arXiv:2303.00949v1 [eess.AS] 2 Mar 2023

使用麦克风阵列和耳机进行实时音视频增强

´Jacob Kealey、Anthony Gosselin、Etienne Deshaies-Samson、Francis Cardinal、F´elix Ducharme-Turcotte、Olivier Bergeron、Am´elie Rioux-Joyal、J´er´emy B´elec、Franc¸ois Grondin

摘要——本文介绍了一套完整的软硬件流水线，可在嘈杂和混响环境中实现实时语音增强。该设备由安装在眼镜上的麦克风阵列和摄像头组成，并连接到一个嵌入式系统。系统对语音进行增强，并通过耳机播放音频，延迟最长为 120 毫秒。所提出的方法依靠人脸检测、跟踪和验证，利用波束形成器和后滤波神经网络增强目标说话者的语音。结果证明了该方法的可行性，并为探索和验证各种实时语音增强波束形成器及语音增强方法打开了大门。

I. 引言

机器人听觉旨在赋予机器人与人类相似或更强的听觉能力。这通常涉及在嘈杂和混响环境中进行语音增强，同时以低成本硬件实现在线处理，并将延迟控制在较低水平。过去，人们提出了 ManyEars [1],ODAS [2]和 HARK [3]等框架，使机器人能够利用麦克风阵列定位、跟踪和分离声源。另一方面，听力受损者在嘈杂环境中（例如存在人声嘈杂噪声时）难以听清特定对话 [4]。本文提出将机器人听觉领域的一些现有方法加以改进，为听力受损听众带来更好的聆听体验。

波束形成基于特定声源的空间特性，将多个音频通道组合起来，以增强感兴趣的特定声音。如果已知目标声源的到达方向（DoA）和麦克风阵列几何结构 [5]，就可以在时域或频域中轻松实现延迟求和波束形成。它提供了一种低复杂度、低延迟的解决方案，但容易受到竞争声源泄漏以及早期反射造成的频谱染色影响。为处理竞争声源，几何声源分离（GSS）在线最小化代价函数，以增强感兴趣的声源并抑制干扰源[6], [7]。然而，该方法需要知道所有声源（包括目标声源和干扰声源）的到达方向，并且通常会缓慢收敛到某个解，尤其是在声源随时间移动的情况下。使用 GSS 时，在高混响环境中进行处理也会变得极具挑战，因为代价函数需要同时满足自由场传播假设，并确保分离后的声源之间具有统计独立性

本研究得到 NSERC-CREATE/CoRoM 和 NSERC探索资助的支持。
作者供职于舍布鲁克大学电气工程与计算机工程系。

。另一方面，最小方差无失真响应（MVDR）波束形成器旨在使目标声源方向上的增益保持恒定，同时降低来自其他方向的任何竞争声源的影响，因此可以未知干扰源的到达方向 [8]。该方法效率较高，但需要分别估计目标声源和噪声的空间协方差矩阵（SCM）。空间协方差矩阵通常使用深度神经网络估计的时频掩码获得 [9], ，而其统计量通常需要在几秒钟的音频上进行计算。这一限制使 MVDR 不太适合实时场景。广义特征值分解（GEV）波束形成在估计空间协方差矩阵时能够更好地应对不准确性 [10], [11], ，但它同样依赖于几秒钟音频上的统计量，并且涉及在线特征向量分解，这在低成本硬件上仍然具有挑战性。

此外，人们还提出了基于端到端多通道神经网络的方法来实现声源分离。深度聚类可以利用语音的空间和频谱特征，通过聚类分离出各个说话者[12]。然而，该方法仅适用于语音信号，并且需要一定的缓冲，才能进行具有统计意义的聚类。另一种方法是调整 Conv-TasNet 模型 [13] ，使其能够处理多个通道 [14]。虽然该策略能实现高质量的语音分离，但需要超过一百万个参数，并且忽略了目标到达方向，因此无法只聆听特定说话者。

近年来，已有研究使用增强现实头戴设备，对说话者 [15]在混响和噪声环境中的语音进行稳健的在线转录。在这种情况下，最终目标是自动语音识别（ASR），因此可以容忍数秒的延迟。然而，对于实时语音增强，延迟应控制在几十毫秒，以避免唇部运动与反馈的增强语音信号之间出现令人不适的延迟。事实上，研究表明，通常语音与唇部运动之间的不同步不应超过 140 毫秒的延迟 [16]。

本文展示了设计一套软硬件流水线的可行性，该流水线利用一副耳机，以及安装在眼镜镜架上的摄像头和麦克风阵列，实现实时语音增强。软硬件均经过专门设计，以尽量降低延迟；所选算法则可在 NVIDIA J etson Xavier NX 嵌入式计算机上进行在线处理。第二节介绍所提系统的算法和定制硬件。

