マイクロホンアレイを用いたサウンドマスキングシステムの検討
－音声入力において指向性を利用したマスキング音レベル低減の試み－

☆小幡将信（上智大院），△秋山あい（上智大），△生田萌人（上智大院），
日岡裕輔（オークランド大），荒井隆行（上智大）

1　はじめに

薬局や病院の待合室，学校の保健室，銀行の窓口などコミュニケーションが不可欠な場面におけるプライベートな内容に対し，プライバシーや機密内容を保護する取り組み，スピーチプライバシーに関する研究が 50 年以上前から欧米で行われている[1]．こうした取り組みの中でマスキング音を周囲に流すことで，漏れてくる会話内容を聞き取りにくくする技術はサウンドマスキングと呼ばれる．しかし，サウンドマスキングの問題としてマスキング音による第三者への会話の妨げや不快感の増加があげられる[2]．そのため，ターゲット音に対しマスキング音は必要以上の音量を流さないことが好ましい．

本研究ではこの点を踏まえ，想定聴取状況において，第三者へ効率的にマスキング音を流すマスキングシステムの開発を試み，作成したシステムの有効性について考察を行う．

2　問題設定と仮説

2.1 問題設定

診察室で医者と患者が会話をしている間，外の待合室では次の診察を待っている患者がおり，医者と患者の会話を待機者には聞かせたくないという状況を例にする．図 1 のように，患者を話者 A, 医者を話者 B，マスキング音を流すスピーカーC, D，待機者を聴取位置 E, F と定める．話者 A と話者 B は交互に発話し，マスキング音をスピーカーC, D から流し聴取位置 E, F においてその会話をマスクしたい．

その状況において，スピーカーC, D から流すマスキング音のレベルをどこまで下げられるかという問題を考える．

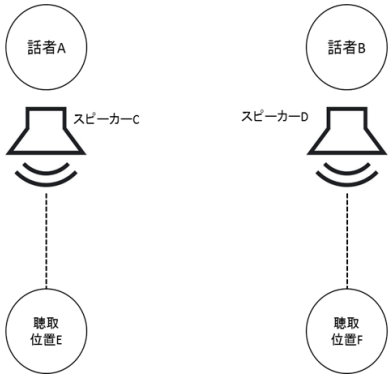

図 1　想定聴取状況

2.2 仮説

N. Marrone ら(2008)の研究[3]において，スピーカーの配置によるマスキング効果の違いが提唱された．この研究から「マスカーの再生位置が変わるとマスキング効果も変わる．特に，話者位置－マスカー－聴取位置が直線上にあるとき一番マスキング効果が高くなる」ということがわかる。このことから，2.1.で想定した状況において話者 A が発話している間，話者 A－スピーカーC－聴取位置 E は直線上にあるため，聴取位置 E においてスピーカーC のマスキング音が最も影響を与えると考えられる．また，話者 A－聴取位置 F の直線上に最も近いスピーカーはスピーカーC であるため，聴取位置 F においてもスピーカーC のマスキング音が最も影響を与えると考えられる．

同様にして話者 B が発話している間，聴取位置 E, F ともにスピーカーD のマスキング音が最も影響を与えると考えられる．

よって，話者 A が発話している間は主にスピーカーC からマスキング音が流れ，逆に話者 B が発話している間は主にスピーカーD からマスキング音が流れるという形にすることで，全体としてマスキング音の音量を低下させつつマスキング効果も損なわないモデルができるのではないかと考えた．

“Investigating the use of microphone array for sound masking system: Towards reducing masking sound level using the directivity of input sounds,” Mochinobu Obata, Ai Akiyama, Moeto Ikuta (Sophia Univ.), Yusuke Hioka (Univ. of Auckland), Takayuki Arai (Sophia Univ.)

