

(12) 特許協力条約に基づいて公開された国際出願

(19) 世界知的所有権機関
国際事務局

(43) 国際公開日
2021年11月18日(18.11.2021)



(10) 国際公開番号

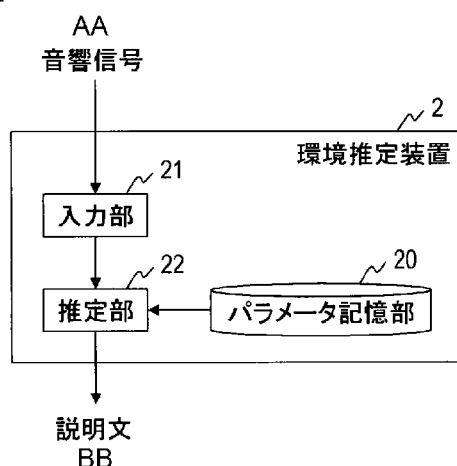
WO 2021/229641 A1

- (51) 国際特許分類:
G10L 15/10 (2006.01) *G10L 25/51* (2013.01) 〒1808585 東京都武蔵野市緑町三丁目9番11号 N T T 知的財産センタ内 Tokyo (JP).
- (21) 国際出願番号: PCT/JP2020/018823 (74) 代理人: 中尾 直樹, 外 (NAKAO, Naoki et al.); 〒1600022 東京都新宿区新宿三丁目1番22号 新宿NSOビル6階 Tokyo (JP).
- (22) 国際出願日: 2020年5月11日(11.05.2020)
- (25) 国際出願の言語: 日本語
- (26) 国際公開の言語: 日本語
- (71) 出願人: 日本電信電話株式会社 (NIPPON TELEGRAPH AND TELEPHONE CORPORATION) [JP/JP]; 〒1008116 東京都千代田区大手町一丁目5番1号 Tokyo (JP).
- (72) 発明者: 小泉 悠馬 (KOIZUMI, Yuma); 〒1808585 東京都武蔵野市緑町三丁目9番11号 N T T 知的財産センタ内 Tokyo (JP). 増村 亮 (MASUMURA, Ryo); 〒1808585 東京都武蔵野市緑町三丁目9番11号 N T T 知的財産センタ内 Tokyo (JP). 齊藤 翔一郎 (SAITO, Shoichiro);
- (81) 指定国(表示のない限り、全ての種類の国内保護が可能): 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, 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.

(54) Title: ENVIRONMENT ESTIMATION METHOD, ENVIRONMENT ESTIMATION DEVICE, AND PROGRAM

(54) 発明の名称: 環境推定方法、環境推定装置、およびプログラム

[図3]



2... ENVIRONMENT ESTIMATION DEVICE
20... PARAMETER STORAGE UNIT
21... INPUT UNIT
22... ESTIMATION UNIT
AA... ACOUSTIC SIGNAL
BB... DESCRIPTIVE TEXT

(57) Abstract: Through the present invention, the environment in which an acoustic signal is collected is estimated with high precision without inputting of auxiliary information. An input unit (21) inputs a subject acoustic signal as an estimation subject. An estimation unit (22) associates an acoustic signal and descriptive text describing the acoustic signal, and thereby estimates the environment in which the subject acoustic signal is collected. The environment is the descriptive text describing the subject acoustic signal obtained from the association. The association is learned so as to minimize the difference between the descriptive text applied to the acoustic signal and the descriptive text obtained by association from the acoustic signal.

(84) 指定国(表示のない限り、全ての種類の広域保護が可能): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), ユーラシア (AM, AZ, BY, KG, KZ, RU, TJ, TM), ヨーロッパ (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).

添付公開書類:

- 国際調査報告 (条約第21条(3))
- 補正された請求の範囲 (条約第19条(1))

(57) 要約: 補助情報を入力することなく、音響信号が集音された環境を高精度に推定する。入力部(21)は、推定対象である対象音響信号を入力する。推定部(22)は、音響信号と音響信号を説明する説明文とを関連付けることで対象音響信号が集音された環境を推定する。環境は、関連付けにより得られる対象音響信号を説明する説明文である。関連付けは、音響信号に付与された説明文と、音響信号から関連付けにより得られた説明文との差を最小化するように学習される。

明 細 書

発明の名称：環境推定方法、環境推定装置、およびプログラム
技術分野

[0001] この発明は、音響信号を説明する自然言語文を生成する技術に関する。

背景技術

[0002] 音響信号などのメディアから説明文を生成する技術は、「キャプションング」と呼ばれる。特に、キャプションングのうち、音声ではなく環境音の説明文を生成するものは、「音キャプションング」と呼ばれる。

[0003] 音キャプションングは、時間領域のL点の観測信号 $x \in \mathbb{R}^L$ を説明するN単語の単語列 $w \in \mathbb{N}^N$ を生成する問題である（例えば、非特許文献1－4参照）。非特許文献1－4では、深層学習を利用したsequence-to-sequence (Seq2Seq) と呼ばれる方式を取っている（例えば、非特許文献5参照）。

[0004] まず、観測信号 x を何らかの音響特徴量系列 $\{\phi_t \in \mathbb{R}^D\}_{t=1}^T$ へ変換する。ここで、 D は音響特徴量の次元であり、 $t=1, \dots, T$ は時間を表すインデックスである。その音響特徴量系列 $\{\phi_t \in \mathbb{R}^D\}_{t=1}^T$ を、パラメータ θ_e を持つエンコーダ E を利用して、別の特徴量空間のベクトルないし行列 $v = E_{\theta_e}(\phi_1, \dots, \phi_T)$ へと埋め込む。そして、 v と、1番目から $n-1$ 番目までに推定された単語 $w_1, \dots, w_{n-1}$ とを、パラメータ θ_d を持つデコーダ D へ入力し、 n 番目の単語 w_n の生成確率 $p(w_n | v, w_1, \dots, w_{n-1})$ を推定する。

[0005] [数1]

$$p(w_n | v, w_1, \dots, w_{n-1}) = D_{\theta_d}(v, w_1, \dots, w_{n-1}) \quad \dots (1)$$

[0006] ただし、デコーダ D の最終出力にはsoftmax関数が適用されている。これを $n=2$ から $n=N$ まで繰り返すことで、文章全体の生成確率 $p(w_1, \dots, w_N | x)$ を推定する。

