人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-A902

ＡＩチャレンジ研究会 (第３０回)

Proceedings of the 30th Meeting of Special Interest Group on AI Challenges

CONTENTS

日　時　2009 年 11 月 19 日　　場　所　慶應義義塾大学 日吉キャンパス 来往舎 中会議室
Keio University Hiyoshi Campus, Yokohama, Nov. 19, 2009

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

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

人工知能学会研究資料
JSAI Technical Report
SIG-Challenge-A902-2

音響テレプレゼンスロボットにおける頭部形状簡略化の音響的・知覚的影響

Acoustical and Perceptual Influence from Simplifying Head Shape of an Acoustical Telepresence Robot: *TeleHead*

○戸嶋 巌樹 (NTT コミュニケーション科学基礎研究所)
青木 茂明（金沢工業大学）
*Iwaki TOSHIMA (NTT CS Lab.), Shigeaki AOKI (Kanazawa Ins. of Tech.)
toshima@brl.ntt.co.jp, aoki_s@neptune.kanazawa-it.ac.jp

Abstract— **We built an acoustical telepresence robot named TeleHead, which has a user-like dummy head and whose movement is synchronized with the user's head movement in real time. An accurate-shape user-like dummy head improves sound localization accuracy, but making an accurate-shape user-like dummy head for all users is not realistic. We are trying to simplify TeleHead's head shape by taking the effect of head movement into consideration. In this work, we made two types of simplified dummy heads, a ball-like dummy head and a ball-like dummy head with a user-like pinna. At first, we compared HRTF between dummy heads. Dummy heads are acoustically different between each other. Effect of pinna is large and in case of using user-like pinna, HRTF is more similar than the other. Then, we used the dummy heads in sound localization experiments. The experimental results show that the pinna is very important for sound localization in the median plane. Head movement can improve sound localization and subjects can localize sound with another person's pinna. The results indicate the possibility of using a ball-like dummy head with a generic pinna for acoustical telepresence robots.**

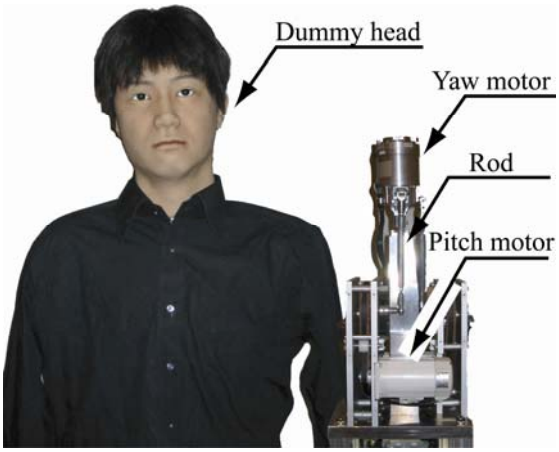


Fig. 1 Acoustical telepresence robot: *TeleHead*. It has a user-like dummy head and synchronizes with user's head movement in three degrees of freedom.

1. はじめに

究極の通信技術は，使用者があたかも遠隔に居るかのように感じる技術であり，それをテレプレゼンス技術と呼ぶ[1]．テレプレゼンスロボットとは，そのテレプレゼンス技術を支えるために，遠隔において使用者の代わりに動くロボットである．使用者が完全に遠隔に居るかのような感覚を得るためには，使用者が遠隔に意図通りの作用を及ぼすことと，使用者に遠隔の感覚を正確に伝えることの，双方向の技術を実現する必要がある．遠隔に身体を持つことは，遠隔において環境に物理的に作用できる可能性を意味する．他のあらゆる信号処理的通信手段は，原理的にこの可能性を持ち得ないか，または，目的とする環境への作用を達成するための新たな装置が必要と考えられる．したがって，テレプレゼンスの実現にはロボット技術が重要な役割を果たすと言える．

これまでに，力覚や触覚を中心として様々なテレプレゼンスが試みられてきた．筆者らは，環境理解という意味において重要な役割を果たし，また，コミュニケーションにおいても重要な位置を占める聴覚・音響のテレプレゼンスロボットの製作を試みている．聴覚は，空間的には全方位に対して機能し，周囲への警戒をはじめとする環境の把握に重要な役割を果たしている．このように，聴覚による音環境の理解は，環境把握にとって重要であり，テレプレゼンス技術によって環境を伝達しようとするとき，音環境の伝達は欠かせない．

身体を持つことの利点は，遠隔において環境へ直接影響を与えることのみではない．人間は自らの身体を通して環境を理解する[2]．音環境の理解においても，自己の音響的性質の果たす役割は大きい．特に自己の頭部形状に由来する頭部の音響的性質である，頭部伝達関数（Head-Related Transfer Function: HRTF）は音源の方向同定，すなわち音像定位において大きな役割を果たすことが知られている[3]．さらに，有効な手がかりは受動的な身体情報だけではない．積極的に頭部を動かすことも音像定位を行う上での重要な手がかりとなることが知られている[4, 5]．頭部運動を考慮したバイノーラルシステムも考案されている[6-8]．このように遠隔に身体を持ち，能動的に動くことが出来るロボット技術は様々な側面でテレプレゼンス技術の実現に欠かせないものである．

筆者らはダミーヘッドを使用者の頭部運動に追従動作させることにより，頭部形状および頭部運動の双方の効果を得ようと考えている．これまでに，使用者と同形状の頭部を持ち，使用者の頭部運動に追従する，音響テレプレゼンスロボット，テレヘッドを製作した（図 1）．また，頭部形状および頭部運動の効果を音像定位実験で確認した[9]．また，ロボットを用いる方法は，音環境の伝達という側面だけを見ても，HRTF やダミーヘッド録音を利用した音環境

の再現手法[10]と比較して，音源の性質や運動を事前に知る必要がなく，メリットがある．しかも，HRTF を測定することの困難さ[11]，およびその困難さを克服し，容易な測定を可能にするために必要な専用の測定系[12]，あるいは HRTF の個人性による定位感の低下問題[13，14]も回避することができる．以上のように音響テレプレゼンスの実現にロボティクスを用いることは，現実的かつ多くのメリットがある．

一方で，ロボティクスを使うことのデメリットも考える必要がある．筆者らはこれまでに，騒音や追従軌道，遅延の問題について，それぞれ議論を行ってきた[15，16]．本稿では頭部形状を使用者個人にカスタマイズして製作しなければならない点に注目し，どの程度の簡略化が現実的に可能であるか検討する．まず，簡略化ダミーヘッドの HRTF について，実頭と比較する．次に，そのダミーヘッドを用いて，音像定位実験を行い，頭部運動の有無や簡略化の条件を変化させ，頭部形状簡略化の可能な範囲について考察する．

2. テレヘッド

2.1 テレヘッドの概要

テレヘッドの外観は図 1 に示したとおりである．使用者の頭部形状を象って製作したダミーヘッドが 3 個のモータによって，Yaw，Roll，Pitch の 3 方向に駆動するようになっている．テレヘッドの概要を図 2 に示す．赤い矢印で表されているのが，音の流れである．ダミーヘッドの外耳道入り口で集音した音信号をアンプとヘッドホンを通じてそのまま使用者に提示する．使用者の HRTF の効果が，使用者に酷似したダミーヘッドによって自動的に加味される仕組みである．青い矢印は頭部運動情報の流れである．ヘッドホンに 3 次元姿勢センサとして Fastrak（Polhmus）を装着し，120 Hz で使用者の頭部姿勢を検出する．テレヘッドはこの情報に基づいてダミーヘッドを PC によって制御した．制御周期は 10 ms である．騒音は人間の聴感度の高い 1-4 kHz の範囲で最大 24 dB SPL であり，静かな図書館程度の騒音である．また，頭部運動に対する追従性能は安定しているが，遅延時間が約 80 ms ある．遅延時間と騒音が現状のテレヘッドではトレードオフの関係にあり，聴覚の心理物理実験を行う都合上，遅延時間を長くし，騒音を最小化した．

2.2 簡略化ダミーヘッド

本稿で論じる簡略化ダミーヘッドの形状について，図 4，5 に示す．図 4 の写真のように，ラグビーボール状の本体の側面に耳介を貼り付けた形で製作した．本体の幅 155mm は図 1 の写真で用いた精密に使用者の頭部形状を再現したダミーヘッドの両耳間の幅と等しい．正面は厚さ約 1mm のシリコン，内壁は厚さ約 4mm 程度の FRP 樹脂である．また，図 5 の中心部分はマイクロホンを接続するための音響インピーダンスの高い素材部分であり，ロボットの振動がマイクロホンに音として伝わることを防止する．耳介部分は取り外しや変更が可能な構造になっており，外した場合は図 5 の様な耳介無しのダミーヘッドとなる．図 1 の写真にあるような精密ダミーヘッドは頭部全体の型取り，MRI や光三次元計測に基づく修

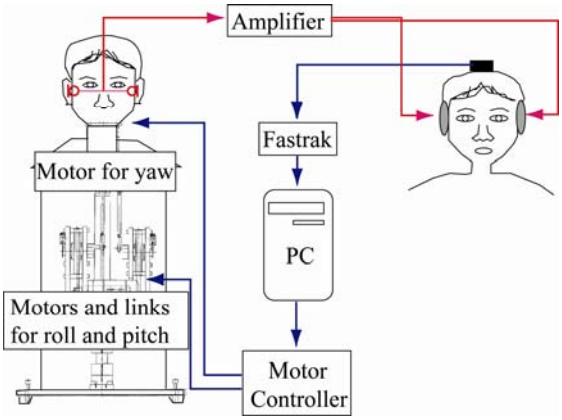


Fig. 2 Outline of *TeleHead*. TeleHead is synchronized with the user's head movement and the sound collected with microphones in the dummy head is transmitted to the user by headphones. Blue lines are the flows of head posture data. Red lines are the flows of acoustical signal.

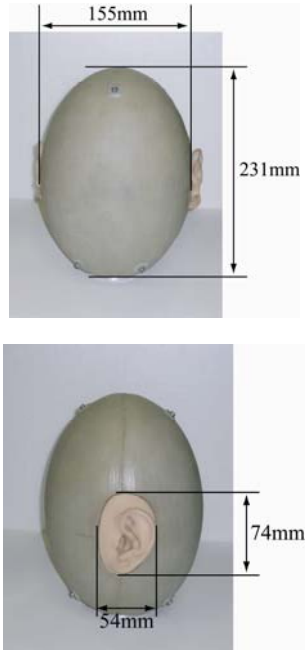


Fig. 4 Simplified dummy head with accurate pinna. Its front view (top panel) and side view (bottom panel).

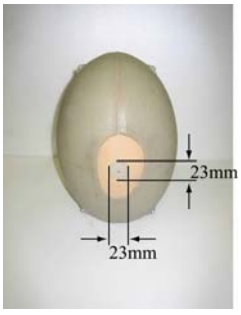


Fig. 5 Ball-like dummy head (no pinna).

正，が必要であり，計測のみでは制作不可能であるし，型取りのみでは十分な効果が得られなかった[17,18]．これらの手順による製作には，高いスキルをもった職人が必要となる．一方，今回制作したダミーヘッドは，耳介つきのもので，特に本人の耳介を模した耳介を使用する場合は耳介の型取りが必要となるが，それ以外には使用者の頭部の写真数枚で十分である．型取りによる耳介の制作は使用者にとっては数 10 分の負担であり，制作者にとっても 1 日程度の負担で制作可能である．従って，本稿で取り上げる簡略化ダミーヘッドは，精密なダミーヘッドと比較して，格段に制作過程が簡略化されたものである．以降，まず音響的性質の再現性について測定し，次に知覚的効果について実験した結果を述べる．

3. HRTF 測定

3.1 測定方法

実頭とダミーヘッドの頭部伝達関数（HRTF）は，頭部の中心から音源までの距離を1.2mとして，無響室内で測定した．HRTFは音源から自由音場における頭部中心位置までの伝達関数 $H_{sp-center}$ と，音源から左右の外耳道入り口までの伝達関数 H_{sp-l} ・ H_{sp-r} の比で表し，周波数 ω と頭部中心位置から見た音源の相対位置（方位角，仰角，距離）＝（θ，ϕ，r）を変数として，式(1)の様に表される．

$$HRTF(\omega,\theta,\phi,r)=\frac{H_{sp-l\,or\,-r}(\omega,\theta,\phi,r)}{H_{sp-center}(\omega,\theta,\phi,r)} \quad (1)$$

実頭の測定では，測定方位は全方位角と仰角-40～90°で合計143点を測定点とした．各測定点は，正中面と水平面は10°おきに，その他の点は隣り合う測定点との間の仰角と水平角が最大でも20°以内に収まるように設定した．これは，HRTF の測定時間を90分以内に納めて，被験者の負担を減らすためである．ダミーヘッドは長時間の測定に耐えられるので，全方位角と仰角-40～90°で5°おきに1873点，ないし10°おきに469点を測定点とした．実頭の測定点はダミーヘッドの測定点のサブセットとなっている．

HRTF の測定には，実頭やダミーヘッドの左右の外耳道入り口付近をシリコン印象材でそれぞれ型取りした耳栓に装着した小型コンデンサマイクロホン（Panasonic，WM62-AT102）を用いた．即ち，実頭もダミーヘッドも外耳道をマイクロホン付きの耳栓で塞ぐ状態で HRTF を測定した[10]． なお，いずれの場合も，測定を開始する前に，レーザポインターを用いて両耳珠と鼻頭の位置が常に同じ位置にくるように位置あわせを行った．ダミーヘッド測定時にはテレヘッドの駆動部等の電源は落とした．また，測定の最初と最後に正面からのHRTFを測定し，大きな誤差が途中に発生していないことを確認した．音源はサンプリング周波数 48 kHzの最適化引き延ばしパルス（TSP）信号[19]を使用し，10回の平均をとった．HRTF は512点のFFTで算出した．

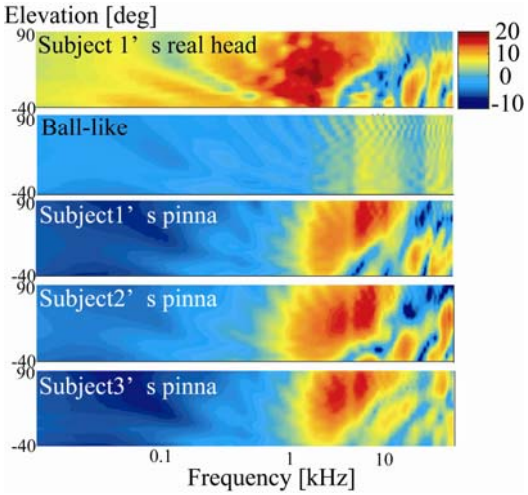


Fig. 6 Results of HRTF measurement for subject 1’s real head, the ball-like dummy head, that with subject 1’s pinna, that with subject 2’s pinna, and that with subject 3’s pinna.

3.2 測定結果

HRTFの測定結果を図6に示す．正面のHRTFを表している．横軸は周波数であり，縦軸は仰角である．正面を0度とし，-40度から90度（真上）までのゲインが色で示されている．暖色の領域でゲインが高い．音像定位の観点からは，6～13kHzあたりに斜めに伸びている低ゲイン領域の形状が重要であるとされている．一番上の実頭の測定結果と比較して，上から2番目の耳介無しの例は大きく異なっている．また，その他の3枚は耳介有りの簡略化ダミーヘッドのHRTFを表しており，3段目のものは，実頭の耳介を象った耳介を装着したものである．耳介形状が等しいダミーヘッド・実頭間のHRTFは比較的近いことが分かる．

次に，HRTFをスペクトル差で表したのが表1である．スペクトル差は式(2)で示される．

$$D_{HRTF}^{FFT}=\sum_d\left(\sqrt{\sum_\omega(|H_i-H_j|)^2/N_\omega}\right)/N_d \quad (2)$$

ここで，d は方向，ωは周波数，[illegible]はそれぞれ，周波数と方向についてそれぞれの測定点数，H_iおよびH_jは$H_i(\omega,d)$，$H_j(\omega,d)$を省略したものである．

実頭を RH，簡略化ダミーヘッド DH で表している．RH や DH の後ろの数字は耳介形状であり，RH1 の耳介形状を模した耳介を装着した簡略化ダミーヘッド

Table 1 Spectral differences between HRTFs [dB]

	RH2	RH3	DH1	DH2	DH3
RH1	7.38	7.90	7.50	7.87	7.66
RH2	–	8.67	8.85	8.48	8.30
RH3		–	8.23	8.30	6.42
DH1			–	6.79	6.63
DH2				–	6.59

が DH1 である．また，耳介無しの簡略化ダミーヘッドは DH0 と表記することにする．表 1 には載せていないが，DH0 は RH と約 10dB の差があり，図 6 から明らかであったが，DH0 は実頭と音響的に大きく異なる形状であると言える．逆に DH1，DH3 は音響的に近い HRTF となった．

4. 音像定位実験

音像定位実験とは，被験者がある条件下において提示された音に関して，その音源の方向を正しく把握しているかどうかを調査する実験である．スピーカ等の実音源を用いる場合は，スピーカ位置と回答位置が近いほど音像定位精度が高いと結論できる．また，HRTF を用いた仮想音源を用いる場合では，実験者が意図した音像位置を音の提示位置と考えて音像定位精度を測定する．聴覚にとって，音像定位機能は最も基本的機能の１つと考えられるため，音像定位実験の結果は音刺激を提示するシステム全体としての質の定量的評価と言える．実音源や仮想音源といった条件に依存せず，人間が直接に音を聴いた場合と同程度の音像定位精度が得られる系があるとすれば，音像定位に関して十分に高品質な音響バーチャルリアリティを実現した系であると結論できる．また，音像定位精度の比較により，系の質を比較することも可能である．

4.1 実験方法

実験の写真を図 7 に示す. 図 7 から分かるように，本稿では，正中面にスピーカを配置した音像定位実験を行った．頭部形状の簡略化は HRTF が使用者とロボットで不一致となるという影響をもたらす．これは主として正中面の音像定位精度に影響を与えると考えられるため，頭部形状簡略化の影響を測るには正中面の音像定位実験を行うことが妥当と考えた．スピーカは 15 度間隔で-45 度から 75 度まで 9 個を設置した．ただし，-40 度から 80 度までの場合や，-50 度から 70 度までの場合もあり，被験者はこれを知らない．ダミーヘッドの中心からスピーカまでの距離は 1.2 m とした．この距離は頭部近傍における HRTF 変化の影響を受けない距離である[17]．刺激音は，持続時間 5 s，音圧 65 dB SPL 程度の白色雑音で，刺激間隔は 8 s とした．再生毎に生成し直し，また，音量も±5dB 程度の範囲で変化させることで，音質，音量等による学習効果を回避した．

実験条件は頭部形状と頭部運動について変化させた．音像定位には頭部形状と共に頭部運動が深く関わっていることが分かっており，頭部形状の簡略化も頭部運動と切り離して考えるのは合理的でない．被験者は 3 名，いずれも成人男性（30 代 2 名，40 代 1 名）で，聴力は事前にオージオメータで測定し正常であることを確かめた．頭部形状の条件は 3 通り．ただし，ダミーヘッドは耳介のみ精密に再現したものを，各被験者に対して製作したため，合計 4 体である．

1. ラグビーボール状のダミーヘッド（DH0）
2. DH0 に被験者 1，2，3 の耳介を接続したもの（DH1/2/3）

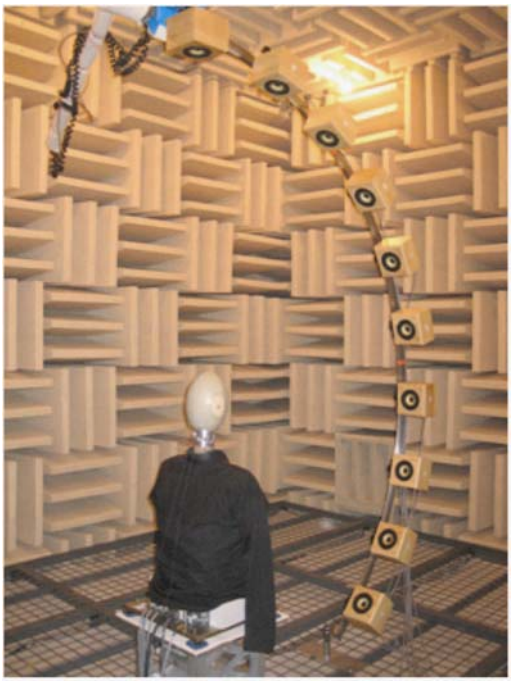


Fig. 7. Photograph of the setup for sound localization experiments. TeleHead with simplified dummy head is set in the anechoic room. Loudspeakers are set in the median plane from -45 deg to 75 deg at intervals of 15 degrees 1.2 m in front of TeleHead.

4.2 実験結果

実験結果の生データを図 8 に示す．頭部運動ありの条件である．最上段は各被験者の耳介形状を模した耳介を接続したダミーヘッドによる結果，最下段は耳介無しのダミーヘッドによる結果である．中段のブロックはそれ以外，つまり，Subject1 にとっては，DH2 と DH3, Subject2 にとっては，DH1 と DH3 など，他者の耳介を装着したダミーヘッドを用いた結果である．横軸は音刺激の提示方位（角度）であり，縦軸は回答方位である．いずれも正面を 0 度とした．マイナスは下方向，プラスは上方向である．従って，提示方位と回答方位が一致する場合，即ち正解は常に対角線上となる．各被験者において，元々の定位精度が異なることもあり，生データにはばらつきが見られるが，傾向として，最下段の DH0 を使用した場合の定位精度が低いことが読み取れる．また，いずれの被験者においても最上段，つまり，被験者の耳介形状を模した耳介を装着した場合の定位精度が高いことも読み取れる．提示方位と回答方位の一致度を相関係数で表したものを図 9，10 に示す．図 9 は被験者毎，図 10 は被験者間をまとめて検定した結果も示しているが，その結果も生データからの印象と一致している．

4.3 精密ダミーヘッドとの比較

テレヘッドにおいては，被験者の頭部形状を精密に模したダミーヘッドも用意している．図 1 は精密ダミーヘッドを装着した写真であり，被験者 1 の精密ダミーヘッドを DH1a と表す．精密ダミーヘッド

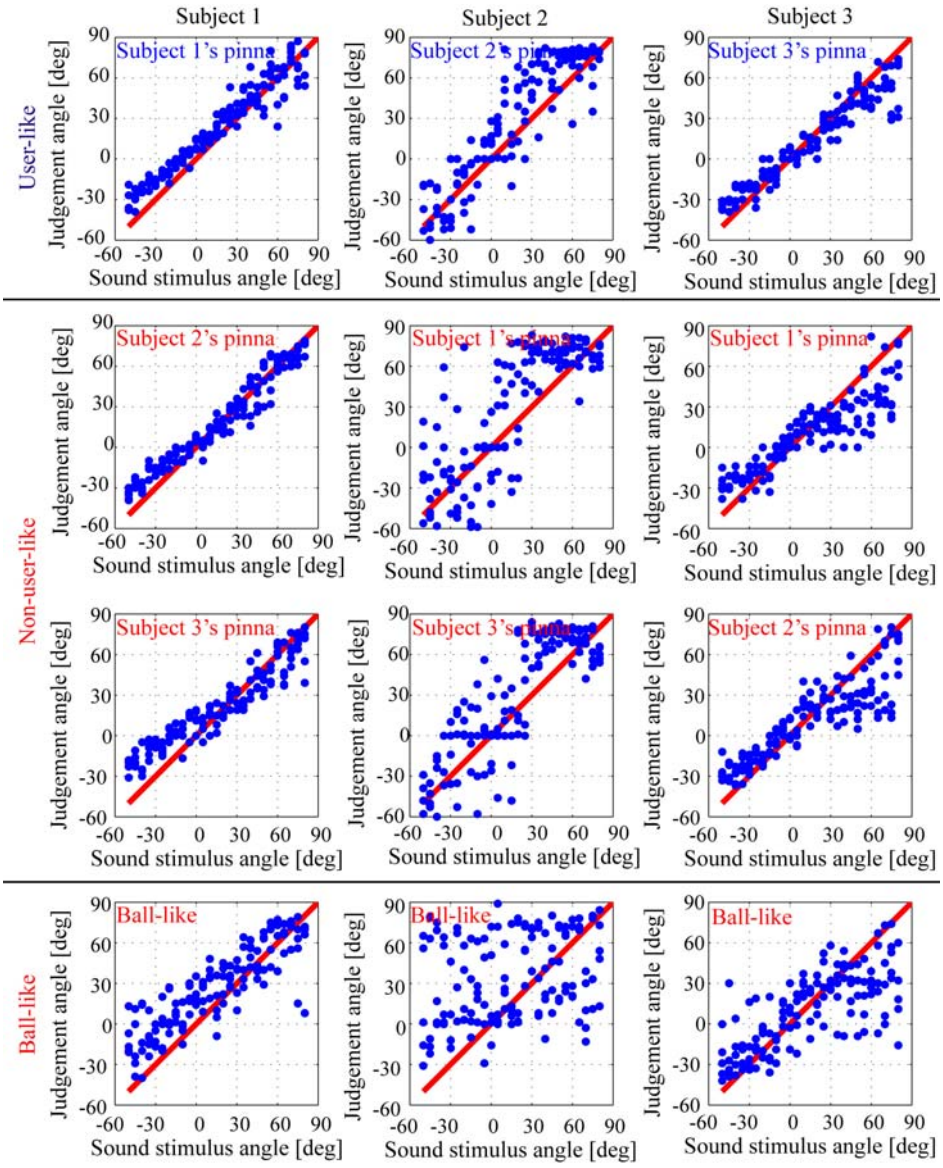


Fig. 8. Results of sound localization experiments in synchronized (head-movement) condition. Top panels show the results using dummy heads with a user-like pinna. Bottom panels show the results using the ball-like dummy head. The others show the results using the dummy heads with a nonuser-like pinna. Left column shows the results for subject 1, center column shows those for subject 2, and right column shows those for subject 3.

と簡略化ダミーヘッドの差を評価する．DH1a は，音響的（HRTF）には RH1 と約 4.6dB の差異であった．また，音像定位実験の生データを図 11 に示す．この時の，正中面音像定位における相関係数は頭部運動無しで 0.94，有りで 0.98 であった．この精度は，実頭で直接聞いた場合の定位精度と概ね同程度である．このように精密ダミーヘッドは確かに効果がある．しかし，簡略化ダミーヘッドは，特に使用者の耳介を装着した場合で，しかも頭部運動がある場合について，精密ダミーヘッドに近い性能を発揮する場合があることが分かる．

5. おわりに

1. ラグビーボール状の簡略化ダミーヘッドを制作した．
2. 簡略化ダミーヘッドには耳介の装着が可能であり，使用者の形状を模した耳介を作成し，HRTF 測定と音像定位実験を行った．
3. HRTF に関して，使用者の耳介を使用した簡略化ダミーヘッドの場合は，実頭の場合と近い数値となった．
4. 音像定位実験の結果は，頭部運動有りかつ被験者と耳介形状が一致している場合に，もっとも高い音像定位精度となった．
5. 耳介無し条件では，常に低い音像定位精度となり，特に頭部運動無しの条件では，実験結果がほとんど意味を成さなかった．
6. 精密ダミーヘッドと比較した場合，音響的には近い精度とはならなかったが，耳介形状一致で，頭部運動有り条件での定位実験の結果は良好であった．これは頭部運動によって，頭部形状を簡略化可能となる可能性を示唆している．

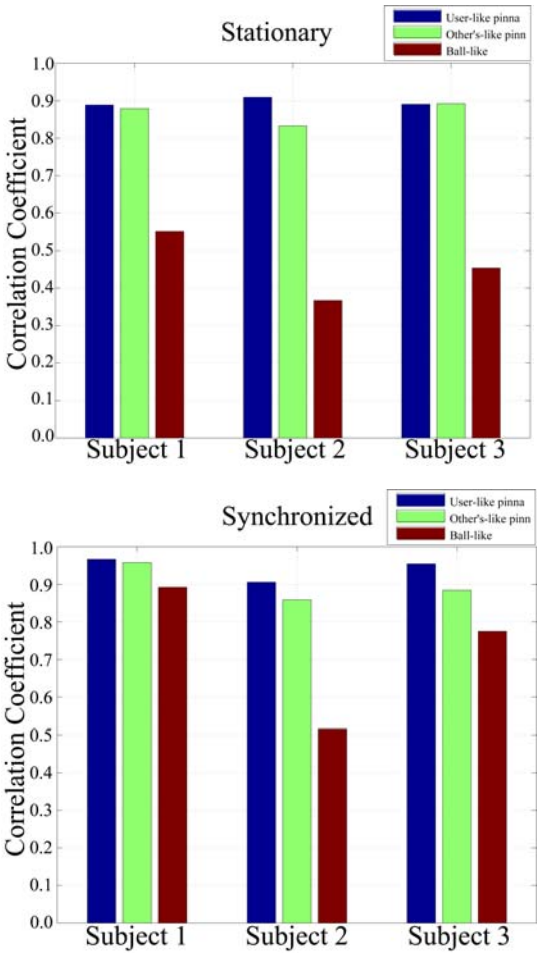


Fig. 9. Correlation coefficient of each result. Upper panel shows the results in the stationary condition and lower panel shows those in the synchronized condition. Blue bars show the results for the ball-like dummy head with user-like pinna. Green bars show the averaged results for two kinds of dummy head with the non-user-like pinna, and brown bars show those for the ball-like dummy head.

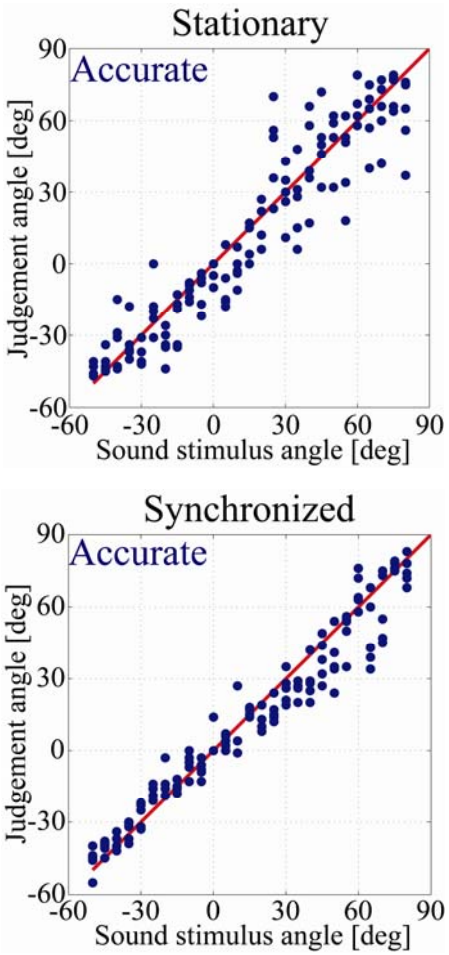


Fig. 11 Results of sound localization experiments using with the accurate user-like dummy head. Correlation coefficient is 0.9427 in stationary condition and is 0.9750 in synchronized condition.

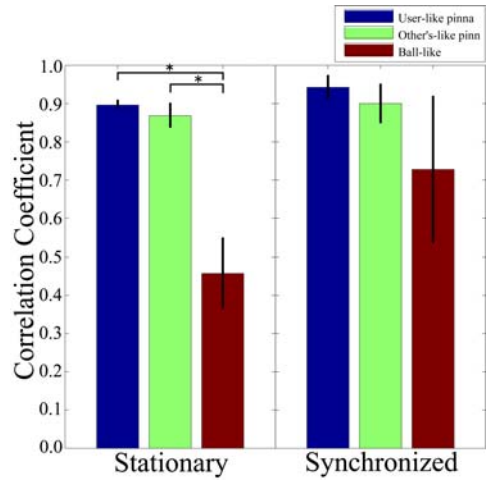


Fig. 10 Averaged correlation coefficient for all subjects. The left graph shows in the stationary condition, and the right graph shows in the synchronized condition. Error bars shows standard deviations of each result.

参考文献

[1] R. M. Held, and N. I. Durlach: "Telepresence", Presence: Teleoperators and Virtual Environments vol. 1, pp. 109 - 112, 1992.

[2] R. Pfeifer, and C. Scheier, “知の創成-身体性認知科学への招待”, 共立出版, 2001.

[3] J. Blauert: "Spacial hearing: The psychophysics of human sound localization", MIT Press, Cambridge, Mass. , 1997.

[4] H. Wallach: "The role of head movements and vestibular and visual cues in sound localization", J. Experimental Psychology, vol. 27, no. 4, pp. 339-368, 1940.

[5] F. L. Wightman, and D. J. Kistler: "Resolution of front-back ambiguity in spatial hearing by listener and source movement", J. Acoust. Soc. Am. , vol. 105, no. 5, pp. 2841-2853, 1999.

[6] W. E. Kock: "Binaural Localization and Masking", J. Acoust. Soc. Am. ,vol. 22, no. 6, pp. 801-804, 1950.

[7] I. Toshima, H. Uematsu, T. Hirahara: "A steerable dummy head that tracks three-dimensional head movement: TeleHead", Acoustical Science and Technology, vol. 24, no. 5, pp. 327-329, 2003.

[8] V. R. Algazi,R. O. Duda, and D. M. Thompson: "Motion-Tracked Binaural Sound", J. Aud. Eng. Soc., vol. 52, no. 11, pp. 1142-1156, 2004.

[9] I. Toshima, S. Aoki, T. Hirahara: "An acoustical tele-presence robot: TeleHead II", Proc. of International conference on intelligent robots and systems 2004(IROS2004), pp. 2105-2110, 2004.

[10] H. Møller: "Fundamentals of binaural technology", Applied Acoustics, vol. 36, pp. 171-218.

[11] K. A. J. Riederere: "Repeatability analysis of head-related transfer function measurements", 105th Audio Eng. Soc. Conv., no. 4846, 1998.

