ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10096105

A KNOWLEDGE-DRIVEN VOWEL-BASED APPROACH OF DEPRESSION CLASSIFICATION FROM SPEECH USING DATA AUGMENTATION

Kexin Feng and Theodora Chaspari

Computer Science and Engineering
Texas A&M University
{kexin, chaspari}@tamu.edu

ABSTRACT

We propose a novel explainable machine learning (ML) model that identifies depression from speech, by modeling the temporal dependencies across utterances and utilizing the spectrotemporal information at the vowel level. Our method first models the variable-length utterances at the local-level into a fixed-size vowel-based embedding using a convolutional neural network with a spatial pyramid pooling layer (*vowel CNN*). Following that, the depression is classified at the global-level from a group of *vowel CNN* embeddings that serve as the input to another 1D CNN (*depression CNN*). Different data augmentation methods are designed for both the training of *vowel CNN* and *depression CNN*. We investigate the performance of the proposed system at various temporal granularities when modeling short, medium, and long analysis windows, corresponding to 10, 21, and 42 utterances, respectively. The proposed method reaches comparable performance with previous state-of-the-art approaches and depicts explainable properties with respect to the depression outcome. The findings from this work may benefit clinicians by providing additional intuitions during joint human-ML decision-making tasks.

Index Terms— Mental health, speech vowel, knowledge-driven, convolutional neural network, data augmentation

1. INTRODUCTION

Depression is a mental health (MH) condition with worldwide prevalence [1]. Depression diagnosis and treatment is challenging due the lack of access to MH care resources and social stigma [2]. Speech-based machine learning (ML) systems have shown promising results in identifying depression due to their ability to learn clinically-relevant acoustic patterns, such as monotonous pitch and reduced loudness [3]. In addition, these systems can potentially increase accessibility to MH care resources, since they can run locally on users' smartphone devices. Various ML models including support vector machines (SVM), convolutional neural network (CNN), and long short-term memory (LSTM) have been explored for depression estimation [4]. However, the majority of these methods are designed independently of MH clinicians. Previous research found that physiological and cognitive impairments associated with depression can influence motor control and consequently, affect the phonological loop [5], producing significant energy variations in speech vowels [6]. Patients with depression depict a reduced frequency range between vowel formants and different formant dynamics compared to healthy individuals [7]. Vowel-based features were found to outperform turn-level acoustic features for detecting depression [8]. Integrating such domain knowledge into the system could potentially enhance the explainability of speech-based ML models for depression detection.

In complex and highly subjective decision-making tasks, such as the ones pertaining to MH care, ML technologies can potentially help clinicians via augmenting their ability to make reliable decisions in a data-driven manner. An explainable ML model of depression estimation would allow clinicians to gain insights into the ML rationale and decision-making processes, and contribute toward better calibrating their trust to the model output [9]. Previously proposed conceptual frameworks for building human-centered explainable ML suggest that users may be able to develop a mental model of the algorithm based on a collection of "how explanations" that demonstrate how the model works based on multiple instances [10]. In addition, it is important to provide both global explanations that describe holistically how the model works, and local explanations that demonstrate the relationship between inputs and outputs [11].

Here, we design an explainable ML model for depression classification based on speech. We leverage knowledge from speech production indicating that depression can influence motor control and consequently the formant frequencies and spectrotemporal variations at the vowel-level [12]. We propose a vowel-dependent CNN (*vowel CNN*) with a spatial pyramid pooling (SPP) layer that learns the spectrotemporal information of short-term speech segments (i.e., 250ms) throughout the utterance. The depression is estimated from

This work is supported by the National Science Foundation (CAREER: Enabling Trustworthy Speech Technologies for Mental Health Care: From Speech Anonymization to Fair Human-centered Machine Intelligence, #2046118). The code is available at: https://github.com/HUBBS-Lab-TAMU/ICASSP-2023-Augmented-Knowledge-Driven-Speech-Based-Method-of-Depression-Detection.

