

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property
Organization
International Bureau

(43) International Publication Date
22 July 2021 (22.07.2021)



(10) International Publication Number
WO 2021/144679 A1

(51) International Patent Classification:
G10L 25/57 (2013.01) *G10L 25/30* (2013.01)

(21) International Application Number:
PCT/IB2021/050167

(22) International Filing Date:
11 January 2021 (11.01.2021)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
16/744,471 16 January 2020 (16.01.2020) US

(71) Applicant: **INTERNATIONAL BUSINESS MACHINES CORPORATION** [US/US]; New Orchard Road, Armonk, New York, 10504 (US).

(71) Applicants (for MG only): **IBM UNITED KINGDOM LIMITED** [GB/GB]; PO Box 41, North Harbour, Portsmouth Hampshire PO6 3AU (GB). **IBM (CHINA) INVESTMENT COMPANY LIMITED** [CN/CN]; 25/

F, Pangu Plaza, No. 27, Cental North 4th Ring Road, Chaoyang District, Beijing 100101 (CN).

(72) Inventors: **ZHANG, Yang**; IBM Corporation, 75 Binney Street, Cambridge, Massachusetts 02142-1123 (US). **GAN, Chuang**; IBM Corporation, 75 Binney Street, Cambridge, Massachusetts 02142-1123 (US). **WANG, Dakuo**; IBM Corporation, 75 Binney Street, Cambridge, Massachusetts 02142-1123 (US).

(74) Agent: **GRAHAM, Timothy**; IBM United Kingdom Limited, Intellectual Property Law, Hursley Park, Hursley, Winchester Hampshire SO21 2JN (GB).

(81) Designated States (unless otherwise indicated, for every kind of national protection available): 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,

(54) Title: GENERATING SYNCHRONIZED SOUND FROM VIDEOS

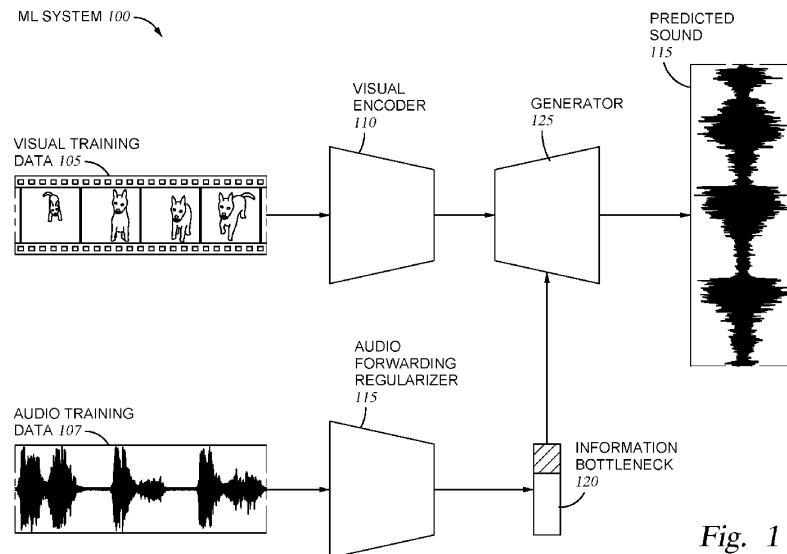


Fig. 1

(57) Abstract: Embodiments herein describe an audio forwarding regularizer and an information bottleneck that are used when training a machine learning (ML) system. The audio forwarding regularizer receives audio training data and identifies visually irrelevant and relevant sounds in the training data. By controlling the information bottleneck, the audio forwarding regularizer forwards data to a generator that is primarily related to the visually irrelevant sounds, while filtering out the visually relevant sounds. The generator also receives data regarding visual objects from a visual encoder derived from visual training data. Thus, when being trained, the generator receives data regarding the visual objects and data regarding the visually irrelevant sounds (but little to no data regarding the visually relevant sounds). Thus, during an execution stage, the generator can generate sounds that are relevant to the visual objects while not adding visually irrelevant sounds to the videos.

[Continued on next page]

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) Designated States (*unless otherwise indicated, for every kind of regional protection available*): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (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).

Published:

— with international search report (Art. 21(3))