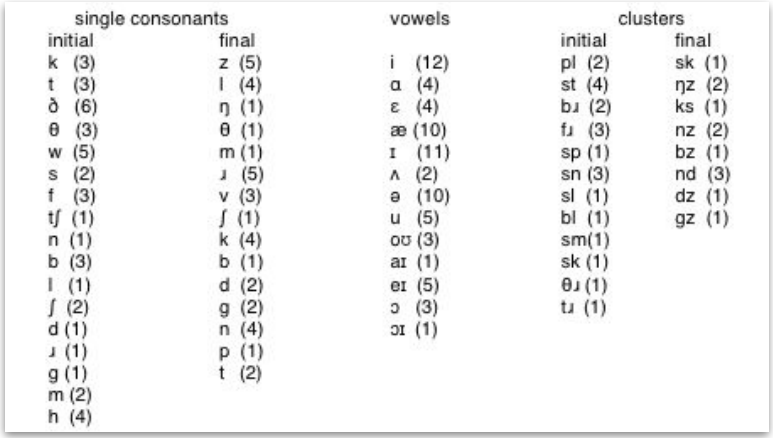
Linguistics for Computer Science

Accent Prediction

Natural Language Processing, University of Bucharest
Sotir Anca-Nicoleta

The task

- A total of 2140 persons of different nationalities have been recorded while reading the same passage of text (written in english)
    - both native and non-native english speakers
    - speakers are from more than 150 different countries
- The ‘Stella’ passage
    - a simple text written in english, formed of only 4 sentences
    - contains most of the consonants, vowels and clusters of standard American English



single consonants		vowels	clusters	
initial	final		initial	final
k (3)	z (5)	i (12)	pl (2)	sk (1)
t (3)	l (4)	ɑ (4)	st (4)	ŋz (2)
ð (6)	ŋ (1)	ɛ (4)	bɹ (2)	ks (1)
θ (3)	θ (1)	æ (10)	fɹ (3)	nz (2)
w (5)	m (1)	ɪ (11)	sp (1)	bz (1)
s (2)	ɹ (5)	ʌ (2)	sn (3)	nd (3)
f (3)	v (3)	ə (10)	sl (1)	dz (1)
tʃ (1)	ʃ (1)	u (5)	bl (1)	gz (1)
n (1)	k (4)	oʊ (3)	sm(1)	
b (3)	b (1)	aɪ (1)	sk (1)	
l (1)	d (2)	eɪ (5)	θɹ (1)	
ʃ (2)	g (2)	ɔ (3)	tɹ (1)	
d (1)	n (4)	ɔɪ (1)		
ɹ (1)	p (1)			
g (1)	t (2)			
m (2)				
h (4)				

Resources:

[Speech Accent Archive (gmu.edu)]

[Speech Accent Archive | Kaggle]

The text passage read by all the speakers:

Please call Stella. Ask her to bring these things with her from the store: Six spoons of fresh snow peas, five thick slabs of blue cheese, and maybe a snack for her brother Bob. We also need a small plastic snake and a big toy frog for the kids. She can scoop these things into three red bags, and we will go meet her Wednesday at the train station.

The text in IPA:

