Journal Pre-proof

Detecting somatisation disorder via speech: introducing the Shenzhen Somatisation Speech Corpus

Kun Qian, Ruolan Huang, Zhihao Bao, Yang Tan, Zhonghao Zhao, Mengkai Sun, Bin Hu, Björn W. Schuller, Yoshiharu Yamamoto

PII: S2667-1026(23)00021-9
DOI: https://doi.org/10.1016/j.imed.2023.03.001
Reference: IMED 73

To appear in: *Intelligent Medicine*

Received date: 7 October 2022
Revised date: 23 February 2023
Accepted date: 7 March 2023

Please cite this article as: Kun Qian, Ruolan Huang, Zhihao Bao, Yang Tan, Zhonghao Zhao, Mengkai Sun, Bin Hu, Björn W. Schuller, Yoshiharu Yamamoto, Detecting somatisation disorder via speech: introducing the Shenzhen Somatisation Speech Corpus, *Intelligent Medicine* (2023), doi: https://doi.org/10.1016/j.imed.2023.03.001

Detecting somatisation disorder via speech: introducing the Shenzhen Somatisation Speech Corpus

Kun Qian[1,2*‡], Ruolan Huang[3,4‡], Zhihao Bao[1,2‡], Yang Tan[1,2†], Zhonghao Zhao[1,2†], Mengkai Sun[1,2], Bin Hu[1,2*], Björn W. Schuller[5,6], and Yoshiharu Yamamoto[7]

1. Key Laboratory of Brain Health Intelligent Evaluation and Intervention, Ministry of Education, Beijing Institute of Technology, Beijing 100081, China
2. School of Medical Technology, Beijing Institute of Technology, Beijing 100081, China
3. The First School of Clinical Medicine, Southern Medical University, Guangzhou 510515, China
4. Department of Neurology, Shenzhen University General Hospital, Shenzhen 518055, China
5. GLAM – Group on Language, Audio, & Music, Imperial College London, London SW7 2AZ, UK
6. Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Augsburg 86159, Germany
7. Educational Physiology Laboratory, The University of Tokyo, Tokyo 113-0033, Japan

Abstract

Objective Speech recognition technology is widely used as a mature technical approach in many fields. In the study of depression recognition, speech signals are commonly used due to their convenience and ease of acquisition. Though speech recognition is popular in the research field of depression recognition, it has been little studied in somatisation disorder recognition. The reason for this is the lack of a publicly accessible database of relevant speech and benchmark studies. To this end, we introduce our somatisation disorder speech database and give benchmark results.
Methods By collecting speech samples of somatisation disorder patients, in cooperation with the Shenzhen University General Hospital, we introduce our somatisation disorder speech database, the Shenzhen Somatisation Speech Corpus (SSSC). Moreover, a benchmark for SSSC using classic acoustic features and a machine learning model is proposed in our work.
Results To obtain a more scientific benchmark, we have compared and analysed the performance of different acoustic features, i. e., the full ComParE feature set, or only MFCCs, fundamental frequency (F0), and frequency and bandwidth of the formants (F1-F3). By comparison. the best result of our benchmark is the 76.0 % unweighted average recall achieved by a support vector machine with formants F1–F3.
Conclusion The proposal of SSSC bridges a research gap in somatisation disorder, providing researchers with a publicly accessible speech database. In addition, the results of the benchmark show the scientific validity and feasibility of computer audition for speech recognition in somatization disorders.

Keywords: Somatisation Disorder, Machine Learning, Healthcare, Computer Audition

1. Introduction

According to the World Health Organisation (WHO) web report, 1 in 8 people worldwide are suffering from mental disorders in 2019 [1]. Since COVID-19 broke out in 2020, all kinds of mental disorders around the world have become more frequent [2–5]. In particular, the number of major depression and anxiety patients worldwide has increased by more than 25 % [6]. Mental disorders have brought severe harm to patients themselves, families, and the community. For the patients themselves, on the one hand, self-harm and attempted suicide are the most harmful behaviours to them, which directly bring physical damage and pain to the patients. As reported in the latest WHO statistics, about 700 000 people worldwide die from suicide every year [7], of which mental disorders account for a large proportion, and many more have attempted suicide. On the other hand, people with psychosis have an increased risk of physical-related illnesses and a shorter life span than the ordinary people [8–10]. For the patients' families, they have to tolerate a series of mental pressures and financial burden brought by the patients. As a result, their life qualities decline fast. For the community, patients with psychotic episodes may cause socially endangering be-

haviours. They may attack people around them or even strangers, out of their control. What is more, neurosis is also characterised by a higher prevalence in children and adolescents [11] and in women than in men [12]. Compared to depression and anxiety, somatisation disorder (SD) is a mental disorder that is less regarded. In fact, it also has a higher prevalence and risk. According to a report, the prevalence rate of SD is 10.1 % in General Hospital Psychiatric Units Tertiary Care Centres in India [13]. In addition, the prevalence rate in women is higher than that in men. In the study of Babu *et al.*, the prevalence rate of SD in adult women was 40.8 % [14]. SD is a neurosis characterised by persistent fear or belief in dominance of various somatic symptoms. Because SD causes patients to shift from emotional and mental distress to the body, they often suffer from unexplained physical discomfort [15]. Such physical discomfort includes stomach pain, back pain, joint pain, and further more. This medically unexplained symptom is characterised by multiple occurrences, persistence, and recurring. Patients are often unaware that they have a SD. Attributing this pain to a physical illness, they repeatedly seek clinical medical advice and treatment, but to no avail [16]. In terms of economic impact, patients frequently seeking medical advice in the wrong direction greatly increase their cost of medical care. The per capita expenditure for health care of patients is up to nine times of the average amount [17]. It also places an indirect burden on the healthcare system [18–20]. According to the report, the annual medical costs of SD in the United States are approximately 2 560 billion dollars [21]. Moreover, unresolved pain and confusion lead to lower quality of life and greater mental stress for patients, which aggravates their condition, or even cause a higher suicide risk [22].

However, the vast majority of patients could not receive effective treatment. This is because the mental health system is severely under-resourced and patients lack of relevant mental health knowledge (especially for some less common mental disorders). Early diagnosis of mental illness is the first step in obtaining beneficial treatment for patients, but the approaches are scarce. As shown in Fig. 1, clinical scales and speech analysis are used as the two main methods for the diagnosis of mental disorders introduced in this paper. The scale, combined with the communication and observation of the patient's behaviour, facial expressions, and speech, among others by doctors, is the common form of psychiatric diagnosis by far. But this approach relies on subjective interviews with patients and the clinical experience of the doctors, which results in a partial deviation of the diagnosis from the reality. With the development of computer technology, it is encouraging to see that, artificial intelligence (AI) has been applied to the classification of mental disorders, which avoids the pitfalls of the scale. In particular, audio signals, because of their 'non-invasive' nature, combined with the rapidly developing computer audition (CA) [23] technology are becoming a popular topic of digital medicine research in the search for new digital phenotypes. Speech signals, as a subclass of audio signals, have been demonstrated to be reliable in the diagnosis of certain mental disorders, such as depression and anxiety [24–27]. In particular, related studies have shown that speech features performed better than visual features or text in depression prediction tasks [28, 29].

