KAISD (Korean language Alcohol Intoxicated Speech Detector)

Project Proposal

SuJin Han 20170841, JeongEon Park 20160811, SeokJun Kim 20160793, HyunChang Oh 20170410

1. Goal

Drunk driving is a serious problem in Korea. In 2019, 15,708 traffic accidents were caused by drunk driving [1], and in 2018, police reported that the number of drunk driving cases averaged 334.2 per day [2]. The problem also prevails in KAIST, where a drunk driving accident was reported in 2016 [3].

While drunk driving is an **important** issue in Korea, methods for drunkenness detection is limited. For example, the frequently used Breathalyzer and blood testing can only work in short distances. We believe that providing a more accessible way of alcohol detection will reduce drunk driving accidents. Thus, we propose an audio detector that can classify whether a person is drunk from speeches.

An **interesting** aspect of our project is that we will be working with Korean speech data. Most of the previous work conducted in the field of intoxicated speech detection have been carried out with the Alcohol Language Corpus(ALC), which is in German [4][5][6][7]. Since there exist language specific differences in emotion recognition [8], we believe similar differences exist in intoxicated speech detection. Hence, we plan to focus on detecting whether a person is drunk solely with Korean speech samples.

We hope to solve this problem by designing a neural network that can determine whether the speaker is under the influence of alcohol given the speaker's speech data. The **final outcome** of this project will be a binary classifier that determines whether the speaker is intoxicated or not.

2. Approach/Baselines

We plan to use the Convolutional Neural Network (CNN) to tackle this problem. Since we take audio data as input, the input matrix is likely to be very large. Also, our input matrix will be variable in size depending on the length of the speech. CNN helps to overcome these challenges because it significantly reduces the number of parameters and can take variable-size input. Previous works on neural network based intoxicated speech detection used CNN as well [9][10].

We selected the AlcoAudio model, which is a CNN-based intoxicated speech detection model trained with the ALC, as our baseline model [10]. This model has an unweighted accuracy of 66.28%. We chose this model because this model tackles the same task.

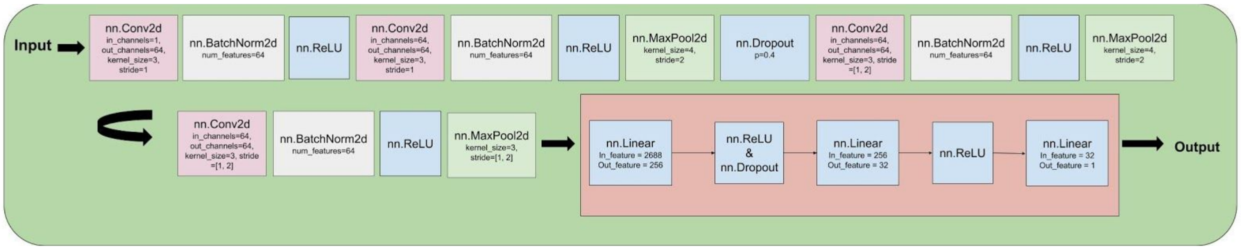
Input
nn.Conv2d in_channels=1, out_channels=64, kernel_size=3, stride=1
nn.BatchNorm2d num_features=64
nn.ReLU
nn.Conv2d in_channels=64, out_channels=64, kernel_size=3, stride=1
nn.BatchNorm2d num_features=64
nn.ReLU
nn.MaxPool2d kernel_size=4, stride=2
nn.Dropout p=0.4
nn.Conv2d in_channels=64, out_channels=64, kernel_size=3, stride =[1, 2]
nn.BatchNorm2d num_features=64
nn.ReLU
nn.MaxPool2d kernel_size=4, stride=2
nn.Conv2d in_channels=64, out_channels=64, kernel_size=3, stride =[1, 2]
nn.BatchNorm2d num_features=64
nn.ReLU
nn.MaxPool2d kernel_size=3, stride=[1, 2]
nn.Linear In_feature = 2688 Out_feature = 256
nn.ReLU & nn.Dropout
nn.Linear In_feature = 256 Out_feature = 32
nn.ReLU
nn.Linear In_feature = 32 Out_feature = 1
Output

Figure 1. AlcoAudio CNN Model for Intoxicated Speech Detection

We plan on making the following modifications. The baseline model uses convolution layers with fully connected(FC) layers. We will make the model fully convolutional so that the model can take variable length speech data as the input. We also hope to improve the model's performance by stacking more layers and adding residual connections and/or inception modules. These modifications will help mitigate the challenges coming from deep networks.

