(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 108877832 B
(45)授权公告日 2022.12.23

(21)申请号 201810532016.5

(22)申请日 2018.05.29

(65)同一申请的已公布的文献号
申请公布号 CN 108877832 A

(43)申请公布日 2018.11.23

(73)专利权人 东华大学
地址 201620 上海市松江区松江新城人民北路2999号

(72)发明人 张逸 韩芳 黄荣

(74)专利代理机构 上海泰能知识产权代理事务所(普通合伙) 31233
专利代理师 宋缨 钱文斌

(51)Int.Cl.
G10L 25/03(2013.01)
G10L 25/27(2013.01)
G10L 15/06(2013.01)

(56)对比文件
CN 107293289 A,2017.10.24
US 2018/0075581 A1,2018.03.15
Volodymyr Kuleshov et al..《AUDIO SUPER-RESOLUTION USING NEURAL NETS》.《arXiv:1708.00853v1》.2017,第1-8页.
朱纯等.《基于深度卷积生成对抗网络的语音生成技术》.《仪表技术》.2018,(第2期),第13-15、20页.

审查员 李海龙

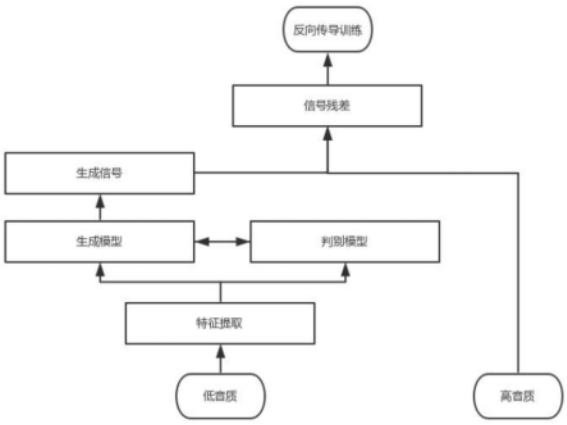
权利要求书1页 说明书4页 附图2页

(54)发明名称

一种基于GAN的音频音质还原系统

(57)摘要

本发明涉及一种基于GAN的音频音质还原系统,包括模型共享区块模块、生成网络模型模块、判别网络模型模块和序列重组模块;所述模型共享区块模块主要用于对于因防止损失信息没有进行频域处理的时域信号进行特征提取,将特征抽象为高层单元;所述生成网络模型模块利用高层抽象单元进行分析和重构;所述判别网络模型模块不断与所述生成网络模型进行对抗训练,不断改进生成效果;所述序列重组模块分析网络对最终生成输出进行序列加权重组。本发明能够生成更加逼真的音频信号。



CN 108877832 B

1.一种基于GAN的音频音质还原系统，其特征在于，包括模型共享区块模块、生成网络模型模块、判别网络模型模块和序列重组模块；所述模型共享区块模块用于对于因防止损失信息没有进行频域处理的时域信号进行特征提取，将特征抽象为高层单元；所述生成网络模型模块利用高层抽象单元进行分析和重构；所述判别网络模型模块不断与所述生成网络模型进行对抗训练，不断改进生成效果；所述序列重组模块分析网络对最终生成输出进行序列加权重组；所述模型共享区块模块使用离散卷积核对离散音频信号进行卷积计算从而提取特征，将信号抽象化，同时使用批标准化，在每次反向传导时，通过送入数组的激活值做规范化操作使得输出信号的均值拟似正态分布，利用线性整流函数的非线性特性拟合模型特性降低网络整体的计算负担；所述生成网络模型模块使用空洞卷积进行跨步输入降低输入维度并同时增大感受野，融合批标准化对输入数据进行标准化减少数据偏移和尺度缩放，使用残差网络分段学习使网络注重学习网络的残差；所述序列重组模块通过分析最终输出单元所依赖的填值单元对于有效传输单元的比例，从而确定单元输出的置信度，最终根据置信度计算权重对被分割的音频信号片段进行重组，其处理方式为：$A_i^o=(1-w)A_i^{pre}+wA_i^{next}$，其中，$w=\frac{\sum_k^c \max(RF_k,i)}{\sum_k^c RF_k}$，$A_i^o$、$A_i^{pre}$、$A_i^{next}$分别为最终输出音频、前合成音频段、后合成音频段，w为两段合成时所使用的权重，c是所有卷积层中所有的通道，RF_k为第k个通道下的感受野长度，i交叠区域数据的索引。