以上を実現するにあたり，マイクロホンアレイを利用した音声分離を用いた方法を提案する．

3　提案手法（システムモデル）

仮説に基づき，それぞれの話者において異なるマスキング音を流すシステムを作成するため，マイクロホンアレイによる音声分離と，それを用いた話者の音量変化に伴い音量レベルが変化するマスキング音[4]の作成及び効果的なスピーカー配置による再生を行う．ただし、本研究ではリアルタイムで録音を行うのではなく、オフラインでの録音データを使う。

3.1 音声分離

図 2 で示すように話者 A と話者 B の音声をそれぞれソース音源s_1, s_2，その音声がマイクロホンアレイに届くまでの伝播経路の伝達関数$a(\theta)$で表わすとき，P個のマイクロホンからなるマイクロホンアレイで録音した音声を録音データ$x_p\ (p = 1, 2, \cdots, P)$に対し任意の角度に指向性を向けたビームフォームのフィルタ係数$w_p(\theta)\ (p = 1, 2, \cdots, P)$得られた出力を$y$とする．ただし，本研究ではマイク数$P = 4$で進めており，図 2 おいても$P = 4$とする．

ここでは日岡ら(2013)によって提案された手法[5]を元に，遅延和ビームフォーマとウィナーフィルタを用いて，録音データ$x_1, x_2, \cdots, x_p$からソース音源s_1, s_2の音声分離を行った。

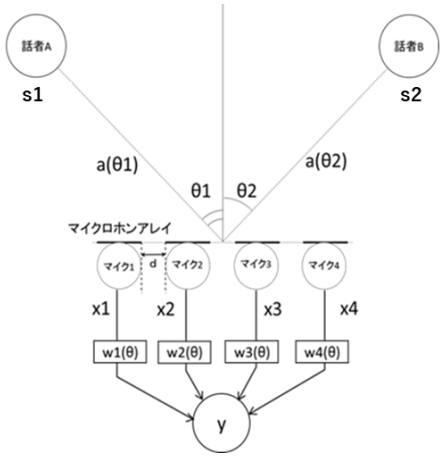

図 2　信号とフィルタ

3.2 マスキング音作成

マスキング音を作成するにあたり川のせせらぎ音をオリジナルとなる音として使用する．これをM_Oとする．また3.1.の音源分離で得た，話者 A 方向から取り出した音声s_1を使用したマスキング音，話者 B 方向から取り出した音声s_2を使用したマスキング音，マイク 1 単体で話者 A, B の会話を収録した信号x_1を使用したマスキング音の 3 種類を作成する．これらをそれぞれM_A, M_B, M_Tとする．これらの作成手順は後の段落に記述する。

先行研究（例えば[6]）において音声マスカーが用いられているが，その特徴の 1 つとしてレベルが追従する点があげられる．本研究では，この追従型のマスキング音が作成しやすいと考え採用した．これ以降M_Aを例にして作成法を示していく．

音源信号$\mathrm{s}(\theta_1)$の最初の 2 秒間のサンプル数をNとし，各サンプルの二乗和をUとする．Uを次の式(2)に示す．

$$U = \sum_{i=1}^{N} |s_1(i)|^2 \tag{2}$$

M_Aの最初の 2 秒間にはそのままM_Oの最初の 2 秒間をあてはめる．これを式(3)に示す．

$$M_A(i) = M_O(i)\ (i = 1 \sim N) \tag{3}$$

その後は，サンプルごとにUと過去 2 秒サンプルの二乗和q_iの比と忘却定数αを使用して以下のように求めていく．これを式(4), (5)に示す．

$$q_i = \sum_{j=i-N}^{i} |s_1(j)|^2 \tag{4}$$
$$(i = N + 1 \sim length(M_O))$$

$$M_A(i) = \alpha \times M_A(i-1) + (1-\alpha) \times \frac{q_i}{U} \times M_O(i) \tag{5}$$
$$(i = N + 1 \sim length(M_O), \text{忘却定数} \alpha = 0.5)$$

同様にしてM_B, M_Tも作成する．

4　従来法との比較

今回のマイクロホンアレイを使用した提案手法では以下の二点を利用することで性能の改善を試みている．

①音声分離で抽出した音の情報，

②目的音との位置関係を考慮した，最適な位置にあるスピーカーの選択

そのため．音声分離を行わない従来の単マイクで集音したデータのマスキング音を二つのスピーカーから同じ音量レベルの音で流した場合を従来法とする．

4.1 想定比較状況

図 1 で示した状況において，提案法として話者 A, B の会話をマイクロホンアレイにより取り出した音声音源信号s_1, s_2を使用して作成したマスキング音M_A, M_BをそれぞれスピーカーC, D から流すパターンと，従来法として話者 A, B の会話を単マイクで録音した音声x_1を使用して作成したマスキング音M_TをスピーカーC, D 同時に流すパターンを比較する．

