# Detection of Depression in Speech

Zhenyu Liu, Bin Hu*, Lihua Yan, Tianyang Wang, Fei Liu, Xiaoyu Li, Huanyu Kang
Ubiquitous Awareness and Intelligent Solutions Lab, Lanzhou University
Lanzhou, China
{liuzhy12, bh, yanlh14, tywang2014, fliu14}@lzu.edu.cn, li461547885@163.com, kanghy11@lzu.edu.cn

***Abstract*—Depression is a common mental disorder and one of the main causes of disability worldwide. Lacking objective depressive disorder assessment methods is the key reason that many depressive patients can't be treated properly. Developments in affective sensing technology with a focus on acoustic features will potentially bring a change due to depressed patient's slow, hesitating, monotonous voice as remarkable characteristics. Our motivation is to find out a speech feature set to detect, evaluate and even predict depression. For these goals, we investigate a large sample of 300 subjects (100 depressed patients, 100 healthy controls and 100 high-risk people) through comparative analysis and follow-up study. For examining the correlation between depression and speech, we extract features as many as possible according to previous research to create a large voice feature set. Then we employ some feature selection methods to eliminate irrelevant, redundant and noisy features to form a compact subset. To measure effectiveness of this new subset, we test it on our dataset with 300 subjects using several common classifiers and 10-fold cross-validation. Since we are collecting data currently, we have no result to report yet.**

***Keywords—depression; speech; acoustic feature; feature selection***

## I. Introduction

Depression is characterized by sadness, loss of interest or pleasure, feelings of guilt or low self-worth, disturbed sleep or appetite, tiredness, and poor concentration. It is a common mental disorder and one of the main causes of disability worldwide，leading to a high impact on individuals, their families and society. The World Health Organization (WHO) estimated that about 400 million people of all ages suffer from this disease[1]. Different from usual mood fluctuations and short-lived emotional responses to challenges in everyday life, depression may become a serious health condition when long-lasting with moderate or severe intensity and even lead to suicide at its worst . According to a 2012 fact sheet released by the WHO, suicide caused by depression results 1 million deaths every year. Moreover, depression is estimated to become the second greatest disease burden in the world by the year 2020 [1].

Although there are effective treatments for depression, fewer than half of those affected in the world (in some countries, fewer than 10%) receive such treatments. Barriers to effective care include a lack of trained health care providers, social discrimination and inaccurate assessment. It is clear that the third element is the key point that we can make a breakthrough. Current depression assessment methods rely on patient self-report and professional interview. Self-report, like Self-Rating Depression Scale (SDS) [2] and Beck Depression Inventory (BDI) [3], risks a range of subjective biases. Similarly, professionals' evaluations vary depending on their expertise and the diagnostic methods used (e.g., Diagnostic and Statistical Manual of Mental Disorders (DSM-IV) [4], Hamilton Rating Scale for Depression (HAMD) [5], etc.). Currently, an objective and effective depression assessment approach for assistance diagnosis is needed.

Developments in affective sensing technology (e.g., facial expression, body gesture, speech, motion, eye movement, etc.) will potentially enable an objective depression evaluation method. Vocal affect refers to the emotional expression and the overall state of the speaker. Comparing with other affective sensing technology, the greatest advantage of speech is that the voice of depressed patients has remarkable characteristics concluded by previous studies: speaking in a low voice, slowly, hesitatingly, monotonously, sometimes stuttering, and whispering [6][7]. Meanwhile, voice signals can be collected easily by a non-invasive and portable instrument. Based on these two reasons above, we focus on depressed patients' speech analysis.

In our research, we mainly focus on following aspects: First, find out a speech feature set which can differentiate depressed people from non-depressed ones by comparison analysis. Second,  figure out a feature set can be evaluated depressive disorder severity and even used for depression prediction through a follow-up study. Third, based on these findings, implement an objective affective sensing system that supports clinicians in their diagnosis of clinical depression.

