Comparative study of estimations of situation types from prosodic features across multiple languages

Nath, Anindita

Thursday, June 29, 2017

Objective

Find commonalities/universal tendencies in prosodic features across languages.

Reminder: We are estimating situation (disaster) types in an audio segment, inferring this from the prosodic features of the training audio segments.

Specifically, we are trying to find out the language which would be the best to train on to predict an unknown low-resource language. We are interested in deducing whether similarities in languages which belong to the same language family have a significant effect on the performance of our algorithm.

Corpus

Number of Languages: 11

Abbreviations of languages:

AMH	-	Amharic
ARA	-	Arabic
FAS	-	Farsi
HAU	-	Hausa
HUN	-	Hungary
RUS	-	Russian
SOM	-	Somalia
SPA	-	Spanish
TUR	-	Turkish
VIE	-	Vietnamese
YOR	-	Yoruba

For training and testing on different language, Data Set: First 10 Audio Directories per language.

For training and testing on the same language,

Training Data Set: First 10 Audio Directories per language.

Test Data Set: Next 10 Audio Directories per language.

The different situation *types* represented as follows:

Type 1	Civil Unrest and Widespread Crime
Type 2	Election and Politics
Type 3	Evacuation
Type 4	Food Supply
Type 5	Infrastructure

Type 6	Medical Assistance
Type 7	Shelter
Type 8	Terrorism and Extreme violence
Type 9	Urgent Rescue
Type 10	Utilities, Sanitation and Energy
Type 11	Water Supply

Method

i. The Relevance graph (indicating the presence of a (any) situation type in the test audio) was generated using the evaluation script. The predicted type confidence scores of each audio of a test language were evaluated against its annotations giving the Precision-Recall curve.

ii. The results are presented below in Table 1: Performance Matrix.

iii. Also, the results are grouped under language families and presented below in Table2: Language Families.

iv. We, then, tested the following hypotheses:

- H1: Prediction in a language is better while trained on a language from the same language family (not the same language) than that belonging to a different one.
- H2: Prediction (AUC) values are generally better than baseline while a language is trained on one from a different language family.

Test Train	AMH		ARA		FAS		HAU		HUN		RUS		SOM		SPA		TUR		VIE		YOR		Utility for training** (%)
	B	M	B	M	B	M	B	M	B	M	B	M	B	M	B	M	B	M	B	M	B	M	
AMH	.74	**.791**	.69	**.695**	.54	**.594**	.76	**.896**	.72	.654	.9	**.945**	.42	**.569**	.06	**.199**	.74	**.862**	.2	**.252**	.5	.446	81.82
ARA	.74	.712	.69	.673	.54	**.598**	.76	.748	.72	.634	.9	**.95**	.42	**.449**	.06	**.247**	.74	.647	.2	**.299**	.5	.491	45.45
FAS	.74	**.875**	.69	.662	.54	**.695**	.76	**.815**	.72	.709	.9	**.955**	.42	**.432**	.06	**.219**	.74	**.781**	.2	**.381**	.5	.413	72.73
HAU	.74	**.787**	.69	.663	.54	**.667**	.76	.431	.72	.638	.9	**.958**	.42	.404	.06	**.197**	.74	.668	.2	**.469**	.5	.409	45.45
HUN	.74	**.833**	.69	**.704**	.54	**.571**	.76	**.862**	.72	.613	.9	**.967**	.42	**.479**	.06	**.204**	.74	**.791**	.2	**.244**	.5	.46	81.82
RUS	.74	**.902**	.69	.679	.54	**.581**	.76	**.872**	.72	.666	.9	**1.00**	.42	**.422**	.06	**.198**	.74	**.755**	.2	**.206**	.5	.43	72.73
SOM	.74	.702	.69	.654	.54	.464	.76	**.815**	.72	.589	.9	**.913**	.42	**.553**	.06	**.230**	.74	.664	.2	.199	.5	.455	36.36
SPA	.74	.708	.69	.670	.54	**.596**	.76	**.772**	.72	.716	.9	.797	.42	**.427**	.06	**.967**	.74	.714	.2	**.220**	.5	.480	45.45
TUR	.74	.661	.69	**.723**	.54	**.629**	.76	**.814**	.72	.658	.9	.893	.42	**.477**	.06	**.248**	.74	.691	.2	.192	.5	**.542**	54.55
VIE	.74	**.805**	.69	.650	.54	.527	.76	**.867**	.72	.683	.9	**.982**	.42	.374	.06	**.213**	.74	.666	.2	**.771**	.5	**.529**	54.55
YOR	.74	.696	.69	**.719**	.54	.527	.76	**.802**	.72	.719	.9	**.905**	.42	.345	.06	**.264**	.74	.712	.2	.189	.5	.423	36.36
Measure of Predictability (%) *	54.55		36.36		72.73		81.82		0		81.82		72.73		100		36.36		72.73		18.18		

