Adaptive Variational AutoEncoder による 3 次元音響の特定個人への知覚ベースキャリブレーション

山本 和彦 *† 五十嵐 健夫 *

概要. VR などに代表される 3 次元映像が一般的となるなか, 音も上下左右あらゆる方向に音を定位させる 3 次元音響を実現することへの要求が高まっている. 左右 2 チャンネルのヘッドフォンでこうした 3 次元音響を実現するためには頭部伝達関数 (Human Related Transfer Function, HRTF) が必要となる. しかし, HRTF は頭や耳の形状による個人差が大きく, 基本的にはユーザ毎に無響室で特別な機材を使って長時間の測定を行わなければ得ることができないという問題がある. そこで, 本稿ではユーザ 1 人 1 人に合った HRTF を得るためのコストを大幅に低減させるために, ユーザの知覚的なフィードバックのみを利用した HRTF のキャリブレーション手法を提案する. 本システムでは, ユーザはシステムによって生成された 2 つの異なる HRTF によるテスト音の “どちらの定位感がどれくらい良いか” を一対比較することを繰り返す. このユーザからのフィードバックを利用することで, システムは内部で個人化パラメータを最適化し, 次第にユーザに合った HRTF を生成できるようになる. 本手法では, ユーザに対して特別な機材や環境を要求することがなく, 実測定と比較して大幅に少ない負担での HRTF のキャリブレーションが可能である. 本システムは事前に一般に公開されている HRTF 実測データセットから個人性を抽出して利用することのできる Adaptive Variational AutoEncoder によって実現されている. 本稿ではユーザスタディや交差検定などの実験によって提案手法の有効性を示す.

1 はじめに

VR やゲームなどに代表される多くのデジタルコンテンツでは, 音の再生に左右 2 チャンネルのヘッドフォンを使用するのが一般的である. こうしたヘッドフォンでは物理的には左右の耳元でしか音を鳴らすことができない. しかし, 3 次元の映像表現が普及するなか, 音も上下左右あらゆる方向に音を定位させる 3 次元音響への要求が高まっている.

ヘッドフォンで 3 次元音響を実現するためには, 物理的には左右の耳元で鳴っている音を, 他の方向から到来した音だと錯覚させるための, 頭部伝達関数 (Human Related Transfer Function, HRTF) が必要となる. これは, 音が両耳に対して到来する方向に応じて, 左右の耳それぞれへの音の到達時間差が生まれたり, 頭の形状や耳の形状による遮蔽や回折によって音のスペクトルが変化する情報を利用して我々の聴覚が音の到来方向を知覚していることを逆に利用したものである. 具体的には, こうした音の方向による特性を再現した, 音の到来方向ごとに左右 2 チャンネル分用意される Finite Impluse Response フィルターが HRTF となる. 実際の使用時には音が到来したと知覚させたい方向に応じてこのフィルターを切り替えてやることでヘッドフォンでもあらゆる方向の音を再現することが可能となる.

しかし, HRTF は個人個人の頭や耳の形状に強く依存し, 非常に個人差が大きい. そのため, 基本的には他人のものを使い回すことができず, ユーザごとに測定をおこなわなくてはいけない. この測定は無響室で特別な機材を使って何時間も要するものであり, 非常にユーザに対する負荷が大きく, 一般のユーザが自分に合った HRTF を利用することは困難であるという問題がある. この問題は, 3 次元の音のレンダリングが映像のレンダリングに比べて遅れている大きな理由にもなっている.

そこで, 本稿ではユーザ 1 人 1 人に合った HRTF を得るためのコストを大幅に低減させるために, ユーザの知覚的なフィードバックのみを利用した HRTF のキャリブレーションを提案する (図 1). 提案手法では, ユーザはシステムによって生成された 2 つの異なる HRTF によるテスト音の “どちらの定位感がどれくらい良いか” を一対比較することを繰り返す. このユーザからのフィードバックを利用することで, システムは内部で個人化パラメータを最適化し, 次第にユーザに合った HRTF を生成できるようになる. 本手法では, ユーザに対して特別な機材や環境を要求することがなく, 実測定と比較して大幅に負担が少ない HRTF のキャリブレーションが可能である.

