RESEARCH ARTICLE OPEN ACCESS

Multilingual speech-to-speech translation System in Mobile Off-line Environment

Humaid Alshamsi[1] ,Veton Kepuska[2]
(*Department of Computer Engineering and Sciences,Florida Institute of Technology,Melbourne FL,USA*)[1,2,3]

ABSTRACT
The process of translating speech-to-speech denotes the conversion of speech signals specifically from the original source language into a distinctive speech signal-bearing identical meaning or resolve into the target language. To achieve such, there is a need to adopt coordinated cooperation of the separate Human Language Technology components. In particular, the most significant elements in a speech translation system encompass the automatic speech recognition, machine translation as well as text to speech. With this understanding, the current paper explores the design as well as architectural building blocks linked to the "Translator" speech-to-speech translation system. In addition, the current paper explores their interactions with each other to ease speech-to-speech translation in terms of reliability, scalability, and potentially distribution.
***Keywords*-**Speech-to-SpeechTranslation;Automatic Speech Recognition;Machine Translation;Text-to-Speech;Multilingual.

Date of Submission: 07-04-2020 Date of Acceptance: 22-04-2020

I. INTRODUCTION

The objective of every computerized speech-to-speech translation system entails bridging the barrier created by verbal language, specifically when speakers use dissimilar first languages. As a result, the speakers struggle to find a common understanding of the underlying auxiliary language. It is important to note that the barriers triggered by verbal language are not solely present in multilingual societies. Nor are they only present in multilingual countries like the United Arab Emirates. On the contrary, these obstacles may be present in monolingual civilizations as well as regions. These obstacles often emerge due to factors such as migration, business, and tourism. The genitive impacts of language-oriented obstacles may create severe and far-reaching penalties on persons and groups. These penalties tend to emerge as a result of interactions such as:

- Healthcare: Patient comprehension can thwart patient satisfaction due to concerns in preventive care as well as primary care [1].
- Policing/Legal Systems: ineffectual communication that triggers delays as well as disappointments [2], [3].
- Commerce: the deterioration of communication within the transnational firms [4].

The use of computerized translation system tends to lessen significant adverse outcomes that are triggered by language barriers. The computerized translation systems attain the benefits above at a reduced cost-effective as well within predetermined timeframes.

Research views the speech-to-speech translation system as an interdisciplinary Human Language Technology (HLT). That is because the construction of the system necessitates the use of talents and expert knowledge that can only be found in specialist linguists, who work in collaboration with the computer scientists as well as engineers. Every speech-to-speech translation system has three distinctive HLT elements. The three elements are separate from the HLT elements that are integrated into the system with the aim of improving its naturalness and/or performance. The first element encompasses automatic speech recognition (ASR), while the second element is machine translation (MT). The last component is the text-to-speech (TTS). The system amalgamates the three technologies to a point where facilitate the speech-to-speech translation system to achieve its work in the way described below:

- The ASR element identifies input speech. Subsequently, it transcribes the speech into a textual form that relates to the input source language.
- The MT element decodes the transcribed message (in textual form) into text that relates to the desired/target language.
- The TTS element produces synthetic speech, specifically in the desired language using the decoded text.

Thus, the current describes the architecture of "Translator", which is characterized by the presence of a scalable as well as multilingual translation pipeline.

"Translator" is agnostic specifically when one considers the underlying HLT apparatuses and, to be precise, has got the ability to accommodate numerous executions of the fundamental technologies simultaneously. For instance, the translator can accommodate the implementation of numerous, but dissimilar TTS engines instantaneously. In such a case, the objective is to:

❖ Advance multilingualism: In most cases, it is possible to see diverse executions of the HLT apparatuses for dissimilar languages.

❖ Create and maintain a unified interface: In particular, the unified interface is offered to the calling applications specifically via polymorphism of the fundamental HLT elements.

While the sections below explore a wide range of necessary and voluntary HLT elements in a speech to speech translation system, the emphasis of the current paper encompasses the design, architecture as well as the application of the speech to speech instant translation that regulates the flow of information between the core elements and the underlying calling applications.

