Meta

Aug 22, 2023

Vimal Manohar
Meta Platforms Inc.
USA

**Doctoral thesis**: Semi-Supervised Speech-To-Text Recognition With Text-To-Speech Critic

**Name of the doctoral student**: Murali Karthick Baskar

Mr. Karthick's thesis explores semi-supervised Automatic Speech Recognition (ASR) using unpaired speech and text. It presents an end-to-end differentiable framework consisting of ASR and TTS models that can be trained on unpaired data, explores various architectures, objective functions, and evaluates the approaches on different data domains. Using the relatively cheap data unpaired speech and text data is an increasingly important topic for training large-scale neural networks in low-resource domains and languages.

The thesis has 7 chapters including introduction and conclusion, and 89 pages in total. This review first deals with the technical content of the thesis, then summarizes its technical quality, comments on the formal points and finally presents the overall conclusion and recommendation to the committee.

## Technical content of the thesis and remarks to chapters

Chapter 1 briefly introduces the field of ASR, and one of its core challenges regarding use of large amount of cheap unpaired speech and text data to improve its models. It presents the motivation for the thesis and puts it in the context of recent related works addressing these problems.

Chapter 2 introduces the seq2seq ASR modeling paradigm and contrasts it with the hybrid model paradigm. The chapter is a good introduction to concepts used in the thesis, which is focused on CTC and AED models. The chapter then somewhat digresses to concepts of learning from more diverse data consisting of unpaired speech and text, cross-domain and cross-modal data. This puts the work of this thesis to use ASR and TTS models in the context of related approaches in semi-/self-supervised learning. *I would have liked to see a more deeper literature review on semi-/self-supervised learning approaches and their strengths and weaknesses, to put more context on where the proposed approaches fit in.*

Chapter 3 introduces the proposed ASR<->TTS training paradigm for training on unpaired speech or text data, while Chapter 4 follows up with experimental analysis conducted on the proposed approach. As noted in the chapter, the approach was proposed by the candidate in a 2019 Interspeech paper. But the candidate has performed deeper analysis and presented it in these chapters. The proposed approach is an end-to-end differentiable pipeline consisting of ASR and TTS modules, which are trained using the cycle-consistency loss. The back-propagation through discrete ASR tokens is achieved through a policy gradient using the TTS model loss for the reward. This makes the full pipeline end-to-end differentiable going from ASR to TTS or TTS to ASR. It is nice to see experimentally that this differentiability adds to performance gain compared to simply taking the top hypothesis from ASR for TTS. The chapter also compares with another closely related approach that is also end-to-end differentiable, but only considers speech encodings for consistency loss instead of spectral features. Using the spectral features for consistency loss is shown to be better, but also is simpler because it can just use regular ASR and TTS models instead of having to use a separate encode training. *One clarification that would be helpful is: There is a bias term used in the score function to reduce the variance. What is the impact of this in terms of training stability?*

Chapter 4 presents experimental analysis of the proposed ASR<->TTS approach using simulated paired and unpaired datasets that are different subsets of WSJ and LibriSpeech. The chapter starts with details about the datasets, models and training hyperparameters. The first set of experimental analysis on WSJ shows that the model mostly improves upon adding more unpaired data, and using both text-only and speech-only data, showing the benefits of both sides of the ASR<->TTS pipeline. There is over-fitting seen with speech-only when there is only 2hr of paired data, which is common when using very little supervised data, but it shows that text-only data is more beneficial in such a situation. Some analysis done shows the kind of errors the speech-only training can correct, but also shows that it is susceptible language errors, something that is addressed in the next chapter. Further analysis shows that improvement comes from being able to train the attention module better. *It might be worth to see if only the attention part can be pre-trained from some sources in order to improve the performance on unpaired data.* An interesting result in the chapter is that the ASR<->TTS pipeline requires freezing the CTC part of the multi-task objective in order to get improvements, and training the CTC part on the unpaired data drastically degrades the performance. *I'm curious to know why there is such a degradation when using the CTC part.* The experiments are repeated on Librispeech and compared against other contemporary work in the literature using unpaired data. The proposed approach is shown to have larger improvements than other works in the literature in both text-only and speech-only settings.

Chapter 5 proposes an improvement in the ASR->TTS pipeline used for speech only data to address some of the language errors seen with speech-only data. A language model (LM) is added as a prior to penalize the language errors in the ASR hypotheses. I like that the chapter motivates the use of LM to be similar to the prior in VAE resulting in an objective analogous to the VAE ELBO objective. The chapter presents a brief foray into VAE training and its connection to the ASR->TTS pipeline to motivate the addition of the LM based regularization

term. In eqn. 5.18, the scaler is applied only on the LM term, but not on the ASR term. *What is the reasoning behind this? Why does both the TTS and ASR log-likelihood terms have scale of 1? How does it affect performance?* The results on WSJ and LibriSpeech with unpaired speech show that the proposed usage of LM during training is complementary to using LM in inference using shallow fusion.

