

DOI: 10.20098/j.cnki.1001-5132.2023.1127

基于核函数领域自适应的语音单声源 DOA 分类

刘明民, 章联军, 叶庆卫, 陆志华*

(宁波大学 信息科学与工程学院, 浙江 宁波 315211)

摘要: 鉴于训练和测试阶段存在不同的噪声或混响环境, 并且由于真实数据的稀缺会降低语音声源到达方向(Direction of Arrival, DOA)的分类准确性, 因此提出一种基于核函数领域自适应的机器学习 DOA 分类算法。通过优化结构风险函数和减小域之间的条件分布差异, 实现对训练数据的适应性学习, 从而提升测试数据的分类准确率。实验结果证明在中小型数据集中, 新算法在各种声学条件下均明显优于对比的深度学习方法。

关键词: 语音信号处理; 分类; 核函数; 领域自适应; 机器学习

中图分类号: TN912.3

文献标志码: A

文章编号: 1001-5132 (2025) 03-0043-09

语音声源到达方向(Direction of Arrival, DOA)分类估计在声学应用中扮演着重要角色, 它包括语音增强、语音识别和会议通话等方面^[1]。在实际环境中, 麦克风阵列通常会接收包含噪声和混响的声源信号, 而这些噪声和混响会对声学的应用性能产生不利影响^[2]。一般情况下, 采用麦克风阵列处理技术来准确估计声源位置, 可减轻这些不利影响, 然而在噪声和混响存在情况下, 实现准确的声源 DOA 分类估计是一项具有挑战性的任务。

近年来, 深度学习在语音声源 DOA 分类中展现出了令人满意的性能^[3]。深度学习中的语音声源 DOA 分类估计将其中声源的空间位置离散成多个对应于不同类别或标签的区域, 例如学习特征(或信号)和 DOA 标签之间的映射关系^[4]。文中算法也采用了这种分类方式来进行 DOA 分类估计, 从输入特征中映射出对应 DOA。数据驱动的深度学习方法在利用神经网络时需要大量真实采集的数据, 但由于涉及实际场景和设备, 真实语音数据采集较为困难, 因此难以获取大量且多样的真实数据, 无法涵盖各种声学场景。所以在中小规模数据下开发一种能够进行鲁棒语音声源 DOA 分类的算法是一项有前景的工作。笔者提出的语音声源 DOA 分

类算法是一种基于核函数领域自适应的机器学习方法, 即 DA-SSL(Domain Adaptation for Sound Source Localization)。在 DOA 分类任务中, 由于数据采集环境的差异或其他因素, 测试数据(目标域)和训练数据(源域)之间都可能存在着分布差异, 也就意味着当将源域模型应用于目标域时, 性能有可能会下降^[5]。文中算法的基本思想是利用核函数在可再生希尔伯特空间(Reproducing Kernel Hilbert Space, RKHS)^[6]来减少源域和目标域之间的分布差异, 通过在 RKHS 中进行特征映射, 将数据映射到高维空间, 使源域和目标域在特征空间中更为接近, 目的是模型能够更好地适应目标域中的特征分布, 并进一步地学习适合目标域中的自适应分类器^[7]。

1 问题描述

文中考虑一个静态语音源被麦克风阵列捕获, 其中, 第 m 个麦克风的接收信号表示为

$$z_m(t) = a_m(t) * s(t) + v_m(t) \quad (1)$$

式中: $z_m(t)$ 表示第 m 个麦克风接收的带噪混响语音信号; $m \in [1, 2]$, 表示麦克风阵列包含 2 个麦克风; $t \in [1, T]$ 表示时间样本索引, T 表示时间样本

收稿日期: 2023-11-23

宁波大学学报(理工版)网址: <http://journalg.nbu.edu.cn/>

基金项目: 国家自然科学基金(61801255)

第一作者: 刘明民, 硕士研究生, 主要研究方向: 语音声源定位。E-mail: 2723144048@qq.com

*通信作者: 陆志华, 副教授, 主要研究方向: 语音信号处理。E-mail: luzhihua@nbu.edu.cn

总数; $a_m(t)$ 表示声源到第 m 个麦克风的声学传递函数; $*$ 表示卷积; $s(t)$ 表示干净语音信号; $v_m(t)$ 表示添加在第 m 个麦克风的噪声。

对式(1)应用短时傅里叶变换(Short-Time Fourier Transform, STFT):

$$z_m(n, f) = a_m(f)s(n, f) + v_m(n, f) \quad (2)$$

式中: $z_m(n, f)$ 、 $a_m(f)$ 、 $s(n, f)$ 、 $v_m(n, f)$ 分别表示时频域中的 $z_m(t)$ 、 $a_m(t)$ 、 $s(t)$ 、 $v_m(t)$; $n \in [1, N]$ 表示时间帧索引; $f \in [1, F]$ 表示频率索引; 其中, N 和 F 分别表示时间帧和频率总数。

根据式(2)的信号模型, 笔者目的是估计语音声源对应的 DOA。相对传递函数(Relative Transfer Function, RTF)^[8]是一种常见的信道间特征, 它完整地编码了语音源位置信息, 因此一般情况下, 瞬时 RTF 可近似为 2 个单时间帧麦克风信号的比值:

$$\hat{g}(f) = \frac{z_2(n, f)}{z_1(n, f)} \quad (3)$$

其中, 第 1 个麦克风被认定为参考麦克风。

协方差减法(Covariance Subtraction, CS)^[9]是一种常用的 RTF 估计方法, 用于获取干净语音的协方差矩阵, 并用以估计带噪语音和噪声的协方差矩阵。然而当噪声非平稳时, 很难估计噪声的协方差矩阵。因此笔者考虑了噪声影响, 并在 CS 估计中用带噪语音的协方差矩阵来替代干净语音的协方差矩阵。当给定连续的 L 个时间帧, 可以得到简化的 CS 估计器, 具体如下:

$$\mathbf{C}_Z = \frac{1}{L} \sum_{n=1}^L \mathbf{Z}(n, f) \mathbf{Z}^H(n, f) \quad (4)$$

$$\hat{\mathbf{g}}_{\text{CS}}(f) = \mathbf{e} \frac{\mathbf{C}_Z \mathbf{e}_1}{\mathbf{e}_1^H \mathbf{C}_Z \mathbf{e}_1} \quad (5)$$

