面向多说话人分离的深度学习麦克风阵列语音增强*

张家扬[1,2],童 峰[1,2,3],陈东升[1,2,3],黄惠祥[1,2]
(1.厦门大学 水声通信与海洋信息技术教育部重点实验室,福建 厦门 361005;
2.厦门大学 海洋与地球学院,福建 厦门 361005;3.厦门大学深圳研究院,广东 深圳 518000)

摘 要:随着近年来人机语音交互场景不断增加,利用麦克风阵列语音增强提高语音质量成为研究热点之一。与环境噪声不同,多说话人分离场景下干扰说话人语音与目标说话人同为语音信号,呈现类似的时、频特性,对传统麦克风阵列语音增强技术提出更高的挑战。针对多说话人分离场景,基于深度学习网络构建麦阵空间响应代价函数并进行优化,通过深度学习模型训练设计麦克风阵列期望空间传输特性,从而通过改善波束指向性能提高分离效果。仿真和实验结果表明,该方法有效提高了多说话人分离性能。

关键词:深度学习;麦克风阵列;波束形成;LSTM
中图分类号:TN912.3 **文献标识码**:A DOI:10.16157/j.issn.0258-7998.212404

中文引用格式:张家扬,童峰,陈东升,等. 面向多说话人分离的深度学习麦克风阵列语音增强[J].电子技术应用,2022,48(5):31-36.
英文引用格式:Zhang Jiayang,Tong Feng,Chen Dongsheng,et al. Deep learning microphone array speech enhancement for multiple speaker separation[J]. Application of Electronic Technique,2022,48(5):31-36.

Deep learning microphone array speech enhancement for multiple speaker separation

Zhang Jiayang[1,2],Tong Feng[1,2,3],Chen Dongsheng[1,2,3],Huang Huixiang[1,2]
(1.Key Laboratory of Underwater Acoustic Communication and Marine Information Technology Ministry of Education,Xiamen University,Xiamen 361005,China;
2.College of Ocean and Earth Sciences,Xiamen Univercity,Xiamen 361005,China;
3.Shenzhen Research Institute of Xiamen Univercity,Shenzhen 518000,China)

Abstract:With the increase of human-computer voice interaction scenes in recent years,using microphone array speech enhancement to improve speech quality has become one of the research hotspots. Different from the ambient noise,the interfering speaker's speech and the target speaker are the same speech signal in the multiple speaker separation scene,showing similar time-frequency characteristics,which poses a higher challenge to the traditional microphone array speech enhancement technology. For the multiple speaker separation scenario,the spatial response cost function of microphone array is constructed and optimized based on deep learning network. The desired spatial transmission characteristics of microphone array are designed through deep learning model training,so as to improve the separation effect by improving the beamforming performance. Simulation and experimental results show that this method effectively improves the performance of multiple speaker separation.
Key words:deep learning;microphone array;beamforming;LSTM

0 引言

随着人与机器之间的语言交互逐渐频繁,更需要考虑噪声、混响和其他说话人的干扰等引起语音信号质量下降的因素对语音识别造成的影响,语音增强技术[1]可以有效地从受干扰的信号中提取纯净的语音,而麦克风阵列比起单个麦克风可以获取更多的语音信息和时空特征,因而麦克风阵列语音增强技术被广泛应用在智能家居、车载系统和音(视)频会议等领域。

麦克风阵列对信号进行空间滤波,可以增强期望方向上的信号并抑制方向性噪声,实现语音增强。传统麦阵语音增强算法:如形成固定波束的滤波累加波束形成算法(Filter-and-Sum Beamforming,FSB)[2],通过一定长度的滤波器系数对多通道信号进行滤波累加,实现了频率无关的空间响应特性,具有低复杂度、硬件容易实现等优点,但是对于具有方向性的噪声效果不佳。

