arXiv:2302.08130v1 [cs.SD] 16 Feb 2023

个性化音频质量偏好预测

Chung-Che Wang、*Yu-Chun Lin*、*Yu-Teng Hsu*、*Jyh-Shing Roger Jang*

国立台湾大学计算机科学与信息工程学系，台湾

摘要

本文提出结合音频输入与受试者信息，预测内容相同但音质不同的两个音频片段之间的个性化偏好。研究采用孪生网络比较输入并预测偏好，并探讨孪生网络两侧的多种不同结构。实验发现，以 PANNs 的 CNN6 作为编码器、以多层感知机模块作为解码器的 LDNet，表现优于仅使用音频输入的基线模型，整体准确率从 77.56% 提升至 78.04%。实验结果还表明，使用所有受试者信息（包括年龄、性别以及头戴式耳机或耳塞式耳机的规格），比仅使用其中一部分更有效。

索引词— 个性化偏好预测、音频质量、孪生网络

1. 引言

自动音频质量测量可应用于许多不同的音频相关任务，包括检测伪劣音频 [1], 监测在线会议中的语音质量[2], 或评估合成语音的自然度 [3]。此外，由于许多人日常生活中都会使用手机，而音频质量是选择手机时的考量因素之一 [4], ，测量手机音频播放的质量也是一项重要任务。

另一方面，虽然有些研究侧重于预测用户评分均值 [5], ，旨在进行个性化预测的方法也得到了研究。Saquil 等人将用户划分为不同类别，并将类别标签作为其网络的输入之一，以对不同视频片段的偏好进行排序 [6]。Xu 等人将受试者的个人信息和人格测试结果加入模型的输入特征，以预测他们对悲伤音乐的偏好 [7]。

Saquil 等人 [6] 和 Xu 等人 [7] 关注不同内容的偏好，而 Leng 等人 [8] 和 Hunag 等人 [9] 则关注对不同质量的相同内容进行偏好预测。Leng 等人提出了 MBNet，该模型由 MeanNet 和 BiasNet 组成：MeanNet 仅接收音频频谱以预测平均意见分（MOS）；BiasNet 与 MeanNet 并行运行，同时接收音频频谱和受试者 ID，以预测个体受试者评定的可能存在偏差的评分 [8]。为改进 MBNet，Hunag 等人 [9]提出了 LDNet，该模型采用编码器—解码器结构来预测听者相关评分，其中编码器接收输入语音，解码器则接收编码向量和受试者 ID，以输出听者相关评分。然而，这些方法最大的局限在于，受试者信息仅以其 ID 的形式输入，因此在推理阶段无法处理未见过的受试者。

据我们所知，尽管受试者信息对听音测试的影响已受到讨论 [10], ，但尚无研究专门探讨将受试者的个人信息加入音频质量偏好预测模型。本文旨在预测不同手机播放的两个音频片段之间的个性化偏好，即对内容相同但音质不同的两个音频片段的偏好。我们研究了采用多种网络结构和不同类型个人信息时的模型性能。本文其余部分安排如下：第 2 节介绍数据集的收集过程，第 3 节介绍我们提出的方法，第 4 节展示实验设置和结果，第 5 节总结本文并探讨未来工作。

2. 数据集

为获取不同音质的音频片段，我们选取了一些歌曲片段，用设备播放并进行录制。本研究选取了英语、普通话、日语和韩语流行歌曲中的七段歌曲片段，由五个不同品牌的手机播放。每段歌曲片段约为 10 至 15 秒，每部手机均以两种不同音量播放。第一种音量为各手机的最大音量，不同手机约为 78 至 84 分贝。第二种音量称为正常音量，不同手机约为 70 至 75 分贝。录音使用 3DIO Free Space XLR 双耳麦克风和 ART USB Dual Pre Project Series 计算机接口。所有歌曲片段和手机的录音配置（包括手机与麦克风之间的距离以及音量增益）均保持一致。

