Aural Language Translation with Augmented Reality Glasses

Ian N. Hovde, Forrest S. Kelley, Ryan J. Kearney, and Douglas E. Dow(✉)

Electrical and Computer Engineering, School of Engineering, Wentworth Institute of Technology, Boston, MA 02115, USA
dowd@wit.edu

Abstract. Communication is core to human activity. Communication across language barriers is necessary for many types of travel, business, diplomacy, environmental or society movements, and friendships. Dictionaries and software tools help with translation, but often are inadequate for in-person practical communication. The natural method is to use a human translator, but that requires sufficient skill in both languages and of the content area being discussed. The human translator as a third party diminishes privacy. A software based mobile system could potentially improve privacy, availability, and knowledge of the content area. Such a system could be a wearable mobile device connected with web services. This project developed and tested an aural translation system using augmented reality (AR) glasses with audio capabilities connected with a smartphone and Amazon Web Services (AWS) for transcription, translation, and conversion of text back to audio. A prototype was developed based on a Bose AR sunglasses system, and was tested for phrases in English, Spanish and German. The results had reasonable accuracy and processing times. Further development and testing are necessary for wide application, but the results support further development.

Keywords: Cloud services · Transcription · German · Spanish · English · Augmented Reality. AR

1 Introduction

Communication is one of the most essential parts of society. It is not only necessary in everyday life, but in travel, finance and politics. The variety of verbal languages used for social interaction and communication globally makes it difficult to communicate in many situations. There are so many languages across the world that learning them all would be impossible. However, business, culture, and education drive people to try and communicate across different languages. Many techniques and tools are available to help translation, but none are fully adequate. A primary hindrance is how intrusive or distracting using the tool is in a physically present person-to-person communication.

The translation method based on natural processes uses a human translator to assist in the translation. The human translator listens to the sentences one person speaks in

T. Nakano (Ed.): BICT 2021, LNICST 403, pp. 61–70, 2021.
https://doi.org/10.1007/978-3-030-92163-7_6

the first language, understands the meaning, translates the meaning to the second language, and speaks the new translated sentences aloud for the second person to hear in the second language. For this natural process to work, the human translator needs to be able understand both languages, including the content. The content often includes jargon, specialized words and implied meanings based on context. Finding a suitable and available translator is often difficult. Not only does the translator have an undue influence in the conversation by the choice and tone of phrases, but also gains an in-depth knowledge of the conversation. Having a third person with influence and knowledge of a conversation may be concerning in negotiations involving politics, business or romance.

Considering the advantages and disadvantages of the natural process of a human translator, an engineered system might be able to alleviate some of the disadvantages while maintaining some benefits. An impersonal tool that would automate and individualize language translation would give power back to people who need to communicate across language barriers. A tool based on discrete mobile microphones and speakers, and software transcription and translation based on web services might be able to lower the price, increase access and have context expertise (if the utilized web services include that context linguistic knowledge).

Several manual or software-based methods for translation exist or are being developed. Traditional or bilingual dictionaries are the most basic tool for language translation. Going from definition to definition can allow for a confident word translation and the elimination of a third person in the conversation. However, using a dictionary for each word would be time consuming, and sometimes leads to poor understanding of grammar and sentence structure. Certain jargon, slang, grammar, and idioms can be improperly translated by standard dictionaries. Such an improperly translated word or phrase may change the entire meaning of a statement, leaving both parties confused, going forward with misunderstanding. Using a dictionary during in-person communication is not a viable solution.

Besides a physical or electronic dictionary, a web service such as Google Translate (Alphabet, Mountain View, CA; https://translate.google.com/) can make translations of text sentences or paragraphs. Nevertheless, such web services have not been routinely utilized for in-person conversations due to the hassle and disruption to the flow of conversation. Google Translate analyzes input text, translates and displays the text back for the user to read. Google translate works relatively well for both singular words and full sentences. As a result, Google Translate is a popular secondary tool for learning a language [6]. For use in real time conversations, Google Translate has drawbacks. With no built-in audio feature and transcription to text, the user would be required to attempt to enter text for the verbal discussion. This would be difficult and not practical. It would also be a burden to enter in long sentences by phonics only in a language that the user does not know.

