2022 10th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)

2022 10th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW) | 978-1-6654-5490-2/22/$31.00 ©2022 IEEE | DOI: 10.1109/ACIIW57231.2022.10086012

Preserving Mental Health Information in Speech Anonymization

1st Vinesh Ravuri
Computer Science and Engineering
Texas A&M University
College Station, TX, USA
vineshr@tamu.edu

2nd Ricardo Gutierrez-Osuna
Computer Science and Engineering
Texas A&M University
College Station, TX, USA
rgutier@cse.tamu.edu

3rd Theodora Chaspari
Computer Science and Engineering
Texas A&M University
College Station, TX, USA
chaspari@tamu.edu

***Abstract*—Widespread speech technologies may change the dynamics of mental health (MH) diagnosis, monitoring, and care by tracking spectrotemporal patterns of speech in a personalized manner. Yet, these technologies have given rise to a large public debate about the ability of users to protect their privacy. This work examines a speech anonymization algorithm that preserves information that is diagnostic of the MH condition of the speaker and representative of the phonological content of speech, while suppressing characteristics related to the speaker identity. The proposed anonymization algorithm relies on an auto-encoder that conducts an identity mapping between the original input, represented as the superposition of Mel-spectrogram coefficients and posteriogram (PPG) vectors, and the reconstructed output. The auto-encoder is trained in an adversarial manner to minimize the loss corresponding to MH-related information and maximize the loss corresponding to the identity (ID) of a speaker. The proposed algorithm is evaluated on data from healthy participants and patients with depression from the Distress Analysis Interview Corpus Wizard of Oz (DAIC-WoZ) dataset. Results indicate that the speech signals synthesized by the proposed anonymization algorithm have higher word error rate (WER), as calculated between the original speech transcripts and transcripts from the anonymized speech signals obtained via an automatic speech recognizer (ASR), and are rated as being less comprehensible by MTurk listeners compared to the original signals. However, the intelligibility of the synthesized speech appears equivalent to baseline speech anonymization algorithms that solely suppress the speaker ID without considering MH-related information, or rely on a cascade of signal processing methods that perform consecutive speech transformations. The proposed algorithm can further effectively suppress information related to the speaker ID, since the corresponding anonymized signals achieve 15.9% reduction in speaker classification accuracy compared to the original speech signals. Finally, the anonymized speech signals yield similar performance to the original speech when they are used as an input for estimating degree of depression severity, and superior performance compared to conventional speech anonymization algorithms that do not consider the preservation of MH-related information as an optimization criterion.**

***Index Terms*—Mental health, depression, speech anonymization, adversarial learning, auto-encoder**

I. Introduction

In 2014, it was estimated that 9.8 million adults aged 18 or older had a serious mental illness (SMI), which includes schizophrenia, bipolar disorder, and major depression. Mental illnesses can impede carrying out daily activities, interacting with family, and fulfill other major functions [1]. Speech technology has great potential as an alternative method to track mental health (MH) and treatment outcomes. Speech signals can reflect changes in muscle tension in articulators, which are indicative of MH disorders, such as depression and post-traumatic stress disorder (PTSD). Speech-based AI algorithms have been heralded as promising solutions to track the severity of MH disorders, since they can learn patient-specific spectrotemporal patterns in speech that are indicative of MH degradation [2]. In addition, speech data is easy to collect using mobile devices, such as smartphones and wearable sensors, and can be captured continuously, unobtrusively, and discreetly over time. As such, MH interventions that rely on speech biomarkers have the potential to be widely adopted [3]. Collecting speech data in a natural setting further makes it easier to capture micro-level behaviors, which are difficult to record in traditional healthcare settings [4], thus allowing to observe factors and antecedents that contribute to MH in real-life. Ambulatory speech-based technologies can also help overcome the limitations of conventional self-report instruments, such as retrospective bias, subjectivity, and temporal sensitivity. Speech data collected from mobile devices has been used in tandem with machine learning (ML) methods to capture individuals' mood, stress, and other well-being markers [5], therefore enabling pervasive computing applications. These can further reduce barriers to accessing MH care, since mobile devices can transcend geographic boundaries, reaching many people otherwise unable to access MH care services [6].

However, speech recordings are a rich source of sensitive data that can be misused to predict personally identifiable information (PII) [7] such as age, gender, race, ethnicity, and speaker identity [8]. Thus, malicious users who gain unauthorized access to non-anonymized speech can potentially misuse PII from this signal, facilitating attacks on other systems and causing safety risks. In fact, recent reports highlight how the increasing use of ambulatory devices is creating new opportunities for cyber-attacks [9], such as identity theft by using another person's voice to fool voice authentication systems and impersonation attacks via utilizing speech synthesis or voice conversion to fake another person [10]. In light of

The authors would like to acknowledge the the National Science Foundation for funding this work (CAREER: Enabling Trustworthy Speech Technologies for Mental Health Care: From Speech Anonymization to Fair Human-centered Machine Intelligence, #2046118).

978-1-6654-5490-2/22/$31.00 ©2022 IEEE

these considerations, speech-based technologies for MH face increasing skepticism and mistrust by patients and consumers. Patients are hesitant to share their data, due to increased concerns over how the data will be used by technology and health care companies alike. According to a recent survey, only 37% of patients feel comfortable in sharing their voice data [11]. Another survey revealed that, while consumers are willing to share health data with their doctors, they are less willing to do so with research institutions, tech companies, and health companies [12]. In fact, public willingness to share health data has declined over time, even when it comes to sharing with physicians [12]. These trends highlight the need to improve speech anonymization methods, particularly in healthcare.

Speech anonymization methods typically rely on the extraction of three different types of information: (1) speaker information that encodes the characteristics of the voice of a person; (2) segmental information that captures the linguistic content of vowels and consonants in speech; and (3) prosodic information that refers to intonation, rhythm, and vocal stress. Prior work in speech anonymization typically seeks to alter information related to the speaker identity, while preserving the segmental and prosodic characteristics of speech [13], [14]. However, disentangling speaker information from segmental and prosodic information is not always straightforward, since the latter is heavily affected by the idiosyncratic characteristics of a person's voice. This task becomes even more challenging when speech anonymization is conducted under the constraint of preserving MH information, since the evidence of MH severity in speech is typically manifested in prosodic and supra-segmental characteristics, such as pitch, loudness, and rhythm [15].