a group of *vowel CNN* embeddings using a 1D CNN (*depression CNN*). The *vowel CNN* captures depression information at the local-level from parts of speech that are theoretically postulated to be most affected by the MH condition [12]. The SPP layer maps utterances of any size into a fix-size embedding that contributes to generating explanations at the utterance-level, which can provide a global view of the depression outcome. To further improve the performance and address challenges related to the small sample size, we use two novel data augmentation methods that are applied during the training of *vowel CNN* (i.e., addressing imbalance between vowel distributions) and *depression CNN* (i.e., alleviating imbalance between healthy and depression classes). Results indicate that the proposed system is comparable to or better than previous depression classification systems, and further provides explanations about its decision-making process to the user. Ablation studies demonstrate the effectiveness of considered augmentation methods.

2. PREVIOUS WORK

End-to-end ML models can effectively detect one's MH condition from speech. Ma *et al.* designed an end-to-end system (DepAudioNet) that uses a 1-dimensional CNN to encode audio features and a LSTM network to model the encoded audio embeddings [13]. Sardari *et al.* utilized a convolutional autoencoder on the raw speech signal to identify depression [14]. Romero *et al.* introduced an ensemble learning approach by training 50 CNN models with different initializations to address the challenge of local optimal [15]. Other recent methods, such as the SpeechFormer, use a hierarchical framework to modeling spectral variations within and across speech frames, phonemes, words, and utterances [16].

Data augmentation can contribute to effectively classifying depression with limited labelled data. As part of CNN-Augm, Lam *et al.* augmented the speech transcripts based on the topic of the clinical interview via manually identifying the most common topics, and added speech samples related to each topic to form an augmented dataset [17]. Ravi *et al.* used different frame-width and frame-shift parameters to obtain data samples at various time-frequency resolutions [18]. Other speech augmentation methods include feature perturbation [19], altering the raw speech signal [20, 21], or generating new data via a GAN-based structure [22].

The contributions of this work are as follows: (1) In contrast to the majority of deep learning models on depression estimation that are not explainable [13, 14, 16], we propose a local explanation of the ML decision via modeling speech patterns at the vowel-level; (2) We extend the limited prior work on explainable ML for depression classification [23] by further providing a global explanation of the decision at the utterance-level via introducing the SPP layer that can model utterances of variable length; and (3) We investigate an oversampling-based and perturbation-based augmentation methods to mitigate the imbalanced distributions of the different vowels and healthy/depression classes.

3. PROPOSED METHODOLOGY

Our system includes three modules: (1) Vowel segmentation module with data augmentation (Section 3.1), which evenly samples 250ms segments from English vowels (/a/, /e/, /i/, /o/, /u/, or not a vowel); (2) Vowel classification module (Section 3.2), that includes a *vowel CNN* trained based on the balanced vowel segments, and can take variable-length utterances as input due to the SPP layer; and (3) Depression classification module with data augmentation (Section 3.3), that trains the *depression CNN* using the *vowel CNN* embeddings.

3.1. Vowel segmentation module with data augmentation

Here, we prepare the training data (i.e., 250ms speech segments with vowel labels) for the *vowel CNN*, similar to [23]. The most common vowel (/a/) occurs over 10 times more than less frequent vowels (e.g., /u/) due to the phonation patterns in the English language. This can hamper the vowel classification module from effectively learning vowel-dependent patterns. Thus, we design a sampling-based data augmentation method by dynamically determining the overlap between segments. If the current segment (0-250ms) belongs to 'not a vowel', a regular overlap length corresponding to half of the analysis window ($250ms * 0.5 = 125ms$)is applied. However, if the current segment is labeled as the commonly occurring vowel /a/, the overlap length is reduced to 75ms ($250ms * 0.3 = 75ms$) so that more segments can be sampled around this vowel. The overlap ratio for each vowel is equal to the vowel frequency in the training data and remains unchanged during sampling. A demonstration of how it works for each vowel is provided in Table 1. A segment is finally assigned to a vowel based on whether the vowel is fully or partially included within this segment, similar to [23].

