, 1]$, $K(\boldsymbol{x}_n, \boldsymbol{x}^*_i; \phi^*_i) \to 0$ $(||\boldsymbol{x}_n - \boldsymbol{x}^*_i||_2 \to \infty)$, $K(\boldsymbol{x}^*, \boldsymbol{x}^*; \phi^*) = 1$ を満たすとき, これは $\nu_{\mathcal{X}} = S(d\boldsymbol{x}^*) Q(d\boldsymbol{\psi}^*) \nu(d\pi, d\omega)$ をレヴィ測度（$\nu(d\pi, d\omega)$ は通常のベータ過程と同様）とするレヴィ過程となることが知られている (Ren et al., 2011).

4.5 互いに似ていないものの同時出現を表現したモデル

時々刻々と更新されていくテキストの中から,（時間に伴って変化する）ニュースの見出しを作る問題を考えてみる. その時々でいくつかのニュースの見出しを作る際, それぞれ一つ一つが重要なものを指しているのと同時に, お互いが違うことを指しているのが望ましい. 例えば, ある時刻に同一のことを指して「火事」と「火災」の二つの見出し

が出現するのは望ましくない. このように似ていないものが同時に出現しやすくなるような機構をモデル化したい場面は少なくない. 近年このような互いに似ていないものを表現するために行列式点過程を用いる方法が提案されている (Kulesza& Taskar, 2010, Affandi et al., 2012).

5 おわりに

各種メディア処理においてベイズ的な手法は有力の選択肢の一つとなってきている. 近年の発展を把握する上で本稿および発表がその概観に役立てば幸いである.

参考文献

[1] T. S. Ferguson, "A Bayesian analysis of some nonparametric problems," *Annals of Statistics*, 1(2): pp. 209-230, 1973.

[2] J. Pitman and M. Yor, "The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator," *Annals of Probability*, 25: pp. 855-900, 1997.

[3] J. Sethuraman, "A constructive definition of Dirichlet priors," *Statistica Sinica*: 4, pp. 639-650, 1994.

[4] C. E. Rasmussen, "The infinite Gaussian mixture model," In *Advances in Neural Information Processing Systems*, vol. 12, 2000.

[5] J. W. Miller and M. T. Harrison, "Dirichlet process mixtures are inconsistent for the number of components in a finite mixture, " in *ICERM*, 2012.

[6] K. Kenichi, M. Welling, and Y. Whye Teh, "Collapsed variational dirichlet process mixture models" in *Proc. IJCAI*, 2007.

[7] D. J. Aldous, "Representations for Partially Exchangeable Arrays of Random Variables," *Journal of Multivariate Analysis*, 11: pp. 581-598, 1981.

[8] J. Paisley, C. Wang and D. Blei, "The discrete infinite logistic normal distribution for mixed-membership modeling", in *Proc. of International Conference on Artificial Intelligence and Statistics*, 2011.

[9] S. G. Walker, "Sampling the Dirichlet mixture model with slices," *Communications in Statistics* - Simulation and Computation, 36:45, 2007.

[10] O. Papaspiliopoulos and G. O. Roberts, "Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models," *Biometrika*, 95(1): pp. 169-186, 2008.

[11] K. Sato, "Lévy processes and infinitely divisible distributions," Cambridge University Press, 1999.

[12] Y. Wang and L. Carin, "Lévy Measure Decompositions for the Beta and Gamma Processes," in *Proc. of ICML*, 2012.

[13] J. F. C. Kingman, "Completely random measure," *Pacific Journal of Mathematics*, vol. 21(1): pp. 59-78, 1967.

[14] M. I. Jordan, "Hierarchical models, nested models and completely random measures," *Frontiers of Sta- tistical Decision Making and Bayesian Analysis*: In Honor of James O. Berger. New York: Springer, 2009.

[15] C. Févotte, N. Bertin and J. L. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis," *Neural Computation*, 21 (3): pp. 793-830, 2009.

[16] M. Hoffman, D. Blei and P. Cook, "Bayesian nonparametric matrix factorization for recorded music" in *Proc. ICML*, pp. 641-648, 2010.

[17] S. Williamson, C. Wang, K. Heller and D. Blei, "The IBP compound Dirichlet process and its application to focused topic modeling," in *Proc. ICML*, 2010.

[18] T. Stepleton, Z. Ghahramani, G. Gordon and T. S. Lee, "The block diagonal infinite hidden Markov model," in *Proc. of the International Conference on Artificial Intelligence and Statistics*, 2009.

[19] J. Van Gael, Y. Saatci, Y. W. Teh and Z. Ghahramani, "Beam sampling for the infinite hidden Markov model," in *Proc. of the International Conference on Machine Learning*, 2008.

[20] R. Thibaux, and M. I. Jordan, "Hierarchical beta processes and the indian buffet process," in *Proc. of International Conference on Artificial Intelligence and Statistics*, 2007.

[21] K. A. Heller, Y. W. Teh and D. Görür, "Infinite hierarchical hidden Markov models," in *Proc. of the International Conference on Artificial Intelligence and Statistics*, 2009.

[22] J. Van Gael, Y. W. Teh and Z. Ghahramani, "The infinite factorial hidden Markov model," in *Advances in Neural Information Processing Systems*, 2009.

[23] D. Wingate, N. D. Goodman, D. M. Roy, D and J. B. Tenenbaum, "The infinite latent events model," in *Proc. of the International Conference on Uncertainty in Artificial Intelligence*, 2009.

[24] F. Doshi-Velez, D. Wingate, N. Roy and J. Tenenbaum, "Infinite dynamic Bayesian networks," in *Proc. of International Conference in Machine Learning*, 2011.

[25] H. Ishwaran and M. Zarepour, "Exact and approximate sum-representations for the Dirichlet process," *Canadian Journal of Statistics*, 30, 269-283, 2002.

[26] Y. W. Teh, M. I. Jordan, M. Beal and D. Blei, "Hierarchical Dirichlet processes," *Journal of the American Statistical Association*, 101, 1566-1581, 2006.

[27] M. Beal, Z. Ghahramani and C. Rasmussen, "The infinite hidden Markov model," in *Advances in Neural Information Processing Systems*, 2002.

[28] Y. W. Teh, D. Görür and Z. Ghahramani, "Stick-breaking construction for the Indian buffet process," in *Proc. of the International Conference on Artificial Intelligence and Statistics*, vol. 11, 2007.

[29] D. M. Blei, A. Y. Ng and M. I. Jordan, "Latent Dirichlet allocation," *Journal of Machine Learning Research*, 3:993-1022, 2003.

[30] P. Liang, S. Petrov, M. I. Jordan, and D. Klein, " The infinite PCFG using hierarchical Dirichlet processes, " in *Proc. of EMNLP*, pp. 688-697, 2007.

[31] H. Kameoka, K. Ochiai, M. Nakano, M. Tsuchiya, S. Sagayama, "Context-free 2D tree structure model of musical notes for Bayesian modeling of polyphonic spectrograms," in *Proc. of ISMIR*, 2012.

[32] M. Nakano, Y. Ohishi, H. Kameoka, R. Mukai, K. Kashino, "Bayesian nonparametric music parser," in *Proc. of ICASSP*, 2011.

[33] D. Mochihashi and E. Sumita, "The infinite Markov model," in *Advances in Neural Information Processing Systems*, 2007.

[34] F. Wood, C. Archambeau, J. Gasthaus, L. F. James and Y.W. Teh, "A Stochastic Memoizer for Sequence Data," in *Proc. of ICML*, 2009.

[35] A. Rodriguez, D. B. Dunson and A. E. Gelfand, "The nested Dirichlet processThe nested Dirichlet process," *Jornal of American Statistics Associaion* 103, 1131-1154, 2008.

[36] A. Rodriguez and K. Ghosh, "Nested partition models," *Jack Baskin School of Engineering*, Technical report, 2009.

[37] E. B. Fox, E. B. Sudderth, M.I. Jordan, A.S. Willsky, "A Sticky HDP-HMM with Application to Speaker Diarization," *Annals of Applied Statistics*, 2011.

[38] A. Kulesza and B. Taskar, "Structured determinantal point processes," in *Proc. of NIPS*, 2010.

[39] R. H. Affandi, A. Kulesza and E. B. Fox, "Markov determinantal point process,", in *Proc. of UAI*, 2012.

[40] S. Williamson, P. Orbanz, and Z. Ghahramani, "Dependent Indian buffet processes," in *Proc. AISTATS*, 2010.

[41] M. Zhou, H. Yang, G. Sapiro, D. Dunson, and L. Carin, "Dependent hierarchical beta process for image interpolation and denoising," in *Proc. AISTATS*, 2011.

[42] Lu Ren, Y. Wang, D. Dunson, and L. Carin, "The Kernel Beta process," in *Proc. NIPS*, 2011.

人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-B202-7

無限混合ガウスモデルを用いた未知クラスに対応可能な実環境音分類法

Nested Infinite Gaussian Mixture Model for Environmental Audio Signal Recognition

○ 佐々木洋子 ⋆, 吉井和佳 †, 加賀美聡 ⋆

Yoko SASAKI, Kazuyoshi YOSHII and Satoshi KAGAMI

産業技術総合研究所 ⋆ デジタルヒューマン工学研究センター/† 情報技術研究部門

Digital Human Research Center/Information Technology Research Institute, AIST

{y-sasaki, k.yoshii, s.kagami}@aist.go.jp

Abstract

The paper proposes a nonparametric Bayesian audio signal modeling based on nested infinite Gaussian mixture model. It can describe a variety of sound sources for recognizing known and unknown sounds in surrounding environment. So far, various methods of audio signal recognition have been proposed for classifying given audio events into a fixed number of categories that are manually defined in advance. Therefore, audio events of unseen classes are forced to be classified into the known classes although those events have distinct acoustic features. To solve this problem, our model can increase the number of classes unboundedly, to represent the given audio signals. Experimental results showed the effectiveness of the proposed model.

1 はじめに

実環境中の様々な音の中で何の音がしたかを理解する技術は，ロボット聴覚分野において主要な課題の一つである．聴覚は視覚に比べ情報量は少ないがより広い範囲に伝播するため，呼びかけに応える，物音に反応して振り返る，など環境変化の初期知覚として有用である．一方で音を認識する技術として，人の声を対象とした音声認識や楽曲中の楽器音推定などが広く発展している．こうした特定種類の音に限らず様々な音が混在する実環境の多様な音響信号を扱うためには，これらの認識技術の前段処理としても，まず何の音なのか理解する技術が役立つ．一例として，一連の音ストリームから人の声，動物の鳴き声，物音などを書き起こす音響イベント検出も盛んに研究が行われている [1] [2] [3].

本研究では，マイクロホンで収録された音響信号が何の音なのか理解するための音源の識別手法について扱う．一般の音を対象とした認識では，既存の音声認識手法を利用したもの [4] や，各種周波数特徴量の GMM [5] [6] など，あらかじめ設計されたモデルの学習が主流である．一方で観測データから直接モデルを設計するアプローチも提案されており [7]，大規模データからベクトル量子化によるデータのランク付け，自動分類が行われている．いずれも正解付のデータに基づく教師あり学習となっている．

様々な条件が想定される実環境に対応するためには，音源の種類数や各音源を表すモデルの次元数など，事前知識は最少であることが望ましい．本稿では，実環境中の多様な音を表現し，未知の音を検出可能なモデルの生成を目指し，ノンパラメトリックベイズ学習 [8] に基づくネスト型無限混合ガウスモデルを提案する．

2 ネスト型混合モデルによる環境音分類法

本章では，実環境中の多様な音を認識することをめざし，複雑さの異なる様々な音を一度に学習可能なモデル生成法を提案する．

2.1 特徴量の抽出

本研究では音響信号の振幅スペクトルを基にした，フレーム単位の特徴量による音源の識別を対象とする．フレーム単位の特徴量を用いることで，時系列情報を扱うことはできない．一方で各フレームの識別結果は，後段のセグメンテーションや移動音源のトラッキングに利用可能である．

振幅スペクトルの局所的な特徴量として，12 次元 MFCC (Mel-Frequency Cepstrum Coefficient), ΔMFCC, 対数エネルギー E, ΔE, ゼロクロス，フラックス, セントロイド, 分散, エントロピー, 歪度 (skewness), 尖度 (kurtosis) の計 33 次元ベクトルを用いる．

2.2 既知の種類の音源への識別

まず音の種類として K 個のクラスが事前に定義されているとして，与えられた特徴量 $\boldsymbol{x}$ がどのクラスに属するかを予測する教師ありクラス分類問題について考える．一般的には学習データ中に含まれる各クラス k $(1 \leq k \leq K)$ の特徴量の分布を，混合ガウスモデル (GMM) $\mathcal{M}_k$ を用いてあらかじめ独立に学習し，与えられた特徴量 $\boldsymbol{x}$ に対する尤度 $\mathcal{M}_k(\boldsymbol{x})$ を計算することで，最も尤度の高いクラスに分類することが行われる．ここで各クラスに対応するモデル $\mathcal{M}_k$ の混合数（ガウス分布の個数）は M であるとしておく．

本研究では各クラスに対応するモデル $\mathcal{M}_k$ をさらに混合したモデル $\mathcal{M}$ を考え，学習データを一挙に与えて**一度に学習**することを提案する．すなわち本モデルは，各クラスの特徴量の分布を M 混合 GMM として表現し，それら K 個をさらに混合したものとして学習する．これは K 混合 GMM の各要素分布が M 混合 GMM となっているものであると言ってもよい．こうすることで各クラス k の出現率（混合比）を加味した分類ができると期待される．

具体的にはあるクラス k $(1 \leq k \leq K)$ の特徴量 $\boldsymbol{x}$ の分布を混合数 M の有限混合ガウス分布

$$\mathcal{M}_k(\boldsymbol{x}) = \sum_{m=1}^{M} \tau_{km} \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_{km}, \boldsymbol{\Lambda}_{km}^{-1}) \tag{1}$$

で表現する．ここでパラメータ $\boldsymbol{\mu}_{km}$ および $\boldsymbol{\Lambda}_{km}$ は，多次元ガウス分布の平均ベクトル，精度行列であり，τ_{km} は足して 1 になるように正規化された各ガウス分布の相対強度（混合比）を表す．さらに K 個のクラスにわたる特徴量の分布を表現するため，$\mathcal{M}_k(\boldsymbol{x})$ をさらに混合することでネスト型 GMM

$$\mathcal{M}(\boldsymbol{x}) = \sum_{k=1}^{K} \boldsymbol{\pi}_k \mathcal{M}_k(\boldsymbol{x}) \tag{2}$$

を得る．すなわち本モデルは，KM 個のガウス分布からなる混合分布として得られる．特徴量 $\boldsymbol{x}$ は，KM 個中のいずれかのガウス分布から生成されることになる．学習データを用いてパラメータ $\boldsymbol{\pi}, \boldsymbol{\tau}, \boldsymbol{\mu}, \boldsymbol{\Lambda}$ を求めることができれば，新たに与えられた特徴量 $\boldsymbol{x}$ がどのクラスから生成されたものであるかの事後分布が計算できるようになる．

2.3 ネスト型混合ガウスモデル

観測データ $\boldsymbol{X}$ の生成過程を表現するネスト型混合ガウスモデルを，ベイズモデルとして定式化する．ベイズモデルでは各パラメータに対して事前分布を導入することで，通常の最尤推定に比べて過学習しにくく，汎化能力の高いモデルを学習が可能である．

いま，学習データに含まれる特徴量は全体で N サンプル（N フレーム）あるとして，観測変数全体を $\boldsymbol{X} = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N\}$ で表す．同様に $\boldsymbol{X}$ に対する潜在変数を $\boldsymbol{Z} = \{\boldsymbol{z}_1, \cdots, \boldsymbol{z}_N\}$ とする．ここで $\boldsymbol{z}_n$ $(1 \leq n \leq N)$ は KM 次元のベクトルであり，クラス k に対応するモデル $\mathcal{M}_k$ を構成する m 番目のガウス分布から $\boldsymbol{x}_n$ が生成された場合に，$z_{km} = 1$ となり，それ以外の要素はゼロ（$z_{k'm'} = 0$ if $k' \neq k, m' \neq m$）となる．

まず完全な同時分布は次式で与えられる．

$$\begin{aligned} & p(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\pi}, \boldsymbol{\tau}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) \\ &= p(\boldsymbol{X} \mid \boldsymbol{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) p(\boldsymbol{Z} \mid \boldsymbol{\pi}, \boldsymbol{\tau}) p(\boldsymbol{\pi}) p(\boldsymbol{\tau}) p(\boldsymbol{\mu}, \boldsymbol{\Lambda}) \end{aligned} \tag{3}$$

ここで右辺の前二項はパラメータが与えられたときの観測変数 X および潜在変数 Z の尤度であり，後ろの三項はパラメータの事前分布である．尤度項はそれぞれ，

$$p(\boldsymbol{X} \mid \boldsymbol{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = \prod_{nkm} \mathcal{N}(\boldsymbol{x}_n \mid \boldsymbol{\mu}_{km}, \boldsymbol{\Lambda}_{km}^{-1})^{z_{nkm}} \tag{4}$$

$$p(\boldsymbol{Z} \mid \boldsymbol{\pi}, \boldsymbol{\tau}) = \prod_{nkm} (\boldsymbol{\pi}_k \boldsymbol{\tau}_{km})^{z_{nkm}} \tag{5}$$

で与えられる．また事前分布は共役事前分布を考える．

$$p(\boldsymbol{\pi}) = \mathrm{Dir}(\boldsymbol{\pi} \mid \alpha \boldsymbol{\nu}) \propto \prod_{\mathrm{k}} \boldsymbol{\pi}_{\mathrm{k}}^{\alpha \nu_{\mathrm{k}} - 1} \tag{6}$$

$$p(\boldsymbol{\tau}) = \prod_{k} \mathrm{Dir}(\boldsymbol{\tau}_{\mathrm{k}} \mid \beta \boldsymbol{\upsilon}) \propto \prod_{\mathrm{k,m}} \boldsymbol{\tau}_{\mathrm{km}}^{\beta \upsilon_{\mathrm{m}} - 1} \tag{7}$$

$$p(\boldsymbol{\mu}, \boldsymbol{\Lambda}) = \prod_{k,m} \mathcal{N}(\boldsymbol{\mu}_{km} \mid m_0, (b_0 \boldsymbol{\Lambda}_{km})^{-1}) \mathcal{W}(\boldsymbol{\Lambda}_{km} \mid \boldsymbol{W}_0, c_0) \tag{8}$$

ここで $p(\boldsymbol{\pi})$ は K 次元ディリクレ分布，$p(\boldsymbol{\tau})$ は M 次元のディリクレ分布の積である．また $p(\boldsymbol{\mu}, \boldsymbol{\Lambda})$ はガウス・ウィシャート分布の積である．超パラメータに関しては，α および β は集中度と呼ばれる正の実数であり，ν および υ はそれぞれ K 次元ベクトルおよび M 次元ベクトルであり，いずれも足して 1 になるよう正規化されている．m_0 および b_0 はガウス分布の平均および精度のスケール，W_0 および c_0 はウィシャート分布のスケール行列および自由度である．本研究では，事前分布ができるだけ無情報になるように超パラメータの値を設定した．

2.4 ベイズモデルの学習

ここでの我々の目的は，観測データ $\boldsymbol{X}$ が与えられたもとでの潜在変数およびパラメータの事後分布 $p(\boldsymbol{Z}, \boldsymbol{\pi}, \boldsymbol{\tau}, \boldsymbol{\mu}, \boldsymbol{\Lambda} | \boldsymbol{X})$ を求めることである．しかし真の事後分布は解析的には求めることはできないため，本研究では変分ベイズ法 (VB) を用いて事後分布を近似的に求めることにする．VB の計算量は，GMM の最尤推定に通常用いられる EM アルゴリズムと同程度であり，非常に効率的である．まず事後分布における潜在変数とパラメータとの独立性を仮定し，因子分解された形の変分事後分布

$$q(\boldsymbol{Z}, \boldsymbol{\pi}, \boldsymbol{\tau}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = q(\boldsymbol{Z}) q(\boldsymbol{\pi}, \boldsymbol{\tau}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) \tag{9}$$

を仮定する．次に $p(\boldsymbol{Z}, \boldsymbol{\pi}, \boldsymbol{\tau}, \boldsymbol{\mu}, \boldsymbol{\Lambda}|\boldsymbol{X})$ の $q(\boldsymbol{Z}, \boldsymbol{\pi}, \boldsymbol{\tau}, \boldsymbol{\mu}, \boldsymbol{\Lambda})$ に対するカルバック・ライブラー (KL) ダイバージェンスが最小化するように，$q(\boldsymbol{Z}, \boldsymbol{\pi}, \boldsymbol{\tau}, \boldsymbol{\mu}, \boldsymbol{\Lambda}|\boldsymbol{X})$ を反復最適化を行えばよい．このとき各ステップでの最適な変分事後分布は，期待値を $\mathbb{E}$ として

$$q(\boldsymbol{Z}) \propto \exp(\mathbb{E}_{q(\boldsymbol{\pi},\boldsymbol{\tau},\boldsymbol{\mu},\boldsymbol{\Lambda})}[\log p(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\pi}, \boldsymbol{\tau}, \boldsymbol{\mu}, \boldsymbol{\Lambda})]) \quad (10)$$

$$q(\boldsymbol{\pi}, \boldsymbol{\tau}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) \propto \exp(\mathbb{E}_{q(\boldsymbol{Z})}[\log p(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\pi}, \boldsymbol{\tau}, \boldsymbol{\mu}, \boldsymbol{\Lambda})]) \quad (11)$$

で与えられる．紙面の都合上，詳細な更新式は省略する．

2.5 モデルの生成と識別

モデル生成の流れを Figure 1 に図示する．まず (1) データの一部分に正解ラベルのついた音響信号を用意する．これに対し (2) 各フレームで求めた特徴量ベクトルを学習データ X として，(3) 一部にのみ正解ラベルが付与された一連の学習データに対し，ラベルが与えられていない部分のクラスを推定しながらモデルの学習を行う．

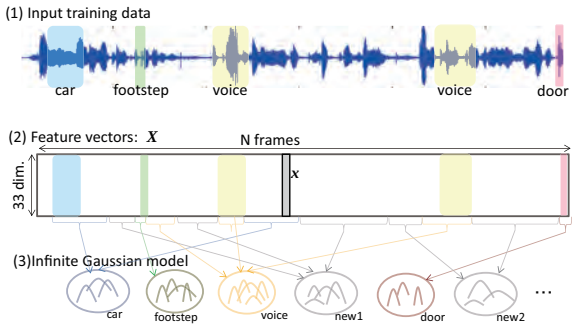


Figure 1: Semi-supervised training of a nested GMM

生成したモデルによる識別では，入力の特徴量ベクトル $\boldsymbol{x}$ がどのガウス分布から生成されたか，KM 次元の離散分布を事後分布として計算する．本研究では，各クラスの次元数 m について足し合わせ，K 次元の離散分布として出力する．

提案手法は一部のラベル付きデータを利用した**半教師あり学習**である．次章では，本章で説明した音源のクラス数 K および各クラスの混合数 M を無限化し，事前にモデルの複雑さを設定する必要のない手法を提案する．観測データのうち正解ラベルのない部分のクラスを無限混合の状態で推定しながら学習することで，観測データに合わせた適切なモデルを生成可能となる．

3 未知クラスを含む音源の分類

特徴量の分布を表現するうえで適切な GMM の混合数は，音源のクラスごとに異なる．たとえば，換気扇など単調な音は少数のガウス分布で表現できるであろう．一方，人の話し声は様々な音素から構成されており，多数のガウス分布が必要であると考えられる．そのため各クラスを表現するためのモデルの混合数は，特徴量分布の複雑さに合わせて自動的に調節可能であることが望ましい．また既知の音源の特徴量分布からは非常に発生しにくいような特徴量をもつ音が発生したら，それを既知のクラスのいずれかに当てはめてしまうのではなく，未知の音として新たにクラスを生成する仕組みも重要である．

これらの問題に対処するため，前章で説明したネスト型有限混合ガウスモデルをノンパラメトリックベイズ理論を用いて無限モデルへ拡張することを提案する．「ノンパラメトリック」とは確率モデルの複雑さを表すパラメータ空間の次元が固定されておらず，無限の複雑さを考えることを意味する．もし観測データが無限であれば，その生成過程を表現するため無限個のパラメータが必要となる．ただし実際には観測データは有限であり，無限個のパラメータのうち一部を使うだけで十分である．

無限混合ガウスモデルでは，本来無限個存在するガウス分布のうち，与えられた観測データを表現するのに必要なガウス分布の個数が推定できる．無限個の異なる混合数のモデルが確率的に重なり合っており，混合数を一意に決定せずに学習や予測ができるため，モデル選択の問題が生じない．以降で提案するネスト型無限混合モデルは，ノンパラメトリックベイズ学習に基づいており，音源のクラス数 K や各クラスにおける混合数 M を事前に指定する必要がない．そのため各クラスの特徴量の複雑さに合わせた GMM を学習できるだけはなく，これまで学習していない未知の音が発生した際に，既知のクラスではない新たなクラスであると識別可能である．

3.1 混合数，クラス数の無限化

式 (2) を拡張し，音源識別のためのモデルを以下のようにネスト型無限混合ガウスモデルで表現する．

$$\mathcal{M}(\boldsymbol{x}) = \sum_{k=1}^{\infty} \boldsymbol{\pi}_k \sum_{m=1}^{\infty} \boldsymbol{\tau}_{km} \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_{km}, \boldsymbol{\Lambda}_{km}^{-1}) \quad (12)$$

まずクラス数 K を無限大にする場合を考える．つまり，式 (6) で各クラス k ごとに無限次元のディリクレ分布を考える．このような分布からサンプルされる混合比 $\boldsymbol{\pi}_k$ は各基底を選ぶ確率を要素として並べた無限次元のベクトルとなる．$\boldsymbol{\pi}_k$ は無限次元の離散分布であるため，無限個の要素の和が 1 となるよう正規化されている．実際には，ごく一部の要素のみが意味のある値をとり，残りの無限個の要素はほぼゼロに等しい．このような確率過程をディリクレ過程 (Dirichlet Process, DP) と呼ぶ．

集中度を α およびガウス・ウィシャート分布を基底測度 G_0 としたディリクレ過程 $\mathrm{DP}(\alpha, \mathrm{G}_0)$ を考える．可算無限個のガウス分布 G は $G \sim \mathrm{DP}(\alpha, \mathrm{G}_0)$ にしたがって生成される．ここで，G_0 は G の期待値となっている．G_0 からサンプルされた G（具体的にはパラメータ μ および Λ）の分布は，α が大きいほど G_0 に近くなる．したがって，α は逆分散のように振る舞う．ディリクレ過程の一つ

