STT 811 Project

Yunting Gu, Shuangyu Zhao, Wenting Liu

1. Project Description:

   The /r/ sound is acoustically complex and can serve as a sociolingusitic variable crosslinguistically. In New Zealand English (NZE) (and numerous Englishes worldwide), two sound variables non-prevocalic /r/ and medial /t/ are sociolinguistically meaningful, but hand code whether a target sound in question is /r/ or /t/ is timing consuming and could be inaccurate since the human coder's decision may be affected by their prior linguistic experience. This project aims to test different models' behavior on classify whether a given sound is /r/ or /t/ in New Zealand English. It could create algorithms to aid future sociolinguistic research and inform acoustic studies about how efficiently distinguish the sounds.

2. Dataset Description:
   - Data Dimension:
     - Dataset 1 /r/: 40,614 rows and 217 columns
     - Dataset 2 /t/: 9,888 rows and 137 columns
   - **Target (y):** a categorical feature, it contains two English sociophonetic variables, non-prevocalic /r/ and word-medial intervocalic /t/, based on tokens' acoustic signatures.

3. Timeline:

Calendar week	**Milestone**
03/16 ~ 03/24	Data preprocess
03/25 ~ 04/18	EDA, Feature extraction
	Classifier selection and evaluation
04/18 ~ 04/22	Report write up

Reference:

[1] Villarreal, D. & Clark, L. & Hay, J. & Watson, K., (2020) “From categories to gradience: Auto-coding sociophonetic variation with random forests”, *Laboratory Phonology* 11(1): 6. doi: https://doi.org/10.5334/labphon.216
[2] https://github.com/nzilbb/How-to-Train-Your-Classifier