3.2. Vowel classification module

The purpose of this module is to train the *vowel CNN* that assigns a 250ms segment, as determined in Section 3.1, into a vowel label (i.e., /a/, /e/, /i/, /o/, /u/, not a vowel). Compared with the vanilla 2D CNN used in previous work [23], the SPP layer increases the model's flexibility by allowing different embedding shapes after the convolutional layers. We use the log-Mel spectrogram as the feature for every 250ms segment. The feature extraction process is completed using the Librosa library [24] and the other parameters include a 512-sample FFT window length, 128-sample hop length, and 128 Mel bands. This leads to a spectrogram patch with size $(128, 28)$ for every segment. The *vowel CNN* includes three convolutional blocks, each block consists of a convolutional, activation, batch normalization, and max-pooling layer. A SPP layer and two fully connected (FC) layers are further added after all the convolutional blocks. We show the detailed *vowel CNN* structure in Table 2. We use Pytorch [25] to implement this model and minimize the cross-entropy loss. A batch size 64 and an Adam optimizer with a learning rate of 0.001 and l2 regularization of 0.001 are also applied.

3.3. Depression classification module with augmentation

This module utilizes the fixed-size embeddings at the utterance level extracted from the *vowel CNN* and classifies de-

Table 1. Overlap between current and next 250ms segments in the vowel classification module per type of vowel.

Vowel in current segment	/a/	/e/	/i/	/o/	/u/	not a vowel
Overlap ratio (between current and next segment)	0.3	0.08	0.1	0.03	0.02	0.5
Start/end time of next segment (ms)	75-325	20-270	25-275	7.5-257.5	5-255	125-375

Table 2. The structure and hyper-parameters of the 2D CNN model that conducts vowel classification.

Layer	Conv block 1	Conv block 2	Conv block 3	SPP	Fully-connected (FC)	Output
Layer setting	conv kernel (3, 1) 64 filters, ReLU pooling kernel (2, 1)	conv kernel (3, 1) 64 filters, ReLU pooling kernel (2, 1)	conv kernel (3, 1) 64 filters, ReLU pooling kernel (2, 1)	3 levels	128 units ReLU	6 units
output dim	(64, 63, 28)	(64, 30, 28)	(64, 14, 28)	1344	128	6

pression for each speaker via soft voting. Because the *vowel CNN* can take an arbitrary-sized input, we extract a 128-dimensional vowel embedding by taking the output of the first FC layer for each utterance (Table 2). Based on this 128-dimensional embedding, we train a *depression CNN* model, which takes a group of embeddings as input, resulting from consecutive utterances, and outputs the depression prediction probability for this group of utterances. Taking decisions based on a set of utterances groups may contribute to a global explanation on how the ML model works. Before training the *depression CNN*, we generate an augmented training set using a novel data augmentation method. Inspired by "Keep-Augment" [26], our method augments the embeddings $\mathbf{X}$ of a given speaker to a revised embedding set $\mathbf{Z}$. We randomly select a subset $\mathbf{X_n}$ of embeddings from n consecutive utterances and obtain a 128-dimensional saliency measure of each utterance via the output of the activation function of the *vowel CNN*. The sum of the absolute value of the saliency measures is further calculated. We then substitute the embeddings of a random subset of these utterances with a constant number c, without substituting the utterance embeddings with the highest saliency measures, thus protecting utterances with rich spectrotemporal information. A detailed description of this procedure is in Algorithm 1. The parameters of the augmentation include $n = 10, 21, 42$ (number of utterances included as an input in the *depression CNN*), $pos = 8, 16, 32$ (number of samples added to the training set if the original sample has depression label), $neg = 4, 8, 16$ (number of samples added to the training set if the original sample has a non-depression label), $p = 1, 2, 6$ (number of utterances that are perturbed for a group of utterances), $r = 21$ (number of utterances with high saliency values that are protected), and $c = 0.001$ (constant value to replace embedding value).

