TasteColorizer: 既存の映像メディアを「味わえる映像」にするシステム

本間 大一優[1,a)]　宮下 芳明[1]

概要：人類は映像技術が発明されて以来，様々な映像を記録してきた．近年，味覚メディアの発達により，映像とともに味を記録し再生することが可能になったが，これまで記録されてきた映像に対して味を付与することはできない．本稿では，白黒映像をカラー化する技術のように，味が記録されていない映像に対しても，味を推定し付与するシステムを提案する．味は GPT4-Vision により映像の全シーンに対して推定される．プロンプトでは，飲食物の名前，材料とその量をまず推測し，その後に味を推定するように指示しており，段階的な推論によって精度の高い推定が可能である．また，推定された味データは映像の各ピクセルに対応する形でファイルに記録される．視聴者は，映像上の飲食物を指定して，対応する味を出力し味わうことができる．本システムによって，既存の映像メディアを「味わえる映像」にすることができる．

1. はじめに

19 世紀後半，人類は連続的に写真を撮影することで，映像を記録できるようになった．当初は白黒映像しか記録できなかったが，20 世紀になってカラー映像を記録できるようになった．映像に色がついているか否かによって，撮影された年代を推定できるほどであったが，AI による Colorization 技術を用いて白黒時代の映像にも着色がなされるようになった．

本稿第二著者は味覚と視聴覚を一緒に記録し編集できる仕組みを提案している [1] が，それ以前の映像に味を付与することは難しい．また，味センサで味を測ること自体が大変であるため，過去の映像に味覚情報を付与する抜本的な提案には至っていない．本稿第二著者が提案した TTTV3 には，画像を入力として AI が味を推定し出力する機能はあるが，対話システムとしての提案にとどまっている [2]．

本稿では，AI による味推定を動画の全シーンに対して行い，推定された味データを動画ファイルに付与する仕組みを提案する．味データと映像は，MP4 を拡張したファイル形式に記録されるため，従来の動画ファイルのようにオフライン環境や，スペックが低い環境でも再生が可能である．また，味データは映像の各ピクセルに対応する形で書き込まれるため，視聴者は映像上の飲食物を指定して，対応する味を出力できる．本システムによって，既存の映像メディアを「味わえる映像」に変換することができる．

[1] 明治大学
[a)] ev220526@meiji.ac.jp

2. 関連研究

本稿第二著者は，視聴覚と味覚を合わせて記録し・編集・再生できる仕組みを提案した [1]．視聴覚はビデオカメラで，味覚は味センサを用いて記録している．しかし，味センサはどれも高価，巨大で使用にもノウハウが必要であり，味の記録を簡単にはできないことが課題であった．編集ソフトウェアでは，複数の動画ファイルを読み込み，それぞれの動画に対する基本五味の分布を設定することができ，視聴覚に味覚を合わせたコンテンツの編集が可能である．

本稿第二著者は，このような味覚を記録・再生する技術に端を発して味覚メディアを考察し，推進をしている．また，視聴覚メディアに起こった，CGM や N 次創作といったムーブメントによる表現の民主化が味覚メディアにおいても起こると推測している [3]．

TTTV3 では, 画像を入力として大規模言語モデル（LLM）を用いて味を推定する手法が提案されている [2]．提案手法では，Bing Search API を用い入力画像の類似画像を検索し，検索結果の要約を入力として LLM に味を推定させている．トマトソースパスタの画像を入力した事例が挙げられており，推定システムは画像がトマトソースパスタであることを理解し，さらにトマトの酸味を表現することができていた．LLM が味覚推定において有用である可能性が示されている．