式中: $\mathbf{C}_Z$ 表示 L 帧带噪语音信号 $\mathbf{Z}(n, f)$ 的协方差矩阵; $\mathbf{Z}(n, f) = [z_1(n, f), z_2(n, f)]^T$; $\mathbf{e} \triangleq [0, 1]$; $\mathbf{e}_1 \triangleq [1, 0]^T$; $[\cdot]^H$ 和 $[\cdot]^T$ 分别表示矩阵共轭转置和矩阵转置。而且, 当 $L=1$ 时, $\hat{\mathbf{g}}_{\text{CS}}(f) = \hat{g}(f)$ 。

在文中的声源 DOA 分类问题中, 式(5)估计的 RTF 作为算法输入样本, 用以学习信号特征和离散 DOA 空间之间的映射。给定一个标签源域 $D_s = \{x_{s_i}, y_{s_i}\}_{i=1}^M$ 和无标签目标域 $D_t = \{x_{t_j}\}_{j=M+1}^{M+U}$, 而且假设特征空间 $X_s = X_t$, 标签空间 $Y_s = Y_t$, 但是源域与目标域的条件概率分布不同, 即 $P_s(y_s | x_s) \neq P_t(y_t | x_t)$ 。

2 基于核函数领域自适应的语音声源 DOA 分类算法

RKHS 是具有映射函数 ϕ 的希尔伯特空间^[10]。为学习适用于目标域的分类器, 文中算法通过在 RKHS 中最小化结构风险函数和条件分布来予以优化, 目的是使用有标签源域 D_s 来学习自适应分类器 $f: x_i \rightarrow y_i$, 以预测目标域 D_t 中的标签 $y_i \in Y_t$ 。根据结构风险最小化理论^[11]和正则化理论, 自适应分类器 f 的目标函数可表示为

$$f = \arg \min_{f \in H_K} \sum_{i=1}^M \ell(f(x_i), y_i) + \sigma \|f\|_{H_K}^2 + \lambda D_{f,P} \quad (6)$$

式中: $f(x)$ 和 y 都是实数; $f(x)$ 为预测标签; y 为真实标签。 $\sigma > 0$, $\lambda > 0$, 表示正则化项系数。式(6)中, 第 1 项用于计算标签样本上的损失, 它衡量了分类器在源域数据上的预测性能; 第 2 项是稀疏性约束的正则化项, 用于控制解规模, 以避免过拟合(前两项共同构成了结构风险函数); 第 3 项是条件分布正则化项, 旨在最小化源域和目标域之间的条件分布差异, 以提高适应目标域数据的能力。

2.1 RKHS 和核函数

RKHS 记作 H , 如数据样本 $x \in X$, $\phi(x) \in H$, 且映射函数 ϕ 与核函数 k 相关联。

RKHS 具有以下 3 个主要性质:

- (1) $\forall x_i \in X, \phi(x_i) = k(\cdot, x_i) \in H$,
- (2) $\forall x_i, x_j \in X, k(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle_H$,
- (3) $\forall x_i \in X, \forall f \in H, f(x_i) = \langle k(\cdot, x_i), f(\cdot) \rangle_H$,

其中, $\langle \cdot, \cdot \rangle_H$ 表示 RKHS 上的内积。

核函数 k 必须满足以下 2 个条件:

- (1) 对称性: $\forall x_i, x_j \in X, k(x_i, x_j) = k(x_j, x_i)$,
- (2) 半正定性: 核矩阵 $\mathbf{K}$ 是半正定, 且 $\mathbf{K}_{ij} = k(x_i, x_j)$ 。

高斯核函数是一种常见核函数, 在满足上述 2 个条件时, 可被定义为

$$k(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{\sigma^2}\right) \quad (7)$$

其中, 核带宽系数 σ 可通过中位数启发式(Median Heuristic)^[12]计算。

根据数据矩阵 $\mathbf{X} = [x_1, \dots, x_M, \dots, x_{M+U}]^T \in \mathbb{C}^{n \times D}$, 定义映射矩阵 $\Phi = [\phi(x_1), \dots, \phi(x_M), \dots, \phi(x_{M+U})]^T$,

$\phi(x_i): \mathbb{C}^D \rightarrow H$ 。文中算法是寻求一个分类矩阵 $W: H \rightarrow \mathbb{R}$, 故式(6)的自适应分类器 $f(x_i): \mathbb{C}^D \rightarrow \mathbb{R}$ 可表达为

$$f(x_i) = W^T \phi(x_i) \quad (8)$$

根据核主成分分析(Kernel Principal Component Analysis, KPCA)^[13], 分类矩阵 W 可表示为 Φ 的线性组合, 即 $W = \Phi^T A$, 其中 $A = [A_1, A_2, \dots]^T \in \mathbb{C}^{(M+U) \times 1}$ 是系数矩阵。

