Deep Neural Networks with Batch Speaker Normalization for Intoxicated Speech Detection

Weiqing Wang*, Haiwei Wu*†, Ming Li*
* Data Science Research Center, Duke Kunshan University, Kunshan, China
E-mail: ming.li369@duke.edu
† School of Electronics and Information Technology, Sun Yat-sen University, Guangzhou, China

***Abstract*—Alcohol intoxication can affect people both physically and psychologically, and one's speech will also become different. However, detecting the intoxicated state from the speech is a challenging task. In this paper, we first implement the baseline model with ComParE feature and then explore the influence of the speaker information on the intoxication detection task. Besides, we apply a ResNet18 based model to this task. The model contains three parts: a representation learning subnetwork with Deep Residual Neural Network(ResNet) of 18-layer, a global average pooling(GAP) layer and a classifier of 2 fully connected layers. Since we cannot perform speaker z-normalization on the variant-length feature input, we employ the batch z-normalization to train the proposed model. It also achieves similar improvement like applying the speaker normalization to the baseline method. Experimental results show that speaker normalization on baseline model and batch z-normalization on ResNet18 based model provides 4.9% and 3.8% improvement respectively. The results show that speaker normalization can improve the performance of both the baseline model and the proposed model.**

Index Terms: intoxicated speech detection, Convolutional Neural Network, computational paralinguistics

I. Introduction

One of the major causes of traffic accidents is alcoholic intoxication. Currently, the known methods to detect the alcoholic state are measuring the breath alcohol concentration(BrAC) and the blood alcohol concentration(BAC). In fact, alcohol can affect one's cognitive and motor function in various ways, leading to obvious changes in behavior. Alcohol has long-term cognitive effects and impairs one's information processing even during decreasing blood-alcohol concentration[1]. These impairment can be reflected in several aspects: vision[2], hand-writing[3, 4] and speech[5].

There are several investigations on how alcohol can affect one's speech. When under the influence of alcohol, people's speech have more filled pauses compared to speech in a sober condition, which means that disfluency events occur in the intoxicated speech more frequently[6]. The accuracy of GMM-UBM speaker recognition(SR) system also degrades when the database contains the intoxicated data[7]. Compared to the baseline SR system without alcohol intoxication, the results indicate a generally negative influence of alcohol intoxication. Additionally, alcohol intoxication can affect one's speaking fundamental frequency(F0). In [8–10] a notable increase in average f0 with intoxication is found, whereas a decrease is reported in [11, 12].

As knowing the effects of alcohol on speech, some listening experiments are proposed to recognize the intoxicated speech manually. Pisoni et al.[13] perform an experiment that college students and State Troopers hear 192 sentences from 8 talkers. The mean accuracy across all of the sentences was 61.5% for the college students, and 64.7% for the State Troopers. The BAC level can affect recognition accuracy significantly. Klingholz et al.[14] shows that when the BAC exceeds 1.0 per mill, the recognition rate can achieve a maximum of 82.0%. When the BAC is lower than 1.0 per mill, however, the accuracy decrease to 54.0%. Recently, Baumeister and Schiel's[15] experiments show that the average overall performance of the listeners on ALC dataset is 61.8%. Since these human performance experiments can achieve a higher recognition rate than chance, the machine learning methods may also be applied to intoxication detection, as other paralinguistic tasks such as fatigue detection.

In the INTERSPEECH 2011 Speaker State Challenge[16], the unweighted accuracy(UA) of official baseline in the Intoxication Sub-Challenge is 65.9% on the test set. Bone et al.[17] perform global and iterative speaker normalization on the feature and finally achieve the UA of 70.54%, which is an improvement of 4.64% absolute over the baseline model. They also refine their work in 2014 and finally achieve the UA of 71.4% on the test set[18].

Since the speaker normalization is so robust and significantly improve the performance of intoxicated speech detection task, we also propose a batch level speaker normalization method with a residual neural network(ResNet). The experiment results show that using speaker normalization improve the performance of both the baseline system and the proposed system.

