# PhotoSpeak

Bocicov Vitalie, Badelita Tiberiu, Soltan Eduard

May 14, 2023

**Abstract**

PhotoSpeak is a cloud computing project that empowers language learners to learn a new language or multiple new languages by uploading photos of objects they encounter in their daily lives. Leveraging Google Vision and Google Translation API, PhotoSpeak provides users with labels translated into different languages, as well as audio pronunciation using text-to-speech and speech-to-text technology. Users can also create accounts to access the full features of the platform, including the ability to upload photos of anything, select languages for translation and audio pronunciation, and see a history of uploaded photos.

## 1 Introduction

Learning a new language is a valuable skill that can open up new opportunities and improve communication with people from different cultures. According to a Eurobarometer survey on attitudes towards foreign languages and multilingualism, nearly half of Romanians say they can have a conversation in a foreign language, English being the most used. The backlash, however, is that only 20% of Romanians use a foreign language during holidays, compared to the European average, i.e. of 50%.

Romanians mostly make use of a foreign language only to watch movies or TV programs and listen to radio. 68% of Romanians, compared to only 44% of Europeans, still prefer subtitled films. While English is the foreign language that most Romanians speak – 31%, followed by French – 17% and German and Spanish – 3%, unfortunately, the results in this area are below the European average.

Internationally, language learning is also a popular activity, with over 1.5 billion people worldwide learning a foreign language. The most commonly studied language is English ( 1.5 billion), followed by French (120 million), Mandarin Chinese (25 million), Spanish (18 million) and German (15 million).

Despite the popularity of language learning, many people struggle to apply traditional language learning methods to their daily lives. PhotoSpeak aims to address this challenge by providing a practical and engaging way for language learners to learn a new language or multiple new languages by uploading photos of anything they encounter in their daily lives.

The project uses Google Vision and Google Translation API to provide users with labels translated into different languages and audio pronunciation using text-to-speech and speech-to-text technology. Users can create accounts to access the full features of the platform, including the ability to upload photos of anything, select languages for translation and audio pronunciation, and see a history of uploaded photos. In this case study, we will provide an overview of the project, including its market analysis, technology, implementation, and results, as well as its impact on language learning in Romania and internationally.

## 2 Technical details about existing solutions

The best-known existing solutions that come close to the solution proposed by our team would be: Google Translate, Word Lens, Duolingo, Rosetta Stone. In the following part, we will write about technical details such as the architecture of the application, technologies and marketing approaches used by the applications mentioned above.

## 2.1 Google Translate

### 2.1.1 Architecture

Google Translate follows a distributed architecture, leveraging Google's vast infrastructure and cloud services. It incorporates a combination of machine learning models, natural language processing algorithms, and neural networks to provide translation services. The architecture includes components such as the translation engine, language detection, and user interface modules. It utilizes Google Cloud Vision's image recognition technology for photo analysis and text extraction from images.

### 2.1.2 Technologies

1.Machine Learning: Google Translate employs machine learning techniques, including neural machine translation (NMT) models, to improve translation quality and accuracy.

2.Natural Language Processing (NLP): NLP algorithms are used to analyze and process text input, identify language patterns, and improve translation results.

3.Optical Character Recognition (OCR): Google Translate utilizes OCR technology to extract text from images, enabling photo-based translations.

4.Cloud Services: Google Translate leverages Google Cloud Platform (GCP) services to handle massive translation workloads, ensuring scalability, reliability, and high performance.

### 2.1.3 Marketing Approaches

1.Online Advertising: Google promotes Google Translate through online advertising platforms, including Google Ads. It displays ads on search engine result pages, websites within the Google Display Network, and other relevant online platforms to reach a wide audience. Ad campaigns are targeted based on keywords, demographics, and user behavior to ensure maximum visibility.

2.App Store Optimization: Google Translate maintains a strong presence in mobile app stores, such as the Apple App Store and Google Play Store. It utilizes app store optimization techniques to improve its visibility in search results within the app stores. This involves optimizing app titles, descriptions, keywords, and screenshots to attract more downloads and users.

