

(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
2025 年 1 月 9 日 (09.01.2025)



(10) 国际公布号
WO 2025/007868 A1

(51) 国际专利分类号:

G10L 19/005 (2013.01) **H04L 65/75** (2022.01)
G10L 25/30 (2013.01) **G10L 25/21** (2013.01)

(21) 国际申请号: PCT/CN2024/103191

(22) 国际申请日: 2024 年 7 月 2 日 (02.07.2024)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:

202310823707.1 2023年7月5日 (05.07.2023) CN

(71) 申请人: 维沃移动通信有限公司(VIVO MOBILE COMMUNICATION CO., LTD.) [CN/CN]; 中国广东省东莞市长安镇维沃路 1 号, Guangdong 523863 (CN)。

(72) 发明人: 张勇(ZHANG, Yong); 中国广东省东莞市长安镇维沃路1号, Guangdong 523863 (CN)。

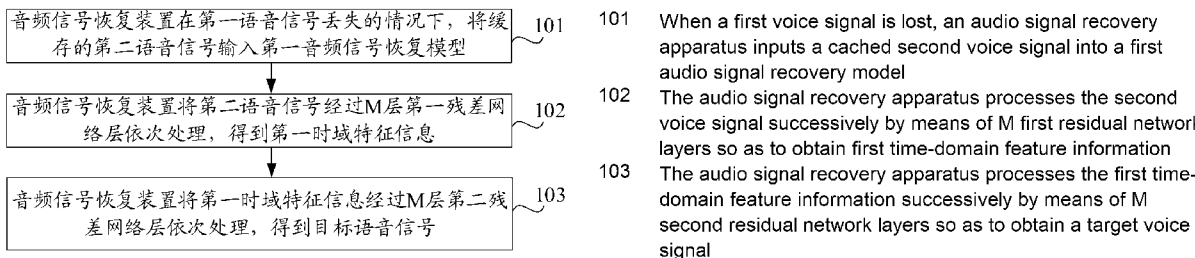
(74) 代理人: 北京远志博慧知识产权代理有限公司(BOHUI INTELLECTUAL PROPERTY); 中国北京市海淀区交大东路 31 号东区 10 号楼等 17 幢31幢108, Beijing 100044 (CN)。

(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, CV, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, 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, MG, MK, MN, MU, 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, CV, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SC, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚

(54) Title: AUDIO SIGNAL RECOVERY METHOD AND APPARATUS, ELECTRONIC DEVICE, AND READABLE STORAGE MEDIUM

(54) 发明名称: 音频信号恢复方法、装置、电子设备及可读存储介质



(57) Abstract: An audio signal recovery method and apparatus, an electronic device, and a readable storage medium, belonging to the field of signal processing. The method comprises: when a first voice signal is lost, inputting a cached second voice signal into a first audio signal recovery model, the first audio signal recovery model comprising an encoding module and a decoding module, the encoding module comprising M first residual network layers, and the decoding module comprising M second residual network layers (101); processing the second voice signal successively by means of the M first residual network layers so as to obtain first time-domain feature information (102); and processing the first time-domain feature information successively by means of the M second residual network layers so as to obtain a target voice signal, the second voice signal comprising at least one of the following: X voice signal frames prior to the first voice signal and Y voice signal frames following the first voice signal, and M, X and Y being positive integers (103).

(57) 摘要: 一种音频信号恢复方法、装置、电子设备及可读存储介质, 属于信号处理领域, 其中该方法包括: 在第一语音信号丢失的情况下, 将缓存的第二语音信号输入第一音频信号恢复模型, 第一音频信号恢复模型包括编码模块和解码模块, 编码模块包括M层第一残差网络层, 解码模块包括M层第二残差网络层 (101); 将第二语音信号经M层第一残差网络层依次处理, 得到第一时域特征信息 (102); 将第一时域特征信息经M层第二残差网络层依次处理, 得到目标语音信号; 其中, 第二语音信号包括以下至少之一: 第一语音信号前的X帧语音信号, 第一语音信号之后的Y帧语音信号, M, X, Y为正整数 (103)。

(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, ME, 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))。
- 在修改权利要求的期限届满之前进行, 在收到该修改后将重新公布(细则48.2(h))。

音频信号恢复方法、装置、电子设备及可读存储介质

相关申请的交叉引用

5 本申请主张在 2023 年 07 月 05 日在中国提交的中国专利申请号 202310823707.1 的优先权，其全部内容通过引用包含于此。

技术领域

本申请属于信号处理领域，具体涉及一种音频信号恢复方法、装置、电子设备及可读存储介质。

背景技术

10 目前，在基于网际协议（Internet Protocol，IP）的语音通信中，语音通常会被分割成较小的语音数据包，并按一定时序进行传输。

15 然而，在相关技术中，在实际通过通信信道传输语音数据包的过程中，由于网络拥塞、抖动延迟等原因，经常会导致在传输过程中出现语音丢包现象。如此，容易导致语音通话质量较差。

发明内容

20 本申请实施例的目的是提供一种音频信号恢复方法、装置、电子设备及可读存储介质，能够提升语音通话的质量。

25 第一方面，本申请实施例提供了一种音频信号恢复方法，该方法包括：在第一语音信号丢失的情况下，将缓存的第二语音信号输入第一音频信号恢复模型，第一音频信号恢复模型包括编码模块和解码模块，编码模块包括 M 层第一残差网络层，解码模块包括 M 层第二残差网络层；将第二语音信号经 M 层第一残差网络层依次处理，得到第一时域特征信息；将第一时域特征信息经 M 层第二残差网络层依次处理，得到目标语音信号；其中，第二语音信号包括以下至少之一：第一语音信号前的 X 帧语音信号，第一语音信号之后的 Y 帧语音信号，M，X，Y 为正整数。

30 第二方面，本申请实施例提供了一种音频信号恢复装置，该装置包括：输入模块和处理模块；输入模块，用于在第一语音信号丢失的情况下，将缓存的第二语音信号输入第一音频信号恢复模型，第一音频信号恢复模型包括编码模块和解码模块，编码模块包括 M 层第一残差网络层，解码模块包括 M 层第二残差网络层；处理模块，用于将第二语音信号经 M 层第一残差网络层依次处理，得到第一时域特征信息；处理模块，还用于将第一时域特征信息经 M 层第二残差网络层依次处理，得到目标语音信号；其中，第二语音信号包括以下至少之一：第一语音信号前的 X 帧语音信号，第一语音信号之后的 Y 帧语音信号，M，X，Y 为正整数。

35 第三方面，本申请实施例提供了一种电子设备，该电子设备包括处理器和存储器，所述存储器存储可在所述处理器上运行的程序或指令，所述程序或指令被所述处理器执行时实现如第一方面所述的方法的步骤。

40 第四方面，本申请实施例提供了一种可读存储介质，所述可读存储介质上存储程序或指令，所述程序或指令被处理器执行时实现如第一方面所述的方法的步骤。

第五方面，本申请实施例提供了一种芯片，所述芯片包括处理器和通信接口，所述通信接口和所述处理器耦合，所述处理器用于运行程序或指令，实现如第一方面所述的方法。

45 第六方面，本申请实施例提供一种计算机程序产品，该程序产品被存储在存储介质中，该程序产品被至少一个处理器执行以实现如第一方面所述的方法。

50 在本申请实施例中，在第一语音信号丢失的情况下，将缓存的第二语音信号输入第一音频信号恢复模型，第一音频信号恢复模型包括编码模块和解码模块，编码模块包括 M 层第一残差网络层，解码模块包括 M 层第二残差网络层；将第二语音信号经 M 层第一残差网络层依次处理，得到第一时域特征信息；将第一时域特征信息经 M 层第二残差网络层依次处理，得到目标语音信号；其中，第二语音信号包括以下至少之一：第一语音信号前的 X 帧语音信号，第一语音信号之后的 Y 帧语音信号，M，X，Y 为正整数。通过该方案，在语音通话时，若丢失了某个语音信号，则可以通过

音频信号恢复模型中的编码模块的残差网络层提取缓存的丢失语音信号前或丢失语音信号后的语音信号的时域特征信息，并通过音频信号恢复模型中的解码模块的残差网络层合成编码模块提取的时域特征信息，恢复出丢失的语音信号。如此，由于可以在语音信号丢失的情况下，通过将缓存的语音信号输入音频信号恢复模型恢复出丢失的语音信号，因此在语音信号的传输过程中，可以通过音频信号恢复模型恢复出丢失的语音信号，从而避免用户感知到语音数据包的丢失，降低语音数据包丢失所带来的负面影响，进而可以有效地提高语音通话的质量。

附图说明

图 1 是本申请实施例提供的一种 AMR-WB 解码器原理的结构示意图；
图 2 是本申请实施例提供的一种 AMR-WB 解码器丢包隐藏的实例流程图；
图 3 是本申请实施例提供的一种 AMR-WB 解码器丢包隐藏状态机判断错误帧与正常帧的实例示意图；
图 4 是本申请实施例提供的一种基于 DNN 的丢包隐藏方法的流程图；
图 5 是本申请实施例提供的一种基于 DNN 的特征提取的流程图；
图 6 是本申请实施例提供的一种基于 DNN 的语音时域信号波形重建的流程图；
图 7 是本申请实施例提供的一种音频信号恢复方法的流程图；
图 8 是本申请实施例提供的一种残差网络的结构示意图；
图 9 是本申请实施例提供的一种编码模块的网络结构示意图；
图 10 是本申请实施例提供的一种对丢失的语音信号进行恢复的流程图；
图 11 是本申请实施例提供的一种解码模块的网络结构示意图；
图 12 是本申请实施例提供的一种训练 GAN 网络模型的流程图；
图 13 是本申请实施例提供的一种第二音频信号恢复模型的训练流程图；
图 14 是本申请实施例提供的一种 MS-STFT 判别器的结构示意图；
图 15 是本申请实施例提供的一种音频信号恢复方法的实例流程示意图；
图 16 是本申请实施例提供的一种音频信号恢复装置的结构示意图；
图 17 是本申请实施例提供的一种电子设备的硬件结构示意图之一；
图 18 是本申请实施例提供的一种电子设备的硬件结构示意图之二。

具体实施方式

下面将结合本申请实施例中的附图，对本申请实施例中的技术方案进行清楚地描述，显然，所描述的实施例是本申请一部分实施例，而不是全部的实施例。基于本申请中的实施例，本领域普通技术人员获得的所有其他实施例，都属于本申请保护的范围。

本申请的说明书和权利要求书中的术语“第一”、“第二”等是用于区别类似的对象，而不适用于描述特定的顺序或先后次序。应该理解这样使用的数据在适当情况下可以互换，以便本申请的实施例能够以除了在这里图示或描述的那些以外的顺序实施，且“第一”、“第二”等所区分的对象通常为一类，并不限定对象的个数，例如第一对象可以是一个，也可以是多个。此外，说明书以及权利要求中“和/或”表示所连接对象的至少其中之一，字符“/”，一般表示前后关联对象是一种“或”的关系。

下面结合附图，通过具体的实施例及其应用场景对本申请实施例提供的音频信号恢复方法、装置、电子设备及可读存储介质进行详细地说明。

本申请实施例提供的音频信号恢复方法、装置、电子设备及可读存储介质适用于基于 IP 进行语音通话的场景中。

目前，当进行基于 IP 的语音通信时，语音通常会被分割成较小的分组包或帧，并按一定时序进行传输。但是，在实际通信信道下，由于网络拥塞、抖动延迟等原因，必然会出现语音丢包现象，从而在很大程度上影响语音通话质量。在出现语音数据包丢失情况下，为了不使用户感觉到有丢包现象的发生，当检测到丢包后，需要对其进行错误隐藏，以便有效地提高最终语音通话的质量。

在相关技术中，针对丢包的错误隐藏方法有两种。一种为基于自适应多速率宽带编码（Adaptive Multi-Rate-Wideband, AMR-WB）解码器的丢包隐藏方法，另一种为基于深度神经网络（Deep Neural Network, DNN）的线性预测编码（Linear Predictive Coding, LPC）方法。

对于基于 AMR-WB 解码器的错误隐藏方法，AMR-WB 解码器首先把接收到的编码流恢复成参数编码，解码后得到 20ms 语音帧对应的编码参数。这些参数为：ISF 参数、4 个基音周期、4 个固定码本矢量、4 个自适应码本增益、4 个固定码本增益。每帧 ISF 参数被转换为 ISP 参数，然后内插并转换得到 4 个子帧的 LPC 合成滤波器系数，每个子帧包含 5ms 的语音帧。AMR-WB 解码器原理框图如图 1 所示，对于每个子帧，其可以在基音周期处内插过去的激励而得到自适应码本，并将自适应码本乘以自适应码本增益 g^p 得到自适应码本激励 $v(n)$ ；并将固定码本乘以固定码本增益 g^c 得到固定码本激励 $c(n)$ ；最后将自适应码本激励 $v(n)$ 和固定码本激励 $c(n)$ 相加得到总激励信号 $u(n)$ ，通过 LPC 合成滤波器得到合成语音。

AMR-WB 解码器语音编码丢包隐藏系统框图如图 2 所示，当编码码流在传输中发生了丢包后，解码端检测到丢包后调用丢包隐藏模块进行错误恢复。AMR-WB 丢包隐藏算法使用状态机来确定错误帧的状态，并根据状态机状态来恢复编码参数，包括 ISF 参数、基音周期、固定码本矢量、自适应码本增益、固定码本增益。得到恢复的编码参数后，再使用 AMR-WB 解码器来合成语音。

AMR-WB 解码器丢包隐藏状态机判断错误帧与正常帧的实例示意图如图 3 所示，当前帧的错帧指示符 (Bad Frame Indicator, BFI) 和前一帧的错帧指示符 PrevBFI 同时作为输入量，根据两个输入量确定当前帧处于状态机模型中的何种状态。图 3 中实线表示错误帧的变化路径，虚线表示正确帧的变化路径。系统从状态 0 开始。每次检测到一个错误语音帧，状态计数器就加 1，当它达到 6 时饱和。每次检测到一个正确语音帧，状态计数器就右移 1。状态表示信道的质量。可以理解，状态计数器的值越大，信道质量越差。

状态机的控制流程可以用下面的代码来描述，“BFI=” 为当前帧错帧指示，“BFI=1” 为错误帧，“BFI=0” 为正确帧，“STATE=” 为状态变量，“PrevBFI=” 为前一帧错帧指示。

```
If (BFI= 0)
    STATE= STATE+ 1;
    If (STATE> 6)
        STATE= 6;
```

```
else
    STATE= STATE >> 1
```

对于图 3 中 STATE=0 或 1、BFI=0 且 PrevBFI=0 的情况，表示当前接收或先前接收的语音帧中没有检测到错误。换句话说，当前接收的正确语音帧将会被正常解码。

对于图 3 中 STATE=0 至 3、BFI=0 且 PrevBFI=1 的情况，表示当前接收到的语音帧中没有检测到错误，但之前接收到的语音帧是错误的。那么可以通过下述公式 (1) 对当前正确接收帧的固定码本增益 $g^c(n)$ 参数进行修改，且其被限制在最后一个正确接收帧的最后一个子帧所使用的固定码本增益 $g^c(n-1)$ 值以下。

$$g^c(n) = \begin{cases} g_{received}^c, & g_{received}^c \leq 100 \text{ or } g_{received}^c \leq g^c(n-1) \times 1.25 \\ 1.25 \times g^c(n-1), & \text{otherwise} \end{cases} \quad \text{公式 (1)}$$

其中， $g_{received}^c$ 为当前正确接收帧解码的固定码本增益， $g^c(n-1)$ 为最后一个正确接收帧的最后一个子帧的固定码本增益， $g^c(n)$ 为当前帧的固定码本增益，1.25 为增益常数。可以理解，当前帧的其余编码参数可以正常解码并用于合成语音。

对于图 3 中 STATE=1 至 6、BFI=1 且 PrevBFI=0 或 1 的情况，表示在当前接收到的语音帧中检测到错误，可以调用错误隐藏恢复编码参数。

对于自适应码本增益和固定码本增益的恢复，自适应码本增益和固定码本增益的恢复中使用的增益因子的大小可以根据状态机状态确定，状态值越大，表示信道质量越差，衰减因子也越大。在 AMR-WB 丢包隐藏算法中，自适应码本增益 $g^p(n)$ 和固定码本增益 $g^c(n)$ 可以由先前子帧的衰减值代替。自适应码本增益 $g^p(n)$ 可以通过下述的公式 (2) 进行计算：

$$g^p(n) = P^p(\text{state}) \times \text{median5} \left(g^p(n-1), \dots, g^p(n-5) \right) \quad \text{公式 (2)}$$

其中， $g^p(n)$ 为当前帧的自适应码本增益， $\left(g^p(n-1), \dots, g^p(n-5) \right)$ 为过去 5 个子帧的自适应码本增益， $\text{median5} \left(g^p(n-1), \dots, g^p(n-5) \right)$ 表示过去 5 个自适应码本增益 $g^p(n)$ 数值的中间值。 $P^p(\text{state})$ 为抑制因子，STATE 为当前状态机模

型的状态, $P^p(1) = 0.98$, $P^p(2) = 0.96$, $P^p(3) = 0.75$, $P^p(4) = 0.23$, $P^p(5) = 0.05$, $P^p(6) = 0.01$ 。可以理解, 状态值越高, 增益衰减得越多。

固定码本增益 $g^c(n)$ 可以通过下述的公式 (3) 进行计算:

$$g^c(n) = \begin{cases} P^c(\text{state}) \times \text{median5}(g^c(n-1), \dots, g^c(n-5)), & VAD_{HIST} \leq 2 \\ \text{median5}(g^c(n-1), \dots, g^c(n-5)), & VAD_{HIST} > 2 \end{cases}$$

5 公式 (3)

其中, $g^c(n)$ 为当前帧的固定码本增益, $g^c(n-1), \dots, g^c(n-5)$ 为过去 5 个子帧的固定码本增益, $\text{median5}(g^c(n-1), \dots, g^c(n-5))$ 表示过去 5 个固定码本增益 $g^c(n)$ 数值的中间值。VAD_HIST 是连续 VAD=0 判决的数量, VAD=0 表示当前帧为非语音帧。 $P^c(\text{state})$ 为抑制因子, STATE 为当前状态机模型的状态, $P^c(1) = 0.50$, $P^c(2) = 0.25$, $P^c(3) = 0.25$, $P^c(4) = 0.25$, $P^c(5) = 0.15$, $P^c(6) = 0.01$ 。可以理解, 状态值越高, 增益衰减得越多。