第三节介绍用于验证所提方法性能的实验设置，第四节讨论实验结果。第五节总结本文，并探讨未来可能的研究方向。

II. 所提系统

图1展示了系统概览。用户佩戴耳机和眼镜，眼镜上配有一个8麦克风阵列（两侧各2个，前方4个）以及一个朝向场景的小型摄像头。摄像头提供视频流，用于进行人脸检测并跟踪目标；用户可以通过在智能手机上运行的Web应用程序选择目标。

A. 人脸检测、跟踪与验证

安装在眼镜上的摄像头从听者的视角拍摄说话者，并提供图像，用于估计目标相对于听者的位置。该过程分为三个步骤，灵感来自DVT方法[17], [18]：1）人脸检测；2）跟踪；3）验证。

人脸检测使用YOLOv5模型扫描光学图像，并返回一个或多个环绕人脸的边界框 [19]。检测可以较低的刷新率（每秒一帧）执行，以发现新加入视觉场景的说话者。跟踪采用核相关滤波器（KCF） [20], ，因为该方法结果稳健且支持高速跟踪（本应用中为每秒20帧），适合计算资源有限的嵌入式系统。听者通过在智能手机上运行的Web应用程序选择要收听的目标人脸。每隔1.5秒，系统使用针对ReSORT算法[21], 训练的ResNet18，将目标说话人的人脸与跟踪到的人脸进行比较，以确保对象恒存性（例如，在人脸暂时被遮挡的情况下）。图像中目标语音对应人脸的跟踪位置以每秒4帧的速率传送至波束形成器模块，以增强目标语音。

B. 延时求和波束形成

音频信号以 16000 个采样点/秒的速率进行采样。每个信号在时域中表示为 $x_m[n]$ ，其中 $m \in \{1,2,\ldots,M\}$ 表示麦克风索引（共 $M \in \mathbb{N}$ 个麦克风）， $n \in \mathbb{N}$ 表示时间索引。这些输入信号通过短时傅里叶变换（STFT）进行变换，帧长为 $N \in \mathbb{N}$，跳长为 $\Delta N \in \mathbb{N}$ 个采样点，并使用正弦窗，变换后的信号记为 $X_m[l,k] \in \mathbb{C}$，其中 $l \in \mathbb{N}$ 表示帧索引， $k \in \{0,1,\ldots,N/2\}$ 表示频率箱索引。在所提系统中，帧长为 $N = 512$，跳长为 $\Delta N = 256$。正弦窗（或根汉宁窗）具有功率互补特性，因此可以在 50% 重叠的情况下进行分析与合成。

每个麦克风 m 与原点（位于麦克风阵列的质心处）之间的到达时间差（TDoA）记为 $\tau_m[l] \in \mathcal{T}$，其中 $\mathcal{T} = [-\tau_{max}, +\tau_{max}]$，而 τ_{max} 表示最大 TDoA 值，该值取决于麦克风之间的距离、采样率和声速。延迟求和波束形成的结果在频域中按如下方式获得：

$$Y[l,k] = \sum_{m=1}^{M} X_m[l,k] \exp\left(j\frac{2\pi k \tau_m[l]}{N}\right), \tag{1}$$

其中 $j = \sqrt{-1}$。虽然仅根据音频信号即可估计到达时间差 [22], [23],但我们仍希望利用视觉反馈来识别所关注的目标。这可以通过使用从光学图像中获得的目标说话人的嘴唇位置，并将像素 xy- 坐标映射到到达时间差来实现。对于由坐标 $(u[l], v[l])$ 表示的每个像素，其中 $u[l] \in \mathcal{U}$，以及 $\mathcal{U} = \{1,2,\ldots,U\}$ 和 $v[l] \in \mathcal{V}$，其中 $\mathcal{V} = \{1,2,\ldots,V\}$，对于宽度为 $U \in \mathbb{N}$ 像素、高度为 $V \in \mathbb{N}$像素的图像，可以用一个函数进行映射。该过程通过简单的校准步骤中的多项式回归实现，如 [24] 所示：