There are a mount of AI researches on speech representation of mental disorders. For anxiety, Wang *et al.* [24] proposed a new Fourier parameter model using the perceptual content of voice quality. Dan *et al.* [25] used K-nearest neighbours as classifier and focused on the fundamental frequency for classification. For depression, much work has been done to detect depression through speech. Pan *et al.* [26] extracted 988 speech features from speech data and established a logistic regression model to achieve a better depression classification rate. Rejaibi *et al.* [27] proposed a deep learning-based method to assess depression and predict its severity. They aimed to extract Mel Frequency Cepstral Coefficients (MFCCs) from speech and used long short-term memory networks. For insomnia, Espinoza *et al.* [30] studied obstructive sleep apnea among sleep disorders using a larger speech database containing 426 participants. They used supervector or i-vector techniques to model speech spectral information and predicted by support vector regression. The above researches demonstrated that speech has the ability to represent mental disorders. For SD, unfortunately, due to its mix-up feature of psychologically and physically, it is challenging for clinicians to recognise this psychiatric disorder masquerading as physical pain [31, 32]. Not only that, the actual clinical condition may also be diagnosed as a SD, thereby hindering the patient's real need for treatment and longer-term pain [33]. However, there are few reports on the application of AI technology in this field. In particular, we found rare work on speech as raw data for classification. According to our search, Idenfors *et al.* [34] analysed brain images for global-brain functional connectivity (GFC). The combination of GFC values and support vector machine (SVM) were used to distinguish patients from the controls. The results showed that the patients' GFC was abnormal. Lv *et al.* [35] constructed an improved bacterial foraging optimisation-based kernel extreme learn-

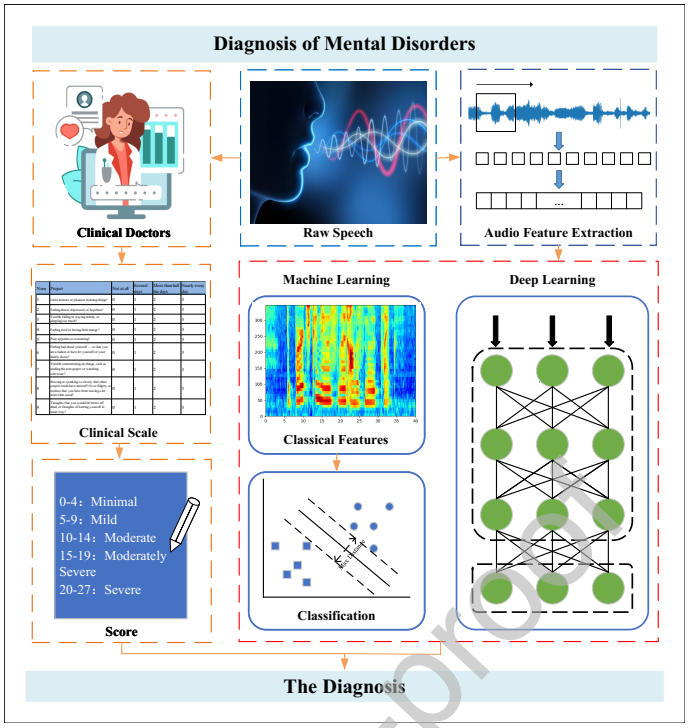


Figure 1: Methods of diagnosing mental disorders.

ing machine (IBFO-KELM) model based on the symbol self-assessment scale (SCL-90) data for the diagnosis of patients with SD. Human communication relies on language communication. From speech, we can feel the others' emotional or mental state. Numerous studies have shown that speech-language pathologists can diagnose mental disorders through speech [36–38]. Therefore, ubiquitous, inexpensive, and easily-acquired speech signals can be used as raw and reliable information for diagnosing or predicting mental disorders.

In this paper, we demonstrate the collected data and publish the speech database for the classification of SD. The name we use is the Shenzhen Somatisation Speech Corpus (SSSC). Unlike the automatic speech recognition technology related to the UASPEECH databases [39], the related research on the SSSC will belong to the category of mental emotion recognition, although these databases are both composed of speech. A conventional and reproducible benchmark for this publicly accessible speech database is also announced. The main contributions of this work are: 1) We demonstrate the feasibility of using speech to classify SD through the benchmark. 2) We provide a standard speech database for the classification of SD. 3) We compare several typical acoustic features to illustrate the usability of this database.

The remainder of this paper is organised as follows: At first, the materials and methods are presented in Section 2. Subsequently, the results of the benchmark work are given in Section 3. We give an experimental discussion, current limitations, and outlooks in Section 4. Finally, we draw a conclusion for our work in Section 5.

2. Methods

In this section, we firstly introduce the proposed publicly somatisation speech database, i. e., SSSC. Next, openSMILE, the toolkit that was used to extract acoustics features in our study is briefly introduced. Then, we give more detailed descriptions of our database work. In the final part, the machine learning method and the optimised strategy we adopted is given.

2.1. SSSC database

2.1.1. Data collection

This study was approved by the ethic committee of the Shenzhen University General Hospital. All the participants involved were informed that their voice data will be used only for research purposes. Their agreements for this study were recorded as one of the five

following original speech phrases. The data was collected in out-patient-department in the Shenzhen University General Hospital. We asked the participants to speak five sentences (with neutral contextual meaning). At the same time, two self-report questionnaires were answered by the participants regarding their anxiety and SD (physical discomfort). The questionnaire is constructed by two widely used separated questionnaires. The two questionnaires are GAD-7 (Patient Anxiety Questionnaire) and PHQ-15 (Patient Health Questionnaire 15) [40–42]. All the data were collected from 12 November 2020 to 5 April 2021. We used a Shinco RV-18 recording pen with 32 GB of storage to record all the participants' voices which have a sample rate of 32 000 Hz and a bit rate of 16 bps.

Following, we give examples of the recorded sentences which were spoken in Chinese:

1. Today, I am in Shenzhen, Guangdong Province, China.
2. I want to know if computer technology can help me improve my life and to what extent.
3. She / He is my friend.
4. Time is money, efficiency is life.
5. I agree to use my voice for emotion recognition.

