臨場感テレワークにおける音処理技術

矢頭　隆　　森戸　誠

普段、我々はオフィス空間にいて何気なく聞く音から、特に努力することもなく無意識に多くの情報を得ている。誰がどのような仕事をしているか、多忙か、在席か、はたまた健康状態は、人物だけではなく機器が動作しているか･･･、会話音に限らず非言語音から得られる情報も決して少なくない。しかし、一般にテレワークでの音通信は会話音が主目的で非言語情報の伝送までは考慮されていない。その会話音でさえもマイクで収録し通信路を介して遠隔地で聞く場合、種々の変形要因によって明瞭性が大きく劣化する。結果として多くの有用な情報量が失われている。テレワークを真に効果的にするためには、雰囲気や様子が自然と伝わる臨場感豊かな音環境の実現が重要である。

本稿では、高品質、高臨場感実現のための音処理技術を概観した上で、要素技術の一つである音源分離技術について紹介する。

高品質、高臨場感実現のための音処理技術

(1) 音臨場感生成技術

遠隔2点間で音によるコミュニケーションを行なうことを考える（**図1**）。特定の場所に設置したマイクから収録した音を相手方のスピーカ等で再生する。限られたマイク数やマイクと音源との位置関係などに影響され、方向感や距離感、個々の音の音量バランスが損なわれる。高臨場感の実現には、遠隔地であっても互いの空間が接しているような音響空間の生成、すなわち空間的な方向感や距離感をも含めて現場の状態を再現できる立体音響の技術が必要である。

バイノーラル再生（**図2（a）**）は、原音場でバイノーラル録音*1) された音をヘッドホンで受聴する。収録音には音源から聴取者の両耳に到達するまでの音響的な影響（頭部による音の反射や回折など）も含めて録音されているため、聴取者はあたかも原音場で音を聞いているような臨場感が得られるとされる。システムが簡易、再生音

