マイクロフォンアレイを利用したサウンドマスキングシステムの実現 ―話者の空間情報を用いた，話者音声とマスカーの到来方向一致の試み―

☆生田萌人（上智大院），小幡将信（上智大院），Hui C. T. Justine（オークランド大），
日岡裕輔（オークランド大），荒井隆行（上智大）

1　はじめに

近年，情報化社会が進む中で個人情報保護法のもと，プライバシー保護の重要性が増している．その中で，自分の話している内容を第三者に聞かれないようにするスピーチプライバシーが注目されている．

これは，聞かれたくない音声（ターゲット音）に，マスカーと呼ばれる別の妨害音を被せることで，音声に含まれる内容を理解できないようにする技術である．サウンドマスキングにおいて重要なことは，いかに高いマスキング効果を保ちつつ，聴取者の不快感を減らすかである．高いマスキング効果とは同じ音圧レベルにおいて，よりマスクする効果が得られることである．

本研究では，マスキング効果を保ちつつ，聴取者の不快感を減らすために，マイクロフォンアレイによる音源分離を利用したサウンドマスキングの構築を試みた．実験では，音源分離を利用し作成したマスカーとそうでないマスカーとの比較を行った．2 人の話者の音声をそれぞれのマスカーでマスキングし，書き取りテストと不快感の心理評価により評価をした．

2　マスキングシステムの提案

2.1 環境適応型マスキングシステム

本研究では環境適応型のマスキングシステムに注目する．従来のマスキングシステムでは，常に予め作成されたマスカーを再生するが[1]，環境適応型のマスキングシステムでは，その場その場の音環境を分析し，音環境に適応したマスカーを作成し再生する．一例として，隠したい話者の音声をマイクロフォンで取得し[2]，その音声をフレームごとに時間反転させてマスカーを作成する時間反転型マスカー[3]はマスキング効果が高いことが知られている．その一方でマスカー自体に不快感があることや，自身の声をマスカーとして使われることに抵抗を感じる場合があるなどの問題点も存在する[4,5]．そこで，発話者と関係のない第三者の音声が収められたデータベースを作成し，マイクロフォンで取得した音声の特徴と類似したマスカーを選択して流すシステムも提案されている[6]．このように環境適応型のマスカーは，隠したい音声に適応したマスカーを流すことができるので，マスキング効果の向上が期待されている．本研究においても，環境適応型のマスキングシステムの考え方を採用したシステムを検討する．特に適応する環境情報として，話者の位置情報に注目する．

2.2 音源位置情報(空間情報)の利用

先行研究[7]によれば，マスキング効果は話者とマスカーの位置に依存することが知られており，特に，話者とマスカーと聴取者が一直線上に位置する場合に最もマスキング効果が高くなるということを示唆している．このことを踏まえれば，話者が複数いる場合において，それぞれの話者と聴取者を結んだ線上にスピーカーを置き，そのスピーカーからマスカーを再生することでマスキング効果を高めることが可能だと考えられる．

2.3 問題設定とシステムの提案

前節による議論を踏まえ，本研究における問題設定と仮説について述べる．いま図 1 のように，話者が２人，聴取者が１人という状況でマスキングシステムを利用する場合を想定し，それぞれの直線上にスピーカーを配置するような環境を考える．この環境において，話者 A と話者 B は高い秘匿性を求められる会話をしており，聴取者は話者から少し離れたところにいる第三者と仮定する．ここで話者

*Realization of sound masking system using microphone array: An attempt to match the speaker's voice and the direction of arrival of a masker using the speaker's spatial information," Moeto Ikuta, Mochinobu Obata, Hui C. T. Justine (Univ. of Auckland), Yusuke Hioka (Univ. of Auckland), Takayuki Arai (Sophia Univ.)