将语音增强构造为有监督学习问题发展出了基于深度学习的语音增强,使用如深度神经网络(DNN)、卷积神经网络(RNN)和长短时记忆网络(LSTM)等利用大数据

* 基金项目:国家自然科学基金项目(11274259);深圳虚拟大学园扶持经费研发机构建设项目(YFJGJS1.0)

量的训练使模型具有语音增强能力。Jiang 等[3]使用 DNN 模型将双耳时间差、双耳水平差和 Gammatone 频率倒谱系数特征输入模型来训练理想二值掩蔽；Xiao 等[4]将多通道信号的广义互相关(GCC)特征送入波束形成网络，得到滤波器权重后作用于信号上获得增强特征，再经过特征提取以及声学模型网络，利用交叉熵函数对各个网络做联合优化，提高自动语音识别(ASR)效果；Ravanelli 等[5]提出新的深度学习框架对标准的联合优化框架做出调整，深度学习框架内的信息可以在语音增强和语音识别模块之间做双向传输，以解决模块不匹配和缺乏沟通问题。

在多说话人分离场景下，目标和干扰同为语音，具有相同的频谱特性，此时可以提取出期望的目标语音的主流方法有波束形成方法、计算听觉场景分析(CASA)、盲源分离和深度学习的分离。其中采用深度学习的分离，如 Huang 等[6]使用 RNN 模型训练两个说话人的分离，在网络模型的输出端连接了时频掩蔽层用于联合训练，同时探讨了区分训练准则，考虑预测信号与其他源信号之间的相似性，获得比 NMF 模型更好的说话人分离效果；Kolbk 等[7]使用 RNN 将说话人跟踪集成到置换不变性训练方法(PIT)中，进一步完成说话人的跟踪和分离，对说话人和语种具有更好的泛化能力。

考虑到麦克风阵列信号具有的空间结构，本文提出了基于深度学习的波束形成器设计和网络框架，通过深度学习实现波束形成，优化期望方向的空间指向特性，减少说话人语音特征的影响，从而对不同方向说话人语音信号进行分离。在多说话人场景下分别进行仿真和实验对所提方法的有效性进行验证。

1 面向多说话人分离的深度学习波束形成器设计

1.1 网络框架

基于深度学习波束形成器的网络框架如图 1 所示，该框架可以分为训练阶段和语音增强阶段。在波束形成器的训练阶段，首先将多通道的两个说话人混合语音通过预处理模块的时频分解和特征提取获取模型的特征输入，将单通道目标语音信号和单通道干扰信号分别做角度的权重控制后叠加，通过时频分解和特征提取后获取模型的训练目标，通过模型训练的方式学习输入和目标的映射函数。基于深度学习的波束形成器训练结束后，在语音增强阶段，对测试语音信号做同样的预处理后输入到训练好的模型中，获得目标方位上的语音幅度谱估计，再经过语音重构模块获得最后的单通道语音增强信号。

模型中的预处理模块包括时频分解和特征提取，首先通过时频分解将时域的多通道混合语音利用短时傅里叶变换(STFT)转化为时频域信号，再转化为傅里叶对数幅度谱(FFT-log-magnitude)以突出高频分量，最后经过 Z-score 标准化保持特征均值为 0，方差为 1，输入模型。

语音增强阶段的语音重构模块的目的为将 FFT-log-magnitude 特征重构回时域信号，即预处理的逆过程，其中由于人耳一般对语音相位不敏感[8]，故可以选择原始信号的相位作为增强信号的相位。

1.2 模型结构

本文在 TensorFlow 开源平台上搭建基于深度学习波束形成器的模型结构如图 2 所示，主要包括输入层、隐藏层、Mask 层和模型输出。多通道信号经过输入层转化为特征送入模型，隐藏层由多层 LSTM 构成，对特征做非线性建模，LSTM 层后级联全连接层，用以估计每个通道的 Mask 函数，全连接层包括多层隐藏层和一层输出层，激活函数为 ReLU 函数，经过全连接层后得到每个通道的 Mask 估计，在 Mask 层将每个通道的 Mask 函数与输入特征相乘后加权平均得到最终模型的预测单通道输出。

