ViVoLAB System Description for the S2TC IberSPEECH-RTVE 2022 challenge

Antonio Miguel, Alfonso Ortega, Eduardo Lleida

ViVoLab, Aragon Institute for Engineering Research (I3A), University of Zaragoza, Spain
amiguel@unizar.es, ortega@unizar.es, lleida@unizar.es

Abstract

In this paper we describe the ViVoLAB system for the IberSPEECH-RTVE 2022 Speech to Text Transcription Challenge. The system is a combination of several subsystems designed to perform a full subtitle edition process from the raw audio to the creation of aligned subtitle transcribed partitions. The subsystems include a phonetic recognizer, a phonetic subword recognizer, a speaker-aware subtitle partitioner, a sequence-to-sequence translation model working with orthographic tokens to produce the desired transcription, and an optional diarization step with the previously estimated segments. Additionally, we use recurrent network based language models to improve results for steps that involve search algorithms like the subword decoder and the sequence-to-sequence model. The technologies involved include unsupervised models like Wavlm to deal with the raw waveform, convolutional, recurrent, and transformer layers. As a general design pattern, we allow all the systems to access previous outputs or inner information, but the choice of successful communication mechanisms has been a difficult process due to the size of the datasets and long training times. The best solution found will be described and evaluated for some reference tests of 2018 and 2020 IberSPEECH-RTVE S2TC evaluations.

Index terms should be included as shown below.

Index Terms: Automatic speech recognition, Recurrent neural networks, Sequence-to-sequence models

1. Introduction

The advances in automatic speech recognition (ASR) in recent years and the current quality of state of the art systems have made possible a number of applications like automatic subtitling of multimedia contents, audio indexation among others

2. System description

2.1. Front End

2.2. Phoneme recognition

2.3. Phonetic subwords

2.4. Subtitle partitioner

2.5. Orthographic token based recognizer

2.6. Diarization subsystem

3. Experiments

4. Conclusions

Authors must proofread their PDF file prior to submission to ensure it is correct. Authors should not rely on proofreading the Word file. Please proofread the PDF file before it is submitted.

5. Acknowledgements

This work was supported in part by the European Union's Horizon 2020 research and innovation program under Marie Skłodowska-Curie Grant 101007666; in part by MCIN/AEI/10.13039/501100011033 and by the European Union "NextGenerationEU" / PRTR underGrants PDC2021-120846-C41 PID2021-126061OB-C44, and in part by the Government of Aragon(GrantGroup T36 20R).

6. References