Vol.2015-MUS-107 No.28
2015/5/23

音響再生方式を考慮した聴覚臨場感の時系列推定モデルの構築

伊藤 将亮[1,a)]　森勢 将雅[2,b)]　小澤 賢司[2,c)]　木下 雄一朗[2,d)]

概要：臨場感は AV 機器やコンテンツの評価に重要であるが，未だ定量的に臨場感を推定する感性モデルは構築されていない．本研究では臨場感は，使用する再生システムに依存するものであると考え，バイノーラル再生方式，ダイオティック再生方式の 2 種の異なる音響再生方式で再生される同一の素材を用い，臨場感の時々刻々の変化を時系列として評価する実験を 20 名の被験者に対して行った．また，実験結果に基づき，音の特徴量から時系列ごとに聴覚臨場感を推定するモデルをニューラルネットワークを使用して構築した．モデルは，バイノーラル再生方式で再生される素材 40 種，ダイオティック再生方式で再生される素材 40 種の計 80 種を学習データとして誤差逆伝搬法を用い学習した．構築されたモデルについて汎用性を検証した結果から，80 種の刺激について実験から得られた臨場感評価値とモデル出力の平均誤差が 7 段階尺度において 0.47 となり，音響再生方式の違いも含めた臨場感を良好に推定できることを示した．

1. はじめに

昨今，臨場感という用語が AV（Audio–Visual）機器やコンテンツの評価で使用されている．音響分野では，5.1 ch サラウンドや 22.2 マルチチャンネルといった 3 次元音響技術が開発されており，映像分野においても，3D テレビや 4K テレビ，そして，それらを超える画質を持つスーパーハイビジョンといった高臨場感を実現する技術が開発されている．高臨場感を伝える機器の開発と同時に，臨場感を評価する研究も多数行われている [1–8]．しかし，未だ定量的に臨場感を推定する感性モデルは確立されていない．つまり，臨場感の程度を数値などの明確な形で表現できないため，いかに高臨場感を感じさせるシステムであってもどれほど優れたシステムなのかを定量的に表示する指標は与えられていない．再生技術は今後も発展していくと考えられるため，臨場感の性質を解明することは重要であると考えられる．

臨場感は，同じの対象であっても録音・再生系によって異なる臨場感を伝える「システム臨場感」と，同一の録音・再生系であっても対象によって異なる臨場感を伝える「コンテンツ臨場感」の 2 つに分けられる [8]．先行研究 [9,10] では，刺激ごとの臨場感を推定するモデルを構築した．具体的には，先行研究 [9] では，コンテンツ臨場感を推定するモデルを構築し，先行研究 [10] では，システム臨場感も推定するモデルを構築した．一方，先行研究 [11] では，臨場感は時々刻々と変化するものであると考え，聴覚のみにおけるコンテンツ臨場感を時系列として推定するモデルを構築した．しかし，完全な臨場感推定モデルを構築するためには，音響再生方式による影響を考慮して時々刻々と変化する臨場感評価値を得る必要がある．そこで，本研究では異なる音響再生方式が時系列としての聴覚臨場感に与える影響について調査し，音響再生方式を考慮した聴覚臨場感の時系列推定モデルを構築することを目的とする．

[1] 山梨大学大学院医学工学総合教育部
University of Yamanashi, 4–3–11 Takeda, Kofu, 400–8511, Japan
[2] 山梨大学大学院総合研究部
University of Yamanashi, 4–3–11 Takeda, Kofu, 400–8511, Japan
[a)] g15mk002@yamanashi.ac.jp
[b)] mmorise@yamanashi.ac.jp
[c)] ozawa@yamanashi.ac.jp
[d)] ykinoshita@yamanashi.ac.jp

2. 聴覚臨場感の実時間評価実験

2.1 実験の目的

本実験は，音響再生方式によって異なる聴覚臨場感の時系列推定モデルの構築に向けて，被験者に実際の音を提示し，それに対する臨場感の程度を実時間測定することを目的とする．45 種の音刺激を用意し，それら全てに対し臨場感を 7 段階で実時間評価を行った．これにより，各刺激に対する実時間臨場感評価値を得ることとした．

2.2 音刺激

本研究で用いる音素材は，先行研究 [12] で用いられて

表 1 刺激一覧（1–40：ダイオティック再生方式，1'–5'：バイノーラル再生方式）.