2.根据权利要求1所述的基于GAN的音频音质还原系统，其特征在于，所述判别网络模型模块使用间隔步长降低网络维度从而防止序列过长导致后续循环神经网络难以训练。

3.根据权利要求1所述的基于GAN的音频音质还原系统，其特征在于，所述判别网络模型模块不断与所述生成网络模型进行对抗训练是指使用小批量数据分批进行训练，训练使用局部梯度下降法进行反向传导，生成网络模型模块和判别网络模型模块交替进行训练，在训练过程调整生成网络模型模块和判别网络模型模块的权重。

4.根据权利要求3所述的基于GAN的音频音质还原系统，其特征在于，使用原音频作为对比文件进行生成训练，使用局部梯度下降法反向传导对整个生成网络模型的参数进行更新，然后调整权重针对生成网络模型进行训练，相对应的生成网络模型的参数训练获得的权重参数更新更大，之后判别网络模型与生成网络模型交替训练直至网络收敛。

一种基于GAN的音频音质还原系统

技术领域

[0001] 本发明涉及音频信号还原处理技术领域，特别是涉及一种基于GAN的音频音质还原系统。

背景技术

[0002] 为了方便网络传输和本地存储需对大量音频文件进行压缩处理，但获取存储空间下降的同时会牺牲相应的音质，为了有效利用网络资源，音乐平台往往提供不同的音乐品质，这在物理空间或计算资源上不是最有效的，使用音频还原技术可以对压缩比较高的音频资源进行还原，从而避免分别存储或者服务端和客户端的解压缩过程。有损压缩不仅用于MP3播放器和个人电脑，还用于视频播放设备、数字电视、网络流媒体、卫星广播等。因为有损压缩抛弃了部分不重要的部分，压缩率远远高于无损压缩(原信号的5％到20％之间)。音频有损压缩是一个不可逆的过程，为了削减信息的冗余性，需要使用符号化、模式识别、线性预测等方法。

[0003] 人类听觉并非能够识别所有的声音信号数据，大多的有损压缩通过识别人耳难以捕捉的信号从而减少知觉的冗余性。这种信号典型的有高频率信号和伴随高能量信号出现的小能量信号。把这些难以识别的信号删除还不足以获得可观的比特削减效果，信号进行编码的位数减少降低了信号比，通过基于心理声学的有损压缩，隐藏不易察觉的高频细节的技术是重要的。例如通过减少分配给高频分量的比特数来完成。这样做并不是因为原始信息只包含一些高频成分，而是人耳感知低频的能力要强于高频。从而高频细节被很好地隐藏并且不被察觉。为了进一步压缩信号，甚至可能降低采样率和通道数。这些有损压缩是一定程度上失真的，对声音的泛音有较大的影响，使得声音不够饱满，降低了人们的听觉感受。音质还原能够在提供较小的容量同时保持较好的音质，同时有利于服务商和用户双方。

[0004] 现在为了向用户提供不同的体验，往往需要存储多种不同音质的文件，这显然是不明智的，使用音频还原技术，可以只保存一份低音质源文件，通过算法，提升音质，既不用保存多份文件，单份文件的大小也相对较小。

发明内容

[0005] 本发明所要解决的技术问题是提供一种基于GAN的音频音质还原系统，能够生成更加逼真的音频信号。

[0006] 本发明解决其技术问题所采用的技术方案是：提供一种基于GAN的音频音质还原系统，包括模型共享区块模块、生成网络模型模块、判别网络模型模块和序列重组模块；所述模型共享区块模块主要用于对于因防止损失信息没有进行频域处理的时域信号进行特征提取，将特征抽象为高层单元；所述生成网络模型模块利用高层抽象单元进行分析和重构；所述判别网络模型模块不断与所述生成网络模型进行对抗训练，不断改进生成效果；所述序列重组模块分析网络对最终生成输出进行序列加权重组。

[0007] 所述模型共享区块模块使用离散卷积核对离散音频信号进行卷积计算从而提取

特征，将信号抽象化，同时使用批标准化，在每次反向传导时，通过送入数组的激活值做规范化操作使得输出信号的均值拟似正态分布，利用线性整流函数的非线性特性拟合模型特性降低网络整体的计算负担，从而使得系统整体显得更为鲁棒。

[0008] 所述生成网络模型模块使用空洞卷积进行跨步输入降低输入维度并同时增大感受野，让每个卷积输出都包括更大感受野的信息，融合批标准化对输入数据进行标准化减少数据偏移和尺度缩放，使用残差网络分段学习使网络注重学习网络的残差，在增加网络深度的同时，防止网络的退化问题，使网络更容易优化且获得更好的准确性，保持较低的复杂度，降低训练误差和测试误差。