の実現方法として，ここでは 棒折り過程 (Stick-Breaking Construction, SBC) を用いる．SBC は変分ベイズ法を適用するうえで都合が良い DP の表現方法である．このとき，混合係数 $\boldsymbol{\pi}_k$ は次式で表現できる．

$$\boldsymbol{\pi}_k = \boldsymbol{\upsilon}_k \prod_{k'=1}^{k-1} (1 - \boldsymbol{\upsilon}_{k'}) \tag{13}$$

$$\boldsymbol{\upsilon}_k \sim \mathrm{Beta}(1, \alpha) \tag{14}$$

K と同様に各クラスの混合数 M を無限大にする場合を考える．下位の $\boldsymbol{\tau}_{km}$ について $\boldsymbol{\pi}_k$ と同様に変数変換し，次式で表現できる．

$$\boldsymbol{\tau}_{km} = \boldsymbol{\upsilon}_{km} \prod_{m'=1}^{m-1} (1 - \boldsymbol{\upsilon}_{km'}) \tag{15}$$

$$\boldsymbol{\upsilon}_{km} \sim \mathrm{Beta}(1, \alpha_{\mathrm{k}}) \tag{16}$$

ここで式 (14), 式 (16) の集中度 α, α_k について考える．ディリクレ過程 $\mathrm{DP}(\alpha, \mathrm{G}_0)$ における集中度 $\alpha > 0$ は無限混合ガウス分布のハイパーパラメータであり，観測データを生成するのに実際に利用されたガウス分布の個数（混合数）に大きく影響する．適切な値は自明ではないので，α の事前分布 $p(\alpha)$ としてガンマ分布をおく．

$$p(\alpha) \propto \mathrm{Gam}(\alpha \mid \mathrm{a}, \lambda), \quad \mathrm{p}(\alpha_{\mathrm{k}}) \propto \mathrm{Gam}(\alpha_{\mathrm{k}} \mid \mathrm{a}, \lambda) \tag{17}$$

このとき，ハイパーパラメータ a, λ に対しては無情報事前分布をおき，特に事前知識がないことを自然に表現する．

3.2 変分事後分布

2.4 節と同様に，観測データ $\boldsymbol{X}$ が与えられたときの事後分布 $q(\boldsymbol{Z}, \boldsymbol{\upsilon}, \boldsymbol{\mu}, \boldsymbol{\Lambda} \mid \boldsymbol{X})$ を求めることを考える．これを解析的に行うことは困難なので変分事後分布 $q(\boldsymbol{Z}, \boldsymbol{\upsilon}, \boldsymbol{\mu}, \boldsymbol{\Lambda})$ を導入し，真の事後分布に近づくよう最適化を行う．

事後分布において潜在変数とパラメータの独立性を仮定し，以下の因子分解を考える．

$$q(\boldsymbol{Z}, \boldsymbol{\upsilon}, \boldsymbol{\mu}, \boldsymbol{\Lambda}, \boldsymbol{\alpha}) = q(\boldsymbol{Z}) q(\boldsymbol{\pi}, \boldsymbol{\upsilon}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) q(\boldsymbol{\alpha}) \tag{18}$$

ここで $\boldsymbol{\alpha}$ は (α, α_k) で表される上位 GMM，下位 GMM の集中度である．右辺の三項についてそれぞれ VB を用いて反復最適化を行えばよい．紙面の都合上詳細な更新式は省略するが，VB 各ステップでの変分事後分布は，

$$q(\boldsymbol{Z}) \propto \exp \mathbb{E}_{\boldsymbol{\upsilon}, \boldsymbol{\mu}, \boldsymbol{\Lambda}, \boldsymbol{\alpha}} [\log p(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\upsilon}, \boldsymbol{\mu}, \boldsymbol{\Lambda}, \boldsymbol{\alpha})] \tag{19}$$

$$q(\boldsymbol{\upsilon}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) \propto \exp \mathbb{E}_{\boldsymbol{z}, \alpha} [\log p(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\upsilon}, \boldsymbol{\mu}, \boldsymbol{\Lambda}, \boldsymbol{\alpha})] \tag{20}$$

$$q(\boldsymbol{\alpha}) \propto \exp \mathbb{E}_{\boldsymbol{\upsilon}, \boldsymbol{\mu}, \boldsymbol{\Lambda}, \boldsymbol{\alpha}} [\log p(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\upsilon}, \boldsymbol{\mu}, \boldsymbol{\Lambda}, \boldsymbol{\alpha})] \tag{21}$$

で与えられる．

4 環境音の分類実験

提案法による環境音の分類例として，自転車で走行しながら周囲の音を IC レコーダで録音し，環境音のモデル生成を行った．収録した音の説明および分類結果を Figure 2 に示す．上段が録音データの時間波形および付与した正解ラベル，下段が分類結果となっている．また時間波形の上部に各部分の主な音源の説明をつけた．

4.1 実験条件

実験用の音源として，自転車走行中にレコーダ (Roland R-09) で録音した，9 分 9 秒のデータを用いた．Figure 2 上部の説明の通り，坂道を下り，静かな通りを走行後，交通量の多い大通りを通り，再び静かな通りを走行した．データは 16bit, 44kHz で録音したものを 16kHz にダウンサンプルした．特徴量はフレーム長 100ms, シフト長 20ms で計算し，計 27490 フレームとなっている．

また Figure 2 上段の時間波形に色つきで示した 3 種類 4 か所のデータに正解ラベルを付与した．ラベル付きの部分は計 4314 フレームあり，残りの 23176 フレーム (白い部分) は未知の音として，モデル生成を行った．

4.2 分類結果

Figure 2 下段に，生成したモデルで求めた各クラスの確率分布を示す．ここではフレームごとに各クラスの混合数 m について和をとった K 次元離散分布となっている．

結果は正解ラベルを付与した 3 クラスに加え，新たに 3 クラス，計 6 クラスに分類された．新クラスのひとつめ (new1) には，大通りに出る手前で遠くから聞こえる車の走行音が主に分類された．二番目 (new2) に分類された 5 分 40 秒あたりと 7 分すぎの部分は，車の走行音がなく，アスファルト上を走る自転車のロードノイズが主な音源であった．三番目 (new3) には，すれ違う人の話し声 (女性) や，自転車が段差を越える際の金属音といった比較的高い音が含まれた．正解ラベルのない新しい音について，ほぼ適切に分類できているといえる．

全体では，冒頭と最後の静かな通りでは，主に自転車の走行音 (bicycle) と風切音 (wind) に分類され，中間の大通りでは主に車の走行音 (car) に分類された．一部に正解を付与した 3 クラスについて，正解ラベルのない部分も適切に推定できているといえる．ただし本稿で提案する混合モデルは，複数のクラスを同時にアクティベートできないため，混合音としてクラスを生成するか，分離された音源に適用するなどの工夫が必要である．

5 識別性能の評価実験

提案する無限混合モデルの識別性能を検証するため，分類・識別実験を行った．ここでは学習データの条件が異なる数種類のモデルを生成し，既知の音，未知の音に対する識別正解率を求め，結果を考察する．

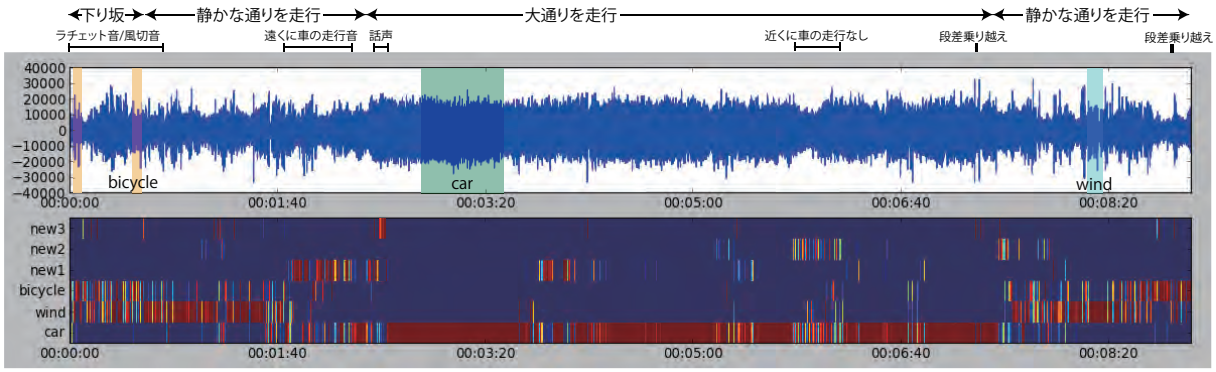


Figure 2: An example of environmental sound modeling

5.1 実験条件

評価のための音源として，7 種類のパーカッション楽器 (ハンドベル，ギロ，木魚，シェイカー，タンバリン，トライアングル，ウッドブロック) および，拍手，話し声の 9 種類を用いた．それぞれ 13 分ずつ録音し，10 分を学習データとしてモデル生成に用い，残り 3 分をテストデータとして音源識別の評価に用いた．

各音の収録にはロボットに搭載した 32ch マイクロホンアレイを用いて，16bit, 16kHz サンプリングで行った．特徴量計算に用いる振幅スペクトルは，フレーム長 256ms, シフト長 128ms, Hamming 窓の短時間フーリエ変換により求めた．

Figure 3 に，各音源の類似度を示す．尺度にはコサイン類似度を用いた．値が大きいほど類似していることを示している．たとえば，ひとつめのハンドベルはマラカス，シェイカー，タンバリンに比較的近い特徴量を持ち，最後から 2 番目の話し声は残り 9 種のどの音源にもあまり似ていないことがわかる．

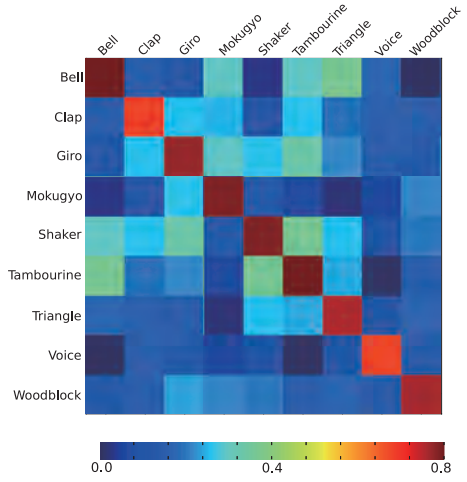


Figure 3: Cosine Similarity of experiment sounds

5.2 楽器音の識別

以下の 2 条件についてそれぞれ正解ラベルを含む割合を変化させて半教師あり学習を行った．

expA 正解ラベルを持つデータで学習した場合

expB 未知音として学習データに含まれる場合

expA では，9 種類のすべての音源について正解ラベル付きのデータでモデルを学習し．既知の音源に対する識別正解率を評価する．**expB** では，**expA** と同じ学習データセットに対し，特定の 1 クラスについて全く正解ラベルがない状態で学習し，このクラスのテストデータが新たなクラスとして学習されるかどうかを評価する．

また **expA**, **expB** それぞれについて学習データの音源ごとにそれぞれ一定の割合 (0, 30, 50, 70%) で正解ラベルをマスクし，正解ラベルを含む割合を変化させた各条件での識別率を比較した．

まず **expA** および **expB** で生成したモデルによる，テストデータの識別正解率を Figure 4 に示す．ここでは K 次元の確率分布に対し，最大尤度のクラスに識別されたとして正解率を求めた．また **expB** については，新しいクラスと識別された場合に正解とした．

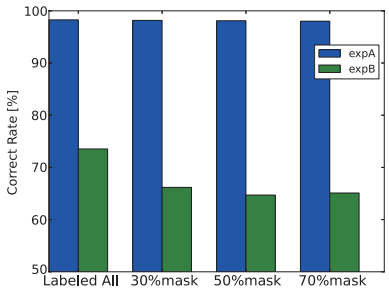


Figure 4: Correct recognition rate of **expA, expB**

expA では正解ラベルを持つデータの割合に関わらず高い正解率が得られており，各クラスの正解ラベルを 70% マスクした場合でも正解率 98.0% となった．**expB** については，既知のクラスが完全に正解ラベルを持つ場合に正解率 73.5%，既知クラスのラベルを一部マスクした場合は 65% 前後の正解率となった．

expB で既知クラスが全て正解ラベル付きだった場合について，テストデータ各 1 分の事後分布の平均値を Figure

5,6 に図示する．ハンドベルを正解ラベルなしの観測データとして学習した結果である Figure 5 では上 3 行のハンドベル (Bell) が高確率で新クラス (new) となっており，ほぼ確実に新クラスと推定されているといえる．また既知のクラスであるその他のテストデータに対しては，確率の高い部分が横軸のクラス名と一致しており，正しく識別されていることがわかる．

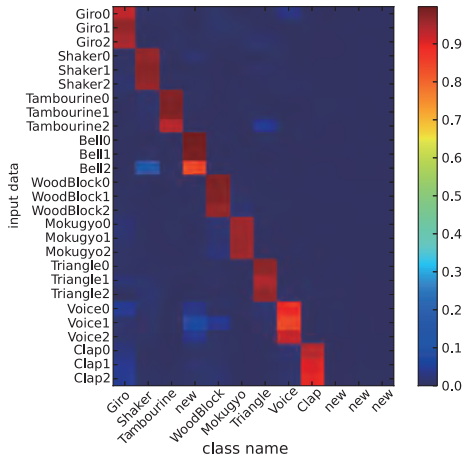


Figure 5: Posterior for unlabeled Bell model

一方，ウッドブロックを正解ラベルなしの観測データとして学習した結果である Figure 6 では，下 3 行のウッドブロック (Woodblock) が新クラス (new) に加え木魚クラス (Mokugyo) にも出現しており，既知クラスとして似たような音がある場合そちらにも分類されていることがわかる．その他の既知クラスについては Figure 5 と同様に正しく識別されている．

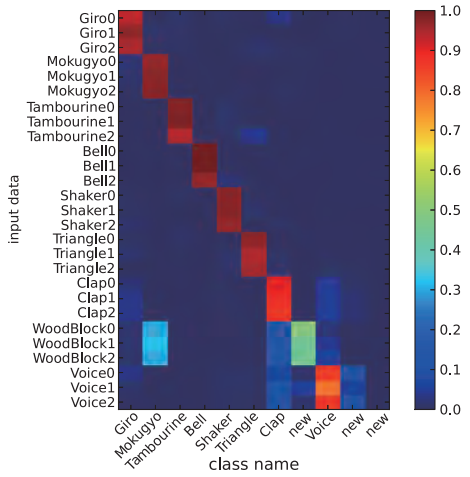


Figure 6: Posterior for unlabeled Woodblock model

6 結言

本稿では，知らない音を含む実環境中の様々な音源を認識することをめざし，観測データに合わせて自動生成可能な音響信号のモデル化方法を提案した．提案法は，従来のように対象の音源ごとに独立にモデルを学習させるのではなく，複雑さの異なる様々な音源を含む学習データを一度に学習させることがひとつめの特徴である．さらにノンパラメトリックベイズに基づき無限混合ガウスモデルをベイズ推定することで，本来未知であるはずの音の種類数や各音を表現するモデルの次元数を事前に決定することなく，観測データに合わせて柔軟なモデル生成が可能である．

実験では，既知の音に対する識別では正解ラベルが付与された割合によらず高い正解率が得られた．また正解ラベルを与えていない未知音に対しては，既知の似た音に分類されることもある一方で，既知の特徴量分布とは離れた音響イベントを新クラスと識別可能であることを確認した．

一方提案した無限混合モデルは，複数のクラスを同時にアクティベートできないため，実環境下での混合音の扱いには工夫が必要である．マイクロホンアレイによる音源定位・分離やロボットの移動を含めたロボット聴覚システムの一部として本手法を組み込むことで，時間・空間的に分離された音源の識別方法としての効果が期待される．また識別結果を利用した時系列方向の音のトラッキングや，他センサとの情報統合など，自律型ロボットシステム全体として提案法を活用することが今後の課題である．

参考文献

[1] Andrey Temko and Climent Nadeu. Acoustic event detection in meeting-room environments. *Pattern Recognition Letters*, Vol. 30, No. 14, pp. 1281–1288, 2009.

[2] Taras Butko and Climent Nadeu. Audio segmentation of broadcast news: A hierarchical system with feature selection for the albayzin-2010 evaluation. In *Proceedings of ICASSP*, pp. 357–360, 2011.

[3] Richard F. Lyon, Martin Rehn, Samy Bengio, Thomas C. Walters, and Gal Chechik. Sound retrieval and ranking using sparse auditory representations. *Neural Computation*, Vol. 22, No. 9, pp. 2390–2416, 2010.

[4] Ramasubramanian V., Karthik R., Thiyagarajan S., and Cherla S. Continuous audio analytics by hmm and viterbi decoding. In *Proceedings of ICASSP*, pp. 2396–2399, May 2011.

[5] G. Valenzise, L. Gerosa, M. Tagliasacchi, F. Antonacci, and A. Sarti. Scream and gunshot detection and localization for audio-surveillance systems. In *Proceedings of Advanced Video and Signal Based Surveillance*, pp. 21–26. IEEE, September 2007.

[6] A. Fleury, N. Noury, M. Vacher, H. Glasson, and J.-F. Serignat. Sound and speech detection and classification in a health smart home. In *EMBS*, pp. 4644–4647. IEEE, August 2008.

[7] Richard F. Lyon, Martin Rehn, Samy Bengio, Thomas C. Walters, and Gal Chechik. Sound retrieval and ranking using sparse auditory representations. *Neural Computation*, Vol. 22, No. 9, pp. 2396–2416, August 2010.

[8] T. S. Ferguson. A bayesian analysis of some nonparametric problems. *Ann. Statist.*, Vol. 1, pp. 209–230, 1973.

アクティブ視聴覚統合による発話区間検出の検討: 因果モデルベースアプローチ

Active Audio-Visual Integration for Voice Activity Detection: a Causal-Model-based Approach

吉田尚水 [1]，中臺一博 [1,2]

Takami YOSHIDA[1], Kazuhiro NAKADAI[1,2]

1. 東京工業大学大学院，2. (株) ホンダ・リサーチ・インスティチュート・ジャパン

1. Tokyo Institute of Technology, 2. Honda Research Institute Japan Co., Ltd.

yoshida@cyb.mei.titech.ac.jp, nakadai@jp.honda-ri.com

Abstract

This paper presents a framework for Active Audio-Visual (AAV) integration which integrates audio, visual and motion information to improve robot's perception, and its application to Voice Activity Detection (VAD) to show the effectiveness of the proposed framework. For the AAV framework, we propose to use a Causal Bayesian Network (CBN) to make a robot predict an optimal active motion in the current situation. We implemented a prototype system based on the proposed AAV integration framework for a humanoid robot and experimental results showed that the proposed system successfully estimated the optimal paths to improve VAD in different conditions.

1 はじめに

人が生活するような日常環境でロボットが音環境を理解するためには，能動的に動作を利用するアクティブ・オーディションが重要である．日常環境では雑音の性質など環境の情報が事前に得られるとは限らないため，周囲の環境をマイクやカメラなどのセンサを使って調べ，その測定に基づき最適な行動を行うことが求められる．

アクティブ・オーディションの従来研究は，大きく以下の 3 種類に分類することができる．

- 音源定位の性能を向上させるため，マイクアレイの最適な姿勢を推定し制御する
- 二次元平面上での音源定位を行うため，与えた軌跡に沿ってマイクアレイを移動させる
- 音源定位の性能やロボット音声の聞き取りやすさを向上させるため，ロボット・マイクアレイの最適な位置を推定し移動させる

従来研究の多くは，一つ目の手法に分類される[Nakadai, 2000; Reid, 2003; Berglund, 2005; Kim, 2007]．複数のマイクを用いて音源定位を行う場合，空間分解能が方向によって異なる場合がある．このとき，空間分解能が最も高い方向に音源が配置されるようマイクアレイを回転することで定位精度が向上する．しかし，これらの従来研究では，マイクアレイの回転しか考慮していないため，遠くの音源をそもそも検出できないという問題がある．

二つ目の手法では，マイクアレイの位置を与えた軌跡に沿って移動させながら音源方向を推定することにより，三角測量の原理で二次元平面上での位置を定位する[Sasaki, 2006]．しかし，Sasaki らの研究では，ロボットの動作は所与であり，その最適化については議論されていない．

三つ目の手法では，雑音の位置情報に基づきロボット・マイクアレイの最適な位置を推定し，移動する[Martinson, 2007]．しかし，この従来研究では，*Signal-to-Noise Ratio: SNR* に基づきロボットの最適な位置を算出している．そのため，音源分離や音声強調など他の処理と組み合わせたシステムにそのまま適用するのは困難である．

我々は，雑音に頑健な音声発話区間検出 (*Voice Activity Detection: VAD*) を実現するため，視聴覚統合を用いた発話区間検出 (*Audio-Visual VAD: AV-VAD*) の研究を行ってきた (例えば[吉田, 2010])．VAD は他の音声処理の前処理として用いられることが多く，人とロボットがインタラクションを行う際に重要な要素技術の一つである．そこで，本稿では，AV-VAD に能動的動作を適用した VAD を**アクティブ視聴覚統合発話区間検出** *(Active Audio-Visual VAD: AAV-VAD)* とし，以降でその実現に向けた課題とアプローチ，AAV-VAD の実装とその評価について述べる．

2 アクティブ視聴覚統合の課題

ロボットには，VAD 性能が最も大きく向上するように動作を行うことが望まれる．これを実現するためには，以下の二つの課題に対処する必要がある．

1. ロボットの能動的動作が VAD 性能に対して与える影響を推定すること，
2. 複数の能動的動作を扱うため高いスケーラビリティを有すること．

ロボットは実際に動作を行う前にその効果を見積もる必要がある．話者や雑音源の情報が事前に得られない環境では，周囲の状況などから間接的に推定する必要がある．スケーラビリティは，ロボットが取りうる動作が複数存在する場合に重要となる．ロボットによる動作を一つしか考慮しないのであれば，その動作と VAD 性能の関連を調べ，詳細にモデル化することができる．しかし，このような手法では，複数の能動的な動作を扱うことが困難となる．

一つの手法として，能動的な動作を観測とみなして，回帰分析を利用することが可能である．VAD 性能を目的変数に，それ以外の周囲の観測などを説明変数とした回帰モデルを構築し，そのモデルを用いて VAD 性能を予測することができる．しかし，説明変数に対して能動的な動作により介入した場合，回帰モデルを用いた予測結果は必ずしも正しいと限らない[宮川, 2004].

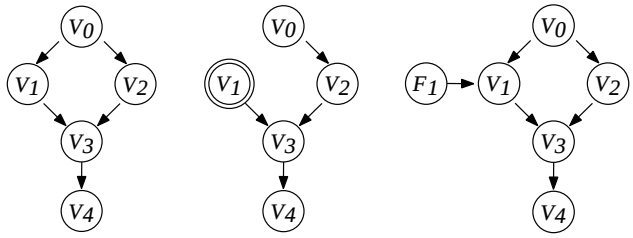
能動的動作による VAD 性能の変化量を予測するためには，観測に含まれる誤差の影響を考慮した確率論的アプローチの方が確定論的アプローチより適している．能動的動作による影響を記述することが可能な確率モデルとして，拡張確率モデル (*Augmented Probabilistic Models: APM*) がある [Pearl, 2009]．APM では，能動的な動作をするかしないかを 2 値の確率変数として新たに追加することにより，能動的な動作を記述する．しかし，この APM では，ロボットの取りうる動作の数が増加した場合に，追加する確率変数の数も増加し，事前に学習が必要な確率分布の数も指数オーダーで増加するため，スケーラビリティに問題がある．

3 因果モデルを用いたアクティブ視聴覚統合

ロボットの能動的動作が VAD 性能に与える影響を推定するため，本稿では因果モデルの一種である，因果ベイジアンネット (*Causal Bayesian Network: CBN* [Pearl, 2009]) を用いる．CBN はベイジアンネットのサブクラスであり，因果関係に基づきネットワーク構造を構築し，かつ他の部分に影響を与えることなく一つの因果関係を変更することができるモデルである．

CBN には “do-計算法” と呼ばれる能動的な動作による影響を動的に計算する手法があり，この do-計算法により，事前に必要な確率分布の数が能動的動作の数に対して線形のオーダーに抑えることができる．そのため，APM に比べてスケーラビリティがあり，本研究の目的に親和性が高い．

本稿では，CBN モデルを構成する確率変数を以下の 3 種類に分類して表記する．



a) Original　b) CBN　c) APM

Figure 1: Example of graphicsl models

- 目的変数 $\boldsymbol{y} = [y_1, \ldots, y_{n_y}]$: 推定を行う対象
- 制御変数 $\boldsymbol{s} = [s_1, \ldots, s_{n_s}]$: 能動的動作を行う対象
- 中間変数 $\boldsymbol{x} = [x_1, \ldots, x_{n_x}]$: 目的変数と制御変数以外

制御変数に対する能動的動作が目的変数へ与える影響は，以下の切断因数分解によって計算できる．

$$P(\boldsymbol{y}|\boldsymbol{x}, do(\boldsymbol{s})) = P(y_1, \ldots, y_{n_y}|x_1, \ldots, x_{n_x}, do(s_1, \ldots, s_{n_s})) = \begin{cases} \prod P(y_i|pa(y_i)) \prod P(x_i|pa(x_i)) \\ \quad \text{if } \boldsymbol{s} \text{ consistent with } do(\boldsymbol{s}), \\ 0, \quad \text{otherwise.} \end{cases} \quad (1)$$

ここで，$pa(\cdot)$ はネットワーク構造上の親である．能動的な動作の影響は，制御変数と因果関係で直接つながっている中間変数・目的変数を通して $P(\boldsymbol{y}|\boldsymbol{x}, do(\boldsymbol{s}))$ に影響を与える．

図 1a), b), c) にグラフィカルモデルの例を示す．図 1a) は動作を行わない場合を表し，同時確率分布は以下の式で求める．

$$P(\boldsymbol{v}) = P(v_0)P(v_1|v_0)P(v_2|v_0)P(v_3|v_1, v_2)P(v_4|v_3) \quad (2)$$

図 1b), c) は図 1a) に対応する CBN，APM に対して $V_1 = v_1'$ と能動的な動作により介入した場合を表し，CBN の場合は式 (3) で，APM の場合は式 (4) により同時確率分布を求める．

$$P(\boldsymbol{v}|do(v_1')) = P(v_0)P(v_2|v_0)P(v_3|v_1', v_2)P(v_4|v_3) \quad (3)$$

$$P(\boldsymbol{v}|v_1', f_1') = P(v_0)P(v_1'|v_0, f_1)P(v_2|v_0) \tilde{P}(v_3|v_1, v_2, f_1')P(v_4|v_3) \quad (4)$$