随后使用录制的文件制作问卷。每首歌曲每种音量的五个录制音频文件

片段组成十对，其中每一对中的两个音频文件分别对应不同的手机，且两个音频文件的顺序是随机分配的。因此，每种音量共有70对。这70对被分为五份问卷，每份问卷包含对应七首不同歌曲的14对。在每份问卷的每一对中，受试者将根据自己的偏好选择哪一个更好。偏好分数为 [1, 5], 中的整数，其中1和2表示受试者更喜欢第一个音频，3表示对两个音频中的任何一个都没有偏好，4和5表示受试者更喜欢第二个音频。为方便起见，我们将把分数转换到 [-2, 2] 的范围内，以便进行后续描述和分析。

共有 31 名受试者参与测试，其中男性 27 名，女性 4 名。他们的年龄介于 21 至 46 岁之间，其中 21 人的年龄在 21 至 25 岁之间。受试者回答的问题总数为 2,856 个。此外，受试者还需提供测试所用头戴式耳机或耳塞式耳机的品牌和型号，而耳机的规格信息（包括阻抗、频率响应的下限和上限以及灵敏度）则从互联网上收集。剔除所用头戴式耳机或耳塞式耳机超过一半规格信息无法查到的受试者数据，以及偏好分数为 0（即对两个音频文件均无偏好）的问题后，剩余受试者人数为 23，回答的音频对数量为 2,000。 ± 2 和 ± 1 的回答数分别为 936 和 1,064。

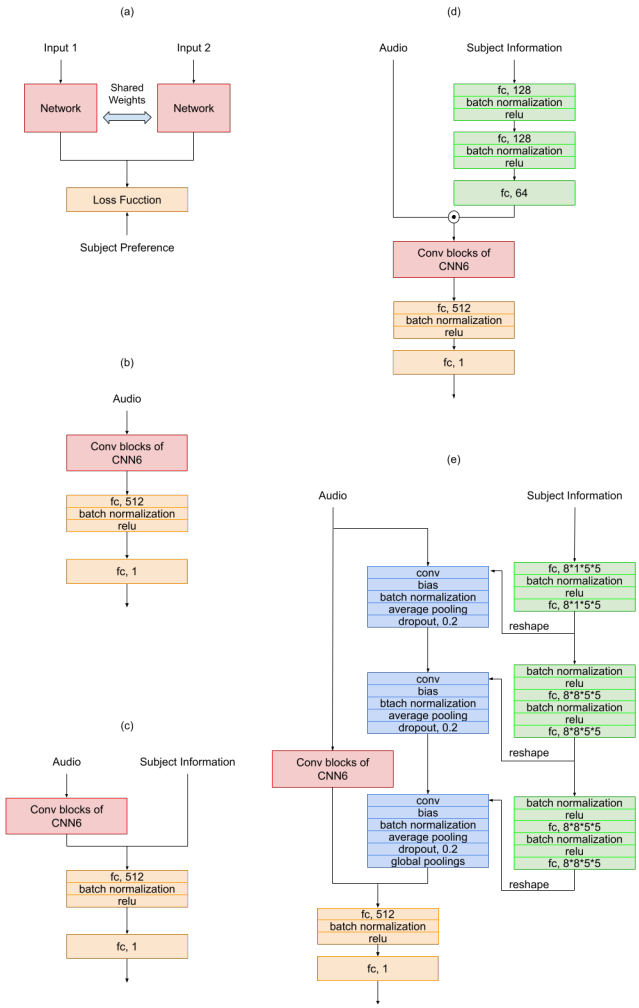

图1。模型结构。(a) 孪生网络。(b) 网络仅使用音频输入（**AO**）。(c) 网络同时使用音频输入和受试者信息，并在CNN6之后进行融合（**A+S-L**）。(d) 音频输入和受试者信息在CNN6之前进行融合（**A+S-E**）。(e) 将受试者信息用作卷积核，与CNN6并行处理音频输入（**A+S-P**）。