## II. Related work

Our research is based on a number of studies, which gathered a considerable amount of evidences that depression can be assessed by speech with different degrees [8]-[12]. Many researchers aimed at the correlations between depression and some particular speech features at early age [10][13][14]. This natural thought came from some observations of clinical depressive patients' voices. Then a lot of experiments have been conducted to reveal the relevance between depression and various acoustic features, like prosodic features (e.g., pitch, jitter, loudness, speaking rate, energy, pause time, intensity, etc.), spectral features (e.g., formants, energy spectrum density, spectral energy distribution, vocal tract spectrum, spectral noise, etc.) and cepstral features (e.g., Mel-Frequency Cepstral Coefficients (MFCC) and Linear Prediction Cepstrum Coefficient (LPCC)) and so on [10][12][14]-[17]. Most of these features are closely

978-1-4799-9953-8/15/$31.00 ©2015 IEEE

related to human cognition and have been widely studied for years. Not only that, some people turned to seek new voice features, like glottal waveform and Teager Energy Operation (TEO) which are presented in recent years and shown strong correlations to depression recognition [18][19]. Low et al. [8] and Mundt et al. [20] illustrated relation between depressive severity and some speech features. Alghowinem et al. summarized and compared different features on depression classification [16][17]. Moore II et al. [15][18] and Low et al. [7][19][21] proposed new feature sets with good classification performance on this disease. Many researchers believe that feature combination optimization may lead to progress of recognition accuracy.

Besides the correlation analysis of acoustic features and depression, some researchers paid attention to special depressive populations, experimental methods, data analysis methods and so on. Bandura et al. discussed children depressive disorder [22] and Low et al. mainly investigated adolescents clinical depression [8]. For illuminating speech changing with depressive disorder, Low et al. followed the group of adolescents patients for two years [8]. For the content of recording, Shankayi et al. suggested that scientific text is working better than emotional ones for depression recognition [23]. Alghowinem et al. figured out that spontaneous speech gives better results than reading [16]. Alghowinem et al. verified 'thin-slicing' approach on reading speech, which means smaller parts of the speech data will perform similarly if not better than using the whole speech data in the same paper [16]. Low et al. proposed that for both type of features (TEO and MFCC), the classification accuracy were higher for female speakers than males [21]. Cohn et al. tried speech and facial expression fusion to improve recognition system performance and showed a good result [24].

A lot of studies contributed to this topic, however, we still have a long way to go. In our research, we emphasize three key points: creating a large sample set (300 subjects), observing the effect of emotion arousal questioning and paying attention to high-risk people. The samples of subjects in previous studies are usually too small to generalize a conclusion with statistical meaning. For instance, Moore et al. interviewed 33 subjects (15 depressive patients, 18 healthy controls) [18]. Alghowinem et al. recruited 40 depressed subjects and 40 healthy controls [16] and Low et al. studied 139 adolescents (68 depressed, 71 healthy) [21]. In our experiment, positive, negative and neutral emotion arousal questions are set for finding the differences among them. In the past, most researchers compared depressive patients with healthy controls, while we add a high-risk of depressed group [25] for follow-up study in our work. We expect to find some speech features which could be used for depressive severity evaluation and prediction by following all three groups participants for several years.

In the remainder of the paper, Section 3 describes the methodology, including participants, interview and observation procedures, data collection, pre-processing and feature extraction, feature selection, classification. Section 4 presents contributions to Affective Computing.

## III. Method

We recruit 300 subjects (100 depressed patients, 100 healthy controls and 100 high-risk people) for this study lasting one year or more. We will focus on four points:

1) To find out an effective acoustic feature set for depression detection through comparative analysis on three subject groups by feature selection methods;

2) To figure out whether depressive disorder severity can be assessed and depression can be predicted by some speech features through a follow-up study for several years;

3) To discover which interview manner is better for depression recognition, emotion arousal questioning or the other way around;

4) To implement an actual auxiliary diagnosis system based on these results above.

### *A. Participants*

Three groups subjects, including depressive patients, healthy controls subjects and high-risk people, are involved for experimental validation. Before experiment, each subject (age range 18-55 years old, both male and female, Chinese native speaker, at least received primary school education) was asked to fill in a pre-assessment booklet (general information like healthy history and demographic information questionnaire contained age, gender, educational level, job and so on), then assessed by psychiatrists following Diagnostic and Statistical Manual of Mental Disorders (DSM-IV) diagnosis rules.