式 (4) の $\tilde{P}(v_3|v_1, v_2, f_1')$ は，式 (2) で示される能動的動作を考慮しない場合における $P(v_3|v_1, v_2)$ に対応する確率分布であり，能動的動作を扱うために変更される．一方，式 (3) では，能動的動作を考慮しない場合の確率分布と同じである．この例の様に，CBN は能動的な動作を簡潔に表すことができる．

3.1 AAV-VAD のための CBN モデル設計

CBN モデルは，我々が提案した**情報量レベル**[吉田, 2011] を能動的動作が VAD 性能に与える影響を推定できるよ

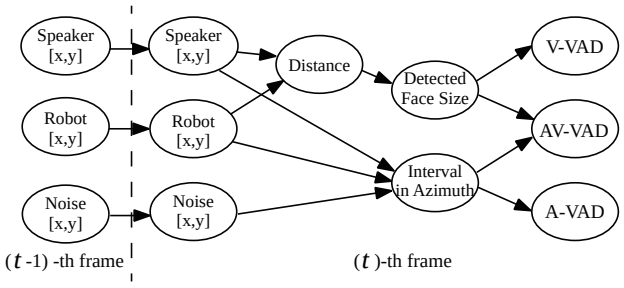


Figure 2: Model structure of the CBN

うに拡張し構築した．情報量レベルは，視覚・聴覚情報が視聴覚統合にどの程度有効であるかを示す尺度として我々が定義した．詳細は[吉田, 2011]を参照されたい．

CBN モデルの構造は，情報量レベルとロボット・話者・雑音源の幾何的情報を統合し，図 2 とした．モデルのパラメータは，[吉田, 2011]の際に予備実験で使用したデータ (話者 3 人，各 60 発話を 14 条件で実機ロボットにより収録) を用いて学習した．話者・ロボット間の距離を 1.5 [m], 2.5 [m] として発話データを収録し，またロボットの頭部伝達関数を用いてロボットから見た話者・雑音源のなす角度が 0 度，30 度，. . . , 180 度となるよう音響雑音を合成し，収録したデータに重畳した．この雑音データに対して AV-VAD を行い，その VAD 性能とロボット・話者・雑音源の位置関係を用いてパラメータの学習を行った．なお，訓練データに含まれない位置関係については，訓練データから補間して補った．

3.2 CBN を用いた移動ロボットナビゲーション

ロボットは do-計算法を用いて，下記のように条件付き期待値を評価関数として最適な能動的動作 s^* を選択する．

$$s^* = \arg\max_s \mathbb{E}\left[\boldsymbol{y}|\boldsymbol{x}, do(\boldsymbol{s})\right] \quad (5)$$

$$= \arg\max_s \sum \boldsymbol{y} P(\boldsymbol{y}|\boldsymbol{x}, do(\boldsymbol{s})) \quad (6)$$

ここで，$\mathbb{E}[\cdot]$ は条件付き期待値を表す．

4 アクティブ視聴覚統合発話区間検出システム

図 3 に提案手法に基づく AAV-VAD システムを示す．本稿では，テストベッドとして図 3 に示すヒューマノイドロボット “Hearbo” を用いる．Hearbo の下半身は全方位台車となっており，その全方位台車の上に上半身が設置されている．

Table 1: CBN モデルに用いる確率変数

意味	分類
ロボットの位置 (x, y [m]) と向き (θ [deg.])	制御変数
話者の位置 (x, y [m])	中間変数
雑音源の位置 (x, y [m])	中間変数
ロボットから見た話者と雑音源のなす角度 [deg.]	中間変数
ロボットから話者までの距離 [m]	中間変数
検出された顔の大きさ [pixels]	中間変数
A-VAD 性能の推定値 [0(悪い) to 1(良い)]	目的変数
V-VAD 性能の推定値 [0(悪い) to 1(良い)]	目的変数
AV-VAD 性能の推定値 [0(悪い) to 1(良い)]	目的変数

全方位台車には，4 つの車輪があり，それぞれの車輪には駆動用とステアリング用の 2 つのモータとエンコーダが備えられており，それぞれを独立に制御することができる．上半身には，首の 3 軸を制御するモータとエンコーダが備えられている．なお，実際には腕や手などにも自由度があるが，今回は使用していない．

Hearbo の頭部には 16 ch のマイクロホンアレイが設置されており，16 kHz，24 bit で同期収録する．また，右目の位置にカメラが一つ設置されており，30 Hz, 8 bit グレースケール，640×480 pixel の画像を収録する．

ソフトウェアは 4 つのブロック (視覚特徴量抽出部，聴覚特徴量抽出部，視聴覚発話区間検出部，ロボット制御部) から構成されている．ロボット制御部以外は状態遷移モデルを用いた視聴覚発話区間検出システム [Yoshida, 2012a] を用いるため，これらについては概略のみを述べる．詳細は [Yoshida, 2012a]を参照されたい．

視覚特徴量抽出部では，カメラで取得した画像から顔検出・唇抽出を行い，抽出された唇の縦横長に基づいた特徴量 [Yoshida, 2012a] を計算する．また同時に，検出された顔の位置とサイズを用いて，ロボットから見た話者の位置を以下の式を用いて推定する．

$$d = c_1 r + c_0, \quad c_1 = -0.0106, \; c_0 = 4.04 \quad (7)$$

なお，顔検出には，MindReader[1] に含まれる顔検出を用いた．

聴覚特徴量抽出部では，マイクロホンアレイの入力から音源定位により話者と雑音源の方向を推定したのち音源分離を行い，分離音から聴覚特徴量を抽出する．音源定位には *Generalized-Eigen Value-Decomposition-based MUltiple SIgnal Classification :GEVD-MUSIC* を，音源分離には，*Geometric High-order Dicorrelation-based Source Separation: GHDSS* を，聴覚特徴量には *Mel-Scale Log Spectrum: MSLS* をそれぞれ用いた．音源定位・音源分離・MSLS 抽出は，ロボット聴覚ソフトウェア HARK [Nakadai, 2010] を基に実装した．これらの処理の詳細は [Nakadai, 2010]を参照されたい．

なお，実装に用いた GEVD-MUSIC では，音源がある方向で大きな値となる空間スペクトルが出力として得られる．この空間スペクトルは，音源位置の方位角に比べ距離の推定が困難であるため，三角測量により二次元座標を算出する．詳細な説明は，[Yoshida, 2012b]を参照されたい．

視聴覚発話区間検出部では，唇の縦横長から求めた特徴量と MSLS から，最大事後確率推定により発話・非発

[1] http://trac.media.mit.edu/mindreader/

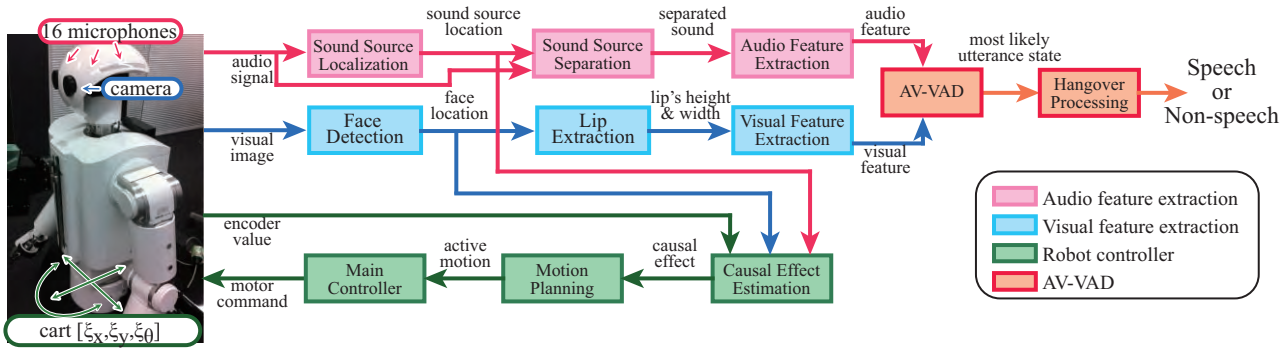


Figure 3: System architecture of AAV-VAD

話を判別する．なお，この確率計算は *Open Probabilistic Network Library: OpenPNL* [2] を基に実装した．最後に判別結果のフラグメンテーションを修正するため，膨張・縮退に基づく Hangover 処理を行い，その結果を VAD 結果として出力する．

ロボット制御部は *Robot Operating System: ROS* [3] を基に実装した．ロボットの位置は台車のエンコーダ値から得られるオドメトリ，話者の位置は視覚特徴量抽出時に行う顔検出の結果，雑音源の位置はオドメトリと音源定位を組み合わせた三角測量を用いて，それぞれ求める．これらの観測値を CBN により統合し，最適な動作を選択する．能動的な動作の候補として，本稿ではロボットの位置を扱う．式 (8) に示すように，現在の位置を中心に半径 Δ の範囲内への移動を候補とし，その範囲内で式 (6) に基づき最適な動作を選択し実行する．

$$\boldsymbol{s} \in [\xi_x + \Delta_x, \xi_y + \Delta_y], \quad \Delta_x^2 + \Delta_y^2 < \Delta^2 \tag{8}$$

システムの実装にあたり，$\Delta = 1$ [m] とし，また計算を簡略化するため，ロボットの移動先を 0.1 [m] 間隔の離散グリッド上に制限した．複数の地点で同じ推定結果となる場合は，その中で最も現在の位置に近い所へ移動することとした．

5 評価

提案手法の有効性を示すため，図 4a), b) に示すように話者と雑音源の距離が近い場合 (condition 1) と遠い場合 (condition 2) の 2 条件で発話区間検出実験を行った．実験室は図 4c) に示すように背景が整っており，視覚情報への雑音は少ない．一方，聴覚情報は，ラウドスピーカーからの音楽やロボット自身のモータやファンからの自己雑音が混入している．

比較のため，以下の手法を用いて実験を行った．

- *Baseline*: 初期位置から移動しない静的な手法，
- *Active (Linear)*: 話者へ直線的に近づく手法，
- *Active (MReg.)*: 重回帰モデルに基づいて VAD 性能を推定する手法，
- *Active (Prop.)*: 因果モデルに基づいて VAD 性能を推定する手法．

Active (Linear) では，初期位置から話者の方向へと近づき，画像から検出される顔のサイズが VAD モデルの学習に用いた画像と同じになったら静止する．*Active (MReg.)* では，重回帰モデル (*Multi Regression model: MReg*) を用いて VAD 性能の予測を行い，一番性能向上が見込める位置へ移動する．重回帰に用いる変数は，多重共線性を考慮しながら実験的に求め，雑音源の位置とロボットから見た話者と雑音源のなす角度を用いるモデルが選択された．なお，この重回帰モデルはモデルの当てはまりの良さを表す決定係数 $R^2 = 0.93$ が *Active (Prop.)* の決定係数 $R^2 = 0.78$ よりも高くなった．

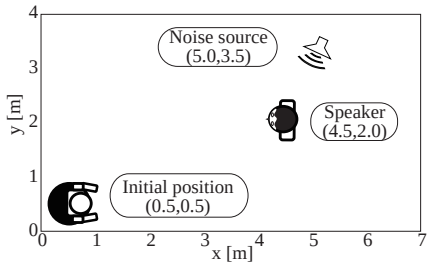
AV-VAD システムのモデル学習には，話者 3 人がロボットから 1.5 [m]，2.5 [m] の位置でそれぞれ 60 単語ずつ発話したデータを用いた．

評価には，“6-word command sentence[4]” と呼ばれる短い命令文を日本語に翻訳して収録した視聴覚データベースを使用した．話者は 2 人であり，各話者は T0–T4 のそれぞれでおよそ 90 [s] の間に 20 文ずつ発話している．雑音源にはラウドスピーカーを用い，音楽 (RWC Music Database Jazz No. 41[5]) を流した．学習データと評価データは収録は同じ部屋で行ったが，話者と発話内容は学習と評価で異なる．

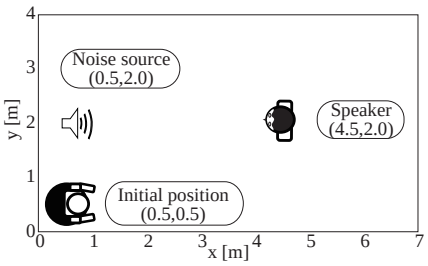
なお，本稿では次のような仮定をおいた．話者と雑音源の数はそれぞれ 1 つずつで，実験中は移動しない．ロボットと人は向かい合っている．また，ロボットの初期位置は $(0.5, 0.5)$ とし，衝突回避のため人と雑音源から 1 [m] 以内には近づかないようにした．

評価指標には，VAD の精度 (実際の発話に対して正しく検出された割合) を用いた．

[2] http://sourceforge.net/projects/openpnl/
[3] http://www.ros.org/wiki/
[4] http://spandh.dcs.shef.ac.uk/gridcorpus
[5] http://staff.aist.go.jp/m.goto/RWC-MDB/



a) Condition 1



b) Condition 2

c) Snapshot of the experiment

Figure 4: Experimental conditions

5.1 実験結果

図 5 は条件 1,2 における各手法による VAD 性能の推定結果を示している．上段 (a,b,c) が条件 1 を，下段 (d,e,f) 条件 2 に対応し，左列 (a,d) は *Active (Linear)*，中央 (b,e) は *Active (MReg.)*，右列 (c,f) は *Active (Prop.)* に対応する．図 6 は各条件における実際の VAD 精度を示している．条件 1 では，初期位置から移動しない *Baseline* での性能は約 60%となっている．これに対し，能動的な動作を利用する 3 つの手法では，移動するに従い徐々に性能が向上している．*Active (Linear)* は雑音源の位置を考慮しないため，話者と雑音源がロボットから見て近い方向に配置され，音源分離性能が劣化しており，VAD 性能は T1 以降 60%で一定となった．一方 *Active (Prop.)* と *Active (MReg.)* では，話者との距離を縮めるだけでなく雑音源の位置も考慮して移動しているため，回り込むような動作となった．その結果，*Active (Linear)* と比べ，さらに 5.0 ポイント性能が向上した．

条件 2 では，ロボットの初期位置が雑音源に近く，T0, T1 では VAD 性能が条件 1 の場合と比べ低い．この条件では，*Active (MReg.)* が話者との距離を考慮していないため，話者・ロボット・雑音源が一直線上にならんだ地点で停止した．この位置では音源分離性能が最高となるため VAD 性能も T0 から 5 ポイント向上している．その一方で，視覚特徴量はまだ向上の余地があるため，視覚特徴量も考慮にいれる *Active (Prop.)* はさらに 7.5 ポイント性能が向上した．

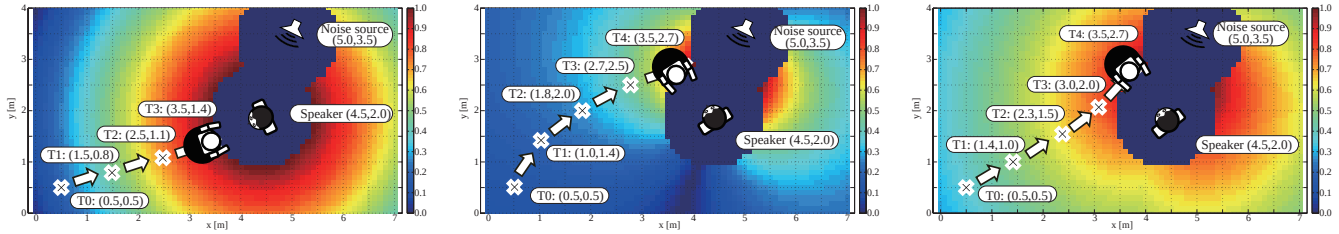
5.2 考察

まず，2 節で述べた課題と提案法について考察する．実験結果から，例え話者に近づくという，非常にシンプルな方針であっても，移動により VAD 性能が向上するということが示された．しかし，条件 1 のような状況へは対処できず，環境への適応という面では，その有効性は限定的である．そのため，VAD 性能の推定を行うことの必要性が改めて示された．また，*Active (MReg.)* については，条件 1 では因果モデルを用いた場合とほぼ同じ推定結果を与えたが，条件 2 では異なる推定結果となり，実際の VAD 性能の向上も限定的であった．このことから 2 で述べた様に回帰モデルは能動的な動作を扱うと必ずしも適切な推定結果が得られるとは限らないということが裏付けられた．提案法では，条件 1，2 の両方で良い推定結果が得られ，本研究の目的に適している．なお，*Active (MReg.)* は，モデルの学習データに対する当てはまりの良さを示す決定係数が *Active (Prop.)* の決定係数より大きい ($R^2_{MReg.} = 0.93, R^2_{Prop.} = 0.78$)．しかし，実験結果では *Active (Prop.)* が *Active (MReg.)* に比べて大きな性能向上を示した．この結果は学習データのサンプルを増やすことで変化する可能性があるものの，決定係数に基づくモデル選択が必ずしも本研究目的には適さないことと，*Active (Prop.)* は *Active (MReg.)* と比べ今回用いたような少ない学習データから妥当なモデルが得られることが分かった．

次に，提案法の音響・画像雑音に対する頑健性について考察する．音響雑音の影響については，定常雑音の場合はその影響を減らすように移動することで，突発性雑音の場合は VAD の後処理である hangover 処理を行うことである程度の対処が可能である．また，この二つの方法で対処できない環境では,「雑音源を取り除く」,「大きな声で発話してもらうようにお願いをする」といった能動的動作を加えることで対処できる可能性がある．画像雑音の影響は，その種類によって影響が大きく変化する．特にテレビ画面に映った人を話者と誤認した場合，提案法では対処することができない．これを解決するためには赤外線カメラやレーザーレンジファインダーを併用するなどの方法が必要となる．

6 終わりに

本稿では，能動的動作を AV-VAD へ適用した AAV-VAD を実現するため因果モデルに基づく手法を提案した．因果モデルには視聴覚情報と能動的な動作を統一的に扱える枠組みをもつ CBN を用い，do-計算法により動作の影響を推定し，その推定結果に基づき最適な行動を行う．提案法に基づく AAV-VAD システムをヒューマノイドロボット Hearbo に実装した．提案法の有効性を検証するため，単純に話者に近づく手法，重回帰分析に基づき動作を選択する手法，能動的な動作を使わない静的な手法と比較を



a) Active (Linear) in the condition 1 b) Active (MReg.) in the condition 1 c) Active (Prop.) in the condition 1

d) Active (Linear) in the condition 2 e) Active (MReg.) in the condition 2 f) Active (Prop.) in the condition 2

Figure 5: Estimation results of AAV-VAD performance

a) VAD performances in the condition 1 b) VAD performances in the condition 2

Figure 6: Actual AAV-VAD performances

行い，提案法がそれぞれの手法より平均で 10.0, 2.5, 3.8 ポイント高いことを示した．

今後の課題は，より詳細な評価を行うこと，話者の顔が検出できない場合への対処，複数話者や移動する話者への対応が挙げられる．また，ロボットが実行可能な能動的動作は移動以外にも，例えば雑音源である音楽を止める，話者に大きな声で発話するように促すなどが考えられる．これらの動作を取り入れることも今後の課題である．

謝辞

本研究の一部は科研費 (24118702，22700165)，特別研究員奨励費の補助を受けた．

参考文献

[Berglund, 2005] E. Berglund and J. Sitte,: Sound source localisation through active audition, *in Proc. of IROS,* pp. 653–658, 2005.

[Kim, 2007] H.D.Kim *et al.*: Human-robot interaction in real environments by audio-visual integration, *Control, Automation, and Systems,* Vol. 5, pp. 61–69, 2007.

[Martinson, 2007] E. Martinson and D. Brock: Improving human-robot interaction through adaptation to the auditory scene, *in Proc. on ACM/IEEE Int. Conf. on Human-Robot Interaction,* pp. 113–120, 2007.

[Nakadai, 2000] K. Nakadai, *et al.*: Active Audition for Humanoid, *Proc. of the 17th National Conf. on Artificial Inteligence,* pp. 832–839, 2000.

[Nakadai, 2010] K. Nakadai *et al.*: Design and implementation of robot audition system 'HARK', *Advanced Robotics,* Vol. 24, Issue. 5-6, pp. 739–761, 2010.

[Pearl, 2009] J. Pearl: Causality. second edition, Cambridge University Press, 2009.

[Reid, 2003] G. L. Reid and E. Milios: Active stereo sound localization, *J. Acoust. Soc. Am,* Vol. 113, pp. 61–69, 2003.

[Sasaki, 2006] Y. Sasaki *et al.*: Multiple sound source mapping for a mobile robot by self-motion triangulation, *in Porc. of IROS,* pp. 380–385, 2006.

[Yoshida, 2012a] T. Yoshida and K. Nakadai: Audio-visual voice activity detection based on an utterance state transition model, *Advanced Robotics,* Vol. 26, Issue 10, pp. 1183–1201, 2012.

[Yoshida, 2012b] T. Yoshida and K. Nakadai: Active audio-visual integration for voice activity detection based on a causal Bayesian network, *in Proc. of Humanoids,* 2012 (to appear).

[宮川, 2004] 宮川雅巳: **統計的因果推論 – 回帰分析の新しい枠組み –**, 朝倉書店，2004.

[吉田, 2010] 吉田他: ロボットを対象とした二階層視聴覚統合音声認識システム，**日本ロボット学会誌**，Vol. 28, No. 8, pp. 970–977, 2010.

[吉田, 2011] 吉田他: ロボットのための情報量レベルに基づくアクティブ視聴覚統合の検討，**第 *29* 回日本ロボット学会学術講演会**，3A3-4，2011.

社団法人 人工知能学会
Japanese Society for
Artificial Intelligence

人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-B202-9

ベイズモデルによるマイクロホンアレイ処理の移動ロボットへの応用

Bayesian Microphone Processing and Its Application to Mobile Robot Audition

大塚 琢馬 †, 石黒 勝彦 ‡, 澤田 宏 ‡, 奥乃 博 †

Takuma Otsuka†, Katsuhiko Ishiguro‡, Hiroshi Sawada‡, Hiroshi G. Okuno†

† 京都大学大学院情報学研究科, ‡ 日本電信電話株式会社 NTT コミュニケーション科学基礎研究所

†**Graduate school of Informatics, Kyoto University,** ‡ NTT Communication Science Laboratories, NTT Corporation

Abstract

This paper addresses a simultaneous sound source localization and separation method using a microphone array and its application to an auditory sensing by a mobile robot. The auditory function embedded in the mobile robot should deal with (1) the uncertainties in the environment such as the unknown reverberation and source number, (2) time-varying sound source locations observed by the robot, and (3) the motor noise caused by the robot motion. Our Bayesian formulation is employed to efficiently cope with the uncertainties. Sound source separation experiments in indoor and outdoor environments confirm encouraging results.