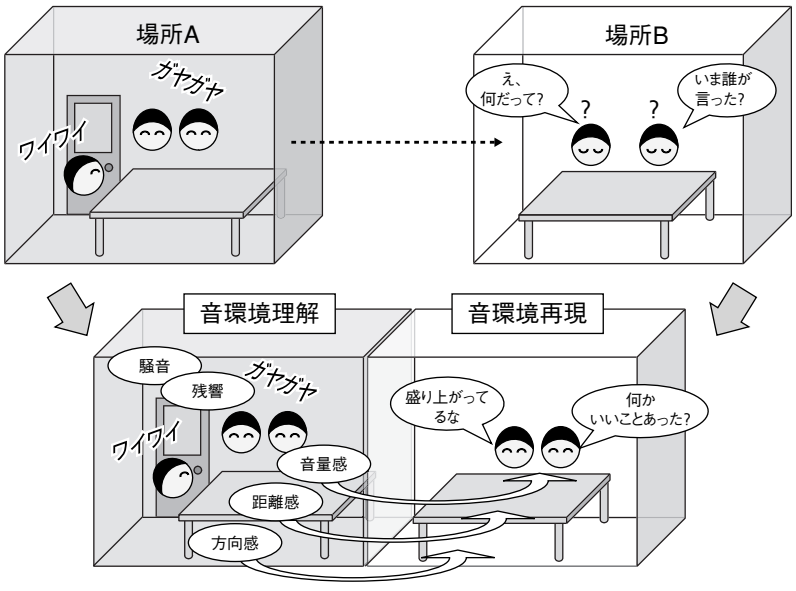


図1　テレワークにおける音臨場感の生成

*1)バイノーラル録音：人間の左右の耳に入る音そのものを収録するため人形（ダミーヘッド）の耳に取り付けられた2つのマイクによって録音する方法。

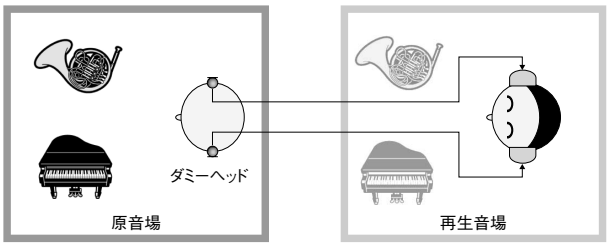


図2　（a）バイノーラル再生

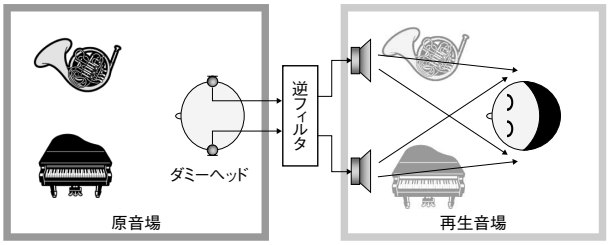


図2　（b）トランスオーラルシステム

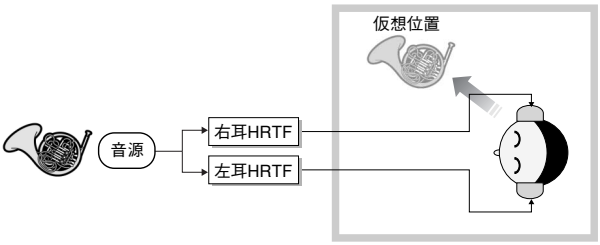


図3　HRTFを用いた音像定位

場の部屋環境の影響を受けないなどの利点がある反面、音像が頭内に定位することが多い、ヘッドホンを装着しなければならず通常のコミュニケーションを阻害するなどの問題・制約がある。一方、ヘッドホンの代わりに複数のスピーカを用いて受聴点耳元の音圧を制御するシステムが提案されている。バイノーラルシステムと区別するためにトランスオーラルシステム（**図2（b）**）と呼ばれる。再生音場において各スピーカから聴取者の両耳までの伝達特性をあらかじめ測定しておき、収録音にその逆特性を施すことで受聴点でのバイノーラル再生を実現する。これらの方法は空間上の固定点を制御するもので、受聴者が頭を動かしたり移動したりすると臨場感や方向感が損なわれる。点ではなく領域を制御する方法も提案されている。

以上は原音場の音響情報を保存し、それを別の場所に再現する音場再現と呼ばれる技術であるが、個々の音源に定位感を与えて立体音を創り出す音像定位技術も広く研究されている。人間は、音源から両耳に到達するまでの音響的な特性（頭部伝達関数:HRTF）の違いから音の方向や距離を認知している。あらゆる方向からのHRTFを予め測定しておき、音源に対し特定の方向のHRTFを適用したバイノーラル信号を耳元で受聴すれば、特定の位置に音源が存在するがごとく音を定位させるができる（**図3**）。これを臨場感テレワークに適用するには、個々の音源の位置が特定され、かつ音源ごとに音が分離されていなければならない。そのため音源位置の推定や音源分離などの技術も必要になる。

（2）高品質化技術

音源から発せられた原信号は空気を媒介として伝達され、耳やマイクロホンに到達する。その間、目的音以外の話し声や環境雑音、残響などが混ざり合いさまざまな変形を受ける。高品位な音コミュニケーションの実現には立体音響だけでなく、これら変形への対策が必要となる。

雑音の種類には、空調音のように比較的定常ではあるが音源が1方向に特定できない拡散性雑音と、音声や音楽のように指向性があって時間変動の大きい指向性雑音がある。性質の違いから対策方法も異なる。拡散性雑音に対しては、スペクトルサブトラクション（通称SS法）やウィーナーフィルタなどの雑音除去方式が、また、主マイクとは別に雑音のみが観測できる参照マイクが利用できる場合には、ノイズキャンセラーの手法が使える。一方、指向性雑音は、このような雑音除去方式を用いて取り除くことは困難である。これには複数の音源が混ざり合った音から目的とする音（主として音声）だけを分離･抽出する音源分離技術が用いられる。マイクロホンアレーを用いて目的音方向に強い指向性を向けるビームフォーマーや、独立成分分析（ICA）を用いた音源分離がよく知られている。

残響は音の了解性を損なうだけでなく音源分離や後述の音場制御に多大な悪影響を及ぼす。残響除去は見過ごすことのできない重要課題である。このほかにも遠隔地側でスピーカから発せられた音がマイクに回り込み、再び発声者側に戻ってくる音響エコーへの対処も必要である。回線エコーと比べ、遅延時間が長い、エコー経路特性が変動するなど難しい課題がある。