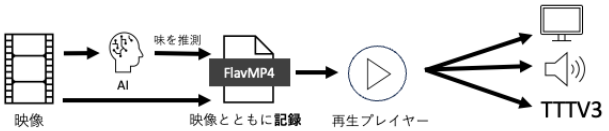

図 1　提案システムの概念図

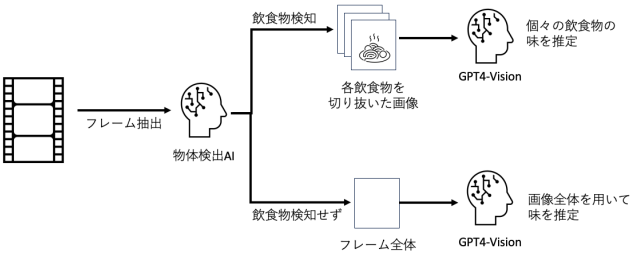

図 2　味推定の流れ

3.　提案システム

システムの概念図を**図 1** に示す．まず，AI が映像に対してどこに飲食物が映っているかを一定間隔で判定し，それぞれの飲食物に対して味を推定する．推定された味の時系列データは，映像や音声とともに MP4 を拡張したファイル形式「FlavMP4」に書き込まれる．再生プレイヤーは FlavMP4 から映像，音声，味データをそれぞれ読み込み出力することで再生を行う．FlavMP4 ファイルには映像のピクセルに対応する形で味データを書き込むことが可能で，視聴者は再生時に任意の映像上の場所を選んで味を出力することができる．

また本稿では，FlavMP4 に記録された味データを編集できるソフトウェアも試作した．AI が提案した味の再現性が不十分なときは，人間が直接調整することも可能である．AI によって推測された味と映像をファイルとして記録し再生できるようにすることで，従来の映像から「味わえる映像」の生成を可能にしている．

3.1　AI による映像の味推定

映像から味を推定する流れを**図 2** に示す．まず，映像から推定対象となるフレームを抽出し，フレーム画像を物体検出 AI に入力する．検出 AI によって飲食物が検出された場合は，フレームから各飲食物を切り抜き，個々に GPT4-Vision[4] を用いて味を推定する．飲食物が検出されなかった場合は，フレーム全体を入力として GPT4-Vision に味を推定させる．GPT4-Vision が飲食物を確認できなかった場合は無味（ゼロベクトル）を推定値とする．この推定処理を一定フレーム間隔で映像に対して行うことで，映像に合った味の時系列データが推定される．

3.1.1　物体検出 AI の利用

GPT4-Vision を用いて物体の正確な位置を求めることは難しいため，一度物体検出 AI を用いて飲食物の位置を検出している．物体検出 AI には，Zero-Shot テキスト条件付き物体検出モデルである OWL-ViT[5] を使用している．OWL-ViT では，テキストによって検出する物体を条件づけすることができ，本システムでは「a photo of food」で条件づけをしている．検出結果に含まれる物体の位置情報は，味データを記録する際に，映像内のピクセルと対応付けるために利用される．

3.1.2　GPT4-Vision による味推定

TTTV3 の画像入力による味推定機能 [2] では，LLM に GPT-4[4] を使用している．しかし，GPT-4 は画像の入力に対応しておらず, Bing Search API を通して画像をテキスト情報に変換する形で画像に関する情報を与えていた．この過程で味推定に影響を与える視覚情報（具材，色，形等）が失われる問題があった．本システムでは，GPT4-V[6] を利用することでこの問題にアプローチした．GPT4-V は，膨大なテキストデータと画像データのペアを学習した視覚言語モデルであり，様々なタスクにおいて視覚情報を使用した推論が可能になっている．料理についても一般的な知識を持っていることが示されている [7].

プロンプトでは, 入力された画像について味を推定し json 形式で結果を返すように指示している．味は，水 200ml に食塩，クエン酸，グルタミン酸ナトリウム，スクロース，炭酸カリウムをそれぞれ何グラムいれるかという形で推定するように指示している．精度の向上を狙い，モデルに段階的に推論を行わせる手法，CoT（Chain Of Thought）[8] を取り入れている．具体的には，飲食物の名前と味，使われている材料とその量についてまず推測し，それらを基に味を推定するように指示している．また，映像の文脈を理解できるように，今まで検出された飲食物の名前をプロンプトに挿入し，その情報も含めて推論するように指示をしている．特定の飲食物に特化した調整を施していないため，様々な種類の飲食物に対して推測が可能である．実際に用いたプロンプトは**図 A·1** に掲載している．

3.2　FlavMP4

推定された味データは映像とともに独自のファイル形式「FlavMP4」に書き込まれる．FlavMP4 は，MP4 ファイルフォーマットを拡張したファイル形式であり，映像データとともに味データを記録することができる．味データは映像データの各ピクセルと対応させることができる．ファイルとして記録することで，オンライン環境や高いマシンパワーを必要とせずに遅延なく再生が可能となる．さらに，MP4 を拡張したファイル形式であるため，映像と音声は従来のプレイヤーで再生可能であり，ランダムアクセスや可変フレームレートも可能である．FlavMP4 は，Flavor と MP4 をかけ合わせた言葉であり，視聴覚と味覚を合わせて「風味」を提示できることを示唆している．

