INTERSPEECH 2021: Show & Tell Contribution
30 August – 3 September, 2021, Brno, Czechia

The INGENIOUS Multilingual Operations App

Joan Codina-Filbà[1], *Guillermo Cámbara*[1], *Alex Peiró-Lilja*[1], *Jens Grivolla*[1], *Roberto Carlini*[1], *Mireia Farrús*[2]

[1]Universitat Pompeu Fabra, Barcelona, Spain
[2]Universitat de Barcelona, Barcelona, Spain

{joan.codina,guillermo.cambara,alex.peiro,jens.grivolla, roberto.carlini}@upf.edu, mfarrus@ub.edu

Abstract

This paper presents the integration of a speech-to-speech translation service into a Telegram bot as a part of the EU funded INGENIOUS project. The bot is thought as a multilingual communication channel where First Responders talk in their own language and receive other's messages in English. The Speech-to-Speech translation system is currently being adapted to the emergency domains, so it will correctly deal with emergency codes and geographical data.

Index Terms: speech-to-speech, speech recognition, machine translation, speech synthesis.

1. Introduction

INGENIOUS is a research project that aims at assisting First Responders (FRs) to be more effective during natural and man-made disasters. Novel technologies are exploited to provide FRs with a high-tech equipment and infrastructure consisting of swarm drones, smart uniform, boots and helmet, a K9(dog) wearable kit, augmented reality, data intelligence, and a suite of mobile applications with: worksite operations, victim triaging, multilingual collaboration and social media monitoring. In this paper, we present the Multilingual Operations Application (MOA) that provides communication support, within the suite, to ease the communication between teams of FRs from different countries. In a cross-border or in a big disaster such as an earthquake, for instance, teams from different countries come to cooperate and they can have different communication problems. The objective of the MOA is to facilitate the communication between international teams, avoiding communication misunderstandings, like the ones that can be caused by emergency codes [1] while keeping the communication safe and fluent as needed in an emergency situation. To this end, the INGENIOUS MOA creates a communication channel for the leaders in the field of each FRs team, consisting of Automatic Speech Recognition (ASR), Machine Translation (MT), and Text-to-Speech (TTS) services integrated within a Telegram Bot.

2. System Description

The core of INGENIOUS is the Command, Control and Coordination (C3) unit, a desktop application managed from the central quarters that gives a global view of all the elements involved in a disaster. C3 manages all the resources (personal and non-personal) and ongoing events. The MOA is a smartphone application connected with the C3 to be used for coordination in the field of different team leaders.

2.1. Telegram Bot

The MOA had to be robust, and easy to install by each team leader. To ensure that the application would support a wide range of devices and operating system versions, we decided to base the application on Telegram[1], a cloud-based instant messaging freeware software. The service is cross-platform, provides end-to-end encrypted VoIP and an API to develop telegram Bots (a software operated account).

The INGENIOUS Bot connects with the C3 unit to get information of the ongoing events and the FRs assigned to each of them. The bot creates a virtual multilingual communication channel for each event. The users connected to the Bot are then assigned to the channel corresponding to the event where they have been assigned in the C3 unit.

The purpose of the communication channel is that users speak (send messages) in their own language and receive other's messages in their own language or in English, when the language of both users is different. The bot also has other capabilities, from which the most prominent are: sending both audio and transcriptions, remembering last messages (so new users can get up to date quickly) and sending them to the C3 unit. These messages can be logged and monitorized by the coordinators. The linguistic capabilities of the bot are provided by specific ASR, MT and TTS web services

2.2. Automatic Speech Recognition

Speech is transcribed into text by means of an ASR system based on deep learning, trained with the wav2letter++/Flashlight [2] toolkit. Specifically, the acoustic model transcribing the Spanish language is mainly constituted by convolutional neural networks with gated linear units , trained with around 200 hours from the Common Voice corpus [3]. Regarding the French model, it is a Transformer-based model trained with 1000 hours from the Multilingual LibriSpeech dataset [4]. Language models for both languages are built with the text from the training corpora, plus additional text from the Gutenberg project in the case of French.

2.3. Machine Translation

Machine translation leverages the ModernMT framework [5], which is to some extent a "convenience layer" on top of state-of-the-art machine translation frameworks, adding functionality, ease-of-use, and production-readiness. This allows us to stay up-to-date with state-of-the-art MT algorithms, even across different underlying frameworks, while maintaining a stable API. E.g., ModernMT switched from a TensorFlow-based Transformer model to Facebook's Fairseq in their latest release (4.x).

