Comparison of read and spontaneous speech in case of Automatic Detection of Depression

Gábor Kiss, Klára Vicsi
Department of Telecommunications and Media Informatics
Budapest University of Technology and Economics
Budapest, Hungary
kiss.gabor@tmit.bme.hu, vicsi@tmit.bme.hu

Abstract*—In this paper, read and spontaneous speech have been compared in the light of automatic depression detection by speech processing. First, statistical analysis was carried out to select those acoustic features that differ significantly between healthy and depressed subjects in case of these two types of speech, separately for both gender. Secondly, statistical examination and classification experiments were prepared to compare the values of the selected features for the two types of speech. We were looking for the answer to which type of speech can be used to achieve better automatic depression detection results. As it was expected, the tempo related features, such as articulation rate, speech rate, and pause lengths are useful in case of spontaneous speech, while formants trajectories can be used only in case of read speech, because their values are mainly influenced by the linguistic content of the speech. Despite the significant differences of the features' values between read and spontaneous speech, there were no major differences in the detection accuracies. 83% detection accuracy was archived with read speech samples, and 86%detection accuracy was achieved with spontaneous speech samples.

Keywords*—depression, classification, speech type comparison, read speech, spontaneous speech, SVM

I. INTRODUCTION

Non-verbal information processing of speech for creation of various IT tools, is an essential area for cognitive infocommunication [1], [2]. Particularly interesting field is the usage of speech as bio-signal and development of different non-invasive diagnostic tools that make automatic assessment of the cognitive and psychological state of the people. Automatic depression detection based on speech processing belongs to this research field. Previous studies indicated that speech can be a promising indicator of depression, and there already are some good results either with depression classification [3], [4], [5], or predicting the severity of depression with regression method [6], [7]. However, the type of used speech material for depression detection/regression is still an open question.

Two types of speech are most commonly used in case of depression detection: read speech or spontaneous speech from monologues or dialogues [8], [9], [10], [11]. Both types have its own advantages and disadvantages. The read speech is easier to preprocess automatically. The content and length of the speech are almost the same, only some misreading or mispronunciation can occur. Because spontaneous speech may reflect the emotional state of the speaker better, using it as the input of a classifier/predictor can provide more accurate depression classification/prediction [8], [9], [10], [11].

In this paper, the differences between read and spontaneous speech will be investigated in the light of depression detection in case of Hungarian language. Although, there are many results in general comparison of read and spontaneous speech, we were focusing strictly how this difference influences the depression detection accuracy, which has not yet been examined in Hungarian language so far. Our earlier results proofed that, using phoneme level segmentation during the feature extraction and using separate models for females and males increase the classification accuracy [12], thus in this investigation separate model will be used for female and male.

The paper is structured as follows. In Section 2, the description of the used database is given. Section 3 describes what methods were used during the investigation. Section 4 contains the results of our statistical analysis and our classification experiments. Section 5 gives some conclusions.

II. DATABASE

Speech samples were collected from healthy and depressed subjects in quiet environment with head microphone. The recordings were recorded at 44,1 kHz with 16-bit sample rate. Two types of speech sample were recorded from each subject, read speech: a short folk tale "The North Wind and the Sun" and spontaneous speech: dialogue between the examined subject and interviewer, both in Hungarian language. A total of 73 subjects were recorded, 42 females and 31 males. From the 73 subject 48 subjects were diagnosed with depression, 30 females and 18 males, and 25 subjects were healthy, 12 females and 13 males (Fig 1.).

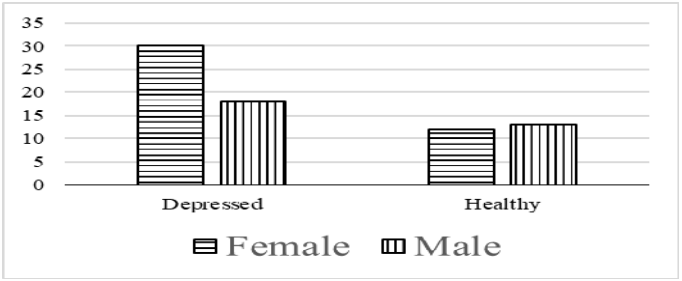

Fig. 1. Distribution of the subjects in the examined database

The research was supported by European Space Agency COALA project: Psychological Status Monitoring by Computerized Analysis of Language phenomena (COALA) (AO-11-Concordia).