提案手法は, 事前に一般に公開されている HRTF 実測データセットから非線形空間上で個人性を抽出して利用することのできる新しい Adaptive Variational AutoEncoder (AVAE) によって実現され

* 東京大学
† ヤマハ株式会社

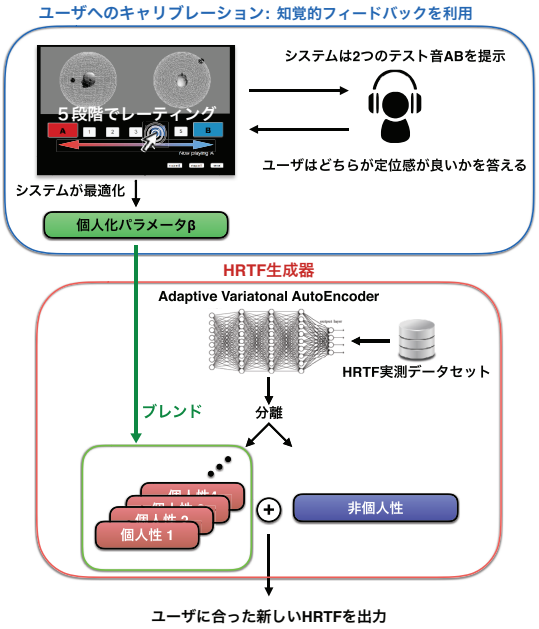

図 1. 提案システム．ユーザはシステムから提示された音を 2 つずつ聞いて一対比較してフィードバックすることを繰り返す．その情報をもとにシステムは事前抽出された学習データセットの個人性をブレンドして，特定ユーザに合った新しい **HRTF** を出力する．

ている．AVAE では，事前に抽出したデータセットの “個人性” をブレンドして，新たな HRTF をデータセットに含まれない特定ユーザのために生成することができる．また，キャリブレーション時にユーザに要求するフィードバック回数を減らすために，CMA-ES [4] のアルゴリズムの内部にガウス過程回帰 [13] を導入する．これにより，サンプリングベースの最適化手法である CMA-ES に勾配の情報を支援的に使うことで高速化することができる．

本稿では，提案手法を検証するために，3 種類の評価実験をおこなった．まず，交差検定では，提案手法がデータセットに含まれない新しいユーザのための HRTF を生成できることを検証し，さらに数値シミュレーションによる実験では，提案手法が 3 次元の音場を適切に予測できることを示す．最後に 20 名の被験者実験によって，実際に比較タスクのみでの特定ユーザへの HRTF のキャリブレーションが可能であることを示す．

本稿の内容は国際会議で採択された内容 [17] を簡潔にまとめたものである．

2 関連研究

測定による手法: ユーザに合った HRTF を得るための直接的な方法は，無響室で音響測定をおこなうことである．測定時，被験者は両耳の中に小さなマイクロフォンプローブをいれ，頭を中心に周囲に球形に設置されたスピーカーアレイの中心に座る．

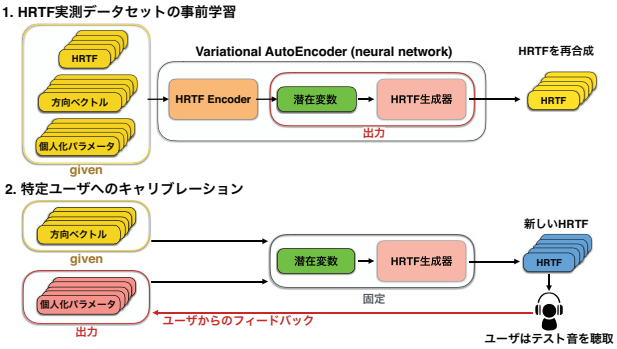

図 2. アルゴリズム概要．まず，**HRTF** 実測データセットの事前学習をおこない，**HRTF** をニューラルネットワークで生成できるようにする．ユーザに対するキャリブレーション時には，ニューラルネットワークで生成された **HRTF** を使ったテスト信号をユーザに聞かせ，知覚的なフィードバックを得る．システムはこの情報を利用して内部の個人化パラメータを最適化し，ユーザに合った **HRTF** を生成できるようにする．