所以式(8)可进一步表达为

$$f(x_i) = W^T \phi(x_i) = A^T K_i \quad (9)$$

其中, K_i 表示核矩阵 K 的第 i 列,

$$K \in \mathbb{R}^{(M+U) \times (M+U)}, K_{ij} = K(x_i, x_j)$$

因此, 下一步目标是找到 A 的解, 从而得到式(9)中的分类器 $f(x_i)$ 。

2.2 结构风险函数

笔者的最终目标是学习一个针对目标域 D_t 的自适应分类器。首先在有标签的源域 D_s 上引导一个自适应分类器 f 。采用结构风险最小化原则对 f 进行最小化结构^[10]:

$$f = \arg \min_{f \in H_K} \sum_{i=1}^M \ell(f(x_i), y_i) + \sigma \|f\|_{H_K}^2 \quad (10)$$

根据表示定理^[14], f 可进一步扩展表示为

$$f = \arg \min_{f \in H_K} \sum_{i=1}^{M+U} E_{ii}(\ell(f(x_i), y_i)) + \sigma \|f\|_{H_K}^2 \quad (11)$$

式中: H_K 表示核空间中的一组分类器; E 表示对角标签指示矩阵, 如果 $x_i \in D_s$, $E_{ii} = 1$, 否则 $E_{ii} = 0$ 。其中目标域的标签不可用, 它们会被对角标签指示矩阵 E 过滤掉。而且 $\|f\|_{H_K}^2$ 表示在 H_K 中 f 的平方范数。 $\sigma > 0$, 表示正则化项系数。 ℓ 表示损失函数, 用来衡量 f 在标签训练样本上预测标签的适应性。

这里选择平方损失作为损失函数, 即

$$\ell(f(x_i), y_i) = (y_i - f(x_i))^2$$

所以式(11)的 f 可表示为

$$f = \arg \min_{f \in H_K} \sum_{i=1}^{M+U} E_{ii} (y_i - f(x_i))^2 + \sigma \|f\|_{H_K}^2 \quad (12)$$

将式(9)代入式(11), 得到

$$f = \arg \min_{f \in H_K} \|(Y - A^T K)E\|_F^2 + \sigma \text{tr}(A^T K A) \quad (13)$$

式中: $\|\cdot\|_F$ 为弗罗贝尼乌斯范数(Frobenious norm); $Y = [y_1, \dots, y_{M+U}] \in \mathbb{C}^{1 \times (M+U)}$, 是来自源域和目标域

的标签向量, 目标域标签可任意, 因为它会被对角标签指示矩阵 E 过滤掉; $\text{tr}(\cdot)$ 表示计算矩阵的迹。

2.3 条件分布正则化项

最小化条件分布 $P_s(y_s | x_s)$ 和 $P_t(y_t | x_t)$ 之间的差异对于鲁棒分布自适应非常重要, 然而还存在 2 个问题: (1) 计算 $P_s(y_s | x_s)$ 和 $P_t(y_t | x_t)$ 的非参数统计量非常困难; (2) 由于目标域中的数据样本是无标签, 因此无法对 $P_t(y_t | x_t)$ 直接建模。针对第 1 个问题, 可使用类条件分布 $P_s(x_s | y_s)$ 和 $P_t(x_t | y_t)$ 近似代替条件分布^[15]。针对第 2 个问题, 可使用伪标签赋予目标域数据。为减少初始伪标签的不可靠性, 只在第 1 次迭代时使用基本分类器, 并在后续迭代中使用上一次迭代得到的自适应分类器对目标域数据的标签进行细化。

初始伪标签是通过在源域标签数据上训练一个基本分类器, 然后在目标域的数据中获取。基本分类器可以是最近邻分类器^[16]。由于文中输入样本是复数, 而最近邻分类器只能处理实数, 因此需要通过由高斯核函数计算出的核矩阵 K 得到最终目标域数据的伪标签。具体地说, 如果第 i 个样本和第 j 个样本相似, 那么核矩阵 K 的第 i 列和第 j 列也相似。通过这种方式可以得到具有真实标签的源域数据和具有伪标签的目标域数据, 此时才能对 $P_s(x_s | y_s = c)$ 和 $P_t(x_t | y_t = c)$ 进行类条件分布匹配, 其中每类 $c \in \{1, \dots, C\}$ 。

为客观地度量源域与目标域的条件分布差异, 可使用最大平均偏差(Maximum Mean Discrepancy, MMD)^[17]在可再生的希尔伯特空间上经验度量 $P_s(x_s | y_s = c)$ 和 $P_t(x_t | y_t = c)$ 之间的距离:

$$D_f^{(c)}(P_s, P_t) = \left\| \frac{1}{M_c} \sum_{x_i \in D_s^{(c)}} \phi(x_i) - \frac{1}{U_c} \sum_{x_j \in D_t^{(c)}} \phi(x_j) \right\|_H^2 \quad (14)$$

式中: $\phi: X \rightarrow H$ 是特征映射; $D_s^{(c)} = \{x_i: x_i \in D_s \wedge y(x_i) = c\}$ 是源域数据中属于类别 c 的样本集; $y(x_i)$ 是 x_i 的真实标签; M_c 表示 $D_s^{(c)}$ 的大小。相应地, $D_t^{(c)} = \{x_j: x_j \in D_t \wedge \hat{y}(x_j) = c\}$ 是目标域数据中属于类别 c 的样本集; $\hat{y}(x_j)$ 是 x_j 的伪标签; U_c 表示 $D_t^{(c)}$ 的大小。

为计算与 f 相关的 MMD, 采用投影 MMD^[18]:

$$D_f^{(c)}(P_s, P_t) =$$

$$\left\| \frac{1}{M^{(c)}} \sum_{x_i \in D_s^{(c)}} f(x_i) - \frac{1}{U^{(c)}} \sum_{x_j \in D_t^{(c)}} f(x_j) \right\|_H^2 \quad (15)$$

因此条件分布正则化项表示为

$$D_{f,P} = \sum_{c=1}^C D_f^{(c)}(P_s, P_t) \quad (16)$$

与结构风险函数一样,将式(9)和(14)代入式(15),可得

$$D_{f,P} = \text{tr}(\mathbf{A}^T \mathbf{K} \mathbf{O} \mathbf{K} \mathbf{A}) \quad (17)$$

