MIT COMPUTER SCIENCE AND ARTIFICIAL INTELLIGENCE LABORATORY

CSAIL

January 24, 2021

RE: Appraisal of "Discovering Acoustic Units from Speech: A Bayesian Approach" by Lucas Ondel

Dear Committee Members:

It is my pleasure to act as an External Examiner for the doctoral thesis of Lucas Ondel entitled "Discovering Acoustic Units from Speech: A Bayesian Approach." As someone who has worked in the area of unsupervised speech processing for many years, I have a much appreciation for the contributions that Mr. Ondel has made over the course of his doctoral research in developing Bayesian formulations for Acoustic Unit Discovery (AUD), and believe the resulting thesis is a great accomplishment.

Mr. Ondel's thesis has three significant contributions, corresponding to Chapters 2-4. First, in Chapter 2, Mr. Ondel extends the Bayesian AUD formulation of one of my former students, Chia-Ying Lee, in a major way by developing a new inference scheme based on the Variational Bayes framework. This is a very substantial contribution that re-casts the optimization problem and objective function resulting in tremendous computational improvements that allow the model to be trained in parallel on much larger corpora than were previously practical. I found the writeup and mathematical derivations very thorough, and thought the experimentation performed to demonstrate the significance of the results on the standard acoustic-phonetic TIMIT and low-resource MOBSHI corpora to be convincing.

The second major contribution, described in Chapter 3, pertains to a generalized subspace model for representing speech. This work is a natural generalization of prior sub-space representations such as the i-vector and subspace Gaussian mixture models. Mr. Ondel thoroughly derives the GSM formulation and shows how it relates to prior work. He then combines it with a hidden Markov model formulation to show how the combined subspace HMM is able to learn a phonetic subspace. Finally, by combining with a Dirichlet Process he shows how the DP-SHMM model is able to improve upon the AUD experiments on TIMIT and MBOSHI, and is also able to leverage prior phonetic information from four European languages of the GLOBALPHONE corpus and improved AUD in terms of the normalized mutual information (NMI) metric he uses throughout the thesis.

The third contribution of Mr. Ondel's thesis is described in Chapter 4, and involves extending the simple "phone-loop" or unigram model of the original AUD by using a Hierarchical Dirichlet Process (HDP) which allows for the incorporation of sequential phonotactic information, in the form of a bigram language model, to be incorporated into the AUD learning process. As in Chapter 2, Mr. Ondel derives a Variational Bayes formulation of the HDP model, and then combines it with an HMM for the AUD task. In his experiments on TIMIT and the MBOSHI corpora he shows that the HDP-HMM is able to achieve modest but consistent gains over the DP-SHMM, and also examines the role of the relative weight of the phonotactic language model vs the acoustic model.

In terms of the thesis document, I found it to be eminently readable and have no major comments. I spotted a few very minor grammatical errors, but nothing significant. The only suggestion I might have, is that there is very little mention of currently popular self-supervised deep learning methods for speech representation learning, and how they might relate to the Bayesian formulations that are the focus of Mr. Ondel's work. Since I believe they are not incompatible with each other, it may be worth at least acknowledging their existence in the Introduction and/or Conclusions, and providing the reader some perspective on these two approaches. I do not feel this is essential, but it is something that I suspect many readers will wonder about, and is probably worth addressing in some fashion.

From an experimental perspective, Mr. Ondel relied almost exclusively on the NMI metric to quantify the performance of his models. While I have no problem with that, it is worth pointing out that there are other metrics that could be considered as well, perhaps to complement NMI. There are some illustrative comparisons between different models that took the form of figures comparing the results on a single utterance. While interesting, these figures were more qualitative in nature, and it might be possible in the future to try to quantify the differences over a larger corpus. I do not view this as critical for the thesis however.

Overall, I am very impressed by Mr. Ondel's accomplishments, and believe his research makes three significant contributions to the speech community. I recommend that his thesis be accepted for the Doctor of Philosophy degree. Please do not hesitate to contact me should you require any additional information.

Sincerely,

James Glass, Ph.D.,
Senior Research Scientist, Massachusetts Institute of Technology
Principal Investigator, MIT Computer Science and Artificial Intelligence Laboratory
Member, Harvard-MIT Health Sciences and Technology Faculty