[12] D. N. Zotokin, R. Duraiswami, E. Grassi, and N. A. Gumerov: "Fast head-related transfer function measurement via reciprocity", J. Acoust. Soc. Am., vol. 120, pp. 2202-2215, 2006.
[13] E. M. Wenzel, M. Arruda, D. J. Kistler, and F. L. Wightman: "Localization using nonindividualized head-related transfer functions", J. Acoust. Soc. Am. Vol. 94, pp. 111-123, 1993.
[14] J. C. Middlebrooks: "Individual difference in external-ear transfer functions reduced by scaling in frequency", J. Acoust. Soc. Am., vol. 106, no. 3, pp. 1480-1492, 1999.
[15] 戸嶋巌樹，青木茂明，"音響テレプレゼンスロボットの頭部運動制御"，ロボティクス・メカトロニクス講演会，1A1-N-057，2005.
[16] 戸嶋巌樹，青木茂明，"音響テレプレゼンスロボットの頭部運動再現における聴覚的時間的余裕の定量的評価"，日本ロボット学会誌，Vol.25，No. 6，pp. 990-996，2007.
[17] 平原達也，植松尚，戸嶋巌樹，"頭部の３次元運動に追従するダミーヘッドシステム － テレヘッド －"，AI チャレンジ研究会 2002
[18] I. Toshima, S. Aoki, and T. Hirahara, "Sound Localization Using an Acoustical Telepresence Robot: *TeleHead II*", Presence, Vol. 17, No. 4, pp. 392-404, 2008.
[19] Y. Suzuki, F. Asano, H. Kim, and T. Sone, "An optimum computer-generated pulse signal suitable for the measurement of very long impulse responses", J. Acoust. Soc. Am. , vol. 97, pp.1119-1123, 1995.

MUSIC空間スペクトログラムを用いた複数音源の発話区間検出の検討

Investigation of utterance detection of multiple sound sources based on the MUSIC spatial spectrogram

○石井カルロス寿憲（ＡＴＲ知能ロボティクス研究所）
梁棟（大阪大学工学部，ＡＴＲ知能ロボティクス研究所）
石黒浩（大阪大学工学部，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, Liang DONG, Hiroshi ISHIGURO, Norihiro HAGITA (Intelligent Robotics and Communication Labs., ATR)

carlos@atr.jp, liang@atr.jp, ishiguro@ams.eng.osaka-u.ac.jp, hagita@atr.jp

Abstract - With the goal of improving human-robot speech communication, the localization of multiple sound sources in the 3D-space based on the MUSIC algorithm was implemented and evaluated in a humanoid robot embedded in real noisy environments. A method for tracking sound intervals of multiple sound sources was then proposed based on the sound directivity inferred from the MUSIC spectrogram. The proposed method achieved good sound interval detection accuracies and low insertion rates compared with previous sound localization results.

1 Introduction

In human-robot speech communication, the microphones on the robot are usually far (more than 1 m) from the human users, so that the signal-to-noise ratio becomes lower than for example in telephone speech, where the microphone is centimeters from the user's mouth. Due to this fact, interference signals, such as voices of other subjects close to the robot, and the background environment noise, would degrade the performance of the robot's speech recognition. Therefore, sound source localization and posterior separation become particularly important in robotics applications.

One of the difficulties that degrade speech recognition performances in the robot's real environment is the lack of accuracy in utterance detection. In the present work we propose the use of sound localization (or more specifically, sound directivity) for improving utterance detection.

There are many works about sound source localization [1]-[9]. The sound localization method adopted in the present work is the MUSIC (MUltiple SIgnal Classification) algorithm, which is a well-known high-resolution method for source localization [1]-[3]. However, there are two issues regarding the MUSIC algorithm, which constrain its application for sound localization in practice. One is the heavy computational cost, while the other is the need of previous knowledge about the actual number of sources present in the input signal.

Regarding evaluation, although there are many works related to sound localization, most of them only evaluate simulation data or laboratory data in very controlled conditions. Also, only a few works evaluate sound localization in the 3D space, i.e., considering both azimuth and elevation directions [8]-[9]. Looking at the user's face while the subject is speaking is also an important behavior for improving human-robot dialogue interaction, and for that, a sound localization in 3D space becomes useful.

Taking the facts stated above into account, in our previous work [10], we constructed a MUSIC-based 3D-space sound localization (i.e., estimation of both azimuth and elevation directions) in the communication robot of our laboratory, "Robovie", and evaluated it in real noisy environments. However, only the raw data (without sound interval segmentation) was evaluated. Also, there still are considerable insertion error rates, for getting high detection rates.

In the present work, we propose and evaluate a sound interval detection method, by tracking sound sources directly on the MUSIC spectrogram and delta-spectrogram.

This paper is organized as follows. In Section 2, descriptions about the hardware and data collection are given. In Section 3, the proposed method is explained, and in Section 4, analyses and evaluation results are presented. Section 5 concludes the paper.

2 Hardware and data description

2.1 The microphone array

A 14-element microphone array was constructed in order to fit the chest geometry of Robovie, as shown in Fig. 1.

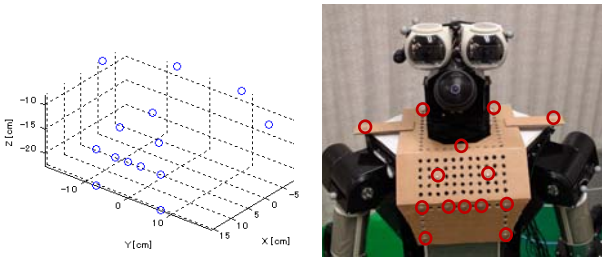


Fig. 1. (a) The geometry of the 14-element microphone array. (b) Robovie wearing the microphone array.

The chest was chosen, instead of the head, due to geometric limitations of Robovie's head. Several 3D array architectures were tested using simulations of the MUSIC algorithm. The array geometries were designed in such a way to cover all three-dimensional coordinate axes, giving emphasis to resolution in azimuth direction, and sounds coming from the front. The array configuration shown in Fig. 1 was chosen since it produced fewer side-lobes and had a fairly good response over different frequency bins.

A 16-channel A/D converter TD-BD-16ADUSB from Tokyo Electron Device Limited was used to capture the signals from the array microphones. Sony ECM-C10 omni-directional electret condenser microphones were used as sensors. Audio signals were captured at 16 kHz and 16 bits.

2.2 Recording setup

The microphone array was set on the robot's chest structure, as shown in Fig. 1. The robot was turned on to account for the noise produced by its internal hardware. The sources (subjects) were positioned around the robot in different configurations and were instructed to speak to the robot in a natural way. Each subject had an additional microphone to capture their utterance. The signals from these additional microphones, which we will call "**source signals**" throughout the paper, will be used only for analysis and evaluation. Nonetheless, the source signals are not required by the proposed method in its final implementation.

2.3 Data collection and environmental conditions

Recoding data using the microphone array was collected in two different environments. One is an office environment (OFC), where the main noise sources are the room's air conditioner and the robot's internal hardware noises. The second environment is a hallway of an outdoor shopping mall (called Universal City Walk Osaka – UCW), where a field trial experiment has been executed [11]. The main noise source in UCW was a loud pop/rock background music coming from the loudspeakers on the hallway ceiling. The ceiling height is about 3.5 meters. Recordings were done with the robot faced to different directions, in several places.

In OFC, four sources (male subjects) are present. At first, each source speaks to the robot for about 10 seconds, as the others remain silent. In the last 15 seconds of the recording, all four sources speak at the same time. For this recording, two of the subjects wore microphones connected to the two remaining channels of the 16-channel A/D device, while the other two subjects wore microphones connected to a different audio capture device (M-audio USB audio). A clap at the beginning of the recording was used to manually synchronize the signals of these two speakers to the array signals. It is worth to mention that a strict synchronization between the source signals was not necessary, because only power information of the source signals will be used, as will be explained in Section 2.4.

In UCW, there are two speech sources (male subjects) present in all recordings. In most of the trials, the sources take turns to speak for about 10 seconds each and then proceed to talk at the same time. In two of the trials (UCW7 and UCW8), one source is moving and the other is static, both speaking at the same time most of time. In five trials (UCW1-4, UCW9), the robot is far from the ceiling loudspeakers, while in four trials, the robot is close (a few meters) to a loudspeaker (UCW5-8), and in another four trials, the robot is right under a loudspeaker (UCW10-13). All trials have different configurations for the robot facing direction and/or source locations.

2.4 Computation of the reference number of sources from the power of the source signals (PNOS)

The number of sources (**NOS**) is an important parameter required by the MUSIC algorithm, which influences on the performance of DOA estimation. For analysis and evaluation of the NOS in the DOA estimation performance, reference NOS were computed from the power of the source signals. These power-based NOS values will be referred as **PNOS**.

Prior to compute the power of each source, a cross-channel spectral binary masking was conducted among the source signals in order to reduce the inter-channel leakage interferences, and get more reliable reference signals. In addition, the signal of the microphone in the center position in the array was used to remove the ambient music noise from all the source signals. Finally, the signal was also manually attenuated in the intervals where interference leakage persisted after the above processing. This resulted in much clearer source signals.

The average power of the signal was computed for each 100 ms, which corresponds to the block interval used in the MUSIC algorithm. A threshold was manually adjusted to discriminate the blocks with sound activity for each source signal. For each block, PNOS is then given by the summation of the source signals with activity.

In the UCW recordings, an additional source (due to the background music) was added to PNOS.

3 Proposed method

3.1 The broadband MUSIC spectrum

Fig. 2 shows the block diagram of the algorithm for computing the broadband MUSIC spectrum. The algorithm structure is similar to a classical approach of the MUSIC algorithm: getting the Fourier transform (FFT) for computation of the multi-channel spectrum, computing the cross-spectrum correlation matrix, making the eigenvalue decomposition of the averaged correlation matrix over a time block, computing the (narrowband) MUSIC responses for each frequency bin using the eigenvectors corresponding to the noise subspace and the steering/position vectors prepared beforehand for the desired search space, and finally computing the broadband MUSIC response by averaging the narrowband responses over a frequency range.

The broadband MUSIC responses are referred as MUSIC spectrum, while the sequence of the MUSIC spectrums along the time is referred as MUSIC spectrogram.

In our previous work, some of the parameters related to the MUSIC response computation were analyzed in order to obtain real-time processing, while keeping the DOA (direction of arrival) estimation performance. These parameters related to the MUSIC computation are described in detail in the following sub-sections 3.1.1 to 3.1.4.

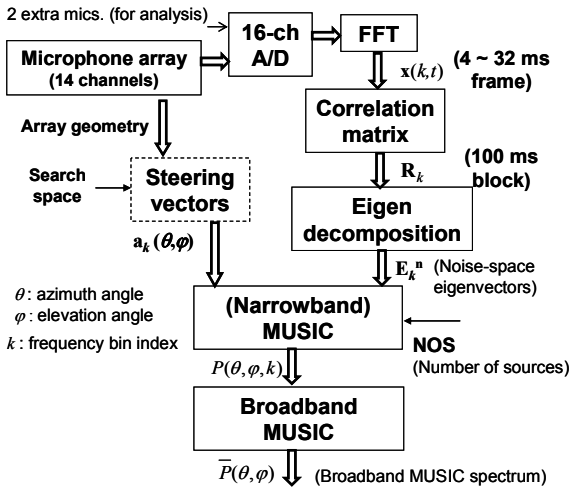


Fig. 2. The MUSIC-based sound localization algorithm, and related parameters.

3.1.1 Search space for DOA (directions of arrival)

The MUSIC algorithm was implemented to obtain not only the azimuth but also the elevation angle of the direction of arrival (DOA) of each source signal. Since the goal of this development is to enhance the human/robot interaction, we considered that it was not necessary to estimate the distance between the robot and the source(s) and that the DOA was the important piece of information. Nonetheless, the MUSIC algorithm can easily be extended to estimate also the distance between the array and the source, by adding the corresponding steering/position vectors. However, this would considerably increase the processing time.

A spherical mesh with a step of 5 degrees was constructed for defining the directions to be searched by the MUSIC algorithm. The mesh was constructed by setting elevations in intervals of 5 degrees, and setting different number of azimuth points for each elevation. The number of azimuths is maximum for 0 degrees elevation (having 5 degrees azimuth intervals), and gradually reduces for higher elevations, in such a way that the arc between two points is kept as close as possible to the arc corresponding to 5 degrees azimuth in 0 degrees elevation. This reduces the number of directions to be scanned by the MUSIC algorithm, reducing computation time. The directions with elevation angles lower than -30 degrees were also removed to speed up the computation, resulting in a total of 1216 directions.

The origin of the coordinate frame is set to the intersection point of the rotational axis of the degrees of freedom of the Robovie's head. This way, the output from the DOA estimation algorithm can be directly used to servo the head.

3.1.2 Frame length and block length

The frame length, which is related to the number of FFT points to be computed in the first stage, is an important parameter that can drastically reduce the computational costs of the MUSIC algorithm. Although FFT of 512 ~ 1024 points is commonly used (corresponding to 32 ~ 64 ms frame length at 16 kHz), we have proposed the use of smaller FFT sizes (64 ~ 128). This allows reducing the computation not only of the FFT stage, but also of the subsequent correlation matrix, eigenvalue decomposition, and MUSIC response computations. We have found that reducing the frame size to 4 ms (or equivalently reducing the FFT size to 64) was effective to allow real-time processing without a big degradation in the estimation of the directions of arrival (DOA) of sound sources.

In the next step of the MUSIC algorithm, a correlation matrix is averaged for the frames within a time block. The block length has to be long enough for getting good estimation of the averaged correlation matrix. On the other hand, it also should be short enough for getting good temporal resolution (considering that a sound source can move) and low latency. In the present work, we decided to use a time block length of 100 ms.

3.1.3 Frequency range of operation

Although speech contains information over a broad frequency band (vowels in 100 – 4000 Hz and fricative consonants in frequencies above 4000 Hz), the frequency range of operation for DOA estimation has to be limited, given the geometric limitations of the array (shown in Fig. 1).

The smallest distance between a pair of microphones is 3 cm, so that on theory the highest frequency of

operation to avoid spatial aliasing would be about 5.6 kHz (according to Rayleigh's Law).

Regarding the lowest frequency boundary, although speech contain important information in frequency bands lower than 1 kHz, the array geometry limitations do not allow good spatial resolution in these low frequency bands. In the present work we use the frequency range of 1 – 6 kHz, for avoiding the issues above.

3.1.4 Number of sources (NOS)

The number of sources is an important parameter necessary for getting a good estimate of the MUSIC spectrum. In theory, there is some relationship between the number of sources and the shapes of the eigenvalue profiles. However, a threshold between strong and weak eigenvalues is difficult to be determined. The environment noise has also a strong impact on the shapes of the eigenvalues, so that both magnitude and slope of the profiles are affected.

Considering the difficulties in estimating NOS from the eigenvalues of the spatial correlation matrix, we have proposed the use of a fixed number of sources for the (narrowband) MUSIC response computation ("fixed NOS"), and establishing a maximum number of sources detectable from the broadband MUSIC response ("max NOS"). Here, we allow the maximum number of sources detectable being larger than the fixed number of sources for the MUSIC response computation. This idea is based on the assumptions that at an instant time, the predominance of different broadband sound sources varies depending on the frequency bins. Therefore, even if the NOS used to the narrowband MUSIC computation is limited to a fixed small number, the combination of frequency bins to compute the broadband MUSIC spectrum may produce more peaks than the fixed number.

Fig. 3 shows examples of MUSIC spectrograms for OFC1, where 4 sources are speaking in front of the robot, UCW8, where one of the sources is speaking in front of the robot, while the other speaker is moving in front of the robot while speaking, and a directional music source is present in the first half of the trial, and UCW13, where the robot is right under a loudspeaker. A suitable plotting of the MUSIC spectrogram is difficult because there are several dimensions: azimuth angle, elevation angle, MUSIC power and time. In the MUSIC spectrogram of Fig. 3, we show the azimuth angle in the y-axis, time in x-axis, different colors for different elevations, and different tonalities according to the MUSIC power. (The colors can be viewed in the electronic version.) Note that in the middle panel of Fig. 3, the music source in the first half appears in green, because the loudspeakers are in higher elevations compared to humans, while there are strong lines in pink/red in the bottom panel of Fig. 3, because there is a loudspeaker over the robot.

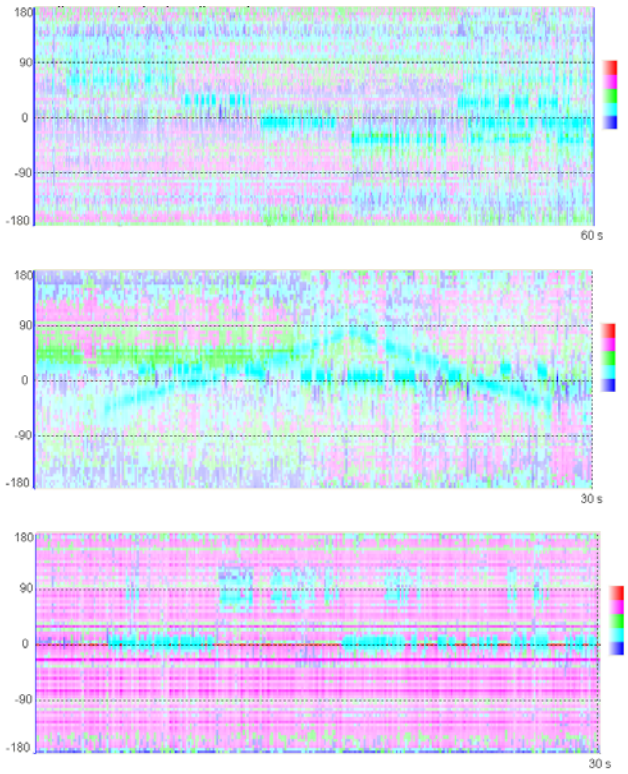


Fig. 3. Examples of MUSIC spectrograms for OFC1 (where 4 sources are present in front of the robot), for UCW8 (where one of the sources is moving in front of the robot), and for UCW13 (where the robot is right under a loudspeaker). The elevation angles are displayed by different colors. (Please refer to the electronic document to see the colors.)

3.2 Sound interval detection from the MUSIC spectrogram

In a classical approach for determining the direction of arrival (DOA) of the sound sources, peak picking is realized on the MUSIC spatial spectrum. In the present work, we proposed a method for detecting sound source intervals, based on a MUSIC spectrogram and delta-spectrogram.

Fig. 4 shows a block diagram of the proposed method for sound interval detection.

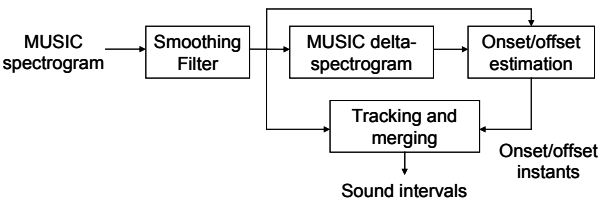


Fig. 4. The proposed sound interval detection based on MUSIC spectrogram and delta-spectrogram.

First, a moving-average smoothing filter is passed through each direction (azimuth vs. elevation) of the raw MUSIC spectrogram, by a Hamming window of 5 taps, for reducing temporal distortions.

Then, a delta-spectrogram is computed by taking, for each direction (azimuth vs. elevation), the minimum difference between the MUSIC power values of the

current block and the neighbor directions in the previous block.

Thresholds are set for estimation of onset and offset instants. First, thresholds are imposed to the positive and negative peaks of the MUSIC delta-spectrogram, to get raw estimates of onset and offset instants. To avoid over-estimation of the number of sources, we set a threshold for the magnitude of the MUSIC power, as proposed in our previous work. Further, the maximum number of onsets per block is constrained, assuming that the probability of multiple sources starting at the same instant is low. This last constraint is in particular important to reduce insertion errors due to sidelobe effects.

Also, to avoid misdetection of the offset instant, we forced offset if the MUSIC power becomes lower than the MUSIC power of the block previous to the onset instant (onset MUSIC power) plus a bias factor alpha. The following summarizes the thresholds involved in the onset/offset detection.

- Onset: [MUSIC delta power > 1.0 dB] and [MUSIC power > 1.8 dB] and [Max. number of onsets per block <= 2]
- Offset: [MUSIC delta power < -1.2 dB] or [MUSIC power < onset MUSIC power + alpha]

Finally, the path with maximum MUSIC power is tracked in the MUSIC spectrogram from the onset to the offset instant. Segments separated by short pauses smaller than 4 blocks (400 ms), and with continuity in the direction are merged.

4 Analysis and experimental results

4.1 The evaluation setup

To measure the performance of the DOA estimation, we used three scalar values. The first represents the percentage of ideal DOA that were detected successfully by the algorithm. We will call this quantity "**DOA accuracy**". The second represents the number of additional sources (insertions) that were detected, on average, per time block. We will call this quantity "**DOA insertion rate**".

To get the ideal DOA of the sources, we used information about the sound source activity (obtained from the power of the source signals – Section 2.4) and raw estimates of the DOA obtained by using the ideal number of sources (PNOS). Piecewise straight lines were fit to the contours of the raw DOA estimates in the intervals where each source is active. Video data were also used to check the instants where a source is moving.

4.2 Analysis of DOA estimation in different trials

Fig. 5 shows the DOA estimation performances (accuracies and insertion rates) for individual trials in office (OFC) and shopping mall (UCW) environments, for several parameter conditions related to the sound interval detection logic: with/without smoothing filter, with/without the threshold for forcing offset, with/without maximum number of onsets, and before/after tracking. For the computation of the MUSIC spectrogram, the following parameters were used: NFFT = 64, frequency range = 1 – 6 kHz, and fixed NOS = 2. Although larger NFFT provide slightly better performances, we used 64, because real-time can be achieved even when running in a 2GHz Centrino CPU.

The average DOA accuracies for speech sources are shown in the left part of the top panel in Fig. 5, and the DOA insertion rates are shown in the middle panel of Fig. 5, for each experimental condition and for each trial in OFC and UCW.

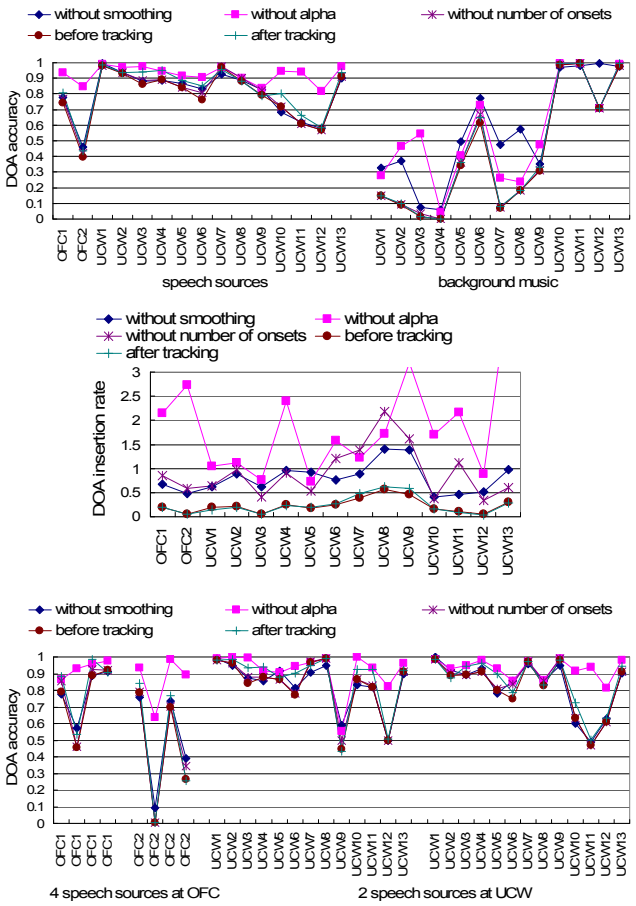


Fig. 5. DOA estimation performances (accuracies and insertion rates) for each source and each trial in OFC and UCW. For all trials, NFFT = 64, frequency range = 1 – 6 kHz, fixed NOS = 2, MUSIC delta-spectrogram Onset threshold = 1.0 dB, Offset threshold = -1.2 dB, and MUSIC power threshold = 2.0 dB.

We can first observe that "without alpha" (forcing an offset according to the onset MUSIC power plus a bias factor alpha) gives the best DOA accuracies. However, it also gives the worst insertion rates. The DOA accuracies of "without smoothing", "without number of onsets" and "before tracking" are very similar. However, a clear reduction in the insertion rate can be observed for

"before tracking", showing the effectiveness of both smoothing procedure and constraining the number of onsets per block. Finally, comparing "before tracking" and "after tracking", a slight improvement is observed in DOA accuracy for UCW3-6 and UCW10-11, while only a very small increasing in DOA insertion rate is observed in UCW8-9.

Regarding the ambient music sources, it can be observed in the right side of the top panel in Fig. 5 that the DOA accuracies were low in UCW1-4 and UCW7-9, since the robot was relatively far from the ceiling loudspeakers. DOA accuracies were almost 100 % in UCW10-13, when the robot was right under one of the loudspeakers, where the background music can be clearly considered as a directional source, while DOA accuracies show intermediate detection rates for UCW5-6, where the robot was relatively closer to one of the ceiling loudspeakers.

Regarding performances for individual sources, it can be observed in the bottom panel of Fig. 5 that the second and fourth sources in OFC2, the first source in UCW9, and the second source in UCW12 show lower DOA accuracy. An explanation is that these sources come from the back side of the robot, so that both power and directivity are lower than the sources coming from the front side.

5 Conclusion and Future works

Sound interval detection of multiple sound sources using a 3D-space sound directivity based on the MUSIC spectrogram was implemented and evaluated in our humanoid robot embedded in real noisy environments.

Evaluation of the proposed method showed lower insertion rates could be achieved. However, the detection accuracies also degraded in some of the trials where the loudspeaker is right over the robot. We are currently investigating the reasons of this degradation.

Finally, although the goal of the present work is to detect speech intervals, the technology here presented is to detect sound intervals. However, in robot applications, other modalities, such as vision, can be used to determine if the detected sound is speech or not. This will be scope of our future work. We are also planning the implementation and evaluation of sound source separation algorithms using the localization and sound interval detection results from the present work.

Acknowledgement

This work was supported in part by the Ministry of Internal Affairs and Communication, and by the Ministry of Education, Culture, Sports, Science and Technology.

References

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.
11) T. Kanda, D. F. Glas, M. Shiomi, H. Ishiguro, and N. Hagita, "Who will be the customer?: A social robot that anticipates people's behavior from their trajectories," *Tenth International Conference on Ubiquitous Computing* (*UbiComp 2008*), 2008.

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

人工知能学会研究資料
JSAI Technical Report
SIG-Challenge-A902-4

境界要素法を用いた音響解析による耳介形状の検討

Examination of Pinna Shape by Acoustic Analysis based on Boundary Element Method

公文誠 石飛光章
Makoto KUMON and Mitsuaki ISHITOBI
熊本大学
Kumamoto University
kumon@gpo.kumamoto-u.ac.jp

Abstract

Pinna plays an important role to provide the direction of the sound source as frequency domain auditory cues, and its shape is said to have strong influence on how it encodes the direction. As the shape of the pinna is complicated, it is not trivial to form the shape to fit the objectives for auditory robots. This paper provides a method to optimize its shape through numerical simulations based on boundary element method that was proposed to estimate Head Related Transfer Functions by Otani et.al. In order to realize efficient optimization procedure, the gradient of the cost function is approximated by Simultaneous Perturbation Stochastic Approximation method instead of computing the gradient itself since it is computationally heavy. An example of this optimized pinna is also shown in this paper.

1 はじめに

人と共存するロボットを実現する上で, 人との対話や電話のベルのような音記号を認識する聴覚機能は必須の機能の一つであり, 音源の位置や方向を認識する音源定位はこの聴覚の重要な能力の一つである. これを実現する方法に, 複数の受聴点で観測された音信号をもとに受聴点間の音信号の到達時間差により方向を推定する方法がある. しかし, マイク数が制限される場合, 到達時間差による方法が適用できないことがある. 例えば, 人や猫などの動物では 2 つしか耳を持たず, 前述のアプローチでは正中面の音源方向を区別できない. 一方, 人間などの動物は耳介に複雑な形状を有し, 耳介が反射や回折によって音信号に対して方向依存性のフィルタとして働くことが知られている [Shaw, 1968]. このため, 音源の周波数特性など一定の情報が得られれば, 周波数領域での特徴量を用いた音源方向の推定が可能である. 特にゲインが急峻に低減する帯域は耳介ノッチと呼ばれ, 音源方向の関数になっていることが知られている.

ところで, ロボットにおいて同様の機構を考えた時, その形状をどのように設計すれば良いかは自明ではない. 本研究では, 耳介が実現すべき機能として周波数特徴量に音源方向を精度良く埋め込むことを目的とし, 所望の周波数特性に近い特性を持つ耳介形状を設計することを考える. 周波数特性から耳介形状を得ることは容易ではないため, 与えられた耳介形状から周波数特性を求め, これを所望の特性に対して最適化することで形状を設計する. このアプローチでは形状から周波数特性を繰り返し計算する必要が生じる. 大谷らの境界要素法を用いた頭部伝達関数 (Head Related Transfer Function, HRTF) 計算手法 [Otani, 2006] は計算に効率が良く, 本研究の目的に適している. また, 耳介の複雑な形状を十分に表現するためには, 大自由度の形状パラメータが必要になるが, これらの値を決定するためには大規模の最適化問題を解くことになる. ここでは大自由度の最適化計算に利点のある同時摂動確率近似 [Spall, 1992] により勾配を近似的に求める最適化手法を適用する.

2 耳介の周波数特性

2.1 耳介ノッチ

耳介はその複雑な形状から, 音到来方向に依って周波数特性が変化することが知られており, 特にゲインが急峻に減少するノッチは音到来方向の手がかりとして動物が活用していると言われる [Batteau, 1967]. しかしながら, 通常耳介ノッチと音源方向との間には複雑な関係があり, また複数のノッチが互いに重なり周波数帯域に存在することなどから, 耳介ノッチそのものの検出は容易ではない. 図 1 に Shimoda ら [Shimoda, 2006] の用いた耳介での周波数応答の一例を示すが, この図の情報だけからノッチを適切に検出することは不可能で, Shimoda らは周波数応答の時間的変化と事前知識を組み合わせる方法を考えている.

ノッチ周波数と音源方向の関係がシンプルで, ノッチ形

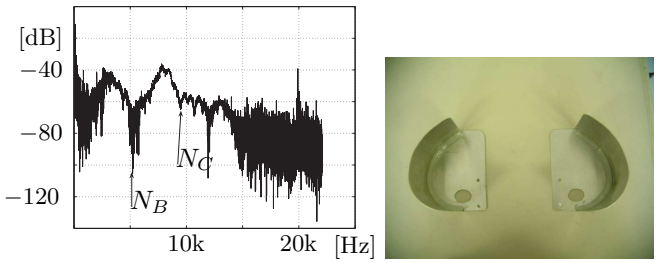


Figure 1: 耳介ノッチの例 (Shimoda らによる [Shimoda, 2006])
周波数応答 (左図) およびロボット用耳介 (右図)

状がはっきりしたものであれば, より簡単な方法で精度良く音源方向が推定できると考えられる. これらの特性は耳介の形状によって定まるため, 適切な耳介形状を得ることが目的に適う耳介を設計することになる.

2.2 境界要素法にもとづく計算法

ここでは Otani らによって提案されている境界要素法による HRTF の計算法 [Otani, 2006] の概略を示す. 演算の詳細等は文献を参照されたい.

HRTF は音源から受聴点の間に頭部などの境界を持つる場において, ヘルムホルツ方程式

$$\nabla^2 P(\boldsymbol{x}) + k^2 P(\boldsymbol{x}) = f \tag{1}$$

を解くことで求めることができる. ここで $P(x)$ は音圧, k は波数, f は外力を表わす. 今,(1) を小要素によって離散化した場合 (要素数を M とする), 各要素での音圧の関係は次のようになる.

$$\left(\frac{1}{2}\boldsymbol{I}_M + \boldsymbol{G}_n + j\omega\rho\boldsymbol{G}\boldsymbol{Y}\right)\boldsymbol{P} = \boldsymbol{g}(\omega) \tag{2}$$

ここで $\boldsymbol{P}$ は要素上の音圧からなる音圧ベクトル, $\boldsymbol{I}_M$ は M 次の単位行列, ρ は媒質密度, ω は角周波数を表わし

$$\begin{aligned}
\boldsymbol{G} &= \left[\int\int_{S_j} G(\boldsymbol{x}|\boldsymbol{x}_j)dS\right] \in \boldsymbol{C}^{M\times M} \\
\boldsymbol{G}_n &= \left[\int\int_{S_j} \frac{\partial}{\partial n}G(\boldsymbol{x}|\boldsymbol{x}_j)dS\right] \in \boldsymbol{C}^{M\times M} \\
\boldsymbol{g} &= [G(\boldsymbol{x}_i|\boldsymbol{x}_j)] \in \boldsymbol{C}^M
\end{aligned}$$

である. ただし, S_j は j 番目の要素を示し, $G(\boldsymbol{x}|\boldsymbol{y})$ は $\boldsymbol{x}$ と $\boldsymbol{y}$ における Green 関数, $\frac{\partial}{\partial n}$ は要素の法線方向への微分を表すものとし, $\boldsymbol{Y}$ は各要素の音響アドミタンスを要素に持つ対角行列とする.