II. RELATED WORK

The NEC Corporation is accredited with creating a speech-to-speech translation system that is currently believed to be amongst the earliest systems made worldwide. It was invented and illustrated in the form of concept exhibit during the 1983 International Telecommunication Union (ITU) Telecom World, which was an international occasion [5]. In the year 1986, Japan's Advanced Telecommunications Research Institute International (ATR) started conducting research focusing on automated speech to speech translation. The initial phase concentrated on conducting the feasibility study involving a narrow vocabulary as well as clear-read speech system. Meanwhile, the second phase concentrated on conversations based on a restricted domain. Currently, they are more concerned with inventing a speech-to-speech translation system that can be applied to real-life contexts [6].

Over the years, joint research initiatives, as well as consortiums, have become the primary vehicles through which the world seeks to advance the state-of-the-art linked to the automated speech to speech translation systems. Some of these entail:

❖ Consortium for Speech Translation Advanced Research (C-STAR): this is a transnational consortium comprising of ATR (Japan), Carnegie Mellon University (United States), Institute for Research in Science and Technology (Italy), the Chinese Academy of Sciences (China), as well as the Electronics and Telecommunications Research Institute (Korea) [5].

❖ Universal Speech Translation Advanced Research (USTAR): A global research group consisting of 33 institutes originating from at least 26 nations or regions [7].

❖ Asian Speech Translation Advanced Research (A-STAR): Though most stakeholders are from Asia, this a global consortium region. The Asian stakeholders include Japan, Vietnam, Korea, Indonesia, Thailand, India, China, and Singapore [8].

❖ Technology and Corpora for Speech to Speech Translation (TC-STAR): A joint research i of The European Union used the 6th Framework Programmed for Research and Development to create this joint research initiative [9].

❖ Vermobil: With the German Ministry for Research and Technology (BMFT) being the financier alongside industrial and international partners, this is a research and development venture [10].

In the year 2004, the International Workshop on Spoken Language Translation (IWSLT) was initiated by the C-STAR. The members linked to the CSTAR as well as ATR work together to formulate sentences relating to tourism issues. These sentences are used, during the workshop, to assess the spoken language translation technologies based on a multilingual speech corpus. It is worth noting that the workshop concentrated on the assessment of technical papers that are authored to discuss spoken language translation technologies [11]. Since its formation in 2004, the workshop is held annually.

The research efforts and development undertakings relating to speech-to-speech translation systems tend to focus on three key factors. The first encompasses focusing on the essential HLT elements within the setting of a speech-to-speech translation. The other two include the architecture and, lastly, the development of the speech-to-speech translation system.

III. SPEECH TO SPEECH TRANSLATION

When approached using a statistical viewpoint, the speech-to-speech translation task may be presented, as shown below [6]:

$$S_T^* = ang \max_{S_T} P(S_T|S_S) \quad (1)$$

Where:

$S_S$andS_T = the respective source language and target language speech signals.

Meanwhile, it is possible to factor the conditional probability $P(S_T|S_S)$, as shown below:

$$P(S_T|S_S) = \sum_{T_T,T_S} P(S_T, T_T, T_S|S_S)$$
$$= \sum_{T_T,T_S} P(S_T|T_T, T_S, S_S)P(T_T|T_S, S_S)P(T_S|S_S)$$
$$\approx \sum_{T_T,T_S} P(S_T|T_T)P\,(T_T, T_S)P(T_S, S_S)$$

(2)

Here:
$T_S$andT_T = the transcriptions of the respective source language and target language speech signals.
Thus, it is possible to simplify the maximization of $P(S_T|S_S)$ in (1) as shown below:

$$\max_{S_T} P(S_T|S_S)$$
$$= \max_{S_T} P(S_T|T_T) \max_{T_T} P(T_T|T_S) \max_{T_s} P(T_s|S_S)$$

(3)

The findings above indicate that it is possible to decompose the automatic speech to speech translation task into three distinctive tasks. The first is $P(T_s|S_S)$, which refers to automatic speech recognition. The second is $P(T_T|T_s)$, which denotes the machine translation. Meanwhile, the last is $P(S_T|T_T)$, which refers to the text-to-speech element.