1 はじめに

ロボットが環境中を移動しながら，自身に備え付けられたセンサを用いてロボットがいる環境から情報を抽出することは，ロボットによる自律的に環境の探索，あるいは，遠隔のロボット操作者のナビゲーションにとって重要である．従来の移動ロボットによるセンシング技術は，カメラからの視覚情報に基づく自己位置推定と環境地図作成 (SLAM; Simultaneous Localization and Mapping) [Thrun *et al.*, 2004; Se *et al.*, 2005] などを中心に発達してきた．これらの視覚情報処理に加えて，ロボットが聴覚情報を扱えるようになると，次のような機能強化が期待できる．(1) 物体のオクルージョンに対する頑健性の獲得，(2) ものの変化の知覚，(3) 音声コミュニケーションの実現．例えば，(1) 視覚のみに頼ると壁の向こうの情報は取得出来ないが，音を聴くことで壁に遮られた場所の知覚を試みることが出来る．(2) 物体が動いたり状況が変化する場合には音を伴うことが多い．例として，Figure 1 上のように，グラスが机から落ちた場合は「ガシャン」と音がする．音環境理解機能を持つロボットであれば，このような出来事に気づきやすくなることが期待できる．(3) もちろん，人間の音声を指令として受け取るなどのコミュニケーションチャネルへの寄与も考えられる．

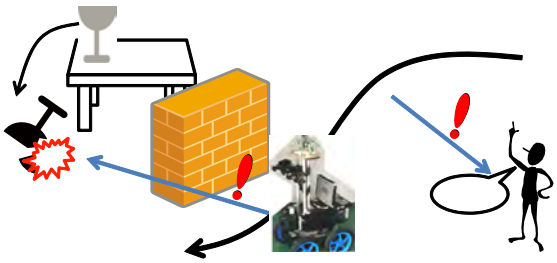


Figure 1: 移動ロボットにおける聴覚機能

ロボットが存在する環境中に複数の音源が同時に存在することがある．そのため，ロボットには複数音源の混合観測音から，個々の音源に分解する「聴き分け」機能が必須である．このような複数音源を扱うため，複数のマイクロホンを利用するマイクロホンアレイ処理 [Benesty *et al.*,] がよく用いられる．本稿では，マイクロホンアレイを用いた混合観測音から，各音源のある方向を推定する音源定位と，各音声信号を抽出する音源分離法を示し，移動ロボットの適用について述べる．マイクロホンアレイ処理に関する最も重要な課題の 1 つは，観測音中に含まれる未知要因を対処するという点である．マイクロホンアレイ処理の性能を左右する環境中の未知要因としては，観測音に含まれる音源数や，音源とマイク間の相対位置や周囲の壁などに依存する残響などが挙げられる．これらの未知要因に対して，本手法では音源数や残響に関する仮定がなく，入力データである観測音からこれらの情報も柔軟に扱うことの出来るベイズモデルに基づく音源定位・分離法を示す．

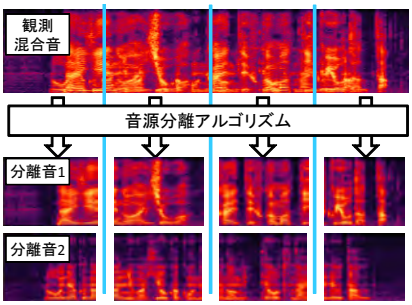


Figure 2: 時分割処理: 各区間内の空間特性は定常と仮定

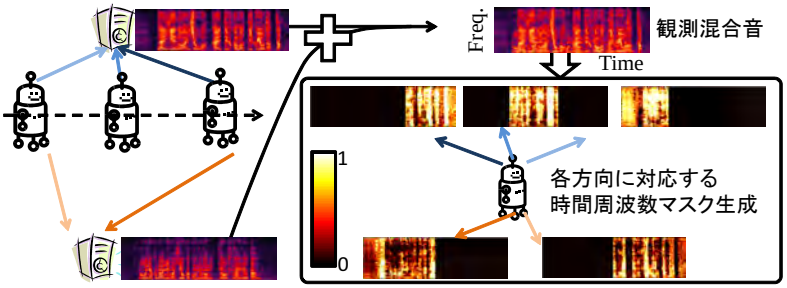


Figure 3: 方向ごとの時間周波数マスク推定: 音源が通過した方向に対応するマスクを生成

2 移動ロボットの聴覚機能における諸課題

ロボットが移動しながら混合音を観測し，個々の音源を分離する際には，音環境の未知要因の存在に加えて，2 つの大きな課題が存在する．

1. ロボットに搭載されたマイクロホンアレイと音源との間の相対的な位置関係 (以下，空間特性と呼ぶ) が時間変化する点と，
2. ロボットの移動などに伴って，車輪を動かすモータ音や地面の段差を乗り越える音が観測音に混入するため，目的となる外部音の signal-to-noise ratio (SNR) が劣化する点である．

マイクロホンアレイは空間フィルタリング機能を持つ．つまり，マイクロホンアレイによる音源分離は原則として，異なる方向から到来する音源の分離が可能である．従って，分離対象の音源は空間的にスパースに存在することを仮定する．また，マイクロホンアレイによる音源定位のためには，音源到来方向と各マイクへの波面到達時間差などの対応付けが必要なため，各マイクロホンの正確な配置や，ステアリングベクトルや伝達関数と呼ばれる波面到達時間差やマイク間の観測音振幅比などの事前情報が必要となる．さらに，多くのマイクロホンアレイによる音源定位や音源分離は定常な空間特性を仮定しているため，空間特性が時変である移動ロボット問題への適用には工夫が必要である．

ロボット聴覚ソフトウェア HARK [Nakadai *et al.*, 2010] は，Figure 2 のように，観測音を時分割し，各区間内では定常な空間特性を仮定した上で分離処理を行う．より具体的には，0.5 [s] 程度の固定窓幅で音源定位を行い，定常方向に各音源が定位された区間ごとに分離処理を行う．このシステムは実時間での音源定位・分離を実現するが，定位が失敗すると分離性能が大きく劣化するほか，精確な音源定位のためには音源数を与える必要や，残響時間など環境に依存したパラメータを設定する必要がある．また，音源分離も環境に依存する伝達関数を事前知識として要すること，環境中の音源数がマイク数未満である必要があるなど，環境依存性が課題として残されている．また，ロボット動作などに伴う自己発生音に対しては，マイクロホンアレイに対する自己発生音源の方向を指定することで，自己発生音を抑圧した信号の抽出を行う．この方法は，自己発生音源の位置がマイクロホンアレイに対して相対的に変わらない場合に有効である．

本稿で示す手法も定常な空間特性を仮定した手法であるが，Figure 3 のように時間変化する空間特性に対応する．Figure 3 左側のように，移動しながら音源を観測すると，ロボットから見た音源方向は時間変化する．このように観測した混合音に対して，十分大きなクラスタ数を用いて本手法による分離を行うと，Figure 3 右側のように，観測音の中で音源が通過した方向に応じた時間周波数マスク (TF マスク) が自動的に生成される．従って，Figure 2 のように定常空間特性を仮定できる十分小さな窓幅などを設定する必要なく，観測音中の各音源の相対的な移動に適した時間幅での音源分離が望める．このように，本手法は環境依存の要素が極力覗かれている点が利点であるが，クラスタリングの反復計算に由来する計算時間の大きさが欠点として挙げられる．

3 統一的音源定位・分離ベイズモデル

本手法は音源定位・分離問題を時間周波数領域でのクラスタリング問題として扱う．音源分離は各時間周波数点の信号ベクトルのクラスタリング問題とし，音源定位は各クラスとステアリングベクトルとの対応付け問題とする．パーミュテーション解決を含む分離処理には Latent Dirichlet Allocation (LDA) [Blei *et al.*, 2003] を利用する．基本的なアイデアは，各時間フレームにおいて特定の音源が多く出現するようモデル化することで，各音源のマスクが周波数ビンをまたいで時間的に同期させるという仕組みである [Otsuka *et al.*, 2012]．通常の LDA は有限混合モデルであるため，本稿では音源数が無制限の HDP [Teh *et al.*, 2006] へと拡張する．

本節で用いる記号を Table 1 にまとめる．Figure 4 は確率変数の条件付き独立性を図示したグラフィカルモデルである．二重円で示された $\mathbf{x}_{tf}$ は観測変数を表し，円で示

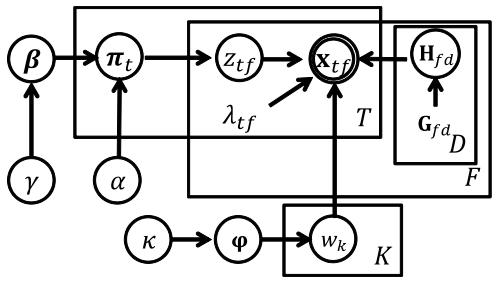


Figure 4: グラフィカルモデル

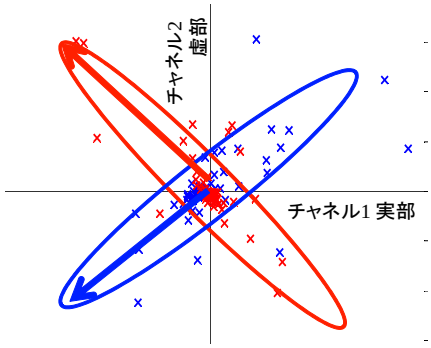


Figure 5: 周波数ビン 3200 Hz での観測信号散布図

Table 1: 記号表

記号	意味
t	時間フレーム ($1 \leq t \leq T$)
f	周波数ビン ($1 \leq f \leq F$)
k	クラスインデックス ($1 \leq k \leq K$)
d	方向インデックス ($1 \leq D \leq D$)
M	マイク数
$\mathbf{x}_{tf}$	観測信号 M 次元列ベクトル
z_{tf}	時間フレーム t 周波数ビン f のクラス
π_t	時間フレーム t でのクラス割合
w_k	クラス k の方向
φ	全クラスに対する方向割合
λ_{tf}	$\mathbf{x}_{tf}$ の逆数スケール
$\mathbf{H}_{fd}$	周波数ビン f 方向 d の精度行列
n_{tk}	時間フレーム t のクラス k サンプル数
n_{fk}, n_{fd}	周波数ビン f のクラス k，方向 d サンプル数
c_d	方向 d に割り当てられたクラス数

された記号は潜在確率変数，囲いのない変数は定数を表す．3.1 節では多チャンネル観測信号とステアリングベクトルの関係を示し，3.2 節に CGS による推論，3.3 節では定位，分離結果の出力法を説明した後，3.4 節で推論の初期化方法を述べる．変数集合は下付添字を略しチルダを付けて表す (例: $\tilde{\mathbf{x}} = \{\mathbf{x}_{tf} | 1 \leq t \leq T, 1 \leq f \leq F\}$)．次節以降で詳細は述べるが，$\tilde{\mathbf{z}}$ の推論が TF マスク推定による分離処理に相当し，$\tilde{\mathbf{w}}$ が定位に相当する．

3.1 多チャンネル複素信号の生成モデル

本節では Figure 4 に示された生成プロセスを説明する．時間周波数領域の多チャンネル信号観測モデルとして covariance mdoel [Duong *et al.*, 2010] を採用する．このモデルでは，各時間周波数点は平均ゼロ，スケール時変共分散の多変量複素正規分布に従うとする．Figure 5 に青と赤で示される 2 つの音源を 2 チャンネルで観測した信号の散布図を示す．これらの点は次のように生成されたと仮定する．時間 t，周波数ビン f の時間周波数点で優勢な信号は方向 d から到来しているとき，観測信号は $\mathbf{x}_{tf} = s_{tf}\mathbf{q}_{fd}$ として表されるとする．但し，s_{tf} はその点における音源成分とし，$\mathbf{q}_{fd}$ は方向 d に対応するステアリングベクトルである．ベクトル $\mathbf{x}_{tf}$ は M 次元ベクトルで，各成分は対応するマイクでの観測に対応する．このベクトルの共分散は $\mathbb{E}[\mathbf{x}_{\approx\upsilon}\mathbf{x}_{\approx\upsilon}^{\mathbb{H}}] = \mathbb{E}[|\sim_{\approx\upsilon}|^{\nvDash}\mathbf{q}_{\upsilon}\mathbf{q}_{\upsilon}^{\mathbb{H}}]$ と書ける．ただし，$\cdot^H$ はエルミート転置を表す．

Figure 5 に楕円で示された各音源の共分散行列は値の大きな固有値を持つ．これに対応する固有ベクトル (図中矢印) は，その音源がある方向に対応するステアリングベクトルと同一方向である．すなわち，各時間周波数点の属する楕円の推定は音源分離に対応し，クラスタの共分散の固有ベクトルを調べることは音源定位に対応する．

各時間周波数点の共分散は時変スケール $|s_{tf}|^2$ と固定の方向行列 $\mathbf{q}_{fd}\mathbf{q}_{fd}^H$ へと分解出来る．行列部分が固定されることが，音源位置が一定であるという仮定に対応する．ここで分布の共役性を取り入れるために共分散行列の逆数である精度行列 $\lambda_{tf}\mathbf{H}_{fd}$ を導入する．λ_{tf} は時変スケールであり，$\mathbf{H}_{fd} \approx (\mathbf{q}_{fd}\mathbf{q}_{fd}^H + \varepsilon\mathbf{I})^{-1}$ とする．$\mathbf{I}$ は単位行列である．λ_{tf} は [Otsuka *et al.*, 2012] では確率変数として推論対象であったが，本手法は $\lambda_{tf} = |\mathbf{x}_{tf}|^{-1}$ として値を固定する．この結果，$\mathbf{H}_{fd}$ を積分消去し，効率的な周辺化推論が可能となる．複素正規分布に従う尤度関数は次のように表される．

$$\mathbf{x}_{tf} | z_{tf}, \tilde{\mathbf{w}}, \lambda_{tf}, \tilde{\mathbf{H}} \sim \mathcal{N}_{\mathbb{C}}(\mathbf{x}_{tf} | \mathbf{0}, (\lambda_{tf}\mathbf{H}_{fw_{z_{tf}}})^{-1}), \quad (1)$$

ここで，z_{tf} と w_k はそれぞれ $\mathbf{x}_{tf}$ のクラス，クラス k の方向を表す．従って，$w_{z_{tf}}$ は $\mathbf{x}_{tf}$ が属する方向となる．平均 μ，精度行列 Λ の複素正規分布 $\mathcal{N}_{\mathbb{C}}(\mathbf{x}|\mu, \Lambda^{-1})$ の確率密度関数の定義は $\frac{|\Lambda|}{\pi^M}\exp\{-(\mathbf{x}-\mu)^H\Lambda(\mathbf{x}-\mu)\}$ である [van den Bos, 1995]．$|\Lambda|$ は行列 Λ の行列式である．方向行列 $\mathbf{H}_{fd}$ は共役事前分布である複素ウィシャート分布 [Conradsen *et al.*, 2003] に従う．

$$\mathbf{H}_{fd} \sim \mathcal{W}_{\mathbb{C}}(\mathbf{H}_{fd} | \nu_{fd}, \mathbf{G}_{fd}), \quad (2)$$

複素ウィシャート分布 $\mathcal{W}_{\mathbb{C}}(\mathbf{H}|\nu, \mathbf{G})$ の確率密度関数は $\frac{|\mathbf{H}|^{\nu-M}\exp\{-\mathrm{tr}(\mathbf{H}\mathbf{G}^{-1})\}}{|\mathbf{G}|^{\nu}\pi^{M(M-1)/2}\prod_{i=0}^{M-1}\Gamma(\nu-i)}$ と定義される．ここで，$\mathrm{tr}(\mathbf{A})$ は行列 $\mathbf{A}$ の跡，$\Gamma(x)$ はガンマ関数である．ハイパーパラメータは $\nu_{fd} = M$，$\mathbf{G}_{fd} = (\mathbf{q}_{fd}\mathbf{q}_{fd}^H + \varepsilon\mathbf{I}_M)^{-1}$ と定める．$\mathbf{G}_{fd}$ は所与のステアリングベクトル $\mathbf{q}_{fd}^H\mathbf{q}_{fd} = 1$ と正規化して利用する．ε は逆行列演算のために導入し，0.01 を用いた．

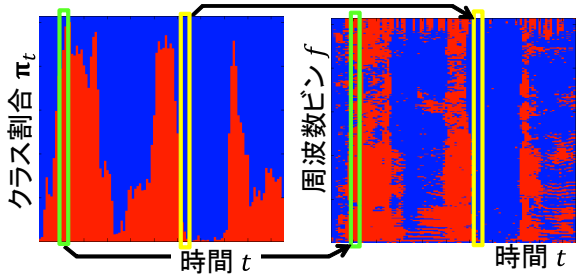


Figure 6: 左: クラス割合，右: 2 音源の TF マスク.

LDA の無限混合拡張である HDP [Teh *et al.*, 2006] を z_{tf} の生成過程として利用する．このモデルはパーミュテーション解決のために導入する [Otsuka *et al.*, 2012]．まず無限個のクラスの大域的な割合 β が生成され，時間ごとのクラス割合 π_t が β を元に生成される．最後に，各時間周波数点において z_{tf} が π_t に基づいて生成される．Figure 6 に示すように，各音源が優勢になる様子は周波数ビンをまたいで時間的に同期する様子が観察される．このことから，π_t を導入することで各クラスが時間ごとに同期する振る舞いをもたせることでパーミュテーション解決に寄与できると期待できる．HDP の棒折り過程による生成は次のように表される [Teh *et al.*, 2006].

$$\beta|\gamma \sim \mathrm{GEM}(\gamma), \quad \pi_t|\alpha,\beta \sim \mathrm{DP}(\alpha,\beta), \quad z_{tf}|\pi_t \sim \pi_t, \quad (3)$$

ただし，GEM(γ) は集中度 γ の Griffiths-Engen-McCloskey 分布，DP(α,β) は集中度 α，基底測度 β のディリクレ過程である．集中度の事前分布としてはガンマ分布を $\gamma \sim \mathcal{G}(\gamma|a_\gamma,b_\gamma)$，$\alpha \sim \mathcal{G}(\alpha|a_\alpha,b_\alpha)$ として用いる．ハイパーパラメータは $a_\gamma = 0.05, b_\gamma = 5, a_\alpha = 0.01, b_\alpha = 1$ とした．

方向変数 w_k は音源定位だけでなく，パーミュテーション解決にも寄与する．なぜなら，全周波数ビンに渡って生成されたクラスを同一方向から到来する信号として選別するためである．この潜在変数はディリクレ分布から生成される方向割合 φ に従う．

$$\varphi|\kappa \sim \mathcal{D}(\varphi|\frac{\kappa}{D}\mathbf{1}_D), \qquad w_k|\varphi \sim \varphi, \qquad (4)$$

ただし，$\mathbf{1}_D$ は要素がすべて 1 の D 次元ベクトルである．$\mathcal{D}(\cdot|\alpha)$ はパラメータ α に従うディリクレ分布を表す．本モデルはマイクロホンアレイの空間解像度は有限であるため，有限の方向数 D を扱う．κ に対してもガンマ分布 $\mathcal{G}(\kappa|a_\kappa,b_\kappa)$，$a_\kappa = 1, b_\kappa = 1$ を事前分布として利用する．

3.2 周辺化ギブズサンプラ (CGS) による推論

音源分離・定位問題には $\tilde{\mathbf{z}}$ と $\tilde{\mathbf{w}}$ の推論が鍵となる．これらの潜在変数は $\pi,\varphi,\tilde{\mathbf{H}}$ を積分消去した次の CGS

$$p(z_{tf}=k|\tilde{\mathbf{x}},\vartheta\backslash z_{tf}) \propto (\alpha\beta_k + n_{tk}^{\neg tf})$$
$$\frac{\Gamma(\hat{\nu}_{fw_k}^{\neg tf}+1)}{\Gamma(\hat{\nu}_{fw_k}^{\neg tf}-M+1)} \frac{\left|\mathrm{inv}(\hat{\mathbf{G}}_{\mathrm{fw_k}}^{\neg \mathrm{tf}})\right|^{\hat{\nu}_{fw_k}^{\neg tf}}}{\left|\mathrm{inv}(\hat{\mathbf{G}}_{\mathrm{fw_k}}^{\neg \mathrm{tf}}) + \lambda_{\mathrm{tf}}\mathbf{x}_{\mathrm{tf}}\mathbf{x}_{\mathrm{tf}}^{\mathrm{H}}\right|^{\hat{\nu}_{fw_k}^{\neg tf}+1}}, \qquad (5)$$

$$p(w_k=d|\tilde{\mathbf{x}},\vartheta\backslash w_k) \propto (\frac{\kappa}{D} + c_d^{\neg k})$$
$$\prod_f \left\{ \prod_{i=0}^{M-1} \frac{\Gamma(\hat{\nu}_{fd}^{\neg k}+n_{fk}-i)}{\Gamma(\hat{\nu}_{fd}^{\neg k}-i)} \frac{\left|\mathrm{inv}(\hat{\mathbf{G}}_{\mathrm{fd}}^{\neg \mathrm{k}})\right|^{\hat{\nu}_{fd}^{\neg k}}}{\left|\mathrm{inv}(\hat{\mathbf{G}}_{\mathrm{fd}}^{\neg \mathrm{k}}) + \sum_{\mathrm{t:z_{tf}=k}} \lambda_{\mathrm{tf}}\mathbf{x}_{\mathrm{tf}}\mathbf{x}_{\mathrm{tf}}^{\mathrm{H}}\right|^{\hat{\nu}_{fd}^{\neg k}+n_{fk}}} \right\}, \qquad (6)$$

によって確率的に生成する．ただし，$\vartheta\backslash z$ は z を除く全ての潜在変数の集合で，上付添字 $\neg tf$ と $\neg k$ は点 tf やクラス k を除いた統計量を表す．また，inv$(\cdot)$ は逆行列である．$\hat{\nu}_{fd}, \hat{\mathbf{G}}_{fd}$ は十分統計量を用いて次のように与えられる．

$$\hat{\nu}_{fd} = \nu_{fd} + n_{fd}, \qquad \hat{\mathbf{G}}_{fd}^{-1} = \mathbf{G}_{fd}^{-1} + \sum_{t:w_{z_{tf}}=d} \lambda_{tf}\mathbf{x}_{tf}\mathbf{x}_{tf}^H, \qquad (7)$$

ここで，$\sum_{t:w_{z_{tf}}=d}\cdot$ は周波数ビン f で方向 d に割り当てられた時間周波数点に関する和である．

K を推論時に生成されているクラス数とする．z_{tf} が未生成のクラス $K+1$ を取る確率を式 (5) 中で考慮するため，β は $K+1$ 次元として操作する [Teh *et al.*, 2006]．$z_{tf} = K+1$ の確率計算のため，$w_{K+1} = d$ を一時的に $\frac{\kappa/D+c_d}{\kappa+\sum_d c_d}$ に従い生成する．もし $z_{tf} = K+1$ が選択された場合，$K \leftarrow K+1$ とし，β の次元も $\beta_K \leftarrow b\beta_K$，$\beta_{K+1} \leftarrow (1-b)\beta_K$ として増やす．ただし，b はベータ分布 $\mathcal{B}(1,\gamma)$ から生成する．

その他のパラメータ $\alpha,\beta,\gamma,\kappa$ の更新は [Teh *et al.*, 2006; Escobar and West, 1995] の詳述されるように更新する．これらの変数は補助変数を導入して確率的に生成される．

3.3 音源定位・分離結果出力

ξ_{tf}^d を推論時にサンプルされた z_{tf}, w_k のうち，$w_{z_{tf}} = d$ である割合，同様に，η_k^d をサンプルされた w_k の中で $w_k = d$ である割合とする．音源分離は，同一方向から到来するクラスを統合した TF マスクを用いる．方向 d から到来する多チャンネル信号を $\hat{\mathbf{x}}_{tf}^d$ とすると，$\hat{\mathbf{x}}_{tf}^d = \xi_{tf}^d\mathbf{x}_{tf}$ に従って分離が可能となる．また，各方向の事後重み $P_d = \sum_{tf}\xi_{tf}^d$ を計算することでどの方向に音源が存在するか推定することが出来る．もし混合音から N 個の音源を抽出したい場合，P_d の大きなものから N 方向を選び，音源を抽出することで，音源の定位と分離が達成される．

3.4 推論初期化

推論の初期化は [Otsuka *et al.*, 2012] に似た方法で行う．推論開始時のクラス数を K とする．まず，w_k を K 個の重複のない領域から一様分布に従って生成する．$p(w_k = d) = \mathcal{U}(\{d|\frac{k-1}{K}D \le d < \frac{k}{K}D\})$，$\mathcal{U}(A)$ は集合 A 上の一様分布である．次に，z_{tf} を初期化された w_k とステアリングベクトルから生成された $\mathbf{G}_{fd}$ を用いて生成する，$p(z_{tf}=k) \propto \exp\left(-\mathbf{x}_{tf}^H\mathbf{G}_{fw_k}\mathbf{x}_{tf}\right)$.

4 実験結果

本節では，分離実験に用いた混合音の収録条件と音源分離結果を示す．Figure 7 に収録環境における音源配置とロボットの移動軌跡，および，ロボットに搭載されたマイク

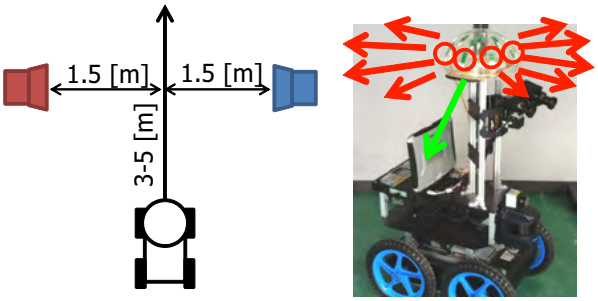


Figure 7: 音源配置とロボットの軌道 (左) とロボットに搭載されたマイクロホンアレイ (右)

ロホンアレイを示す．Figure 7 左のように，実験では 2 音源の間を直線的に移動しながら収録した．その際，片方の音源は常にロボットの左側，もう片方の音源はロボットの右側となるように動いた．これは，Figure 3 のように，様々な方向から分離された複数の分離音を同一の音源にまとめる処理を簡単化するためである．収録にはロボット上部に備え付けられた 8 チャネルマイクロホンアレイを用い，2 種類の環境で行った．1 つは屋外で，残響時間 $RT_{60} = 150$ [ms]，もう 1 つは屋内で，$RT_{60} = 800$ [ms] 以上の環境である．収録音は音源の種類に対する頑健性を示すため，Figure 7 中，青で示された右側スピーカからはピアノやギターから成る音楽音響信号を，赤で示された左側スピーカからは人間の音声や，鈴虫，カエルの鳴き声などを再生した．録音データの長さはおよそ 10 [s] であった．いずれの環境においても，多少の凹凸はあるがおおむね平坦な床面での走行を行い，ロボット本体の揺れを極力抑えた．

音源定位処理での方向の候補数 D の決定はマイクロホンアレイの持つ空間解像度などを考慮して行う．例えば，水平面上を 5° の解像度で定位を行う場合は，$D = \frac{360}{5} = 72$ と設定する．本実験では，ロボットの移動時に発生する車輪音を抑圧するために，上記の水平面上 72 方向 (Figure 7 右の赤矢印) のステアリングベクトルに加えて，ロボットの荷台方向 (Figure 7 右の緑矢印) のステアリングベクトルを用いて $D = 73$ とした．これにより，モータノイズなどはロボットの荷台方向として定位される音源に分離されることが期待できる．

5　実験結果

Figure 8，9 に，屋内，屋外それぞれの環境の観測音，分離音，再生された原音のスペクトログラムを示す．混合音と分離音に示された緑の枠は，その時間区間でロボットが移動していたことを示す．音源分離結果は Figure 3 のように同一音源でも様々な方向に分割された結果が得られるが，ロボットからみて左右どちらの方向に定位されているかに基いて各方向の音源を復元した．

ロボットの荷台方向に定位された車輪分離音について Figure 8，9 を比較すると，屋外環境については走行時以外に抽出された音はあまりないが，屋内環境については走行時以外も音声などが含まれている．さらに，屋内環境での車輪分離音の低周波領域では，右側分離音に含まれるべき成分を多く含んでいる．このように，残響の多い環境においては，直接音のみを対処しようとする音源分離手法の性能は特に低周波領域において劣化する．

Figure 8 での屋外環境での左右分離音は，左側音源の前半の虫の鳴き声は右側分離音や車輪分離音に埋もれてしまったが，ロボット移動中でもある程度音源分離が達成されている．虫の鳴き声の分離が特に困難な一因としては，この音源が比較的狭い周波数帯域のみにエネルギが集中しているため，他音源がこの領域でエネルギが大きかった場合に時間周波数マスクの推定が影響を受けるためと考えられる．一方，Figure 9 に示された左右分離音については，特にロボットが移動中の右側分離信号の抽出精度が劣化している．分離精度低下の要因としては残響の他，観測音中に含まれる右側の音楽音響信号の割合が少ない (SNR，signal to noise ratio が低い) ことが挙げられる．

以上のように，本手法には残響，狭帯域音，低 SNR 音などに伴う分離性能の劣化という限界はあるものの，(1) 異なる残響環境に対してパラメータなどの手動設定なしに，(2) ロボット自身の移動に伴って空間特性が時間変化する観測信号および，(3) 自己発生音の抑圧，を扱うことが可能であることが示された．

6　考察と今後の課題

本稿では，移動ロボットが備えるべき聴覚機能について，マイクロホンアレイが持つ空間フィルタリング機能の中で最も基本的な，音源定位・分離問題を扱った．移動ロボットを用いた音源分離実験を通じて，マイクロホンアレイを用いることで空間特性が事変である混合音の分離や，車輪音などの自己発生音の抑圧がある程度対処可能であることを示した．ただし，残響の大きな環境における分離性能低下が確認されたため，残響抑圧を取り入れた音源分離 [Yoshioka *et al.*, 2011] など，マイクロホンアレイの空間フィルタリング機能をさらに活用することが今後の課題の 1 つである．また，本手法は HARK [Nakadai *et al.*, 2010] などの環境に対するチューニングが必要であるが，短いターンアラウンド時間で高速処理可能な手法と違い，音環境の違いには頑健ながら処理に時間を要する手法である．したがって，ロボットなどへの応用のためには，これら 2 つの手法を状況に応じて効果的に使い分ける枠組みなども必要となる．