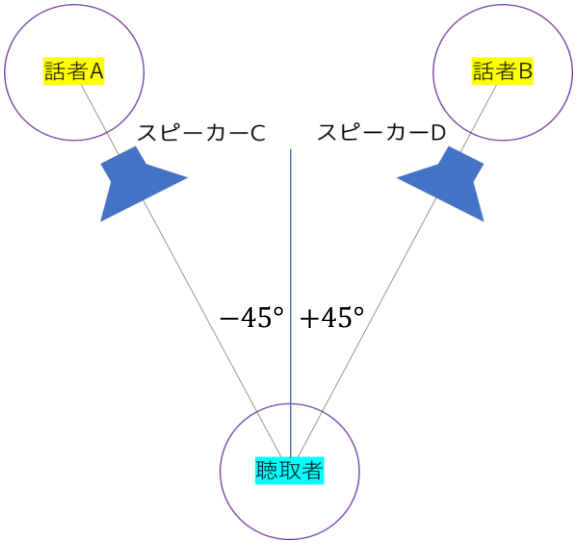

図 1　想定環境

A，話者 B は，聴取者からみて左右 45° の直線上に存在する．また，話者―スピーカー―聴取者が直線状に並ぶようにスピーカーC，スピーカーD をそれぞれ設置する．この状況下において，聴取点でのマスカーの音圧レベルを一定に保った場合に最もマスキング効果が高くなると予想されるのは，話者 A の発話中にスピーカーC からマスカーが再生され，逆に話者 B の発話中にスピーカーD からマスカーが再生される場合である．一方，話者 A の発話中に，スピーカーC とスピーカーD から同じ音圧レベルで同時にマスカーを再生することを考える．この場合，それぞれのスピーカーから再生されるマスカーの音圧レベルは，聴取点での音圧レベルを同じにしたならば，スピーカーC からのみマスカーを再生する場合に比べて小さくなることがわかる．またこの際，スピーカーD より再生されるマスカーのマスキング効果は，その再生方向が離れることから小さくなる．

従って，それぞれの話者からの音声を，その直線上スピーカーを使ってマスキングすることで，従来よりも小さい音圧レベルのマスカーで同等のマスキング効果を実現できる可能性があるのではないか，との仮説が立てられる．

3　マイクロフォンアレイを利用した空間別話者音声取得

2.3 節で述べたようなマスキングシステムを構築するためには，話者ごとの音量を観測する必要がある．話者が異なる方向に位置することからマイクロフォンアレイの利用を考える．マイクロフォンアレイは，空間の異なる位置に複数のマイクロフォンを設置し集音することにより，単一マイクロフォンでは得られない空間情報を信号処理によって得ることができる．この技術を利用することにより，異なる地点に位置する話者の音声を，別々に取得することが可能となる．

そのために，話者 A と話者 B の会話について音源分離処理を行った．1 段階目として，遅延和ビームフォーマを利用した．さらに 2 段階目として，パワースペクトル密度推定によるウィナーフィルタを利用した[8]．本研究ではマイクロフォンの数を 4 つにし，10 cm の等間隔で直線上に並べることとする．紙面の都合上，音源分離の詳細な説明は参考文献[8]に譲り，割愛する．

4　実験条件

提案したマスキングシステムの有効性を検証するために，聴取実験を実施した．本実験では，以下の 3 通りのマスキングシステムを比較し，それぞれのマスキング効果，およびマスカーが聴取者に生ずる不快感について調査した．

1. Same : マイクロフォンアレイを使わないもの
2. BF : マイクロフォンアレイを使用して遅延和ビームフォーマまでの処理を行ったもの
3. WF : BF に加えてさらに PSD 推定によるウィナーフィルタを使用したもの

上記の 3 つの実現方法の違いの総称を以後 PlayStyle と呼ぶことにする．異なる PlayStyle において，TMR（Target to Masker Ratio）を+3, 0, -3, -6, -12 dB と変化させた場合のマスキング効果，不快感を調査した．マスキング効果の評価には書き取りテストを，不快感については心理評価により行った．なお，本実験ではマスカーの作成はオフライン処理で行った．

4.1 実験に使用した音声

ターゲット音声(以後，話者をターゲットと呼ぶ)には，あらかじめ，2 名の 20 代男性話者で録音した音声を使用した．ターゲット文は図 2 のように，キャリア文と 4 モーラの単語から構成される．単語は FW03 コーパス[9]の親密度 5.5~7.0 の単語群の中から 75 単語を

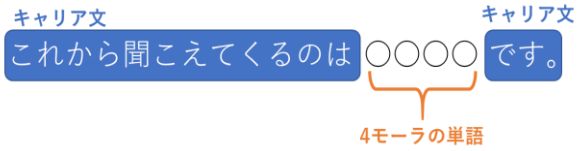

図 2.ターゲット文

選択した.