其中, $\mathbf{O} = \sum_{c=1}^C \mathbf{O}_c$ 是 MMD 矩阵,即

$$(\mathbf{O}_c)_{ij} = \begin{cases} \frac{1}{M_c^2}, & x_i, x_j \in D_s^{(c)} \\ \frac{1}{U_c^2}, & x_i, x_j \in D_t^{(c)} \\ -\frac{1}{M_c U_c}, & \begin{cases} x_i \in D_s^{(c)}, x_j \in D_t^{(c)} \\ x_j \in D_s^{(c)}, x_i \in D_t^{(c)} \end{cases} \\ 0, & \text{其他} \end{cases} \quad (18)$$

2.4 目标函数优化

在得到式(6)目标函数的各项后,通过最小化目标函数来进行求解。将式(13)和(17)整合到式(6),式(6)中 f 重新表述为

$$f = \arg \min_{f \in H_K} \left\| (\mathbf{Y} - \mathbf{A}^T \mathbf{K}) \mathbf{E} \right\|_F^2 + \sigma \text{tr}(\mathbf{A}^T \mathbf{K} \mathbf{A}) + \lambda \text{tr}(\mathbf{A}^T \mathbf{K} \mathbf{O} \mathbf{K} \mathbf{A}) \quad (19)$$

设置一阶微分 $\partial f / \partial \mathbf{A} = 0$, 得到 $\mathbf{A} \in \mathbb{C}^{(M+U) \times 1}$ 解:

$$\mathbf{A} = ((\mathbf{E} + \lambda \mathbf{O}) \mathbf{K} + \sigma \mathbf{I})^{-1} \mathbf{E} \mathbf{Y}^T \quad (20)$$

综上所述,为获取自适应分类器 f , 首先需要根据数据矩阵 $\mathbf{X}$ 和核函数构造核矩阵 $\mathbf{K}$, 然后利用标签源域数据训练一个基本分类器,并对无标签目标域数据进行标签估计,在得到其初始化伪标签后,通过 ρ 次迭代来估计目标域数据的伪标签以及矩阵 $\mathbf{O}$ 。

3 实验结果

3.1 实验数据集

实验使用合成数据,数据通过干净语音信号与真实声脉冲响应卷积而得。干净语音信号选择 LibriSpeech 语料库开发集^[19],真实声脉冲响应则选择 MIR 声脉冲响应数据集^[20]。

MIR 声脉冲响应数据集的房间脉冲响应(Room Impulse Response, RIR)是在一个混响房间内进

测量,房间尺寸为 $6.0 \text{ m} \times 6.0 \text{ m} \times 2.4 \text{ m}$ 。测量时的混响时间分别为 160、360 和 610 ms,采样频率为 48 kHz。使用均匀线性阵列中间的 2 个相邻麦克风进行信号接收,该均匀线性阵列总共包含 8 个麦克风,麦克风间距为 8 cm。声源位于空间网格上的 $-90^\circ \sim 90^\circ$ 范围内,每个间隔 15° ,即总计 13 个不同的 DOA。此外,声源与麦克风阵列间距离为 2 m,麦克风和声源的实验场景如图 1 所示。

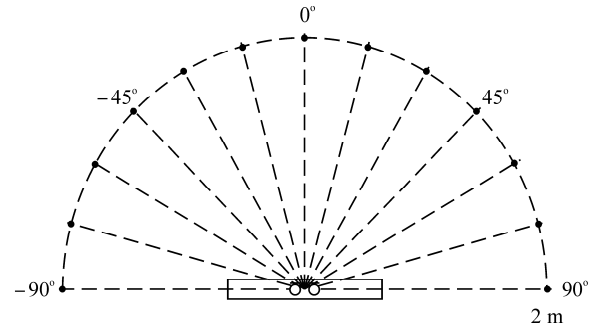


图1 实验场景示意图

LibriSpeech 语料库开发集包含 40 位说话者,具体为 20 位男性和 20 位女性。语音采样频率为 16 kHz。为保持采样频率一致性,将 RIR 降采样到 16 kHz,且 RIR 被截断至 1 s。因为沉默的语音帧不能提供有效 DOA 信息,所以在干净语音与 RIR 卷积前,需要将干净语音中的沉默片段予以删除。使用 WebRTC 语音活动检测(Voice Activity Detection, VAD)方法删除语音中的沉默片段,该方法使用高斯混合模型对语音信号和背景噪声进行建模,并根据相应概率来确定语音和噪声。

选取 40 条去除沉默段的纯净语音,每条时长约 3 s。随机选择其中 20 条,然后将它们与 MIR 声脉冲响应数据集中的 RIR 进行卷积,生成用于训练的混响语音。另外 20 条纯净语音以同样方式生成用于测试的混响语音。接着根据实验设置对训练和测试语音分别加上特定信噪比的背景噪声,其中背景噪声分别来自 NOISEX92^[21]和 DEMAND^[22]数据集。此外还对加噪的训练语音进行短时傅里叶变换,提取一定数量的 RTF 向量特征作为训练集。测试集获取方式与训练集相同。

3.2 对比算法及输入样本

相应对比算法及算法类型比较见表 1。输入样本为 RTF 幅度与相位串联,其具体样本大小均为 $2D \times 1$ 。表 1 中最近邻分类器是传统机器学习分类

算法, 近邻数为 1, 记 1-NN(1-Nearest Neighbor)^[16]。CNN(Convolutional Neural Networks)是基于深度学习的完全监督分类算法^[23], 输入数据是标签数据。VAE-SSL(Variational Auto Encoder-Semi Supervised Learning)是基于深度学习的半监督分类算法^[23], 输入数据是标签数据和无标签数据。