In this paper, we aim to anonymize speech signals while preserving MH information, specifically for the task of estimating depression severity from speech. We propose a auto-encoder architecture that is trained to preserve acoustic characteristics related to depression severity and linguistic information, while reducing information that could be used to predict PII, and specifically the identity (ID) of a user. The encoder module takes in as input mel-spectrograms and phonetic posteriorgrams (PPGs) to generate a low-dimensional embedding. This embedding serves as an input to three additional modules: (1) a decoder module that is trained to reconstruct the original signal; (2) a depression-estimation module that is trained to estimate the degree of depression severity; and (3) a speaker-classification module that is trained to identify the speaker. To anonymize the speech signal, the speaker-classification module is trained in an adversarial manner (i.e., via gradient reversal), which forces the encoder to learn an embedding that removes speaker-dependent information, while preserving information that is diagnostic of MH. Via quantitative and qualitative experiments, we evaluate the anonymized speech signals in terms of their ability to capture depression severity information and speaker information, as well as in terms of speech intelligibility.

II. Prior Work

A. Speech-based detection of MH conditions

A large number of studies have investigated speech as a key behavioral marker for diagnosing and tracking the severity of MH conditions. According to a recent systematic review [16], the majority of studies focus on depression and schizophrenia, while some studies also consider PTSD and anxiety. In a seminal study, Mundt *et al.* tracked the progression of voice acoustic measures from 35 patients with depression for 6-weeks, while they were receiving pharmacotherapy and/or psychotherapy treatment [17]. Results indicated significant differences in terms of fundamental frequency (F0) variability, speech pause duration, and speaking rate between patients responding to treatment and patients not responding to treatment. Cummins also examined potential relationships between changes in speech cues and depression severity and found that the spectral feature space becomes more concentrated for patients with depression compared to their healthy counterparts [18]. Trevino *et al.* reported that this reduction in the spectral feature space appears to be stronger when extracted at the phoneme-level, rather than at the global utterance-based level [19]. Williamson *et al.* quantified the degree of coordination in vocal tract articulation via a set of correlation structure features computed across formant frequencies and of delta-functions of the Mel-cepstral coefficients [20]. Results indicate that these features computed over 4-minute speech excerpts provide promising performance in estimating the degree of depression severity.

Much of the recent work has focused on examining ML models that use speech samples from clinics or real-life environments for estimating MH conditions. Scherer *et al.* showed that the combination of voice quality characteristics capturing the capturing the tenseness of the speaker's voice with ML algorithms can yield 75% accuracy in automatically classifying between patients with depression and health participants, and and 69% accuracy in classifying between patients suffering from PTSD and health participants [21]. Khorram *et al.* further explored personalized ML models that rely on a small portion of labelled samples from a target speaker using rhythm statistics and i-vector based features [22]. The authors showed that the combination of personalized and general models (i.e., the latter modeling general speech patterns from all speakers) achieve superior results in estimating the degree of depression severity compared to each method separately. He & Cao showed examined features learned by CNN-based architectures from the speech spectrogram [23]. They found that these deep-learned features can estimate the degree of depression severity more accurately compared to hand-crafted spectral and energy-related features. However, the two types of features yields a reduction of approximately one unit of absolute error in estimating the degree of depression severity, as compared to the deep-learned features alone. A detailed review of ML algorithms that have been used to detect MH conditions can be found in [24].

B. Speech anonymization

Speech anonymization has become an increasingly important topic. A variety of approaches have been used, ranging from accent conversion [25], sanitization [26], and watermarking methods [27]. Initial speech anonymization approaches have employed voice conversion methods to modify the perceived attributes of speech. For instance, Patino *et al.* and Gupta *et al.* proposed to transform the spectral envelope of a speech signal by altering the position and radius of the poles of the linear prediction spectral envelope, obtained via the McAdams-based solution [28]. Adversarial learning has been also proposed for speech anonymization. Champion *et al.* designed an end-to-end deep encoder-decoder for the task of automatic speech recognition (ASR) [29]. The bottleneck features of the deep auto-encoder were learned so that they minimize the ASR loss and maximize the speaker classification loss. The deep auto-encoder was trained in a semi-adversarial manner, according to which the weights of the network were first optimized based on the ASR loss only and then learned based on the speaker classification loss. Following that, the weights of the encoder were frozen and the decoder was trained based on both the ASR and speaker classification loss. Auto-encoders have been also proposed for converting the identify of one speaker to that of another target speaker (i.e., voice conversion), while preserving the linguistic content of speech data. Yoo *et al.* proposed a variational auto-encoder that encodes speaker information as a one-hot-encoding vector [30]. The encoder is trained to extract a latent vector which corresponds to the linguistic information of the input speech, and the decoder is trained to reconstruct the input speech from the latent vector and the source speaker identity vector. Bahmaninezhad *et al.* also used an auto-encoder architecture to encode spectral information and excitation features of speech [31]. The auto-encoder mapped the bottleneck features to a target speaker that was characterized by the average of other speakers.