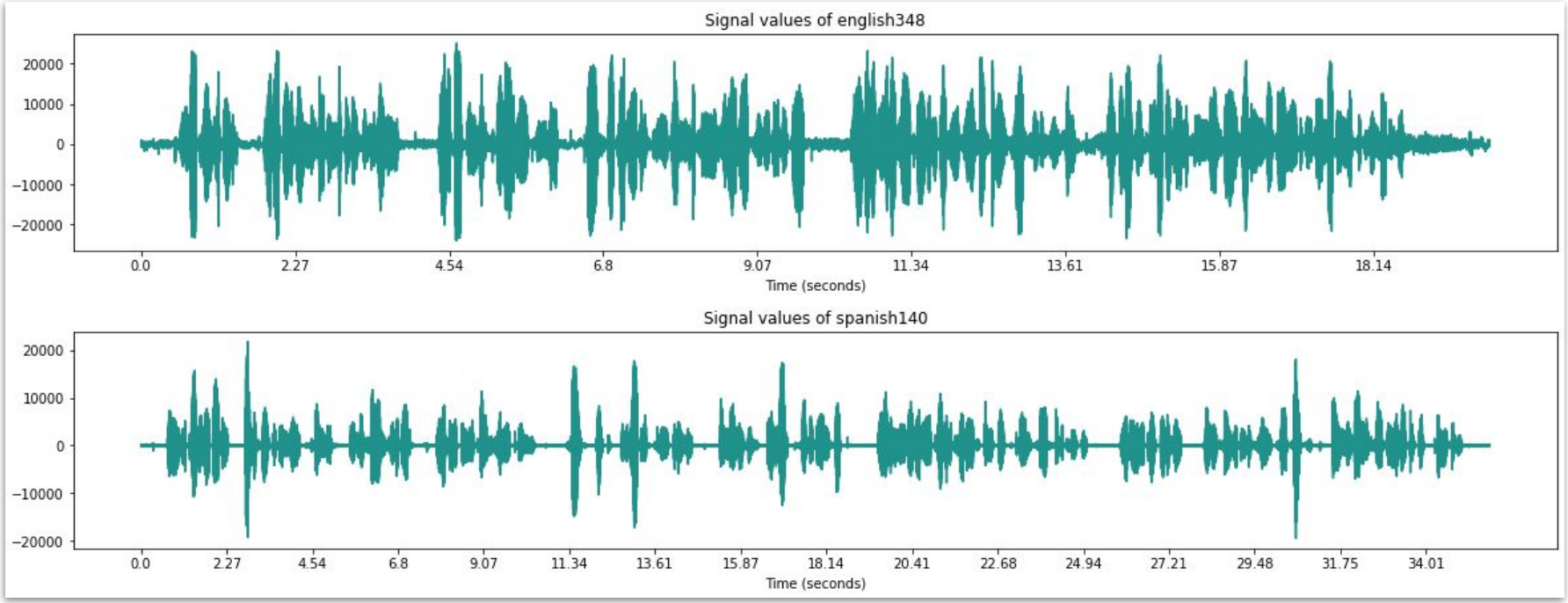
---

pliz kɔl ˈstɛlə. æsk hər tɪ brɪŋ ðiz θɪŋz wɪθ hər frəm ðə stɔr: sɪks spunz əv frɛʃ snoʊ piz, faɪv θɪk slæbz əv blu ʧiz, ənd ˈmeɪbi ə snæk fər hər ˈbrəðər bɑb. wi ˈɔlsoʊ nid ə smɔl ˈplæstɪk sneɪk ənd ə bɪg tɔɪ frɑg fər ðə kɪdz. ʃi kən skup ðiz θɪŋz ˈɪntu θri rɛd bægz, ənd wi wɪl goʊ mit hər ˈwɛnzˌdeɪ æt ðə treɪn ˈsteɪʃən.

Dataset description

- A total of 2138 .mp3 files, each containing the recording of a person reading the passage
  - different file size, different audio signal lengths
- A .csv file containing information about the speakers (age, gender, country, native language, the age they began speaking english)
- The files belong to a total of 200 different accents
- The dataset is very unbalanced:
  - many samples for some accents (namely english)
  - as little as one sample per class in some cases (more than 150 accents with less than 10)