$$f : \mathcal{U} \times \mathcal{V} \mapsto \mathcal{T}^M. \tag{2}$$

通过这一映射，可以按如下方式获得到达时间差：

$$\{\tau_1[l], \tau_2[l], \ldots, \tau_M[l]\} = f((u[l], v[l])), \tag{3}$$

然后将其代入（1）。目标说话人可能会移动，或者听者可能会转动头部：对于每一帧 l，随着像素坐标随时间变化，系统会对此进行处理，并相应地更新到达时间差。

C. 后滤波

虽然波束形成提高了目标声源的增益，但也会受到来自其他方向的干扰泄漏影响。我们提出在最后一步进行后滤波，以实现时频滤波。对于每个麦克风 m，目标语音和干扰噪声分别表示为表达式 $s_m[n] \in \mathbb{R}$ 和 $b_m[n] \in \mathbb{R}$，使得$x_m[n] = s_m[n] + b_m[n]$。语音和噪声干扰的 STFT 分别对应表达式 $S_m[l,k] \in \mathbb{C}$ 和$B_m[l,k] \in \mathbb{C}$，理想比率掩蔽可定义为：

$$C[l,k] = \frac{\sum_{m=1}^{M} |S_m[l,k]|^2}{\sum_{m=1}^{M} |S_m[l,k]|^2 + |B_m[l,k]|^2}, \tag{4}$$

其中 $C[l,k] \in [0,1]$。

目标是从波束形成后的信号中估计该理想比率掩蔽。当两个或更多说话者同时发声时，我们需要解决排列歧义，以区分目标声源与干扰声源。我们将时频域中的总功率定义为：

$$|\hat{Y}[l,k]|^2 = \sum_{m=1}^{M} |X_m[l,k]|^2. \tag{5}$$

与波束形成信号的功率$|Y[l,k]|^2$ 不同，总功率信号 $|\hat{}Y[k,l]|^2$ 忽略了相位信息，其表现如同一个不存在相长干涉或相消干涉的参考信号。这里的策略

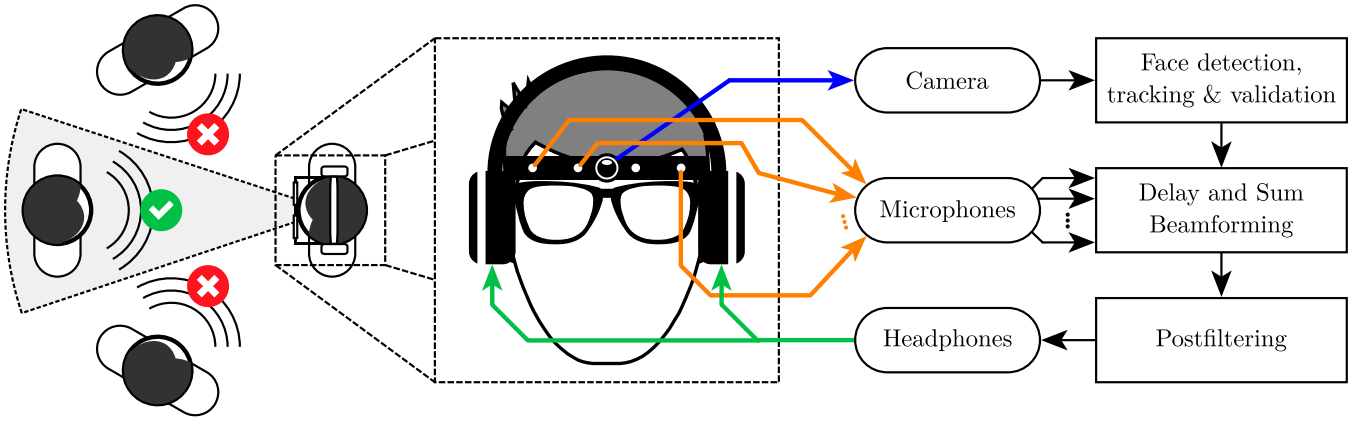


图1：所提系统概览。眼镜上的摄像头捕捉视觉场景，并定位目标人脸。随后将其位置转换为到达方向，供延迟求和波束形成器使用，以增强目标语音。接着进行后滤波，以提高语音质量，并将处理后的信号作为音频反馈传送至耳机。

是使用深度神经网络来估计理想比率掩蔽，具体如下：

$$\hat{\mathbf{C}} = g_{DNN}\{\log(|\mathbf{Y}|^{\circ 2}), \log(|\hat{\mathbf{Y}}|^{\circ 2})\} \tag{6}$$