The augmented dataset consists of $(n, 128)$-dimensional samples, where n is the number of utterances and 128 is the dimension of the vowel embeddings. The structure of the *depression CNN* is in Table 3. This model minimizes the cross-entropy loss, trained using a batch size of 16, and an Adam optimizer with 0.001 learning rate and 0.01 l2-regularization. For a test speaker, we segment the speech into windows of n utterances without overlap and perform soft voting.

Algorithm 1 Data augmentation in the augmented depression classification module

Input: *vowel CNN* f, utterance embeddings $\mathbf{X}$ and depression label y, window size n, # augmented positive (negative) samples pos (neg); # utterances perturbed p; # utterances protected r; constant c
Output: set Z of augmented data
$Z \leftarrow \{\}$ ▷ Initialize the augmented set
▷ Define number of augmented samples based on label
if $\mathbf{y} = 1$ **then**
  $aug\_num = pos$
else
  $aug\_num = neg$
end if
for i from 0 to aug_num **do**
  Randomly select subset $\mathbf{X_n}$ of n consecutive utterance embeddings from $\mathbf{X}$
  Get saliency measure for each utterance in $\mathbf{X_n}$ via f
  Create $\mathbf{X_r}$ with the r highest saliency utterances in $\mathbf{X_n}$
  Create $\mathbf{X}_p$ via randomly selecting p utterances from $\mathbf{X_n} \setminus \mathbf{X_r}$
  Create $\mathbf{X_p'}$ via substituting the $\mathbf{X}_p$ embeddings with c
  Create perturbed utterance embeddings $\mathbf{X_n'} = \mathbf{X_p'} \cup (\mathbf{X_n} \setminus \mathbf{X_p})$
  $Z \leftarrow Z \cup \{(\mathbf{X_n'}, y)\}$
end for
return Z

Table 3. The structure and hyper-parameters of final classification CNN in augmented depression classification module.

Layer	Conv block 1	Conv block 2	FC	Output
layer Setting	kernel 7 32 filters ReLU pool size 2	kernel size 7 32 filters ReLU pool size 2	64 units ReLU	2 units
Output dim	(32, 18)	(32, 6)	64	2

3.4. Evaluation

We measure the vowel classification performance of the *vowel CNN* (Section 3.2) and depression classification performance of the *depression CNN* (Section 3.3). We also explore the acoustic descriptors that correlate with the system's output.

Table 4. Vowel classification F1-score.

	/a/	/e/	/i/	/o/	/u/	Not vowel	Macro F1
random oversampling	0.64	0.4	0.5	0.34	0.33	0.66	0.48
augmented *vowel CNN*	0.66	0.41	0.51	0.36	0.35	0.69	0.50

For every n utterances that serve as the input of the *depression CNN*, we extract 7 interpretable features related to depression: (1) speech percentage; (2) mean fundamental frequency (mean F0); (3) standard deviation of fundamental frequency (std F0); (4) mean jitter; (5) mean shimmer; (6) mean loudness. These acoustic measures are obtained using the Librosa library and the openSMILE toolkit [24, 27]. We then quantify the association of each acoustic measure with the model prediction, using the Pearson's correlation.

4. EXPERIMENTS

4.1. Data description

We use the Wizard-of-Oz part of the Distress Analysis Interview Corpus (DAIC-WoZ) [28] that includes 142 clinical interviews split into a training set (30 depression, 77 healthy), a development set (12 depression, 23 healthy), and a test set (labels are unknown to the public per the AVEC 2016 challenge) [4]. A participant is considered to have depression if they have a PHQ-8 score equal to or larger than 10 [29]. We report the performance on the development set, consistent with previous research.

