Smart Health 36 (2025) 100562

Contents lists available at ScienceDirect

Smart Health

journal homepage: www.elsevier.com/locate/smhl

Intoxication detection from speech using representations learned from self-supervised pre-training

Abigail Albuquerque *, Samuel Chibuoyim Uche, Emmanuel Agu

Department of Computer Science, Worcester Polytechnic Institute (WPI), Worcester, MA, United States of America

ARTICLE INFO

Keywords:
Transformers
Speech classification
Wav2Vec 2.0

ABSTRACT

Alcohol intoxication is one of the leading causes of death around the globe. Existing approaches to prevent Driving Under the Influence (DUI) are expensive, intrusive, or require external apparatus such as breathalyzers, which the drinker may not possess. Speech is a viable modality for detecting intoxication from changes in vocal patterns. Intoxicated speech is slower, has lower amplitude, and is more prone to errors at the sentence, word, and phonological levels than sober speech. However, intoxication detection from speech is challenging due to high inter- and intra-user variability and the confounding effects of other factors such as fatigue, which may also impair speech. This paper investigates Wav2Vec 2.0, a self-supervised neural network architecture, for intoxication classification from audio. Wav2Vec 2.0 is a Transformer-based model that has demonstrated remarkable performance in various speech-related tasks. It analyzes raw audio directly by applying a multi-head attention mechanism to latent audio representations and was pre-trained on the Librispeech, Libri-Light and EmoDB datasets. The proposed model achieved an unweighted average recall of 73.3%, outperforming state-of-the-art models, highlighting its potential for accurate DUI detection to prevent alcohol-related incidents.

1. Introduction

Motivation: Alcohol is one of the most widely abused substances (Jacob & Wang, 2020), often consumed as a coping mechanism and recreational substance. However, alcohol impairs neuronal transmission and significantly alters consciousness (Taylor et al., 2010), attention and behavioral control of drinkers (Lin et al., 2022). Driving Under the Influence (DUI) of alcohol is particularly dangerous as it also interferes with visual accuracy, perception, and psychomotor functions (Jovanovic, Jovanovic, Vukovic, & Jevremovic, 2000). While the legal driving limit is 0.08% in most states in the US, a Blood Alcohol Concentration (BAC) of as small as 0.02 can slow down a driver's reaction times and decision-making processes (Hingson, 1996). Over 30% of vehicular accidents are caused by drivers intoxicated by alcohol (National Highway Traffic Safety Administration, 2023). Initiatives to reduce alcohol-related traffic fatalities include a minimum legal drinking age and legal action against DUIs (Hingson, 1996). However, existing measures are intrusive, reactive, or require the acquisition of additional apparatus to identify intoxicated drivers. Alcohol impairs the drinkers' speech (Hollien, DeJong, Martin, Schwartz, & Liljegren, 2001); intoxicated speech is slower, has lower amplitude, and is more prone to errors at the sentence, word and phonological levels than sober speech (Pisoni & Martin, 1989).

Specific Problem: We propose a neural networks framework for detecting alcohol-induced speech impairment and inebriation by using machine learning methods, transforming it to an audio classification problem.

Challenges: Include person-to-person variability in alcohol tolerance and speech impairment, and confounding by other factors such as fatigue or substance use (*e.g.*, marijuana). Additionally, there are few publicly available intoxicated speech datasets.

* Corresponding author.
E-mail addresses: aralbuquerque@wpi.edu (A. Albuquerque), scuche@wpi.edu (S.C. Uche), emmanuel@wpi.edu (E. Agu).

https://doi.org/10.1016/j.smhl.2025.100562
Received 10 March 2025; Accepted 15 March 2025

Available online 27 March 2025

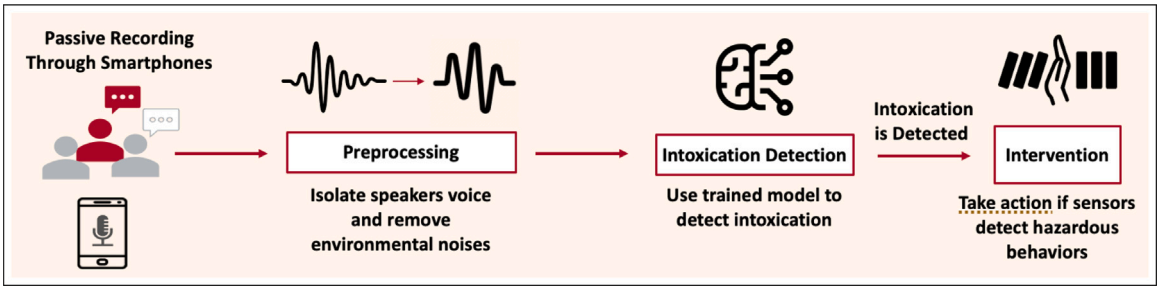

Fig. 1. Overview of pipeline for speech intoxication.