An example of audio input and output for Google Translate is available with the Google Pixel 2 earbuds [9]. Google's earbuds connect with the Google Pixel mobile phone, providing a translation pathway from spoken audio in one language. They have the added advantage by Google Translate and text-speech to play through earbuds, which allows for almost "real-time" translation to the user [1]. This audio-to-audio translation

functionality is good but is currently available only on a Google Pixel phone, reducing accessibility.

Another product that could help cross the language barrier of communication is the Vuzix Augmented Reality Glasses [7]. The system inputs verbally spoken audio, translates and displays the translated text on AR glasses display for the user to read [7]. This works well for translation, but the user would have to read display text each time, which may be distracting during a conversation.

A translation system that goes from verbal audio in one language to verbal audio in another language using web-based tools would be beneficial. Some models of AR glasses have capabilities beyond visual to include audio, with built in microphones and speakers. Figure 1 shows an illustration of such an AR glasses based aural translation system. The goal of this project was to develop and test the use of AR audio glasses with mobile device software communicating to web services for translation between languages.

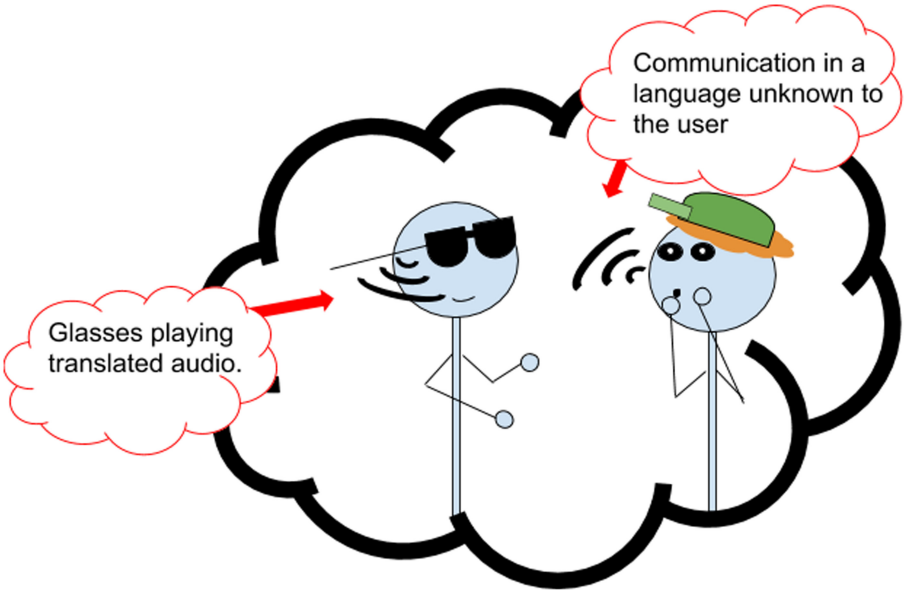


Fig. 1. Illustration of 2 people needing to communicate and using the proposed AR audio classes with web services for translation.

2 Design

Bose (Framingham, Massachusetts) has developed prototypes of AR Sunglasses that have microphones and speakers to record and play audio back to the user. Prototypes of these AR audio glasses were used for the prototype of this project. Using the Bose AR Glasses and commercially obtained mobile phone, the primary development work for this project was done with software development.

The envisioned design involved an app that ran on the user's smartphone, and connected with the Bose AR sunglasses. Amazon Web Services (AWS, Seattle, WA) was utilized for the transcription and translation services.

The user would use built in AR features of the glasses as the human computer interface (HCI) to signal that the system should start recording the speech and initialize the translation procedure. For the prototype, the application was started when the user double tapped on the side of the glasses. They heard a small acknowledgement ping that their action was successful and that the program was ready to record. Then whatever was spoken or heard was recorded. When the verbal sentence or phrase was done the user would again double tap to let the system know the recording should stop. The application would receive the recording as an audio file via Bluetooth from the AR glasses. The app would send the audio file to AWS to transcribe the audio to text. The transcribed text would then be translated to the desired language using AWS. The translated text would then be converted back to audio and then played to the user through the built-in speaker on the glasses. The whole process from the AR audio glasses through the smartphone to AWS is shown in Fig. 2.