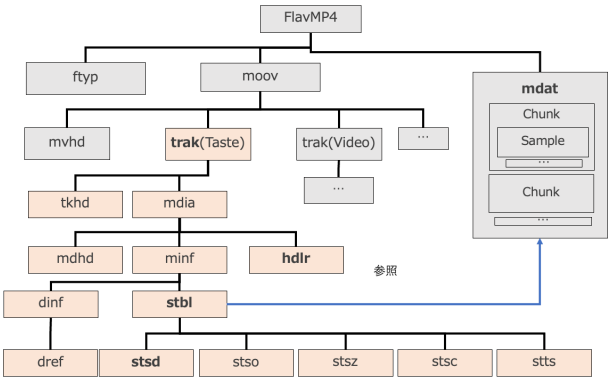

図 3 FlavMP4 のファイル構造．オレンジ色のノードが FlavMP4 で MP4 に対して新たに付け加えられている

3.2.1 ファイル構造

FlavMP4 の基本的なファイル構造の例を**図 3** に示す．FlavMP4 のファイル構造は，MP4 の基となっている ISO Base Media File Format[9] をベースとしている．FlavMP4 は，MP4 と同じくボックスと呼ばれる単位で情報を管理し，それらを入れ子構造にしていくことで情報を記録している．図 3 の各ノードはボックスを示している．trak ボックスでは各メディアデータに対応する様々なメタデータを格納しており，映像，音声等異なるメディアデータごとに 1 つずつ存在する．FlavMP4 では，味データに関するメタデータを管理する trak ボックス（図 3，オレンジ色のノード）を新たに定義し加えることで，味データの記録と再生を可能にしている．

3.2.2 味の符号化方式

味データは，TTTV3 などの液体混合式のデバイスで再生できるように符号化されている．TTTV3 では，より細かな風味の違いを再現するために，20 種類の溶液を搭載することができるようになっている．例えば酸味の表現には，リンゴ酸，クエン酸，酢酸，乳酸を使い分けている．FlavMP4 では，このような細かな風味の違いも再生できるようにするため，溶液の種類を記録できるようにしている．

味データ本体は，符号なし 8bit 整数（0〜255）として記録される．動画と同じくフレーム単位で管理され，1 フレームには各ピクセルに対応する味データが記録されている．ファイル内には，溶液の濃度（100ml 中何 g 溶質が含まれているか）と，溶液の最大混合量（一度に混合する溶液の最大量）が記録されており，味データは溶液の最大混合量に基づいて変換される．具体的には，味データの値を最大混合量で乗じ，その結果を 256 で割ることで，最大混合量に対する味データの相対的な値を求めている．この手法によって，少ないデータ量でより大きな範囲と細かい精度でデータを表現できる．また，溶液の濃度を記録することで，溶液の濃度が異なるデバイスで再生する際に再生プレイヤー側で補正をすることが可能になっている．

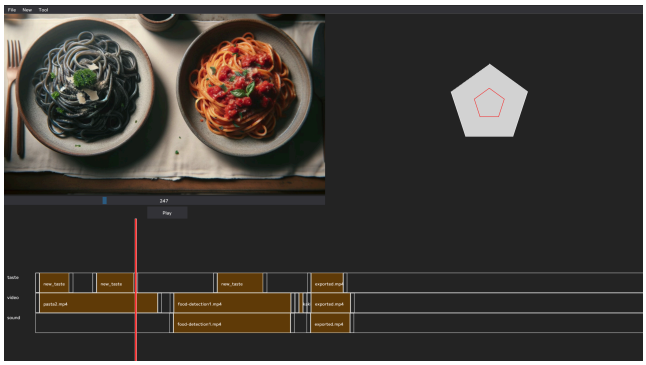

図 4 編集ソフトウェアの編集画面．上には映像と味のプレビュー画面，下にはタイムラインがある

3.3 再生プレイヤー

本稿では，FlavMP4 ファイルを再生できる再生プレイヤーも合わせて試作した．映像と音声は通常のプレイヤーと同じように再生されるが，映像上のピクセルを指定することで，そのピクセルに対応した味を出力することができる．視聴者は，映像の中で味わいたい飲食物があったら一時停止ボタンを押し，飲食物をクリックすることでその味を出力して味わうことができる．

