A Modular Approach for Romanian-English Speech Translation

Andrei-Marius Avram[0000−0001−7045−038X], Vasile Păiș[0000−0002−0019−7574], and Dan Tufiș[0000−0002−8280−9852]

Research Institute for Artificial Intelligence, Romanian Academy
{andrei.avram,vasile,tufis}@racai.ro

Abstract. Automatic speech to speech translation is known to be highly beneficial in enabling people to directly communicate with each other when they do not share a common language. This work presents a modular system for Romanian to English and English to Romanian speech translation created by integrating four families of components in a cascaded manner: (1) automatic speech recognition, (2) transcription correction, (3) machine translation and (4) text-to-speech. We further experimented with several models for each component and present several indicators of the system's performance. Modularity allows the system to be expanded with additional modules for each of the four components. The resulting system is currently deployed on RELATE and is available for public usage through the web interface of the platform.

Keywords: Speech translation · Romanian-English · Bidirectional · Cascaded System.

1 Introduction

The recent significant advances in automatic speech recognition (ASR), machine translation (MT) and text-to-speech (TTS) have been mainly driven by the development of deep learning models, higher computational power and greater data availability. These advancements have also aroused interest in converging them and creating more efficient speech to speech translation (S2ST) systems, thus further breaking down communication barriers between people that do not speak the same language.

However, the S2ST problem is far from being solved and there are currently two methods in approaching it: (1) cascaded systems and (2) end-to-end models. Although cascaded systems still outperform end-to-end models [11], they usually propagate the error from one module to another, making the whole system brittle and hard to analyse. End-to-end S2ST models do not have this issue and there are recent developments that try to shrink the gap between the two [13]. Yet, their performance is limited by the lack of speech translation datasets in comparison to the rich resources that are available for each individual field: ASR, MT or TTS.

The Romanian resources available for S2ST are quite scarce, being far from enough for training a competitive end-to-end model. Thus, in the context of the ROBIN project[1], we opted to create a cascaded system for Romanian to English and English to Romanian speech translation, by combining other existing components in a modular framework, allowing a similar low-latency S2ST mechanism. Our main contribution is the creation of this open source framework which allows different modules to be easily integrated into the system. We further analysed the end-to-end latency of various combinations of models and found out that in some cases, the whole system can obtain a near real-time performance, with a response time of around one second for Romanian to English speech translation. The system was also integrated in the RELATE platform [17].

The rest of the paper is organised as follows. The next section presents a review of the cascaded and end-to-end models, and of the previous S2ST systems for Romanian-English speech translation. Section 3 presents the models used for each module and how they were integrated into the RELATE platform. Finally, the paper ends with the conclusion and possible directions for future work in the Section 4.

2 Related Work

Extensive research was put into combining different modules within cascaded S2ST and some of the early work on speech translation used an ASR followed by a MT module [14]. However, this kind of approach makes the MT access the errors produced by the ASR and in [19] the authors propose to integrate the acoustic and the translation modules into a transducer that can decode the translated text directly from the audio signal. In addition, because a cascaded system is not naturally capable of maintaining the paralinguistic information, [1] proposed a model that can find the F0-based prosody features in an unsupervised manner and transfer the intonation to the synthesized speech.

One of the earliest attempts to create an end-to-end speech translation system was proposed by [7]. The model obtained a worse performance than a cascaded system, but since then several methods have been applied in order to boost their accuracy from which we can enumerate pre-training, multitask learning and attention passing [10, 6]. In [13] the authors showed that by combining multitask training with synthetic data, the model slightly underperforms a cascaded baseline for Spanish-English S2ST.

The bidirectional Romanian-English speech translation has been also attempted in [9] by using a cascaded approach composed of three modules (ASR, MT, TTS) with additional textual corrections like spell checking and diacritic restoration of the ASR transcript, or letter-to-sound conversion and syllabification of the MT translation. They used the Google ASR API for transcribing the audio signal, and an in-house developed MT and TTS. However, their approach

[1] http://aimas.cs.pub.ro/robin/en/

was highly coupled and the system was not able to easily adapt to new modules as they became available with improved performance.