This study uses the Alcohol Language Corpus (Schiel, Heinrich, Barfüßer, & Gilg, 2008), which contains 15,180 recordings of conversational and scripted speech in both sober and intoxicated states. Despite being one of the largest datasets available, its size remains a challenge for training deep learning models effectively.

Proposed Approach: We fine-tune Wav2Vec 2.0, a state-of-the-art, self-supervised model pre-trained on large speech corpora, for intoxication detection. Fig. 1 illustrates its real-world application. Wav2Vec 2.0 has excelled in speech tasks such as Automatic Speech Recognition (ASR), speaker verification, and mispronunciation detection (Baevski, Zhou, Mohamed, & Auli, 2020; Hsu et al., 2021; Peng, Fu, Lin, Ke, & Zhan, 2021). Fine-tuning allows adaptation to domain-specific problems, particularly for small datasets. Its self-supervised design enhances generalizability (Zhang et al., 2024), crucial given our dataset size. The model extracts features from audio waveforms via transformer blocks optimized for temporal sequences (Baevski et al., 2020). A classification head then predicts intoxication. Pre-trained on large datasets such as Librispeech and Libri-Light, and fine-tuned on EmoDB for emotion recognition, Wav2Vec 2.0 is further fine-tuned for intoxication detection.

2. Background and related work

Human-Level Accuracy in Detecting Alcohol Intoxication: Baumeister and Schiel evaluated human performance in recognizing alcohol intoxication through speech, and the accuracy averaged at 63.1% (Baumeister, Heinrich, & Schiel, 2012). Humans usually utilize visual and scent clues to determine intoxication.

Unweighted Average Recall (UAR) is a metric commonly used for this dataset. It balances performance across imbalanced classes expressed as $\text{UAR} = \frac{\text{Sensitivity}+\text{Specificity}}{2}$, where sensitivity (true positive rate) and specificity (true negative rate) measure correct class identifications.

Audio Features used in baseline models include those from OpenSMILE, Praat, and Mel spectrograms (Berninger, Hoppe, & Milde, 2016; Bone et al., 2011; Bonela et al., 2023; Hönig, Batliner, & Nöth, 2011). OpenSMILE extracts features such as MFCCs, PLPCCs, pitch, and formant frequencies. Mel spectrograms represent audio visually, enhancing classification.

2.1. Prior machine learning approaches

2.1.1. Intoxication detection machine learning models proposed at interspeech 2011

Bone et al. achieved a UAR of 70.54% (Bone et al., 2011) using a pipeline with pause removal, feature extraction via OpenSMILE and Praat, iterative speaker normalization, and classifying GMM supervectors with an SVM.

2.2. Prior deep learning approaches

Deep learning research on this dataset is still limited. Table 1 summarizes related work. Bonela et al. developed the Audio Based Deep Learning Algorithm to Identify Alcohol Inebriation (ADLAIA). They transformed audio into log Mel spectrograms for a ResNet-18 model pre-trained on ImageNet, applying a weighted loss for class imbalance, achieving a UAR of 68.09% (Bonela et al., 2023). Berninger et al. used 40-D FBANK features with a BiRNN, achieving the highest UAR (71.03%) (Berninger et al., 2016).

3. Methodology

In this paper, we fine-tune a pre-trained Wav2Vec 2.0 model for speech representation learning. Wav2Vec 2.0 is a self-supervised framework designed to extract meaningful representations from raw speech audio without requiring extensive labeled data. It comprises a convolutional feature encoder, a Transformer-based context network, and a quantization module (Baevski et al., 2020) as shown in Fig. 2. The masked latent representations are passed through a Transformer-based context network, which builds rich contextual embeddings of the audio. The model is trained using contrastive loss, where the objective is to distinguish the true latent representation from distractors. Further, discrete speech units are learned via the Gumbel Softmax approach (Baevski et al., 2020).