Table-1: Performance Matrix[1]

[1] In Table-1:

- B stands for Baseline, M stands for Model
- Figures under column M, for each language, represent Area under Curve (AUC) values of the Precision-Recall Curve in the Relevance Graph.
- Figures in bold represent better performance than baseline.
- *The proportion of times the corresponding column language performed better than baseline.
- **The proportion of times the corresponding row language effectively trained the column language to help it perform above baseline.
- Figures highlighted in green correspond to the best training language while those in yellow correspond to the language easier to predict.

Language Family	Afro-Asiatic				Indo-European			Uralic	Turkic	Austro-asiatic	Niger-Congo
Test Train	AMH	ARA	HAU	SOM	FAS	RUS	SPA	HUN	TUR	VIE	YOR
AMH		0.005	0.136	0.149	0.054	0.045	0.139	-0.066	0.122	0.052	-0.054
ARA	-0.028		-0.012	0.029	0.058	0.05	0.187	-0.086	-0.093	0.099	-0.009
HAU	0.047	-0.027		-0.016	0.127	0.058	0.137	-0.082	-0.072	0.269	-0.091
SOM	-0.038	-0.036	0.055		-0.076	0.013	0.17	-0.131	-0.076	-0.001	-0.045
FAS	0.135	-0.028	0.055	0.012		0.055	0.159	-0.011	0.041	0.181	-0.087
RUS	0.162	-0.011	0.112	0.002	0.041		0.138	-0.054	0.015	0.006	-0.07
SPA	-0.032	-0.02	0.012	0.007	0.056	-0.103		-0.004	-0.026	0.02	-0.02
HUN	0.093	0.014	0.102	0.059	0.031	0.067	0.144		0.051	0.044	-0.04
TUR	-0.079	0.033	0.054	0.057	0.089	-0.007	0.188	-0.062		-0.008	0.042
VIE	0.065	-0.04	0.107	-0.046	-0.013	0.082	0.153	-0.037	-0.074		0.029
YOR	-0.044	0.029	0.042	-0.075	-0.013	0.005	0.204	-0.001	-0.028	-0.011	

Table-2: Language Families[2]

[2] In Table-2,

- Figures in each cell represent the difference of Area under Curve (AUC) values from the baseline corresponding to the train-test language pair.
- The regions highlighted in yellow contain results of training on the test language family.
- The prediction values of a language trained on itself have been omitted from the table.

Hypotheses Test

H1: Prediction in a language is better while trained on a language from the same language family (not the same language) than that belonging to a different one.

The two populations considered are as follows:-

P1: Trained and Tested on the same language family (but not the same language)

P2: Trained and Tested on different language family

The populations are estimated by the two samples, **S1** and **S2**, respectively, with values representing the difference of AUC from baseline. However, the values corresponding to the same train-test language pair have not been taken into consideration.

Sample

S1**:** size=18, mean=0.0339, std. dev=0.0742

S2: size=92, mean=0.0250, std. dev=0.0806

Test

Null Hypothesis, **H10**: (mean of S1) - (mean of S2) = 0

Alternative Hypothesis, **H11** :(mean of S1) - (mean of S2) >0

Statistical Test performed: **2-sample T-test** (both using separate variances and assuming equal variances)

Rejection Criteria of Null Hypothesis, H10:-

Since this is a right-tailed T-test, we reject Null Hypothesis if the *p*-value is either less than or equal to level of significance, *α*.

Parameters	**Un-pooled Variance T-test (samples have separate variances)**	**Pooled Variance T-test (assuming samples have equal Variances)**
Test-statistic, t*=	**0.46**	**0.433**
Degrees of freedom, df	**17**	**108**
p-value (at α=0.05)	**0.326**	**0.333**

As is evident from the table, *p*-value (=0.326 or 0.333) > α(=0.05) and so, at 5% level of significance, null hypothesis cannot be rejected. Therefore, we accept the null hypothesis H10, i.e. there is no significant difference between the mean of the two populations.