音圧ベクトル $\boldsymbol{P}$ を用いれば, 受聴点 $\boldsymbol{s}$ における音圧 $P_s(\omega)$ は

$$P_s(\omega) = g_s(\omega) - (\boldsymbol{G}_{ns} + j\omega\rho\boldsymbol{G}_s\boldsymbol{Y})\,\boldsymbol{P} \tag{3}$$

と表わせる. ここで

$$\begin{aligned}
\boldsymbol{G}_s &= \left[\int\int_{S_j} G(\boldsymbol{x}|\boldsymbol{s})dS\right] \in \boldsymbol{C}^M \\
\boldsymbol{G}_{ns} &= \left[\int\int_{S_j} \frac{\partial}{\partial n}G(\boldsymbol{x}|\boldsymbol{s})dS\right] \in \boldsymbol{C}^M \\
\boldsymbol{g}_s &= [G(\boldsymbol{o}|\boldsymbol{s})] \in \boldsymbol{C}
\end{aligned}$$

であり, $\boldsymbol{o}$ は音源位置を表す.

音源の各位置について以上を計算すれば, HRTF を得ることができる. (2) における線形方程式を繰り返し解くことになり計算コストが大きいが, Otani らによれば, 以下の方法でこの計算を効率化できる. HRTF の計算では境界要素と受聴点の幾何学的関係は不変であるので, (2) において境界要素・受聴点の関係のみに依存するベクトル $\boldsymbol{Q}$ を

$$\boldsymbol{Q} = (\boldsymbol{G}_{ns} + j\omega\rho\boldsymbol{G}_s\boldsymbol{Y})\left(\frac{1}{2}\boldsymbol{I}_M + \boldsymbol{G}_n + j\omega\rho\boldsymbol{G}\boldsymbol{Y}\right)^{-1} \tag{4}$$

とおくと, この $\boldsymbol{Q}$ を用いて (3) を

$$P_s(\omega) = g_s(\omega) - \boldsymbol{Q}\boldsymbol{g}(\omega) \tag{5}$$

と表わせる. (4) は $\boldsymbol{Q}$ について M 個の線形方程式から成るので, $\boldsymbol{Q}$ を求めるには右辺の逆行列を求める必要はなく, 適当な求解法を用いることで演算量を減らすことができる. また, $\boldsymbol{Q}$ は音源位置に依らず, g_s および $\boldsymbol{g}$ は容易に計算できることから, 様々な音源位置に対して $P_s(\omega)$ を計算する際, 線形方程式を一度解けば十分で計算効率が良い.

本研究ではこの HRTF 計算法における境界要素の配置を耳介に合わせることで, 耳介の周波数特性の計算法とする.

2.3 設計パラメータに対する感度 : 勾配

一般に境界要素法などを用いた形状最適化には, 形状等を指定するパラメータに対する評価関数の勾配 (感度) を計算し, 適切なパラメータ更新を行う. 勾配を求めることを考え, (5) を各境界要素の位置 x_i について微分すると

$$\frac{\partial}{\partial x_i}P_s(\omega) = \frac{\partial}{\partial x_i}g_s(\omega) - \left(\frac{\partial}{\partial x_i}\boldsymbol{Q}\right)\boldsymbol{g}(\omega) - \boldsymbol{Q}\left(\frac{\partial}{\partial x_i}\boldsymbol{g}(\omega)\right)$$

となる. g_s および $\boldsymbol{g}$ の微分は容易に実行できるが, 右辺第 2 項は計算量の観点から注意を払う必要があるが,

$$\begin{aligned}
\boldsymbol{A} &= \frac{1}{2}\boldsymbol{I}_M + \boldsymbol{G}_n + j\omega\rho\boldsymbol{G}\boldsymbol{Y} \\
\boldsymbol{B} &= \boldsymbol{G}_{ns} + j\omega\rho\boldsymbol{G}_s\boldsymbol{Y}
\end{aligned}$$

とすれば,

$$\frac{\partial}{\partial x_i}\boldsymbol{Q}\boldsymbol{g}(\omega) = \left\{\frac{\partial}{\partial x_i}\boldsymbol{B} - \boldsymbol{Q}\frac{\partial}{\partial x_i}\boldsymbol{A}\right\}\boldsymbol{A}^{-1}\boldsymbol{g}(\omega)$$

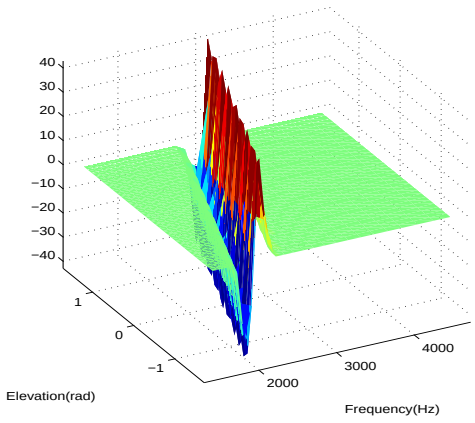


Figure 2: dP_d の例

となり, 再度 (4) を解くのと同等の計算量が生じることが分かる. また, 上記をパラメータの数だけ繰り返し計算することになる. このことから, 大自由度のパラメータを持つ最適化問題に対して上記の勾配を直接計算する方法は計算コストが大きく現実的ではない. 以下では, 勾配を直接求めない計算方法を示す.

3 形状の最適化

3.1 設計目的と評価関数

本研究では耳介形状を適切に選ぶことで周波数特性を所望の特性に近づけることを考える. ここでは, 理想とする周波数特性として明瞭な耳介ノッチを有するものを考え, 以下のように周波数帯域 f_l から f_h の間で音源方向に対して線形に耳介ノッチ周波数 f_n を持つものとする.

$$f_n(\theta) = \frac{f_l + f_h}{2} + \frac{f_h - f_l}{\pi}\theta \tag{6}$$

θ は音源方向を表す.

また, 所望の音圧特性 $P_d(\omega)$ は以下のモデルでノッチを表現する.

$$P_d(\omega, \theta) = 1 - \exp\left(-\frac{(\omega - f_n(\theta))^2}{2\sigma^2}\right) \tag{7}$$

ここで σ はノッチの広がりを与える定数. ただし, 音圧の大きさそのものは重要ではないため, (7) を ω についての微分に比例する次の

$$dP_d(\omega, \theta) = C_{dP}(\omega - f_n(\theta)) \exp\left(-\frac{(\omega - f_n(\theta))^2}{2\sigma^2}\right) \tag{8}$$

を規範とする (dP_d の一例を図 2 に示す). ただし, C_{dP} は適当に定める係数. 同様に音圧の周波数についての微分を $dP(\omega, \theta)$ と表わせば, 形状最適化のための評価関数として

$$J = \sum_{\theta \in \Theta} \sum_{\omega \in \Omega} \|dP(\omega, \theta) - dP_d(\omega, \theta)\| \tag{9}$$

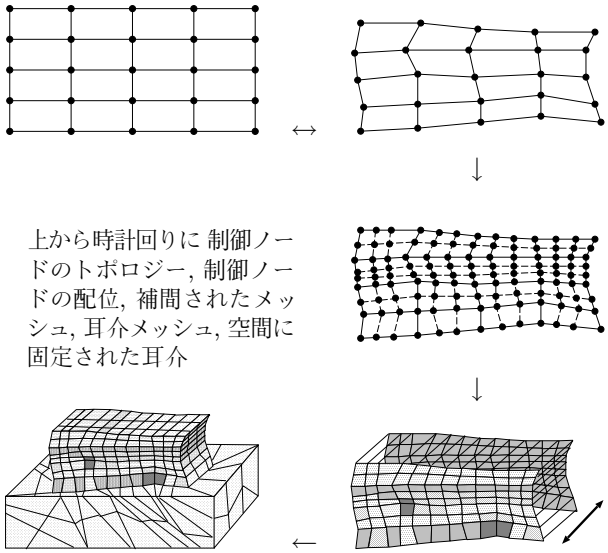


Figure 3: 耳介の形状モデル

を考えることができる. ここで Θ および Ω はそれぞれ考慮する音源方向および周波数成分からなる集合を示す.

3.2 耳介形状のモデル

十分な精度で周波数特性を計算するには, 耳介の境界要素に多数の小さな要素を用いる必要があるが, 要素数が増えれば形状パラメータ数が増えるため計算量の点で望ましくない.

ここでは, 少ない点数の粗い 2 次元格子を成す制御ノードを形状パラメータと考え, 周波数特性の計算には制御ノードを補間した後, 3 次元形状に変形することで耳介を表現することとした. 詳細を以下に示す.

1. 2 次元格子上 i, j の点における制御ノードを c_{ij} と表し, c_{ij} はノードの 3 次元座標を保持するものとする. このため, 最適化する形状パラメータは $3N$ 個 (N は制御ノードの数) である.
2. 周波数特性を計算するための細かい境界要素には制御ノード間を適当な間隔で細分化し, 多項式近似によって滑らかに補間した小要素からなる 2 次元メッシュを作成する.
3. 耳介は薄板を折り曲げた構造と考えられるため, 得られたメッシュを各要素の法線方向の正負に適当な距離移動させた点を耳介表面上の点とし, 耳介表面を小要素で表現する (耳介メッシュ).
4. 耳介を適当な位置に配置するため, 空間に固定した長方形型のメッシュと耳介メッシュを適当な方法で接続する.

3.3 勾配の同時摂動確率近似と形状最適化

前述の通り, 評価関数 J の設計パラメータに対する勾配を直接求めることは, 計算量の観点から現実的ではない. わずかに異る設計パラメータの組に対して周波数特性を評

価し, これらの差分によって勾配を近似的に求めることが考えられるが, $3N$ 個のパラメータ全てについて差分を求めることは, 結局 $3N$ 回 J の差分を評価することになり, 勾配を求めるのと同程度の計算量が必要で, 問題の解決にならない. そこで, 少ない評価回数でパラメータ更新を行う勾配の同時摂動確率近似 (Simultaneous Perturbation Stochastic Approximation) による最適化手法 (以下 SPSA と略記) を利用する.

SPSA では, まず以下のようにパラメータの全ての要素に同時に摂動を与える.

$$\boldsymbol{c}_{ij,k}^{\pm} = \boldsymbol{c}_{ij,k} \pm c_{p,k}\boldsymbol{\delta}$$

ここで, k は最適化計算のステップ数,

$$c_{p,k} = \frac{c_{p,0}}{(k+1)^{\gamma}}$$

であり, $c_{p,0}$, γ は適当な定数, $\boldsymbol{\delta}$ はどの要素も 0 にならないような適当な確率分布に従う確率変数からなる摂動ベクトルである.

この一対の摂動を受けたパラメータの組に対してパラメータ要素 i, j についての J_k の差分 $dJ_{ij,k}$ を

$$dJ_{ij,k} = \frac{J(\{c_{ij,k}^{+}\}) - J(\{c_{ij,k}^{-}\})}{c_{ij,k}^{+} - c_{ij,k}^{-}}$$

とする. これらを用いてパラメータを

$$\boldsymbol{c}_{ij,k+1} = \boldsymbol{c}_{ij,k} - a_{p,k}dJ_{ij,k} \tag{10}$$

と更新する. ここで

$$a_{p,k} = \frac{a_{p,0}}{(k+A+1)^{\alpha}}$$

であり, $a_{p,0}$, A, α は適当な定数.

以上のように最適化の各ステップでは 2 回 J を評価すれば良く, 大自由度の最適化問題や評価関数の計算コストの大きい場合有効な方法である. 最適解に辿りつくには多くのステップが必要になることがあるが, 本研究で扱う問題のように, 必ずしも最適解が必要なのではなく, 仕様を満たす解が求まれば十分である場合や, 局所的な改善を得ることが目的である場合は, SPSA に利点があると考えられる.

なお, SPSA の各パラメータは問題にあわせて適当に設定する必要がある. α, γ については [Spall, 1998, SPSA] によればそれぞれ 0.602, 0.101 にするのが良いとの指針があるので, 本研究でもこれらの値を用い, 残るパラメータについては試験的な最適化計算の結果をもとに適当な値を設定することとした. また, (10) より摂動が小さい場合更新ステップが大きくなり過ぎること, 対象とする問題によっては摂動が大きすぎると正しい勾配が求められないことが考えられる. 本研究では過大な更新を防ぐため (10) に替え

$$\boldsymbol{c}_{ij,k+1} = \boldsymbol{c}_{ij,k} - a_{p,k}\tanh(\beta dJ_{ij,k})$$

を更新則として用いることとした (β は適当な定数).

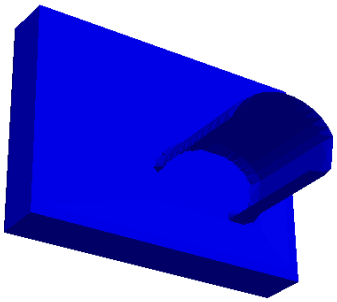


Figure 4: 耳介の初期形状

4 数値例

上述の方法で形状最適化を行った例を示す.

4.1 初期形状

Shimoda ら [Shimoda, 2006] に倣って, 対数らせん型の薄板状の耳介を初期形状として用いた (図 4). 耳介の高さは 61mm, 開口部が約 46mm 程度であり, 受聴点はらせんの中央付近の台座上 1mm の点とした. また, 板厚は 2mm である.

4.2 条件

ここでは, (6) における f_l, f_h をそれぞれ 1700Hz, 2300Hz とし, 2kHz を中心に方向に対して線形にノッチを得ることを目的とした.

Table 1: 最適化パラメータ

Θ :	[±1.5708, ±1.4708, ±1.3709, ±1.2707, ±1.1708, ±1.0708, ±0.9709, ±0.8708, ±0.7708, ±0.6706, ±0.5709, ±0.4708, ±0.3708, ±0.2702, ±0.1708, ±0.0709, 0](rad) (33 点)
Ω :	[1.5523, 1.6299, 1.7114, 1.7970, 1.8869, 1.9812, 2.0803, 2.1843, 2.2935, 2.4082, 2.5286, 2.6550, 2.7877, 2.9271, 3.0735](kHz) (15 点)
A : 2,	$a_{p,0}$: 9.0^{-4} $c_{p,0}$: 4.0×10^{-5} β :2×10^{-3}
C_{dP}: 0.5	σ : 100
繰り返し回数上限 : 100 回	

音源は耳介の開口部方向の半径 0.5m の半円上に配置し, 計算対象の周波数は最適化対象とする帯域を含むよう対数刻みで 1.5kHz から 3.0kHz 程度の帯域とした. 最適化計算に用いたパラメータ の値を表 1 にまとめた. また, 摂

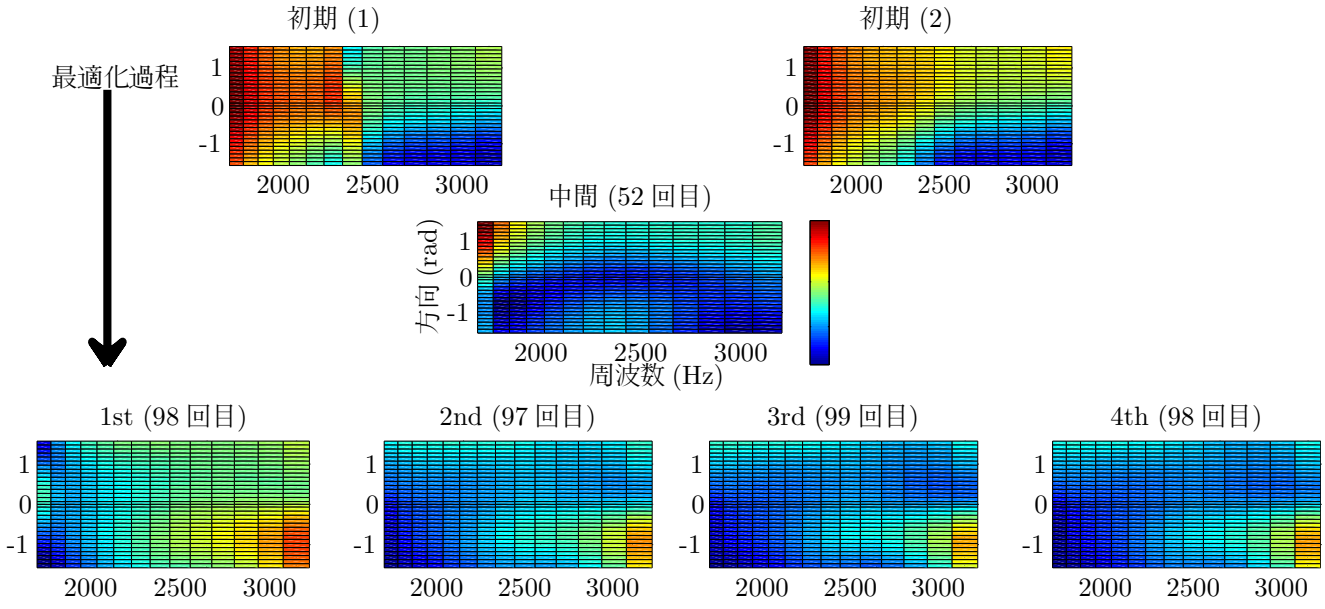


Figure 5: 最適化計算による周波数特性の変化

初期の周波数特性 (上段 2 図) から提案法による最適化計算を行ったところ, 中段 (52 回目における特性) のような特性を経て, 下段の 4 図のような特性が得られた. 下段の図は評価関数の最も良いものから順に示しており, 濃青がノッチ部分に相当する. 1700Hz から 2300Hz にかけて方向に対して右上がりの直線に沿ってノッチを得ることを目的に近づくよう計算されたことが分かる.

動ベクトル δ は 等確率のベルヌーイ分布に従う確率変数からなるベクトルとし, 制御メッシュは 4×20 とし, このうち 3×20 の点について最適化対象とする.

計算環境は Intel Core^{TM}2Quad Q9400 (2.66GHz) プロセッサ, 4GB メモリにおいて Ubuntu9.10 上の gcc-4.4.1 を用いてコンパイルしたコードを利用した. 線形方程式の演算には ATLAS 3.9.14[ATLAS] を用い, ATLAS の関数以外の一部のコードも OpenMP[OpenMP] によってを並列化している. 各要素の積分には 1 次三角要素について 7 点の Gauss-Legendre 積分を行ったが, 必要に応じて境界要素を再分割し, より小さな三角要素について計算した. また, 周波数特性の差分 dP は 3 点の中心差分によった.

4.3 結果と考察

提案法による設計計算を行い, 得られた周波数特性を図 5 に示す. 提案法では摂動を与えた耳介形状に対して周波数特性を計算するため, 耳介の初期形状に対して 2 つの特性が得られる. 評価関数値は摂動に伴って揺動するので, 100 回の繰り返し計算のうち, 最も評価値の良好のものから順に 4 つを示しているが, いずれも繰り返し計算の終盤 (97, 98, 99 回目) に得られたものである. なお, 摂動によって稀に良い評価を得られることもあり, 52 回目に得られた特性 (図中段) は最適化計算中 5 番目に良い評価値であった. また, 評価が最も良かった耳介形状の形状を図 6 に示す.

初期形状では 2kHz 付近にピークがあり周波数とともにゲインが漸減する特性であった. 提案法の計算によって,

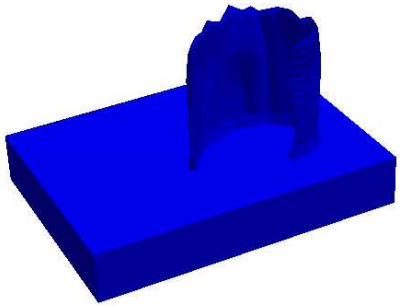


Figure 6: 得られた耳介形状

図の右下部分の領域にあった低ゲイン領域が左下部分に遷移し (52 回目の特性参照), さらに右下部分のゲインが向上する変化があった. この結果, 得られた特性では $-\frac{\pi}{2}$ の方向に対して 1700Hz 付近でゲインが下がっており, 評価値が 2 番目から 4 番目に良好であった特性では方向の増分に応じノッチ様の帯域が高周波領域へと遷移する構造となった. これは所望の特性の傾向を反映した結果である. しかし, 高周波数帯域での低ゲイン領域が広がっており, 耳介ノッチとして十分な形状を得ているとは言い難い. これに対し最も評価の良かった特性では, 方向が 0 から $\frac{\pi}{2}$ の領域でノッチ周波数は減少してはいるものの, 低ゲイン領域が狭く保たれノッチの形成に成功している.

設計過程の性能として演算時間を考えることができる. 図 7 に各繰り返しステップにおける CPU 利用時間および実際の計算時間を示す. 計算開始時はメモリアロケーショ

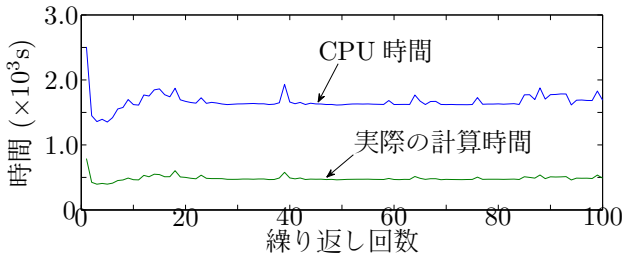


Figure 7: 演算時間

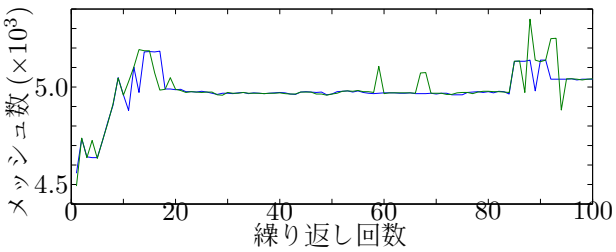


Figure 8: 境界要素数の変化

ンなどに伴うオーバーヘッドがあり計算時間を要しているが, 以降はほぼ均一の計算時間で処理している. 15 の周波数点で特性を求めているため, 1 回の更新に 30 回の周波数特性を計算する. 計算時間の変動があまり見られなくなった 20 回目以降の計算について考えれば, 各繰り返しに必要な計算時間は実時間で平均 390.2 秒, 並列化率は 3.45 であった.

また本方法の必要とする記憶領域は境界要素のメッシュ数の二乗のオーダに支配される. 制御メッシュの更新に伴い境界要素法のメッシュ数が変化するが, メッシュ数を示した図 8 によれば, 大幅なメッシュ数の変動はなく, 限られた記憶容量のもとで計算を実行できたことが分かる. より大規模な対象に対しても, 耳介メッシュの生成方法を工夫することで改善を図ることが可能であると考えられるが, メッシュの再計算には時間コストがかかるためトレードオフを考慮する必要がある.

5 おわりに

本研究では, 境界要素法を用いた HRTF 計算法を用いて耳介の周波数特性を計算し, 同時摂動確率近似による最適化手法によって所望の周波数特性に近い耳介形状を求める方法を提案した. 簡単な構造の耳介を例に提案法を適用したところ, 所望の特性の傾向を反映するような形状が得られることが分かった.

しかしながら, 十分な結果が得られたとは言い難く今後の改良が必要である. まず, 今回の例は 180 個 (60 点 ×3 次元) のパラメータの最適化問題であったが, 実際の繰り返し計算を 100 回で打ち切っている. 設計仕様を満足する形状を得るのに必ずしも最適解を得る必要はないため, この条件がただちに不当な設定とは言えないが, より精度の良い結果を得るためにはより多くの繰り返しを実行する必要があると考えられる. 計算時間の制約も考慮すると, 繰り返し回数を増すには計算アルゴリズムの見直しと高速化, より大規模な並列計算機の活用なども合わせて必要であろう. また, 仕様を満足する解は唯一であるとは限らないこと, 勾配法に基づく方法では局所解に収束する可能性もあるなど, 注意を払う必要もある.

参考文献

[Shaw, 1968] Shaw, E.A.G. and Teranishi, R.: Sound pressure generated in an external-ear replica and real human ears by a nearby point source, J. of Acoust. Soc. Am. 44 (1), 240–249, (1968).

[Otani, 2006] Otani, M. and Ise, S.: Fast calculation system specialized for head-related transfer function based on boundary element method, J. of Acoust. Soc. Am. 119 (5), 2589–2598, (2006).

[Spall, 1992] Spall, J.C.: Multivariate stochastic approximation using a simultaneous perturbation gradient approximation, IEEE Trans. on Automatic Control, 37, 332–341, (1992)

[Batteau, 1967] Batteau, D.W.: The role of the pinna in human localization, Proc. of Royal Soc. of London, B 158, 158–180,(1967)

[Shimoda, 2006] Shimoda, T., Nakashima, T., Kumon, M., Kohzawa, R., Mizumoto I., and Iwai, Z.: Spectral cues for robust sound localization with pinnae, Proc. of 2006 IEEE/RSJ Int'l Conf. Intell. Robot. and Sys., 386–391,(2006).

[Spall, 1998] Spall, J.C.: Implementation of the simultaneous perturbation algorithm for stochastic optimization, IEEE Trans. on Aerospace and Electronic Systems, 34, 817–823, (1998)

[SPSA] Simultaneous Perturbation Stochastic Approximation(website), *http://www.jhuapl.edu/SPSA/*

[ATLAS] Automatically Tuned Linear Algebra Software(ATLAS)(website), *http://math-atlas.sourceforge.net/*

[OpenMP] OpenMP, The OpenMP API specification for parallel programming(website), *http://openmp.org/*

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

人工知能学会研究資料
JSAI Technical Report
SIG-Challenge-A902-5

Blind signal separation with selective post filtering: application to speech enhancement

Jani Even, Hiroshi Saruwatari, Kiyohiro Shikano, Tomoya Takatani[+]
Nara Institute of Science and Technology, Ikoma, JAPAN
[+]Toyota motor corporation, Toyota, JAPAN
even@is.naist.jp

Abstract

In this paper, we consider the human/machine hands-free speech interface where the user voice is picked at a distance with a microphone array. The proposed method aims at suppressing the diffuse background noise efficiently without distorting the speech estimate. This method is a modification of a method combining frequency domain blind signal separation (FD-BSS) and Wiener filter based post-processing. Contrary to the conventional approach, the Wiener post filter is only applied to a selected number of the components separated by FD-BSS. Simulation results show that the proposed approach can achieve a better speech enhancement, measured in term of word recognition in a speech recognition task, than the conventional Wiener filter based post-processing.

1 Introduction

In hands-free speech recognition, microphone array techniques are used to improve the captured speech by reducing the effect of noise and reverberation ([7, 4]). Among these techniques, in recent years, frequency domain blind signal separation (FD-BSS) has been used with success for recovering the speech by separating the observed signals in their different components (see review paper [13]). FD-BSS is efficient for speech/speech separation [11]. But in the human/machine communication where the user's voice has to be extracted from a diffuse background noise, FD-BSS gives a better estimate of the diffuse background noise than of the target speech. Consequently FD-BSS has to be combined with some nonlinear post-filtering techniques in order to improve the quality of the captured speech [18, 11, 16, 17, 9]. An efficient approach suppresses the diffuse background noise estimated by FD-BSS via Wiener filtering [16].

In this paper, our goal is to improve the speech recognition performance for the human/machine hands-free speech interface. The user is assumed to be close to the microphone array and thus is modeled as a point source whereas the other sources create a diffuse background noise. We use a similar approach as in [16] where the noise estimate is obtained by FD-BSS and noise suppression is performed via Wiener filtering. But we propose a modified noise estimate and we do not apply the Wiener filtering directly to the observations.

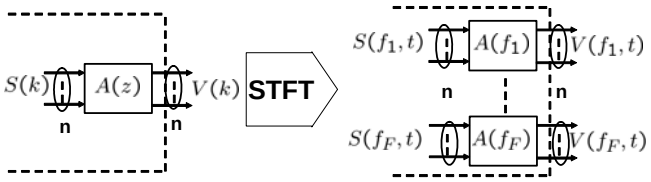


Figure 1: Equivalent mixtures in frequency domain.

The main idea is that if some of the sources from the diffuse background noise are efficiently canceled by the FD-BSS (linear processing), it is better not to include them in the noise estimate used by the post-filter (non linear processing) in order to keep the distortion of the estimated speech low. This is particularly important for speech recognition tasks where the the post-filter should give a good trade-off between high SNR and low distortion [16]. In the proposed approach, after the FD-BSS, we exclude from the noise estimate the estimated noise components that are the least correlated with the speech estimate. Using this modified noise estimate in the Wiener filter based post-filter also requires the modification of the observation before filtering.

Experimental results show the impact of the proposed method on the quality of the speech estimate in a speech recognition task. In particular, the proposed method achieves better performance than the conventional Wiener filter based post-processing.

2 Preliminaries

2.1 Frequency Domain Blind Signal Separation

In blind signal separation of acoustic signals, the propagation of the sounds from their locations of emission

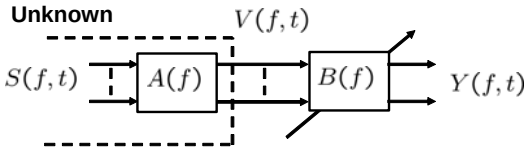


Figure 2: BSS at frequency bin f.

to the microphone array is modeled by a convolutive mixture. After applying a F points short time Fourier transform (STFT) to the observed signals, the convolutive mixture is equivalent to F instantaneous mixtures in the frequency domain (see Fig. 1). At the fth frequency bin, the observed signals are

$$V(f,t) = A(f)S(f,t)$$

where the $n \times n$ complex valued matrix $A(f)$ represents the instantaneous mixture received by the n microphone array and

$$S(f,t) = [s_1(f,t), \ldots, s_n(f,t)]^T$$

are the emitted signal components at the fth frequency bin. t denotes the frame index.

In each frequency bin, the blind estimation of the emitted signal components is possible using BSS [15]. The estimates

$$Y(f,t) = [y_1(f,t), \ldots, y_n(f,t)]^T$$

are obtained by applying an unmixing matrices $B(f)$ to the observed signals (see Fig.2)

$$Y(f,t) = B(f)X(f,t) = B(f)A(f)S(f,t). \qquad (1)$$

If the components of $S(f,t)$ are statistically independent (and at most one is Gaussian) then it is possible to recover the components of $S(f,t)$ up to scale and permutation indeterminacy by finding the separation matrix $B(f)$ such that the components of $Y(f,t)$ are statistically independent [3]. Namely $B(f)$ is such that

$$Y(f,t) = P(f)\Lambda(f)S(f,t)$$

where $P(f)$ is a $n \times n$ permutation matrix and $\Lambda(f)$ is a diagonal $n \times n$ matrix.

Consequently several FD-BSS methods adapt the matrices $B(f)$ in order to minimize a cost function measuring the statistical dependence between the components of the estimate $Y(f,t)$ (see [13]).

Because of the unknown order of the estimated components $y_i(f,t)$, in order to achieve separation in the time domain , it is necessary to match the components from the same signal in all the frequency bins before transforming back the signals in time. This is referred to as *permutation resolution*. After resolving the permutation, the estimated signals are still filtered by an indeterminate filter because of the scaling indeterminacy $\Lambda(f)$. A solution is to *project back* the estimated signals to the microphone array [12]. The projection back of the ith estimate is a n component signal defined by

$$Z_i(f,t) = B(f)^{-1} D_i Y(f,t)$$

where D_i is a matrix having only one non null entry $d_{ii} = 1$. If we assume perfect separation $B(f)A(f) = P(f)\Lambda(f)$ and the estimated signal is $s_j(f,t)$ then $P(f)$ is such that

$$P(f)^{-1} D_i P(f) = D_1$$

and

$$Z_i(f,t) = A(f,t)^{(:,j)} s_j(f,t)$$

where $A(f,t)^{(:,j)}$ is the j^{th} column of $A(f,t)$. Namely $Z_i(f,t)$ is equal to the contribution of the jth estimated signal at the microphone array because the projection back replaces the indeterminate filtering of the estimated signal by the estimate of the room impulse response between the location of the jth signal and the microphone array (represented by $A(f,t)^{(:,j)}$). Note that the observation is the sum of all the projected back components

$$X(f,t) = \sum_{i=1}^{n} Z_i(f,t).$$

3 Proposed method

The block diagram in Fig 3 shows the proposed processing in the frequency domain. The different blocks are explained in the following sections.

3.1 Speech and Diffuse Background Noise Blind Separation

In [14], the authors showed that for speech/speech separation (cocktail party model) FD-BSS is equivalent to a set of adaptive null beamformers (ANBF) each having its null toward different speakers. Thus the separation is achieved because FD-BSS is able to cancel the speeches that are point sources. In our case, FD-BSS gives a good estimate of the diffuse background noise by placing a null in the direction of the speech. But it is not possible to get a good speech estimate since with a limited number of microphones it is not possible to cancel the diffuse background noise [18].

Another problem of the separation of speech and diffuse background noise is the permutation resolution. The methods developed for the speech/speech separation are often not efficient for the case of speech in diffuse background noise [5]. Here, in order to find the speech component in each of the frequency bins, we rely on the fact that the speech distribution is spikier than that of the diffuse background noise. To measure the 'spikedness' of the distribution, we use the average of the modulus of the $y_i(f,t)$

$$\alpha_i(f) = \mathcal{E}\{|y_i(f,t)|\}$$

under the constraint

$$\mathcal{E}\{|y_i(f,t)|^2\} = 1$$

where $\mathcal{E}\{\cdot\}$ denotes the expectation operator. The component with the smallest parameter is selected as the target speech (for details see [6]). After this first step of permutation resolution, we assume that the components are permuted such that $y_1(f,t)$ is the speech component in the fth bin.

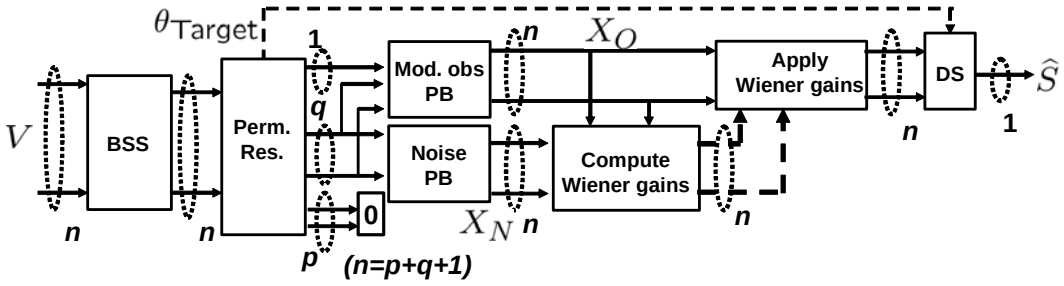


Figure 3: Overview of the proposed architecture.

