論文 / 著書情報
Article / Book Information

題目(和文)	2次元画像からの後部残響特性推定に関する研究
Title(English)	
著者(和文)	今誉
Author(English)	Homare Kon
出典(和文)	学位:博士(工学), 学位授与機関:東京工業大学, 報告番号:甲第11481号, 授与年月日:2020年3月26日, 学位の種別:課程博士, 審査員:小池 英樹,篠田 浩一,徳永 健伸,三宅 美博,齋藤 豪
Citation(English)	Degree:Doctor (Engineering), Conferring organization: Tokyo Institute of Technology, Report number:甲第11481号, Conferred date:2020/3/26, Degree Type:Course doctor, Examiner:,,,,
学位種別(和文)	博士論文
Category(English)	Doctoral Thesis
種別(和文)	要約
Type(English)	Outline

博士論文　要約
2 次元画像からの後部残響特性推定に関する研究

東京工業大学 情報理工学院 情報工学系

今 誉

指導教官 小池 英樹 教授

博士論文要約

本研究の背景と目的

近年、XR と称される仮想現実（Virtual Reality, VR）、拡張現実（Augmented Reality, AR）や複合現実（Mixed Reality, MR）といった技術は、スマートフォンの普及、デバイスの小型化や省電力化によりウェアラブルデバイスも容易に実現できるようになり、新たなサービスを提供するプラットフォームとして期待されている技術である [1, 2]。VR は使用者が別の仮想空間へ行ったかのような体験を提供するのに対し、AR/MR は現実世界に情報を重ね合わせ、人間の現実認識を仮想的なデジタルコンテンツと複合や拡張させるものである。

特に MR は仮想的なオブジェクトの 3D CG 映像が投影されたシースルーヘッドマウントディスプレイ越しに現実空間を見ることで、本来ならば存在していない物体が本当に現実空間にあるかのように見ることができる。製造業においては建築物や機械部品の設計、生産ラインのレイアウトや室内設備の環境イメージの検討など、直感的なシミュレーションツールとしての利用が期待されている。

一方で、エンターテイメントへの利用としては、現実を拡張したゲームコンテンツや、遠隔地にいる人がこの場にいるかのようにコミュニケーションができる人と人とをつなぐ体験への応用が期待されている [3, 4]。これらにおいて、仮想的なデジタルコンテンツが本物と区別がつかないほどリアルに再現し、現実空間と融合させることができれば、製造業においてはより正確なシミュレーションに基づいた深い議論ができ、エンターテイメントにおいてもより没入感のあるコンテンツとして楽しむことができる。

ここで AR/MR 特有の重要な研究課題の 1 つは現実空間の環境に適合したデジタルコンテンツのレンダリング手法である [5, 6]。VR は仮想コンテンツへの没入を実現することが重要であったが、AR/MR は現実空間にいながら、現実と仮想の両方を同時に違和感なく体験させるという違いがあるためである。

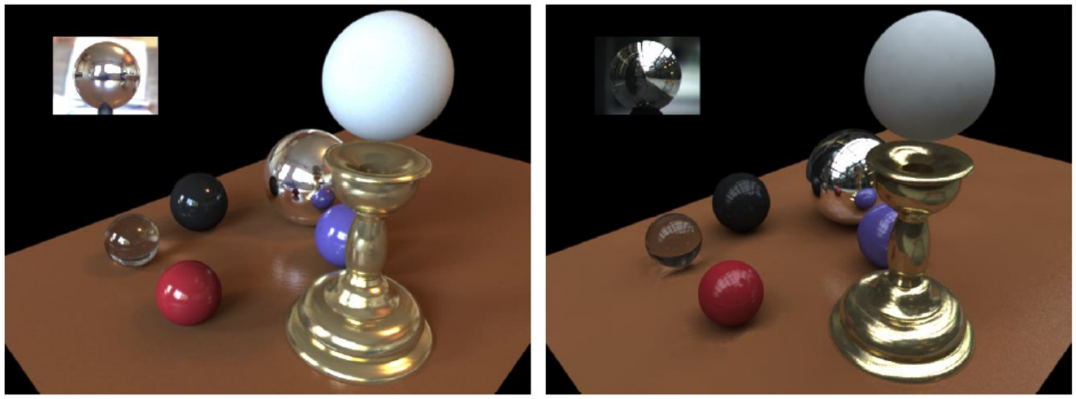
AR/MR においてデジタルコンテンツを環境に適合させる手法は視覚情報を中心に数

多くの研究開発がなされてきている。例えば、仮想的なオブジェクトを現実空間の床や壁に配置するための部屋のレイアウト認識と画像処理の手法 [1] や、質感を再現するために現実空間の光源位置と光源の特性を推定し、仮想オブジェクトがなすであろう床への影や物体の陰影を再現する手法などがある [5, 6](Figure 1)。

一方で、環境適合すべき聴覚情報はどうであろうか？ 空間伝搬する音の要素は Figure 2 に示されるように、大きく分けると直接音と反射音に分けられ、前者はその名の通り、音源から発せられた音が耳に直接届く音である。後者は音源から発せられた音が壁や床などから反射した音であり、この繰り返し反射された音を残響と言う [7]。 残響は空間の容積、境界面である壁や床の形状・反射率によって変化する。通常の壁や床などの建材は音響エネルギーを非常によく反射しており、人間は常に残響音を感じながら生活をしている。そのため、残響は人間が音から感じる空間的印象と密接な関係がある。例えば、長く続く残響音はコンサートホールのような広い部屋という印象を与えることから空間の大きさの知覚に影響を与える。また、高音が強調された残響音を聞くと、吸音率が低いコンクリートのような壁で囲まれた空間を想起させることから、残響音の音色は空間を構成する境界面の材質を想起させることができる。このように仮想的な音があたかも現実空間で鳴り響いているようにレンダリングするためには、現実空間に適合した残響を付加することが非常に重要であることがわかる [8]。

以上のような背景において、本研究の目的は環境に適合した残響の生成手法を示すことである。

(a)



(b)

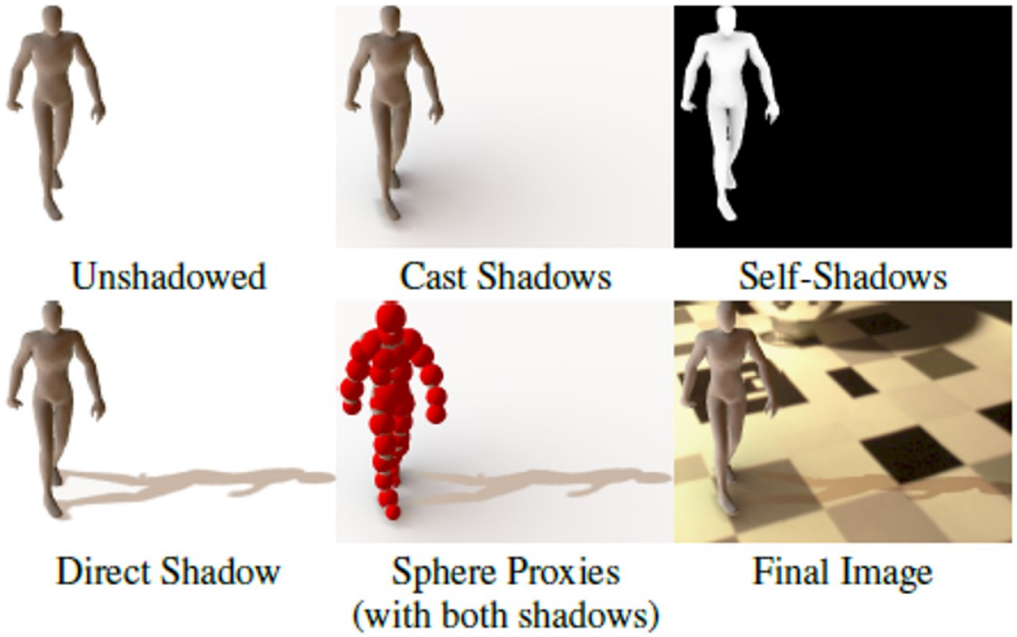
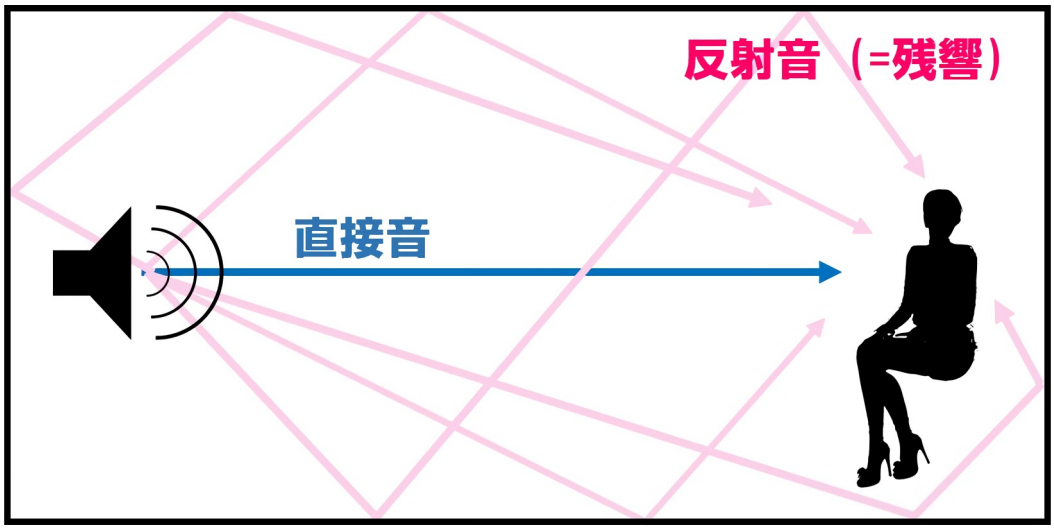


