Master Thesis Review

Optimizing End-to-End Neural Speaker Diarization for Swedish Customer Service Conversations
Silke Kylberg

Reviewer: Lukas Gardberg

1 Paper Summary

The thesis covers the problem of speaker diarization, which consists of determining when each person in a recorded conversation spoke. More specifically it concerns two-speaker customer service conversations, as well as situations where several people can be speaking at the same time. The data sets used to train the chosen diarization model are created from two single-speaker data sets via simulation, as well as a real-world customer service data set. The overall goal is to evaluate the diarization performance in a real-life setting based mainly on simulated data as large-scale annotations are not readily available, as well as to investigate how multi-language data as well as fine-tuning affect performance.

The family of models considered is limited to End-to-End Neural Speaker Diarization models, and in particular the type of models proposed by Kinoshita et al. These integrate traditional vector clustering methods with transformer-based neural networks in order to achieve better performance for both longer recordings as well as overlapping speech. Such a model is trained on a range of different synthetic data sets, as well as one real, and then evaluated using the Diarization Error Rate (DER) on respective test sets and a separate evaluation set.

From the data set evaluation study the thesis finds that having the same spoken language in the training data as the evaluation data to be critical for good performance without fine-tuning. Furthermore it is highlighted that adding additional data in another language does not impact the model positively, and that there is risk for overfitting when using simulated data. Lastly it is also hypothesized that the model becomes better at distinguishing speakers when training on simulated data compared to real data indicated by lower speaker-confusion rates.

For fine-tuning the best performance is obtained when training on a combined simulated Swedish and English data set, followed by tuning on real data, compared to just training on real data. The increased performance is hypothesized to both stem from a better ability to distinguish between speakers as a result of pre-training, as well as a larger data set size with more variability. Lastly the amount of data needed for fine-tuning is also investigated, where

results indicate that very little data is needed to achieve a performance increase, where the model only marginally improves when using 90 hours of data compared to 10 minutes.

2 Summary of Strengths

The thesis is of particular interest because it investigates the applicability of modern speaker diarization methods on the relatively low-resource language Swedish. This is especially important as the majority of speech-related machine learning research is English-focused, and a larger research focus on smaller languages is needed in order to democratize the development of new technologies. In addition to this, special focus is put on being able to achieve good performance without large amounts of manually annotated data, i.e. using simulated data. This is important because manually annotating data is expensive, which leaves the ability to create sophisticated models based on large annotated data sets to those who can afford it. Being able to use raw data, such as audio, in order to improve the performance of machine learning algorithms is essential both for democratizing the technology, and being able to fully utilize the large amounts of unlabeled data which exists today.

Furthermore, it is experimentally shown that fine-tuning a pre-trained diarization model can successfully increase its performance, which again shows promise for usage on low resource languages. This is taken a step further by experimental results which show that a considerable performance increase can be achieved with very limited amounts of data. It is also useful that data set differences such as utterance lengths is discussed, since it is a factor which can have an impact on performance.

Another strength of the thesis lies in its focus on real-life applications. It mainly uses available public data sets for training, and importantly evaluates the model on data collected in practice in the form of an evaluation set. This is important because only evaluating a machine learning method on an academic data set often times does not provide enough evidence of how the model would generalize to a real-world setting. The thesis also considers different types of diarization errors in the context of a real application which gives a more detailed view of the model's performance. In addition to this, that the thesis considers the models applicability in the real world is key, since it is essential to be able to deploy such technologies towards users in order for them to have an actual impact.

Overall the thesis provides a thorough investigation on how simulated data can be used to effectively create a diarization model. It highlights challenges in applying such a model in the real-world, puts weight on being able to do it using few resources, and demonstrates its effectiveness on a lower-resource language like Swedish.

3 Summary of Weaknesses

What are the concerns that you have about the paper that would cause you to favor prioritizing other high-quality papers that are also under consideration for publication? These could

include concerns about correctness of the results or argumentation, limited perceived impact of the methods or findings (note that impact can be significant both in broad or in narrow sub-fields), lack of clarity in exposition, or any other reason why interested readers may gain less from this paper than they would from other papers under consideration. Where possible, please number your concerns so authors may respond to them individually.

Additional metrics?

good that several parts of DER are used - but are there other metrics which could be useful? It is seldom the case that a single metric captures all aspects of a performance that we want from an ML system.

confidence intervals

often glosses over more technical details which can have a large impact on the performance

No introduction to machine learning, might be hard for readers with no such experience to follow

ethical concerns only touched on briefly - more?

4 Recommended Changes

If you have any comments to the authors about how they may improve their paper, other than addressing the concerns above, please list them here.

Better description of why clustering is needed? And how it works when a network is estimating diarization for one speaker in one block and another in the next. Why does this happen? How does each speaker get assigned to a specific network? Is it just through who speaks first?

Add an overview of all the different data sets used

Maybe a larger section about ethical concerns using recorded data of customers? Legal aspect?

Add a graph of the different models performance on the evaluation set per epoch to get a better view of how quickly the different models generalize? Is it fair to use 60 epochs on 60k vs 40 on 30k and call the 60-epoch one the best one?

Add confidence intervals to results

Add information about what overfitting is?

5 Questions

List any questions you would like the author to clarify

- Why are model parameters of the last 10 epochs averaged? What assumptions are we making when we let a model train with more steps on a larger data set? It is presented in the results that it resulted in better models, compared to just picking performing model on the validation set.
- Since Voice Activity Detection is used to automatically label the simulated data the model is trained on, did you investigate how well the VAD performed? Does it make a lot of mistakes? Was the threshold manually optimized for your specific data? Can you think of any ways to improve the energy-based VAD that might increase the model's performance?
- Assumption of only two speakers per block. Why is this assumption made, and is it true in reality? Can the assumption result in worse performance on the real data set? What other methods are there to combat the problem? Block size is set to 15 seconds.
- Why was 60 epochs used for the larger Libri-Simu-NST-Simu data set, compared to 40 for the rest?
- Was overfitting monitored in any way during training?

In addition to this, you can consider the following issues:

- Limitations and Societal Impact
- Ethical Concerns
- Reproducibility
- Release of Datasets & Software