对于 ISF 矢量参数的恢复, AMR-WB 标准使用的谱系数是 16 阶 ISF 参数。在发生帧错误时, 错误帧 ISF 参数的估计值通过 ISF 参数的自适应平均值插值得到, 其可以通过下述公式 (4) 计算得到:

$$15 \quad ISF_q(i) = \alpha \times \text{past_ISF}_q(i) + (1 - \alpha) \times ISF_{\text{mean}}(i) \quad \text{公式 (4)}$$

其中, i 为帧索引, $i=0$ 至 15, $\alpha = 0.9$, $ISF_q(i)$ 为当前帧的 ISF 矢量参数, $\text{past_ISF}_q(i)$ 为前一帧的 ISF 矢量参数, $ISF_{\text{mean}}(i)$ 表示自适应均值和常量 ISF 的加权值, 其可以通过公式 (5) 计算得到:

$$20 \quad ISF_{\text{mean}}(i) = \beta \times ISF_{\text{const_mean}}(i) + (1 - \beta) \times ISF_{\text{adaptive_mean}}(i) \quad \text{公式 (5)}$$

其中, $\beta = 0.75$, $ISF_{\text{const_mean}}(i)$ 是一个包含长时间平均 ISF 的矢量。 $ISF_{\text{adaptive_mean}}(i)$ 是最近 3 个正确接收帧的 $\text{past_ISF}_q(i)$ 的平均值, 只有当 BFI = 0 时 (即正确接收帧), $ISF_{\text{adaptive_mean}}(i)$ 才进行更新。 $ISF_{\text{adaptive_mean}}(i)$ 可以通过公式 (6) 计算得到:

$$25 \quad ISF_{\text{adaptive_mean}}(i) = \frac{1}{3} \sum_{j=0}^2 \text{past_ISF}_q(i) \quad \text{公式 (6)}$$

对于基音周期参数的恢复, 可以将 5 个最后正确接收语音帧的基音周期和自适应码本增益分别被存入缓冲 buffer, 令其分别为 T_{buffer} 和 g^p_{buffer} , 它们将用于寻找当前帧最佳的基音周期。基音周期恢复算法首先对过去子帧基音周期的可用性进行判断, 如果浊音性和稳定性强, 说明该错误帧和过去帧比较变化不大, 丢失帧的基音周期可以用过去一帧的代替; 否则, 该错误帧的基音周期值在一定范围内随机产生。基音周期参数恢复算法分为下述的步骤 1 和步骤 2:

步骤 1: 计算过去基音周期可用性因子 $Q_{\text{lag_t-1}}$, 其可以通过公式 (7) 计算得到:

$$30 \quad Q_{\text{lag_t-1}} = \begin{cases} 1 & g^p_{\min} > 0.5 \\ 1 & g^p(n-1) > 0.5 \text{ 且 } g^p(n-2) > 0.5 \\ 0 & \text{otherwise} \end{cases} \quad \text{公式 (7)}$$

其中, $g^p_{\min}$ 表示自适应码本增益缓冲 buffer 中的最小值, $g^p(n-1)$ 表示前一个正确接收帧的自适应码本增益, $g^p(n-2)$ 是前一个正确接收帧之前的正确帧的自适应码本增益。

步骤 2: 估计当前帧基音周期 T 。步骤 1 得到 $Q_{\text{lag_t-1}}$ 后, 当前帧基音周期 T 可以通过下述的公式 (8) 计算得到:

$$35 \quad T = \begin{cases} T(n-1), & Q_{\text{lag_t-1}} = 1 \\ \frac{1}{3} \sum (T_{\text{max}} + T_{\text{max-1}} + T_{\text{max-2}}) + \text{RND}(T_{\text{max}} - T_{\text{max-2}}), & Q_{\text{lag_t-1}} = 0 \end{cases} \quad \text{公式 (8)}$$

其中, $T(n-1)$ 为前一个正确接收帧的基音周期, T_{max} 为基音周期缓冲 buffer 中的最大值, $T_{\text{max-1}}$ 为基音周期缓冲 buffer 中的第二大值, $T_{\text{max-2}}$ 为基音周期缓冲 buffer 中的第三大值。 $\text{RND}(T_{\text{max}} - T_{\text{max-2}})$ 为一个随机数生成函数。

对于固定码本矢量的恢复, 错误帧的固定码本矢量可以由随机产生的序列代替。

由于语音信号是一种短时平稳的准周期信号, 相邻的语音帧之间存在着一定的相关性, 因此现有的丢包隐藏方法通常利用编码参数的帧内与帧间的相关性来对丢包的语音编码参数进行恢复。在相关技术中, 针对丢包的错误隐藏方法的语音编码通常基于码激励线性预测模型 (Code Excited Linear Prediction, CELP) / 代数码激励线性预

测 (Algebraic Code Excited Linear Prediction, ACELP), 在编码过程中通常提取丢包隐藏 (Packet Loss Concealment, PLC) 参数、基音周期、自适应码本及其增益、固定码本及其增益等参数将其编码发送到解码端, 解码端基于上述参数合成语音。

然而, 由于 CELP/ACELP 编码方法固有的较紧密的帧间相关性, 使得一个完整语音帧信息的丢失不仅会对当前丢失帧的合成语音质量造成严重影响, 还会影响其后连续的几帧甚至十几帧。现在的丢包隐藏方法在连续丢帧时, 其恢复的编码参数不准确, 严重降低了合成语音的质量。同时, 现有方法只能实现丢帧语音的声学特征参数恢复, 尽量保持恢复语音波形的连续性, 无法恢复丢帧语音的语义信息。

对于 DNN 的 PLC 方法, 随着深度学习技术的发展, 利用对数功率谱 (Log Power Spectrum, LPS) 和频谱相位。在相关技术中, DNN 的 PLC 方法的流程图如图 4 所示, 图 4 中 $\angle Y^l$ 和 $\angle Y^f$ 分别代表过去 P 帧信号的 LPS 特征和频谱相位, $\hat{X}^l$ 和 $\angle \hat{X}^f$ 分别代表 DNN 网络预测的 LPS 特征和频谱相位。在训练阶段, 该方法利用干净语音训练 2 个 DNN 网络 (均为包含 4 层隐藏层的全连接网络), 分别用于预测丢失语音帧的 LPS 和频谱相位。在丢包恢复阶段, 首先判断当前接收帧是否为丢失信号帧, 如果不是丢失帧, 则进行正常解码。如果是丢失帧, 则调用丢包隐藏方法进行恢复。

在执行丢包恢复时, 其首先提取历史 buffer 中缓存的过去 P 帧信号, P 为正整数。然后计算每帧信号的 LPS 和频谱相位 (如图 5 所示), 并将过去 P 帧信号的 LPS 特征和频谱相位分别输入对应的 DNN 网络, 分别预测当前丢失语音帧的 LPS 特征和频谱相位, 得到预测的 LPS 特征和频谱相位后执行波形重建 (如图 6 所示), 即可得到重建的时域信号帧。

其中, 图 5 中 Y^t 表示输入的时域信号, Y^f 表示对输入时域信号进行离散傅里叶变换 (Discrete Fourier Transform, DFT) 后得到的频谱, $|\cdot|^2$ 和 $\log(\cdot)$ 分别代表平方运算和对数运算。特征提取模块的输出为 LPS 特征 Y^l 和频谱相位 $\angle Y^f$ 。图 6 中 $\hat{X}^l$ 和 $\angle \hat{X}^f$ 分别代表 LPS 特征和 DNN 网络预测的频谱相位, $\exp(\cdot)$ 代表指数运算。波形重建模块的输出为重建的时域信号 $\hat{X}^t$ 。

对于深度学习而言, 训练集数据本身的结构性对网络训练的准确性十分重要, 结构性在概率的角度分析就是数据服从某种概率分布。语音信号在频域中呈现复高斯分布, 在该条件下, 可以证明语音信号频谱的幅度服从瑞利分布, 语音信号频谱的相位服从均匀分布。LPS 通过幅度值变换而来, 保留了结构性, 可以被 DNN 网络学习。语音信号频谱的相位服从均匀分布, 其随机性较强, 很难被 DNN 网络学习。因此, 通过 DNN 网络预测的频谱相位并不准确, 其降低了合成语音信号的质量。

而本申请实施例提供的音频信号恢复方法、装置、电子设备及可读存储介质中, 在语音通话时, 若丢失了某个语音信号, 则可以通过音频信号恢复模型中的编码模块的残差网络层提取缓存的丢失语音信号前或丢失语音信号后的语音信号的时域特征信息, 并通过音频信号恢复模型中的解码模块的残差网络层合成编码模块提取的时域特征信息, 恢复出丢失的语音信号。如此, 由于可以在语音信号丢失的情况下, 通过将缓存的语音信号输入音频信号恢复模型恢复出丢失的语音信号, 因此在语音信号的传输过程中, 可以通过音频信号恢复模型恢复出丢失的语音信号, 从而避免用户感知到语音数据包的丢失, 降低语音数据包丢失所带来的负面影响, 进而可以有效地提高语音通话的质量。

本申请实施例提供的音频信号恢复方法的执行主体可以为音频信号恢复装置。示例性地, 该音频信号恢复装置可以为电子设备, 也可以为该电子设备中的部件, 例如集成电路或芯片。以下将以音频信号恢复装置为例对本申请实施例提供的音频信号恢复方法进行示例性说明。

本申请实施例提供一种音频信号恢复方法, 图 7 示出了本申请实施例提供的一种音频信号恢复方法的流程图, 该方法可以应用于音频信号恢复装置。如图 7 所示, 本申请实施例提供的音频信号恢复方法可以包括下述的步骤 101 至步骤 103。

步骤 101、音频信号恢复装置在第一语音信号丢失的情况下, 将缓存的第二语音信号输入第一音频信号恢复模型。

其中, 上述第二语音信号包括以下至少之一: 第一语音信号前的 X 帧语音信号, 第一语音信号之后的 Y 帧语音信号, X, Y 为正整数。

在本申请的一些实施例中, 第一语音信号可以为当前语音通话中的一帧或多帧语音信号, 也可以为一段录制好的音频中的一帧或多帧音频信号。

在本申请的一些实施例中, 第一语音信号丢失可以表示音频信号接收端接收到的

语音信号不是连续的语音信号，即音频信号接收端接收到的语音信号中存在缺失的语音信号帧。

示例性地，音频信号发送端可以发送一段包含 10 帧的语音信号给音频信号接收端，音频信号接收端只接收到了第 1-8 帧和第 10 帧的语音信号，没有接收到第 9 帧语音信号。则第 9 帧语音信号为丢失的语音信号，即第一语音信号。

在本申请的一些实施例中，音频信号恢复装置可以缓存第一语音信号的前几帧语音信号，也可以缓存第一语音信号的后几帧语音信号，以便后续使用。具体的，音频信号恢复装置至少缓存了第一语音信号前的 X 帧语音信号和第一语音信号之后的 Y 帧语音信号中的至少一项。

示例性地，假设第一语音信号为 5 帧语音信号中的第 4 帧语音信号，则音频信号恢复装置可以缓存第 1-3 帧语音信号，也可以缓存第 5 帧语音信号，还可以缓存第 1-3 帧语音信号和第 5 帧语音信号。

可以理解，在实时的语音通话中，音频信号恢复装置至少缓存了第一语音信号前的 X 帧语音信号；在非实时的音频中，如语音留言，音频信号恢复装置至少缓存了第一语音信号前的 X 帧语音信号和第一语音信号之后的 Y 帧语音信号。

本申请的一些实施例中，上述第一音频信号恢复模型包括编码模块和解码模块，其中，上述编码模块包括 M 层第一残差网络层，上述解码模块包括 M 层第二残差网络层，M 为正整数。

在本申请的一些实施例中，上述第一音频信号恢复模型中的编码模块和解码模块均可以采用卷积编解码网络（Convolutional Encoder-Decoder Network, CEDN）对语音信号的时域特征信息进行卷积。如此，可以仅基于语音信号的时域特征信息对丢失的语音信号进行恢复，而无需进行时频变换，从而可以避免频域处理过程中由于相位预测不准确而造成的音频失真，进而可以恢复出符合语音特征、与前后文语音连贯的高质量语音。

在本申请的一些实施例中，CEDN 网络是一种对称性网络，由预卷积层（pre-conv）、残差网络层、采样层和后卷积层（post-conv）组成。

其中，预卷积层和后卷积层可以用来改变输入第一音频信号恢复模型的信号的通道数，从而使得其输出的时域特征信息可以输入其他的网络层；采样层可以包括上采样层或下采样层，用来升高或降低时域特征信息的维度并保留有效信息；残差网络层可以用于提取时域特征信息。

需要说明的是，音频信号恢复装置在对音频信号恢复模型中的任一层进行输入，得到对应的时域特征信息后，还可以将得到的时域特征信息经过归一化处理，以提升采用反向传播算法训练的网络的收敛性。

步骤 102、音频信号恢复装置将第二语音信号经 M 层第一残差网络层依次处理，得到第一时域特征信息。

在本申请的一些实施例中，上述第一时域特征信息可以包括第二语音信号的时域波形图的特征信息，如时域波形图的周期、幅度。进一步地，时域特征信息可以以第二语音信号的时域波形特征图的形式在 M 层第一残差网络层间进行传输。

在本申请的一些实施例中，第二语音信号经 M 层第一残差网络层依次处理的详细描述，可以参见下述步骤 102b1 和步骤 102b2 中的相关描述，此处不予赘述。

可选地，在本申请的一些实施例中，上述步骤 102 可以包括下述的步骤 102a 和步骤 102b。

步骤 102a、音频信号恢复装置通过编码模块中的预卷积层，提取第二语音信号的第二时域特征信息。

在本申请的一些实施例中，音频信号恢复装置可以通过编码模块中的预卷积层，改变第二语音信号的通道数，从而将第二语音信号转换成多维度的时域特征信息，即第二时域特征信息。

示例性地，假设预卷积层的输入通道数为 1，输出通道数为 32，那么音频信号恢复装置可以将第二语音信号输入预卷积层，得到通道数为 32 的第二时域特征信息。也就是说，音频信号恢复装置通过编码模块中的预卷积层，将第二语音信号转换成 32 维时域特征信息，即第二时域特征信息。

步骤 102b、音频信号恢复装置将第二时域特征信息经 M 层第一残差网络层依次处理，得到第一时域特征信息。

在本申请的一些实施例中，第一时域特征信息相较于第二时域特征信息为具有更

多维度且包含更高语义的时域特征信息。如此，音频信号恢复装置可以将第二时域特征信息经 M 层第一残差网络层依次进行特征提取处理，得到第一时域特征信息。

如此，由于音频信号恢复装置可以先通过编码模块中的预卷积层，提取出第二语音信号的时域特征信息，再将该第二时域特征信息经 M 层第一残差网络依次处理，得到第一时域特征信息。因此，音频信号恢复装置可以通过逐层提取语音信号的时域特征信息，最终得到具有更多维度且包含更高语义的时域特征信息，从而可以通过该包含更高语义的时域特征信息完成对丢失语音信号的恢复，提高语音通信的质量。

可选地，在本申请的一些实施例中，上述步骤 102b 可以包括下述的步骤 102b1 和步骤 102b2。

步骤 102b1、音频信号恢复装置将输入第 i 层第一残差网络层的时域特征信息，通过第 i 层第一残差网络层的第一卷积层处理，得到第三时域特征信息，并通过第 i 层第一残差网络层的第二卷积层以及第三卷积层处理，得到第四时域特征信息。

其中，上述第一卷积层、第二卷积层以及第三卷积层均为一维卷积层，第一卷积层的输入通道数和输出通道数相同，第二卷积层的卷积核与第一卷积层的卷积核不同、且第三卷积层的卷积核与第一卷积层的卷积核相同， $i \in \{1, \dots, M\}$ 。

在本申请的一些实施例中，第三时域特征信息为包含第一卷积层的卷积核大小的信息尺度的时域特征信息，第四时域特征信息为包含第二卷积层的卷积核大小的信息尺度的时域特征信息。即，第三时域特征信息和第四时域特征信息为尺度不同的时域特征信息。

可以理解，第 i 层第一残差网络层可以为 M 层第一残差网络层中的任一层第一残差网络层。

示例性地，第 i 层第一残差网络层可以为 M 层第一残差网络层中的第一层第一残差网络层，也可以为最后一层第一残差网络层，还可以为中间的任一层第一残差网络层。

在相关技术中，为了提高神经网络的性能，传统的经典卷积神经网络通常是通过增加神经网络的层数来更好的提取特征。然而，一味的增加神经网络的层数不仅不会提高神经网络的性能，反而会出现梯度消失问题，神经网络也会不可避免地发生退化从而阻碍神经网络模型的训练，此时层数少的神经网络在性能上反而优于层数多的网络。此外，神经网络模型层数的加深通常会使得神经网络模型的复杂度变高，容易导致神经网络发生过拟合。因此，本申请实施例提供的音频信号恢复方法还提供了一种残差网络（Modified-Resnet）层，其结构示意图如图 8 所示。

可选地，本申请的一些实施例中，如图 8 所示，上述残差网络层可以由并行的两条支路构成。其中一条支路包含卷积层 81 和其所对应的归一化层；另一条支路包含卷积层 82 和其所对应的归一化层、激活函数，卷积层 83 和其所对应的归一化层、激活函数。其中，卷积层 81、卷积层 82 和卷积层 83 均为一维卷积层，卷积层 81 的输入通道数与输出通道数相同，卷积层 82 的卷积核与卷积层 81 的卷积核不同、且卷积层 83 的卷积核与卷积层 81 的卷积核相同。如此，音频信号恢复装置可以通过多个一维卷积层增加通道间的信息交互，同时可以采用两条支路并行的结构，通过分别使用卷积核大小不同的卷积层对输入的特征进行卷积，并将得到的不同尺度的时域特征信息进行融合，因此可以使得每一层第一残差网络都可以提取到更加丰富的时域特征信息。