The rest of our paper is organized as follows. Section 2 describes the information of Alcohol language corpus dataset. Section 3 is our proposed method. Experiments and results are presented in Section 4. Section 5 is the conclusion.

II. Method

In this section, we introduce our baseline system using ComParE acoustic feature set followed by the SVM. Besides, we propose a deep neural network system applying ResNet as pattern extractor on ALC dataset.

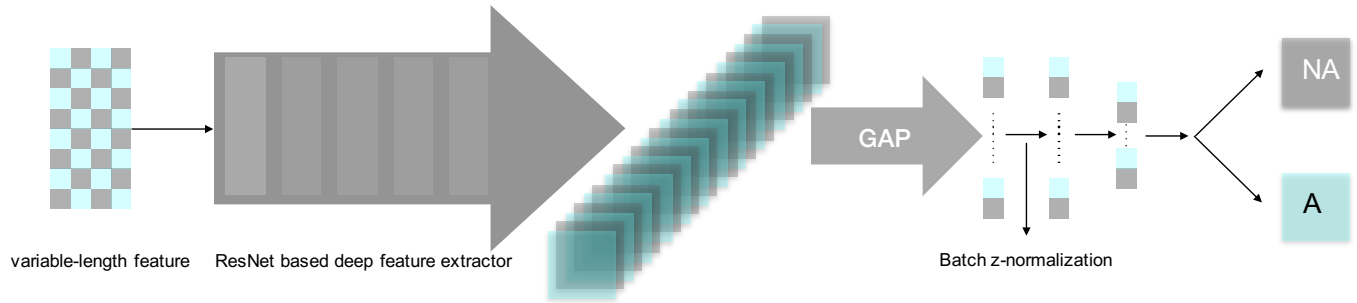

Fig. 1. Architecture of proposed model with batch z-normalization

A. ComParE Baseline

The baseline system is based on the recent ComParE challenge[19], using the Support Vector Machines (SVM) with the linear kernel as the classifier.

We employ the ComParE acoustic feature set for the baseline system. This feature set is the official baseline feature which has been used in ComParE Challenge since 2013[20]. It contains 6373 static features which are generated by various functionals on low-level descriptors (LLDs). Features like voice quality features, F_0, energy, cepstral, spectral, HNR (logarithmic harmonic-to-noise ratio) are included in the set. We extract the feature set and LLDs with the OPENSMILE toolkit[21, 22].

Practically, we scale the features to zero mean and unit standard deviation within all data and apply the linear Support Vector Classification (linearSVC) to the intoxication detection task. Additionally, speaker normalization has been proved to be robust on this task [17, 18]. Differences between speakers could affect the detection of intoxication. Intuitively, eliminating the speaker information would help improve accuracy. Therefore, we also perform speaker normalization on both training and testing data, which means that the z-normalization is applied to the features within each speaker.

B. ResNet based framework

ResNet has been proved efficient on image classification task [23]. The network learns with the reference of the information from each input layer. With this structure, the model can be easier to optimize.

In recent years, Cai et al.[24, 25] apply ResNet to the areas of speaker verification and language identification task and achieve a much more favorable performance than the traditional methods. In many paralinguistic speech attribute recognition tasks, ResNet is also widely applied to learn the attributes.

In the light of previous works, we adopt ResNet as our deep pattern extractor. Our proposed network architecture is illustrated in Fig 1. In this section, we introduce our framework as follows.

1) Frame-level feature extraction: First, we extract frame-level spectral features from raw wavform using STFT as well as other hand-crafted filters. The shape of the output feature is $D \times T$, where D means the dimension on the frequency axis and T denotes the number of frames. The filter bank is used as our input feature here.

2) CNN representations: We use ResNet structure acting as a local pattern extractor to fetch the abstract representations of the frame-level features. The shape of feature matrix $D \times T$ is then transformed into $C \times H \times W$. C, H and W denote the number of channels, the height, and width of the ResNet feature maps.