今回の音源分離実験では，分離対象の音源はロボットの左右に分かれるという仮定のもとで分離音の復元を行った．ロボットがより一般的な軌道で移動する場合はこのような

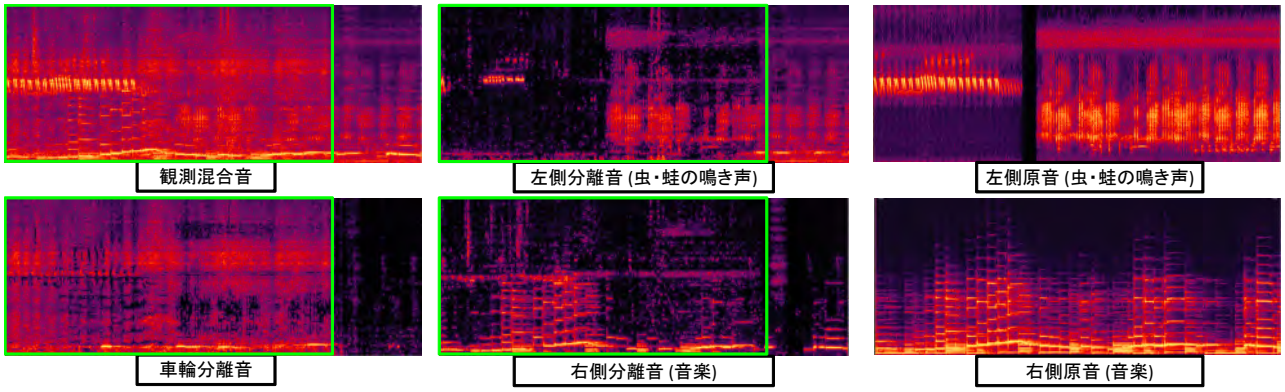


Figure 8: 屋外混合音分離結果: 残響時間 $RT_{60} = 150$ [ms]

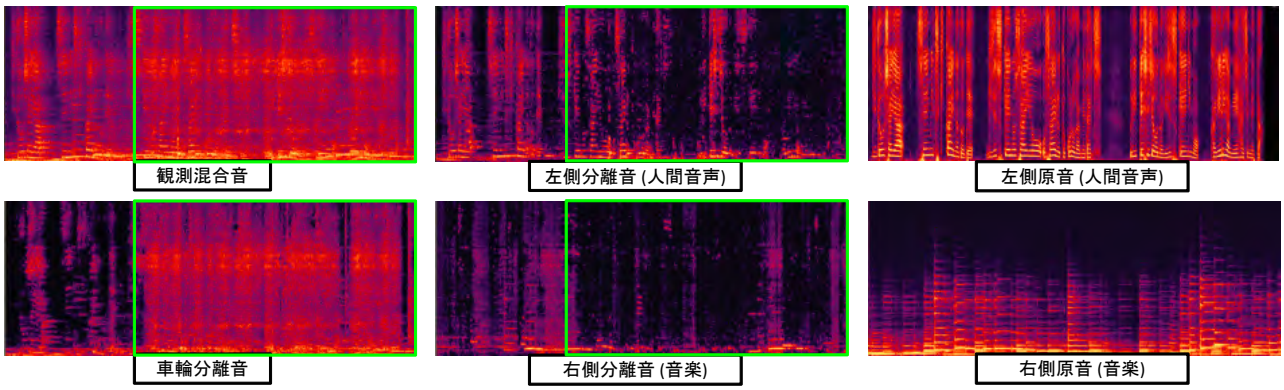


Figure 9: 屋内混合音分離結果: 残響時間 $RT_{60} = 800$ [ms]

方法は用いることが出来ないため，音源定位結果の時間連続性を考慮したトラッキング [Otsuka *et al.*, 2011] や，分離音の持つ音色などの特徴による同一音源の識別 [Sasaki *et al.*, 2009] を通じて，同一音源から発せられた分離音を集約する必要がある．状況をさらに一般化し，音源そのもののが移動する場合や，音が断続的に発せられ一時的に消失しうる場合では，これらの手法や視覚情報など異なるモダリティの統合など，マイクロホンアレイの枠組みを越えた手法が必要となる．

本稿でのロボット自己発生音の対処は，音源のマイクロホンアレイに対する相対位置は不変であることを仮定して行った．ヒューマノイドロボットなど，複雑な動作を行うロボットの自己発生音では音源位置の定常性の仮定が成り立たないこともありうる．解決策としては，自己発生音源の近くにマイクやセンサを設置し，マイクロホンアレイ処理に組み込んで抑圧する手法 [Sawada *et al.*, 2010] や，ロボットを動かすモータ指令値から自己発生音を予測し，抑圧する手法 [Ince *et al.*, 2011] などが挙げられる．

さらなる今後の展望としては，今回取り扱った音源定位・分離という汎用的ながら低次な問題から発展させ，分離結果を用いた複雑なタスクをこなすロボット (たとえば音を頼りにしたレスキューロボットや警備ロボット) などが考えられる．これらの高度なタスクを手がけるロボットを実現する要素技術の取捨選択や研究の加速には，データセットの整備も重要な今後の課題として挙げられる．

謝辞: 本研究の一部は科研費特別研究員奨励金/基盤 (S) の支援を受けた．

参考文献

[Benesty *et al.*,] J. Benesty, J. Chen, and Y. Huang. *Microphone Array Signal Processing*. Springer Topics in Signal Processing.

[Blei *et al.*, 2003] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[Conradsen *et al.*, 2003] K. Conradsen, A. A. Nielsen, J. Schou, and H. Skiriver. A Test Statistic in the Complex Wishart Distribution and Its Application to Change Detection in Polarimetric SAR Data. *IEEE Trans. on Geoscience and Remote Sensing*, 41(1):4–19, 2003.

[Duong *et al.*, 2010] N. Q. K. Duong, E. Vincent, and R. Gribonval. Under-Determined Reverberant Audio Source Separation Using a Full-Rank Spatial Covariance Model. *IEEE Trans. on ASLP*, 18(7):1830–1840, 2010.

[Escobar and West, 1995] M. Escobar and M. West. Bayesian Density Estimation and Inference Using Mixtures. *Journal of the American Statistical Association*, 90:577–588, 1995.

[Ince *et al.*, 2011] G. Ince, K. Nakamura, F. Asano, H. Nakajima, and K. Nakadai. Assessment of General Applicability of Ego Noise Estimation. In *Proc. of International Conference on Robotics and Automation*, pages 3517–3522, 2011.

[Nakadai *et al.*, 2010] K. Nakadai, T. Takahashi, H. G. Okuno, H. Nakajima, Y. Hasegawa, and H. Tsujino. Design and Implementation of Robot Audition System "HARK". *Advanced Robotics*, 24(5–6):739–761, 2010.

[Otsuka *et al.*, 2011] T. Otsuka, K. Nakadai, T. Ogata, and H. G. Okuno. Bayesian Extension of MUSIC for Sound Source Localization and Tracking. In *Proc. of International Conference on Spoken Language Processing*, pages 3109–3112, 2011.

[Otsuka *et al.*, 2012] T. Otsuka, K. Ishiguro, H. Sawada, and H. G. Okuno. Bayesian Unification of Sound Source Localization and Separation with Permutation Resolution. In *Proc. of AAAI Conf. on Artificial Intelligence*, pages 2038–2045, 2012.

[Sasaki *et al.*, 2009] Y. Sasaki, M. Kaneyoshi, S. Kagami, H. Mizoguchi, and T. Enomoto. Daily Sound Recognition Using Pitch-Cluster-Maps for Mobile Robot Audition. In *Proc. of International Conference on Intelligent Robots and Systems*, pages 2724–2729, 2009.

[Sawada *et al.*, 2010]
H. Sawada, J. Even, H. Saruwatari, K. Shikano, and T. Takatani. Improvement of Speech Recognition Performance for Spoken-Oriented Robot Dialog System using End-fire Array. In *Proc. of IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 970–975, 2010.

[Se *et al.*, 2005] S. Se, D. G. Lowe, and J. J. Little. Vision-Based Global Localization and Mapping for Mobile Robots. *IEEE Trans. on Robotics*, 21(3):364–375, 2005.

[Teh *et al.*, 2006] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet Processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

[Thrun *et al.*, 2004] S. Thrun, M. Montemerlo, D. Koller, B. Wegbreit, J. Nieto, and E. Nebot. FastSLAM: An Efficient Solution to the Simultaneous Localization and Mapping Problem with Unknown Data Association. *Journal of Machine Learning Research*, 2004.

[van den Bos, 1995] A. van den Bos. The Multivariate Complex Normal Distribution–A Generalization. *IEEE Trans. on Information Theory*, 41(2):537–539, 1995.

[Yoshioka *et al.*, 2011] T. Yoshioka, T. Nakatani, M. Miyoshi, and H. G. Okuno. Blind Separation and Dereverberation of Speech Mixtures by Joint Optimization. *IEEE Trans. on ASLP*, 19(1):69–84, 2011.

社団法人 人工知能学会
Japanese Society for
Artificial Intelligence

人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-B202-10

MULTI-MODAL SOUND LOCALIZATION FROM A MOBILE PLATFORM

Jani Even, Nagasrikanth Kallakuri, Yoichi Morales, Carlos Ishi, Norihiro Hagita

ATR Intelligent Robotics and Communication Laboratories, Kyoto, Japan
even@atr.jp

ABSTRACT

This paper presents a multi-modal sound source localization method for mobile platforms. The sound source localization is performed while the robot is autonomously navigating through the environment by combining the power and bearing estimation given by a steered response power (SRP) algorithm with the range estimation obtained from the on-board laser range finders (LRF). First the positions of the sound sources in the environment are determined by taking into account the robot pose which is estimated with a particle filter and the estimated power is accumulated in the cells of a grid map covering the environment. Finally, a local maxima search is performed on this grid map to find the area with higher estimated sound power that correspond to the sound source locations.

1. INTRODUCTION

Sound source localization has been a topic of interest in the audio processing community for a long time (see [1]). The most effective techniques that emerged are either based on the estimation of the time delay of arrival at microphone pairs, or on the estimation of a steered response power (SRP) or on spectral decomposition techniques like the MUSIC algorithm. All these approaches rely on the use of microphone arrays. Using a robot, it is possible to explore the space and effectively extend the operational range of sound localization. Thus a natural framework for sound source localization from a robot is to use a conventional sound localization algorithm at different locations and combine the results from all these different locations [2, 3, 4, 5]. Since these localization results are obtained for different times, it is important to distinguish between fixed sound sources and moving sound sources. In this paper, we are interested in the localization of the environmental noises that are fixed sources.

The authors in [5] rely on triangulation to estimate the positions of the sound sources using audio scans taken by an autonomous mobile robot. One of the very interesting approaches in this area is the use of evidence grids in [3, 6]. The space to be explored is partitioned into grid cells of fixed size. Then the probabilities of having a sound source in each of the cells are estimated during the exploration. To achieve this, at a given location, an SRP with phase transform (SRP-PHAT [7]) is estimated for a grid centered on the robot and these estimated powers are used to update the evidence grid. In that method, the robot is tele-operated to gather sound data in the vicinity of the sound sources [6].

In this paper, we present a framework for localizing sound sources using an autonomous mobile robot equipped with a microphone array. The novelty of the present work is in obtaining the audio information about the environment using a multi-modal approach. In particular, the laser range finder (LRF) data are explicitly used during sound source localization. Most autonomous mobile robots are equipped with LRFs and odometry (obtained from the encoders on the wheels) to localize themselves in the environment. This ability to estimate precisely the range of the objects around the robot is exploited in this paper to solve the problem of poor range estimation from audio localization techniques. In the proposed approach, the audio modality is used to estimate the bearing of the sound sources whereas their ranges are obtained using laser range finders. Consequently, our approach assumes that the geometric coordinates of sound sources are detectable by the LRFs on the robot. While this assumption is a bit restrictive when a two dimensional horizontal plane is scanned, it will be reasonable when extended to a three dimensional scan. In the proposed framework, sound source localization is performed while the robot is autonomously navigating through the environment. During navigation, the two on-board LRFs (front and back) provide range scans and a steered response power (SRP) algorithm generates audio scans. The SRP gives the bearing of the candidate sound sources and an estimate of the received audio power. Combining the bearing, the received power and the range information, an estimate of the emitted power from candidate sound sources is computed. The audio and LRF scans are acquired at regular intervals and for each of the audio scan, the power of the most powerful emitting candidate sound source is accumulated on a grid map that covers the environment. The cells from that grid map contains an average of the estimated emitted power and a count of the number of visits (the number of time a cell has been selected). This procedure requires to transform the sound source positions from the

THIS RESEARCH WAS FUNDED BY THE MINISTRY OF INTERNAL AFFAIRS AND COMMUNICATIONS OF JAPAN UNDER THE STRATEGIC INFORMATION AND COMMUNICATIONS R&D PROMOTION PROGRAMME (SCOPE).

robot referential to the room referential. This is performed by taking into account the robot pose which is estimated with a particle filter. Finally, a local maxima search on this grid map finds the locations of the sound sources in the environment by selecting the cells with higher power.

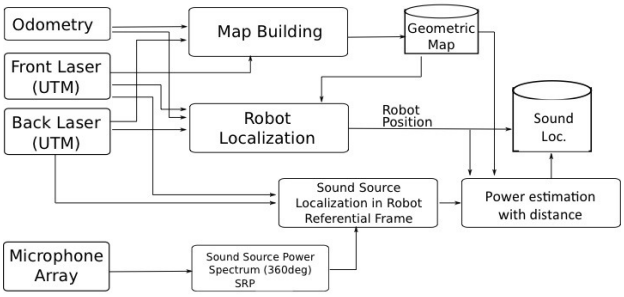


Fig. 1. Block diagram of the system.

2. PROPOSED APPROACH

A block diagram describing the proposed approach is shown in Fig. 1. The main processing blocks will be described in the following sections.

2.1. Map Building

The map building is performed in advance. The aim is to create a map that describes the environment in which the robot is due to navigate autonomously. The environment is represented by an occupancy grid, namely a grid which cells are either empty (open space) or occupied (walls and structures). The occupancy grid map is referred to as the *geometric map* in the remainder.

In this work we use them for building grid maps. To build the map, we controlled the robot with a joystick through the environment gathering odometry and laser sensor information. Then we used iterative closest point based SLAM to correct the trajectory of the robot and to align the laser sensor scans [8] using the $3DToolkit$ library framework [9], [10]. With the resulting aligned scans the occupancy grid map was created [11], [12]. The map obtained for one of the test environments is shown in Fig. 2.

2.2. Robot Localization

The goal of the robot localization task is to precisely estimate the pose (location and orientation) of the robot in the geometric map representing the environment. We used a particle filter approach to estimate the robot position with a weighted set of M particles. Each particle has a pose given by the state vector $\{x_m(k), y_m(k), \theta_m(k), w_m(k)\}$ containing a candidate position and orientation of the robot and the associated weight. While the robot moves, each particle also moves based on the odometry readings and the probabilistic motion model, which describes the uncertainty in the robot motion (prediction step). In the correction step, the particle filter estimates the posterior density by considering measurement likelihood. This likelihood is estimated from the LRF scans using the ray casting approach likelihood model in [13]. The map update depends on the state of the particle dispersion and the matching of the laser scans. The particles, which are more likely to be correct after ray casting, have a higher likelihood score, and therefore, more weight. Particle re-sampling is performed regularly and the robot pose $\{x(k), y(k), \theta(k)\}$ is given by the average weight of the particles.

2.3. Audio scanning

In order to present this framework in greater detail, let us first briefly describe the SRP approach to sound source localization (see [7] and references herein). The goal of sound source localization is to estimate the position of sound sources in a search space using the audio observation. At the sampled time k, the observed signals from the Q microphones of the array are $v_1(k), \cdots, v_Q(k)$. Because the geometry of the microphone array is precisely known, it is possible to *focus* the array using spatial beam forming to estimate the sound from a spatial location. The beam forming output is denoted by

$$s(k, [x, y, z]) = \mathcal{F}(v_1(k), \cdots, v_Q(k), [x, y, z]), \quad (1)$$

where $[x, y, z]$ are the coordinates of the focus point in the array referential. The SRP is obtained by computing the power of this output over T samples

$$J(k, [x_n, y_n, z_n]) = \frac{1}{T} \sum_{\tau=0}^{T-1} s^2(k - \tau, [x_n, y_n, z_n]) \quad (2)$$

for a set of N locations $[x_n, y_n, z_n]_{n \in [1,N]}$ in the search space. The locations corresponding to the peaks of the SRP gives the sound sources' positions. There are several ways to obtain the beam forming output, compute the power and select the set of locations. In the remainder, the SRP obtained at the time k that contains the power from the N locations is referred to as the kth audio scan.

In this paper, the SRP processing is done in the frequency domain after applying a short time Fourier transform (STFT) to the observed signals sampled at 48kHz (the analysis window is 25 ms long and the shift of the window is 10 ms). Then the SRP is computed for the frequency band $[1000, 6000]$ Hz using 10 STFT frames for averaging the power. Thus a new audio scan is available every 100 ms. A delay and sum beamformer is used to *focus* the observations.

A particularity of sound source localization algorithm is that the estimation of a source range is imprecise whereas its bearing estimate is accurate. Thus spherical coordinates $[\rho, \theta, \phi]$ are often used to describe the search space. Contrary to [3], we assume the far field conditions hold (ρ large compared to array aperture) and a bearing only scan is performed. Namely the kth scan is a set of N angles $\theta_n(k) \in [0, 2\pi]$ with their associated power $J_n(k)$. In our approach the distances are obtained using the LRF scans as explained in Sect. 2.4.

Note that these scans are computed in search spaces in the array's frame of reference (as the position of the focus point have to be known in the array's frame of reference). Thus it is necessary to transform the poses of the robot at these locations to a global coordinate system to localize the sound sources in the global referential.

2.4. Emitted power estimation

The audio scans $\{\theta_n(k), J_n(k)\}$ are in the robot coordinate frame and the goal of the fusion procedure is to combine them with the range estimation from the LRFs and the knowledge about the robot pose in the global referential in order to estimate the position of the sound sources in the geometric map.

The main idea is to use the range estimation in the directions given by the SRP and combine it with the estimate of the received powers to estimate the powers that was emitted by the potential sound source candidates. For this purpose, the phase transform is not used as it discards the amplitude of the signals of interest.

For each of the directions $\theta_n(k)$, an estimated range $\rho_n(k)$ is given by the LRFs (the closest ray in the LRF scans is selected). Then, for estimated ranges in $[d_{\text{min}}, d_{\text{max}}]$, the estimated emitted power is

$$C_n(k) = J_n(k) \left(\frac{\rho_n(k)}{d_{\text{min}}} \right)^{\alpha} \tag{3}$$

where α controls the effect of the distance on the power (in free field $\alpha = 2$). Namely, the received power is corrected by the estimated distance to the sound source candidate in order to compensate for the power drop during propagation between the source and the array, see the circles representing the propagation in Fig. 3. A maximum distance d_{max} is set because the audio power decreases rapidly with the distance and sound sources are covered by the background noise for long distances.

For each of the audio scan, only the largest emitted power estimate $C_m(k) = \max_n C_n(k)$, obtained for $\{\theta_m(k), \rho_m(k)\}$, is considered. By combining the robot pose with the maximum power location $\{\theta_m(k), \rho_m(k)\}$ a position in the global referential is obtained. That position correspond to a cell $\{i, j\}$ of a grid map covering the room. This transform is illustrated in Fig. 4.

The average estimated power $P_{ij}(k)$ of that cell is obtained by taking

$$P_{ij}(k) = \frac{P_{ij}(k-1)K_{ij}(k-1) + C_m(k)}{K_{ij}(k-1) + 1} \tag{4}$$

$$K_{ij}(k) = K_{ij}(k-1) + 1 \tag{5}$$

where $K_{ij}(k)$ denotes the number of time for which the cell $\{i, j\}$ is visited ($K_{ij}(0) = 0$). This count is also used to remove cells that have been seen very few times. The grid map containing the average power is referred to as *power map* in the remainder.

Then the sound source localization is performed by finding the cells that have higher power in the power map. Namely, a local maxima search algorithm is used on the power map to find the candidate sound sources.

In practice, a small neighborhood of the cell $\{i, j\}$ is selected and the power $C_m(k)$ is distributed in that neighborhood. The size of the neighborhood is taken as $\Delta\theta\rho_n(k)$ where $\Delta\theta$ is the angular resolution of the audio scan. The N cells in this neighborhood receive the power $\frac{C_m(k)}{N}$ and are counted as visited one time. This smearing of the power is performed in order to take into account the larger uncertainty for longer ranges.

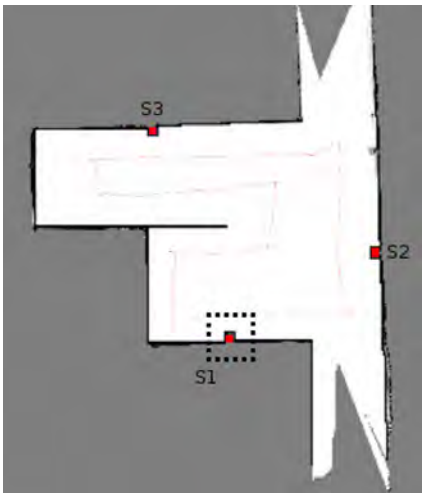


Fig. 2. Geometric map of the corridor with sound sources and robot trajectory.

3. EXPERIMENTS

For experimental validation we used a pioneer robot. This robot has a differential drive configuration and was equipped with two motor encoders and two laser range finder (UTM-30LX from Hokuyo, maximal range 30 m). The experimental platform can be observed in Fig. 5.

The microphone array is composed of 16 Sony ECM-C10 microphones mounted on a circular frame (diameter 31 cm). The audio capture interface is a Tokyo Electron Device

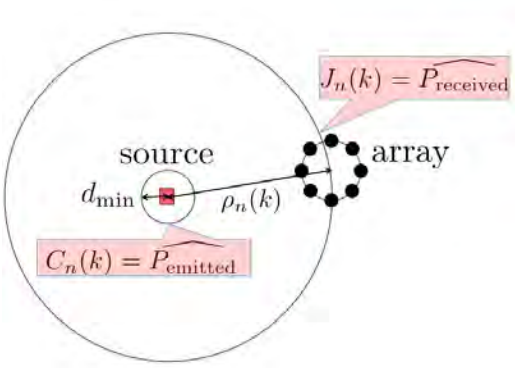


Fig. 3. Received and emitted powers.

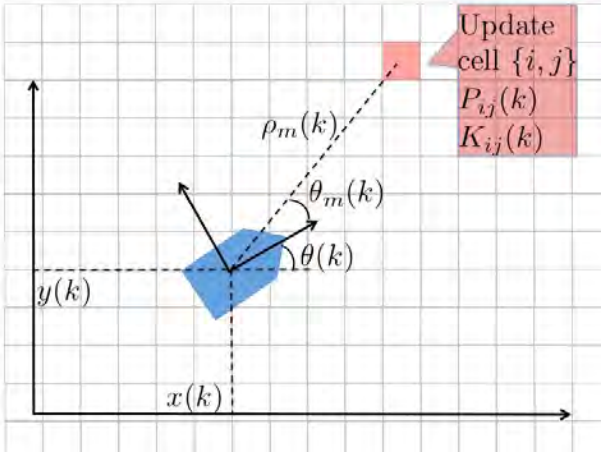


Fig. 4. Cell update using robot referential coordinates and robot pose.

Limited TDBD16AD-USB that samples the signals at 48kHz. The experimental evaluation of the approach was conducted in a corridor. Different sound sources of known intensities were setup in the environment.

Fig. 2 depicts the geometric map of the environment. The dimension of the cells in this map is 5 cm x 5 cm. The grid map used for localization has also 5 cm x 5 cm cells.

The robot navigates autonomously in the corridor using a set of way points that defined a loop covering all parts of the corridor In the remainder, a *run* corresponds to the robot performing one loop in the corridor. The sound source localization is performed during these runs.

Several (up to three) sound sources were placed in the environment (these locations are in the scan plane of the LRFs). These sound sources are loudspeakers playing recorded sounds. There was the sound of an air conditioning unit ($S1$ with a sound pressure of 78.5 dBA measured at 5 cm), the sound of a desktop computer fan ($S2$ at 77.5 dBA) and the sound of a server rack ($S3$ at 77 dBA). The sound pressure in the quiet corridor was around 42 dBA. The activation pattern

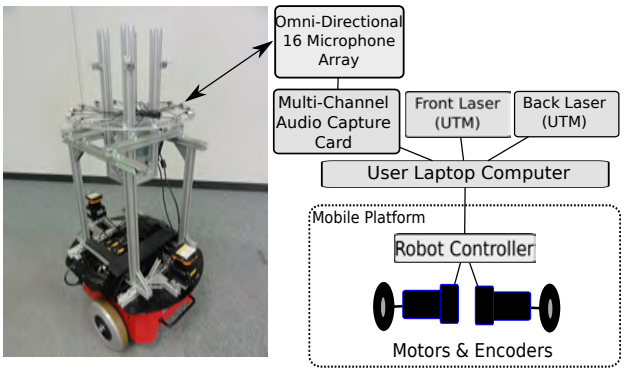


Fig. 5. Experimental robot platform with omni-directional microphone array and two laser range sensors.

Table 1. Parameters used during all runs.

$\Delta\theta$	α	$d_{\min}$	$d_{\max}$
3^o	0.5	0.3m	3m

of the sound sources for the runs can be observed in Table 2 and their positions are reference in Fig. 2. The parameters are given in Table 1. Note that α is set to a small value in order to avoid far estimate concentrating the power.