可选地，在本申请的一些实施例中，上述第一残差网络层可以包括并行的两条支路，一条支路包括第一卷积层，另一条支路包括第二卷积层和第三卷积层。也就是说，音频信号恢复装置在将输入第 i 层第一残差网络层的时域特征信息输入第 i 层第一残差网络层之后，可以通过该两条支路同时得到第三时域特征信息和第四时域特征信息。

可选地，在本申请的一些实施例中，以图 8 所示残差网络层的结构为例，卷积层 81 为第 i 层第一残差网络层的第一卷积层，卷积层 82 为第 i 层第一残差网络层的第二卷积层，卷积层 83 为第 i 层第一残差网络层的第三卷积层。音频信号恢复装置可以将输入第 i 层第一残差网络层的时域特征信息，通过卷积层 81 和归一化层处理，得到第三时域特征信息；并通过第 i 层第一残差网络层的卷积层 82 及其对应的归一化层、卷积层 83 及其对应的归一化层处理，得到第四时域特征信息。

需要说明的是，对于第 i 层第一残差网络层的第二卷积层和第三卷积层，音频信

号恢复装置可以先将输入第 i 层第一残差网络层的时域特征信息，通过激活函数层处理，使得输入第 i 层第一残差网络层的时域特征信息增加非线性特征，从而使得第二卷积层和第三卷积层可以更好的对输入第 i 层第一残差网络层的时域特征信息进行特征提取。

在本申请的一些实施例中，可以采用指数线性单元（Exponential Linear Unit, ELU）激活函数以及权重归一化来保证第一音频信号恢复模型的稳定性。在实际实现中，还可以通过其他的激活函数和归一化方法来保证第一音频信号恢复模型的稳定性。本申请实施例不作具体限定。

步骤 102b2、音频信号恢复装置基于第三时域特征信息和第四时域特征信息，得到第 i 层第一残差网络层输出的时域特征信息。

在本申请的一些实施例中，音频信号恢复装置可以将第三时域特征信息和第四时域特征信息中不同尺度的时域特征信息相加，得到第 i 层第一残差网络层输出的时域特征信息。

可以理解，第 M 层第一残差网络层输出的时域特征信息则为第一时域特征信息。

需要说明的是，对于 M 层第一残差网络层中的每层第一残差网络，音频信号恢复装置均可以对其执行上述步骤 102b1 和步骤 102b2，最终得到第一时域特征信息。

在本申请的一些实施例中，可以为第二卷积层和第三卷积层设置不同的输入通道数、输出通道数以及卷积核，以使得得到的第四时域特征信息的通道数可以与第三时域特征信息的通道数相同但尺度不同，从而使得第三时域特征信息可以与第四时域特征信息相加可以得到更加丰富的时域特征信息。

示例性地，如图 8 所示，音频信号恢复装置可以在卷积层 82 的输入通道数设置为输出通道数的 2 倍，卷积核设置为 3，再将卷积层 83 的输出通道数设置为输入通道数的 2 倍，卷积核设置为 1，以使得经卷积层 82 和卷积层 83 处理后输出的时域特征信息的通道数，与经卷积层 81 处理后输出的时域特征信息的通道数相同但尺度不同。

例如，假设输入第 i 层第一残差网络层的时域特征信息的通道数为 128，则第一卷积层的输入通道数为 128，输出通道数为 128，卷积核为 1；第二卷积层的输入通道数为 128，输出通道数为 64，卷积核为 3；第三卷积层的输入通道数为 64，输出通道数为 128，卷积核为 1。如此，输入第 i 层第一残差网络层的时域特征信息通过第一卷积层处理，可以得到通道数为 128，卷积核为 1 的第三时域特征信息；并通过第 i 层第一残差网络层的第二卷积层以及第三卷积层处理，得到通道数为 128，卷积核为 3 的第四时域特征信息。音频信号恢复装置在将第三时域特征信息与第四时域特征信息相加之后，可以得到包含卷积核为 1 和卷积核为 3 两个尺度的时域特征信息。

在实际实现中，可以设置其他大小的卷积核和输入输出通道数，以得到不同尺度的时域特征信息。本申请实施例不作具体限定。

如此，由于第一残差网络层可以采用两条支路并行的结构，通过分别使用卷积核大小不同的卷积层对输入的特征进行卷积，并将得到的不同尺度的时域特征信息进行融合，因此可以使得每一层第一残差网络都可以提取到更加丰富的时域特征信息。

可选地，在本申请的一些实施例中，上述编码模块还可以包括 M 层下采样层， M 层下采样层与 M 个第一残差网络层交替排列； M 层下采样层的输出通道数按照层级从高至低依次增大。

示例性地，上述步骤 102b 可以包括下述的步骤 102b3 和步骤 102b4。

步骤 102b3、音频信号恢复装置将输入第 x 层第一残差网络层的时域特征信息，通过第 x 层第一残差网络层处理，得到第五时域特征信息。

步骤 102b4、音频信号恢复装置将第五时域特征信息通过与第 x 层第一残差网络层连接的第 x 层下采样层输出。

其中， $x \in \{1, \dots, M\}$ 。

在本申请的一些实施例中，下采样层可以用来增加输入的时域特征信息的通道数，以降低时域特征信息的维度并保留有效信息，防止时域特征信息丢失现象的发生。

示例性地，假设第 x 层第一残差网络层的通道数为 32，则第 x 层下采样层的输入通道数为 32，音频信号恢复装置可以设置第 x 层下采样层的输出通道数为 62，在经过第 x 层残差网络提取时域特征信息后，增加时域特征信息的通道数，防止时域特

征信息的丢失。

在本申请的一些实施例中，编码模块中的下采样层的输出通道数可以按照层级从高至低依次增大。其中，M层下采样层的输出通道数可以为预先设置好的，也可以为按照固定数值按照层级从高至低依次增大。

5 示例性地，以编码模块包括4层下采样层，且该4层下采样层的输出通道数为预先设置好的为例。第一层下采样层的输入通道数可以预先设置为32，输出通道数可以预先设置为64；第二层下采样层的输入通道数可以预先设置为64，输出通道数可以预先设置为128；第三层下采样层的输入通道数可以预先设置为128，输出通道数可以预先设置为256；第四层下采样层的输入通道数可以预先设置为256，输出通道数可以预先设置为512。

10 示例性地，以编码模块包括4层下采样层，且该4层下采样层的输出通道数按照固定数值按照层级从高至低依次增大为例。假设固定数值为64，则第一层下采样层的输入通道数可以为32，输出通道数可以为96；第二层下采样层的输入通道数可以为96，输出通道数可以为160；第三层下采样层的输入通道数可以为160，输出通道数可以为224；第四层下采样层的输入通道数可以为224，输出通道数可以为288。

15 需要说明的是，M层下采样层的输入通道数随层级高低依次从高至低逐渐增大是因为M层下采样层的输出通道数在增大，为了可以下采样层之间的连接与时域特征信号的传输，因此M层下采样层的输入通道数也会随之增大。

20 在本申请的一些实施例中，音频信号恢复装置可以通过M层下采样层对时域特征信息进行逐层下采样。具体地，音频信号恢复装置可以通过采用渐进式采样和具有不同感受野的滤波器滤波对时域特征信息进行逐层下采样，从而可以有效缓解大范围直接下采样造成的时域特征图波形细节的损失，并且可以通过对对不同时域的时域特征信息的通道数进行调整，增加时域特征信息的表征性。

25 本申请的一些实施例中，网络层的层级高低可以根据该网络层在第一音频恢复模型中的位置确定。具体地，第一音频恢复模型中最上层的网络层层级最高，第一音频恢复模型中处于中间位置的网络层层级根据位置的下移而降低，第一音频恢复模型中最下层的网络层层级最低。

30 在本申请的一些实施例中，第一音频信号恢复模型的编码模块中可以包含M层第一残差网络层和M层下采样层，编码模块中各网络层可以由第一残差网络层和下采样层交替进行排布。即，输入第x层第一残差网络层的时域特征信息在通过第x层第一残差网络层进行时域特征信息提取之后，可以随即通过第x层下采样层增加时域特征信息的通道数，防止时域特征信息丢失。

35 示例性地，假设编码模块包含一层预卷积层、一层后卷积层、4层第一残差网络层以及4层下采样层，编码模块的网络结构如图9所示。音频信号恢复装置可以将第二语音信号输入编码模块，经预卷积层901处理，将第二语音信号先通过一个输入通道数为1、输出通道数为32、卷积核大小为7、步长为1的一维卷积，并将预卷积层901输出的时域特征信息通过归一化函数处理，得到第二语音信号的第二时域特征信息；然后将第二时域特征信息通过第一残差网络层902处理，即第一层第一残差网络层，得到第一层第一残差网络层输出的通道数为32的时域特征信息；将第一残差网络层902输出的时域特征信息输入下采样层903，即第一层下采样层，先通过一个输入通道数为32、输出通道数为64、卷积核大小为4、步长为2的一维卷积，并将一维卷积输出的时域特征信息经归一化函数处理，得到下采样层903输出的通道数为64的时域特征信息；再将下采样层903输出的时域特征信息经第一残差网络层904处理，即第二层第一残差网络层，得到第一残差网络层904输出的通道数为64的时域特征信息；然后将该通道数为64的时域特征信息输入下采样层905，通过一个输入通道数为64、输出通道数为128、卷积核大小为8、步长为4的一维卷积，并将该一维卷积输出的时域特征信息通过归一化函数处理，得到通道数为128的时域特征信息；再将该通道数为128的时域特征信息输入第一残差网络层906，即第三层第一残差网络层，得到更高语义的通道数为128的时域特征信息；然后将第一残差网络层906输出的时域特征信息输入下采样层907，即第三层下采样层，通过一个输入通道数为128、输出通道数为256、卷积核大小为10、步长为5的一维卷积，并将该一维卷积输出的时域特征信息通过归一化函数，得到通道数为256的时域特征信息；再将该通道数为256的时域特征信息输入第一残差网络层908，即第四层第一残差网络层，得到更高语义的通道数为256的时域特征信息；然后将第一残差网络层908输出

的时域特征信息输入下采样层 909, 即第四层下采样层, 通过一个输入通道数为 256、输出通道数为 512、卷积核大小为 16、步长为 8 的一维卷积, 并将该一维卷积输出的时域特征信息通过归一化函数, 得到通道数为 512 的时域特征信息; 最后将该通道数为 512 的时域特征信息输入后卷积层 910, 先通过激活函数, 再通过一个输入通道数为 512、输出通道数为 128、卷积核大小为 7、步长为 1 的一维卷积, 并将该一维卷积输出的时域特征信息通过归一化函数, 得到通道数为 128 的第一时域特征信息, 并将该第一时域特征信息从编码模块中输出。

如此, 由于音频信号恢复装置可以通过 M 层下采样层, 对语音信号的时域特征信息进行逐层的特征提取和下采样, 因此可以在增加提取到的时域特征信息的表征性的同时, 避免时域特征信息的损失。

步骤 103、音频信号恢复装置将第一时域特征信息经 M 层第二残差网络层依次处理, 得到目标语音信号。

在本申请的一些实施例中, 音频信号恢复装置将第一时域特征信息经 M 层第二残差网络层依次处理的过程, 实际上是根据第一时域特征信息恢复第一语音信号, 得到目标语音信号的过程。

在本申请的一些实施例中, 上述音频信号恢复装置在得到目标语音信号之后, 可以将目标语音信号作为恢复的第一语音信号进行输出。

示例性地, 假设音频信号恢复装置缓存有第 1、2、3 帧的语音信号, 即第二语音信号, 在第 4 帧语音信号丢失的情况下, 即第一语音信号丢失的情况下, 音频信号恢复装置可以将第 1、2、3 帧的语音信号输入第一音频信号恢复模型中, 通过第一音频信号恢复模型中的预处理模块将第 1、2、3 帧的语音信号映射为 1 帧语音信号, 将该 1 帧语音信号经编码模块提取出其所对应的时域特征信息; 并将提取出的时域特征信息经解码模块生成 1 帧语音信号, 即目标语音信号, 该生成的 1 帧语音信号即为恢复的第 4 帧语音信号。

在本申请的一些实施例中, 音频信号恢复装置在完成对第一语音信号的恢复之后, 可以对缓存的语音信号进行更新。

示例性地, 音频信号恢复装置可以缓存第一语音信号前的 3 帧语音信号, 即缓存有第 1、2、3 帧的语音信号, 在音频信号恢复装置完成对第 4 帧语音信号的恢复之后, 可以对第 4 帧语音信号进行缓存, 使得音频信号恢复装置中缓存的语音信号更新为第 2、3、4 帧语音信号。

在本申请的一些实施例中, 音频信号恢复装置对缓存的语音信号进行更新时, 可以将每一帧语音信号覆盖前一帧语音信号的方式更新缓存, 也可以以最后一帧语音信号覆盖最前一帧语音信号的方式更新缓存。本申请实施例不作具体限定。

在本申请的一些实施例中, 在上述第一语音信号没有丢失的情况下, 音频信号恢复装置可以对其进行正常解码并更新缓存。

示例性地, 以音频信号恢复装置缓存有第 1 帧语音信号 X_1 、第 2 帧语音信号 X_2 、第 3 帧语音信号 X_3 、第 4 帧语音信号 X_4 、第 5 帧语音信号 X_5 、第 6 帧语音信号 X_6 、第 7 帧语音信号 X_7 、第 8 帧语音信号 X_8 、第 9 帧语音信号 X_9 、第 10 帧语音信号 X_{10} , 正在接收第 11 帧语音信号 X_{11} 为例。如图 10 所示, 为音频信号恢复装置接收语音信号时对丢失的语音信号进行恢复的流程图。音频信号恢复装置首先判断第 11 帧语音信号 X_{11} 是否丢失。在第 11 帧语音信号 X_{11} 没有丢失的情况下, 音频信号恢复装置可以对其进行正常解码, 输出第 11 帧语音信号 X_{11} 并更新缓存; 在第 11 帧语音信号 X_{11} 丢失的情况下, 音频信号恢复装置可以将缓存的第 1 帧语音信号 X_1 、第 2 帧语音信号 X_2 、第 3 帧语音信号 X_3 、第 4 帧语音信号 X_4 、第 5 帧语音信号 X_5 、第 6 帧语音信号 X_6 、第 7 帧语音信号 X_7 、第 8 帧语音信号 X_8 、第 9 帧语音信号 X_9 、第 10 帧语音信号 X_{10} , 输入第一音频信号恢复模型, 生成目标语音信号, 将目标语音信号作为恢复的第 11 帧语音信号 X_{11} 进行输出并更新缓存。可以理解, 更新后的缓存包含第 2 帧语音信号 X_2 、第 3 帧语音信号 X_3 、第 4 帧语音信号 X_4 、第 5 帧语音信号 X_5 、第 6 帧语音信号 X_6 、第 7 帧语音信号 X_7 、第 8 帧语音信号 X_8 、第 9 帧语音信号 X_9 、第 10 帧语音信号 X_{10} 、第 11 帧语音信号 X_{11} 。

可选地, 在本申请的一些实施例中, 上述步骤 103 可以包括下述的步骤 103a 和步骤 103b。

步骤 103a、音频信号恢复装置将第一时域特征信息经 M 层第二残差网络层依次处理, 得到第六时域特征信息。

在本申请的一些实施例中，第一时域特征信息相较于第六时域特征信息为具有更多维度且包含更高语义的时域特征信息。换句话说，第一时域特征信息与第六时域特征信息为不同尺度的时域特征信息。

进一步可选地，在本申请的一些实施例中，上述步骤 103a 可以包括下述的步骤 5 103a1 和步骤 103a2。

步骤 103a1、音频信号恢复装置将输入第 j 层第二残差网络层的时域特征信息，通过第 j 层第二残差网络层的第四卷积层处理，得到第七时域特征信息，并通过第 j 层第二残差网络层的第五卷积层以及第六卷积层进行处理，得到第八时域特征信息。

其中，第四卷积层、第五卷积层以及第六卷积层均为一维卷积层，第四卷积层的输入通道数和输出通道数相同，第五卷积层的卷积核与第四卷积层的卷积核不同、且第六卷积层的卷积核与第四卷积层的卷积核相同， $j \in \{1, \dots, M\}$ 。

步骤 103a2、音频信号恢复装置基于第七时域特征信息和第八时域特征信息，得到第 j 层第二残差网络层输出的时域特征信息。

对于第二残差网络层的第四卷积层、第五卷积层和第六卷积层的详细描述，可以参见上述步骤 102b1 和步骤 102b2 中关于第一卷积层、第二卷积层和第三卷积层的相

关描述，此处不再赘述。如此，由于第二残差网络层可以采用两条支路并行的结构，通过分别使用卷积核大小不同的卷积层对输入的特征进行卷积，并将得到的不同尺度的时域特征信息进行融合，因此可以使得每一层第一残差网络都可以提取到更加丰富的时域特征信息。

可选地，在本申请的一些实施例中，上述解码模块还可以包括 M 层上采样层， M 层上采样层与 M 个第二残差网络层交替排列； M 层上采样层的输出通道数按照层级从高至低依次降低。

示例性地，上述步骤 103a 还可以包括下述的步骤 103a3。

步骤 103a3、音频信号恢复装置将第 y 层上采样层输出的时域特征信息，通过与第 y 层上采样层连接的第 y 层第二残差网络层处理，得到第 y 层第二残差网络层输出的时域特征信息。

其中， $y \in \{1, \dots, M\}$ 。

在本申请的一些实施例中，上采样层可以用来减少输入的时域特征信息的通道数，以通过对第一时域特征信息逐层上采样，将时域特征信息重建为原始大小，并增加时域特征信息的维度，降低时域特征信息的尺度，实现语音信号的恢复。

示例性地，假设第 y 层第二残差网络层的通道数为 128，则与第 y 层第二残差网络层连接的第 y 层下采样层的输入通道数为 128，音频信号恢复装置可以设置第 y 层下采样层的输出通道数为 62，在经过第 y 层残差网络提取时域特征信息前，减少时域特征信息的通道数，降低时域特征信息的维度，实现语音信号的恢复。

