(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国 际 局

(43) 国际公布日
2023 年 11 月 23 日 (23.11.2023)

WIPO | PCT

(10) 国际公布号
WO 2023/221237 A1

(51) 国际专利分类号:
G10L 25/51 (2013.01)

(21) 国际申请号: PCT/CN2022/101361

(22) 国际申请日: 2022 年 6 月 27 日 (27.06.2022)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
202210528373.0 2022年5月16日 (16.05.2022) CN

(71) 申请人: 江 苏 大 学 (**JIANGSU UNIVERSITY**) [CN/CN]; 中 国 江 苏 省 镇 江 市 京 口 区 学 府 路 301号, Jiangsu 212013 (CN)。

(72) 发明人: 毛启容 (**MAO, Qirong**); 中国江苏省镇江市京口区学府路301号, Jiangsu 212013 (CN)。 高利剑 (**GAO, Lijian**); 中国江苏省镇江市京口区学府路301号, Jiangsu 212013 (CN)。 沈雅馨 (**SHEN, Yaxin**); 中国江苏省镇江市京口区学府路301号, Jiangsu 212013 (CN)。 任庆桦 (**REN, Qinghua**); 中国江苏省镇江市京口区学府路301号, Jiangsu 212013 (CN)。 詹永照 (**ZHAN, Yongzhao**); 中国江苏省镇江市京口区学府路301号, Jiangsu 212013 (CN)。 成科扬 (**CHENG, Keyang**); 中国江苏省镇江市京口区学府路301号, Jiangsu 212013 (CN)。

(74) 代理人: 南 京 智 造 力 知 识 产 权 代 理 有 限 公 司 (**NANJING DOYES INTELLECTUAL PROPERTY CO., LTD.**); 中国江苏省南京市江宁区芝兰路18号4幢1202室, Jiangsu 211100 (CN)。

(81) 指定国(除另有指明，要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB,

(54) **Title:** METHOD AND SYSTEM FOR WEAKLY-SUPERVISED SOUND EVENT DETECTION BY USING SELF-ADAPTIVE HIERARCHICAL AGGREGATION

(54) 发明名称: 自适应层次聚合的弱监督声音事件检测方法及系统

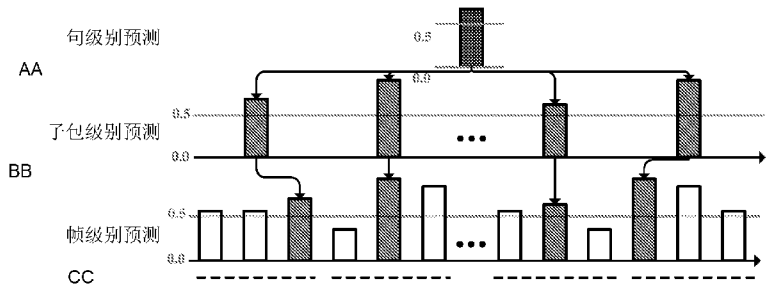

图 2

AA Sentence level prediction
BB Subpacket level prediction
CC Frame level prediction

(57) **Abstract:** A method and system for weakly-supervised sound event detection by using self-adaptive hierarchical aggregation. The system comprises an acoustic model and a self-adaptive hierarchical aggregation algorithm module (5); the acoustic model inputs an audio signal that has undergone preprocessing and feature extraction, the acoustic model predicts to obtain a frame level prediction probability, and the self-adaptive hierarchical aggregation algorithm module (5) aggregates the frame level prediction probability to obtain a sentence level prediction probability; the acoustic model and relaxation parameters are combined and optimized, optimal model weight and optimal relaxation parameters are obtained, and an optimal aggregation policy is formulated for each type of sound event according to the optimal relaxation parameters; and an unknown audio signal that has undergone preprocessing and feature extraction is inputted, frame level prediction probabilities of all target sound events are obtained, and sentence level prediction probabilities of all target sound event categories are obtained according to the optimal aggregation policy for each type of target sound event. The described method is applicable for complex acoustic environments, being applicable for audio classification and positioning in weakly-supervised

WO 2023/221237 A1

[见续页]

GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IQ, IR, IS, IT, JM, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW。

(84) 指定国(除另有指明，要求每一种可提供的地区保护): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SC, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

本国际公布:
— 包括国际检索报告(条约第21条(3))。

sound event detection while also demonstrating strong versatility.

(57) 摘要: 一种自适应层次聚合的弱监督声音事件检测方法及系统，该系统包括声学模型和自适应层次聚合算法模块(5)，声学模型输入预处理和特征提取的音频信号，声学模型预测得到帧级别预测概率，自适应层次聚合算法模块(5)将帧级别预测概率聚合得到句级别预测概率；联合优化声学模型和松弛化参数，得到最优模型权重和最优松弛化参数，根据最优松弛化参数为每类声音事件制定最优聚合策略；输入预处理和特征提取的未知音频信号，得到所有目标声音事件的帧级别预测概率，并根据每类目标声音事件的最优聚合策略，得到所有目标声音事件类别的句级别预测概率。适用于复杂的声学场景，同时适用于弱监督声音事件检测中的音频分类和定位，具有良好的通用性。