Recent approaches to speech anonymization focus on replacing the identity of the original speaker with that of another speaker via voice conversion. In that direction, as part of the VoicePrivacy 2020 Challenge [14], Fang *et al.* used an ASR system based on a deep neural network architecture to capture segmental information in the form of a phoneme posteriorgram (PPG) [13]. They further leveraged a pre-trained x-vector system to encode the speaker identity. Prosodic information related to pitch was captured by extracting the F0 from the original waveform. Anonymization was conducted by replacing the x-vector of the original speaker with another x-vector corresponding to a different set of arbitrary speakers. The anonymized speech waveform was generated by using the PPG and F0 of the original speech and the anonymized x-vector. Following that, Srivastava *et al.* showed that the quality of anonymization using the x-vector methodology is affected by the choice of the pseudo-speaker [32], [33]. The authors examined various design choices for choosing a pseudo-speaker, including different distance metrics between speakers and the region of x-vector space where the pseudo-speaker is picked. Results indicate that effective speech anonymization is accomplished when the candidate x-vectors are selected using the probabilistic linear discriminant analysis (PLDA) distance between the original speaker and the pseudo-speaker. In addition, robust privacy protection was achieved by randomly selecting a subset of pseudo-speakers from the cluster of x-vectors with most members (i.e., the most dense cluster). To effectively disentangle speaker information from language and prosody, Shamsabadi *et al.* integrated a differential privacy method to the existing x-vector system [34]. The authors employed differentially private feature extractors based on an auto-encoder and an ASR system. The F0 and PPG extractors were trained to retain the desired prosodic and segmental information, respectively, while adding noise via a Laplace noise layer. The pitch extractor consisted of a pitch estimator followed by an auto-encoder network with Laplace noise layer trained to reconstruct the global pitch dynamics using a custom loss function. Similarly, the PPG extractor utilized a deep ASR acoustic model, in which a Laplace noise layer was integrated. The Laplace noise layer provided a provably upper bound of the amount of residual speaker information embedded in the F0 and PPG features, thus resulting in a ϵ-differential privacy (ϵ-DP) guarantee.

While there has been a lot of work on speech anonymization algorithms that preserve segmental information, to the best of our knowledge, this is the first work that attempts to anonymize speech signals while also preserving information related to the speaker's MH condition. In addition to typically used evaluation methods that assess intelligibility and speaker identification of the anonymized speech, we also assess the extent to which the anonymized speech can be used to estimate the speaker's degree of depression severity.

III. Methodology

We discuss steps that were taken for data pre-processing (Section III-A) and feature extraction (Section III-B), as well as the architecture and the evaluation methods (Section III-C) for the proposed anonymization algorithm.

A. Data Description and Pre-processing

We used the Distress Analysis Interview Corpus Wizard of Oz (DAIC-WoZ) dataset [35] for our experiments. The dataset consists of audio interviews of 107 participants (63 male and 44 female), and each participant was classified as depressed or not based on their responses to questions on the Patient Health Questionnaire (PHQ-8) [36]. Thirty participants were classified as depressed, and the remaining as healthy [35]. Each audio recording was converted to a sample rate of $22,050Hz$. Each recording contained speech from the interviewer and the interviewee; however, for the purpose of this study, the interviewer's speech segments were omitted based on the start and end timestamps of the corresponding turns that were provided as part of the DAIC-WoZ dataset. Each recording from the interviewee was then split into individual utterances, that ranged from 1 second to 30 seconds long, resulting in a total of $11,993$ utterances.

B. Feature Extraction

The input of the speech anonymization model (Figure 1) comprises of Mel-spectrograms and phonetic posteriorgrams (PPGs). The Mel-spectrograms are used to preserve acoustic properties of speech, whereas the PPGs are used to extract the linguistic content in the form of the posterior probability for each of 40 phonemes occurring in a given analysis window. The Mel-spectrograms are extracted at an analysis window of 2048 with a hop length of 256, and 80 coefficients are used. The PPGs are extracted using an acoustic model based on deep neural networks (DNN), as described in Zhao *et al.* [37]. The Mel-spectrograms are stacked such that each input contains the previous ($n-1$), current (n), and next ($n+1$) analysis window, which will be denoted as $\mathbf{s_{n-1}}$, $\mathbf{s_n}$, and $\mathbf{s_{n+1}}$. Following that, the three stacked 80-dimensional Mel-spectrograms are concatenated with the 40-dimensional PPG that corresponds to the n^{th} analysis window, denoted as $\mathbf{p_n}$. The final vector $\mathbf{x_n} = [\mathbf{s_{n-1}}, \mathbf{s_n}, \mathbf{s_{n+1}}, \mathbf{p_n}]$ is used as a 280-dimensional input to the auto-encoder based anonymization system (Section III-C).

C. Anonymization System

Our proposed system uses an auto-encoder architecture that aims to reconstruct the speech at the output (Figure 1). Bottleneck features from the auto-encoder are used as inputs to two auxiliary classifiers, one that aims to predict depression, and another that aims to predict speaker identity.

The auto-encoder consists of a 5-layer feed forward neural network (FFN), which serves as the encoder, $f_{\mathbf{w}}$, followed another 5-layer FFN, which acts as the decoder, $f'_{\mathbf{w}}$. The auto-encoder implements an identity function, whose parameters $\mathbf{w}$ are learned from data. The auto-encoder takes in a 280-dimensional input that results from the Mel-spectrograms and the PPG, $\mathbf{x_n} = [\mathbf{s_{n-1}}, \mathbf{s_n}, \mathbf{s_{n+1}}, \mathbf{p_n}]$ (Section III-B). The encoder part of the auto-encoder has 5 fully-connected layers with 220, 160, 100, 70, and 40 nodes, respectively, which implement non-linear transformations of the input signal. Finally, this last 40-dimensional vector is further transformed into a 10-dimensional bottleneck layer $\mathbf{y_n} = f_{\mathbf{w}}(\mathbf{x_n})$ for each analysis window n, that serves as the latent space of the auto-encoder. Similarly, the decoder part of the auto-encoder comprises of 5 fully-connected layers with 40, 70, 100, 160, and 220 nodes, respectively. A total of $331,890$ parameters are being learned while training the auto-encoder. The auto-encoder loss L_{error} is used to preserve the acoustic and phonetic information of the original speech signal and is defined as the mean-square error between the actual input $\mathbf{x_n}$ and reconstructed signal, $\mathbf{x'_n} = f'_{\mathbf{w}}\left(f_{\mathbf{w}}\left(\mathbf{x_n}\right)\right)$, summed over all analysis windows ($n = 1, \ldots, N$):