被験者には頭を動かさないようにしてもらい，各スピーカーから 1 方向ずつ，総計約数千方向からテスト音 (インパルス音やスイープ音) を鳴らし，その音を耳にいれたマイクで収録する．本稿で HRTF の事前学習に使ったデータセット [1] は実際にこうして集音されたものであり，1 人につき 1404 方向，45 人分の HRTF が収められている．

数値シミュレーションによる手法: HRTF を得るために，無響室での測定を避けるための方法としては，数値シミュレーションを利用するものがある．これらの手法では，3D スキャンしたユーザの頭と耳のメッシュを使い，境界要素法 [6][8] や時間領域差分法 [16] で音響シミュレーションをおこなうことで HRTF を得る．しかし，こうした手法には，1: 一般的な PC では 10 時間 ～ 数日の長時間の計算時間がかかる，2: 高精細な 3D スキャンをユーザ自身がおこなうことが難しい，という問題がある．DeepEarNet [7] では 3D スキャンを避けつつ数値シミュレーションで HRTF を得るために，ユーザの耳を 2 方向から撮影した写真から畳み込みニューラルネットワークを使って，耳の 3 次元形状を推定している．しかし，この手法では高周波数域の HRTF を得るために十分な精度の耳形状までは推定できないという問題がある．

機械学習による手法: Zotkin ら [18] は複数人のデータセットの中から最もユーザにフィットする HRTF を探し出すために，耳の形状パラメータのユークリッド距離が最も近いものを採用した．しかし，この手法では，選ばれた HRTF が本当にそのユーザに合っているという保証は無い．Holzl [5] は HRTF 実測データセットを主成分分析し，合成音をユーザに聞かせながら，各主成分をユーザ自身に直

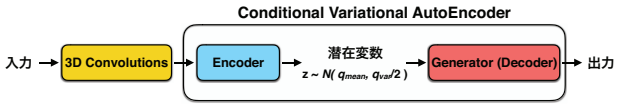

図 3. 使用するニューラルネットワークでは，通常の **AutoEncoder** と同様に入力と同じデータを潜在変数を介して復元できるように学習される．

接スライダーで直接調整させることで HRTF を得ることを試みた．しかし，各主成分を直接調整することは直感的ではなく，非常に難しいという問題がある．Luo ら [12] は人間の聴覚システムを，入力された音に対して知覚方向を返すブラックボックス関数として仮定し，ガウス回帰モデルでモデリングすることで “virtual” なユーザをつくりだし，HRTF をその “virtual” なユーザに対して AutoEncoder を使うことでフィッティングした．しかし，彼らの手法を実際の人間のユーザに適用するための手段は述べられていない．機械学習を利用して HRTF を新しいユーザに対して最適化する主要なアプローチとしては，低次元の耳の形状パラメータから HRTF を回帰する手法がいくつか提案されている [3]．しかしこれらの手法で使用されている低次元の耳形状パラメータでは本来の耳形状に対して情報量が少なすぎるため，HRTF を十分な精度で回帰することは難しいという問題がある．