表 1 对比算法及输入样本

对比算法	算法类型
1-NN ^[16]	机器学习
CNN ^[23]	深度学习
VAE-SSL ^[23]	深度学习

由于 VAE-SSL 网络是实值网络, 只能处理实数, 因此需进行适当变换。现有 4 种类型的实值需输入, 可由 L 个连续帧计算得到: (1) 在 VAE-SSL 中将 RTF 堆叠为矩阵, 即堆叠 RTF 相位(大小为 $D \times L$); (2) 堆叠 RTF 的幅值与相位串联(大小为 $2D \times L$); (3) 文中 RTF 是向量, 即 $\hat{g}_{cs}$ RTF 相位(大小为 $D \times 1$); (4) $\hat{g}_{cs}$ RTF 的幅值与相位串联(大小为 $2D \times 1$)。实验表明, 将 $\hat{g}_{cs}$ RTF 的幅值与相位串联作为输入样本时, VAE-SSL 性能最好, 此时 1-NN 和 CNN 的输入与 VAE-SSL 相同。因此相比之下, 笔者提出的 DA-SSL 算法能直接处理复值数据, 可将 $\hat{g}_{cs}$ RTF 直接作为输入样本。

由于输入样本的维度发生变化, 因此需要对深度学习算法中的网络参数进行调整。具体方式是将原先的二维卷积操作转换为一维卷积操作, 其他参数则保持不变, 此做法可以确保网络适应新的输入维度。

3.3 参数设置

用于算法训练的标签数据量为 988, 且 VAE-SSL 训练的无标签数据量为 10 000。根据式(19), 提出基于核函数领域自适应声源 DOA 分类算法中包含 2 个超参数 λ 和 σ , 分别设置为 $\lambda = 5$, $\sigma = 1$ 。算法迭代次数设置为 $\rho = 5$ 。实验中使用帧长为 256、帧移为 128 的汉明窗进行 STFT, 即 $D = F = 129$, 其中, F 为频点数。而且采用 $L = 31$, 和 VAE-SSL 保持一致。

3.4 评价指标

文中使用平均准确率(Accuracy, ACC)来评估这些算法的性能。ACC(参数 R_{ACC})定义为正确预测

的 DOA 标签数量与测试样本总数之比:

$$R_{ACC} = \frac{100\%}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} 1_{\hat{y}_i = y_i} \quad (21)$$

式中: n_{test} 表示测试样本总数; 1 表示指示函数, 当预测 DOA 标签 $\hat{y}_i$ 和真实标签 y_i 相同时, 它等于 1, 否则为 0。测试样本量为 10 000, 对应语音总时长约为 81 s。

3.5 实验结果

实验在训练数据与测试数据声学条件不匹配情况下进行, 不匹配的声学条件包括 3 个类别: 信噪比不匹配、混响时间不匹配、噪声类型不匹配。

在信噪比不匹配实验中(源域和目标域具有不同信噪比), 选择混响时间为 160 ms 的 MIR 数据集, 即 {10, 15, 20} dB, 噪声为高斯噪声, 结果见表 2。表 2 中比如任务 10 dB→15 dB 表示源域数据的信噪比是 10 dB, 目标域数据的信噪比是 15 dB。从平均准确率可见文中 DA-SSL 算法性能最好, 具有较强噪声鲁棒性。此外当源域和目标域的声学环境存在较大差异时, 它们的分布差异也会增大。例如在目标域信噪比为 10 dB 情况下, 选择源域信噪比为 15 dB 时, CNN、VAE-SSL 和 DA-SSL 算法性能会优于选择源域信噪比为 20 dB 的情况。DA-SSL 算法主要通过减少分布差异来提高算法在不匹配信噪比下的性能, 由表 2 可见, 其平均准确率只降低了 1.18 个百分点, 而 CNN 和 VAE-SSL 算法分别降低了 4.14 个百分点和 1.79 个百分点。从平均准确率来看, CNN 和 VAE-SSL 算法较文中 DA-SSL 算法分别降低了 16.78 个百分点和 11.22 个百分点。

表 2 信噪比不匹配下算法的 ACC 单位: %

信噪比任务	1-NN	CNN	VAE-SSL	DA-SSL
10 dB→15 dB	59.31	85.11	87.26	97.92
15 dB→10 dB	44.04	69.17	76.51	94.52
15 dB→20 dB	76.55	93.03	94.43	99.44
20 dB→15 dB	63.84	77.49	87.52	97.84
10 dB→20 dB	74.02	91.93	94.69	99.38
20 dB→10 dB	45.32	65.03	74.72	93.34
平均准确率	60.51	80.29	85.85	97.07

在混响时间不匹配实验中(源域和目标域具有不同的混响时间), 选择 3 种混响时间的 MIR 数据集, 即 {160, 360, 610} ms, 同时添加 20 dB 的高斯

噪声, 实验结果见表 3。从表中可见文中 DA-SSL 算法性能最好, 平均准确率分别高出 CNN 和 VAE-SSL 算法 14.07 个百分点和 13.23 个百分点, 明显优于两种深度学习算法, 表明具有较强的混响鲁棒性。此外, VAE-SSL 算法在混响时间不匹配下的性能跟 CNN 算法相差不多, 不像在信噪比不匹配实验中明显优于 CNN 算法。

表 3 混响时间不匹配下算法的 ACC 单位: %

混响时间任务	1-NN	CNN	VAE-SSL	DA-SSL
160 ms→360 ms	49.73	71.32	72.30	92.24
360 ms→160 ms	58.91	80.55	78.48	81.76
360 ms→610 ms	44.29	73.94	77.30	86.73
610 ms→360 ms	47.12	77.00	77.44	97.32
160 ms→610 ms	37.99	66.53	65.94	90.72
610 ms→160 ms	51.90	68.86	71.74	73.83
平均准确率	48.32	73.03	73.87	87.10