$$\begin{aligned} L_{error}(\mathbf{w}) =& \alpha * \frac{1}{N} \sum_{n=0}^{N} \Big(c_1 \cdot \left(\mathbf{s_{n-1}} - \mathbf{s'_{n-1}}\right)^2 \\ &+ c_2 \cdot \left(\mathbf{s_n} - \mathbf{s'_n}\right)^2 + c_3 \cdot \left(\mathbf{s_{n+1}} - \mathbf{s'_{n+1}}\right)^2 \\ &+ c_4 \cdot (\mathbf{p_n} - \mathbf{p'_n})^2)\Big) \end{aligned} \tag{1}$$

where $\mathbf{s_{n-1}}$ and $\mathbf{s'_{n-1}}$ represent the original and reconstructed vector of the Mel-spectrogram at the $n-1$ analysis window, $\mathbf{s'_n}$ and $\mathbf{s_n}$ are the original and reconstructed vector of the Mel-spectrogram at the n analysis window, $\mathbf{s_{n+1}}$ and $\mathbf{s'_{n+1}}$ are the original and reconstructed vector of the Mel-spectrogram at the $n+1$ analysis window, and $\mathbf{p_n}$ and $\mathbf{p'_n}$ are the PPG of the n analysis window (Section III-B). The weights $c_1 = 0.7$, $c_2 = 0.7$, $c_3 = 0.7$, and $c_4 = 10$ were empirically selected to assign different importance in the reconstruction error of the Mel-spectrogram and PPG vectors.

The 10-dimensional latent space $\mathbf{y_n}$ of the auto-encoder serves as an input to the depression-estimation module $g_{\mathbf{v}}$, where $\mathbf{v}$ are the parameters of a 2-layer FFN with 5 nodes at the inner layer and 1 node that the output layer, such that the model outputs the degree depression severity, as obtained by the PHQ (Section III-A), i.e., $\mathbf{d_n} = g_{\mathbf{v}}(\mathbf{y_n})$. A total of 61 parameters are being learned while training the depression model. The loss for the depression model is defined as the mean-square error between the actual d_n and estimated depression label d'_n, summed over all the analysis windows (i.e., $n = 1, \ldots, N$):

$$L_{depression}(\mathbf{v}) = \frac{1}{N} \sum_{n=0}^{N} (d_n - d'_n)^2 \tag{2}$$

The 10-dimensional latent space $\mathbf{y_n}$ also comprises the input of another 2-layer FFN that serves as the speaker-classification module. The speaker classifier has 50 nodes at the internal layer and 107 nodes at the outer layer, the latter being the same as the number of speakers, and is implemented by function $h_{\mathbf{u}}$ i.e., $\mathbf{i'_n} = h_{\mathbf{u}}(\mathbf{y})$, where $\mathbf{u}$ is the vector that contains the parameters of the 2-layer FNN and is the estimated $\mathbf{i'_n}$ speaker identity at every analysis window n, represented via a 1-hot encoding vector. A total of $6,007$ parameters are being learned while training the speaker classification model. The loss for the speaker classification model is defined as the cross-entropy error (CCE) between the actual $\mathbf{i_n}$ and estimated $\mathbf{i'_n}$ speaker identity at every analysis window n:

$$L_{speaker}(\mathbf{u}) = \sum_{n=1}^{N} CCE(i_n, i'_n) \tag{3}$$

The three models are trained and their parameters $\mathbf{w}$, $\mathbf{v}$, and $\mathbf{u}$ are updated according to their respective losses (i.e., L_{error} for the auto-encoder, $L_{depression}$ for the depression-estimation module, $L_{speaker}$ for the speaker-classification module). The training is conducted in an adversarial manner such that the loss of the auto-encoder and depression estimation model is minimized, while the corresponding loss of the speaker classification model is maximized:

$$\begin{aligned} L(\mathbf{w}, \mathbf{v}, \mathbf{u}) =& \alpha \cdot L_{error}(\mathbf{w}) \\ &+ L_{depression}(\mathbf{v}) - \beta \cdot L_{speaker(\mathbf{u})} \end{aligned} \tag{4}$$

where variables α and β weight the importance of the depression severity error compared to the signal reconstruction error and speaker cross-entropy loss, and were empirically set to

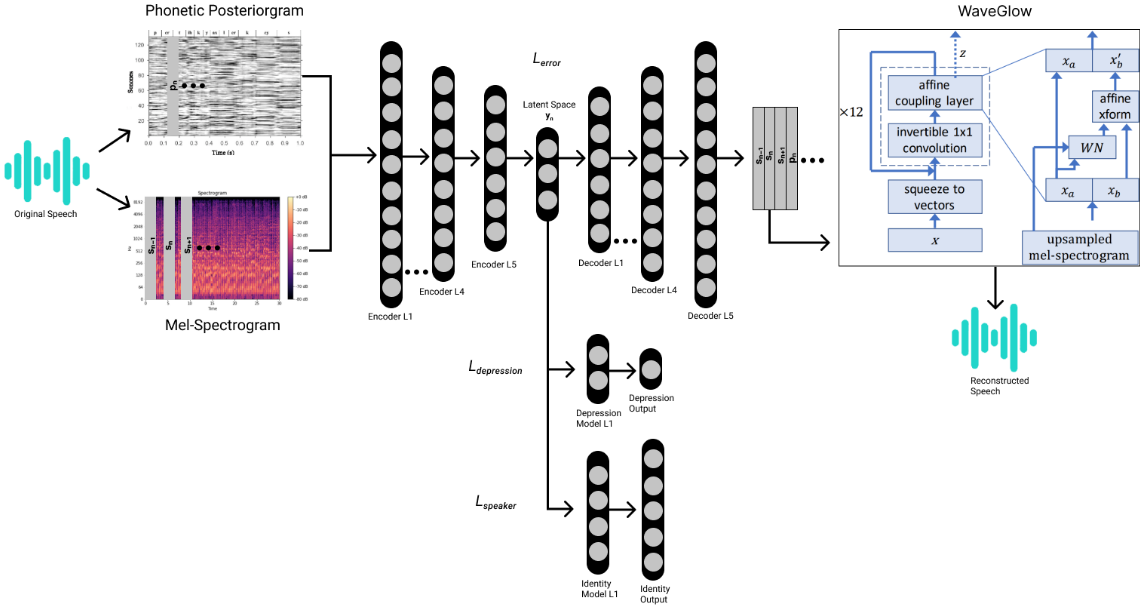

Fig. 1. The proposed mental health (MH) preserving speech anonymization system. The auto-encoder performs an identify mapping between the original and anonymized Mel-spectrograms and posteriograms (PPGs) so that depression information is preserved and speaker information is reduced.