Figure 1　仮想オブジェクトの自然な質感再現の研究例。(a) 光源と光量を推定し仮想オブジェクトに自然な陰影をつける研究 (Debevec 1998)[5]。(b) 仮想オブジェクトの影を自然に生成する研究 (Nowrouzezahrai ら 2011)[6]。

(a) 音の空間伝搬



(b) 残響の時間波形構造

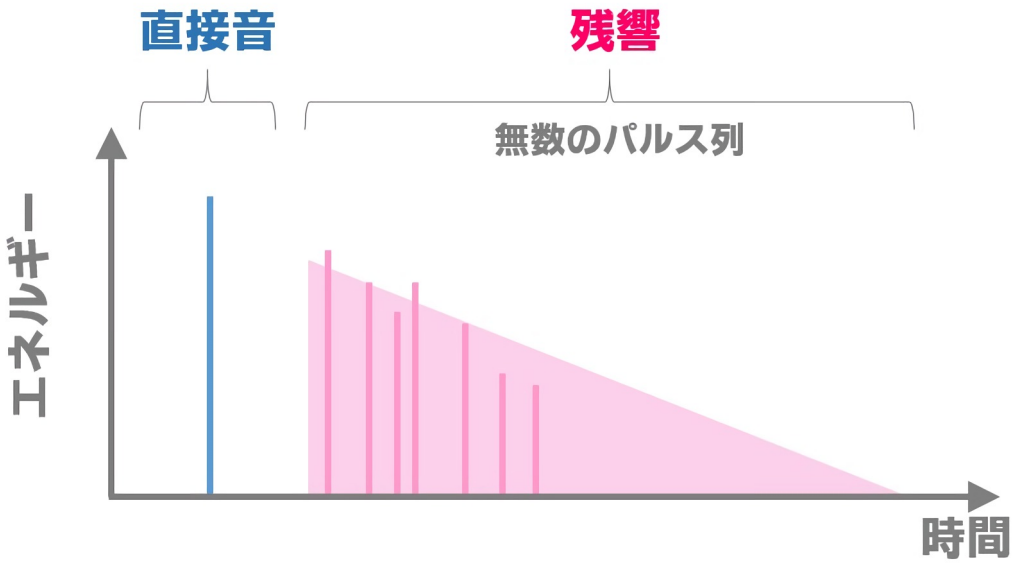


Figure 2　音の空間伝搬

想定されるアプリケーションと実現のための残響特性

想定されるアプリケーション

関連するアプリケーション例として、2015 年にソニーからリリースされた音声エンタメアプリ AELU(アエル) [9] がある。これはヘッドホンやイヤホンで体験するアプリケーションであり、キャラクターが周囲に居るような感覚を創り、さらには GPS の位置情報やユーザーの動きによってキャラクターがユーザーに話しかけたり、さりげなく動作をするというものである。通常のヘッドホンで音を聞くと、音源が頭の中にあるような感覚になる頭内定位が発生するが、立体音響技術である頭部伝達関数（Head-Related-Transfer-Function:HRTF）をヘッドホンで利用することで、キャラクターがあたかも 3 次元空間にいて、動き回っているような体験を実現することが可能である [10, 11]。

このアプリケーションはユーザーが常にヘッドホンやイヤホンをつけて生活すると、見えないキャラクター（本アプリ中では女の子の幽霊）が自分の周りにいて、ともに生活しているかのような体験を狙った音の AR/MR アプリである。

このアプリケーションではユーザーの位置を GPS から取得し、その情報を元にユーザーがいるであろうコンビニやデパートなどのランドマークに応じた会話や動作、独り言などがあるにも関わらず、その空間に適した音の響きが反映されていなかった。つまり、HRTF によってキャラクターが自分の周囲にいて、動きまわっているような位置的な感覚は創れたが、そのキャラクターと空間を共有しているような体験ができていなかった。

また、Figure 3 に示すように holoportation [4] は遠隔地にいる人がまさにこの場にいるかのような視覚体験を提供する MR アプリケーションである。今後、holoportation のような視覚体験と他の感覚刺激の再現とが組み合わされて発展すると、仮想的に作り出されたキャラクターや遠隔地の人とより一層リアルにインタラクション可能な MR アプリケーションが登場すると予想される。将来的に仮想的に再現される人や物体のリアルさだけではなく、それらがまさにこの場に存在していて、空間を共有しているような感覚をユーザーに提供できれば、より一層自然で、現実と仮想の見分けがつかない体験が可能となるだろう。つまり、環境に適合した感覚刺激を創り出すということは、仮想的な物体と現実空間との相互作用の実現であり、仮想的な物体との場 [12] の共有体験の実現をも意味する。よって、環境に適合した残響再現は、空間を共有しているような感覚を創り出す聴覚刺激として非常に重要である。