Table 1
Previous work related to the ALC dataset.

Authors	Name of paper	Summary of approach	Best UAR
Bone et al. (2011)	Intoxicated speech detection by fusion of speaker normalized hierarchical features and GMM supervectors	The winning INTERSPEECH paper combined speaker-normalized hierarchical features and Gaussian Mixture Model (GMM) supervectors as their features. They used an SVM with linear kernel, L2 regularization, and L2 loss.	70.54%
Bonela et al. (2023)	Audio-based Deep Learning Algorithm to Identify Alcohol Inebriation (ADLAIA)	This paper used log Mel spectrogram representations of the audio recordings and trained a Convolutional Neural Network with a ResNet-18 model pre-trained on the ImageNet dataset.	68.09%
Berninger et al. (2016)	Classification of Speaker Intoxication Using a Bidirectional Recurrent Neural Network	The paper achieving the highest accuracy so far used 40-dimensional FBANK features by the CMU Sphinx speech recognition toolkit and a Bi-directional Neural Network.	71.03%

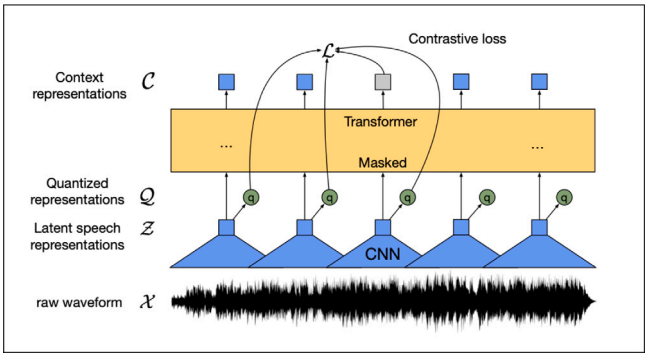

Fig. 2. Illustration of the Wav2Vec 2.0 framework showing the self-supervised pre-training process (Baevski et al., 2020).

3.1. Feature encoder

The feature encoder is a multi-layer convolutional neural network which processes raw audio inputs and outputs latent speech representations (Baevski et al., 2020). The raw input is denoted as X, and L is the length of the audio signal. The audio is first normalized to have zero mean and unit variance: $\widetilde{X} = \frac{X-\mu}{\sigma}$ where μ and σ are the mean and standard deviation of the signal. The normalized signal $\widetilde{X}$ is then passed through convolutional blocks. Each block consists of:

(1) A temporal convolution layer with kernel size k, stride s, and number of channels C.
(2) Layer normalization to stabilize training.
(3) A GELU activation function for non-linearity.

The total stride of the encoder determines the number of time steps T given by L/s which are input to the transformer. The output of latent speech representation is represented by $Z = z_1, \ldots, z_T$. The pre-trained model used in this study contains 7 convolutional blocks with 512 channels and respective strides, s, of (5, 2, 2, 2, 2, 2, 2) and kernel sizes, k, of (10, 3, 3, 3, 3, 2, 2) (Baevski et al., 2020).

3.2. Transformer context network

The output of the feature encoder is fed to a context network, which follows the Transformer architecture. The context network stacks 12 Transformer blocks, each with model dimension $d_{model} = 768$, inner dimension $d_{inner} = 3072$ and 8 attention heads. Instead of fixed positional embeddings which encode absolute positional information, the model uses a convolutional layer, which acts as relative positional embedding. The output of the convolution is added followed by a GELU to the inputs and then layer normalization is applied (Baevski et al., 2020). Each of the 12 Transformer blocks processes these normalized inputs. First, the input sequence needs to go through a feature projection layer, to increase the dimension from 512 to 768. The transformer uses attention to boost the speed in which models can be trained. The transformer contextualizes the masked representations and generates context representations C (Baevski et al., 2020).

3.3. Quantization module

The quantization module is used to discretize latent speech representations Z into Q using product quantization. Product quantization involves choosing quantized representations from multiple codebooks and concatenating them. Given G codebooks

(or groups), with V entries $e \in \mathbb{R}^{V \times \frac{d}{G}}$, one entry is chosen from each codebook and concatenated with resulting vectors $e_1, \ldots, e_G$ and a linear transformation is applied $\mathbb{R}^d \rightarrow \mathbb{R}^f$ to obtain $q \in \mathbb{R}^f$ (Baevski et al., 2020).

