Check for updates

Trans. Korean Soc. Noise Vib. Eng., 35(1) : 47~55, 2025
https://doi.org/10.5050/KSNVE.2025.35.1.047

한국소음진동공학회논문집 제 35 권 제 1호, pp. 47~55, 2025
ISSN 1598-2785(Print), ISSN 2287-5476(Online)

정확한 스펙트럴 큐 예측을 위한 간소 피나전달함수 기반 머리전달함수 개인화

HRTF Individualization using Compact PRTF for Accurate Spectral Cues

고 병 윤* · 민 덕 기* · 남 현 욱* · 박 용 화†

Byeong-Yun Ko*, Deokki Min*, Hyeonuk Nam* and Yong-Hwa Park†

(Received September 4, 2024 ; Revised October 25, 2024 ; Accepted November 4, 2024)

Key Words : Head-related Transfer Function(머리전달함수), Spatial Audio(공간음향), Deep Learning(딥러닝), Spectral Cues(스펙트럼 큐)

ABSTRACT

Spatial audio rendering relies on accurate localization perception, which necessitates individualized head-related transfer functions (HRTFs). Previous deep neural network (DNN) approaches for predicting HRTF magnitude spectra from pinna images primarily used HRTF log-magnitude as the training output. However, because HRTFs encompass the acoustic properties of the head and torso, reconstructing the spectral cues essential for elevation localization remains a significant challenge. To address this issue, we introduce PRTFNet, an end-to-end convolutional neural network (CNN) designed to reconstruct individual spectral cues in HRTFs while minimizing the influence of head and torso reflections. PRTFNet utilizes the compact pinna-related transfer function (PRTF) as its output, which isolates the pinna's contribution by excluding head and torso reflections from the head-related impulse response (HRIR). Evaluation using the HUTUBS dataset demonstrates that PRTFNet effectively reconstructs critical spectral cues, such as the first and second spectral notches. It surpasses previous deep learning models in performance, achieving lower log spectral distortion (LSD) and effective LSD (LSD_E), thereby enhancing the precision of spectral cue reconstruction for spatial audio rendering.

기 호 설 명

f : 주파수
H : 머리 전달 함수
i : 주파수 인덱스
j : 음원 방향 인덱스
N_f : 주파수 빈 개수
N_d : 음원 방향 개수
ϕ : 고도각
θ : 방위각

1. 서 론

머리 전달 함수(HRTF)는 임의의 방향에서 귀로 소리가 전달되는 것을 나타내는 주파수 전달함수로 공간음향(spatial audio)를 생성하는데 주요하게 사용되

† Corresponding Author ; Member, KAIST, Professor
E-mail : yhpark@kaist.ac.kr
* Member, KAIST, Department of Mechanical Engineering, Student

A part of this paper was presented and selected as one of best papers at the KSNVE 2024 Annual Spring Conference
‡ Recommended by Editor Nam Keun Kim
© The Korean Society for Noise and Vibration Engineering

며[1], 정확한 음원 위치 인식을 위해서는 개인의 인체 음향 특성을 드러내고 방위 및 고도 위치 파악을 위한 큐를 제공하는 개별 HRTF가 요구된다[2]. 그러나 개별 HRTF는 복잡한 인체 음향 효과와 특히 고주파수 범위에서 스펙트럼 큐(spectral cue)가 귀의 형상에 따라 민감하게 변화하는 특성으로 인해 정확히 예측하는 것이 어렵다. 음향 측정 및 FEM, BEM 시뮬레이션 기술을 포함하여 개별 HRTF를 계측 및 예측하기 위한 다양한 방법이 제안되었지만[3,4], 이러한 방법은 고주파수 큐를 복원하기 위해 고가의 측정 장비가 사용되고 계산 시간이 많이 요구되는 제한점이 존재했다. 반면 물리적 모델을 통해 귀와 머리 형상 치수에 따른 HRTF의 상관관계를 규명하는 방법도 제안되었지만[5], 인체 형상을 매 실험마다 측정하는 과정이 요구되고 이를 위한 별도의 장비가 필요하거나 사람을 통해 측정할 때 측정 오차가 발생되는 문제가 있었다.

최근 이미지 기반 딥러닝 기술의 발전으로 귀 사진을 사용한 HRTF 개별화기술이 제안되었다. 딥러닝 기반 HRTF 개인화 기법은 귀 사진과 같은 인체 형상 이미지를 직접 활용하여 기존의 HRTF 개인화 방식 실용성을 개선하는 것을 목표로 한다[6]. 이를 통해 DNN 모델이 귀 사진만을 활용하여 개별 HRTF를 예측하도록 학습시킴으로써 별도의 인체 측정 과정이 생략된다. 하지만 기존 딥러닝 기반 HRTF 개인화 방법은 DNN 모델을 하위 네트워크로 분할 및 학습함으로써 개인화 성능이 저하되거나 HRTF 학습 데이터의 적은 개수와 수많은 HRTF 주파수 빈으로 인해 과적합 문제가 발생되어 스펙트럼 큐가 손실되는 결과가 나타났다.