Fig. 2 shows the spectrograms (extracted from *dev0136.wav, dev0003.wav, dev0010.wav*, and *dev0004.wav*, respectively) corresponding to the second sentence in Chinese.

2.1.2. *Data pre-processing*

As described in [43], we executed a series of data pre-processing stages before establishing the 'standard' SSSC, which includes data cleansing, voice activity detection, speaker diarisation, and speech transcription. First, we excluded recordings with low quality (e. g., the level of the speech is low compared to the background noise). Then, we removed the non-speech parts (e. g., non-subjects' speech, breathing, and coughing) from each recording, which resulted in maintaining only the segments including voice from the recordings. The segments containing solely the target patient and scripted content (e. g., excluding laughing) were kept. Finally, we obtained 705 audio recordings from 141 participants. To attenuate the effects of the audio recording equipment, the background noise condition and the level of the recording, all files were first high-pass filtered to eliminate low-frequency background noise (cut-off frequency: 120 Hz, 10^{th}-order Chebyshev filter) and then their waveforms were normalised individually (peak amplitude set to -3 dB).

Table 1: Subjects information in data splits.

	Average Age	**# Male**	**# Female**
Train	38.2	46	33
Dev	35.0	20	12
Test	36.7	18	12
$\sum$	37.1	84	57

2.1.3. *Tasks definition*

We define two tasks for the benchmark setup: First, the Anxiety Degree should be grouped into: Yes (labelled as "1" when GAD-7 $\geq$ 11), No (labelled as "0" when GAD-7 $<$ 11). Then, an estimation of the Physical Discomfort Disorder Degree should be made as: Yes (labelled as "1" when PHQ-15 $\geq$ 10), No (labelled as "0" when PHQ-15 $<$ 10).

2.1.4. *Data partitioning*

Totally, the database contains as mentioned audio samples of 705 speech events from 141 subjects who were checked without organic disease by experts. The number of males in all subjects is 84, and the number of females is 57. The average age of all subjects is 37.1 ± 13.2 years (range 15 to 70). The total duration of audio in the database is 3 039.192 s, equalling roughly 50 minutes. The average sample duration is 4.311 ± 2.297 s (range 0.864 to 16.920 s). In order to carry out the experiment, we randomly partitioned the data into a train, a development (dev) and a test set, which are subject independence. Table 1 gives more information about data partitions in detail. Fig. 3 shows data distribution details.

2.2. openSMILE *and acoustic features*

openSMILE [44] is an open source toolkit which is widely applied in the field of acoustic representation extractions. openSMILE can provide features commonly used in classic acoustic signal processing methods, such as short-time zero-crossing rate, energy spectrum features, and Mel Frequency Cepstrum Coefficients. To get the statistical information of an audio signal sample, openSMILE firstly extracts the low-level descriptors (LLDs) from the original frame-level audio signals, then performs the statistical information extraction to the frame-based LLDs by functionals. By this method, the limitations of static machine learning models such as SVM are unlocked from the inconsistency of sample duration.

To train the baseline system, we use the ComParE feature set, which includes 65 LLDs (see Table 2). The

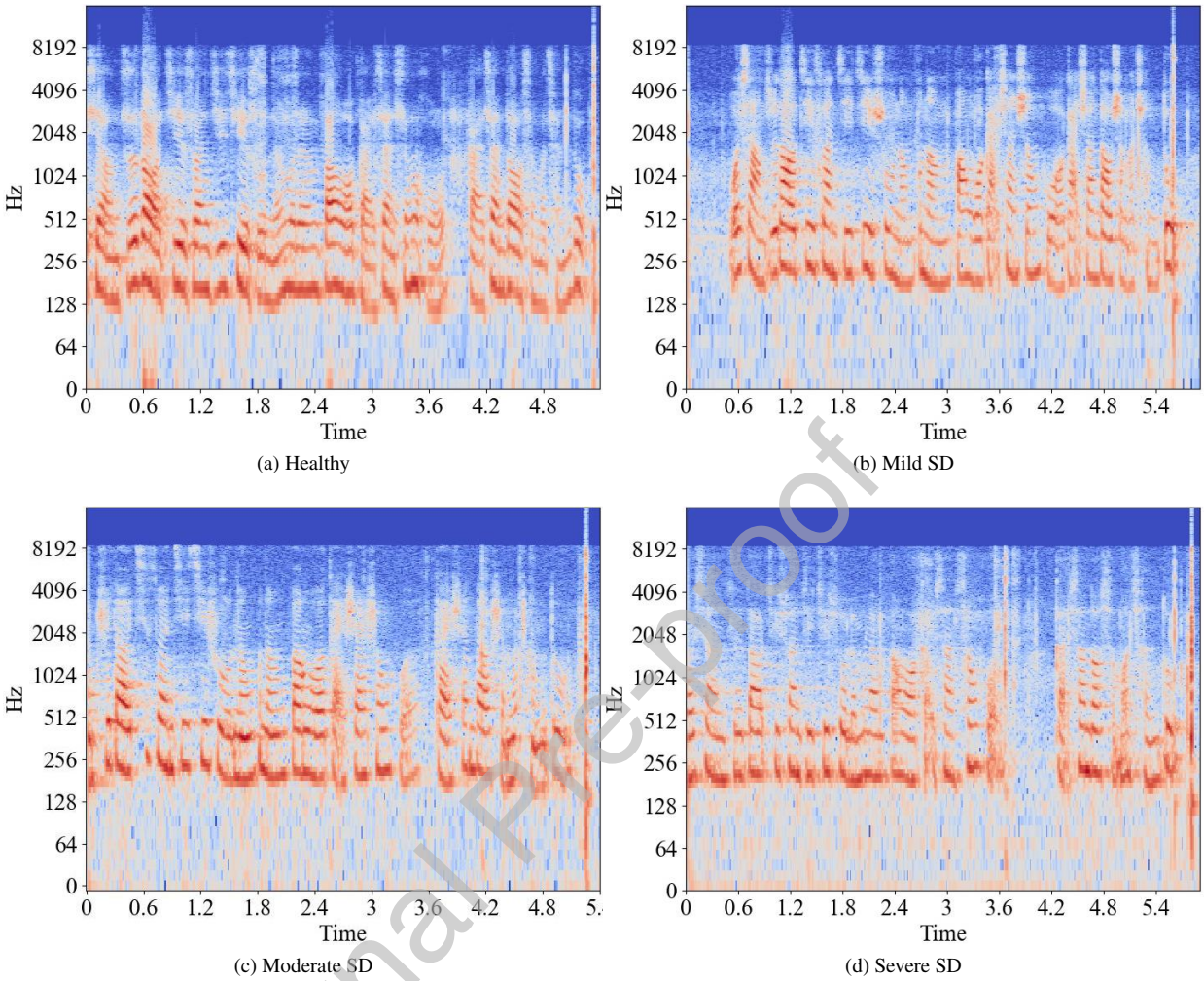


Figure 2: Audio spectrogram examples of different PHQ-15 outcomes.

configuration of the 2.3 version openSMILE is ComParE_2016. The more specific information and details are given in [45]. Moreover, [45] also features a detailed introduction of the used functionals (see Table 3). The ComParE feature set includes in total 6 373 features. The mechanism of functionals is to map the time series based LLDs to a scalar value per each applied functional (e. g., mean, standard deviation, maximum); then, a single, fixed dimension vector which is independent of the audio signal sample's time duration is generated [45].