自适应层次聚合的弱监督声音事件检测方法及系统

技术领域

本发明涉及人工智能技术中的声音事件检测技术领域，具体涉及一种自适应层次聚合的弱监督声音事件检测方法及系统。

背景技术

弱监督声音事件检测中，最重要的任务之一是设计聚合函数。聚合函数的作用是从模型预测的帧级别概率序列中推断句级别概率，即从预测的“定位信息”推断事件的“类别信息”，从而有效建模弱标注的音频样本。当前主流的聚合函数大致可以分为两类：最大值聚合和加权平均聚合。最大值聚合捕捉信号中最显著的信息，从而为音频分类带来出色的性能。然而，由于最大值聚合检测事件的定位边界总是不完整的，造成较差的定位性能，体现在较多的漏检测；为了解决最大值聚合在音频定位任务上的缺陷，各种类型的加权平均聚合算法发展了起来。加权平均聚合对所有帧级别的概率进行加权平均，以获得句级别的预测，这种考虑所有帧级别概率而不是仅关注显著信息的聚合方式能够召回更多的正样本帧（即事件激活帧），在音频定位子任务中能够取得较好的性能。但与此同时，加权平均聚合也将事件无关的信息考虑进来，给音频分类带了干扰信息，造成次优的音频分类性能。实际上，没有任何单一的聚合方法可以为所有类型的事件提供最优策略。例如，加权平均聚合更适合于持续时间较长的连续事件（如音乐），而对于一些较短的事件（如狗叫），应该考虑使用最大值聚合来关注最显著的音频帧。显然，聚合策略的设计应该自适应声音事件的自然特性。

近年来，研究自适应聚合的方法逐渐被提出，如 McFee 等人提出的自动聚合及 Zhang 等人提出的阶乘聚合都采用自适应的加权 Softmax 聚合方法，即在 Softmax 聚合中将可学习参数乘以帧级别概率，其中不同类别的事件权重不同。然而，这两类自适应方法本质上利用不同的权重调和最大值聚合和加权平均聚合，无法同时有效兼顾音频分类和音频定位任务，且无法高效地为每类事件自适应学习定制的聚合策略，从而导致弱监督声音事件检测性能欠佳。

发明内容

针对现有技术中存在不足，本发明提供了一种自适应层次聚合的弱监督声音事件检测方法及系统，利用分层结构和连续松弛法自动为每类事件学习最优聚合策略，既能够捕捉多片段的显著信息又能保留完善的定位边界，实现同时提高弱监督声音事件检测中音频分类和音频定位的性能。

本发明是通过以下技术手段实现上述技术目的的。

自适应层次聚合的弱监督声音事件检测方法，具体为：

提取预处理音频信号的声学特征，并输入声学模型，将声学模型预测的帧级别预测概率序列分成若干个连续的子包，利用最大值聚合计算每个子包的显著信息，得到子包级预测集合，利用均值聚合取子包级预测集合的平均概率作为句级别预测概率；

联合优化声学模型和松弛化参数，直至收敛，得到最优模型权重和最优松弛化参数，根据最优松弛化参数为每类声音事件制定最优聚合策略；

给定未知的音频信号，进行预处理和特征提取，送入训练后的声学模型，得到所有目标声音事件的帧级别预测概率，实现音频定位任务，并根据每类目标声音事件的最优聚合策略，得到所有目标声音事件类别的句级别预测概率，实现音频分类任务。

进一步，所述制定最优聚合策略具体为：利用$\lambda = \{softmax(\alpha_k); \forall k\}$计算最优松弛化参数下选择不同 R 的概率λ^*，对于第 k 类声音事件，λ_k^*中最大选择概率对应的 R 即为当前类别最优子包数量R_k^*，其中：λ为概率集合，R 为子包数量，α_k为第 k 维松弛化参数，$softmax()$为运算符。

更进一步，所述当子包数量为 R 时，第 k 类声音事件的句级别预测概率表示为：

$$\hat{Y}_k = \phi_{hi}(F_w(X)_k, R) = \phi_{avg}(\{\phi_{max}(b_r); \forall r\})$$

其中：$\hat{Y}_k$为第 k 类声音事件的句级别预测概率，ϕ_{hi}为自适应层次聚合算法，F_w表示声学模型，ϕ_{avg}表示均值聚合，ϕ_{max}表示最大值聚合，b_r为子包集合B中的第 r 个元素。

更进一步，所述第 k 类声音事件的句级别预测期望概率$\bar{Y}_k$为：

$$\bar{Y}_k = \sum_{R \in N^+} \{\lambda_k^R * \phi_{hi}(F_w(X)_k, R)\}$$

其中：λ_k^R表示第 k 类事件选择 R 个子包数量的概率，N^+为所有可选的子包数量的集合。

更进一步，所述联合优化声学模型和松弛化参数采用反向传播进行的：

$$L(W, \alpha; X, Y) = \frac{1}{K} \sum_{k}^{K} BCELoss(\bar{Y}_k, Y_k)$$

其中：L所有声音事件类别的平均预测误差，W、α分别为模型参数和松弛化参数，X、Y分别为模型输入的梅尔频谱特征和句级别标签，Y_k为第 k 类声音事件的句级别标签，$BCELoss$表示二进制交叉熵函数，K 为声音事件的类别总数。

进一步，所述声学模型为任意主流的深度学习模型，声学模型的基准模型为卷积循环神经网络模型。

进一步，提取的特征为梅尔频谱特征。

更进一步，所述声学模型训练和验证采用 DCASE2017 数据集。

更进一步，所述音频信号下采样至 16kHz，帧长和帧移分别设置为 1024、664，分帧后每条信号得到 240 帧样本，梅尔频谱特征为 64 维。

一种自适应层次聚合的弱监督声音事件检测系统，包括依次相连的声学模型和自适应层次聚合算法模块，所述声学模型输入预处理和特征提取的音频信号，所述声学模型预测得到帧级别预测概率，所述自适应层次聚合算法模块将帧级别预测概率聚合得到句级别预测概率。