3.4 編集ソフトウェア

本稿ではさらに，FlavMP4 を読み込み，編集し，書き出す事ができる編集ソフトウェアを試作した．編集ソフトウェアを用いることで，AI が提案した味の不十分な点を手で細かく修正することができる．また，修正だけでなく任意の新しい味を追加することも可能である．編集画面を**図 4** に示す．ユーザインタフェースには動画編集ソフトウェアで広く用いられるタイムラインを取り入れており，ユーザは従来の動画編集と同じ感覚で味覚コンテンツの編集ができる．タイムラインには映像トラック，音声トラックに加えて味トラックを新しく追加している．編集した結果は，再度 FlavMP4 ファイルとして書き出す形で共有できる．また，他人が共有した FlavMP4 ファイルを編集ソフトウェアに読み込み，自分好みに編集し再び書き出して共有することで，コンテンツのリミックスが可能である．

4. 事例と検証

4.1 複数の食品が映っている映像

本システムを用いて，イカスミパスタとトマトソースパスタが並んだ映像（**図 5** 上）を入力として味を推定した．結果，図 5 下の 2 つの矩形で囲まれた範囲を飲食物として検出し，それぞれ正しくイカスミパスタとトマトソースパスタであると認識できた．味の推定結果は，イカスミパスタは水 200ml に対して，食塩 0.8g，クエン酸 0.1g，グルタミン酸ナトリウム 0.9g，スクロース 0.2g であり，トマトソースパスタは食塩 0.3g，クエン酸 0.5g，グルタミン酸ナトリウム 0.2g，スクロース 0.2g であった．

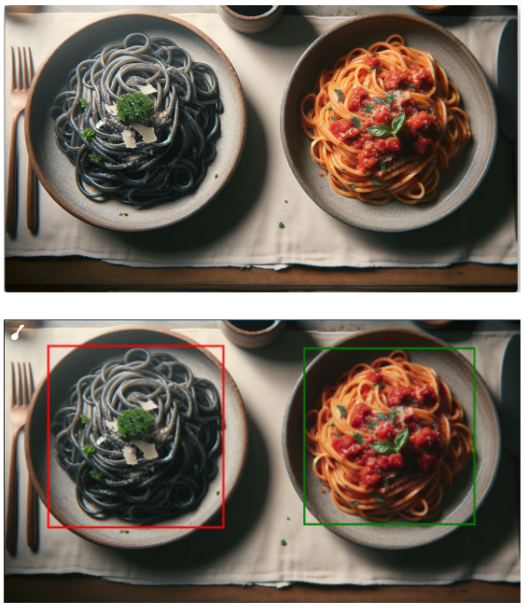

図 5　イカスミパスタとトマトソースパスタの映像（上）と画像検出 AI による検出結果（下）

図 6　コーヒーにミルクを入れていく映像

推定された味を実際に出力し味わってみたところ，イカスミパスタは旨味と塩味を強く感じ，トマトソースパスタは強い酸味に加えて塩味と旨味を感じた．イカスミ特有の旨味の強さと，トマト特有の酸味の強さをうまく表現できているように感じた．両者は完全に味を再現できているわけではないが，特徴を捉えた味の違いをはっきりと提示できていた．また，トマトソースパスタについては，チーズがかかっていることを GPT4-Vision が考慮し，塩味と旨味を加えていた．推定された結果は，図 5 下の 2 つの矩形内にそれぞれ埋め込まれるため，再生時に矩形内をクリックすることでその味を味わうことができる．

4.2　味が変化する映像

4.2.1　コーヒーにミルクを入れる映像

カップにコーヒーを入れ，その後ミルクを入れる映像（**図 6**）に対して味推定を行った．結果，コーヒーを入れていないとき（図 6 左）には無味が推定され，コーヒーを入れた後（図 6 中）では，食塩 0g，クエン酸 0.1g，グルタミン酸ナトリウム 0.05g，スクロース 0g，炭酸カリウム 0.2g と推定された．コーヒーにミルクを入れた後（図 6 右）では食塩 0g，クエン酸 0g，グルタミン酸ナトリウム 0.1g，スクロース 0.5g，炭酸カリウム 0.1g と推定された．

