Modular Design of a Front-End and Back-End Speech-to-Speech Translation Application for Psychiatric Treatment of Refugees

Enes Yavuz Ugan
Interactive Systems Lab
Karlsruhe Insitut of Technology
Karlsruhe, Germany
enes.ugan@kit.edu

Mohammed Mediani [§]
College of Information Technology
United Arab Emirates University
Al-Ain, Abu-Dhabi, UAE
mohammed.mediani@uaeu.ac.ae

Omar Al Jawabra
Interactive Systems Lab
Karlsruhe Insitut of Technology
Karlsruhe, Germany
uwyos@student.kit.edu

Aya Khader
Interactive Systems Lab
Karlsruhe Insitut of Technology
Karlsruhe, Germany
Khdar.aya@gmail.com

Yining Liu
Interactive Systems Lab
Karlsruhe Insitut of Technology
Karlsruhe, Germany
hk0206@partner.kit.edu

Alexander Waibel
Interactive Systems Lab
Carnegie Mellon University
Pittsburgh, USA
alexander.waibel@cmu.edu

2023 IEEE Global Humanitarian Technology Conference (GHTC) | 979-8-3503-2132-6/23/$31.00 ©2023 IEEE | DOI: 10.1109/GHTC56179.2023.10354809

***Abstract*—One of the inevitable impacts happening in areas with political conflicts is the significant influx of displaced individuals. The psychological consequences on individuals enduring such events are profound. Therefore, the imperative of providing adequate mental health care to refugees coming from conflict areas becomes apparent. However, providing this necessary care faces two obstacles. On the one hand, not all this target population is expected to have an acceptable level of proficiency of the hosting country's local language. On the other hand, finding enough number of suitable interpreters is a very challenging task. Moreover, even when the availability of the human interpreters is no problem, the refugees may hesitate to share their experiences with interpreters due to the associated stigma. To address these challenges and enhance mental health care for refugees, we propose the design of a modular front-end and back-end Speech-to-Speech translation system, with a focus on safeguarding patient data privacy. As our system is Speech-to-Speech, it also enables dialogue with dyslexic people and removes barriers for their treatment as well.**

***Index Terms*—mental health, application, artificial intelligence, speech recognition, machine translation, speech synthesis, speech-to-speech system**

I. Introduction

In recent years, several new conflict hotbeds have emerged, resulting in an augmented number of civilians seeking refuge abroad. For example, Germany witnessed a notable increase in the number of refugees from the Middle East, particularly among Arabic speakers. Considering only Iraq and Syria, for instance, the number of migrants rose from 502.955 to 1.208.400 during the period of 2015 to 2022 [1]. Even worse, the number of Ukrainian citizens in Germany has witnessed a more substantial rise from 133.775 to 1.164.200 in the same time period [1].

Many individuals who have fled these conflicts have experienced traumatic events that continue to affect their daily lives even after seeking refuge in safer countries. Providing appropriate mental health care is crucial to help these traumatized individuals cope with their experiences. However, a significant challenge in delivering effective treatment arises from the language barrier between patients and doctors, as many refugees do not speak the same language as the healthcare professionals.

Although hiring interpreters may seem like a potential solution, there are several reasons that render this approach infeasible. Firstly, the costs associated with employing interpreters pose a major financial burden for clinics and hospitals, making it difficult for them to sustain such arrangements over extended periods. Additionally, the presence of a third-party interpreter during the treatment process can impede the patient's comfort and hinder the establishment of a therapeutic relationship. Traumatized individuals often experience feelings of shame related to their experiences and may find it challenging to open up in the presence of an intermediary.

The use of publicly available translation tools is not a viable alternative either, as there is a lack of trust in these systems. Concerns arise due to the transmission of sensitive patient data to large corporations, compromising data privacy. Furthermore, these translation tools are not tailored to the specific vocabulary and nuances of the psychological domain, which often requires specialized terminology that differs from general domain data.

Based on our experience from previous projects aiming at providing technologies benefiting humans [2], [3], [4], [5] and [6] we address these new challenges, and propose the development of a modular front-end and back-end Speech-to-Speech translation system powered by advanced artificial intelligence (AI) models. These AI models can be easily extended, adjusted, or replaced as needed, enabling more accessible and readily available treatment for a larger number of traumatized refugees.