[0009] 所述判别网络模型模块使用间隔步长降低网络维度从而防止序列过长导致后续循环神经网络难以训练。

[0010] 所述判别网络模型模块不断与所述生成网络模型进行对抗训练是指使用小批量数据分批进行训练，训练使用局部梯度下降法进行反向传导，生成网络模型模块和判别网络模型模块交替进行训练，在训练过程调整生成网络模型模块和判别网络模型模块的权重。

[0011] 使用原音频作为对比文件进行生成训练，使用局部梯度下降法反向传导对整个生成网络模型的参数进行更新，然后调整权重针对生成网络模型进行训练，相对应的生成网络模型的参数训练获得的权重参数更新更大，之后判别网络模型与生成网络模型交替训练直至网络收敛。

[0012] 所述序列重组模块通过分析最终输出单元所依赖的填值单元对于有效传输单元的比例，从而确定单元输出的置信度，最终根据置信度计算权重对被分割的音频信号片段进行重组。

[0013] 有益效果

[0014] 由于采用了上述的技术方案，本发明与现有技术相比，具有以下的优点和积极效果：

[0015] 本发明使用深度神经网络针对经不可逆压缩后的源音频文件进行还原，允许信号在传输之前进一步压缩，方便存储和传输，同时保证经不可逆压缩后可以还原的音频信号的质量较同等存储容量大小的音频文件更好。

[0016] 本发明使用对抗生成网络模型，结合生成模型和判别模型进行对抗训练，生成网络针对源信号进行再理解和重构，并通过估计结果针对信号对象进行修复，判别网络不断将生成模型的输出和实际源信号进行比较，促进生成模型进一步拟合源信号，使其能够生成更加逼真的音频信号。

[0017] 本发明使用双向循环网络综合时域中的所有特征，使用残差网络、空洞卷积、步长、长短时记忆单元解决网络在深度和广度上的退化问题，优化网络训练，减少网络收敛时间，使系统更加鲁棒。

附图说明

[0018] 图1是训练过程流程图；

[0019] 图2是客户端使用过程流程图。

具体实施方式

[0020] 下面结合具体实施例，进一步阐述本发明。应理解，这些实施例仅用于说明本发明而不用于限制本发明的范围。此外应理解，在阅读了本发明讲授的内容之后，本领域技术人员可以对本发明作各种改动或修改，这些等价形式同样落于本申请所附权利要求书所限定的范围。

[0021] 本发明的实施方式涉及一种基于GAN的音频音质还原系统，包括模型共享区块模块、生成网络模型模块、判别网络模型模块和序列重组模块；所述模型共享区块模块主要用于对于因防止损失信息没有进行频域处理的时域信号进行特征提取，将特征抽象为高层单元；所述生成网络模型模块利用高层抽象单元进行分析和重构；所述判别网络模型模块不断与所述生成网络模型进行对抗训练，不断改进生成效果；所述序列重组模块分析网络对最终生成输出进行序列加权重组。该音频音质还原系统基于深度学习，使用局部梯度下降法进行分批训练，并使用局部失效法减少网络的过拟合现象。

[0022] 所述的共享模型区块模块主要是指使用离散卷积核对离散音频信号进行卷积计算从而提取特征，将信号抽象化，同时使用批标准化，在每次反向传导时，通过送入数组的激活值做规范化操作使得输出信号的均值拟似正态分布，而最后的归一化和偏移操作则是为了让因训练而使用的批标准化有更大的可能模拟输入，利用线性整流函数的非线性特性拟合模型特性，其微分计算大大降低了网络整体的计算负担，从而使得系统整体显得更为鲁棒。

[0023] 所述的生成网络模型模块主要使用空洞卷积，即对一般卷积核的输入进行跨步输入，且不改变参数数量，不需要使用池化层而损失大量信息的情况下依旧能够降低输入维度并同时增大感受野，让每个卷积输出都包括更大感受野的信息，有利于解决音频信号序列过长的问题，融合批标准化对输入数据进行标准化减少数据偏移和尺度缩放，使用残差网络能够使网络注重学习网络的残差，细化学习结果，在增加网络深度的同时，防止网络的退化问题，使网络更容易优化且获得更好的准确性，保持较低的复杂度，降低训练误差和测试误差。

[0024] 生成网络模型模块使用残差网络分段学习降低每一个部分所要训练的难度，也同时减小了单个单元和整体网络的耦合性，注重于针对残差进行学习，使用空洞卷积能够让每个网络单元获得更广的感受野，更好的兼顾单元周围的元素，进而理解频域信息。