3) Utterance-level embedding extraction: Since the audio signals are variable, the shapes of output CNN representations are not constant. We need to further extract a fixed size features for the back end classifier. To achieve this goal, we adopt the Global Average Pooling (GAP) layer on top of the ResNet structure. The GAP layer accumulates the statistics by taking the means along with the time-frequency axis. Given an output feature map $\boldsymbol{F}$ with a size of $C \times H \times W$, the process can be formulated as:

$$u_k = \frac{1}{H \times W} \times \sum_{H}^{i=1} \sum_{W}^{j=1} \boldsymbol{F}_{i,j,k} \tag{1}$$

With the GAP layer, we can get an utterance-level feature $\boldsymbol{u} = [u_1, u_2, \ldots, u_C]$ for each sample.

In the training stage, the input frame-level features are fed into the ResNet with different sizes of batches. Truncating and padding are applied to ensure every input feature in a batch is in the same size. In the evaluation stage, we directly input the frame-level features to the network to obtain fixed size utterance-level representations.

4) DNN classifier: We construct a two fully-connected layer structure as our back end classifier. The two output units of the final layer represent alcoholised(A) and non-alcoholised(NA). Through this classifier, we can finally obtain the decisions from the network.

C. Batch z-normalization

Z-normalization, also known as "Normalization to Zero Mean and Unit of Standard Deviation", is defined as:

$$x_i^{'} = \frac{x_i - \mu}{\sigma} \tag{2}$$

where x_i denotes the feature of each dimension.

Since ComParE feature set has a constant length of 6373, we can easily apply speaker normalization to this feature.

However, when training a neural network, the length of the feature is different, and the content in each utterance is also distinct. To reduce the influence of speaker information, we perform z-normalization after the GAP layers, but it also means that the data in one batch must belong to the same speaker, as shown in Fig 1.

Although the batch z-normalization can reduce the speaker information, it means that in each batch the training data should be extracted from the same speaker. In the testing step, we should also know the speaker label, but sometimes the speaker labels are unknown. Therefore, we employ the clustering method to predict the speaker label for testing data.

III. Experiments and results

A. Dataset Description

Alcohol language corpus(ALC) comprises intoxicated and sober speech of 162 German speakers(84 male and 78 female) within the age range 21 to 75[26, 27]. The type of speech ranges from reading single digits to full conversation style, and the level of intoxication was measured by BAC or BrAC before the speech record. Each recording contains several speech items: monologues, dialogues, numbers, command&control, addresses, and tongue twisters.

Speakers can choose the BAC he/she wanted to reach during the intoxication test. After consuming the alcohol, the actual level of intoxication was measured by breath alcohol, and blood samples were taken immediately before the speech recording. If the breath alcohol concentration is above 0.05%, the participant is eligible for the speech test, and his/her recordings can be labeled as alcoholished(A). Then, at least one week later, the speaker is recorded again, and his/her recordings will be labeled as non-alcoholished(NA). Meanwhile, the information of each speaker, such as gender, age, and weight, was also documented for the additional research and investigation.

In our experiment, we only use the data in BLOCK 10 to 40. BLOCK10 and BLOCK30 are intoxicated data, and BLOCK20 and BLOCK40 are sober data. Since some speakers with small BAC are also grouped to intoxication, we label speakers with BAC equal or below 0.5 per mill to non-alcoholised(NA) and those with BAC exceeding 0.5 per mill to alcoholised(A), which is the same as the setup of Intoxication Sub-Challenge in 2011[16].

B. Data preprocessing

All waves in ALC are 44100 Hz and too long for paralinguistic tasks. We first downsample audio files to 16000 Hz and apply voice activity detection(VAD) to the downsampled audio. Then, we randomly split all wav to shorter utterances(duration: 3.50 ± 1.48) and drop the original utterances shorter than 2 seconds, finally producing 29267 utterances from 162 speakers. For our experiments, we randomly picked 22758 utterances from 132 speakers for training and 6509 utterances from 30 speakers for testing. After preprocessing, we can obtain 20372 non-alcoholished(NA) utterance and 8895 alcoholished(A) utterance.