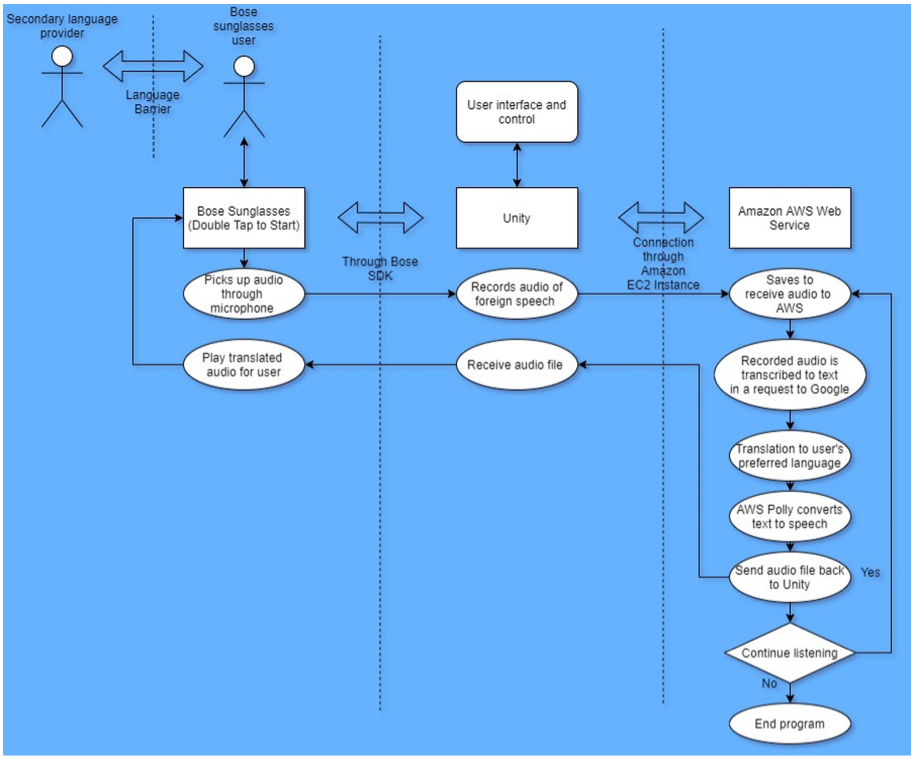


Fig. 2. Block diagram of audio file path. Shows the full cycle the application goes through to provide the user an accurate translation.

Several software modules were developed for the prototype. The modules included the code that ran on the mobile phone and the code that ran on the server. The code running on the phone handled all the communication with the glasses and also included the user interface. The server code handled the transcription and translation. The communication between these two modules occurred with HTTP Restful API requests. The overall workflow of this design is illustrated in Fig. 3.