[0007] [数2]

$$p(w_1, \dots, w_N | x) = \prod_{n=2}^N p(w_n | v, w_1, \dots, w_{n-1}) \quad \dots (2)$$

- [0008] ただし、 w_1 は文章の開始を表すトークンに対応するインデックスに固定されている。実際には、エンコーダEやデコーダDは、長短期記憶（LSTM: Long short-term memory）などの再帰型ニューラルネットワークや、Transformerと呼ばれる注意機構（Attention）を利用したニューラルネットワークにより実装される。また、 w_n は $p(w_n|v, w_1, \dots, w_{n-1})$ が最大となるインデックスとするか、 $p(w_n|v, w_1, \dots, w_{n-1})$ からサンプリングすることで生成する。
- [0009] 音キャプションの難しさは、音に対応する生成文が無数に存在し得ることにある。まず、機械翻訳の例を挙げて説明する。例えば“A powerful car engine running with wind blowing.”という英文を和訳することを考える。この英文には“powerful”, “car”, “wind”, “blowing”という単語が含まれているため、それらに対応する日本語の単語「力強い／パワフル」「車」「風」「吹く」を使いながら翻訳するのが妥当であろう。ゆえに、人間がこの英文を和訳するのであれば、翻訳文は「風に吹かれながら走る、車のパワフルなエンジン」や「車のエンジンが、力強く風に吹かれながら走っている」などとなるのが妥当であろう。つまり、文章の鍵となる単語（キーワード）を外しながら翻訳することは妥当ではない。
- [0010] 一方で、エンジン音と正体不明の環境音とが混在する音が含まれる音を説明するときには、その音だけでは、その環境音が例えば風の音なのか水の音なのかわからないこともあるし、そのエンジンが何のエンジンなのかも（専門のエンジニアでない限り）わからないこともあろう。そのため、その音から生成し得る説明文は多様であり、例えば、“A powerful car engine running with wind blowing.”の他にも、“A speedboat is traveling across water.”、“A small motor runs loudly.”、“An engine buzzing consistently.”など、文章の翻訳以上に多様な説明文が許容されてしまう。
- [0011] そこで、従来の音キャプションでは、音響信号とキーワードの両方を入力に使ってよい、という問題設定で解かれることが多い。音キャプションのデータセットは、YouTube（登録商標）やフリーの音素材集（例えば、非特許文献6）から収集されたものが用いられることが多い。そういったデ

ータには、そのメディアに関するタグが付与されていることがほとんどである。例えば、機関車の通過音や汽笛には、“railway”, “trains”, “horn”などのタグが付与される（例えば、非特許文献7）。また、非特許文献4で用いられるデータセットは、YouTube（登録商標）から収集された音に人手でオントロジーラベルが付与されたものである。非特許文献4や音キャプションングのコンペティションでは、こういったタグをキーワードとして利用し、それをエンコーダEやデコーダDに同時に入力することで、説明文の多様性を抑制したり、文章生成の精度を向上させたりしている。つまり、従来技術では、音キャプションングの難しさを、問題設定を変えることで回避している。

先行技術文献

非特許文献

[0012] 非特許文献1: K. Drossos, et al., “Automated audio captioning with recurrent neural networks,” Proc. of WASPAA, 2017.

非特許文献2: S. Ikawa, et al., “Neural audio captioning based on conditional sequence-to-sequence model,” Proc. of DCASE, 2019.

非特許文献3: M. Wu, et al., “Audio Caption: Listen and Tell,” Proc. of ICASSP, 2019.

非特許文献4: C. D. Kim, et al., “AudioCaps: Generating Captions for Audios in The Wild,” Proc. of NAACL-HLT, 2019.

非特許文献5: I. Sutskever, O. Vinyals, and Q. V. Le, “Sequence to Sequence Learning with Neural Networks,” Proc. of NIPS, 2014.

非特許文献6: K. Drossos, et al., “Clotho: An Audio Captioning Dataset,” Proc. of ICASSP, 2020.

非特許文献7: Freesound, “Train passing by in the Wirikuta Desert (Mexico, SLP)”、[online]、[令和2年4月27日検索]、インターネット<URL: <https://freesound.org/people/felix.blume/sounds/166086/>>

発明の概要

発明が解決しようとする課題

[0013] しかしながら、キーワード入力を許容する問題設定では、音キャプショニング技術の利用可能範囲を著しく制限してしまう。例えば、機械が発する異常な音を検出し、それをユーザに報告するシステムを考える。この場合、単に「異常」というラベルを出力するよりも、「普段は聞かない、金属がこすれるような高い音がしている」といった説明文を出力した方が、ユーザには理解しやすいし、故障箇所の発見などにも役立つだろう。しかしながら、説明文生成にキーワードが必要な場合、人間が「金属」「こすれる」などのキーワードを入力する必要がある。そこまで音の特徴がわかっているならば、もはや説明文を生成する必要はない。音キャプショニングの利用範囲を広げるためには、キーワードなどの補助情報を入力することなく、高精度な説明文生成が可能な技術が必要とされる。

[0014] この発明の目的は、上記のような技術的課題に鑑みて、補助情報を入力することなく、音響信号が集音された環境を高精度に推定することである。

課題を解決するための手段

[0015] 上記の課題を解決するために、この発明の一態様の環境推定方法は、推定対象である対象音響信号を入力する入力ステップと、音響信号と音響信号を説明する説明文とを関連付けることで対象音響信号が集音された環境を推定する推定ステップと、を含む。

発明の効果

[0016] この発明によれば、補助情報を入力することなく、音響信号が集音された環境を高精度に推定することができる。

図面の簡単な説明

[0017] [図1]図1は推定器学習装置の機能構成を例示する図である。

[図2]図2は推定器学習方法の処理手順を例示する図である。

[図3]図3は環境推定装置の機能構成を例示する図である。

[図4]図4は環境推定方法の処理手順を例示する図である。

[図5]図5は実験結果を説明するための図である。

[図6]図6はコンピュータの機能構成を例示する図である。

発明を実施するための形態

[0018] [発明の概要]

本発明は、キーワードなどの補助情報を用いずに、音キャプションを行う技術である。つまり、学習データとして「音」と「説明文」のペアが与えられ、テストデータとして「音」のみが与えられる、最もシンプルかつ実用的な問題設定を考える。