3.2 Modified Noise Estimate and Modified Observation

Assuming that the FD-BSS method achieved the best possible separation, the estimated noise components $y_2(f,t),\cdots,y_n(f,t)$ contain no speech however the speech estimate $y_1(f,t)$ is still contaminated by the noise.

The noise estimate for the conventional Wiener filter based post-processing is obtained by projecting back the $n-1$ components $y_2(f,t),\cdots,y_n(f,t)$ to the microphone array [16]. But in our approach we do not project back the $n-1$ noise components.

The noise estimate is composed of several components and these components may contribute at different levels in the noise still present in the speech estimate. In particular some of these estimated noise components may have a very small contribution in the noise contaminating the speech estimate. Meaning that FD-BSS suppressed some part of the diffuse background noise (the diffuse background noise may contain contributions of point sources for example). In such case, we propose to exclude these components from the noise estimate used by the Wiener filter post-processing. The reason is that it is better in term of speech distortion to suppress these components with the FD-BSS filter that is linear than with the nonlinear post-processing.

To determine the noise components that have few contribution in the noise contaminating the speech estimate, we compute the correlation between the speech estimate $y_1(f,t)$ and the estimated noise components $y_2(f,t),\cdots,y_n(f,t)$. This correlation is denoted by

$$C_i = \mathcal{E}\{y_1(f,t)y_i(f,t)^*\}$$

where * denotes the complex conjugation.

The noise components are sorted according to the absolute value of these correlations. In the remainder, the components are permuted such that $C_2 > \cdots > C_n$. The p components with smallest correlation are not projected back (see the p components set to 0 in Fig. 3).

Thus the noise estimate $X_N(f,t)$ is only composed of the projection back of $y_2(f,t),\cdots,y_{n-p}(f,t)$

$$X_N(f,t) = B(f)^{-1}D_N Y(f,t) = \sum_{i=2}^{n-p} Z_i(f,t)$$

where D_N is a matrix selecting $y_2(f,t),\cdots,y_{n-p}(f,t)$ (the *Noise PB* block in Fig. 3).

Since the last p components are not projected back the Wiener filtering has to be applied to the modified observation $X_O(f,t)$ obtained by projecting back all the components except these p last ones

$$X_O(f,t) = B(f)^{-1}D_O Y(f,t) = \sum_{i=1}^{n-p} Z_i(f,t)$$

where D_O is a matrix selecting $y_1(f,t),\cdots,y_{n-p}(f,t)$ (the *Mod. obs PB* block in Fig. 3).

3.3 Wiener post-filter and delay and sum beamformer

The modified noise estimate $X_N(f,t)$ and the modified observation $X_O(f,t)$ both have n components. The Wiener filtering is applied component wise and the Wiener gain for the ith component is

$$G^{(i)}(f,t) = \frac{|\widehat{X}_O^{(i)}(f,t)|^2}{|\widehat{X}_O^{(i)}(f,t)|^2 + \gamma|\widehat{X}_N^{(i)}(f,t)|^2}$$

where the subscript (i) denotes the ith component and γ is a parameter controlling the noise reduction. The ith component of the filtered target speech is

$$\widehat{S}^{(i)}(f,t) = \sqrt{G^{(i)}(f,t)|\widehat{X}_O^{(i)}(f,t)|^2}\,\frac{\widehat{X}_O^{(i)}(f,t)}{|\widehat{X}_O^{(i)}(f,t)|}.$$

finally the n components of the Wiener filtered speech estimate are merged into one by applying a delay and sum (DS) beamformer in the direction θ_{target} of the target speech

$$\widehat{S}(f,t) = \sum_{i=1}^{n} G_{\mathrm{DS}\theta}^{(i)}(f,t)\widehat{S}^{(i)}(f,t)$$

where $G_{\mathrm{DS}\theta}^{(i)}(f,t)$ the gain of the DS beamformer at the ith microphone (the target DOA is estimated during the permutation resolution step. It is an average over all bins of the estimated DOA of the separated speech component).

4 Experimental Results

To demonstrate the effectiveness of the proposed post-processing based on selective projection back, we compare it to the conventional Wiener filter based post-processing, to the FD-BSS with no post-processing and to a delay and sum beamformer (DS).

A four ($n = 4$) microphone array (inter microphone spacing of 2.15cm) was used to record a diffuse background noise (a vacuum cleaner at two meters from the array and -60^o), the impulse responses at one meter from the array in front of the array and at an angle of 60^o (see Fig. 4). The recorded noise is mixed with the convolution of the impulse response at an angle of 60^o with a recorded fan noise. The SNR of this mixture is 0dB. Then this mixture of noises is mixed with the convolution of the impulse responses and a clean speech (100 signals from the JNAS database of Japanese sentences [8]). A second set of data is obtain by mixing only the diffuse background noise with the filtered speeches. The first data set is referred to by 'fan' whereas the second is referred to by 'no fan'. The SNR values between noise and speech are adjusted to be the same for both datasets.

For the frequency domain processing, the short time Fourier transform uses a 512 point hamming window with 50% overlap. The separation is performed by 300 iterations of a BSS method with adaptation step of 0.1 divided by two every 100 iterations (the method is adapted from [2, 19]).

The proposed approach is tested with two modified noise estimates corresponding to $p = 1$ and $p = 2$. The result are compared to the delay and sum beamformer in front of the array (DS), the FD-BSS with no post processing (BSS) and the conventional Wiener filter $p = 0$ (note: the FD-BSS with no post processing can be seen as discarding all the noise components $p = 3$). Several values of the coefficient γ of the Wiener filter were tested for each method: $\gamma \in \{1, 5, 10, 15, 20, 25\}$.

Since our goal is speech recognition, a 20K-word Japanese dictation task from JNAS is used as performance measure. The word accuracy achieved by the recognizer is function of both the SNR and the amount of distortion of the speech estimate. The recognizer is JULIUS [1] using Phonetically Tied Mixture (PTM) model [10]. The open test set is composed of 100 utterances (female speakers). The conditions used in recognition are given in Table 1. The acoustic model is a clean model with super-imposed noise (office noise 25dB SNR).

Figure 5 shows the word accuracy achieved by the different methods on the two data sets ('fan' and 'nofan') for the different SNR values. For each case the result is the one obtain with the parameter γ giving the best

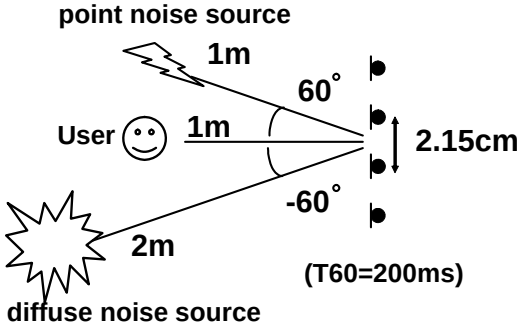


Figure 4: Experimental setup.

Table 1: *System specifications.*

Sampling frequency	16 kHz
Frame length	25 ms
Frame period	10 ms
Pre-emphasis	$1 - 0.97z^{-1}$
Feature vectors	12-order MFCC, 12-order ΔMFCCs 1-order ΔE
HMM	PTM , 2000 states
Training data	Adult and Senior (JNAS)
Test data	Adult and Senior female (JNAS)

word accuracy (also see first row of Table 2).

We can see that, at the same SNR, the performances are better for the 'fan' dataset that contains a point source in addition to the diffuse background noise. In particular for the lower SNRs (5dB and 10dB), the improvement of the word accuracy with the proposed method over the conventional method is better for the 'fan' dataset. This shows that if some components of the noise are canceled by the FD-BSS (the point source fan noise), modifying the noise estimate improves the performance.

There is also a performance gain on the 'nofan' dataset showing that some of the noise components of the diffuse background noise contributed less to the noise contaminating the speech estimate given by FD-BSS. We can also notice that for $p = 2$ the performance is better than for $p = 1$. Meaning that discarding more noise components lead to better results on these datasets.

These results also show the necessity of the nonlinear post-processing as in all cases there is an improvement over the FD-BSS.

The effect of the coefficient γ is depicted in Fig. 6 (the three plots share same color scale). For the proposed post-processing, like for the conventional Wiener filter there is a trade-off between SNR and distortion, the word accuracy is better with a larger γ at low SNR and a smaller γ at high SNR.

Table 2 shows the difference of word accuracy between the proposed method with $p = 2$ and the conventional method $p = 0$ for different choice of γ ()A positive value indicates that the proposed method is superior to the conventional method). The row 'best γ' is obtained by selecting for each method at each SNR the parameter γ from the list $\{1, 5, 10, 15, 20, 25\}$ that gives the best word accuracy. This row shows the improvement for the proposed method for $p = 2$ over conventional method ($p = 0$) in Fig. 5. The other rows show the improvement for fixed values of γ. Note that for larger γ there is no performance improvement on the 'fan' dataset as the proposed method perform best with small γ (see bottom of Fig. 6). This shows that at high SNR it is important to choose a smaller γ for the proposed method. We can also notice that for the 'nofan' dataset the performance difference is larger than for the 'fan' dataset at high SNR.

	'fan' dataset				'nofan' dataset			
	5dB	10dB	15dB	20dB	5dB	10dB	15dB	20dB
best γ	4.92	5.1	2.74	0.91	2.18	2.28	3.53	1.36
$\gamma = 1$	13.3	15.03	6.79	0.91	9.3	12.47	7.13	4
$\gamma = 15$	6.19	5.08	0.77	−0.63	0.72	4.12	3.71	2.18
$\gamma = 25$	1.81	4.26	−1.3	−1.77	3.58	3.98	1.74	2.44

Table 2: Word Accuracy differences for $p = 2$ and $p = 0$ versus γ

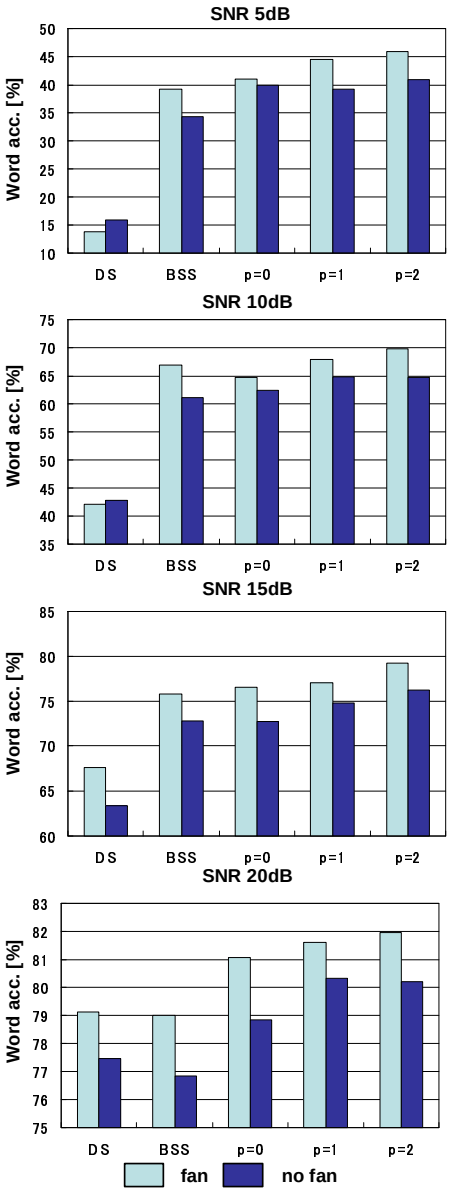


Figure 5: Word accuracy for different SNR values with the different methods for both datasets.

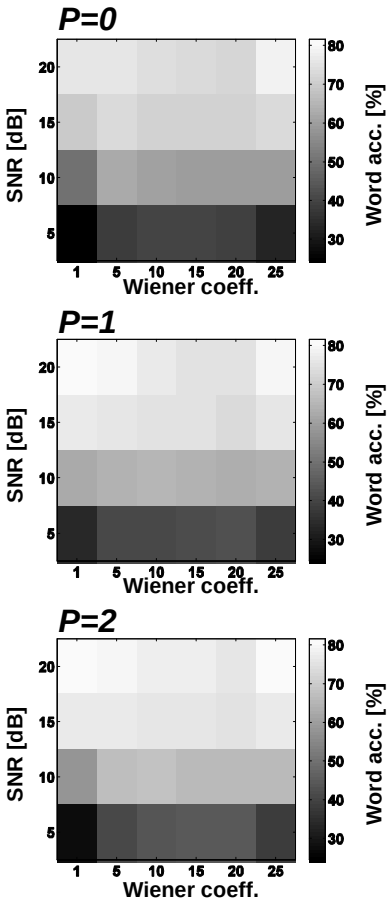


Figure 6: Effect of Wiener coefficient on word accuracy for the different methods ('fan' dataset only).

5 Conclusion

In this paper, we consider the suppression of the diffuse background noise in the human/machine communication scenario. We proposed a modification of the noise estimation given by FD-BSS. This modification leads to a more efficient Wiener filter based post-processing of the speech estimate. Some experimental results showed that this approach increases the word accuracy in a dictation task.

References

[1] Julius, an open-source large vocabulary csr engine - http://julius.sourceforge.jp.

[2] A. J. Bell and T. J. Sejnowski. An information maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7(6):1129–1159, 1995.

[3] P. Comon. Independent component analysis, a new concept ? *Signal Processing*, 36:287–314, 1994.

[4] S. Doclo, A. Spriet, and M. Moonen. Efficient frequency-domain implementation of speech distortion weighted multi-channel wiener filtering for noise reduction. *in Proc. EUSIPCO, Vienna, Austria*, pages 2007–2010, 2004.

[5] J. Even, H. Saruwatari, and K. Shikano. An improved permutation solver for blind signal separation based front-ends in robot audition. *2008 IEEE/RSJ International Conference on Intelligent Robots and Systems, Nice, France*, pages 2172–2177, 2008.

[6] J. Even, H. Saruwatari, and K. Shikano. Blind signal extraction based speech enhancement in presence of diffuse background noise. *2009 IEEE Workshop on Statistical Signal Processing (SSP2009), Cardiff, Wales, UK*, pages 513–516, 2009.

[7] L.J. Griffiths and C.W. Jim. An alternative approach to linearly constrained adaptive beamforming. *IEEE Trans. Antennas Prapagation*, AP-30:27–34, 1982.

[8] K. Ito et al. Jnas: Japanese speech corpus for large vocabulary continuous speech recognition research. *The Journal of Acoust. Soc. of Japan*, 20:196–206, 1999.

[9] J. Kocinski. Speech intelligibility improvement using convolutive blind source separation assisted by denoising algorithms. *Speech Communication*, 50:29–37, 2008.

[10] A. Lee, T. Kawahara, K. Takeda, and Shikano K. A new phonetic tied-mixture model for efficient decoding. *In Proceedings of ICASSP*, pages 1269–1272, 2000.

[11] Y. Mori, T. Takatani, H. Saruwatari, K. Shikano, T. Hiekata, and T Morita. Blind source separation combining simo-ica and simo-model-based binary masking. *ICASSP 2006, Toulouse,France*, pages 81–84, 2006.

[12] N. Murata, S. Ikeda, and A. Zieh. An approach to blind source separation based on temporal structure of speech signals. *Neurocomputing*, 41(1-4):1–24, 2001.

[13] M.S. Pedersen, J. Larsen, U. Kjems, and L.C. Parra. *A Survey of Convolutive Blind Source Separation Methods*. Springer, 2007.

[14] H. Saruwatari et al. Blind source separation combining independent component analysis and beamforming. *EURASIP Jour. on Appl. Sig. Proc.*, 2003(11):1135–1146, 2003.

[15] P. Smaragdis. Blind separation of convolved mixtures in the frequency domain. *Neurocomputing*, 22(1-3):21–34, 1998.

[16] Y. Takahashi, K. Osako, H. Saruwatari, and K. Shikano. Blind source extraction for hands-free speech recognition based on wiener filtering and ica-based noise estimation. *Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSMCA)*, pages 164–167, 2008.

[17] Y. Takahashi, T. Takatani, K. Osako, H. Saruwatari, and K. Shikano. Blind spatial subtraction array for speech enhancement in noisy environment. *IEEE Transaction on Audio, Speech and Language Processing*, 17(4):650–664, 2009.

[18] Y. Takahashi, T. Takatani, H. Saruwatari, and K. Shikano. Blind spatial subtraction array with independent component analysis for hands-free speech recognition. *International Work Shop on Acoustic Echo and Noise Control (IWAENC) (CD-ROM)*, 2006.

[19] N. Vlassis and Y. Motomura. Efficient source adaptivity in independent component analysis. *IEEE Trans. Neural Networks*, 12(3):559–566, 2001.

内部雑音抑圧型ロボット音声対話システムにおけるマイクロホンアレー配置の検討

A Study on Microphone Array Layout for Spoken-Oriented Robot Dialogue System with Internal-Noise Suppressor

† 澤田紘志，†Jani Even，† 猿渡洋，† 鹿野清宏，‡ 高谷智哉

†Hiroshi Sawada, †Jani Even, †Hiroshi Saruwatari, †Kiyohiro Shikano, and ‡Tomoya Takatani

† 奈良先端科学技術大学院大学

‡ トヨタ自動車株式会社

†Nara Institute of Science and Technology

‡TOYOTA MOTOR CORPORATION

hiroshi-s@is.naist.jp

Abstract

本論文では，内部雑音抑圧型ロボット音声対話システムのために有効なマイクロホンアレー配置について提案する．我々は既に，ハンズフリー音声認識のための内部雑音抑圧手法として，独立成分分析 (independent component analysis: ICA) に基づくセミブラインド音源分離 (semi-blind source separation: SBSS) と wiener filter (WF) を統合した手法を提案している．本稿でははじめに，従来のマイクロホンアレー配置では SBSS の雑音推定性能が劣化し，音声認識性能があまり高くないという問題を示す．次に，内部雑音の音源到来方位 (direction of arrival: DOA) を解析し，内部雑音が 0 °に定位することを示す．この解析結果に基づき，マイクロホンアレーをブロードサイドアレーからエンドファイアアレーに回転させることを提案する．最後に，提案するマイクロホンアレー配置の有効性を確かめるためにシミュレーション実験を行った．従来に比べて高い音声認識性能を有することを示す．

1 はじめに

人と音声コミュニケーション可能なパートナーロボットでは，ユーザから離れた位置にマイクロホンを設置して音声認識を行うハンズフリー音声認識が必要不可欠である．しかし，実環境下においては，周囲に存在する環境雑音や残響，さらにはファンノイズやモータ音などのロボット自身が発する内部雑音によって，音声認識性能が低下する問題がある．そこで，著者らの一人である Even らによって，内部雑音測定用センサの観測信号を教師情報に用いる内部雑音抑圧手法として，セミブラインド音源分離 (semi-blind source separation: SBSS) [1]と wiener filter (WF) [2]を統合した手法が提案されている [3]．SBSS と WF を統合した手法を Fig. 1 に示す．この手法では，はじめに独立成分分析 (independent component analysis: ICA) [4]に基づく SBSS によって環境雑音と内部雑音を動的に推定する．次に，推定した環境雑音と内部雑音を基に，マイクロホンの観測信号にＷＦを適用することによって，各チャネル毎に目的音声抽出を実現する．最後に，各チャネル毎の目的音声抽出結果を遅延和アレー (delay-and-sum: DS) [5]によって目的音声強調し，最終出力音声信号を得る．しかし，ユーザがロボットの正面に立ち，かつロボットに設置されているマイクロホンアレーがブロードサイドアレーのとき，SBSS の雑音推定性能が劣化し音声認識性能があまり高くないという問題があった．

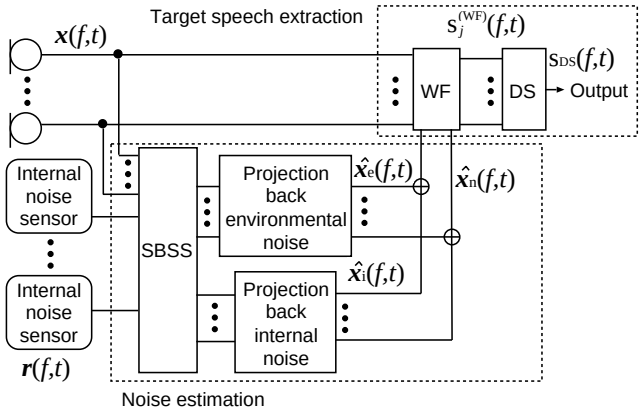


Figure 1: Block diagram of speech extraction method.

そこで本論文では，まず，従来のマイクロホンアレー配置における SBSS の雑音推定性能があまり高くないことを示す．次に，この原因を検証するために，内部雑音の音源到来方位 (direction of arrival: DOA) について解析を行い，内部雑音が 0 °に定位していることを示す．この結果から，ユーザがロボット正面に立ちロボットに設置されているマイクロホンアレーがブロードサイドアレーのとき，目的音声と内部雑音の DOA が 0 °で重なることにより十分な SBSS の雑音推定性能が得られていなかったことを確かめる．この知見に基づき，本稿では，SBSS の

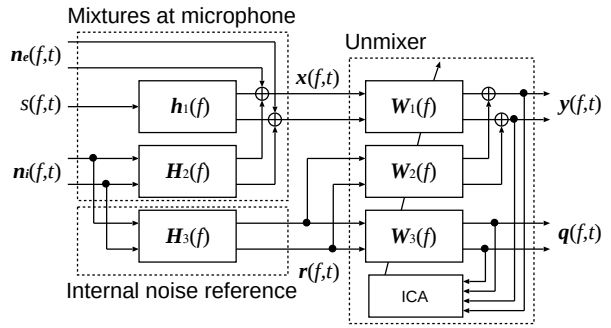


Figure 2: Block diagram of mixing and unmixing at f-th frequency bin.

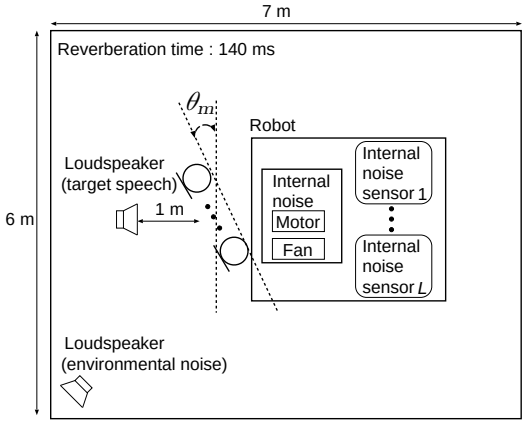


Figure 3: Layout of reverberant room used in our simulation.

雑音推定性能を高めるためにロボットに搭載する直線上マイクロホンアレーをブロードサイドアレーからエンドファイアアレーに回転させ，目的音声と内部雑音の DOA を空間的に異なる方位にすることを提案する．最後に，提案するマイクロホンアレー配置の有効性を確かめるために，シミュレーション実験を行う．提案法よって得られた最終出力音声信号を Julius [6]によって機械音声認識を行い，客観評価から提案法の有効性を示す．

2 セミブラインド音源分離を用いた目的音声抽出

2.1 混合過程

本稿では，点音源で近似できる目的音声信号，点音源で近似されない (非点音源) 環境雑音信号，ロボットが発する内部雑音信号が同時に存在する環境を想定する．ここで内部雑音信号とは，ロボットの筐体振動であると推測される．ロボット内部に搭載されているサーボモータや CPU ファン等がロボットの筐体を振動させ，筐体に設置されているマイクロホンアレーが振動し，その振動が雑音として観測される．そこで，内部雑音測定用センサをロボットの筐体に設置し，筐体振動である内部雑音信号を測定する．想定する環境の混合過程と分離過程を Fig. 2 に示す．以下では混合過程について説明する．

マイクロホン数を J，内部雑音測定用センサ数を L としたとき，マイクロホンと内部雑音測定用センサにおけるそれぞれの観測信号は以下のように表現できる．

$$\boldsymbol{x}(f,t) = \boldsymbol{h}_1(f)s(f,t) + \boldsymbol{n}_{\mathrm{e}}(f,t) + \boldsymbol{H}_2(f)\boldsymbol{n}_{\mathrm{i}}(f,t) \quad (1)$$

$$\boldsymbol{r}(f,t) = \boldsymbol{H}_3(f)\boldsymbol{n}_{\mathrm{i}}(f,t) \quad (2)$$

ここで，f は周波数帯域番号，t は時間フレーム番号を表す．また，$\boldsymbol{x}(f,t) = [x_1(f,t),\ldots,x_J(f,t)]^{\mathrm{T}}$ はマイクロホンにおける観測信号ベクトル，$\boldsymbol{n}_{\mathrm{e}}(f,t) = [n_1^{(\mathrm{e})}(f,t),\ldots,n_J^{(\mathrm{e})}(f,t)]^{\mathrm{T}}$ は環境雑音ベクトル，$\boldsymbol{n}_{\mathrm{i}}(f,t) = [n_1^{(\mathrm{i})}(f,t),\ldots,n_K^{(\mathrm{i})}(f,t)]^{\mathrm{T}}$ は内部雑音ベクトル (K は内部雑音の音源数)，$\boldsymbol{r}(f,t) = [r_1(f,t),\ldots,r_L(f,t)]^{\mathrm{T}}$ 内部雑音測定用センサにおける観測信号ベクトル，$s(f,t)$ は目的音声信号を表す．さらに，$\boldsymbol{h}_1(f) = [h_1^{(1)}(f),\ldots,h_J^{(1)}(f)]^{\mathrm{T}}$ は目的音声音源から各マイクロホンへの伝達関数ベクトル，$\boldsymbol{H}_2(f) = [\boldsymbol{h}_1^{(2)}(f),\ldots,\boldsymbol{h}_K^{(2)}(f)]$ は各内部雑音源から各マイクロホンへの伝達関数行列，$\boldsymbol{H}_3(f) = [\boldsymbol{h}_1^{(3)}(f),\ldots,\boldsymbol{h}_K^{(3)}(f)]$ は各内部雑音源から各内部雑音測定用センサへの伝達関数行列を表す．$\boldsymbol{h}_k^{(2)}(f) = [h_{1k}^{(2)}(f),\ldots,h_{Jk}^{(2)}(f)]^{\mathrm{T}}$ $(k = 1,\ldots,K)$ は k 番目の内部雑音源から各マイクロホンへの伝達関数ベクトル，$\boldsymbol{h}_k^{(3)}(f) = [h_{1k}^{(3)}(f),\ldots,h_{Lk}^{(3)}(f)]^{\mathrm{T}}$ $(k = 1,\ldots,K)$ は k 番目の内部雑音源から各内部雑音測定用センサへの伝達関数ベクトルを表す．また内部雑音測定用センサは筐体振動のみを観測することができ，空気振動である目的音声信号や環境雑音信号をほとんど観測しないという特徴を持つ．そのため混合過程モデルでは，内部雑音測定用センサは空気振動の成分を全く観測しないものとして定式化される．

2.2 セミブラインド音源分離に基づく雑音推定

ICA に基づく SBSS では分離行列 $\boldsymbol{W}_i(f)$ $(i = 1,2,3)$ を用いて，目的音声と環境雑音の分離信号ベクトル $\boldsymbol{y}(f,t) = [y_{\mathrm{s}}(f,t), y_{\mathrm{n}}(f,t)]^{\mathrm{T}}$ と内部雑音の分離信号ベクトル $\boldsymbol{q}(f,t) = [q_1(f,t),\ldots,q_K(f,t)]^{\mathrm{T}}$ を，各周波数毎に以下の式で求めることができる [1].

$$\boldsymbol{y}(f,t) = \boldsymbol{W}_1(f)\boldsymbol{x}(f,t) + \boldsymbol{W}_2(f)\boldsymbol{r}(f,t) \quad (3)$$

$$\boldsymbol{q}(f,t) = \boldsymbol{W}_3(f)\boldsymbol{r}(f,t) \quad (4)$$

ここで，$y_{\mathrm{s}}(f,t)$ は目的音声推定信号，$y_{\mathrm{n}}(f,t)$ は環境雑音推定信号を示す．観測信号を分離する最適な分離行列 $\boldsymbol{W}_i(f)$ $(i = 1,2,3)$ は以下の反復学習式によって求めることができる．

$$\boldsymbol{W}_i^{[k+1]}(f) = \boldsymbol{W}_i^{[k]}(f) - \mu\Delta\boldsymbol{W}_i^{[k]}(f) \tag{5}$$

$$\Delta\boldsymbol{W}_1^{[k+1]}(f) = (\boldsymbol{I} - \langle\boldsymbol{\phi}(\boldsymbol{y}(f,t)^{[k]})\boldsymbol{y}^{\mathrm{H}}(f,t)^{[k]}\rangle_t)\boldsymbol{W}_1^{[k]}(f) \tag{6}$$

$$\Delta\boldsymbol{W}_2^{[k+1]}(f) = (\boldsymbol{I} - \langle\boldsymbol{\phi}(\boldsymbol{y}(f,t)^{[k]})\boldsymbol{y}^{\mathrm{H}}(f,t)^{[k]}\rangle_t)\boldsymbol{W}_2^{[k]}(f) - (\langle\boldsymbol{\phi}(\boldsymbol{y}(f,t)^{[k]})\boldsymbol{q}^{\mathrm{H}}(f,t)^{[k]}\rangle_t)\boldsymbol{W}_3^{[k]}(f) \tag{7}$$

$$\Delta\boldsymbol{W}_3^{[k+1]}(f) = (\boldsymbol{I} - \langle\boldsymbol{\phi}(\boldsymbol{q}(f,t)^{[k]})\boldsymbol{q}^{\mathrm{H}}(f,t)^{[k]}\rangle_t)\boldsymbol{W}_3^{[k]}(f) \tag{8}$$

$\boldsymbol{\phi}(\cdot)$ は非線形関数ベクトルを表し，分離信号 $\boldsymbol{y}(f,t)$, $\boldsymbol{q}(f,t)$ のサンプルデータから，カーネルに基づくスコア関数の推定により求める [7]．また，$\boldsymbol{I}$ は単位行列，μ は更新係数，$\langle\cdot\rangle_t$ は時間平均演算子，$\boldsymbol{M}^{\mathrm{H}}$ は $\boldsymbol{M}$ の複素共役転置，$\cdot^{[k]}$ は k 回目の反復学習であることを表す．また，ICA における permutation 問題は，確率密度分布推定と DOA 推定を組み合わせた手法を用いて解決している [8]．

ここで，非点音源雑音環境下において，ICA の目的音声信号推定精度はあまり優れていないのに対して，雑音信号推定精度は非常に高いことが知られている[9]．このような事実に基づき，ICA を雑音推定器として用いる．よって，最適化された分離フィルタを基に以下の式を用いて，環境雑音信号とマイクロホンで観測される内部雑音信号の推定を行う．

$$\hat{\boldsymbol{x}}_{\mathrm{e}}(f,t) = \boldsymbol{W}_1^{+}(f)[0, y_{\mathrm{n}}(f,t)]^{\mathrm{T}} \tag{9}$$

$$\hat{\boldsymbol{x}}_{\mathrm{i}}(f,t) = -\boldsymbol{W}_1^{+}(f)\boldsymbol{W}_2(f)\boldsymbol{W}_3^{+}(f)\boldsymbol{q}(f,t) \tag{10}$$

ここで $\hat{\boldsymbol{x}}_{\mathrm{e}}(f,t) = [\hat{x}_1^{(\mathrm{e})}(f,t),\ldots,\hat{x}_J^{(\mathrm{e})}(f,t)]^{\mathrm{T}}$ は推定環境雑音ベクトル，$\hat{\boldsymbol{x}}_{\mathrm{i}}(f,t) = [\hat{x}_1^{(\mathrm{i})}(f,t),\ldots,\hat{x}_J^{(\mathrm{i})}(f,t)]^{\mathrm{T}}$ はマイクロホン地点での推定内部雑音ベクトルを表す．

2.3 目的音声抽出

SBSS によって求めた推定環境雑音ベクトル，推定内部雑音ベクトル，マイクロホンの観測信号ベクトルを用いて，各チャネル毎に WF のゲイン係数設計と適用を行い，目的音声抽出を実現する．以下でその処理の詳細を述べる．

抑圧すべき全ての雑音信号は推定環境雑音信号と推定内部雑音信号の和で表され，これを用いて各チャネル毎に WF のゲイン係数を設計する．

$$\hat{\boldsymbol{x}}_{\mathrm{n}}(f,t) = \hat{\boldsymbol{x}}_{\mathrm{e}}(f,t) + \hat{\boldsymbol{x}}_{\mathrm{i}}(f,t) \tag{11}$$

$$g_j(f,t) = \frac{|x_j(f,t)|^2}{|x_j(f,t)|^2 + \beta|\hat{x}_j^{(\mathrm{n})}(f,t)|^2} \tag{12}$$

ここで，$\hat{\boldsymbol{x}}_{\mathrm{n}}(f,t) = [\hat{x}_1^{(\mathrm{n})}(f,t),\ldots,\hat{x}_J^{(\mathrm{n})}(f,t)]^{\mathrm{T}}$ は全ての推定雑音ベクトル，$g_j(f,t)$ は j チャネルにおけるゲイン係数，β は雑音抑圧の処理強度パラメータを表す．

最終的に，各チャネル毎にゲイン係数 $g_j(f,t)$ をマイクロホンの観測信号に適用することで，以下のように j チャネルにおける推定目的音声信号 $s_j^{(\mathrm{WF})}(f,t)$ を得る．

$$s_j^{(\mathrm{WF})}(f,t) = \sqrt{g_j(f,t)|x_j(f,t)|^2}\frac{x_j(f,t)}{|x_j(f,t)|} \tag{13}$$

最後に，WF によって得られた各チャネル毎の推定目的音声信号に対して DS をすることで，目的音声強調を行い最終出力音声信号を得る．

$$s_{\mathrm{DS}}(f,t) = \boldsymbol{w}_{\mathrm{DS}}(f,\theta_{\mathrm{U}})^{\mathrm{T}}[s_1^{(\mathrm{WF})}(f,t),\ldots,s_J^{(\mathrm{WF})}(f,t)]^{\mathrm{T}} \tag{14}$$

$$\boldsymbol{w}_{\mathrm{DS}}(f,\theta) = [w_1^{(\mathrm{DS})}(f,\theta),\ldots,w_J^{(\mathrm{DS})}(f,\theta)]^{\mathrm{T}} \tag{15}$$

$$w_j^{(\mathrm{DS})}(f,\theta) = \frac{1}{J}\exp\left(-i2\pi(f/N)f_{\mathrm{s}}d_j\sin\theta/c\right) \tag{16}$$

ここで $s_{\mathrm{DS}}(f,t)$ は最終出力音声信号，$\boldsymbol{w}_{\mathrm{DS}}(f,\theta)$ は DS のフィルタ係数ベクトル，θ_{U} は DS の目的音声方位を表し，ICA によって学習した分離行列から推定することが可能である[10]．ここで f_{s} はサンプリング周波数，d_j $(j = 1,\ldots,J)$ はマイクロホン位置，N は DFT 長，c は音速を表す．

2.4 セミブラインド音源分離の問題点

従来のマイクロホンアレー配置では，SBSS の雑音推定性能があまり高くないという問題があった．このことを確かめるために，予備実験を行った．Figure 3 に示す音響環境で収録したインパルス応答をクリーン音声に畳み込んだ信号を目的音声信号とした．この信号に対して，実収録内部雑音信号における静止区間 (CPU ファン雑音のみ) と目的音声信号の信号対雑音電力比 (signal-to-noise ratio: SNR) が 20 dB になるように内部雑音信号を付加した．目的音声信号と動作区間を含んだ内部雑音信号全体における入力 SNR は 16.6 dB である．さらに，目的音声信号と環境雑音信号の入力 SNR が 10 dB になるように実収録の環境雑音信号を付加した．マイクロホンの素子数は 4，内部雑音測定用センサの素子数は 3，$\theta_{\mathrm{m}} = 0\,^{\circ}$ とした．

Figure 4 に目的音声信号 10 文において，SBSS を適用したときの環境雑音信号と内部雑音信号による全ての雑音信号におけるスペクトル歪み (spectral distortion: SD) $e(f)$ の平均値を示す．$e(f)$ は以下の式で与えられる．

$$e(f) = 10\log_{10}\left(\frac{1}{J}\sum_j\sum_t|x_j^{(\mathrm{n})}(f,t) - \hat{x}_j^{(\mathrm{n})}(f,t)|^2\right) \tag{17}$$

ここで，$x_j^{(\mathrm{n})}$ は j 番目のマイクロホン地点における真の雑音信号ベクトルを表す．SD はスペクトルドメインにおける歪みを表しており，値が小さいほど歪みが小さく雑音推定性能が良いことを表す指標である．Figure 4 より，低域の SD が大きく，SBSS の雑音推定性能が高くないことが分かる．

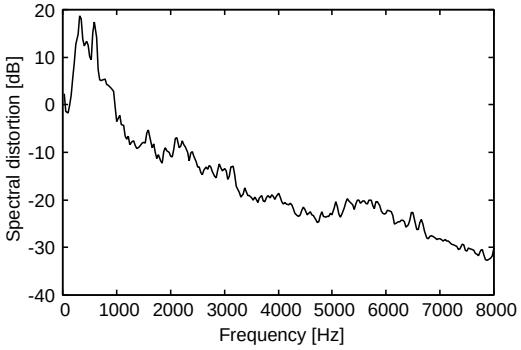


Figure 4: Spectral distortion between components of true noise and estimated noise (averaged on channels and utterances).