Word error rate denotes a kept performance parameter commonly used in assessing ASR system. The findings presented via [12] indicates that optimizing the ASR-MT element improves the quality of translation. That means converting the ASR-MT element into a performance-oriented. It should embrace machine translation like the bilingual evaluation understudy (BLEU). However, such efforts ultimately sacrifice the WER on the ASR element.

When operating in multilingual environments and contexts, there should be extra input parameters or alternatively, consider investing in alternative methods for informing the ASR element that the source language expects. In multilingual settings, the ASR element possesses the capability of identifying multiple source languages. Attaining these goals requires the use of extra input parameter or, alternatively, by using a spoken language identification system. There are times when joint language identification is necessary [13]
Human speech is an intricate signal that disseminates a wide range of information such as identity, styled, intent, emotion as well as the state of the speaker, [14]. In recent times, the research has focused on enhancing the naturalness of speech to speech translation systems [15], emotion [16], punctuation, as well as segmentation [17].

A. System Overview

In figure No. 1, there is an illustration with architecture linked to the currently proposed automated customer satisfaction assessment system. Thus, the "Translator" App. Denotes an offline-based multilingual speech-to-speech translation system that is designed for use via the smartphones. The voices spoken by the users tends to be processed via speech recognition while machine translation converts the input data into machine language. Eventually, the text to speech translates the machine language data to use readable data in the form of foreign languages. It is important to understand that all three processes, despite being cumbersome, happen via the server-side. As such, it is possible to make use of the large-scale speech recognition models as well as translation models. The objective of the model is to accomplish more precise speech translations that would be the case when using the stand-alone systems that are designed to handle all the processes within the underlying terminal units.
For the current task, speech to speech translation is the primary case study. It entails the use of a mobile application specifically for use using the android platform. The mobile app translates the English voice/speech to Chinese voice/speech based on the corpus-based database.

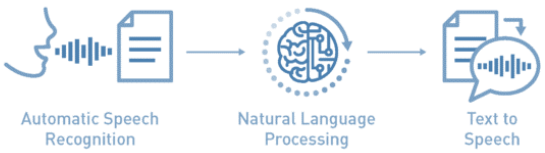

Fig. 1. Architecture of Speech-to-Speech Translation Model.

Speech to Speech Translation demonstrated in Figure No 1 covers three key modules. These include Automatic Speech Recognition Module, the Machine Translation Module and Speech Synthesis module Module.

B. Feature Extraction

Speech recognition module – this module captures a speech, mostly in the form of voice, from the digital or rather mobile device using a speaker. The module detects the language that the user has used and, subsequently, transforms the voice message into text form. Then, the same module sends the converted text to the next module. For compact, precise model from corpora based on a narrow size, the MDL-SSS is mostly used [18] as well as the composite multi-class N-gram models [19] to

conduct the modeling of the acoustic and language. With the MDL-SSS, the appropriate size of the metric is determined automatically to match the volume of the training data based on the Maximum Description Length (MDL) criterion. A Minimum mean square error (MMSE) estimator for log Mel-spectral energy coefficients using a GMM (Gaussian Mixture Model) [20] suppress interference and noise and for offsetting echo. Speech detection errors can be triggered by factors including the variability of speakers, disparities in the training and testing channels, as well as interference from environmental noise. Such speech detection errors may be precluded by tagging them with significantly low confidence value, which is possible when one issuing word posterior probability (GWPP)-based detection error rejection specifically when it comes to post-processing of the voice data [21, 22].

Machine Translation module is also known as Natural language processing. The module is concerned with the translation activity. In this module, there is a library for the language and text received by the module. The module converts the received textual data, which is in a single language, to a different dialect based on user choice. Eventually, the module sends its output to the last module.

The translation modules that are often used are automatically constructed. In addition, they are a phrase-based SMT module as well as an Example-based machine translation (EBMT). This is a memory-based module. The EMBT is a vital component given that it helps in aligning the source sentence with the source language elements linked to the translation. When the perfect matches are found, the end result ultimately becomes the corresponding target language sentence.

Speech Synthesis module is the last module that is also known as the Text To Speech module. It helps the system to convert all pieces of the translated text from digital into analog form. The module ultimately sends the output to the user.