3.Partnerships: Google collaborates with various partners, such as smartphone manufacturers, to preinstall Google Translate on devices. This strategic partnership approach expands the reach of Google Translate, making it readily available to users without requiring separate installation or downloads.

4.Word-of-Mouth and User Advocacy: Google Translate benefits from positive word-of-mouth marketing as satisfied users recommend it to others. Users share their positive experiences using Google Translate, especially when it helps them overcome language barriers during travel, communication, or language learning. User testimonials and reviews contribute to building trust and credibility for the service.

## 2.2 Word Lens

### 2.2.1 Architecture

Word Lens utilized a client-server architecture. The client was the mobile application installed on users' devices, while the server handled the heavy processing required for optical character recognition (OCR) and translation. The client app captured live video using the device's camera and sent it to the server for analysis and translation. The server processed the video frames, performed OCR on the text, and sent back translated text overlays to be displayed on the client's screen in real-time.

### 2.2.2 Technologies

1.Optical Character Recognition (OCR): Word Lens employed advanced OCR technology to recognize and extract text from the captured video frames. It involved techniques such as image preprocessing, character segmentation, and character recognition to convert the visual text into editable and translatable content.

2.Neural Machine Translation (NMT): Word Lens used NMT models to translate the recognized text into the desired language. NMT models are built on artificial neural networks and are known for providing more accurate and fluent translations compared to traditional statistical machine translation approaches.

3.Mobile Development: The Word Lens app was developed for mobile platforms, such as iOS and Android, utilizing the respective development frameworks (e.g., Swift for iOS, Java/Kotlin for Android).

### 2.2.3 Marketing Approaches

1.App Store Presence: Word Lens was promoted through app store optimization techniques, making it discoverable on platforms like the Apple App Store and Google Play Store. It focused on optimizing app titles, descriptions, keywords, and visual assets to increase visibility and attract potential users.

2.Freemium Model: Word Lens followed a freemium model where the app was available for free, allowing users to experience basic translation functionality. Additional premium features or language packs were offered as in-app purchases to generate revenue.

3.Partnerships: Word Lens sought partnerships with organizations, publishers, or travel-related companies to expand its user base and reach. Collaborations with entities that aligned with the app's purpose helped promote and integrate Word Lens into travel apps, guidebooks, or other relevant platforms.

## 2.3 Duolingo

### 2.3.1 Architecture

Duolingo follows a client-server architecture. The client side consists of the mobile application or web interface used by learners, while the server side handles the processing, storage, and delivery of lessons and user data. The server handles user authentication, progress tracking, content delivery, and exercises generation. The client app communicates with the server through APIs to retrieve lessons, submit user responses, and synchronize progress across devices.

### 2.3.2 Technologies

1.Mobile Development: Duolingo's mobile applications are developed natively for iOS and Android platforms using programming languages like Swift or Objective-C for iOS and Java/Kotlin for Android.

2.Web Development: Duolingo's web interface is developed using web technologies such as HTML5, CSS, and JavaScript, ensuring cross-platform compatibility.

3.Backend: The server-side infrastructure of Duolingo is built using various technologies, including programming languages like Python, frameworks like Django, and databases like PostgreSQL or MySQL. It also leverages cloud services and scalable architecture to handle a large number of users and deliver lessons efficiently.

### 2.3.3 Marketing Approaches

1.Gamification: Duolingo's gamified approach to language learning has been a key marketing strategy. It incorporates game-like elements such as achievements, levels, streaks, and rewards to engage users,

create a sense of progress, and make the learning experience enjoyable.

2.Viral Growth: Duolingo has relied on viral marketing, with a strong focus on word-of-mouth recommendations and user referrals. The platform encourages users to invite friends and share their achievements on social media, spreading awareness and attracting new users.

3.Free-to-Play Model: Duolingo follows a freemium model, offering basic language courses and features for free. Additional premium features, such as ad-free experience, offline access, and advanced lessons, are available through a subscription-based service called Duolingo Plus.