[§]The author worked on the project while pursuing his Post-Doc at KIT

II. Innovation

We developed a comprehensive application to facilitate therapeutic sessions, comprising a front-end interface, back-end infrastructure comprising integrated automatic speech recognition (ASR), machine translation (MT), and text to speech synthesis (TTS) models. The front-end component operates independently, allowing for flexibility in customization and design. Both users have the freedom to choose their preferred headsets for the sessions.

Once a user finishes their utterance, the front-end seamlessly communicates with the back-end, displaying the corresponding transcript and translation for the other participant. Moreover, we provide a back-translation as a reference for speakers to gauge the quality of the translation. For patients who are unable to read, we incorporate a speech synthesis function that enables them to listen to the translations.

The back-end architecture is designed to efficiently handle concurrent clients through parallelization and batching techniques for transcription, translation, and speech synthesis requests. The back-end establishes a distinct communication channel with the AI models, facilitating easy and fast updates, changes, or additions to the AI model repertoire. All these concurrent communications are managed by an internal mediator.

III. Application Design

In this section, we briefly elaborate the design choices for the front-end and the back-end of our application. We aim at keeping both parts as independent as possible from each other. This has multiple benefits, like helping to develop different front-ends or newer ones in the future, as well as using the back-end for different applications in different domains. An abstract overview is given in Fig. 1.

A. The Front-end

The front-end of the current system is created for Windows-based operating systems. We select the flutter framework for development, which employs the object-oriented Dart programming language. Since Flutter is often a cross-platform framework, we intend to add support of the application on other operating systems like IOS. In our scenario, it is not as straight forward to create a cross-platform application for different operating systems because our application requires some particular utilities, such as the ability to use and choose numerous microphones at once. A screenshot of our front-end is depicted in Fig. 2. In the image, we can see the current application scenario, in which the patient and the doctor sit a cross from each other. While the Arabic side has the back-translation feature activated (blue text box) the German side has it deactivated.

B. The Back-end

Our back-end server, described as Mediator in Fig. 1, makes use of the Python-based Django REST framework, which is simple to read and modify by all our team-members who are not all experienced developers. The back-end provides for API endpoints which only allow for encrypted requests by the front-end. The front-end application has to send the audio to the correct endpoint specifying the language spoken and the back-end will forward it to the correct ASR model and return the transcript to the front-end. For translations and TTS the client side sends the text to the respective API and specifies the target language. The Mediator will again forward the data to the right model and return the appropriate results. Our Mediator as well as the AI models we use are implemented in a way to support parallel processing of requests and thus reduce the waiting time in case of simultaneous usage of the application. One has to bear in mind that there might be multiple front-end sessions run by different doctors. If the users agree, we also store the communication on the server for retrieval in the future. The doctor can store a session conversation as a text file locally as well. Modern transformer models for autonomous machine translation, Long Short-Term Memory LSTM-based models for speech synthesis, and recognition utilizing attention mechanisms are the AI models we deploy.

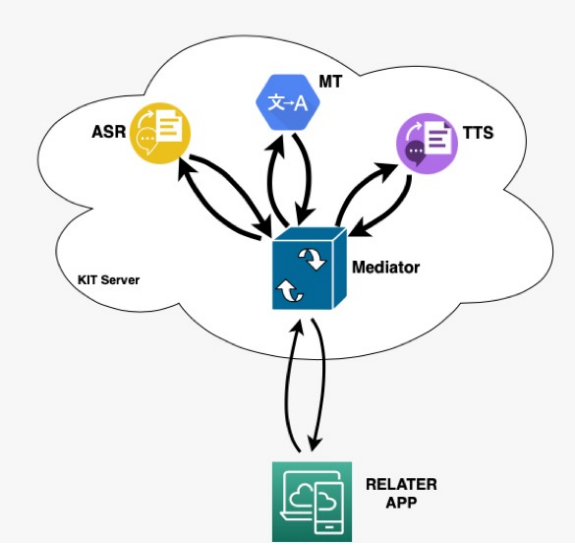

Fig. 1. Abstract depiction of our current application workflow.