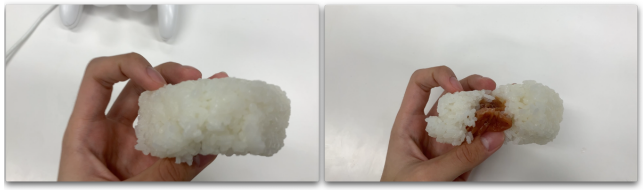

図 7　梅干し入りのおにぎりを食べていく一人称映像

両者を実際に出力し味わってみたところ，コーヒーの推定値では，苦みと酸味をしっかりと感じることができた一方，ミルクコーヒーの推定値では苦みが抑えられており，甘みを感じることができた．ミルクコーヒーやコーヒーの味を完全に再現できてはいないが，ミルクを入れたことによる苦みの抑制と甘みの増加をうまく推定できている．

また，ミルクを入れた後の画像単体で入力すると，GPT4-Vision はミルクティーであるとして推定したが，映像として入力するとミルクコーヒーであるとして推定できた．これは，3.1.2 項で述べた，今まで検出された飲食物を踏まえて推定させる手法によって，映像の文脈を理解できたためと考える．

4.2.2　梅入りおにぎりを食べていく一人称映像

梅干しが入ったおにぎりを食べ，食べていくうちにおにぎりの中から梅干しが現れる一人称映像（**図 7**）に対して味推定を行った．結果，おにぎりの映像（図 7 左）では，食塩 0.3g，クエン酸 0g，グルタミン酸ナトリウム 0.1g，スクロース 0g，炭酸カリウム 0g と推定された．おにぎりを食べ進め，梅干しが現れた映像（図 7 右）では，梅干し入りのおにぎりと認識され，食塩 0.4g，クエン酸 0.8g，グルタミン酸ナトリウム 0.5g，スクロース 0.1g，炭酸カリウム 0g と推定された．

実際に出力し味わってみたところ，おにぎりの推定値は塩むすびのような塩味と旨味を感じた．一方，梅干し入りのおにぎりの推定値には強い酸味を感じたが，酸味だけでなく旨味も感じることができた．おにぎりを食べていき，梅干しに達したときの味変化をうまく提示できているように感じた．連続的変化を提示できる味ディスプレイを出力装置として用いれば，梅干しが現れた瞬間に酸味が現れるという体験が可能になるだろう．

4.3　食品が映った最古の映像

カラーの映像でなくても，飲食物を推定できるような情報が含まれていれば味を付与することができる．食品が映った最古の映像と思われる，1895 年に公開されたリュミエール兄弟による白黒サイレント映画 Repas de bébé（Feeding the Baby）[10] の 1 シーン（**図 8**）に対して推定を行った．結果，紅茶が写っていることを認識し，紅茶の苦みとミルクや砂糖による甘みを映像から推測した．推

図 8 Repas de bébé（Feeding the Baby）（1895）の 1 シーン

定値は，食塩 0g，クエン酸 0g，グルタミン酸ナトリウム 0g，スクロース 0.5g，炭酸カリウム 0.1g であった．実際に味わってみたところ若干の苦味と甘みを感じることができた．19 世紀の白黒無音映像についても，映像に合った味を蘇らせて味わうという体験をすることができた．

5. おわりに

本稿では，AI を用いて映像から味を予測し，映像と味を一緒に記録し再生することができるシステムを構築した．これにより，どのような時代の映像でも映像を「味わえる映像」として蘇らせることを可能にした．AI の推定結果は，完全に味を再現できてはいないが，人間のイメージにある程度合った味と味変化を推定することができた．

AIが推定した味の再現性の不十分さについては, YouTube のような FlavMP4 のオンライン配信システムを作り，Wikipedia のように味がおかしいところを視聴者によって編集できるようにすることで解決できると考える．また，映像から予測する味というのは，人により異なり，人それぞれの好みが現れると予想される．視聴者が自分好みに味を編集することで，１つの映像から味の異なる「味わえる映像」が，無数に派生していくだろう．

映像は，早送り・巻き戻しができる．速度も変えられる．スキップもできる．逆再生もできる．いわば，映像操作というのは，自在に時間を操る体験にほかならない．たとえば，コーヒーにミルクが拡散しミルクコーヒーとなるのは，いわゆる「時間の矢」とよばれる不可逆過程である．我々の物理世界はエントロピー増大の法則を超えられない．しかし，映像世界は，これを逆再生することで不可逆な法則を容易に超えることができる．本稿は，映像と味を合わせる単純な仕組みの提案に思えるかもしれないが，これによって将来，我々の味覚に関する時間感覚が解き放たれる可能性があると筆者らは考えている．

参考文献

[1] 宮下芳明：画面に映っている食品の味を再現して味わえる味ディスプレイの開発，第 28 回インタラクティブシステムとソフトウェアに関するワークショップ (WISS2020) 論文集，pp. 103–108 (2020).