3 アルゴリズム概要

提案手法は，HRTF 生成器の事前学習と特定ユーザへのキャリブレーションという 2 段階から成り立つ (図 2)．まず，事前学習では，HRTF 生成器を一般に公開されている HRTF 実測データセット [1] で機械学習する．この HRTF 生成器は Variational AutoEncoder [9][14] を拡張したニューラルネットワークの生成モデル (AVAE) であり，HRTF のためにデザインされた 3 次元コンボリューションレイヤーと，学習データセットから個人性と非個人性を分離できる適応レイヤーという拡張がなされている (図 3)．AVAE では通常の AutoEncoder と同様に，大きく Encoder と Decoder の 2 つのブロックから成り立っており，事前学習時には Encoder に，左右 2 チャンネルの HRTF のパワースペクトルと位相ベクトル，対応する方向ベクトルを入力し，潜在変数 z を抽出する．この潜在変数と方向ベクトルを Decoder に入力することで再び Encoder に入力された元の HRTF を再合成する．一方，キャリブレーション時には Decoder 部分だけを使用し，潜在変数と方向ベクトルから新しい HRTF を生成する．さらに，Encoder，Decoder 両方を通して共通の入力パラメータとして個人化パラメータというベクトルを持つ．これは抽出された個人性のブレンド量を調整して新たな HRTF を生成するためのものである．

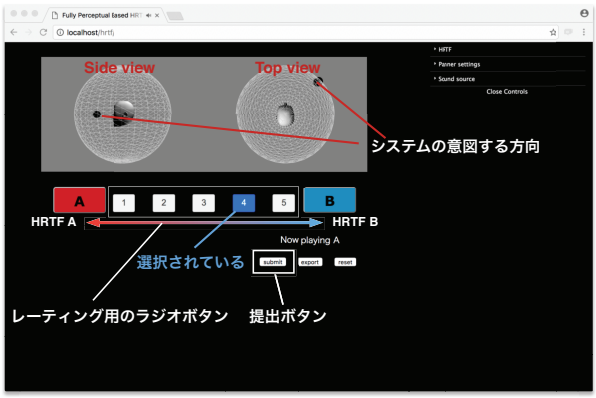

図 4. ユーザインタフェース画面．赤青の **AB** ボタンを押すと，それぞれシステムが生成した異なる **HRTF** によるテスト音が再生される．システムが意図している音の方向は頭の **3DCG** の周囲で回転する球体として可視化されている．ユーザは **A** と **B** の信号のどちらが定位感が良いか，を **5** 段階で評価し，ラジオボタンでシステムにフィードバックする．

次に，キャリブレーション時には，HRTF 生成器における個人化パラメータを特定ユーザに対して最適化して，そのユーザに合った HRTF を生成できるようにする．最適化は，HRTF 生成器で生成した 2 つの異なる HRTF によるテスト信号を，ユーザに一対比較させることを繰り返して得られた情報を使うことによりおこなう．キャリブレーション結果は任意の HRTF ファイルフォーマットで書き出され，ゲームエンジンなどで 3 次元音響を実現するために使うことができる．

4 ユーザインタフェース

図 4 がユーザが自分のために HRTF のキャリブレーションをおこなうための画面である．赤と青の AB ボタンを押すことにより，それぞれシステムが生成した異なる HRTF を使ったテスト音が再生される．提示される音はどちらも頭の周りの水平方向 360 度，垂直方向 360 度を交互に回り続けるように聞こえるように設定されており，システムが意図する方向は頭の 3D グラフィックスの周りを周回移動する球体で表現されている．ユーザはそれぞれの音を聞き，どちらのほうが定位感が良いか，を例えば，UI 画面ラジオボタンの “1” ならば A のほうが明らかによく，“5” ならば B のほうが明らかによく，“3” ならば違いが判別できない，といった具合に 5 段階で評価する．ユーザがこの作業を繰り返すうちにシステムは内部で個人パラメータの最適化をおこない，次第にこのユーザに合った HRTF を生成できるようになっていく．ユーザが A と B の音の違いを全く判別できなくなった時点でキャリブレーションは終了する．実験では，100~200 回ほど繰り返したところで終了となり，20~30 分程度要した．