3 System Overview

To approach the problem of S2ST, we used a cascaded system composed of four modules: ASR, textual correction (TC), MT and TTS. Each module contains one or more configurable models for both Romanian and English. This type of architecture allows us to easily integrate new modules and models into the platform and also to select a specific pipeline with respect to a potential problem demands. The overall architecture is depicted in Figure 1. From the four modules, the TC can be skipped and it is marked with a dotted arrow.

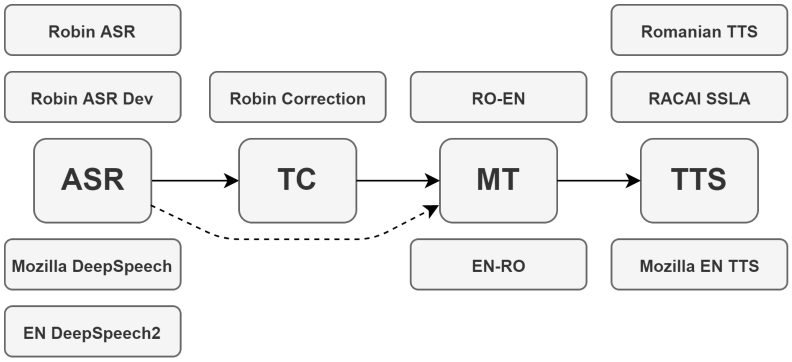

Fig. 1. The proposed S2ST cascaded architecture with the four modules. The Romanian models are depicted in the upper part and the English models in the lower part.

3.1 Modules Description

Automatic Speech Recognition The models used for the Romanian speech recognition use deep neural networks and their architecture were based on DeepSpeech2 [2]. The models were developed in the ROBIN project in order to improve the transcription latency [3] of the system of that period. They also obtained a satisfying word error rate (WER) of 9.91% on a customized test set when combined with a language model. We provide two variants for this module: (1) the base version - `RobinASR` - presented in [4] and a development version that we continue to improve - `RobinASR Dev`.

For transcribing the English speech, we used the latest version of the speech-to-text model offered by Mozilla[2] that was based on DeepSpeech [12] - `Mozilla`

[2] https://github.com/mozilla/DeepSpeech

`DeepSpeech` -, and also a DeepSpeech2 model[3] that was trained on LibriSpeech [15] - `EN DeepSpeech2`. Although the DeepSpeech2 architecture provided a deeper neural network with more parameters for each layer, `Mozilla DeepSpeech` turned out to be better and outperformed `EN DeepSpeech2` with almost 3% on the LibriSpeech clean test set, obtaining a WER of 7.06%.

Transcription Correction We currently offer only a version for Romanian textual correction that consists of (1) capitalizing the first letter of words from the transcription that are present on a known named entity list and (2) replacing the words with words from a vocabulary. In addition, we also employ a hyphen restoration for the `RobinASR` variant based on bi-gram and uni-gram statistics.

Machine Translation Both Romanian to English and English to Romanian MT systems are based on eTRANSLATION platform that was additionally trained and enhanced with a neural network layer, under the coordination of TILDE in the project "CEF Automated Translation toolkit for the Rotating Presidency of the Council of the EU", TENtec no. 28144308. The translation module is a component of a larger system for the Presidency of the Council of the European Union[4].

Text-to-Speech The English version of the TTS uses the pretrained Tacotron2 with Dynamic Convolution Attention [5] offered by MozillaTTS[5] - `Mozilla EN TTS`. The system obtained a median opinion score (MOS) naturalness of 4.31±0.06 with a 95% confidence interval.

To synthesize the Romanian speech, we integrate two models in our pipeline: (1) `Romanian TTS` developed in [18] and (2) `RACAI SSLA` developed in [8], that are based on Hidden Markov Models (HMM) to compute the most probable sequence of spectrograms. However, they offer a trade-off between speed and speech quality, with the `Romanian TTS` being the version that is slower, but with a higher quality of synthesis (3.15±0.73 MOS with a 95% confidence interval) and the `RACAI SSLA` being the version that is faster but with a lower quality of the produced speech.

