Project Proposal

Speech to Text Engine

Saurabh Mathur
samathur@iu.edu

Ankit Mathur
anmath@iu.edu

1 Abstract

For our final project, we intend to implement a deep-learning based Automatic Speech Recognition (ASR) system that can convert spoken words into computer-readable text. For the baseline, we'll be using the DeepSpeech model[1, 2], which is an example of such a system that uses deep-learning to perform end-to-end speech recognition.

However, since DeepSpeech requires different models for different accents, it suffers from scalability issues. Hence, we want to develop a single model that adapts itself based upon different accents.

In order to achieve this, we intend to use a conditioning layer [3] that combines two neural networks, trained on different features of the dataset, such that the final output is conditioned on the features learned by both the neural networks. We will evaluate the models on Word Error Rate for each accent.

2 Datasets

1. *TIMIT Corpus:* It has 6300 sentences, spoken by speakers from 8 major dialect regions of the United States[6].

2. *The Mozilla Common Voice English dataset:* It has 780 hours of speech and corresponding text labels for 8 accents. Additionally, there is 300 hours of unlabeled speech data (voice.mozilla.org, published under CC-0).

3 Goals

1. Accent classification with uncertainty quantification [5].

2. Speech recognition for American English.

3. Speech recognition with one-hot accent conditioning vector.

4. Speech recognition using accent classifier for conditioning.

References

[1] Hannun, Awni, et al. "Deep speech: Scaling up end-to-end speech recognition." *arXiv preprint arXiv:1412.5567* (2014).

[2] Amodei, Dario, et al. "Deep speech 2: End-to-end speech recognition in english and mandarin." *International conference on machine learning.* 2016.

[3] Perez, Ethan, et al. "Film: Visual reasoning with a general conditioning layer." *Thirty-Second AAAI Conference on Artificial Intelligence.* 2018.

[4] Connectionist Temporal Classification : Labelling Unsegmented Sequence Data with Recurrent Neural Networks Graves, A., Fernandez, S., Gomez, F. and Schmidhuber, J., 2006. *Proceedings of the 23rd international conference on Machine Learning*, pp. 369–376. DOI: 10.1145/1143844.1143891

[5] Gal, Yarin, and Zoubin Ghahramani. "Dropout as a bayesian approximation: Representing model uncertainty in deep learning." *International Conference on Machine Learning.* 2016.

[6] Garofolo, John S. et al. "DARPA TIMIT: : acoustic-phonetic continuous speech corpus CD-ROM, NIST speech disc 1-1.1." (1993).