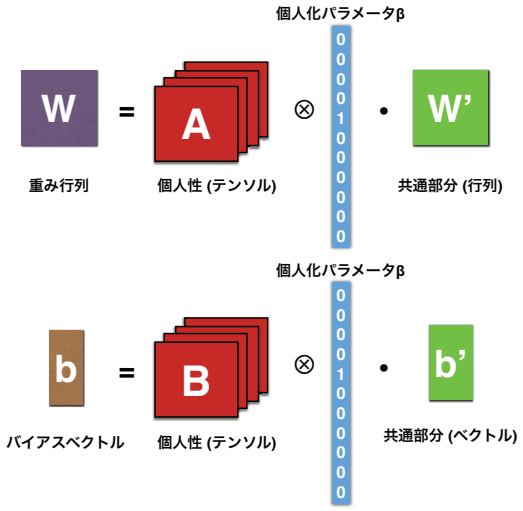

図 5. 適応レイヤー．重み行列とバイアスベクトルを個人化パラメータをはさんで個人依存の特徴量を表すテンソルとそれ以外の共通部分に分解する．

5 HRTF 生成器の事前学習

第 3 章で述べたように，提案手法ではまず，一般に公開されている HRTF 実測データセット [1] を学習データとして，HRTF 生成器 AVAE が HRTF を再合成できるように機械学習する．ここではニューラルネットワークの各層において，HRTF の中の個人に依存する特徴と，個人に依存しない共通した特徴を分離することをおこなう．通常ニューラルネットワークの各層は線形関数と非線形関数の合成関数となっている．

$$y = f(\boldsymbol{W} \cdot x + b). \tag{1}$$

ここで，$f()$ は任意の非線形関数，$\mathbf{W}$, b は重み行列とバイアスベクトルである．AVAE では各層のこの重み行列とバイアスを，

$$y = f(\boldsymbol{A} \otimes_3 \beta \cdot \boldsymbol{W}' \cdot x + \boldsymbol{B} \otimes_3 \beta \cdot b'), \tag{2}$$

のようにそれぞれ分解する (図 5)．ここで，β は個人化パラメータ (ベクトル)，$\boldsymbol{A}$ と $\boldsymbol{B}$ はテンソル，$\boldsymbol{W}'$ は行列，b' はベクトルである．また，$\otimes_d$ は d モードを展開したテンソル積である．事前学習時，β は学習データセットに含まれる被験者の人数と同じ次元の one-hot ベクトルとなり，現在 Encoder に入力されている被験者を表す要素だけが 1 で他の要素が全て 0 となる．つまり，新たな学習データが入力されるたび，β がスイッチのように切り替わり，テンソル $\boldsymbol{A}$, $\boldsymbol{B}$ の対応する一部分だけで学習がおこなわれる．これを確率的最適化手法 [10] と組み合わせることで，各テンソルにデータセットの個人依存の特徴が分離されていく．

6 キャリブレーション

事前学習後に実際に特定のユーザに対してキャリブレーションをおこなうときには，式 2 の個人化パラメータ β は one-hot ベクトルではなく，複数の要素に連続値をとるベクトルとなる．この β は個人性のブレンド量としての役割をはたし，キャリブレーション時にはこの β をユーザに対して最適化することによって，ユーザに合った新たな HRTF を生成する．これは，ユーザのもつ個人性が，学習データセットに含まれていた多数の個人性のブレンドとして表現できるとモデル化したことに相当する．キャリブレーションでは生成される HRTF のユーザによる絶対評価値が最大となるように β の最適化をおこなう．

この β の最適化はサンプリングベースの最適化手法である CMA-ES [4] と勾配ベースの最適化手法である準ニュートン法のハイブリッド法 [2] でおこなう．流れとしては，まず CMA-ES はある分布から N 個の異なる β をサンプリングし，これを使ってシステムは N 個の HRTF を生成する．ユーザへはこの HRTF をテスト信号と畳み込んだ音を図 4 の AB ボタンに割り当てて，2 つずつ $N/2$ 回提示する．第 4 章で述べた方法でユーザがこの $N/2$ 回の相対的なレーティングを終えると，次にシステムはこの相対スコアから N 個の β の評価値の絶対値を計算する．この絶対値の算出は Koyama ら [11] とほぼ同様の手法でおこなう [17]．計算された β の絶対評価値から，システムはガウス回帰過程 [13] によって局所的な評価関数の形状を推定し，勾配を求める．この勾配情報を使って準ニュートン法で β の局所解への最適化をおこなったのち，CMA-ES の分布を更新し，再び β を N 個サンプリングする，ということを繰り返す．