2.3. *Database related questionnaires*

In our study, we carry out measurement work on two scales as mentioned (i. e., GAD-7 and PHQ-15). Based on the scores of these two scales, the Shenzhen University General Hospital's psychologists enrolled in the study classified the conditions into four classes. We give a separate introduction of each scale in this part. Specifically, Table 4 shows the data distribution in details.

GAD-7 Generalised anxiety disorder (GAD) is a common and disabling illness that is often underdiagnosed and undertreated [46]. With the influence of the COVID-19 in China [47], more and more people are suffering from symptoms of anxiety. Therefore, we want to get more information of the GAD for further research. GAD-7 is a brief clinical measure for assessing GAD, which consists of 7 items about anxiety self-report [48]. The score of each item is between 0-3 and the GAD-7's total score is in the range of 0 to 21. Based on the total scores, the Shenzhen University General Hospital's mental health experts classified the

Table 2: The LLDs for ComParE feature set.
RASTA: Relative Spectral Transform; HNR: Harmonics to Noise Ratio; RMSE: Root Mean Square Energy. Details can be found in [45].

4 Energy related LLDs	**Group**
RMSE, zero-crossing rate	Prosodic
Sum of auditory spectrum (loudness)	Prosodic
Sum of RASTA-filtered auditory spectrum	Prosodic
6 Voicing related LLDs	**Group**
Probability of voicing	Voice Quality
F_0 (SHS and Viterbi smoothing)	Prosodic
log HNR, jitter (local and δ), shimmer (local)	Voice Quality
55 Spectral LLDs	**Group**
MFCCs 1–14	Cepstral
Spectral energy 250–650 Hz, 1 k–4 kHz	Spectral
Spectral flux, centroid, entropy, slope	Spectral
Spectral roll-off point 0.25, 0.5, 0.75, 0.9	Spectral
Spectral variance, skewness, kurtosis	Spectral
Psychoacoustic sharpness, harmonicity	Spectral
RASTA-filtered auditory spectral bands 1–26 (0–8 kHz)	Spectral

participants into 4 types (i. e., **no-anxiety** for 0-4, **mild anxiety** for 5-9, **moderate anxiety** for 10-14, and **severe anxiety** for 15-21).

PHQ-15 SD is a prevalent condition which is not well treated by many psychiatrists [49]. Patients usually seek care in the medical setting convinced that they suffer from physical discomfort rather than a mental disorder. Based on this situation, the SD patients often encounter ineffective care and even harm. That somatising patients may represent 40 % or more of the ambulatory medicine patient population greatly magnifies the problem [50]. To reduce unnecessary expenses, it is very significant for patients with SD to be fully recognised and treated. The classification of SD by the speech of the patients is not only feasible, but also appears efficient and convenient [51]. The PHQ-15 comprises 15 somatic symptoms from the Patient Health Questionnaire (PHQ), each symptom scored from 0 ("not bothered at all") to 2 ("bothered a lot") [52]. According to the PHQ-15 scale's total score ranging from 0 to 30, the participants are divided into 5 types (i. e., **minimal** for 0-4, **low** for 5-9, **medium** for 10-14, and **high** for 15-30).

2.4. Machine learning method and optimising strategy

Support Vector Machine (SVM) is a stable and by now 'traditional' classifier. To make this study comparable and reproducible, we use an SVM classifier with linear kernel to conduct all experiments. We train an SVM model with the complexity parameter in the range of $\{10^{-8},10^{-7},...,10^{-1},1.0\}$. Then, we choose the complexity that performs best on the development set to classify the test set. Moreover, both the training set and the development set train-devel set were joined to predict on test data. Upsampling the training set and the train-devel set was used for balancing the dataset. At the same time, we processed the feature sets with feature normalisation. In order to obtain more specific results, we evaluate the performance of different acoustic features, i. e., the full ComParE feature set, or selectively only MFCCs, fundamental frequency (F0), and frequency and bandwidth of the formants (F1-F3). All the features are extracted by the openSMILE feature extraction and audio analysis tool.

3. Results

Although the database includes two scales, our experiment series focuses on the PHQ-15 outcome. This can reflect the order of severity of the subjects with SD. Moreover, we conduct related experiments to analyse if there is any connection between GAD-7 and PHQ-15. Classification results are shown in Table 5 and Table 6 with feature type and the final feature number. The best mean **unweighted average recall (UAR)** per feature on the development set and test set are highlighted.

In order to obtain a better understanding of the data set, we make some changes for the listed labels in Ta-

Table 3: The functionals applied to LLDs in the ComParE feature set. Note that, the LLDs listed in Table 2 may or may not use all of the functionals of this table, which is described in details in [45].

Functionals
Arithmetic or positive arithmetic mean
Inter-quartile ranges 1–2, 2–3, 1–3,
Linear regression slope, offset
Linear regression quadratic error
Linear Prediction gain and coefficients 1–5
Mean and std. dev. of peak to peak distances
Peak-valley-peak slopes mean and std. dev.
Peak and valley range (absolute and relative)
Peak mean value and distance to arithmetic mean
Quadratic regression coefficients
Quadratic regression quadratic error
Root-quadratic mean, flatness
Rise time, left curvature time
Relative position of max. and min. value
Range (difference between max. and min. values)
Segment length mean, min., max., std. dev.
Standard deviation, skewness, kurtosis, quartiles 1–3
Temporal centroid
Up-level time 25 %, 50 %, 75 %, 90 %
99-th and 1-st percentile, range of these

ble 4: We set a threshold on the scores. We consider the participants whose scores higher than the threshold as aïňÄected by the condition and the healthy when their scores lower than the threshold. According to different thresholds, we set two discrimination modes named "A" and "B" for PHQ-15 (i. e., threshold of each mode, GAD-7: 10, PHQ-15A: 5, PHQ-15B: 10). This modification helps us distinguish the participants with speciïňĄc condition from the healthy subjects. As described above, PQH-15 reflects the degree of physical discomfort. We process the labels based on the mentioned threshold above, and then operate 'AND' or 'OR' on them given their relatedness under the umbrella of being psychological disorders.