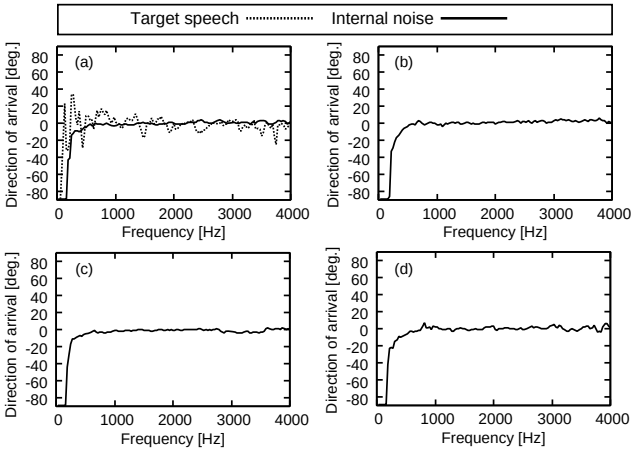


Figure 5: DOAs of internal noise: (a) type 1 target speech, (b) type 2, (c) type 3 and (d) type 4.

3 提案法

3.1 概要

本章では，SBSS の雑音推定性能が高くない原因を検証し，その改善策を提案する．まず，内部雑音信号の DOA を最小分散法 (minimum variance method: MV) によって解析する．この解析より，内部雑音信号の DOA が 0 ° に定位していることを示す．この結果から，従来のマイクロホンアレー配置では正面の目的音声信号と内部雑音信号の DOA が 0 °で重なり、十分な SBSS の雑音推定性能が得られていなかったことを示す．この知見に基づき，SBSS の雑音推定性能を改善するために有効なマイクロホンアレー配置を提案する．

3.2 DOA に基づく内部雑音の解析

はじめに，内部雑音信号の DOA を解析する予備実験を行った．実験における音響環境と条件は 2.4 節と同じである．ただしロボットの内部雑音に関しては，機械音やモータ音を含む異なる動作によって発生した 4 種類について検討を行う．

目的音声信号と内部雑音信号の DOA を解析するために MV を用いる．はじめに，マイクロホンの受音信号に対して MV を適用したときの出力パワー $P(f,\theta)$ を以下のように求める．

$$P(f,\theta) = \frac{1}{\boldsymbol{a}^{\mathrm{H}}(f,\theta)\boldsymbol{R}^{-1}(f)\boldsymbol{a}(f,\theta)} \tag{18}$$

$$\boldsymbol{R}(f) = \mathrm{E}[\boldsymbol{z}(f,t)\boldsymbol{z}^{\mathrm{H}}(f,t)] \tag{19}$$

$$\boldsymbol{a}(f,\theta) = [a_1(f,\theta),\ldots,a_J(f,\theta)]^{\mathrm{T}} \tag{20}$$

$$a_j(f,\theta) = \exp\left(i2\pi(f/N)f_{\mathrm{s}}d_j\sin\theta/c\right) \tag{21}$$

ここで，$\boldsymbol{a}(f,\theta)$ はステアリングベクトル，$\boldsymbol{R}(f)$ は共分散行列，$\boldsymbol{z}(f,t) = [z_1(f,t),\ldots,z_J(f,t)]^{\mathrm{T}}$ は受音信号ベクトルを表し，予備実験においては目的音声信号，もしくは内部雑音信号を表す．さらに，$\mathrm{E}[\cdot]$ は 平均値操作を表し，ここでは時間フレームに対して平均値操作を行う．また，各周波数毎に θ を -90 ° から 90 ° まで 1 °ずつ変化させ，$|P(f,\theta)|$ が最も大きくなったときの θ を受音信号の DOA とする．

4 種類の内部雑音信号における DOA の結果を Figs. 5(a), (b), (c), (d) に示す．また，Fig. 5(a) には目的音声信号の DOA の結果も示す．Figure 5 より，内部雑音タイプによらず内部雑音信号の DOA が 0 °に定位していることが確かめられる．これは，ロボットに取り付けられたマイクロホンアレーで観測される内部雑音信号が，固体中を伝わる筐体振動であることが原因だと考えられる．一般的に，固体中の音速は空気中の音速よりも 10 倍以上速い．そのため，各マイクロホンに入力する内部雑音信号の時間差は非常に小さくなる．その結果，内部雑音信号の DOA は 0 °になると考えられる．よって，従来のマイクロホンアレー配置では Fig. 5(a) に示すように目的音声信号と内部雑音信号の DOA が同じ方位になってしまう．

3.3 マイクロホンアレー配置の提案

ロボット音声対話システムでは，ロボットの正面にユーザが立っている状況を想定するのが一般的である．よって，ロボットに搭載するマイクロホンアレーがブロードサイドアレーの場合，Fig. 6(a) に示すように目的音声信号と内部雑音信号の DOA が 0 °で同じ方位になる．Figure 6(a) の場合，ICA が目的音声信号に対して死角を形成すると，同時に内部雑音信号も抑圧することになり，適切に内部雑音信号を推定することができない．そこで，マイクロホンアレー配置をブロードサイドアレーからエンドファイアアレーに回転させることを提案する．これによって，マイクロホンアレーに同位相で入力する内部雑音信号の DOA は，Fig. 6(b) のように，目的音声信号とは異なる方位にマッピングされる．その結果，ロボット音声対話システムにおけるユーザとロボットの対面関係を保ったまま，目的音声信号と内部雑音信号の DOA を空間的に異なる方位にすることができ，SBSS の雑音推定性能が改善すると考え

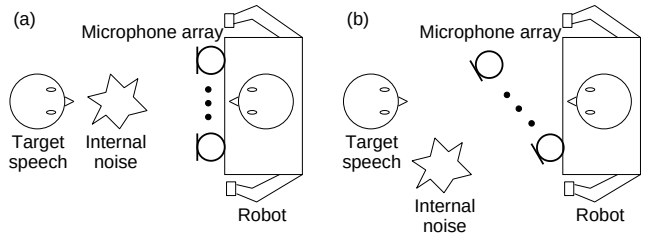


Figure 6: (a) Conventional and (a) proposed microphone array structures.

られる．本実験で用いている内部雑音抑圧型ロボットにおける最適なマイクロホンアレー配置 (θ_{m}) を以下の実験により評価する．

4 実験

4.1 実験条件

マイクロホンアレー配置の有効性を検証するために，シミュレーション実験を行った．実験における音響環境は，2.4 節で示した環境と同じである．目的音声信号には，Fig. 3 で収録したインパルス応答を畳み込んだ日本語新聞記事読み上げ音声コーパス (Japanese newspaper article sentences: JNAS) のテストセット 100 文 (女性発声) を用い，内部雑音信号は前述の 4 種類を用いた．目的音声信号と内部雑音信号の入力 SNR は，内部雑音タイプ 1 が 16.6 dB，タイプ 2 が 4.54 dB，タイプ 3 が 0.39 dB，タイプ 4 が 5.37 dB である．さらに，目的音声信号と環境雑音信号の入力 SNR が 10 dB になるように実収録の環境雑音信号を付加した．音声認識実験の条件を表 1 に示す．WF の処理強度パラメータ β は内部雑音タイプによらず 5 で固定した．以上の実験条件において，Fig. 3 に示す θ_{m} を -90 °から 90 °の範囲で計 11 パターン変化させて実験を行った．

4.2 実験結果

マイクロホンアレー配置を回転させたときの単語認識精度，SD，雑音抑圧量 (noise reduction rate: NRR)，ケプストラム歪み (cepstral distortion: CD) の比較を行った．NRR は出力 SNR [dB]− 入力 SNR [dB] で定義され，値が大きくなるほど雑音抑圧性能が高いことを表す指標である．また，CD はスペクトル包絡の歪み具合を表す尺度で，値が小さいほど処理による歪みが小さいことを表す指標である [11].

Figure 7 に音声認識実験の結果を示す．Figure 7 より，全ての内部雑音タイプにおいて $\theta_{\mathrm{m}} = 60$ °のときのほうが $\theta_{\mathrm{m}} = 0$ °よりも単語認識精度が優れていることが分かる．$\theta_{\mathrm{m}} = 60$ °のとき $\theta_{\mathrm{m}} = 0$ °に比べて，単語認識精度が内部雑音タイプ 1 では 14%，タイプ 2 では 11%，タイプ 3 では 7%，タイプ 4 では 13%改善された．また，SBSS 地点における全ての雑音信号における SD を Figs. 8(a), (b), (c), (d)，最終出力音声信号における NRR と CD の結果を Figs. 9(a), (b) に示す．Figure 8 より，$\theta_{\mathrm{m}} = 60$ °のときのほうが SD が小さくなっており，SBSS の雑音推定性能が改善されていることが分かる．また Fig. 9(a) より，NRR はタイプ 3 以外の内部雑音タイプではほとんど性能が変わらないことが示されている．さらに Fig. 9(b) より，CD は全ての内部雑音タイプにおいて小さくなっていることが分かる．これらの結果から，従来のマイクロホンアレー配置の場合，雑音推定性能が高くないことにより，WF において雑音信号を多く見積もり雑音抑圧されたため，目的音声信号の CD が大きかったと考えられる．しかし提案するマイクロホンアレー配置では，雑音推定性能が改善されたことにより適切に雑音抑圧を実現することができ，雑音抑圧性能を低下させずに CD 値を改善することができたと考えられる．

Table 1: 音声認識実験の条件

テストデータ	JNAS テストセット (女性話者 23 名 100 文)
音声認識タスク	新聞記事読み上げ (語彙数: 20 k)
音響モデル	音素内タイドミクスチャーモデル (phonetic-tied mixture model: PTM) [12]に基づく 25 dB オフィス雑音重畳モデル
音響モデルの学習データ	JNAS 260 話者 (1 話者あたり 150 文)
認識デコーダ	Julius ver. 3.5.1

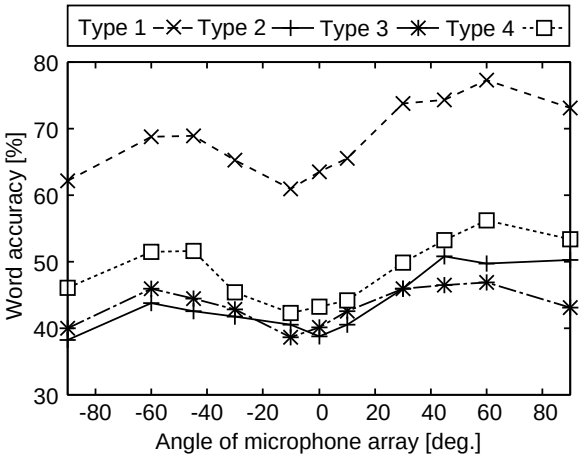


Figure 7: Word accuracy for all internal noise types with different angles of microphone array.

5 まとめ

本論文では，内部雑音の DOA が固体を伝わる振動であることによって 0 °に定位することを示した．この知見に基づき，マイクロホンアレーの配置を回転させることを提案し，音声認識実験によりその有効性を確認した．この提案はロボットにおける内部雑音に限らず，同じように固

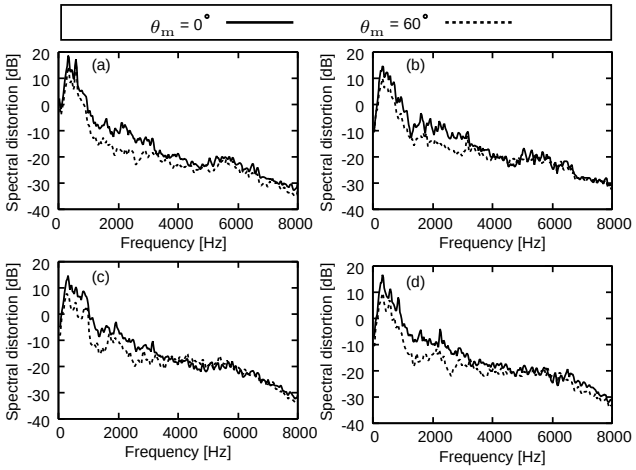


Figure 8: Experimental results of spectral distortion of internal noise: (a) type 1, (b) type 2, (c) type 3 and (d) type 4 for cases of $\theta_{\mathrm{m}} = 0$ ° and $\theta_{\mathrm{m}} = 60$ °.

体中を振動して伝わる車のロードノイズにも応用することができると考えられる [13].

謝辞

本研究の一部は総務省・戦略的情報通信研究開発推進制度 (SCOPE) の支援を受けた.

参考文献

[1] J. Even, et al., "Frequency domain semi-blind signal separation: application to the rejection of internal noises," *Proc. International Conference on Acoustic Speech and Signal Processing,* pp. 157–160, 2008.

[2] P. Loizou, *Speech Enhancement: Theory and Practice,* CRC Press, 2007.

[3] J. Even, et al., "Semi-blind suppression of internal noise for hands-free robot spoken dialog system," *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems,* pp. 658–663, 2009.

[4] P. Comon, "Independent component analysis, a new concept?," *Signal Processing,* vol.36, pp. 287–314, 1994.

[5] M. Brandstein, et al., *Microphone Arrays Signal Processing Techniques and Applications,* Springer-Verlag, 2001.

[6] A. Lee, et al., "Julius? An open source real-time large vocabulary recognition engine," *Proc. Eur. Conf. Speech Commun. Technol.*, pp. 1691–1694, 2001.

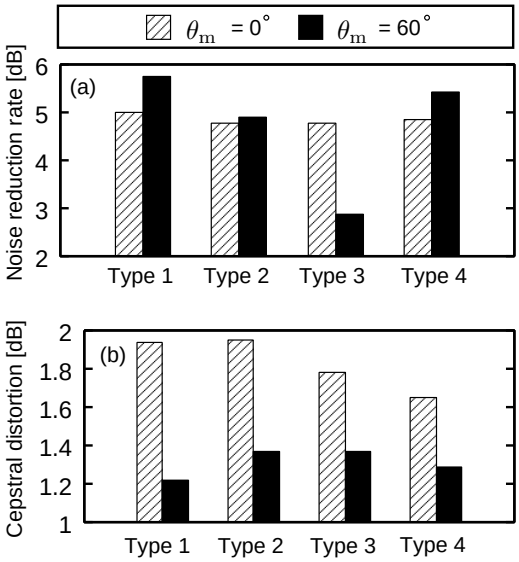


Figure 9: Experimental results of (a) noise reduction rate (b) cepstral distortion for cases of $\theta_{\mathrm{m}} = 0$ °and $\theta_{\mathrm{m}} = 60$ °.

[7] N. Vlassis, et al., "Efficient source adaptivity in independent analysis," *IEEE Trans. Neural Networks,* vol.12, no.3, pp. 559–566, 2001.

[8] J. Even, et al., "An improved permutation solver for blind signal separation based front-ends in robot audition," *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems,* pp. 2172–2177, 2008.

[9] Y. Takahashi, et al., "Blind spatial subtraction array for noisy environment," *IEEE Trans. Audio, Speech, and Language Processing,* vol.17, no.4, pp. 650–664, 2009.

[10] H. Saruwatari, et al., "Blind source separation combining independent component analysis and beamforming," *EURASIP J. Applied Signal Porc.,* vol.2003, no.11, pp. 1135–1146, 2003.

[11] L. Rabiner, et al., *Fundamentals of speech recognition,* Upper Saddle River, NJ: Prentice Hall PTR, 1993.

[12] A. Lee, et al., "A new phonetic-tied mixture model for efficient decoding," *In Proceedings of ICASSP.*, pp. 1269–1272, 2000.

[13] H. Saruwatari, et al., "Speech enhancement in car environment using blind source separation," *Proc. International Conference on Spoken Language Processing,* pp. 1781–1784, 2002.

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

人工知能学会研究資料
JSAI Technical Report
SIG-Challenge-A902-7

Automatic Speech Recognition Under Ego-motion Noise of a Robot

Gökhan Ince [1,3], Kazuhiro Nakadai [1,3], Tobias Rodemann [2], Yuji Hasegawa [1], Hiroshi Tsujino [1] and Jun-ichi Imura [3]

[1]Honda Research Institute Japan Co., Ltd. 8-1 Honcho, Wako-shi, Saitama 351-0188, Japan,
{gokhan.ince, nakadai, yuji.hasegawa, tsujino}@jp.honda-ri.com

[2]Honda Research Institute Europe GmbH, Carl-Legien Strasse 30, 63073 Offenbach, Germany,
tobias.rodemann@honda-ri.de

[3]Dept. of Mech. and Env. Informatics, Graduate School of Information Science and Eng.,
Tokyo Institute of Technology 2-12-1-W8-1, O-okayama, Meguro-ku, Tokyo, 152-8552, Japan,
imura@mei.titech.ac.jp

Abstract

Active auditory perception related tasks like sound localization and speech recognition have to be performed with high accuracy even while the robot is moving. However, the joints of the robot inevitably generate noise because of the active motors, i.e. ego-motion noise. This problem is very critical, especially in humanoid robots, because they tend to have a lot of joints and the motors are located relatively closer to the microphones than the sound sources. In this work, we investigate methods for the prediction and suppression of the ego-motion noise. In the first part, we analyze the performance of different noise subtraction strategies, assuming that the noise prediction problem has been solved. In the second part, we present some results for a noise prediction scheme based on the current robot joint status. Performance is evaluated for a number of criteria, including Automatic Speech Recognition (ASR). We demonstrate that our method improves recognition performance during ego-motion considerably.

1 Introduction

An active auditory perception system is very essential for robots to be able to interact with their environment. Tasks like sound localization and speech recognition have to be performed with high accuracy even when the head (or whole robot) is moving. Unfortunately, the research done in the field of active audition suffers highly from this additive motor noise, which deteriorates the quality of the recorded sounds considerably. Therefore two restricting assumptions are made very often: Either the sounds are selected loud enough to ignore the motor noises generated during the body motion, or the sound processing is performed without movement at all [1]. An alternative method that overcomes the noise problem is utilization of a separate close-talk microphone [2], nevertheless it limits human-robot interaction.

In our research, the goal is to tackle the noise problem directly. We propose to utilize a biologically-inspired method for learning and suppressing the ego-noise that *weakly-electric fishes* exploit in the nature. They have evolved sensory systems that make use of copies of their self-generated dynamic electric wave patterns to decode the temporal characteristics of incoming sensory signals from the surrounding waves [3]. Localization and scene analysis procedures involve the computation of the spatial map of sensory expectations from recent inputs, and removal of the ego-motion effects, namely the spike events, from the total input image [4]. The ego-noise cancellation on a robot could be accomplished by autonomous mechanisms similar to the electrosensory system of the electric fishes, just like the way the animal learns what kind of noise template it has to subtract in case of the execution of a certain motor plan. In this paper, we first deal with fixed motion patterns that follow known trajectories. This approach is suitable for focusing on the noise suppression problem explicitly. Then, we generalize the ego-noise problem for freely moving robots by showing methods how the noise could be predicted. We demonstrate that the proposed methods can eliminate motor noise by evaluating them qualitatively in terms of ASR results.

1.1 Comparison to Related Work

In the field of "Robot Audition", noise suppression is mostly carried out using sound source seperation techniques with a microphone array [5]. However, in our case, the motors are located in the near field of the microphones and produce more like diffuse rather than directional sounds. In a standard task with robot motions where acoustic conditions such as power, frequencies and locations of the motor noise sources dynamically change at each time instance, the performance of sound source separation and ASR deteriorates drastically even when a microphone array is used. Nakadai et al [6] proposed a noise cancellation method with two pairs of microphones. One pair in the inner part of the shielding body records only internal motor noise and helps the sound localizer to distinguish between the spectral subbands that are

noisy and not noisy, and to ignore the ones where the noise is dominant. In contrary to our approach, this technique does not suppress the noise. Nishimura et al [7] estimates the ego-noise using robot's gestures and motions. With the help of the motion command, the pre-recorded correct noise template matching to the recent motion is selected from the template database and subtracted. Compared to their small set of noise template database of limited motions, we target to deal with the whole ego-noise space that is generated by any possible motor combination of the robot. Ito et al [8] developed a new approach of frame-by-frame based prediction with a neural network (NN) to cope with unstable walking noise. The trained network had to predict the noise spectrum from angular velocities of the joints of the robot. However, they concentrated on a small robot with limited degrees of freedom. For a huge dataset, NN will have a slow training speed and online adaptation is difficult to achieve. Therefore we rather propose the usage of a template database due to its efficiency and additionally enhance the accuracy of the templates further by incorporating more information related to the joints. Besides, both Nishimura [7] and Ito [8] based their research mainly on the estimation of templates for different motions, but neither focused on the possibility of quality improvement by utilizing spectral enhancement optimization factors nor evaluated the performance with any other criteria except ASR.

2 Blockwise Template Subtraction

This section gives an outline of the noise reduction strategy that we followed. Main point of investigation in this section is clearly not the prediction of the noise, but the suppression of it. Therefore, we concentrated especially on a single motion (quick horizontal motion of the neck) generated by the experimental robot head which looks like the head of ASIMO.

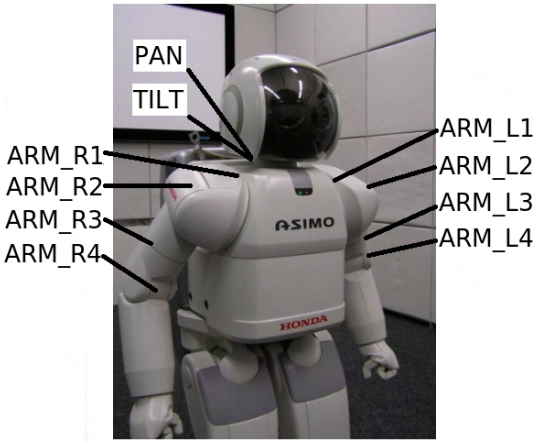


Figure 1: Hardware setup of ASIMO.

2.1 Template Generation

Spectro-temporal investigations conducted on the recorded ensemble of noise data for the same motion (same origin, target, velocity and onset time) revealed following results:

- The regions of the spectrum where noise power is densely distributed, correspond to the increased rotational velocity of the motor (see Fig. 2 for the case of one active joint). Most critical phases are acceleration and breaking.
- The energy distribution remains nearly the same during the constant velocity phase.
- The duration of the signals does not change by more than a few samples.
- Envelope shape does not deviate much from the mean envelope of the same motions.

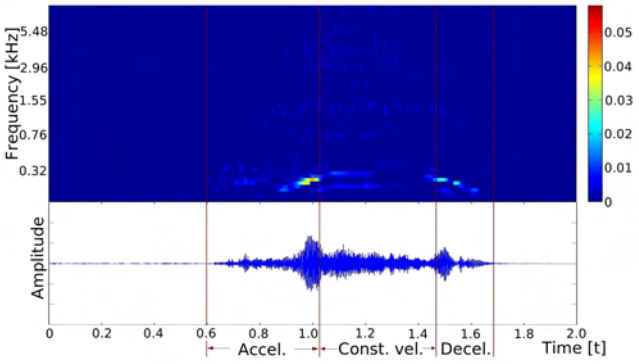


Figure 2: Envelope spectrum of the head motor noise for a rotation from -70° to 70° in the horizontal plane

The underlying notion for our first method, *blockwise template estimation*, relies on the idea that the motor noise can be predicted, if the motion performed by the robot has a pattern of a prior known duration and onset time. Noise spectra of different motions can be recorded by repeating the same motion M times. An important preprocessing step after short-time spectral decomposition is the removal of stationary background noise, which involves an adapted version of Cohen's Minimum Controlled Recursive Averaging [9]. Furthermore, the electrical noise of the motors (static noise caused by the electrical circuits) is also suppressed by this background noise reduction scheme, so that only non-stationary mechanical noise remains as a final product of the processing chain. Template generation follows as the consequent stage. Time alignment of recorded motor noise is required before calculating the templates. The synchronization point regarding each element is determined at the sample number where the cross-correlation function of each spectrum and pilot (one specific pre-selected instance) spectrum gets its maximum value. Let $D(n, \Omega)$ be the short-time basis frequency spectrum of the distortion (motor noise), where Ω stands for the discrete frequency representation and n for the current frame. A single template is represented by an average matrix $\overline{D}(n, \Omega)$ and a standard deviation matrix $\sigma_D(n, \Omega_i)$ such as follows:

$$\overline{D}(n,\Omega) = \frac{1}{M}\sum_{k=1}^{M} D(n,\Omega) \quad (1)$$

$$\sigma_D(n,\Omega) = \sqrt{\frac{1}{M}\sum_{k=1}^{M}\left(D_k(n,\Omega) - \overline{D}(n,\Omega)\right)^2} \quad (2)$$

2.2 Template Subtraction

Let $S(n,\Omega)$ and $D(n,\Omega)$ be the spectrum of useful signal and motor noise, respectively. Then the spectrum of the observed signal $Y(n,\Omega)$ is defined by

$$Y(n,\Omega) = S(n,\Omega) + D(n,\Omega). \quad (3)$$

The spectrum of the useful signal can be estimated by using the inverse operation:

$$Y_r(n,\Omega) = Y(n,\Omega) - \overline{D}(n,\Omega), \quad (4)$$

where $Y_r(n,\Omega)$ stands for the spectral magnitude comprising the magnitudes of useful sound and residual motor noise. The reason for the existence of this residual magnitude is that the original magnitudes of the motor noise $D(n,\Omega)$ deviate from their arithmetic mean $\hat{D}(n,\Omega)$. To compensate this error, we further suggest to use spectral subtraction approach that exploits *overestimation factor*, α, and *spectral floor*, β. α, also termed *aggressiveness factor*, allows a compromise between perceptual signal distortion and noise reduction level. On the other hand, β is required to deal with the problem called *musical noise*. The cause of musical noise is a nonlinear mapping of the negative or small-valued spectral estimates, producing a metallic noise sounding like the sum of tone generators with random fundamental frequencies which are turned on and off constantly [10]. β reduces the effect of the sharp valleys and peaks in the spectrum which is caused by the smaller attenuations of the frequencies compared to relatively larger attenuations of their neighboring frequencies due to the random fluctuations in the magnitude estimations. *Overestimated template subtraction* is introduced such as in the following formula:

$$\hat{H}_{SS}(n,\Omega) = max\left(1 - \alpha(n,\Omega)\frac{\hat{\sigma}_D(n,\Omega)}{Y_r(n,\Omega)}, \beta(n,\Omega)\right), \quad (5)$$

Finally, the template is conceptually 'subtracted', by weighting the signal $Y_r(n,\Omega)$ with the gain coefficients $\hat{H}_{SS}(n,\Omega)$:

$$\hat{S}(n,\Omega) = Y_r(n,\Omega)\cdot\hat{H}_{SS}(n,\Omega) \quad (6)$$

3 Parameterized Template Subtraction

In this section, we explain the techniques that are necessary to extend the proposed solution of the ego-noise reduction problem from a stereotyped motion level towards complicated motions with higher degrees of freedom. So far disregarded subjects like synchronization of templates, effect of increased number of motors and noise prediction are inspected further in this section.

Note that the *blockwise template subtraction* had several shortcomings, e.g. it could be performed properly only after the detection of the exact starting moment of the template, which is a very hard task to achieve. Another drawback was that it would require a large collection of signal representations consisting of the motor noise statistics like average values and standard deviations of the whole dataset of a given motion. Besides, it requires a huge amount of data for each possible motion. Considering the impossibility to collect and produce templates for each joint of different combinations of origin, target, position, velocity and acceleration parameters, the former approach was simply not feasible to be applied in a realistic scenario.

To overcome these deficits, a new technique is proposed that parameterizes a discrete audio segment under consideration using motor status and get a spectral energy vector to represent the ego-noise at that time instant. The experiments for parameterized template subtraction are conducted on Honda (humanoid robot) ASIMO (Fig. 1) due to the necessity of additional body joints beside the head motors. ASIMO has sensors that measure the angular positions of all of its joints separately.

3.1 Template Generation

For that purpose, joint status information provided by the sensors on the motors will be utilized, with the following assumptions:

- Current motor noise depends on position, velocity and acceleration of that specific motor.
- Similar combinations of joint status will result in similar motor noise spectral vectors at any time instance.
- The superposition of single joint motor noises at any arbitrary time equals to the whole body noise at that specific time instance.

Figure 3 illustrates the proposed template generation scheme. During the motion of the robot, actual position (θ) information regarding each motor is gathered regularly. Using the difference between consecutive sensor outputs, velocity ($\dot{\theta}$) and acceleration ($\ddot{\theta}$) values are calculated. Considering that N joints are active, feature vectors consisting of $3N$ attributes are generated. Each feature is normalized to [-1 1] so that all features have the same contribution on the prediction. The resulting feature vector has the form of $F = [\theta_1, \dot{\theta}_1, \ddot{\theta}_1, \theta_2, \dot{\theta}_2, \ddot{\theta}_2, \ldots, \theta_N, \dot{\theta}_N, \ddot{\theta}_N]$. At the same time, motor noise is recorded and spectrum of the motor noise is calculated by the sound processing branch running in parallel. Both feature vectors and spectra are continuously labeled with time tags so that templates are generated when their time tags match. Finally, a large noise template database that consists of short noise templates for many joint configurations is created.

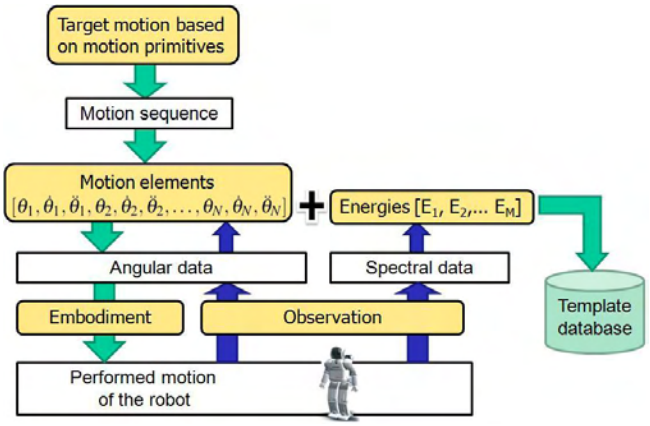


Figure 3: Flowchart of the proposed template generation and database creation.