이러한 문제를 해결하기 위해 이 연구에서는 PRTFNet이라는 새로운 종단간(end-to-end) 컨볼루션 신경망(CNN) 기반 딥러닝 모델을 제안한다. PRTFNet은 귀 사진으로부터 개별 HRTF를 예측하고 정확한 스펙트럼 큐를 복원하는 것을 목표로 한다. 이를 위해 딥러닝 전처리 기법으로써 귀에 의한 음향 효과만을 HRTF로부터 추출하고자 머리 충격 함수(HRIR)로부터 윈도우 함수를 적용하는 방법, HRTF의 주파수 빈 개수를 최소화하기 위해 다운 샘플링 기법을 통해 간소 PRTF를 추출하여 학습하는 방법을 제안한다. 또한 음원 방향에 따른 스펙트럼 큐의 변화 특성을 DNN에 정확히 학습하기 위해 음원 방향 인덱스를 입력단에 넣어 딥러닝 모델을 음원 방향별로 학습하는 방식을 제안했다.

이 연구의 증명으로 개별 HRTF 스펙트럼 큐의 복원 결과를 평가하고자 개별 귀 형상 및 HRTF가 포함된 HUTUBS HRTF 데이터베이스를 활용하여 PRTFNet을 검증한다[7].

2. HRTF의 스펙트럼 큐

HRTF의 크기(magnitude) 스펙트럼은 고도 방향에 대한 음원의 공간지각을 결정하는 데 중요한 역할을 한다. HRTF 크기 스펙트럼의 전반적인 패턴은 음원의 고도각 인식을 하는 데 있어 스펙트럼상의 세부 패턴보다 더 중요하다[8]. 이때 HRTF 스펙트럼의 메인 피크(peak)과 노치(notch)를 스펙트럼 큐라고 정의한다. 특히, 5 kHz 이상의 스펙트럼 큐는 음원의 고도각 인식을 담당하며 16 kHz 이상 3.8 kHz 미만의 HRTF의 주파수 성분은 음원의 고도각 인식에 영향을 미치지 않는다[8]. 고도각 인식에서 스펙트럼 큐의 중요성을 검증하기 위해 수행된 실험에서 Fig. 1(a)와 같이 HRTF의 스펙트럼 큐만 추출한 파라메트릭 HRTF를 활용하여 합성된 공간음향을 청취자가 들을 때 음원 고도각 인식 결과가 기존의 HRTF와 동일하게 응답하는 것을 보였다[9].

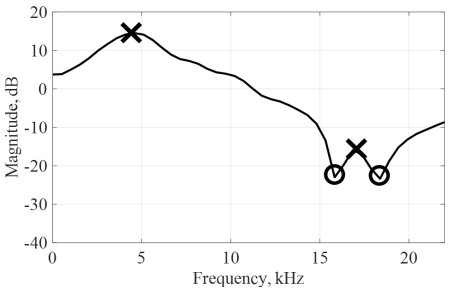
정중면(median plane)에서 음원 방향에 따라 달라지는 스펙트럼 큐의 분포는 수평면(horizontal plane)에 비해 상대적으로 더 명확하며[8], 청취자는 스펙트럼 큐의 주파수와 크기 정보를 통해 음원의 고도각을 추정하게 된다. 실제로 음원이 청취자의 정면에서 머리 위로 이동함에 따라 스펙트럼 큐 첫번째 노치의 주파수가 고주파수로 이동한다[10]. 스펙트럼 큐를 분석하기 위해 수행된 이전 실험에서 B&K 헤드 앤 토르소 시뮬레이터(HATS) Type 4100의 HRTF를 측정했으며 Fig. 1(b)와 Fig. 1(c)에 수평면 및 정중면에 대한 결과를 나타냈다[11]. 해당 결과에서 노치의 주파수뿐 아니라 4 kHz 이상의 피크의 주파수도 소리 상승에 따라 변화가 발생하는 것을 알 수 있다. 추가적인 스펙트럼 큐의 특성으로서 주파수 및 크기가 귀의 형상에 따라 상당하게 변화된다[12]. 이는 스펙트럼 큐의 분포가 개개인에 따라 민감하게 달라지며 높은 개별성을 갖는 것을 의미한다.

음원 방향에 대한 민감도 및 높은 개별성 같은 스펙트럼 큐의 특성을 고려할 때, HRTF 개인화를 위한 DNN 모델은 다음 두 가지 문제를 해결해야한다. (1) 음원의 방향에 따라 민감하게 변화하는 HRTF에

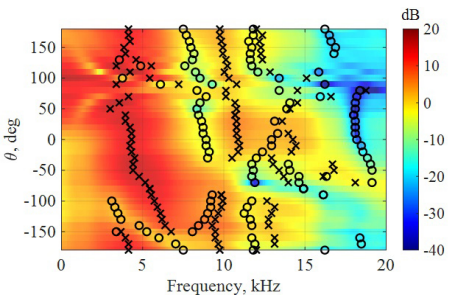
서 스펙트럼 큐 패턴을 정확하게 추출해야 하며, (2) 개인의 귀 모양과 스펙트럼 큐 간의 연관성을 포착하는 딥러닝 모델을 설계해야 한다.