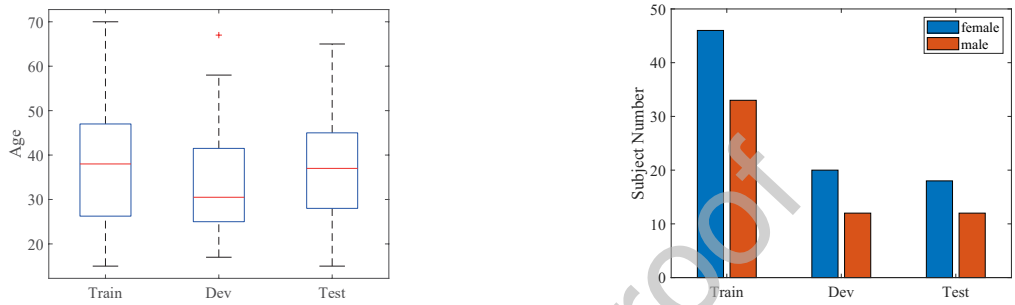
According to the UAR indicators, it appears remarkable that the formants F1-F3 perform best on most of the classification models. This means, F1-F3 will provide more information in our tasks. Except for PHQ-15 labelled with four types scores, the classification results of other tasks achieve the UAR higher than 50.0 %. Fig. 4 and Fig. 5 show confusion matrices of the best-performing setup for the test set. From Fig. 4, we can find that the predicted labels focus on label 0 and label 3. As could be suspected, the classifiers tend to prefer predictions of healthy subjects or such with severe conditions. Although the classifier performs better on this task, unlike humans, the machine requires more samples to understand the features.

4. Discussion

4.1. Classification performance

Refering to Table 5 and Table 6, we can see that, although F1–F3 has the least number of dimensions, it performs better than the other feature sets. However, the performance is relatively poor when we fed all the features in the ComParE feature set into the model. We have noted that some redundant features led to decrease model performance. In future work, we will analysis the contribution of features. Overall, the formant-based features can represent the phenomenon of interest efficiently.

On the other hand, we only use an SVM classifier to test the representational ability of the extracted features. Furthermore, the number of feature types selected are few. In the future, we expect other work on the database to increase the number of features and compare more models to improve the ability to represent this task.



(a) Age Distribution.

(b) Gender Distribution.

Figure 3: Demographic statistic of the SSSC.

4.2. *Features and meaning of SSSC*

SSSC is the first speech dataset available for SD classification. Of great interest, SD could only be recognised shortly when the doctor is skilful and major physical disorders are excluded. Moreover, some symptoms similar to SD should be excluded carefully. The training set in this study had been reviewed beforehand to exclude severe physical disorders such as stroke or coronary artery diseases. The models in current benchmarks achieve over 50 % of UAR, yet more would be needed to expand the reliance. We provide a public database for the study and use of AI and CA. Recognition and primary estimation of the mental state or mood by CA could be the first step. Adjusting the response or reaction could be the next step follow up, which may generate more importance. When different international standards are used to classify categories, the standard we use generates different results. As the first database, the future of SSSC is undoubted. One of the key features of mental disorders is that they are diagnosed without an obvious objective criterion or examinations. Therefore, independent and skilful doctors with psychological experiences are of great importance. Unfortunately, it is becoming more difficult and expensive for doctors likewise to diagnose SD. AI recognition would be the first step for screening before searching for medical help. Similar to AI assistance in the medical imaging or pathological field, CA could surely be of helpful assistance for the primary and unskillful individuals to recognise SD, and even reduce unnecessary anxiety.

4.3. *Current limitations and outlook*

SSSC is the first speech dataset available for SD classification – a heretofore untapped resource for such public data. Most of the models in current benchmarks achieve over 50 % of UAR on the development set. However, SSSC also inevitably underlies limitations:

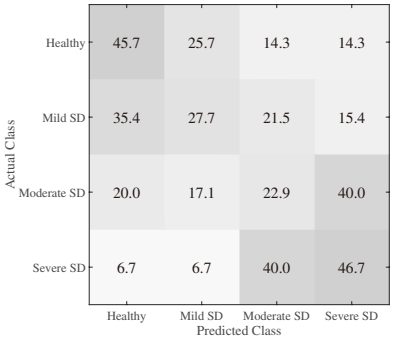


Figure 4: Normalised confusion matrices (values are shown in [%]) of the best-performing (development) model for the PHQ-15 test set.

(a) PHQ-15A

(b) PHQ-15B

(c) GAD-7&PHQ-15A

(d) GAD-7&PHQ-15B

(e) GAD-7|PHQ-15A

(f) GAD-7|PHQ-15B

Figure 5: Normalised confusion matrices (values are shown in [%]) of the best-performing (development) model for the test set.

Table 4: Number of per scales and score in the data splits.

	Score	Label	# Train	# Dev	# Test	∑
GAD-7	0~4	0	130	40	45	215
	5~9	1	125	55	50	230
	10~14	2	80	25	30	135
	15~21	3	60	40	25	125
		∑	395	160	150	705
PHQ-15	0~4	0	90	15	35	140
	5~9	1	135	80	65	280
	10~14	2	110	45	35	190
	≥15	3	60	20	15	95
		∑	395	160	150	705

Table 5: Classification results of different feature types and tasks. (Part A)
#: Number of features; PHQ-15: Labels by way of four scores. PHQ-15A: Labels by way of A; PHQ-15B: Labels by way of B; GAD-7&PHQ-15A: AND operation on GAD-7 and PHQ-15A labels; */*: The first one means the best result of unweighted average recall (UAR) on the development set in all experiments, the last one means the test result by the model that performed best on the development set. (Unit: %.)

Feature type	#	PHQ-15	PHQ-15A	PHQ-15B	GAD-7&PHQ-15A
ComParE	6373	24.4/22.1	53.4/45.9	50.8/49.0	49.8/50.0
MFCCs(all)	1400	28.3/29.5	**58.4**/**64.7**	54.8/61.5	51.6/53.5
MFCCs(only coef)	756	26.4/**35.7**	58.2/62.7	57.4/68.0	50.4/47.0
MFCCs(only delta)	644	25.7/27.1	52.3/52.9	53.2/56.5	**53.8**/50.5
F0	24	23.9/24.7	47.2/49.1	54.8/69.5	47.8/50.5
F1-F3	14	**33.1**/28.3	49.3/52.0	**57.6**/**76.0**	51.0/**54.5**

1. *Limited dataset*. Due to monocentricity, it is difficult for us to recruit a sufficient number of well-represented subjects. This limited amount of data is not conducive to the application of deep learning on it. But in fact, that SSSC can be trusted enough to be adopted for the study of SD was demonstrated through our benchmark experiments. Therefore, we encourage researchers to give more consideration to the application of traditional machine learning which may be more suitable for the small-scale databases. More importantly, in future works, data augmentation is worth looking into and researching, which should be expected to reduce the model's lack in generalisation. An easy way is to use some toolboxes for data enhancement. For instance, [53] published a matlab toolkit that provides 15 different augmentation algorithms for raw audio data and 8 for spectrograms.