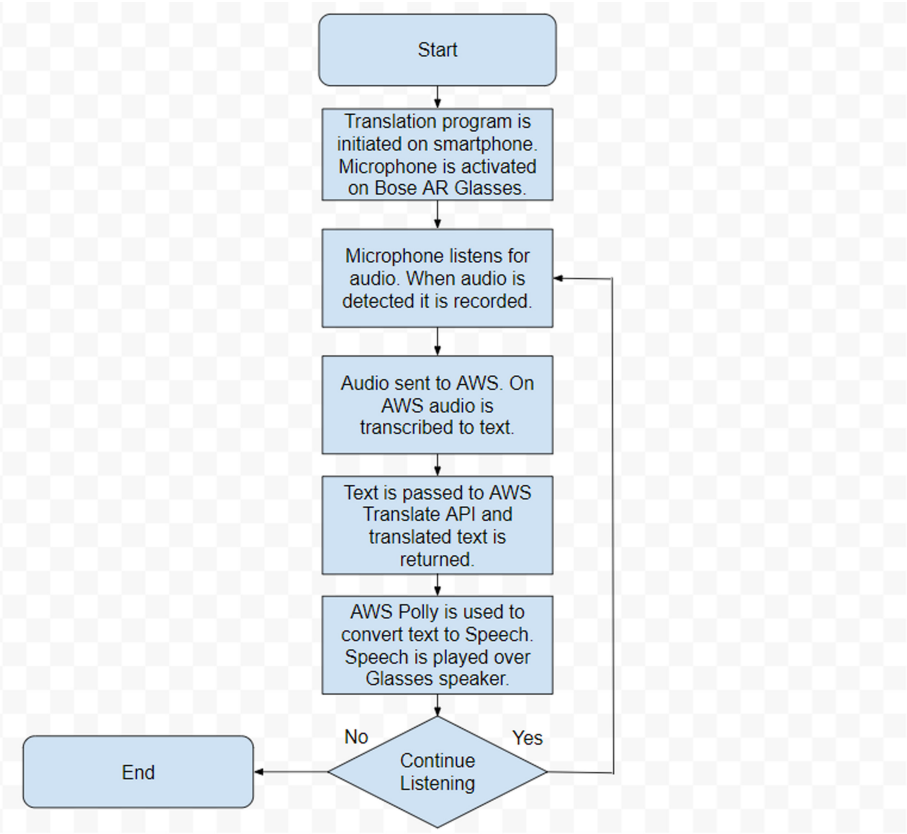
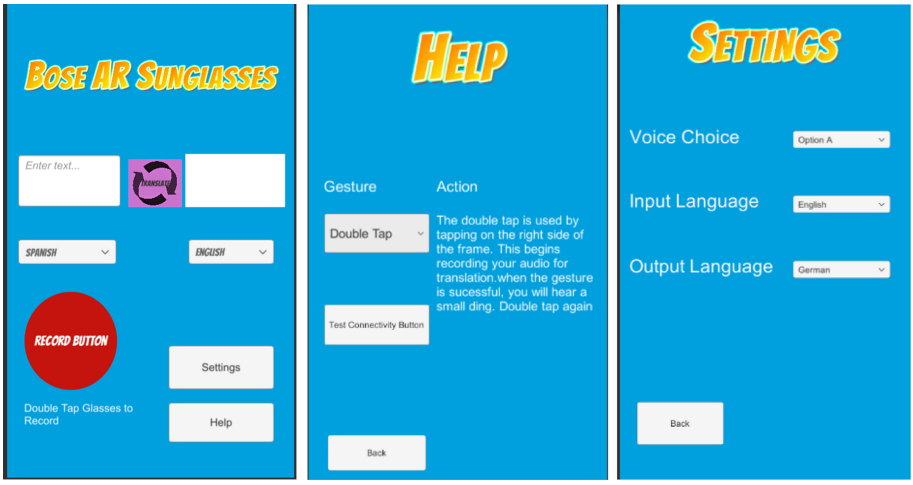


Fig. 3. Flow chart highlights the full cycle of the program. This shows steps of translation through the Amazon Web Services (AWS) for transcription and translation.

The rationale for this design includes the available technologies contained in the Bose glasses. These features included a built-in microphone and a pair of Bose quality speakers [4]. It also had AR touch and gesture features, which were used in the proposed solution. This allowed for several gestures to facilitate the HCI, allowing the user to keep their phone in their pocket, and their focus on who they are talking with. The front-end app allowed for a visual interface the user could navigate the system with. While it is not the primary method of navigation it helped the user to comprehend the product features. Examples of the mobile interface are shown in Fig. 4 and follow a path of connecting to

the glasses over Bluetooth, bringing the user to the home screen, then opening up paths for settings, language alterations, and a help page.



(a) **(b)** **(c)**

Fig. 4. Unity Interface. (A) is the main screen of the application. Here the language settings can be set and the features of the translator can be used. (B) is the help page. On this page you can see if the glasses are connected and information about AR gestures are displayed. (C) is the settings pages where you can set additional settings like voice preferences.

The front-end for the application was built in Unity. Unity is a platform that supports AR game development [2]. The user interface allowed the user to connect their Bose AR glasses (Fig. 5). The user was able to select the language to translate to from a provided dropdown list. It also connected to the microphone on the glasses and begin recording using scripts when the conversation started. The audio files were sent to AWS in an MPEG Audio Layer-3 (mp3) file format over the Internet for transcription and translation. Unity then received the translated mp3 audio file back from AWS. The audio file was then over the speakers in the Bose AR glasses. An example of a gesture used would be shaking your head "no" to have the translated text repeated. Double tapping the frames was used to start and stop recording.