Doctors who used the application in therapeutic test sessions provided comments as the front-end was being created. The AI models are also always being improved in accordance with user feedback. The modular and independent structure of our AI models, depicted as ASR, MT and TTS in Fig. 1, enable us to easily exchange models without changing any back-end code.

IV. Current AI Models

In this section, we briefly describe the models used by the back-end to perform ASR, MT, and speech synthesis.

A. Automatic Speech Recognition

We conducted experiments using various ASR models with and without batch-weighting [7]. Our study focuses on developing a speech-to-speech translation application for psychiatric diagnosis, specifically targeting three languages: Arabic, German, and Ukrainian. For Arabic, we utilize the Alj.1200h corpus [8] and test our models on a 2-hour

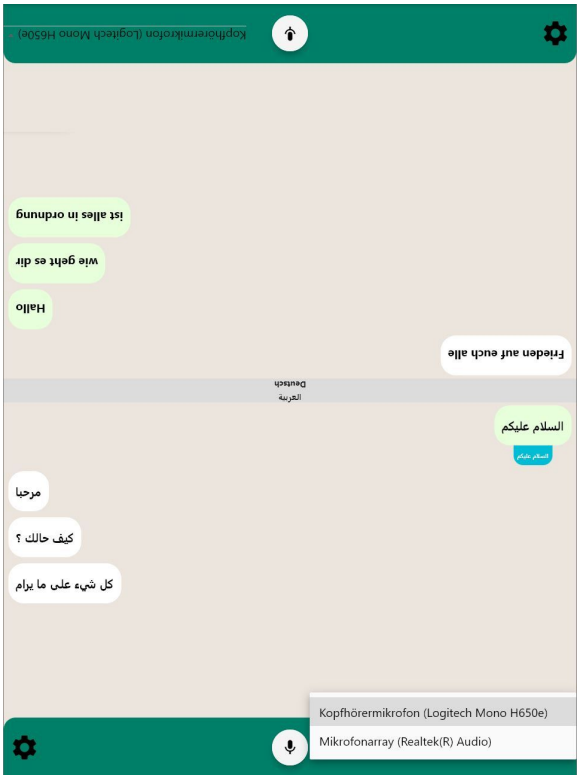

Fig. 2. The UI of our Speech-to-Speech Translation Application for psychiatric treatment.

subset of the modern standard Arabic speech, extracted from the Alj.1200h test set, as described in [9] (Alj.1200h-2h). Additionally, we collect 10.53 hours of interview data using the M.I.N.I [10] questionnaire, with 9 hours allocated for training and 1.53 hours for testing. The data collection was done using the TEQST[§] tool. After user feedback, we decided that adapting our models to Levantine dialect is crucial for our use-case. As such, we collect 71.06 hours of dialogue data through native Levantine dialect speakers using the ASROT Website[§]. Of this data, 66.57 hours are used for training, and 4.49 hours are held out for testing.

For the German language, we follow a similar data collection process, with 3.7 hours used for training and 55.18 minutes of M.I.N.I interview recordings used for testing.

For Ukrainian, we are still in the process of collecting the appropriate M.I.N.I. data. Please refer to Table I for a detailed distribution of our training and test data.

Our current results are presented in Table II. Achieving satisfactory results, especially for Arabic, has been a non-trivial task, requiring extensive research. Notably, the results obtained on Levantine dialectal data have been less favorable. We consider this an ongoing research area and aim to improve our performance on Levantine dialects without compromising our results on M.I.N.I or modern standard Arabic. In our baseline models the performance on the M.I.N.I. data was considerably higher with a Word

[§]https://github.com/teqst

[§]https://transcript-corrector.dataforlearningmachines.com/login

TABLE I
TRAINING (UPPER PART) AND TEST (LOWER PART) DATA DISTRIBUTION FOR ASR MODELS

Language	Data-sets	utterance length
German	CommonVoice [11], Europarl [12], Lectures,M.I.N.I	867.7 h
Arabic	Alj.1200h, M.I.N.I, Levantine	1202.57 h
Ukrainian	Various data-sets [13]	1655.08 h
German	CommonVoice, M.I.N.I	25 h
Arabic	Alj.1200h-2h , M.I.N.I, Levantine	8.02 h
Ukrainian	CommonVoice	8.56 h

Error Rate (WER) of 23.69%. These results show the importance of researching adaptation techniques of ASR models to specific domains of use, as well as different dialects, which are underrepresented in publicly available datasets.