在本申请的一些实施例中，输入的时域特征信息可以被解码模块中的上采样层的输出通道数按照层级从高至低依次降低。其中， M 层上采样层的输出通道数可以为预先设置好的，也可以为按照固定数值按照层级从高至低依次降低。

示例性地，以解码模块包括 4 层上采样层，且该 4 层上采样层的输出通道数为预先设置好的为例。第一层上采样层的输入通道数可以预先设置为 512，输出通道数可以预先设置为 218；第二层上采样层的输入通道数可以预先设置为 218，输出通道数可以预先设置为 128；第三层上采样层的输入通道数可以预先设置为 128，输出通道数可以预先设置为 64；第四层上采样层的输入通道数可以预先设置为 64，输出通道数可以预先设置为 32。

示例性地，以编码模块包括 4 层上采样层，且该 4 层上采样层的输出通道数按照固定数值按照层级从高至低依次降低为例。假设固定数值为 64，则第一层上采样层的输入通道数可以为 288，输出通道数可以为 224；第二层上采样层的输入通道数可以为 224，输出通道数可以为 160；第三层上采样层的输入通道数可以为 160，输出通道数可以为 96；第四层上采样层的输入通道数可以为 96，输出通道数可以为 32。

需要说明的是， M 层上采样层的输入通道数随层级高低依次从高至低逐渐降低是因为 M 层上采样层的输出通道数在降低，为了时域特征信号能在两个上采样层之间传输，因此 M 层上采样层的输入通道数也会随之降低。

在本申请的一些实施例中，音频信号恢复装置可以通过 M 层上采样层对时域特征信息进行逐层上采样，通过采用渐进式采样和具有不同感受野的滤波器滤波可以有效缓解大范围直接上采样造成的时域特征图波形细节的损失，并且可以通过对对

尺度的时域特征信息的通道数进行调整,实现语音信号的恢复。

在本申请的一些实施例中,第一音频信号恢复模型的解码模块中可以包含M层上采样层和M层第二残差网络层,解码模块中各网络层可以由上采样层和第二残差网络层交替进行排布。即,输入第i层上采样层的时域特征信息在通过第i层上采样层减少时域特征信息的通道数之后,可以随即通过第i层第二残差网络层进行时域特征信息的特征提取,以恢复语音信号。

示例性地,假设解码模块包含一层预卷积层、一层后卷积层、4层上采样层以及4层第二残差网络层,解码模块的网络结构如图11所示。音频信号恢复装置可以先将第一时域特征信息输入解码模块,经预卷积层1101处理,将第一时域特征信息先通过一个输入通道数为128、输出通道数为512、卷积核大小为7、步长为1的一维卷积,并将预卷积层1101输出的时域特征信息通过归一化函数处理,得到通道数为32的时域特征信息;然后将该通道数为32的时域特征信息输入上采样层1102,即第一层上采样层,通过一个输入通道数为512、输出通道数为256、卷积核大小为16、步长为8的一维卷积,并将该一维卷积输出的时域特征信息通过归一化函数,得到通道数为256的时域特征信息;再将该通道数为256的时域特征信息输入第二残差网络层1103,即第一层第二残差网络层,得到通道数为256的时域特征信息;然后将该通道数为256的时域特征信息输入上采样层1104,即第二层上采样层,通过一个输入通道数为256、输出通道数为128、卷积核大小为10、步长为5的一维卷积,并将该一维卷积输出的时域特征信息通过归一化函数,得到通道数为128的时域特征信息;再将该通道数为128的时域特征信息输入第二残差网络层1105,即第二层第二残差网络层,得到通道数为128的时域特征信息;然后将该通道数为128的时域特征信息输入上采样层1106,即第三层上采样层,通过一个输入通道数为128、输出通道数为64、卷积核大小为8、步长为4的一维卷积,并将该一维卷积输出的时域特征信息通过归一化函数,得到通道数为64的时域特征信息;再将该通道数为64的时域特征信息输入第二残差网络层1107,即第三层第二残差网络层,得到通道数为64的时域特征信息;然后将该通道数为64的时域特征信息输入上采样层1108,即第四层上采样层,通过一个输入通道数为64、输出通道数为32、卷积核大小为4、步长为2的一维卷积,并将该一维卷积输出的时域特征信息通过归一化函数,得到通道数为32的时域特征信息;再将该通道数为32的时域特征信息输入第二残差网络层1109,即第四层第二残差网络层,得到通道数为32的时域特征信息;最后将该通道数为32的时域特征信息输入后卷积层1110,通过一个输入通道数为32、输出通道数为1、卷积核大小为7、步长为1的一维卷积,并将该一维卷积输出的时域特征信息通过归一化函数,得到通道数为1的目标语音信号,并将该目标语音信号从解码模块中输出。

如此,由于音频信号恢复装置可以通过M层上采样层,对语音信号的时域特征信息进行逐层的上采样和特征提取,因此可以根据编码模块提取到的时域特征信息,恢复出丢失的语音信号。

步骤103b、音频信号恢复装置将第六时域特征信息输入解码模块的后卷积层处理,得到目标语音信号。

在本申请的一些实施例中,音频信号恢复装置可以通过解码模块中的后卷积层,改变第六时域特征信息的通道数,从而将第六时域特征信息进行合成,得到目标语音信号。

示例性地,假设后卷积层的输入通道数为32,输出通道数为1,那么音频信号恢复装置可以将第六时域特征信息输入后卷积层,将第六时域特征信息的32个通道合为1个通道,从而得到目标语音信号。可以理解,音频信号恢复装置可以通过解码模块中的后卷积层,将时域特征信息转换成语音信息。

如此,由于音频信号恢复装置可以先将第一时域特征信息经M层第二残差网络层依次处理,得到第六时域特征信息,再通过解码模块中的后卷积层,得到目标语音信号。因此,音频信号恢复装置可以通过经M层第二残差网络逐层还原出丢失语音信号的时域特征信息,并通过解码模块中的后卷积层将时域特征信息转换为语音信号,最终会恢复出丢失的语音信号。

在本申请实施例提供的音频信号恢复方法中,在语音通话时,若丢失了某个语音信号,则可以通过音频信号恢复模型中的编码模块的残差网络层提取缓存的丢失语音信号前或丢失语音信号后的语音信号的时域特征信息,并通过音频信号恢复模型中的

解码模块的残差网络层合成编码模块提取的时域特征信息，恢复出丢失的语音信号。如此，由于可以在语音信号丢失的情况下，通过将缓存的语音信号输入音频信号恢复模型恢复出丢失的语音信号，因此在语音信号的传输过程中，可以通过音频信号恢复模型恢复出丢失的语音信号，从而避免用户感知到语音数据包的丢失，降低语音数据包丢失所带来的负面影响，进而可以有效地提高语音通话的质量。

可选地，在本申请的一些实施例中，在第二语音信号为多帧语音信号的情况下，在上述步骤 102 之前，本申请实施例提供的音频信号恢复方法还可以包括下述的步骤 104，结合步骤 104，上述步骤 102 可以包括下述的步骤 102c。

步骤 104、音频信号恢复装置将多帧语音信号输入第一音频信号恢复模型之后，将多帧语音信号映射为一帧语音信号。

在本申请的一些实施例中，在第二语音信号为多帧语音信号的情况下，该多帧语音信号可以包括第一语音信号前的至少两帧语音信号，也可以包含第一语音信号之后的至少两帧语音信号，还可以包含第一语音信号前的至少一帧语音信号和第一语音信号之后的至少一帧语音信号。本申请实施例不作具体限定。

在本申请的一些实施例中，上述第一音频信号恢复模型还可以包括预处理模块，音频信号恢复装置可以通过预处理模块将多帧语音信号映射为一帧语音信号。

示例性地，音频信号恢复装置可以提取该多帧语音信号的时域特征信息，映射至一帧语音信号中。

步骤 102c、音频信号恢复装置将一帧语音信号经 M 层第一残差网络层依次处理，得到第一时域特征信息。

对于音频信号恢复装置将一帧语音信号经 M 层第一残差网络层依次处理的过程的描述，可以参见上述步骤 102 中音频信号恢复装置将第二语音信号经 M 层第一残差网络层依次处理的过程的相关描述，此处不再赘述。

如此，由于频信号恢复装置可以在第二语音信号为多帧语音信号的情况下，通过将多帧语音信号映射为一帧语音信号，对该一帧语音信号经 M 层第一残差网络层依次处理，得到第一时域特征信息。因此可以只对该多帧语音信号中有用的部分进行处理，避免对多帧语音信号中无用的部分的处理，节省的频信号恢复模型的资源。

在本申请的一些实施例中，在上述步骤 101 之前，本申请实施例提供的音频信号恢复装置还可以包括下述的步骤 105 和步骤 106。

步骤 105、音频信号恢复装置获取至少一组训练语音样本。

其中，每组训练语音样本包括标记为正确接收的语音包和标记为丢失的语音包。

在本申请的一些实施例中，上述每组训练语音样本中标记为丢失的语音包可以为根据真实语音模拟语音丢包得到的。可以理解的是，实际实现中，一个完整的语音包可以包括上述正确接收的语音包和标记为丢失的语音包。

示例性地，假设真实语音包含 5 帧语音信号，则包含丢失语音信号的语音包可以包含丢失第 4 帧语音信号的第 1-3 帧和第 5 帧语音信号。

步骤 106、音频信号恢复装置基于至少一组训练语音样本，对第二音频信号恢复模型训练，得到第一音频信号恢复模型。

其中，第二音频信号恢复模型中的损失函数包括以下至少两项：对抗损失参数、多分辨率时频域信号损失参数以及相对特征匹配损失参数。

可以理解，上述第二音频信号恢复模型为不具备音频信号恢复功能的原始模型。即，第二音频信号恢复模型为未经训练的第一音频信号恢复模型。

在本申请的一些实施例中，上述第二音频信号恢复模型可以为基于生成对抗网络（Generative Adversarial Network, GAN）的深度学习网络模型。其中，GAN 网络通常由两个同时训练、相互竞争的网络组成，即生成器（Generator）和判别器（Discriminator）。

示例性地，在 GAN 网络的训练过程中，如图 12 所示，判别器的输入由两部分组成，分别是真实样本数据和生成器生成的数据，其输出通常是一个概率值，表示判别器认定输入是真实数据的分布概率。若输入来自真实数据，则输出为 1，否则输出 0。生成器可以接收一个随机变量，生成假样本数据，判别器判断该假样本数据为真实数据的分布概率；并且判别器的输出还会反馈给生成器，用于指导生成器的训练。在理想情况下，判别器无法判别输入数据是来自真实样本数据还是生成器生成的数据，即判别器每次的输出概率都为 0.5，此时生成器生成的数据无限接近真实样本数据，模型达到最优。在实际应用中，生成器和判别器通常采用深度神经网络来实现。

在实际实现中，GAN 网络对生成器输入的随机变量基本没有限制，随机变量可以为随机噪声或者为符合某种分布的变量。生成器理论上可以逐渐学习任何概率分布，训练后的生成器可以生成和真实数据无限接近的、但又不是和真实数据完全一样的数据，即生成器实际上是学习了训练数据的一个近似分布。判别器的主要目的是判断输入的数据是否为真实的样本数据，并提供反馈以指导生成器进行训练。在 GAN 网络的训练过程中，生成器和判别器会各自更新自身的参数使得损失最小，通过不断迭代优化，最终达到纳什均衡状态。可以理解，上述第二音频信号恢复模型为 GAN 网络中的生成器。

在本申请的一些实施例中，音频信号恢复装置可以将获取到的训练语音样本中包含丢失语音信号的语音包，输入第二音频信号恢复模型，生成一个语音信号；音频信号恢复装置可以通过判别器判断生成的语音信号与真实语音的相似度，判断第二音频信号恢复模型恢复丢失语音信号是否成功，并根据判别器的判断结果对生成器模型参数进行调整，以提升生成的语音信号与真实语音的相似度。

具体地，音频信号恢复装置可以缓存语音包中未丢失的语音信号。第二音频信号恢复模型可以通过编码模块提取缓存的语音信号的时域特征信息，并通过解码模块根据提取的时域特征信息，生成语音信号。音频信号恢复装置可以通过判别器判断生成的语音信号与真实语音的相似度，判断第二音频信号恢复模型恢复丢失语音信号是否成功。音频信号恢复装置可以在判别器判断生成的语音信号与真实语音的相似度大于或等于相似度阈值的情况下，判断第二音频信号恢复模型成功恢复丢失的语音信号；在判别器判断生成的语音信号与真实语音的相似度小于相似度阈值的情况下，判断第二音频信号恢复模型恢复丢失的语音信号失败，音频信号恢复装置可以根据判别器的判断结果调整生成器模型参数，以提高再次生成的语音信号与真实语音的相似度。

图 13 示出了基于 GAN 网络的丢包隐藏方法训练流程图，即第二音频信号恢复模型的训练流程图。生成器由预处理模块、编码模块、解码模块组成，其中，编码模块用于提取语音信号的时域特征信息，解码模块通过编码模块提取的时域特征信息合成为语音信号。在训练过程中，生成器生成语音信号后，可以将其与真实语音信号一起输入到判别器。判别器对输入的语音信号进行判决，并将判别结果反馈给生成器。生成器根据反馈结果对其模型参数进行更新，逐渐使其生成的语音信号接近真实语音信号。在训练过程中，判别器也会根据判别器损失函数对判别器的网络参数进行更新，在生成器提高语音信号生成质量的同时，判别器也可以同时提高其判别能力。

在本申请的一些实施例中，音频信号恢复装置可以使用多尺度短时傅里叶变换（Multi-Scale Short Time Fourier Transform, MS-STFT）判别器对生成器生成的语音信号和真实语音信号进行判别。实际实现中，还可以使用 Multi-Period Discriminator 判别器、Multi-Scale Discriminator 判别器等判别器对生成器生成的语音信号和真实语音信号进行判别。本申请实施例不作具体限定。

示例性地，以采用 MS-STFT 判别器进行判别为例。如图 14 所示，MS-STFT 判别器包括 5 个短时傅里叶变换（Short Time Fourier Transform, STFT）鉴别器。在对输入语音信号进行判别时，5 个 STFT 鉴别器分别对输入的语音信号进行 FFT 变换，其对输入语音信号执行 FFT 变换的帧长为：[128, 256, 512, 1024, 2048]，对应的帧移为：[32, 64, 128, 256, 512]，得到输入的语音信号对应的频谱数据。然后将频谱数据输入判别器中进行处理，不同尺度的判别器可以从不同层次的语音结构中学习特征，从而促使生成器提高合成语音质量。

在实际应用中，STFT 鉴别器通常选择复数谱作为输入，其有助于判别器学习语音中更精细的细节特征。同时，由于语音的上下文信息对丢帧语音的恢复是极其重要的，因此在 STFT 鉴别器中增加了空洞卷积（Dilated Convolution）来提高卷积层对语音上下文信息的提取效率，空洞卷积允许卷积增加感受野大小从而捕获更大的上下文信息，并且不同的感受野可以在序列建模中相互补充，从而提高判别器性能。

在 GAN 网络训练时，判别器损失函数可以通过公式（9）计算得到：

$$L_D = \frac{1}{K} \sum_{k=1}^K [\max(0, 1 - D_k(x)) + \max(0, 1 + D_k(\hat{x}))] \quad \text{公式}$$

其中， L_D 为判别器损失函数， K 为判别器个数， $\hat{x}$ 为生成器生成的语音信号， $D_k(\hat{x})$ 为输入生成的语音信号时第 k 个判别器的判别输出， x 为真实语音， $D_k(x)$ 为输入真实语音时第 k 个判别器的判别输出， $\max(\cdot)$ 为取最大值操作。在 GAN 训练时，

可以通过最小化判别器损失函数 L_D 来更新判别器的网络参数。

在本申请的一些实施例中，第二音频信号恢复模型中的损失函数可以为一个由多约束项构成的损失函数。在第二音频信号恢复模型训练过程中，可以通过该多约束项构成的损失函数，提高生成语音信号的质量，从而更准确的对丢失语音信号进行恢复，提高语音通话的质量。

在本申请的一些实施例中，上述多约束项构成的损失函数可以包括以下至少两项：对抗损失、多分辨率时频域信号损失、相对特征匹配损失。

在本申请的一些实施例中，上述对抗损失参数可以通过下述的公式（10）计算得到：

$$L_g(\hat{x}) = \frac{1}{K} \sum_{k=1}^K \max(0, 1 - D_k(\hat{x}))$$

公式（10）

其中， $L_g(\hat{x})$ 为对抗损失，K为判别器个数， $\hat{x}$ 为生成器生成的语音信号， $D_k(\hat{x})$ 为第k个判别器的判别输出， $\max(0, 1 - D_k(\hat{x}))$ 为取最大值操作。

在本申请的一些实施例中，上述多分辨率时频域信号损失参数可以通过下述的公式（11）计算得到：

$$L_{t-f}(x, \hat{x}) = \sum_{K \in \{2^6, \dots, 2^{11}\}} \left\{ \sum_n \sum_k \|S(n, k) - \hat{S}(n, k)\|_1 + \alpha_{t-f} \sum_n \sum_k \|\log S(n, k) - \log \hat{S}(n, k)\|_2 \right\}$$

公式（11）

其中， $L_{t-f}(x, \hat{x})$ 为多分辨率时频域信号损失， $S(n, k)$ 表示对真实语音信号做快速傅里叶变换（Fast Fourier Transform, FFT）变换后第n帧第k个频点的频谱幅度，其FFT变换长度为K，其中 $K \in \{2^6, 2^7, 2^8, 2^9, 2^{10}, 2^{11}\}$ ，其帧移为K/4， $\hat{S}(n, k)$ 表示对生成的合成语音信号做FFT变换后第n帧第k个频点的频谱幅度， α_{t-f} 为加权系数， $\alpha_{t-f} = \sqrt{K/2}$ 其用于平衡真实语谱和合成语音谱的1-范数失真和2-范数失真的比重， $\|\cdot\|_1$ 表示计算输入变量的1-范数， $\|\cdot\|_2$ 表示计算输入变量的2-范数。