[0019] 従来技術からも明らかなように、音キャプションにキーワードを入力することは、高精度に文章生成をすることに有用である。そこで、本発明では、まず、学習データの説明文からキーワードを抽出し、それを学習データに追加する。次に、観測音からキーワードを推定し、それが上記で抽出したキーワードと一致するようにディープニューラルネットワーク（DNN: Deep Neural Network）などの統計モデル（以下、「推定器」とも呼ぶ）を学習する。そして、推定したキーワードとキーワードを埋め込んだ音響特徴量とをデコードに同時に入力し、説明文を生成する。

[0020] <キーワード学習データの作成>

キーワード学習データの作成手順を説明する。実行手順は以下である。

1. 正解の説明文を単語分割（tokenize）する。
2. 分割された単語それぞれの品詞を推定する。
3. 事前に指定した品詞の単語のみについて語幹を求め、それをキーワードとする。ここで利用する品詞は何でもよいが、例えば、名詞、動詞、副詞、形容詞などを利用すればよい。

[0021] 上記手順で得られるキーワードが多すぎる場合、すべての学習データについて上記手順を繰り返し、抽出された単語それぞれの出現回数をカウントし、頻出する単語のみをキーワードとして選択してもよい。

[0022] 例として“A muddled noise of broken channel of the TV.”という文を考える。この文中で名詞、動詞、副詞、形容詞のいずれかに該当する単語は、”

muddled”, “noise”, “broken”, “channel”, “TV”である。これらの語幹はそれぞれ“muddle”, “noise”, “break”, “channel”, “TV”である。したがって、“A muddled noise of broken channel of the TV.”のキーワードは“muddle”, “noise”, “break”, “channel”, “TV”となる。

[0023] <キーワード推定>

キーワードの種類を K とし、観測音 x に付与されているキーワードを表すベクトル（以下、「キーワードベクトル」とも呼ぶ）を $h \in \mathbb{N}^K$ とする。ここで、 h は k 番目のキーワードが含まれていれば k 次元目が1、含まれていなければ0となるバイナリベクトルである。

[0024] ここでの目的は、観測音 x から抽出した音響特徴量系列 $\{\phi_t \in \mathbb{R}^D\}_{t=1}^T$ から、キーワードベクトル h を高精度に推定することである。最も単純には、キーワードを推定するパラメータ θ_a を持つ推定器 A を考え、 $h = A_{\theta_a}(\phi_1, \dots, \phi_T)$ とする方法が考えられる。しかし、 T の大きさは入力音に依存して可変であり、それを固定長のベクトルへと変換することは容易ではない。また、“background”や“engine”といったキーワードに対応する音は長い時間にわたって鳴っている一方で、“door”や“knock”などのキーワードに対応する音は短い時間しか鳴っていないと想定される。つまり、音の信号処理／統計処理としては、時刻 t 毎にキーワードを推定する方が理にかなっている。しかし、時刻 t 毎にキーワードのラベルが与えられているのであれば、その推定器の学習は容易であるが（「強ラベル学習」と呼ばれる）、本発明の問題設定では、一つのキーワードのラベルは時系列全体に対してしか与えられていない。そこで本発明では、弱ラベル付き音響イベント検出問題を参考に推定方法を設計する。

[0025] まず、推定器 A を利用して、時刻 t において k 番目のキーワードに関する音が存在する確率 $p(z_{t,k} | x)$ を以下のように推定する。

[0026] [数3]

$$p(z_{t,k} | x) = A_{\theta_a}(\phi_1, \dots, \phi_T) \quad \dots (3)$$

[0027] ここで、 $A: \mathbb{R}^{D \times T} \rightarrow \mathbb{R}^{K \times T}$ であり、出力層の活性化関数はsigmoid関数となる。

[0028] 次に、音響特徴量系列内に k 番目のキーワードが存在する確率 $p(z_k | x)$ を、 $p($

$z_{t,k}|x)$ を時間方向に集約することで推定する。集約方法は時間方向の平均など様々なものが考えられ、どのようなものでもよいが、例えば、以下の最大値を取る方法などを推奨する。

[0029] [数4]

$$p(z_k|x) = \max_t p(z_{t,k}|x) \quad \cdots (4)$$

[0030] そして、パラメータ θ_a を以下の重み付きバイナリクロスエントロピーを最小化するように最適化する。

[0031] [数5]

$$\mathcal{L}_{\text{key}} = -\frac{1}{K} \sum_{k=1}^K \lambda h_k \ln(p(z_k|x)) + \gamma(1-h_k) \ln(1-p(z_k|x)) \quad \cdots (5)$$

[0032] ここで、重み係数 λ と γ は出現頻度の逆数として定義され、以下となる。

[0033] [数6]

$$\lambda = \frac{1}{p(h_k)}, \gamma = \frac{1}{1-p(h_k)} \quad \cdots (6)$$

[0034] また、 $p(h_k)$ は学習データに k 番目のキーワードが出現する比率として以下のよう求める。

[0035] [数7]

$$p(h_k) = \frac{\text{学習データ中の} k \text{ 番目のキーワードの出現回数}}{\text{学習データ数}} \quad \cdots (7)$$

[0036] なお、式 (5) は一個のサンプルから求めるように記載されているが、実際には、数個のサンプル（ミニバッチ）の平均値を最小化するように学習する。

[0037] <推定したキーワードを併用した説明文生成>

キーワードを説明文生成の入力とするために、 $p(z_k|x)$ から M 個のキーワードを選択する。この実現方法はいくつか考えられるが、何を利用してもよい。例えば、 $p(z_k|x)$ をソートし、確率の高いものから M 個を選択する、などが考えられる。このようにして選択した M 個のキーワードを、何らかの単語埋め込み

方法を利用して特徴量空間に埋め込み、M個のキーワード特徴量ベクトル $\{y_m \in \mathbb{R}^D\}_{m=1}^M$ を得る。そして、キーワード特徴量ベクトル $\{y_m \in \mathbb{R}^D\}_{m=1}^M$ をエンコーダEやデコーダDの入力とすることで説明文を生成する。この実装法には、簡便なものには例えば以下を利用できる。

[0038] [数8]

$$p(w_n | v, w_1, \dots, w_{n-1}) = D_{\theta_d}(v, y_1, \dots, y_M, w_1, \dots, w_{n-1}) \quad \dots (8)$$