4.2. Baseline methods

We use two baseline methods that have considered data augmentation for depression estimation; the CNN-Augm [17], which requires content information, and the Ensemble-CNN *et al.* [15], which includes 50 CNN models rendering it challenging to run locally on wearable or portable devices and difficult to explain. We also report results from DepAudioNet [13] and SpeechFormer [16], as outlined in Section 2.

4.3. Results

We demonstrate the effect of data augmentation in the vowel segmentation module (Section 3.1). We report the F1-score of vowel classification on the development set using random oversampling and our proposed vowel augmentation method in Table 4. Our proposed method leads to significantly better performance identified via a McNemar test ($p < 0.001$).

We also report the depression F1-score and the macro F1-score obtained using the proposed augmented depression classification model (Section 3.3) (Table 5). The proposed method provides competitive performance with different utterance window sizes n, which adds extra flexibility in decision granularity (i.e., fine-grain decisions for $\mathbf{n} = 10$; and coarse-grain for $\mathbf{n} = 42$). We verify the effectiveness of the data augmentation method by removing the random replacement (i.e., set $p = 0$) and observe that the performance declined for all the experimental settings ($n = \{42, 21, 10\}$). Using $n = 42$ and $p = 0$, we obtain a macro-F1 at 0.64, a performance slightly better than DepAudioNet, which applies a sampling-based augmentation without replacement.

Table 5. Depression classification performance scores.

Method	Precision	Recall	F1	Macro F1
DepAudioNet (2016) [13]	0.35	1.0	0.52	0.61
CNN-Aug (2019) [17]	0.78	0.58	0.67	-
Ensemble-CNN (2020) [15]	0.55	0.79	0.65	0.73
SpeechFormer (2022) [16]	-	-	-	0.69
Fraug (2022) [18]	-	-	-	0.66
Proposed ($n = 10$)	0.55	0.50	0.52	0.64
Proposed ($n = 21$)	0.80	0.33	0.47	0.65
Proposed ($n = 42$)	0.70	0.58	0.64	0.73

Table 6. Pearson's correlation between acoustic measures and depression probability for different number n of input utterances.

n	speech percentage	mean F0	std F0	jitter	shimmer	loudness
10	**-0.119**	**0.396**	0.043	**0.13**	**-0.088**	**-0.324**
21	**-0.195**	**0.377**	0.088	0.15	**-0.138**	**-0.411**
42	-0.122	**0.236**	0.151	0.113	-0.156	**-0.213**

Bold font indicates significant correlation ($p < 0.05$).

Finally, we report the association between acoustic attributes and the probability of the depression classification system (Section 3.4) in Table 6. Jitter and speech percentage are significantly associated with the depression probability, which might indicate that the proposed explainable model captures perceptually intuitive information about depression. Mean F0 and loudness are the most significantly correlated acoustic features with the model output for $n = 10$. Prior work indicates that these features are not always correlated with depression [30, 31], but are highly indicative of a speaker's demography, including gender [32]. This may indicate that the model captures gender information, as a result of the difference in depression prevalence between female and male speakers (i.e., 16 out of the 45 female speakers and 14 out of 63 male speakers with depression). While prior work evidences gender bias in this data [33], a more detailed analysis is needed to fully understand this.

5. CONCLUSION

We explored an explainable ML that integrated vowel-based information at the local level and modeled speech utterances with variable lengths with an SPP layer at the global level. A dynamic sampling-based data augmentation method was designed to address the distribution difference among vowels. Another data augmentation method with random substitutions was applied to further mitigate the class imbalance between patients with depression and healthy participants. Our proposed system depicted better or comparable performance to multiple baselines, along with increased explainability. However, the robustness of our system toward other physical, physiological, and psychological disorders beyond depression remains unexplored. As part of our future work, we plan to evaluate the proposed approach via user studies with MH experts and consider confounding factors such as gender to better disentangle their interplay with MH information.