1.3 训练目标和损失函数

利用目标信号与干扰信号的方位可以学习期望方向上的空间传输响应。根据方位的不同，对语音信号做权重控制，构建空间传输特性。假设模型对准方向为 θ_{model}，亦表示目标语音方向，干扰语音方向为 $\theta_{interference}$，则干扰信号相对模型对准方向的角度偏差为：

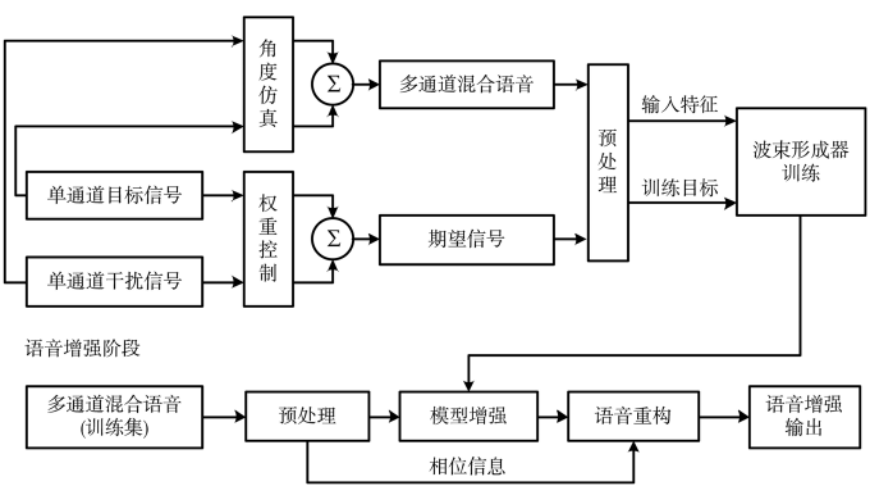


图 1 深度学习波束形成器的网络框架

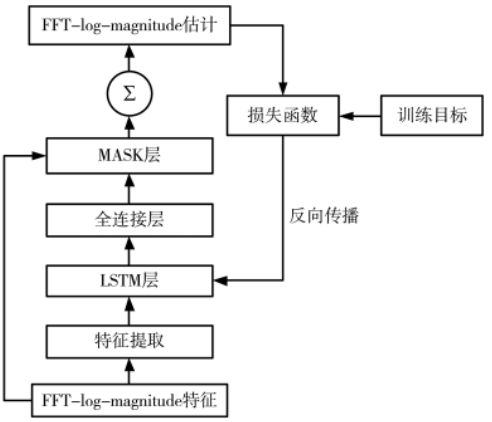


图 2　模型训练框图

$$\Delta\theta=|\theta-\theta_{\text{interference}}| \tag{1}$$

利用得到的角度偏差，依照表 1 进行权重控制。

表 1　期望的空间传输响应

$\Delta\theta$/(°)	$F_d(\Delta\theta)$
0	1
30	0.707
60	0.5
其他	0.001

则期望方向上的语音信号 $s_d(t)$ 为：

$$s_d(t)=F_d(0)s_{\text{target}}(t)+F_d(\Delta\theta)s_{\text{interference}}(t) \tag{2}$$

再通过相同的预处理后得到 Z-score 标准化后的特征作为训练目标 A_d。

$$S_d(t,f)=\int_{-\infty}^{+\infty}s_d(t)W(t-k)e^{-j2\pi ft}dt \tag{3}$$

$$P_d=10\log_{10}(|S_d(t,f)|^2) \tag{4}$$

$$A_d=\frac{P_d-\overline{P_d}}{\sigma(P_d+\text{eps})} \tag{5}$$

其中，$S_d(t,f)$ 表示 $s_d(t)$ 经过 STFT 变换后得到的第 t 个时间帧第 f 个频点的 STFT 系数，$W(t-k)$ 表示对信号的加窗处理，P_d 为 FFT-log-magnitude 特征，A_d 由 P_d 经过 Z-score 标准化得到，σ 表示特征方差，eps 为一个极小常数(避免分母为 0)。