[1]https://telegram.org/

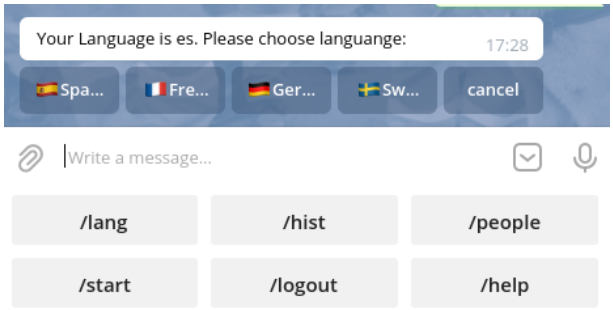

Figure 1: *Global menu and language selection buttons.*

We trained the translation models on openly available data from OPUS[2], in particular the OpenSubtitles and MultiUN corpora, providing good coverage of general-domain speech. Additional domain-specific vocabulary such as emergency codes can be injected dynamically through the use of specific translation memories.

2.4. Speech Synthesis

To generate synthetic speech we integrated PyTorch versions of Tacotron2 [6] and Multi-Band MelGAN (MB-MG) [7] models. The former was trained with some modifications with repect to the original publication: (1) we substituted dropout in the prenet module by batch normalization, as dropout required to be activated also in the inference originally, causing random behavior in the synthesis, and (2) we added a so called Double Decoder Consistency[3] as it is shown that helps Tacotron2 alignment even for longer sentences without losing quality. A pretrained MB-MG model was used as a vocoder. From our experiments, this neural vocoder performs 55% faster than its antecesor, MelGAN, both on CPU and on GPU. So the synthetic response is much faster. Both models were trained using the open-access LJSpeech dataset[4].

2.5. App Interface

The application, takes the form of a chat from the user point of view. It includes some simple buttons to: select the language, get last messages, list of connected people, start the communication with the bot, logout and get help (Figure 1). Apart from this set of basic commands, the user can send audio messages that the bot will broadcast to all the users, and receive audio messages from other users. Figure 2 shows a possible sequence of message interchanges between two users.

3. Conclusions and Future Work

We have presented the integration of a speech-to-speech translation service with the messaging platform Telegram. The service is thought to ease the international collaboration between First Responders on big disasters. The tool is simple and easy to use for people that already use Telegram or similar messaging platforms. The current version of the system includes some general purpose linguistic models. As a future work we are working on adapting the these models to the FR needs and domain, including: the expansion of emergency codes to natural language, the extraction of geographical names of the area where the disaster happens and their integration in the vocabulary to be recognized by the ASR and MT systems and finally to train the ASR in a noisy environment as the one that the users may be working it.

[2]https://opus.nlpl.eu/

[3]https://erogol.com/solving-attention-problems-of-tts-models-with-double-decoder-consistency/

[4]https://keithito.com/LJ-Speech-Dataset/

Figure 2: *Data flow between different users. User on the left sends a message in Spanish (1), then it receives the transcription (2) while the other user (on the right) gets the English translation and transcription (3). The right user then sends a message in French (4) that is transcribed by the system (5) and broadcasted to Spanish users in English (6).*

4. Acknowledgements

This work is part of the INGENIOUS project, funded by the European Union's Horizon 2020 Research and Innovation Programme and the Korean Government under Grant Agreement No 833435. The last author has been funded by the Agencia Estatal de Investigación (AEI), Ministerio de Ciencia, Innovación y Universidades and the Fondo Social Europeo (FSE) under grant RYC-2015-17239 (AEI/FSE, UE). This work has been carried out using an NVIDIA GPU Titan Xp generously provided by NVIDIA Company.

5. References

[1] B. W. Dauksewicz, "Hospitals should replace emergency codes with plain language," *Journal of Healthcare Risk Management*, vol. 38, no. 3, pp. 32–41, 2019. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/jhrm.21346

[2] V. Pratap, A. Hannun, Q. Xu, J. Cai, J. Kahn, G. Synnaeve, V. Liptchinsky, and R. Collobert, "wav2letter++: The fastest open-source speech recognition system. arxiv 2018," *arXiv preprint arXiv:1812.07625*.

[3] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," *arXiv preprint arXiv:1912.06670*, 2019.

[4] V. Pratap, Q. Xu, A. Sriram, G. Synnaeve, and R. Collobert, "Mls: A large-scale multilingual dataset for speech research," *arXiv preprint arXiv:2012.03411*, 2020.

[5] U. Germann, E. Barbu, L. Bentivogli, N. Bertoldi, N. Bogoychev, Buck *et al.*, "Modern mt: a new open-source machine translation platform for the translation industry," in *Proceedings of EAMT-2016 conference: Projects/Products*, 2016.

[6] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. J. Skerry-Ryan, R. A. Saurous, Y. Agiomyrgiannakis, and Y. Wu, "Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions," *CoRR*, vol. abs/1712.05884, 2017.

[7] G. Yang, S. Yang, K. Liu, P. Fang, W. Chen, and L. Xie, "Multi-band melgan: Faster waveform generation for high-quality text-to-speech," *CoRR*, vol. abs/2005.05106, 2020. [Online]. Available: https://arxiv.org/abs/2005.05106