SPEECH TASKS RELEVANT TO SLEEPINESS DETERMINED WITH DEEP TRANSFER LEARNING

Bang Tran[1]*, Youxiang Zhu*[1]*, Xiaohui Liang*[1]*, James W. Schwoebel*[2]*, Lindsay A. Warrenburg*[2]

[1]University of Massachusetts Boston
[2]Sonde Health Inc.

ABSTRACT

Excessive sleepiness in attention-critical contexts can lead to adverse events, such as car crashes. Detecting and monitoring sleepiness can help prevent these adverse events from happening. In this paper, we use the Voiceome dataset to extract speech from 1,828 participants to develop a deep transfer learning model using Hidden-Unit BERT (HuBERT) speech representations to detect sleepiness from individuals. Speech is an under-utilized source of data in sleep detection, but as speech collection is easy, cost-effective, and non-invasive, it provides a promising resource for sleepiness detection. Two complementary techniques were conducted in order to seek converging evidence regarding the importance of individual speech tasks. Our first technique, *masking*, evaluated task importance by combining all speech tasks, masking selected responses in the speech, and observing systematic changes in model accuracy. Our second technique, *separate training*, compared the accuracy of multiple models, each of which used the same architecture, but was trained on a different subset of speech tasks. Our evaluation shows that the best-performing model utilizes the *memory recall* task and *categorical naming* task from the Boston Naming Test, which achieved an accuracy of 80.07% (F1-score of 0.85) and 81.13% (F1-score of 0.89), respectively.

Index Terms— Sleepiness detection, acoustic features, transfer learning, deep learning

1. INTRODUCTION

Sleepiness results in human cognitive and behavioral changes that may cause a variety of negative outcomes, including automobile crashes, poor work performance, accidents at work, and other long-term physical and mental health consequences [1, 2]. As one example, the U.S. National Highway Traffic Safety Administration reported that in 2019 alone, 697 deaths were due to sleepiness-related automobile crashes [3]. The following year, a poll conducted by the National Sleep Foundation showed that nearly half of Americans feel sleepy on an average of three days a week [4]. Given these statistics, it is clear that excessive sleepiness negatively impacts safety and threatens our working and living environment.

By detecting sleepiness, it is possible to alert people to unsafe conditions and allow them to evaluate their own capacity to engage in dangerous behavior, such as driving. Analyzing speech for signs consistent with sleepiness is an ideal solution, as speech collection is non-invasive, low-cost, scalable, and can be performed quickly and easily [5]. One practical barrier with speech-based sleepiness detection has been the limited availability of human speech datasets. The ComParE 2019 challenge, for example, utilized speech recordings from 915 German speakers with two types of speech elicitations (reading and spontaneous speech). While this challenge resulted in significant contributions in sleepiness detection, the models and results were limited due to the low number of speech tasks and participants [6].

The current study used speech samples from the Voiceome dataset, which includes speech data from 6,650 participants. In the Voiceome dataset, each participant responded to 12 types of speech tasks (e.g., picture description, category naming, memory recall) for a total of 48 speech utterances. The participants also answered questions about their mental and physical health, including sleepiness, depression, anxiety, and smoking status.

The primary goal of the current paper was to infer a person's self-reported sleepiness from their speech. Secondary goals included (1) evaluating the importance of each of the 12 speech tasks in the detection process using a *masking* technique and (2) confirming and evaluating the potential of using a single speech task for sleepiness detection using a *separate training* technique.

2. RELATED WORKS

Detecting sleepiness has been addressed in past research by investigating acoustic speech factors. Previous methods have used a large number of general-purpose low-level descriptors (LLDs) such as short-term spectrum, short-term energy, and other voice-related features. Schuller et al. summarized these previous methods from the Interspeech 2011 Speaker State Challenge on sleepiness estimation [7]. Histogram representations of clustered LLDs, known as bag-of-audio-words, and

This research is funded by the US National Institutes of Health National Institute on Aging, under grant No. R01AG067416.

melspectrogram feature representations were studied in the Interspeech 2019 ComParE challenge [5]. For example, Yeh et al. presented a system that uses eGeMAPS features as the input of a Bidirectional Long Short-Term Memory network with attention to estimate sleepiness levels from speech. Gosztolya et al. created utterance-level Fisher vectors by training a Gaussian Mixture Model on frame-level MFCC features and used the vectors for sleepiness classification with Support Vector Machine [8]. To address acoustic-phonetic changes in sleepy speech, Fritsch et al. investigated the differences in speech production from a phonetic perspective by inheriting the knowledge of a pre-trained Convolutional Neural Network model to extract articulatory features from the speech data [9]. Egas-López et al. addressed the same problem by adopting a pre-trained x-vector model to estimate sleepiness level [10].