模型估计出 Mask 函数后与多通道混合语音特征做掩蔽再加权平均后得到单通道预测结果，通过该结果与训练目标计算损失函数。本文针对幅度谱的估计任务，采用欧氏距离计算输出与目标之间的损失，并利用 Adam 优化器进行模型参数的更新。

1.4 模型训练设置

1.4.1 数据库描述

模型语料库分为目标语料库和干扰语料库两部分：目标语料库采用文献[9]原始语音数据库数据，含男女各 55 人共 110 人语音信号，语句数目大约为 42 000 句，总时长约为35 h，其中 90 人语句 28.7 h 作为训练集，20 人语句 6.3 h 作为测试集。为了提高模型对不同性别声音的泛化能力，保持训练集和测试集中的男女比例为 1∶1，以减少因男女声音基频差异对模型学习能力的影响。干扰语料库为 TIMIT 语音信号库，包含了 630 人的英文录音数据，每人 10 句共 6 300 句英文语音数据。

1.4.2 仿真参数设置

仿真声源个数为 2，分别作为目标声源和干扰声源。仿真麦克风阵列为直径为 65 mm 的 6 麦圆阵，麦克风均匀分布在圆周上，将空间分为 24 个方向，每 15°一个方向。混响条件下，利用 IMAGE 模型[10]模拟11 m×11 m×3 m 典型办公室尺寸下不同反射强度(0.2、0.4、0.6、0.8)的房间冲激响应，与目标语音和干扰语音分别卷积后得到不同混响强度的语音信号。

以训练对准 0°方向的波束形成器为例，目标语料库单通道信号由 0°方向入射，干扰语料库单通道信号则随机仿真一个角度入射，采样率均为 16 kHz，根据麦克风阵列的时延关系分别仿真出多通道目标语音和多通道干扰语音，与不同房间反射强度的冲激响应做卷积后再按照 0 dB、3 dB、5 dB 的不同信干比叠加，构成多通道混合语音信号。

1.4.3 模型参数设置

滤波累加波束形成器的滤波器阶数设置为 128 阶，方向传输响应设计与表 1 一致。

深度学习波束形成器的 STFT 帧移和帧长分别为 256 个采样点和 512 个采样点，模型输入 257×6 维的傅里叶对数幅度谱特征。模型具有 3 层 LSTM 层，每层具有 256 个细胞；2 层全连接层，每层 512 个神经元；输出层的输出维度为 257×6。模型学习率为 0.001，每经过 100 个 epoch 时学习率衰减为 0.95。

2 深度学习波束形成器仿真结果与分析

本节在多说话人训练集的仿真数据下，从滤波累加波束形成器(FS beamformer)和混合说话人混响条件深度学习波束形成器(DL beamformer)两种不同算法的波束指向性图和识别率结果进行对比，分析算法性能。

2.1 波束指向性

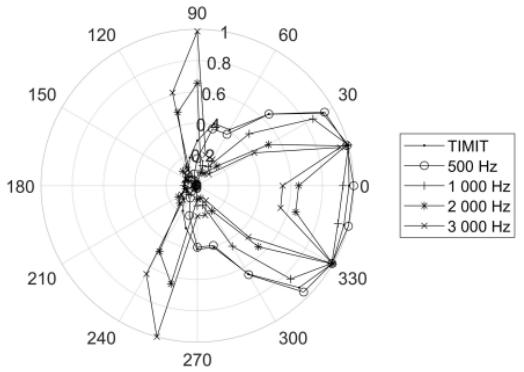
基于 TIMIT 仿真多通道信号得到 FS beamformer 和 DL beamformer 两种算法的各频点波束指向性图如图 3 所示。可以看出，FS beamformer 在非期望方向上依然保持较大的能量，不同频段的抑制一致性不好，在 2 000 Hz 和 3 000 Hz 高频段呈现明显的旁瓣；而 DL beamformer 在期望方向上能量最大，在非期望方向上有明显的抑制，并且展现出更好的不同频段的抑制一致性，也更接近于期望空间传输响应。