3. 方法

由于本文的主要任务是预测两个音频片段之间的偏好，因此采用孪生网络结构，以确保即使改变输入顺序，输出的偏好也保持不变。图1a展示了孪生网络的整体结构，其中输入分别经过两个共享相同权重的网络，两个网络的输出随后进行拼接，并经过softmax函数处理，用于进一步的损失计算。对于孪生网络的两侧，输出较大的一侧表示受试者更偏好相应的输入音频。需要注意的是，该网络经过训练，用于预测一名受试者对两个音频片段音质的偏好，因此输入到两侧的音频片段具有不同的音质，但输入到两侧的受试者信息相同。

图1b展示了该网络，其两侧均仅使用音频特征。我们采用预训练模型“CNN6” inPANNs[11], 作为基本结构，该模型以包含64个梅尔频带的对数梅尔频谱图作为输入。CNN6原有的全连接层被一个多层感知机（MLP）模块取代，该模块包含一个输出512个节点的线性层、一个批归一化层、一个ReLU激活函数，以及一个输出节点的线性层。本文将这一结构命名为**音频质量** 。

图1c展示了第一个在两侧同时使用音频输入和受试者信息的网络。受试者信息包含一个维度的年龄、一个维度的性别，以及四个维度的头戴式耳机或耳塞式耳机规格，即阻抗、频率响应的下限和上限，以及灵敏度。对于头戴式耳机或耳塞式耳机规格的缺失值，将使用 -1 。六维向量与 CNN6 卷积块的输出拼接后，一同输入后续

层。该网络可视为以 CNN6 为编码器、以多层感知机模块为解码器的 LDNet，不同之处在于，输入的受试者信息不只是其 ID；此外，该网络还将与另一个完全相同的网络组合，并视为一个孪生网络。本文将这种结构命名为**A+S-L**（L 表示晚期融合）。

为探究同时使用音频输入和受试者信息的不同方法的可行性，本文还提出了另外两种网络结构。图1d和图1e分别展示了第二种和第三种同时使用音频输入与受试者信息的网络，两者均在网络两侧分别处理这两类信息。在图1d所示的网络中，受试者信息经过多层感知机模块处理后，与音频输入相乘，再送入后续部分。我们将这种结构命名为**A+S-E**，其中 E 代表早期融合。在图1e所示的网络中，受试者信息同样经过若干多层感知机模块处理，但这些模块的输出会经过重塑，并作为输入音频的卷积核；卷积模块的设计与 CNN6 类似。该路径与原始 CNN6 并行，其输出与 CNN6 的输出将拼接在一起，以便进一步处理。我们将这种结构命名为**A+S-P**，其中 P 代表并行。

音频文件的采样率原为 44.1 kHz，使用 CNN6 时重采样为 32 kHz。帧大小和跳步大小分别为 32 ms 和 10 ms。在训练过程中，SpecAugment[12] 在线应用于训练数据进行数据增强，其中时间轴和频率轴的条带数量均为 2，时间轴和频率轴的最大丢弃宽度分别为 64 和 4。负对数似然用作损失函数，并使用 ADAM 优化器优化网络。初始学习率为 $5e-4$，每个 epoch 后将降低至原来的 0.95 倍。批大小为 64。在所有模型中，CNN6 都将进行微调。

4. 实验结果

4.1. 实验设置

为测试预测未见过的受试者偏好的能力，我们将全部 23 名受试者划分为 7 折。所有受试者按年龄排序；当使用索引为 k 的受试者进行测试时，使用索引为 $k+1 \ mod \ 7$ 的受试者进行验证，其余受试者用于训练。训练的 epoch 数为 50；如果验证损失连续 10 个 epoch 未达到新的更低值，训练过程将提前停止。评估指标为准确率，其定义为正确预测偏好的配对数量除以测试配对总数。

实验在一台运行容器的机器上进行，该机器配备 Intel(R) Xeon(R) Gold 6154 CPU，占用 4 个 CPU核心、90 GB 主机内存以及一块 Tesla V100-SXM2 GPU.