6. REFERENCES

[1] Kerri Smith and IBC De Torres, "A world of depression," *Nature*, vol. 515, no. 181, pp. 10–1038, 2014.

[2] Wayne J Katon, "Depression research in under-resourced populations: An academic–community partnership," *Journal of general internal medicine*, vol. 28, no. 10, pp. 1255–1257, 2013.

[3] Jeffrey F Cohn, Nicholas Cummins, Julien Epps, Roland Goecke, Jyoti Joshi, and Stefan Scherer, "Multimodal assessment of depression from behavioral signals," *The Handbook of Multimodal-Multisensor Interfaces: Signal Processing, Architectures, and Detection of Emotion and Cognition-Volume 2*, pp. 375–417, 2018.

[4] Michel Valstar, Jonathan Gratch, Björn Schuller, Fabien Ringeval, Denis Lalanne, Mercedes Torres Torres, Stefan Scherer, Giota Stratou, Roddy Cowie, and Maja Pantic, "AVEC 2016: Depression, mood, and emotion recognition workshop and challenge," in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, 2016, pp. 3–10.

[5] FC Murphy, BJ Sahakian, JS Rubinsztein, A Michael, RD Rogers, TW Robbins, and ES Paykel, "Emotional bias and inhibitory control processes in mania and depression," *Psychological medicine*, vol. 29, no. 6, pp. 1307–1321, 1999.

[6] Murray Alpert, Enrique R Pouget, and Raul R Silva, "Reflections of depression in acoustic measures of the patient's speech," *Journal of affective disorders*, vol. 66, no. 1, pp. 59–69, 2001.

[7] Stefan Scherer, Gale M Lucas, Jonathan Gratch, Albert Skip Rizzo, and Louis-Philippe Morency, "Self-reported symptoms of depression and PTSD are associated with reduced vowel space in screening interviews," *IEEE Transactions on Affective Computing*, vol. 7, no. 1, pp. 59–73, 2015.

[8] Bogdan Vlasenko, Hesam Sagha, Nicholas Cummins, and Björn Schuller, "Implementing gender-dependent vowel-level analysis for boosting speech-based depression recognition," 2017.

[9] Deepti Saraswat, Pronaya Bhattacharya, Ashwin Verma, Vivek Kumar Prasad, Sudeep Tanwar, Gulshan Sharma, Pitshou N Bokoro, and Ravi Sharma, "Explainable AI for Healthcare 5.0: Opportunities and challenges," *IEEE Access*, 2022.

[10] Tania Lombrozo, "Explanation and categorization: How "why?" informs "what?"," *Cognition*, vol. 110, no. 2, pp. 248–253, 2009.

[11] Sina Mohseni, Niloofar Zarei, and Eric D Ragan, "A multidisciplinary survey and framework for design and evaluation of explainable ai systems," *ACM Transactions on Interactive Intelligent Systems (TiiS)*, vol. 11, no. 3-4, pp. 1–45, 2021.

[12] Bt Atal and J Remde, "A new model of lpc excitation for producing natural-sounding speech at low bit rates," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 1982, vol. 7, pp. 614–617.

[13] Xingchen Ma, Hongyu Yang, Qiang Chen, Di Huang, and Yunhong Wang, "Depaudionet: An efficient deep model for audio based depression classification," in *Proceedings of the 6th international workshop on audio/visual emotion challenge*, 2016, pp. 35–42.

[14] Sara Sardari, Bahareh Nakisa, Mohammed Naim Rastgoo, and Peter Eklund, "Audio based depression detection using convolutional autoencoder," *Expert Systems with Applications*, vol. 189, pp. 116076, 2022.

[15] Adrián Vázquez-Romero and Ascensión Gallardo-Antolín, "Automatic detection of depression in speech using ensemble convolutional neural networks," *Entropy*, vol. 22, no. 6, pp. 688, 2020.