7 評価

7.1 交差検定

学習データセットの中から一人分のデータだけ抜き出したうえで HRTF 生成器を学習し，取り出したデータを回帰で精度よく再現できるかを評価した．これは，データセットの中の個人性のブレンドで新しい個人のデータをつくることができる，という仮説を実証するための検証である．図 6 が結果で，比較対象として主成分分析 (PCA) での結果も併せて示す．全ての角度において提案手法は PCA より誤差が小さい結果が出ており，高い精度で回帰が可能であることが示された．

7.2 シミュレーションでの実験

数値シミュレーションを使った実験をおこなった．頭の形を模した簡単な楕円球の障害物形状 1~4 を 4 種類用意し，その周囲でインパルス音を鳴らしたと

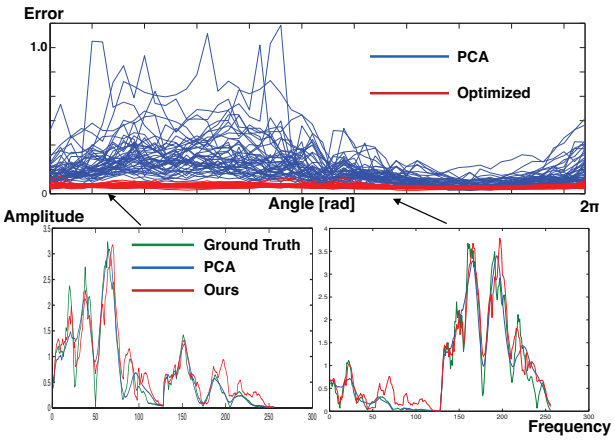

図 6. 交差検定の結果．上段：横軸が水平面上の方向 (角度)，縦軸が誤差である．青が PCA での回帰結果，赤が提案手法の結果を示している．下段：2 方向における HRTF パワースペクトルの比較．

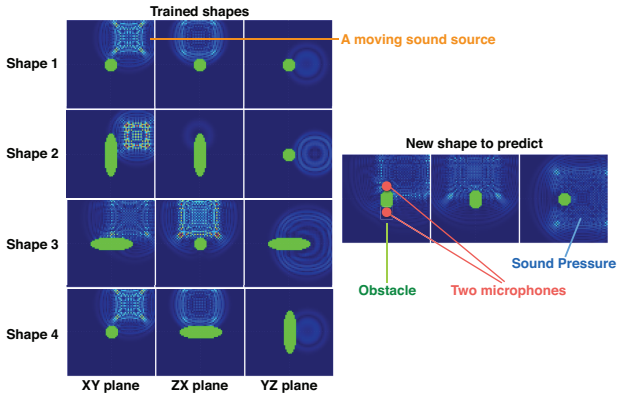

図 7. シミュレーション実験条件．左図が事前学習に使った形状である．緑の球体が障害物であり，その両端で音圧を測定した (右図赤丸)．

きの障害物両端の音圧を, 時間領域差分法で 3 次元音響数値シミュレーションし, 実測データセットの代わりに提案手法の学習データとして事前学習させた 図 7. 次に形状 1 と形状 2 のちょうど中間の障害物形状 5 を用意し, この形状 5 でのシミュレーション結果を提案手法でどのくらい予測できるかを検証した．図 8 に各方向における誤差の結果を示す．比較対象として形状 1 と形状 2 のシミュレーション結果を単純に線形補間したものも示している．ここでも, 全ての方向において提案手法の性能が線形補間したものより優っていることがわかる．