刺激番号	コンテンツ名
1	自動車の通過
2	電車の通過
3	ローラコースタ A の通過 (前方)
4	ローラコースタ A の通過 (後方)
5	キャンパス内の施設前 (鳥の声・自動車など)
6	夕方の学校 (チャイム・鳥の声など)
7	甲府駅前の歩行者・バスなど
8	公園で落ち葉が舞う風景
9	トンネル内での自動車の通過
10	バスケットボールの試合
11	バッティングセンタ
12	花火 (穏やか)
13	花火 (激しい)
14	自分が廊下を歩行
15	自分がロール式滑り台を滑降
16	自分が乗車中のロープウェイ
17	自分が乗車中の自動車のエンジン音
18	メリーゴーランド
19	夜店の様子
20	弓道
21	研究室の一席
22	ボートの通過
23	ローラコースタ B の通過 (後方)
24	キャッチボール
25	スピーカから鳴る音楽
26	滝
27	木々のざわめき
28	噴水 (夜景)
29	料理風景
30	公園での雑踏
31	夜景
32	湖畔
33	携帯電話の着信
34	タイピング
35	湖面に映る夜景と雨
36	合唱
37	吹奏楽の演奏
38	静かな森林
39	静かな廊下
40	小川のせせらぎ
1'	自動車の通過
2'	電車の通過
3'	ローラコースタ A の通過 (前方)
4'	ローラコースタ A の通過 (後方)
5'	キャンパス内の施設前 (鳥の声・自動車など)

いたものと同じものを使用した．音刺激の一覧を表 1 に示す．これらの刺激の再生時間は約 20〜40 秒間である．

素材の録音は，音場の情報を可能な限り正確に再現するために，ダミーヘッド（高研，SAMRAI）を用いたバイノーラル録音で行なわれている．記録方式は非圧縮（PCM: Pulse Code Modulation，16 bit 量子化，48 kHz 標本化）である．

先行研究 [12] では，本実験と同様の素材をバイノーラル再生方式で再生した刺激を用い臨場感評価を行った．一方，本実験では，刺激番号 1 から 40 をダイオティック再生方式で再生することとした．さらに，刺激番号 1' から 5' を先行研究 [12] で用いられたバイノーラル再生方式で再生される刺激とし，被験者に提示した．ここでバイノーラル再生方式を本実験に用いた理由として，本研究で構築する聴覚モデルへの教師信号として先行研究 [12] で行われた実験結果も使用することになるため，本実験と先行研究の実験の結果が符合していることを確認するためである．

2.3 2 種の音響再生方式

先行研究 [10] において音響再生方式が臨場感の評価に影響を与えることが示されているため，音像の動きや，音像定位の再現精度が異なる音響再生方式を選定した．先行研究 [12] で使用された音響再生方式は，バイノーラル再生方式である．この音響再生方式と比較するため，本研究では，先行研究 [10] で使用された 5 種の音響再生方式の実験結果を参照してダイオティック再生方式を選択し，実験を行った．以下に使用した音響再生方式について述べる．

本研究でバイノーラル再生方式と呼称するのは，ダミーヘッドを用いて録音したバイノーラル音源についてヘッドホン順特性を打ち消す補正 [13] を施した後に，ヘッドホンにより再生する方式である．つまり，バイノーラル再生方式は，理論上ダミーヘッドが原音場で音を聞いている状態を忠実に再現したことになる．ただし，ダミーヘッドと聴取者の間には頭部伝達関数と呼ばれる音源から外耳までの音響伝達関数に差異があるため，聴講者が原音場で音を聴く状態を完璧に再現できるわけではない．