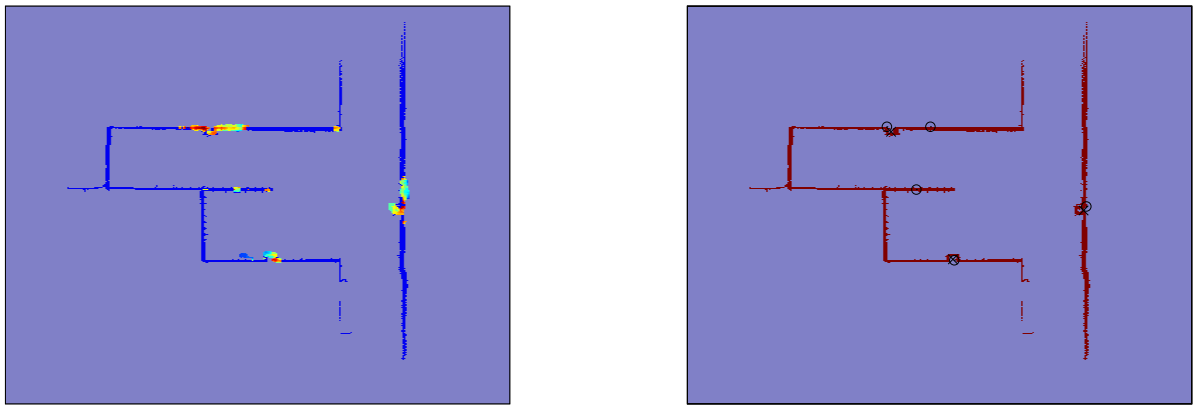
4. RESULTS

The power map is obtained by taking the accumulated power of the cells $P_{ij}(k)$ for which the number of visits $K_{ij}(k)$ is greater than 10% of the maximal number of visits. Thus an updated power map is available after each audio scan. Fig. 6-(a) shows the power map obtained at the end of the run 1 and Fig. 6-(b) shows the result of the local maxima search. The locations of the local maxima appear as black circles. The ground truth, i.e. the real positions of the sound sources are given as black crosses. For each of the sources the errors are given in Table 2. The results for run 2 are also in the table.

Figs. 7-(a) presents the power map difference (in dB) that is obtained by taking the difference of the power map for run 1 (three sound sources) and run 3 (no sound source). Figs. 7-(b) shows the same results for run 2 (two sound sources) and run 3 (no sound source).

5. DISCUSSION

The power map in Fig. 6-(a) illustrates the fact that large areas of higher power appear around the locations of the sources. Note that a few small areas with high power are also present in the power map. After local maxima search, the proposed approach successfully estimated the positions of the sound sources see Fig. 6-(b) (the three local maxima close to the true sources' location are the one with higher



(a) (b)

Fig. 6. (a) Power map for the run 1 , (b) local maxima search results (x) are the ground truth and (o) are the local maxima.

Table 2. Sound source localization results.

Run	Active Sources	Detected Sources	Error(m)
1	$S1$	$S1$	0.29
	$S2$	$S2$	0.22
	$S3$	$S3$	0.07
2	$S1$	$S1$	0.11
	$S3$	$S3$	0.18
3			

values). The average localization error was 0.17 m and the maximum error 0.29 m. Considering that the loudspeakers are not point sources but may span several cells and the localization was performed while moving, an error in the obtained range indicates a precise localization.

Another interesting point of the proposed approach is that the power maps contain estimates of the emitted power (these estimates are obtained by correcting the SRP without phase transform with a function of the estimated range, see Eq.3). Consequently, it makes sense to estimate sound source localization by using difference of power map in dB (equivalent to a power ratio). A *background power map* obtained when there is no sound source of interest (here the run 3) can be subtracted to a power map obtained when some sources of interest are present. In Figs. 7-(a) and (b), the difference of power maps clearly show the locations of the active sound sources. The local maxima search proved to be more easily conducted on the difference of power maps as they have larger dynamics and contains less false alarms (spurious local maxima). When it is not possible to obtain a good *background power map*, the local maxima search is to be applied on the power map.

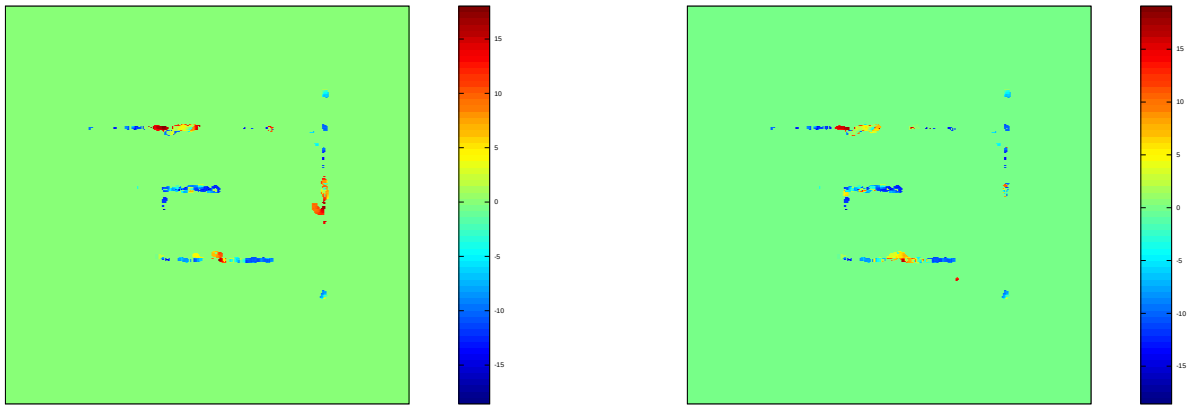
6. CONCLUSIONS

This paper presented a framework for localizing environmental sound sources using an autonomous mobile robot equipped with encoders, laser sensors and a 16 channel microphone array. The sound source localization results obtained in the experiments had an average distance error of 0.17 m using local maxima search, showing that the proposed framework is capable of localizing sound sources. Up to 3 sources within an 8m x 8m space were localized. The method is also robust towards false positive detections and noise effects produced by echoes. The novelty of the approach is in the combination of the audio scans with the LRFs scans. These kind of sound localization approach will aid the robot in attaining a better knowledge about environmental noise. It can be used for better speech recognition (suppressing the known environmental noise), effective human-robot interaction, and also for surveillance of environments. With the available framework, it is possible to extend the work to 3-Dimensional sound source localization that is more informative.

7. REFERENCES

[1] H. DiBiase, J. nad Silverman and M. Brandstein, *Microphone arrays : Signal Processing Techniques and Applications*, Springer-Verlag, 2007.

[2] J.-M. Valin, J. Rouat, and F. Michaud, "Enhanced robot audition based on microphone array source separation with post-filter," in *Intelligent Robots and Systems, 2004. (IROS 2004). Proceedings. 2004 IEEE/RSJ International Conference on*, sept.-2 oct. 2004, vol. 3, pp. 2123 – 2128 vol.3.



(a) (b)

Fig. 7. (a) Difference of power maps for run 1 and run 3 (scale in dB) , (b) Difference of power maps for run 2 and run 3 (scale in dB).

[3] Eric Martinson and Alan C. Schultz, "Auditory evidence grids.," in *IROS*. 2006, pp. 1139–1144, IEEE.

[4] K. Nakadai, H.G. Okuno, H. Nakajima, Y. Hasegawa, and H. Tsujino, "An open source software system for robot audition hark and its evalation," in *IEEE-RAS International Conference on Humanoid Robots*, 2008, pp. 561–566.

[5] Y. Sasaki, S. Thompson, M. Kaneyoshi, and S. Kagami, "Map-generation and identification of multiple sound sources from robot in motion," in *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2010*, 2010, pp. 437–443.

[6] Eric Martinson and Alan C. Schultz, "Robotic discovery of the auditory scene," in *ICRA*, 2007, pp. 435–440.

[7] M.S. Brandstein and H.F. Silverman, "A robust method for speech signal time-delay estimation in reverberant rooms," in *IEEE Conference on Acoustics, Speech, and Signal Processing, ICASSP 1997*, 1997, pp. 375–378.

[8] P.J. Besl and H.D. McKay, "A method for registration of 3-d shapes," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 14, no. 2, pp. 239–256, Feb. 1992.

[9] D. Borrmann, J. Elseberg, K. Lingemann, Andreas Nüchter, and J. Hertzberg, "The Efficient Extension of Globally Consistent Scan Matching to 6 DoF," in *Proceedings of the 4th International Symposium on 3D Data Processing, Visualization and Transmission (3DPVT '08)*, Atlanta, USA, June 2008, pp. 29–36.

[10] slam6d, "Slam6d - simultaneous localization and mapping with 6 dof," Retrieved December May, 20 2011 from `http://www.openslam.org/slam6d.html`, 2011.

[11] Hans Moravec and A. E. Elfes, "High resolution maps from wide angle sonar," in *Proceedings of the 1985 IEEE International Conference on Robotics and Automation*, March 1985, pp. 116–121.

[12] A. Elfes, "Using occupancy grids for mobile robot perception and navigation," *Computer*, vol. 22, no. 6, pp. 46–57, June 1989.

[13] Sebastian Thrun, Wolfram Burgard, and Dieter Fox, *Probabilistic Robotics (Intelligent Robotics and Autonomous Agents)*, The MIT Press, 2005.

人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-B202-11

複数のマイクロホンアレイおよび空間情報と反射音を利用した音源定位の検討

Investigation on sound localization using multiple microphone arrays, reflection and spatial information

○石井カルロス寿憲（ＡＴＲ知能ロボティクス研究所）
Jani Even（ＡＴＲ知能ロボティクス研究所）
萩田紀博（ＡＴＲ知能ロボティクス研究所）

* Carlos Toshinori ISHI, Jani EVEN, Norihiro HAGITA (Intelligent Robotics and Communication Labs., ATR)

carlos@atr.jp, even@atr.jp, hagita@atr.jp

Abstract -複数のマイクロホンアレイにおいて音源方向推定を行い、空間の情報と反射音の方向の情報を利用し音源定位（3次元空間の位置推定）に利用する枠組みを提案し、人の声とスピーカから再生した音声を音源とした評価実験をおこなった。マイクロホンアレイの位置、壁の位置、音源の種類、音源の位置と向きに応じて、観測される直接音や反射音が変化し、反射音が重要となる条件を分析した。

1　はじめに

家庭, オフィス, 商店街など, 異なった環境では, 場所や時間によって多様な雑音特性を持つため，音声などの特定の音を対象としたアプリケーションでは，使用される環境の雑音の種類や度合いにより，期待した性能が得られないという問題がある．

本研究では，音環境の事前知識の習得およびその利用を総称して「音環境知能」と呼ぶ．また，のことを「音環境地図」と呼ぶ．実環境では，異なった場所で発生する複数の音が混合して観測されるため，音環境地図の生成において，騒音計で空間をスキャンするような従来の単純な方法は不十分である．音環境の事前知識として役立つと考えられる音源の位置や種類を特徴付けた音環境地図の生成には，空間的情報（通常の地図）に加え，音源の定位，分離及び分類が必要となる．そこで，本研究では，複数の音源を定位するため，複数のマイクロホンアレイを連携させ，空間内の特定の音源に対する音環境地図を生成し，音環境を構造化することを目的としている．本論文では、この目的を達成するための第一ステップとして、複数のマイクロホンアレイによる音源位置推定の問題に焦点を当てる。

マイクロホンアレイ処理における一つの問題として、アレイの周りに壁や天井やガラス窓やディスプレイなどの音を反射する表面が存在する場合、音源の直接音と同時に音源の反射音も観測されることがある。我々はマイクロホンアレイを天井に取り付けて集音を試みているが、特に音源との距離が大きい場合、強い反射音も頻繁に観測している。

これまでの音源定位や音源分離に関するほとんどの研究[1 ~ 10]では、反射音は悪影響を与えるものとして扱われてきたが、本研究では、反射音を利用して、音源位置推定に役立てる枠組みを提案し、その効果を評価した。

本論文は以下のように構成される。次ぐ2章では、提案手法を説明する。3章では、データ収集と提案手法による音源位置推定における分析結果を述べる。4章でまとめと今後の課題を記す。

2　提案手法

提案手法では、複数のマイクロホンアレイを用いて複数の音源方向を推定し、空間の情報を用いて反射音の方向を推定し、これらの情報を統合して音源定位（3次元空間の位置推定）を行う。音源方向推定においては、先行研究で提案した手法を採用し、2.1節で述べる。空間情報と反射音を利用した音源定位の提案手法は2.2節で述べる。

2.1 MUSIC スペクトル

MUSIC（Multiple Signal Classification）とは、音源定位において分解能が高い特徴を持つ手法の一種である。Fig. 1 にMUSICスペクトルの推定法のブロック図を示す。まずフーリエ変換(FFT)により多チャンネルのスペクトル X(k,t) をフレーム毎に求め、スペクトル領域でチャンネル間の空間的相関行列 R_k をブロック毎に求め、相関行列の固有値分解により指向性の成分と無指向性の成分のサブ空間を分解し、無指向性のサブ空間に対応する固有ベクトル

E_k^n と、対象の検索空間に応じて予め用意した方向ベクトル a_k を用いて（狭帯域の）MUSICスペクトルP(k)を周波数ビンごとに求め、特定の周波数帯域内の周波数ビン毎のMUSICスペクトルを統合して広帯域MUSICスペクトルが求まる。アルゴリズムの詳細は付録に記載している。

ここでは、広帯域MUSICスペクトルを単に「MUSICスペクトル」と呼び、MUSICスペクトルの時系列を「MUSICスペクトログラム」を呼ぶ。

音源定位においては、MUSICスペクトルのピークを探索することにより、音源の方向が求まる。

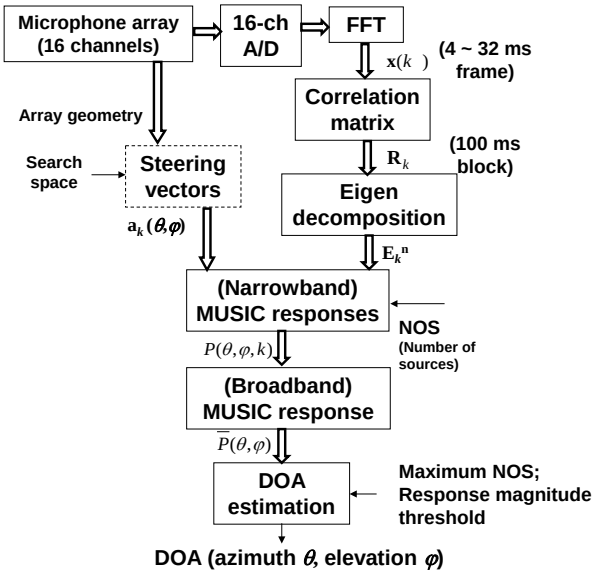


Fig. 1. The MUSIC-based sound localization algorithm, and related parameters.

ただし、MUSIC法を用いた音源定位の実用化においては、主に2つの問題が挙げられる。一つ目は、チャンネルの数および探索空間が大きくなるにつれて、処理時間が重くなり、通常のパソコンでは、実時間処理が追いつかないことである。もう一つは、MUSICスペクトルを求めるには、音源数を予め与える必要があることである。

著者らの先行研究[10]で、実時間処理を可能にするため、MUSICスペクトルの推定においていくつかのパラメータを分析した。その結果、FFTのフレーム長を64～128点（4～8msに対応）、ブロック長を100msに設定することにより、2GHzのCore2DuoのCPUを用いて、音源方向推定の精度を保ちつつ、実時間処理が可能であることを示した。

狭帯域MUSICスペクトルの推定において、その時刻に発している指向性を持つ音源数（NOS）を与える必要があるが、音源数の推定は難しいため、先行研究[10]で提案した通り、固定数を与え、MUSICスペクトル上で、特定の閾値を超えたピークのみを指向性のある音源とみなす方法を用いる。

2.2　空間情報と反射音を利用した複数アレイによる音源定位

本節では、複数のアレイにおいて、2.1節で説明したMUSICスペクトルによる音源方向推定を行い、空間情報とアレイの位置情報を用いて反射音の方向も推定し、これらの情報を統合して複数の音源位置の推定を行う手法を説明する。概要図をFig. 2に示す。

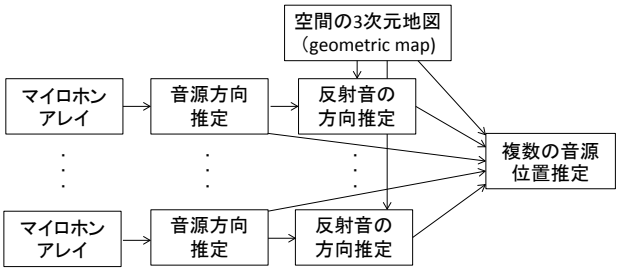


Fig. 2. The proposed sound localization system using multiple arrays and sound reflections.

複数のマイクロホンアレイを用いて音源方向を推定し、空間内のアレイの位置と向きが既知である場合、それぞれのアレイで推定された音源方向が重なった位置に音源が存在する確率が高いというのが本手法の基本的な概念である。

また、空間内のアレイの位置、音が反射しやすい天井や壁やディスプレイなどとの位置関係によって、アレイで反射音が測定される場合があり、一つのアレイでも反射音と直接音の方向が検出された場合、反射音を壁や天井で反転させた方向と直接音が重なった位置に音源が存在する確率が高いと予想される。従来のマイクロホンアレイ処理では、反射音は音源定位や音源分離に悪影響を与えるが、本手法では、逆に反射音の情報を利用することとなる。

定位された音源が反射音であるか否かは予め分からないため、まず推定されたすべての音源方向を壁や天井で反転させる。反射は空間内で複数生じ得るが、本研究では、2度目以降の反射は強度も指向性も衰える可能性があるため、反転は1度のみ行うこととする。

また3次元空間を考慮し、方位角および仰角で音源方向を表現する。

推定された方向には、角度の誤差（Angle uncertainty: AU）があり、アレイからの距離に応じて推定位置の誤差（Position uncertainty: PU）が大きくなる。幾何学的に、推定位置誤差を以下の式で求めることができる。

$$PU(d) = \pm AU / 360 * 2\pi * d \quad (1)$$

d はアレイの中心からの距離で、AUは推定角度の誤差を度単位で表したものである。例えば、球面上で5度の分解能で音源方向が検知された場合（AU = 5）、アレイから1メートル離れた位置に音源がある場合

（d = 1m）、その方向に直線を1メートル伸ばした際の推定位置誤差は ±8.7 cm となる。2メートルの場合、誤差はその倍の ±17.4 cm となる。

検知された2つの方向が上述の誤差を考慮して空間上で重なっているか否かを判定する方法として、それぞれの方向に直線を引き、2直線の最短距離を幾何学の公式を用いて推定する。この最短距離がそれぞれの直線における誤差（PU）を足した値よりも小さい場合、これらの直線は重なっていると判定する。また、検出された方向の重なりが生じた位置に音源が存在する可能性が高いとみなす。

検出されたすべての直接音と反射音の方向に引いた直線の距離をペア毎に求め、方向の重なりを複数探索する。重なりがあった場合は、平均位置を音源の推定位置とする。重なりがない場合は方向情報を保留とし、重なりが生じた時点で、位置を割り当てる。

3　データ収集および分析結果

3.1 マイクロホンアレイと音源方向推定の設定

本実験に用いた 16 素子のマイクロホンアレイの形状を Fig. 3 に示す。3 次元空間における方位角および仰角を求めるため、マイクは直径 30cm の半球面上に Fig. 3 に示すように配置した。

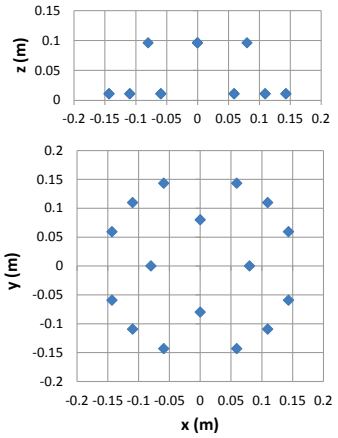


Fig. 3 The geometry of the 16-element microphone array.

多チャンネルオーディオキャプチャデバイスとして、東京エレクトロンデバイス社の16-channel A/D 変換機 TD-BD-16ADUSB を使用した。マイクは、Sony の ECM-C10 を用い、16 kHz/16 bitsでサンプリングを行った。

MUSICスペクトルによる音源方向推定のパラメータとして、音源の固定数を3、MUSICパワーの閾値を2.5dB、同時に発する音源の最大数を6 に設定した。

また、MUSICスペクトルを求める際に用いる周波数帯域に関しては、空間的歪み（spatial aliasing）と低周波数帯域における低い分解能を避けるため、1000 ~ 5000 Hzの帯域を用いた。

音源方向推定の探索空間は、3 次元空間で球面上 5 度間隔の分解能に設定し、アレイを天井に取り付けるため、方位角は 0 ～ 360 度、仰角は-5 度～-80 度に制限した。-85 ~ -90 度（アレイの真上の方向）には、アレイの形状より音源が存在しない場合にも MUSIC スペクトルにピークが生じるため、その領域を探索空間から除外している。これは使用したキャプチャの特性により、すべてのチャンネルで同位相の雑音が生じるためである。

3.2 評価データの収集

本実験では、Fig. 4 に示すように、2つのアレイを天井に取り付けた。アレイと天井の間には吸音素材を入れ込み、天井での反射は扱わないこととした。また床は反射しにくいタイルカーペットであり、反射が生じたとしても天井に設置したアレイへの距離が大きいため、床での反射も扱わないこととした。従って本研究では、推定された音源方向を壁で一回のみ反転させることとした。

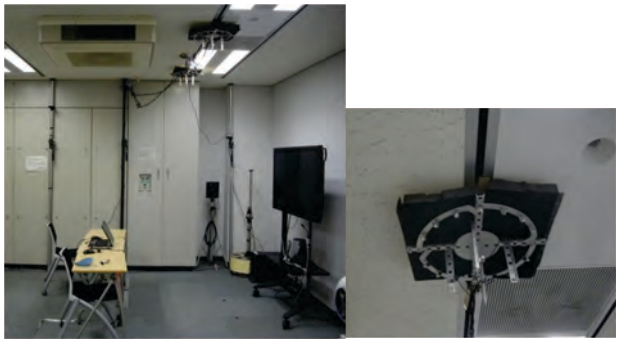


Fig. 4 The microphone arrays attached in the ceiling.

音源の向きによって、その音源の指向性が変化し、同じ位置でもアレイに対する向きによってアレイで観測される指向性の強度が変化することが考えられる。また、音源の種類によっても、指向性が異なることが予想され、本研究では、人が発した音声と、音声をスピーカから流した場合の2つの種類を対象音源とした（これらの音源をこれ以降それぞれ “Human” および “Loudspeaker” と呼ぶ）。また、環境に固定されたエアコン（Fig. 4 の左上）もアレイに対して指向性を持つ雑音源となる。

対象音源の位置として、Fig. 4の机の周り6か所を固定し、各位置において、前後左右の4つの向きでデータを収録した。エアコンはスイッチオンの状態にした。正確に音源の位置を固定することは難しいが、向きを変えた際に、口の位置ができるだけ変わらないようにした。

スピーカとして、ONKYOのGX-77Mを用いた。スピーカの高さは、話者が椅子に座った時の口の高さと一致するようにした。話者には各位置および各向きで同じ文を同じような発話スタイルで発声する

よう指示した。スピーカからは同じ話者の声を再生した。スピーカの音量は人の声の強度に近づけるよう調整した。表 1 に、設定した音源の位置とマイクロホンアレイの位置を記す。Fig. 5 に音源の位置および向きとアレイの位置を部屋の上面図に重ねて示す。x = 0 および y = 0 の平面には壁が存在する。x = 7400 mm および y = -5600 mm にも壁が存在するが、アレイから離れているため、本実験の反射音推定には用いなかった。

表 1．対象音源の位置およびマイクロホンアレイの位置情報

	1	2	3	4	5	6
x (mm)	1000	2000	3000	3000	2000	1000
y (mm)	-2700	-2700	-2700	-1200	-1200	-1200
z (mm)	1160	1160	1160	1160	1160	1160

	array1	array2
x (mm)	1410	3560
y (mm)	-1430	-1430
z (mm)	2630	2630

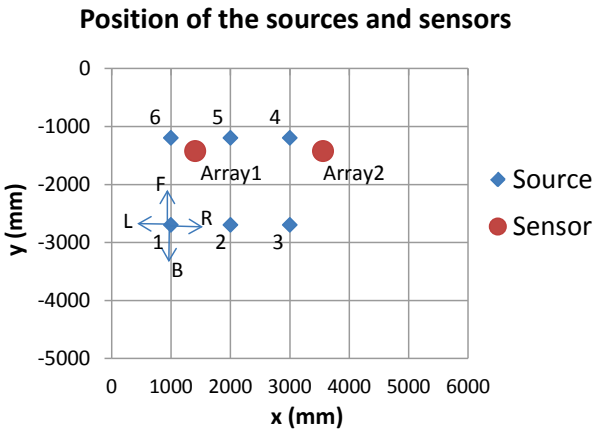


Fig. 5 Position (1 ~ 6) and orientation (F: front, L: left, B: back, R: right) of the target sources and the microphone array sensors (Array1, Array2) in the room.