[0039] そして、各パラメータ $\theta_e, \theta_d, \theta_a$ を、何らかの目的関数を最小化するように学習する。これには、例えば、生成した説明文と人手で与えた正解の説明文とのクロスエントロピーと、式(5)のキーワード推定誤差 $\mathcal{L}_{key}$ との和などが利用できる。

[0040] [数9]

$$\mathcal{L} = \frac{1}{N-1} \sum_{n=2}^N \text{CE}(w_n, p(w_n | v, w_1, \dots, w_{n-1})) + \mathcal{L}_{key} \quad \dots (9)$$

[0041] ただし、 $\text{CE}(a, b)$ はaとbから計算されるクロスエントロピーである。

[0042] 本発明のポイントは、以下の三点である。

[0043] ポイント1. キーワードは、正解の説明文からヒューリスティックルールにより抽出される。正解の説明文を単語分割および品詞推定した上で、「名詞、動詞、形容詞、副詞」など所定の品詞のみを利用する。そして、各単語を語幹に変形して、全学習データでカウントし、頻出する単語をキーワードとして利用する。

[0044] ポイント2. キーワードの推定は、弱ラベル付き音響イベント検出問題として解く。キーワードに関連する音情報は、入力された音データの全体に含まれているかもしれないし、一部にしか含まれていないかもしれない。一部にしか含まれていない音（例えばドア音など）を高精度に検知するためには、何秒目から何秒目までその音が鳴っていたか、という強ラベル（音の種類＋発生時刻）が必要であるが、ポイント1の方法では時刻を知ることはできない。そこで、キーワードに関連する音の発生確率は時刻毎に計算するが、学習は時間を集約する形で行う、弱ラベル付き音響イベント検出の問題設定

を踏襲する。そして、学習方法を工夫することで、キーワードを高精度に推定できるようにする。

[0045] ポイント3. 推定したキーワードは、事前に学習した特徴空間に埋め込み、デコーダへ入力される。この埋め込み方法は複数種類あるが、実施形態ではそのうちの一つを示す。

[0046] [実施形態]

以下、この発明の実施の形態について詳細に説明する。なお、図面中において同じ機能を有する構成部には同じ番号を付し、重複説明を省略する。

[0047] この発明の実施形態は、音響信号と正解の説明文とのペアからなる学習データセットを用いて、音響信号から説明文を生成する推定器のパラメータを学習する推定器学習装置および方法と、推定器学習装置および方法により学習された推定器を用いて音響信号からその音響信号が集音された環境を推定する環境推定装置および方法とからなる。

[0048] <推定器学習装置>

図1に示すように、実施形態の推定器学習装置1は、例えば、学習データ記憶部10、キーワード生成部11、初期化部12、ミニバッチ選択部13、特徴量抽出部14、キーワード推定部15、キーワード埋込部16、説明文生成部17、パラメータ更新部18、収束判定部19、およびパラメータ記憶部20を備える。この推定器学習装置1が、図2に示す各ステップを実行することにより、実施形態の推定器学習方法が実現される。

[0049] 推定器学習装置1は、例えば、中央演算処理装置（CPU: Central Processing Unit）、主記憶装置（RAM: Random Access Memory）などを有する公知又は専用のコンピュータに特別なプログラムが読み込まれて構成された特別な装置である。推定器学習装置1は、例えば、中央演算処理装置の制御のもとで各処理を実行する。推定器学習装置1に入力されたデータや各処理で得られたデータは、例えば、主記憶装置に格納され、主記憶装置に格納されたデータは必要に応じて中央演算処理装置へ読み出されて他の処理に利用される。推定器学習装置1の各処理部は、少なくとも一部が集積回路等のハードウ

エアによって構成されていてもよい。推定器学習装置 1 が備える各記憶部は、例えば、RAM (Random Access Memory) などの主記憶装置、ハードディスクや光ディスクもしくはフラッシュメモリ (Flash Memory) のような半導体メモリ素子により構成される補助記憶装置、またはリレーショナルデータベースやキーバリューストアなどのミドルウェアにより構成することができる。

[0050] 以下、図 2 を参照して、実施形態の推定器学習装置 1 が実行する推定器学習方法について説明する。

[0051] 学習データ記憶部 10 には、複数の学習データからなる学習データセットが記憶されている。各学習データは、予め収集した音響信号と、各音響信号に対して人手で付与した正解の説明文とからなる。

[0052] ステップ S 10 において、推定器学習装置 1 は、学習データ記憶部 10 に記憶されている学習データセットを読み出す。読み出された学習データセットは、キーワード生成部 11 へ入力される。

[0053] ステップ S 11 において、キーワード生成部 11 は、入力された学習データセットから各説明文に対応するキーワードを生成する。キーワードの生成方法は、上記の〈キーワード学習データの作成〉で説明した手順に従う。キーワード生成部 11 は、生成したキーワードを学習データの音響信号および説明文に付与して学習データ記憶部 10 に記憶する。

[0054] ステップ S 12 において、初期化部 12 は、推定器のパラメータ θ_e , θ_d , θ_a を乱数などで初期化する。

[0055] ステップ S 13 において、ミニバッチ選択部 13 は、学習データ記憶部 10 に記憶されている学習データセットからランダムに 100 サンプル程度のミニバッチを選択する。ミニバッチ選択部 13 は、選択したミニバッチを特徴量抽出部 14 へ出力する。

[0056] ステップ S 14 において、特徴量抽出部 14 は、ミニバッチに含まれる各音響信号から音響特徴量を抽出する。特徴量抽出部 14 は、抽出した音響特徴量をキーワード推定部 15 へ出力する。

[0057] ステップ S 15 において、キーワード推定部 15 は、入力された音響特徴

量からキーワードを推定する。キーワードの推定方法は、上記の＜キーワード推定＞で説明した手順に従う。キーワード推定部 15 は、生成したキーワードをキーワード埋込部 16 へ出力する。

[0058] ステップ S 16 において、キーワード埋込部 16 は、入力されたキーワードを特徴量空間に埋め込み、キーワード特徴量ベクトルを生成する。キーワード埋込部 16 は、生成したキーワード特徴量ベクトルを説明文生成部 17 へ出力する。