7.3 ユーザスタディ

23~62 歳の男女 20 名を対象に評価実験をおこなった．実験は 3 段階である．まず, 学習データセットの中から最も各被験者にとって一番良い HRTF を選んでもらった．これには 漸進比較法 [15] を利用した．次にシステムを使って実際にキャリブレーションを

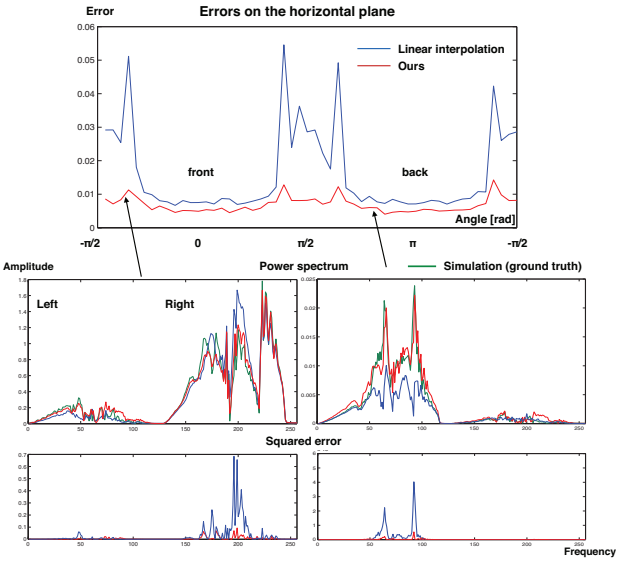

図 8. シミュレーションでの実験結果．上段：横軸が水平面上の方向 (角度)，縦軸が誤差である．中段：2 方向でのパワースペクトルの比較．下段：中段の 2 方向でのパワースペクトル誤差．

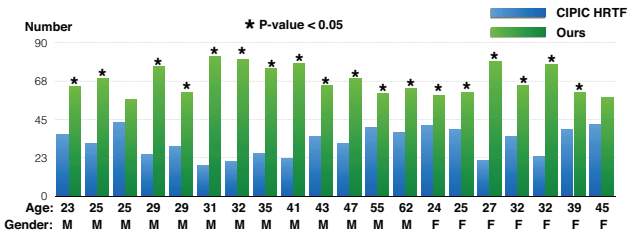

図 9. ユーザスタディ結果．下段では，データベースの HRTF (青) と提案手法で最適化された HRTF (緑) において，それぞれの被験者が定位感が良いと選択した回数 (それぞれの被験者につき合計 100 回) を示している．また p 値が 5%以下のものに ∗ をつけている．

おこなってもらった．最後に一つ目の実験で得られた HRTF と、システムによって得られた HRTF を比較した．比較の方法はブラインドで 2 つの HRTF によってつくられた音をどちらがどちらなのかを被験者には知らせずに提示し, 被験者にはどちらが定位感が良いか, を答えてもらうことを 100 回繰り返した．図 9 が実験結果である．カイ二乗検定で有意差があるもの (p 値 5%以下) に ∗ をつけている．20 人中 18 人で有意差がみられた． これにより, 提案手法がユーザに対して有意にフィットした HRTF をキャリブレーションできることが示された．

8 結論と今後の課題

本稿では, 特定のユーザに対してフィットした HRTF を得るために, Adaptive Variational AutoEncoder で事前に実測データセットから個人性を抽出したの

ちに, その個人性のブレンド量を特定のユーザに対して最適化する, という二段階のアルゴリズムを提案した. 提案手法では, 特別な機材を必要とせず, 無響室での測定と比較して非常に短時間で, ユーザはヘッドフォンも含めた自分自身の環境で, 音を聞いて比較するだけで容易に 3 次元音響のキャリブレーションをおこなうことができる.