本发明的有益效果为：

（1）本发明自适应层次聚合算法首先获取多个音频片段中的显著信息，打破了最大值聚合方法只能够捕捉信号中最显著的片段的局限，扩大的定位时的感知区域；其次，仅对多片段的显著信息加权平均得到最终预测，解决了加权平均聚合考虑所有帧信号而带来噪声问题；因此，自适应层次聚合算法具备捕捉多片段显著信息的同时保证完整的定位边界的能力，使其同时适用于弱监督声音事件检测的两个子任务-音频分类和音频定位。

（2）本发明自适应层次聚合利用连续松弛法联合学习模型最优权重及每类声音事件最优聚合策略；较短的声音事件（如“枪声”）通常仅持续一个短时片段，这种情况下最大值聚合往往优于加权平均聚合，此时自适应层次聚合能够自动学习较小的子包数量，即大部分信号帧属于同一个子包，增加最大值聚合的作用比例；而相对于持续时间较长或周期性声音时间（如“音乐”或“警报声”），此时噪声片段较少、事件信息分布于整个长序列，这种情况下加权平均聚合往往优于最大值聚合，此时自适应层次聚合能够自动分配较多的子包数量，即一个子包包含较少的帧信息，增加加权平均聚合的作用比例；自适应聚合实现了根据声音事件的自然属性来定制的最优聚合策略，从而适用于更加复杂的声学场景。

（3）本发明的自适应层次聚合算法设计轻便，仅依赖一组可学习的参数实现，易于高效地嵌入任何声学模型完成弱监督声音事件检测任务。

附图说明

图 1 为本发明所述基于自适应层次聚合的弱监督声音事件检测系统框架图；

图 2 为本发明所述自适应层次聚合算法流程图；

图 3（a）为本发明所述弱监督声音事件检测可视化结果的对比图一；

图 3（b）为本发明所述弱监督声音事件检测可视化结果的对比图二；

图中：1、原始音频信号，2、信号预处理，3、梅尔频谱特征，4、卷积循环神经网络，5、自适应层次聚合算法模块，6、长短期记忆网络，7、卷积层，8、标准化层，9、ReLU 激活层。

具体实施方式

下面结合附图以及具体实施例对本发明作进一步的说明，但本发明的保护范围并不限于此。

如图 1 所示，本发明基于自适应层次聚合的弱监督声音事件检测系统包括依次相连的声学模型和自适应层次聚合算法模块 5，按照音频信号数据流传递过程，依次为信号预处理→声学特征提取→声学模型→层次聚合。信号预处理 2 的过程是将高维、复杂的原始音频信号 1 处理成较低维度、便于后续处理的短时、连续的信号帧序列。声学特征提取过程为每帧样本提取符合人耳特性的梅尔频谱特征 3，初步过滤冗余信息，提升声学模型建模效率。声学模型可以为任意主流的深度学习模型，声学模型的基准模型为卷积循环神经网络模型，本实施例中选用卷积循环神经网络（CRNN）4，卷积循环神经网络 4 由 6 个卷积块和 3 层长短期记忆网络（LSTM）组成，每个卷积块包含卷积层 7、标准化层 8、ReLU 激活层 9。将提取好的梅尔频谱特征 3 序列送入卷积循环神经网络 4 中，即可得到帧级别预测概率序列，即声音事件的定位信息，之后通过自适应层次聚合算法模块 5 计算句级别预测概率，即声音事件的分类信息。

图 2 自下而上表示帧级别预测概率逐渐聚合得到句级别预测概率的过程，自上而下表示反向传播时梯度传播路径，其中灰度表示梯度的大小。

信号预处理 2 首先将原始信号按照特定采样率重采样，采样后首先进行预加重处理，弥补高频分量的能量，之后按照指定帧长进行分帧，得到若干连续的较短的帧样本，最后对每帧样本加窗处理，平滑帧信号，防止能量泄露，得到短时连续的信号帧序列，完成信号预处理过程。具体过程如下：从 DCASE2017 挑战赛提出的大规模弱标注声音事件数据集中选取信号 s，原始信号 s 的采样率为 22.5kHz，下采样至 16kHz 降低复杂度，在卷积循环神经网络 4 接收之前，信号 s 需要进行预处理，增加高频分辨率。DCASE17 数据集中数据时长均为 10 秒，即上述信号 s 共有 160000 个采样点，此时需要进行分帧处理以降低计算复杂度。在本发明中，帧长设置为 1024 个采样点（64 毫秒）、帧移为 664 个采样点（41.5 毫秒），即每帧前后保留 22.5 毫秒的重叠部分以保证帧信号的平滑性。分帧后每条 10 秒的信号包含 240 帧样本：$s' = \{s'_1, s'_2, \ldots, s'_{240}\}$。最后，对每帧样本加窗处理，完成信号的预处理过程。随后，利用短时傅里叶变换将每帧时域信号转换到频率上，利用 64 个梅尔滤波器对每帧信号进行过滤，得到 64 维梅尔频谱特征 3，即，对于每条信号，卷积循环神经网络 4 的输入特征维度为 240*64。

用F_w表示卷积循环神经网络 4，给定输入特征X，即可得到帧级别预测概率：$\hat{y} = F_w(X)$。按照图 2 所示流程，还需利用自适应层次聚合算法ϕ_{hi}将$\hat{y}$聚合成句级别预测概率$\hat{Y}$，从而构造预测误差（公式（6））、训练模型。而卷积循环神经网络 4 测试阶段则根据已经确定的最优模