3. PRTFNet

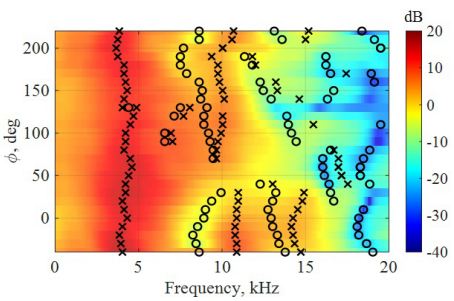
이 연구는 HRTF 개인화 DNN 모델인 PRTFNet을 제안한다. PRTFNet의 주요 목표는 HRTF의 스펙트럼 큐를 귀 사진만을 이용해 개인별로 정확하게 예측하는 것이다. PRTFNet은 종단간 CNN 구조로 구성되며, 딥러닝 모델 학습 절차는 아래 세 단계로 구성된다. 첫째, 머리와 토르소의 소리 반사의 영향을 제거하기 위해 윈도우 함수를 사용하여 HRIR을 필터링한다. 둘째, 윈도우 처리된 HRIR에서 영의 값을 갖는 샘플들을 제거하고, FFT(fast Fourier transform)를 사용하여 HRIR을 간소 PRTF로 변환한다. 마지막으로, CNN 모델을 음원 방향별로 학습시키며 이를 위해 네트워크 입력으로써 귀 사진과 원앤핫(one-and-hot) 인코딩의 방향 인덱스를 사용하며, 해당 방향에 대한 간소 PRTF는 네트워크 출력으로 사용한다. PRTFNet의 각 구성 요소에 대한 자세한 설명은 다음의 내용에서 기술한다.



(a) In the HRTF for 0° azimuth and 60° elevation



(b) In the horizontal plane HRTFs depending on sound azimuth(-180° ~ 180°)



(c) In the median plane HRTFs depending on sound elevation(-40° ~ 220°). Here, the symbol legend means

Fig. 1 Example of the distribution of the frequency spectral cues in the HRTFs(X, prominent peak; O, prominent notch)

3.1 귀 효과만을 추출하기 위한 윈도우 기법

이전 연구의 위치 추정 테스트에서 인간의 청각 시스템이 HRTF 크기를 로그 스케일에 기반해서 인식한다는 것이 나타났다[13]. 이를 통해 이 연구는 로그 스케일의 HRTF 스펙트럼 큐를 정확하게 복원하고자 한다. 이전 개인화 DNN 모델은 네트워크 입력으로 귀 사진을 사용하고 네트워크 출력으로 HRTF를 사용했다. 하지만 HRTF는 귀의 음향 특성뿐 아니라 머리와 토르소의 음향 특성도 포함되어 있다. 이로 인해 네트워크 입력인 귀 사진에서 추측할 수 없는 머리 토르소 정보는 네트워크 입력과 출력 사이의 상관관계를 손상시키는 문제를 발생시키며 결과적으로 HRTF 추정 정확도가 저하된다.

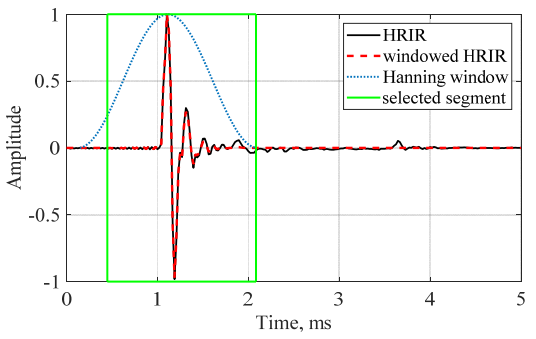
앞에서 언급한 입력과 출력 간의 손상된 상관 관계 문제를 확인하고자 무향실에서 단방향 스피커로 구성된 반원 형상의 스피커 어레이를 통해 B&K HATS Type 4100의 HRIR를 취득 및 분석했다[11]. Fig. 2에서 FFT를 통해 HRIR을 HRTF로 변환할 때 전 주파수 범위에 걸쳐 작은 고조파(HRTF 스펙트럼 윤곽선을 따라 나타나는 국소 변동)가 존재하는 것을 알 수 있다. 이러한 고조파는 머리와 토르소에서 발생하는 음향 반사에서 기인한다[14]. 머리와 토르소에 대한 정보가 없는 입력 데이터, 즉 귀 사진에서 이러한 스펙트럼 요소를 예측하도록 딥러닝 모델을 학습하는 것은 불가능하다. 또한 이러한 고조파로 인해 HRTF 크기 스펙트럼을 네트워크 출력으로 사용할 때 스펙트럼상의 세부적인 패턴에 과적합(overfitting)되는 문제

가 발생할 수 있다. 이러한 과적합은 HRTF 개인화 딥러닝 네트워크가 HRTF의 전반적인 패턴, 즉 스펙트럼 큐를 학습하는 것을 방해한다.