在本申请的一些实施例中，上述相对特征匹配损失参数可以通过下述的公式（12）计算得到：

$$L_{feat}(x, \hat{x}) = \frac{1}{KL} \sum_{k=1}^K \sum_{l=1}^L \frac{\|D_k^l(x) - D_k^l(\hat{x})\|_1}{\text{mean}(\|D_k^l(x)\|_1)}$$

公式（12）

其中， $L_{feat}(x, \hat{x})$ 为相对特征匹配损失，K为判别器网络的个数，L为判别器的层数， $D_k^l(x)$ 表示输入真实语音信号时第k个判别器第l层的输出， $D_k^l(\hat{x})$ 表示输入生成的语音信号时第k个判别器第l层的输出。 $\text{mean}(\cdot)$ 表示取均值操作。

在本申请的一些实施例中，在训练过程中，第二音频信号恢复模型中的损失函数可以通过下述的公式（13）计算得到：

公式（13）

其中， L_G 为第二音频信号恢复模型的损失函数， λ_g 、 λ_{t-f} 、 λ_{feat} 分别为对抗损失、多分辨率时频域信号损失、相对特征匹配损失的加权系数，均为常数。在第二音频信号恢复模型的训练过程中，音频信号恢复装置可以通过最小化第二音频信号恢复模型损失函数 L_G 来更新第二音频信号恢复模型的模型参数，以提高生成的语音信号的质量。

如此，由于可以采用包含标记为正确接收的语音包和标记为丢失的语音包的训练语音样本，通过包含对抗损失、多分辨率时频域信号损失、相对特征匹配损失的多约束项损失函数，对第二音频信号恢复模型训练，得到第一音频信号恢复模型，因此可以使得训练后的第一音频信号恢复模型可以生成与真实语音较为接近的语音信号，从

而使得训练后的第一音频信号恢复模型具有较好的音频信号恢复能力。

以下以具体示例来对本申请实施例提供的音频信号恢复方法进行示例性说明。

在实际应用中，在完成对第二音频信号恢复模型的训练，得到第一音频信号恢复模型之后，可以将训练好的第一音频信号恢复模型集成到 AMR-WB 解码器中。当语音数据

包传输发生丢包时，AMR-WB 解码器可以通过本申请实施例提供的音频信号恢复方法高质量地恢复丢失的语音信号。

假设一段 16kHz 采样宽带语音信号通过 AMR-WB 编码器编码，AMR-WB 编码器的编码码率设定为 12.65 kbit/s。假定一个 IP 数据包只存储 1 帧编码语音信号，解码端已正确接收第 1、2、...、100 帧，第 101 帧丢失。此时，解码器可以对第 101 帧语音信号进行恢复。AMR-WB 编解码算法一帧长度为 20ms，一帧信号包含 320 个样本点。假设缓存的大小为 5 帧，即存储过去 5 帧的解码语音信号，包括：96 帧、97 帧、98 帧、99 帧、100 帧。

如图 15 所示，该音频信号恢复方法可以包括下述的步骤 201 至步骤 203。

步骤 201、音频信号恢复装置通过第一音频信号恢复模型中的预处理模块，将输入的 5 帧长度的语音信号映射为 1 帧固定长度的语音信号。

在本申请的一些实施例中，频信号恢复装置可以先从缓存中取出过去 5 帧的解码语音信号，组成一个输入张量 x ，其维度为 $\text{torch.size}([1, 1, 1600])$ ，其中第 1 个维度为 batch 大小，第 2 个维度为 channel 大小，第 3 个维度为信号长度（ $1600=320*5$ ）。

在本申请的一些实施例中，第一音频信号恢复模型中的预处理模块可以由一个全连接层网络和“Tanh()”激活函数组成，输入信号可以依次通过全连接层网络层和非线性激活函数层，然后得到输出信号，其定义为：

$\text{nn.Linear}(1600, 320)$

$\text{nn.Tanh}()$

将输入张量 x 通过上述预处理模块，可以将输入的 5 帧长度的信号映射为 1 帧固定长度的信号。令其输出张量为 x_1 ，其维度为 $\text{torch.size}([1, 1, 320])$ 。

步骤 202、音频信号恢复装置将 1 帧固定长度的语音信号输入第一音频信号恢复模型，通过第一音频信号恢复模型中的编码模块和第一音频信号恢复模型中的解码模块生成目标语音信号。

在本申请的一些实施例中，音频信号恢复装置可以将编码模块输出张量 x_1 （其维度为 $\text{torch.size}([1, 1, 320])$ ）通过编码模块进行处理，令其输出特征为 z ，其维度为 $\text{torch.size}([1, 128, 1])$ 。然后将编码模块输出的特征 z 输入到解码模块进行处理，令其输出张量为 x_2 ，其维度为 $\text{torch.size}([1, 1, 320])$ ，解码模块输出的 320 个样本点即为恢复的第 101 帧信号。

步骤 203、音频信号恢复装置更新缓存。

在本申请的一些实施例中，音频信号恢复装置可以将存储的第 96 帧的 320 个样本点从缓存中移除，然后将新生成的第 101 帧的 320 个样本点存入缓存，此时缓存 buffer 中存储的 5 帧信号为 97 帧、98 帧、99 帧、100 帧、101 帧。

在本申请的一些实施例中，音频信号恢复装置在更新完缓存之后，可以继续开始第 102 帧的解码，依次循环直至所有语音信号均解码完成。

在本申请实施例提供的音频信号恢复方法中，在传输语音信号的过程中，若丢失了某段语音信号，则可以通过音频信号恢复模型中的预处理模块将缓存的至少一帧语音信号映射为固定长度的语音信号，然后通过音频信号恢复模型中的编码模块和解码模块恢复出丢失的语音信号。如此，可以通过音频信号恢复模型恢复出丢失的语音信号，从而避免用户感知到语音数据包的丢失，降低语音数据包丢失所带来的负面影响。同时，还可以在完成对丢失语音信号的恢复之后，缓存恢复得到的语音信号，以使得恢复之后丢失的语音信号，从而可以提高语音通话的质量。

本申请实施例提供的音频信号恢复方法，执行主体可以为音频信号恢复装置。本申请实施例中以音频信号恢复装置执行音频信号恢复方法为例，说明本申请实施例提供的音频信号恢复装置。

图 16 示出了本申请实施例中涉及的音频信号恢复装置的一种可能的结构示意图。如图 16 所示，该音频信号恢复装置 160 可以包括：输入模块 161 和处理模块 162。

其中，输入模块 161，用于在第一语音信号丢失的情况下，将缓存的第二语音信

号输入第一音频信号恢复模型，第一音频信号恢复模型包括编码模块和解码模块，编码模块包括 M 层第一残差网络层，解码模块包括 M 层第二残差网络层；处理模块 162，用于将第二语音信号经 M 层第一残差网络层依次处理，得到第一时域特征信息；以及用于将第一时域特征信息经 M 层第二残差网络层依次处理，得到目标语音信号；其中，第二语音信号包括以下至少之一：第一语音信号前的 X 帧语音信号，第一语音信号之后的 Y 帧语音信号，M，X，Y 为正整数。

在一种可能的实现方式中，上述处理模块 162，具体用于：

通过编码模块中的预卷积层，提取第二语音信号的第二时域特征信息；

将第二时域特征信息经 M 层第一残差网络层依次处理，得到第一时域特征信息。

在一种可能的实现方式中，上述处理模块 162，具体用于：

将输入第 i 层第一残差网络层的时域特征信息，通过第 i 层第一残差网络层的第一卷积层处理，得到第三时域特征信息，并通过第 i 层第一残差网络层的第二卷积层以及第三卷积层处理，得到第四时域特征信息， $i \in \{1, \dots, M\}$ ；

基于第三时域特征信息和第四时域特征信息，得到第 i 层第一残差网络层输出的时域特征信息；

其中，第一卷积层、第二卷积层以及第三卷积层均为一维卷积层；

第一卷积层的输入通道数和输出通道数相同；

第二卷积层的卷积核与第一卷积层的卷积核不同、且第三卷积层的卷积核与第一卷积层的卷积核相同。

在一种可能的实现方式中，上述编码模块还包括 M 层下采样层，M 层下采样层与 M 个第一残差网络层交替排列；M 层下采样层的输出通道数按照层级从高至低依次增大；

上述处理模块 162，具体用于：

将输入第 x 层第一残差网络层的时域特征信息，通过第 x 层第一残差网络层处理，得到第五时域特征信息；

将所述第五时域特征信息通过与所述第 x 层第一残差网络层连接的第 x 层下采样层输出， $x \in \{1, \dots, M\}$ 。

在一种可能的实现方式中，上述处理模块 162，具体用于：

将第一时域特征信息经 M 层第二残差网络层依次处理，得到第六时域特征信息；

将第六时域特征信息输入解码模块的后卷积层处理，得到目标语音信号。

在一种可能的实现方式中，上述处理模块 162，具体用于：

将输入第 j 层第二残差网络层的时域特征信息，通过第 j 层第二残差网络层的第四卷积层处理，得到第七时域特征信息，并通过第 j 层第二残差网络层的第五卷积层以及第六卷积层进行处理，得到第八时域特征信息， $j \in \{1, \dots, M\}$ ；

基于第七时域特征信息和第八时域特征信息，得到第 j 层第二残差网络输出的第六时域特征信息；

其中，第四卷积层、第五卷积层以及第六卷积层均为一维卷积层；

第四卷积层的输入通道数和输入通道数相同；

第五卷积层的卷积核与第四卷积层的卷积核不同、且第六卷积层的卷积核与第四卷积层的卷积核相同。

在一种可能的实现方式中，上述解码模块还包括 M 层上采样层，M 层上采样层与 M 个第二残差网络层交替排列；M 层上采样层的输出通道数按照层级从高至低依次降低；

上述处理模块 162，具体用于：

将第 y 层上采样层输出的时域特征信息，通过与第 y 层上采样层连接的第 y 层第二残差网络层处理，得到第 y 层第二残差网络层输出的时域特征信息， $y \in \{1, \dots, M-1\}$ 。

在一种可能的实现方式中，上述处理模块 162，还用于在第二语音信号为多帧语音信号的情况下，再将第二语音信号经 M 层第一残差网络层依次处理得到第一时域特征信息之前，将多帧语音信号输入第一音频信号恢复模型之后，将多帧语音信号映射为一帧语音信号；

上述处理模块 162，具体用于将一帧语音信号经 M 层第一残差网络层依次处理，

得到第一时域特征信息。

在一种可能的实现方式中，上述装置还包括：获取模块；

获取模块，用于在输入模块 161 在第一语音信号丢失的情况下，将缓存的第二语音信号输入第一音频信号恢复模型之前，获取至少一组训练语音样本，每组训练语音样本包括标记为正确接收的语音包和标记为丢失的语音包；

上述处理模块 162，还用于基于获取模块获取的至少一组训练语音样本，对第二音频信号恢复模型训练，得到第一音频信号恢复模型；

其中，第二音频信号恢复模型中的损失函数包括以下至少两项：对抗损失参数、多分辨率时频域信号损失参数以及相对特征匹配损失参数。

本申请实施例提供一种音频信号恢复装置，在语音通话时，若丢失了某个语音信号，则可以通过音频信号恢复模型中的编码模块的残差网络层提取缓存的丢失语音信号前或丢失语音信号后的语音信号的时域特征信息，并通过音频信号恢复模型中的解码模块的残差网络层合成编码模块提取的时域特征信息，恢复出丢失的语音信号。如此，由于可以在语音信号丢失的情况下，通过将缓存的语音信号输入音频信号恢复模型恢复出丢失的语音信号，因此在语音信号的传输过程中，可以通过音频信号恢复模型恢复出丢失的语音信号，从而避免用户感知到语音数据包的丢失，降低语音数据包丢失所带来的负面影响，进而可以有效地提高语音通话的质量。

本申请实施例中的音频信号恢复装置可以是电子设备，也可以是电子设备中的部件，例如集成电路或芯片。该电子设备可以是终端，也可以为除终端之外的其他设备。示例性的，电子设备可以为手机、平板电脑、笔记本电脑、掌上电脑、车载电子设备、移动上网装置（Mobile Internet Device, MID）、增强现实(augmented reality, AR)/虚拟现实(virtual reality, VR)设备、机器人、可穿戴设备、超级移动个人计算机（ultra-mobile personal computer, UMPC）、上网本或者个人数字助理（personal digital assistant, PDA）等，还可以为服务器、网络附属存储器（Network Attached Storage, NAS）、个人计算机（personal computer, PC）、电视机（television, TV）、柜员机或者自助机等，本申请实施例不作具体限定。

本申请实施例中的音频信号恢复装置可以为具有操作系统的装置。该操作系统可以为安卓（Android）操作系统，可以为 ios 操作系统，还可以为其他可能的操作系统，本申请实施例不作具体限定。

本申请实施例提供的音频信号恢复装置能够实现如上述音频信号恢复方法实施例实现的各个过程，达到相同的技术效果，为避免重复，这里不再赘述。

可选地，如图 17 所示，本申请实施例还提供一种电子设备 1700，包括处理器 1701 和存储器 1702，存储器 1702 上存储有可在所述处理器 1701 上运行的程序或指令，该程序或指令被处理器 1701 执行时实现上述音频信号恢复方法实施例的各个步骤，且能达到相同的技术效果，为避免重复，这里不再赘述。

需要说明的是，本申请实施例中的电子设备包括上述所述的移动电子设备和非移动电子设备。

图 18 为实现本申请实施例的一种电子设备的硬件结构示意图。

该电子设备 1800 包括但不限于：射频单元 1801、网络模块 1802、音频输出单元 1803、输入单元 1804、传感器 1805、显示单元 1806、用户输入单元 1807、接口单元 1808、存储器 1809、以及处理器 1810 等部件。

本领域技术人员可以理解，电子设备 1800 还可以包括给各个部件供电的电源（比如电池），电源可以通过电源管理系统与处理器 1810 逻辑相连，从而通过电源管理系统实现管理充电、放电、以及功耗管理等功能。图 18 中示出的电子设备结构并不构成对电子设备的限定，电子设备可以包括比图示更多或更少的部件，或者组合某些部件，或者不同的部件布置，在此不再赘述。

其中，处理器 1810，用于在第一语音信号丢失的情况下，将缓存的第二语音信号输入第一音频信号恢复模型，第一音频信号恢复模型包括编码模块和解码模块，编码模块包括 M 层第一残差网络层，解码模块包括 M 层第二残差网络层；以及，用于将第二语音信号经 M 层第一残差网络层依次处理，得到第一时域特征信息；并且用于将第一时域特征信息经 M 层第二残差网络层依次处理，得到目标语音信号；其中，第二语音信号包括以下至少之一：第一语音信号前的 X 帧语音信号，第一语音信号之后的 Y 帧语音信号，M，X，Y 为正整数。

在一种可能的实现方式中，上述处理器 1810，具体用于：

通过编码模块中的预卷积层，提取第二语音信号的第二时域特征信息；
将第二时域特征信息经 M 层第一残差网络层依次处理，得到第一时域特征信息。

在一种可能的实现方式中，上述处理器 1810，具体用于：

5 将输入第 i 层第一残差网络层的时域特征信息，通过第 i 层第一残差网络层的第一卷积层处理，得到第三时域特征信息，并通过第 i 层第一残差网络层的第二卷积层以及第三卷积层处理，得到第四时域特征信息， $i \in \{1, \dots, M\}$ ；

基于第三时域特征信息和第四时域特征信息，得到第 i 层第一残差网络层输出的时域特征信息；

10 其中，第一卷积层、第二卷积层以及第三卷积层均为一维卷积层；

第一卷积层的输入通道数和输出通道数相同；

第二卷积层的卷积核与第一卷积层的卷积核不同、且第三卷积层的卷积核与第一卷积层的卷积核相同。

15 在一种可能的实现方式中，上述编码模块还包括 M 层下采样层， M 层下采样层与 M 个第一残差网络层交替排列； M 层下采样层的输出通道数按照层级从高至低依次增大；

上述处理器 1810，具体用于：

将输入第 x 层第一残差网络层的时域特征信息，通过第 x 层第一残差网络层处理，得到第五时域特征信息；

20 将第五时域特征信息通过与第 x 层第一残差网络层连接的第 x 层下采样层输出， $x \in \{1, \dots, M\}$ 。

在一种可能的实现方式中，上述处理器 1810，具体用于：

将第一时域特征信息经 M 层第二残差网络层依次处理，得到第六时域特征信息；

25 将第六时域特征信息输入解码模块的后卷积层处理，得到目标语音信号。

在一种可能的实现方式中，上述处理器 1810，具体用于：

将输入第 j 层第二残差网络层的时域特征信息，通过第 j 层第二残差网络层的第四卷积层处理，得到第七时域特征信息，并通过第 j 层第二残差网络层的第五卷积层以及第六卷积层进行处理，得到第八时域特征信息， $j \in \{1, \dots, M\}$ ；

30 基于第七时域特征信息和第八时域特征信息，得到第 j 层第二残差网络输出的第六时域特征信息；

其中，第四卷积层、第五卷积层以及第六卷积层均为一维卷积层；

第四卷积层的输入通道数和输出通道数相同；

35 第五卷积层的卷积核与第四卷积层的卷积核不同、且第六卷积层的卷积核与第四卷积层的卷积核相同。

在一种可能的实现方式中，上述解码模块还包括 M 层上采样层， M 层上采样层与 M 个第二残差网络层交替排列； M 层上采样层的输出通道数按照层级从高至低依次降低；

上述处理器 1810，具体用于：

40 将第 y 层上采样层输出的时域特征信息，通过与第 y 层上采样层连接的第 y 层第二残差网络层处理，得到第 y 层第二残差网络层输出的时域特征信息， $y \in \{1, \dots, M-1\}$ 。

45 在一种可能的实现方式中，上述处理器 1810，还用于在第二语音信号为多帧语音信号的情况下，在将第二语音信号经 M 层第一残差网络层依次处理得到第一时域特征信息之前，将多帧语音信号输入第一音频信号恢复模型之后，将多帧语音信号映射为一帧语音信号；

上述处理器 1810，具体用于将一帧语音信号经 M 层第一残差网络层依次处理，得到第一时域特征信息。

50 在一种可能的实现方式中，上述处理器 1810，还用于在第一语音信号丢失的情况下，将缓存的第二语音信号输入第一音频信号恢复模型之前，获取至少一组训练语音样本，每组训练语音样本包括标记为正确接收的语音包和标记为丢失的语音包；以及用于基于至少一组训练语音样本，对第二音频信号恢复模型训练，得到第一音频信号恢复模型；