しかし, まだ 20〜30 分のキャリブレーション時間というのはユーザにとって負担が大きく, これを短縮することは重要な今後の課題である. これを解決するアプローチとしては, 個人化パラメータのスパース性を利用することが考えられる. 実験では, 最適化後の個人化パラメータがほぼ全ての被験者に対して疎 (ベクトルのいくつかの要素だけが大きな値となり他はゼロに近くなる) となることが観測された. これは, あるユーザが, 多くの人の特徴を兼ね揃えているわけではなく, より少数の人の特徴から表現できるということを示している. この特性を利用することで, 最適化途中でも徐々にゼロに近づいていく成分を切り捨てていったり, スパースコーディングなどの手法を使うことで, 将来的にキャリブレーション時間の短縮が見込まれる.

謝辞

本研究は, JST, ACT-I の支援を受けたものである.

参考文献

[1] Algazi, Duda, Thompson, and Avendano. The CIPIC HRTF Database. In *IEEE Workshop on Applications of Signal Processing to Audio and Electroacoustics*, pp. 99–102, 2001.

[2] Chen, Liu, and Jia. Combining Evolution Strategy and Gradient Descent Method for Discriminative Learning of Bayesian Classifiers. In *Proceedings of the 11th Annual Conference on Genetic and Evolutionary Computation*, No. ¯ 8, pp. 507–514, 2009.

[3] Grijalva, Martini, Goldenstein, and Florencio. Anthropometric-Based Customization of Head-Related Transfer Functions using Isomap in The Horizontal Plane. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2014.

[4] Hansen, Muller, and Koumoutsakos. Reducing the time complexity of the derandomized evolution strategy with covariance matrix adaptation (CMA-ES). In *Evolutionary Computation*, No.11, pp. 1–18, 2003.

[5] Holzl. A Global Model for HRTF Individualization by Adjustment of Principal Component Weights. In *Diploma Thesis*, 2014.

[6] Kahana and Nelson. Boundary element simulations of the transfer function of human heads and baffled pinnae using accurate geometric model. In *Journal of sound and vibration*, pp. 552–579, 2007.

[7] Kaneko, Suenaga, and Sekine. DeepEarNet: individualizing spatial audio with photography, ear shape modeling, and neural networks. In *AES Conference on Audio for Virtual and Augmented Reality*, 2016.

[8] Katz. Boundary element method calculation of individual head-related transfer function. i. rigid model calculation. In *The Journal of the Acoustical Society of America*, 2001.

[9] Kingma and Welling. Auto-encoding variational Bayes. In *The International Conference on Learning Representations (ICLR)*, 2014.

[10] Kingma and Ba. Adam: A method for stochastic optimization. In *CoRR abs/1412.6980*, 2014.

[11] Koyama, Sakamoto, and Igarashi. Crowd-powered parameter analysis for visual design exploration. In *Proceedings of the 27th annual ACM symposium on User interface software and technology (UIST)*, pp. 65–74, 2014.

[12] Luo, Zotkin, and Duraiswami. Virtual AutoEncoder Based Recommendation System for Individualizing Head-Related Transfer Functions. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2013.

[13] Matheron. Principles of geostatistics. In *Economic Geology*, pp. 1246–1266, 1963.

[14] Rezende, Mohamed, and Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *The International Conference on Machine Learning (ICML)*, 2014.

[15] Takahama, Kamishima, and Kashima. Progressive Comparison for Ranking Estimation. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-16)*, 2016.

[16] Xiao and Liu. Finite difference computation of head-related transfer function for human hearing. In *The Journal of the Acoustical Society of America*, 2003.

[17] Yamamoto and Igarashi. Fully Perceptual-Based 3D Spatial Sound Indivisualization with An Adaptive Variational AutoEncoder. In *ACM Transaction on Graphics (SIGGRAPH Asia)*, 2017.

[18] Zotkin, Duraiswami, and Davis. Rendering localized spatial audio in a virtual auditory space. In *IEEE Transactions on Multimedia, vol. 6(4)*, 2004.