[0059] ステップ S 17 において、説明文生成部 17 は、入力されたキーワード特徴量ベクトルを用いて説明文を生成する。説明文の生成方法は、上記の＜推定したキーワードを併用した説明文生成＞で説明した手順に従う。説明文生成部 17 は、生成した説明文をパラメータ更新部 18 へ出力する。

[0060] ステップ S 18 において、パラメータ更新部 18 は、式 (9) のコスト関数 L のミニバッチでの平均を減少させるように、推定器のパラメータ θ_e , θ_d , θ_a を更新する。

[0061] ステップ S 19 において、収束判定部 19 は、あらかじめ設定した終了条件を満たしているか否かを判定する。収束判定部 19 は、終了条件を満たしていればステップ S 20 へ処理を進め、満たしていなければステップ S 13 へ処理を戻す。終了条件は、例えば、パラメータ更新を所定の回数実行したこと、などと設定すればよい。

[0062] ステップ S 20 において、推定器学習装置 1 は、学習済みのパラメータ θ_e , θ_d , θ_a をパラメータ記憶部 20 へ記憶する。

[0063] <環境推定装置>

図 3 に示すように、実施形態の環境推定装置 2 は、推定対象の音響信号を入力とし、その音響信号を説明する説明文を出力する。環境推定装置 2 は、例えば、パラメータ記憶部 20、入力部 21、および推定部 22 を備える。この環境推定装置 2 が、図 4 に示す各ステップを実行することにより、実施形態の環境推定方法が実現される。

[0064] 環境推定装置 2 は、例えば、中央演算処理装置 (CPU: Central Processing

Unit)、主記憶装置(RAM: Random Access Memory)などを有する公知又は専用のコンピュータに特別なプログラムが読み込まれて構成された特別な装置である。環境推定装置2は、例えば、中央演算処理装置の制御のもとで各処理を実行する。環境推定装置2に入力されたデータや各処理で得られたデータは、例えば、主記憶装置に格納され、主記憶装置に格納されたデータは必要に応じて中央演算処理装置へ読み出されて他の処理に利用される。環境推定装置2の各処理部は、少なくとも一部が集積回路等のハードウェアによって構成されていてもよい。環境推定装置2が備える各記憶部は、例えば、RAM(Random Access Memory)などの主記憶装置、ハードディスクや光ディスクもしくはフラッシュメモリ(Flash Memory)のような半導体メモリ素子により構成される補助記憶装置、またはリレーショナルデータベースやキーバリューストアなどのミドルウェアにより構成することができる。

[0065] 以下、図4を参照して、実施形態の環境推定装置2が実行する環境推定方法について説明する。

[0066] パラメータ記憶部20には、推定器学習装置1により学習された推定器のパラメータ θ_e , θ_d , θ_a が記憶されている。

[0067] ステップS21において、入力部21へ、推定対象である対象音響信号が入力される。ここでは、対象音響信号は音声ではなく、環境音であるものとする。入力部21は、入力された対象音響信号を推定部22へ出力する。

[0068] ステップS22において、推定部22は、入力された対象音響信号を、パラメータ記憶部20に記憶された学習済みのパラメータ θ_e , θ_d , θ_a を用いる推定器に入力し、対象音響信号を説明する説明文を推定する。ここでは、対象音響信号は環境音であるため、推定される説明文は対象音響信号が集音された環境を説明する自然言語文となる。なお、環境とは、対象音響信号が集音された場所の周囲で発生していた音響イベントおよび／または非音響イベントを含むものを意味する。推定部22は、推定した説明文を環境推定装置2の出力とする。

[0069] [実験結果]

図5に、本発明の有効性を確認するためにClotho dataset（非特許文献6）を利用して実験を行った結果を示す。評価指標は、DCASE2020 Challenge task 6と同じもの（ $BLEU_1$, $BLEU_2$, $BLEU_3$, $BLEU_4$ （図中ではB-1, B-2, B-3, B-4と記載）, CIDEr, METEOR, $ROUGE_L$ （図中ではROUGE-Lと記載）, SPICE, SPIDER）を利用した。比較手法は、DCASE2020 Challenge task 6のベースラインシステム（図中のBaseline）と、LSTMとTransformerを利用したSec2Secモデルとした（図中のLSTM, Transformer）。本発明は、Transformerのエンコーダの後段にキーワード推定を利用し、式（8）のようにデコーダへ同時に入力するものである。“# of param”は各手法で用いたパラメータ数である。各手法の行に記載された数値は各評価指標によるスコアであり、値が大きいほど精度が高いことを表す。

[0070] 図5からもわかるように、本発明はベースラインシステムよりも少ないパラメータ数で高精度に説明文推定ができています。また、キーワード推定を用いないLSTMやTransformerと比べて、同程度のパラメータ数で高精度に説明文推定ができていたことがわかる。

[0071] 以上、この発明の実施の形態について説明したが、具体的な構成は、これらの実施の形態に限られるものではなく、この発明の趣旨を逸脱しない範囲で適宜設計の変更等があっても、この発明に含まれることはいうまでもない。実施の形態において説明した各種の処理は、記載の順に従って時系列に実行されるのみならず、処理を実行する装置の処理能力あるいは必要に応じて並列的にあるいは個別に実行されてもよい。

[0072] [プログラム、記録媒体]

上記実施形態で説明した各装置における各種の処理機能をコンピュータによって実現する場合、各装置が有すべき機能の処理内容はプログラムによって記述される。そして、このプログラムを図6に示すコンピュータの記憶部1020に読み込ませ、演算処理部1010、入力部1030、出力部1040などに動作させることにより、上記各装置における各種の処理機能がコンピュータ上で実現される。

