First Workshop on Speech Technologies for Code-switching in Multilingual Communities: Shared Task Description

We are organizing a shared task for the workshop which will consist of two sub-tasks for Spoken language Identification (LID) of code-switched audio. The two sub-tasks will consist of:

- Part A: Utterance-level LID
- Part B: Frame-level LID

The shared task will consist of two phases: a training phase and a testing phase. During the training phase, participants will register on the workshop website to receive a link to download training and development data. This document contains descriptions of how to build baseline systems using the training and development data. During the testing phase, participants will receive test data for which they will have to submit results in the specified format. In addition, participants will also have to submit their systems for verification. We will maintain a leaderboard of best performing systems in the testing phase. All participants will be invited to submit a paper describing their system and results, which if accepted can be presented during the Workshop.

Please see the Workshop website for up-to-date information about the shared task timeline.

1 Data Description

You will receive both a training data set ("participant Training Data") and a Dev data set ("participant Dev Data") during the training phase, and a Test data set ("participant Test Data") during the testing phrase. Participants are expected to train their models on the training data set and use Dev data set for validation. The Dev data can be used for the final models if desired.

The train and dev sets consist of audio with corresponding language labels. For the test set, only audios shall be provided without corresponding language labels.

The data set consists of set of three code-switched language pairs - Gujarati-English, Tamil-English and Telugu-English. In addition, we will also be using some data from monolingual datasets released by us in Gujarati, Tamil and Telugu which can be found here. You may use the training and dev data provided below for your systems. Please do not use the test data for training or fine-tuning your systems.

`https://msropendata.com/datasets/7230b4b1-912d-400e-be58-f84e0512985e`

2 Evaluation

For task A, participants need to submit an utterance level label (Monolingual or code-switched) for a particular audio file. For task B, participants need to submit a frame-level (200ms) label for each frame in the audio. We will use accuracy and Equal Error Rate as evaluation metrics.

2.1 Metrics

- Accuracy

$$Accuracy = \frac{N}{T}$$

Where,
N is the total no. of correctly predicted data samples
T is the total no. of data points

- Equal Error Rate (EER)

$$EER = \frac{FRR + FAR}{2}$$

$$FRR = \frac{TFR}{T}$$

$$FAR = \frac{TFA}{T}$$

where,
EER Equal Error Rate
FRR False Rejection Rate
FAR False Acceptance Rate
TFR Total No. of False Rejects
TFA Total No. of False Accepts
T Total No. of Datapoints

3 Baseline Systems

3.1 Sub-task A: Classify audio into Code-switched or Monolingual at utterance-level

Feature extraction: We use python's librosa library for feature extraction. Following are the steps followed for feature extraction.

- Take Short Term Fourier Transform of the audio with *windosize*=0.02, *windostride*=0.01. (https://librosa.github.io/librosa/generated/librosa.core.stft.html)
- We convert the output matrix into magnitude and phase using magphase function of librosa library. (librosa.magphase)
- We take log of 1 + the output matrix to avoid errors popping in for features values which are almost equal to 1 using numpy library of python.

Model: Our baseline system is made up of an end-to-end multi-layer model consisting of 5 layers of LSTM, each consisting of 1024 neurons. The model is based on deepspeech 2 (https://arxiv.org/pdf/1512.02595v1.pdf). The model is trained using the CTC loss function. We have trained our model for 40 epochs and the reported Accuracies and EERs are on the Dev set.

Class 1	Class 2	Accuracy (%)	EER
Gu-En	Gu-Mono	76.8%	11.6%
Te-En	Te-Mono	71.2%	14.4%
Ta-En	Ta-Mono	74.0%	13.0%

Tabelle 1: Baseline Accuracy and Equal Error Rate (EER) for Part A. The Numbers are being reported on the Dev Set.

3.2 Sub-task B: Classify audio into English or other language (Gujarati/Telugu/Tamil) at frame-level

Feature extraction: We use python's librosa library for feature extraction. Following are the steps followed for feature extraction.

