# Paper Analysis and Review: Improving Performance of End-to-End ASR on Numeric Sequences

## I. Product Design (Mini Project 2): Speech Processing

[1] Boston University B̈oston Massachusetts
e-mail: olutayo@bu.edu
[2] ...
e-mail: olutayo@knights.ucf.edu [⋆]

October 15, 2019

**ABSTRACT**

*Aims.* To explore, in depth, the solutions offered for Out of vocabulary words in Automatic Speech Recognition systems (ASR). Specifically, to understand the methods used, in the paper to be analyzed, to gain significant improvements in long numeric sequence speech recognition such as including more long numeric data sequences in the training set for the Language Model (LM) processing, performing a pre and post processing of the RNN-T written word hypothesis using neural networks, etc.
*Methods.* Baseline implementation without improving the training set for long numerical sequences is compared with a new model in which the training set now includes long numeric sequences thus improving its predictive accuracy.
*Results.* Neural networks and RNN-T form the basic architectures for all the tests implemented. Also, tests are performed using TTS and FST and results are compared in the paper.

**Key words.** Finite State Transducer (FST) – Accoustic Model (AM) – Language Model (LM) – Pronunciation Model (PM)– Recurrent Neural Network (RNN)

## 1. Introduction

The goal of this report is to do a detailed analysis on state of the art methods in speech processing with a specific focus on automatic speech recognition(ASR). ASR techniques have been of huge interest in speech processing as evidenced by the advent of now ubiquitous speech recognition platforms such as SIRI (Apple), ALEXA (Amazon), CORTANA (microsoft) etc. There is a growing demand for this technology and for better implementation. Challenges ranging from the analog domain(signal integrity preservation and noise effects) to the digital domain(processing models (DTW vs Neural networks vs End-to-end ASR) have been tackled for decades by engineers and scientists and this has led to groundbreaking advancemnst in the field of speech processing. For example, the use of Multi-channel ASR technologies that use arrays of microphones to receive the speech signal have been incredible useful in tackling challenges that arise from speech recognition in noisy environments. Also significant advancements in processing power and machine learning algorithms have paved the way for novel techniques in the implementation of ASR technologies.

The main focus of this paper is an indepth analysis of a paper titled **Improving Performance of End-to-End ASR on Numeric Sequences** by ML researchers at Google. End-to-End ASR has become increasingly adopted as it tackles, quite effectively, some of the problems that arises from the traditional Hidden Markov Models (HMM). The basic architecture for HMMs involve separating the components for training. The acoustics, pronunciation, model etc. are trained seperately as opposed to the end-to-end model where the components are treated holistically and a single training module is used. The challenge this paper seeks to address are the issues that arise from recognizing speech that are classified as Out of Vocabulary words (OOV). A subclass of this class of words include long numeric sequences which is the focus of this paper.

Long numeric Sequences pose a challenge in ASR systems because they tend not to be included in the training data. As such, the obvious solution would be to just include more long numeric sequences in the training data. However, this approach has been tried and there are some issues that arise from doing it this way **?**. The method proposed in this paper involves not just adding more long sequence numeric data to the training set but also applying RNN-T in both the written domain and in the spoken domain. This way errors can be corrected and the overall model can be improved. The written domain correction is made via neural network processing performed after the RNN-T hypothesis (original written text) is taken as an input and the correct version, the output.

## 2. Results, Pros and Cons

The results of the experiment show that the regular FST based models, the predictive accuracy of the model declines with increasingly long numeric sequences, as expected. However, as the corrections are implemented the errors slowly vanish but still remain relatively high compared to non numeric data sets. The baseline model, W0, exhibits errors for long numeric data sequences due to the fact that it still associates the words with non numeric string. The first correction, W1, to the baseline model includes training the model with long numeric sequence data to improve its accuracy. This showed some improvement

[⋆] Secondary Email, Expect Slower responses from the baseline model. Another correction was implemented, W2, where the RNN-T hypothesis (written domain prediction) and the actual output(obtained from the spoken domain) is corrected using a "post-processing" neural network . This, plus the more informed data set, also resulted in an improvement from the base line model and W1 thus making the experiment a success. However, the positive correlation between longer numeric sequences was still present however it was less pronounced in W1 and lesser in W2.

The Pros of this method is that the predictive model becomes more accurate. However, this comes at the cost of more data required for training and more processing power required for the neural network. Also, the errors in the written domain prediction still exist even with all the methods applied to the predictive model. The errors also get progressively worse with longer numeric sequences despite the implemented corrective measures. As such, it is up to the designer to make necessary trade-offs to meet required design specifications.

Another potential drawback of this application is that it makes it harder to perform on device processing for ASR due to the amount of computing involved in the correction process. However, if the application is such that an on device ASR is not required this method presents a great way to ensure better predictive accuracy.

## 3. Conclusions

1. The errors arising from using the traditional FTS method, using the corrective measures implemented in this paper, cannot be entirely removed but can be minimized. Furthermore, the positive correlation between the length of the numeric sequence and the error still persists albeit, attenuated.
2. Further improvement in ASR prediction is possible if there is sufficient processing power and a well trained model (data set includes long numeric sequences).
3. Using Text To Speech (TTS) with neural networks resulted in fewer errors than with FST for long numeric sequences. The performance of the corrective methods implemented in this paper when using non numeric sequences exceeded expectations. The errors in the RRN-T hypothesis were severely reduced.

## References

Baker, N. 1966, in Stellar Evolution, ed. R. F. Stein,& A. G. W. Cameron (Plenum, New York) 333
Balluch, M. 1988, A&A, 200, 58
Cox, J. P. 1980, Theory of Stellar Pulsation (Princeton University Press, Princeton) 165
Cox, A. N.,& Stewart, J. N. 1969, Academia Nauk, Scientific Information 15, 1
Mizuno H. 1980, Prog. Theor. Phys., 64, 544
Tscharnuter W. M. 1987, A&A, 188, 55
Terlevich, R. 1992, in ASP Conf. Ser. 31, Relationships between Active Galactic Nuclei and Starburst Galaxies, ed. A. V. Filippenko, 13
Yorke, H. W. 1980a, A&A, 86, 286
Zheng, W., Davidsen, A. F., Tytler, D. & Kriss, G. A. 1997, preprint