One of my goals to learn computer science is to implement a translator that can translate PutongHua to ChaoshanHua. In the project of my C++ course, when I was a freshman, I implemented a prototype that translates PutongHua to characters by [Xunfei]() API and then converted the characters to ChaoshanHua according to the recorded audio of each word (Lol, how naïve it was). It's far from satisfactory because our program doesn't consider the semantics of the language! Due to its complexity, I postponed this project and continued my studies in computer science. Thanks to the deep learning I learned under the supervision of my advisor Professor Si Wu, I am reconsidering this project with serious commitment!

I am from the Chaoshan region, Guangdong. There is a dialect called ChaoshanHua widely spoken in the Chaoshan region. ChaoshanHua is significantly different from Putonghua (i.e., a widely-accepted, standard language spoken in China) due to its isolation geologically and cultural traditions. For example, PutongHua is believed as one of the most challenging languages because it's based on hieroglyphs, and it contains 21 consonants, 13 vowels and four tones. While ChaoshanHua has no characters, it contains even 18 consonants, 64 vowels, and eight tones! It's insanely difficult for non-local residents to learn how to speak ChaoshanHua. Indeed, there are dialects everywhere around the world. However, ChaoshanHua, with more than 30 million users, brought negligible negative impacts to the Chaoshan Region, economically and culturally. Honestly, the fundamental reasons for its complexity are as follows:

- Geologically, the Chaoshan region is located in the mountainous area where communities and villages are separated by mountains. In ancient times, people hardly communicated with outsiders due to geological inconvenience. It's discovered that many ancient pronunciations that date back to Qin Dynasty (200 B.C.) are found in ChaoshanHua!

- Culturally: The geological isolation exacerbates the cultural isolation of the Chaoshan region. People married, commercialized with local residences, which endows the anti-diversity social attitude. The undeveloped situation of the Chaoshan region is related to the anti-diversity attitude. When I was young, adults believed it was a shame to marry a non-Chaoshanese.

- Ideologically: It was pretty feudal in the Chaoshan region 30-40 years ago; For example, women are incredibly less literate than men, and family culture is widely accepted. It was believed that a qualified woman should take good care of the family.

By reading so far, you may understand my feelings about my hometown. Even though we have been developing in recent fifteen years along with the development of China, many people, especially middle-aged women, are unable to recognize the PutongHua. They can't listen to the news, read the book, surf the internet or use smartphones, most of which are non-separable parts of our daily lives. They are unwelcomed to the new era. As a typical woman in the Chaoshan region, my mom took good care of her four children, and we are all college students now. But she remains illiterate as many women out there do.

One of my goals to learn computer science is to implement a translator that can translate PutongHua to ChaoshanHua. In the project of my C++ course, when I was a freshman, I implemented a prototype that translates PutongHua to characters by [Xunfei]() API and then converted the characters to ChaoshanHua according to the recorded audio of each word (Lol, how naïve it was). It's far from satisfactory because our program doesn't consider the semantics of the language! Due to its complexity, I postponed this project and continued my studies in computer science. Thanks to the deep learning I learned under the supervision of my advisor Professor Si Wu, I am reconsidering this project with serious commitment!

Therefore, this blog will show you how this project is working on!

This blog will introduce how I prepare a website [PT-CS-Dataset](<https://demo.guanyueli.com/model/pt_cs_translator>) for collecting audio translations from Chaoshanese. The header will display the number of visitors and translations. By two days after the website was published, the header displayed that there were already 23 visitors and 414 translations, with every visitor donating about 20 translations on average.

![]()

After inputting your basic personal information, you are encouraged to input your email to receive our advancement report of the project. The checkbox informs you whether you consent to make your translations open-source or not. If you consent to participate in the open-source dataset, then not only your translations would be used for training the model but also publicized as an open-source Chaoshanese ASR dataset.

The website requires you to grant recording permission for us, so you need to check if your PCs or smartphones are capable of recording. If the button becomes green after a click, you are ready to record and contribute to our dataset.

The website will generate a random sentence for you to translate. You need to click the "Record" button before you can speak. You are required to speak fluently in a normal tone as if you were talking. You can use the "Play" button to preview your recording. The system allows you to record as many times as you like before submitting your final version. The first sentence is used for the user to get familiar with the UI so it won't be uploaded to the system. The system will upload your second and later sentences. Once you submit your translation, it will be uploaded to our dataset, and the system will return you the following sentence. You don't have to retype your information if you continue. Therefore I believe you will have a smooth recording experience. Usually, every donator records 20-50 translations on average.

About the dataset

========

I chose a Chinese Audio Speech Recognition (ASR) dataset [ST-CMDS](https://link.ailemon.net/?target=http://www.openslr.org/resources/38/ST-CMDS-20170001\_1-OS.tar.gz) for my convenience. The dataset contains approximately 100 thousand sentences and the associate Chinese translations. Therefore, I can use the sentences to prompt users to upload their Chaoshanese translations.

About the website

========

The website is based on [gunicorn](<https://gunicorn.org/>) for it's developed in Python environment. Gunicorn enables simple implementation, light server dependencies and fairly speediness.

About the User Interface Design

=======

I used the [Bulma](<https://bulma.io/>) as my CSS framework that renders beautiful and user-friendly UI design. Bulma is a free, open-source framework that provides ready-to-use frontend components that you can easily combine to build responsive web interfaces.

About the Database

=======

I used the [sqlite3]() as the database for storing the paths of uploaded translations.