- Take Short Term Fourier Transform of the audio with *windo-size*=0.02, *window-stride*=0.01. (https://librosa.github.io/librosa/generated/librosa.core.stft.html)
- We convert the output matrix into magnitude and phase using magphase function of librosa library. (librosa.magphase)
- We take log of 1 + the output matrix to avoid errors popping in for features values which are almost equal to 1 using numpy library of python.

Model: Our baseline system for Part B is similar to Part A. The only difference is that instead of having a softmax classifier we output the label having the highest probability for each time slice of "200ms"which is our *window-size*.

Language Pair	Accuracy (%)	EER
Gu-En	76.7%	6.7%
Te-En	76.5%	6.7%
Ta-En	77.6%	6.5%

Tabelle 2: Baseline Accuracy and Equal Error Rate (EER) for Part B. The Numbers are being reported on the Dev Set.

3.3 Reproducing the baselines

- Install Deepspeech 2 (Pytorch) using the following link: https://github.com/SeanNaren/deepspeech.pytorch
- Use the following command line parameters to train the model
  - –rnn-type lstm
  - –hidden-size 1024
  - –hidden-layers 5
  - –epochs 40 (Sub-Task A) 60 (Sub-task B)
  - –learning-anneal 1.01
  - –batch-size 32
  - –window-size 0.02 (for Sub-Task A) 0.2 (for Sub-Task B)
  - –window-stride 0.01 (for Sub-Task A) 0.1 (for Sub-Task B)
- Optimize your model to maximize Accuracy instead of WER / CER which is usually done for Automatic Speech Recognition.

c) Evaluation

- Part A: For the both the sub-parts, the participants can report their scores in Accuracy and EER as shown in Table 2
  Example of Evaluating tasks in Part A

Example for Sub-Task A:
Ground Truth (submitted file)
fname1,0,0
fname2,0,1
fname3,0,0
fname4,1,0
fname5,1,1
fname6,1,1
fname7,1,0
fname8,0,1
N (Total No. of correct predictions): 4
T (Total no. of data points): 8

$$Accuracy = \frac{N}{T} = \frac{4}{8} = 50\%$$

$$EER(P=0) = \frac{FRR + FAR}{2} = \frac{\frac{TFR}{T} + \frac{TFA}{T}}{2}$$

$$EER(P=0) = \frac{\frac{2}{8} + \frac{2}{8}}{2} = \frac{0.5}{2} = 0.25$$

Similarly,

$$EER(P=1) = \frac{\frac{2}{8} + \frac{2}{8}}{2} = \frac{0.5}{2} = 0.25$$

Average EER = 0.25

Example for Sub-Task B:
Ground Truth (submitted file)
fname1,0,0
fname2,0,0
fname3,1,2
fname4,4,0
fname5,3,3
fname6,2,2
fname7,4,1
fname8,2,1
N (Total No. of correct predictions): 3
T (Total no. of data points): 8

$$Accuracy = \frac{N}{T} = \frac{4}{8} = 37.5\%$$

$$EER(P=0) = \frac{FRR + FAR}{2} = \frac{\frac{TFR}{T} + \frac{TFA}{T}}{2}$$

$$EER(P=0) = \frac{\frac{0}{8} + \frac{1}{8}}{2} = \frac{0.125}{2} = 0.0625$$

Cal EER for all labels and take an average.

- Part B: For Part B also the participants can report their scores in Accuracy and EER.

Example for Part B Task
Ground Truth Language Tag Sequence: SSTTTTTTTTTSSSSSEETTSSTTTETTTTSTTTTS
Predicted Language Tag Sequence: SSSTTTSTSSSSSSSEESSSSSSSSTSTSTSSSTS
The above two sequences are of Language Tags for every 200ms of audio
N (Total No. of correct Language Tags): 16
T (Total no. of data points): 36

$$Accuracy = \frac{N}{T} = \frac{16}{36} = 44.4\%\%$$

$$EER(P=S) = \frac{FRR + FAR}{2} = \frac{\frac{TFR}{T} + \frac{TFA}{T}}{2}$$

$$EER(P=S) = \frac{\frac{0}{36} + \frac{15}{36}}{2} = \frac{0.416}{2} = 0.208$$

Cal EER for 'T' and 'E' and take an average.