Data visualization

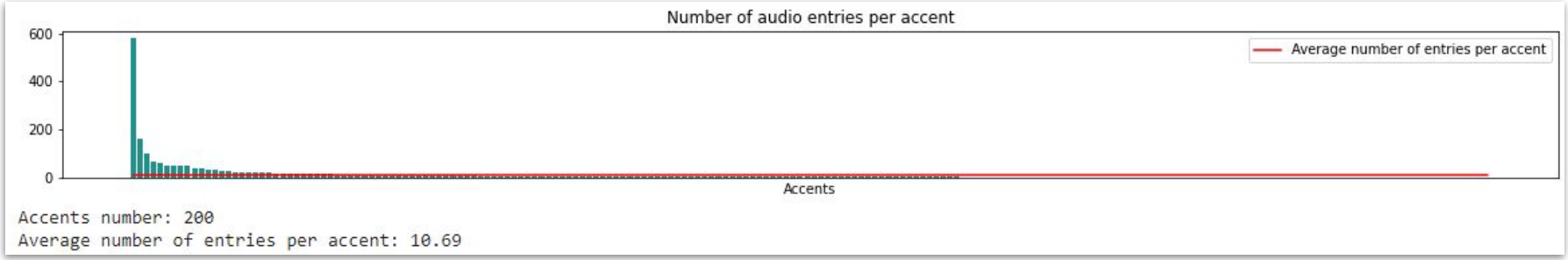


Data visualization

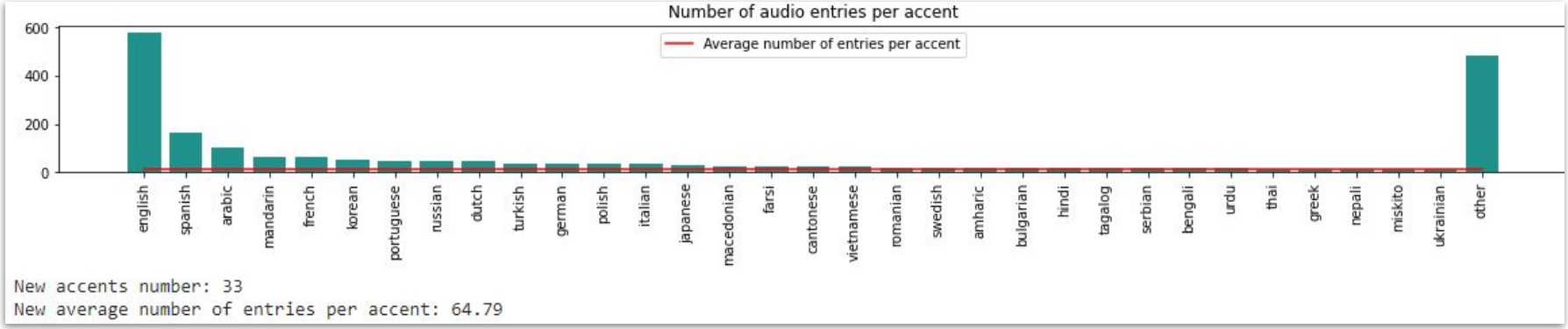
	age	age_onset	filename	native_language	sex	country
1154	23.0	4.0	hebrew9	hebrew	male	israel
1411	20.0	8.0	macedonian8	macedonian	male	macedonia
1240	29.0	13.0	japanese11	japanese	female	japan
523	27.0	0.0	english243	english	male	canada
1535	25.0	5.0	nepali4	nepali	male	nepal
93	30.0	12.0	arabic28	arabic	male	united arab emirates
1254	40.0	13.0	japanese24	japanese	female	japan
498	18.0	0.0	english220	english	female	usa
1226	21.0	8.0	italian29	italian	female	italy
651	32.0	0.0	english359	english	male	canada
2153	35.0	15.0	xasonga1	xasonga	female	senegal