$\alpha = 300$ and 10^{-14} in our experiments to avoid an exploding loss.

We used the WaveGlow vocoder [38] to synthesize the audio using the reconstructed 80-dimensional Mel-spectrogram vector $\mathbf{s'_n}$ for every analysis window n. The WaveGlow vocoder was pre-trained on the LJ speech dataset [39] that consists of $13,100$ short audio clips of a single female speaker reading passages from 7 non-fiction books. The speech output of WaveGlow is converted into a PCM-16 format *wav* file.

D. Evaluation

We evaluate the proposed system in terms of its ability to preserve MH information, eliminate speaker-dependent information, and generate intelligible anonymized speech. The proposed anonymization model (Section III-C) was trained and tested using 7-fold cross-validation. Similar to the feature extraction process (Section III-B), the original dataset was split into analysis windows of 2048 points length with a hop length of 256. The analysis windows were randomly assigned to one of the 7 folds. While the same analysis window was not part of more than one fold, analysis windows from the same speaker could be present at multiple folds, which allowed us to evaluate the speaker classification performance.

The 2-FNN depression estimation model $g_{\mathbf{v}}$ (Section III-C) was used to evaluate the ability of the anonymization system to preserve MH information on the test data of each fold. Evaluation was conducted at each analysis window belonging to the current test set by taking as an input the entire 280-dimensional vector $\mathbf{x_n}$ consisting of the Mel-spectrogram coefficient of the previous, current, and next window, as well as the current PPG. The input was first passed through the encoder part of the auto-encoder function $f_{\mathbf{w}}$, followed by the depression model $g_{\mathbf{v}}$, thus rendering the estimated degree of depression severity, i.e., $d'_n = g_{\mathbf{v}}\left(f_{\mathbf{w}}\left(\mathbf{x_n}\right)\right)$. We note that the parameters $\mathbf{v}$ of the depression model were learned based on the train data, thus no leakage occurs between the train and the test sets. The Pearson's correlation between the predicted and actual depression severity scores is computed over all windows of the current fold, and is then averaged across all folds, to yield a final evaluation metric of depression ($C_{depression} \in [-1, 1]$). A metric $C_{depression}$ closer to 1 indicates a good ability of an input speech signal to preserve depression-related information.

In a similar manner, the 2-FNN speaker classification model $h_{\mathbf{u}}$ (Section III-C) was used to evaluate the extent to which speaker-dependent information is preserved in the audio signal. According to this, a sample $\mathbf{x_n}$ belonging to the test set serves as the input to the speaker classification model, which provides a speaker decision in the form of a 1-hot encoding, , i.e., $s'_n = h_{\mathbf{u}}\left(f_{\mathbf{w}}\left(\mathbf{x_n}\right)\right)$. The speaker classification accuracy, computed as the number of correctly estimated speaker IDs divided by the total of samples, was estimated for each fold, and then averaged across all folds, serving as an evaluation metric for speaker classification (i.e., $A_{speaker} \in [0, 100]$). The metric $A_{speaker}$ would be low for a high-performing speech anonymization system.

The intelligibility of the reconstructed speech signals was further assessed in a quantitative and qualitative manner. Quantitative assessment involved the use of the Google Speech to Text API, which is a cloud-based ASR system implemented via the SpeechRecognition 3.8.1 toolbox. The ASR system takes as an input the speech signal and outputs the corresponding speech transcript, which is then compared against the original one to yield the word error rate (WER) that computes the Levenshtein distance between the two transcripts. Qualitative evaluation was performed via perceptual listening tests administered via the MTurk. To increase the reliability of the experiments, MTurkers were restricted to people who resided in the U.S., are Turk masters, and have an approval rate higher than 98%. The MTurkers were asked to rate the audio signals according to three dimensions: naturalness, comprehension, and pleasantness, similar to the original evaluation from WaveGlow [38] and Tacotron's [40]. MTurkers were asked the following questions for each file: "*The voice was easy to comprehend*", "*The voice sounded natural*," and "*The voice was pleasant*" and provided their answer on a 5-point Likert scale (i.e., 1: Strongly Disagree, 5: Strongly Agree). Each MTurker was given a batch of 10 speech files either from the original signals or generated with the same anonymization method, and did not have access to audio files from other conditions. Each audio file was rated by 5 MTurkers.

The above evaluation metrics were employed based on the original speech signal, as well as the generated speech signals that were synthesized based on the following methods: (1) *WaveGlow Baseline*: The Mel-spectrogram was extracted from the original speech signal and served as an input to the WaveGlow synthesizer. This baseline was used to better understand the potential amount of noise that the Waevglow synthesizer introduces on the original non-anonymized speech signal; (2) *Speech Anonymization 1*: An auto-encoder that contains only the speaker classification module and does not contain the depression estimation module, trained to minimize the following loss function $L(\mathbf{w},\mathbf{u}) = \alpha \cdot L_{error}(\mathbf{w}) - \beta \cdot L_{speaker(\mathbf{u})}$; (3) *Speech anonymization 2*: A voice anonymization method that suppresses speaker information while maintaining speech intelligibility relying on a cascade of signal processing methods [41], that include vocal tract length normalization [42], McAdams transformation [43] that modifies the resonance frequencies, smoothing of the modulation spectrum, signal resampling to stretch the speech signal, and signal distortion via waveform clipping; and (4) *Proposed:* The MH-preserving speech anonymization method based on (4).