3.2 Template Prediction and Selection

The prediction phase starts with a search in the database for the best matching template of motor noise for the current time instance. Finding the correct template involves a search among all the templates in the database for most similar joint configuration. We implemented a nearest neighbor (1-NN) search to accomplish this task. The spectral vector associated with the point in the database that has the shortest distance to the query point is used as the template. The prediction process is applied for every frame. In that sense, the block template for an arbitrary motion (e.g. neck motion template in Sec. 2) can be regarded as the concatenation of smaller templates that are predicted according to the abovementioned approach on a frame-by-frame basis with the following setting:

For a given database $\mathbf{S}$ of templates in $3N$-dimensional feature space V and a query point $\mathbf{q} \in V$, find the closest point in $\mathbf{S}$ to $\mathbf{q}$. V is taken to be the $3N$-dimensional Euclidean space and the distance is measured by the Euclidean distance between two points $\mathbf{q} = (q_1, q_2, ..., q_{3N})$ and $\mathbf{s} = (s_1, s_2, ..., s_{3N})$, where $\mathbf{s}$ is an element of the set $\mathbf{S}$.

$$d(\mathbf{q}, \mathbf{s}) = ||\mathbf{q} - \mathbf{s}|| = \sqrt{\sum_{i=1}^{3N} (q_i - s_i)^2} \qquad (7)$$

The spectral vector associated with $\mathbf{s}$ having the shortest distance to $\mathbf{q}$ is used as the template.

3.3 Template Subtraction

On contrary to blockwise template subtraction, there is no ready-to-use average template for parameterized template subtraction. Occasionally, the prediction accuracy could even become very low. In this respect, we employ a slightly changed version of weight calculation formula for spectral subtraction:

$$\hat{H}_{SS}(n, \Omega) = max\left(1 - \alpha(n, \Omega)\frac{D_{pr}(n, \Omega)}{Y(n, \Omega)}, \beta(n, \Omega)\right), \qquad (8)$$

where $D_{pr}(n, \Omega)$ stands for predicted template. This operation is followed by Eq. 6 to finish the noise reduction operation.

4 Results

For the first part of our experiments, we evaluated the blockwise template subtraction. Tests are done on the robot head which is a close derivative of the actual ASIMO head. It is equipped with Sennheiser DPA 4060-BM omni-directional microphones for recording. We used only one microphone on the left side. For more information regarding the ears and pinnae refer to [11]. The head motor is an Amtec Robotics PowerCube070. Data was recorded in a noisy, very echoic room ($T_{60} = 1100ms$). The tests for blockwise template subtraction are focused on motor noise signals generated by a horizontal motion of 140° with a very high angular velocity ($v_{max} = 200°/sec$). Sampling rate was set to 48kHz. We used a Gammatone filterbank with 60 channels where center frequencies are increasing quasi-logarithmically from 100Hz to 10 kHz.

The obtained noise signals are added to clean male speech. Not only the signal-to-noise ratio (SNR) is very low (nominalSNR=-5.7dB and segSNR=-2.8dB), but also the frequency bins with high energy content of both speech and noise are overlapping. These signals and their spectrally enhanced versions after noise reduction are evaluated using Perceptual Evaluation of Speech Quality (PESQ, ITU-T P.862 Standard). It is designed to calculate an index value of quality that correlates to a mean opinion score (MOS) given by human subjects in evaluation sessions. It predicts subjective opinion scores of a degraded audio sample in a range from 4.5 to -0.5, with higher scores indicating better quality. Results in relation with α and β are given in Tab. 1. When the aspect of *intelligibility* is considered, overestimated subtraction with low spectral floor is not appropriate for speech enhancement, because the human ear is especially sensitive to musical noise. Therefore, high spectral floor values ($\beta > 0.3$) are desirable. The best score is achieved when mean template subtraction is applied.

Table 1: PESQ results for Magnitude Spectral Subtraction

MOS for noisy signal: 0.361		MOS after mean template subtraction: 2.681					
MOS values		β					
		0.0	0.2	0.3	0.5	0.8	0.9
α	1	0.329	0.309	0.277	0.221	1.526	2.429
	1.5	0.322	0.291	0.216	0.249	1.535	2.606
	2	0.313	0.312	0.262	1.448	1.541	2.592
	2.5	0.331	0.248	0.244	1.464	1.545	2.194
	3	0.35	0.241	0.291	1.476	1.546	2.619
	4	0.255	0.338	1.371	1.439	1.547	2.62

The second evaluation criteria we utilize, exploits the *Precedence Effect* [12], which makes localization in echoic

environments possible for humans. Using this model, the detection of noise and sound signals is to be verified on their onset points. Onsets are the points where a position measurement for sound localization is done. They are frames where the signal amplitude increases and the effect of echoes is still small. Therefore we assume, the larger the energy of the onset, the larger the impact on localization. They are used particularly for sound localization in order to suppress the onsets caused by the echoes of the same sound source, by introducing the inhibition of the local echo onset points other than these particular desired signal onsets (See [13]).

Provided that the noisy signal consists of the superposition of the noise and speech signals, the onsets of both signals can be extracted separately by giving only the interested signal to the input. That way, the energies and positions of the onsets are saved individually. A likelihood method is introduced so that the onsets of the degraded signal can be compared with the onsets of its noise and speech components assessed before. Given a certain confidence area (explained below), it should return an objective measure how likely the onsets of the degraded signal are to its nearest onset belonging to either one of those classes.

Considering that the onsets are computed for each channel, two parameters are selected to tune the confidence area, namely the *timing* and the *energy* of the onsets. An optimized *timing* confidence interval of 60 ms defines the limit of interest for the corresponding onsets. The onsets beyond the limits are considered as completely dissimilar onsets. The second parameter in the confidence area is the *energy* level of the onset. The onset of a class whose energy seems to be reduced far more than the other class is rewarded more. The total confidence value (product of the position and energy confidence) acts as an indicator for the competition between the noise and speech onsets in the reference onset set. The candidate which has the greatest confidence value is selected as the winner and the onset is assigned to belong to either speech, motor noise, or indecisive category. This method gives out a measurement bench how many onsets from the noise are suppressed, how much energy has remained in the onsets of the noise (see Fig. 4).

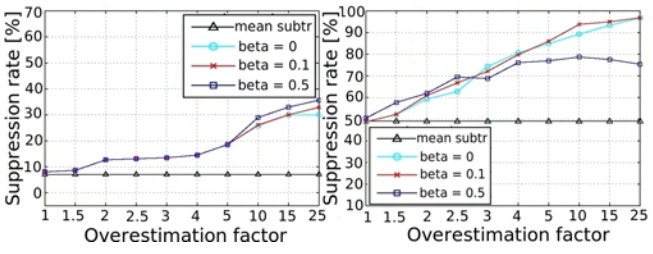


Figure 4: Onset based results using magnitude spectral subtraction for (a) voice onset energy suppression rates (b) noise onset energy suppression rates

The results demonstrated that the higher the overestimation factor is selected, the more the noise reduction is achieved. Template reduction can suppress 76% of the total energy of noise onsets, while keeping voice suppression in low rates like 15%. (see Tab. 1 and Fig. 4 for $\alpha = 4$ and $\beta = 0.5$)

We also evaluated the speech recognition results with Sphinx-4 to inspect the *qualitative* aspects of our noise suppression scheme. Totally 200 evaluation word sequences (Resource Management Speech Corpus) are selected each comprising of 5 to 12 words chosen randomly. Utterances belong to both male and female speakers. The recognition is performed speaker- and gender- independent. No grammar is used in the tests. The results will be evaluated for 7 different SNR values between approximately -10 and 40 dB.

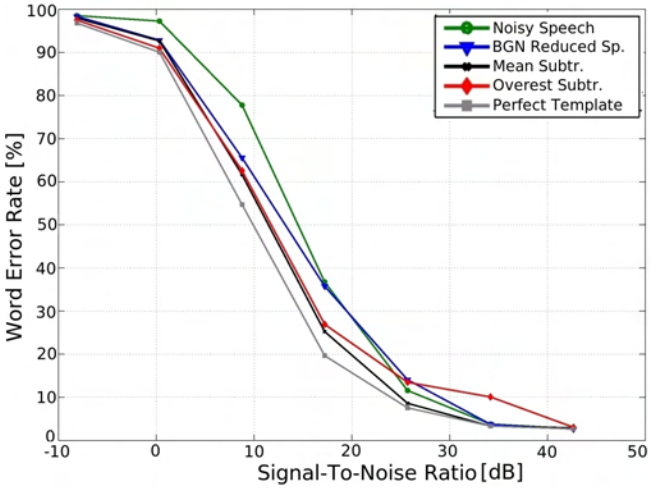


Figure 5: ASR results

The experiments carried out with ASR show that the word error rate after mean template subtraction decreases substantially in the sensible region between -5dB and 30dB compared to both the reference recognition results with noisy signals and to the results after applying stationary background noise reduction scheme as shown in Fig. 5. For an SNR value of 17dB, the improvement is 12% and for 8dB case 16% improvement is achieved. For an additional test bench, the recognition performance of an perfect template is introduced as well. This perfect template is in fact nothing but the identical spectrogram that the motor noise has. This defines the upper limit of performance and defines a benchmark for the comparison of all methods by providing a best case scenario.

It is also clear that the templates generated by variance weighting (Overestimated Temp. Subtraction) are not suitable to be applied to the signals with high SNR. They worsen the Word Error Rate (WER), which is an expected consequence coinciding with the results obtained from the previous PESQ and onset measurement tests. However, recognition for low SNRs (below 0dB) yields better performance if an overestimation of the noise variance is used (within a certain range). For moderate SNR levels, usage of variance weighting techniques reduces WERs by up to 10%.

The second part of the experiments is carried out on ASIMO. Experiment involves random motions of 10 different joints simultaneously. We rotated the head of

ASIMO (elevation = [-30° 30°], azimuth = [-90° 90°]) randomly, while the arms were performing a random grasp motion in the reaching space of the body without moving its torso or hip. Status information of the motors are gathered from the joints with an average acquisition rate of 7.3 ms. ASIMO also has a circular array consisting of 8 microphones mounted on the head. We made evaluations using the data recorded from the third microphone that corresponds to a spatial position of 90° counterclockwise with respect to the front. The training data was a joint database consisting of 30 minutes of motor noise and the corresponding feature vectors stored during this time span. The probability was very high that no similar motions could be generated for this scenario with another arbitrarily generated random trajectory. In that case, the performance of the experiments would be biased by the inappropriately selected test set. Therefore, we followed a similar trajectory used in the training session but with a sequence of slightly different destination points as before that deviated in their final positions by a certain random displacement. This distance is determined by a Gaussian distribution with a variance of σ=0.1. We stored a test database of 10 minutes long. Data is recorded in a noisy and echoic room (reverberation time (RT20) was about 0.2 seconds). Data was sampled on 16kHz and frame shift was 12.5 ms. Hamming window of 16 ms was used.

We evaluated the effectiveness of the proposed approach using Julius which is an open-sourced ASR. For this experiment, we created 35 different motor noise patterns from the test set. The length of each test set is kept flexible so that it matches the duration of the utterances used in the wordset. As speech corpus, ATR phonemically-balanced wordset (ATR-PB) was used. This word-set includes 216 Japanese words and average word correct rate was calculated as depicted in Fig. 6. Please note that the signals were formerly subject to stationary background noise reduction. Hence, SNR values are given for signal-to-motor noise ratio (background noise of the room was approx. 5dB).

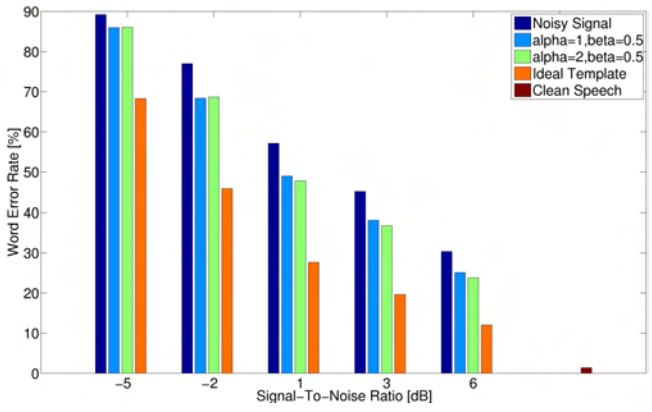


Figure 6: Recognition performance for different spectral subtraction settings

The graph shows that a template subtraction with $\alpha = 2$ and $\beta = 0.5$ is slightly better than a subtraction with $\alpha = 1$ and $\beta = 0.5$ for high SNR values. Latter set of parameters allow us to obtain improvement rates of up to 10% for SNR conditions that can be observed in a realistic human-robot interaction. Fig. 7 illustrates the ASR performance distribution on 35 different noise test sets. With the exception of two test cases (#12 and #29), high improvement rates are achieved. We also depicted the recognition rates for an ideal template subtraction for comparison. Ideal template represents the template that is constructed for the current test motor noise using the predictions from the test set. The reason of the gap between ideal template subtraction performance and the results for overestimated template subtraction with optimal settings is due to the incorrect predictions of the template. Nearest neighbor search does not make a decision on whether the final prediction is a reasonably correct template, that is why it is called a *lazy learning algorithm*. The errors are mostly caused by the absence of similar templates that are available for the current motor status combination.

Because the feature set has big impact on the prediction accuracy, we also tested the influence of the feature vector selection. For that purpose, we reduced the number of features from 30 to 20 by excluding the acceleration values. In the second condition, we eliminated the angular velocities and provided only the position and acceleration features (20 in total) for the prediction. We found out (See Tab. 2) that angular velocity and acceleration information do not provide independent features. The combination of $(\theta,\dot{\theta})$ has outperformed the other feature combinations. Additional benefit of this feature reduction is that the search algorithm works now considerably faster and is less affected by the curse of dimensionality.

Table 2: ASR performances for three feature sets ($\alpha = 2$ and $\beta = 0.5$)

	$(\theta,\dot{\theta},\ddot{\theta})$	$(\theta,\dot{\theta})$	$(\theta,\ddot{\theta})$
SNR = 1dB	48.0	**47.8**	47.9
SNR = 3dB	37.2	**36.8**	37.1
SNR = 6dB	24.2	**23.8**	24.2

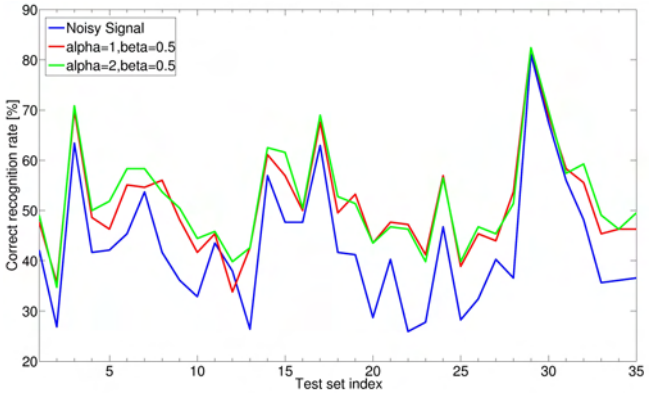


Figure 7: Distribution of the recognition performance over the noise set data

5 Summary and Outlook

In this paper, we have presented methods for removing ego-motion noise from sound signals. We showed that there is a trade-off between quality and intelligibility of the speech. Results are very promising in the sense that high suppression rates are achieved while keeping the speech as untouched as possible. We also demonstrated methods to maintain the same intelligibility while improving the quality of the speech by tuning the spectral subtraction parameters, α and β. We suggest to choose these parameters depending on one's purpose in using the enhancement algorithm: If the aim is sound localization, template subtraction can be used aggressively to remove the onsets originating from motor noise. For speech recognition, however, no harm to the speech signal can be tolerated, hence only milder suppression is recommended. We have also investigated methods for noise prediction based on joint status information. Results are preliminary, but they show that described concept works.

In its current form, our system has difficulties in achieving precise prediction of templates. Therefore, additional features that utilize cues about time series expansion of consecutive motion elements and incorporate information on motion primitives and motion-sequences would improve the reliability and performance of the predictions. Next steps involve an online implementation of the template subtraction scheme on ASIMO that performs motions using more joints. Besides, more sophisticated online compatible learning and indexing techniques will increase the speed of our approach and endow the system with a capability of online adaptation. An important advantage of parameterized approach would be that it can update the database on the fly making the prediction more adaptive and accurate in case any change in the characteristics of the motor noise (e.g. due to heating or alterations in the material) occurs at any time. Moreover, it can run online on the background while the robot is performing its duties and tasks. In order to improve the robustness, we plan to embed the current single-channel ego-noise reduction stage into a general multi-channel microphone array processing framework for speech recognition that utilizes geometric source separation and post filtering.

6 Acknowledgements

We thank Dr. Björn Schölling for his fruitful discussions.

References

[1] T. Rodemann, M. Heckmann, B. Schölling, F. Joublin and C. Goerick "Real-time sound localization with a binaural head-system using a biologically-inspired cue-triple mapping", *Proc. of the IEEE/RSJ International Conference on Robots and Intelligent Systems (IROS)*, 2006.

[2] M. Nakano, A. Hoshino, J. Takeuchi, Y. Hasegawa, T. Torii, K. Nakadai, K. Kato and H. Tsujino,"A Robot that Can Engage in Both Task-oriented and Non-task-oriented Dialogues", *Humanoids*, pp.404-411, 2006.

[3] B. Rasnow and J. M. Bower, "Imaging with electricity: how weakly electric fish might perceive objects", *Proceedings of the annual conference on Computational neuroscience : trends in research*, 1997.

[4] P. D. Roberts, "Modeling Inhibitory Plasticity in the Electrosensory System of Mormyrid Electric Fish", *The Journal of Neurophysiology*, vol. 84, No.4, 2000.

[5] S. Yamamoto, K. Nakadai, M. Nakano, H. Tsujino, J. M. Valin, K. Komatani, T. Ogata, and H. G. Okuno, "Real-time robot audition system that recognizes simultaneous speech in the real world", *Proc. of the IEEE/RSJ International Conference on Robots and Intelligent Systems (IROS)*, 2006.

[6] K. Nakadai, H.G. Okuno, H. Kitano, "Humanoid Active Audition System Improved by The Cover Acoustics", *PRICAI 2000 Topics in Artificial Intelligence (Sixth Pacific Rim International Conference on Artificial Intelligence)*, 544–554, Springer Lecture Notes in Artificial Intelligence No. 1886, 2000.

[7] Y. Nishimura, M. Nakano, K. Nakadai, H. Tsujino and M. Ishizuka, "Speech Recognition for a Robot under its Motor Noises by Selective Application of Missing Feature Theory and MLLR", *ISCA Tutorial and Research Workshop on Statistical And Perceptual Audition*, 2006.

[8] A. Ito, T. Kanayama, M. Suzuki, S. Makino, "Internal Noise Suppression for Speech Recognition by Small Robots", *Interspeech 2005*, pp.2685-2688, 2005.

[9] I. Cohen, "Noise Estimation by Minima Controlled Recursive Averaging for Robust Speech Enhancement", *IEEE Signal Processing Letters*, vol. 9, No.1, 2002.

[10] S. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-27, No.2, 1979.

[11] T. Rodemann, G. Ince, F. Joublin and C. Goerick "Using Binaural and Spectral Cues for Azimuth and Elevation Localization", *Proceedings of the IEEE/RSJ International Conference on Robots and Intelligent Systems (IROS)*, 2008.

[12] B.C.J. Moore, *An introduction to the psychology of hearing*, 5th ed. London: Academic Press, 2003.

[13] M. Heckmann, T. Rodemann, B. Schölling, F. Joublin and C. Goerick "Auditory Inspired Binaural Robust Sound Source Localization in Echoic and Noisy Environments", *Proc. of the IEEE/RSJ International Conference on Robots and Intelligent Systems (IROS)*, 2006.

24ｂｉｔオーディオボードを使ったワイドレンジ耐雑音音声認識

Wide-Range Noise Robust Speech Recognition using a 24-bit Audio Board

○荒川 隆行、田中 大介、西沢 俊弘、山下 信行
(ＮＥＣ共通基盤ソフトウェア研究所)

* Takayuki ARAKAWA, Daisuke TANAKA, Toshihiro NISHIZAWA, Nobuyuki YAMASHITA,
(NEC Common Platform Software Research Laboratories)

t-arakawa@cp.jp.nec.com, d-tanaka@rf.jp.nec.com, nishizawa@bk.jp.nec.com, n-yamashita@ax.jp.nec.com

Abstract— For some voice input systems; such as robots or car-navigation systems, distances between speakers and microphones are variable. In these cases, SNR often becomes worse and the dynamic range of voice volumes becomes wide. To deal with these noisy and wide-range voices, this paper proposes a combined method with a 24-bit audio board and noise suppressor. Speech recognition experiments shows that the performance with 24-bit audio recording is better than 16-bit audio recording and that, for noisy environment, 24-bit audio recording with noise suppressor achieves much better performance than without it.
Key Words: Speech Recognition, Gain Control, Robot.

1　はじめに

近年、ロボットやカーナビ等の様々な機器のインターフェースとして、音声対話機能が注目されている。しかしながら、これらの機器は離れた位置から音声コマンドで制御を行うため、ヘッドセットを使った近接発話に較べると性能が下がってしまうという問題がある。

この性能劣化の主な原因として、周囲雑音の影響とマイクゲインの不一致という２つの原因が考えられる。前者の周囲雑音の影響は、マイクと発話者との距離が離れるため周囲の雑音の影響を受け易くなりＳＮＲが低くなる事に起因する。後者のマイクゲインの不一致は、話者毎の発声音量の違いや発話者とマイクとの距離が一定でないために音量のレンジが広くなることに起因する。マイクゲインが合っていないと、音割れや量子化誤差による性能劣化が起こる。

前者の周囲雑音の影響を軽減する方法としては、雑音成分を推定し抑圧するノイズサプレッサ[1]や、複数マイクを用いて対象となる音声を強調するマイクロホンアレー[2]など様々な方法が提案されている。

後者のマイクゲインの不一致に対しては、マイクゲインを動的に変更するAutomatic Gain Control (AGC)が用いられてきた[3][4]。しかしながら、動的にマイクゲインを変更することは、特徴量の差分成分に悪影響を及ぼすなど音声認識の性能劣化につながることが知られている。

本稿では、上記周囲雑音の影響とマイクゲインの不一致という２つの性能劣化の原因に対し、２マイクノイズキャンセラと、24ビットオーディオボードを組み合わせたワイドレンジ耐雑音音声認識を提案する。本稿では、まず、２節で提案法の全体の構成について説明する。次に、３節で複数チャンネルのマイクに対して24ビットのオーディオデータを収録できるオーディオボードについて説明する。次に、４節で対象話者方向以外から来る妨害音を除去するノイズキャンセラについて説明する。次に、５節で24ビットのオーディオデータを音声認識向けに16ビットに変換する方法について説明する。次に６節で、今回用いた音声認識について説明し、最後に７節で複数の音量、複数のマイク距離で収録した音声データに対する音声認識の評価について説明する。

2　全体構成

全体の構成をFigure.1に示す。雑音のある環境で対象話者の音声を強調し認識するために、以下のような手順で処理を行う。

1. まず、音声マイクと雑音マイクを用いて、対象話者方向の音声と、対象話者方向以外から来る妨害音を取得する。
2. 次に、24ビットオーディオボードでこれら２つのマイクから取得されたアナログデータを24ビットのデジタルデータに変換する。
3. 次に、ノイズキャンセラで、この２つのデジタルデータから対象となる音声のみを強調し取り出す。
4. 次に、24→16ビット変換で、音声認識の入力に合うように24ビットのデジタルデータを16ビットに変換する。
5. 最後に音声認識を行う。

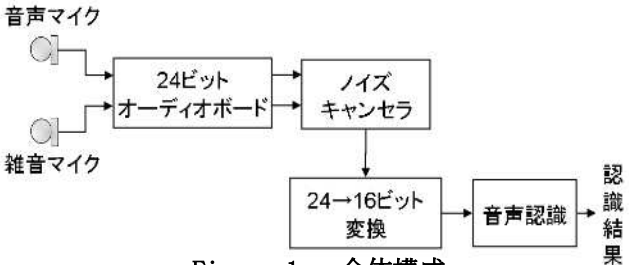


Figure 1.　全体構成

3　24 ビットオーディオボード

通常、マイクより取得されたアナログ音声データは、16ビットのA／D変換器によりデジタルデータに変換される。しかしながら、機器から数十ｃｍに近づいて入力される音声と、機器から数m離れたところから入力される音声とでは、音量のダイナミックレンジが異なるため、16ビットの範囲に収まらない。そこでよりレンジの広い音声を収録できる24ビットオーディオボード(DS-BD-24ADUSB)を試作した。今回開発したオーディオボードのスペックをTable. １に示す。

Table.1 24ビットオーディオボードのスペック

サンプリング周波数	48kHz, 32kHz, 16kHz, 8kHz, 44.1kHz, 22.05kHz, 11.025kHz（本稿では11kHzを使用）
チャンネル数	8 ch（本稿では2chのみ使用）
分解能	24 bit
サイズ	W100mm x D60mm x H20mm
インターフェース	USB 2.0
ノイズ性能	105dB

ノイズ性能とは、マイクを接続しない状態で計測したノイズレベルと最大入力レベルとの比を意味する。24ビットA／D変換の理論上の最大値は120dBである。参考までにほぼ同等の構成の16ビットオーディオボードでのノイズ性能は70dBである[5]。今回試作したボードは、マイクゲインを変更することなく35dB広いレンジの音声を扱うことができる。

4　ノイズキャンセラ

ノイズキャンセラは、音声マイクに混入する雑音成分を雑音マイクを用いて推定し、消去することで対象とする音声のみを強調する。一般的なノイズキャンセラの構成を Fig. 2 に示す。音声マイクに混入する雑音成分を推定するために、雑音マイクに入力された信号に対し適応フィルタの処理を行う。音声が大きいときは雑音の推定にとって音声は妨害信号となるため、適応フィルタのステップサイズを小さくし、更新を抑える。音声が小さく雑音が大きいときは雑音への追従性能を高めるため、適応フィルタのステップサイズを大きくする。

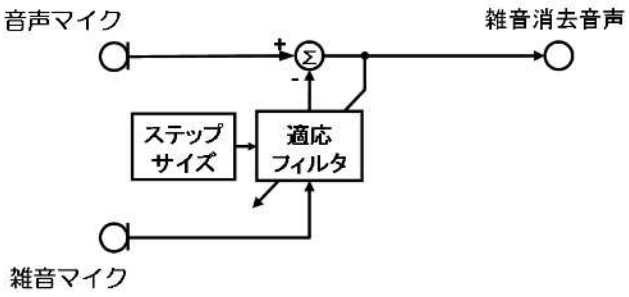


Figure 2.　一般的なノイズキャンセラの構成

しかしながら、このような一般的なノイズキャンセラでは、雑音マイクに混入する音声信号（クロストーク）を誤って雑音としてしまうために、対象とする音声を消去してしまう。

本稿では、このようなクロストークの影響を軽減するために、クロストークの推定、消去を行うノイズキャンセラを用いた。Fig. 3 にノイズキャンセラの構成を、Fig. 4 に推定 SN 比と係数更新ステップについて示す。提案法のノイズキャンセラは、クロストーク推定用の適応フィルタも備えている。雑音用フィルタは音声用マイクロホンから雑音を消去し、音声用フィルタは雑音用マイクロホンに混入する音声を消去する。上段にある２つのフィルタがメインフィルタ、下段にある２つのフィルタがサブフィルタである。雑音成分用と音声成分用の適応フィルタは、それぞれ専用のサブフィルタを用いて推定した SN 比でステップサイズを制御する。SN 比が大きく音声が支配的なときは音声用フィルタのステップサイズを大きくし、雑音用フィルタのステップサイズを大きくする。反対に SN 比が小さく座右音が支配的なときは雑音用フィルタのステップサイズを大きくし、音声用フィルタのステップサイズを小さくする。このような構成とすることにより、大きな消去量と小さな音声歪を両立する[6]。

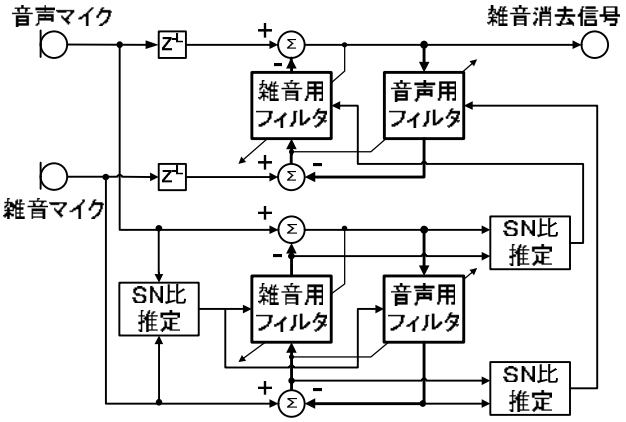


Figure 3.　提案法のノイズキャンセラの構成

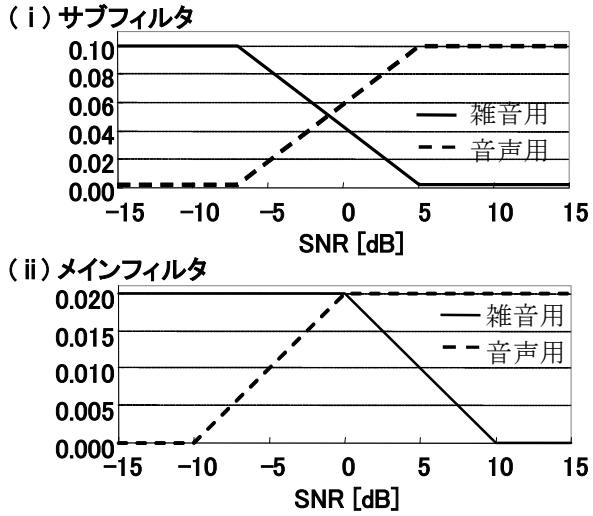


Figure 4.　推定SN比と係数更新ステップサイズ

5　24 ビットから 16 ビットへの変換

一般的な音声認識システムは、入力が16ビットであるため、本稿では16ビットのオーディオデータを入力とする音声認識システムを用いた。この為、24ビットのオーディオデータを16ビットに変換し、音声認識システムへの入力とした。変換には以下の３つの方法を実装した。

- **下位８ビットを削る**
  24ビットのオーディオデータに対し、下位８ビットを削り、上位16ビットのみを用いる。この場合は最大入力レベルを揃え16ビットで音声を収録するのと同等である。下位ビットを削ってしまうために、音量の小さな音声は量子化誤差の影響を受けやすくなる。

- **非線形処理を行い、ビット数を削減**
  Figure 5 に示す非線形関数を用いて、入力値に対して出力値を計算する。図の横軸が入力値（振幅の値）、縦軸が出力値である。24ビットのデータから符号ビットを除いた下位４ビットを削り、上位５ビット分に対し音量抑圧（コンプレッサ）を行った。この方法では、前記下位８ビットを削る方法に較べ扱える音量のレンジは広くなるが、音量の大きな音声が歪んでしまう。

- **適切な16ビットを選択する**
  環境と話者毎に予め音量の最大値を求め、その最大値が16ビットの最大値に収まるように適切な16ビットを選択する。この場合はマイクゲインを理想的に正しく設定した16ビットでの音声収録と同等となる。

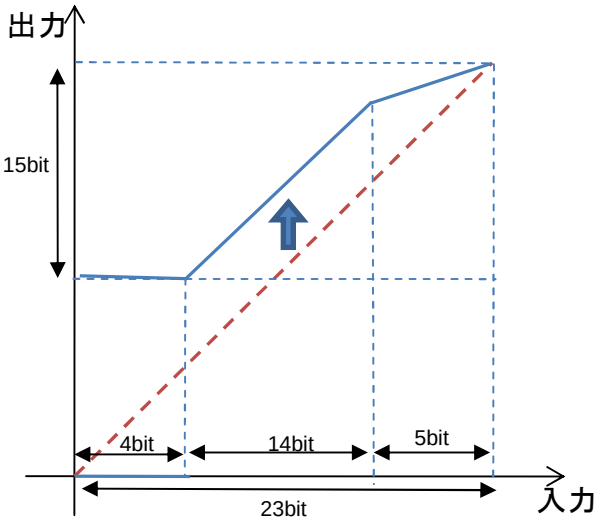


Figure 5. 非線形関数

入力は符号ビットを除く 23 ビット、出力は符号ビットを除く 15 ビットである

6　音声認識

音声認識には、単語認識エンジンを用いた。音声認識辞書は，W3C 勧告済の記述仕様である SRGS (Speech Recognition Grammar Specification)[7]のサブセットに対応している[8]。本稿では５０単語の辞書を用いた。特徴量にはケプストラムの１次から 10 次までの成分、およびその差分成分、パワーの差分、調波性特徴量およびその差分成分の計２３次元の特徴量を用いている。

7　評価実験

7.1 音量の異なる音声の収録

発声音量およびマイクとの距離の異なる音声をスピーカーより再生し、24ビットで収録した。以下の評価においてマイクゲインは一定である。スピーカー及びマイクの配置をFig. 6に、収録風景をFig.7に示す。

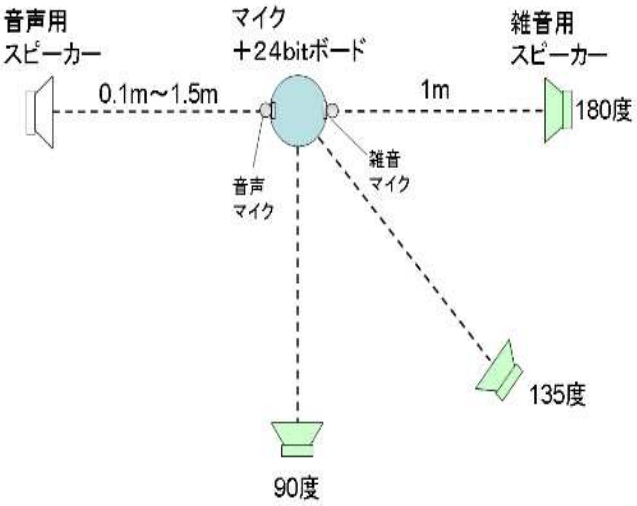


Figure 6. スピーカー、マイク配置

Figure 7. 撮影風景

7.2 音量の変化に対する評価

まず、音量の変化に対する24ビットでの音声収録単体の効果を見るために、雑音が無い環境で、ノイズキャンセラをオフにした評価を行った。評価条件をTable 2に示す。評価には音声認識を用い、単語正解率を求めた。