ダイオティック再生方式とは，左右チャネルに同じ音を提示する方式である．具体的には，バイノーラル再生方式における左右チャネルの信号の平均値を両耳に提示する．左右チャネルに同じ音を提示することにより，両耳間の音圧レベル差や時間差が消失するため音像が全て頭内に定位して動かない．そのため，音像の定位感が大きく損なわれる．

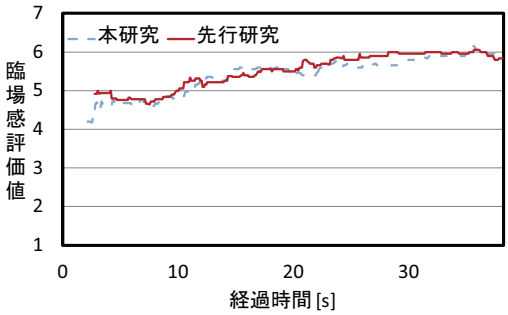
2.4 実験概要

実験では，表 1 に示した 45 種の刺激をランダムな順序で被験者に提示した．まず被験者は，各刺激の提示中に「臨場感を感じている」という項目を 7 段階で評価した．ここで，1 が「臨場感を全く感じていない」，7 が「臨場感を非常に感じている」の評価に対応している．この評価は，キーボードの 1〜7 のキーを使用して，評価値に対応するキーを押し続けることで評価した．実験に先立ち，被験者に対し臨場感の定義を「その場にいる感じ」と教示した．

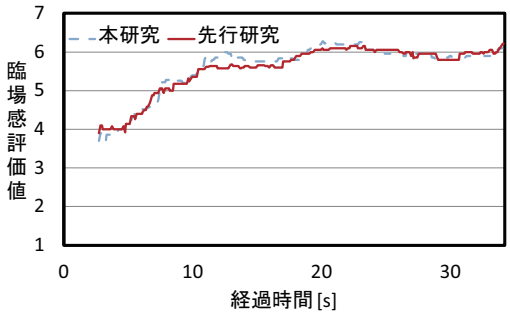
各刺激の提示後に，被験者は，刺激全体に対して「臨場感を感じた」という項目を 7 段階のリッカート尺度で評価した．この評価は，各刺激の提示後に表示される評価フォームにマウスでクリックすることにより行った．なお，10 種の刺激に対する評価が終わるごとに休憩時間を 10 分程度設けた．

実験は，実際にテレビを視聴する環境を想定して，一般的な部屋（内寸: W: 3.8 × D: 6.3 × H: 2.8 m）において，被験者はソファーに座ってコンテンツを視聴することで実施した．なお，ディスプレイと被験者の距離は，ITU-R Rec. Bt.710-4 [14] で勧告されている 3H（2.4 m）とした．