3.2 RELATE Integration

All the modules are implemented as server processes, exposing their specific functionality as HTTP-based APIs. This allows for hosting the modules on different computing nodes and then integrate them via API calls into a single, unified framework. Furthermore, in order to allow easy interaction with the aggregated pipeline for speech to speech translation, we integrated it in the RELATE platform. We followed the approach described in [16] that allowed us to develop a platform component invoking each module as needed.

[3] https://github.com/SeanNaren/deepspeech.pytorch
[4] https://ro.presidencymt.eu/#/text
[5] https://github.com/mozilla/TTS

The user is offered the possibility to either record in real-time using a microphone and have it translated or start by uploading an already existing sound file. Furthermore, the user is in complete control of the processing chain, being able to select for each step the desired module. However, default settings are pre-loaded, thus the user may use the framework without having to consider individual modules. Only modules available according to the user's choice of primary language are presented in the interface, as depicted in Figure 2 for English to Romanian S2ST.

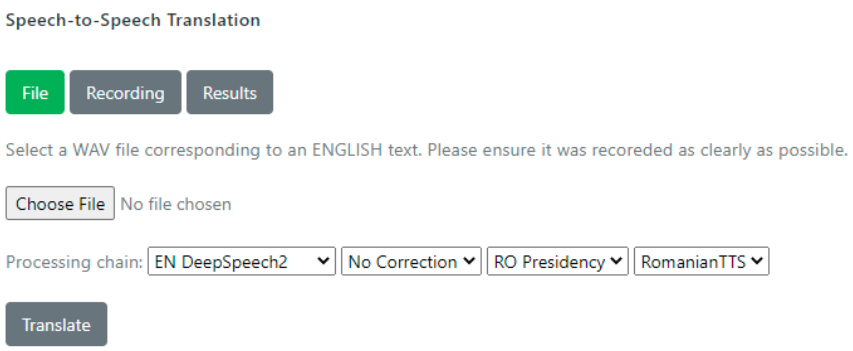

Fig. 2. Web interface for choosing pipeline modules.

When a translation process is started, the platform will call all the selected modules in order and aggregate the results. When this process is done, the user is presented with the final synthesized sound, obtained from the selected TTS module, and intermediate texts, obtained from the ASR and translation modules. The average response time of the whole cascaded system is around 1 second for Romanian to English and around 5 seconds for English to Romanian[6], while audio waves with less than 10 seconds are given as inputs.

4 Conclusions

This paper presented our work on creating a modular system for bidirectional Romanian-English speech translation that is composed of four modules that are put in a cascaded manner. Each module comes with a series of configurable models that allows a higher flexibility in choosing a specific processing pipeline. Furthermore, our architecture can be easily scaled by integrating new modules and models into the cascaded system. The whole system and its components were made publicly available for use on the RELATE platform[7].

[6] This slow down in latency is mostly caused by the Romanian TTS models that are based on HMMs.

[7] RO → EN: https://relate.racai.ro/index.php?path=translate/speech_ro_en
EN → RO: https://relate.racai.ro/index.php?path=translate/speech_en_ro

One direction for possible future work is to develop and integrate a neural based TTS for the Romanian language in order to reduce the latency of the current component, without compromising the speech synthesis quality. Another possible work is to make the source speech and target speech sound more alike by transferring the intonation from the source speech to the target speech.

5 Acknowledgement

This work was realized in the context of the ROBIN project, a 38 months grant of the Ministry of Research and Innovation PCCDI-UEFISCDI, project code PN-III-P1-1.2-PCCDI-2017-734 within PNCDI III.