In this paper, we employ the Hidden-Unit BERT [11] speech representation technique to extract acoustic-phonetic and linguistic features from speech data for sleepiness classification.

3. MATERIALS

3.1. Voiceome protocol and dataset

The Voiceome protocol is a high-fidelity, longitudinal, and scalable protocol that can be used in digital settings to advance speech and language biomarkers research [12]. The corresponding Voiceome dataset consists of responses from 6,650 participants who completed the Voiceome protocol. All participants were residents of the United States, were 18 years of age or older, and indicated that they felt comfortable reading and writing in English. Additionally, all participants were required to have access to a device with a microphone and internet connection. The participants were broadly representative of the U.S. population with respect to age, gender, race and ethnicity, and general health behaviors [12]. More details can be found on the Voiceome GitHub[1].

The Voiceome protocol consists of twelve speech tasks, which result in a total of 48 speech utterances. The tasks are as follows:

1. *Microphone test*: read the sentence, "The quick brown fox jumps over lazy dog" (10 sec).

2. *Free speech*: talk about a recent happy memory based on experiences from the past month (60 sec).

3. *Picture description*: look at an image and describe everything they see going on in the picture (60 sec).

4. *Category naming*: given a category, such as "animals," speak as many words in that category as they can (60 sec).

5. *Phonemic fluency*: given a letter, such as "F," speak as many words that start with that letter as they can (60 sec).

6. *Phonetically-balanced paragraph reading*: read the Caterpillar passage [13].

7. *Sustained phonation*: make the vowel sound "/*a*/" (as in hallelujah) for as long as they can (30 sec).

8. *Diadochokinetic task*: repeat *puh-puh-puh* as quickly and accurately as they can (10 sec).

9. *Diadochokinetic task*: repeat *puh-tuh-kuh* as quickly and accurately as they can (10 sec).

10. *Confrontational naming*: look at an image and speak the name of the image within 10 seconds; 25 images total.

11. *Non-word pronunciation*: pronounce the "non-word" (e.g., plive, fwov) shown on the screen within 10 seconds; 10 non-words total.

12. *Memory recall*: listen to a short audio clip (15 sec) and repeat the sentence that they just heard.

3.2. Sleepiness operationalization

Voiceome participants completed the Stanford Sleepiness Scale (SSS), which is a single question that asks people to self-report their current state. The SSS is a Likert scale that ranges from *1–Feeling active, vital, alert, or wide awake* through *7–No longer fighting sleep, sleep onset soon; having dream-like thoughts*. The middle of the scale is biased towards sleepiness, with an indicator of *4–Somewhat foggy, let down*. For the purposes of the current study, we operationalized sleepiness as a binary category, where sleepiness ratings of 1-3 indicate non-sleepy states and ratings of 4-7 represent sleepy states. Note that only half of the Voiceome participants included health annotations in the Voiceome dataset; of those participants, 1,828 participants completed every speech task.

3.3. Speech pre-processing

The 1,828 participants submitted a total of 186.2 hours of speech responses. Voice samples went through similar preprocessing steps, including being converted to mono recordings at a sampling rate of 16,000 Hz. The Voiceome GitHub page provides all code used for audio pre-processing, feature extraction, and automatic transcription.

4. SYSTEM DESIGN

Deep transfer learning has been proven effective across a wide range of domains, outperforming traditional machine learning approaches. We used the HuBERT model recently released by Facebook to featurize speech files to create a fixed-length embedding (1024 dimensions per audio file). We also separately benchmarked this deep learning model across a TPOT-trained classifier [14] with the OpenSMILE GeMAPSv01a [15] aggregate feature embedding (62 dimensions per audio file). These approaches are discussed in the sections that follow.