TABLE I
THE DETAILED NETWORK STRUCTURE OF OUR RESNET18 BASED NETWORK

Layer	Input size	Output size	Structure
Conv1	$1 \times 64 \times L$	$16 \times 64 \times L$	3×3, stride 1
Res1	$16 \times 64 \times L$	$16 \times 64 \times L$	block $\times$ 2
Res2	$16 \times 64 \times L$	$32 \times 32 \times \frac{L}{2}$	block $\times$ 2
Res3	$32 \times 32 \times \frac{L}{2}$	$64 \times 16 \times \frac{L}{4}$	block $\times$ 2
Res4	$64 \times 16 \times \frac{L}{4}$	$128 \times 8 \times \frac{L}{8}$	block $\times$ 2
GAP	$128 \times 8 \times \frac{L}{8}$	128	pooling
z-norm	128	128	z-normalization
FC	128	64	fully-connected
Output	64	2	fully-connected

The acoustic feature is filter-bank of 64 dimensions with 25ms frame length and 10 frame shift; then an utterance level mean subtraction is applied to all feature.

C. Baseline method

The description of the baseline feature is in Section 3.1, which is a 6373-dimensional feature set. We train several linearSVC models on the training data and test on the testing data, producing three baseline due to different data processing.

For baseline 1, we scale the training data with the MINMAXSCALER of SCIKIT-LEARN[28] and use the parameters from the training set to scale the testing data.

For baseline 2, we perform the z-normalization for each speaker on both training and testing data.

However, in most situation, we do not know the information about the speakers. Therefore, for baseline 3, we extract the i-vector of each utterance in testing data with a pretrained GMM-UBM based speaker recognition system, which is trained on the voxceleb[29] with kaldi toolkit[30]. Then we perform the L2-normalization on all i-vector and employ the spectral clustering on testing data to obtain the clustered speaker label. Therefore we can perform the z-normalization again on training and testing data even though we do not know the speaker label of testing data.

D. Proposed method

There are also three ResNet models due to different data processing. In our experiment, we use ResNet18 for training. The detailed network structure is in Table I and Fig 1.

For the first proposed method without the speaker normalization, we can randomly select data from 132 speakers for training and test each utterance from 30 speakers, as mentioned in Section 4.1.

For the second proposed method that uses the true speaker label for speaker normalization, we have to pick up the data from the same speaker for training each batch and perform the batch z-normalization as speaker normalization. Also, the testing data have to be grouped by the speaker label; then we test each group with batch z-normalization.

For the third proposed method that uses the spectral clustered label for speaker normalization, the setup of the training

TABLE II
UAR OF EACH METHOD

System	UAR
baseline 1(without normalization)	0.616
baseline 2(with speaker normalization)	0.669
baseline 3(with spectral clustering and speaker normalization)	0.665
ResNet(without normalization)	0.633
ResNet(with batch z-normalization)	0.677
ResNet(with spectral clustering and batch z-normalization)	0.671

step does not change. For the testing step, the only difference is that we grouped the data by the label from spectral clustering.

E. Results

The metric we use in out experiments is unweighted average recall(UAR). The recall of each class is defined as:

$$recall = \frac{1}{T}\sum_{t=1}^{T}\frac{TP[t]}{TP[t]+FN[t]} \tag{3}$$

where T is the number of samples and TP, FP, and FN denote true positive, false positive, and false negative, respectively. Then UAR can be calculated by:

$$UAR = \frac{\sum_{i=1}^{n} r_i}{n} \tag{4}$$

where n is the number of class and r_i is the recall of each class.