B. Machine Translation

Currently, we deploy multiple Transformer based models. As it was previously found out that cascaded models still outperform direct speech translation models [14], [15] we employ the translation after the initial speech recognition. For MT a pivoting through English is used as it yielded better results in our experiments. The system architecture and data are similar to the ones presented in [9]. Currently, the German-Ukrainian MT system is realized in a pivoting fashion through English as well. For the Ukrainian $\rightarrow$ English, we mainly train models similar to the Arabic $\rightarrow$ English, using data from public sources (such as OPUS[§]) and LDC. This amounted to a total number of 1.7 million sentence pairs. To complete the full path between Ukrainian and German, we exploit the same English $\rightarrow$ German models formerly developed for the Arabic–German language pair. Our results are depicted in BLEU in Table III.

C. Text to Speech Synthesis

For our TTS model, we currently use the model described in [9]. However, we also develop new Arabic TTS models by using models like VITS [16]. In contrast to previous other models, VITS adopts a more end2end structure without predicting spectrograms first and models the intermediate representations to be subjected to Gaussian distributions with unfixed-variance. This enables the model to have a more generalised learning capacity and thus enables better speech naturalness. Besides, VITS keeps the non-auto-regressive structure and replaces the time-consuming transformer decoder, like in FastSpeech [17], with light-weighted flow layers thus gains more speed bonus. As we want a more seamless user experience the speed and naturalness of TTS models are very important, as well.

V. EXPECTED IMPACT

Refugees from war-torn areas may suffer serious traumas. This can result in depression, drug abuse, and affecting

[§]https://opus.nlpl.eu/

TABLE II
WER PERFORMANCE OF CURRENT MONOLINGUAL ASR MODELS ON RESPECTIVE TEST SET

Language	Data-set	WER
German	CommonVoice	16.53
	M.I.N.I.	0.21
Arabic	Alj.1200h-2h	11.48
	M.I.N.I.	8.03
	Levantine	31.68
Ukrainian	CommonVoice	3.57

TABLE III
BLEU PERFORMANCE OF OUR CURRENT MT SYSTEMS

Language pair	General Domain	M.I.N.I Domain
German to Arabic	24.49	38.59
Arabic to German	31.35	64.23
Ukrainian to English	37.99	-
English to Ukrainian	30.11	-

the person's daily life even after they have fled to safer places.

As such, it is essential to offer these "casualties" the right mental health-care in order to aid them to recover from their traumatic experiences. However, it is not uncommon for refugees not being able to speak the language of their target country. In these situations, interpreters are required to help patients and professionals communicate. However, communication with a doctor through an interpreter has its own shortcomings. First, its cost is very high. Second, the interpreters could influence the transferred message due to their cultural background. Finally, the patient may feel a lack of privacy and will not be able to express themselves freely with the presence of the interpreter.

We remove the language barrier for the mental health treatment of refugees by offering a modular front-end and back-end application with suitable AI models. Since the back-end can be installed in every mental health facility, it eliminates the requirement for an internet connection and addresses privacy concerns of the patients.

It is worth mentioning that at first, the application targeted the Arabic-speaking refugees. However, while being in an advanced state of the application, the war broke out in Ukraine; thanks to the modular design of our application, we could incorporate the communication with Ukrainian patients in a short amount of time Additionally, our approach also removes barriers for dyslexic patients by enabling speech input and speech output and as such there is no need to read or write anything in order to communicate. This project enables faster communication and diagnosis of mental health issues with refugees of all kinds of backgrounds. By providing a faster diagnosis and treatment for these people, we hope to relieve and prevent people who experienced traumas from their struggles and habits like drug abuse. This in return will help them integrate into their new communities in a faster and more enjoyable way.

This project aligns with the United Nations Sustainable Development Goals (SDGs) of Good Health and Well Being (SDG3), as well as Reduced Inequalities (SDG10).

ACKNOWLEDGMENT

The project on which this report is based was funded by the Federal Ministry of Education and Research (BMBF) of Germany under the number 01EF1803B (RELATER).