2.2 识别率

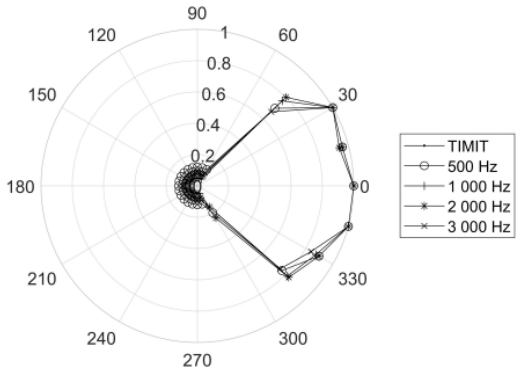
将模型输出的分离语音送入语音识别软件[11]进行文本识别，对比算法处理后的识别文本与标准文本可以得到文本识别率结果，作为评估语音质量的标准，同时可以测试语音增强模型与后端识别系统的适配性。识别率(Recognition Rate，RR)与字错误率(Word Error Rate，WER)的关系为：

$$RR=1-WER \tag{6}$$

在不同信干比条件下(房间反射强度 0.8)，两种算法的识别率结果如表 2 所示。



(a)FS beamformer



(b)DL beamformer

图 3 不同算法波束指向性图(仿真)

表 2 不同信干比下不同算法的识别率结果(仿真)

(%)

SIR/dB	原始混合语音	FS beamformer	DL beamformer
5	42.59	69.93	71.03
3	27.52	51.70	58.76
0	8.45	31.80	35.99

在不同房间反射强度下(信干比为 0 dB),两种算法的识别率结果如表 3 所示。

表 3 不同反射强度下不同算法的识别率结果(仿真)

(%)

反射强度	原始混合语音	FS beamformer	DL beamformer
0.2	10.76	37.49	42.64
0.4	9.59	35.02	42.52
0.6	9.09	31.94	40.81
0.8	8.45	31.80	35.99

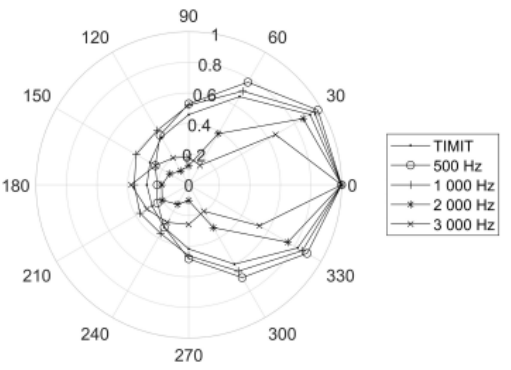
从表 2 可以看出在同一混响条件下,原始信号识别率很低,低信干比下几乎无法识别,FS beamformer 对比原始有较大的提升,而本文的 DL beamformer 模型的识别率结果最好,且信干比越低提升越显著。这是由于随着信干比的降低,FS beamformer 对非期望方向的干扰抑制能力较弱,与具有更好的波束形成能力的 DL beamformer 拉开了差距,DL beamformer 模型学习到了多说话人的语音空间信息,能有效处理多说话人场景。从表 3 可以看出同一信干比条件下,随着房间反射强度的增强,混响程度加大,语音识别率降低,DL beamformer 识别率结果最高,不同反射强度和不同信干比条件下均优于 FS beamformer 算法。