被験者は，10 代および 20 代の男子大学生 15 名，女子大学生 5 名の計 20 名である．なお，本実験の被験者は，先



(a) 自動車の通過：バイノーラル再生方式



(b) ローラコースタ A の通過 (後方)：バイノーラル再生方式

図 1 本実験の信頼性の検証．バイノーラル再生方式で再生された刺激の各実験の比較．

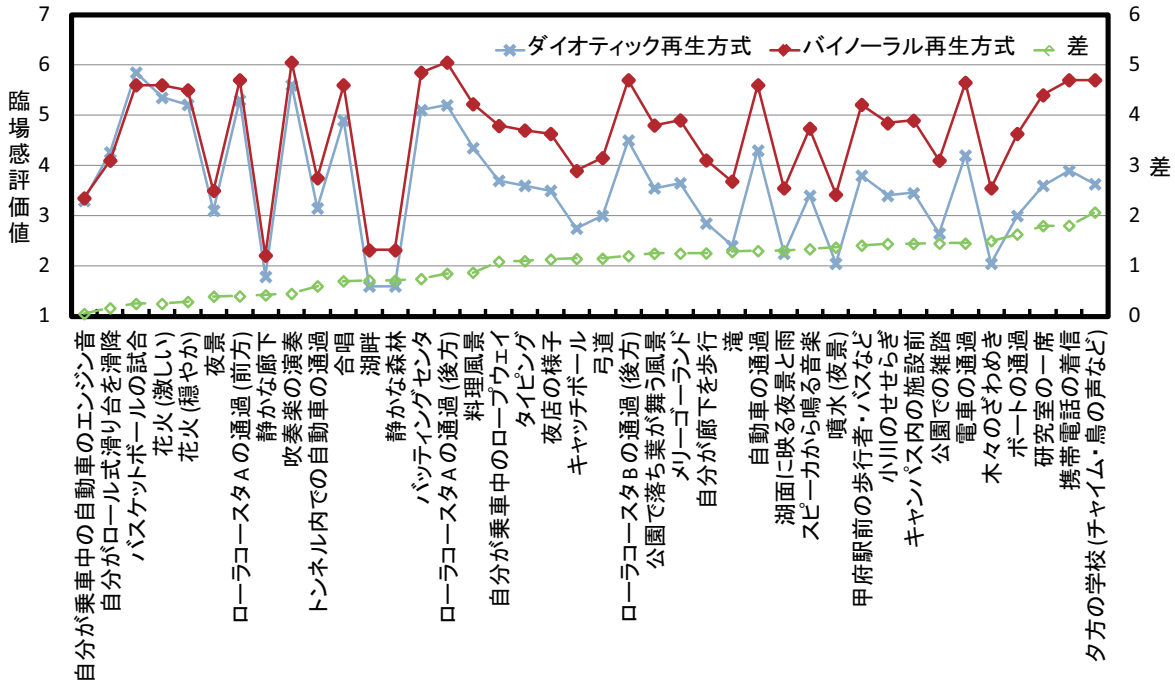


図 2 異なる音響再生方式による素材全体における聴覚臨場感の比較

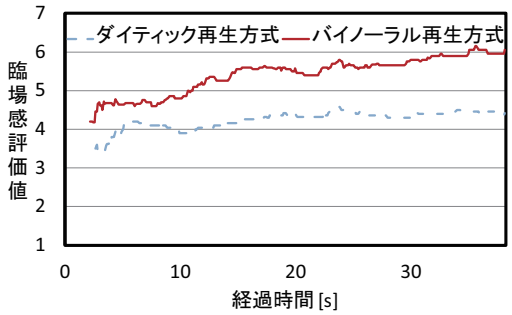
行研究 [12] の実験に参加した被験者とは異なる．

2.5 実験結果と考察

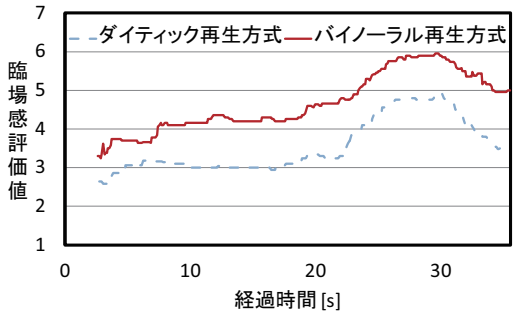
図 1 に先行研究から得られた実験結果と，本実験から得られた実験結果の一例を示す．グラフから，各実験から得られた臨場感評価値は同様の推移傾向が見られ，その差は小さいことがわかる．ここから，先行研究で行われた実験と，本実験の結果が符合していることを確認できる．