在噪声类型不匹配实验中(源域和目标域具有不同类型的噪声), 选择混响时间为 160 ms 的 MIR 数据集, 即{Pink, m109, Park, Station}, 信噪比为 10 dB, 实验结果见表 4。从表中可见文中 DA-SSL 算法性能最好, 并且明显优于 CNN 和 VAE-SSL 算法, 分别高出 28.43 个百分点和 21.17 个百分点。但 DA-SSL 算法在目标域噪声类型为 Pink 噪声实验中, 算法准确率相比其在该其他噪声类型不匹配的实验中较低, 表明 Pink 噪声可能对该声源 DOA 分类算法性能具有较高干扰性。需注意的是, 1-NN

表 4 噪声类型不匹配下算法的 ACC 单位: %

噪声类型任务	1-NN	CNN	VAE-SSL	DA-SSL
Pink→m109	94.52	61.92	87.14	99.98
m109→Pink	68.76	71.80	65.66	78.84
Pink→Station	94.82	69.07	65.36	99.86
Station→Pink	70.23	57.83	64.43	76.68
Pink→Park	91.64	45.72	71.88	99.20
Park→Pink	74.52	61.02	67.55	86.24
m109→Park	88.58	64.21	69.97	95.06
Park→m109	94.92	66.26	66.35	99.40
m109→Station	94.10	74.00	76.17	98.18
Station→m109	93.97	65.13	78.61	97.98
Park→Station	94.82	69.64	66.33	99.18
Station→Park	89.22	77.38	91.69	94.54
平均准确率	87.53	65.33	72.59	93.76

算法性能明显超过 CNN 和 VAE-SSL 算法。

为验证文中不同算法在中小型数据集中进行 DOA 分类的优势, 进行了不同数量训练样本的实验。实验在不同混响声学条件下进行, 源域混响时间为 160 ms, 目标域为 360 ms。图 2 展示了训练样本数量对算法性能的影响。由图可见, 随训练样本数量的增加, 总体上所有算法的平均准确率都相应提高。当样本数量从 247 增加到 988 时, CNN 算法性能提升明显, DA-SSL 和 1-NN 算法也有较为明显的提升。然而当数量继续增加到 5 200 时, 各算法性能改善并不显著。说明当合成样本数量在实验达到一定上限时, 提供的 DOA 信息已经足够。总体来说, 文中 DA-SSL 算法在不同数量训练样本上表现出的性能都超过了比较的 3 种深度学习算法。

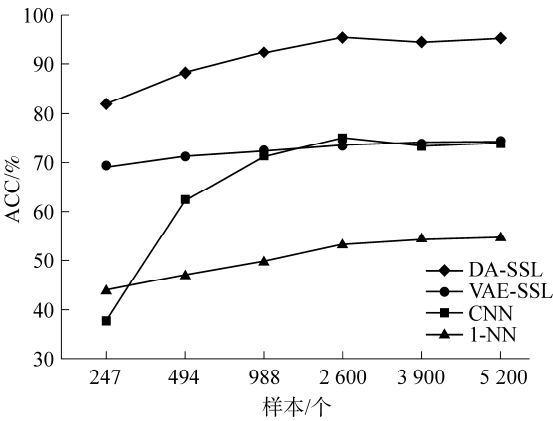


图 2 算法在不同训练样本数量下的 ACC

在信噪比不匹配情况下, 评估了不同算法在不同 DOA 数量下的分类性能。实验采用包含 19 个分辨角度的 DTU 数据集^[24], 该数据集的麦克风阵列由 3 个麦克风组成, 麦克风间距 8.5 cm, 选择其中 2 个相邻麦克风进行实验。声源位于空间网格 -90°~90°范围内, 间隔 10°, 共 19 个不同的 DOA, 声源与麦克风阵列间距离 1.5 m。训练和测试样本的生成方式和 MIR 相同。实验结果如图 3 所示, 其中折线图表示在信噪比不匹配情况下, 算法在 DTU 数据集上的准确率; 其中柱状图表示在信噪比不匹配情况下, 算法在 DTU 数据集与 MIR 数据集上的准确率差异, 即图 3 的折线图和表 2 进行对比。由图可见, 所有算法在 DTU 数据集上的平均准确率明显低于 MIR 数据集, 这是由于 MIR 数据集角度间隔大于 DTU 数据集, 有利于算法进行 DOA 分类。总而言之, 即使分类的 DOA 数量增加, 文中

DA-SSL 算法在 DOA 分类性能上仍然优于对比的其他算法。

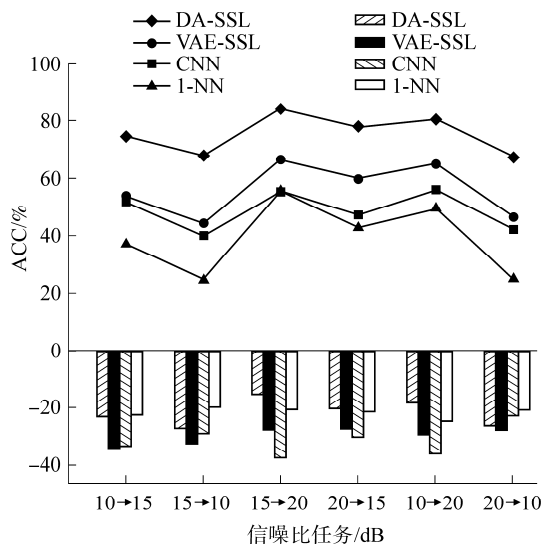


图3 算法在 DTU 数据集中的 ACC 变化

为评估不同算法在 DOA 失配情况下的容忍度, 使用绝对平均误差(Mean Absolute Error, MAE)^[23]进行量化, MAE 越低, 表示算法对于 DOA 失配的容忍度越高。选取 DTU 数据集对算法进行类外测试实验。首先对于类内语音添加 20 dB 高斯噪声, 然后从中提取 988 个特征样本, 构成源域数据。同时, 对类外语音也添加 20 dB 高斯噪声, 并提取特征样本以形成目标域数据。从表 5 结果看, DA-SSL 算法对于 DOA 失配情况下的容忍度依然高于其他对比算法。