型权重和最优聚合策略进行前向计算，即可完成未知数据的声音事件检测。具体过程如下：

1）首先，由于每条信号长度为 240 帧，因此，自适应层次聚合中可选的子包数量 R 为 240 的所有因数集合N^+，即$N^+ = \{1,2,3,\ldots,120,240\}$；利用连续松弛法将可选的子包数量的离散搜索空间转换成可优化的连续搜索空间，能够与卷积循环神经网络 4 联合优化，实现自动为每类声音事件选择最优子包数量，即自适应地定制特定事件的最优聚合策略；

2）为每类声音事件设置一组低维的、可学习的松弛化参数对应离散搜索空间中所有元素，利用 Softmax 激活求得搜索空间中所有备选项选择的概率，根据此概率遍历该离散空间，得到每类声音事件的激活期望，实现将离散的搜索空间连续松弛化；具体地：

假设 DCASE2017 数据集中共有 K 种声音事件，搜索空间大小为 N（即N^+中元素个数），利用一组可学习的参数$\alpha \in R^{K\times N}$将该离散搜索空间松弛化，得到选择不同 R 的概率集合λ：

$$\lambda = \{softmax(\alpha_k); \forall k\} \quad (1)$$

其中，α_k为第 k 维松弛化参数，$softmax()$为运算符；

3）以第 k 类声音事件为例，选择N^+中某一个元素作为当前确定的子包数量 R，将帧级别预测概率$\hat{y}$分割至连续的 R 个子包中，得到子包集合$B = \{b_1, b_2, \ldots, b_R\}$，如图 2 所示，每条黑色虚线所包含的帧样本属于同一个子包；利用最大值聚合ϕ_{max}计算每个子包中最大概率值，即最显著的信息，得到子包级预测集合$\hat{p}$：

$$\hat{p} = \{\phi_{max}(b_r); \forall r\} \quad (2)$$

其中，b_r为子包集合B中的第 r 个元素；

4）随后，利用均值聚合ϕ_{avg}取子包级预测集合的平均概率作为最终句级别预测概率 $\hat{Y}$：

$$\hat{Y} = \phi_{avg}(\hat{p}) \quad (3)$$

最终，当子包数量为 R 时，第 k 类声音事件的激活概率（即句级别预测概率）可以表示为：

$$\hat{Y}_k = \phi_{hi}(F_w(X)_k, R) = \phi_{avg}(\{\phi_{max}(b_r); \forall r\}) \quad (4)$$

其中，子包数量 R 决定聚合策略，当子包数量越多时，最大值聚合作用比例越小，均值聚合比例越大，即更多地关注全局信息；当子包数量较少时，最大值聚合作用比例较大，均值聚合比例较少，即更多地关注局部显著信息；自适应层次聚合为每类声音事件自动学习子包数量，实现特定事件定制化的聚合策略；

5）重复过程 3）和 4），遍历N^+中所有可选的子包数量，即$R \leftarrow N^+$，得到所有情况下第 k 类声音事件的句级别预测概率，结合公式（1）所得概率，得到第 k 类声音事件期望的激活概率：

$$\bar{Y}_k = \sum_{R\in N^+}\{\lambda_k^R * \phi_{hi}(F_w(X)_k, R)\} \quad (5)$$

其中，λ_k^R表示第 k 类事件选择 R 个子包数量的概率；

6）利用二进制交叉熵函数（$BCELoss$）计算所有类别的预测误差L，并完成反向传播，联合优化模型参数和松弛化参数（即训练阶段，训练的模型为现有技术）直至收敛：

$$L(W, \alpha; X, Y) = \frac{1}{K}\sum_k^K BCELoss(\bar{Y}_k, Y_k) \quad (6)$$

其中，W、α分别为模型参数和松弛化参数，X、Y分别为模型输入的梅尔频谱特征和句级别标签，Y_k为第 k 类声音事件的句级别标签；

7）以上基于连续松弛法联合优化过程完成后，即可得到最优模型权重W^*以及最优松弛化参数α^*，利用公式（1）即可计算最优松弛化参数下选择不同 R 的概率λ^*；此时，对于第 k 类声音事件，λ_k^*中最大选择概率对应的 R 即为当前类别最优子包数量R_k^*。至此，自适应层次聚合算法完成了为每类声音事件定制一个最优聚合策略ϕ_{hi}^*。利用连续松弛法联合优化，仅需引入一组低维松弛化参数，即可高效完成联合优化；相比于模型参数，松弛化参数数量微乎其微，且优化只需关注声音事件的自然属性，如时长、周期等全局特性，而无需关注高维的信号内容，因此易于收敛，从而引导模型参数在最优聚合策略下快速收敛直全局最优。人工选择子包数量并重复训练卷积循环神经网络也能够找到最优聚合策略，但其计算复杂度高达$O(N^K)$，其中 N 表示搜索空间中可选子包数量大小，K 表示声音事件种类数。本发明利用连续松弛法联合优化，仅需引入一组低维松弛化参数，即可将计算复杂度降低至$O(N)$。

自适应层次聚合是一个由一组独立的、可学习参数控制的独立模块，模块的输入是模型预测的音频定位结果，输出为音频分类结果，自适应层次聚合方法可以方便、有效地嵌入到任何声学模型中实现弱监督声音事件检测；以卷积循环神经网络为基准模型能够同时学习空间和时序上下文特征的多尺度声学特征，为声音事件检测领域的主流模型框架。