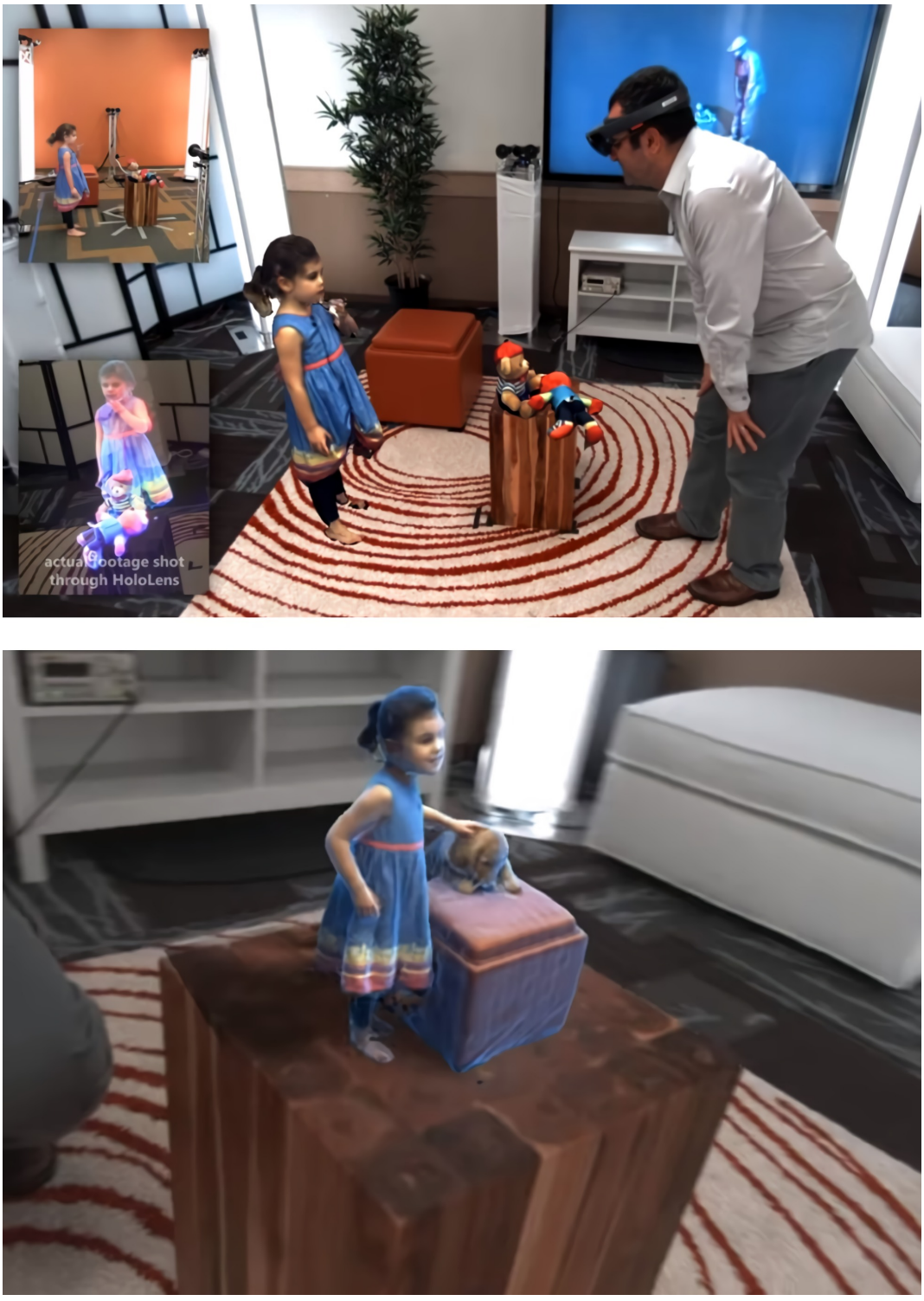


Figure 3　holoportation (Orts-Escolano ら 2016) [4]。遠隔地の子供がこの場にいるかのような体験ができる MR アプリケーション。

実現に必要な残響特性

Figure 4 の階層に示されるように、人が受ける音の印象や趣向は発する音自体の音色、音量感、複数音のアンサンブル、録音されたものであるならば、その環境や機材のノイズや歪など、さまざまな要因から影響を受ける [13]。その中でも残響は空間印象に影響を与えるものと位置づけられている。

そして、空間印象を構成する階層が Mason に示唆されており、Figure 5 に示されるように、大きく音源 (Source) の印象と環境 (Environment) の印象がある。音源側の構成要素としては、音源自体の Position（位置）、Dimensions(サイズ)、Focus/diffuseness（焦点・ぼやけ）が挙げられている。ここで Position は、直接音と残響音の比率によって音源との距離感が変化することや、音の到来方向により変化する人間の両耳の時間差や音圧差、または個人の身体的特徴から影響を受ける頭部伝達関数である [10, 11]。Dimensions は音源自体の大きさなどである。一方で環境印象の方は、Envelopment(音の包まれ感) や Dimensions(部屋の大きさ) などである。

下層にある Perceived dimensions は、見かけの音源幅 (apparent source width) [14] とも言われ、響きによって実際とは違った音源の大きさの印象を受ける。そしてこれは音源印象だけではなく環境印象の要因でもある Focus/diffuseness にも影響を与えているが、主に音源要素への影響が支配的であると言われている [15]。

究極的には残響特性の完全再現が理想ではあるが、想定するアプリケーションで目指す残響は、音源自体の印象ではなく、その場の空間を共有しているかのような環境的な印象の残響であると考えられる。そして、後部残響特性が環境印象に影響を与える [14, 16]。

以上から、本研究では残響特性の中でも後部残響特性の推定を目指す。

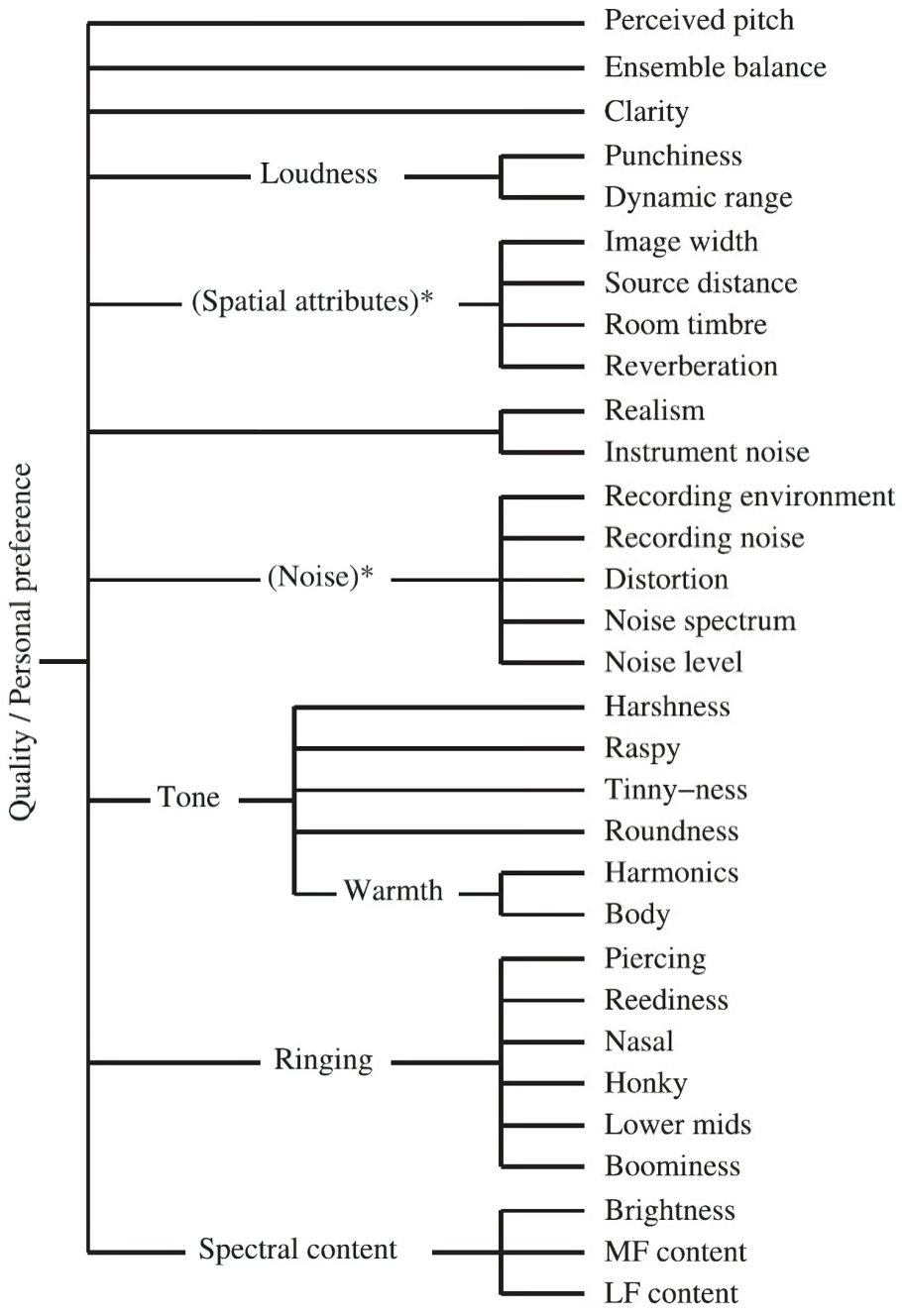


Figure 4　趣向に影響を与える音の要因と階層 [13]