其中，$\hat{\mathbf{C}} \in [0,1]^{L\times N/2+1}$ 表示对所有 L 帧和 $N/2+1$ 频率箱估计得到的理想比率掩蔽。表达式 $\log(|\mathbf{Y}|^{\circ 2})$ 和 $\log(|\mathbf{Y}|^{\circ 2})$由频谱图构成，其中包含每帧 l 和频率箱 k对应的信号 $\log\ |Y[l,k]|^2$ 和 $\log\ |\hat{Y}[l,k]|^2$。本文提出一种具有 $H = 512$个隐藏单元的单向双层门控循环单元（GRU）神经网络架构。GRU的输出输入到一个线性层，该层将输出映射至与STFT所选帧长相匹配的频率箱数量，并使用Sigmoid函数，因为理想比率掩蔽值位于区间 $[0,1]$内。请注意，为确保此应用中的延迟最小，网络必须具有因果性，这也是采用单向结构的原因。网络通过最小化由频谱图加权的均方误差（MSE）损失函数进行训练：

$$loss = \left\| \mathbf{C} \odot |\mathbf{Y}|^{\circ \mathbf{2}} - \hat{\mathbf{C}} \odot |\mathbf{Y}|^{\circ \mathbf{2}} \right\|_2^2, \tag{7}$$

其中，$\odot$表示Hadamard积，$\|\ldots\|_2^2$对应 l^2-范数。这种加权方式更加重视由大声语音主导的时频区域。

随后可将估计得到的理想比率掩蔽转换为增益 $G[l,k] \in [0,1]$，如下所示：

$$G[l,k] = \sqrt{\hat{C}[l,k]}, \tag{8}$$

其中，$\sqrt{\hat{C}[l,k]}$表示$\hat{\mathbf{C}}$中的每个时频元素。增强信号 $Z[l,k]$ 则对应于：

$$Z[l,k] = G[l,k]Y[k,l]. \tag{9}$$

最后，可以使用逆短时傅里叶变换将其转换回时域，以生成 $z[n] \in \mathbb{R}$。随后，这路音频流会作为增强信号播放到耳机中。

D. 硬件

我们设计了定制硬件，以最低延迟提供音频反馈。如图2所示，音频采集由安装在3D打印眼镜前部和侧面的印刷电路板（PCB）上的八个MEMS麦克风完成。

麦克风通过在印制电路板上使用时分复用（TDM）以菊花链方式连接。实时处理由 Jetson Xavier NX 执行，该设备可安装在听者背负的背包中。因此，需要将高频信号通过约 1 米的距离，从眼镜传输到背包中的 Jetson Xavier NX。为了使信号通过长线传输时具备良好的抗噪性，眼镜上的印制电路板将数字音频信号转换为低压差分信号（LVDS），再由安装在 Jetson 设备上的另一块印制电路板接收。Jetson 印制电路板将 LVDS 信号转换为 TDM，并直接连接到 Jetson Xavier NX 的通用输入输出引脚（I2S通用输入输出引脚，重新配置为支持 TDM）。由于 Jetson 的音频通用输入输出引脚专用于使用 TDM 协议的输入，处理后的输出音频会通过 USB 传输到 USB转AUX适配器，再连接至耳机。系统使用被动噪声隔离（PNI）耳机，因此用户只能听到播放的增强信号。图3总结了从麦克风到耳机的音频信号路径。

从麦克风到播放信号（未经处理）的整体延迟测得为 80 毫秒，而算法延迟为 40 毫秒，总延迟为 120 毫秒。虽然对于实时增强而言，这一延迟仍然较大（但仍低于前文提到的 140 毫秒上限），如果使用通用输入输出引脚生成播放信号，而不是通过 USB 输出将信号转换为发送至耳机的辅助信号，则可以降低 80 毫秒的延迟。目前，这一限制源于 Jetson Xavier NX 在音频通用输入输出引脚数量方面有所不足。

