Center for Language and Speech Processing

The Johns Hopkins University
3400 N. Charles Street
304 Barton Hall
Baltimore, MD 21218-2686

September 15, 2023

Review of doctoral dissertation
Semi-Supervised Speech-to-Text Recognition with Text-to-Speech Critic,
submitted by **Murali Karthick Baskar** to Brno University of Technology, Faculty of Information Technology

With the advent of end-to-end trained deep neural networks and large-scale machine learning, the inherent limitation of the supervised methods for having the teacher information, i.e., in the context of speech recognition, transcribed speech, became laid bare. This need has become the center of attention of the state-of-the-art research. The proposed thesis follows in this direction and aims to contribute method(s) for reduction of the amount of the transcribed data needed by exploiting unpaired speech and text data (both available in large quantities, if we are not considering implications and limitations of intellectual property and ownership issues) by virtue of using the machinery of cycle consistency methods and Variational Bayes methods. I consider both the subject of the thesis and the chosen research direction to be felicitous.

Technical content of the thesis and remarks to chapters

Formally, the thesis is divided into seven chapters.

The **Chapter 1** introduces the problem at glance, briefly discusses the related works and describes the structure and claims of the Thesis.

The **Chapter 2** is dedicated to laying ground of Sequence-to-sequence ASR (and TTS) including some aspects of learning from unpaired data. It does not attempt to explain GMM-HMM and hybrid approaches to ASR (except brief mention in the preface of the chapter). My general comment on this chapter is that author sacrificed clarity and scientific precision in expressing themselves for the sake of jumping directly to the subject of interest, i.e., sequence-to-sequence ASR.
One example particularly standing out might be the following citation (page 11, first paragraph):

> "*Hybrid systems perform classification either by using the Gaussian Mixture Model (GMM) or Deep Neural Network (DNN) models.*"

To readers from the ASR field, the meaning of the shortcut will be clear but formally it's a very abbreviating and not fully scientifically/terminologically legitimate statement, due to "hybrid systems" having a particular meaning (i.e., systems where neural networks are used to produce features for GMM/HMM

system). The same might be to an extent said about the description of the sequence-to-sequence methods (where the author does not attempt to fully chart the sequence-to-sequence ASR field and ultimately also by including Text-to-speech discussion into a chapter titled "Sequence-to-sequence ASR". These comments might be seen as very pedantic and probably are.

In **Chapter 3**, the cycle consistency concept is discussed in the context of ASR and TTS. The chapter carefully delineates the current research in the area (and references properly other works) and the author's contribution in this area. I consider this chapter very informative and well executed.

Chapter 4 contains corpora description and experiments using the cycle consistency framework introduced in the previous chapter. Authors performs experiments on two English language corpora of different size – Wall Street Journal (~80 hours) and Librispeech (~1000 hours). Both are read speech corpora and thus relatively easy tasks in ASR. For TTS, author is using LJSpeech. Firstly, the pipeline is introduced and described. The pipeline was implemented in ESPNet. Author starts with experiments on WSJ – firstly demonstrating saliency of the adage "there is no data like more data" by training baseline on several amounts of supervised training data and observing systematic improvement with growing amount of training data. Then, the experiments with adding additional unpaired (but in-domain) data follows together with detailed discussion and analysis (and examples) of types of errors and dependency of performance on gender of the speaker and other aspects of sequence-to-sequence systems (attention alignment). Next, the experimental results on Librispeech are introduced, together with baselines and state-of-the-art results from the literature. The proposed method clearly outperforms the earlier-published works of other teams working on the same task and corpus. *I cannot help and observe that the section named "Experiments on Librispeech" contains table containing also baseline and SotA numbers from literature for the WSJ corpus.* The chapter is concluded by a short summary of the results.

Chapter 5 discusses incorporation language model (LM) prior during the training as way for providing additional, important, information about the language (or perhaps more specifically about the surface form of language and orthography). Author choses VAE (variational auto encoder) as a way how to incorporate the LM penalty (for e.g., un-grammatical surface forms) into the cycle consistency training. This was motivated by analysis of the errors done in the previous chapter. After formulating the problem and the training process formally (mathematically), the experiments are presented to (successfully) demonstrate efficacy of this method. *I miss more careful explanation about where the text comes from, especially for the WSJ data and how LM-1 and LM-2 in this chapter relates to the LM-1 and LM-2 in the previous chapter. Is the same LM-1 used both during training and during decoding, or is the LM-1 used during decoding the LM-1 from the Chapter 4? How does the Table 4.5 and Figure 5.3 and Figure 5.4 relate? You mark the method introduced as ASR->TLM – is there a reason why you didn't demonstrate it in full cycle, i.e., ASR<->(TTS/TLM)?*

Chapter 6 introduces closes the full cycle and introduces TTS->AFV method and introduces EAT as a combination of TTS->AFV and ASR->TLM. TTS->AFV and EAT is motivated by low performance of previous methods on out-of-domain data. The topology of the networks is different than previous chapters for performance reason. Experiments are presented on Swahili (using Babel Swahili corpus and the ALFFA corpus) and Librispeech. The chapter feels very hastily put together. *In Section 6.3.3. two experimental setups are introduced – Libri-TO and Swahili-TO. Libri-TO contains new data (from TED dataset, which doesn't even have reference citation. Is it TEDLIUM?) and Swahili comes from the corpus ALFFA (also*

without reference). But it's not clear to me how these relate to the results in Table 6.4 which even states different amount of paired and unpaired data used. Neither Swahili-TO nor Libri-TO is ever mentioned again. Author also states these were used in a manner described in previous chapter (Chapter 3), but more specific pointer could be used (such as Section 3.2). Author states the setups for the experiments are published on GitHub, but no link is provided. The chapter is ended by discussing related work and discussion of the results and in what way they relate to the methods introduced in this chapter.

Chapter 7 contains conclusion and statement of (possible) future work.

Summary on the technical content of the thesis

From the technical point of view, the presented framework (ASR-TLM, TTS-AFV and EAT) is sound, and experiments performed confirm its utility. I commend the author on performing experiments on variety of databases and two different languages, especially for using very difficult Swahili corpus.

Comments on the formal aspects

The thesis is written in a good English (to the extent I, as an L2 speaker, can judge). While the work is structured in formally appropriate way (from simpler to more complex, first introducing the methods to evaluating the performance), on a more detailed level I would appreciate more careful job. The chapters are very independent and feel more like 3 different papers put together and lightly edited. Each chapter introduces new pretrained models, new corpora, and new model topologies and it's sometimes not very easy to follow up the authors train of thought.

The mathematical writing is correct. I found small inconsistencies with respect to which direction the neural network should be drawn and with respect to adhering to the common practices – "figure 3", "chapter 3" and similar should be always written with first word capitalized, i.e., "Figure 3", "Chapter 3" and so on. But perhaps different manual of style was used for writing the thesis? I don't consider it a serious issue. Ditto the choice of "modelling" vs "modeling", albeit the used "modelling" is generally considered UK spelling, but the Thesis is written in US English.

Summary and recommendation

I have carefully examined the doctoral thesis of Mr. Murali Karthick Baskar. Despite my criticism raised above (several points are rather recommendations than critique), in my opinion, it is a solid work that contributes to progress in our research field. I also appreciate his publication record, including substantial body of papers published in top conferences.

To conclude, I propose the committee to accept the Thesis unconditionally as a partial fulfillment of the requirements for granting Mr. Murali Karthick Baskar the Doctoral degree at Brno University of Technology.

In Baltimore, Maryland, USA on 15th of September 2023

Jan Trmal
Center For Language and Speech Processing