3.3 音源の種類およびアレイに対する音源の向きの影響

本実験は、それぞれのアレイで測定された音源方向と反射音が実際発した音源位置をどの程度精度よく推定可能であるかを評価することを目的とする。そのため、評価尺度として、各アレイで検出された各音源方向に対する直線と、表 1 に示した対象音源の座標位置との距離を測定した。ここでは、音源方向推定誤差による位置推定誤差の他に、スピーカの直径が 9 cm で、対象音源の位置が正確ではないことも考慮し、位置推定誤差が40 cm以内であれば、その方向は対象音源が発しているものとみなすこととした。各アレイで観測された各方向に対し、上述の条件を満たしたブロックの数を発話区間のブロック数で割ったものを検出率とする。

Fig. 6 に “human” および “loudspeaker”の2種類の音源に対する結果を音源の位置（“1” ~ “6”）と向き（“F”, “L”, “B”, “R”）の条件ごとに表示している。それぞれのアレイ（“Array1”, “Array2”）で検出された音源方向は、直接音（“d”）、平面 y=0 での反射音（“ry”）および平面 x=0 での反射音（“rx”）に分けて結果を表示している。

Fig. 6 の結果より、まず音源の位置と方向によってそれぞれのアレイで直接音（d）および反射音（ry, rx）が観測される率が変化することが分かる。これは、音源の位置と向きによって、アレイが「見えている」のか、壁が「見えている」のかに依存する。例えば “human” 音源におけるFig. 6の上図の “1L” の条件では、Array1の直接音 d と反射音 rx がおよそ 0.8 の率で検出されている。また、Array2では、反射音 rx がおよそ 0.6 の率で検出され、直接音はほとんど検出されていない。

“human”と “loudspeaker”の結果を比較すると、全体的に人が発声した場合の方向の検出率が高い結果が得られた。これは人よりもスピーカの方が、指向性が強いことが原因である。

音源位置推定においては、同じ音源に対し、複数（少なくとも2つ）の方向が検出されれば、その重なった位置に音源が存在すると判定することができる。例えば、“human”音源の “6R”の条件で、0.9 以上の率で両アレイの直接音が重なって観測されている。 “loudspeaker”音源の場合でも、0.8前後の検出率が得られている。

“human”音源で、直接音が高い率で上位を占めている条件は、{2F, 3F, 4F, 5F, 6F, 3L, 4L, 4B, 5B, 1R, 2R 5R, 6R} で、全条件のおよそ半分を占めいている。平面x=0での反射音（rx）が上位に入っている条件は、Array1の場合{1L, 2L, 5L, 6L, 6B}となっている。これらの条件は、平面x=0 の壁に近く、その方向を向いている条件である。また、平面y=0での反射音（ry）が上位に入っている条件は、 Array1の場合は{1F, 6F}で、Array2の場合は{3F, 4F, 4R}となっている。

その一方、 “loudspeaker”音源では、直接音が高い率で上位を占めている条件は、{6R} のみの条件となっている。反射音（rx もしくは ry）が最も高い率で検出されている条件は、{4F, 5F, 6F, 1L, 2L, 5L, 6L} となっている。そのうち、{4F, 1L, 6L}の条件では、直接音がほとんど観測されず、反射音のみが上位を占めている。これらは、両アレイに背いているが、壁が近いため反射音が直接音よりも強く観測される条件である。従って、音源の指向特性に応じて、反射音の特定は音源定位に大きな役割を果たすことが示されている。

{1B, 2B, 3B} の条件では、音源が両アレイに背いている状態であるため、両アレイで直接音も反射音

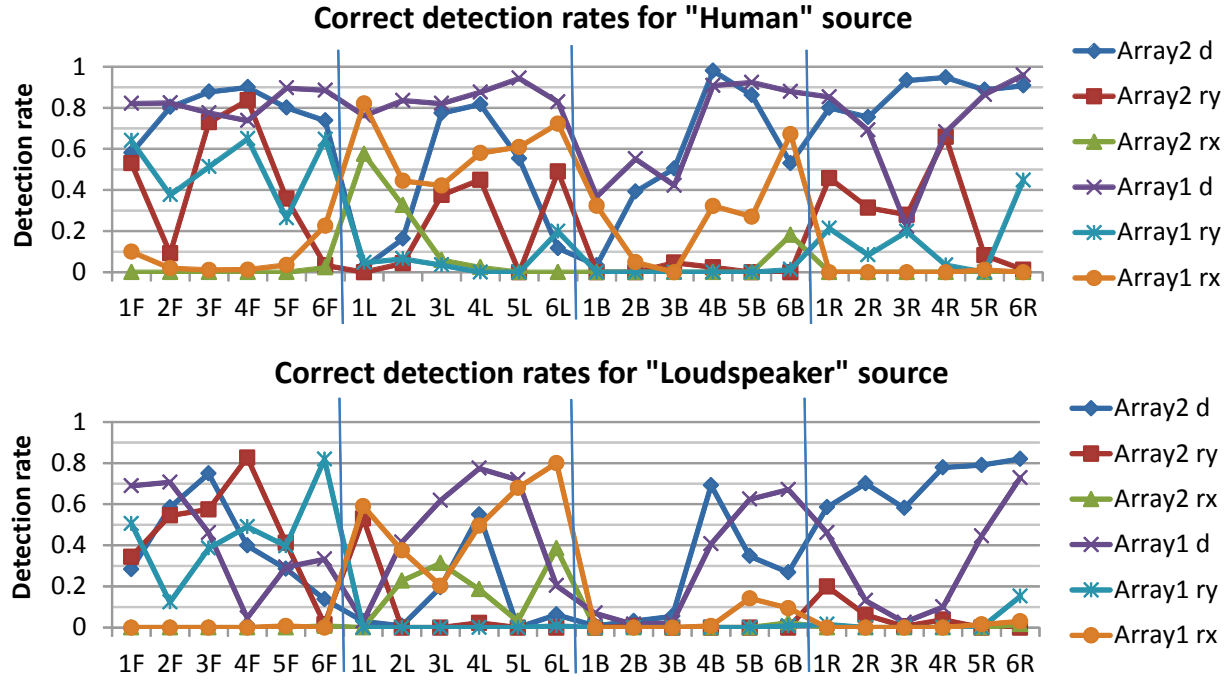


Fig. 6 Correct detection rates for direct path (d) reflection at plane y=0 (ry) and reflection at plane x=0 (rx) by each array (Array1, Array2), for each position (1 ~ 6) and orientation (F: front, L: left, B: back, R: right) of the target sources ("human" and "loudspeaker").

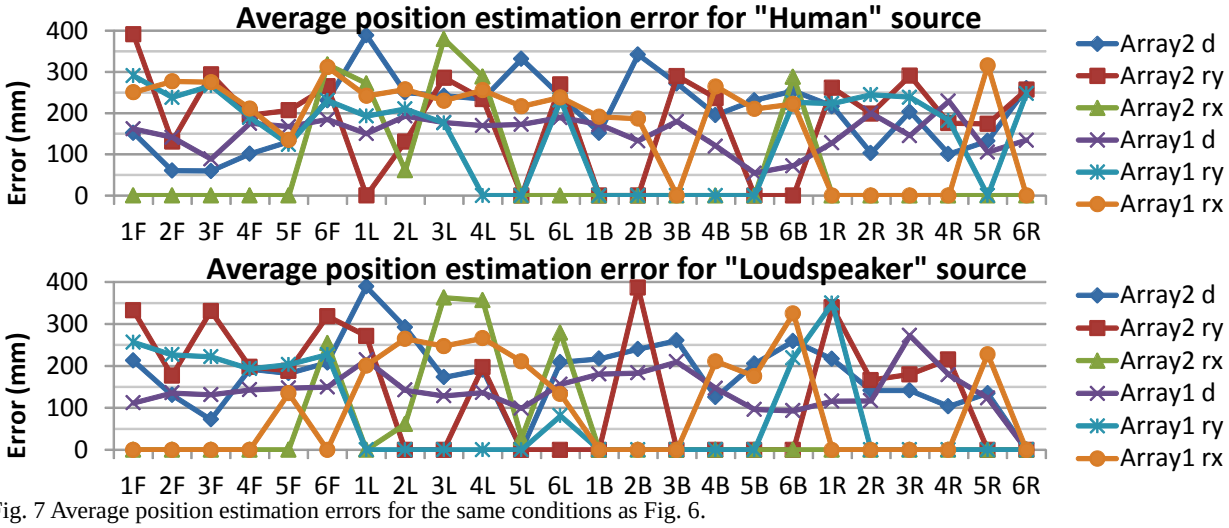


Fig. 7 Average position estimation errors for the same conditions as Fig. 6.

も検出率が低い（0.5 前後）。指向性の高いスピーカでは、検出率が 0 となっている。これらの条件の対処として、部屋に設置するアレイの数を増やす必要があると考えられる。

Fig. 7 に、Fig. 6 の条件に対応する平均位置推定誤差を示している。ただし、これらのグラフで誤差が0の点は、その条件で方向が検出されなかった場合を示している。

Fig. 7 の結果より、直接音でも反射音でも平均誤差は100 ~ 300 mmの範囲で検出されていることが読み取れる。いくつかの条件（例えば "Array2 ry 1F", "Array2 d 1L", "Array2 rx 3L 4L"）では、誤差が300 mm 以上となっているが、これらの条件は、音源とアレイとの距離が長く（直接音は 1 番目の位置からおよそ3 m、反射音の場合折り返しの距離も含めて 1 番目の位置からおよそ 5 m、3 番目と4 番目の位置からおよそ 7 m となり）、方向推定の誤差が大きくなることが原因と考えられる。この誤差を小さくする対処法として、局所的に分解能を上げることが考えられるが、今後の課題とする。

4　おわりに

本研究では、複数のマイクロホンアレイにおいて音源方向推定を行い、空間の情報と反射音の方向の

情報を利用し音源定位（3 次元空間の位置推定）に利用する枠組みを提案した。

人の声とスピーカから再生した音声を音源とした評価実験を行った結果、マイクロホンアレイの位置、壁の位置、音源の種類、音源の位置と向きに応じて、観測される直接音や反射音が変化し、反射音が重要となる条件を明らかにした。スピーカは人よりも指向性が強いことが分かり、アレイとの距離よりも壁とアレイに対する向きに応じて、反射音のみが観測される場合があり、人の声とは異なる観測パターンが得られた。また、このような結果は、通常スピーカを用いて「実験室実験」を行う研究が多いが、人が実際に発声した際の指向性特定が異なることを考慮すべきであることを示している。

今後の課題として、異なった音源や向きのより詳細な分析を行い、本研究の音源定位法を LRF などによる人位置検出の結果と統合させ、誰がいつどこで発話したのかを記述する音環境知能技術に発展させる予定である。

付録：MUSIC 法

M個のマイク入力のフーリエ変換$X_m(k,t)$は、式(1)のようにモデル化される。

$$\mathbf{x}(k,t) = [X_1(k,t),...,X_M(k,t)]^{\mathrm{T}} = \mathbf{A}_k\,\mathrm{s}(k,t) + \mathbf{n}(k,t) \quad (1)$$

ベクトルs(k,t)はN個の音源のスペクトル$S_n(k,t)$から成る：s$(k,t) = [S_1(k,t), \ldots, S_N(k,t)]^{\mathrm{T}}$ 。kとt はそれぞれ周波数と時間フレームのインデックスを示す。ベクトル$\mathbf{n}(k,t)$は背景雑音を示す。行列 $\mathbf{A}_k$ は変換関数行列であり、(m,n)要素はn番目の音源からm番目のマイクロホンへの直接パスの変換関数である。$\mathbf{A}_k$ のn列目のベクトルをn番目の音源の位置ベクトル（steering vector）と呼ぶ。

まず、式(2)で定義される空間相関行列$\mathbf{R}_k$を求め、式(3)に示す$\mathbf{R}_k$の固有値分解により、固有値の対角行列$\mathbf{\Lambda}_k$および固有ベクトルから成る$\mathbf{E}_k$が求められる。

$$\mathbf{R}_k = E[\mathbf{x}(k,t)\mathbf{x}^H(k,t)] \quad (2)$$

$$\mathbf{R}_k = \mathbf{E}_k\mathbf{\Lambda}_k\mathbf{E}_k^{-1} \quad (3)$$

固有ベクトルは $\mathbf{E}_k=[\mathbf{E}_k^{\mathrm{s}} \mid \mathbf{E}_k^{\mathrm{n}}]$ のように分割出来、$\mathbf{E}_k^{\mathrm{s}}$と$\mathbf{E}_k^{\mathrm{n}}$はそれぞれ支配的な$N$個の固有値に対応する固有ベクトルと、それ以外の固有ベクトルである。

MUSIC空間スペクトルは式(4)と(5)で求める。rは距離、θとφはそれぞれ方位角と仰角を示す。式(5)は、スキャンされる点 (r,θ,φ) における正規化した位置ベクトルである。

$$P(r,\theta,\varphi,k) = \frac{1}{|\,\tilde{\mathbf{a}}_k^H(r,\theta,\varphi)\mathbf{E}_k^n\,|^2} \quad (4)$$

$$\tilde{\mathbf{a}}_k(r,\theta,\varphi) = \frac{\mathbf{a}_k(r,\theta,\varphi)}{\left\|\mathbf{a}_k(r,\theta,\varphi)\right\|} \quad (5)$$

空間スペクトル（本稿ではMUSIC応答と呼ぶ）は、MUSIC空間スペクトルを式(6)のように平均化したものである。

$$\overline{P}(r,\theta,\varphi) = \frac{1}{K}\sum_{k=k_L}^{k_H} P(r,\theta,\varphi,k) \quad (6)$$

k_Lとk_Hは、周波数帯域の下位と上位の境界のインデックスであり、$K = k_H - k_L + 1$。音源の方位は、MUSIC応答のN個のピークから求められる。

謝辞

本研究は総務省の戦略的情報通信研究開発推進制度（SCOPE）の研究委託により実施したものである。

参考文献

1) F. Asano, M. Goto, K. Itou, and H. Asoh, “Real-time sound source localization and separation system and its application on automatic speech recognition,” in *Eurospeech 2001*, Aalborg, Denmark, 2001, pp. 1013–1016.
2) K. Nakadai, H. Nakajima, M. Murase, H.G. Okuno, Y. Hasegawa and H. Tsujino, "Real-time tracking of multiple sound sources by integration of in-room and robot-embedded microphone arrays," in *Proc. of IROS 2006*, Beijing, China, 2006, pp. 852–859.
3) S. Argentieri and P. Danès, "Broadband variations of the MUSIC high-resolution method for sound source localization in Robotics," in *Proc. of IROS 2007*, San Diego, CA, USA, 2007, pp. 2009–2014.
4) M. Heckmann, T. Rodermann, F. Joublin, C. Goerick, B. Schölling, "Auditory inspired binaural robust sound source localization in echoic and noisy environments," in *Proc. of IROS 2006*, Beijing, China, 2006, pp.368–373.
5) T. Rodemann, M. Heckmann, F. Joublin, C. Goerick, B. Schölling, "Real-time sound localization with a binaural head-system using a biologically-inspired cue-triple mapping," in *Proc. of IROS 2006*, Beijing, China, 2006, pp.860–865.
6) J. C. Murray, S. Wermter, H. R. Erwin, "Bioinspired auditory sound localization for improving the signal to noise ratio of socially interactive robots," in *Proc. of IROS 2006*, Beijing, China, 2006, pp. 1206–1211.
7) Y. Sasaki, S. Kagami, H. Mizoguchi, "Multiple sound source mapping for a mobile robot by self-motion triangulation," in *Proc. of IROS 2006*, Beijing, China, 2006, pp. 380–385.
8) J.-M. Valin, F. Michaud, and J. Rouat, "Robust 3D localization and tracking of sound sources using beamforming and particle filtering," *IEEE ICASSP 2006*, Toulouse, France, pp. IV 841–844.
9) B. Rudzyn, W. Kadous, C. Sammut, "Real time robot audition system incorporating both 3D sound source localization and voice characterization," Procs.of *ICRA 2007*, Roma, Italy, 2007, pp. 4733–4738.
10) C. T. Ishi, O. Chatot, H. Ishiguro, N. Hagita, “Evaluation of a MUSIC-based real-time sound localization of multiple sound sources in real noisy environments,” in *Proc. of the 2009 IEEE/RSJ Intl. Conf. on Intelligent Robots and System*, St. Louis, USA, 2009, pp. 2027–2032.

人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-B202-12

A Two Microphone-Based Approach for Multiple Speaker Localization on the SIG-2 Humanoid Robot

Ui-Hyun Kim and Hiroshi G. Okuno
Dept. of Intelligence Science and Technology
Graduate School of Informatics, Kyoto University
Kyoto, Japan
{euihyun, okuno}@kuis.kyoto-u.ac.jp

Abstract—A system based on the generalized cross-correlation (GCC) method weighted by the phase transform (PHAT) has been developed for multiple speaker localization. In real environments with binaural robot audition, speaker localization is degraded by the interference created when the speech waves arrive at a microphone from two directions around the robot head and by impaired performance when there are multiple speakers. This paper presents a new time difference of arrival (TDOA) factor for the GCC-PHAT method to compensate multipath interference on the assumption of spherical robot head and a multisource speech tracking method consisting of voice activity detection and K-means clustering for multiple speaker situations. The standard K-means clustering algorithm was improved for the purpose of multisource speech tracking by adding two additional steps that increase the number of clusters automatically and eliminate clusters containing incorrect DOA estimations. Experiments conducted on the SIG-2 humanoid robot in a real environment show that our method improved the localization accuracy and can track multiple speakers in real-time with tracking error below 5.3°.

I. INTRODUCTION

'Binaural' literally means having two sound inputs. For a robot, it means having two microphones, one on each side of its head (like ears). Recently, many researchers and engineers have conventionally used lots of microphones for robots to improve the hearing performance [1]–[2]. However, using numerous microphones causes some problems: rising maintenance costs for microphones and computational power, and losing a general-purpose software interface due to the different microphone array configuration for each robot. The cost for a binaural audition device is substantially less than that for a multichannel audition device. Binaural audition hardware and its software can be easily ported to various kinds of robot platforms and embedded in information and communication technology (ICT) devices. Moreover, research on binaural audition can contribute to understanding the human hearing mechanism [3]. For these reasons, binaural audition is particularly important for robots.

Among the various functions required for robot audition, sound source localization (SSL) is one of the most important techniques to achieve more natural and intelligent human-robot interaction (HRI). For example, a robot estimates the directions of sound sources to understand the acoustic scene. Then it faces or tracks the person speaking and signals him/her that it is ready to listen and thereby appear to express its interest in the conversation. SSL has been extensively studied by a number of researchers and the primary clues have revealed. They include the interaural level difference (ILD), the interaural time difference (ITD), and the spectral modifications caused by parts of the body (the pinna, head, shoulders, etc.). These clues are implicitly included in the head related transfer function (HRTF) [4]. The ITD, more commonly referred to as the time difference of arrival (TDOA), plays an important role in SSL; the sound signals arrive at each microphone at different times for directions other than front and back. One of the most well-known methods to estimate TDOA for SSL with binaural sound inputs is the generalized cross-correlation (GCC) method with phase transform (PHAT) weighting [5].

The use of a microphone array with many microphones improves localization performance in noisy and reverberant environments. On the other hand, localization performance generally drops as the number of microphones is reduced. Since a binaural robot audition system uses only two microphones embedded on each side of the robot's head, there are difficulties in obtaining a performance as good as that when using the microphone array. The localization performance with only two microphones must be improved to enable robots with a binaural audition system to be deployed in various acoustic environments.

In this paper, we addressed two problems affecting the accuracy of the direction-of-arrival (DOA) estimation based on the GCC-PHAT method in binaural robot audition:

1) *Multipath interference due to diffraction of sound waves caused by shape of robot head:* Sound waves easily bend around the robot's head, resulting in a difference in TDOA between the waves that travel around the front of the head and those that travel around the back of the head.

2) *Correlation between multisource sound sources in real environments:* The accuracy in SSL deteriorates when multiple sound sources are correlated, which is generally the case in real environments, i.e., when the sound sources are speech.

Multipath interference severely degrades localization performance especially for sound sources in the lateral direction (around ±90°) and the correlation between sources limits the number of sound sources that the binaural system can localize to a single source. Our solutions to these two problems are twofold:

1) We incorporate a new TDOA factor for multipath interference into the GCC-PHAT method along with

assuming that the robot's head is spherical.

2) We devised a multisource speech tracking method consisting of voice activity detection (VAD) and K-means clustering in order to eliminate incorrect DOA estimations due to the correlation between multiple sound sources.

Our proposed methods were implemented as a real-time system using the 'HARK' open-source robot audition software [6] and evaluated experimentally in the binaural audition system of the SIG-2 humanoid robot.

The paper is outlined as follows: Section II summarizes the ML-based DOA estimation using the GCC-PHAT method for a multisource situation and describes the two problems related to the SSL accuracy in real environments. Section III gives our solutions to the two problems: a new TDOA factor to compensate for multipath interference and a multisource speech tracking method consisting of VAD and K-means clustering algorithms to eliminate incorrect DOA estimations. Section IV presents the experimental results. Section V concludes the paper.

II. Multisource Direction-Of-Arrival Estimation

In this section, we summarize the ML-based DOA estimation using the GCC-PHAT method for multiple sound sources. Then two problems related to the SSL accuracy with binaural robot audition in real environments are explained.

A. Acoustic Model

This paper employs a *F*-point short-time Fourier transform (STFT) under a far-field assumption [7]. The observed signals from the left and right microphones in a situation with *K* sound sources can be mathematically modeled as

$$X_l[f,n] = \sum_{k=1}^{K} \alpha_{lk}[f]\,|S_k[f,n]|\exp\left(-j2\pi\frac{f}{F}fs\tau_{lk}\right) + N_l[f,n] \quad (1)$$
$$X_r[f,n] = \sum_{k=1}^{K} \alpha_{rk}[f]\,|S_k[f,n]|\exp\left(-j2\pi\frac{f}{F}fs\tau_{rk}\right) + N_r[f,n],$$

where $X_{l,r}[f,n]$, $S_k[f,n]$, and $N_{l,r}[f,n]$ are the *f*-th elements of the STFT of the measured signals from the two microphones (*l* and *r*), the sound sources (*k* denotes the index of each sound source), and uncorrelated additive noise, respectively, on the *n*-th time frame index; the $f \in \{1, \ldots, F\}$ denotes a frequency bin, *F* is the time frame size of the STFT, and *fs* is the sampling frequency; $\alpha_{l,r}$ and $\tau_{l,r}$ are the attenuation factor and time delay from the position of the sound source to each microphone, respectively.

B. ML-Based DOA Estimation for Multiple Sound Sources

The ML-based DOA estimation for multiple sound sources is basically defined by the GCC-PHAT method as follows:

$$\hat{\theta}_{mle_k}[n] = \arg\max_{\theta} \frac{1}{F}\sum_{f=1}^{F} G^{PHAT} X_l[f,n]X_r^*[f,n]\exp\left(j2\pi\frac{f}{F}fs\tau_{lr}(\theta)\right), \quad (2)$$

where

$$G^{PHAT} = \frac{1}{|X_l[f,n]X_r^*[f,n]|}, \quad (3)$$

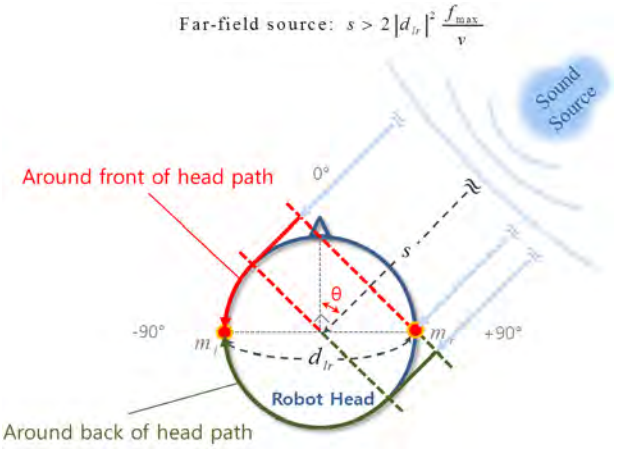


Figure 1. Multipath interference due to diffraction of sound waves with spherical-head assumption.

$$\tau_{lr}(\theta) = \frac{d_{lr}}{v}\sin\left(\frac{\theta}{180}\pi\right). \quad (4)$$

$X_l[f,n]X_r^*[f,n]$, G^{PHAT}, and τ_{lr} are the cross-power spectrum, the PHAT weighting that preserves only the phase information in the cross-power spectrum, and a steering factor for TDOA derived from the free space environment, respectively; $\theta \in \{-90°, \ldots, +90°\}$ is an angle of sound incidence, * is the complex conjugate, d_{lr} is the distance between two microphones, and *v* is the speed of sound (340.5 m/s, at 15 °C, in air). The estimated DOAs θ_{mle_k} of the multiple sound sources in (2)–(4) are obtained by finding several expected angles of sound incidence θ that equally maximizes the sum of the characteristic function obtained from the cross-power spectrum with PHAT weighting in the frequency domain.

C. Problem: Multipath Interference Due to Diffraction of Sound Waves Caused by Shape of Robot Head

Basically, TDOAs are estimated under the assumption that the microphones are located in free space, e.g. in (4). However, this assumption is not applicable to TDOA estimation using two microphones in a robot head because the sound waves easily bend and spread along the shape of the robot head, which creates a difference in TDOA between the waves that travel around the front of the head path and those that travel around the back of the head path. Figure 1 illustrates the two paths created by the diffraction of the sound waves with the assumption that the robot head is spherical. It clearly shows that these two sound-wave paths and multipath interference must be considered if sound source localization in binaural robot audition is to be more accurate.

D. Problem: Correlation between Multiple sound sources in Real Environments

The multisource DOA estimation using (2)–(4) can produce accurate estimates of DOAs in the ideal case that the multiple sound sources S_k are uncorrelated with each other and with additive noise $N_{l,r}$; i.e., $S_1[f,n]S_2[f,n]=0$,

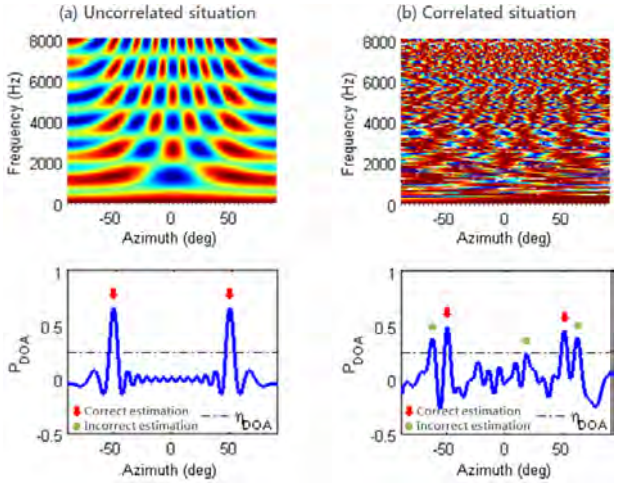


Figure 2. Frequency spectrum and the peak distributions of the ML-based DOA estimation for two sound sources coming from angles of -50° and +50°: (a) A situation with two uncorrelated sources. (b) A situation with two highly correlated sources.