[0025] 所述的判别网络模型模块使用间隔步长降低网络维度从而防止序列过长导致后续循环神经网络难以训练。双向循环神经网络能够兼顾上下文序列信息，分析时域特征，综合信息，长短时记忆单元能够防止梯度弥散和梯度爆炸，有效地兼顾网络深度不同的单元，最终使用softmax进行分类。

[0026] 判别网络模型模块主要以循环神经网络为主，关键是兼顾整体信息从而做出判断为送入信号进行评分，前置的卷积网络则是使用步长缩小维度，让循环神经网络方便作为输入进行训练，同时避免了序列过长的问题。

[0027] 所述的序列重组模块通过分析最终输出单元所依赖的填值单元对于有效传输单元的比例，从而确定单元输出的置信度，最终根据置信度计算权重对被分割的音频信号片段进行重组。

[0028] 本发明使用时，需要在服务器端对训练数据进行预处理，对压缩后音频和原音频

进行分段和匹配，之后将音频转化为比特流，构建生成网络模型和判别网络模型，将训练数据进行分组，压缩后音频样本作为网络的输入，原音频作为网络输出的对比，同时标记音频是否为原文件用于判别网络模型的训练。如图1所示，首先使用对比文件进行生成训练，使用局部梯度下降法反向传导对整个生成网络模型的参数进行更新，然后调整权重针对生成网络模型进行训练，相对应的生成网络模型的参数训练获得的权重参数更新更大，之后判别网络模型与生成网络模型交替训练直至网络收敛，初始化参数可针对压缩比做相应估计。

[0029] $l=l^c+\lambda^d l^d$，

[0030] 式中：

[0031] $$l^c=\frac{MSE(A^{HQ},G(A^{LQ}))}{MSE(A^{HQ},A^{LQ})+bias},$$

[0032] $l^d=\mathrm{MSE}(P_{softmax},\mathrm{Label})$，

[0033] $$\lambda^d\approx\frac{GE}{l^d},$$

[0034] $l^d=-\mathrm{Avg}(p_{HQ}(A^{SQ}))$，

[0035] 其中，l为总损失，l^c为生成网络输出结果相较于原音频的损失，l^d为判别网络损失，λ^d是平衡l^c和l^d的系数，可以根据样本集压缩误差做相应调整，GE是对预计收敛损失的估计(可去除判别网络进行预估)，A^{HQ}、A^{LQ}、A^{SQ}分别为高音质音频信号、低音质音频信号、生成音频信号，MSE为均方误差，G为生成模型，$P_{softmax}$为通过softmax层产生的概率输出结果，与样本标签Label的均方误差作为判别网络损失，p_{HQ}为判别网络将生成音频信号判别为高音质信号的概率。

[0036] 考虑到每个输入样本拥有不同压缩误差，所以使用训练输出与样本目标的差异和样本输入和目标之间的差异的比值作为目标函数，bias是为防止0除值的偏置。λ^d是平衡l^c和l^d的系数，可以根据样本集压缩误差做相应调整，GE是对预计收敛损失的估计(可去除判别网络模型进行预估)，因为此次训练偏重于判别网络，适当提高λ^d的权重。第二次训练固定判别网络的参数，使它们不参加训练(包括共享的部分参数)，用第一次训练完成的判别网络来判别生成网络的生成结果，生成网络的参数直接继承上次的训练参数做初始化，需要训练判别网络无法区分生成网络和原高音质信号，达到以假乱真的效果。

[0037] 训练完成时保存网络参数，生成网络参数保存为比特流，可经过无损压缩作为软件数据供客户端使用。如图2所示，客户端在接收完音频文件后，通过判别网络，甄别音频信号质量，可根据需求使用保存的生成网络数据对音频进行还原操作。因为还原单位一般不超过一秒，生成网络总共包括20个卷积块，每个卷积块平均拥有64个卷积核，总参数大约为十万，通过网络处理后进行拼接，拼接处理方式为：

[0038] $A_i^o=(1-w)A_i^{pre}+wA_i^{next}$，其中$w=\frac{\sum_k^c\max(RF_k,i)}{\sum_k^c RF_k}$，

[0039] 式中，A_i^o、A_i^{pre}、A_i^{next}分别为最终输出音频、前合成音频段、后合成音频段，w为两段合成时所使用的权重，c是所有卷积层中所有的通道，RF_k为第k个通道下的感受野长度，i交叠区域数据的索引。总体时间复杂度可以实现实时解码。