To allow the selection of codebook entries in a fully differentiable way, the Gumbel softmax is used. The module uses the straight-through estimator to compute gradients and setup G hard Gumbel softmax operations. The feature encoder output z is mapped to $l \in \mathbb{R}^{G \times V}$ logits and the probabilities for choosing the v-th codebook entry for group g are

$$p_{g,v} = \frac{\exp\left(l_{g,v} + n_v\right)/\tau}{\sum_{k=1}^{V} \exp\left(l_{g,k} + n_k\right)/\tau},$$

where τ is a temperature parameter controlling the smoothness of the distribution, $n = -\log(-\log(u))$, and $u \sim \mathcal{U}(0,1)$ are uniform random samples. During the forward pass, the selected codeword i for group g is determined by $i = \arg\max_j p_{g,j}$, while in the backward pass, the gradients of the Gumbel softmax outputs are used (Baevski et al., 2020). There are $G = 2$ codebooks in the Quantization module, each containing $V = 320$ with a size of 128.

3.4. Pre-training

To pre-train the model, certain proportions of time steps in the latent feature encoder space are masked. The representations in the audio are learned through a contrastive task L_m, which needs to identify the accurate quantized latent audio representation in a set of distractors. L_m is augmented by a codebook diversity loss L_d that encourages the model to use the codebook entries equally often. The loss is defined as $L = L_M + \alpha L_D$. Given context network output c_t centered over masked time step t, the model needs to identify the true quantized latent speech representation q_t in a set of $K + 1$ quantized candidate representations $\tilde{q}_t \in Q_t$ which includes q_t and K distractors, the contrastive loss is defined as:

$$L_M = -\log \frac{\exp(\mathrm{sim}(c_t, q_t)/k)}{\sum_{\tilde{q} \sim Q_t} \exp(\mathrm{sim}(c_t, \tilde{q}))/k} \tag{1}$$

where the cosine similarity between context representations and quantized latent speech representations is computed. And the diversity loss is defined as:

$$L_D = \frac{1}{GV} \sum_{g=1}^{G} \sum_{v=1}^{V} \bar{p}_{g,v} \log \bar{p}_{g,v}, \tag{2}$$

which maximizes the entropy of the average softmax distribution 1 over the codebook entries for each codebook $\bar{p}_g$ across a batch of utterances and encourages of the use of V entries in each of the G codebooks.

3.5. Adaptation for fine-tuning

The pre-trained models are fine-tuned for speech recognition. This is performed by adding a randomly initialized linear projection on top of the context network representing the vocabulary. The model used in this study was trained on LibriSpeech and Librilight, and further fine-tuned for classification on the German EmoDB dataset (Burkhardt et al., 2005). Fine-tuning Wav2Vec 2.0 is essential as the pre-trained model captures general speech features, while fine-tuning adapts it to German and task-specific speech characteristics. Prior research (Baevski et al., 2020) showed fine-tuning significantly improves domain adaptation and performance. Finally, for our classification purposes, a lightweight classifier head consisting of a dropout layer, regularizing the model to prevent overfitting, followed by a linear layer maps the encoder's output to the desired single output.

4. Evaluation

The evaluation metrics used in evaluating the model and baselines include accuracy, unweighted average recall (UAR), sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV). UAR was specifically chosen as the dataset is heavily imbalanced, and the metric has been utilized to evaluate prior ALC work.

4.1. Baseline models

Models containing numerical and image-based features, as well as self-supervised models, were trained in combination with traditional machine learning methods as well as deep learning models such as CNNs, which are used in cutting edge speech classification studies. A list of these baselines as well as their inputs are highlighted in Table 2.

4.2. Datasets

4.2.1. Pre-training datasets

To generate a robust model, Wav2Vec 2.0 is pre-trained on a diverse mix of datasets:

(1) **LibriSpeech** – 1000 h of audiobook speech, aiding phonetic and linguistic pattern learning.
(2) **Libri-Light** – 60,000+ hours of unlabeled speech with a smaller labeled subset, enhancing robustness to accents and noise.
(3) **EmoDB** – 535 German utterances across seven emotions, improving prosody, pitch, and emotional variation modeling relevant to intoxication detection.