III. METHODS

A. Preprocessing and Segmentation

The pressure amplitude of wave form was normalized to peak, and was segmented and annotated on phoneme level using Sampa character set. The annotation of the read speech samples was done with forced alignment method [13]. In case of dialogues, the speech parts of the examined subject were annotated and segmented only with a speech recognition tool [14], then the output of the speech recognizer (annotation and segmentation) were manually checked and corrected.

B. Feature Extraction

Those low-level descriptors (LLD) were extracted from the speech samples that were considered as good indicators of the depressed state based on our previous works [3], [15] and on the literature [8], [16]. The following 37 LLD features were measured: fundamental frequency (f0), intensity, formant frequencies and their bandwidths (F1, F2, F3, B1, B2, B3), jitter, shimmer, Mel-filtered band energies (27 band). These LLD features were extracted in two different ways: with 25 ms frame size and 10 ms time step on the whole voiced parts of the speech sample and with 25 ms frame size measured only at the middle of the same vowel "E" (in SAMPA). Vowel "E" was selected because it was the most common vowel in the speech samples. Functional features (mean, variance and percentile range) were derived from these two groups of LLD features with three statistical functions: mean, variance and percentile range (5-95%); the mean value of f0 was not calculated because the mean value of f0 is largely speaker dependent. The functional feature set was expanded with the articulation rate, speech rate and the ratio of pauses and speech (rpl) [16]. This functional feature set was calculated both on the read speech samples and the spontaneous speech samples.

C. Statistical Analysis and Feature Selection

The features showed normal distribution by Kolmogorov-Smirnov test on 95% significant level.

Two type of statistical analysis were conducted. For feature selection independent samples t-test was used by analyzing which functional features show significant differences between healthy and depressed subjects. This examination was performed both on the read and on the spontaneous feature set separately for both gender.

Paired samples t-test was used to analyze which functional features show significant difference between the read and the spontaneous speech. Those functional features were analyzed which showed significant differences between healthy and depressed subjects in both read and spontaneous speech at least in one gender or which showed significant differences in both gender at least in one speech type.

D. Classification and Evaluation

For classification, Support Vector Machine (SVM) was used [17], which was implemented in LIBSVM [18]. For kernel function linear kernel was selected with the default value of the hyper parameter (cost:1) to avoid the overfitting. Separate models were trained for read and spontaneous speech separately for both genders, thus four classification scenarios were performed. The feature vector for each model was derived from the result of the independent samples t-test. The values of each selected functional feature in the feature vector were normalized between -1 and 1.

Leave one out cross validation was used for training and testing without any overlap between the training and testing samples. This testing method is beneficial with relatively small sized datasets. To evaluate the classification experiments, the accuracy value (acc) of the classification.

IV. RESULTS

A. Results of Statistical Analysis

The comparison of the features for healthy and depressed subjects by independent samples t test, can be seen in Table 1. Only those functional features are shown that differ significantly at 95% significance level between healthy and depressed subjects at least in one case of speech type.

The comparison of the features for read and spontaneous speech type separately for healthy and depressed subjects by paired samples t-test can be seen in Table 3.

In each table, the "Meas. Location" column refers to the location of the feature measurement.

The "Direction of Difference" column in Table 1 refer how the value of the examined feature differed with depression, and the "Significant Speech Type" column shows the type of speech in which the significant difference was measured.

The "Difference in Healthy/Depressed Subjects" columns in Table 2 show the difference between the mean values of the examined functional features in the case of read and spontaneous speech and the confidence interval of the difference at significance level 95% in the bracket. Features that have shown significant differences are shown in bold, and the degree of significance was marked with "*", if it was above 95%, with "**", if above it was 99%, and with "***", if it was above 99.9%.