[2] 宮下芳明, 村上崇斗, 大友千宙, 深池美玖：TTTV3 (Transform The Taste and reproduce Varieties): 産地や品種の違いも再現する調味機構と LLM による味覚表現，エンタテインメントコンピューティングシンポジウム論文集，Vol. 2023 (2023).

[3] 宮下芳明：TTTV2 (Transform The Taste and Visual appearance): 飲食物の味と見た目を変える調味家電によるテレイート，エンタテインメントコンピューティングシンポジウム 2022 論文集，Vol. 2022, 情報処理学会，pp. 143–150 (2022).

[4] OpenAI: GPT-4 Technical Report (2023).

[5] Minderer, M., Gritsenko, A., Stone, A., Neumann, M., Weissenborn, D., Dosovitskiy, A., Mahendran, A., Arnab, A., Dehghani, M., Shen, Z., Wang, X., Zhai, X., Kipf, T. and Houlsby, N.: Simple Open-Vocabulary Object Detection with Vision Transformers (2022).

[6] OpenAI: GPT-4 Vision System Card (2023).

[7] Yang, Z., Li, L., Lin, K., Wang, J., Lin, C.-C., Liu, Z. and Wang, L.: The Dawn of LMMs: Preliminary Explorations with GPT-4V(ision) (2023).

[8] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q. and Zhou, D.: Chain-of-Thought Prompting Elicits Reasoning in Large Language Models (2023).

[9] International Organization for Standardization: ISO Base Media File Format, ISO/IEC Standard 14496-12, International Organization for Standardization (2022).

[10] Lumiere, A. and Lumiere, L.: Repas de bébé (Feeding the Baby), Film (1895).

付　　録

A.1 GPT4-Vision による味推定に使用したプロンプト

「写真に写っている飲食物」の味と同じ塩味・酸味・うま味・甘味・苦みに感じられるように，水 200ml に食塩・クエン酸・グルタミン酸ナトリウム・スクロース．炭酸カリウムを入れたい．なお，食塩は塩味，クエン酸は酸味，グルタミン酸はうま味，砂糖は甘味，苦みは炭酸カリウムを用いて表現せよ．苦みについては，どうしても苦みを表現したい場合のみ炭酸カリウムを入れてもよい．各物質を何グラム入れればいいかを記述せよ．提供される写真はある映像の一部であり，これまで映像には「検出された飲食物の名前」の順番に飲食物が登場してきました．これらの映像の流れから提供される文脈も考慮に入れなさい．はっきりと何の飲食物とかわからない場合も，周りの環境からできるだけ味を推定するようにしてください．
なお，推論結果は，以下の json 形式で返してください．また，画像が不鮮明であったり飲食物が確認できない場合は，{"satus" : 1} と返答しなさい．
json 形式:

```
{
"status":0,
"message":<なにかメッセージがあればここに>
"name":<写真に写っている飲食物の名前>,
"description":<飲食物とその味の説明（塩味・酸味・うま味・甘味・苦みについてそれぞれ分析）>,
"ingredients": [
    { "name": <飲食物を作る際の材料の名前> ,
  "amount": <予想される量>,},...
],
"inference": {
    "食塩" : <以上の情報を踏まえて，再現する際の水 200ml に含むべき食塩の量を推測する．単位は g(g は含まなくて良い, float)>,
    "クエン酸" : <水 200ml に含むべきクエン酸の量>,
    "グルタミン酸ナトリウム" : <水 200ml に含むべきグルタミン酸ナトリウムの量>,
    "砂糖" : <水 200ml に含むべき砂糖の量>,
    "炭酸カリウム" : <水 200ml に含むべき炭酸カリウムの量>
    }
}
```

返答はすべて json 形式とする．余計な忠告は必要ない．聞き返すことは認めない．返答は日本語とする．あなたの返答はそのまま JSON パースされる．JSON パースエラーが起きないようにせよ．

図 A·1 プロンプト