IV. Results

We first examine the ability of the model to preserve the linguistic information of the input speech signal via the quantitative evaluation using the ASR system (Table I). The WER of the original speech is 0.32 and increases when using the Waveglow baseline to 0.44. This indicates that the WaveGlow speech synthesizer adds noise to the Original signal, rendering it less intelligible, even when no other transformation has been conducted. The baseline anonymization methods further increase the WER to 0.51 and 0.62 for Speech Anonymization 1 and Speech Anonymization 2, respectively. This indicates that anonymizing the speech renders its content less intelligible compared to the original speech. Finally, the proposed method achieves WER equal to 0.51, which is comparable to the best speech anonymization method (i.e., Speech Anonymization 1). In terms of the listening experiments conducted via MTurk (Table II), both the proposed and baseline methods produce speech that is perceived significantly less natural and pleasant (i.e., $p \simeq 0$) than the original speech, a result that has been found in prior work [44]. However, the perceptual results achieved between the two anonymization methods were equivalent, since naturalness, comprehensibility, and pleasantness was not significantly different between Speech anonymization 1 and the Proposed anonymization method (i.e., $p = 0.11$, $p = 0.11$, $p = 0.23$ for each of the three dimensions, respectively). We also report these perceptual dimensions in terms of gender (Table II). Listeners thought that the original utterances expressed by female speakers were overall more natural, comprehensible, and pleasant compared to the ones provided by the male speakers ($p \simeq 0$), a finding which is consistent with prior work [45]. This difference was preserved when transforming the original signals via Speech Anonymization 1 ($p \simeq 0$) and the Proposed method ($p \simeq 0$). A potential reason for this might also be the fact that the WaveGlow vocoder was trained on data from a female speaker.

The speech anonymization baselines have decreased ability to preserve depression information (Table I), since the depression estimation model has significantly lower Pearson's correlation (i.e., $r = 0.38$ and $r \simeq 0$ for Speech Anonymization 1 and 2, respectively) compared to the original signal (i.e., $r = 0.52$) (i.e., $p \simeq 0$ when comparing Speech Anonymization 1 and 2 with the Original speech). However, the proposed MH-preserving speech anonymization achieves depression estimation performance similar to the original signal (i.e., $r = 0.55$). We suspect that the slightly higher Pearson's correlation of the proposed method compared to the original speech might be the result of minimizing the depression loss function, therefore potentially producing spectrotemporal characteristics of the signal that preserve the degree of depression severity.

In terms of speaker classification, the speaker classifier achieves moderate accuracy (i.e., 51.8%) on the original speech, which is significantly reduced when using the proposed MH-preserving speech anonymization baseline methods (i.e., 31.2% and 2.2% for Speech Anonymization 1 and 2, respectively) (Table I). The proposed MH-preserving speech anonymization method achieves similar degradation in speech classification accuracy (i.e., 35.9%) compared to Speech Anonymization 1, which is also based on an adversarial learning approach. A potential reason why Speech Anonymization 2 is so effective in removing speaker-specific properties from the original speech might be the fact that the transformations conducted as part of this method completely distort the spectrotemporal patterns of the original speech signal, therefore significantly reducing evidence from both the speaker ID and the degree of depression severity. Overall, these indicate that the proposed system can degrade speaker-specific information to some extent, although it is not able to fully eliminate this information, potentially because MH- and speaker-specific information are highly interconnected.

V. Discussion

We proposed a model that can be used to increase the anonymity of speech signals while preserving information related to the language content and the speaker's MH information. Our results indicate that the proposed model is able to remove information that can be used to recognize the speaker while pronouncing information that is indicative of depression severity. We further found that the linguistic

TABLE I
WORD ERROR RATE (WER), PEARSON'S CORRELATION BETWEEN ACTUAL AND ESTIMATED DEPRESSION SEVERITY, SPEAKER CLASSIFICATION ACCURACY OF ORIGINAL AND SYNTHESIZED SPEECH.

Method	**WER**	**Depression Correlation** $C_{depression}$	**Speaker Classif. Accuracy (%)** $A_{speaker}$
Original speech	0.322	0.521	51.8
WaveGlow Baseline	0.446	0.003	1.8
Speech Anonymization 1	0.511	0.382	31.2
Speech Anonymization 2	0.622	-0.007	2.2
Proposed	0.515	0.553	35.9

TABLE II
AVERAGE LISTENER RATINGS OF NATURALNESS, COMPREHENSIBILITY, AND PLEASANTNESS.

Method	**Perceptual Dimension**	**Listener's Rating**		
		All	**Female**	**Male**
Original	Naturalness	4.36	4.48	4.27
	Comprehension	4.31	4.40	4.23
	Pleasantness	3.82	4.00	3.66
Speech Anonymization 1	Naturalness	2.822	3.16	2.52
	Comprehension	3.40	3.66	3.16
	Pleasantness	2.77	3.16	2.41
Proposed	Naturalness	2.76	3.01	2.48
	Comprehension	3.34	3.56	3.14
	Pleasantness	2.72	3.1	2.38

content of a sentence gets degraded after the conversion, which is both a result of the vocoder that is being used, and the transformation that is implemented via the auto-encoder. Based on the perceptual experiments, the anonymized speech was still perceived as comprehensible, but its comprehensibility was reduced in comparison to the original speech.

The results of this work should be considered accounting for the following limitations. We modeled inherent temporal dependencies between audio frames by superimposing three consecutive vectors of the Mel-spectrogram, which served as an input to the auto-encoder. While this preliminary approach rendered promising results, we will examine state-of-the-art speech-to-speech approaches that rely on 1D-CNN and LSTM networks to explicitly model the spectrotemporal evolution of speech signals. The small size of DAIC-WoZ dataset might further prevent the results of this paper to be generalizable to other speakers and different audio recording settings. DAIC-WoZ consisted of short utterances from 107 participants, and depicted an uneven split of depressed and non-depressed participants. Many utterances were short and did not contain much information. Having a larger dataset in natural conditions could have potentially yielded different results. As part of our future work, we will consider a more broad range of ages to account for different vocal cord characteristics and additional MH conditions, such as PTSD. Finally, additional evaluation approaches could be considered. The model evaluation in terms of the extent to which speaker identity is preserved is being conducted via a speaker classification module. Instead, a large part of the current literature focuses on speaker verification systems that are pre-trained on large corpora, such as the VoxCeleb dataset, and relies on audio features known to effectively preserve the speaker characteristics, such as the x-vector. Qualitative evaluation was performed only in terms of speech intelligibility. It would be worthwhile to conduct additional experiments with MH clinicians in order to understand the extent to which the proposed anonymization approach preserves clinically-relevant information at the perceptual level, as well as with users and patients in order to quantify perceived re-identification risks.