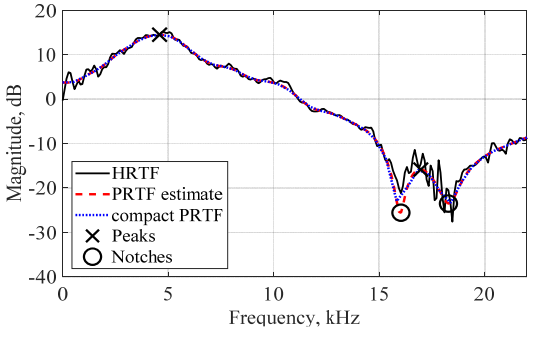
Fig. 2(b)에서 HRTF의 스펙트럼 큐가 4 kHz, 17 kHz에서 피크와 16 kHz, 18 kHz에서 노치로 나타난다. 머리와 토르소에서 음향 반사 효과를 제거함으로써 스펙트럼 단서를 추출할 수 있으며 머리와 토르소에서 발생한 반사는 일반적으로 직접음(direct sound)이 외이도에 도달된 후에 1 ms 이후에 도착한다[15]. 머리와 토르소 음향 효과를 제거하기 위해 2 ms 길이의 해닝(hanning) 윈도우를 적용하여 중심을 HRIR의 최대값에 정렬한다. 윈도우 처리된 HRIR의 FFT 스펙트럼은 주로 귀의 음향 효과만 반영되어 PRTF 추정치를 나타낸다. Fig. 2(b)에서 PRTF 추정치를 빨간색 점선으로 표시했다.

3.2 간소 PRTF를 추출을 위한 다운 샘플링

앞서 언급한 PRTF 추정치는 수백 개의 주파수 빈으로 구성되며 HRTF 개인화 모델을 통해 예측되어야 하는 네트워크 출력이다. 하지만 CIPIC[16], ITA[17] 및 HUTUBS와 같이 HRTF 개인화에 사용되는 데이터베이스에는 딥러닝 네트워크 매개변수에 비해 상대적으로 적은 수의 학습 데이터 샘플이 있다[7]. 이러한 제한된 수의 학습 데이터셋으로부터 PRTF 추정치와 귀 사진을 사용하여 네트워크 모델을 학습할 경우 방대한 수의 네트워크 출력 포인트로 인한 과적합 문제가 발생하게 된다.



(a) HRIRs



(b) HRTFs

Fig. 2 Comparison of HRIRs and HRTFs of artificial head-torso simulator at azimuth 0° and elevation 60°

이 연구에서는 FFT를 적용하기 전에 HRIR에서 불필요한 시간 영역 샘플을 제거하여 스펙트럼 큐를 보존하면서 PRTF 추정치의 주파수 빈 수를 줄이는 방법을 제안한다. Fig. 2(a)에서 윈도우 처리된 HRIR는 윈도우 범위 밖에서는 영의 값을 갖는다. 제로 패딩 이론(zero-padding theorem)에 의하면[18], 시간 영역의 제로 패딩은 주파수 영역 데이터 보간 역할을 한다. 이에 따라 윈도우 HRIR에서 영의 값을 갖는 샘플을 제거하면 전반적인 스펙트럼 패턴(스펙트럼 큐)을 유지하면서 PRTF 추정치의 주파수 빈 수를 효과적으로 줄이게 된다. sinc 함수로 표시되는 HRIR의 직접음은 사이드로브(sidelobe)를 수반한다. sinc 함수의 길이가 48 kHz 샘플링 주파수에서 1.3 ms를 이상일 때 sinc 함수의 사이드로브 통과대역(passband) 리플(ripple)은 1 dB 미만으로 형성된다[19]. 따라서 귀의 효과가 존재하지 않는 HRIR 최대 진폭으로부터 0.6 ms 이상 떨어진 왼쪽 사이드로브를 제거할 수 있다. 이러한 과정을 거친 후 선택된 구간은 Fig. 2(a)에서 실선 녹색 선으로 표시되며 해당 구간의 FFT는 Fig. 2(b)에서 점선 파란색 선으로 표시된다. 이 결과를 통해 주파수 빈의 수가 효과적으로 감소되었지만 스펙트럼의 전반적인 패턴이 PRTF 추정치와 거의 동일한 것을 알 수 있다. 이 연구에서는 이 스펙트럼을 간소 PRTF라고 정의하여 딥러닝 모델의 출력으로 사용한다.

3.3 딥러닝 모델 및 방향별 HRTF 개인화 학습법

이전 연구에서 개별 귀 사진을 사용한 HRTF 개인화 모델은 세 가지 하위 네트워크, 즉 variational autoencoder(VAE), fully connected layer(FC) 및 conditional VAE(CVAE)를 사용했다[20,21]. VAE 및 CVAE 모델은 일반적으로 다양한 분야에서 데이터

복원에 사용되지만, 2개 이상의 네트워크를 통한 다단계 개인화 학습 방법은 각 네트워크가 개별적으로 학습되기 때문에 비효율적이며 개인화 성능이 감소되는 문제가 발생할 수 있다. 이러한 한계를 해결하기 위해 이 연구에서는 귀 사진에서 HRTF 크기 스펙트럼에 이르는 HRTF 개인화 과정을 종단간 네트워크로 통합하는 딥러닝 모델을 제안한다. 주요 네트워크 모델로는 이미지 인식과 같은 패턴 추출 영역에서 효과가 입증된 CNN을 활용한다. 이때 귀는 복잡한 형상학적 구조를 포함하며 스펙트럼 큐의 근원인 귀의 공진 모드는 국소적인 형상뿐 아니라 귀의 폭 및 깊이와 같은 전반적인 형상의 영향을 받게된다[5]. 귀의 복잡한 구조 패턴을 효과적으로 추출하고 스펙트럼 큐와 귀 형상 간의 상관관계를 정확하게 학습하기 위해 ResNet 아키텍처에서 영감을 받은 잔차 블록(residual block)[22]을 CNN 구조에 적용했다. Fig. 3은 설계된 네트워크 모델인 PRTFNet를 묘사한다.