4.Social Integration: Duolingo allows users to connect with friends, compete in leaderboards, and share achievements, fostering a sense of community and encouraging users to engage with the app regularly. This social integration helps in user retention and organic promotion.

## 2.4 Rosetta Stone

### 2.4.1 Architecture

Rosetta Stone follows a client-server architecture. The client side consists of the language learning software installed on users' devices (e.g., computer, tablet, or smartphone), while the server side manages content delivery, user authentication, and progress tracking. The client application communicates with the server through APIs to retrieve lessons, submit user responses, and synchronize progress across devices. The server infrastructure handles the storage of language content, user data, and manages the logic for lesson progression and tracking.

### 2.4.2 Technologies

1. Web and Mobile Development: Rosetta Stone offers both web-based and mobile applications to provide language learning experiences across different platforms. These applications are developed using technologies such as HTML5, CSS, and JavaScript for web, and languages like Swift, Objective-C, Java, or Kotlin for mobile platforms.

2. Speech Recognition: Rosetta Stone utilizes speech recognition technology to evaluate and provide feedback on users' pronunciation and speaking skills. This technology involves converting spoken words into text and analyzing the accuracy of pronunciation.

3. Multimedia Integration: The software incorporates multimedia elements, such as audio recordings, images, and interactive exercises, to enhance the learning experience. These multimedia components are seamlessly integrated into the lessons to facilitate language acquisition.

### 2.4.3 Marketing Approaches

1.Online Advertising: Rosetta Stone utilizes online advertising channels, including display ads, search engine marketing (SEM), and social media advertising, to reach its target audience. They create targeted campaigns to showcase their products and generate interest among potential customers.

2.Subscription: Rosetta Stone has primarily offered its language learning programs through a paid subscription or one-time purchase model, where users gain access to the full range of content and features upon payment.

3.Educational Institutions: Rosetta Stone collaborates with schools, colleges, and universities to offer language learning programs as part of their curriculum or as supplemental resources. These partnerships help reach a wide range of students and educators.

4.Corporate Partnerships: Rosetta Stone works with businesses and organizations to provide language training programs for their employees. This includes customized solutions, language assessments, and integration with corporate learning management systems.

5.Language Certification: Rosetta Stone partners with language certification bodies and institutions to offer preparation courses for language proficiency exams. This helps learners prepare for internationally recognized language tests and certifications.

# 3 Overview of our technologies

Both frontend and backend technologies are used in the PhotoSpeak application. Thus, in what follows, the purpose of using the technologies within this application will be described.

## 3.1 Front-end technologies

1. React - is used in our application, because follows a component based approach allowing you to break your UI into reusable and self-contained components. This modularity makes it easier to manage and maintain our codebase, as well as promote code reusability across our application. Another reason would be the fact that provides various performance optimization techniques, such as virtualization, memoization, and code splitting. React can be used to build not only web applications but also native mobile apps using frameworks like React Native.

## 3.2 Back-end technologies

We implement the back-end in Node.js and use Cloud services from Google and Microsoft.

1. Google Cloud Vision - is used in our project to extract the labels corresponding to the objects in the images uploaded or taken by the user.

2. Google Cloud Translation - is used in our project for translating labels in selected languages by the user.

3. Microsoft Azure Storage - is used in our project to store images and labels corresponding to the images. Microsoft Azure Storage is used instead of Google Cloud Storage due to low costs.

4. Microsoft Azure SQL - is used in our project to insert data into Microsoft Azure Storage and to query Microsoft Azure Storage to obtain the required data.

5. Google Cloud Text-to-Speech - is used in our project to obtain the ArrayBuffer corresponding to each individual label. From ArrayBuffer we get the sound that we play in the front-end when a label is pressed.

6. Google Cloud Speech-to-Text - is used in our project to obtain the words from an audio recording. Then we check if the words obtained from the audio recording correspond to the labels obtained from the loaded image.