More details are described as follows:

1) Depressive patients: outpatients or inpatients diagnosed with depression and recommended by psychiatrists;

2) Healthy controls: persons with no mental disorder past or current and matching the depressive group in items on demographics;

3) High-risk people: persons whose first-degree relative was diagnosed with depression or whose child was disabled or diagnosed with intellectual disability and matching the depressive group in items on demographics.

There are some exclusion criteria for subjects:

1) Psychotic disorder in the past or current;

2) Severe somatic disease;

3) Alcohol or drug abusers;

4) Suicide attempters;

5) Pregnant or lactating woman.

### *B. Interview and Observational Procedures*

*1) Interview:* This task contained 18 questions with positive, neutral and negative meanings. These topics came from DSM-IV and some depression scales which are often used in this field. For examples: If you have a vacation, please describe your travel plans [21]. What is your best gift you have ever received and how did you feel [26]? Please describe one of your friends, including age, job, characters and hobbies.

978-1-4799-9953-8/15/$31.00 ©2015 IEEE

How do you evaluate yourself? What would you like to do when you are unable to fall asleep? What makes you desperate?

*2) Reading:* This part consists of a short story named “The North Wind and the Sun”, which is from the booklet “The Principles of the International Phonetic Association” [11]. and often used in acoustic analysis in international, multilingual clinical research, and three groups words with positive (e.g., outstanding, happy), neutral (e.g., center, since) and negative (e.g., depression, wail) emotion. Positive and negative words are selected from affective ontology corpus created by Hongfei Lin [27], and neutral ones are picked out from Chinese affective words extremum table [28]. All these words are often-used words in Chinese to avoid the impact of educational level and three groups words have close stroke number. Subjects are told to read story and these words in their common ways.

*3) Picture description:* The materials for this task include four pictures in all. Three pictures, which express positive, neutral and negative faces, are selected from Chinese Facial Affective Picture System (CFAPS) and the last one with a “crying woman” came from Thematic Apperception Test (TAT) [26]. TAT is created by Murray in 1935, which is used in psychological counseling and psychotherapy at present. In this task, subjects are told to describe these four pictures freely.

### *C. Data Collection*

The laboratory for this experiment is a clean, quiet, soundproof room without electromagnetic interference. Each subject is invited to complete all three experimental tasks on a comfortable chair. During the course of experiment, it is necessary that the ambient noise of lab is less than 60dB. Meanwhile, participants can not touch the microphone and should keep the distance between mouth and microphone less than 25cm. The audio signals are picked up via a NEUMANN TLM102 microphone and recorded by a RME FIREFACE UCX audio card with 44.1KHz sampling rate and 24-bit sampling depth. All recording data are saved as uncompressed WAV format. Ambient noise signals are measured before experiment starting in case influences on subjects’ audio signals.

### *D. Pre-processing and Feature extraction*

All recordings are segmented and labeled manually. Only subjects’ voice signals are reserved for analysis. Pre-processing mainly consists of filtering, framing, windowing and sometimes endpoint detection for some particular feature extraction. Each frame is 25ms length with 50% overlap. Voice features could be divided into two categories: acoustic and linguistic features [21][29]. Compared to acoustic features, linguistic features are usually influenced by subjective consciousness and languages. The linguistic features will not be analyzed since we are aiming at general characteristics for depressed speech regardless of the language used. Acoustic features could also be categorized into two branches: low-level descriptors (LLD), which usually are calculated frame-by-frame, and statistical features, which could be calculated based on the LLD over certain units (e.g. words, syllables, sentences, etc.) [16]. Several software tools are available for extracting sound features for free. We used the open-source software ‘openSMILE’ [30] to extract speech features (e.g., pitch, jitter, shimmer, energy, Mel-Frequency Cepstral Coefficients (MFCC), loudness, intensity) and supplement it with Glottal Waveform, Teager Energy Operation (TEO), Harmonic-to-Noise Ratio (HNR), Signal-to-Noise Ratio (SNR), spectral balance, Pitch Period Perturbation Quotient (PPQ), Amplitude Perturbation Quotient (APQ), peaking rate and so on. It has been verified that most of these features are useful to depression classification.

