

(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



(10) 国际公布号
WO 2022/156479 A1

(43) 国际公布日
2022 年 7 月 28 日 (28.07.2022)

(51) 国际专利分类号:
G10L 13/08 (2013.01)

(21) 国际申请号: PCT/CN2021/140858

(22) 国际申请日: 2021 年 12 月 23 日 (23.12.2021)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
202110076168.0 2021 年 1 月 20 日 (20.01.2021) CN

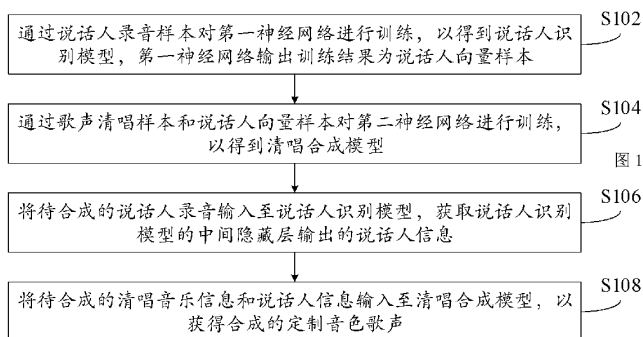
(71) 申请人: 北京沃东天骏信息技术有限公司 (BEIJING WODONG TIANJUN INFORMATION TECHNOLOGY CO., LTD.) [CN/CN]; 中国北京市北京经济技术开发区科创十一街 18 号院 2 号楼 4 层 A402 室, Beijing 100176 (CN)。北京京东世纪贸易有限公司 (BEIJING JINGDONG CENTURY

TRADING CO., LTD.) [CN/CN]; 中国北京市大兴区北京经济技术开发区科创十一街 18 号院 C 座 2 层 201 室, Beijing 100176 (CN)。

(72) 发明人: 张政臣 (ZHANG, Zhengchen); 中国北京市北京经济技术开发区科创十一街 18 号院 2 号楼 4 层 A402 室, Beijing 100176 (CN)。吴俊仪 (WU, Junyi); 中国北京市北京经济技术开发区科创十一街 18 号院 2 号楼 4 层 A402 室, Beijing 100176 (CN)。蔡玉玉 (CAI, Yuyu); 中国北京市北京经济技术开发区科创十一街 18 号院 2 号楼 4 层 A402 室, Beijing 100176 (CN)。袁鑫 (YUAN, Xin); 中国北京市北京经济技术开发区科创十一街 18 号院 2 号楼 4 层 A402 室, Beijing 100176 (CN)。宋伟 (SONG, Wei); 中国北京市北京经济技术开发区科创十一街 18 号院 2 号楼 4 层 A402 室, Beijing 100176 (CN)。何晓冬 (HE, Xiaodong); 中国北京

(54) Title: CUSTOM TONE AND VOCAL SYNTHESIS METHOD AND APPARATUS, ELECTRONIC DEVICE, AND STORAGE MEDIUM

(54) 发明名称: 定制音色歌声的合成方法、装置、电子设备和存储介质



- S102 Train a first neural network by means of a speaker record sample to obtain a speaker recognition model, the output training result of the first neural network being a speaker vector sample
- S104 Train a second neural network by means of an unaccompanied vocal singing sample and the speaker vector sample to obtain an unaccompanied singing synthesis model
- S106 Input a speaker record to be synthesized into the speaker recognition model to obtain speaker information output by the intermediate hidden layer of the speaker recognition model
- S108 Input unaccompanied singing music information to be synthesized and the speaker information into the unaccompanied singing synthesis model to obtain a synthesized custom tone and vocal

(57) Abstract: A custom tone and vocal synthesis method and apparatus, an electronic device, and a storage medium. The synthesis method comprises: training a first neural network by means of a speaker record sample to obtain a speaker recognition model, the output training result of the first neural network being a speaker vector sample (S102); training a second neural network by means of an unaccompanied vocal singing sample and the speaker vector sample to obtain an unaccompanied singing synthesis model (S104); inputting a speaker record to be synthesized into the speaker recognition model to obtain speaker information output by the intermediate

[见续页]

市北京经济技术开发区科创十一街18号院2
号楼4层A402室, Beijing 100176 (CN)。

(74) 代理人: 北京律智知识产权代理有限公司 (BEIJING INTELLEGAL INTELLECTUAL
PROPERTY AGENT LTD.); 中国北京市朝阳区慧忠路5号B1605、B1606、B1607, Beijing
100101 (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,
GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, IT,
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, 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))。

hidden layer of the speaker recognition model (S106); and inputting unaccompanied singing music information to be synthesized and the speaker information into the unaccompanied singing synthesis model to obtain a synthesized custom tone and vocal (S108). By means of the method, the efficiency and effect of custom tone and vocal synthesis are improved, and the model training time and response time of custom tone and vocal synthesis are shortened.

(57) 摘要: 一种定制音色歌声的合成方法、装置、电子设备和存储介质, 该合成方法包括: 通过说话人录音样本对第一神经网络进行训练, 以得到说话人识别模型, 第一神经网络输出训练结果为说话人向量样本 (S102); 通过歌声清唱样本和说话人向量样本对第二神经网络进行训练, 以得到清唱合成模型 (S104); 将待合成的说话人录音输入至说话人识别模型, 获取说话人识别模型的中间隐藏层输出的说话人信息 (S106); 将待合成的清唱音乐信息和说话人信息输入至清唱合成模型, 以获得合成的定制音色歌声 (S108)。通过该方法, 提高了定制音色歌声合成的效率和效果, 缩短了定制音色歌声合成的模型训练时间和响应时间。