REFERENCES

[1] Statistische Bundesamt (2022) Ausländische Bevölkerung nach Bundesländern und Jahren. DESTATIS. https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Bevoelkerung/Migration-Integration/Tabellen/rohdaten-auslaendische-bevoelkerung-zeitreihe.html#fussnote-1-586248 Cited 31 Mai 2023

[2] Hourin, S., Binder, J., Yaeger, D., Gamerdinger, C., Wilson, K. & Torres-Smith, K. undefined. *SPEECH-TO-SPEECH TRANSLATION TOOL (S2ST2) Limited Utility Assessment Final Report*. pp. 1-77 (2013)

[3] Waibel, A. Speech translators for humanitarian projects. *Spoken Language Technologies for Under-Resourced Languages*. pp. 4-5 (2010)

[4] Schultz, T., Alexander, D., Black, A., Peterson, K., Suebvisai, S. & Waibel, A. A Thai speech translation system for medical dialogs. *Demonstration Papers At HLT-NAACL 2004*. pp. 34-35 (2004)

[5] Scheytt, P., Geutner, P. & Waibel, A. Serbo-Croatian LVCSR on the dictation and broadcast news domain. *Proceedings Of The 1998 IEEE International Conference On Acoustics, Speech And Signal Processing, ICASSP'98 (Cat. No. 98CH36181)*. **2** pp. 897-900 (1998)

[6] Schultz, T. & Waibel, A. Experiments on cross-language acoustic modeling.. *INTERSPEECH*. pp. 2721-2724 (2001)

[7] Huber, C., Hussain, J., Nguyen, T., Song, K., Stüker, S. & Waibel, A. Supervised adaptation of sequence-to-sequence speech recognition systems using batch-weighting. *Proceedings Of The 2nd Workshop On Life-long Learning For Spoken Language Systems*. pp. 9-17 (2020)

[8] Ali, A., Bell, P., Glass, J., Messaoui, Y., Mubarak, H., Renals, S. & Zhang, Y. The MGB-2 challenge: Arabic multi-dialect broadcast media recognition. *2016 IEEE Spoken Language Technology Workshop (SLT)*. pp. 279-284 (2016)

[9] Hussain, J., Mediani, M., Behr, M., Cheragui, M., Stüker, S. & Waibel, A. German-Arabic Speech-to-Speech Translation for Psychiatric Diagnosis. *Proceedings Of The Fifth Arabic Natural Language Processing Workshop*. pp. 1-11 (2020)

[10] Sheehan, D., Lecrubier, Y., Sheehan, K., Amorim, P., Janavs, J., Weiller, E., Hergueta, T., Baker, R., Dunbar, G. & Others The Mini-International Neuropsychiatric Interview (MINI): the development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. *Journal Of Clinical Psychiatry*. **59**, 22-33 (1998)

[11] Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., Morais, R., Saunders, L., Tyers, F. & Weber, G. Common voice: A massively-multilingual speech corpus. *ArXiv Preprint ArXiv:1912.06670*. (2019)

[12] Koehn, P. Europarl: A parallel corpus for statistical machine translation. *Proceedings Of Machine Translation Summit X: Papers*. pp. 79-86 (2005)

[13] Snakers4 open_stt. *GitHub Repository*. (2022), https://github.com/snakers4/open_stt

[14] Nguyen, T., Nguyen, T., Huber, C., Pham, N., Ha, T., Schneider, F. & Stüker, S. KIT's IWSLT 2021 offline speech translation system. *Proceedings Of The 18th International Conference On Spoken Language Translation (IWSLT 2021)*. pp. 125-130 (2021)

[15] Zhang, W., Ye, Z., Tang, H., Li, X., Zhou, X., Yang, J., Cui, J., Deng, P., Shi, M., Song, Y. & Others The USTC-NELSLIP Offline Speech Translation Systems for IWSLT 2022. *Proceedings Of The 19th International Conference On Spoken Language Translation (IWSLT 2022)*. pp. 198-207 (2022)

[16] Kim, J., Kong, J. & Son, J. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. *International Conference On Machine Learning*. pp. 5530-5540 (2021)

[17] Ren, Y., Ruan, Y., Tan, X., Qin, T., Zhao, S., Zhao, Z. & Liu, T. Fastspeech: Fast, robust and controllable text to speech. *Advances In Neural Information Processing Systems*. **32** (2019)