2. *Overlapping symptoms*. A study has shown overlap between somatic symptoms, anxiety, and depression [54]. In other words, the three are co-morbid and triggering each other. Thus, different mental illnesses may show the same features. This may be challenging to model, yet, on the contrary, may facilitate better results. As an emerging trend, multi-label learning [55] requires more discovery and exploration in this field.

3. *Reliability of labels*. The labels stem from scoring participants' scale questionnaires. As mentioned above, a small number of labels generated by the scale score may be deviating from the actual situation, as these are greatly influenced by the subjective interviews of participants and the clinical experience of doctors. This is an inevitable mistake caused by the lack of objective factors. In the future, one can involve more professional doctors in the scoring, set more differential versions to recognise SD and use multi-modality to ensure reliability of the label. Nevertheless, manual annotation of speech data is an expensive and time-consuming task. In order to overcome this difficulty, in the AI filed, we think that self-supervised learning [56, 57] could be used to reduce the reliance on labels, as well as introducing active learning [58] on this database.

4. *Availability of samples*. There are incomplete or redundant samples of participants' speech content in the database. Therefore, those samples affect the SD audio analysis. We will exhaustively screen and remove these

Table 6: Classification results of different feature types and tasks. (Part B)
#: Number of features; GAD-7&PHQ15B: AND operation on GAD-7 and PHQ-15B labels; GAD-7|PHQ-15A: OR operation on GAD-7 and PHQ-15A labels; GAD-7|PHQ-15B: OR operation on GAD-7 and PHQ-15B labels; */*: The first one means the best result of unweighted average recall (UAR) on the devolpment set in all experiments, the last one means the test result by the model that performed best on the development set. (Unit: %.)

Feature type	#	GAD-7&PHQ-15B	GAD-7|PHQ-15A	GAD-7|PHQ-15B
ComParE	6373	51.4/47.6	56.8/50.0	53.9/54.5
MFCCs(all)	1400	52.1/59.2	59.8/53.8	54.3/57.2
MFCCs(only coef)	756	51.2/60.4	**61.4**/56.2	54.3/56.6
MFCCs(only delta)	644	50.7/51.2	53.3/48.8	53.2/60.5
F0	24	**55.9**/56.8	47.9/50.0	**55.3**/57.4
F1-F3	14	55.0/**74.8**	51.0/**67.5**	53.5/**61.9**

samples to provide researchers with a more scientific and credible new version.

5. Conclusion

In this study, we firstly introduced a publicly available speech database, namely SSSC. Then, we described the current techniques in somatisation speech classification. A benchmark experiment was given based on the approaches proposed in this work. Moreover, we discussed the results and the limitations, and pointed out some future directions. In this work, an SVM model trained with features based on the first three formants, i. e., spectral maxima, performed best in the classification leading to promising results. We will consider using self-supervision plus fine tuning as strategy in future work.

Conflicts of interest statement

The authors declare that they have no conflicts of interest.

Funding

This work was partially supported by the Ministry of Science and Technology of the People's Republic of China with the STI2030-Major Projects (Grant No. 2021ZD0201900), the National Natural Science Foundation of China (Grant Nos. 62227807 and 62272044), the Teli Young Fellow Program from the Beijing Institute of Technology, the Shenzhen Municipal Scheme for Basic Research (Grant Nos. JCYJ20210324100208022 and JCYJ20190808144005614), China, the JSPS KAKENHI (Grant No. 20H00569), the JST Mirai Program (Grant No. 21473074), the JST MOONSHOT Program (Grant No. JPMJMS229B).

Author contributions

Kun Qian: Idea generation, Funding acquisition, Writing - Conceptualisation, Supervision. **Ruolan Huang**: Data collection & annotation, Funding acquisition, Writing. **Zhihao Bao**: Writing - original draft. **Yang Tan**: Investigation, Writing. **Zhonghao Zhao**: Methodology, Writing. **Mengkai Sun**: Writing - review & editing. **Bin Hu**: Writing - review, Supervision, Funding acquisition. **Björn W. Schuller**: Writing - review. **Yoshiharu Yamamoto**: Writing - review.

References

[1] World Health Organisation. Mental disorders. https://www.who.int/news-room/fact-sheets/detail/mental-disorders, 2022 (accessed 1 July 2022).

[2] Nina Vindegaard and Michael Eriksen Benros. Covid-19 pandemic and mental health consequences: systematic review of the current evidence. *Brain Behav Immun*, 89:531–542, 2020. https://doi.org/10.1016/j.bbi.2020.05.048.

[3] Nicole Racine, Jessica E Cooke, Rachel Eirich, et al. Child and adolescent mental illness during covid-19: a rapid review. *Psychiatry Res*, 292:113307, 2020. https://doi.org/10.1016/j.psychres.2020.113307.

[4] Wanderson Carneiro Moreira, Anderson Reis de Sousa, and Maria do Perpétuo Socorro de Sousa Nóbrega. Mental illness in the general population and health professionals during covid-19: a scoping review. *Texto e Contexto Enferm*, 29, 2020. https://doi.org/10.1590/1980-265x-tce-2020-0215.

[5] Matthew J Carr, Sarah Steeg, Roger T Webb, et al. Effects of the covid-19 pandemic on primary care-recorded mental illness and self-harm episodes in the uk: a population-based cohort study. *Lancet Public Health*, 6(2):e124–e135, 2021. https://doi.org/10.1016/S2468-2667(20)30288-7.

[6] World Health Organisation. Mental health and covid-19: early evidence of the pandemicâĂŹs impact: scientific brief, 2 march 2022. https://www.who.int/publications/i/item/WHO-2019-nCoV-Sci_Brief-Mental_health-2022.1, 2022 (accessed 1 July 2022).

[7] World Health Organisation. World health statistics 2022: monitoring health for the sdgs, sustainable development goals. https://www.who.int/publications/i/item/9789240051157, 2022 (accessed 1 July 2022).

[8] Naomi Launders, Leiah Kirsh, David PJ Osborn, et al. The temporal relationship between severe mental illness diagnosis and chronic physical comorbidity: a uk primary care cohort study of disease burden over 10 years. *Lancet Psychiatry*, 9(9):725–735, 2022. https://doi.org/10.1016/S2215-0366(22)00225-5.