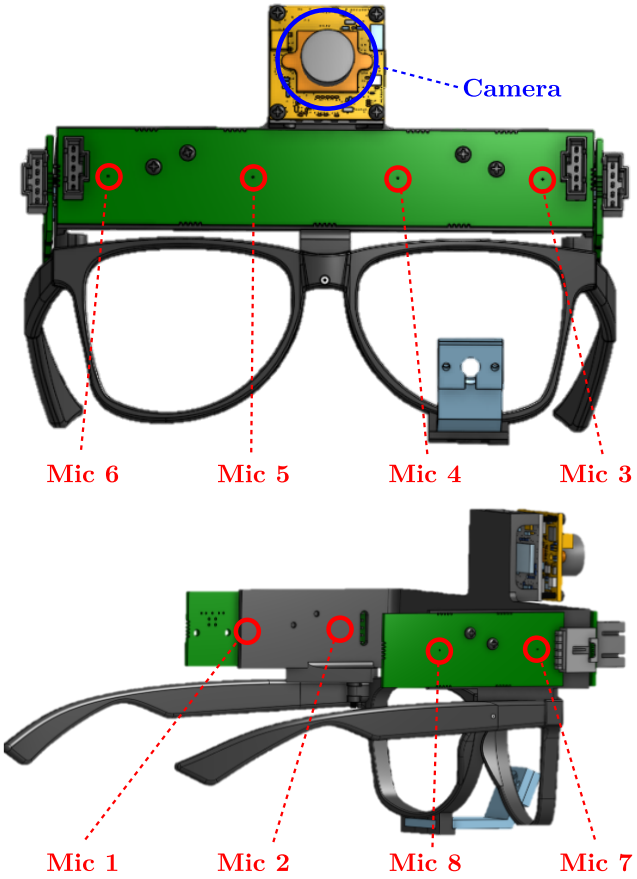


图2：带有印刷电路板和传感器的3D打印眼镜的计算机辅助设计

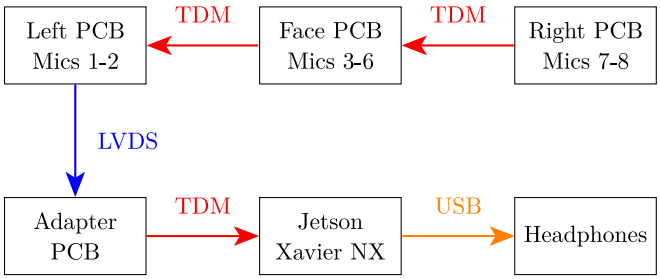


图3：各部分之间的连接方式：将音频信号从麦克风传输到计算机，再将增强信号传输到耳机

摄像头可以通过标准 USB端口连接，并直接接入 Jetson Xavier NX。Jetson模块还运行着一个服务器，提供一个网页，用户可以通过智能手机使用 Wi-Fi 连接到该网页。此应用会显示由安装在眼镜上的摄像头捕获的视频流，并允许用户选择需要随时间进行跟踪且其语音需要增强的说话者面部。

III. 实验设置

所提系统安装在仿真人头上，并在五个不同的房间中分别播放声源，以生成训练数据集。仿真人头在每个房间中分别放置于三个不同的位置。在其中四个房间里，扬声器播放语音（来自 MS-SNSD数据集 [25]），并放置在摄像头视野内的五个随机位置；另一个扬声器播放噪声源（同样来自 MS-SNSD数据集），并放置在仿真人头周围的十三个不同位置。在第五个房间里，语音声源有六个随机位置，噪声源有十个位置；这些声音由参与者产生，而不是通过扬声器播放。每个语音和噪声声源持续十秒，并分别录制，以便之后在离线状态下以不同增益混合，生成信噪比处于 $[0.5, 10]$ dB 区间内的混合信号。总共在5个房间中进行了语音和噪声录制，房间配置各不相同（麦克风阵列、语音声源和噪声源的位置不同），共录得560秒语音和1920秒噪声。将这些录音随机组合，生成包含1至3个干扰源且增益随机的场景，共得到63389组、每组时长为10秒的组合，总计176小时。

在验证过程中，将声源布置在仿真人头周围的16个不同位置，并在图4中用字母 A和 P 表示。

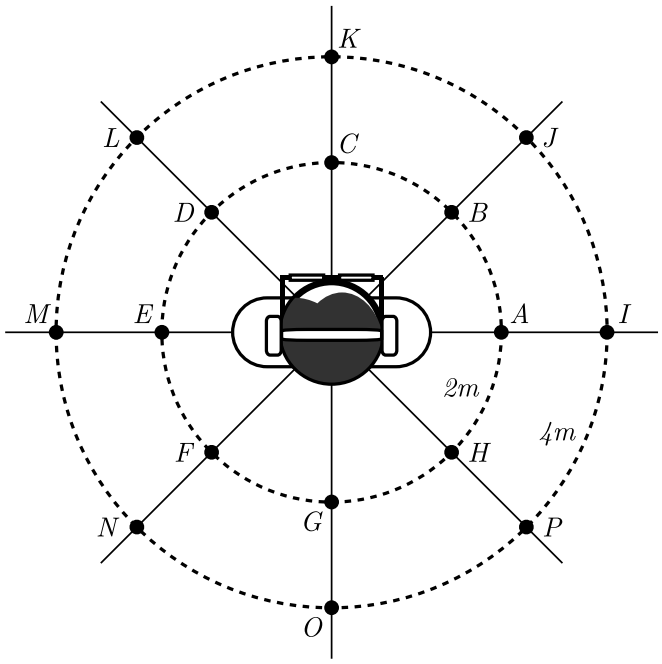


图4：用户周围声源位置的实验设置

扬声器设置在 165 cm 的高度，并在每个位置播放五分钟随机选取的 Librispeech [26]语音样本和 MS-SNSD 非语音样本（A 至 P）。随后，通过混合各段录音生成不同的 10 秒场景，其中目标