音　源　分　離

テレワークにおける音処理の要素技術の1つとして、音源分離技術の研究開発に取り組んでいる。小林らが提案した本方式[1)]はコンパクトなマイクロホン配置で、かつ少ない演算コストで実現可能である。2つのマイクを用いた基本的な方式の構成を**図4**（次ページ）に、また本方式に用いる空間フィルタの原理を**図5**（次ページ）に示す。

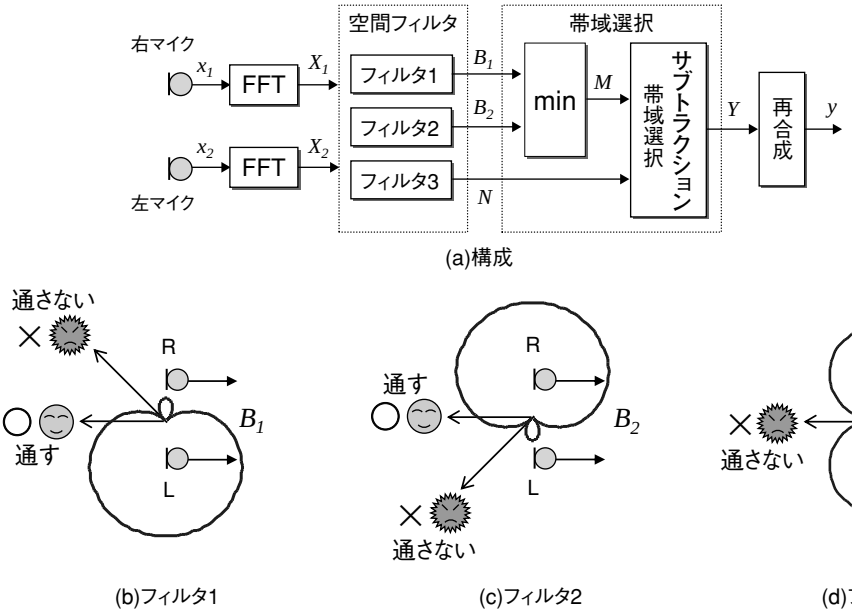


図4　音源分離方式

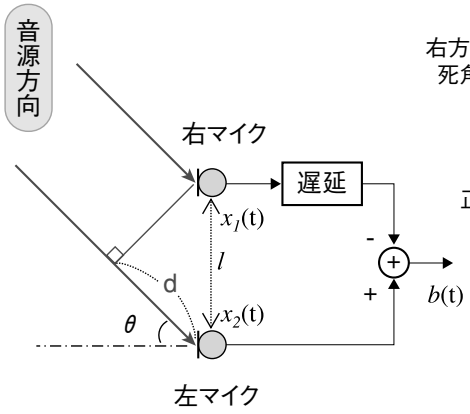


図5　（a）空間フィルタの原理

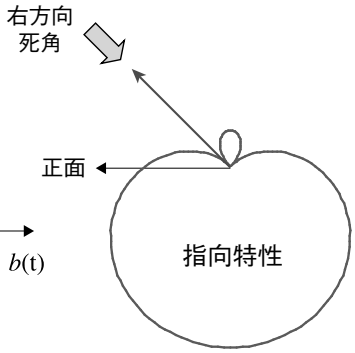


図5　（b）指向特性

はじめに空間フィルタの原理について説明する。**図5**（**a**）において、θ方向から到来する平面波を距離lだけ離れて設置された左右2つのマイクロホンで受音することを考える。θ方向から到来した音波は、まず音源に近い右マイクに受音される。次に音波は距離dだけ進んで左マイクに到達する。距離dは

$$d = l \sin\theta \qquad (1)$$

と表される。したがって、左マイクでの受音信号$x_2(t)$は右マイクでの受音信号$x_1(t)$と比べて音波が距離dだけ進行するのに要する時間τだけ遅れた信号となっている。すなわち

$$x_2(t)=x_2(t-\tau) \qquad (2)$$

$$\tau = d/c = l \sin\theta / c \qquad c：音速 \qquad (3)$$

の関係が成立する。したがって$x_1(t)$にτなる遅延を与え$x_2(t)$から減算（逆位相で加算）すれば（式（4））、信号同士が相殺され、特定方向θに死角が形成される。

$$b(t)=x_2(t)-x_2(t-\tau) \qquad (4)$$

このときの指向特性の例を**図5**（**b**）に示す。

時間軸上での空間フィルタ形成操作は、周波数領域でも同様に行うことができる。時間軸をτだけ遅らせた信号のフーリエ変換は、もとの信号をフーリエ変換した結果に$e^{-j\omega\tau}$を乗じたものになることが知られている。時間軸上の式（4）は、$x_1(t)$、$x_2(t)$の短時間フーリエ変換$X_1(\omega)$、