以上过程完成了卷积循环神经网络 4 和自适应层次聚合算法的联合优化，在卷积循环神经网络 4 测试阶段，给定未知的声音信号，进行预处理和特征提取后，送入训练后的卷积循环神经网络F_{w^*}中，得到所有待检测目标事件的定位输出（帧级别预测概率），实现音频定位任务，并根据每类事件特定的最优聚合策略ϕ_{hi}^*得到所有类别的激活概率（句级别预测概率），实现音频分类任务。

图 3（a）、（b）为两条典型音频信号的声音事件定位结果可视化，对比方法为最大值聚合和加权平均聚合。其中，最大值聚合仅能够捕捉显著区域而造成定位边界不完整，尤其在检测时长较长的声音事件时（如“火车声”），而加权平均聚合总是带来较多的误检测，尤其当检测较短的或多片段的声音事件时（如“尖叫声”和“鸣笛”）。图 3（a）、（b）中三种典

型声音事件的定位效果均证实：自适应层次聚合不仅能够捕捉多片段的显著信息从而丢弃冗余信息，还能够在降低误检测率同时保证定位边界的完整，实现最优的声音事件检测性能。

基于与自适应层次聚合的弱监督声音事件检测方法相同的发明构思，本申请还提供了一种电子设备，该电子设备包括一个或多个处理器和一个或多个存储器，存储器中存储了计算机可读代码，其中，计算机可读代码当由一个或多个处理器执行时，进行本发明一种基于自适应层次聚合的弱监督声音事件检测方法的实施。其中，存储器可以包括非易失性存储介质和内存储器；非易失性存储介质可存储操作系统和计算机可读代码。该计算机可读代码包括程序指令，该程序指令被执行时，可使得处理器执行任意一种基于自适应层次聚合的弱监督声音事件检测方法。处理器用于提供计算和控制能力，支撑整个电子设备的运行。存储器为非易失性存储介质中的计算机可读代码的运行提供环境，该计算机可读代码被处理器执行时，可使得处理器执行任意一种基于自适应层次聚合的弱监督声音事件检测方法。

应当理解的是，处理器可以是中央处理单元（Central Processing Unit，CPU），该处理器还可以是其他通用处理器、数字信号处理器（Digital Signal Processor，DSP）、专用集成电路（Application Specific Integrated Circuit，ASIC）、现场可编程门阵列（Field-Programmable Gate Array，FPGA）或者其他可编程逻辑器件、分立门或者晶体管逻辑器件、分立硬件组件等。其中，通用处理器可以是微处理器或者该处理器也可以是任何常规的处理器等。

其中，所述计算机可读存储介质可以是前述实施例所述电子设备的内部存储单元，例如所述计算机设备的硬盘或内存。所述计算机可读存储介质也可以是所述电子设备的外部存储设备，例如所述电子设备上配备的插接式硬盘、智能存储卡（SmartMedia Card，SMC）、安全数字（Secure Digital，SD）卡、闪存卡（Flash Card）等。

所述实施例为本发明的优选的实施方式，但本发明并不限于上述实施方式，在不背离本发明的实质内容的情况下，本领域技术人员能够做出的任何显而易见的改进、替换或变型均属于本发明的保护范围。

权 利 要 求 书

1.自适应层次聚合的弱监督声音事件检测方法，其特征在于：

提取预处理音频信号的声学特征，并输入声学模型，将声学模型预测的帧级别预测概率序列分成若干个连续的子包，利用最大值聚合计算每个子包的显著信息，得到子包级预测集合，利用均值聚合取子包级预测集合的平均概率作为句级别预测概率；

联合优化声学模型和松弛化参数，直至收敛，得到最优模型权重和最优松弛化参数，根据最优松弛化参数为每类声音事件制定最优聚合策略；

给定未知的音频信号，进行预处理和特征提取，送入训练后的声学模型，得到所有目标声音事件的帧级别预测概率，实现音频定位任务，并根据每类目标声音事件的最优聚合策略，得到所有目标声音事件类别的句级别预测概率，实现音频分类任务。

2.根据权利要求 1 所述的弱监督声音事件检测方法，其特征在于，所述制定最优聚合策略具体为：利用$\lambda = \{softmax(\alpha_k); \forall k\}$计算最优松弛化参数下选择不同 R 的概率λ^*，对于第 k 类声音事件，λ_k^*中最大选择概率对应的 R 即为当前类别最优子包数量R_k^*，其中：λ为概率集合，R 为子包数量，α_k为第 k 维松弛化参数，$softmax()$为运算符。

3.根据权利要求 2 所述的弱监督声音事件检测方法，其特征在于，所述当子包数量为 R 时，第 k 类声音事件的句级别预测概率表示为：

$$\hat{Y}_k = \phi_{hi}(F_w(X)_k, R) = \phi_{avg}(\{\phi_{max}(b_r); \forall r\})$$

其中：$\hat{Y}_k$为第 k 类声音事件的句级别预测概率，ϕ_{hi}为自适应层次聚合算法，F_w表示声学模型，ϕ_{avg}表示均值聚合，ϕ_{max}表示最大值聚合，b_r为子包集合B中的第 r 个元素。