References

1. Aguero, P., Adell, J., Bonafonte, A.: Prosody generation for speech-to-speech translation. In: 2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings. vol. 1, pp. I–I. IEEE (2006)
2. Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., Casper, J., Catanzaro, B., Cheng, Q., Chen, G., et al.: Deep speech 2: End-to-end speech recognition in english and mandarin. In: International conference on machine learning. pp. 173–182. PMLR (2016)
3. AVRAM, A.M., PĂIȘ, V., TUFIȘ, D.: Romanian speech recognition experiments from the robin project. ISSN 1843-911X p. 103
4. Avram, A.M., Vasile, P., Tufis, D.: Towards a romanian end-to-end automatic speech recognition based on deepspeech2. In: Proc. Rom. Acad. Ser. A. vol. 21, pp. 395–402 (2020)
5. Battenberg, E., Skerry-Ryan, R., Mariooryad, S., Stanton, D., Kao, D., Shannon, M., Bagby, T.: Location-relative attention mechanisms for robust long-form speech synthesis. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 6194–6198. IEEE (2020)
6. Bérard, A., Besacier, L., Kocabiyikoglu, A.C., Pietquin, O.: End-to-end automatic speech translation of audiobooks. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 6224–6228. IEEE (2018)
7. Bérard, A., Pietquin, O., Servan, C., Besacier, L.: Listen and translate: A proof of concept for end-to-end speech-to-text translation. arXiv preprint arXiv:1612.01744 (2016)
8. Boros, T., Dumitrescu, S.D., Pais, V.: Tools and resources for romanian text-to-speech and speech-to-text applications. arXiv preprint arXiv:1802.05583 (2018)
9. BOROȘ, T., TUFIȘ, D.: Romanian-english speech translation. Proceedings of the Romanian Academy, Series A **15**(1), 68–75 (2014)
10. Duong, L., Anastasopoulos, A., Chiang, D., Bird, S., Cohn, T.: An attentional model for speech translation without transcription. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 949–959. Association for Computational Linguistics, San Diego, California (Jun 2016). https://doi.org/10.18653/v1/N16-1109, https://www.aclweb.org/anthology/N16-1109

11. Federico, M., Waibel, A., Knight, K., Nakamura, S., Ney, H., Niehues, J., Stüker, S., Wu, D., Mariani, J., Yvon, F. (eds.): Proceedings of the 17th International Conference on Spoken Language Translation. Association for Computational Linguistics, Online (Jul 2020), https://www.aclweb.org/anthology/2020.iwslt-1.0
12. Hannun, A., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., Prenger, R., Satheesh, S., Sengupta, S., Coates, A., et al.: Deep speech: Scaling up end-to-end speech recognition. arXiv preprint arXiv:1412.5567 (2014)
13. Jia, Y., Weiss, R.J., Biadsy, F., Macherey, W., Johnson, M., Chen, Z., Wu, Y.: Direct speech-to-speech translation with a sequence-to-sequence model. Proc. Interspeech 2019 pp. 1123–1127 (2019)
14. Ney, H.: Speech translation: Coupling of recognition and translation. In: 1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No. 99CH36258). vol. 1, pp. 517–520. IEEE (1999)
15. Panayotov, V., Chen, G., Povey, D., Khudanpur, S.: Librispeech: an asr corpus based on public domain audio books. In: 2015 IEEE international conference on acoustics, speech and signal processing (ICASSP). pp. 5206–5210. IEEE (2015)
16. Păiș, V., Tufiș, D., Ion, R.: Integration of romanian nlp tools into the relate platform. In: International Conference on Linguistic Resources and Tools for Natural Language Processing (2019)
17. Păiș, V., Tufiș, D., Ion, R.: A processing platform relating data and tools for romanian language. In: Proceedings of The 12th Language Resources and Evaluation Conference. pp. 81–88. European Language Resources Association, Marseille, France (May 2020), https://lrec2020.lrec-conf.org/media/proceedings/Workshops/Books/IWLTP2020book.pdf
18. Stan, A., Yamagishi, J., King, S., Aylett, M.: The romanian speech synthesis (rss) corpus: Building a high quality hmm-based speech synthesis system using a high sampling rate. Speech Communication **53**(3), 442–450 (2011)
19. Vidal, E.: Finite-state speech-to-speech translation. In: 1997 IEEE International Conference on Acoustics, Speech, and Signal Processing. vol. 1, pp. 111–114. IEEE (1997)