- [0073] この処理内容を記述したプログラムは、コンピュータで読み取り可能な記録媒体に記録しておくことができる。コンピュータで読み取り可能な記録媒体は、例えば、非一時的な記録媒体であり、磁気記録装置、光ディスク等である。
- [0074] また、このプログラムの流通は、例えば、そのプログラムを記録したDVD、CD-ROM等の可搬型記録媒体を販売、譲渡、貸与等することによって行う。さらに、このプログラムをサーバコンピュータの記憶装置に格納しておき、ネットワークを介して、サーバコンピュータから他のコンピュータにそのプログラムを転送することにより、このプログラムを流通させる構成としてもよい。
- [0075] このようなプログラムを実行するコンピュータは、例えば、まず、可搬型記録媒体に記録されたプログラムもしくはサーバコンピュータから転送されたプログラムを、一旦、自己の非一時的な記憶装置である補助記録部1050に格納する。そして、処理の実行時、このコンピュータは、自己の非一時的な記憶装置である補助記録部1050に格納されたプログラムを一時的な記憶装置である記憶部1020に読み込み、読み込んだプログラムに従った処理を実行する。また、このプログラムの別の実行形態として、コンピュータが可搬型記録媒体から直接プログラムを読み込み、そのプログラムに従った処理を実行することとしてもよく、さらに、このコンピュータにサーバコンピュータからプログラムが転送されるたびに、逐次、受け取ったプログラムに従った処理を実行することとしてもよい。また、サーバコンピュータから、このコンピュータへのプログラムの転送は行わず、その実行指示と結果取得のみによって処理機能を実現する、いわゆるASP (Application Service Provider) 型のサービスによって、上述の処理を実行する構成としてもよい。なお、本形態におけるプログラムには、電子計算機による処理の用に供する情報であってプログラムに準ずるもの（コンピュータに対する直接の指令ではないがコンピュータの処理を規定する性質を有するデータ等）を含むものとする。

[0076] また、この形態では、コンピュータ上で所定のプログラムを実行させることにより、本装置を構成することとしたが、これらの処理内容の少なくとも一部をハードウェア的に実現することとしてもよい。

請求の範囲

- [請求項1] 推定対象である対象音響信号を入力する入力ステップと、
音響信号と前記音響信号を説明する説明文とを関連付けることで前記対象音響信号が集音された環境を推定する推定ステップと、
を含む環境推定方法。
- [請求項2] 請求項1に記載の環境推定方法であって、
前記環境は、前記対象音響信号が集音された場所の周囲で発生していた音響イベントおよび／または非音響イベントを含む、
環境推定方法。
- [請求項3] 請求項1または2に記載の環境推定方法であって、
前記環境は、前記関連付けにより得られる前記対象音響信号を説明する説明文であり、
前記関連付けは、前記音響信号に付与された説明文と、前記音響信号から前記関連付けにより得られた説明文との差を最小化するように学習される、
環境推定方法。
- [請求項4] 請求項3に記載の環境推定方法であって、
前記関連付けは、前記説明文から抽出された前記音響信号を説明するためのキーワードを前記音響信号のラベルとして用いて行われる、
環境推定方法。
- [請求項5] 請求項4に記載の環境推定方法であって、
前記関連付けは、前記関連付けを用いて得られた、前記音響信号を説明するためのキーワードの候補である候補キーワードが、前記音響信号に含まれる確率を用いて行われる、
環境推定方法。
- [請求項6] 推定対象である対象音響信号を入力する入力部と、
音響信号と前記音響信号を説明する説明文とを関連付けることで前記対象音響信号が集音された環境を推定する推定部と、

を含む環境推定装置。

[請求項7] 請求項6に記載の環境推定装置としてコンピュータを機能させるためのプログラム。

補正された請求の範囲
[2021年7月5日(05.07.2021)国際事務局受理]

[請求項 1] (補正後)

推定対象である対象音響信号を入力する入力ステップと、
音響信号と前記音響信号を説明する説明文とを関連付けることで前記対象音響信号が集音された環境を推定する推定ステップと、
を含み、
前記環境は、前記対象音響信号が集音された場所の周囲で発生していた音響イベントおよび／または非音響イベントを含む、
環境推定方法。

[請求項 2] (削除)

[請求項 3] (補正後)

請求項 1 に記載の環境推定方法であって、
前記環境は、前記関連付けにより得られる前記対象音響信号を説明する説明文であり、
前記関連付けは、前記音響信号に付与された説明文と、前記音響信号から前記関連付けにより得られた説明文との差を最小化するように学習される、
環境推定方法。

[請求項 4]

請求項 3 に記載の環境推定方法であって、
前記関連付けは、前記説明文から抽出された前記音響信号を説明するためのキーワードを前記音響信号のラベルとして用いて行われる、
環境推定方法。

[請求項 5]

請求項 4 に記載の環境推定方法であって、
前記関連付けは、前記関連付けを用いて得られた、前記音響信号を説明するためのキーワードの候補である候補キーワードが、前記音響信号に含まれる確率を用いて行われる、
環境推定方法。

[請求項 6] (補正後)

推定対象である対象音響信号を入力する入力部と、

音響信号と前記音響信号を説明する説明文とを関連付けることで前記対象音響信号が集音された環境を推定する推定部と、

を含み、

前記環境は、前記対象音響信号が集音された場所の周囲で発生していた音響イベントおよび／または非音響イベントを含む、

環境推定装置。

[請求項 7]

請求項 6 に記載の環境推定装置としてコンピュータを機能させるためのプログラム。

[図1]

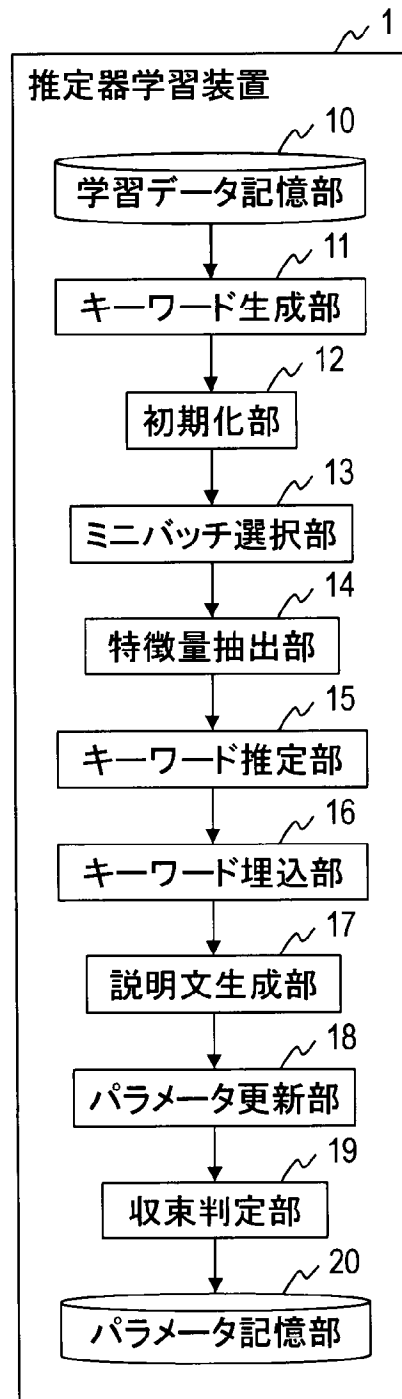


図1

[図2]