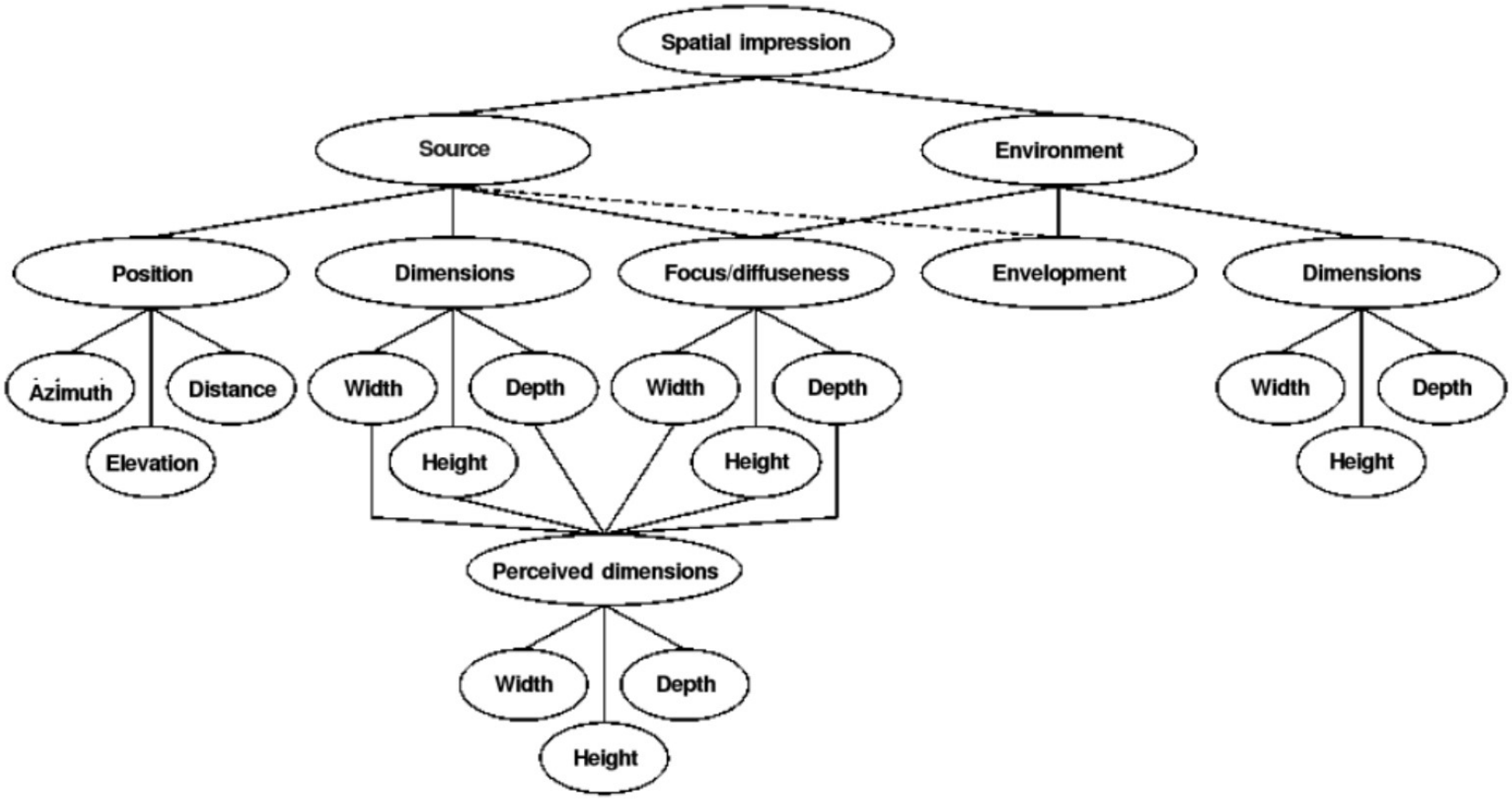


Figure 5　音の空間印象の階層 (Mason, 1999) [17]

関連研究と本研究の目的

残響に関する研究は 50 年以上の歴史があり、大きく 2 つの物理的アプローチ (Physically-Based Approach) と感覚的アプローチ (Perceptually-based Approach) がある [18, 7]。以下にそれぞれの特徴を挙げ、本研究の位置づけを述べる。

物理的アプローチ

物理的アプローチは音の反射を波形レベルで完全再現しようというアプローチである。まず、残響再現手法の基本的な処理手順は、再現したい空間の仮想音源位置にスピーカー、受聴位置にマイクロホンをそれぞれ設置する。そのスピーカーからノイズや時間引き延ばしパルス（Time Stretched Pulse, TSP）などの計測信号を出力し、マイクロホンでその信号を収録する。収録された信号をインパルス応答に変換し、得られたインパルス応答を有限インパルス応答（Finite Impulse Response, FIR) フィルタ の係数として仮想音源を畳み込みレンダリングする。このような手順を踏むことで、計測された空間で実際に音源を再生したかのような残響音が得られる [19]。この手法の再現精度は、マイクやスピーカー、アナログ・デジタル変換（AD 変換）などの装置のダイナミックレンジや S/N 比、計測時と再現時のサンプリングレート、そして畳み込み演算するフィルタとインパルス応答の長さなどで影響を受けるが、理想的な装置と計算資源を用いることができれば、理論上、完全に物理現象を再現できる手法である。

この手順が示すように、さまざまな空間の残響を再現しようとした場合、あらゆる空間のあらゆる再生位置、受聴位置での計測が必要となるため、計測に大きな労力を要する。そのため、実際に計測をせずに残響インパルス応答を取得しようとする音響シミュレーションの研究が注目されている。音響シミュレーションによる残響インパルス応答の取得には、コンサートホールなど残響を再現しようとする 3 次元空間を計算機上で構成し、壁・床・天井などの反射係数や吸音係数を境界条件として設定する。その後、反射経路を幾何学的に演算する音線法 (Ray-Tracing) [20, 11, 21, 22] や鏡像法 (Image Method)[23, 24, 25]、数値解析手法である時間領域差分法（Finite Difference Time Domain, FDTD） [26]、境界要素法 (Boundary Element Method, BEM) [27] などを用いて空間伝搬特性シミュレートし、残響特性としてのインパルス応答を得る (Figure 6)。この想定される 3 次元空間は手動で計算機上で設計されるか、想定される空間の複数の 2 次元画像から 3 次元再構成する技術 [28, 29] を用いる必要がある。この手法は 3 次元空間とその境界条件を完全に

再現することができたならば、理論上、実計測した残響特性であるインパルス応答と波形レベルで完全一致する。2017 年に Schissler らによって、3 次元形状だけではなく壁や床などの吸音率も画像から推定する Acoustic Classification 手法が提案され、より精度の高い残響波形を生成できるようになっている [30](Figure 7)。

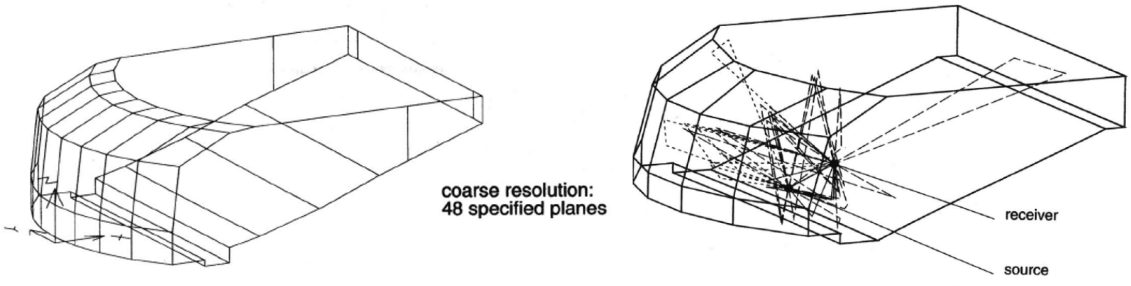


Figure 6　音線法 (Ray-Tracing) による残響特性シミュレーション [11]

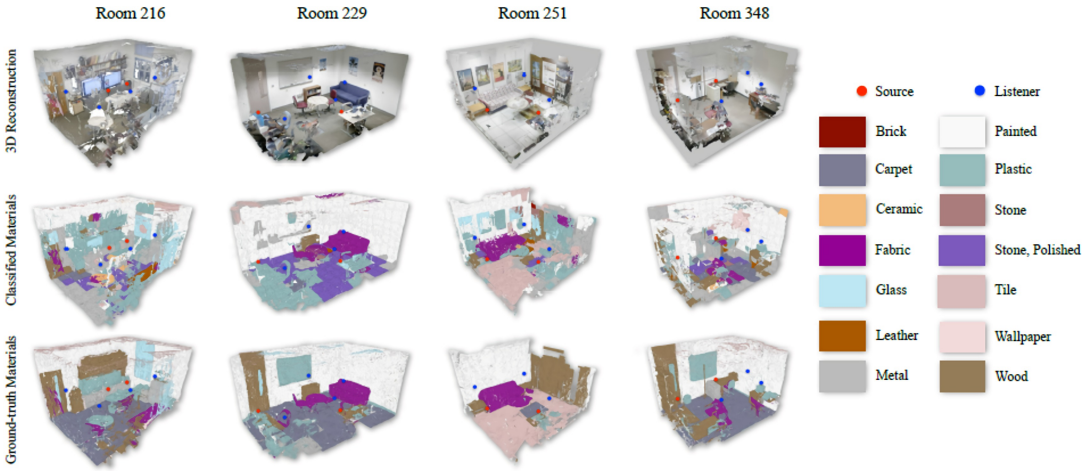


Figure 7　3 次元再構成と音響素材推定を用いた残響特性シミュレーション (Schissler ら 2017) [30]