전구면 HRTF의 데이터 크기는 주파수 빈 개수 × 방위각 개수 × 고도각 개수로 결정된다. 이때 PRTFNet을 사용하여 전구면 간소 PRTF의 크기 스펙트럼을 네트워크 출력으로 사용하게 되면 네트워크 모델은 상당한 출력 차원으로 인해 스펙트럼 큐의 방향별 특성 학습이 이루어지지 않는다. 이를 해결하고자 이 연구에서는 방향 인덱스로서 원앤핫 인코딩을 추가 네트워크 입력으로 도입하고 해당 방향에 대한 간소 PRTF의 크기의 로그 스케일을 네트워크 출력으로 사용함으로써 PRTFNet의 방향별 학습을 제안한다. Fig. 3은 PRTFNet의 네트워크 입력 및 출력을 나타낸다. 네트워크 입력단에서 2D 원 핫 인코딩은 평탄화와 FC 레이어를 통해 1×256 임베딩 벡터로 변환되고 피나 사진과 통합하게 된다. 이 통합된 이미지 결과를 PRTFNet 입력으로 사용한다.

음원 위치 인식 특성을 고려하여 HRTF 개인화를 달성하기 위해 이 연구에서 PRTFNet을 위한 손실 함수를 정의한다. 청각 시스템이 로그 스케일로 HRTF 크기 인식하는 특성을 토대로 정의된 HRTF 개인화 성능지수인 로그 스펙트럼 왜곡(LSD)은 식 (1)과 같이 정의된다[23].

$$\mathrm{LSD}=\sqrt{\frac{1}{N_d N_f}\sum_{j=1}^{N_d}\sum_{i=1}^{N_f}\left(20\log\frac{\left|H_{\phi_j,\theta_j}(f_i)\right|}{\left|\tilde{H}_{\phi_j,\theta_j}(f_i)\right|}\right)^2} \tag{1}$$

여기서 $H_{\phi j\theta j}(f_i)$는 음원 j번째 방향에서 i번째 주파수 빈의 HRTF의 참값을 나타내며 $\hat{H}_{\phi j\theta j}(f_i)$는 HRTF의 예측값을 나타낸다. PRTFNet의 경우 모든 방향에 대한 모델 학습이 동시에 수행되는 것이 아닌 방향별로 학습이 수행되므로 PRTFNet 최적화를 위해 음원 방향별 LSD으로 식 (2)와 같이 손실값을 정의하여 학습한다.

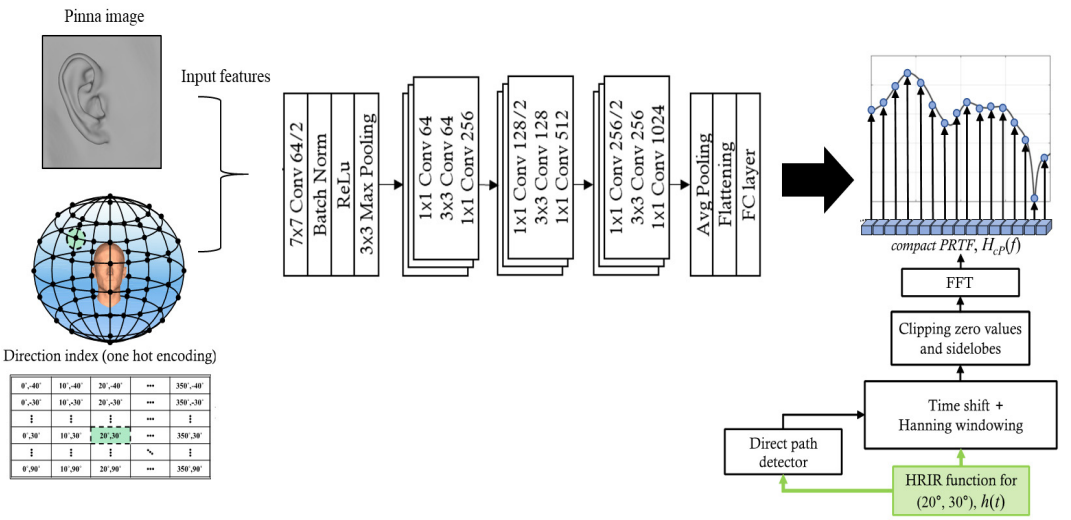


Fig. 3 Example of training of proposed PRTFNet using pinna image and one hot encoding for direction index. PRTFNet is trained to predict compact PRTF from the concatenated input

$$\text{Loss}=\sqrt{\frac{1}{N_f}\sum_{i=1}^{N_f}\left(20\log\frac{|H_{\phi,\theta}(f_i)|}{|\tilde{H}_{\phi,\theta}(f_i)|}\right)^2} \quad (2)$$