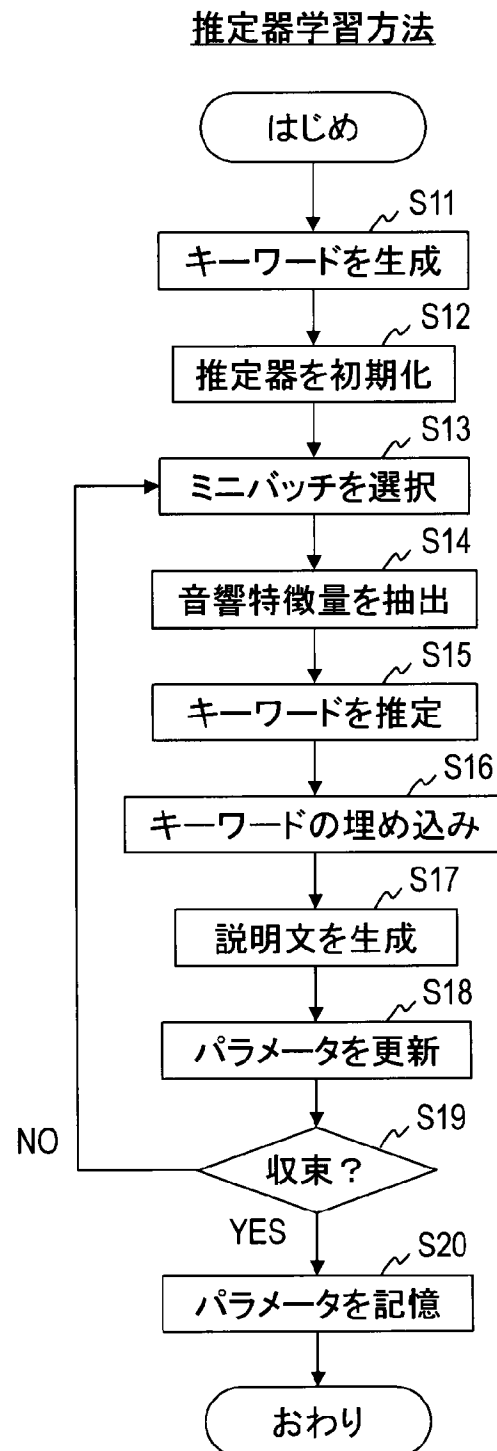


図2

[図3]

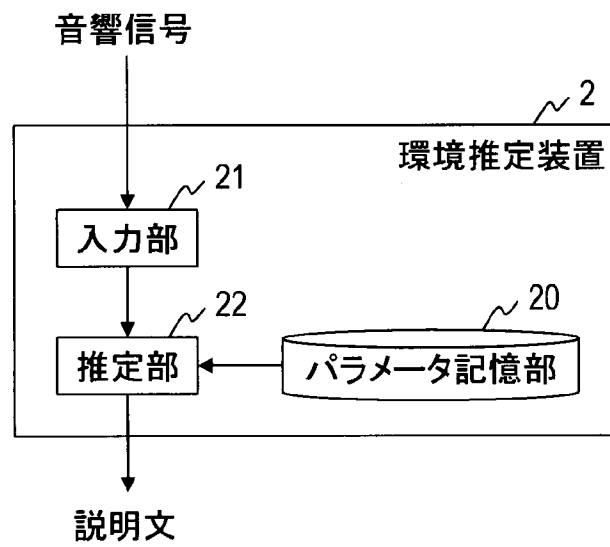


図3

[図4]

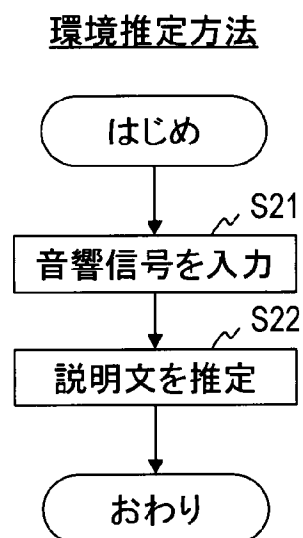


図4

[図5]

Table 1: Experimental results on Clotho dataset and DCASE2020 Challenge metrics.

Method	# of param.	B-1	B-2	B-3	B-4	CIDEr	METEOR	ROUGE-L	SPICE	SPIDEr
Baseline	4.64M	41.9	12.4	4.0	1.0	2.5	7.8	25.5	2.3	2.4
LSTM	1.14M	49.2	27.5	15.8	8.7	20.6	13.9	32.2	9.2	14.9
Transformer	1.11M	50.2	29.9	18.3	10.2	23.3	14.1	33.7	9.1	16.2
Ours	1.13M	52.1	30.9	18.8	10.7	25.8	14.9	34.2	9.7	17.7

[図5]

[図6]

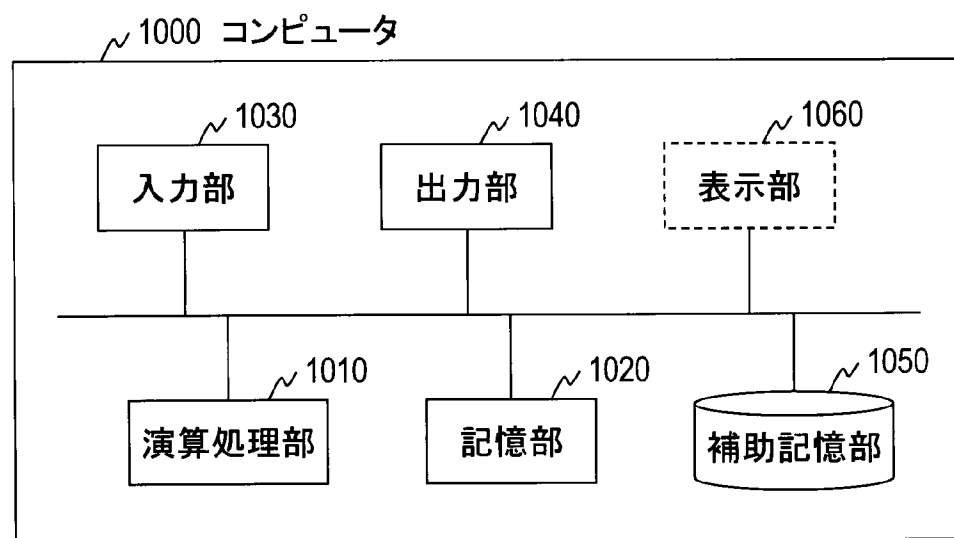


図6

INTERNATIONAL SEARCH REPORT

International application No.