Table 2. 評価条件

話者	男性2名、女性2名、子供3名 （スピーカー再生）
発声内容	50単語
発声音圧※	50dBA，60dBA，70dBA
音声マイクとの距離	0.1m，0.5m，1.0m，1.5m

※発声音圧は、スピーカーとマイクを1.0m離した状態で、音声マイクの位置で計測した。

- **下位 8 ビットを削った場合の評価**

収録した音声に対し、まず下位8ビットを削った場合の評価を行った。結果をFig. 8に示す。50dBAの発声ではマイクとの距離が長くなるにつれて大きく性能が劣化している。これは音量が小さくなり16ビットでは量子化誤差が大きくなった為と考えられる。また、70dBAの発声においてマイクとの距離が0.1mの時に性能が悪くなっている。先程とは逆に音声が大きくなりすぎたために音割れが起きた為と考えられる。

- **非線形処理を行った場合の評価**

次に、非線形処理を行ってビット数を削減した場合について評価を行った。結果をFig.9に示す。50dBAの音声の劣化がなくなっていることがわかる。しかし、70dBAの音量が大きい音声の認識性能が若干劣化している。これは、非線形処理の影響により音量の大きい成分に歪みが生じたためであると考えられる。

- **適切な 16 ビットを選択した場合の評価**

次に、適切な16ビットを選択した場合の評価を行った。結果をFig.10に示す。上記２つ(Fig.8, Fig.9)に較べ音声認識率の劣化がなく最も性能が高くなっていることがわかる。

以上の評価から、24ビットで音声収録を行うことで、従来の16ビットの音声収録では対応できなかった広いレンジの音声に対してマイクゲインの変更無しに対応できることが確認できた。

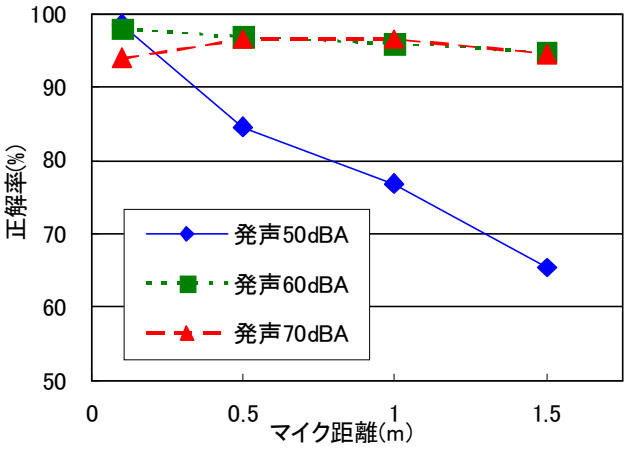


Figure 8. 下位８ビットを削った場合

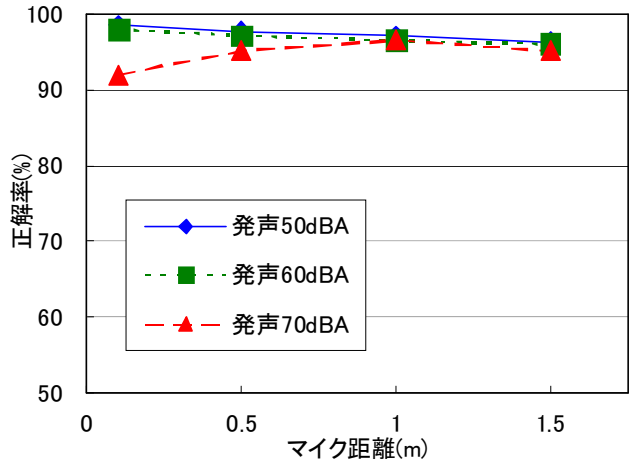


Figure 9. 非線形処理を行った場合

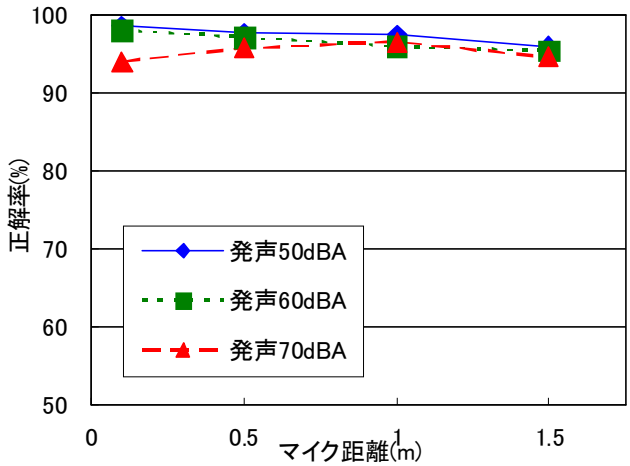


Figure 10. 適切な16ビットを選択した場合

7.3 ノイズキャンセラの効果

次に、ノイズキャンセラの効果を見るため、ノイズキャンセラを有効にし、雑音のある環境と雑音の無い環境での評価を行った。評価条件をTable 3に示す。以下の評価では、対象となる発声の音圧70dBAとし、雑音の音圧を60dBAに固定している。対象となる発声とマイクとの距離および雑音方向を変化させて評価を行った。先程と同様に評価には音声認識を用い、単語正解率を求めた。

Table 3. 評価条件

話者	男性2名、女性2名、子供3名 （スピーカー再生）
発声内容	50単語
発声音圧※	70dBA
音声マイクとの距離	0.1m，1.0m，1.5m
雑音内容	テレビCM（スピーカー再生）
雑音音圧※	60dBA
雑音マイクとの距離	1m
雑音方向	180度、135度、90度 （対象発話方向を0度とする）

※発声音圧および雑音音圧は、スピーカーとマイクを1.0m離した状態で、音声マイクの位置で計測した。

- **下位 8 ビットを削った場合の評価**

下位8ビットを削った場合の評価結果をFig. 11に示す。発声用スピーカーと音声マイクとの距離を1.5m, 1.0m, 0.1mと変えたものをプロットしている。図中の中塗りの印はノイズキャンセラを行った結果を示し、白抜きの印はノイズキャンセラを行っていないことを示す。雑音の有る環境では、３つの方向全てに対してノイズキャンセラの効果がはっきりと現れている。しかしながら、雑音が無い環境では、マイク距離1.5mの条件でノイズキャンセラを行うと性能劣化が見られる。これは、雑音マイクに回り込んだ音声を誤って雑音と判定してしまったためであると考えられる。

- **非線形処理を行った場合の評価**

非線形処理を行ってビット数を削減した場合の評価結果をFig.12に示す。雑音が有る環境では全ての条件に対しノイズキャンセラの効果が現れている。

- **適切な16ビットを選択した場合の評価**

適切な16ビットを選択した場合の評価結果をFig.13に示す。上２つ(Fig.11, Fig.12)と同様、雑音が有る環境で全ての条件に対しノイズキャンセラの効果が現れている。この場合が最も性能が高くなっている。

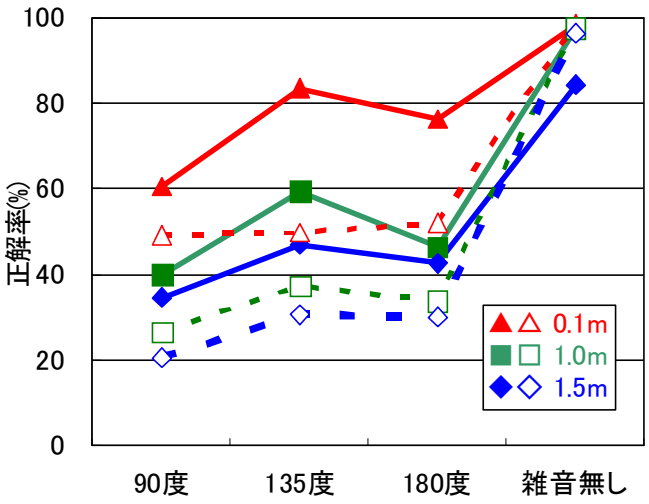


Figure 11. 下位８ビットを削った場合
中塗りの印はNCあり、白抜きの印はNCなしを示す

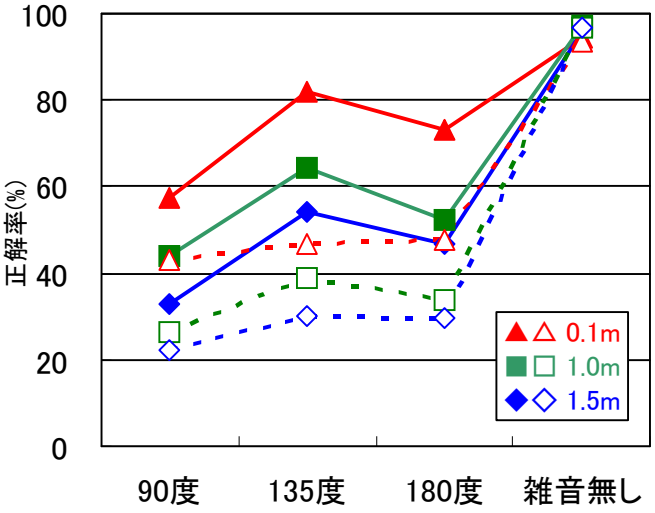


Figure 12. 非線形処理を行った場合
中塗りの印はNCあり、白抜きの印はNCなしを示す

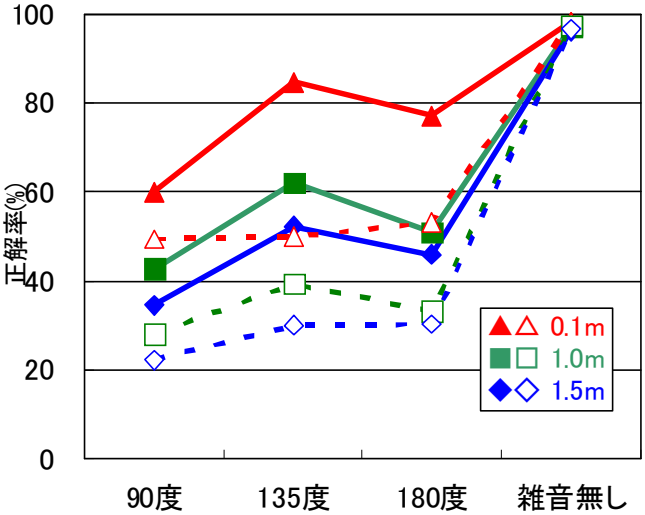


Figure 13. 適切な16ビットを選択した場合
中塗りの印はNCあり、白抜きの印はNCなしを示す

8　まとめ

離れたマイクを用いて音声認識を行う際に問題となる、雑音とマイクゲインの不一致に対して、２マイクノイズキャンセラと24ビットオーディオボードを用いたワイドレンジ耐雑音音声認識を提案し、評価を行った。

24ビットで音声を収録することで、16ビットでは対応できなかった広いレンジの音声に対応できることを確認した。また、ノイズキャンセラを用いた評価では、雑音環境ではどの条件であってもノイズキャンセラを用いなかった場合に較べて大幅な性能改善が見られた。さらに適切な16ビットを選択した場合には、雑音の無い環境でノイズキャンセラを用いても性能劣化の無いことが確認できた。

9　今後の予定

本稿では、16 ビット入力に対応する音声認識での評価を行ったが、今後は 24 ビット入力に対応する音声認識で評価を行う予定である。今回評価した適切な 16 ビットを選択する処理はバッチ処理であり一発話分処理が遅延してしまう。24 ビット入力に対応した音声認識を用いることで 24 ビットから 16 ビットに変換する必要がなくなるために、遅延の無いオンラインでの処理が可能となる。

また、話者方向推定や音声検出と組み合わせるなど、さらに雑音環境での性能向上を行う予定である。

謝辞

本研究は、平成 20, 21 年 NEDO 委託研究『次世代ロボット知能化技術開発プロジェクト』の一環として行った。本研究をご支援いただいた関係各位に感謝する。

参考文献

1) M. Kato, A. Sugiyama and M. Serizawa, “A Family of 3GPP-standard Noise Suppressors for the AMR Codec and the Evaluation Results,” ICASSP ’03 SP-P5.14, 2003.
2) M. Brandstein and D. Ward, "Microphone Arrays," Springer Verlag, Berlin, 2001.
3) 小林他, “信学論”, J87A(12), 1491-1501, 2004.
4) 寺澤, 竹山, “マルチゲイン関数自動選択型音声 AGC”, 松下電工技法(Feb), 70-74, 2003.
5) 東京エレクトロンデバイス, “16 チャンネル専用 A/D・D/A ボード”, http://www.inrevium.jp/pm/image_audio/16adusb.html
6) M. Sato, A. Sugiyama, S. Ohnaka, "An Adaptive Noise Canceller with Low Signal-Distortion Based on Variable Stepsize Subfilters for Human-Robot Communication", IEICE Trans. Fundamentals, Vol.E88-A, No.8, pp.2055-2061, Aug. 2005.
7) http://www.w3.org/TR/speech-grammar/
8) 岩沢, ”組込み機器への搭載を可能にする小型音声対話モジュールの開発,” 機械設計, 第 51 巻, 第 17 号, pp.114-118, (2007).

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

人工知能学会研究資料
JSAI Technical Report
SIG-Challenge-A902-9

SPEECH ENHANCEMENT OPTIMIZATION BASED ON ACOUSTIC MODEL LIKELIHOOD FOR NOISY AND REVERBERANT ENVIRONMENT

Randy Gomez and Tatsuya Kawahara

Kyoto University, ACCMS,
Sakyo-ku, Kyoto 606-8501, JAPAN

ABSTRACT

Noise and channel contamination acoustically degrade the speech signal. To suppress the effects of degradation and recover the original signal, speech enhancement techniques are employed. In this paper, we focus on two simple and low-computational methods: Wiener filtering (WF) and spectral subtraction (SS). Conventionally, these are formulated with no relation with automatic speech recognition (ASR). We propose to optimize the conventional speech enhancement technique in relation with likelihood of the acoustic model. We also exploit these simple speech enhancement techniques that are originally designed for denoising, to address reverberation as well. In the experiment with real noisy and reverberant environments, we have achieved significant improvement in recognition performance using the proposed approach.

Index Terms— Robustness in ASR, Dereverberation, Denoising, Spectral Subtraction, Wiener Filtering

1. INTRODUCTION

Acoustic degradation is a common problem in speech recognition applications. There have been a lot of research involving speech enhancement that are specifically designed to suppress acoustic degradation of the speech signal caused by channel and noise. One of the widely used approaches is Wiener filtering (WF) [1] where short term estimates of the noise and speech are used in defining an adaptive filter to reduce as much noise energy while removing little speech energy as possible. A number of variants have been proposed and implementations in different domains such as time, frequency and wavelet [1] [2] are investigated. Another popular enhancement technique based on spectral subtraction (SS) [3] which removes the magnitude spectrum of noise from that of the noisy speech. The noise is assumed to be uncorrelated and additive to the speech signal. A modification is given in [4] where multi-band is considered to deal with different effects of noise in different frequencies. Although these simple methods are widely used, they are formulated totally independent of the backend ASR systems.

Another approach which is linked with ASR or acoustic model likelihood is the feature transformation and adaptation [5] [6] [7]. Although these methods work well, they require a sufficient amount of adaptation data, and need some training to derive mapping parameters. These methods cannot be easily deployed in arbitrary environments especially when information of the room acoustics is not available.

In this paper, we focus on the simple enhancement algorithms: Wiener filtering (WF) and spectral subtraction (SS). Although there exist more sophisticated approaches, the enhancement schemes based on WF and SS are simple and fast to implement, which make its adoption to be effective in ASR applications. We first extend the original formulation of WF and SS to work in reverberant environments and then optimize the enhancement process in relation with ASR.

The paper is organized as follows; in Section 2, we show the method of extending both WF and SS to address reverberant conditions. In Section 3, we discuss the optimization of the scaling parameters used in WF and SS in the context of ASR followed by the RIR estimation in Section 4. Experimental conditions and results are given in Section 5, and we will conclude this paper in Section 6.

2. METHODS

The classical noisy speech model is given as,

$$y(n) = s(n) + d(n) \tag{1}$$

where $s(n)$ and $d(n)$ are the uncorrelated speech and noise signal respectively. To make use of the classical speech enhancements to work in reverberant scenario, we treat the reverberant signal analogous to that of Eq. (1). Thus, the reverberant model is given as,

$$x(n) = x_E(n) + x_L(n) \tag{2}$$

where $x_E(n)$ and $x_L(n)$ are the uncorrelated early and late reflections. The early reflections are composed of the direct signal and reflections in earlier time while the latter renders itself as coloration due to multiple reflections. In this paper, we consider both speech $s(n)$ and noise $d(n)$ are reverberant in nature. Assuming we can access the room impulse

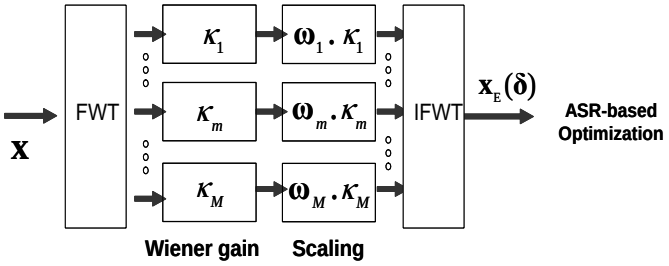


Fig. 1. Speech enhancement using Wiener filtering (WF)

response (RIR) $h(n) = [h_E(n)h_L(n)]$ and effectively identify its early and late components $h_E(n), h_L(n)$ [8][9] respectively, we can further rewrite Eq. (2) as,

$$x(n) = (s(n) + d(n)) * h_E(n) + (s(n) + d(n)) * h_L(n) \quad (3)$$

The power spectrum of the reverberant model in Eq. (2) can be estimated as:

$$|X(f)|^2 \approx |X_E(f)|^2 + |X_L(f)|^2 \quad (4)$$

where $X_E(f)$ and $X_L(f)$ are the magnitude spectra of the early and late reflections. By convention, we denote both $X_E(f)$ and $X_L(f)$ to contain both filtered speech and noise. Also, when referring to reverberant data $x(n)$, we assume a reverberant speech and reverberant noise as depicted in Eq. (3). In dealing with reverberation (both reverberant speech and noise), we are interested only in suppressing the effects of the late reflection since the early reflection is sensitive to microphone-speaker location. Moreover, the effect of early reflection is mostly mitigated with cepstral mean normalization (CMN) [8][9].

2.1. Wiener Filtering

The proposed Wiener filtering in the wavelet domain is a form of compression of the wavelet coefficients. By way of compression, the thresholding of the wavelet coefficients is avoided. The wavelet-based Wiener filtering [2] which is used in suppressing additive noise requires the calculation of Wiener gains given as,

$$\kappa_m = \frac{S(a)_m^2}{S(a)_m^2 + D(a)_m^2}, \quad (5)$$

where $S(a)_m^2$ and $D(a)_m^2$ are the speech and noise power respectively, calculated from the wavelet coefficients at scale a. Noise segments were detected using a voice activity detector (VAD). For the j^{th} contaminated wavelet coefficient in band m w_{mj}, the denoised wavelet coefficient is given as,

$$\tilde{w}_{mj}(denoised) = w_{mj}.\kappa_m, \quad (6)$$

The Wiener weighting κ_m dictates the degree of suppression of the contaminant to the observed signal. The enhanced wavelet coefficients are used to reconstruct the speech signal by inverse fast wavelet transform (IFWT).

This work of [2] is originally designed to suppress additive noise only. We expand it to deal with reverberant channel by suppressing the late reflections. Thus, the Wiener gain given in Eq. (5) is modified to,

$$\kappa_m = \frac{X_E(a)_m^2}{X_E(a)_m^2 + \delta_m X_L(a)_m^2}, \quad (7)$$

where $X_E(a)_m^2$ and $X_L(a)_m^2$ are the early and late reflection power respectively, calculated from the wavelet coefficients at scale a. Although $X_E(a)$ has relatively high power values than $X_L(a)$, the VAD method to select the correct segments may not be sufficient. Thus, a scaling parameter δ_m is introduced to minimize the error in calculating $X_E(a)_m^2$ and $X_L(a)_m^2$. We note that we can synthetically generate data using the clean speech and noise database together with the RIR [8][9]. Thus, we can calculate $\boldsymbol{\delta} = [\delta_1 ..., \delta_m, ..., \delta_M]$ that minimize the error between $\{X_E(a)_m^2, X_L(a)_m^2\}$ with the VAD and $\{X_E(a)_m^2, X_L(a)_m^2\}$ for the synthetically generated data. This process is similar to that in [8][9]. By applying the Wiener gains to the reverberant wavelet coefficients w_{mj} (analogous to Eq. 6), the enhanced wavelet coefficients are given as,

$$\tilde{w}_{mj}(enhanced) = w_{mj}.k_m. \quad (8)$$

The enhanced wavelet coefficients are converted back to the time domain through IFWT and we denote this as $x_E(\boldsymbol{\delta})$ to signify that only the early reflections are retained using $\boldsymbol{\delta}$. Fig. 1 illustrates the implementation of the modified WF where the reverberant and noisy speech signal is processed using a Fast Wavelet Transform (FWT). M subbands are created through FWT decomposition [11]. In our application we use five subbands to reflect that of [8][9]. Each of these subbands outputs a wavelet coefficient as a result of the fast wavelet structure. Then, the Wiener gains are calculated and the contaminated data is scaled by the Wiener gains. The early reflections (enhanced data) are then recovered through IFWT. Optimization

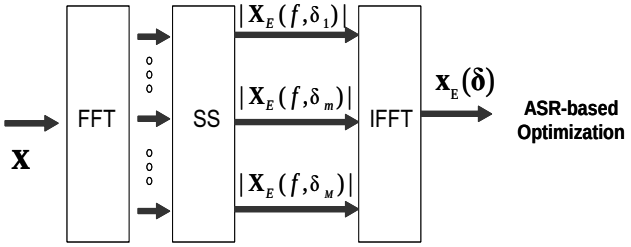


Fig. 2. Speech enhancement using spectral subtraction (SS)

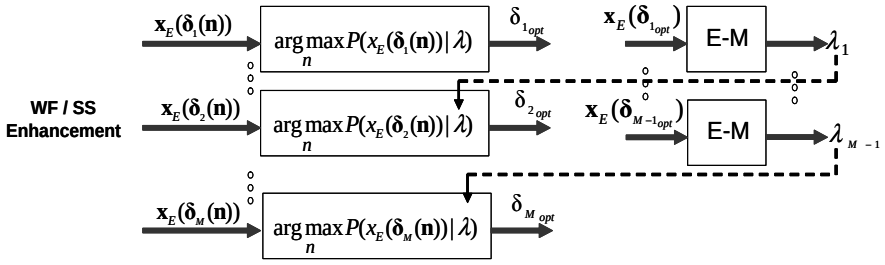


Fig. 3. ASR-based optimization of the scaling parameters.

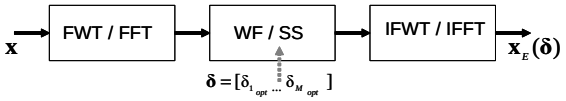


Fig. 4. Overall block diagram of the speech enhancement utilizing ASR-optimized scaling parameters.

of the scaling parameters based on ASR follows, which will be discussed in Section 3.

2.2. Spectral Subtraction

We will show the expansion of the conventional SS to address reverberation problems. As previously mentioned, we are interested in recovering only the early reflection and suppressing the late reflection. This can be done with multi-band SS [8][9]. Thus, the mth band power spectra of $X_E(f)$ is achieved through,

$$|X_E(f,\tau)| = \begin{cases} |X(f,\tau)|^2 - \delta_m |X_L(f,\tau)|^2 \\ \quad \text{if } |X(f,\tau)|^2 - \delta_m |X_L(f,\tau)|^2 > 0 \\ \\ \beta |X_L(f,\tau)|^2 \quad otherwise \end{cases} \tag{9}$$

where β the flooring coefficient, $|X(f,\tau)|^2$ and $|X_L(f,\tau)|^2$ are the power spectra of the reverberant signal and power of the late reflection respectively, with a window period of τ. δ_m denotes the mth band scaling parameter. The multi-band scaling factors $\boldsymbol{\delta} = [\delta_1 ..., \delta_m, ..., \delta_M]$ are derived through an offline training which minimizes the error of the estimate $|X_L(f,\tau)|$ under the MMSE criterion. The values of $\boldsymbol{\delta}$ coefficients (through offline training), and the effective identification of the late components of the impulse response $h_L(n)$ are discussed in [8] [9]. Fig. 2 shows the block diagram of the SS implementation. First, the early reflection X_E are recovered as discussed in Eq. (9) and reverted back to $x_E(\boldsymbol{\delta})$ by IFFT.

Table 1. System specification used in evaluating the system

Sampling frequency	16 kHz
Frame length	25 ms
Frame period	10 ms
Pre-emphasis	$1 - 0.97z^{-1}$
Feature vectors	12-order MFCC, 12-order ΔMFCCs 1-order ΔE
HMM	8000 Gaussian pdfs

3. OPTIMIZATION BASED ON ACOUSTIC LIKELIHOOD

In Section 2, the multi-band scaling parameters $\boldsymbol{\delta}$ are all set to initial MMSE-based values and in effect serve as a global weighting. In this section, we will discuss the optimization of $\boldsymbol{\delta}$, fine-tuning both WF and SS to be directly linked with ASR.

In Fig. 3, we show the ASR-based optimization of $\boldsymbol{\delta}$ where the scaling parameters in each band is sequentially optimized from band m=1 to m=M. The band coefficient to be optimized is allowed to change within a close neighborhood $n\triangle$ from its initial $MMSE$ value, where $n = \pm 1 ... N$ and $\triangle = 0.02$. The reverberant data $\boldsymbol{x}$ is enhanced using either multi-band WF/SS. Initially, we fix the rest of the scaling parameters to MMSE-based estimates except for the band to be optimized. Thus, for optimizing band $m = 1$, we generate $\boldsymbol{\delta}_1(\boldsymbol{n}) = [\, \delta_{1\,MMSE} + \boldsymbol{n}\,\triangle,\, \delta_{2\,MMSE},\, \delta_{m\,MMSE}, ...,\, \delta_{M\,MMSE}]$, and execute WF/SS using the generated coefficients. The resulting enhanced data $x_E(\boldsymbol{\delta}_1(\boldsymbol{n}))$ are evaluated using the HMM-based acoustic model which is trained with data processed with MMSE-based WF/SS parameters, denoted as $\lambda = \lambda_{MMSE}$. A likelihood score is computed for each of the data processed with different WF/SS conditions. Based on this result, $\delta(1)_{opt}$ that has the corresponding highest likelihood score is selected. Right after $\delta(1)_{opt}$ is found, the acoustic model is updated with data processed by WF/SS using $\delta(1)_{opt}$. The newly updated model λ_1 is then used in calculating the likelihood score for the next band

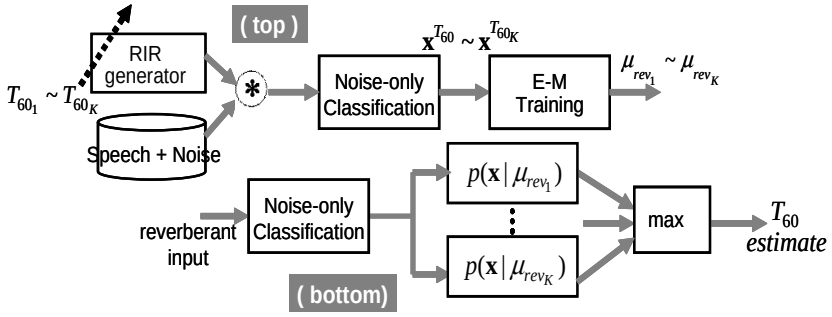


Fig. 5. Robust RIR Estimation.

and the process is repeated until the complete set of parameters δ_{1opt},...,δ_{Mopt} are optimized. After the optimization, the reverberant data are processed with the proposed ASR-optimized WF/SS as shown in Fig. 4.

4. ROBUST RIR ESTIMATION

Since we need the RIR, we implement an automatic estimation of the RIR as opposed to physically measure it [8][9]. We have shown that due to the low resolution characterization of HMM to the speech signal compared to the RIR, rough estimate of the RIR is sufficient in HMM applications. The RIR can be modeled as having a decaying exponential energy,

$$h^2(n) \approx e^{(6\ \ln(10)/T_{60})\ l}, \tag{10}$$

where l is the discrete time sample, and T_{60} is the reverberation time. To effectively identify T_{60} in the presence of both convolutive speech and noise, we designed a GMM-based T_{60} classifier as shown in Fig. 5 (top). Reverberant speech and noise are synthetically generated $x^{T_{60_k}}$ with variable T_{60_k} to train GMMs μ_{rev_k}. To attain robustness, we employed the following; first, reverberant noise-only frames (occur in block segments during silence part of the clean speech) are used to train the GMM. This avoids the variability caused by the convolutive speech. From these reverberant noise-only block segments, we select only the frames that have low power to capture only the late reflection of the reverberant noise signal. We note that the late reflection renders itself as coloration in frequency due to multiple overlapping. This results in less sensitivity to noise types and SNR since noise information is smeared by the coloration effect. Finally, we use a larger mixture for the GMM (i.e. 256 mix). The use of a large number of mixture components makes the GMM sensitive to the higher resolution RIR. Fig. 5 (bottom) shows the actual identification of T_{60}. The reverberant speech and noise input is processed to classify noise-only frames. Then, likelihood is calculated given all of the GMMs with different T_{60_k}. The corresponding T_{60} that results in the highest likelihood score is selected and from this, the RIR is estimated using Eq. (10).

5. EXPERIMENTAL EVALUATION

5.1. Training and Testing Data

The training database is from the Japanese Newspaper Article Sentence (JNAS) corpus. The open test set is composed of 200 utterances. Recognition experiments are carried out on the Japanese dictation task with 20K vocabulary. The language model is a standard word trigram model. The acoustic model is a triphone HMMs of 8000 Gaussian pdfs. A summary of the system specification is shown in Table 1.

We experimented using T_{60}=200 msec reverberation time. Reverberant training data are synthetically produced with the automatically generated RIR discussed in Section 4. The test data were recorded in a room with known reverberation time : T_{60}=200 msec. Thus, we used actual reverberant data for evaluation. Three types of noise are considered; office, vacuum cleaner, and white Gaussian noise. The signal-to-noise ratio (SNR) are 15 dB, 20 dB and 25 dB. The microphone-to-speaker distance is approximately 1.5 m. The noise source is also placed 1.5 m from the microphone with a 30 degrees angle relative to the microphone-to-speaker distance. In the experiments we use a total number of bands $M = 5$ which is consistent that of the former work [8][9].

5.2. Acoustic Model Training

We have shown the incremental optimization of the multi-band scale parameters in Section 3. This process selects the optimal scale factors $\boldsymbol{\delta}_{opt} = [\delta(1)_{opt}, ...\delta(m)_{opt}, ..., \delta(M)_{opt}]$. The acoustic model training is carried out as,

$$\lambda_{opt} = \arg\max_{\lambda_M} \prod_{r=1}^{R} P(\mathbf{x}_r^{\boldsymbol{\delta}_{opt}}|\mathbf{w}; \lambda_M),$$

where λ_{opt} is the desired acoustic model to be trained and later used by the ASR. λ_M is the Mth updated model which is the last model update in a series of model re-estimation as part of the optimization process discussed in Section 3. $\mathbf{x}_r^{\boldsymbol{\delta}_{opt}}$ is the enhanced utterance processed by WF/SS using the opti-

Table 2. Recognition Results in Word Accuracy

	office noise			vacuum cleaner noise			white gaussian noise		
Methods	15dB	20dB	25dB	15dB	20dB	25dB	15dB	20dB	25dB
$Testing$: Unprocessed $Training$: clean	23.4%	34.6%	40.3%	19.3%	32.2%	37.5%	25.6%	38.7%	42.0%
$Testing$: Unprocessed $Training$: Unprocessed	37.1%	43.5%	48.6%	35.4%	38.7%	42.6%	39.4%	45.1%	50.3%
$Testing$: SS $Training$: SS	51.8%	58.6%	63.2%	49.1%	57.3%	60.1%	52.8%	59.9%	64.7%
$Testing$: ASR-optimized SS $Training$: ASR-optimized SS	**61.4%**	**72.1%**	**75.9%**	**58.3%**	**70.1%**	**73.6%**	**63.4%**	**73.2%**	**77.1%**
$Testing$: WF $Training$: WF	52.3%	57.4%	61.8%	50.6%	56.4%	58.2%	53.6%	58.7%	62.9%
$Testing$: ASR-optimized WF $Training$: ASR-optimized WF	**62.5%**	**71.4%**	**74.1%**	**59.4%**	**68.3**%	**70.3%**	**64.7%**	**72.8%**	**76.5%**

mal scale parameters while $\mathbf{w}$ refers to its transcription. The training database has a total $r = R$ training utterances.

5.3. Recognition Performance

In Table 1, we show the recognition performance of the different methods. It is observed that enhancing the reverberant data using WF and SS is better than not processing the reverberant data at all. However, when WF and SS are optimized in relation with the ASR, further improvement in recognition performance is achieved. This is attributed to the fact that the ASR-optimized variants are capable of improving the model likelihood used by the ASR. The superior performance of the proposed method is consistent to all of the different SNRs and noise types in our experiment. We note that we test using real recording noisy and reverberant data.