3.4 방향 인덱스의 공간 분해능

공간음향 분야에서는 청취자와 음원 모두 자유롭게 이동할 수 있으며, 이는 음원 방향이 어느 방향으로든 달라질 수 있음을 의미한다. 따라서 공간음향 렌더링의 주요 목표 중 하나는 모든 방향에 대해 연속적인 HRTF를 얻는 것이다. 이 논문에서는 PRTFNet은 공간음향 렌더링의 완결성을 위해 대상 청취자의 연속적인 HRTF를 생성하고자 한다. 연속적인 HRTF의 공간 해상도를 높이는 것은 공간 이산화로 인해 발생하는 공간음향 아티팩트를 완화할 수 있지만, PRTFNet을 학습하는 데 요구되는 계산 시간이 방향 인덱스의 공간 해상도에 비례해서 증가한다. 따라서 계산량과 청각 인식 특성을 고려하여 방향 인덱스의 적절한 공간 해상도로 PRTFNet을 학습해야 한다. 사람의 공간음향 방향 분해능에 대한 이전 연구에 의하면 최소 공간음향 방향 분해능은 5.4° 이상이다[24]. 또한 매니폴드 학습에 기반한 HRTF 보간 방법은 20° 미만의 간격으로 공간 샘플링된 HRTF를 사용하여 HRTF를 재구성할 수 있다[25]. 공간음향 방향 해상도와 HRTF 보간법 성능을 고려하여 Fig. 3과 같이 방위각(총 36개의 방위각)과 고도각(총 14개의 고도각)에 대해 10°의 공간 해상도를 설정하여 학습에 사용했다.

4. 실험 구성

HUTUBS HRTF 데이터베이스를 사용하여 이 연구에서 제안한 PRTFNet의 검증을 수행했다[7]. 데이터베이스에는 0°~350° 범위의 방위각과 -90°~90° 범위의 고도각에 대해 HRIR이 피험자별로 측정되었다(방위각 고도각에서 10° 해상도로 샘플링됨). 또한 데이터베이스는 귀, 머리 및 토르소의 인체 측정뿐 아니라 머리 및 귀 모양이 스캔된 3D 메쉬 데이터를 포함한다. 개별 귀 사진을 생성하기 위해 3D 메쉬의 측면 뷰를 2D 회색조 이미지로 변환하고 Fig. 3과 같이 사진을 256 × 256 픽셀 해상도로 다운 샘플링하여 딥러닝 학습에 사용했다.

HUTUBS 데이터셋에서 사용할 수 있는 116개의 귀 샘플 중 90개의 귀 샘플(45명의 피험자에 해당)을 PRTFNet 모델 학습에 사용했다. 테스트 데이터셋을 위해 14개의 귀 샘플(7명의 피험자)이 사용되었다. 검증하고자 하는 음원 방향으로 총 36개의 방위각(해상도 10°, 0°에서 350°)과 14개의 고도각(해상도 10°, -40°에서 90°)을 설정했다.

Table 1 Objective performance comparison of different methods for HRTF magnitude individualization in terms of LSD and LSD_E

Methods	LSD, dB	LSD_E, dB
Baseline	10.4	7.5
Baseline with compact PRTF	6.1	6.3
CNN with full grid HRTF	12.3	14.5
CNN with direction-wise training	8.8	8.5
PRTFNet	**5.0**	**5.1**

5. 성능 평가

5.1 LSD를 이용한 개인화 성능 검증

PRTFNet의 성능을 객관적으로 평가하기 위해 HRTF 크기 스펙트럼의 복원 성능 지수로 LSD를 사용했다. 추가적인 성능 지수로 음원 고도각 추정에 사용되는 주파수 범위(4 kHz ~ 16 kHz)의 LSD인 effective LSD(LSD_E)를 정의하여 사용했다. 성능 지수를 계산하기 위해 HRTF 참값으로써 데이터셋의 HRIR 로부터 파생된 HRTF를 사용했다.

이 연구에서 제안한 간소 PRTF, CNN 및 방향별 학습의 개별 기여도를 평가하기 위해 제거 연구(ablation study)를 수행했으며 테스트 데이터셋에서 얻은 평균 LSD 및 LSD_E 값을 Table 1에 나타냈다. 간소 PRTF을 통해 머리와 토르소에서 발생하는 음향 반사 효과가 제거된 출력으로 딥러닝 모델을 학습함으로써 baseline과 비교했을 때 LSD가 3 dB 이상 개선되었다[21]. 전구형 HRTF 학습과 비교했을 때 방향별 학습법은 네트워크 출력 차원을 줄임으로써 LSD 및 LSDE가 각각 약 4 dB 및 6 dB으로 개선된 것을 보였다. PRTFNet은 간소 PRTF, CNN 및 방향별 학습을 결합하여 LSD 및 LSD_E가 각각 5.0 dB 및 5.1 dB를 달성하여 baseline 성능을 능가하는 것을 보였다.