刺激全体に対する臨場感の評価結果を図 2 に示す．グラフ内の「ダイオティック再生方式」は本実験から得られたダイオティック再生方式で再生された刺激全体の臨場感評価値，「バイノーラル再生方式」は先行研究 [12] において得られたバイノーラル再生方式で再生された刺激全体の臨場感評価値を示し，差は各音響再生方式間の差の絶対値を表す．差の大きい素材の特徴として，音像の動きが大きく，音像の広がり感が感じられる素材であることが挙げられる．また，ほとんどの素材についてバイノーラル再生方式が，ダイオティック再生方式より高い臨場感評価値であることがわかる．ここから，音響再生方式によって臨場感が変化することがわかる．バイノーラル再生方式は，ダイオティック再生方式よりも原音場をより精密に再現できるため，臨場感が高くなっていると推測できる．

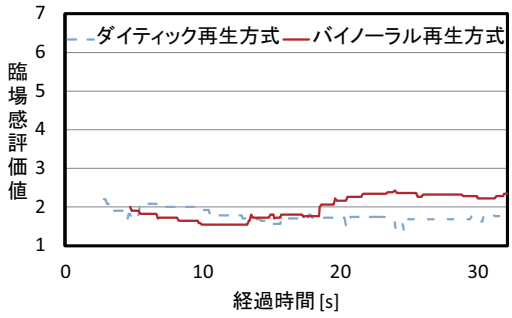
図 3 にバイノーラル再生方式とダイオティック再生方式の 2 種の音響再生方式で再生された場合の実験結果の一例を示す．図 3(a)，図 3(b) は，音像の動きが大きい素材（刺激番号 1, 2）の実験結果を表し，図 3(c) は，音像の動きが小さい素材（刺激番号 39）の実験結果を示す．グラフ内のバイノーラル再生方式の結果は，先行研究 [12] で観測された臨場感評価値を使用している．音像の動きが大きい素材では，音響再生方式の違いによる観測値に大きな差がある．



(a) 自動車の通過（音像の動きがある素材）



(b) 電車の通過（音像の動きがある素材）



(c) 静かな廊下（音像の動きがない素材）

図 3 異なる音響再生方式で再生される素材の時系列としての臨場感評価結果

しかし，音像の動きが小さい素材では，音響再生方式の違いによる観測値の差が小さいことがわかる．音像が動くことで物体の移動がイメージできるため，より原音場の状況を把握しやすくなったと推測できる．このことから，高い臨場感を得るためには，再生システムは音像の動きを正確に伝えることが重要であるといえる．

3. 聴覚臨場感推定モデルの構築と評価

本章では，第 2 章における実験結果と，先行研究 [12] から得られた臨場感評価値を基に，臨場感を時系列として推定するモデルをニューラルネットワークを用いて構築する．

3.1 モデルへの入力とする特徴量

以下に示す特徴量を聴覚モデルへの入力とした．

(1) 500 ms ごとのラウドネス [sone]
(2) 500 ms ごとのシャープネス [acum]
(3) 500 ms ごとの過去 3 s における 95 %時間率騒音レベル L_{95} に対する 5 %時間率騒音レベル L_5 の相対レベル [dB]
(4) 500 ms ごとの過去 3 s における上述の相対レベルにおける標準偏差 [dB]
(5) 500 ms ごとの両耳間レベル差 [dB]
(6) 500 ms ごとの両耳間相関度 [無名数]

各入力値となる特徴量は全ての刺激を通じて各特徴量における最大値を 1，最小値を 0 とする [0, 1] の範囲の実数値に正規化した．なお，各素材の 500 ms ごとにおける $\frac{1}{3}$ オクターブバンドごとの平均音圧レベルを基に，ラウドネス [15] を算出した．シャープネスの算出には Zwicker のモデル [15] を用いた．

3.2 モデルの構築

モデルの構築には階層型ニューラルネットワークを用いた．学習には，誤差逆伝搬法を適応し，学習パラメータは試行錯誤的に決定した．上記した特徴量を聴覚臨場感の時系列推定モデルへの入力とし，ニューラルネットワークの入力層におけるユニットと対応付けた．そのため，入力層のユニット数は特徴量の数と等しくなるため，6 とした．中間層のユニット数は試行錯誤的に 15 とした．出力層のユニット数は 1 であり，出力値となる臨場感推定値もまた，[1, 7] の範囲から [0, 1] の範囲の実数値に正規化して扱う．