Native language	Countries
afrikaans	['south africa']
agni	['ivory coast']
akan	['ghana']
albanian	['kosovo', 'albania']
amazigh	['morocco']
amharic	['ethiopia']
arabic	['saudi arabia', 'egypt
armenian	['armenia', 'iran', 're
ashanti	['ghana']
azerbaijani	['azerbaijan']
bafang	['cameroon']
baga	['guinea']
bai	['china']
bambara	['mali', 'senegal']
bamun	['cameroon']
bari	['sudan']
basque	['spain']
bavarian	['germany']
belarusan	['belarus']
bengali	['bangladesh', 'india',
bosnian	['bosnia and herzegovin
bulgarian	['bulgaria']
burmese	['myanmar']
cantonese	['china']
carolinian	['northern mariana isla
catalan	['spain', 'chile']

Data visualization




```
Accents number: 200
Average number of entries per accent: 10.69
```




```
New accents number: 33
New average number of entries per accent: 64.79
```

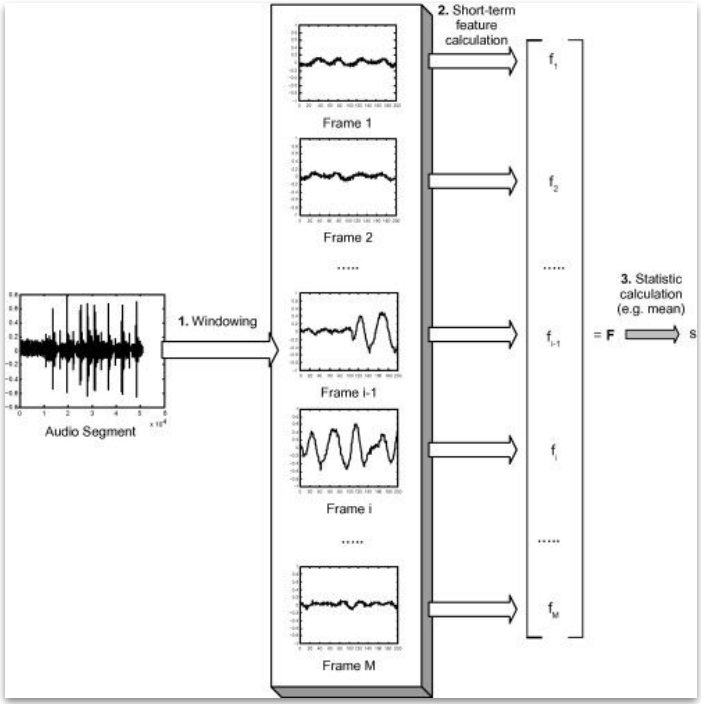
What data was chosen for machine learning

- For the case of most accents in the dataset, there are very few samples
  - this could be useful for people who want to learn the specific accent (for example)
  - for machine learning, a lot of classes with almost no samples complicated things
- For experimenting with machine learning on this dataset it is possible to:
  - downsampling classes that have more entries to balance the dataset
  - group similar accents to have fewer of them and to combine their entries
  - consider only some accents (discard the others)
    - this, along (with downsampling english category) will be the case for this study
    - the top three accents by sample number: english, spanish, arabic

Feature extraction

- The **`pyAudioAnalysis`** and **`librosa`** packages were used for audio manipulation and analysis
- The audio signal can be split by using a sliding window-like approach
  - window size and step size (the windows can also overlap if step is less than window size)
  - **Short Term Features** are defined by computing statistics on individual windows sector size and step size can be specified (similar to the smaller windows)
  - **Mid Term Features** are defined for sectors by aggregating on the short term features
  - A final aggregation can be applied on the mid term features (this helps with regard to the audio samples having different lengths, but this can also be solved with padding or interpolation and trimming)