Table 2
A list of models used as baselines, along with their model type and inputs.

Input	Model	Model type
Self-Supervised	MERT-v1-95M (Li et al., 2024)	Transformer
	HuBERT-base-ls960 (Hsu et al., 2021)	Transformer
Image	AST-base384 (Gong, Chung, & Glass, 2021)	Vision Transformer
	EfficientNet (Tan & Le, 2020)	CNN
	ResNet (He, Zhang, Ren, & Sun, 2015)	CNN
Numerical	Naive Bayes (Ng & Jordan, 2001)	Classical ML
	SVM (Hearst, Dumais, Osuna, Platt, & Scholkopf, 1998)	Classical ML
	Random Forest (Breiman, 2001)	Classical ML

Table 3
Distribution of ALC Classes showing heavy class imbalance.

Class	Number of recordings
Sober	10 540
Intoxicated	4640
Total	15 180

4.2.2. The Alcohol Language Corpus (ALC) dataset

The ALC dataset (2007–2009) contains recordings of 162 German speakers (78 female, 84 male, aged 21–75) in intoxicated and sober states. Speech types include digit reading and conversation. Intoxication levels were self-chosen and measured via breath and blood samples. Each speaker contributed 30 intoxicated and 60 sober samples, totaling 15,180 validated recordings. As we see in Table 3, the classes are heavily imbalanced.

We opted for a stratified sampling approach. This method allowed us to maintain similar distributions across the training, test, and validation sets based on Blood Alcohol Concentration (BAC) levels which were classified as follows:

- 0: $BAC = 0.0$
- 1: $0.01 \leq BAC \leq 0.019$
- 2: $0.02 \leq BAC \leq 0.049$
- 3: $0.05 \leq BAC \leq 0.079$
- 4: $0.08 \leq BAC$

The subsets are speaker-disjoint, ensuring no speaker appears in both training and test/validation sets. A speaker is intoxicated if their BAC is over 0.05. The data was divided into training, validation, and testing sets in a 60:20:20 ratio, respectively. While BAC level distributions remain similar, exact matching is unattainable due to data constraints. To mitigate class imbalance, the minority class was upsampled to a 7:10 ratio with the majority class, selected based on validation performance.

The Wav2Vec 2.0 model was pre-trained on LibriSpeech, Libri-Light, and EmoDB. Audio was preprocessed by removing pauses and noise, then sampled at 16 kHz. A binary classification head was added for intoxication detection. Fine-tuning was performed using AdamW and Binary Cross Entropy with Logits loss, incorporating weighted loss to prioritize the minority class. A low learning rate of $1e-6$ was chosen after extensive hyperparameter tuning.

5. Results

The performance on the test split and the comparison to a variety of baselines are highlighted in Table 4.

Wav2Vec 2.0 outperforms all models and baselines, achieving superior UAR compared to prior research on the ALC dataset. However, its performance in FPR and PPV is lower, indicating confusion likely due to class imbalance in the binary task, where the minority class is underrepresented. This leads the model to prioritize accuracy and recall over precision and false positive control.

Confusion matrices show good performance with slight confusion. As seen in Fig. 3, the diagonal is strong, indicating solid performance, with more True Positives and True Negatives than False Positives and False Negatives. However, False Positives and False Negatives occur at nearly equal rates.

6. Discussion

The Wav2Vec 2.0 model outperforms baselines, achieving a UAR of 73.3%, surpassing the best baseline by 7%. This highlights the effectiveness of self-supervised learning and pre-training on large speech datasets, facilitating transfer to intoxicated speech classification. In spite of class imbalance and fine-grained class boundaries, the model maintains balanced performance with sensitivity (0.73) and specificity (0.74). It outperforms prior deep learning and machine learning approaches, surpassing the best reported UAR on the ALC dataset by 2%. Unlike previous studies relying on hand-engineered features, self-supervised learning enhances feature extraction and adaptation to complex datasets.

Challenges remain in fine-grained intoxication detection. While achieving high UAR, the model exhibits nearly equal False Positives and False Negatives, likely due to class imbalance and subtle speech variations.