The findings from this study can lay a foundation towards trustworthy speech-based technologies for accessible MH diagnosis, monitoring, and prevention. MH-preserving speech anonymization can reduce public resistance towards ambulatory technologies and can potentially decrease user skepticism in data sharing, since users can consent in sharing only the anonymized speech, while the original speech can be permanently deleted. This can also have positive implications on people who are bounded by legal constraints, such as undocumented immigrants, inmates, and military veterans who might need MH services. An important component for increasing the feasibility of such technologies would be to create lightweight edge computing algorithms that can work closer to the sources of the data (e.g., on the mobile device), therefore minimizing identity leakage risks that yield from transferring and storing the speech data to online servers. While the proposed auto-encoder based algorithm is computationally lighter compared to large-scale deep learning models, additional improvement is needed for this algorithm to be truly ubiquitous and implementable on edge devices.

VI. CONCLUSION

This paper demonstrates the feasibility of designing speech anonymization algorithms that preserve MH information, while reducing evidence of speaker identity on the speech signal. Our proposed algorithm relies on an auto-encoder architecture that generates the anonymized speech at its output with two constraints; one constraint is to learn the identity mapping between original and anonymized signal while reducing the loss of the depression-estimation module, while the other constraint aims to maximize the loss of the speaker-classification module. The acoustic and linguistic information is preserved by conducting the identity mapping with respect to the Mel-spectrogram coefficients and the PPG for each analysis window. Quantitative and qualitative results indicate that the speech signals synthesized by the proposed anonymization algorithm are less intelligible compared to the original speech, but equally intelligible to the considered baseline speech anonymization algorithms. The proposed anonymization algorithm can further effectively suppress information related to the speaker ID, since the anonymized signals depict 15.9% reduction in speaker classification accuracy compared to the original speech, and can effectively preserve MH-related information, since the anonymized speech signals yield similar performance to the original speech when they are used as an input for estimating one's degree of depression severity.

REFERENCES

[1] C. f. B. H. S. SAMHSA and Quality, "Behavioral health trends in the united states:results from the 2014 national survey on drug use and health," 2015. [Online]. Available: https://www.samhsa.gov/data/sites/default/files/NSDUH-FRR1-2014/NSDUH-FRR1-2014.htm

[2] N. Cummins, J. Epps, M. Breakspear, and R. Goecke, "An investigation of depressed speech detection: Features and normalization," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[3] J. Gideon, K. Matton, S. Anderau, M. G. McInnis, and E. M. Provost, "When to intervene: Detecting abnormal mood using everyday smartphone conversations," *arXiv preprint arXiv:1909.11248*, 2019.

[4] L. Marzano, A. Bardill, B. Fields, K. Herd, D. Veale, N. Grey, and P. Moran, "The application of mhealth to mental health: opportunities and challenges," *The Lancet Psychiatry*, vol. 2, no. 10, pp. 942–948, 2015.

[5] A. C. Timmons, B. R. Baucom, S. C. Han, L. Perrone, T. Chaspari, S. S. Narayanan, and G. Margolin, "New frontiers in ambulatory assessment: Big data methods for capturing couples' emotions, vocalizations, and physiology in daily life," *Social Psychological and Personality Science*, vol. 8, no. 5, pp. 552–563, 2017.

[6] A. Ramsey, S. Lord, J. Torrey, L. Marsch, and M. Lardiere, "Paving the way to successful implementation: identifying key barriers to use of technology-based therapeutic tools for behavioral health care," *The journal of behavioral health services & research*, vol. 43, no. 1, pp. 54–70, 2016.

[7] E. McCallister, T. Grance, and K. Scarfone, "Identifiable information (pii)," *NIST Special Publication*, vol. 800, p. 122, 2010.

[8] B. Kaplan, "Selling health data: de-identification, privacy, and speech," *Cambridge Quarterly of Healthcare Ethics*, vol. 24, no. 3, pp. 256–271, 2015.

[9] F. T. Commission *et al.*, "Internet of things: Privacy & security in a connected world," *Washington, DC: Federal Trade Commission*, 2015.

[10] C. B. Tan, M. H. A. Hijazi, N. Khamis, Z. Zainol, F. Coenen, A. Gani *et al.*, "A survey on presentation attack detection for automatic speaker verification systems: State-of-the-art, taxonomy, issues and future direction," *Multimedia Tools and Applications*, vol. 80, no. 21, pp. 32 725–32 762, 2021.

[11] R. L. German and K. S. Barber, "Consumer attitudes about biometric authentication," *The University of Texas at Austin*, 2018.

[12] S. Day, C. Seninger, J. Fan, K. Pundi, A. Perino, and M. Turakhia, "Digital health consumer adoption report 2019." 2019.

[13] F. Fang, X. Wang, J. Yamagishi, I. Echizen, M. Todisco, N. Evans, and J.-F. Bonastre, "Speaker anonymization using x-vector and neural waveform models," *arXiv preprint arXiv:1905.13561*, 2019.

[14] N. Tomashenko, X. Wang, E. Vincent, J. Patino, B. M. L. Srivastava, P.-G. Noé, A. Nautsch, N. Evans, J. Yamagishi, B. O'Brien *et al.*, "The voiceprivacy 2020 challenge: Results and findings," *Computer Speech & Language*, vol. 74, p. 101362, 2022.

[15] Y. Yang, C. Fairbairn, and J. F. Cohn, "Detecting depression severity from vocal prosody," *IEEE transactions on affective computing*, vol. 4, no. 2, pp. 142–150, 2012.

[16] D. M. Low, K. H. Bentley, and S. S. Ghosh, "Automated assessment of psychiatric disorders using speech: A systematic review," *Laryngoscope Investigative Otolaryngology*, vol. 5, no. 1, pp. 96–116, 2020.