Feature extraction

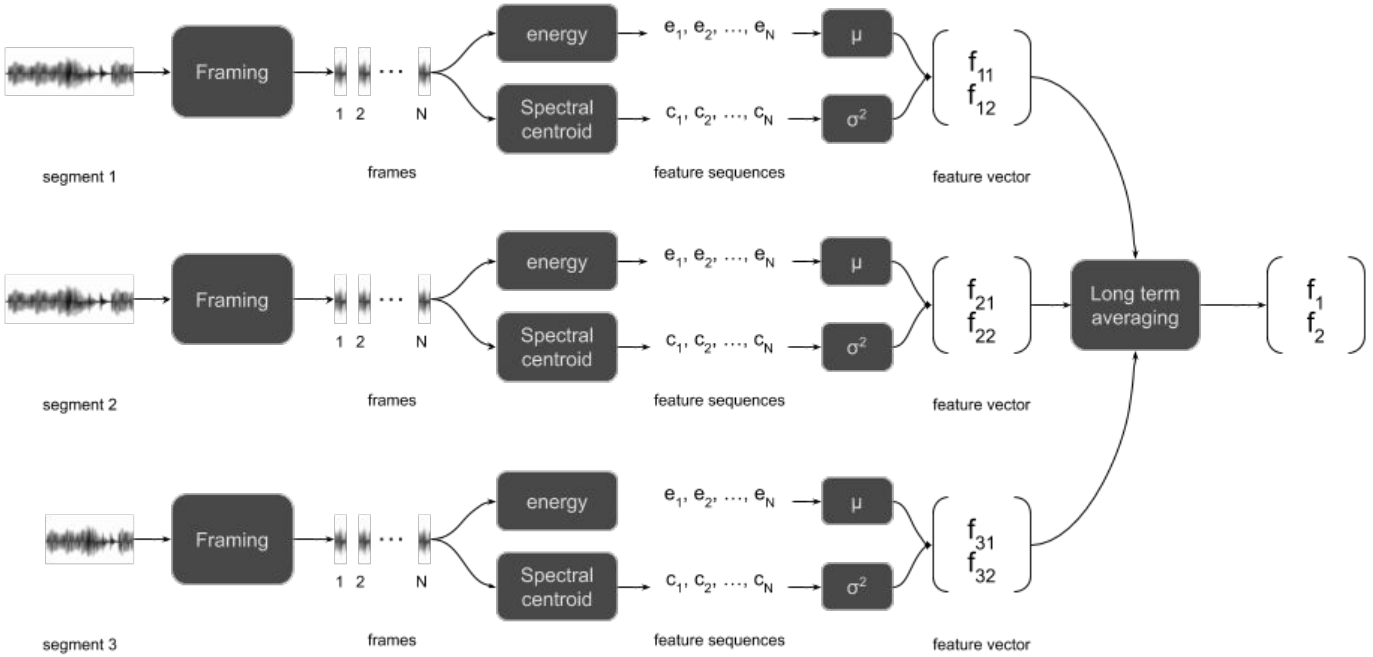


Reference:

Audio Features
Theodoros Giannakopoulos,
Aggelos Pikrakis,
Introduction to Audio Analysis, 2014

Intro to Audio Analysis: Recognizing Sounds Using Machine Learning
Theodoros Giannakopoulos

Feature extraction



Feature extraction

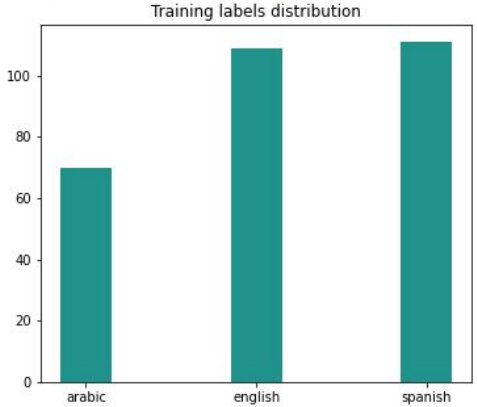
```
Duration: 21 seconds

539 frames, 68 short-term features

Feature names:
0       zcr
1       energy
2       energy_entropy
3       spectral_centroid
4       spectral_spread
5       spectral_entropy
6       spectral_flux
7       spectral_rolloff
8       mfcc_1
9       mfcc_2
```