[9] Joseph Firth, Najma Siddiqi, AI Koyanagi, et al. The lancet psychiatry commission: a blueprint for protecting physical health in people with mental illness. *Lancet Psychiatry*, 6(8):675–712, 2019. https://doi.org/10.1016/S2215-0366(19)30132-4.

[10] Mark Rodgers, Jane Elizabeth Dalton, Melissa Harden, et al. Integrated care to address the physical health needs of people with severe mental illness: a mapping review of the recent evidence on barriers, facilitators and evaluations. *Int J Integr Care*, 18(1), 2018. https://doi.org/10.5334/ijic.2605.

[11] Kathleen Ries Merikangas, Erin F. Nakamura, and Ronald C. Kessler. Epidemiology of mental disorders in children and adolescents. *Dialogues Clin Neurosci*, 11(1):7–20, 2009. https://doi.org/10.31887/DCNS.2009.11.1/krmerikangas.

[12] Walter R Gove. The relationship between sex roles, marital status, and mental illness. *Soc Forces*, 51(1):34–44, 1972. https://doi.org/10.1093/sf/51.1.34.

[13] Kalaivanan Rakesh Chander, Narayana Manjunatha, B Binukumar, et al. The prevalence and its correlates of somatization disorder at a quaternary mental health centre. *Asian J Psychiatr*, 42:24–27, 2019. https://doi.org/10.1016/j.ajp.2019.03.015.

[14] Arjun Rajendra Babu, Alexander John Aswathy Sreedevi, and Vijayakumar Krishnapillai. Prevalence and determinants of somatization and anxiety among adult women in an urban population in kerala. *Indian J Community Med*, 44(Suppl 1):S66–S69, 2019. https://doi.org/10.4103/ijcm.IJCM_55_19.

[15] Michael Witthöft, Alexander L Gerlach, and Josef Bailer. Selective attention, memory bias, and symptom perception in idiopathic environmental intolerance and somatoform disorders. *J Abnorm Psychol*, 115(3):397–407, 2006. https://doi.org/10.1037/0021-843X.115.3.397.

[16] Arthur J Barsky, E John Orav, and David W Bates. Distinctive patterns of medical care utilization in patients who somatize. *Med Care*, 44(9):803–811, 2006. https://doi.org/10.1097/01.mlr.0000228028.07069.59.

[17] G Richard Smith Jr, Roberta A Monson, and Debby C Ray. Psychiatric consultation in somatization disorder. *N Engl J Med*, 314(22):1407–1413, 1986. https://doi.org/10.1056/NEJM198605293142203.

[18] Frauke Dorothee Weiss, Winfried Rief, and Maria Kleinstäuber. Health care utilization in outpatients with somatoform disorders: descriptives, interdiagnostic differences, and potential mediating factors. *Gen Hosp Psychiatry*, 44:22–29, 2017. https://doi.org/10.1016/j.genhosppsych.2016.10.0.

[19] Zbigniew Jerzy Lipowski et al. Somatization: the concept and its clinical application. *Am J Psychiatry*, 145(11):1358–1368, 1988. https://doi.org/10.1176/ajp.145.11.1358.

[20] Per Fink. The use of hospitalizations by persistent somatizing patients. *Psychol Med*, 22(1):173–180, 1992. https://doi.org/10.1017/S0033291700032827.

[21] Arthur J Barsky, E John Orav, and David W Bates. Somatization increases medical utilization and costs independent of psychiatric and medical comorbidity. *Arch Gen Psychiatry*, 62(8):903–910, 2005. https://doi.org/10.1001/archpsyc.62.8.903.

[22] Andrea P Chioqueta and Tore C Stiles. Suicide risk in patients with somatization disorder. *Crisis*, 25(1):3, 2004. https://doi.org/10.1027/0227-5910.25.1.3.

[23] Kun Qian, Xiao Li, Haifeng Li, et al. Computer audition for healthcare: opportunities and challenges. *Front Digit Health*, 2:5, 2020. https://doi.org/10.3389/fdgth.2020.00005.

[24] Kunxia Wang, Ning An, Bing Nan Li, et al. Speech emotion recognition using fourier parameters. *IEEE Trans Affect Comput*, 6(1):69–75, 2015. https://doi.org/10.1109/TAFFC.2015.239210110.1109/TAFFC.2015.2392101.

[25] Marius Dan Zbancioc and Silvia Monica Feraru. A study about the automatic recognition of the anxiety emotional state using emo-db. In *2015 E-Health and Bioengineering Conference (EHB)*, pages 1–4, Iasi, Romania, 2015. IEEE. https://doi.org/10.1109/EHB.2015.7391506.

[26] Wei Pan, Jingying Wang, Tianli Liu, et al. Depression recognition based on speech analysis. *Chin Sci Bull*, 63(20):2081–2092, 2018. https://doi.org/10.1360/N972017-01250.

[27] Emna Rejaibi, Ali Komaty, Fabrice Meriaudeau, et al. Mfcc-based recurrent neural network for automatic clinical depression recognition and assessment from speech. *Biomed Signal Process Control*, 71:103–107, 2022. https://doi.org/10.1016/j.bspc.2021.103107.

[28] James R Williamson, Elizabeth Godoy, Miriam Cha, et al. Detecting depression using vocal, facial and semantic communication cues. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, pages 11–18, Amsterdam, The Netherlands, 2016. https://doi.org/10.1145/2988257.2988263.

[29] Le Yang, Hichem Sahli, Xiaohan Xia, et al. Hybrid depression classification and estimation from audio video and text information. In *Proceedings of the 7th annual workshop on audio/visual emotion challenge*, pages 45–51, New York, the United States, 2017. https://doi.org/10.1145/3133944.3133950.

[30] Fernando Espinoza-Cuadros, Rubén Fernández-Pozo, Doroteo T Toledano, et al. Reviewing the connection between speech and obstructive sleep apnea. *Biomed Eng Online*, 15(1):1–20, 2016. https://doi.org/10.1186/s12938-016-0138-5.

[31] Robert C Smith and Francesca C Dwamena. Classification and diagnosis of patients with medically unexplained symptoms. *J Gen Intern Med*, 22(5):685–691, 2007. https://doi.org/10.1007/s11606-006-0067-2.

[32] Alessia Raffagnato, Caterina Angelico, Perla Valentini, et al. Using the body when there are no words for feelings: alexithymia and somatization in self-harming adolescents. *Front Psychiatry*, 11:262–261, 2020. https://doi.org/10.3389/fpsyt.2020.00262.

[33] L De Jonge, S Petrykiv, J Fennema, et al. Misdiagnosis of loin pain hematuria syndrome as a somatization disorder. *Eur Psychiatry*, 41(S1):S491–S491, 2017. https://doi.org/10.1016/j.eurpsy.2017.01.597.