其中，第二音频信号恢复模型中的损失函数包括以下至少两项：对抗损失参数、

多分辨率时频域信号损失参数以及相对特征匹配损失参数。

本申请实施例提供一种电子设备，在语音通话时，若丢失了某段语音信号，则可以通过音频信号恢复模型中的编码模块的残差网络层提取缓存的丢失语音信号前或丢失语音信号后的语音信号的时域特征信息，并通过音频信号恢复模型中的解码模块的残差网络层合成编码模块提取的时域特征信息，恢复出丢失的语音信号。如此，由于可以在语音信号丢失的情况下，通过将缓存的语音信号输入音频信号恢复模型恢复出丢失的语音信号，因此在语音信号的传输过程中，可以通过音频信号恢复模型恢复出丢失的语音信号，从而避免用户感知到语音数据包的丢失，降低语音数据包丢失所带来的负面影响，进而可以有效地提高语音通话的质量。

应理解的是，本申请实施例中，输入单元 1804 可以包括图形处理器（Graphics Processing Unit, GPU）18041 和麦克风 18042，图形处理器 18041 对在视频捕获模式或图像捕获模式中由图像捕获装置（如摄像头）获得的静态图片或视频的图像数据进行处理。显示单元 1806 可包括显示面板 18061，可以采用液晶显示器、有机发光二极管等形式来配置显示面板 18061。用户输入单元 1807 包括触控面板 18071 以及其他输入设备 18072 中的至少一种。触控面板 18071，也称为触摸屏。触控面板 18071 可包括触摸检测装置和触摸控制器两个部分。其他输入设备 18072 可以包括但不限于物理键盘、功能键（比如音量控制按键、开关按键等）、轨迹球、鼠标、操作杆，在此不再赘述。

存储器 1809 可用于存储软件程序以及各种数据。存储器 1809 可主要包括存储程序或指令的第一存储区和存储数据的第二存储区，其中，第一存储区可存储操作系统、至少一个功能所需的应用程序或指令（比如声音播放功能、图像播放功能等）等。此外，存储器 1809 可以包括易失性存储器或非易失性存储器，或者，存储器 1809 可以包括易失性和非易失性存储器两者。其中，非易失性存储器可以是只读存储器(Read-Only Memory, ROM)、可编程只读存储器(Programmable ROM, PROM)、可擦除可编程只读存储器(Erasable PROM, EPROM)、电可擦除可编程只读存储器(Electrically EPROM, EEPROM)或闪存。易失性存储器可以是随机存取存储器(Random Access Memory, RAM)、静态随机存取存储器(Static RAM, SRAM)、动态随机存取存储器(Dynamic RAM, DRAM)、同步动态随机存取存储器(Synchronous DRAM, SDRAM)、双倍数据速率同步动态随机存取存储器(Double Data Rate SDRAM, DDRSDRAM)、增强型同步动态随机存取存储器(Enhanced SDRAM, ESDRAM)、同步连接动态随机存取存储器(Synch link DRAM, SLDRAM)和直接内存总线随机存取存储器(Direct Rambus RAM, DRRAM)。本申请实施例中的存储器 1809 包括但不限于这些和任意其它适合类型的存储器。

处理器 1810 可包括一个或多个处理单元；可选的，处理器 1810 集成应用处理器和调制解调处理器，其中，应用处理器主要处理涉及操作系统、用户界面和应用程序等的操作，调制解调处理器主要处理无线通信信号，如基带处理器。可以理解的是，上述调制解调处理器也可以不集成到处理器 1810 中。

本申请实施例还提供一种可读存储介质，所述可读存储介质上存储有程序或指令，该程序或指令被处理器执行时实现上述音频信号恢复方法实施例的各个过程，且能达到相同的技术效果，为避免重复，这里不再赘述。

其中，所述处理器为上述实施例中所述的电子设备中的处理器。所述可读存储介质，包括计算机可读存储介质，如计算机只读存储器 ROM、随机存取存储器 RAM、磁碟或者光盘等。

本申请实施例另提供了一种芯片，所述芯片包括处理器和通信接口，所述通信接口和所述处理器耦合，所述处理器用于运行程序或指令，实现上述音频信号恢复方法实施例的各个过程，且能达到相同的技术效果，为避免重复，这里不再赘述。

应理解，本申请实施例提到的芯片还可以称为系统级芯片、系统芯片、芯片系统或片上系统芯片等。

本申请实施例提供一种计算机程序产品，该程序产品被存储在存储介质中，该程序产品被至少一个处理器执行以实现如上述音频信号恢复方法实施例的各个过程，且能达到相同的技术效果，为避免重复，这里不再赘述。

需要说明的是，在本文中，术语“包括”、“包含”或者其任何其他变体意在涵盖非排他性的包含，从而使得包括一系列要素的过程、方法、物品或者装置不仅包括那些要素，而且还包括没有明确列出的其他要素，或者是还包括为这种过程、方法、

物品或者装置所固有的要素。在没有更多限制的情况下，由语句“包括一个……”限定的要素，并不排除在包括该要素的过程、方法、物品或者装置中还存在另外的相同要素。此外，需要指出的是，本申请实施方式中的方法和装置的范围不限按示出或讨论的顺序来执行功能，还可包括根据所涉及的功能按基本同时的方式或按相反的顺序来执行功能，例如，可以按不同于所描述的次序来执行所描述的方法，并且还可以添加、省去、或组合各种步骤。另外，参照某些示例所描述的特征可在其他示例中被组合。

通过以上的实施方式的描述，本领域的技术人员可以清楚地了解到上述实施例方法可借助软件加必需的通用硬件平台的方式来实现，当然也可以通过硬件，但很多情况下前者是更佳的实施方式。基于这样的理解，本申请的技术方案本质上或者说对现有技术做出贡献的部分可以以计算机软件产品的形式体现出来，该计算机软件产品存储在一个存储介质（如 ROM/RAM、磁碟、光盘）中，包括若干指令用以使得一台终端（可以是手机，计算机，服务器，或者网络设备等）执行本申请各个实施例所述的方法。

上面结合附图对本申请的实施例进行了描述，但是本申请并不局限于上述的具体实施方式，上述的具体实施方式仅仅是示意性的，而不是限制性的，本领域的普通技术人员在本申请的启示下，在不脱离本申请宗旨和权利要求所保护的范围情况下，还可做出很多形式，均属于本申请的保护之内。

权 利 要 求 书

1.一种音频信号恢复方法,所述方法包括:

在第一语音信号丢失的情况下,将缓存的第二语音信号输入第一音频信号恢复模型,所述第一音频信号恢复模型包括编码模块和解码模块,所述编码模块包括 M 层

第一残差网络层,所述解码模块包括 M 层第二残差网络层;

将所述第二语音信号经所述 M 层第一残差网络层依次处理,得到第一时域特征信息;

将所述第一时域特征信息经所述 M 层第二残差网络层依次处理,得到目标语音信号;

其中,所述第二语音信号包括以下至少之一:所述第一语音信号前的 X 帧语音信号,所述第一语音信号之后的 Y 帧语音信号, M, X, Y 为正整数。

2.根据权利要求 1 所述的方法,其中,所述将所述第二语音信号经所述 M 层第一残差网络层依次处理,得到第一时域特征信息,包括:

通过所述编码模块中的预卷积层,提取所述第二语音信号的第二时域特征信息;

将所述第二时域特征信息经所述 M 层第一残差网络层依次处理,得到所述第一时域特征信息。

3.根据权利要求 2 所述的方法,其中,所述将所述第二时域特征信息经所述 M 层第一残差网络层依次处理,得到所述第一时域特征信息,包括:

将输入第 i 层第一残差网络层的时域特征信息,通过所述第 i 层第一残差网络层的第一卷积层处理,得到第三时域特征信息,并通过所述第 i 层第一残差网络层的第二卷积层以及第三卷积层处理,得到第四时域特征信息, $i \in \{1, \dots, M\}$;

基于所述第三时域特征信息和所述第四时域特征信息,得到所述第 i 层第一残差网络层输出的时域特征信息;

其中,所述第一卷积层、所述第二卷积层以及所述第三卷积层均为一维卷积层;

所述第一卷积层的输入通道数和输出通道数相同;

所述第二卷积层的卷积核与所述第一卷积层的卷积核不同、且所述第三卷积层的卷积核与所述第一卷积层的卷积核相同。

4.根据权利要求 2 或 3 所述的方法,其中,所述编码模块还包括 M 层下采样层,所述 M 层下采样层与所述 M 个第一残差网络层交替排列;

所述 M 层下采样层的输出通道数按照层级从高至低依次增大;

所述将所述第二时域特征信息经所述 M 层第一残差网络层依次处理,得到所述第一时域特征信息,包括:

将输入第 x 层第一残差网络层的时域特征信息,通过所述第 x 层第一残差网络层处理,得到第五时域特征信息;

将所述第五时域特征信息通过与所述第 x 层第一残差网络层连接的第 x 层下采样层输出, $x \in \{1, \dots, M\}$ 。

5.根据权利要求 1 所述的方法,其中,所述将所述第一时域特征信息经所述 M 层第二残差网络层依次处理,得到目标语音信号,包括:

将所述第一时域特征信息经所述 M 层第二残差网络层依次处理,得到第六时域特征信息;

将所述第六时域特征信息输入所述解码模块的后卷积层处理,得到所述目标语音信号。

6.根据权利要求 5 所述的方法,其中,所述将所述第一时域特征信息经 M 层第二残差网络层依次处理,得到第六时域特征信息,包括:

将输入第 j 层第二残差网络层的时域特征信息,通过所述第 j 层第二残差网络层的第四卷积层处理,得到第七时域特征信息,并通过所述第 j 层第二残差网络层的第五卷积层以及第六卷积层进行处理,得到第八时域特征信息, $j \in \{1, \dots, M\}$;

基于所述第七时域特征信息和所述第八时域特征信息,得到所述第 j 层第二残差网络层输出的时域特征信息;

其中,所述第四卷积层、所述第五卷积层以及所述第六卷积层均为一维卷积层;

所述第四卷积层的输入通道数和输出通道数相同;

所述第五卷积层的卷积核与所述第四卷积层的卷积核不同、且所述第六卷积层的卷积核与所述第四卷积层的卷积核相同。

7.根据权利要求 5 或 6 所述的方法,其中,所述解码模块还包括 M 层上采样层,所述 M 层上采样层与所述 M 个第二残差网络层交替排列;

所述 M 层上采样层的输出通道数按照层级从高至低依次降低;

5 所述将所述第一时域特征信息经 M 层第二残差网络层依次处理,得到第六时域特征信息,包括:

将第 y 层上采样层输出的时域特征信息,通过与所述第 y 层上采样层连接的所述第 y 层第二残差网络层处理,得到所述第 y 层第二残差网络层输出的时域特征信息, $y \in \{1, \dots, M\}$ 。

10 8.根据权利要求 1 所述的方法,其中,在所述第二语音信号为多帧语音信号的情况下,所述将所述第二语音信号经所述 M 层第一残差网络层依次处理,得到第一时域特征信息之前,所述方法还包括:

将所述多帧语音信号输入所述第一音频信号恢复模型之后,将所述多帧第二语音信号映射为一帧语音信号;

15 所述将所述第二语音信号经所述 M 层第一残差网络层依次处理,得到第一时域特征信息,包括:

将所述一帧语音信号经所述 M 层第一残差网络层依次处理,得到所述第一时域特征信息。

9.根据权利要求 1 所述的方法,其中,所述在第一语音信号丢失的情况下,将缓存的第二语音信号输入第一音频信号恢复模型之前,所述方法还包括:

20 获取至少一组训练语音样本,每组训练语音样本包括标记为正确接收的语音包和标记为丢失的语音包;

基于所述至少一组训练语音样本,对第二音频信号恢复模型训练,得到所述第一音频信号恢复模型;

25 其中,所述第二音频信号恢复模型中的损失函数包括以下至少两项:对抗损失参数、多分辨率时频域信号损失参数以及相对特征匹配损失参数。

10.一种音频信号恢复的装置,所述装置包括:输入模块和处理模块;

所述输入模块,用于在第一语音信号丢失的情况下,将缓存的第二语音信号输入第一音频信号恢复模型,所述第一音频信号恢复模型包括编码模块和解码模块,所述编码模块包括 M 层第一残差网络层,所述解码模块包括 M 层第二残差网络层;

30 所述处理模块,用于将所述第二语音信号经所述 M 层第一残差网络层依次处理,得到第一时域特征信息;

所述处理模块,还用于将所述第一时域特征信息经所述 M 层第二残差网络层依次处理,得到目标语音信号;

35 其中,所述第二语音信号包括以下至少之一:所述第一语音信号前的 X 帧语音信号,所述第一语音信号之后的 Y 帧语音信号, M, X, Y 为正整数。

11.根据权利要求 10 所述的装置,其中,所述处理模块,具体用于:

通过所述编码模块中的预卷积层,提取所述第二语音信号的第二时域特征信息;

40 将所述第二时域特征信息经所述 M 层第一残差网络层依次处理,得到所述第一时域特征信息。

12.根据权利要求 11 所述的装置,其中,所述处理模块,具体用于:

将输入第 i 层第一残差网络层的时域特征信息,通过所述第 i 层第一残差网络层的第一卷积层处理,得到第三时域特征信息,并通过所述第 i 层第一残差网络层的第二卷积层以及第三卷积层处理,得到第四时域特征信息, $i \in \{1, \dots, M\}$;

45 基于所述第三时域特征信息和所述第四时域特征信息,得到所述第 i 层第一残差网络层输出的时域特征信息;

其中,所述第一卷积层、所述第二卷积层以及所述第三卷积层均为一维卷积层;

所述第一卷积层的输入通道数和输出通道数相同;

50 所述第二卷积层的卷积核与所述第一卷积层的卷积核不同、且所述第三卷积层的卷积核与所述第一卷积层的卷积核相同。

13.根据权利要求 11 或 12 所述的装置,其中,所述编码模块还包括 M 层下采样层,所述 M 层下采样层与所述 M 个第一残差网络层交替排列;

所述 M 层下采样层的输出通道数按照层级从高至低依次增大;

所述处理模块,具体用于:

将输入第 x 层第一残差网络层的时域特征信息,通过所述第 x 层第一残差网络层

处理, 得到第五时域特征信息;

将所述第五时域特征信息通过与所述第 x 层第一残差网络层连接的第 x 层下采样层输出, $x \in \{1, \dots, M\}$ 。

14. 根据权利要求 10 所述的装置, 其中, 所述处理模块, 具体用于:

5 将所述第一时域特征信息经 M 层第二残差网络层依次处理, 得到第六时域特征信息;

将所述第六时域特征信息输入所述解码模块的后卷积层处理, 得到所述目标语音信号。

15. 根据权利要求 14 所述的装置, 其中, 所述处理模块, 具体用于:

10 将输入第 j 层第二残差网络层的时域特征信息, 通过所述第 j 层第二残差网络层的第四卷积层处理, 得到第七时域特征信息, 并通过所述第 j 层第二残差网络层的第五卷积层以及第六卷积层进行处理, 得到第八时域特征信息, $j \in \{1, \dots, M\}$;

基于所述第七时域特征信息和所述第八时域特征信息, 得到所述第 j 层第二残差网络输出的时域特征信息;

15 其中, 所述第四卷积层、所述第五卷积层以及所述第六卷积层均为一维卷积层;

所述第四卷积层的输入通道数和输入通道数相同;

所述第五卷积层的卷积核与所述第四卷积层的卷积核不同、且所述第六卷积层的卷积核与所述第四卷积层的卷积核相同。

20 16. 根据权利要求 14 或 15 所述的装置, 其中, 所述解码模块还包括 M 层上采样层, 所述 M 层上采样层与所述 M 个第二残差网络层交替排列;

所述 M 层上采样层的输出通道数按照层级从高至低依次降低;

所述处理模块, 具体用于:

25 将第 y 层上采样层输出的时域特征信息, 通过与所述第 y 层上采样层连接的所述第 y 层第二残差网络层处理, 得到所述第 y 层第二残差网络层输出的时域特征信息, $y \in \{1, \dots, M-1\}$ 。

17. 根据权利要求 10 所述的装置, 其中, 所述处理模块, 还用于在所述第二语音信号为多帧语音信号的情况下, 在将所述第二语音信号经所述 M 层第一残差网络层依次处理得到第一时域特征信息之前, 将所述多帧语音信号输入所述第一音频信号恢复模型之后, 将所述多帧语音信号映射为一帧语音信号;

30 所述处理模块, 具体用于将所述一帧语音信号经所述 M 层第一残差网络层依次处理, 得到所述第一时域特征信息。

18. 根据权利要求 10 所述的装置, 其中, 所述装置还包括: 获取模块;

35 所述获取模块, 用于在所述输入模块在第一语音信号丢失的情况下, 将缓存的第二语音信号输入第一音频信号恢复模型之前, 获取至少一组训练语音样本, 每组训练语音样本包括标记为正确接收的语音包和标记为丢失的语音包;

所述处理模块, 还用于基于所述获取模块获取的所述至少一组训练语音样本, 对第二音频信号恢复模型训练, 得到所述第一音频信号恢复模型;

其中, 所述第二音频信号恢复模型中的损失函数包括以下至少两项: 对抗损失参数、多分辨率时频域信号损失参数以及相对特征匹配损失参数。

40 19. 一种电子设备, 包括处理器和存储器, 所述存储器存储可在所述处理器上运行的程序或指令, 所述程序或指令被所述处理器执行时实现如权利要求 1 至 9 任一项所述的音频信号恢复方法的步骤。

20. 一种可读存储介质, 所述可读存储介质上存储程序或指令, 所述程序或指令被处理器执行时实现如权利要求 1 至 9 任一项所述的音频信号恢复方法的步骤。

45 21. 一种芯片, 所述芯片包括处理器和通信接口, 所述通信接口和所述处理器耦合, 所述处理器用于运行程序或指令, 实现如权利要求 1 至 9 任一项所述的音频信号恢复方法的步骤。