Besides the user interface, back-end function on the web servers was necessary to support the translation application. AWS was used to host the back-end code and data for the system. AWS offered the capability to host servers and AI services that were accessed through API [8]. The back-end modules were hosted on AWS. There were three main modules used in this process. They were transcribing speech to text, translating the text to the desired language, and converting the translated text back to speech. Each of these three modules required their own API to communicate with AWS. Speech to text utilized AWS Transcribe API, and the use of an S3 bucket to save the original audio file. Text translation utilized the AWS Translation API. Text to Speech utilized the AWS Polly API. These modules were integrated with each other to provide seamless speech to speech translation. The backend code was written in PHP, since AWS has many features

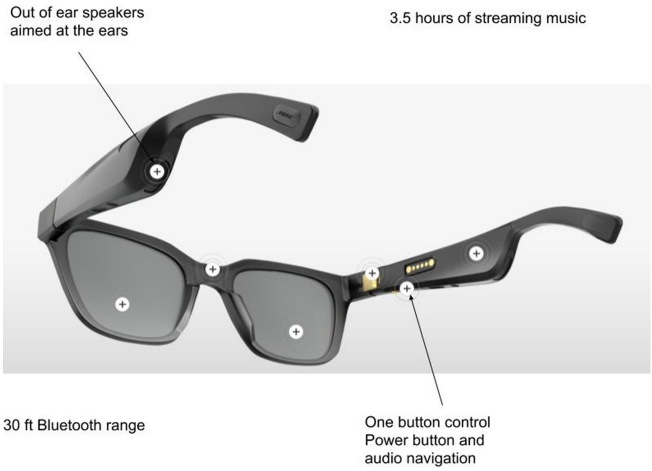


Fig. 5. (Bose AR Sunglasses Interface) Modified from www.Bose.com. This image shows the important features present in the sunglasses. The sunglasses have powerful AR features for the user to use.

and tutorials on how to manage the infrastructure in apps, with the example code being PHP [3].

3 Testing

Testing of the design was primarily focused on the accuracy of the transcription and translation, and the delay time. The method used was putting sample sentences and phrases through the system as a user would, which captured the outputs of the transcription, translation, and timestamp in a log file. Then this data was analyzed for accuracy and delay time. An important distinction to make was that a translation error would occur if there was a contextual error in the translation. People knowledgeable of the output languages were used to verify the validity of the foreign language phrases. The scoring only compared with the prior step. For example, if there was an error in transcription and the translation module correctly translated the input text, then an error would be counted for the transcription step, but no error would be counted for the translation step. The process for measuring input and output translation success was performed four times with different language combinations. The different combinations were English to German, English to Spanish, Spanish to English, and German to English. Another measurement was the time required, such as whether different language translations or varying complexity of sentences had a significant impact on translation time, or if there was little apparent impact.

4 Results

Table 1 shows examples of phrases that were tested in the system and the resulting translation. The system correctly picked up names as proper nouns. The system also

correctly would get the context of the voice inflection for questions and would say the question with a recognizable questioning tone. When translating "Where is the bathroom?" from English to German, there was a contextual error. The German language would more likely use the English equivalent of toilet instead of the word "bathroom". The translation in our system translated bathroom into the German word Badezimmer, which is literally the combination of the German words bath and room. Normally this phrase would not be used in German, but it is not so far off where the understanding would be lost. The results of the testing appeared to show that the prototype worked in an effective way.

Table 1. Sample data fed through the system. This is a few of many sentences that went through each language translation. Four tables were recorded like this one capturing all the data for the system handling each translation.

Input	Expected	Transcribed Output	Score	Translated Output	Score	Time
Hello my name is…	Hallo, mein Name ist…	Hello my name is Forrest	1	Hallo, mein Name ist Forrest	1	4.052641
Where is the bathroom?	Wo ist die Toilette?/Wo ist das Klo?/Wo ist das WC?	Where is the bathroom	1	Wo ist das Badezimmer	0	3.83589
How much does that cost?	Wie viel kostet das?	How much does that cost	1	Wie viel kostet das?	1	3.860418
Where is the closest transportation?	Wo gibt es Verkehrsmittel in der Nähe?	Where is the closest transportation	1	Wo ist der nächste Transport	0	4.122393
What do you do for work?	Was machen Sie zum Beruf?	What do you do for work	1	Was tun Sie für die Arbeit	1	3.86144
Thank you, goodbye	Vielen Dank, tschüss	Thank you goodbye	1	Danke auf Wiedersehen	1	3.885468
I'm sorry (apology)	Es tut mir leid	I am sorry	1	Es tut mir leid	1	3.663461
The car is red	Das Auto ist rot	The car is red	1	Das Auto ist rot	1	3.991076

The results for speed and accuracy showed consistency in the system no matter which language combinations or phrases were used. The system's transcription success rate was 100% for English, 93.3% for Spanish, and 73.3% for German (Fig. 6). Part of the speaking and phrasing errors for Spanish and German transcription may have resulted due to the fact that the subjects speaking the phrases were not native speakers of those languages and the errors might have been mitigated if native speakers had been

speaking into the system. The translation success rate for the system was also satisfactory at 80%–86% for the four language combinations.