3 实验与结果分析

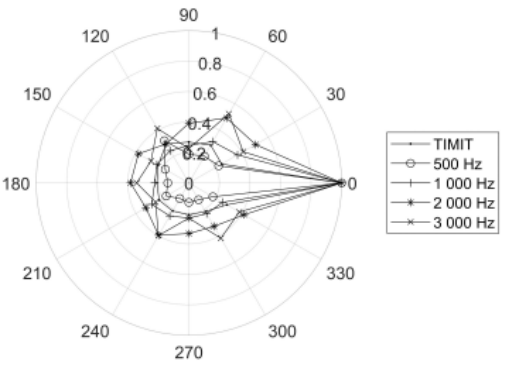
在某大厅采集实测语音数据作为测试集,以评估模型对实际信号的语音增强能力,大厅尺寸为30 m×20 m×6 m,早期混响时间约为 30 ms。实际实验使用 ReSpeaker Far field Mic Array 圆形 6 麦麦克风阵列采集信号,直径为 65 mm,麦克风型号为 STMP34DT01-M。

3.1 波束指向性

为测试模型的实际波束形成能力,将目标声源放置在距离麦阵 5 m 处,使用 Marshall Kilburn 移动音箱播放一段测试集信号,旋转麦克风阵列使得每 30°采集一次信号,全空间采集到的 12 个角度的信号分别经过 FS beamformer 和 DL beamformer 进行处理。以对准 0°方向为例,每种算法得到 12 个角度增强信号后计算总能量和分频点能量如图 4 所示。可以看出在实际测试条件下,两种算法都在期望方向上保持了最大,但 DL beamformer 的波束最窄,体现出了更好的旁瓣抑制效果,相比 FS beamformer 的频率无关性更好。



(a)FS beamformer



(b)DL beamformer

图 4 不同算法波束图(实验)

3.2 识别率

将目标声源和干扰声源分别放置在麦阵 5 m 处的 0°和 180°方位,分别播放目标语料库的测试集和干

扰语料库的 TIMIT 信号，采用 Marshall Kilburn 移动音箱播放语音。采集不同信干比条件下的信号，最后实际得到8.52 dB、5.67 dB 和 2.07 dB 3 种信干比信号。将采集到的信号分别经过 FS beamformer 和 DL beamformer 处理后得到不同信干比下不同算法的识别率结果，如表 4 所示。在 3 种信干比下，DL 波束形成算法相对原始语音识别率分别提升 47.35%、53.43%、48.58%，相对 FS 波束形成语音识别率分别提升 0.6%、0.74%、3.42%。可以看出，DL beamformer 结果略高于 FS beamformer，体现了 DL beamformer 算法在实际测试环境下的语音增强性能。

表 4 不同信干比下不同算法的识别率结果(实验)