Chapter 6 proposes an improvement in the TTS->ASR pipeline used for text only data, when dealing with text that's out-of-domain to the pre-trained model. The chapter is based on a ICASSP 2021 publication by the candidate. The problem is that when such out-of-domain text is used with TTS, the synthesized speech is expected to be poor, and the investigation in the chapter shows that it indeed has a very different distribution from real speech or speech synthesized with in-domain text. The first approach to mitigate the poorer synthesized speech is to scale down the attention context vector to reduce the input from the ASR encoder. This way the model can still benefit by training the ASR decoder on the text. The analysis with Fig 6.3 shows an optimal scaler at 0.3. *One thing that would be helpful is to have a comparison with simpler approaches 1) where only the decoder is trained/fine-tuned on the text or 2) use shallow-fusion or 3) where the context vector is replaced with zeros or a constant similar to Google's MUTE paper (Peidong Wang, Tara Sainath et al.)*

The second approach proposed in the chapter is to add a self-supervised model with a contrastive loss as an auxiliary loss, with the hypothesis that it will aid the TTS to fix imperfections when prediction acoustic contexts. It looks like feat2vec model is also pretrained on only English data. *Are there any intuitions on how this helps even on out-of-domain language? What type of errors does it help fix?* The chapter considers different language data from Swahili as the out-of-domain data. The overall ablation results are presented in Table 6.4, consisting of all the proposed works in this thesis. *I find the table a little hard to follow since there are many results. It would be useful to add more detailed analysis of the observations.* In most of the datasets, the alpha scaler gives a big jump in performance when using text-only data, but it's not the case in the 5hr supervised Swahili setting. Specaug also seems to help in some settings more than others. *Are there any insights on these? I am curious to know how the performance varies with other out-domain-text. The chapter initially mentions that Tedlium can be used as out-of-domain text. Are there any results using this to see how the performance vary with out-of-domain text compared to using Librispeech in Table 6.4? How does the results with Swahili compare with about another language like Pashto which was mentioned in the EAT paper?* The chapter ends with a comparison of the proposed approach against other approaches in the literature on a librispeech benchmark, but some are slightly older works and it might be infeasible to compare with all newer approaches. But the results demonstrate that the proposed approach learning from both speech only and text-only data can compete with the SoTA approaches.

Chapter 7 summarizes the findings of the thesis and the proposed approaches. *In my view, the conclusion could be longer with some more key insights from the experiments, and more outlook for future research in these directions.*

## I. Doctoral Thesis

Appropriateness and Relevance

Scaling the training of ASR models to larger models and new domains is a challenging topic of high interest in the field of Automatic Speech Recognition. Mr. Karthick's thesis explores the problems in this important topic and presents solutions to address these problems through approaches to use unpaired speech and text data.

A summary of the Contributions of the Thesis

The goal of the thesis is to improve ASR performance using unpaired speech and text data by developing an end-to-end trainable pipeline. It tries to keep the pipeline simple by re-using pre-existing models. It aims to explore the pipeline under different data domain settings and improve it for harder out-of-domain settings.

Some of the main contributions of the thesis are:

1) An end-to-end differentiable pipeline, ASR<->TTS, consisting of ASR and TTS modules to train on unpaired speech and text data.
2) In comparison to previous works, this is an improvement through the pipeline being differentiable through the ASR tokens and using cycle-consistency loss through the spectral features and not just speech encodings. This is confirmed through ablative studies in the thesis.
3) The experimental results show that the proposed ASR<->TTS out-performs other approaches in literature for learning from unpaired data.
4) A mathematics basis for adding an LM penalty inspired by VAE to fix language errors and improve the ASR<->TTS pipeline.
5) Enhancements to the pipeline to deal with out-of-domain text data using a contrastive loss through feat2vec model, and an attention scaler. Experiments show that these improvements help to learn from unpaired text better, and achieve results competitive with SoTA approaches.
6) An analysis of the best experimental procedures with the pipeline through architecture, data annealing and augmentation strategies.

Evaluation of the Formal Aspects of the Thesis:

The thesis has a logical structure and is easily readable. It is written in well readable English (except for a few minor typos or issues that do not impact the understandability). The mathematical arguments, figures and tables are mostly well presented and formatted, and showcase the experiments performed and results achieved. Abbreviations and notations are mostly consistent across the thesis. I have highlighted a few minor issues and inconsistencies in the manuscript copy and shared with the candidate.

Quality of Publications

The core of the Mr. Karthick's thesis has been published in two leading conferences in Automatic Speech Recognition namely, the ISCA Interspeech conference and the International Conference on Acoustic Speech and Signal Processing (ICASSP). These publications have been well cited by the research community, receiving nearly 100 citations overall according to Google Scholar. One of these was nominated for a Best Student paper award at Interspeech 2019. This shows that the candidate has significantly contributed to the field of Automatic Speech Recognition, with many researchers building on the candidate's work for using unpaired speech and text to improve speech recognition performance.

## II. Candidate's Overall Achievements

Overall R&D Activities Evaluation:

Apart from the publications that were the core of the Mr. Karthick's thesis, he has also contributed to 19 research publications in leading conferences and journals in Automatic Speech Recognition, including 8 publications as the first author. Overall, these have nearly 500 citations combined according to Google Scholar. He has also engaged in research activities collaborating with multiple groups including at industry through internships and as a visiting researcher at another university. These publications and activities demonstrate Mr. Karthick's abilities as a researcher.

---

## III. Conclusion

I have carefully examined the thesis, and in my opinion, Mr. Karthick's thesis and his overall achievements meet the requirements for the Doctoral degree at the Brno University of Technology.

Baltimore, 22-08-2023

Dr. Vimal Manohar
Research Scientist
Meta Platforms Inc.