7. Google Cloud Functions - is a function that runs on NodeJS runtime, and gets triggered by a http call and works as intermediate level between backend application and Cloud Storage and SQL Database.

8. Firebase - is used in our project for user registration and authentication.

# References


[1] Lexington, *Foreign Languages Spoken by Romanians*, https://www.lexington.ro/foreign-languages-spoken-by-romanians/?lang=en, n.d.

[2] Newsdle, *The Most Studied Foreign Languages in the World*, https://www.newsdle.com/blog/most-studied-foreign-languages, accessed May 14, 2023.

# The Business Model Canvas

| Designed for: | Designed by: | Date: | Version: |
|---|---|---|---|
| | | | |

## Key Partners

**Cloud computing providers**

**Advertisers**

**Education institutions**

**Influencers**

**Payment providers**

## Key Activities

**Software Development**

**Marketing and promotion**

**Partnerships and collaborations**

## Key Resources

**Software Developers**

**Cloud computing infrastructure**

**Cloud computing services**

## Value Propositions

**Free and fun language learning**

**Explore new languages with your eyes in a playful manner**

## Customer Relationships

**Mass Customers**

## Channels

**photospeak.com**

**Paid Advertising**

**Influencer Marketing**

**Education institutions**

## Customer Segments

**Web Users Globally**

**Language learners**

**Tourists**

**Language teachers and tutors**

**Multilingual professionals**

## Cost Structure

**Software Development**

**Cloud Computing Providers**

**Marketing**

**Payment processing**

## Revenue Streams

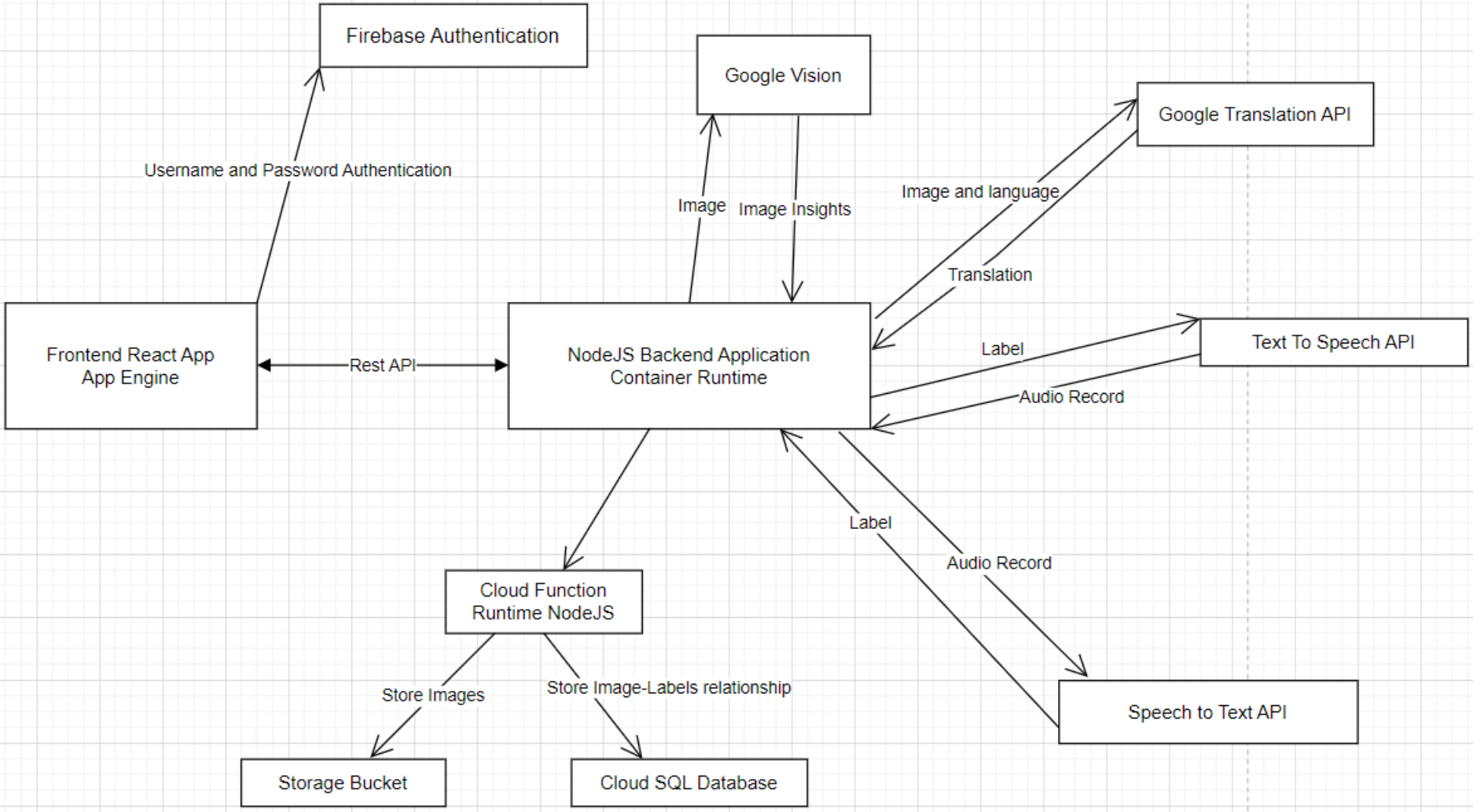