声源位于系统视野内的随机位置（位置 B、 C、 D、 J、 K 或 L），而干扰声源则位于另一个位置，这次可能在视野内，也可能在视野外。混合声源时，我们特别注意确保目标声源与干扰声源位于不同位置，且对应的声音片段来自不同的 Librispeech 样本。

IV. 结果与讨论

图 5 展示了麦克风 1 处输入信号（$\mathbf{X}_1$）、麦克风 1 处仅包含目标信号的信号（$\mathbf{S}_1$）以及后滤波信号（$\mathbf{Z}$）的频谱图。部分干扰音素经波束形成和后滤波得到衰减。后滤波后的频谱图中仍然存在干扰，这表明，在干扰源方向具有更小旁瓣的波束形成器和/或效率更高的后滤波神经网络，能够进一步提升性能。

为评估后滤波步骤的影响，我们估计了在存在语音干扰（如表 I 所示）以及语音和非语音干扰（如表 II 所示）的场景下，输入信号和后滤波信号的尺度不变信号失真比（SI-SDR）、语音质量感知评估（PESQ）和短时客观可懂度（STOI）值。

表 I：语音干扰下的分离指标

指标	输入	后滤波
尺度不变信号失真比（dB）	0.70	2.39
PESQ	1.28	1.43
STOI	0.61	0.70

表 II：语音和非语音干扰的分离指标

指标	输入	后滤波
尺度不变信号失真比（dB）	1.27	3.09
PESQ	1.40	1.58
STOI	0.59	0.69

结果表明，该系统改善了语音质量，但仍有进一步改进的空间。例如，可以在假设每个竞争说话者都位于摄像头视野内的情况下，利用每个干扰语音源的到达时间差来探索 GSS 方法。此外，还可以扩大数据集规模，以提升后滤波神经网络的性能。事实上，尽管混合信号总计达 176 小时，但它们全部来自同一组 42 分钟的录制信号。

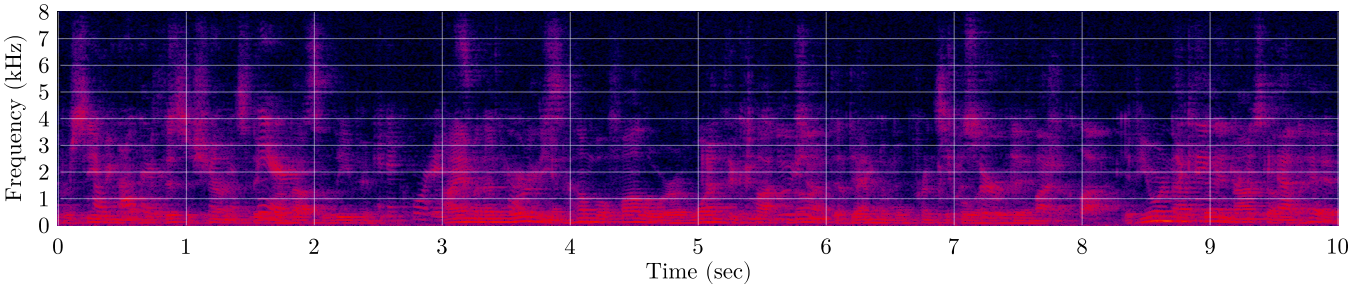
V. 结论

本文展示了在嵌入式硬件上实现完整实时语音增强系统的可行性。所提出的方法使用人脸检测、跟踪和验证来获取目标声源的到达时间差，然后将其输入延迟求和波束形成器。随后，使用神经网络预测的增益对音频信号进行后滤波。

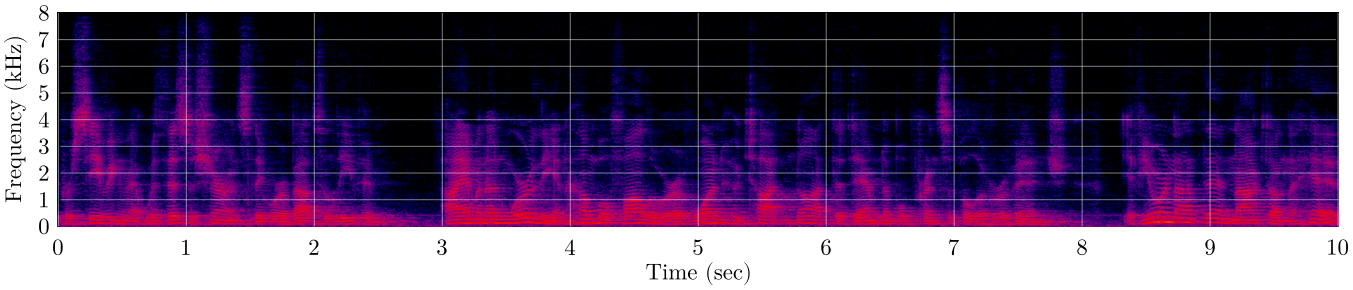
结果表明，作为未来工作，应探索其他类型的波束形成器（例如 GSS），并收集更多数据以提升神经网络的性能。还可以探索用于后滤波的不同神经网络架构（例如长短期记忆网络、残差网络、Transformer），前提是其延迟与当前网络相当或更低。当前的硬件和软件可用于在真实环境中验证这些方法。还可以重新设计硬件，以便为音频播放提供直接输出，从而减少当前使用的 USB连接所引入的延迟。