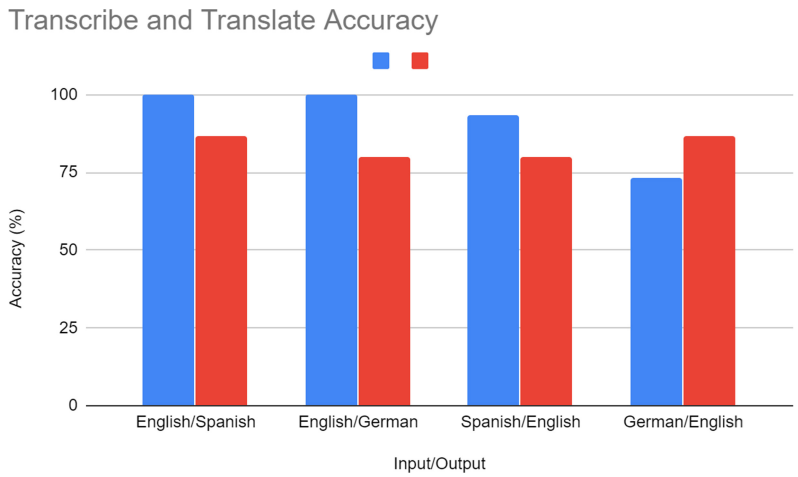


Fig. 6. Graphical representation of the accuracy of the system transcribing and translating each language.

The time for total translation was also consistent which makes the prototype a viable solution for real time communication. Table 2 show times required for translations during the testing process. The four language combinations had average translation times for the sample data that were similar. The graph also displays the consistency for the one sample of English to German translation.

Table 2. Table with average translation time.

Input/Output	Avg. time (s)	Standard deviation
English/Spanish	3.90	0.12
English/German	3.90	0.13
Spanish/English	3.5	0.70
German/English	3.92	0.60

5 Conclusion and Future Direction

The translation system of passing audio files through an Amazon server to a Unity app showed promise. More development and testing is required toward having a system ready for wide application. The development of such a product would improve language translation for hands-free auditory language translation. The prototype showed reasonable response speeds and response accuracy, encouraging further development.

References

1. Min, C., Mathur, A., and Kawsar, F.: Exploring audio and kinetic sensing on earable devices, in Jun 10, 2018. https://doi.org/10.1145/3211960.3211970
2. Fowler, A.: Beginning iOS AR Game Development: Developing Augmented Reality Apps with Unity and C#. Apress, Berkeley, CA (2019). https://doi.org/10.1007/978-1-4842-3618-5
3. Wadia, Y., Udell, R., Chan, L., Gupta, U.: Implementing AWS: Design, Build, and Manage your Infrastructure (2019)
4. Bose Frames Alto: https://www.bose.com/en_us/products/frames/bose-frames-alto.html
5. Vuzix M4000 Smart Glasses: https://www.vuzix.com/products/m4000-smart-glasses
6. Valijärvi, R., Tarsoly, E.: Language Students as Critical Users of Google Translate: Pitfalls and Possibilities, Practitioner Research in Higher Education (2019)
7. McKelvey, C., Dreyer, R, Zhu, D., Wang, W., Quarles, J.: Energy-oriented designs of an augmented-reality application on a VUZIX blade smart glass. In: 2019 Tenth International Green and Sustainable Computing Conference (IGSC), pp. 1–8 (2019)
8. Tripuraneni, S., Song, C.: Hands-On Artificial Intelligence on Amazon Web Services (2019)
9. Google launches Pixel Buds, its Apple AirPods rival, Bennett, Coleman & Co. Ltd, New Delhi. Oct. 16, (2019)