22. 一种计算机程序产品, 所述计算机程序产品被存储在非易失的存储介质中, 所述计算机程序产品被至少一个处理器执行以实现如权利要求 1 至 9 任一项所述的音频信号恢复方法的步骤。

50 23. 一种电子设备, 所述电子设备被配置成用于执行如权利要求 1 至 9 任一项所述的音频信号恢复方法的步骤。

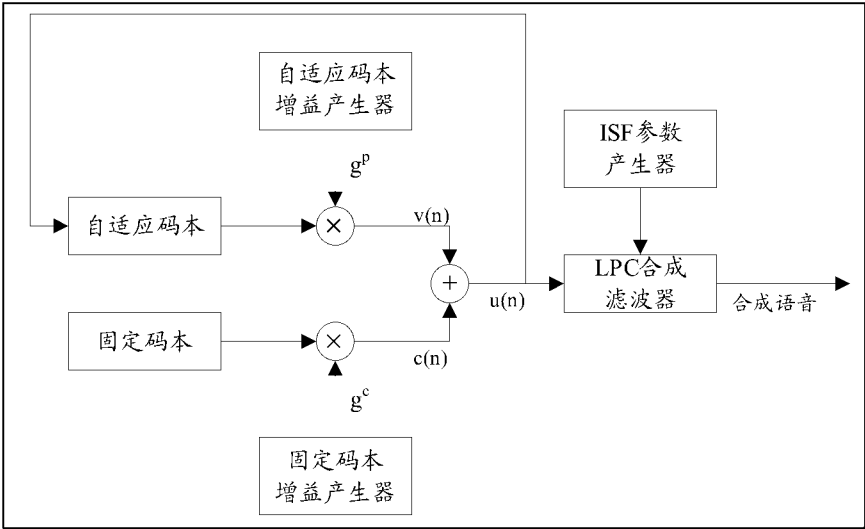


图 1

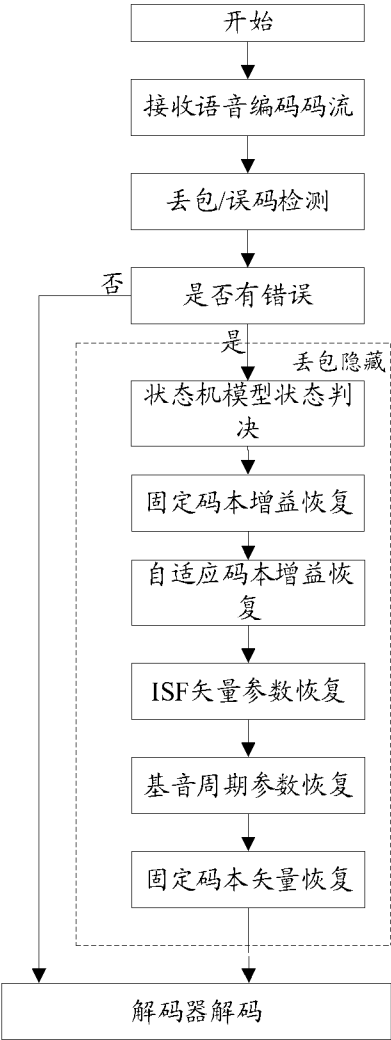


图 2

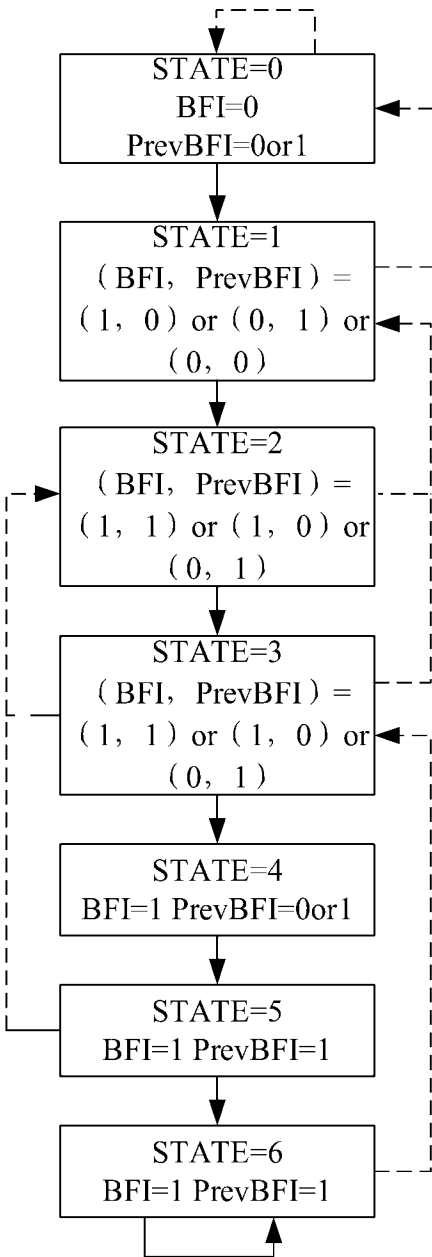


图 3

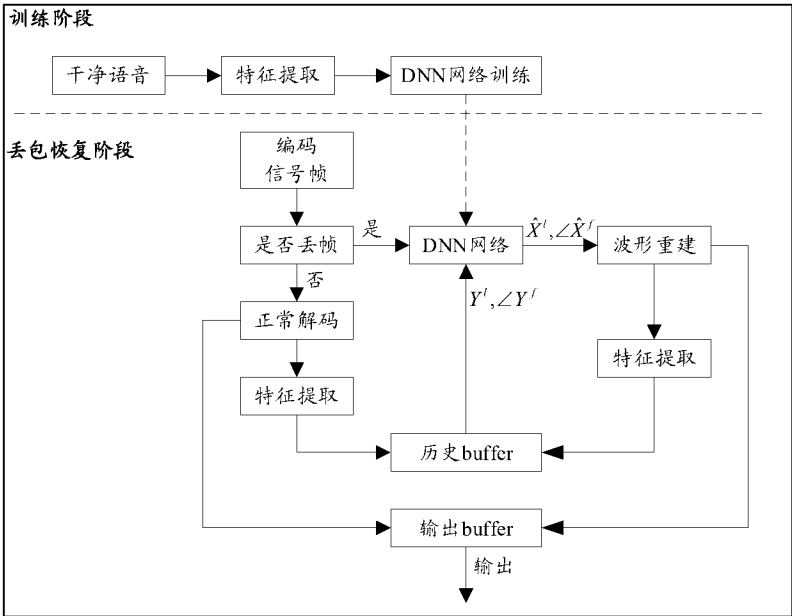


图 4

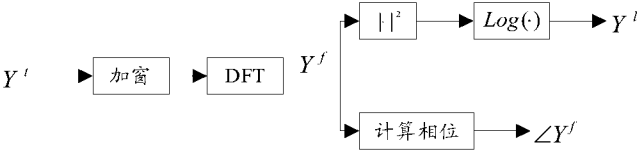


图 5

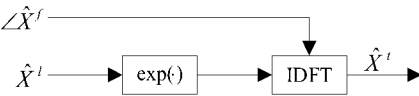


图 6

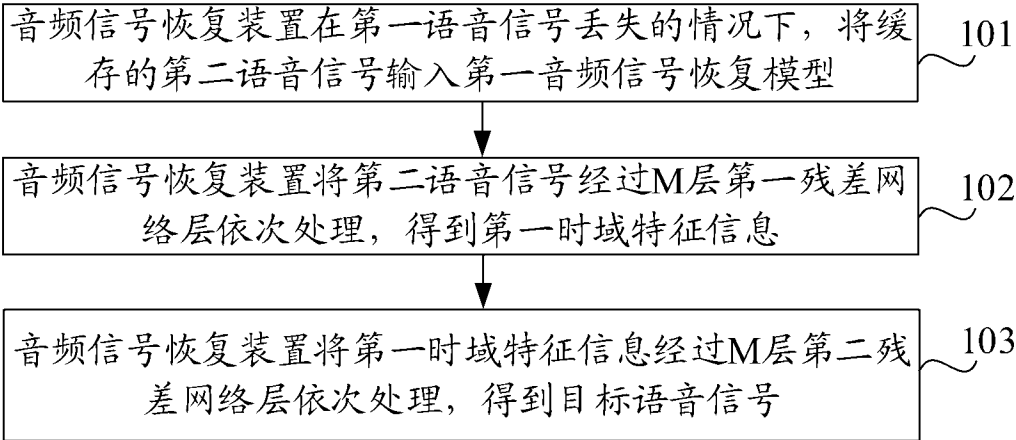


图 7

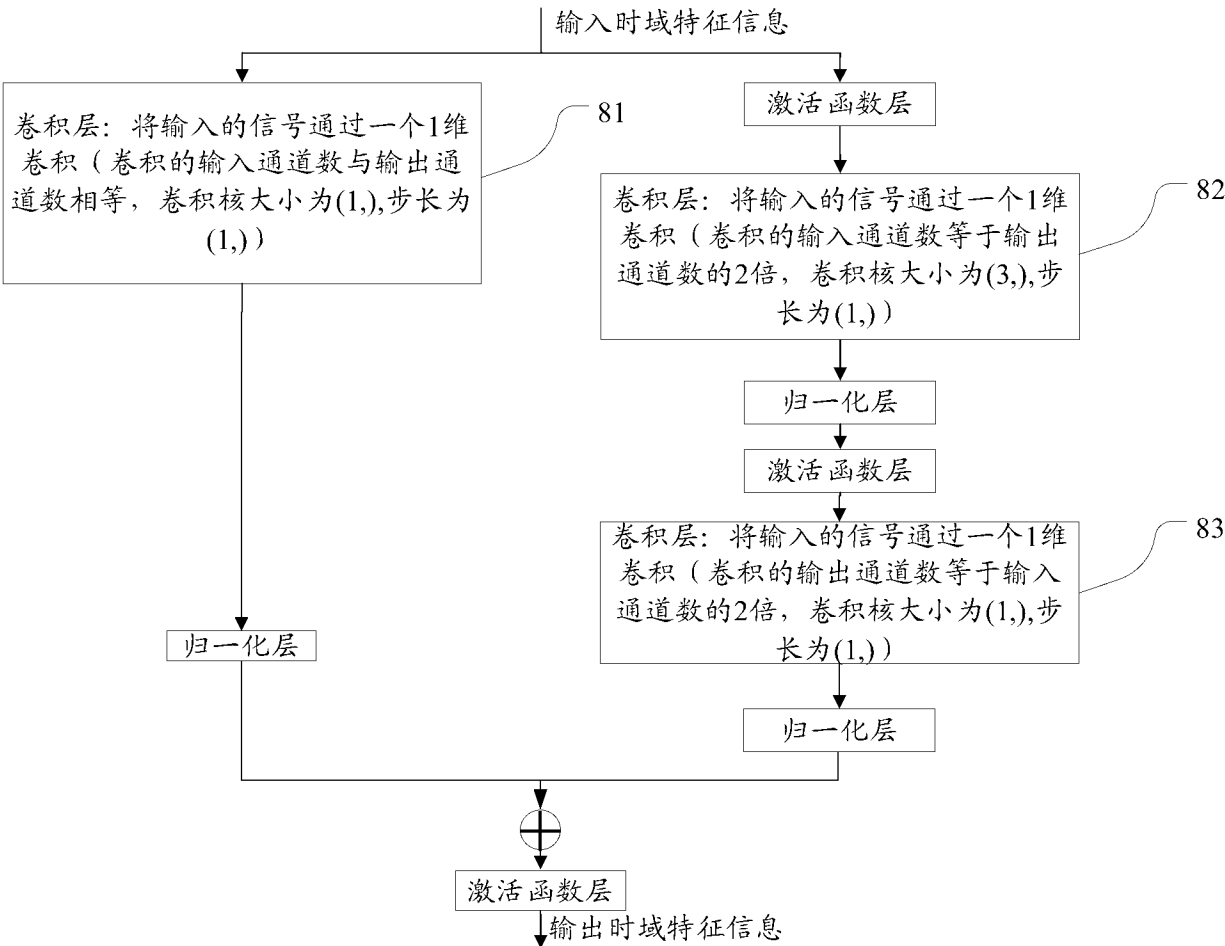


图 8

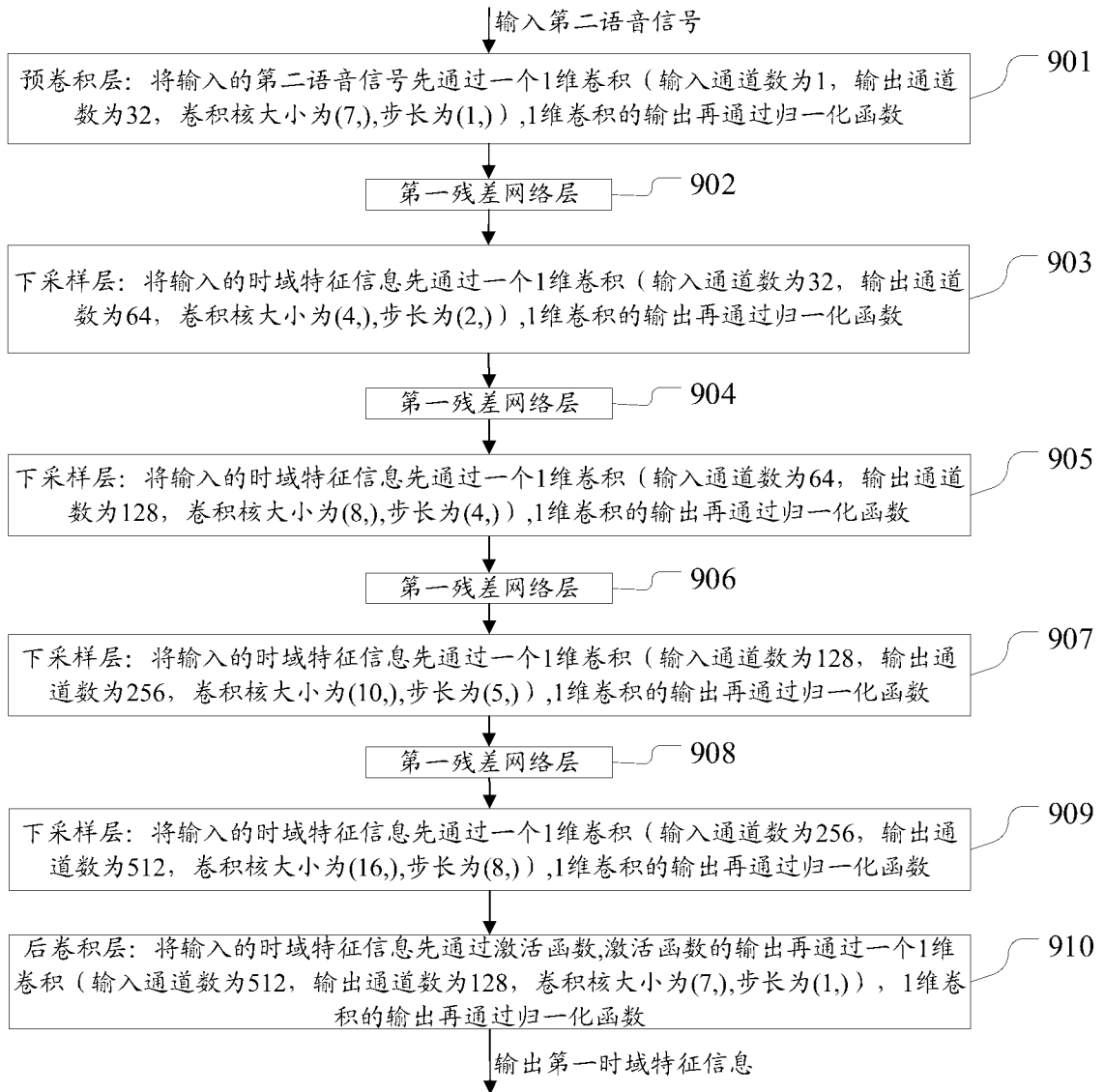


图 9

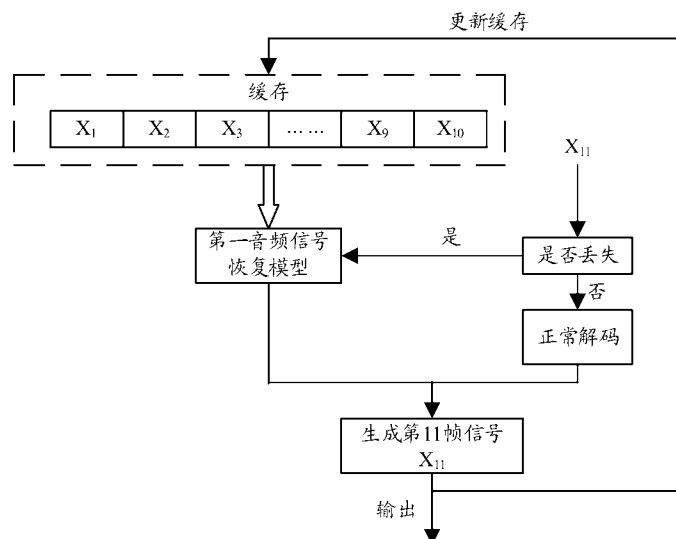


图 10

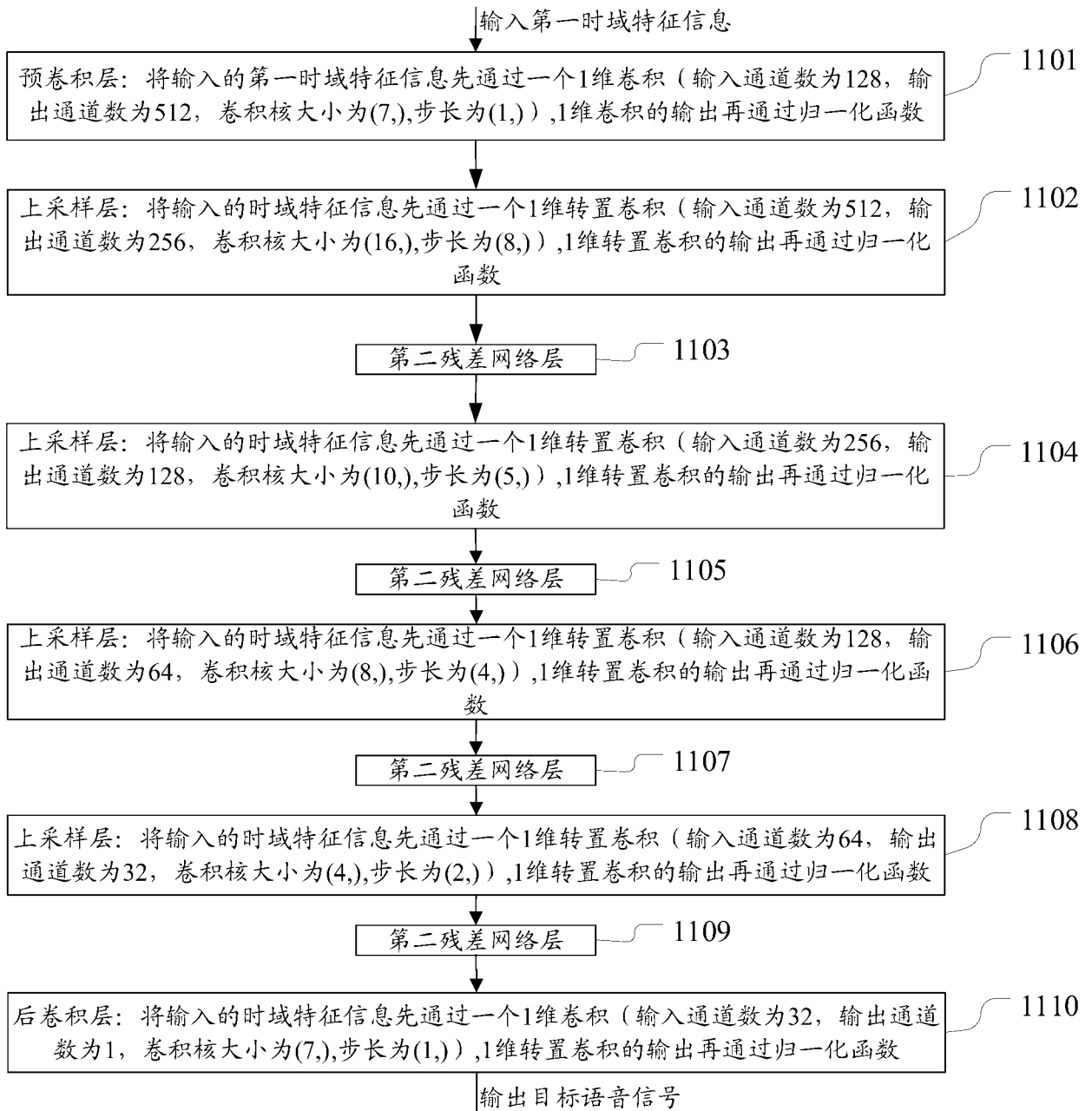


图 11

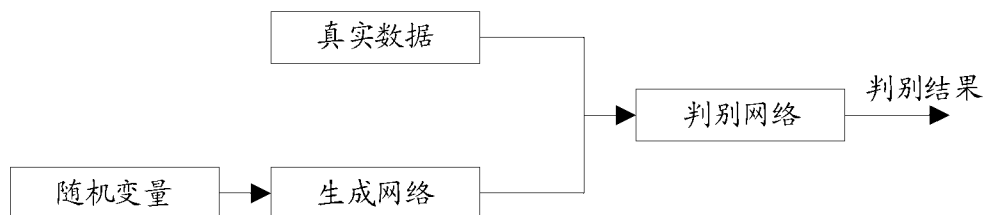


图 12

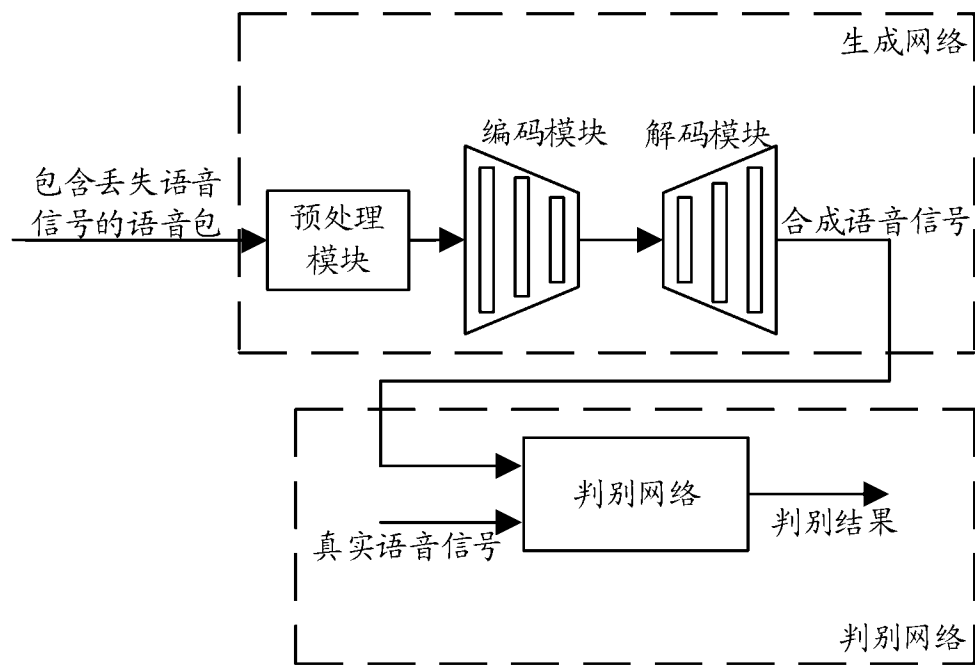


图 13

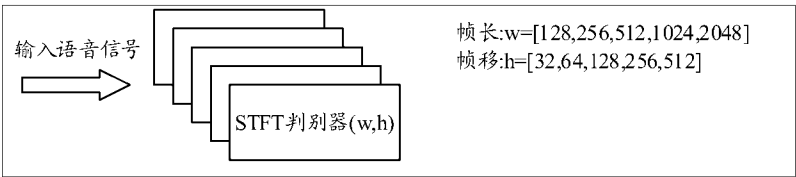


图 14

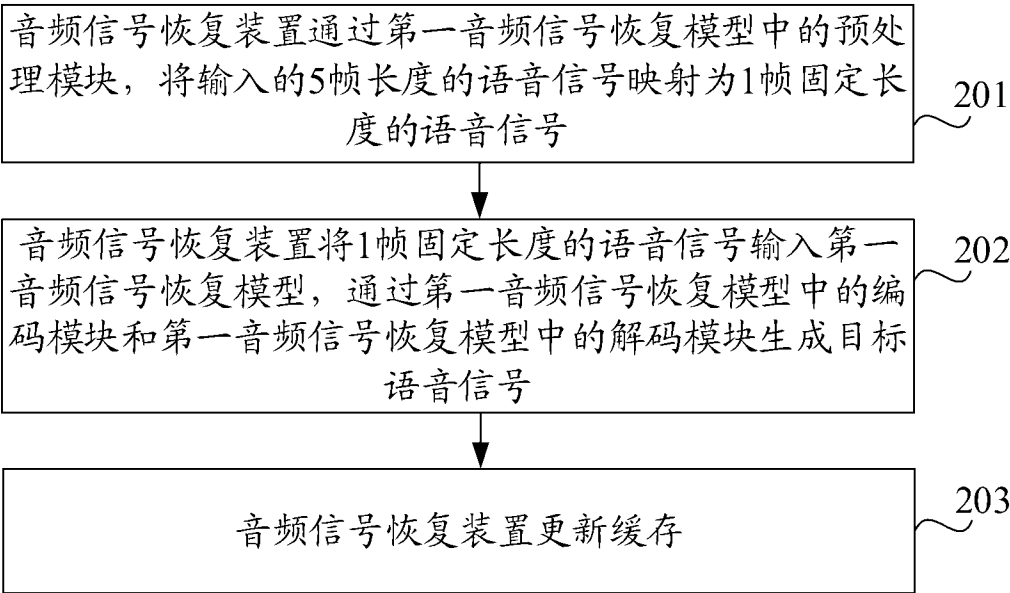


图 15

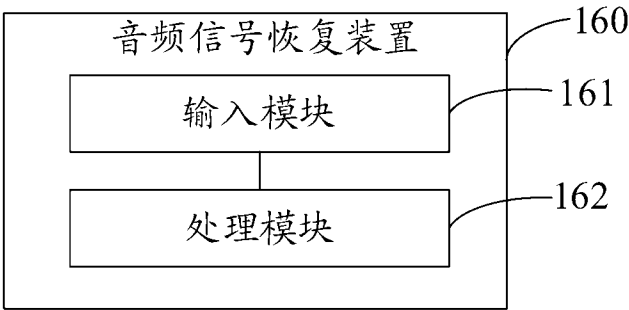


图 16

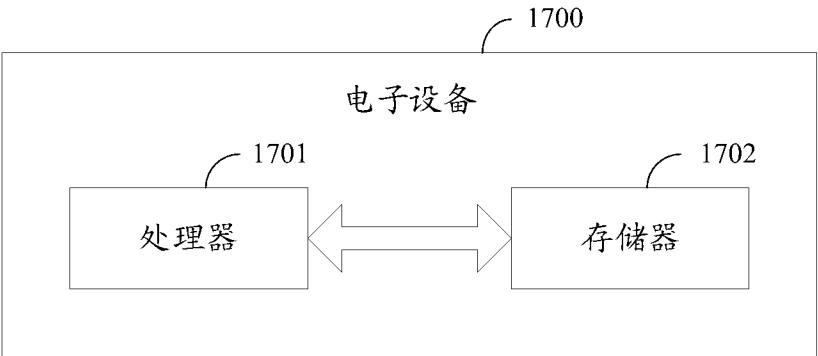


图 17

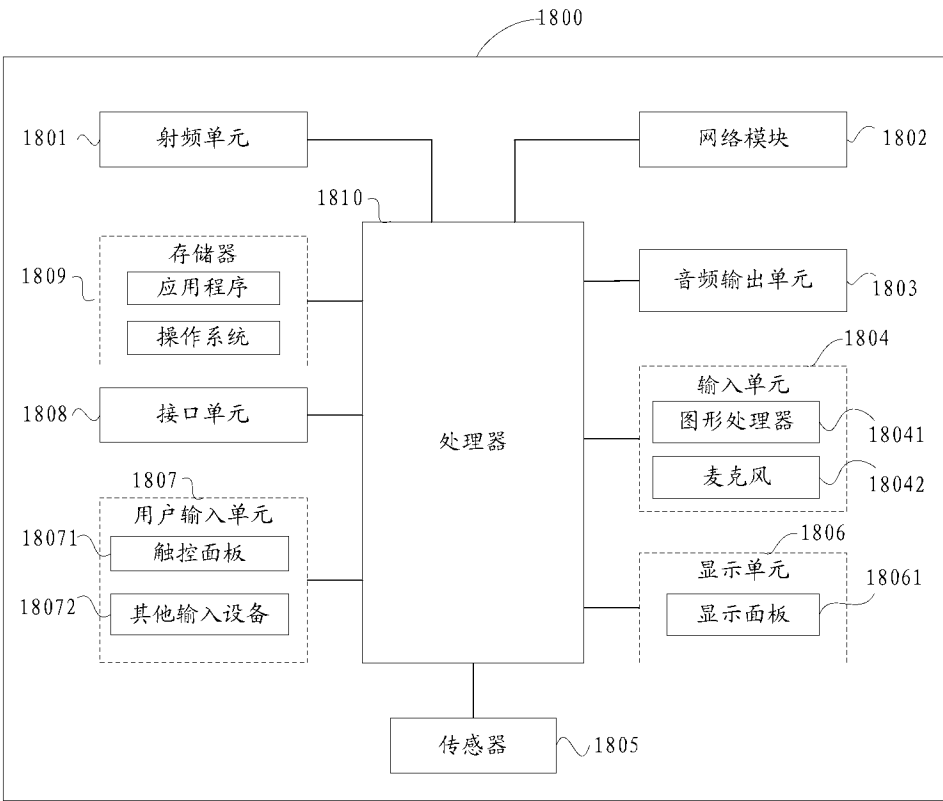


图 18

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2024/103191

A. CLASSIFICATION OF SUBJECT MATTER

G10L19/005(2013.01)i; G10L25/30(2013.01)i; H04L65/75(2022.01)i; G10L25/21(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)

IPC: G10L 19, G10L 25, H04

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)

CNABS, CNTXT, WPABSC, ENTXTC, VEN: 丢包, 丢帧, 丢, 缺, 失, 包, 帧, 语音, 音频, 声音, 通话, 编码, 解码, 卷积, 残差, 重建, 重构, 恢复, 补偿, 隐藏, loss+, lost, miss+, audio, voice, sound, speech, VOIP, recover+, compens+, conceal+, reconstruct+, reconstitute+, encoder+, decoder+, residual, Resnet, CED?

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 116705040 A (VIVO MOBILE COMMUNICATION CO., LTD.) 05 September 2023 (2023-09-05) claims 1-20, description, paragraphs 5-312, and figures 1-18	1-23
X	观云升 (GUAN, Yunsheng). "语音时频域丢失信息重建方法研究 (Research on Reconstruction Method of Speech Time Domain and Frequency Domain Lost Information)" 中国优秀硕士学位论文全文数据库 (信息科技辑) (Information Science & Technology, China Master's Theses Full-Text Database), No. 11, 15 November 2022 (2022-11-15), 26-31 theses, pages 26-31, chapter 3, section 3.3	1-23
X	CN 115171705 A (ALIBABA CLOUD COMPUTING (BEIJING) CO., LTD.) 11 October 2022 (2022-10-11) description, paragraphs 67-100	1-23
A	CN 113035205 A (ALIBABA (CHINA) CO., LTD.) 25 June 2021 (2021-06-25) entire document	1-23
A	CN 115171707 A (BIGO TECHNOLOGY PTE. LTD.) 11 October 2022 (2022-10-11) entire document	1-23

☒ 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

"D" document cited by the applicant in the international application

"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

25 October 2024

Date of mailing of the international search report

28 October 2024

Name and mailing address of the ISA/CN

China National Intellectual Property Administration (ISA/
CN)
China No. 6, Xitucheng Road, Jimenqiao, Haidian District,