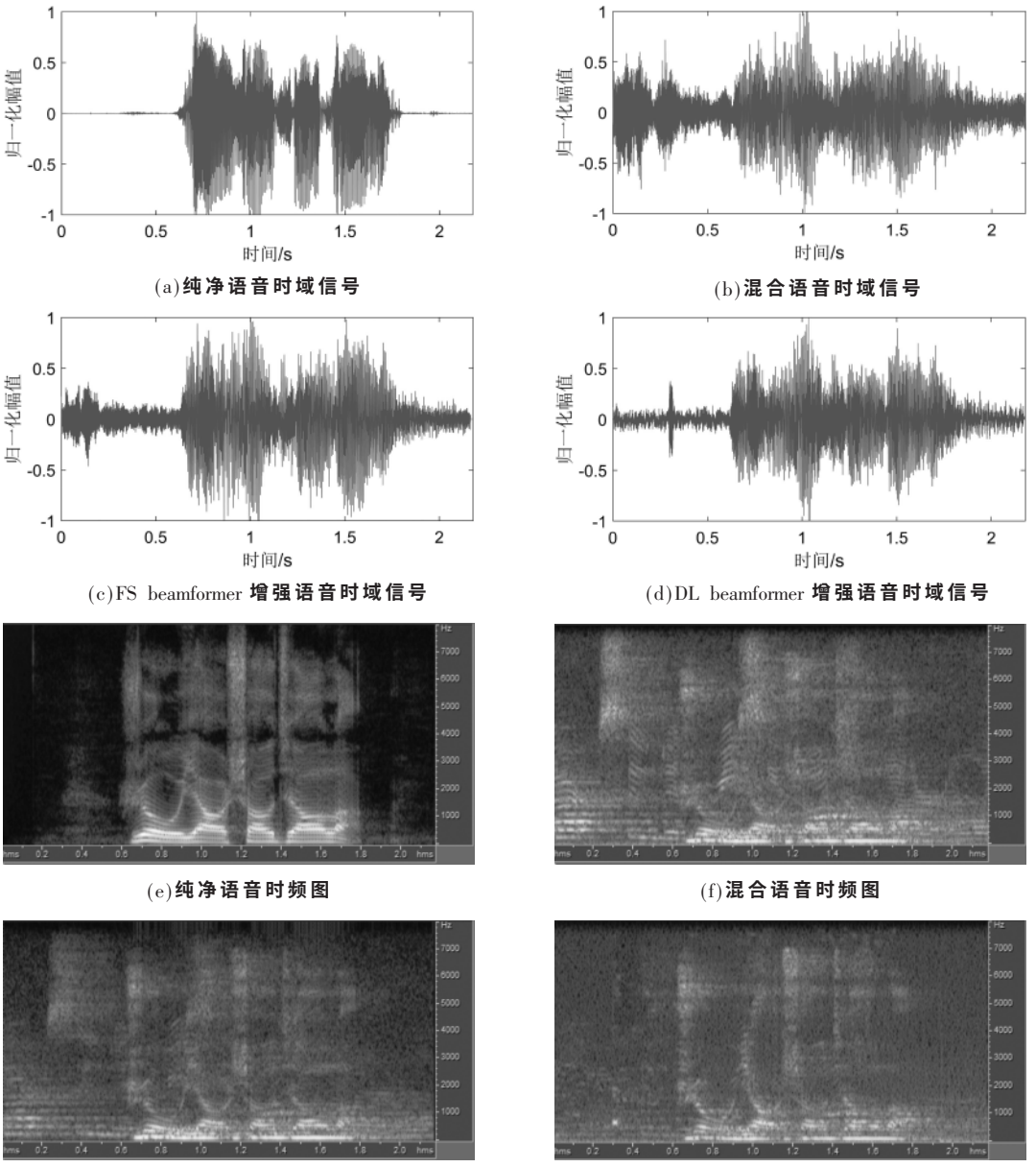
(%)

SIR/dB	原始混合语音	FS beamformer	DL beamformer
8.52	26.46	73.21	73.81
5.67	17.29	69.98	70.72
2.07	9.95	55.11	58.53

但是，与仿真数据结果比较，实际数据结果的提升并不明显，原因可能在于模型训练采用的是仿真数据，且本文训练量较小，对处理实际采集数据时 LSTM 模型的泛化能力造成一定影响。

3.3 时域波形图与时频图结果

以实际采集到的信干比 2.07 dB 的混合语音为例，分别通过 FS beamformer 和 DL beamformer 处理后得到的语音信号时域频域如图 5 所示。由于干扰声源的能量较大，时域和频域上都基本无法识别目标语音，FS beamformer 处理后的信号仍存在干扰语音，而 DL beamformer 对于



(a)纯净语音时域信号　(b)混合语音时域信号

(c)FS beamformer 增强语音时域信号　(d)DL beamformer 增强语音时域信号

(e)纯净语音时频图　(f)混合语音时频图

(g)FS beamformer 增强语音时频图　(h)DL beamformer 增强语音时频图

图 5 不同算法时域波形图与时频图(实验)