Furthermore, we also considered Bone et al.'s model, which was the winner of the 2011 INTERSPEECH Intoxication Sub-Challenge, as the baseline model. This model was also developed with the ALC and had unweighted accuracy of 70.5% [5]. However, we did not choose this model as our baseline because the model was developed 9 years ago and thus did not utilize neural networks.

3. Plan for Training

Since the only accessible alcohol voice dataset is the ALC and its price is €1,020.00, we decided to collect our own data. Our data collecting procedure will be as follows:

1. **Determine the Korean sentences to collect**
   We plan to come up with 10 sentences in Korean language to use as the template for each recording sample. It has been proven that intoxicated people struggle with pronouncing sentences that contain a sequence of consonants without a vowel in between [11]. An example in Korean could be a tongue twister like “간장공장공장장” (Soy sauce factory). In addition, there are certain phonetics in a sentence that are emphasized or altered when speaking in intoxicated condition [12]. An example is pronouncing a "r" sound as "l". Thus, considering these ideas, the selected sentences will be noticeably different when spoken in sober and intoxicated situations.

2. **Obtain recordings of people in both sober and intoxicated conditions**
   We plan to obtain audio recordings of people speaking the selected Korean sentences in both sober and intoxicated conditions. The most ideal method is to obtain recordings of a single person in both sober and intoxicated conditions, but this could be very impractical. Hence, we first plan to get sober recordings from ourselves and our close neighbors. For intoxicated recordings, we plan to visit the pub and ask people for the recording. There will be background noises, but we believe that they will be beneficial as training with them can contribute to adversarial defense.

3. **Standard for intoxicated or sober in each recording**
   In order to classify whether the speaker is either intoxicated or sober, we need a concrete classification boundary. From June 25th, the standard blood alcohol concentration for an intoxicated person in Korea reduced from 0.05% to 0.03% [14]. Thus, before recording the sentences, we plan to ask the speaker to breathe into a breathalyzer so that we can successfully label the sentences as either "intoxicated" or "sober."

4. **Data augmentation**
   If we record 10 Korean sentences from 100 people, that will be a total of 1000 recordings. This might not be enough for training our model. Thus, we plan to apply the following data augmentation techniques to increase our data size.

- Add mute / pauses in sober recordings: This will not only increase the input data size but also prevent the model from learning that a mute/pause is an aspect of intoxicated recording.
- Adjust the speed of a section in sober recordings: This is also to prevent the model from overfitting to the pace of speech

4. Plan for Evaluation

1. **Evaluation Dataset**

We plan on dividing our data to 10 sections and create 10 different models. For each model, we will use 90% of data to train the model and use the remaining 10% for validation.

2. **Evaluation Metric**

We will use the highest unweighted accuracy (i.e. unweighted average recall) from the aforementioned 10 models as our evaluation metric. The same metric was used for the 2011 INTERSPEECH Intoxication Sub-Challenge.

3. **Additional Evaluation Plan**

The model may take free speech or pronounciation of pre-selected sentences as input. We hope to create a model that will be able to predict drunkenness given free speech. Thus, we will conduct an additional user-study for evaluation.

5. Risk Management

Since we are manually collecting data, we may not be able to collect enough data for training. In this section, we propose augmentary methods to fully utilize the raw training data, so that the system may show satisfactory performance even with limited training data set.

A set of phonemes—the smallest atomic unit of pronunciation — that are especially error-prone has been identified in intoxicated speeches [13]. However, errors occur not only in the phonemes themselves, but rather in the course of devolution to the next proximate successor. Therefore, the search space should involve the possible combinatorial interactions between phonemes. Hence, we should include a variety of combinations in the sentences to be enunciated during data collection. KoNLPy [15], a public NLP package for the Korean language, will be used to retrieve the phonemes from independent words, which will be combined with stochastic algorithms to generate the final sentences.