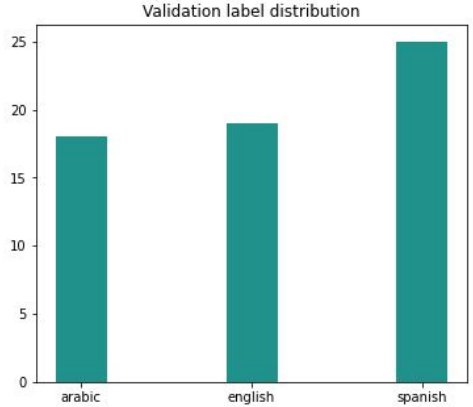
```
Duration: 24 seconds

488 x 68 short-term features
25  x 136 mid-term features

Feature names:
0       zcr_mean
1       energy_mean
2       energy_entropy_mean
3       spectral_centroid_mean
4       spectral_spread_mean
5       spectral_entropy_mean
6       spectral_flux_mean
7       spectral_rolloff_mean
8       mfcc_1_mean
```

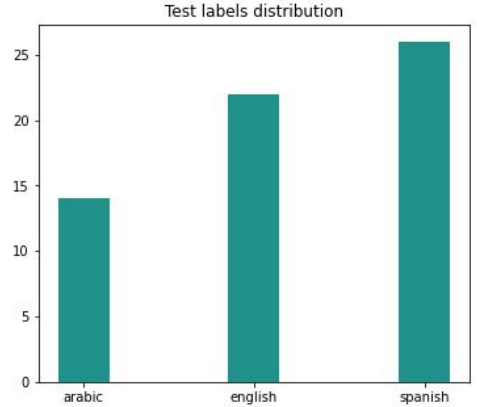
Machine Learning Approaches

Dataset splitting

- As stated before, only the top three accents by sample number were considered (english: 579, spanish: 162 and arabic: 102). English was downsampled to 150 to balance it with the other accents







- The dataset was split into training, validation and test sets (70-15-15 ratio)

Support Vector Classifier

- The data is standardized
- The sklearn package was used for the SVC model
- Parameters chosen according to grid search results

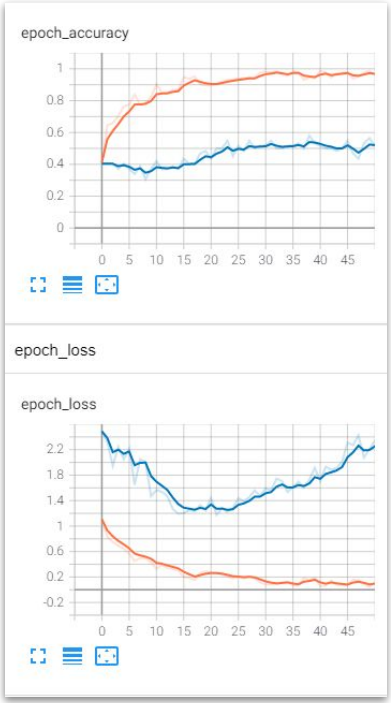
	C	kernel	Accuracy
0	0.1	linear	0.403448
1	0.1	rbf	0.396552
2	1.0	linear	0.331034
3	1.0	rbf	0.448276
4	5.0	linear	0.324138
5	5.0	rbf	0.382759
6	10.0	linear	0.324138
7	10.0	rbf	0.393103
8	15.0	linear	0.324138
9	15.0	rbf	0.393103