The results in Table II shows that the UAR of baseline 1 and baseline 2 are 0.616 and 0.669, respectively. The baseline 3, which speaker labels on testing data are given by the result of spectral clustering, achieves the UAR of 0.665. The UAR of baseline 2 and baseline 3 are better than the UAR of baseline 1, and it is a surprise to see that the UAR of baseline 2 and baseline 3 are very similar. The result shows that even we do not know the speaker label of testing data, the speaker normalization using clustering result as the label can also contribute to the intoxication detection task.

For the proposed method, we train each network for 50 epochs, and the result is the mean of 3 times training. If we do not perform speaker normalization on the network, the UAR is only 0.633. When performing the batch z-normalization with true speaker label, the UAR increase to 0.677. In addition, when given the speaker label from clustering, the performance also improves and the UAR is 0.671. All results show that the proposed methods perform better than baseline, and the improvement on both the baseline and the proposed methods prove that speaker normalization can improve the accuracy of intoxication detection task.

IV. Conclusions and Future Work

In this paper, we propose a ResNet based model for intoxication detection task. First, the model can receive variant-length acoustic feature and produce a time-invariant deep embedding feature for the last two fully-connected layers to classify. Second, batch z-normalization is proposed. The results show that ResNet with batch z-normalization achieve a similar improvement as the baseline method with speaker normalization. The improvement of batch z-normalization on the proposed model shows that the embeddings after the TAP layer still contain some speaker information, even though the model is used for classifying the intoxicated state. After batch z-normalization, the speaker information decreases, thus we can obtain a better performance.Third, we extract the i-vector and use the spectral clustering predicted label as the speaker label. Both the baseline and proposed methods do not degrade too much. All the results show that speaker information is an essential factor to be considered in the intoxication detection task.

In the future, we will further investigate the influence of speaker information on intoxication detection and explore the speaker normalization in an end-to-end framework.

V. Acknowledgements

This research was funded in part by the National Natural Science Foundation of China (61773413), Natural Science Foundation of Guangzhou City (201707010363), Six Talent Peaks project in Jiangsu Province (JY-074), Science and Technology Program of Guangzhou City (201903010040) and DiDi chuxing.

References

[1] T. A. Schweizer, P. Jolicœur, M. Vogel-Sprott, and M. J. Dixon, "Fast, but error-prone, responses during acute alcohol intoxication: effects of stimulus-response mapping complexity," *Alcoholism: Clinical and Experimental Research*, vol. 28, no. 4, pp. 643–649, 2004.

[2] B. D. Abroms and M. T. Fillmore, "Alcohol-induced impairment of inhibitory mechanisms involved in visual search." *Experimental and clinical psychopharmacology*, vol. 12, no. 4, p. 243, 2004.

[3] N. Galbraith, "Alcohol: its effect on handwriting," *Journal of Forensic Science*, vol. 31, no. 2, pp. 580–588, 1986.

[4] F. Aşıcıoğlu and N. Turan, "Handwriting changes under the effect of alcohol," *Forensic science international*, vol. 132, no. 3, pp. 201–210, 2003.

[5] B. Schuller, S. Steidl, A. Batliner, F. Schiel, J. Krajewski, F. Weninger, and F. Eyben, "Medium-term speaker statesa review on intoxication, sleepiness and the first challenge," *Computer Speech & Language*, vol. 28, no. 2, pp. 346–374, 2014.

[6] F. Schiel and C. Heinrich, "Disfluencies in the speech of intoxicated speakers." *International Journal of Speech, Language & the Law*, vol. 22, no. 1, 2015.

[7] J. T. Geiger, B. Zhang, B. Schuller, and G. Rigoll, "On the influence of alcohol intoxication on speaker recognition," in *Proc. AES 53rd International Conference Semantic Audio, AES, London, UK*, 2014.

[8] B. Baumeister, C. Heinrich, and F. Schiel, "The influence of alcoholic intoxication on the fundamental frequency of female and male speakers," *The Journal of the Acoustical Society of America*, vol. 132, no. 1, pp. 442–451, 2012.