Successful training, and consequently successful evaluation ought to cover a large portion of the network. Perturbation can be randomly introduced to the training data to not only increase the size of the training set, but also to build a model resistant to adversarial attacks. The system is expected to be used in a noisy environment, so having a good adversarial defense is necessary for usability. Addition of white noise or a different speech at a low magnitude to the target audio can be an example of such perturbation, and the waveform itself can be manipulated in wav format as well.

The random generation of perturbation can be guided to cover large parts of the network with genetic algorithms. The selection pressure will be the coverage of the training data of each perturbation, and whether the perturbation has actually caused the model to yield a different outcome. Such a pressure setting, however, is assured to make naively large changes as there is no competing pressure. A heuristic must be developed to limit the magnitude of the perturbation and preserve the integrity of the phonetic features. Human-in-the-loop approach will be employed to directly exert such a heuristic.

When in the course of preparing for a failure, we believe that the plans will be useless, but planning is indispensable. If the collected training data turns out to be deficient even with all the augmentary measures, the only viable option is to rely on external data. There exists a model pre trained with German speeches [AlcoAudio]. The auditory aspect of a language is merely combinations of phonemes, and as our work focuses on the phonetic characters rather than the lexical characters, we believe the model can still perform to decent satisfaction even if trained with a different language.

6. References

[1] KOSIS. 음주운전교통사고비율. http://kosis.kr/statHtml/statHtml.do?orgId=101&tblId=DT_1YL14001

[2] Kim, Arin. 153 Caught Drunk Driving Nationwide on First Day of Toughened DUI Law. http://www.koreaherald.com/view.php?ud=20190625000693

[3] 임성민. 서측 쪽문에서 교통사고 발생.http://times.kaist.ac.kr/news/articleView.html?idxno=3452

[4] Bavarian Archive for Speech Signals. Alcohol Language Corpus - ALC. https://www.phonetik.uni-muenchen.de/Bas/BasALCeng.html

[5] Bone, et al. Intoxicated Speech Detection by Fusion of Speaker Normalized Hierarchical Features and GMM Supervectors. INTERSPEECH 2011.

[6] Biopsy, et al. Intoxication Detection using Phonetic Phonotactic and Prosodic Cues. INTERSPEECH 2011.

[7] Bocklet, Tobias, Korbinian Riedhammer, and Elmar Noth. Drink and Speak: On the Automatic Classification of Alcohol Intoxication by Acoustic, Prosodic and Text-Based Features. INTERSPEECH 2011.

[8] Rajoo, Rajesvary, and Ching Chee Aun. Influences of Languages in Speech Emotion Recognition: A Comparative Study Using Malay, English and Mandarin Languages. ISCAIE 2016.

[9] Miller, Joshua, Jillian Donahue and Benjamin Schmitz. Speech Emotion and Drunkenness Detection Using a Convolutional Neural Network. http://www2.ece.rochester.edu/~zduan/teaching/ece477/projects/2018/JoshuaMiller_JillianDonahue_BenjaminSchmitz_ReportFinal.pdf

[10] ShreeshaN. AlcoAudio. https://github.com/ShreeshaN/AlcoAudio

[11] 이원희 *et al.* 발성문장 종류에 따른 음주여부 판단에 관한 연구. http://journal.hsst.or.kr/DATA/pdf/v8_7_77.pdf

[12] Keith Johnson *et al.* Do Voice Recordings Reveal whether a Person Is Intoxicated? A Case Study https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3524529/

[13] Florian Schiel *et al.* Alcohol language corpus: the first public corpus of alcoholized German speech. https://link.springer.com/article/10.1007/s10579-011-9139-y

[14] 정환봉. '딱 한 잔만 마셔도 잡힌다' 25일부터 음주운전 기준 0.03%로 강화. http://www.hani.co.kr/arti/society/society_general/898919.html

[15] Eunjeong L. Park, Sungzoon Cho. "KoNLPy: Korean natural language processing in Python" https://konlpy.org/en/latest/