3.3 推定精度の検証方法

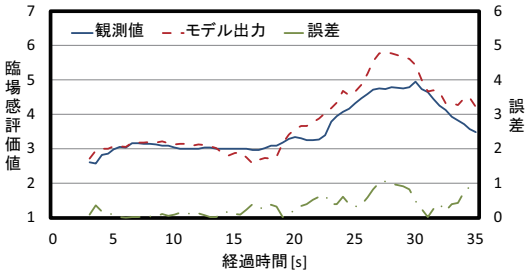
ダイオティック再生方式で再生された 40 種と，先行研究 [11] により得られたバイノーラル再生方式で再生される 40 種の刺激，計 80 種をモデル構築・評価のサンプルとして用いる．これらの刺激を，モデル構築のための学習用サンプルと構築後の検証用サンプルのどちらかに分割すると，十分に学習が行なわれない，または，汎化性の検証が不十分になる可能性がある．そこで，検証の手法として leave-one-pair-out cross-validation を用いた．

この方法では，サンプルが N 個の場合，それを $N-2$ 個の学習用サンプルと 2 個の検証用サンプルに分割し，$N-2$ 個のサンプルを用いた学習結果から残り 2 個のサンプルを評価する．この場合の 2 個のテスト用サンプルは，同じ素材で異なる音響再生方式で再生された刺激をペアとして取り出したものとなる．leave-one-pair-out cross-validation を使用することで同じ内容の素材を含まずにモデルの検証を行えるため，モデルの検証に適していると考えた．

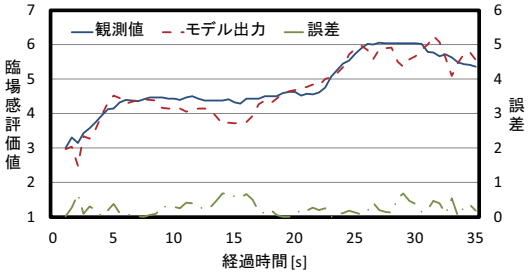
本研究では，39 素材分のサンプルで学習し，1 素材分のサンプルで検証することを 40 通りの分割全てに対して行い，汎化性能を検証した．

3.4 推定精度の検証結果と考察

前項で述べた 6 種の特徴量を用いて構築したモデルにつ

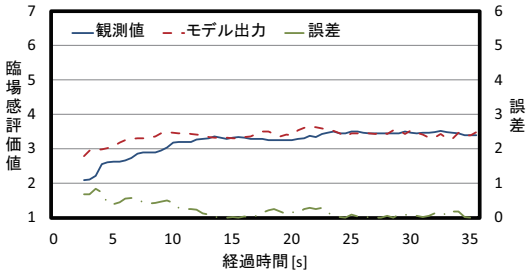


(a) ダイオティック再生方式

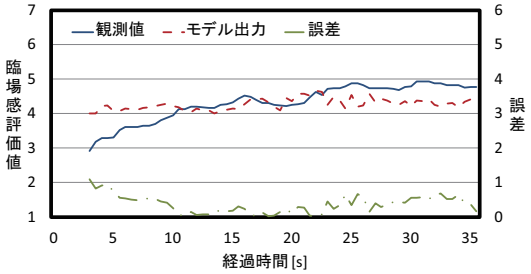


(b) バイノーラル再生方式

図 4 モデルの有効性の検証：電車の通過



(a) ダイオティック再生方式



(b) バイノーラル再生方式

図 5 モデルの有効性の検証：キャンパス内の施設前

いて，検証結果の一例を図 4，図 5 に示す．グラフの横軸は経過時間を表し，縦軸は臨場感評価値を示す．また，モデルの出力結果と観測値の差の絶対値を誤差としてグラフ内に示す．それぞれの平均誤差は，図 4(a) が 0.33，図 4(b) が 0.26，図 5(a) が 0.21，図 5(b) が 0.37 となった．図 4，図 5 から，音響再生方式によらず，観測値とモデル出力の誤差が小さく，モデルは時々刻々の臨場感を推定していることがわかる．「電車の通過」，「キャンパス内の施設前」という素材は，音像の動きのあるコンテンツである．つまり，本研究で構築されたモデルは音響再生方式によって音像の定位感が損なわれる場合でも，時系列としての臨場感を推定できていることがわかる．