感覚的アプローチ

残響のインパルス応答は無数の密なパルス列として表現されるが、これを物理的アプローチの信号処理である FIR フィルタで再現するには非常に多くの演算量とメモリ量が必要である。例えば、一般的な音質のサンプリングレートである 48kHz で 3 秒の残響を生成するためにはおよそ 14 万点のサンプリングデータが必要である。近年ではフーリエ変換を用いることで演算効率を上げたり、フィルタ処理で発生する時間遅延（群遅延）を小さくするような研究がなされている [31, 32] 一方で、残響の研究が始まった 1960 年ごろの計算機では困難であったため、人間の聴覚特性を考慮し、少ない演算量とメモリ量で自然な残響を作り出そうというのが感覚的アプローチの始まりである。

長い残響時間を再現するために時間経過とともに徐々に減衰する密なパルス列を生成しなくてはならず、FIR フィルタで再現するとフィルタと係数が大きくなってしまう。1962 年に Schroeder らはフィードバックループを含む Comb-Filter と Allpass-Filter を用いることで、少ない演算量とメモリ量で長時間の残響生成手法を提案し、Schroeder's Reverberator として知られる [33]（Figure 8）。

Schroeder の残響手法は平坦な周波数特性をもち、残響時間のみを制御していたが、Jot らはこれに加えて周波数特性を制御できる Feedback Delay Network(FDN) へと発展させた [34]（Figure 9）。所望の周波数特性を制御できるようになったため、単なる時間的な音の響きだけではなく部屋の印象をも制御できるようになったこの手法は現在でもホームシアターの残響アルゴリズムとして広く使われている。それ以降もパルス列をノイズに置き換えるなど、より一層少ない演算量で違和感のない残響信号処理手法が研究されている [35, 36]。

このように感覚的アプローチは物理的な波形レベルでは一致することはないが、人間の感覚に影響を与えている残響時間や周波数特性などの残響特性を再現することで、物理的アプローチで生成された残響と主観的に等価な体験を可能とする手法である。

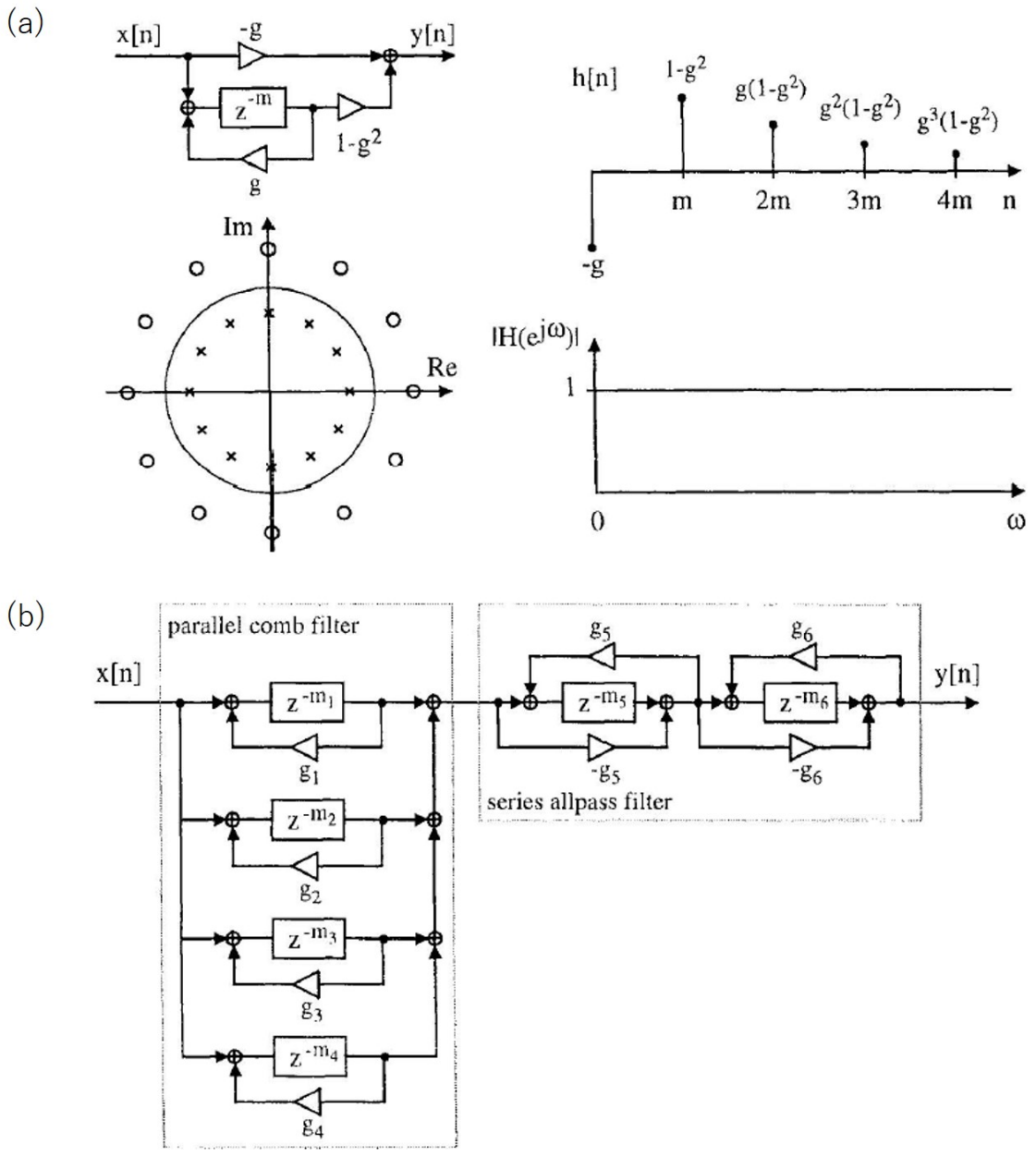


Figure 8　Schroeder's Reverberator (Schroeder ら 1961)[33]. (a)Allpass-Filter の特性。パルス列を生成し、平坦な周波数特性を有する。(b) 並列 Comb-Filter と直列 Allpass-Filter が組み合わされた Schroeder's Reverberator。

