Topic Title: Exploring Transfer Learning for Arabic Dialect Identification

Student Name: Madeline Younes **Student ID:** z5208494

A. **Problem statement**

Although voice enabled technology is used globally advancements have not benefited all languages equally. With more diverse languages with several dialects limited in advancements due to data scarcity and the complexity that comes from shared features between the dialects. One language that this problem has become apparent is Arabic. Dialectal Identifiers enable more accurate systems in speech recognition and transcription by allowing a speech model tuned for a particular dialect to be selected at the start of the process. In order to build a high performance Arabic Dialectal Identifier, desirable methods should require low amounts of data and leverage existing speech models. However, no work has been done using this modern method of transfer learning and so there is a need to apply contemporary techniques to Arabic Dialectical Identifiers.

B. **Objective**

Implement an Arabic Dialectal Identifier for both four umbrella dialects and its adaptability to a finer grain task of identifying seventeen regional dialects using transfer learning techniques. Assess the performance of using various pretrained models, utterance lengths, amount of training files, the effect of inserting downstream models and finetuning the encoder layers within the pretrained model. Analyse the results, provide conclusions and next steps.

C. **My solution**

Arabic Dialectal Identifier using XLSR Arabic pretrained model, finetuning the encoder layers within the pretrained model.
Fine tuning with dialectal Arabic grouped into four umbrella dialects and then seventeen regional dialects.
Explored inserting DNN and LSTM downstream models into the model structure.

D. **Contributions** (at most one per line, most important first)

Novel application of modern speech representation architecture.
Novel downstream task using pretrained models.
Detailed assessment of the preposed model's performance for both broader and more fine grained Arabic dialect groups.

E. **Suggestions for future work**

Train & test with datasets that contain codeswitching between dialects and languages.
Apply methods to use contextual information to improve performance.
Explore inserting other downstream models architectures into the model.

While I may have benefited from discussion with other people, I certify that this report is entirely my own work, except where appropriately documented acknowledgements are included.

Signature: __*Madeline Younes*_____ Date: 20 /11 / 2022

Pointers

List relevant page numbers in the column on the left. Be precise and selective: Don't list all pages of your report!

5	Problem Statement
5	Objective

Theory (up to 5 most relevant ideas)

7	Introduction into Dialectical Identification
9	Pretrained models
11	Context Network: Transformer Encoder-Decoders
12	Introduction into Dialectal Arabic

Method of solution (up to 5 most relevant points)

23-25	Preposed architecture
25-28	Implementation of architecture
21	Dataset Design

Contributions (most important first)

25	Novel application of modern speech representation architecture.
29	Novel downstream task using pretrained models.
28-40	Detailed assessment of the preposed model's performance for both broader and more fine grained Arabic dialect groups.
41	Summary of key results of novel application

My work

7-9	Description of key metrics
23-25	Preposed architecture
41	Summary of results

Results

41	Summary of key results of novel application

Conclusion

43	Statement of insights gained through the experiments
43	Suggestions for future research

Literature: (up to 5 most important references)

18	[11] Babu et al. 2021
16	[17] Biadsy et al. 2009
15	[38] Lin et al. 2020
20	[59] Shon et al. 2018