6. CONCLUSION

We have extended two popular denoising techniques (WF and SS) to address reverberant speech and noise, and optimize each of these to be effectively used in ASR applications. Moreover, we have shown the process of embedding optimized enhancement in the acoustic model training. Improvement in performance is achieved as the enhancement procedure is closely linked to the improvement of the acoustic model likelihood. We have shown that the this concept works in both frequency and wavelet domain. In general, the optimization in relation to ASR is applicable to any speech enhancement algorithms and in any domains.

7. REFERENCES

[1] S. Vaseghi "Advanced Signal processing and Digital Noise reduction", Wiley and Teubner, 1996.

[2] E. Ambikairajah et. al., "Wavelet Transform-based Speech Enhancement" *In Proceedings International Conference on Speech and Language Processing ICSLP*, 1998

[3] S.F. Boll, "Suppression of acoustic noise in speech using spectral subtraction" *In Proceedings IEEE Int. Conf. Acoust., Speech, Signal Proc. ICASSP* pp 208-211, Apr. 1979.

[4] S. Kamath and P. Loizou,"A Multi-Band Spectral Subtraction Method for Enhancing Speech Corrupted by Colored Noise" *In Proceedings IEEE Int. Conf. Acoust., Speech, Signal Proc. ICASSP* 2002.

[5] H. Hermansky et.al., "Data-driven Nonlinear Mapping for Feature Extraction in HMM" *ASRU Workshop*, 1999.

[6] T. Hwang et. al., "Feature Adaptation Using Deviation Vector for Robust Speech Recognition in Noisy Environment" *In Proceedings IEEE Int. Conf. Acoust., Speech, Signal Proc. ICASSP*, 1997.

[7] A. Torre et.al., "Non-linear Transformation of the Feature Space for Robust Speech Recognition" *In Proceedings IEEE Int. Conf. Acoust., Speech, Signal Proc. ICASSP*, 2002.

[8] R. Gomez et.al. , "Distant-talking Robust Speech Recognition Using Late Reflection Components of Room Impulse Response" *In Proceedings IEEE Int. Conf. Acoust., Speech, Signal Proc. ICASSP*, 2008

[9] R. Gomez et.al., "Fast Dereverberation for Hands-Free Speech Recognition" *IEEE Workshop HSCMA*, 2008

[10] R. Gomez, T. Kawahara, "Optimization of Dereverberation Parameters based on Likelihood of Speech Recognizer" *Interspeech*, 2009.

[11] G. Strang, and T. Nguyen, "Wavelets and Filter Banks" *Wellesley-Cambridge Press*, 1996.

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

人工知能学会研究資料
JSAI Technical Report
SIG-Challenge-A902-10

音情報を用いたロボットハンドによるタスク達成判別および水量推定

Applications of the acoustic information
in the task achievement and the water volume estimation by a robot hand

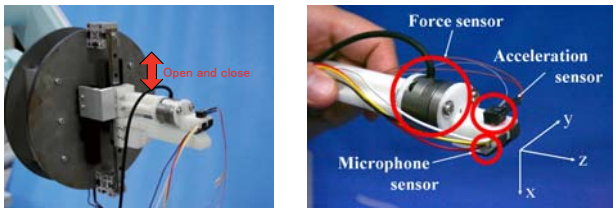
栗田 雄一，池田 篤俊，祖父江 厚志，小笠原 司
Yuichi KURITA, Atsutoshi IKEDA, Atsushi SOBUE, and Tsukasa OGASAWARA
奈良先端科学技術大学院大学 情報科学研究科
Graduate School of Information Science, Nara Institute of Science and Technology (NAIST)
kurita@is.naist.jp

Abstract

Sounds during a work include information of task achievement and material condition. In this paper, we developed a robotic gripper that equips a sound sensor (microphone). We used this robot gripper to judge the task achievement conditions and estimate the water volume in various cups. For applications in the task achievement judgment, the sound information produced during the cable installing task is measured and the usefulness of the sound information is confirmed. For applications in water volume estimation, the glass harp acoustics is used and the experimental results show the system can estimate the water volume with high accuracy.

1 はじめに

人は視覚・力覚・触覚などさまざまな感覚器官から得た情報を有効活用して作業を行っている．そこでこれらの情報を取得するセンサを使い，作業実行や環境認識に用いるロボットやシステムも数多い [1, 2, 3]．一方，音情報の利用に関しては，ディーゼル機関の運転状態を排気音により監視する手法 [4]，音情報を利用して飲み口部分が欠けたビンと正常のビンを区別する手法 [5]，生産ラインの工作機械の監視を音情報を利用して行う手法 [6] などが提案されているものの，これらは監視を主な目的としており，対象が発生する音情報を積極的に利用して作業遂行を行うロボットはほとんどない．人の聴覚能力は力覚や触覚に比べて分解能や情報処理能力が高く，また特に注意を向けなくても情報が得られるため，日常生活や生産現場においても人は極めて自然に音を利用している．将来的にロボットが人と同じ環境で動いたり，人が扱うものと同じ製品を扱ったりすることを考えた場合，音はもっと積極的かつ有効に使われてしかるべき情報である．



(a) Developed gripper (b) Attached sensors

Figure 1: Microphone embedded robot gripper

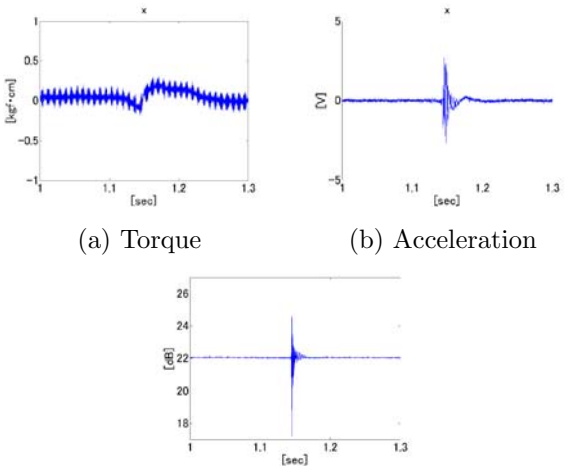
音情報を利用するにあたっては，対象物や環境から音が発生することが前提になる．ここで作業遂行にあたって発生する音は大きく (1) 作業進行にあたって自然に発生する音，(2) アクティブに働きかけて発生させる音，の 2 種類がある．(1) はケーブルの差し込み時やプラモデルの組み立て時に生じる「カチッ」という音が該当し，(2) は打鍵検査のように中身の状態を知るために対象物を叩いたときの音が該当する．本稿ではこれら 2 種類の音情報についてそれぞれの効果的な利用方法を検討することを目的とし，音計測と周波数解析による作業の正誤判定実験およびグラス内の水量推定実験を行った結果について報告する．

2 音情報を用いたタスク達成判別 [7]

作業中に発生する音は作業状態や結果に関する情報を含んでおり，作業状態管理や達成状況の把握に有用である．そこで本章ではグリッパの指先に音センサ（マイクロフォン）を搭載し，作業で発せられる音情報を用いてタスク達成判別が可能かを調べた．

2.1 マイクロフォンを内蔵したロボットグリッパ

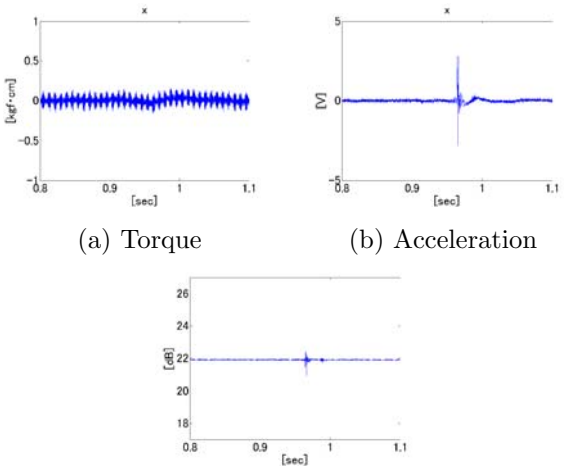
Fig. 1 に本研究で使用したロボットグリッパを示す．このグリッパは 2 本の指が開閉する平行グリッパであり，片



(a) Torque (b) Acceleration

(c) Sound

Figure 2: Measured data from the normal cable



(a) Torque (b) Acceleration

(c) Sound

Figure 3: Measured data from the broken cable

方の指に 6 軸力覚センサ（ビーエルオートテック社 MICRO5/50），指先の外側の部分に小型の加速度センサ（Endevco 社 Model 23）が内蔵されている．また小型アンプ内蔵シリコンマイクロフォン（Knowles 社 SP0103NC3-3）を指先に埋め込む．マイクロフォンの仕様を Table 1 に示す．

2.2 音を用いた作業の正誤判定

2.2.1 実験条件

ここでは LAN ケーブルの挿し込み作業が成功したか否かを，音情報を用いて判定することを試みる．一般的な市販 LAN ケーブルは，そのコネクタ部にツメがついており，正常に挿さると「カチッ」という音が鳴るように作られている．そこで音センサを用いてこの正常挿入時の音がなったかを判定することで，作業の正誤判定を行う．

実験では 2 種類のケーブルを用いた．1 つはコネクタ部のツメが正常であり，挿した後に機器に固定がされる（正常ケーブル）．もう 1 つはツメが折れており，挿した後も固定がされず，引っ張るとすぐに抜けてしまう（異常ケーブル）．このケーブル挿し込み作業を，Fig. 1 に示したグリッパを用いて行い，このときの力センサ，加速度センサ，音センサ情報をそれぞれ測定し，正誤判定が可能かを調べた．

Table 1: Specification of the microphone sensor

Dimension	6.15 × 3.76 × 1.45 [mm]
Sensitivity	-22±4 [dB]
Frequency range	100〜10,000 [Hz]

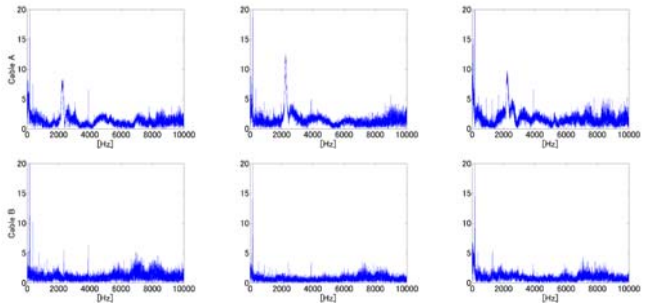
2.2.2 実験結果

正常ケーブルと異常ケーブルを使って挿し込み作業を行ったときの x 軸のトルク，加速度および音データをそれぞれ Fig. 2 と Fig. 3 に示す．これらの図から，ケーブルが挿し込まれたとき，それぞれのセンサに何らかの反応が観測されることが分かる．そこでこれらセンサデータを用いて正常ケーブルと異常ケーブルの違いを判別することを試みたが，力センサの情報は試行ごとにばらつきが大きく，また加速度センサの情報は周波数解析からも両者を精度良く判別することとができなかった．

一方，音データについて周波数解析を行った結果を Fig. 4 に示す．今回の実験設定におけるケーブル挿し込み作業においては，正常ケーブルについては 2250[Hz] 付近にピークを持つスペクトルが得られた一方，異常ケーブルについてはこのような特徴は見られなかった．このことは，周波数のピーク値付近のスペクトル形状を用いることで，正常・異常ケーブルの判別ができる可能性を示唆する．

そこでこのスペクトル形状の特徴を用いることにより正常・異常ケーブルの判別を行えるかを調べた．ケーブル挿し込み作業の達成判別のフローチャートを Fig. 5 に示す．まずケーブル挿し込み時の音情報を測定し，それに対してフーリエ変換をかける．得られたフーリエ変換データをあらかじめ取得してあった正常挿入音のフーリエ変換データと比較し，ピーク周波数± 300[Hz] の範囲で相関値を計算する．相関値が一定値以上であれば成功，それ未満であれば失敗と判断する．

判定の閾値を 0.8 として，Fig. 6 に示すようにマニピュレータを使った差し込み作業実験を行ったところ，正常ケーブルについては 10 回中 8 回において作業が成功したと判定し，異常ケーブルについては 10 回中 10 回を作業が失敗したと判定した．このことから本手法により 90 %の精度で作業達成判別が可能であることが分かった．な



1st trial　2nd trial　3rd trial

Figure 4: FFT of the sound data (top: normal cable, bottom: broken cable)

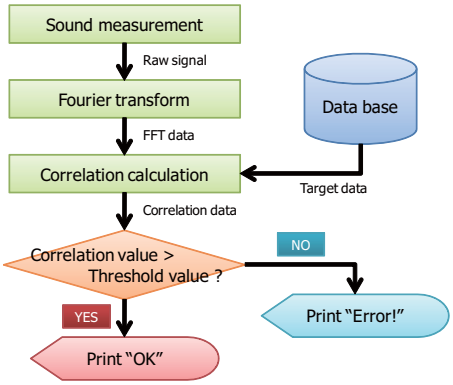


Figure 5: Flowchart of the task achievement judgment

お正常ケーブルについて挿入が成功したにもかかわらず失敗と判定した理由は，スムーズな挿し込み作業が行えず別の音が発生していたり，爪部分の剛性が弱く差し込み時の音が小さかったことが原因であった．

3　音情報を用いたグラス内の水量推定 [8]

本章では対象にアクティブに働きかけて音を発生させることで，対象物の状態推定が可能であるかを検証することを目的として，グラス内の水量推定実験を行う．

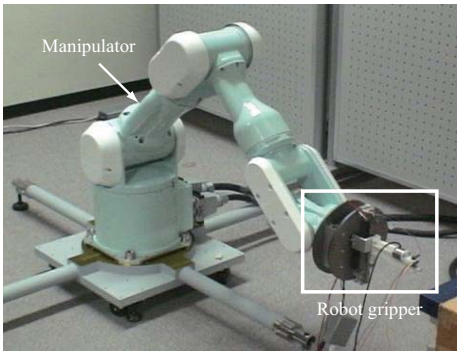
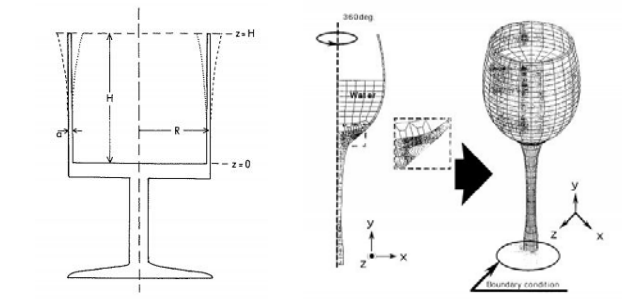


Figure 6: Cable insertion experiment



(a) French's moel [9]　(b) Oku's model [10]

Figure 7: Glass models

3.1　グラスハープの音響特性に関する従来研究

物理学の分野ではグラスハープの音響特性を解析的に明らかにしようとする研究が行われており，例えば French は Fig. 7(a) に示すグラスの基本振動モードを力学的性質から解析的に調べ，最終的に次の関係が成り立つことを示している [9]：

$$\left(\frac{f_0}{f_h}\right)^2 \approx 1 + \frac{\alpha}{5}\frac{\rho_l R}{\rho_g a}\left(\frac{h}{H}\right)^4 \tag{1}$$

ここで H はグラスの高さ，h は水で満たされている高さ，f_0, f_h はそれぞれ空のときおよび高さ h まで水で満たされたグラスの振動周波数，ρ_l, ρ_g はそれぞれ水およびグラスの密度，a, R はそれぞれグラスの厚みと半径，α は定数である．このような解析モデルに基づく音響特性の導出結果は元のモデルと同一の形状・特性をもつグラスに対してはよく一致するものの，異なる形状・特性を持つグラスに対しては必ずしも一致せず，また複雑な形状のグラスに対する一般的な解析モデルを導出することは難しい．そこで Oku らは有限要素解析 (FEM) により実験的によく一致する音響特性の導出法を提案している [10]．Oku らは Fig. 7(b) に示す形状のグラスの音響特性について，次式の関係を得たとした：

$$\frac{f_h}{f_0} = 1 - 0.5\left(\frac{V_h}{V_H}\right)^3 \tag{2}$$

ここで V_h, V_H はそれぞれ高さ h まで満たされた水量と一杯まで満たされた水量である．

そこで本稿ではこれら従来研究の結果を用いて，次章に示すマイクロフォンを内蔵したグリッパと 5 種類のグラスを使って周波数の計測からグラス内水量の推定が可能であるかを調べる．

3.2　音を用いた水量推定

3.2.1　実験条件

Fig. 8 に実験に利用するグラスを示す．本稿では，ガラス製（glass1, glass2），陶器製（ceramic1, ceramic2），ステンレス製（stainless）の 5 種類を使用した．これらのグ

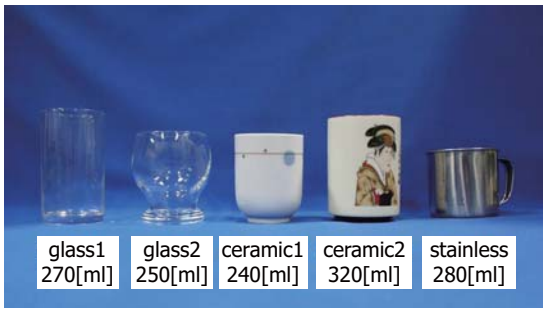
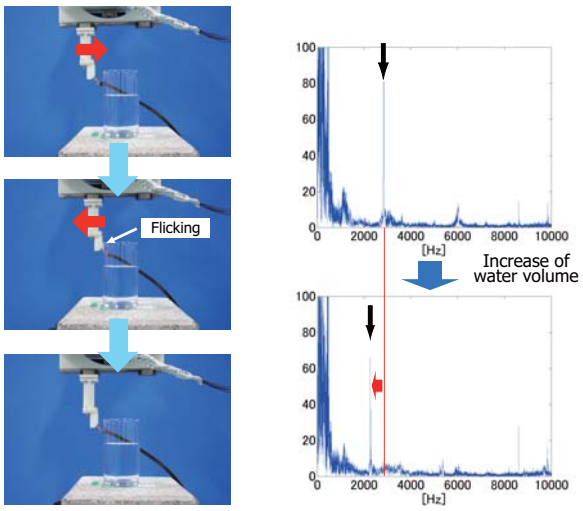


Figure 8: Glasses used in the experiment and their full capacity



(a) Flicking motion (b) Change of peak frequenc

Figure 9: Flicking by the robot gripper

ラスの縁に Fig. 9 に示すようにロボットグリッパの先端をはじくようにあてることで振動を発生させ，そのときの音をマイクロフォンにより計測する．計測した尾情報にフーリエ変換をかけた上で，本稿では振動の 1 次モードに相当するピーク値をグラスの振動周波数として，まず空の状態における振動周波数 f_0 を手動で設定する．次に水を 20 [mℓ] ずつグラスにそそぎ，同様にロボットグリッパによるはじき動作を行うことで振動周波数の変化を計測する．ここで Oku らの式 Eq. 2 によれば，水を入れた状態のグラスの振動周波数 f_h は，空の状態の振動周波数 f_0 の半分以下にはならない．そこで f_h は $[f_0/2 \sim f_0]$ の範囲における最も大きな周波数のピーク値として自動取得する．

3.2.2 実験結果

前節で説明した方法により各グラスの音響特性変化を計測した．各グラスの水量と周波数ピーク値との関係を Fig. 10 に示す．グラスによって基本振動周波数が異なること，そして水量が増えるにつれて周波数のピークが徐々に下がっている様子が確認できる．

こうして得られたグラスの音響特性が，従来の解析的・

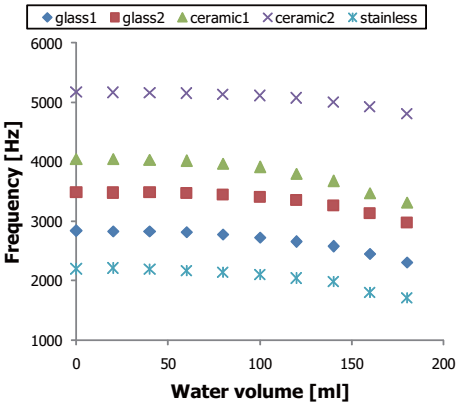
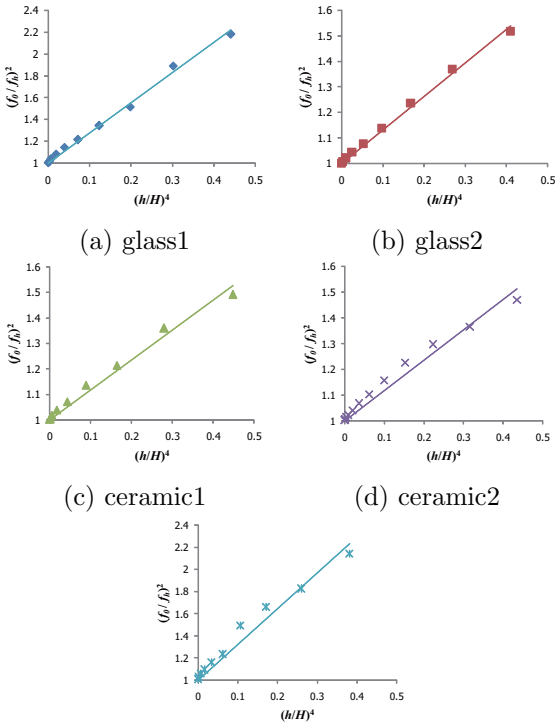


Figure 10: Relationship between the water volume and the peak frequency



(a) glass1 (b) glass2

(c) ceramic1 (d) ceramic2

(e) stainless

Figure 11: Results of French's estimation

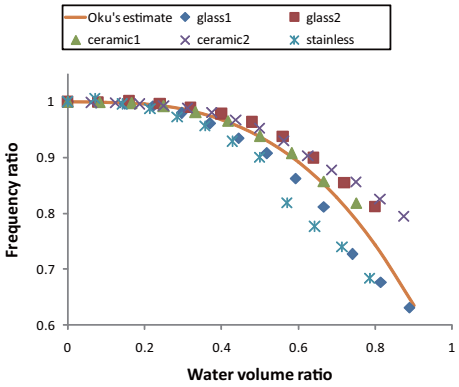


Figure 12: Relationship between the volume ratio and the frequency ratio based on Oku's model

Table 2: RMSE[%] between the measured data and the fit curve based on the proposed method

	Estimation by all points			Estimation by 3 points		
	a	b	RSME	a	b	RMSE
glass1	0.52	2.49	3.70	0.50	2.54	5.88
glass2	0.38	3.06	1.04	0.38	3.15	1.32
ceramic1	0.41	2.75	1.00	0.41	2.84	1.40
ceramic2	0.29	2.47	1.47	0.29	2.61	2.13
stainless	0.57	2.31	3.94	0.58	2.50	5.59

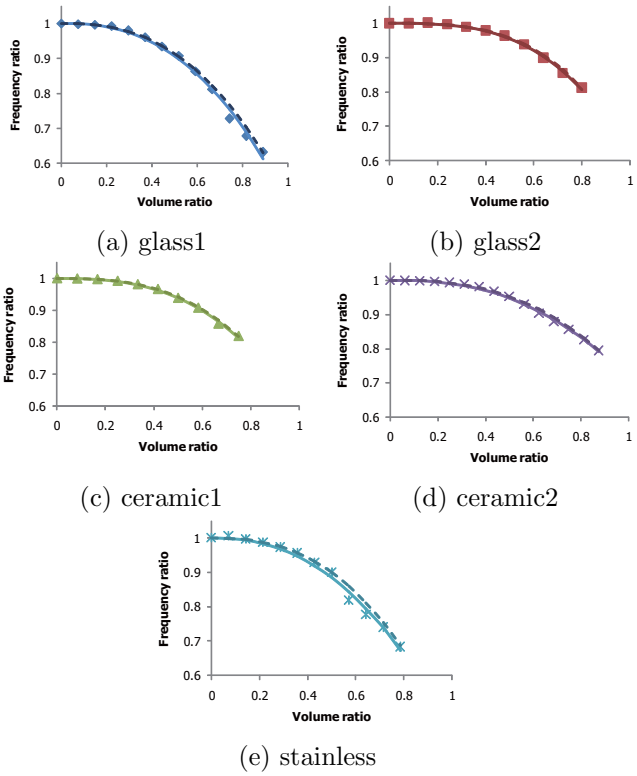
実験的な音響特性関係式と一致するかを確認する．Fig. 11 は各グラスの $(\frac{f_0}{f_h})^2$ と $(\frac{h}{H})^4$ との関係を示した図であり，また図中の直線は最小 2 乗法で計算した縦軸との切片が 1 の近似直線である．French の解析式が当てはまるのであれば，これらの関係は近似直線上にのるはずである．図を見ると，ガラス製のグラス（glass1, glass2）は比較的よく一致するものの，それ以外のグラスでは French の式の適用は困難であることが予想される．

次に Fig. 13 は横軸に水量変化率 $\frac{V_h}{V_H}$，縦軸に周波数変化率 $\frac{f_h}{f_0}$ をとって両者の関係を示した図であり，図中の曲線は Oku の関係式 Eq. 2 を当てはめた結果を示している．水量が少ないうちは周波数変化が少ないが，半分を超えた辺りから急激に周波数変化が起こる性質が共通して確認できる．Oku らの式はこの性質をよく表現できている一方，グラスの形状や特性を補正する変数が式中に存在せず，個々のグラスの違いをそのままでは表現できない．

そこで本稿では，Oku らの式を拡張し，変数 a, b を使って以下の関係式をたて，これら変数を各グラスがもつパラメータとして最小 2 乗法により推定することを提案する．

$$\frac{f_h}{f_0} = 1 - a\left(\frac{V_h}{V_H}\right)^b \tag{3}$$

この関係式に基づき，それぞれのグラスのパラメータを推定した結果を適用した図を Fig. 13 に，推定結果および RMSE 誤差を Table 2 に示す．ここで図中の実線は計測したすべての点を利用してパラメータ推定した結果を，破線は水量がゼロ，半分，いっぱいの状態の 3 点のみを利用してパラメータ推定した結果を意味する．このように提案手法によりグラスごとの音響特性をよく表現できていることが確認できる．またパラメータ推定においては 3 点のみの情報だけを使っても精度に大きな違いがないことも確認できる．これは，あらかじめ水量がゼロ，半分，いっぱいの状態の 3 点の振動周波数さえわかっていれば，音情報計測から水量を精度よく推定できることを意味する．



(a) glass1 (b) glass2

(c) ceramic1 (d) ceramic2

(e) stainless

Figure 13: Results of the proposed estimation method

3.3 ロボットフィンガのはじき動作による水量推定

3.3.1 マイクロフォン内蔵ロボットフィンガ

ここでは Fig. 14 に示す多指ロボット [11] の指にマイクロフォンを装着し，グラスの縁をはじいた時の音を計測することで，他のセンサ情報を一切使わずにグラス内の水量を推定する実験を行う．

3.3.2 実験条件

Fig. 8 に示した 5 種類のグラスに対して，20 [mℓ] ずつ水量を増やしながらはじき動作を行って音情報計測およびフーリエ変換を行い，3.2.1 節と同様の手法でピーク周波

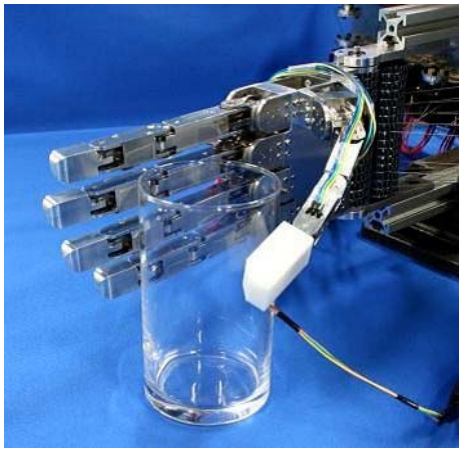
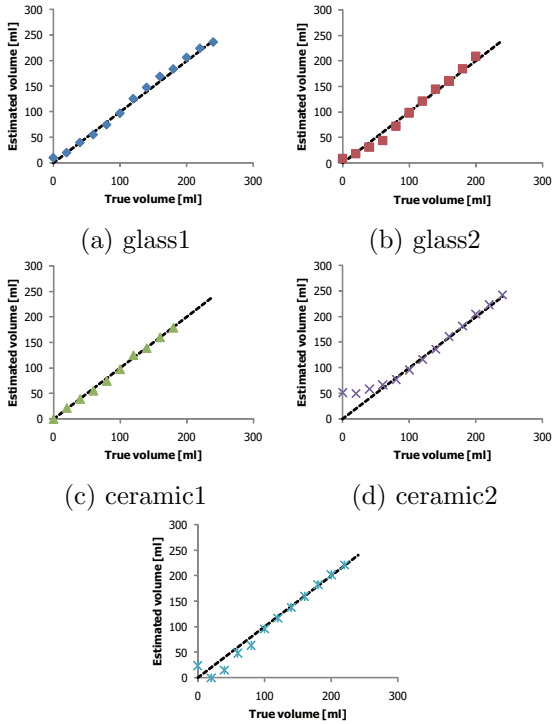


Figure 14: Anthropomorphic robot hand with a microphone sensor



(a) glass1 (b) glass2

(c) ceramic1 (d) ceramic2

(e) stainless

Figure 15: Experimental results of the water volume estimation

数の自動取得を行った．また得られたピーク周波数から，Table 2 に示したパラメータのうち，水量がゼロ，半分，いっぱいの 3 状態の周波数から推定したパラメータ値と，Eq. 3 とを利用して水量推定を行った．

3.3.3 実験結果

水量の真値と推定結果の関係を Fig. 15 に，真値との誤差の平均およびその最大容量との比を Table 3 に示す．これらの結果から，水量にして 2~10 [mℓ] 程度，割合にして 1~3 [%] 程度の誤差で水量推定が可能であることが確認できた．なお推定誤差は水量が少ないときに比較的大きくなりやすい．これは Fig. 13 に示したように水量が少ないときはグラスによらず周波数変化が少ないため，振動周波数の計測誤差が水量推定結果に大きく影響することが原因と考えられる．

Table 3: Estimation error by the robot hand

	Average [mℓ]	volume ratio [%]
glass1	4.85	1.80
glass2	5.74	2.30
ceramic1	2.15	0.90
ceramic2	10.15	3.17
stainless	9.32	3.33

4 おわりに

本稿では (1) 作業進行にあたって自然に発生する音，(2) アクティブに働きかけて発生させる音に着目し，これら 2 種類の音情報についてそれぞれの効果的な利用方法を検討した．実験の結果，作業時に発する音のスペクトル波形を用いることで，ケーブル差し込みの正常・異常を 90 %の精度で判別できることが分かった．またグラスハープの音響特性を用いることで，3 %程度の誤差でグラス内の水量推定が行えることが分かった．今後は，音のアクティブセンシング技術を用いることで，作業遂行をより柔軟かつロバストに行えるロボットシステムの開発に取り組んでいきたい．

参考文献

[1] J. Park and G. Kim, “Development of the 6-axis Force/Moment Sensor for an Intelligent Robot’s Gripper,” Sensor and Actuators A, vol.118, No.3, pp.127-134, 2005.

[2] A. Morales, P. J. Sanz, A. P. del Pobil, and A. H. Fagg, “Vision-based Three-finger Grasp Synthesis Contrained by Hand Geometry,” Robotics and Autonomous Systems, vol.54, No.6, pp.496-512, 2006.

[3] P. A. Schmidt, E. Mael, and R. P. Wurtz, “A Sensor for Dynamic Tactile Information with Applications in Human-robot Interaction and Object Exploration,” Robotics and Autonomous Systems, vol.54, No.12, pp.1005-1014, 2006.

[4] 水谷博，木村隆一，濱本宏：“音による舶用ディーゼル機関の燃料弁噴射圧の異常検出”，日本音響学会誌，Vol.43, No.8, pp553-563, 1987.

[5] 岡田賢，上條哲平，石川稜威男：“ニューラルネットワークを用いた打音の特徴抽出”，電子情報通信学会技術研究報告，Vol.102, No.398, pp.1-6, 2002.

[6] 大澤拓也，陳連怡，中村隆，藤本英雄：“切削異常検出のための加工音の解析”，日本機械学会東海支部総会講演会講演論文集，Vol.2003, No.52, pp.187-188, 2003.

[7] 祖父江厚志, 池田篤俊, 栗田雄一, 高松淳, 小笠原司，“音による作業タスク達成判別のためのマイクロフォングリッパ”, 第 26 回日本ロボット学会学術講演会, RSJ2008AC1K3-06, 2008.

[8] 栗田雄一，祖父江厚志，池田篤俊，小笠原司，“グラスハープの音響特性を利用したはじき動作による水量推定”, 第 27 回ロボット学会学術講演会, 1A3-04, 2009.

[9] A. P. French: “In Vino Veritas: A study of wineglass acoustics” American Journal of Physics, 51, 8, pp.688–694, 1982.

[10] K. Oku, A. Yarai, and T. Nakanishi: “A New Tuning Method for Glass Harp Based on a Vibration Analysis that Uses a Finite Element Method”, Journal of the Acoustical Society of Japan (E), 21, 2, pp.97-104, 2000.

[11] Y. Kurita, Y. Ono, A. Ikeda, and T. Ogasawara, “Human-sized Anthropomorphic Robot Hand with Detachable Mechanism at the Wrist,” 2009 IEEE/RSJ International Conference on Intelligent Robots and Systems, pp.2271-2276, 2009

〒162 東京都新宿区津久戸町 4-7 OS ビル 402 号室 03-5261-3401 Fax: 03-5261-3402

(本研究会についてのお問い合わせは下記にお願いします.)

SIG-AI-Challenges home page (WWW): `http://winnie.kuis.kyoto-u.ac.jp/SIG-Challenge/`