Hence, our original hypothesis H1, i.e. prediction in a language is better while trained on a language from the same language family (not the same language) than that belonging to a different one, cannot be claimed to be true.

H2: Prediction (AUC) values are generally better than baseline while a language is trained on one from a different language family.

Population: Trained and Tested on different language family.

This is estimated by a pair of sample values, one of them consist the AUC values of the population while the other represents the corresponding baseline.

Sample

S_d: size=92, mean=0.0250, std. dev=0.0806

Test

Null Hypothesis, **H20**: mean of S_d= 0

Alternative Hypothesis, **H21**: mean of S_d >0, where d is the difference between the sample pairs.

Statistical Test performed: **Matched pair T-test**

Rejection Criteria of Null Hypothesis, H20:-

Since this is a right-tailed T-test, we reject Null Hypothesis if the *p*-value is either less than or equal to level of significance, *α*.

Parameters	**Matched-pair T-test**
Test-statistic, t*=	**2.98**
Degrees of freedom, df	**91**
p-value (at α=0.05)	**0.002**

As is evident from the table, *p*-value (=0.002) < α(=0.05) and so, at 5% level of significance, we reject the null hypothesis. Therefore,we accept the alternate hypothesis, H21, i.e. the mean of the difference values between the sample pairs is positive.

Hence, our original hypothesis H2, i.e., prediction (AUC) values are generally better than baseline while a language is trained on one from a different language family, can be claimed to be true.

H3: Prediction (AUC) values are generally better than baseline while a language is trained on one from the same language family.

Population: Trained and Tested on same language family (but not the same language)

This is estimated by a pair of sample values, one of them consist the AUC values of the population while the other represents the corresponding baseline.

Sample

S_d: size=18, mean=0.0339, std. dev=0.0742

Test

Null Hypothesis, **H30**: mean of S_d= 0

Alternative Hypothesis, **H31**: mean of S_d >0, where d is the difference between the sample pairs.

Statistical Test performed: **Matched pair T-test**

Rejection Criteria of Null Hypothesis, H30:-

Since this is a right-tailed T-test, we reject Null Hypothesis if the *p*-value is either less than or equal to level of significance, *α*.

Parameters	**Matched-pair T-test**
Test-statistic, t*=	**1.94**
Degrees of freedom, df	**17**
p-value (at α=0.05)	**0.035**

As is evident from the table, *p*-value (=0.035) < α(=0.05), we can reject the null hypothesis at 5% level of significance. Therefore, we accept the alternate hypothesis, H31, i.e. the mean of the difference values between the sample pairs is positive.

Hence, our original hypothesis H3, i.e., prediction (AUC) values are generally better than baseline while a language is trained on one from the same language family can be claimed to be true.

H4: Prediction (AUC) values are, in general, better than baseline, irrespective of the language family

Population: Trained on a language not the same as test. (i.e. includes all values other than those belonging to the same language train-test pair).

This is estimated by a pair of sample values, one of them consists the AUC values of the population while the other represents the corresponding baseline.

Sample

S_d: size=110, mean=0.02646, std. dev=0.07933

Test

Null Hypothesis, **H40**: mean of S_d= 0

Alternative Hypothesis, **H41**: mean of S_d >0, where d is the difference between the sample pairs.

Statistical Test performed: **Matched pair T-test**

Rejection Criteria of Null Hypothesis, H40:-

Since this is a right-tailed T-test, we reject Null Hypothesis if the *p*-value is either less than or equal to level of significance, *α*.

Parameters	**Matched-pair T-test**
Test-statistic, t*=	**3.49**
Degrees of freedom, df	**109**
p-value(at α=0.05)	**.00034**

As is evident from the table, *p-value* (=0.00034) < α(=0.05), we reject null hypothesis at 5% level of significance. Therefore, we accept the alternate hypothesis, H41, i.e. the mean of the difference values between the sample pairs is positive.

Hence, our original hypothesis H4, i.e., prediction (AUC) values are, in general, better than baseline, irrespective of the language family can be claimed to be true.

Conclusions

i. Similarities in the prosodic features of the same language family could not be established.

ii. Performance of our algorithm is generally better than baseline while predicting types in language whether trained on a language from the same or different family.

iii. It follows from pt. (ii) above that our algorithm performs better than baseline, in general.

iv. The search for the best training language to predict an unknown low-resource language still continues.