図 6 に横軸をコンテンツ番号，縦軸を平均誤差として表示した，ダイオティック再生方式とバイノーラル再生方式で再生されるそれぞれの刺激の検証結果を示す．誤差棒は各時系列データにおける標準偏差を表す．図 6 から，ほとんどの刺激における平均誤差は 1 未満であることがわかる．80 刺激全体の平均誤差は 0.47 となり，全ての刺激に対して良好に臨場感を推定しているといえる．しかし，「木々のざわめき」，「ボートの通過」といった一部の素材では，平均誤差が 1 を超えている．これらの素材では，音だけで原音場を正確にイメージすることは難しい．そのため，モデル出力との誤差が大きくなってしまったのではないかと考えられる．

4. おわりに

本研究では，音響再生方式を考慮した聴覚臨場感の時系列推定モデルを構築した．モデルの検証を行った結果，構築した聴覚モデルが観測値を十分に模擬していることを確認した．

今後の発展として，この手法を視聴覚条件にも応用したモデルの構築を行う予定である．

謝辞 本研究は，NICT 委託研究「革新的な三次元映像技術による超臨場感コミュニケーション技術の研究開発」の一環として行った．モデル構築にご助力いただいた，山梨大学工学部コンピュータ．メディア工学科卒業生 塚原将太氏に深謝する．

参考文献

[1] B. G. Witmer and M. J. Singer: Measuring presence in virtual environments: A presence questionnaire, Presence: Teleorepators and Virtual Environments, Vol. 7, pp. 225–240 (1998).
[2] M. Meehan, B. Insko, M. Whitton and F. P. Brooks, Jr.: Physiological measures of presence in stressful virtual environments, ACM Transactions on Graphics, Vol. 21, pp. 645–652 (2002).
[3] K. Ozawa, Y. Chujo, Y. Suzuki and T. Sone: Contents which yield high auditorypresence in sound reproduction, *Kansei* Engineering International, Vol. 3, No. 4, pp. 25–30 (2002).
[4] K. Ozawa, S. Ohtake, Y. Suzuki and T. Sone: Effects of visual information on auditory presence, Acoustical Science and Technology, Vol. 24, No. 2, pp. 97–99 (2003).
[5] E. Emoto, E. Masaoka, M. Sugawara and F. Okano: Viewing angle effects from wide video projection images on the human equilibrium, Displays, Vol. 26, pp. 9–14 (2005).
[6] T. Jari, N. Gote and L. Leif: Components of human experience in virtual environments, Computers in Human Behavior, Vol. 24, pp. 1–15 (2008).
[7] M. Iizuka, K. Ozawa, Y. Kinoshita and K. Fukue: Factor analysis of the sense of presence in daily scenes, Proc. 20th Inter. Cong. on Acoustics, No. 490 (5 pages on CD-