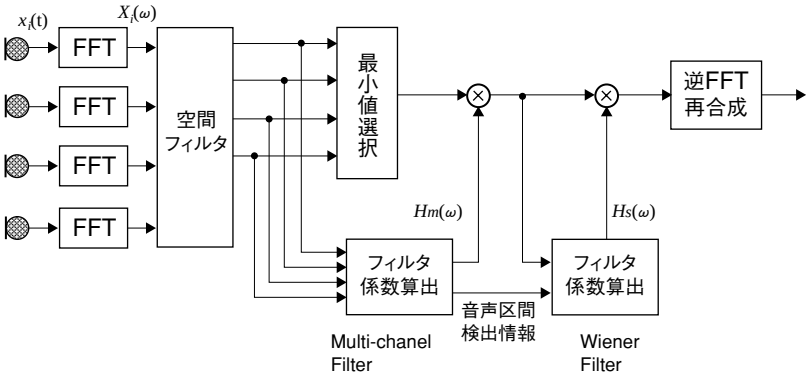


図6　拡散性雑音を考慮した音源分離の構成

$X_2(\omega)$を用いて周波数軸上では式（5）のように表される。

$$B(\omega) = X_2(\omega) - e^{-j\omega\tau} X_1(\omega) \qquad (5)$$

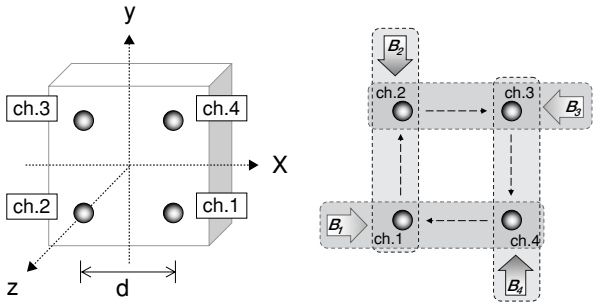
次に音源分離方式について説明する。この方式では**図4**（a）に示すように2つのマイクからの入力を用いて3つの空間フィルタを形成する。空間フィルタ1は右方向に死角が設定されており、右方向から到来する妨害音を抑圧する。目的音は、ある利得を持って出力される。この出力を$B_1(\omega)$とする（**図4**（b））。空間フィルタ2は左方向に死角が設定されており、左方向から到来する妨害音を抑圧する。空間フィルタ1と同様、目的音はある利得を持って出力される。出力を$B_2(\omega)$とする（**図4**（c））。空間フィルタ3は、正面方向に死角が設定され（**図4**（d））、目的音以外の成分を抽出する働きを有する。出力を$N(\omega)$とする。空間フィルタ1の出力の振幅成分$|B_1(\omega)|$と空間フィルタ2の出力の振幅成分$|B_2(\omega)|$の小さい方を選択する。

$$M(\omega) = \min\left[|B_1(\omega)|, |B_2(\omega)|\right] \qquad (6)$$

右方向に妨害音音源が存在した場合、右方向に死角を持つ空間フィルタ1の出力$B_1(\omega)$は、妨害音が抑圧されて振幅が小さくなる。これに対し妨害音が存在しない方向に死角を持つ空間フィルタ2の出力$B_2(\omega)$には振幅に大きな変化はないと考えられる。逆に、左方向に妨害音源があれば$B_2(\omega)$は小さくなるが$B_1(\omega)$の変化は少ない。したがって最小値選択されたMは、最大の妨害音を抑圧した目的音候補成分である。最後に$M(\omega)$と$N(\omega)$によって以下のように帯域選択とスペクトル・サブトラクションを行い出力$Y(\omega)$を決定する。

$$Y(\omega) = \begin{cases} \sqrt{|M(\omega)|^2 - \alpha|N(\omega)|^2} & \text{if } |M(\omega)| > \alpha|N(\omega)| \\ 0 & otherwise \end{cases} \qquad (7)$$