[16] Weidong Chen, Xiaofen Xing, Xiangmin Xu, Jianxin Pang, and Lan Du, "SpeechFormer: A hierarchical efficient framework incorporating the characteristics of speech," in *Interspeech*, 2022, pp. 346–350.

[17] Genevieve Lam, Huang Dongyan, and Weisi Lin, "Context-aware deep learning for multi-modal depression detection," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 3946–3950.

[18] Vijay Ravi, Jinhan Wang, Jonathan Flint, and Abeer Alwan, "Fraug: A frame rate based data augmentation method for depression detection from speech signals," in *2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6267–6271.

[19] Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le, "SpecAugment: A simple data augmentation method for automatic speech recognition," in *Interspeech*, 2019, pp. 2613–2617.

[20] Tom Ko, Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur, "Audio augmentation for speech recognition," in *Interspeech*, 2015, pp. 3586–3589.

[21] Xiaodong Cui, Vaibhava Goel, and Brian Kingsbury, "Data augmentation for deep neural network acoustic modeling," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 9, pp. 1469–1477, 2015.

[22] Zengrui Jin, Mengzhe Geng, Xurong Xie, Jianwei Yu, Shansong Liu, Xunying Liu, and Helen Meng, "Adversarial data augmentation for disordered speech recognition," in *Interspeech*, 2021, pp. 4803–4807.

[23] Kexin Feng and Theodora Chaspari, "Toward knowledge-driven speech-based models of depression: Leveraging spectrotemporal variations in speech vowels," in *2022 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*. IEEE, 2022, pp. 01–07.

[24] Brian McFee, Colin Raffel, Dawen Liang, Daniel P Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto, "Librosa: Audio and music signal analysis in python," in *Proceedings of the 14th Python in Science Conference*. Citeseer, 2015, vol. 8, pp. 18–25.

[25] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al., "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.

[26] Chengyue Gong, Dilin Wang, Meng Li, Vikas Chandra, and Qiang Liu, "Keepaugment: A simple information-preserving data augmentation approach," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 1055–1064.

[27] Florian Eyben, Martin Wöllmer, and Björn Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM International Conference on Multimedia*, 2010, pp. 1459–1462.

[28] Jonathan Gratch, Ron Artstein, Gale Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel Wood, Jill Boberg, David DeVault, Stacy Marsella, et al., "The distress analysis interview corpus of human and computer interviews," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 2014, pp. 3123–3128.

[29] Simon Gilbody, David Richards, Stephen Brealey, and Catherine Hewitt, "Screening for depression in medical settings with the patient health questionnaire (phq): a diagnostic meta-analysis," *Journal of General Internal Medicine*, vol. 22, no. 11, pp. 1596–1602, 2007.

[30] Takaya Taguchi, Hirokazu Tachikawa, Kiyotaka Nemoto, Masayuki Suzuki, Toru Nagano, Ryuki Tachibana, Masafumi Nishimura, and Tetsuaki Arai, "Major depressive disorder discrimination using vocal acoustic features," *Journal of affective disorders*, vol. 225, pp. 214–220, 2018.

[31] Larry Zhang, Joshua Driscol, Xiaotong Chen, and Reza Hosseini Ghomi, "Evaluating acoustic and linguistic features of detecting depression sub-challenge dataset," in *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, 2019, pp. 47–53.

[32] Chao-Yang Lee, Lauren Dutton, and Gayatri Ram, "The role of speaker gender identification in relative fundamental frequency height estimation from multispeaker, brief speech segments," *The Journal of the Acoustical Society of America*, vol. 128, no. 1, pp. 384–388, 2010.

[33] Andrew Bailey and Mark D Plumbley, "Gender bias in depression detection using audio features," in *2021 29th European Signal Processing Conference (EUSIPCO)*. IEEE, 2021, pp. 596–600.