### *E. Feature Selection*

Feature selection is essential to us for several reasons:

1) To eliminate irrelevant, redundant or noisy features for a more compact feature set, which is easy to explain;

2) To reduce the dimensionality of data for computing efficiency improvement;

3) To gain a more effective expression (with less features), which is convenient to feature fusion and new feature discovery.

Feature selection is a very critical step in pattern recognition to cope with the curse of dimensionality. It improves prediction performance and reduces computation cost [31]. This requires an efficient search algorithm to find a candidate subset and a suitable criterion function to evaluate the candidate subset. With methodologies applied typically, the problem of feature selection is grouped into three main categories: filters [32], wrappers [33], and embedded solutions [34]. The filter approach utilized the data alone to decide which feature should be kept. In general, the filter approach usually has an efficient searching strategy with a result tradeoff. One popular algorithm of this class is Relief [35]. The wrapper approach is conditioned by the usefulness of the subset. The wrapper approach may lead to a better performance compared to a filter approach. However, it may show a bias that performing better with given classifier but worse with others, because the process is adjusted to specific characteristic of the chosen classifier. Meanwhile, it has a relative high computational cost. Assessing a subset by its classification rate is a representative evaluation of the wrapper approach. The embedded solution requires a predictor has its own inherent mechanism dedicated to feature selection. It is more efficient that adding feature selection in the training process instead of retraining. For instance, Decision Trees have a built-in mechanism to perform feature selection [36, 37]. In general, most people select features by combining these solutions for achieving better performance.

In our experiment, we utilize a two-stage feature selection method by combining the minimal-redundancy-maximal-relevance (mRMR) criterion [38] as the filter approach and the Sequential Forward Floating Selection (SFFS) algorithm [39] as the search algorithm of the wrapper approach. The mRMR algorithm is very efficient and the SFFS algorithm has good performance for classification. This can help us to find a compact feature set efficiently.

Here are the details about our two-stage feature selection. On the first stage, mRMR algorithm tries to find a candidate feature subset which has maximum mutual information with labels, as well as minimum relevance in terms of mutual information among the chosen features. And we employ sequential forward selection (SFS) [40] as the search algorithm. In this process, we followed a 10-fold feature selection procedure: in each fold 90% of the samples are used to rank the 1745 features, and the remaining samples are used to evaluate the performance of the feature subset; then we choose the top k features as a candidate subset in each fold, k value is the minimum dimension of the subset with best performance in classification accuracy; at last, we gather the 10 subsets together as the candidate subset and eliminated the repeating features. On the second stage, SFFS algorithm has proven to be superior to the SFS algorithm in many comparisons [41]. The primary idea of this algorithm is that the algorithm involves a backtracking phase in chosen feature set after selecting a feature to cope with the "nesting effect" problem. This backtracking is carried on until no better subset is found, then the SFFS algorithm goes back to seek next feature until the corresponding subset has already been determined. This strategy may help us to find a more compact feature subset. In this paper, the SFFS algorithm is carried on the candidate subset to select a more compact subset to differentiate depressed people from non-depressed people.

### *F. Classfication*

We employ the k nearest neighbors (k-NN) prediction rules [42] and Support Vector Machine (SVM) [43] to assess feature set chosen above. In this experiment, we employed the 10-fold cross validation scheme in testing and applied the original data as the inputs for classifiers without discretization.

## IV. CONTRIBUTIONS TO AFFECTIVE COMPUTING

We aim to detect depression degree via speech is a part of affective computing. The samples studied by us can enrich the affective computing library and provide more experimental evidence. It is most important that all the results may lead to an efficient actual computer-assisted depression diagnosis system.