Accuracy: 0.5

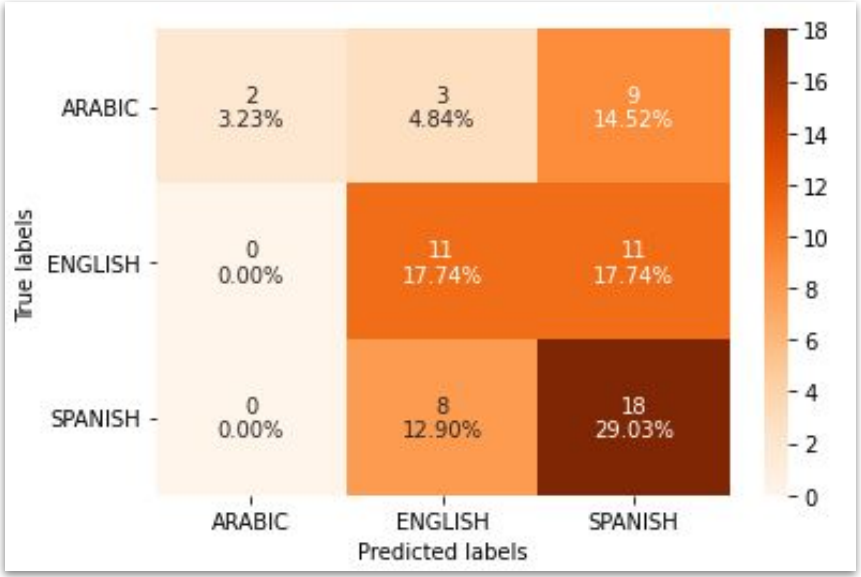
(using C=1 and rbf kernel)

Neural Network

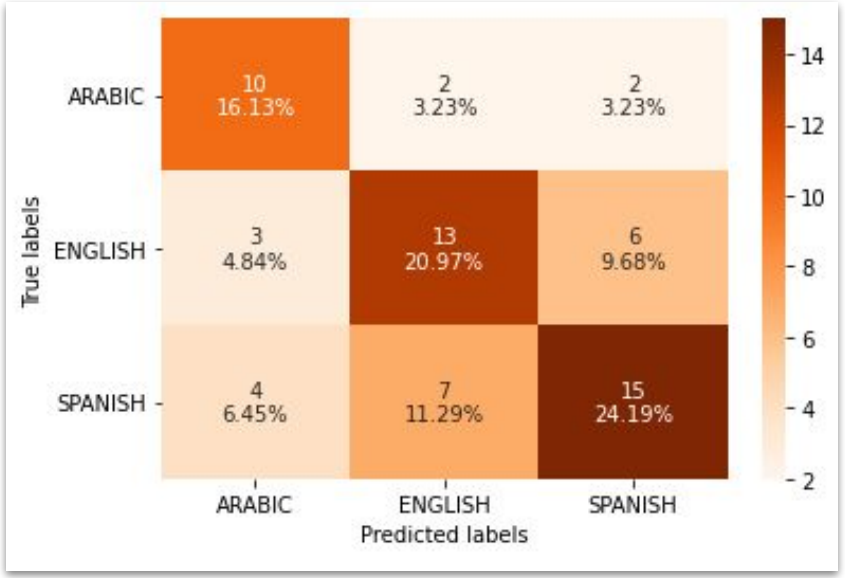
- `tensorflow.keras` was used for the model
- Batch normalization used as a first layer of the network to speed up convergence
- Multiple fully connected layers with ReLU activation
- Regularization: dropout was used to prevent overfitting
- A last fully connected layer with 3 neurons were used (softmax activation)
- Adam optimizer



Results - confusion matrix



Support Vector Classifier: **0.5** accuracy



Neural Network: **0.61** accuracy

Conclusion

- There is still a lot of room for improvement (Similar work - *Deep Learning Approach to Accent Classification,* Leon Mak An Sheng, Mok Wei Xiong Edmund obtained 88% using a convolutional neural network)

- Possible improvements:
  - more preprocessing on the audio samples (for instance, remove the silent portions from the speech)
  - improve the models to predict more classes (maybe choose a bigger, balanced dataset)
  - audio samples of female speakers can have very different statistics than the ones of male speakers (pitch and amplitude); additional gender detection based on audio samples can be done before applying a suitable model for the predicted gender
  - selecting only the most relevant audio features to apply machine learning on

Thank you!