**Advertising**

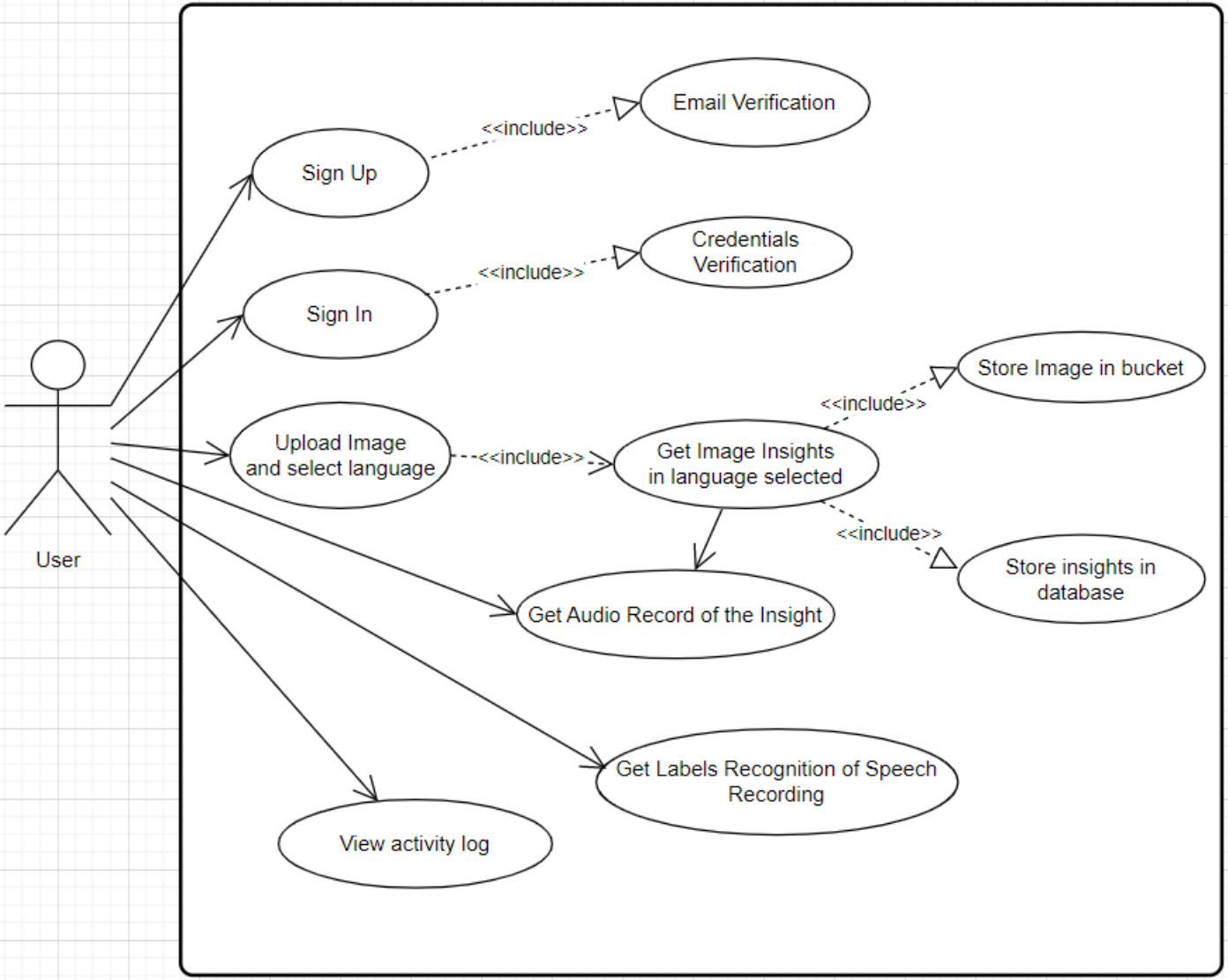
**Premium Subscription (Ad-free)**



**DESIGNED BY:** Strategyzer AG
*The makers of Business Model Generation and Strategyzer*

Strategyzer
strategyzer.com


Firebase Authentication
Google Vision
Google Translation API
Username and Password Authentication
Image
Image Insights
Image and language
Translation
Frontend React App
App Engine
Rest API
NodeJS Backend Application
Container Runtime
Label
Text To Speech API
Audio Record
Label
Audio Record
Cloud Function
Runtime NodeJS
Store Images
Store Image-Labels relationship
Speech to Text API
Storage Bucket
Cloud SQL Database


Email Verification
<<include>>
Sign Up
Credentials Verification
<<include>>
Sign In
Store Image in bucket
<<include>>
Upload Image and select language
<<include>>
Get Image Insights in language selected
<<include>>
Store insights in database
User
Get Audio Record of the Insight
Get Labels Recognition of Speech Recording
View activity log

# PhotoSpeak

1.0.0 OAS3

The API for the Photospeak - CloudComputing Project

Servers

https://photospeak-project-dot-photospeak.lm.r.appspot.com/ - SwaggerHub API A...

## Translate

POST /translate Translate text

Parameters Try it out

No parameters