参考文献

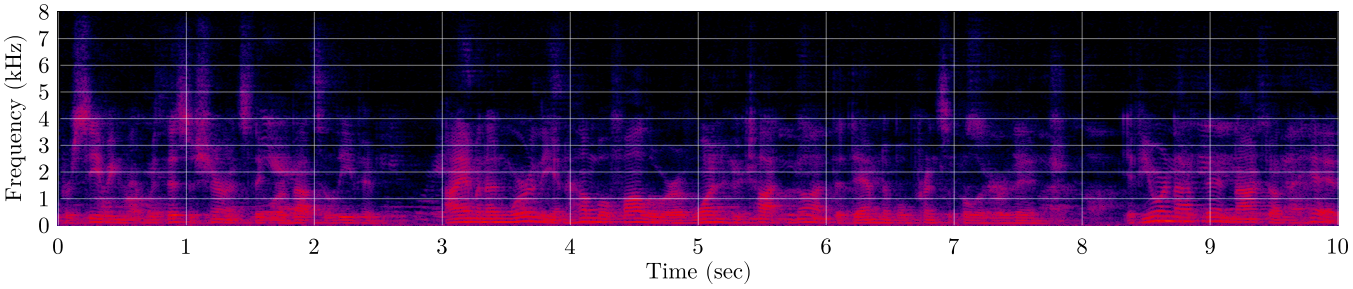
[1] F. Grondin、D. L´etourneau、F. Ferland、V. Rousseau 和 F. Michaud，《ManyEars开放框架：面向机器人应用的麦克风阵列开放软件与开放硬件系统》，《*Autonomous Robots*》，第34卷，第217–232页，2013年。[2] F. Grondin、D. L´etourneau、C. Godin、J.-S. Lauzon、J. Vincent、S. Michaud、S. Faucher 和 F. Michaud，《ODAS：开放式嵌入式听觉系统》，《*Frontiers in Robotics and AI*》，第1–9页。[3] K. Nakadai、T. Takahashi、H. G. Okuno、H. Nakajima、Y. Hasegawa 和 H. Tsujino，《机器人听觉系统“HARK”的设计与实现——用于聆听三位同时说话者的开源软件》，《*Advanced Robotics*》，第24卷，第5–6期，第739–761页，2010年。[4] I. Brons、R. Houben 和 W. A. Dreschler，《噪声抑制对听力受损者语音可懂度、感知聆听努力度和个人偏好的影响》，《*Trends in Hearing*》，第18卷，第1–10页，2014年。[5] V. Perrot、M. Polichetti、F. Varray 和 D. Garcia，《所以你认为自己会做DAS? ——关于延迟求和波束形成的观点》，《*Ultrasonics*》，第111卷，第1–13页，2021年。[6] L. C. Parra 和 C. V. Alvino，《几何源分离：将卷积源分离与几何波束形成相结合》，《*IEEE*语音与音频处理汇刊》，第10卷，第6期，第352–362页，2002年。[7] J.-M. Valin、J. Rouat 和 F. Michaud，《基于麦克风阵列声源分离和后滤波器的增强型机器人听觉》，载于《*IEEE/RSJ*智能机器人与系统国际会议论文集》，第3卷，2004年，第2123–2128页。[8] E. A. P. Habets、J. Benesty、I. Cohen、S. Gannot 和 J. Dmochowski，《室内声学中 MVDR 波束形成器的新见解》，《*IEEE*音频、语音与语言处理汇刊》，第18卷，第1期，第158–170页，2009年。[9] H. Erdogan、J. R. Hershey、S. Watanabe、M. I. Mandel 和 J. Le Roux，《使用单通道掩码预测网络改进MVDR波束形成》，载于《*INTERSPEECH*论文集》，2016年，第1981–1985页。[10] J. Heymann、L. Drude、A. Chinaev 和 R. Haeb-Umbach，《用于第3届CHiME挑战赛的双向长短期记忆网络支持的广义特征值波束形成器前端》，载于《*IEEE*自动语音识别与理解研讨会论文集》，2015年，第444–451页。[11] F. Grondin、J.-S. Lauzon、J. Vincent 和 F. Michaud，《利用在麦克风对上生成的基于到达方向的掩码支持广义特征值波束形成》，载于《*INTERSPEECH*论文集》，2020年，第3341–3345页。[12] Z.-Q. Wang、J. Le Roux 和 J. R. Hershey，《多通道深度聚类：用于说话者无关语音分离的判别性谱和空间嵌入》，载于《*IEEE*国际声学、语音与信号处理会议论文集》，2018年，第1–5页。[13] Y. Luo 和 N. Mesgarani，《Conv-TasNet：超越语音分离的理想时频幅度掩码》，《*IEEE/ACM*音频、语音与语言处理汇刊》，第27卷，第8期，第1256–1266页，2019年。[14] D. Lee、S. Kim 和 J.-W. Choi，《通道间Conv-TasNet用于多通道语音增强》，《*arXiv*预印本 *arXiv:2111 04312*》，2021年。