マスカーの作成には，オリジナルの音として水の音[10]を用い，これを加工した．オフライン処理によるマスカーの作成にあたり，あらかじめインパルス応答を計測しておき，録音した話者音声に畳み込むことで，各マイクロフォンの出力とした．これを上記の PlayStyle に沿って処理を行い，得られた音声を，先行文献[11] で述べられている方法で音声レベルに追従するように加工した.

4.2 実験環境

実験は上智大学荒井研究室の防音室で行い，4 つのスピーカー（GENELEC 8020A）とオーディオインターフェース（Roland OCTA-CAPTURE）を用いた．各機材の配置は図 3 のようにした.

提示条件は 3 つの PlayStyle と 6 つの TMR（+3, 0, -3, -6, -9, -12）で計 18 条件である．カウンターバランスを考え PlayStyle 間，および単語間においてランダマイズをした.

実験参加者の耳の位置における騒音レベルはターゲットが約 50dB，マスカーが約 47～62 dB (A) （TMR の+3 から-12 dB に対応）となるように設定した.

4.3 実験参加者

日本語を母語とし，健聴である大学生 15 名(男性 5 名，女性 10 名)が実験に参加した．平均年齢は約 21.5 歳で，健聴であるかどうかは自己申告とした.

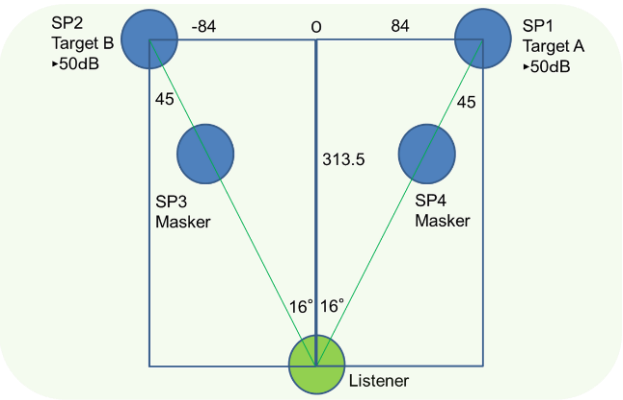

図 3. 実験環境

4.4 評価基準

書き取りテストの採点は，1 文字ずつ行った. 心理評価においては, 特に何も感じない，少し不快感がある，不快感がある，とても不快感がある，の中から 4 段階で評価をしてもらった.

5 結果と考察

実験結果は，実験参加者の回答を採点後，統計的手法の 1 つである線形混合モデル(LMM)[12]を利用し分析を行った．以後，書き取りテストの得点を Score とし，不快感を以後 Annoyance とする.

Score に関する分析では，PlayStyle と TMR と Talker の三つの相互作用を固定効果に入れ，単語によるばらつきを変量効果へ入れた．参加者におけるばらつきは分散が少なかったため変量効果から除外した．LMM による分析結果を図 4 に示す.

話者 A では PlayStyle による差が見られたが，話者 B では PlayStyle による差が見られなかった．ポストホックテストで有意差を調べたところ，話者 A 間においては，WF と Same において（t (790) = -5.227, p = <.0001），WF と BF において（t (879) = -4.256, p = 0.0003）で確認ができた．話者 B 間においては有意差を確認することができなかった．話者 A と話者 B 間では Same において（t (1023) = -4.059, p = 0.0008）のみ確認できた．この結果より，WF は話者 A に対しては効果があるが，話者 B に対しては WF の効果は薄いと考えられる．しかし, 話者 A と話者 B のそれぞれの WF を比べた場合，有意差が見られなかったことから，話者 B においても十分マスキングができていることが考えられた.

Annoyance に関する分析では，PlayStyle と TMR を固定効果に入れ，ID（実験参加者）を変量効果に入れた. Word は分散が小さかったため，変量効果から除外した．分析結果を図 5 に示す.