4.2 録音実験

まず，図 3 で示すように 4 つのマイクロホンアレイ話者 A と話者 B の座標をそれぞれ定めた。

次に，話者 A と話者 B の位置から流したインパルス音を OCTA-CAPTURE を使用して 4 つのマイクロホンアレイで録音した．また，事前に 21 歳男性と 23 歳女性の音声をそれぞれ話者 A, B の発話音声として録音した．これらの音声をたたみ込むことでシミュレーション音声を作成し[7]，話者 A と話者 B の会話をマイクロホンアレイで録音した信号x_1, x_2, x_3, x_4として使用した．

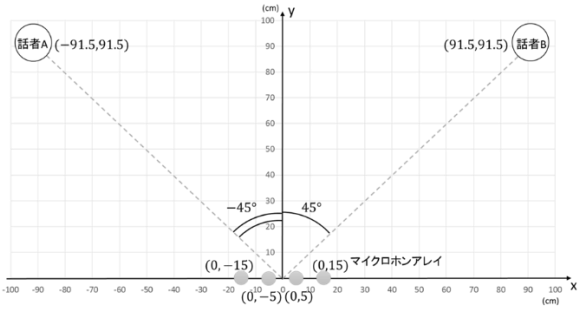

図 3　話者とマイクの座標

4.3 結果

4.2.の実験から得た録音データx_1, x_2, x_3, x_4を使用し，3.1.で示した分離過程を経て得られた音源信号s_1, s_2，3.2.で示した作成過程を経て得られたマスキング音M_A, M_B, M_T，比較のための単マイク録音信号x_1の波形を図 5 に示す．ただし，音声波形の横軸の単位は$\times 10^5$サンプルである（サンプリング周波数は 16 kHz）．

それぞれの波形の形から，音源信号s_1, s_2や単マイクで録音した音声信号x_1の振幅が大きくなったとき，マスキング音M_A, M_B, M_Tの振幅も追従して大きくなっていることがわかる．

また，音声分析ソフト Pratt を使用し，それぞれのマスキング音のインテンシティの平均レベルを測ると，M_Aが約 58.0 dB, M_Bが約 57.0 dB であるのに対し，M_Tは約 57.3 dB となった。近しい値になったが，想定状況において従来法では両方のスピーカーからM_Tが常時流れるのに対し，提案法ではスピーカーが交互に音が流れる形となっているため総合的にみると従来法のレベルが大きいと考えられる。

さらに，図 5 にs_1の音声波形とM_Aの音声波形を重ね，話者 A が発話する一部分を拡大した音声波形を示す．

この図のより話者 A が発話してからマスキング音が十分なレベルになるまで約 1500 サンプル，つまり約 0.1 秒の遅延が発生していることがわかる．

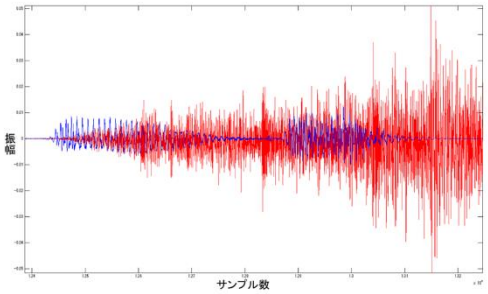

図 4　s_1とM_Aを重ねた音声波形と一部拡大したもの

5　考察

4.3.で得られた結果から，4.1.で想定した状況においてマイクロホンアレイを使用することで話者 A が発話している間はスピーカーC から，話者 B が発話している間はスピーカー D から発話者の音声に合わせた音量レベルのマスキング音が流れる状況ができており，これは仮説で考えたモデルが実現できていると考えられる．

話者が発話を開始したタイミングにおいて遅延の影響でマスキング音のレベルが十分になっていないという問題については，話者の発話を感知した瞬間にある程度のレベルのマスキング音を流し，そこからレベル変化の調整をしていくというシステムによって改善できると予想される．