(a) 输入信号$\mathbf{X}_1$



(b) 目标信号$\mathbf{S}_1$



(c) 后滤波信号$\mathbf{Z}$

图5：输入信号和处理后信号的频谱图。在此示例中，输入信号和后滤波信号的尺度不变信号失真比分别为6.9 dB和12.1 dB。输入信号的感知语音质量评估为1.34，后滤波信号为2.17。输入信号的短时客观可懂度为0.72，后滤波信号为0.88。

[15] K. Sekiguchi、A. A. Nugraha、Y. Du、Y. Bando、M. Fontaine 和 K. Yoshii，《在真实嘈杂对话环境中使用增强现实头戴设备的方向感知自适应在线神经语音增强》，载于《*IEEE/RSJ*智能机器人与系统国际会议论文集》，2022年，第9266–9273页。[16] Q. Summerfield，《唇读与视听语音感知》，《伦敦皇家学会哲学汇刊·*B*辑：生物科学》，第335卷，第1273期，第71–78页，1992年。[17] Z. Kalal、K. Mikolajczyk 和 J. Matas，《Face-TLD：应用于人脸的跟踪—学习—检测》，载于《*IEEE*国际图像处理会议论文集》，2010年，第3789–3792页。[18]K. Zhang、E. Rashedi、E. Barati 和 X.-w. Chen，《使用深度学习进行野外长期人脸跟踪》，《*arXiv*预印本 *arXiv:1805.07646*》，2018年。[19] Q. Xu、Z. Zhu、H. Ge、Z. Zhang 和 X. Zang，《基于YOLOv5和超分辨率重建的有效人脸检测器》，《医学中的计算与数学方法》，第2021卷，第1–9页，2021年。[20] J. F. Henriques、R. Caseiro、P. Martins 和 J. Batista，《使用核相关滤波器进行高速跟踪》，《*IEEE*模式分析与机器智能汇刊》，第37卷，第3期，第583–596页，2014年。[21] T. M. Tran、N. H. Tran、S. T. Duong、H. D. Ta、C. D. T. Nguyen、T. Bui 和 S. Q. Truong，《Resort：一种身份恢复多人人脸跟踪方法》面向监控摄像头”，载于*IEEE*国际自动人脸与手势识别会议论文集，2021年，第1–8页。[22] X. Anguera、C. Wooters和J. Hernando，“用于会议说话人日志的声学波束形成”，《*IEEE*音频、语音与语言处理汇刊》，第15卷，第7期，第2011–2021页，2007年。[23] F. Grondin和J. Glass，“基于SVD-PHAT的多声源定位”，载于*INTERSPEECH*论文集，2019年，第2698–2702页。[24] F. Grondin、H. Tang和J. Glass，“基于SVD-PHAT的二维投影视听校准多项式回归方法”，载于*IEEE*国际声学、语音与信号处理会议论文集，2020年，第4856–4860页。[25] C. K. Reddy、E. Beyrami、J. Pool、R. Cutler、S. Srinivasan和J. Gehrke，“可扩展的噪声语音数据集与在线主观测试框架”，*arXiv*预印本 *arXiv:1909.08050*，2019年。[26] V. Panayotov、G. Chen、D. Povey和S. Khudanpur，“Librispeech：基于公有领域有声读物的自动语音识别语料库”，载于*IEEE*国际声学、语音与信号处理会议论文集，2015年，第5206–5210页。