4.根据权利要求 3 所述的弱监督声音事件检测方法，其特征在于，所述第 k 类声音事件的句级别预测期望概率$\bar{Y}_k$为：

$$\bar{Y}_k = \sum_{R \in N^+} \{\lambda_k^R * \phi_{hi}(F_w(X)_k, R)\}$$

其中：λ_k^R表示第 k 类事件选择 R 个子包数量的概率，N^+为所有可选的子包数量的集合。

5.根据权利要求 4 所述的弱监督声音事件检测方法，其特征在于，所述联合优化声学模型和松弛化参数采用反向传播进行的：

$$L(W, \alpha; X, Y) = \frac{1}{K}\sum_{k}^{K} BCELoss(\bar{Y}_k, Y_k)$$

其中：L所有声音事件类别的平均预测误差，W、α分别为模型参数和松弛化参数，X、Y分别为模型输入的梅尔频谱特征和句级别标签，Y_k为第 k 类声音事件的句级别标签，$BCELoss$表示二进制交叉熵函数，K 为声音事件的类别总数。

6.根据权利要求 1 所述的弱监督声音事件检测方法，其特征在于，所述声学模型为任意

主流的深度学习模型，声学模型的基准模型为卷积循环神经网络模型。

7.根据权利要求 1 所述的弱监督声音事件检测方法，其特征在于，提取的特征为梅尔频谱特征。

8.根据权利要求 7 所述的弱监督声音事件检测方法，其特征在于，所述声学模型训练和验证采用 DCASE2017 数据集。

9.根据权利要求 8 所述的弱监督声音事件检测方法，其特征在于，所述音频信号下采样至 16kHz，帧长和帧移分别设置为 1024、664，分帧后每条信号得到 240 帧样本，梅尔频谱特征为 64 维。

10.一种实现权利要求 1-9 任一项所述的弱监督声音事件检测方法的系统，其特征在于，包括依次相连的声学模型和自适应层次聚合算法模块，所述声学模型输入预处理和特征提取的音频信号，所述声学模型预测得到帧级别预测概率，所述自适应层次聚合算法模块将帧级别预测概率聚合得到句级别预测概率。

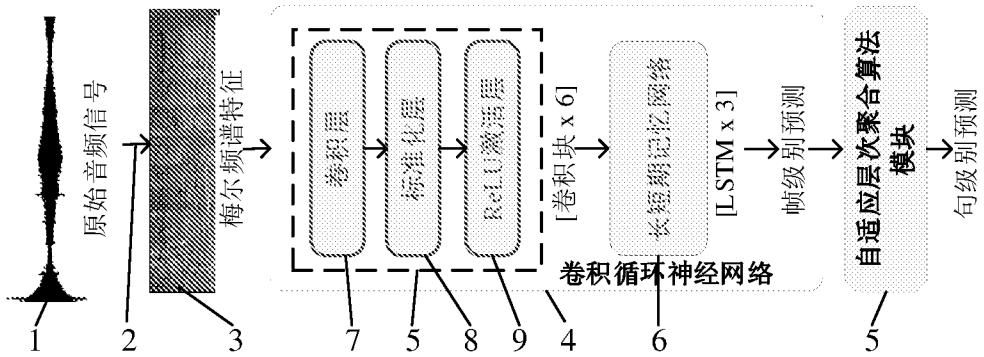
原始音频信号
梅尔频谱特征
卷积层
标准化层
ReLU激活层
[卷积块 x 6]
卷积循环神经网络
长短期记忆网络
[LSTM x 3]
帧级别预测
自适应层次聚合算法模块
句级别预测
1
2
3
7
5
8
9
4
6
5

图 1

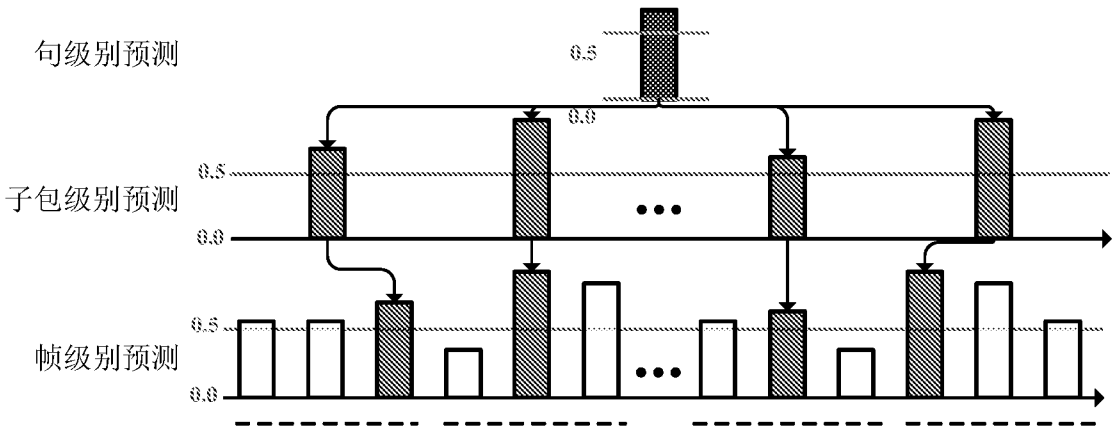
句级别预测
子包级别预测
帧级别预测
0.5
0.0
0.5
0.0
0.5
0.0