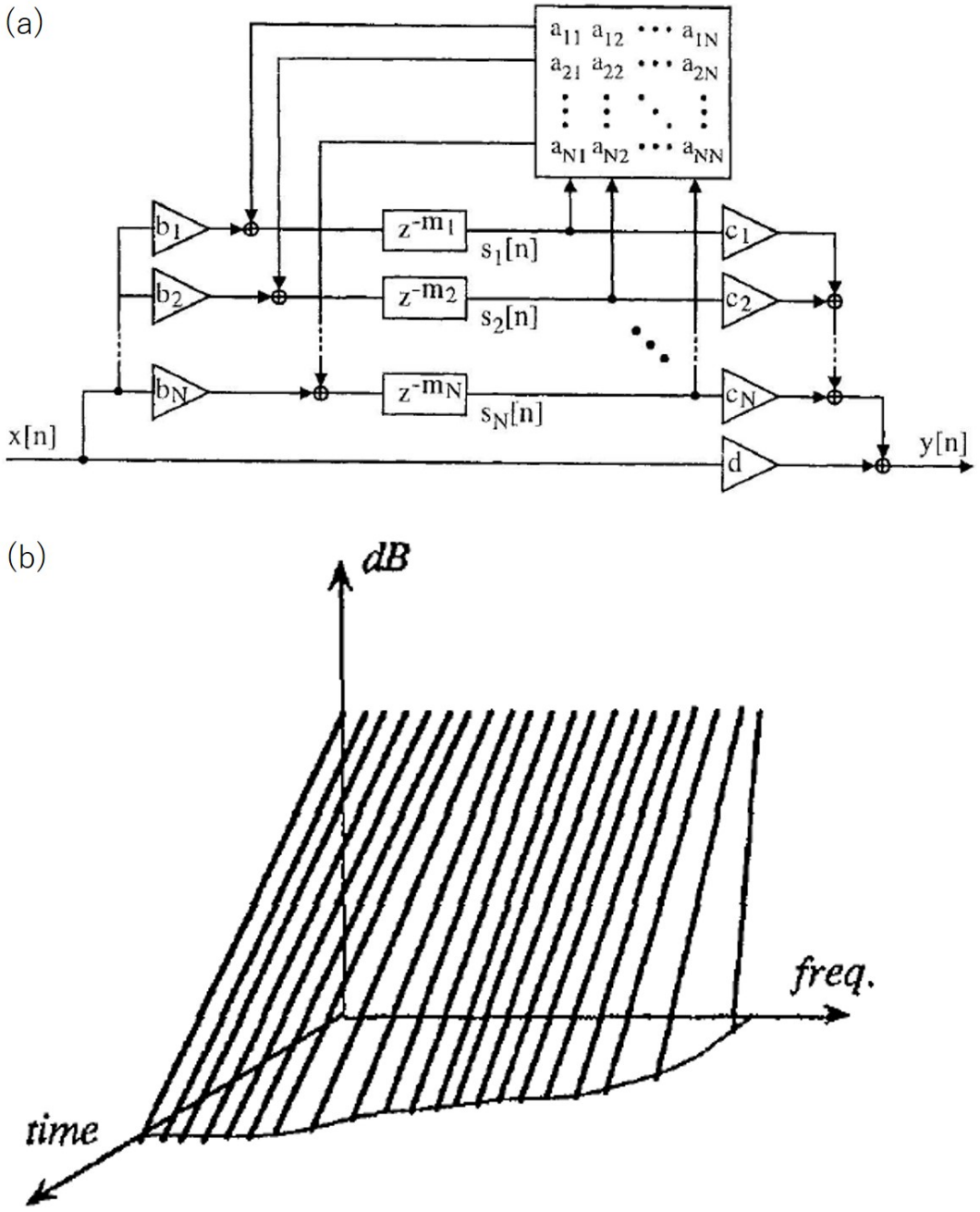


Figure 9　Feedback Delay Network(Jot ら 1991)[34].(a) 信号処理ブロックの例、(b) 生成された残響の時間周波数特性の例。

本研究の位置付けと目的

AR や MR に応用できる環境に適合した残響生成の既存手法としては、物理的アプローチの音響シミュレーションによる手法である。しかしながら、この手法は 2 次元画像からの 3 次元再構成と残響特性シミュレーションの演算量が非常に大きい。特に残響時間は一般的に Sabine の式 [37] に従い、残響時間を $RT60$、部屋の容積を V、吸音面積を A とすると、

$$RT_{60} \simeq 0.16 \cdot \frac{V}{A} \tag{0.0.1}$$

と示され、対象となる空間が大きくなるにつれて残響時間が長くなる。残響時間が長くなるということはシミュレーションの演算ステップも増えてしまうため、シミュレーションする空間の大きさに比例して演算量も増大してしまう。例えば、前出の Schissler らの Acoustic Classification 手法 [30] では、小さなリビングルーム程度の広さの残響を 1 箇所シミュレーションするのに、複数の GPU を用いても 6 時間以上かかってしまうという難点がある。AR や MR での利用を考えると、スマートフォンやウェアラブル機器の演算器では不十分である上、ユーザーがどの空間で利用するかもわからないため、クラウドコンピューティングを用いたと仮定してもリアルタイム性が損なわれ、ユーザビリティが確保できない。

具体的な場面を想定すると、前出の AELU のようなアプリケーションにおいて、各ユーザーが初めていく場所や空間でなにか仮想的なキャラクターとのコミュニケーションが発生すると仮定する。その際、スマートフォンや MR のウェアラブルディスプレイなどに装着されたカメラで空間画像が撮影できたとする。本来であれば、数秒程度で理想的な効果がある残響音が生成されることが望ましいが、それまでに数時間かかってしまうのでは遅すぎる。また、このようなアプリケーションは通常モバイル機器で利用されるため、演算量が少なければバッテリー消費を節約でき、ユーザーが長くアプリケーションを楽しむことができる。

一方で、演算結果をクラウドで蓄えておく手法なども考えられる。つまり、一度作られた残響特性を蓄えておけば、ある別のユーザーがその場所を訪れた際に再度演算する必要がない。しかし、世界には多くの人が訪れる場所からそうでない場所まで、様々な空間があり、多数のユーザーが利用するようなアプリであれば、はじめての場所が多数存在し、結局その都度残響特性の演算することになってしまうだろう。この場合は演算時間だけではなく、サーバーの演算負荷に加えて、運用コストにも影響を与えてしまうことが予想される。そのため、低演算処理であることもユーザーエクスペリエンスの観点とサービス提

供者側の観点からも共に重要である。

ところで、熟練したレコーディングエンジニアや音響エンジニアは、非常に聴覚に優れており、聴いた音や残響の周波数特性や残響時間を言い当てたりと人並外れた能力を有している者たちがいる。そして、彼らはその経験から、行ったことのない場所であっても、2 次元の写真を見ただけで空間の残響特性を推測することができる。これは人間の脳が視覚と聴覚情報を相互に想起できるように、画像から音の響きである残響特性を推測する何らかのモデルを脳内に構築できる可能性を示唆している。 Figure 10 のように、映像から音を想起する人間の脳活動の例も示されている [38]。 よって脳機能に見られる特性に類似した数理モデルであるニューラルネットワーク [39, 40] を用いることで、2 次元画像からの残響特性推定の感覚的アプローチ手法が実現ができると考えられる (Figure 11)。この手法が実現されると、空間の広さに演算量が左右されず、短時間演算で主観的に違和感のない残響特性を推定でき、様々なアプリケーションへの応用が期待できる。

以上を踏まえ、本研究では感覚的アプローチとしてニューラルネットワークを用いた残響特性の推定を行う。Figure 12 と Figure 13 にそれぞれ本研究の概念図と想定される信号処理ブロックダイアグラムを示し、Table 1 に本研究の概念図と位置づけを記す。

本論文概要と成果

以上に基づいて、本論文では、人間の脳が視覚から聴覚刺激を想起できることに着目した 2 次元画像から後部残響特性を推定するニューラルネットワークを提案し、実現への課題を明確にし、課題の解決手法を提案、実施するとともに客観的な側面と主観的な側面からの評価を行い、推定精度を向上させている。本研究は、AR/MR で空間に適合した後部残響をリアルタイムに再現する低演算手法を提案し，その効果を示したものであり、本研究成果は音を用いた複合現実のアプリケーションの高臨場感実現に貢献するものである。