To understand the entire process, one can look at the procedural illustration below.

```
Click (View ) {
API used for TTS is Android.speech.tts.TextTo Speech
android.speech.tts.TextToSpeech.and get the text entered
EditTextenteredText as select the text entered
Change text to string;
Speak the words;
}
speak the user text
speak the word(String speech)
{
Use TTS.speak(speech, TextToSpeech.);
}
Check status {
check for successful instantiation
if text to speech is success {
the corresponding language is available
}
else if error {
text to speech failed;
}
}
```

Immediately after effective configuration of the utterance object, the system passes it to the speech synthesizer object. The main objective is to give the system input that it can use to generate speech. One does not need to make endless efforts to speak many utterances. On the contrary, one only needs to set the desired statements to the synthesizer based on the order that ought to be observed. The synthesizer, in return, will queue the statements automatically.

The AV Speech Synthesizer Delegate protocol always accompanies the AV Speech Synthesizer. It is vital for the well-functioning of the system because it holds valuable delegate methods. When used effectively, these delegate methods make it easy for the system to track the progress of the underlying speech as well as the presently spoken text. The act of tracking progress can be unnecessary. However, for those who do so, they get to see the best ways of delivering positive results. This is a problematic process, though one can embrace it once he or she gets to understand the overall functioning of the system.

IV. SYSTEM EVALUATION

The findings will be linked to the speech to speech translator together with an application diagram or the screen shorts showing all modules linked to the speech-to-speech translator system. Hence, the findings will demonstrate how the module is essentially working within a specific context.

The key sections of the findings include:

- Speech Recognition Module
- Machine Translation Module
- Speech Synthesis Module

The Speech recognition module provides effective measures of the work that has been completed in the past. The key feature linked to speech recognition encompasses:-

a. Word accuracy – This helps to detect the percentage of correctness linked to a certain word. In particular, speech recognition helps to determine the accuracy of the word.

b. Utterance correctness – This determines the speed and the percentage of correctness lined to the pronunciation of the underlying word.

Machine Translation module is essential to the system. Its main features include:

a. Translation Accuracy – used to determine the accuracy of the machine translator in terms of translating the necessary words as well as sentences.

b. Translation Speed – It helps to measure the speed of retrieving the corresponding words as well as sentences from the primary database. It also helps to measure the speed of the entire translation process.

Speech synthesis module is integral to the system in that it makes it possible to convert the underlying text into speech. It is the speech synthesis module that gives the translator system the foundation to serve as a better communication avenue. The main features of this module entail:

a. Word accuracy – helps in determining the correctness in word spelling.

b. Utterance correctness – helps to assess the speed at which pronunciation is done accurately.

V. CONCLUSION

"Translator" Application denotes a unique android mobile app that enables a user to engage in the translation of one language to a different dialect using an off-line environment thus creating a foundation upon which humans can talk with each other in dissimilar languages. In addition to helping humans communicate amongst themselves, it helps commercial ventures to engage in conferences and meetings in diverse lingual settings. Thus, humans can interact amongst themselves because they can understand one another easily. The mobile app eradicates the language-related obstacles between people using multiple dissimilar languages, thereby limiting their ability to converse with their friends, associates, and partners. The mobile app helps people to bridge the social gap amongst themselves as a result of language-oriented obstacles. Users can learn different languages, including English, Chinese, and others through interactions with foreigners. Through the app, people can interact easily and work better as well as easily.

REFERENCES

[1]. E. A. Jacobs, D. S. Shepard, J. A. Suaya, and E.-L. Stone, "Overcoming language barriers in health care: costs and benefits of interpreter services," American journal of public health, vol. 94, no. 5, pp. 866– 869, 2004.

[2]. L. Herbst and S. Walker, "Language barriers in the delivery of police services: A study of police and hispanic interactions in a Midwestern city," Journal of Criminal Justice, vol. 29, no. 4, pp. 329–340, 2001.

[3]. R. C. Davis, E. Erez, and N. Avitabile, "Access to justice for immigrants who are victimized: The perspectives of police and prosecutors," Criminal Justice Policy Review, vol. 12, no. 3, pp. 183–196, 2001.