[1] https://github.com/jim-schwoebel/voiceome

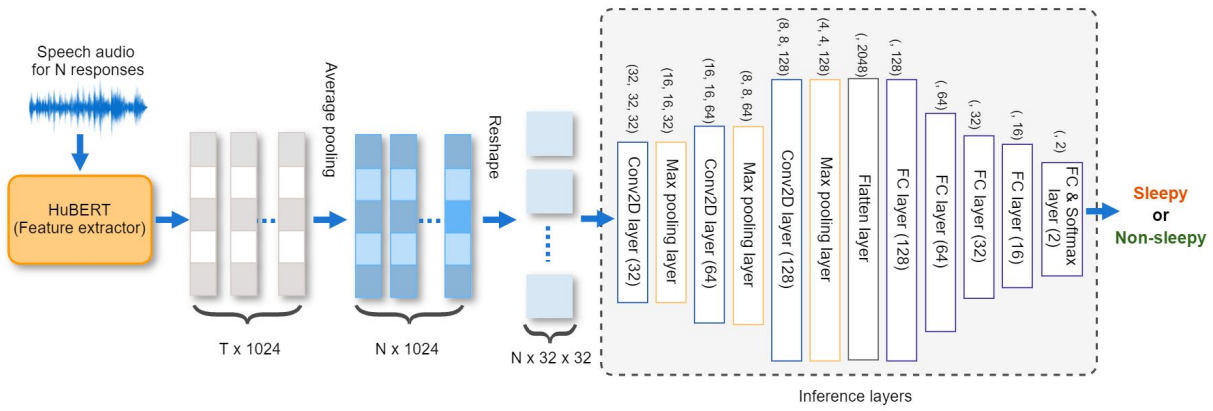

Fig. 1. HuBERT-Sleep deep transfer learning model.

4.1. HuBERT-Sleep Architecture/Baseline model

The HuBERT-Sleep baseline model was developed by incorporating the HuBERT pre-trained model as a fixed-length feature extractor (1024 dimensions), with added inference layers for the downstream task as shown in Figure 1. A fixed feature extractor was selected to minimize overfitting because the dataset of the downstream task was relatively small [16]. The HuBERT embedding was selected because it is the state-of-the-art approach for speech representations [17, 18, 19].

For each session, a participant has N speech utterances, with a total possibility of 48 responses. The N responses of a session were concatenated into a single speech input. If the speech input has T frames, it would be denoted by $X = [x_1, \ldots, x_T]$. The fixed-length HuBERT feature extractor outputs 1024 hidden units for each frame x_i. The hidden units for all frames of X were denoted as $Z = [z_1, \ldots, z_T]$, where z_i is a 1024-dimension vector. An average pooling layer was applied to the speech frames from the same response in order to reduce the dimensionality of Z from $T \times 1024$ to $N \times 1024$. The N vectors of dimension 1024 were mapped onto N matrices of dimensions 32×32, using convolutional 2D layers for faster convergence. The inference layers included 3 convolutional 2D layers interweaving with 3 max-pooling layers. The numbers of neurons in the 3 convolutional layers are (32, 64, 128), respectively. The kernel size is 3x3 and the stride is 1. The numbers of neurons in the 4 fully-connected layers were (128, 64, 32, 16), respectively. Finally, a softmax layer was added to produce the inferences for the two dependent variable classes, non-sleepy and sleepy.

The baseline HuBERT-Sleep model was trained on all speech tasks with an 80-20 split for training and testing.

4.2. HuBERT-Sleep Masking Experiment

In addition to the HuBERT-Sleep baseline model, we used two additional model training techniques. The first training technique was a masking paradigm where certain sets of speech elicitations were eliminated from the training set embeddings. For example, if the Microphone-test task was masked, then the embedding size would be 48×1024, but one of the 1024 layers would be zeroed out (e.g. $[0, 0, \ldots, 0]$). In other words, all *non-microphone-test* task data were used in model training, but the *microphone-test* data were masked. We used this masking technique to train and test models across 12 tasks, as described in Section 4.1. By comparing the baseline model–trained with all speech tasks–with a model with a masked speech task, we can evaluate the importance of the masked task on sleepiness inference.

4.3. HuBERT-Sleep Separate Training Experiment

The second model training technique was to use a small subset of speech task(s) for model training. For example, only the *microphone-test* task was used for model training. For speech tasks that included more than one elicitation, such as the two elicitations in the memory recall task, the embedding size would be $N = 2 \times 1024$. Once again, the separate training models were compared with the baseline model in order to evaluate the importance of speech tasks on sleepiness.

4.4. The OpenSMILE GeMAPS01a/TPOT Experiment

An independent third-party validation was conducted as a way to benchmark the models reported in sections 4.1 - 4.3. Instead of using HuBERT embeddings, aggregate acoustic features were used from the OpenSMILE GeMAPSv01a embeddings (62 dimensions per audio file). An automated machine learning tool, TPOT-light, was used to generate the classification models. The TPOT configuration iterates over a range of classification techniques (naive Bayes, decision trees, k-nearest-neighbors, logistic regression). The performance of the best one out of the four models via TPOT for each task is reported in Table 1 as a baseline.

5. EVALUATION

In this section, the implementation details are presented and the results on sleepiness detection are reported.