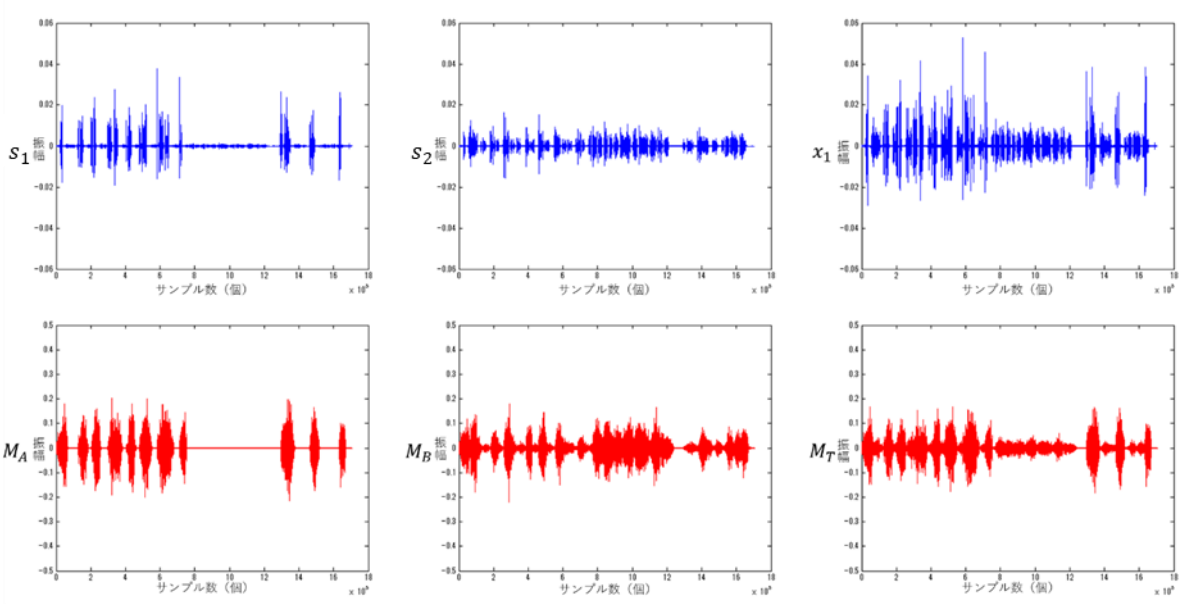

図 5　音源信号とマスキング音の音声波形

6　おわりに

本研究では，マイクロホンアレイを用いて音声分離を行い，マスキング音を作成することでマスキング効果を損なわずに音量低下も見込めるようなシステムについて提案し，その可能性を探った．

今後は，聴取者を対象にしてマスキング効果を評価するために明瞭度や Annoyance を測る聴取実験を行う必要がある．また，今回考えた仮説は定量的な議論が出来ていないため，それを考慮したシステムモデルも今後検討していく．

謝辞

荒井研究室の Justine Hui さんには，実験の指導や論文の執筆等アドバイスしていただき感謝申し上げます．

参考文献

[1] W. J. Cavanaugh, W. R. Farrell, P. W. Hirtle and B. G. Watters, “Speech privacy in buildings,” J. Acoust. Soc. Am., 34, pp. 475–492, 1962.

[2] 王循，藤田佑一郎，木庭洋介，石川諭，雉本信哉, “各種マスカーによる音声マスキング効果及び心理的影響の比較, ” 第 25 回環境工学総合シンポジウム論文集, pp. 14–17, 2015.

[3] N. Marrone, C. R. Mason, G. Kidd, Jr, “Tuning in the spatial dimension: Evidence from a masked speech identification task,” J. Acoust. Soc. Am., 124, pp.1146–1158, 2008.

[4] 李孝珍，上野佳奈子，坂本慎，藤原舞，清水寧, “レベル適応型マスキングシステムの有効性に関する検討, ” 日本音響学会研究発表会講演論文集, pp. 1077–1080, 2010.

[5] Y. Hioka, K. Furuya, K. Kobayashi, K. Niwa and Y. Haneda, "Underdetermined Sound Source Separation Using Power Spectrum Density Estimated by Combination of Directivity Gain," in IEEE Transactions on Audio, Speech, and Language Processing, vol. 21, no. 6, pp. 1240-1250, 2013.

[6] T. Arai, “Masking speech with its time-reversed signal,” Acoust. Sci. & Tech., 31(2), pp. 188–190, 2010.

[7] 橘秀樹，日高新人，“実物及び模型ホールのインパルス応答の測定, ” 日本音響学会誌，48 巻，4 号，pp.244–249, 1992.