Request body required application/json

Example Value | Schema

```
{
  "text": "string",
  "language": "string"
}
```

Responses

| Code | Description | Links |
|---|---|---|
| 200 | OK | No links |
| 500 | Internal Server Error | No links |

## What-Is

POST /what-is Get labels from a photo and translate them

Parameters Try it out

No parameters

Request body required multipart/form-data

photo
string(binary)

language
string

Responses

| Code | Description | Links |
|---|---|---|
| 200 | OK | No links |
| 400 | Bad Request | No links |
| 500 | Internal Server Error | No links |

## UserPhotos

GET /getUserPhoto Get user's photos

Parameters Try it out

| Name | Description |
|---|---|
| username<br>string<br>(query) | username |

Responses

| Code | Description | Links |
|---|---|---|
| 200 | OK | No links |
| 404 | Not found | No links |
| 501 | Internal Server Error | No links |

## TextToSpeech

GET /speech Convert text to speech

Parameters Try it out

| Name | Description |
|---|---|
| text<br>string<br>(query) | text |
| lang<br>string<br>(query) | lang |

Responses

| Code | Description | Links |
|---|---|---|
| 200 | OK | No links |
| 500 | Internal Server Error | No links |

## SpeechToText

POST /speech Check speech pronunciation

Parameters Try it out

No parameters

Request body required application/json

Example Value | Schema

```
{
  "text": "string",
  "language": "string",
  "speech": "string"
}
```

Responses