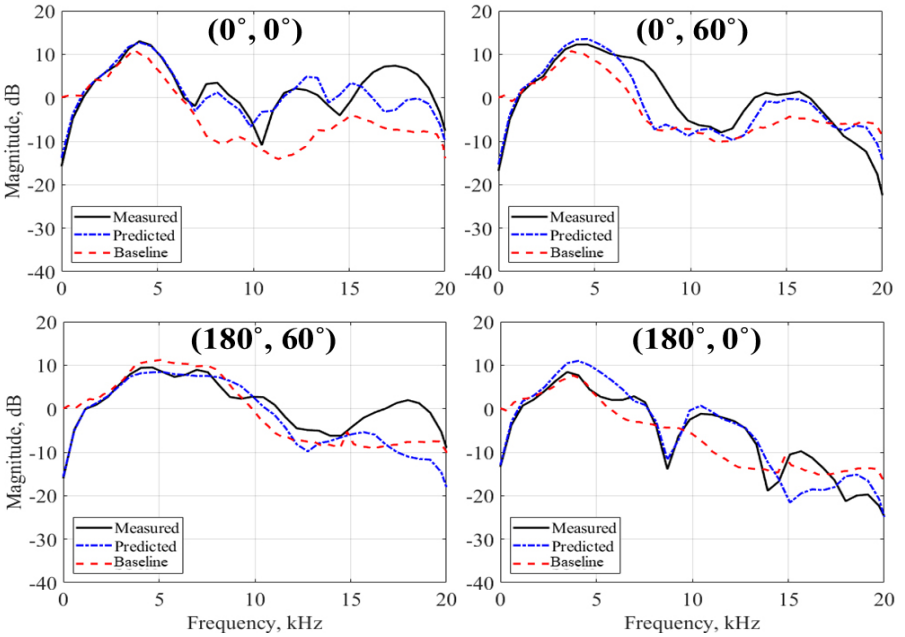


Fig. 4 Comparison of true parametric HRTFs and network outputs from PRTFNet and baseline. The title(ϕ, θ) denotes azimuth angle of ϕ and elevation angle of θ for DoA of sound source

5.2 HRTF 개인화 네트워크 출력 비교

Fig. 4에서 PRTFNet 및 baseline의 네트워크 출력과 함께 매개변수 HRTF(데이터베이스의 HRIR으로부터 계산된)를 정중면 방향에 대해 나타냈다. 이때 정중면에 대한 결과를 도시한 이유는 해당 방향에서 스펙트럼 큐의 변화가 우세하게 나타나 스펙트럼 큐 복원 성능 비교에 용이하기 때문이다[10]. 네트워크 출력 비교에서 PRTFNet은 정면 방향($\phi = 0°$, $\theta = 0°$) 4 kHz ~ 8 kHz의 첫 번째 및 두 번째 피크와 첫 번째 노치를 예측하였고 후면 방향($\phi = 180°$, $\theta = 0°$) 약 8 kHz에서 발생한 첫 번째 노치를 정확하게 복원했다. 그러나 귀 사진으로 예측하기 어려운 고주파수에서 고차 모드가 우세하게 발생하기 때문에 10 kHz 이상의 주파수 영역에서 스펙트럼 왜곡이 증가하게 된다. 하지만 음원 고도각 인식에 대한 주요 큐가 첫 번째 두 번째 스펙트럼 큐에 의해 결정되므로[9], PRTFNet은 이러한 피크와 노치를 성공적으로 예측함으로써 청취자에게 정확한 고도 현지화를 가능하게 함을 알 수 있다.

6. 결 론

이 논문에서는 귀 사진을 네트워크의 입력으로 활용하여 개인화된 HRTF 크기 스펙트럼을 예측하는 방법을 제안했다. 이를 위해 음원 고도각 인식에 중요한 스펙트럼 큐를 정확하는 PRTFNet이라는 개인화 딥러닝 모델을 새롭게 개발했다. PRTFNet은 간소 PRTF를 네트워크의 출력으로 사용함으로써 머리와 토르소의 음향 반사 효과를 배제해 학습한다. 이 접근 방식은 네트워크 입력과 출력 간의 보다 정확한 상관 관계를 보장하는 동시에 수백 개의 주파수 빈에 의한 과적합 문제를 최소화한다. PRTFNet의 네트워크 구조로써 종단간 CNN과 residual block을 사용하고 방향별 학습법을 적용함으로써 스펙트럼 큐의 방향 특성을 추출하고 네트워크 출력의 차원을 최소화하여 각 개인별로 강건하게 PRTF를 예측하는 방법을 제안했다.

HUTUBS 데이터 세트를 사용하여 이 논문에서 제안한 PRTFNet을 검증했으며, LSD 및 LSD_E 지수를 기반으로 개별화 성능을 평가했다. PRTFNet은 이전 딥러닝 모델에 비해 개인화 성능이 LSD 및 LSD_E에서 각각 5 dB 및 2 dB가 개선된다. PRTFNet의 네트워크 출력인 간소 PRTF 스펙트럼을 분석했을 때 첫 번째 및 두 번째 피크와 8 kHz 미만의 첫 번째 노치를 정확히 예측하는 것을 확인했다. 이러한 결과는

PRTFNet이 음원 고도각 인식을 위한 스펙트럼 큐를 효과적으로 제공함을 입증한다.

후 기

이 연구는 대한민국 정부(산업통상자원부 및 방위사업청) 재원으로 민군협력진흥원에서 수행하는 민군기술협력사업(협약번호 UM22409RD4)과 해양수산부 재원으로 선박해양플랜트연구소의 기본사업인 '스마트 해양안전 및 기업지원을 위한 오픈플랫폼 기술개발(PES5230)'에 의해 수행되었습니다.

References

(1) Son, D., Park, Y., Park, Y. and Jang, S., 2014, Building Korean Head-related Transfer Function Database, Transactions of the Korean Society for Noise and Vibration Engineering, Vol. 24, No. 4, pp. 282~288.