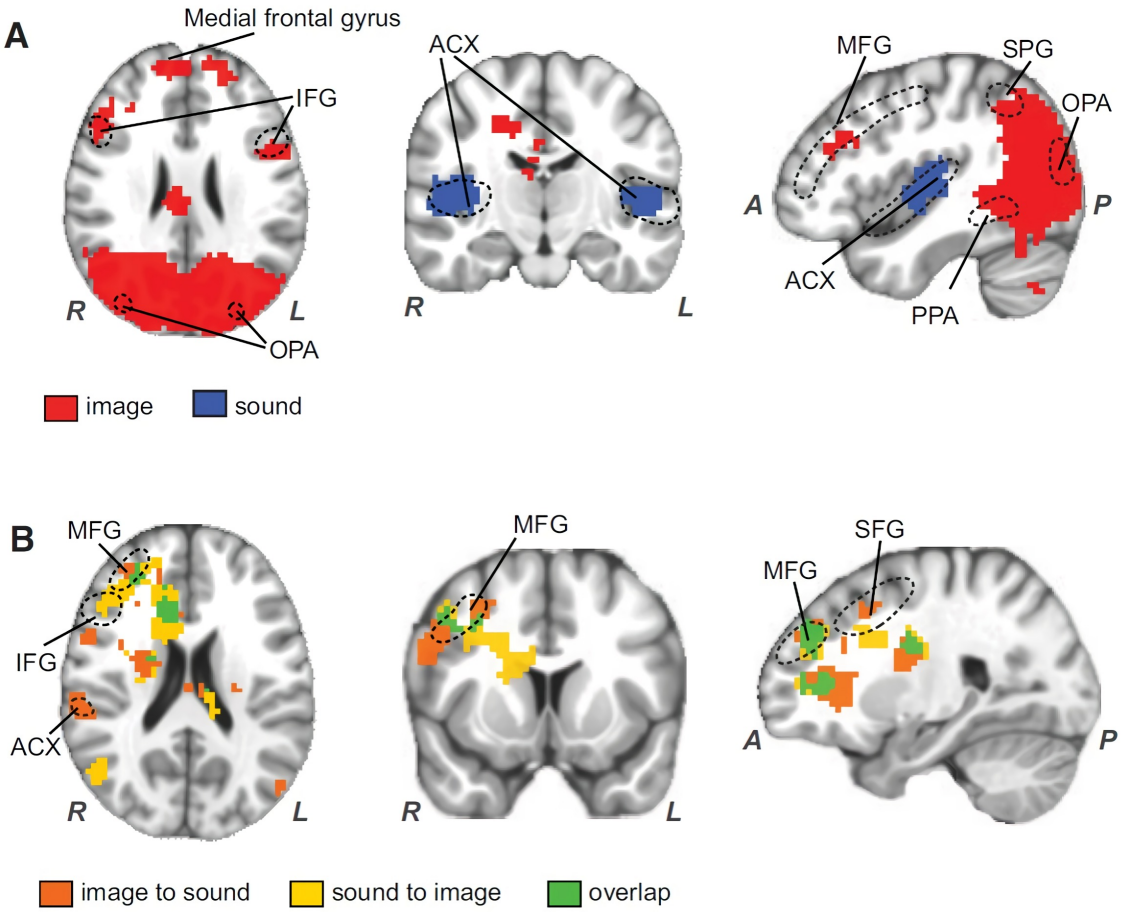


Figure 10　映像と音を相互に想起する脳活動の研究例 (Jung ら 2018) [38]。音から映像、映像から音を想起する脳活動の例を示している。

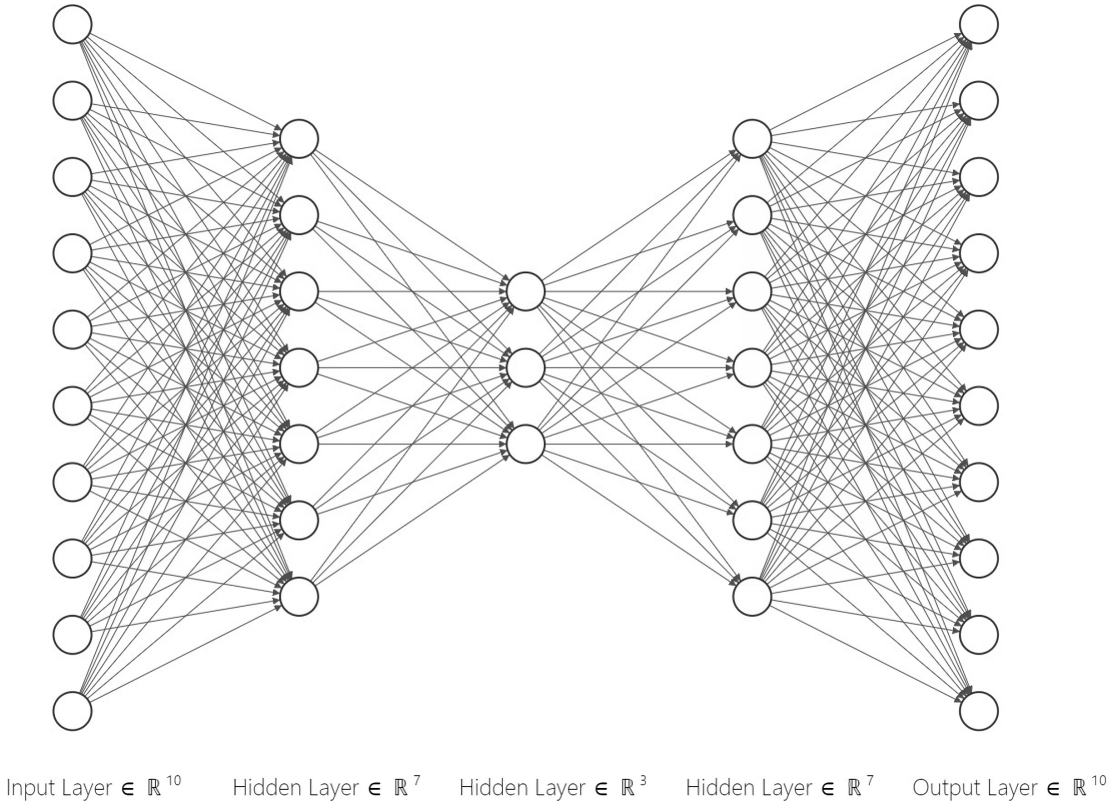


Figure 11　ニューラルネットワークの例 [41]

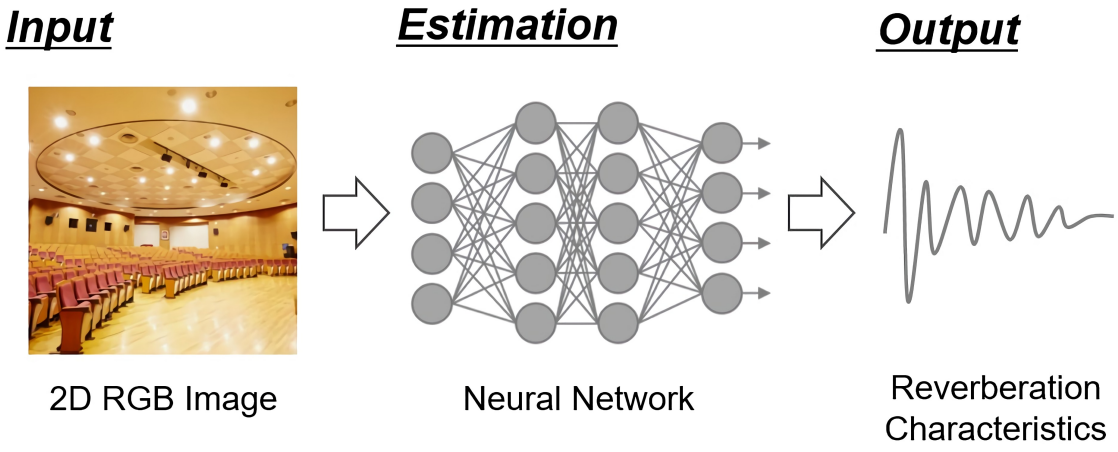


Figure 12 Concept of reverb estimation by perceptual approach

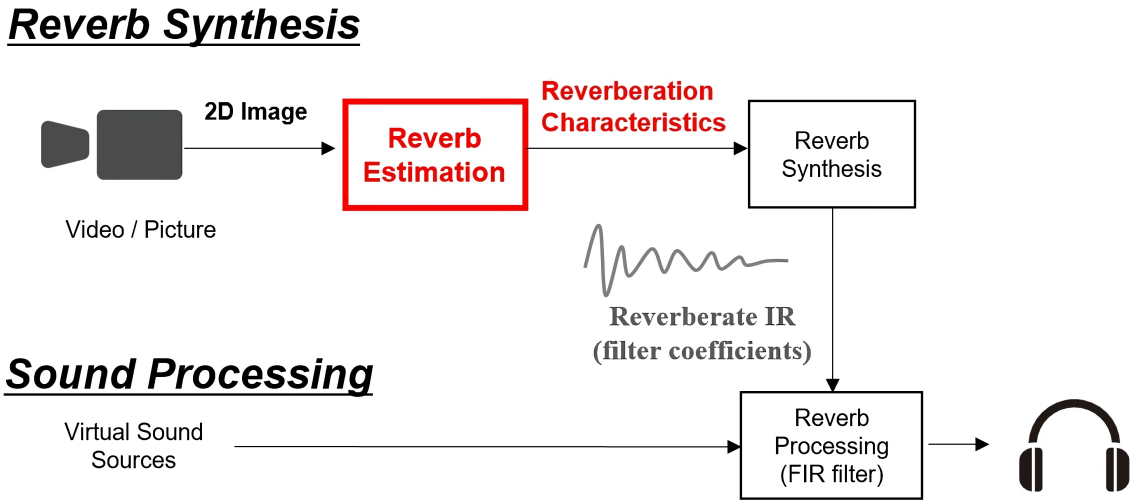


Figure 13 Block diagram

Table 1　物理的アプローチと感覚的アプローチからの本研究の位置づけ

Field / Approach	Phycically-Based	Perceptual-Based
Signal Processing Methods	- FIR Filter [19] - FFT Convolution [31] - Fast Zero-Latency [32] Convolution	- Schroeder Reverb [33] (Allpass/Comb) - Feedback Delay Network (FDN) [34]
IR / Parameter Simulation Methods	- 3D Reconstruction w/ Image Method)[23], FDTD [26], BEM [27], and Ray-Tracing [20, 11, 21, 22] - Acoustic Classification [30] w/ Ray-Tracing	**None : My Target**

参考文献

[1] Ronald T. Azuma. A survey of augmented reality. *Presence: Teleoper. Virtual Environ.*, 6(4):355–385, 1997.