$S_k[f,n]N_{l,r}[f,n]=0$, *and* $N_l[f,n]N_r[f,n]=0$. However, the accuracy deteriorates when multiple sound sources are correlated, which is generally the case in real environments, i.e., when the sound sources are speech corrupted by noise and reverberation. For example, if two correlated sound sources are assumed to come from different directions, (1) can be rewritten as:

$$\begin{aligned} X_l[f,n] &= \alpha_{l1}[f]\left|S_1[f,n]\right|\exp\left(-j2\pi\frac{f}{F}fs\tau_{l1}\right) \\ &\quad + \alpha_{l2}[f]\left|S_2[f,n]\right|\exp\left(-j2\pi\frac{f}{F}fs\tau_{l2}\right) + N_l[f,n] \\ X_r[f,n] &= \alpha_{r1}[f]\left|S_1[f,n]\right|\exp\left(-j2\pi\frac{f}{F}fs\tau_{r1}\right) \\ &\quad + \alpha_{r2}[f]\left|S_2[f,n]\right|\exp\left(-j2\pi\frac{f}{F}fs\tau_{r2}\right) + N_r[f,n], \end{aligned} \tag{5}$$

and their cross-power spectrum can be expressed as

$$\begin{aligned} & X_l[f,n]X_r^*[f,n] \\ &= \alpha_{l1}[f]\alpha_{r1}[f]\left|S_1[f,n]\right|^2\exp\left(j2\pi\frac{f}{F}fs(\tau_{r1}-\tau_{l1})\right) \\ &\quad + \alpha_{l2}[f]\alpha_{r2}[f]\left|S_2[f,n]\right|^2\exp\left(j2\pi\frac{f}{F}fs(\tau_{r2}-\tau_{l2})\right) \\ &\quad + \alpha_{l1}[f]\alpha_{r2}[f]\left|S_1[f,n]\right|\left|S_2[f,n]\right|\exp\left(j2\pi\frac{f}{F}fs(\tau_{r2}-\tau_{l1})\right) \\ &\quad + \alpha_{l1}[f]\alpha_{r2}[f]\left|S_1[f,n]\right|\left|S_2[f,n]\right|\exp\left(j2\pi\frac{f}{F}fs(\tau_{r1}-\tau_{l2})\right). \end{aligned} \tag{6}$$

We can verify that there are two more incorrect TDOAs produced by the correlation between two sound sources in (6). Moreover, if we assume a situation in which there are more than two sources or in which additive noise and reverberation are correlated with other sound sources, the number of incorrect TDOAs will increase geometrically. This phenomenon causes ambiguity in multisource DOA estimation because there will be many peaks in incorrect directions as well in correct ones. Figure 2 shows examples of peak distributions in multisource DOA estimation for two sound signals coming from angles of -50° and +50°. In the uncorrelated situation (a), two sound sources were virtually generated; in the correlated situation (b), two speech signals (one for a male and one for a female) recorded at the same time by the SIG-2 humanoid robot in an experiment room were used. When the sound sources were correlated, the ML-based DOA estimation inaccurately estimated multiple DOAs because of the numerous peaks spread in all directions. In addition, since the intensity of each peak changed over time because of the attenuation factors in (1) applied to each sound source, the ML-based DOA estimation may select peaks in incorrect directions as correct DOAs when the peak intensities in the correct directions are lower than those in the incorrect directions. Furthermore, the ML-based DOA estimation with a threshold η_{DOA} frequently fails to produce the same number of DOAs as sound sources, especially in the absence of information on how many sound sources are active.

These results show that a function is needed to filter out the incorrect DOA estimations in order to get accurate multisource sound localizations.

III. Improved Multisource Direction-Of-Arrival Estimation in Binaural Robot Audition

Our solutions to the two problems in real environments described above are presented here. In binaural DOA estimation, the two problems cause inaccurate and unreliable localizations. We have proposed a new TDOA factor and devised a multisource speech tracking method consisting of voice activity detection (VAD) and K-means clustering. The standard K-means clustering algorithm was extended to enable tracking of an unknown time-varying number of speakers by adding two additional steps that increase the number of clusters automatically and eliminate clusters containing incorrect DOA estimations.

A. New TDOA Factor for Multipath Interference

To solve the problem of multipath interference due to the sound waves traveling along two paths around the robot head, we first apply a simplified formula to these two paths under the assumption that the head is spherical:

$$Path_{front}(\theta) = \frac{d_{lr}}{2v}\left(\frac{\theta}{180}\pi + \sin\left(\frac{\theta}{180}\pi\right)\right), \tag{7}$$

$$Path_{back}(\theta) = \frac{d_{lr}}{2v}\left(\mathrm{sgn}(\theta)\pi - \frac{\theta}{180}\pi + \sin\left(\frac{\theta}{180}\pi\right)\right), \tag{8}$$

where $Path_{front}$ and $Path_{back}$ are respectively the time delays for the path around the front of the head and that around the back of the head for each sound incidence direction, and *sgn* is a signum function that extracts the sign of θ, i.e., if θ has a negative sign, $sgn(\theta)$ is −1. After the formulas for the two paths are derived, the time delay between them for each sound direction is obtained using

$$Diff_{front-back}(\theta) = Path_{back} - Path_{front} = \frac{d_{lr}}{2v}\left(\mathrm{sgn}(\theta)\pi - \frac{2\theta}{180}\pi\right), \tag{9}$$

where $Diff_{front-back}$ is 0 when θ is −90° or +90°. Suppose that the intensity of the multipath interference from $Path_{back}$ for each sound direction complies with that of the ILD ratios

between two microphones located in the robot head and this intensity of the ILD ratios shows the sine function in the ideal condition. We use $Diff_{front\text{-}back}$ multiplied by the absolute sine function with attenuation factor β_{multi} as a factor to compensate for multipath interference:

$$Multi_{front-back}(\theta) = \frac{d_{lr}}{2v}\left(\mathrm{sgn}(\theta)\pi - \frac{2\theta}{180}\pi\right)\cdot\left|\beta_{multi}\sin\left(\frac{\theta}{180}\pi\right)\right|, \quad (10)$$

where $Multi_{front\text{-}back}$ is the compensation factor for multipath interference in binaural robot audition. The final time delay factor for the binaural DOA estimation can be derived using $Path_{front}$ and $Multi_{front\text{-}back}$:

$$\begin{aligned}\tau_{multi}(\theta) &= Path_{front}(\theta) - Multi_{front-back}(\theta) \\ &= \frac{d_{lr}}{2v}\left(\frac{\theta}{180}\pi + \sin\left(\frac{\theta}{180}\pi\right)\right) \\ &\quad - \frac{d_{lr}}{2v}\left(\mathrm{sgn}(\theta)\pi - \frac{2\theta}{180}\pi\right)\cdot\left|\beta_{multi}\sin\left(\frac{\theta}{180}\pi\right)\right|.\end{aligned} \quad (11)$$

This new TDOA factor, τ_{multi}, is used instead of τ_{lr} in (4) with the ML-based DOA estimation.

B. Multisource Speech Tracking

Our approach to eliminate incorrect DOAs estimations due to the correlation between multiple sound sources described above is to use data mining in each time frame. For this purpose, we devised a multisource speech tracking module based on two methods:

- *Statistical model-based Voice Activity Detection*

If the target sound sources are localized speech in noisy environments, all DOA estimations during the noisy periods can be eliminated by using the VAD method to differentiate speech from background noise. We used the statistical model-based VAD algorithm proposed by Sohn et al [8]. This algorithm uses the log likelihood ratio (LLR) between the Gaussian statistical models of speech and background noise for low signal-to-noise (SNR) ratio cases to indicate with high accuracy the presence or absence of speech.

Each time frame is determined to be "speech-present" or "speech-absent" by using a decision procedure based on LLR with a threshold:

$$\begin{aligned}&if\ \hat{P}_{VAD}[n] = \frac{1}{F}\sum_{f=1}^{F}(\gamma[f,n] - \log\gamma[f,n] - 1) > \eta_{VAD} \\ &then\quad n = speech\text{-}present\ frame \\ &else\quad n = speech\text{-}absent\ frame\end{aligned}, \quad (12)$$

$\gamma[f,n]=|X[f,n]|^2/\lambda_N[f,n]$ is the *a posteriori* SNR and $\lambda_N[f,n]$ is the estimated variance of $(N_l[f,n]+N_r[f,n])/2$.

- *Improved K-means clustering*

K-means clustering is a commonly used data mining algorithm featuring computational simplicity and high speed. We improved the standard *K*-means clustering algorithm to work well for multisource sound tracking in real situations. If the multiple DOA estimations in the given time frames are the observations to be clustered and if their cluster centers represent the tracked DOAs for a specific time frame, i.e., given the initial sets of observations $(\theta_{mle\_1}, \theta_{mle\_2}, \ldots, \theta_{mle\_p})$ and *K*-clusters $(\Theta_{track\_1}, \Theta_{track\_2}, \ldots, \Theta_{track\_k})$ with their center means $(\theta_{track\_1}, \theta_{track\_2}, \ldots, \theta_{track\_k})$, the standard *K*-means algorithm proceeds by alternating between two steps:

[Assignment Step] Assign each observation to the cluster with the closest mean:

$$\Theta_{track_k}^{(i)} = \{\hat{\theta}_{mle_p} : |\hat{\theta}_{mle_p} - \theta_{track_k}^{(i)}|^2 \le |\hat{\theta}_{mle_p} - \theta_{track_j}^{(i)}|^2\ \forall 1 \le j \le K\}, \quad (13)$$

where p denotes the index of all estimated DOA in the given time frames, i denotes the iteration number. Each initial center mean is randomly assigned and each DOA estimation θ_{mle_p} goes into exactly one cluster Θ_{track_k}.

[Update Step] Calculate the new means to be the centroid of the observations in each cluster:

$$\theta_{track_k}^{(i+1)} = \frac{1}{\left\langle\Theta_{track_k}^{(i)}\right\rangle}\sum_{\hat{\theta}_{mle_p}\in\Theta_{track_k}^{(i)}}\hat{\theta}_{mle_p}, \quad (14)$$

where $<\Theta_{track\_k}>$ is the number of estimated DOAs belonging to cluster Θ_{track_k}. These two steps are repeated until the assignments no longer change.

There are two problems with the standard *K*-means clustering when it is to be used for multisource speech tracking:

1) Fixed number of clusters: The number of clusters is fixed from the beginning to the end of the standard K-means clustering calculations. This means that the number of speech sources needs to be known in advance for exact clustering. Furthermore, the number of clusters cannot be automatically changed in the observation period for clustering even though speech signals independently appear and disappear over time.

2) Absence of a function for filtering out incorrect DOA estimations: In the standard *K*-means clustering, the tracked directions of the speech signals are not correct because even incorrect direction estimations are used for calculating the center of each cluster.

These two problems cause errors in the results of multisource speech tracking. For accurate multisource speech tracking, we improved the standard *K*-means clustering by including two additional steps with new criteria:

[Increase Step] Increase the number of clusters automatically:

$$\begin{aligned}&if\ \frac{1}{\left\langle\Theta_{track_k}^{(i)}\right\rangle}\sum_{\hat{\theta}_{mle_p}\in\Theta_{track_k}^{(i)}}\left|\hat{\theta}_{mle_p} - \theta_{track_k}^{(i)}\right|^2 > \eta_{C1} \\ &then\ K^{(i+1)} = K^{(i)} + 1\ and\ move\ to\ \text{Assignment Step} \\ &else\qquad\qquad move\ to\ \text{Elimination Step}.\end{aligned} \quad (15)$$

The *K*-means clustering algorithm begins with one cluster *(K=1)*. After executing the assignment step and the update step, it adds another cluster *(K=K+1)* if the variance of observations in each cluster is more than a given threshold η_{C1}.

[Elimination Step] Eliminate clusters containing incorrect direction estimations:

$$\begin{aligned}&if\ \frac{\left\langle\Theta_{track_k}^{(i)}\right\rangle}{\sum_{k=1}^{K}\left\langle\Theta_{track_k}^{(i)}\right\rangle} < \eta_{C2}\quad then\quad \text{eliminate } cluster\ \Theta_{track_k}^{(i)} \\ &else\qquad\qquad \text{keep } cluster\ \Theta_{track_k}^{(i)}\end{aligned}. \quad (16)$$

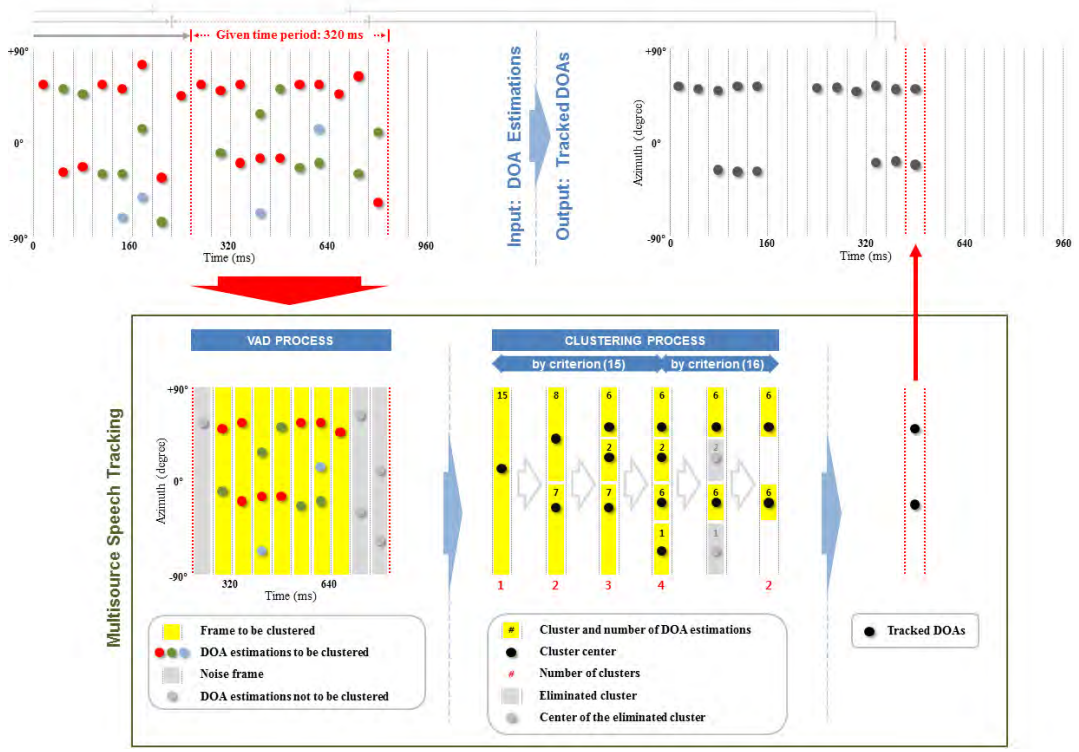


Figure 3. Multisource speech tracking by VAD and *K*-means clustering.

The increase step maximizes the number of clusters by using the variance of DOA estimations in each cluster. In this case, some clusters will likely contain few DOA estimations that are all incorrect. The elimination step filters out the clusters containing incorrect direction estimations by checking the ratio between the number of DOA estimations in each cluster and the number of all DOA estimations in the given time frames with a given threshold η_{C2}.

The process of the improved *K*-means clustering algorithm for multisource speech tracking is thus as follows:

1) The standard *K*-means algorithm (the assignment step and the update step) is executed with $K=1$.

2) The standard *K*-means algorithm is repeated with $K=K+1$ on the basis of Criterion (15).

3) All clusters containing incorrect DOA estimations are eliminated on the basis of Criterion (16).

The process of multisource speech tracking with multisource DOA estimations by VAD and K-means clustering is shown in Fig. 3.

IV. Evaluation

We evaluated our ML-based SSL method with the new TDOA factor τ_{multi} to verify that it makes fewer localization errors than with the conventional TDOA factor τ_{lr} in binaural robot audition and tested it with the multisource speech tracking method in a time-varying two or three number of speakers situation. The subject of the experiment was the SIG-2 humanoid robot equipped with two Sennheiser ME 104 omnidirectional microphones and operated by the 'HARK' open-source robot audition software in the real-time. Figure 4 shows the flow of the implemented robot audition system. The tracked DOAs were used to make the robot turn at its neck and waist in order to look in the speaker's directions.

A. Experimental Setup

The experiments were conducted in a room with a reverberation time of about 120 ms and noise from air conditioners and personal computers. To create a noisier environment, background music with lyrics was played as additive noise. The average sound pressure level (SPL) of the background music and the average SNR of the target speech signals were about 70.1 dB and 19.2 dB, respectively. The SIG-2 humanoid robot was placed at the center of the room, and the speakers were located 1.5–2.5 m from the robot. The attenuation factor (β_{multi} in (11)) in the ML-based DOA estimation was set to 0.1 and the values for the thresholds (η_{VAD} in (12), η_{C1} in (15), and η_{C2} in (16)) used in the multisource speech tracking method were set to 50.0, 3.0, and 0.25, respectively. The system recorded the background noise for 2 s before each trial to estimate the noise variance and used the variance as the *a priori* noise variance for the VAD used in the multisource speech tracking method.

To obtain an accurate estimate of the performance improvement with our ML-based DOA estimation with the new TDOA factor in binaural robot audition, we estimated it in a single-speaker situation first. A male and then a female speaker stood at points along the azimuth from −90° to +90° in 10° steps and spoke to the robot five times at each point. Then we evaluated the ML-based DOA estimation with the

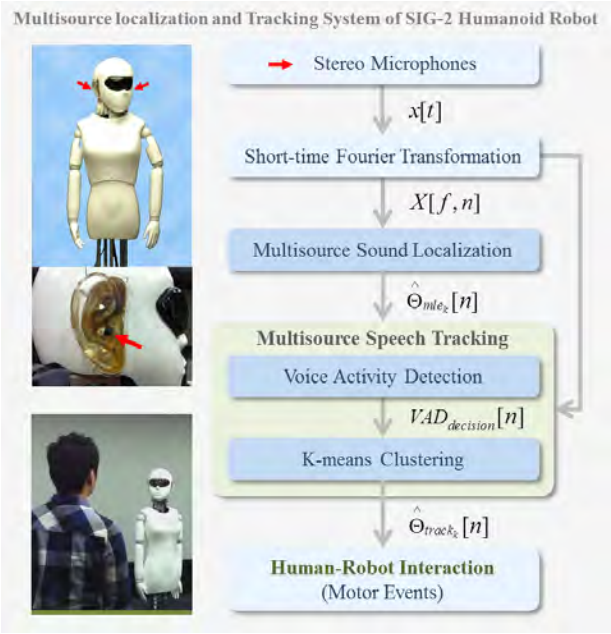


Figure 4. Flowchart of multissource localization and tracking.

multisource speech tracking method in time-varying two- and three-speaker situations for 6 s.

B. *Experimental Results*

Figure 5 shows the root mean square error (RMSE) for the 190 trials (19 points × 5 speech signals × 2 speakers) for the two experimental methods in a single-speaker situation. As shown in the figure, the ML-based SSL methods with the new TDOA factor τ_{multi} had fewer localization errors than with the conventional TDOA factor τ_{lr}. The new TDOA factor τ_{multi} was particularly effective—it reduced the average RMSE by 18.1° and the RMSEs for the side directions by over 37°.

Our tacking method consisting of statistical model-based VAD and improved *K*-means clustering showed good overall performance even though it sometimes failed in tracking with the exact number of directions. Figure 6 shows the experimental results of two- and three-speaker localization and tracking for each 6 s period. Even though the multisource DOA estimation produced many incorrect DOA estimations (shown by (c)), the multisource speech tracking method filtered them out and tracked the direction of each speaker in the running-time domain regardless of changes in the number of speakers over time (shown by (d)). The root mean square error (RMSE) of each tracked DOA for 6 s was less than 5.3° for two- and three-speaker situations.

As a result, despite the use of only two microphones, the robot audition system showed good overall performance for binaural multi-speaker localization in a real environment.

V. Conclusion

We addressed two accuracy problems with the binaural DOA estimation using the GCC-PHAT method in real environments. To solve the problem of multipath interference due to diffraction of the sound waves around the robot head, a

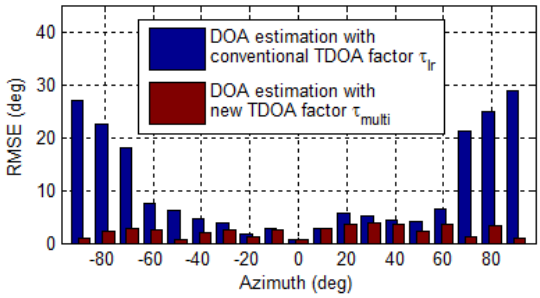


Figure 5. RMSEs for single-speaker localization.

new TDOA factor that takes into account multipath interference is applied to the GCC-PHAT method with the assumption that the robot head is spherical. To overcome the correlation between multiple sound sources, a multisource speech tracking method consisting of statistical model-based VAD and K-means clustering was devised. To make multisource speech tracking more effective, the standard K-means clustering algorithm was improved by adding two additional steps increasing the number of clusters automatically and eliminating clusters containing incorrect direction estimations.

Experimental results demonstrated that taking multipath interference into account when estimating the time delay caused by the diffraction of the sound waves is a key to improving localization performance in binaural robot audition. Doing this with the multisource speech tracking method enabled our real-time binaural robot audition system to correctly track the directions of multiple speakers regardless of the periods during which they spoke and changes in the number of speakers below in tracking error 5.3°.

Future work includes extending our multisource speech tracking method so that it can deal with even more moving speakers. Several problems can occur in a moving-speaker situation, such as incorrect tracking due to ambiguity of speaker identification when moving speakers cross paths or when they are speaking in the same direction. We are planning to implement a blind source separation technique with independent vector analysis [9] in our multisource speech tracking method to handle this problem.

References

[1] U. H. Kim, J. Kim, D. Kim, H. Kim, and B. J. You, "Speaker Localization Using the TDOA-based Feature Matrix for a Humanoid Robot," in *Proc. IEEE Int. Symp. on Robot and Human Interactive Communication (RO-MAN)*, pp. 610-615, Munich, Germany, August 2008.

[2] K. Nakadai, H. Nakajima, G. Ince, and Y. Hasegawa, "Sound Source Separation and Automatic Speech Recognition for Moving Source," in *Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, pp. 976-981, Taipei, Taiwan, October 2010.

[3] J. Blauert and J. Braasch, "Binaural Signal Processing," in *Proc. IEEE Int. Conf. on Digital Signal Processing (DSP)*, pp. 1-11, Greece, July 2011.

[4] C. I. Cheng and G. H. Wakefield, "Introduction to Head-Related Transfer Functions (HRTFs): Representations of HRTFs in Time, Frequency, and Space," *Audio Engineering Society*, vol. 49, pp. 231–249, April 2001.

[5] C. H. Knapp and G. C. Carter, "The Generalized Correlation Method for Estimation of Time Delay," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 24, no. 4, pp. 320-327, 1976.

[6] K. Nakadai, T. Takahashi, H. G. Okuno, H. Nakajima, Y. Hasegawa, and H. Tsujino, "Design and Implementation of Robot Audition System 'HARK' - Open Source Software for Listening to Three Simultaneous Speakers," *Advanced Robotics*, vol.24, pp.739-761, 2010.

[7] J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization (Revised Edition)*, Cambridge, MA: MIT Press, 1997.

[8] J. Sohn, N. S. Kim, and W. Sung, "A Statistical model-based voice activity detection," *IEEE Signal Processing Letters*, vol. 6, no. 1, pp. 1-3, January 1999.

[9] T. Kim, T. Eltoft, and T. W. Lee, "Independent Vector Analysis: An Extension of ICA to Multivariate Components" *International Conference on Independent Component Analysis and Blind Source Separation (ICA)*, LNCS 3889, pp. 165–172, 2006.

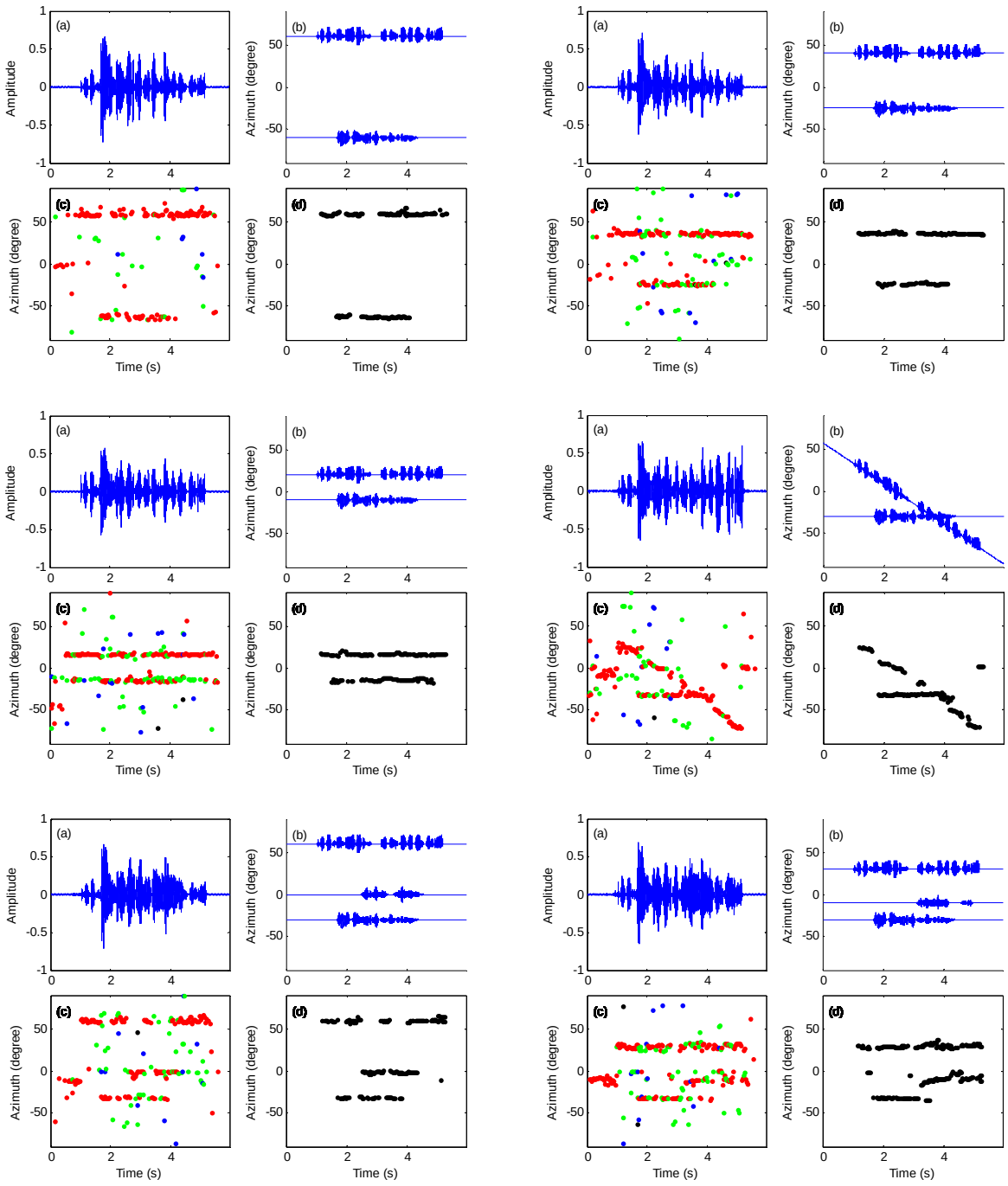


Figure 6. Results of two- and three-speaker tracking with 128-ms time frame, 32-ms time shift, and 320-ms time duration for clustering (10 time frames). (a) Signal input to left microphone consisting of male speech signal and female speech signal. (b) Actual directions and speech durations of two speakers. (c) Results of multisource sound localization, where colors (red, green, and blue) indicate peaks heights in ascending order. (d) Results of multisource speech tracking.