(2) Wenzel, E. M., Arruda, M., Kistler, D. J. and Wightman, F. L., 1993, Localization using Nonindividualized Head-related Transfer Functions, Journal of the Acoustical Society of America, Vol. 94, No. 1, pp. 111~123.

(3) Kahana, Y. and Nelson, P. A., 2006, Numerical Modelling of the Spatial Acoustic Response of the Human Pinna, Journal of Sound and Vibration, Vol. 292, No. 1-2, pp. 148~178.

(4) Mokhtari, P., Takemoto, H., Nishimura, R. and Kato, H., 2007, Comparison of Simulated and Measured HRTFs: FDTD Simulation using MRI Head Data, Audio Engineering Society Convention, p. 7240.

(5) Shaw, E. A. G., 1997, Binaural and Spatial Hearing in Real and Virtual Environments, Psychology Press, Chapter 2, Acoustical Features of the Human External Ear, NY, United States, pp. 25~47.

(6) Grijalva, F., Martini, L., Florencio, D. and Goldenstein, S., 2016, A Manifold Learning Approach for Personalizing HRTFs from Anthropometric Features, IEEE/ACM Transactions on Audio, Speech and Language Processing, Vol. 24, No. 3, pp. 559~570.

(7) Brinkmann, F., Dinakaran, M., Pelzer, R., Grosche, P., Voss, D. and Weinzierl, S., 2019, A Cross-evaluated Database of Measured and Simulated HRTFs Including 3D Head Meshes, Anthropometric Features and Headphone Impulse Responses, Journal of the Audio Engineering Society, Vol. 67, No. 9, pp. 705~718.

(8) Asano, F., Suzuki, Y. and Sone, T., 1990, Role of Spectral Cues in Median Plane Localization, Journal of the Acoustical Society of America, Vol. 88, No. 1, pp. 159~168.

(9) Iida, K., Itoh, M., Itagaki, A. and Morimoto, M., 2007, Median Plane Localization using a Parametric Model of the Head-related Transfer Function based on Spectral Cues, Applied Acoustics, Vol. 68, No. 8, pp. 835~850.

(10) Asano, F., Suzuki, Y. and Sone, T., 1990, Role of Spectral Cues in Median Plane Localization, Journal of the Acoustical Society of America, Vol. 88, No. 1, pp. 159~168.

(11) Lee, G.-T., Choi, S.-M., Ko, B.-Y. and Park, Y.-H., 2022, HRTF Measurement for Accurate Sound Localization Cues, Research Gate, Version 2, pp. 1~39.

(12) Musicant, A. D. and Butler, R. A., 1984, The Influence of Pinnae-based Spectral Cues on Sound Localization, Journal of the Acoustical Society of America, Vol. 75, No. 4, pp. 1195~1200.

(13) Romigh, G. D., Brungart, D. S., Stern, R. M. and Simpson, B. D., 2015, Efficient Real Spherical Harmonic Representation of Head-related Transfer Functions, IEEE Journal of Selected Topics in Signal Processing, Vol. 9, No. 5, pp. 921~930.

(14) Algazi, V. R., Avendano, C. and Duda, R. O., 2001, Elevation Localization and Head-related Transfer Function Analysis at Low Frequencies, Journal of the Acoustical Society of America, Vol. 109, No. 3, pp. 1110~1122.

(15) Iida, K. and Oota, M., 2018, Median Plane Sound Localization using Early Head-related Impulse Response, Applied Acoustics, Vol. 139, pp. 14~23.

(16) Algazi, V. R., Duda, R. O., Thompson, D. M. and Avendano, C., 2001, The CIPIC HRTF Database, Proceedings of the IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics, pp. 99~102.

(17) Bomhardt, R., de la Fuente Klein, M. and Fels, J., 2016, A High-resolution Head-related Transfer Function and Three-dimensional Ear Model Database, Proceedings of Meetings on Acoustics, Vol. 29, No. 1, 050002.

(18) Smith, J. O III., 2007, Mathematics of the Discrete Fourier Transform(DFT): With Audio Applications, 2nd Edition, BookSurge Publishing, SC, United States.

(19) Viswanathan, M., 2020, Wireless Communication Systems in Matlab, 2nd Edition, Independent Publisher, IL, United States.

(20) Miccini, R. and Spagnol, S., 2020, HRTF Individualization using Deep Learning, Proceedings of the IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops, pp. 390~395.

(21) Miccini, R. and Spagnol, S., 2021, A Hybrid Approach to Structural Modeling of Individualized HRTFs, Proceedings of the IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops, pp. 80~85.

(22) He, K., Zhang, X., Ren, S. and Sun, J., 2016, Deep Residual Learning for Image Recognition, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770~778.

(23) Bilinski, P., Ahrens, J., Thomas, M. R. P., Tashev, I. J. and Platt, J. C., 2014, HRTF Magnitude Synthesis via Sparse Representation of Anthropometric Features, Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 4468~4472.

(24) Scarpaci, J. W., Colburn, H. S. and White, J. A., 2005, A System for Real-time Virtual Auditory Space, Proceedings of the International Conference on Auditory Display, pp. 241~246, Limerick, Ireland

(25) Grijalva, F., Martini, L. C., Florencio, D. and Goldenstein, S., 2017, Interpolation of Head-related Transfer Functions using Manifold Learning, IEEE Signal Processing Letters, Vol. 24, No. 2, pp. 221~225.