In Table 1, the direction of difference can be seen in depression-specific parameters. Lower dynamic of fundamental frequency and intensity, lower mean value of first formant frequency measured on vowel "E", and higher energy values in case of low Mel-filtered bands measured on vowel "E" are characteristic for depression regardless of the type of speech:

TABLE I. COMPRAISON OF THE MEAN VALUEOS OF ACOUSTIC FEATURES FOR HEALTHY AND DEPRESSED SUBJECTS SEPERATELY FOR READ AND SPONTANEOUS SPEECH

Feature Name	Meas. Location	Direction of Difference	Significant Speech Type
Percentile range of f0	voiced parts	↓	both
Mean of F1	vowel "E"	↓	both
Mean of F1	voiced parts	↑	read
Mean of F2	voiced parts	↑	read
Variance of F2	voiced parts	↑	read
Mean of F3	voiced parts	↑	read
Mean of B1	vowel "E"	↑	both
Variance of B1	voiced parts	↑	read
Percentile range of B1	voiced parts	↑	read
Mean of B2	vowel "E"	↑	spontaneous
Mean of B3	vowel "E"	↑	spontaneous
Variance of intensity	voiced parts	↓	both
Percentile range of intensity	voiced parts	↓	both
Speech rate	whole speech	↓	spontaneous
Rpl	whole speech	↑	spontaneous
Mean of MelFiler1	whole speech	↑	both
Mean of MelFiler2	whole speech	↑	both
Mean of MelFiler3	whole speech	↑	both

TABLE II. COMPRAISION OF THE MEAN VALUES OF ACOUSTIC FEATURES FOR READ AND SPONTANEOUS SPEECH SEPERATELY FOR HEALTHY AND DEPRESSED SUBJECTS

Feature name	Meas. Location	Difference in Healthy Subjects	Difference in Depressed Subjects
Percentile range of f0	voiced parts	8,2 (±13,8)	0,25 (±7,85)
Mean of F1	vowel "E"	3,9 (±10,5)	**15,5 ** (±10,0)**
Mean of F1	voiced parts	**-14,0 * (±13,9)**	**42,1 *** (±15,0)**
Mean of F2	voiced parts	**65,7 ** (±28,3)**	**36,8 ** (±27,3)**
Variance of F2	voiced parts	**25,1 ** (±18,2)**	**-52,7 *** (±17,0)**
Mean of F3	voiced parts	**59,5 *** (±34,1)**	**-71,6 *** (±33,4)**
Mean of B1	vowel "E"	5,6 (±38,6)	**-58,9 *** (±18,7)**
Variance of B1	voiced parts	3,9 (±19,1)	**-21,4 * (±17,4)**
Percentile range of B1	voiced parts	-14,0 (±62,2)	**-152,9 *** (±34,7)**
Mean of B2	vowel "E"	**80,6 *** (±69,0)	-29,1 (±46,5)
Mean of B3	vowel "E"	-8,4 (±56,2)	**-110,0 *** (±61,1)**
Variance of intensity	whole speech	0,89 (±0,22)	**0,18 *** (±0,35)**
Percentile range of intensity	whole speech	-0,15 (±1,05)	**2,12 *** (±0,74)**
Speech rate	whole speech	**1,21 *** (±0,54)**	**3,82 *** (±0,66)**
Rpl	whole speech	**-0,01 *** (±0,03)**	**-0,33 *** (±0,05)**
Mean of MelFiler1	vowel "E"	**1,8 * (±1,3)**	**2,8 *** (±1,2)**
Mean of MelFiler2	vowel "E"	**2,6 ** (±1,8)**	**3,5 *** (±1,2)**
Mean of MelFiler3	vowel "E"	**3,0 ** (±1,8)**	**4,0 *** (±1,2)**

In Table 2, we can see that almost all the examined features differ significantly between read and spontaneous speech. Those features that show significant difference in case of healthy subjects are content and speech type dependent. However, those features that show no significant difference in case of healthy subjects but show significant difference in case of depressed subjects indicate that the depressed state influences some of the acoustic features differently depending on the speech type.

One of the most promising feature for depression detection, the variability of f0 shows no significant differences at all between read and spontaneous speech. The other very promising feature for depression detection, the variability of the intensity shows no significant difference in case of healthy subjects, but show significant difference in case of depressed subjects. Because the lower variability of intensity is typical of depressed speech, the spontaneous speech seems more usable in case of this feature.

B. Results of Classifications

The confusion matrices of the classification experiments can be seen in Table 3 in case of read speech and in Table 4 in case of spontaneous speech. The summarized results of the four classification experiments can be seen in Table 5.