## ACKNOWLEDGMENT

This work was supported by the National Basic Research Program of China (973 Program) (No.2014CB744600, No.2011CB711000), the Program of International S&T Cooperation of MOST (No.2013DFA11140), the National Natural Science Foundation of China (grant No.61210010, No.61300231).

## REFERENCES

[1] http://www.who.int/mediacentre/factsheets/fs396/en/.

[2] W. W. Zung, "A self-rating depression scale," *Archives of general psychiatry,* vol. 12, pp. 63-70, 1965.

[3] A. T. Beck, R. A. Steer, and G. K. Brown, "Beck depression inventory-II," *San Antonio,* 1996.

[4] A. American Psychiatric Association and A. P. Association, "Diagnostic and statistical manual of mental disorders," 1980.

[5] M. Hamilton, "A rating scale for depression," *Journal of neurology, neurosurgery, and psychiatry,* vol. 23, p. 56, 1960.

[6] E. Kraepelin, "Manic depressive insanity and paranoia," *The Journal of Nervous and Mental Disease,* vol. 53, p. 350, 1921.

[7] L.-S. A. Low, N. C. Maddage, M. Lech, L. Sheeber, and N. Allen, "Influence of acoustic low-level descriptors in the detection of clinical depression in adolescents," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, 2010, pp. 5154-5157.

[8] K. E. B. Ooi, L.-S. A. Low, M. Lech, and N. Allen, "Early prediction of major depression in adolescents using glottal wave characteristics and teager energy parameters," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, 2012, pp. 4613-4616.

[9] M. Alpert, E. R. Pouget, and R. R. Silva, "Reflections of depression in acoustic measures of the patient's speech," *Journal of affective disorders,* vol. 66, pp. 59-69, 2001.

[10] Å. Nilsonne, J. Sundberg, S. Ternström, and A. Askenfelt, "Measuring the rate of change of voice fundamental frequency in fluent speech during mental depression," *The Journal of the Acoustical Society of America,* vol. 83, pp. 716-728, 1988.

[11] D. J. France, R. G. Shiavi, S. Silverman, M. Silverman, and D. M. Wilkes, "Acoustical properties of speech as indicators of depression and suicidal risk," *Biomedical Engineering, IEEE Transactions on,* vol. 47, pp. 829-837, 2000.

[12] N. Cummins, J. Epps, M. Breakspear, and R. Goecke, "An Investigation of Depressed Speech Detection: Features and Normalization," in *Interspeech*, 2011, pp. 2997-3000.

[13] E. Scripture, "A study of emotions by speech transcription," *Vox,* vol. 31, pp. 179-183, 1921.

[14] A. J. Flint, S. E. Black, I. Campbell-Taylor, G. F. Gailey, and C. Levinton, "Abnormal speech articulation, psychomotor retardation, and subcortical dysfunction in major depression," *Journal of psychiatric research,* vol. 27, pp. 309-319, 1993.

[15] E. Moore, M. Clements, J. W. Peifer, and L. Weisser, "Critical analysis of the impact of glottal features in the classification of clinical depression in speech," *Biomedical Engineering, IEEE Transactions on,* vol. 55, pp. 96-107, 2008.

[16] S. Alghowinem, R. Goecke, M. Wagner, J. Epps, M. Breakspear, and G. Parker, "Detecting depression: a comparison between spontaneous and read speech," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, 2013, pp. 7547-7551.

[17] S. Alghowinem, R. Goecke, M. Wagner, J. Epps, T. Gedeon, M. Breakspear, *et al.*, "A comparative study of different classifiers for detecting depression from

spontaneous speech," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, 2013, pp. 8022-8026.

[18] E. Moore, M. Clements, J. Peifer, and L. Weisser, "Investigating the role of glottal features in classifying clinical depression," in *Engineering in Medicine and Biology Society, 2003. Proceedings of the 25th Annual International Conference of the IEEE*, 2003, pp. 2849-2852.

[19] L.-S. A. Low, N. C. Maddage, M. Lech, L. B. Sheeber, and N. B. Allen, "Detection of clinical depression in adolescents' speech during family interactions," *Biomedical Engineering, IEEE Transactions on,* vol. 58, pp. 574-586, 2011.