Task	Accuracy % (Mean F1-scores)		
	TPOT w/ GeMAPs	**Separate w/ HuBERT**	**Masking w/ HuBERT**
T1. Microphone test	45.10 (.440)	***69.70 (.805)***	81.13 (.893)
T2. Free speech	54.25 (.578)	77.24 (.867)	80.31 (.888)
T3. Picture description	49.67 (.516)	***70.66 (.799)***	80.69 (.891)
T4. Category naming	42.48 (.333)	75.00 (.852)	81.35 (.895)
T5. Phonemic fluency	48.37 (.484)	78.34 (.876)	80.02 (.893)
T6. Paragraph reading	45.39 (.443)	73.14 (.840)	81.62 (.897)
T7. Sustained phonation	50.98 (.483)	77.68 (.870)	81.56 (.897)
T8. Diadochokinetic (puh-puh-puh)	58.17 (.579)	***67.61 (.822)***	82.00 (.900)
T9. Diadochokinetic (puh-tuh-kuh)	50.33 (.531)	***69.83 (.796)***	80.36 (.889)
T10. Confrontational naming	44.44-60.13 (.317-.643)	**81.13 (.894)**	**78.36 (.889)**
T11. Non-word pronunciation	45.10-58.17 (.339-.623)	78.66 (.877)	80.85 (.893)
T12. Memory recall	46.41-50.98 (.312-.529)	**80.07 (.853)**	**72.76 (.886)**
Baseline (all tasks)	**54.90 (.566)**	**81.29 (.895)**	

Table 1. Accuracy and mean of F1-scores

5.1. Implementation details

The 1,828 sessions from the Voiceome dataset were split into 80% training (1,462 sessions) and 20% testing (366 sessions) sets. The training and testing were conducted in 5 non-overlapping rounds by a 40GB of memory Graphics Processing Unit (NVIDIA TESLA A100). The average result of the 5 rounds is reported. In order to adapt our training process to limited memory resources, the learning rate was set to 10^{-4}, the maximum epoch was set to 200, and the batch size was set to 32.

5.2. HuBERT-Sleep model

The baseline HuBERT-Sleep model was trained and tested on all speech data. The accuracy of this model was 81.29%, meaning that a person's sleepiness category (sleepy or non-sleepy) was determined correctly 81% of the time.

The TPOT classification model used MaxAbs scaling for pre-processing the GeMAPS features and a KNN model architecture. The accuracy of this model was 54.90%, suggesting that the HuBERT-based neural-net architecture was better for detecting sleepiness than other classification methods.

5.3. Speech task evaluation

Both masking and separate training techniques were used to evaluate the importance of a single speech task. Table 1 shows the average accuracy of the test data over 5 rounds; the masking technique and separate training techniques are reported separately. Recall that the masking technique consisted of using all speech data except for the masked task. In Table 1, the masking accuracy score for the *microphone test* indicates the accuracy of the model using all speech data except the *microphone test*. A lower accuracy score for a task indicates that the task was more important for sleepiness detection, as the removal of that task's data hinders model performance.

On the other hand, the separate training technique uses only the indicated task. The separate training accuracy for the *microphone test* corresponds to the accuracy of the model using only the data of *microphone test*. For the separate training task, a higher accuracy score is consistent with the idea that the task was more important for sleepiness classification, as that task was the only predictor of sleepiness.

In addition to the masking and separate training methods, we ran non-deep learning tasks with the aggregate GeMAPS features. The first main result is that the HuBERT architecture–both with the masking and separate training techniques–performed with higher accuracy and mean F1-scores than did the non-deep learning tasks. The accuracy scores of non-deep learning tasks are ranged from 42.48% to 60.13% which are low and hardly used to identify the relevant speech tasks.

The accuracy and mean of F1-scores in Table 1 are consistent with the idea that the *memory recall* and *confrontational naming* tasks were the most important for predicting sleepiness, as they resulted in the lowest masking scores and highest separate training scores. The results are also consistent with the idea that diadochokinetic tasks (puh-tuh-kuh; puh-puh-puh), picture descriptions, and the sentence-reading based microphone test are not good predictors of sleepiness, as they produce the lowest separate training scores and high masking scores.

6. CONCLUSION

In this paper, we developed a deep transfer learning model to classify sleepy versus non-sleepy speech samples in the Voiceome dataset. We furthermore utilized two techniques in order to identify the most relevant speech tasks in sleepiness detection. The results of these two training approaches converge: both training techniques suggest that memory recall and categorical naming tasks (from the Boston Naming Test) are the most important tasks for detecting sleepiness from speech data. Future work may wish to employ these two speech tasks in production systems where sleepiness detection can lead to safer and healthier environments.