非期望方向上的干扰语音抑制更加明显，也因此识别率结果更好。

4 结论

本文针对多说话人分离场景，以期望方向语音信号的 FFT-log-magnitude 作为目标，用来训练出模型在期望方向上的空间传输特性。分别在仿真数据和实测数据测试下与传统波束形成算法对比，本文所提深度学习波束形成器具有更好的波束形成能力，不同信干比和混响条件下语音识别率更高，在非期望方向上展现了更好的干扰抑制效果，验证了深度学习波束形成提高多说话人分离效果的有效性。

也需要指出，由于当前尚缺乏具有一定代表性、数量较大的麦阵数据库，本文模型训练数据较少，影响了所提方法性能改善的充分发挥及评估。下一步将以不同方式进一步扩充训练数据，提高模型泛化性。

参考文献

[1] 朱民雄，闻新.计算机语音技术[M].北京：北京航空航天大学出版社，2002.

[2] KHALIL F，JULLIEN J P，GILLOIRE A.Microphone array for sound pickup in teleconference systems[J].Journal of the Audio Engineering Society，1994，42(9)：691-700.

[3] JIANG Y，WANG D L，LIU R S，et al.Binaural classification for reverberant speech segregation using deep neural networks[J].IEEE/ACM Transactions on Audio Speech & Language Processing，2014，22(12)：2112-2121.

[4] XIAO X，WATANABE S，ERDOGAN H，et al.Deep beamforming networks for multi-channel speech recognition[C]//IEEE International Conference on Acoustics，Speech and Signal Processing，2016：5745-5749.

[5] RAVANELLI M，BRAKEL P，OMOLOGO M，et al.A network of deep neural networks for Distant Speech Recognition[C]//2017 IEEE International Conference on Acoustics，Speech and Signal Processing(ICASSP)，2017.

[6] HUANG P S，KIM M，HASEGAWA-JOHNSON M，et al.Deep learning for monaural speech separation[C]//ICASSP IEEE International Conference on Acoustics，2014.

[7] KOLBK M，Yu Dong，Tan Zhenghua，et al.Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks[J].IEEE/ACM Transactions on Audio，Speech，and Language Processing，2017，25(10)：1901-1913.

[8] WANG D L，CHEN J.Supervised speech separation based on deep learning：An overview[J].IEEE/ACM Transactions on Audio，Speech，and Language Processing，2018，26(10)：1702-1726.

[9] 章宇栋.面向语音交互的麦克风阵列声源定位及波束形成研究[D].厦门：厦门大学，2019.

[10] ALLEN J B，BERKLEY D A.Image method for efficiently simulating small-room acoustics[J].The Journal of the Acoustical Society of America，1998，65(4)：943-950.

[11] HONG Q，LI L，LI M，et al.Modified-prior PLDA and score calibration for duration mismatch compensation in speaker recognition system[C]//Interspeech，2015.

(收稿日期：2021-11-29)

作者简介：

张家扬(1998-)，男，硕士研究生，主要研究方向：麦克风阵列、语音增强。

童峰(1973-)，通信作者，男，博士，教授，主要研究方向：水声通信与网络、声探测与感知、智能语音处理，E-mail：ftong@xmu.edu.cn。

陈东升(1975-)，男，硕士，助理教授，主要研究方向：水声通信、声信号处理。

扫码下载电子文档

(上接第 30 页)

Journal of the Royal Statistical Society，1985，47(3)：528-539.

[15] KARLIN S，TAYLOR H M.A first course in stochastic processes[M].Second Edition Academic Press，1975：15-30.

[16] COPPERSMITH D，WINOGRAD S.Matrix multiplication via arithmetic progressions[J].Journal of Symbolic Computation，1990，9(3)：251-280.

(收稿日期：2021-07-09)

作者简介：

周明升(1981-)，男，博士，高级工程师，主要研究方向：智慧城市、决策支持。

刘抒扬(2000-)，男，本科，主要研究方向：大数据、数据分析和预测。

扫码下载电子文档

版权声明