表5 DOA 失配时算法的 MAE

项目	MAE	
	1-NN	CNN
类内→类外	13.09	9.74

项目	MAE	
	VAE-SSL	DA-SSL
类内→类外	6.28	5.83

为进一步说明算法的稳健性, 选择在信噪比不匹配情况下的 2 个任务(20 dB→15 dB 和 15 dB→20 dB), 并在 DTU 数据集上进行 100 次实验, 每次实验中随机选取 988 个特征样本作为源域, 同时目标域的特征样本则是通过在测试集语音中引入一定信噪比的随机高斯噪声来生成。结果如图 4 所示。对于 20 dB→15 dB 的任务(图 4(a)), 平均准确

率为 80.26%, 标准差为 1.75%; 对于 15 dB→20 dB 任务(图 4(b)), 平均准确率为 84.61%, 标准差为 1.24%。结果表明, 文中 DA-SSL 算法在信噪比不匹配情况下同样具有一定的稳健性。

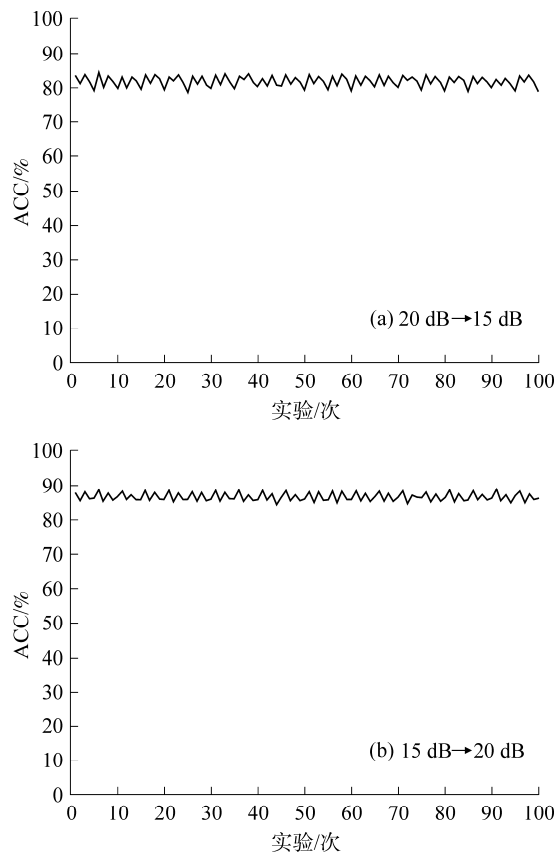


图4 蒙特卡洛实验

3.6 消融实验

为验证条件分布正则化项对文中新算法的重要性, 选择混响时间不匹配时的声学场景进行消融实验, 结果如图 5 所示(图中横坐标对应于表 3 中

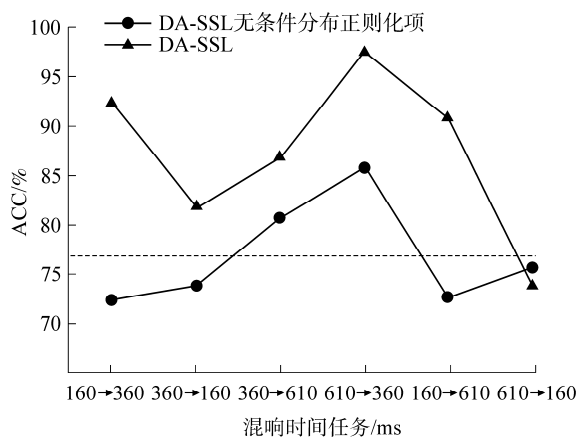


图5 条件分布正则化项的有效性分析

的不同任务)。由图可见,对于每个任务,除源域混响时间为 610 ms,目标域混响时间为 160 ms 任务以外,其他任务在使用条件分布正则化项时,非常有利于 DA-SSL 算法性能的发挥。此外,图 5 中虚线表示 2 种算法的平均准确率,上面虚线表示 DA-SSL 算法,下面虚线表示无条件分布正则化项的 DA-SSL 算法,同样可以看出条件分布正则化项对算法性能影响的重要性。

4 结论

提出了基于核函数领域自适应的机器学习的语音声源 DOA 分类算法,该算法具有鲁棒性,适用于中小规模数据集,可以降低混响和噪声对 DOA 分类的影响,而且减少了对数据量的需求。此外,新算法还利用核函数在可再生希尔伯特空间(RKHS)来减小源域和目标域之间的分布差异,从而使模型能够更好地适应目标域的特征分布。通过学习自适应分类器应用于目标域,新算法能够实现更准确的语音声源 DOA 分类估计。实验结果表明,与其他算法(包括深度学习算法)对比,文中 DA-SSL 算法在 DOA 分类方面表现更优秀。

参考文献:

- [1] GANNOT S, VINCENT E, MARKOVICH-GOLAN S, et al. A consolidated perspective on multimicrophone speech enhancement and source separation[J]. IEEE/ACM transactions on audio, speech, and language processing, 2017, 25(4):692-730.
- [2] WANG D L, CHEN J T. Supervised speech separation based on deep learning: an overview[J]. IEEE/ACM transactions on audio, speech, and language processing, 2018, 26(10):1702-1726.
- [3] SCHMIDT R. Multiple emitter location and signal parameter estimation[J]. IEEE transactions on antennas and propagation, 1986, 34(3):276-280.
- [4] DIBIASE J H, SILVERMAN H F, BRANDSTEIN M S. Robust localization in reverberant rooms[M]// BRANDSTEIN M, WARD D. Microphone arrays. Berlin: Springer, 2001:157-180.
- [5] HE W P, MOTLICEK P, ODOBEZ J M. Adaptation of multiple sound source localization neural networks with weak supervision and domain-adversarial training[C]// IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brighton, UK. IEEE, 2019: 770-774.
- [6] MUANDET K, FUKUMIZU K, SRIPERUMBUDUR B, et al. Kernel mean embedding of distributions: a review and beyond[J]. Foundations and trends in machine learning, 2017, 10(1/2):1-141.
- [7] LONG M S, WANG J M, DING G G, et al. Adaptation regularization: a general framework for transfer learning [J]. IEEE transactions on knowledge and data engineering, 2014, 26(5):1076-1089.
- [8] GANNOT S, BURSHTEIN D, WEINSTEIN E. Signal enhancement using beamforming and nonstationarity with applications to speech[J]. IEEE transactions on signal processing, 2001, 49(8):1614-1626.
- [9] COHEN I. Relative transfer function identification using speech signals[J]. IEEE transactions on speech and audio processing, 2004, 12(5):451-459.
- [10] MÜLLER K, MIKA S, RÄTSCHE G, et al. An introduction to kernel-based learning algorithms[J]. IEEE transactions on neural networks, 2001, 12(2):181-201.
- [11] VAPNIK V N. An overview of statistical learning theory [J]. IEEE transactions on neural networks, 1999, 10(5): 988-999.
- [12] GHIFARY M, BALDUZZI D, KLEIJN W B, et al. Scatter component analysis: a unified framework for domain adaptation and domain generalization[J]. IEEE transactions on pattern analysis and machine intelligence, 2017, 39(7):1414-1430.
- [13] SCHÖLKOPF B, SMOLA A, MÜLLER K R. Nonlinear component analysis as a kernel eigenvalue problem[J]. Neural computation, 1998, 10(5):1299-1319.
- [14] BELKIN M, NIYOGI P, SINDHWANI V. Manifold regularization: a geometric framework for learning from labeled and unlabeled examples[J]. Journal of machine learning research, 2006, 7(1):2399-2434.
- [15] LONG M S, WANG J M, DING G G, et al. Transfer feature learning with joint distribution adaptation[C]// IEEE International Conference on Computer Vision. Sydney, Australia. IEEE, 2013:2200-2207.
- [16] COVER T, HART P. Nearest neighbor pattern classification[J]. IEEE transactions on information theory, 1967, 13(1):21-27.
- [17] SCHÖLKOPF B, PLATT J, HOFMANN T. Analysis of representations for domain adaptation[C]//Advances in Neural Information Processing Systems: Proceedings of the 2006 Conference, Neural Information Processing Systems. 2007:137-144.

- [18] QUANZ B, HUAN J. Large margin transductive transfer learning[C]//Proceedings of the 18th ACM Conference on Information and Knowledge Management. Hong Kong, China. ACM, 2009:1327-1336.
- [19] PANAYOTOV V, CHEN G G, POVEY D, et al. Librispeech: an ASR corpus based on public domain audio books[C]//IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). South Brisbane, Australia. IEEE, 2015:5206-5210.
- [20] HADAD E, HEESE F, VARY P, et al. Multichannel audio database in various acoustic environments[C]//IEEE 14th International Workshop on Acoustic Signal Enhancement (IWAENC). Juan-les-Pins, France. IEEE, 2014:313-317.
- [21] VARGA A, STEENEKEN H J M. Assessment for automatic speech recognition: II. NOISEX-92: a database and an experiment to study the effect of additive noise on speech recognition systems[J]. *Speech communication*, 1993, 12(3):247-251.
- [22] THIEMANN J, ITO N, VINCENT E. The diverse environments multi-channel acoustic noise database (demand): a database of multichannel environmental noise recordings[C]//Proceedings of Meetings on Acoustics. Montreal, Canada. ASA, 2013:19-33.
- [23] BIANCO M J, GANNOT S, FERNANDEZ-GRANDE E, et al. Semi-supervised source localization in reverberant environments with deep generative modeling[J]. *IEEE access*, 2021, 9:84956-84970.
- [24] VERBURG-RIEZU S A, FERNANDEZ-GRANDE E. Room impulse response dataset - ACT, DTU Elektro (011, IEC; plane, sphere)[DB/OL]. [2022-07-12]. https://data.dtu.dk/articles/dataset/Room_Impulse_Response_Dataset_-_ACT_DTU_Elektro_011_IEC_plane_sphere_/14320166/1.

DOA classification of single speech sound sources based on kernel function domain adaptation

LIU Mingmin, ZHANG Lianjun, YE Qingwei, LU Zhihua

(Faculty of Electrical Engineering and Computer Science, Ningbo University, Ningbo 315211, China)

Abstract: Given the existence of different noise or reverberant environments in the training and testing phases, and due to the scarcity of real data, the classification accuracy of the Direction of Arrival (DOA) of speech sound sources shall be reduced. In this study, we propose a machine learning DOA classification algorithm based on kernel function domain adaptation, which achieves adaptive learning of training data by optimizing the structural risk function and reducing the conditional distribution difference between domains, thus improving the classification accuracy of test data. The experimental results demonstrate that the algorithm significantly outperforms the contrasting deep learning algorithms under various acoustic conditions in small and medium-sized datasets.

Keywords: speech signal processing; classification; kernel function; domain adaptation; machine learning

(责任编辑 章践立)