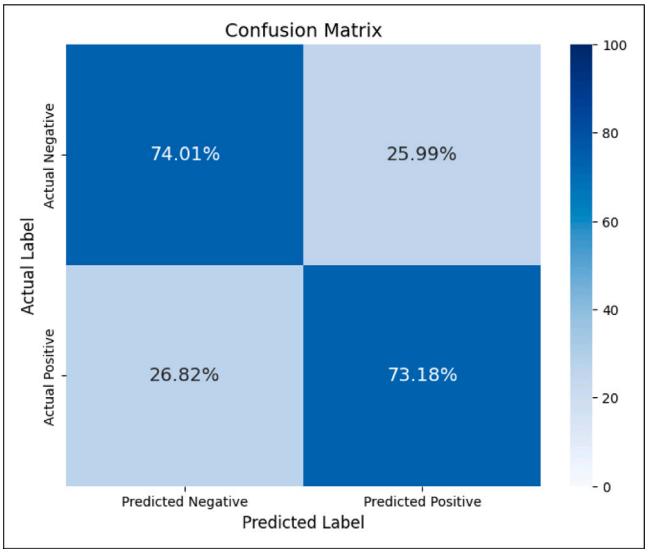

Fig. 3. The model's confusion matrix on the test dataset.

Table 4
Comparison of baselines with the Wav2Vec 2.0 results. The entries in this table are sorted by accuracy in descending order. The best entry for each metric is bolded.

Model	Accuracy	UAR	Sensitivity	PPV	NPV
Wav2Vec 2.0	**73.31%**	**0.736**	**0.732**	0.574	**0.848**
MERT-v1-95M	66.21%	0.669	0.688	0.491	0.809
AST-base384	63.76%	0.619	0.566	0.459	0.76
HuBERT-base-ls960	60.35%	0.599	0.536	0.575	0.625
EfficientNetB3	59.80%	0.592	0.52	0.569	0.619
EfficientNetB0	59.90%	0.59	0.475	0.578	0.612
ResNet18	59.50%	0.588	0.501	0.568	0.614
ResNet50	58.07%	0.575	0.503	0.548	0.604
Naive Bayes	58.10%	0.555	0.235	0.617	0.573
SVM	57.20%	0.54	0.145	0.656	0.562
Random Forest	55.60%	0.522	0.096	**0.607**	0.552

Limitations include dataset size and computational demands. The controlled recording environment and limited speaker diversity may reduce real-world robustness. Additionally, Wav2Vec 2.0's size impacts real-time feasibility and raises overfitting concerns on small datasets.

7. Conclusion

Current DUI prevention methods are costly, intrusive, or require external devices like breathalyzers. This study explored Wav2Vec 2.0 for intoxication detection from audio, leveraging self-supervised learning and pre-training on large speech datasets. The model achieved a UAR of 73.3%, surpassing state-of-the-art methods, demonstrating its potential for DUI detection to prevent alcohol-related incidents. Challenges remain, including class imbalance, dataset limitations, and computational complexity. Future work should focus on expanding dataset diversity, addressing imbalance, and exploring smaller, efficient models. This research contributes to safer communities by advancing intoxication detection technology.

CRediT authorship contribution statement

Abigail Albuquerque: Formal analysis, Methodology, Software, Writing – original draft. **Samuel Chibuoyim Uche:** Methodology, Writing – review & editing. **Emmanuel Agu:** Conceptualization, Methodology, Resources, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

References

Baevski, Alexei, Zhou, Yuhao, Mohamed, Abdelrahman, & Auli, Michael (2020). Wav2vec 2.0: A framework for self-supervised learning of speech representations. *vol. 33*, In *NeurIPS* (pp. 12449–12460).

Baumeister, B., Heinrich, C., & Schiel, F. (2012). The influence of alcoholic intoxication on the fundamental frequency of female and male speakers. *J. Acoust. Soc. America*, *132*(1), 442–451.

Berninger, Kim, Hoppe, Jannis, & Milde, Benjamin (2016). Classification of speaker intoxication using a bidirectional recurrent neural network. In *Proc TSD* (pp. 435–442). Springer.

Bone, Daniel, et al. (2011). Intoxicated speech detection by fusion of speaker normalized hierarchical features and GMM supervectors. In *Annual conf. int'l speech comm. assoc*.

Bonela, A. A., He, Z., Nibali, A., Norman, T., Miller, P. G., & Kuntsche, E. (2023). Audio-based deep learning algorithm to identify alcohol inebriation (ADLAIA). *Alcohol*, *109*, 49–54, Epub 2022 Dec 28.