[4]. H. Tenzer, M. Pudelko, and A.-W. Harzing, "The impact of language barriers on trust formation in multinational teams," Journal of International Business Studies, vol. 45, no. 5, pp. 508–535, 2014.

[5]. S. Nakamura, "Overcoming the Language Barrier with Speech Translation Technology," NISTEP Quarterly Review, no. 31, pp. 35–48, 2009.

[6]. S. Nakamura, K. Markov, H. Nakaiwa, G. ichiro Kikui, H. Kawai, T. Jitsuhiro, J.-S. Zhang, H. Yamamoto, E. Sumita, and S. Yamamoto, "The ATR Multilingual Speech-to-Speech Translation System," IEEE Transactions on Audio, Speech, and Language Processing, vol. 14, no. 2, pp. 365–376, 2006.

[7]. (2018, Jun) U-STAR - The Universal Speech TranslationAdvanced Research Consortium. [Online]. Available: http://www.ustarconsortium.com/qws/slot/u50227/index.html

[8]. S. S. Watiasri, M. Paul, A. Finch, S. Sakai, T. T. Vu, N. Kimura, C. Hori, E. Sumita, S. Nakamura, J. Park, C. Wutiwiwatchai, B. Xu, H. Riza, K. Arora, C. M. Luong, and H. Li, "A-STAR: Toward Translating Asian Spoken Languages," Computer Speech & Language, vol. 27, no. 2, pp. 509–527, 2011.

[9]. G. Lazzari and V. Steinbiss, "Human language technologies for europe," ITC IRST/TC-Star project report, 2006.

[10]. W. Wahlster, "Verbmobil," in Grundlagen und anwendungen der k¨unstlichen intelligenz. Springer, 1993, pp. 393–402.

[11]. (2018, Jun) International Workshop on Spoken Language Translation (IWSLT) 2004. [Online]. Available: https://www.iscaspeech. org/archive/iswlt 04/

[12]. X. He, L. Deng, and A. Acero, "Why word error rate is not a good metric for speech recognizer training for the speech translation

task?" in Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Prague, Czech Republic, May 2011,pp. 5632–5635.

[13]. D. C. Y. Lim, I. Lane, and A. Waibel, "Real-Time Spoken Language Identification and Recognition for Speech-to-Speech Translation," in International Workshop on Spoken Language Translation (IWSLT), Paris, France, 2010, pp. 307–312.

[14]. S. Johar, Emotion, affect and personality in speech: The Bias of language and paralanguage. Springer, 2015.

[15]. Q. T. Do, S. Takamichi, S. Sakti, G. Neubig, T. Toda, and S. Nakamura, "Preserving word-level emphasis in speech-to-speech translation using linear regression HSMMs," in Sixteenth Annual Conference of the International Speech Communication Association, Dresden, Germany, 2015.

[16]. M. Akagi, X. Han, R. Elbarougy, Y. Hamada, and J. Li, "Toward affective speech-to-speech translation: Strategy for emotional speech recognition and synthesis in multiple languages," in Signal and Information Processing Association Annual Summit and Conference (APSIPA), Asia-Pacific, Chiang Mai, Thailand, 2014, pp. 1–10.

[17]. E. Cho, J. Niehues, and A. Waibel, "Domain-independent Punctuation and Segmentation Insertion," in International Workshop on Spoken Language Translation (IWSLT), Tokyo, Japan, 2017, pp. 90–97.

[18]. T .Hirokawa and K.Hakoda," Segment select ion and pitch modificat ion for high quality speech synthesis using wave form segment s", proceeding of the First International Conference on Spoken Language Processing, 1990.

[19]. R.Donovan,"Trainable speech synthesis", P h.D. Dissert at ion, Cambridge University, 1996.

[20]. A. Breen and P .Jackson," Non uniform unit select ion and the similarit y met ricswithin BT 's laureat e T T S system", Proceedings of third International Workshop on Speech Synthesis, Jenolan, 1998.

[21]. Y.Sagisaka, N.Kaiki, N.Iwahashi and K.Mimura,"ATR talk speech synthesis system", The Second International Conference on Spoken Language Processing, ICSLP 1992.

[22]. A.W.Black and P .T aylor, " Chat r: a genet ic speech synthesissystem",P roceedings of Coling, Japan, 1994.