[17] J. C. Mundt, P. J. Snyder, M. S. Cannizzaro, K. Chappie, and D. S. Geralts, "Voice acoustic measures of depression severity and treatment response collected via interactive voice response (ivr) technology," *Journal of neurolinguistics*, vol. 20, no. 1, pp. 50–64, 2007.

[18] N. Cummins, "Automatic assessment of depression from speech: paralinguistic analysis, modelling and machine learning," *School of Electrical Engineering and Telecommunications, PhD Thesis, UNSW Australia, Sydney, Australia*, 2016.

[19] A. C. Trevino, T. F. Quatieri, and N. Malyska, "Phonologically-based biomarkers for major depressive disorder," *EURASIP Journal on Advances in Signal Processing*, vol. 2011, no. 1, pp. 1–18, 2011.

[20] J. R. Williamson, T. F. Quatieri, B. S. Helfer, R. Horwitz, B. Yu, and D. D. Mehta, "Vocal biomarkers of depression based on motor incoordination," in *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*, 2013, pp. 41–48.

[21] S. Scherer, G. Stratou, J. Gratch, and L.-P. Morency, "Investigating voice quality as a speaker-independent indicator of depression and ptsd." in *Interspeech*, 2013, pp. 847–851.

[22] S. Khorram, J. Gideon, M. G. McInnis, and E. M. Provost, "Recognition of depression in bipolar disorder: Leveraging cohort and person-specific knowledge." in *INTERSPEECH*, 2016, pp. 1215–1219.

[23] L. He and C. Cao, "Automated depression analysis using convolutional neural networks from speech," *Journal of biomedical informatics*, vol. 83, pp. 103–111, 2018.

[24] J. F. Cohn, N. Cummins, J. Epps, R. Goecke, J. Joshi, and S. 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.

[25] S. Ding, G. Zhao, and R. Gutierrez-Osuna, "Accentron: Foreign accent conversion to arbitrary non-native speakers using zero-shot learning," *Computer Speech & Language*, vol. 72, p. 101302, 2022.

[26] J. Qian, H. Du, J. Hou, L. Chen, T. Jung, and X. Li, "Speech sanitizer: Speech content desensitization and voice anonymization," *IEEE Transactions on Dependable and Secure Computing*, 2019.

[27] C. O. Mawalim and M. Unoki, "Speech watermarking method using mcadams coefficient based on random forest learning," *Entropy*, vol. 23, no. 10, p. 1246, 2021.

[28] J. Patino, N. Tomashenko, M. Todisco, A. Nautsch, and N. Evans, "Speaker anonymisation using the mcadams coefficient," in *Interspeech*, 2020.

[29] P. Champion, D. Jouvet, and A. Larcher, "Speaker information modification in the voiceprivacy 2020 toolchain," in *Interspeech*, 2020.

[30] I.-C. Yoo, K. Lee, S. Leem, H. Oh, B. Ko, and D. Yook, "Speaker anonymization for personal information protection using voice conversion techniques," *IEEE Access*, vol. 8, pp. 198 637–198 645, 2020.

[31] F. Bahmaninezhad, C. Zhang, and J. H. Hansen, "Convolutional neural network based speaker de-identification." in *Odyssey*, 2018, pp. 255–260.

[32] B. M. L. Srivastava, N. Tomashenko, X. Wang, E. Vincent, J. Yamagishi, M. Maouche, A. Bellet, and M. Tommasi, "Design choices for x-vector based speaker anonymization," in *Interspeech*, 2020.

[33] B. M. L. Srivastava, M. Maouche, M. Sahidullah, E. Vincent, A. Bellet, M. Tommasi, N. Tomashenko, X. Wang, and J. Yamagishi, "Privacy and utility of x-vector based speaker anonymization," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 2021.

[34] A. S. Shamsabadi, B. M. L. Srivastava, A. Bellet, N. Vauquier, E. Vincent, M. Maouche, M. Tommasi, and N. Papernot, "Differentially private speaker anonymization," *arXiv preprint arXiv:2202.11823*, 2022.

[35] J. Gratch, R. Artstein, G. Lucas, G. Stratou, S. Scherer, A. Nazarian, R. Wood, J. Boberg, D. DeVault, S. 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.

[36] K. Kroenke and R. L. Spitzer, "The phq-9: a new depression diagnostic and severity measure," pp. 509–515, 2002.

[37] G. Zhao, S. Ding, and R. Gutierrez-Osuna, "Foreign accent conversion by synthesizing speech from phonetic posteriorgrams." in *INTERSPEECH*, 2019, pp. 2843–2847.

[38] R. Prenger, R. Valle, and B. Catanzaro, "Waveglow: A flow-based generative network for speech synthesis," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 3617–3621.

[39] K. Ito and L. Johnson, "The lj speech dataset," https://keithito.com/LJ-Speech-Dataset/, 2017.

[40] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio *et al.*, "Tacotron: Towards end-to-end speech synthesis," *arXiv preprint arXiv:1703.10135*, 2017.

[41] H. Kai, S. Takamichi, S. Shiota, and H. Kiya, "Lightweight voice anonymization based on data-driven optimization of cascaded voice modification modules," in *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021, pp. 560–566.

[42] L. Lee and R. Rose, "A frequency warping approach to speaker normalization," *IEEE Transactions on speech and audio processing*, vol. 6, no. 1, pp. 49–60, 1998.

[43] S. E. McAdams, *Spectral fusion, spectral parsing and the formation of auditory images*. Stanford university, 1984.

[44] S. E. Stern, J. W. Mullennix, and S. J. Wilson, "Effects of perceived disability on persuasiveness of computer-synthesized speech." *Journal of Applied Psychology*, vol. 87, no. 2, p. 411, 2002.

[45] W. Ji, R. Liu, and S. Lee, "Do drivers prefer female voice for guidance? an interaction design about information type and speaker gender for autonomous driving car," in *International Conference on Human-Computer Interaction*. Springer, 2019, pp. 208–224.