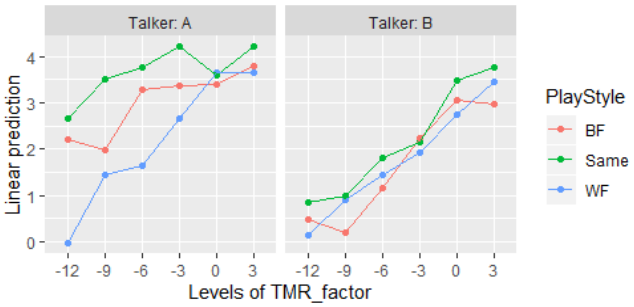

図 4. Score に関する LMM 分析結果

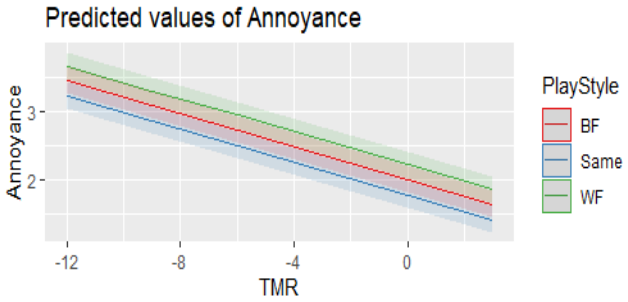

図 5.Annoyance に関する LMM 分析

図 5 より, TMR が低くなるほど不快感が増え，Same，BF，WF の順に不快と感じる度合いが増えていくことが確認できた．エネルギーを一方向から集中させることによるうるささが原因として考えられる．その一方で，書き取りテストの正答率がある値である場合の TMR を PlayStyle ごとに求め, その TMR をもとに不快感を比べると少し異なる結果になった．PlayStyle 間で差が見られた話者 A においては，WF，BF，Same の順に Annoyance が低いことが分かった．マスキング効果の向上に伴い，話者 A においては不快感を軽減することに成功していると考えられる．

6　おわりに

マイクロフォンアレイを用いて話者の空間情報を利用したマスキングシステムを提案し，聴取実験によりその効果を検討した．実験では, 話者 A と話者 B の会話をマスキングすることを想定し，3 つの PlayStyle におけるマスキング効率の調査を行った．マスキング効果については，話者 A に対しては WF が有意差を得られたが，話者 B に対しては有意差が得られなかった. また, 不快感に関しては, Same, BF，WF の順に不快感が増すような結果が得られたが、話者 A について見れば WF の不快感は他より少なかった．話者 A，話者 B の両方に対して高いマスキング効果が期待でき，不快感も話者 A に対しては少ない点から、WF は優れていると考察を行った．話者によるマスキング効果の違いの原因調査と，不快感の軽減が今後の課題である．

謝辞

本研究について, 上智大学のユンさんには, 頻繁に実験のデザインや統計分析についてのご相談に乗っていただきました．大変感謝しております．

参考文献

[1] 佐伯 徹郎 他, “マスキングノイズによるスピーチプライバシー保護に関する一考察,” 電子情報通信学会技術研究報告, EA103 (398), 43-48, 2003.

[2] 赤木正人，入江佳洋，“ 音情景解析の概念にもとついた 音声プライバシー保護，” 信学 Ek, J97-A , 247-255 (2014).

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

[4] Jiang et al., “Sound Masking Performance of Time-Reversed Masker Processed from the Target Speech,” Acta Acustica united with Acustica, Volume 98, Number 1, January/February 2012, pp. 135-141(7).

[5] Hioka et al., “Effect of adding artificial reverberation to speech-like masking sound,” Acoustics 2016, Volume 114, 15 December 2016, Pages 171-178.

[6] 三戸 武大 他, “サウンドマスキングシステムにおけるデータベースを用いた音声マスカ作成法の提案,” 日本音響学会誌, Vol. 71, No. 8, pp. 382–389, 2015.

[7] N. Marrone et al., “Tuning in the spatial dimension: Evidence from a masked speech identification task,” J. Acoust. Soc. Am., 124, pp.1146–1158, 2008.

[8] Y. Hioka et al., "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, June 2013.

[9] 国立情報学研究所, “NTT・東北大 親密度別単語了解度試験用音声データセット（FW03）”, http://research.nii.ac.jp/src/FW03.html.

[10] “効果音自然 2，”King Record Co. Ltd, 2014

[11] 小幡将信, 秋山あい, 生田萌人, 日岡裕輔, 荒井隆行, “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－,” 音講論（秋）2019.

[12] McCulloch et al., "Generalized linear mixed models." *Encyclopedia of biostatistics* 4 2005.