Breiman, Leo (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Burkhardt, Felix, et al. (2005). A database of german emotional speech. *5*, (pp. 1517–1520).

Gong, Yuan, Chung, Yu-An, & Glass, James (2021). Ast: Audio spectrogram transformer. arXiv preprint arXiv:2104.01778.

He, Kaiming, Zhang, Xiangyu, Ren, Shaoqing, & Sun, Jian (2015). Deep residual learning for image recognition.

Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J., & Scholkopf, B. (1998). Support vector machines. *IEEE Intell. Sys. Apps.*, *13*(4), 18–28. http://dx.doi.org/10.1109/5254.708428.

Hingson, Ralph (1996). Prevention of drinking and driving. *Alcohol Health Res World*, *20*(4), 219–226.

Hollien, Harry, DeJong, Gea, Martin, Cynthia A., Schwartz, Ronald, & Liljegren, Karin (2001). Effects of ethanol intoxication on speech suprasegmentals. *J. Acoust. Soc. America*, *110*(6), 3198–3206.

Hönig, Florian, Batliner, Anton, & Nöth, Elmar (2011). Does it groove or does it stumble: Automatic classification of alcoholic intoxication using prosodic features.

Hsu, Wei-Ning, et al. (2021). HuBERT: Self-supervised speech representation learning by masked prediction of hidden units.

Jacob, Asha, & Wang, Ping (2020). Alcohol intoxication and cognition: Implications on mechanisms and therapeutic strategies. *Frontiers in Neuroscience*, *14*, http://dx.doi.org/10.3389/fnins.2020.00102, URL https://www.frontiersin.org/journals/neuroscience/articles/10.3389/fnins.2020.00102.

Jovanovic, J., Jovanovic, M., Vukovic, N., & Jevremovic, J. (2000). Vozacka sposobnost alkoholisanih vozaca motornih vozila. *Facta Universitatis—Ser Med Biol*, *7*(1), 81–85.

Li, Yizhi, et al. (2024). MERT: Acoustic music understanding model with large-scale self-supervised training.

Lin, Hsin-Ann, et al. (2022). Evaluating the effect of drunk driving on fatal injuries among vulnerable road users in Taiwan: a population-based study. *BMC Public Health*, *22*(1), 2059. http://dx.doi.org/10.1186/s12889-022-14402-3.

National Highway Traffic Safety Administration (2023). Drunk driving. https://www.nhtsa.gov/risky-driving/drunk-driving#:~:text=If%20you%20drive%20while%20impaired (Accessed 16 April 2024).

Ng, Andrew, & Jordan, Michael (2001). On discriminative vs. Generative classifiers: A comparison of logistic regression and naive Bayes. *vol. 14*, In *NeurIPS*. MIT Press.

Peng, Linkai, Fu, Kaiqi, Lin, Binghuai, Ke, Dengfeng, & Zhan, Jinsong (2021). A study on fine-tuning wav2vec 2.0 model for the task of mispronunciation detection and diagnosis. In *Interspeech 2021* (pp. 4448–4452).

Pisoni, David B., & Martin, Christopher S. (1989). Effects of alcohol on the acoustic-phonetic properties of speech: Perceptual and acoustic analyses. *Alcoholism, Clinical and Experimental Research*, *13*(4), 577–587.

Schiel, Florian, Heinrich, Christian, Barfüßer, Sabine, & Gilg, Thomas (2008). ALC: Alcohol language corpus. In *Proc. int'l conf. lang. res. and eval.*. Marrakech, Morocco: ELRA.

Tan, Mingxing, & Le, Quoc V. (2020). EfficientNet: Rethinking model scaling for convolutional neural networks.

Taylor, B., et al. (2010). The more you drink, the harder you fall: A systematic review and meta-analysis of how acute alcohol consumption and injury or collision risk increase together. *Drug and Alcohol Dependence*, *110*(1), 108–116. http://dx.doi.org/10.1016/j.drugalcdep.2010.02.011, URL https://www.sciencedirect.com/science/article/pii/S0376871610000712.

Zhang, Xiaoming, et al. (2024). Improving speech depression detection using transfer learning with wav2vec 2.0 in low-resource environments. *Scientific Reports*, *14*, 9543.