[20] 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," *J Neurolinguistics,* vol. 20, pp. 50-64, Jan 2007.

[21] L.-S. A. Low, N. C. Maddage, M. Lech, L. Sheeber, and N. Allen, "Content based clinical depression detection in adolescents," in *Signal Processing Conference, 2009 17th European*, 2009, pp. 2362-2366.

[22] A. Bandura, C. Pastorelli, C. Barbaranelli, and G. V. Caprara, "Self-efficacy pathways to childhood depression," *Journal of Personality and social Psychology,* vol. 76, p. 258, 1999.

[23] R. Shankayi, M. Vali, M. Salimi, and M. Malekshahi, "Identifying depressed from healthy cases using speech processing," in *Biomedical Engineering (ICBME), 2012 19th Iranian Conference of*, 2012, pp. 242-245.

[24] J. F. Cohn, T. S. Kruez, I. Matthews, Y. Yang, M. H. Nguyen, M. T. Padilla*, et al.*, "Detecting depression from facial actions and vocal prosody," in *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on*, 2009, pp. 1-7.

[25] I. M. Goodyer, A. Tamplin, J. Herbert, and P. Altham, "Recent life events, cortisol, dehydroepiandrosterone and the onset of major depression in high-risk adolescents," *The British Journal of Psychiatry,* vol. 177, pp. 499-504, 2000.

[26] F. Hönig, A. Batliner, E. Nöth, S. Schnieder, and J. Krajewski, "Automatic Modelling of Depressed Speech: Relevant Features and Relevance of Gender," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[27] http://ir.dlut.edu.cn/Group.aspx?ID=4.

[28] http://www.datatang.com/data/43216.

[29] G. Zhou, J. H. Hansen, and J. F. Kaiser, "Nonlinear feature based classification of speech under stress," *Speech and Audio Processing, IEEE Transactions on,* vol. 9, pp. 201-216, 2001.

[30] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the international conference on Multimedia*, 2010, pp. 1459-1462.

[31] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *The Journal of Machine Learning Research,* vol. 3, pp. 1157-1182, 2003.

[32] J. Jelonek, K. Krawiec, and J. Stefanowski, "Comparative study of feature subset selection techniques for machine learning tasks," *Proceedings of International Symposium Intelligent Information Systems,* 1998.

[33] G. H. John, R. Kohavi, and K. Pfleger, "Irrelevant Features and the Subset Selection Problem," *Machine Learning Proceedings,* pp. 121–129, 1994.

[34] U. Stańczyk and L. C. Jain, "Feature Selection for Data and Pattern Recognition: An Introduction," *Studies in Computational Intelligence,* pp. 1-7, 2015.

[35] Y. Sun and D. Wu, "A RELIEF Based Feature Extraction Algorithm," in *SDM*, 2008, pp. 188-195.

[36] L. Breiman and Friedman, "JH, Olshen RA, et al. Classification and Regression Trees," *Wiley Interdisciplinary Reviews Data Mining & Knowledge Discovery,* vol. 1, pp. 14–23, 1984.

[37] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of Machine Learning Research,* vol. 3, pp. 1157--1182, 2003.

[38] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* vol. 27, pp. 1226-1238, 2005.

[39] P. Pudil, J. Novovičová, and J. Kittler, "Floating search methods in feature selection," *Pattern recognition letters,* vol. 15, pp. 1119-1125, 1994.

[40] N. A. S. I. o. P. Recognition, S. Processing, and C. H. Chen, "Pattern recognition and signal processing," *Nato Advanced Study Institutes,* vol. 93, pp. 231-244, 1978.

[41] J. Reunanen, "Overfitting in making comparisons between variable selection methods," *The Journal of Machine Learning Research,* vol. 3, pp. 1371-1382, 2003.

[42] R. J. Schalkoff, "Pattern recognition : statistical, structural, and neural approaches," *Pattern Recognition Statistical Structural & Neural Approaches,* 1992.

[43] T. M. Mitchell, "Machine learning. WCB," ed: McGraw-Hill Boston, MA:, 1997.