[34] Pan Pan, Yangpan Ou, Qinji Su, et al. Voxel-based global-brain functional connectivity alterations in first-episode drug-naive patients with somatization disorder. *J Affect Disord*, 254:82–89, 2019. https://doi.org/10.1016/j.jad.2019.04.099.

[35] Xinen Lv, Huiling Chen, Qian Zhang, et al. An improved bacterial-foraging optimization-based machine learning framework for predicting the severity of somatization disorder. *Algorithms*, 11(2):17–34, 2018. https://doi.org/10.3390/a11020017.

[36] Cynthia B Solot, Debbie Sell, Anne Mayne, et al. Speech-language disorders in 22q11. 2 deletion syndrome: best practices for diagnosis and management. *Am J Speech Lang Pathol*, 28(3):984–999, 2019. https://doi.org/10.1044/2019_AJSLP-16-0147.

[37] Arthur Brito-Marcelino, Edmea Fontes Oliva-Costa, Salvyana Carla Palmeira Sarmento, et al. Burnout syndrome in speech-

language pathologists and audiologists: a review. *Rev Bras Med Trab*, 18(2):217, 2020. https://doi.org/10.47626/1679-4435-2020-480.

[38] Alyssa M Lanzi, James M Ellison, and Matthew L Cohen. The âĂIJcounseling+âĂİ roles of the speech-language pathologist serving older adults with mild cognitive impairment and dementia from alzheimer's disease. *Perspect ASHA Spec Interest Groups*, 6(5):987–1002, 2021. https://doi.org/10.1044/2021_PERSP-20-00295.

[39] Heejin Kim, Mark Hasegawa-Johnson, Adrienne Perlman, Jon Gunderson, Thomas S Huang, Kenneth Watkin, and Simone Frame. Dysarthric speech database for universal access research. In *Ninth Annual Conference of the International Speech Communication Association*, 2008.

[40] Yen-Cheng Shih, Chien-Chen Chou, Yi-Jiun Lu, et al. Reliability and validity of the traditional chinese version of the gad-7 in taiwanese patients with epilepsy. *J Formos Med Assoc*, pages 1–7, 2022. https://doi.org/10.1016/j.jfma.2022.04.018.

[41] Francisco Cano-García, Javier, Roger Muñoz-Navarro, Albert Abad, Sesé, et al. Latent structure and factor invariance of somatic symptoms in the patient health questionnaire (phq-15). *J Affect Disord*, 261:21–29, 2020. https://doi.org/10.1016/j.jad.2019.09.077.

[42] Xiao-Jie Huang, Hai-Yan Ma, Xue-Mei Wang, et al. Equating the phq-9 and gad-7 to the hads depression and anxiety subscales in patients with major depressive disorder. *J Affect Disord*, 311:327–335, 2022. https://doi.org/10.1016/j.jad.2022.05.079.

[43] Jing Han, Kun Qian, Meishu Song, et al. An early study on intelligent analysis of speech under covid-19: severity, sleep quality, fatigue, and anxiety. In *Proceedings of INTERSPEECH*, pages 4946–4950, Shanghai, China, 2020. https://doi.org/10.1109/EHB.2015.7391506.

[44] 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*, pages 1459–1462, Firenze, Italy, 2010. https://doi.org/10.1145/1873951.1874246.

[45] Florian Eyben. *Real-time Speech and Music Classification by Large Audio Feature Space Extraction*. Springer International Publishing, Cham, Switzerland, 2015. Doctoral Thesis. https://doi.org/10.1007/978-3-319-27299-3.

[46] Jeremy DeMartini, Gayatri Patel, and Tonya L Fancher. Generalized anxiety disorder. *Ann Intern Med*, 170(7):ITC49–ITC64, 2019. https://doi.org/10.7326/AITC201904020.

[47] Yeen Huang and Ning Zhao. Generalized anxiety disorder, depressive symptoms and sleep quality during covid-19 outbreak in china: a web-based cross-sectional survey. *Psychiatry Res*, 288:112954, 2020. https://doi.org/10.1016/j.psychres.2020.112954.

[48] Robert L Spitzer, Kurt Kroenke, Janet BW Williams, et al. A brief measure for assessing generalized anxiety disorder: the gad-7. *Arch Intern Med*, 166(10):1092–1097, 2006. https://doi.org/10.1001/archinte.166.10.1092.

[49] François Mai. Somatization disorder: a practical review. *Can J Psychiatry*, 49(10):652–662, 2004. https://doi.org/10.1177/070674370404901002.

[50] Robert C Smith. Somatization disorder. *J Gen Intern Med*, 6(2):168–175, 1991. https://doi.org/10.1007/BF02598318.

[51] Thomas E Oxman, Stanley D Rosenberg, Paula P Schnurr, et al. Linguistic dimensions of affect and thought in somatization disorder. *Am J Psychiatry*, 1985. https://doi.org/10.1176/ajp.142.10.1150.

[52] Kurt Kroenke, Robert L Spitzer, and Janet BW Williams. The phq-15: validity of a new measure for evaluating the severity of somatic symptoms. *Psychosom Med*, 64(2):258–266, 2002. https://doi.org/10.1097/00006842-200203000-00008.

[53] Gianluca Maguolo, Michelangelo Paci, Loris Nanni, and Ludovico Bonan. Audiogmenter: a matlab toolbox for audio data augmentation. *Applied Computing and Informatics*, 2021. https://doi.org/10.1108/ACI-03-2021-0064.

[54] Xiaoya Fu, Fengyu Zhang, Feng Liu, et al. Brain and somatization symptoms in psychiatric disorders. *Front Psychiatry*, 10:146, 2019. https://doi.org/10.3389/fpsyt.2019.00146.

[55] Weiwei Liu, Haobo Wang, Xiaobo Shen, et al. The emerging trends of multi-label learning. *IEEE Trans Pattern Anal Mach Intell*, 44(11):7955–7974, 2021. https://doi.org/10.1109/TPAMI.2021.3119334.

[56] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. Big self-supervised models are strong semi-supervised learners. *Advances in neural information processing systems*, 33:22243–22255, 2020.

[57] Hankook Lee, Sung Ju Hwang, and Jinwoo Shin. Self-supervised label augmentation via input transformations. In *International Conference on Machine Learning*, pages 5714–5724. PMLR, 2020. https://doi.org/10.1109/TPAMI.2021.3119334.

[58] Kun Qian, Zixing Zhang, Alice Baird, and Björn Schuller. Active learning for bird sound classification via a kernel-based extreme learning machine. *The Journal of the Acoustical Society of America*, 142(4):1796–1804, 2017. https://doi.org/10.1121/1.5004570.