图 2

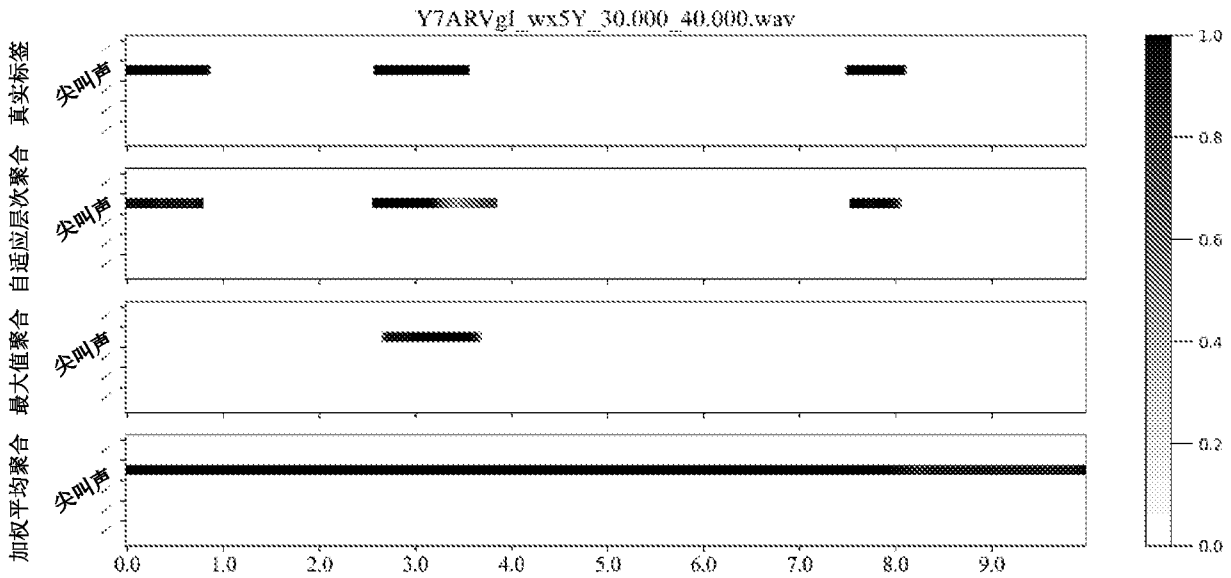
Y7ARVgI_wx5Y_30.000_40.000.wav
真实标签
自适应层次聚合
最大值聚合
加权平均聚合
尖叫声
尖叫声
尖叫声
尖叫声
0.0
1.0
2.0
3.0
4.0
5.0
6.0
7.0
8.0
9.0
1.0
0.8
0.6
0.4
0.2
0.0

图 3（a）

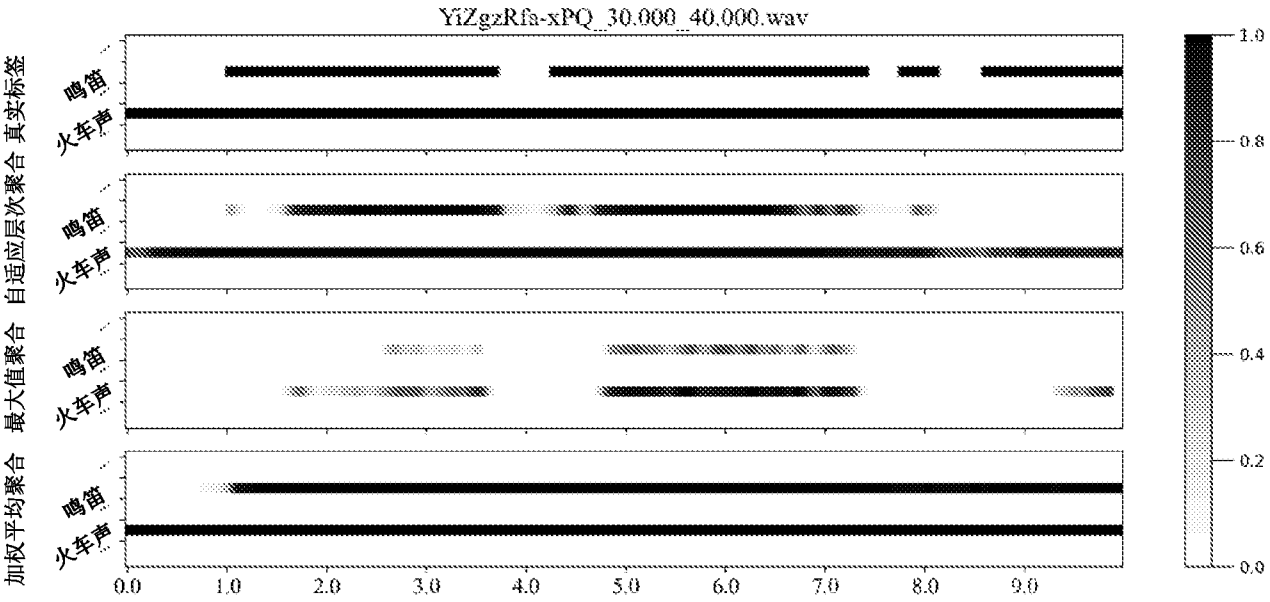
YiZgzRfa-xPQ_30.000_40.000.wav
加权平均聚合
最大值聚合
自适应层次聚合
真实标签
鸣笛
火车声
鸣笛
火车声
鸣笛
火车声
鸣笛
火车声
0.0
1.0
2.0
3.0
4.0
5.0
6.0
7.0
8.0
9.0
1.0
0.8
0.6
0.4
0.2
0.0

图 3（b）

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2022/101361

A. CLASSIFICATION OF SUBJECT MATTER

G10L 25/51(2013.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G10L; G06N

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CNPAT, WPI, EPODOC, CNKI: 声音, 事件, 检测, 自适应, 聚合, 弱, 监督, 帧, 预测, 概率, 连续, 最大值, 均值, 声学模型, 收敛, 最优, 分类, sound, event, detection, SED, adaptive, aggregate, weakly, supervision, frame, prediction, probability, continuous, maximum, mean, acoustic model, convergence, optimum, classification