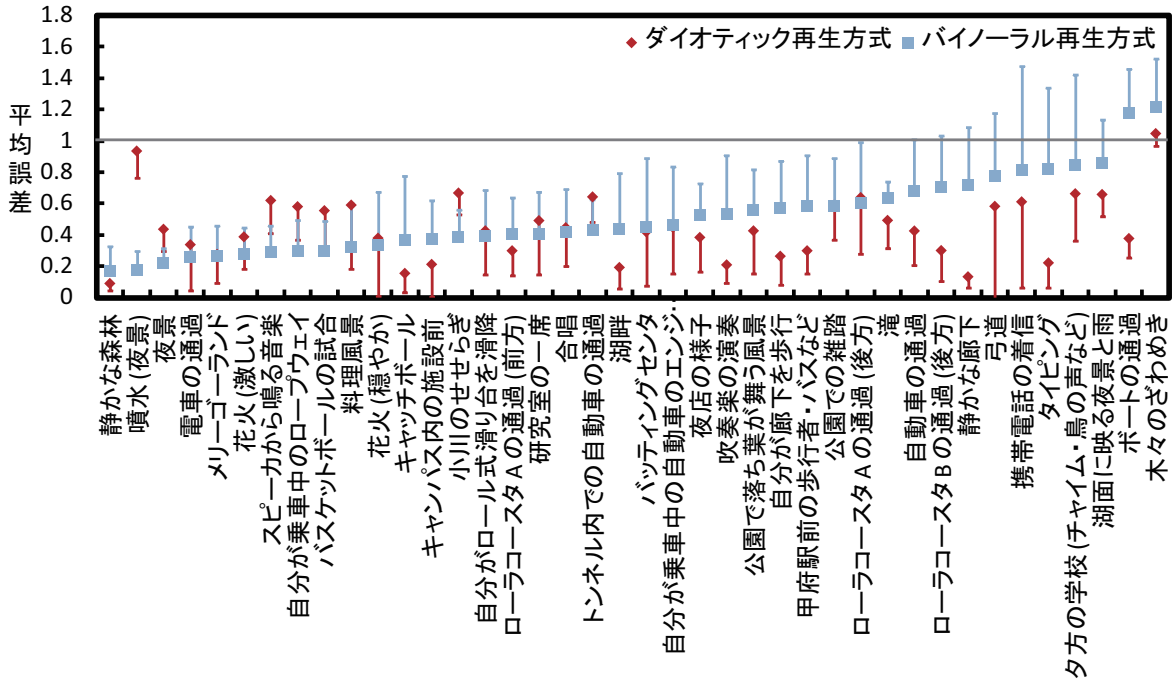


図 6 各刺激に対する聴覚臨場感の時系列推定モデルの平均誤差．誤差棒は時系列ごとの誤差の標準偏差を示す．

ROM) (2010).
[8] K. Ozawa and Y. Chujo: Content presence vs. system presence in audio reproduction systems, Proc. International Symposium on Universal Communication, pp. 50–55 (2008).
[9] Y. Kinoshita, K. Fukue and K. Ozawa: Development of *Kansei* estimation models for the sense of presence in audio-visual content, Proc. IEEE Inter. Conf. on Systems, Man, and Cybernetics (SMC 2011), pp. 3280–3285 (2011).
[10] K. Ozawa, M. Obinata and Y. Kinoshita: *Kansei* estimation models for the sense of presence in audio-visual content with different audio reproduction methods, Proc. ACIS Inter. Conf. on Software Engineering, Artificial Intelligence Networking and Parallel/Distributed Computing (SNPD 2012), pp. 567–573 (2012).
[11] K. Ozawa, S. Tsukahara, Y. Kinoshita and M. Morise: Development of an estimation model for instantaneous presence in audio content, Proc. of 2014 IEEE Fourth Inter. Conf. on Consumer Electronics - Berlin (IEEE ICCE-Berlin 2014), pp. 238–242 (2014).
[12] K. Ozawa, S. Tsukahara, Y. Kinoshita and M. Morise: Instantaneous evaluation of the sense of presence in Audio-Visual Content, IEICE Trans. on Information and Systems, Vol.E98-D, No.1, pp. 49–57 (2015).
[13] H. Møller: Fundamentals of binaural technology, Applied Acoustics, Vol.36, No.3–4, pp. 171–218 (1992).
[14] ITU-BT.710-4: Subjective assessment methods for image quality in high-definition television (1998).
[15] H. Fastl and E. Zwicker: Psychoacoustics–Facts and models, p.470, Springer, NewYork (2006).