ここでは空間フィルタゲイン補正係数である。帯域選択は、信号$M(\omega)$に目的音の成分が含まれているかどうかを判定するために行なう。$N(\omega)$は目的音方向以外からの周囲雑音と考えられるから$N(\omega)$が$M(\omega)$より大きい場合は、そもそも目的音の成分が存在しない区間とみなして棄却する。$M(\omega)$に目的音の成分があると判断されれば、サブトラクションを行なって正面方向に鋭い指向性を向け、目的音を分離する。



(a)マイク配置　　(b)マイクの組み合わせ

図7　マイク配置

簡単のため、ここでは2マイクでの構成を示したが左右水平方向だけでなく上下垂直方向にもマイクを配置すれば、空間中の種々の方向からの指向性雑音に対応可能になる。

拡散性雑音を考慮した音源分離

実際の使用環境では指向性雑音だけが存在することはごく稀であり、指向性および拡散性雑音が混在して存在する。ここでは拡散性雑音も同時に抑圧する音源分離システムについて述べる[2)]。システムは**図6**に示すように、指向性雑音抑圧部、拡散性雑音抑圧部、残留雑音抑圧部から構成される。本システムでは、**図7**（a）で示すように平面上に4個の無指向性マイクを正方形に配置する。目的音は正面（Z軸方向）から到来するものとする。

(1) 指向性雑音抑圧

始めに本システムにおける指向性雑音抑圧から説明する。先に記した空間フィルタの原理を用いて、4つのマイクのうちの2個ずつを**図7（b）**のように組み合わせ4通りのマイクペアから4方向の空間フィルタを構成する。それぞれの空間フィルタは、式（8）～（11）で実現され上下左右4方向に指向性を持つ。

$$B_1(\omega) = X_1(\omega) - e^{-j\omega\tau} X_4(\omega) \quad (8)$$

$$B_2(\omega) = X_2(\omega) - e^{-j\omega\tau} X_1(\omega) \quad (9)$$

$$B_3(\omega) = X_3(\omega) - e^{-j\omega\tau} X_2(\omega) \quad (10)$$

$$B_4(\omega) = X_4(\omega) - e^{-j\omega\tau} X_3(\omega) \quad (11)$$

これら4つの空間フィルタの出力の振幅成分のうち最も小さな成分を選択し出力とすることで、指向性雑音の成分を最も小さくした出力を得ることができる。

$$|B_{\min}| = \min\,[\,|B_i|\,]\ (i=1,2,3,4) \quad (12)$$

(2) 拡散性雑音抑圧

拡散性雑音抑圧は指向性雑音の抑圧と同じ4つの空間フィルタ出力を用いたマルチチャンネルウィーナーフィルタで実現する。目的音である話者の声は各マイクロホンで観測される信号の相関が高いが、拡散性の雑音は各信号間で相関が低い。この性質を利用し、対向する方向に指向性を持った信号（B_1とB_3、B_2とB_4 ）を組み合わせ、互いの相関の程度を反映した係数を持つフィルタを構成する。

$$H_m(\omega) = \frac{|B_1(\omega)B_3^*(\omega)| + |B_2(\omega)B_4^*(\omega)|}{\frac{1}{2}\sum_{i=1}^{4}|B_i(\omega)|^2} \quad (13)$$

上式は分子のクロススペクトルを分母のパワースペクトルで正規化する形になっており、相関が高ければ1に、低ければ0に近づく特性を持つ。このフィルタを前記の指向性雑音を抑圧した信号$|B_{\min}|$に乗じることにより、相関が低い成分を抑圧し拡散性雑音を低減する。

$$\hat{S}_m(\omega) = H_m(\omega)|B_{min}(\omega)| \quad (14)$$

(3) 残留雑音抑圧

指向性雑音、および拡散性雑音を抑圧した信号$\hat{S}_m(\omega)$に対し、さらにシングルチャンネルのウィーナーフィルタを適用して残留する定常雑音を抑圧する。ウィーナーフィルタは、信号や雑音を確率過程とみなし平均二乗誤差を最小にするフィルタであり、信号と雑音が無相関であると仮定するとゲイン関数は次式のように与えられる。

$$H_s(\omega) = \frac{SNR_{prio}(\omega)}{SNR_{prio}(\omega)+1} \quad (15)$$

ここで、事後SN比$SNR_{post}(\omega)$、事前SN比$SNR_{prio}(\omega)$をそれぞれ、以下に定義する。

$$SNR_{post}(\omega) = \frac{|\hat{S}_m(\omega)|^2}{E[\,|N(\omega)|^2\,]} \quad (16)$$

$$SNR_{prio}(\omega) = \frac{E[\,|S(\omega)|^2\,]}{E[\,|N(\omega)|^2\,]} \quad (17)$$