7. REFERENCES

[1] Mathieu Nollet, William Wisden, and Nicholas P Franks, "Sleep deprivation and stress: a reciprocal relationship," *Interface focus*, vol. 10, no. 3, pp. 20190092, 2020.

[2] Marco Hafner, Martin Stepanek, Jirka Taylor, Wendy M. Troxel, and Christian Van Stolk, *Why sleep matters the economic costs of insufficient sleep: A cross-country comparative analysis*, RAND Corporation, Santa Monica, CA, 2016.

[3] National Highway Traffic Safety Administration, "Traffic safety fact - research note," December 2020, [Online; accessed 05-September-2021].

[4] National Sleep Foundation, "2020 sleepiness and low levels of action," 2020, [Online; accessed 23-August-2021].

[5] Björn Schuller, Anton Batliner, Christian Bergler, Florian B Pokorny, Jarek Krajewski, Margaret Cychosz, Ralf Vollmann, Sonja-Dana Roelen, Sebastian Schnieder, Elika Bergelson, et al., "The interspeech 2019 computational paralinguistics challenge: Styrian dialects, continuous sleepiness, baby sounds &orca activity," 2019.

[6] Michelle M Mielke, Prashanthi Vemuri, and Walter A Rocca, "Clinical epidemiology of alzheimer's disease: assessing sex and gender differences," *Clinical epidemiology*, vol. 6, pp. 37, 2014.

[7] Björn Schuller, Stefan Steidl, Anton Batliner, Florian Schiel, Jarek Krajewski, Felix Weninger, and Florian Eyben, "Medium-term speaker states—a review on intoxication, sleepiness and the first challenge," *Computer Speech & Language*, vol. 28, no. 2, pp. 346–374, 2014.

[8] Gábor Gosztolya, "Using fisher vector and bag-of-audio-words representations to identify styrian dialects, sleepiness, baby & orca sounds," 2019.

[9] Julian Fritsch, S Pavankumar Dubagunta, and Mathew Magimai Doss, "Estimating the degree of sleepiness by integrating articulatory feature knowledge in raw waveform based cnns," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6534–6538.

[10] José Vicente Egas-López and Gábor Gosztolya, "Deep neural network embeddings for the estimation of the degree of sleepiness," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 7288–7292.

[11] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *arXiv preprint arXiv:2106.07447*, 2021.

[12] James W Schwoebel, Joel Schwartz, Lindsay Warrenburg, Roland Brown, Ashi Awasthi, Austin New, Monroe Butler, Mark Moss, and Eleftheria K Pissadaki, "A longitudinal normative dataset and protocol for speech and language biomarker research," *medRxiv*, 2021.

[13] Rupal Patel, Kathryn Connaghan, Diana Franco, Erika Edsall, Dory Forgit, Laura Olsen, Lianna Ramage, Emily Tyler, and Scott Russell, ""the caterpillar": A novel reading passage for assessment of motor speech disorders," 2013.

[14] Randal S Olson and Jason H Moore, "Tpot: A tree-based pipeline optimization tool for automating machine learning," in *Workshop on automatic machine learning*. PMLR, 2016, pp. 66–74.

[15] Florian Eyben, Klaus R Scherer, Björn W Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y Devillers, Julien Epps, Petri Laukka, Shrikanth S Narayanan, et al., "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE transactions on affective computing*, vol. 7, no. 2, pp. 190–202, 2015.

[16] Youxiang Zhu, Xiaohui Liang, John A Batsis, and Robert M Roth, "Exploring deep transfer learning techniques for alzheimer's dementia detection," *Frontiers in computer science*, vol. 3, 2021.

[17] Shu-wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y Lin, Andy T Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, et al., "Superb: Speech processing universal performance benchmark," *arXiv preprint arXiv:2105.01051*, 2021.

[18] Kushal Lakhotia, Evgeny Kharitonov, Wei-Ning Hsu, Yossi Adi, Adam Polyak, Benjamin Bolte, Tu-Anh Nguyen, Jade Copet, Alexei Baevski, Adelrahman Mohamed, et al., "Generative spoken language modeling from raw audio," *arXiv preprint arXiv:2102.01192*, 2021.

[19] Adam Polyak, Yossi Adi, Jade Copet, Eugene Kharitonov, Kushal Lakhotia, Wei-Ning Hsu, Abdelrahman Mohamed, and Emmanuel Dupoux, "Speech resynthesis from discrete disentangled self-supervised representations," *arXiv preprint arXiv:2104.00355*, 2021.