4.2 结果

我们首先比较不同模型的性能。7 折的测试准确率和总体准确率如表 1 所示，其中每个模型在每一折上的结果均为 21 次运行的均值，总体准确率则通过对所有折的准确率进行加权平均计算得出。如表所示，**A+S-L** 的平均表现优于**AO** ，准确率从 77.56% 提高到 78.04%。对于各折，7 折中有 5 折的均值更高，6 折的标准差更低，4 折的均值更高且标准差更低。另一方面，**A+S-E** 的表现显著低于其他模型；**A+S-P** 的均值与**AO** 和**A+S-L**相近，但标准差更高，这可能是因为网络更复杂，而数据量并不大。我们没有与采用不同编码器和解码器的其他LDNet结构进行比较，因为目前的数据量并不大，因此可能不适合训练更复杂的网络结构。

为了观察音频输入和受试者信息各自的重要性，我们分别将**A+S-L** 和**A+S-P** 最后一个多层感知机模块中线性层的权重矩阵相乘，并在图 2 中展示其平均绝对值。其中，前 512 个值对应由 CNN6 处理的音频信息，剩余的 6 个或 8 个值则对应**A+S-L**中的原始受试者信息，或**A+S-P**中与音频信息卷积后的受试者信息。对于**A+S-L**，前 512 个值约为 0.4，剩余值约为 0.02。这种尺度差异可能只是因为 CNN6 处理后的音频信息值经过了归一化，而原始受试者信息的值没有经过归一化。因此，音频信息和受试者信息的重要性可能相近。对于**A+S-P**，音频信息和受试者信息都经过了归一化，但相乘后的权重矩阵的平均绝对值处于不同的水平（音频信息为 0.03，受试者信息为 0.01），因此**A+S-P**可能没有很好地利用受试者信息；不过，收集更多数据来训练更复杂的模型或许能提高准确率。

使用不同受试者信息子集时的测试准确率如表 2 所示。当仅使用年龄和性别作为受试者信息时，在大多数折中，模型的表现略逊于使用其他受试者信息子集的情况。另一方面，当使用头戴式耳机或耳塞式耳机的所有规格作为受试者信息时，准确率略高于使用所有可用受试者信息的情况，但标准差也略高。上述结果表明，使用所有可用受试者信息时，模型的表现更加稳定。

折 (#Q) \ 模型	AO	A+S-L	A+S-E	A+S-P
1 (403)	0.8029 ± 0.0151	**0.8056 ± 0.0069**	0.5757 ± 0.0462	0.8098 ± 0.0188
2（269）	0.7836 ± 0.0121	**0.7946 ± 0.0062**	0.6887 ± 0.1071	0.7874 ± 0.0132
3 (318)	**0.7843 ± 0.0118**	0.7808 ± 0.0066	0.6499 ± 0.1241	0.7719 ± 0.0117
4 (215)	**0.6742 ± 0.0134**	0.6670 ± 0.0118	0.5826 ± 0.0698	0.6684 ± 0.0151
5（289）	0.7569 ± 0.0157	**0.7670 ± 0.0091**	0.6753 ± 0.0537	0.7161 ± 0.0539
6（305）	0.7761 ± 0.0063	**0.7864 ± 0.0076**	0.5633 ± 0.1340	0.7775 ± 0.0069
7 (201)	0.8316 ± 0.0171	**0.8420 ± 0.0141**	0.5714 ± 0.0538	0.8380 ± 0.0164
总体	0.7756	**0.7804**	0.6155	0.7699

表 1。不同模型在 7 折上的测试准确率及总体准确率。受试者回答的偏好问题数量以“#Q”表示。每个模型在每折上的结果均为 21 次运行的均值。均值和标准差以“均值±标准差”的形式表示。

