Exercise 4 Output

Output of the work from DeepSpeech

Language	File	WER
English	checkin.wav	20%
English	checkin_child.wav	40%
English	parents.wav	20%
English	parents_child.wav	0%
English	suitcase.wav	0%
English	suitcase_child.wav	50%
English	what_time.wav	20%
English	what_time_child.wav	20%
English	where.wav	0%
English	where_child.wav	0%
English	my_sentence1.wav	14%
English	my_sentence2.wav	0%
Italian	checkin_it.wav	25%
Italian	parents_it.wav	20%
Italian	suitcase_it.wav	14%
Italian	what_time_it.wav	100%
Italian	where_it.wav	42%
Spanish	checkin_es.wav	25%
Spanish	parents_es.wav	0%
Spanish	suitcase_es.wav	16%
Spanish	what_time_es.wav	83%
Spanish	where_es.wav	57%

Further development

Output of the work from Pocketsphinx

Language	File	WER
English	checkin.wav	100%
English	checkin_child.wav	100%
English	parents.wav	100%
English	parents_child.wav	100%
English	suitcase.wav	100%
English	suitcase_child.wav	100%
English	what_time.wav	100%
English	what_time_child.wav	100%
English	where.wav	100%
English	where_child.wav	100%

Command line output of Pocketsphinx: Ground truth and best 4 hypothesses of each audio

```
where is the checkin desk [(None, -22642), ('huh', -22777), ('ah', -22930), ('a', -23047)]
where is the checkin desk []
i have lost my parents [(None, -22030), ('ah', -22404), ('a', -22482), ('the', -22630)]
i have lost my parents []
please i have lost my suitcase []
please i have lost my suitcase []
what time is my plane []
what time is my plane [(None, -22451), ('ah', -22803), (None, -788597), (None, -788777)]
where are the restaurants and shops []
where are the restaurants and shops []
where can i get on the taxi []
how do i get to london from here []
```

Output of the work from Vosk

Language	File	WER
English	checkin.wav	0%
English	checkin_child.wav	100%
English	parents.wav	20%
English	parents_child.wav	100%
English	suitcase.wav	16%
English	suitcase_child.wav	100%
English	what_time.wav	0%
English	what_time_child.wav	100%
English	where.wav	0%
English	where_child.wav	0%

Command line output of Vosk: a hypothesis of each audio

```
{
  "text" : "where is the check in desk"
}
{"text": ""}
{
  "text" : "i lost my parents"
}
{"text": ""}
{
  "text" : "please have lost my suitcase"
}
{"text": ""}
{
  "text" : "what time is my plane"
}
{"text": ""}
{
  "text" : "where are the restaurants and shops"
}
{
  "text" : "where are the restaurants and shops"
}
```

Exercise 4 Report

Description of DeepSpeech and its configuration

Mozilla DeepSpeech I used for my speech recognition application is the offline ASR library. It provides the speech-to-text engine which can run in real time on devices ranging from a Raspberry Pi4 to high power GPU servers. DeepSpeech is the deep learning model using Recurrent Neural Network. It has been configured by setting different model and scorer for different languages. In this exercise, I configured for English, Italian, and Spanish using the following models and scorers:

English
scorer = "Models/deepspeech-0.9.3-models.scorer"
model = "Models/deepspeech-0.9.3-models.pbmm"

Italian
scorer = "Models/kenlm_it.scorer"
model = "Models/output_graph_it.pbmm"

Spanish
scorer = "Models/kenlm_es.scorer"
model = "Models/output_graph_es.pbmm"

Analysis of the solution for the noisy environment

To solve the issue of the noisy environment, I applied the algorithm to find the best process method. In this algorithm, it tries 3 options: 1. process the original audio as it is, 2. process the low-pass filtered audio, and 3. process the noise reduced audio. After tried 3 options, choose the result with the best WER.

The following screenshot shows the best WER, the best output, and the used option, as the result in each language. For all languages, with my solution, DeepSpeech performed best on an original audio file in most cases, and low-pass filter or noise reduction worked better for some sentences. For particular sentences such as "What time is my plane?" and "Where are the restaurants and shops?", DeepSpeech had difficulty in recognition if they are spoken in Italian or Spanish, but had performed well on English.

In summary, I can analyze that DeepSpeech can handle noise in many cases even though in the pretty loud environment. However, some particular sentences where noise reduction or filter does not work need further research and experiment in order to be recognized well. If integrate the predict() function, the performance can be improved. For example, based on the output hypothesis from DeepSpeech, the function predicts a phrase from a given list of phrases that are commonly used in airport (I did not implement this as the exercise required to test how well it recognizes phrases).

Analysis of the test result: Provided audios vs. Own recordings

The test result shows that DeepSpeech performs well on my own recordings even though spoken by me as a non-native speaker. This might be because the audio files do not contain any loud noise. Also, if spoken by a native speaker, WER could be zero. Listening that the English audio sounds seem to be spoken by native English speakers, I can say that noise is affecting the performance and decreasing WER.

However, DeepSpeech performed better on original audio rather than on filtered or noise reduced one. It might be insufficient for the airport environment just to apply a low pass filter or noise reduction as my solution. This suggests I will need further analysis on the type of noise in the airport and improve the protocol. For example, I may want to try to remove the airport noise using Audacity, record the process, and test the edited audio with DeepSpeech. Once I found the acceptable WER, I can implement the same process on my application.

Analysis of the evaluation of other ASR libraries

I evaluated the results from 2 other ASR systems, Pocketsphinx and Vosk. Since DeepSpeech performed well on oroginal audio files in most cases, I applied both systems to the original ones only without using any filter so I can compare the performance on the same measure.

1. Pocketsphinx

Observe the result table shown below, Pocketsphinx does not recognize any of the English sentences including my sentences without noise for AudioFile recognition. However, in the class, Matthew was using LiveSpeech with mic input and it recognized pretty well. For this reason, I would not recommend the company to use Pocketsphinx, but would suggest to try the mic input version with LiveSpeech.

2. Vosk

I applied Vosk as same as I did for Pocketsphinx. Observe the result table shown below, it recognizes pretty well on adult voice. However, when it comes to child voice, it does not recognize any word except the last sentence. Since there are so many children are in airport, Vosk would probably not strong enough for the company to use for an airport assistance.

In conclusion, I would recommend that the company should use DeepSpeech for their airport assistance system because it has more potential to improve the performance in the airport environment as it still produces sentences as the result, not an empty string or a few words. As long as it produces natural sentences, we can improve the hypothesis by predicting based on the result.