$E[\cdot]$は期待値を、$S(\omega)$は目的音信号表す。事前SN比$SNR_{prio}(\omega)$は、$E[\,|S(\omega)|^2\,]$を含むため直接測定できない。そこで事後SN比と前フレームの推定信号$\hat{S}_{-1}(\omega)$を用いて近似的に計算する。

$$\hat{SNR}_{prio}(\omega) = \beta\frac{|\hat{S}_{-1}(\omega)|^2}{E[\,|N(\omega)|^2\,]} + (1-\beta)P[SNR_{post}(\omega)-1] \quad (18)$$

ここで、$P[\cdot]$は半波整流、βは忘却係数を示す。

一方、雑音レベルの推定は、非発話区間の信号から以下のように忘却的に行う。

$$|N(\omega)|^2 = (1-\lambda)|\hat{S}_m(\omega)|^2 + \lambda|N_{-1}(\omega)|^2 \quad (19)$$

忘却係数λは、0.95～0.99程度に選ばれる。また、目的音成分の混入を防ぐために、音声発声区間中は雑音学習を停止する。

(4) 音源分離装置の試作

開発した音源分離方式を実環境で評価するため、4チャンネルのMEMSマイク、CPUボード、AD変換ボードを搭載した小型端末を試作した（**写真1**）。演算処理をすべて固定小数点化した上で、FFT、平方根、倍長除算などの演算処理を高速化、前記のすべての処理を試作機内に実装した。マイク間の距離は縦横ともに3cmと非常に小型であり、リモコンや携帯電話などの小型の機器にも実装可能である。

ま と め

臨場感テレワークにおける音処理技術について概観し、要素技術である音源分離技術について説明した。音による臨場感生成技術は、音楽演奏などの固定コンテンツを対象にした音場再現技術として、あるいは音源そのもの

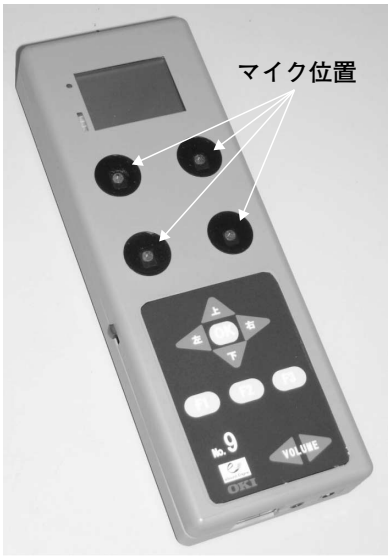


写真1　音源分離試作機

に定位感を付加して仮想的な音像を作り出す音像定位技術として研究されてきた。しかし、実時間遠隔コミュニケーションにおける臨場感の生成には、まだ多くの課題が残されている。あたかも職場にいるような音環境の実現を目指し音処理技術の研究を行っていく。なお音源分離方式の開発・試作は、経済産業省、平成18、19年度戦略的技術開発委託費「音声認識基盤技術の開発」の一部として、早稲田大学からの委託により実施されたものである。 ◆◆

■参考文献

1）高田晋太郎，他：少数のマイクロホンを用いた携帯端末向け音源分離，日本音響学会講演論文集，3-1-8，2006年9月
2）高田晋太郎，他：指向性雑音と拡散性雑音の混在する環境を対象とした携帯端末向け音声強調の検討，日本音響学会講演論文集，3-P-3，2007年9月

●筆者紹介

矢頭隆：Takashi Yazu. 研究開発本部 ヒューマンコミュニケーションラボラトリ スペシャリスト
森戸誠：Makoto Morito. 研究開発本部 ヒューマンコミュニケーションラボラトリ シニアスペシャリスト