TOWARDS END-TO-END SPEECH-TO-TEXT SUMMARIZATION

Trường Đại học Công nghệ Thông tin
(University of Information Technology – UIT)

Tác giả: NGUYỄN HOA

What ?

We propose new improved models to summarize a long speech to text as following:

- To design and propose an **End-to-end models for Speech-to-text (S2T) summarization** that use transformer or transfer learning methods
- To conduct experiments on languages of spoken documents, such as instructional videos, news broadcasts, and podcasts, and compare with conventional **SOTA ASR/T2T cascade models** and baselines.

Why ?

- S2T summarization can help users access and digest large amounts of audios in a time-efficient manner to summerize of spoken documents, such as podcasts, lectures, or news broadcasts, …
- Current **SOTA ASR/T2T models** introduces errors and inefficiencies due to the mismatch between the speech and text domains.
- **E2E modeling of S2T summarizer** is a promising alternative but how can its performance compare to the conventional **SOTA ASR/T2T models ?**

Overview

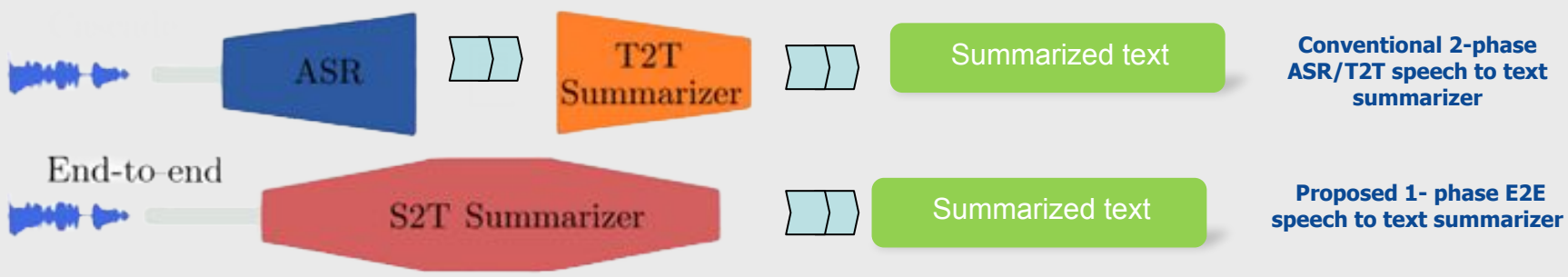

Figure 1. The Conventional text summarizer infrastructure and proposed one.

Description

Speech to text summarization

Speech-to-text (S2T) summarization is a task that aims to address this need by producing concise and informative summaries of spoken documents. S2T summarization can be seen as a combination of two subtasks: speech recognition, which converts speech signals into text transcripts; and text summarization, which generates summaries from text inputs.

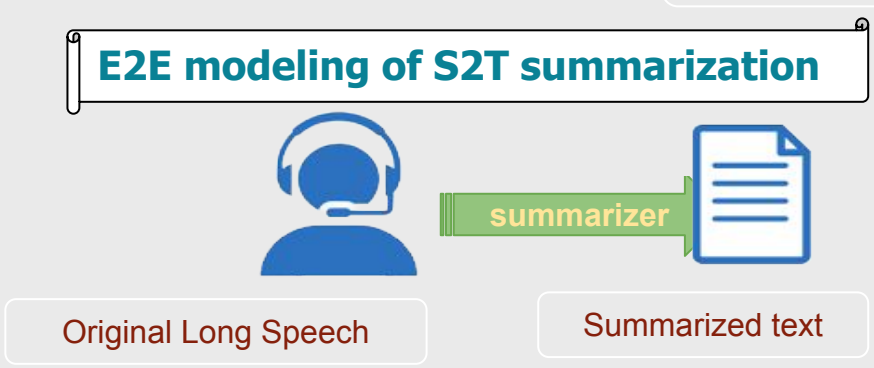

Figure 3. Speech to text E2E summarizer

End-to-end (E2E) modeling of S2T summarization *is a promising alternative which can potentially leverage acoustic and prosodic features from speech, such as pitch, intensity, or pauses; as well as learn rich latent representations that capture the semantic and pragmatic aspects of the spoken content. Moreover,* ***E2E models*** *can avoid the errors propagated by speech recognition models, and optimize for the final summarization objective.*

ASR/T2T S2T summarization models

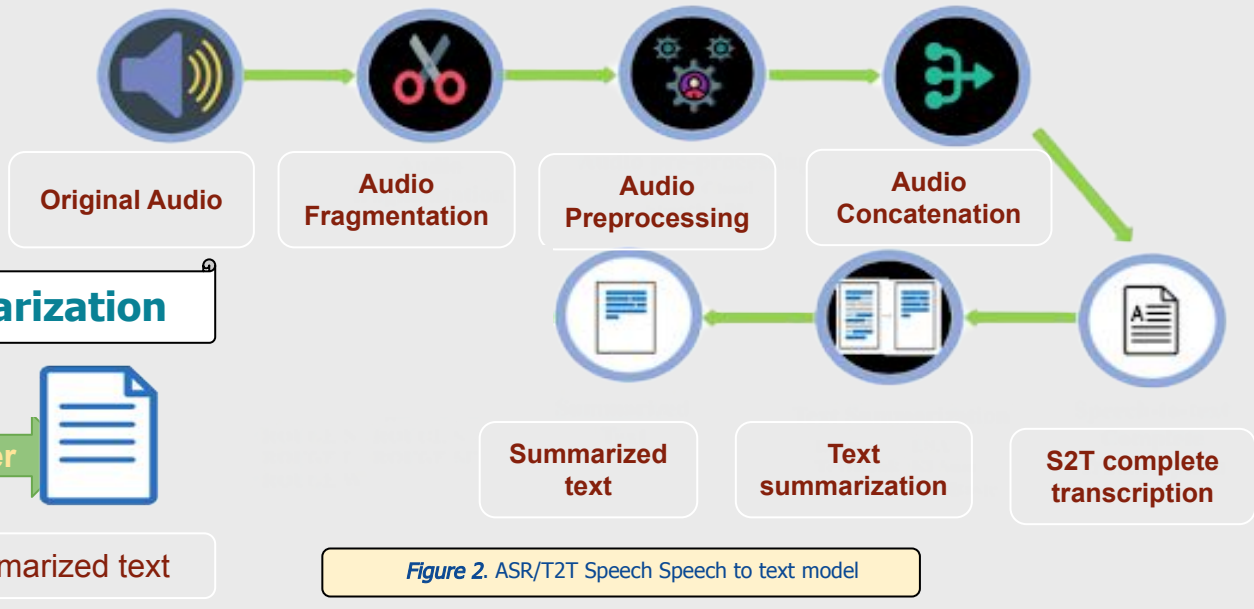

Figure 2. ASR/T2T Speech Speech to text model

S2T content summarization systems use the ***ASR/T2T cascad model*** *which uses an automatic speech recognition (ASR) model to generate intermediate dialogue, in series with a model. text-to-text (T2T) summaries to generate speech content summaries. They are trained on two separate datasets to summarize the content of the dialogue. However, this pipeline introduces errors and inefficiencies and our* ***End-to-end (E2E) modeling of S2T summarization*** *is a promising alternative.*

Nguyễn Hoa– Trường Đại học Công nghệ Thông tin (University of Information Technology – UIT)
TEL : 84.909.517.174 **Email : hoan.17@grad.uit.edu.vn**