PCT/JP2020/018823

A. CLASSIFICATION OF SUBJECT MATTER

Int. Cl. G10L15/10 (2006.01) i, G10L25/51 (2013.01) i
FI: G10L15/10 500Z, G10L25/51

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)

Int. Cl. G10L13/00-25/93

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

Published examined utility model applications of Japan 1922-1996

Published unexamined utility model applications of Japan 1971-2020

Registered utility model specifications of Japan 1996-2020

Published registered utility model applications of Japan 1994-2020

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

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X A	IKAWA, Shota et al. Neural Audio Captioning Based on Conditional Sequence-to-Sequence Model. Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop. October 2019, pp. 99-103, in particular, section 3.1, fig. 1, table 2	1-3, 6-7 4-5
A	KIM, Chris Dongjoo et al. AudioCaps: Generating Captions for Audios in The Wild, Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technology, June 2019, vol. 1, pp. 119-132, in particular, chapter 4, fig. 4	4-5

☒ 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
02.09.2020

Date of mailing of the international search report
15.09.2020

Name and mailing address of the ISA/
Japan Patent Office
3-4-3, Kasumigaseki, Chiyoda-ku,
Tokyo 100-8915, Japan

Authorized officer

Telephone No.

INTERNATIONAL SEARCH REPORT

International application No.

PCT/JP2020/018823

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	岩月道生ほか. Listen and Tell: 深層学習を用いた音響シーンのキャプション生成. 情報処理学会第 81 回全国大会講演論文集. March 2019, pp. 2-407 to 2-408, entire text, all drawings, non-official translation (IWATSUKI, Michio et al. Listen and Tell: Audio Scene Caption Generation using Deep Learning. Proceedings of the 81st National Convention of IPSJ.)	1-7
A	JP 2011-250100 A (SONY CORP.) 08 December 2011, entire text, all drawings	1-7
A	US 2018/0061439 A1 (DIAMOS, Gregory Frederick et al.) 01 March 2018, entire text, all drawings	1-7

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/JP2020/018823

Patent Documents referred to in the Report	Publication Date	Patent Family	Publication Date
JP 2011-250100 A	08.12.2011	(Family: none)	
US 2018/0061439 A1	01.03.2018	(Family: none)	

A. 発明の属する分野の分類（国際特許分類（IPC）） G10L 15/10(2006.01)i; G10L 25/51(2013.01)i FI: G10L15/10 500Z; G10L25/51		
B. 調査を行った分野		
調査を行った最小限資料（国際特許分類（IPC）） G10L13/00-25/93		
最小限資料以外の資料で調査を行った分野に含まれるもの 日本国実用新案公報 1922-1996年 日本国公開実用新案公報 1971-2020年 日本国実用新案登録公報 1996-2020年 日本国登録実用新案公報 1994-2020年		
国際調査で使用した電子データベース（データベースの名称、調査に使用した用語）		
C. 関連すると認められる文献		
引用文献の カテゴリー*	引用文献名 及び一部の箇所が関連するときは、その関連する箇所の表示	関連する 請求項の番号
X	IKAWA, Shota et al., Neural Audio Captioning Based on Conditional Sequence-to-Sequence Model, Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop, 2019.10, pp.99-103 特に3.1節, 図1, 表2	1-3, 6-7
A		4-5
A	KIM, Chris Dongjoo et al., AudioCaps: Generating Captions for Audios in The Wild, Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technology, 2019.06, Vol. 1, pp.119-132 特に4章, 図4	4-5
A	岩月道生ほか, Listen and Tell: 深層学習を用いた音響シーンのキャプション生成, 情報処理学会第81回全国大会講演論文集, 2019.03, pp.2-407~2-408 全文, 全図	1-7
A	JP 2011-250100 A (ソニー株式会社) 08.12.2011 (2011-12-08) 全文, 全図	1-7
☒ C欄の続きにも文献が列挙されている。 ☒ パテントファミリーに関する別紙を参照。		
* 引用文献のカテゴリー “A” 特に関連のある文献ではなく、一般的技術水準を示すもの “E” 国際出願日前の出願または特許であるが、国際出願日以後に公表されたもの “L” 優先権主張に疑義を提起する文献又は他の文献の発行日若しくは他の特別な理由を確立するために引用する文献（理由を付す） “O” 口頭による開示、使用、展示等に言及する文献 “P” 国際出願日前で、かつ優先権の主張の基礎となる出願の日の後に公表された文献 “T” 国際出願日又は優先日後に公表された文献であって出願と抵触するものではなく、発明の原理又は理論の理解のために引用するもの “X” 特に関連のある文献であって、当該文献のみで発明の新規性又は進歩性がないと考えられるもの “Y” 特に関連のある文献であって、当該文献と他の1以上の文献との、当業者にとって自明である組合せによって進歩性がないと考えられるもの “&” 同一パテントファミリー文献		
国際調査を完了した日 02.09.2020		国際調査報告の発送日 15.09.2020
名称及びあて先 日本国特許庁(ISA/JP) 〒100-8915 日本国 東京都千代田区霞が関三丁目4番3号		権限のある職員（特許庁審査官） 中村 天真 5Z 1786 電話番号 03-3581-1101 内線 3591

C. 関連すると認められる文献		
引用文献の カテゴリー*	引用文献名 及び一部の箇所が関連するときは、その関連する箇所の表示	関連する 請求項の番号
A	US 2018/0061439 A1 (DIAMOS, Gregory Frederick et al.) 01.03.2018 (2018 - 03 - 01) 全文, 全図	1-7

国際調査報告
パテントファミリーに関する情報

国際出願番号

PCT/JP2020/018823

引用文献			公表日	パテントファミリー文献	公表日
JP	2011-250100	A	08.12.2011	(ファミリーなし)	
US	2018/0061439	A1	01.03.2018	(ファミリーなし)	