[2] 日経ビジネス ON LINE. Mixed Reality から始まる産業革命. *https://special.nikkeibp.co.jp/atclh/NBO/17/microsoft0419/*, 2017.

[3] Magic Leap, Inc. Magic leap one. *https://www.magicleap.com/*, 2018.

[4] Sergio Orts-Escolano, Christoph Rhemann, Sean Fanello, Wayne Chang, Adarsh Kowdle, Yury Degtyarev, David Kim, Philip L. Davidson, Sameh Khamis, Mingsong Dou, Vladimir Tankovich, Charles Loop, Qin Cai, Philip A. Chou, Sarah Mennicken, Julien Valentin, Vivek Pradeep, Shenlong Wang, Sing Bing Kang, Pushmeet Kohli, Yuliya Lutchyn, Cem Keskin, and Shahram Izadi. Holoportation: Virtual 3d teleportation in real-time. *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*, pages 741–754, 2016.

[5] Paul Debevec. Rendering synthetic objects into real scenes: bridging traditional and image-based graphics with global illumination and high dynamic range photography. *Proceedings of the 25th annual conference on Computer graphics and interactive techniques*, pages 189–198, 1998.

[6] Derek Nowrouzezahrai, Stefan Geiger, Kenny Mitchell, Robert Sumner, Wojciech Jarosz, and Markus Gross. Light factorization for mixed-frequency shadows in augmented reality. *2011 10th IEEE International Symposium on Mixed and Augmented Reality*, pages 173–179, 2011.

[7] Vesa. Välimäki, Julian D. Parker, Lauri. Savioja, Julius O. Smith, and Jonathan S. Abel. Fifty years of artificial reverberation. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(5):1421–1448, 2012.

[8] Jean-Marc Jot and Keun Sup Lee. Augmented reality headphone environment

rendering. *Audio Engineering Society Conference: 2016 AES International Conference on Audio for Virtual and Augmented Reality*, 2016.

[9] Sony corporation. 新感覚音声エンタメ「AELU （アエル） β version」. *https://twitter.com/aelu_official*.

[10] Bosun Xie. *Head-Related Transfer Function and Virtual Auditory Display (Second Edition)*. J. Ross Publishing, 2013.

[11] Jens Blauert and Robert Butler. Spatial hearing: The psychophysics of human sound localization by jens blauert. *Journal of The Acoustical Society of America - J ACOUST SOC AMER*, 77:334–335, 1985.

[12] 久米是志 三宅美博 清水博, 三輪敬之. **場と共創**. NTT 出版, 2000.

[13] Andy Pearce, Tim Brookes, Martin Dewhirst, and Russell Mason. Eliciting the most prominent perceived differences between microphones. *The Journal of the Acoustical Society of America*, 139:2970–2981, 2016.

[14] Michael Barron and A. H. Marshall. Spatial impression due to early lateral reflections in concert halls: The derivation of a physical measure. *Journal of Sound and Vibration*, 77(2):211–232, 1981.

[15] J. S. Bradley, R. D. Reich, and S. G. Norcross. On the combined effects of early- and late-arriving sound on spatial impression in concert halls. *The Journal of the Acoustical Society of America*, 108(2):651–661, 2000.

[16] Jens Ahrens. *Analytic Methods of Sound Field Synthesis*. T-Labs Series in Telecommunication Services. Springer, 2012.

[17] Francis Rumsey. *Spatial Audio*. Focal Press, 2012.

[18] Vesa Välimäki, Julian Parker, Lauri Savioja, Julius O. Smith, and Jonathan Abel. More than 50 years of artificial reverberation. *Audio Engineering Society Conference: 60th International Conference: DREAMS (Dereverberation and Reverberation of Audio, Music, and Speech)*, 2016.

[19] William G. Gardner. *Reverberation Algorithms*, pages 85–131. Springer US, Boston, MA, 2002.

[20] A. Krokstad, S. Strom, and S. Sørsdal. Calculating the acoustical room response by the use of a ray tracing technique. *Journal of Sound and Vibration*, 8(1):118–125, 1968.

[21] Niklas Röber, Ulrich Kaminski, and Maic Masuch. Ray acoustics using computer graphics technology. *Proceedings of the 10th International Conference on Digital*

Audio Effects (DAFx-07), 2007.

[22] Andrzej Kulowski. Algorithmic representation of the ray tracing technique. *Applied Acoustics*, 18(6):449–469, 1985.

[23] Jont B. Allen and David A. Berkley. Image method for efficiently simulating small - room acoustics. *The Journal of the Acoustical Society of America*, 65(4):943–950, 1979.

[24] James A Moorer. About this reverberation business. *Computer music journal*, pages 13–28, 1979.

[25] Jeffrey Borish. Extension of the image model to arbitrary polyhedra. *The Journal of the Acoustical Society of America*, 75(6):1827–1827, 1984.

[26] Abdelghani Gramez, Jean-Louis Guyader, and Boubenider Fouad. Modeling and simulation of the sound propagation by the fdtd method. 2012.

[27] Stephen Kirkup. *The Boundary Element Method in Acoustics*, volume 8. 2007.

[28] Alex Flint, David Murray, and Ian Reid. Manhattan scene understanding using monocular, stereo, and 3d features. In *2011 International Conference on Computer Vision*, pages 2228–2235.

[29] Hansung Kim, Luca Remaggi, Philip J. B. Jackson, Filippo Maria Fazi, and Adrian Hilton. 3d room geometry reconstruction using audio-visual sensors. pages 621–629, 2017.

[30] Carl Schissler, Christian Loftin, and Dinesh Manocha. Acoustic classification and optimization for multi-modal rendering of real-world scenes. *IEEE Transactions on Visualization and Computer Graphics*, PP(99):1–1, 2017.

[31] Alan V. Oppenheim and Ronald W. Schafer. *Discrete-time Signal Processing*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1989.

[32] William G. Gardner. Efficient convolution without input/output delay. In *Audio Engineering Society Convention 97*, Nov 1994.

[33] Manfred R. Schroeder and Benjamin F. Logan. 'colorless' artificial reverberation. *J. Audio Eng. Soc*, 9(3):192–197, 1961.

[34] Jean-Marc Jot and Antoine Chaigne. Digital delay networks for designing artificial reverberators. *90th Conv. Audio Engineering Society*, 1991.

[35] Keun-Sup Lee, Jonathan S. Abel, Vesa Välimäki, Timothy Stilson, and David P. Berners. The switched convolution reverberator. *J. Audio Eng. Soc*, 60(4):227–236, 2012.

[36] Vesa Välimäki, H. M. Lehtonen, and M. Takanen. A perceptual study on velvet noise and its variants at different pulse densities. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(7):1481–1488, 2013.

[37] Wallace Clement Sabine. *Collected papers on acoustics*. Harvard University Press, Cambridge, 1922.

[38] Yaelan Jung, Bart Larsen, and Dirk B. Walther. Modality-independent coding of scene categories in prefrontal cortex. *The Journal of Neuroscience*, 38(26):5969–5981, 2018.

[39] Md Zahangir Alom, Tarek M. Taha, Christopher Yakopcic, Stefan Westberg, Paheding Sidike, Mst Shamima Nasrin, Brian C Van Esesn, Abdul A S. Awwal, and Vijayan K. Asari. The history began from alexnet: A comprehensive survey on deep learning approaches, 2018.

[40] Mahbubul Alam, Manar D. Samad, Lasitha Vidyaratne, Alexander Glandon, and Khan M. Iftekharuddin. Survey on deep neural networks in speech and vision systems, 2019.

[41] Alexander LeNail. Nn-svg: Publication-ready neural network architecture schematics. *Journal of Open Source Software*, 4:747, 2019.