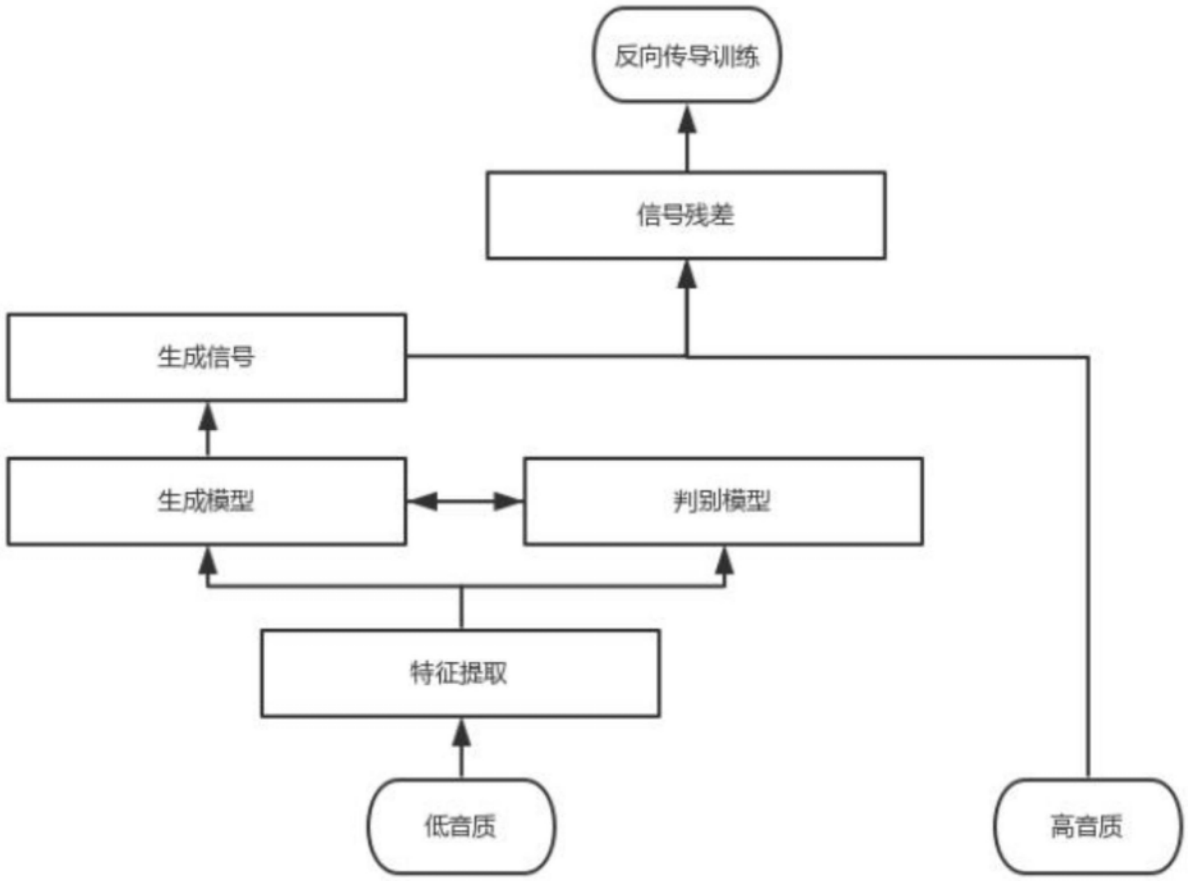


图1

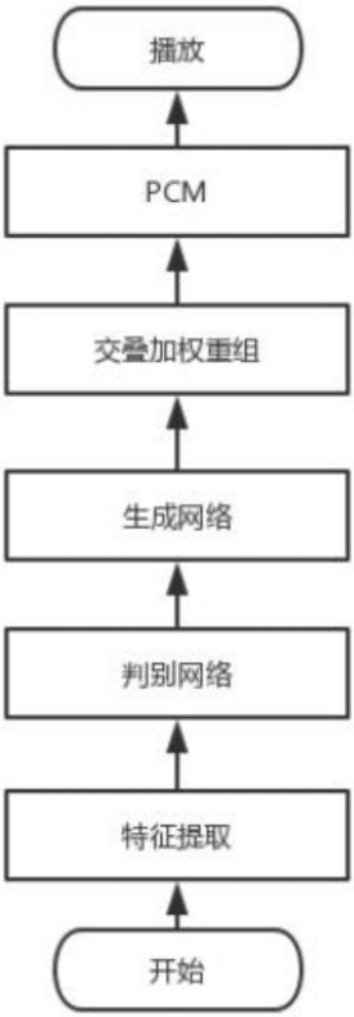


图2