折（#Q） \ 主体 Info	all	年龄和性别	所有规格参数 头戴式耳机	阻抗和灵敏度 头戴式耳机的	频率响应 的头戴式耳机
1 (403)	0.8056 ± 0.0069	0.8117 ± 0.0177	0.8060 ± 0.0103	**0.8134 ± 0.0197**	0.8073 ± 0.0154
2 (269)	0.7946 ± 0.0062	0.7775 ± 0.0134	**0.7959 ± 0.0098**	0.7809 ± 0.0104	0.7864 ± 0.0059
3 (318)	0.7808 ± 0.0066	0.7811 ± 0.0108	0.7767 ± 0.0089	**0.7848 ± 0.0224**	0.7774 ± 0.0081
4 (215)	0.6670 ± 0.0118	**0.6712 ± 0.0107**	0.6709 ± 0.0114	0.6668 ± 0.0183	0.6633 ± 0.0135
5 (289)	0.7670 ± 0.0091	0.7330 ± 0.0113	**0.7690 ± 0.0108**	0.7631 ± 0.0255	0.7681 ± 0.0093
6 (305)	0.7864 ± 0.0076	0.7807 ± 0.0071	**0.7866 ± 0.0064**	0.7672 ± 0.0121	0.7831 ± 0.0070
7 (201)	**0.8420 ± 0.0141**	0.8224 ± 0.0138	0.8398 ± 0.0094	0.8403 ± 0.0181	0.8400 ± 0.0138
总体	0.7804	0.7721	**0.7806**	0.7771	0.7782

表 2。使用不同受试者信息子集得到的**A+S-L** 测试准确率。**impd**：阻抗。**sensit**：灵敏度。**freq**：频率。

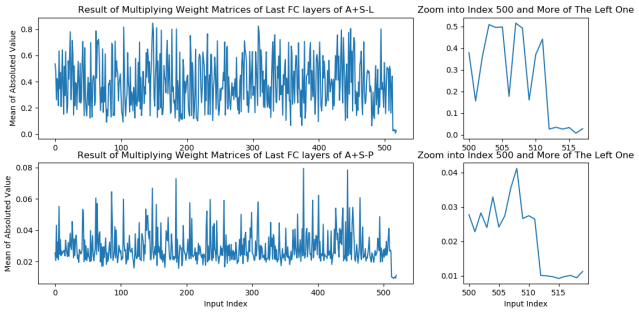

图 2。**A+S-L** （上图）和**A+S-P** （下图）最后一个多层感知机模块中线性层权重矩阵乘积结果的平均绝对值，左图和右图分别显示所有数值，以及索引 500 及之后的数值。

5. 结论与未来工作

本文提出结合音频输入和受试者信息，对内容相同但质量不同的两个音频片段进行偏好个性化预测。本文使用孪生网络比较两组输入，输出较大的一侧表示对应的输入音频更受偏好。对于每一侧的在孪生网络中，研究了几种不同结构的性能，其中，采用 CNN6 [11] 作为编码器、采用多层感知机模块作为解码器的 LDNet [9] 相较于仅使用音频输入的基线模型表现最佳，整体准确率从 77.56% 提升至 78.04%。同时，还研究了不同类型受试者信息的有效性，结果发现，使用所有可获得的受试者信息（包括年龄、性别以及头戴式耳机或耳塞式耳机的规格）时，模型表现最为稳定。由于目前大多数受试者的年龄集中在较小的范围内，未来招募更多不同年龄的受试者，可能有助于更好地讨论年龄对于预测个性化音频质量偏好的有效性。此外，由于聆听设备的环境及其他信息（包括头戴式耳机或耳塞式耳机的使用年限和电量）也可能影响音频偏好，收集此类信息也可能有助于更准确地预测音频偏好。