[9] H. Hollien, G. DeJong, C. A. Martin, R. Schwartz, and K. Liljegren, "Effects of ethanol intoxication on speech suprasegmentals," *The Journal of the Acoustical Society of America*, vol. 110, no. 6, pp. 3198–3206, 2001.

[10] F. Klingholz, R. Penning, and E. Liebhardt, "Recognition of low-level alcohol intoxication from speech signal," *The Journal of the Acoustical Society of America*, vol. 84, no. 3, pp. 929–935, 1988.

[11] H. Watanabe, T. Shin, H. Matsuo, F. Okuno, T. Tsuji, M. Matsuoka, J. Fukaura, and H. Matsunaga, "Studies on vocal fold injection and changes in pitch associated with alcohol intake," *Journal of Voice*, vol. 8, no. 4, pp. 340–346, 1994.

[12] G. A. Alderman, H. Hollien, C. Martin, and G. DeJong, "Shifts in fundamental frequency and articulation resulting from intoxication," *The Journal of the Acoustical Society of America*, vol. 97, no. 5, pp. 3363–3364, 1995.

[13] D. B. Pisoni and C. S. Martin, "Effects of alcohol on the acoustic-phonetic properties of speech: perceptual and acoustic analyses," *Alcoholism: Clinical and Experimental Research*, vol. 13, no. 4, pp. 577–587, 1989.

[14] F. Klingholz, R. Penning, and E. Liebhardt, "Recognition of low-level alcohol intoxication from speech signal," *The Journal of the Acoustical Society of America*, vol. 84, no. 3, pp. 929–935, 1988.

[15] B. Baumeister and F. Schiel, "Fundamental frequency and human perception of alcoholic intoxication in speech." in *ICPhS*, 2015.

[16] B. Schuller, S. Steidl, A. Batliner, F. Schiel, and J. Krajewski, "The interspeech 2011 speaker state challenge," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[17] D. Bone, M. P. Black, M. Li, A. Metallinou, S. Lee, and S. Narayanan, "Intoxicated speech detection by fusion of speaker normalized hierarchical features and gmm supervectors," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[18] D. Bone, M. Li, M. P. Black, and S. S. Narayanan, "Intoxicated speech detection: A fusion framework with speaker-normalized hierarchical functionals and gmm supervectors," *Computer speech & language*, vol. 28, no. 2, pp. 375–391, 2014.

[19] B. W. Schuller, A. Batliner, C. Bergler, F. B. Pokorny, J. Krajewski, M. Cychosz, R. Vollmann, S.-D. Roelen, S. Schnieder, E. Bergelson10 *et al.*, "The interspeech 2019 computational paralinguistics challenge: Styrian dialects, continuous sleepiness, baby sounds & orca activity."

[20] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi *et al.*, "The interspeech 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism," in *Proceedings INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France*, 2013.

[21] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*. ACM, 2010, pp. 1459–1462.

[22] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proceedings of the 21st ACM international conference on Multimedia*. ACM, 2013, pp. 835–838.

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[24] W. Cai, Z. Cai, X. Zhang, X. Wang, and M. Li, "A novel learnable dictionary encoding layer for end-to-end language identification," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5189–5193.

[25] W. Cai, J. Chen, and M. Li, "Exploring the encoding layer and loss function in end-to-end speaker and language recognition system," *arXiv preprint arXiv:1804.05160*, 2018.

[26] F. Schiel and C. Heinrich, "Laying the foundation for in-car alcohol detection by speech," in *Tenth Annual Conference of the International Speech Communication Association*, 2009.

[27] F. Schiel, C. Heinrich, and S. Barfüsser, "Alcohol language corpus: the first public corpus of alcoholized german speech," *Language resources and evaluation*, vol. 46, no. 3, pp. 503–521, 2012.

[28] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *Journal of machine learning research*, vol. 12, no. Oct, pp. 2825–2830, 2011.

[29] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: a large-scale speaker identification dataset," *arXiv preprint arXiv:1706.08612*, 2017.

[30] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," IEEE Signal Processing Society, Tech. Rep., 2011.