Beijing 100088

Authorized officer

Telephone No.

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	CN 101471073 A (HUAWEI TECHNOLOGIES CO., LTD.) 01 July 2009 (2009-07-01) entire document	1-23
A	CN 114866856 A (BEIJING DAJIA INTERNET INFORMATION TECHNOLOGY CO., LTD.) 05 August 2022 (2022-08-05) entire document	1-23
A	US 2019051310 A1 (INDUSTRY-UNIVERSITY COOPERATION FOUNDATION HANYANG UNIVERSITY) 14 February 2019 (2019-02-14) entire document	1-23

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2024/103191

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	116705040	A	05 September 2023	None			
CN	115171705	A	11 October 2022	None			
CN	113035205	A	25 June 2021	None			
CN	115171707	A	11 October 2022	None			
CN	101471073	A	01 July 2009	None			
CN	114866856	A	05 August 2022	None			
US	2019051310	A1	14 February 2019	KR	20190017242	A	20 February 2019
				KR	102051891	B1	04 December 2019
				US	10861466	B2	08 December 2020

A. 主题的分类

G10L19/005(2013.01)i; G10L25/30(2013.01)i; H04L65/75(2022.01)i; G10L25/21(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)

IPC: G10L 19, G10L 25, H04

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))

CNABS, CNTXT, WPABSC, ENTXTC, VEN: 丢包, 丢帧, 丢, 缺, 失, 包, 帧, 语音, 音频, 声音, 通话, 编码, 解码, 卷积, 残差, 重建, 重构, 恢复, 补偿, 隐藏, loss+, lost, miss+, audio, voice, sound, speech, VOIP, re-cover+, compens+, conceal+, reconstruct+, reconstitute+, encoder+, decoder+, residual, Resnet, CED?

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
PX	CN 116705040 A (维沃移动通信有限公司) 2023年9月5日 (2023 - 09 - 05) 权利要求1-20, 说明书第5-312段, 附图1-18	1-23
X	观云升. "语音时频域丢失信息重建方法研究" 中国优秀硕士学位论文全文数据库 (信息科技辑), 第11期, 2022年11月15日 (2022 - 11 - 15), 26-31 论文第26-31页, 第3章第3.3节	1-23
X	CN 115171705 A (阿里巴巴云计算(北京)有限公司) 2022年10月11日 (2022 - 10 - 11) 说明书第67-100段	1-23
A	CN 113035205 A (阿里巴巴(中国)有限公司) 2021年6月25日 (2021 - 06 - 25) 全文	1-23
A	CN 115171707 A (百果园技术(新加坡)有限公司) 2022年10月11日 (2022 - 10 - 11) 全文	1-23
A	CN 101471073 A (华为技术有限公司) 2009年7月1日 (2009 - 07 - 01) 全文	1-23

☒ 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

“D” 申请人在国际申请中引证的文件

“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

2024年10月25日

国际检索报告邮寄日期

2024年10月28日

ISA/CN的名称和邮寄地址

中国国家知识产权局
中国北京市海淀区蓟门桥西土城路6号 100088

授权官员

杨士林

电话号码 (+86) 010-62085717

PCT/ISA/210 表(第2页) (2022年7月)

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 114866856 A (北京达佳互联信息技术有限公司) 2022年8月5日 (2022 - 08 - 05) 全文	1-23
A	US 2019051310 A1 (UNIV. HANYANG IND. UNIV. COOP. FOUND) 2019年2月14日 (2019 - 02 - 14) 全文	1-23

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2024/103191

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	116705040	A	2023年9月5日	无			
CN	115171705	A	2022年10月11日	无			
CN	113035205	A	2021年6月25日	无			
CN	115171707	A	2022年10月11日	无			
CN	101471073	A	2009年7月1日	无			
CN	114866856	A	2022年8月5日	无			
US	2019051310	A1	2019年2月14日	KR	20190017242	A	2019年2月20日
				KR	102051891	B1	2019年12月4日
				US	10861466	B2	2020年12月8日