6. 致谢

本研究得到了华硕电脑股份有限公司的支持。我们还感谢国家高速网络与计算中心（NCHC）提供计算和存储资源。

7. 参考文献

[1] Jinglei Zhou、Rangding Wang、Chao Jin 和 Diqun Yan，《基于相位差检测伪劣音质 WAV 音频》，载于 国际数字水印研讨会。施普林格，2014年，第 525–534 页。[2] Gaoxiong Yi、Wei Xiao、Yiming Xiao、Babak Naderi、Sebastian M¨oller、Wafaa Wardah、Gabriel Mittag、Ross Cutler、Zhuohuang Zhang、Donald S Williamson 等，《ConferenceSpeech 2022 挑战赛：在线会议应用的非侵入式客观语音质量评估（NISQA）挑战赛》，*arXiv* 预印本 *arXiv:2203.16032*，2021年。[3] Brian Patton、Yannis Agiomyrgiannakis、Michael Terry、Kevin Wilson、Rif A Saurous 和 D Sculley，《AutoMOS：学习一种非侵入式语音自然度评估器》，*arXiv* 预印本 *arXiv:1611.09207*，2016年。[4] KV Sriram、H Mahesh Prabhu 和 Aditi Ajay Bhat，《手机易用性及其对品牌忠诚度和再次购买意愿的影响：一项实证研究》，载于 *2019*年 *IEEE* 国际 *WIE* 电气与计算机工程会议（*WIECON-ECE*）。IEEE，2019年，第 1–4 页。[5]Pranay Manocha、Zeyu Jin 和 Adam Finkelstein，《Sqapp：通过成对偏好进行无参考语音质量评估》，载于 *ICASSP 2022-2022 IEEE* 声学、语音与信号处理国际会议（*ICASSP*）。IEEE，2022年，第 891–895 页。[6] Yassir Saquil、Da Chen、Yuan He、Chuan Li 和 Yong-Liang Yang，《用于个性化视频摘要的多成对排序网络》，载于 *IEEE/CVF* 国际计算机视觉会议论文集，2021年，第 1718–1727 页。

[7] Liuchang Xu、Ye Zheng、Dayu Xu 和 Liang Xu，《预测对悲伤音乐的偏好：性别、人格与音频特征的作用》，*IEEEAccess*，第 9 卷，第 92952–92963 页，2019年。[8]Yichong Leng、Xu Tan、Sheng Zhao、Frank Soong、Xiang-Yang Li 和 Tao Qin，《MBNet：采用均值偏差网络预测合成语音的 MOS》，载于*ICASSP 2021-2021 IEEE* 国际声学、语音与信号处理会议（*ICASSP*）。IEEE，2019年，第 391–395 页。[9] Wen-Chin Huang、Erica Cooper、Junichi Yamagishi 和 Tomoki Toda，《LDNet：用于合成语音 MOS 预测的统一听众依赖建模》，载于 *ICASSP 2022-2022 IEEE* 国际声学、语音与信号处理会议（*ICASSP*）。IEEE，2021年，第 896–900 页。[10] Tim Walton 和 Michael Evans，《人类影响因素对整体聆听体验的作用》，《质量与用户体验》，第 3 卷，第 1 期，第 1–16 页，2018 年。[11] Qiuqiang Kong、Yin Cao、Turab Iqbal、Yuxuan Wang、Wenwu Wang 和 Mark D Plumbley，《PANNs：用于音频模式识别的大规模预训练音频神经网络》，*IEEE/ACM* 音频、语音与语言处理汇刊，第 28 卷，第 2880–2894 页，2020 年。[12] Daniel S Park、William Chan、Yu Zhang、Chung-Cheng Chiu、Barret Zoph、Ekin D Cubuk 和 Quoc V Le，《SpecAugment：一种简单的自动语音识别数据增强方法》，*arXiv* 预印本 *arXiv:1904.08779*，2019年。