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	CN 111933188 A (UNIVERSITY OF ELECTRONIC SCIENCE AND TECHNOLOGY OF CHINA) 13 November 2020 (2020-11-13) description, paragraphs [0017]-[0037], and figures 1-3	1-10
A	CN 110827804 A (FUZHOU UNIVERSITY) 21 February 2020 (2020-02-21) entire document	1-10
A	CN 108648748 A (SHENYANG UNIVERSITY OF TECHNOLOGY) 12 October 2018 (2018-10-12) entire document	1-10
A	CN 112036477 A (TSINGHUA UNIVERSITY) 04 December 2020 (2020-12-04) entire document	1-10
A	CN 112786029 A (AI SPEECH LTD.) 11 May 2021 (2021-05-11) entire document	1-10
A	CN 113707175 A (SHANGHAI NORMAL UNIVERSITY et al.) 26 November 2021 (2021-11-26) entire document	1-10

☑ Further documents are listed in the continuation of Box C. ☑ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
12 December 2022	**26 December 2022**
Name and mailing address of the ISA/CN **China National Intellectual Property Administration (ISA/CN)** **No. 6, Xitucheng Road, Jimenqiao, Haidian District, Beijing 100088, China** Facsimile No. **(86-10)62019451**	Authorized officer Telephone No.

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2022/101361

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	US 2017372725 A1 (PINDROP SECURITY INC.) 28 December 2017 (2017-12-28) entire document	1-10

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2022/101361

Patent document cited in search report	Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
CN 111933188 A	13 November 2020	None	
CN 110827804 A	21 February 2020	None	
CN 108648748 A	12 October 2018	None	
CN 112036477 A	04 December 2020	None	
CN 112786029 A	11 May 2021	None	
CN 113707175 A	26 November 2021	None	
US 2017372725 A1	28 December 2017	WO 2018005620 A1	04 January 2018
		US 2019096424 A1	28 March 2019
		US 2021134316 A1	06 May 2021
		US 10141009 B2	27 November 2018
		US 10867621 B2	15 December 2020

国际检索报告

国际申请号
PCT/CN2022/101361

A. 主题的分类

G10L 25/51(2013.01)i

按照国际专利分类(IPC)或者同时按照国家分类和IPC两种分类

B. 检索领域

检索的最低限度文献(标明分类系统和分类号)

G10L; G06N

包含在检索领域中的除最低限度文献以外的检索文献

在国际检索时查阅的电子数据库(数据库的名称，和使用的检索词（如使用））

CNPAT, WPI, EPODOC, CNKI：声音，事件，检测，自适应，聚合，弱，监督，帧，预测，概率，连续，最大值，均值，声学模型，收敛，最优，分类，sound, event, detection, SED, adaptive, aggregate, weakly, supervision, frame, prediction, probability, continuous, maximum, mean, acoustic model, convergence, optimum, classification

C. 相关文件

类 型*	引用文件，必要时，指明相关段落	相关的权利要求
A	CN 111933188 A (电子科技大学) 2020年11月13日 (2020-11-13) 说明书第[0017]-[0037]段，附图1-3	1-10
A	CN 110827804 A (福州大学) 2020年2月21日 (2020-02-21) 全文	1-10
A	CN 108648748 A (沈阳工业大学) 2018年10月12日 (2018-10-12) 全文	1-10
A	CN 112036477 A (清华大学) 2020年12月4日 (2020-12-04) 全文	1-10
A	CN 112786029 A (苏州思必驰信息科技有限公司) 2021年5月11日 (2021-05-11) 全文	1-10
A	CN 113707175 A (上海师范大学 等) 2021年11月26日 (2021-11-26) 全文	1-10
A	US 2017372725 A1 (PINDROP SECURITY INC.) 2017年12月28日 (2017-12-28) 全文	1-10

☐ 其余文件在C栏的续页中列出。

☑ 见同族专利附件。

* 引用文件的具体类型：

“A” 认为不特别相关的表示了现有技术一般状态的文件

“E” 在国际申请日的当天或之后公布的在先申请或专利

“L” 可能对优先权要求构成怀疑的文件，或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件（如具体说明的）

“O” 涉及口头公开、使用、展览或其他方式公开的文件

“P” 公布日先于国际申请日但迟于所要求的优先权日的文件

“T” 在申请日或优先权日之后公布，与申请不相抵触，但为了理解发明之理论或原理的在后文件

“X” 特别相关的文件，单独考虑该文件，认定要求保护的发明不是新颖的或不具有创造性

“Y” 特别相关的文件，当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时，要求保护的发明不具有创造性

“&” 同族专利的文件

国际检索实际完成的日期	国际检索报告邮寄日期
2022年12月12日	2022年12月26日
ISA/CN的名称和邮寄地址 中国国家知识产权局(ISA/CN) 中国北京市海淀区蓟门桥西土城路6号 100088 传真号 (86-10)62019451	受权官员 陈成 电话号码 86-(10)-53962548

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2022/101361

检索报告引用的专利文件			公布日 (年/月/日)	同族专利			公布日 (年/月/日)
CN	111933188	A	2020年11月13日	无			
CN	110827804	A	2020年2月21日	无			
CN	108648748	A	2018年10月12日	无			
CN	112036477	A	2020年12月4日	无			
CN	112786029	A	2021年5月11日	无			
CN	113707175	A	2021年11月26日	无			
US	2017372725	A1	2017年12月28日	WO	2018005620	A1	2018年1月4日
				US	2019096424	A1	2019年3月28日
				US	2021134316	A1	2021年5月6日
				US	10141009	B2	2018年11月27日
				US	10867621	B2	2020年12月15日