TABLE III. CONFUSION MATRIX OF THE CLASSIFICATION EXPERIMENT IN CASE OF READ SPEECH

	Healthy Classified	Depressed Classified
Healthy	**71%** **female: 66,7%** **male: 76,9%**	29% female: 33,3% male: 23,1%
Depressed	11,7% female: 12,5% male: 10,5%	**88,3%** **female: 87,5%** **male: 89,5%**

In case of read speech, the overall accuracy was 83%, which is comparable to the similar results in the literature (70-90%) [8][9][10][11][16]. The achieved accuracy was slightly better for male subjects 84% than for female subjects 82% (Tables 3).

TABLE IV. CONFUSION MATRIX OF THE CLASSIFICATION EXPERIMENT IN CASE OF SPONTANEOUS SPEECH

	Healthy Classified	Depressed Classified
Healthy	**77,6%** **female: 66,7%** **male: 92,3%**	22,4% female: 33,3% male: 7,7%
Depressed	11,7% female: 12,5% male: 10,5%	**88,3%** **female: 87,5%** **male: 89,5%**

In case of spontaneous speech, the overall accuracy was even better (86%). The achieved accuracy was better for male subjects 91% than for female subjects 82% (Tables 4).

TABLE V. SUMMERIZED RESULTS OF THE FOUR CLASSIFICATION EXPERIMENTS

	Accuracy (female/male)
Read	**83%** female: 82% male: 84%
Spontaneous	**86%** female: 82% male: 91%

From the summarized results (Table 5), it can be seen that there are only slight differences in the accuracy between the results of read and spontaneous classification experiments. Using the spontaneous speech, a relative improvement of 3.6% was observed with respect to using the read speech. This difference is not significant, especially in the light of the database size. However, this result fits in with the other results on different languages in the literature [8][9][10][11]. It is important to note, that while there were significant differences in the statistical analysis between the two types of speech, this was no longer observable in the results of the classification experiments.

V. CONCLUSIONS AND FUTURE WORK

The way of the pronunciation of read speech and spontaneous speech different, thus the two type of speech, differ from each other phonetically and phonologically [19][20]. The consequence is that the characteristic speech parameters are also differ. Therefore, it is important to examine how these differences influence the detection of depression by speech processing.

In this paper, two types of speech were compared: read and spontaneous.

From the results, it can be clearly stated that some features differ significantly between healthy and depressed state only in case of one type of speech. For example, the tempo related features, such as articulation rate, speech rate, and pause lengths are useful in case of spontaneous speech, while formants trajectories can be used only in case of read speech, because their values are mainly influenced by the content of the speech.

It was very interesting that some features showed no significant differences between read and spontaneous speech in case of healthy subjects, but showed significant differences in case of depressed subjects. For example, the bandwidths of the formant frequencies and the variability of the intensity. This phenomenon would require further investigation and suggests that the speech production of depressed people is fundamentally altered in reading and spontaneous speech according to some acoustic features.

However, one of the most well-known indicator of depressed speech, the variability of fundamental frequency, showed no significant difference between the read and spontaneous speech at all.

Despite the significant differences of the feature values between read and spontaneous speech, there were no major differences in the detection accuracies. 83% detection accuracy was achieved with the read speech samples, and 86% accuracy was achieved with the spontaneous speech samples, which is 3.6% relative difference, but both results are promising enough compared to the literature [8][9][10][11][16]. Thus, both speech types can be used for a well-designed depression detection tool. However, it is important to note, that the preprocessing of the spontaneous speech samples was not fully automatic and the recording of this type of speech is harder, thus we believe that read speech can be the better choice as an input of an automatic depression detection tool.

ACKNOWLEDGMENT

The research was supported by European Space Agency COALA project: Psychological Status Monitoring by

Computerized Analysis of Language phenomena (COALA) (AO-11-Concordia).

REFERENCES

[1] P. Baranyi and A. Csapo, "Definition and synergies of cognitive infocommunications," in Acta Polytechnica Hungarica , vol. 9, no. 1, pp. 67–83, 2012.

[2] P. Baranyi, A. Csapo, and G. Sallai, "Cognitive Infocommunications (CogInfoCom),"Springer International, 2015.

[3] Kiss, G., Tulics, M. G., Sztahó, D., Esposito, A., & Vicsi, K., "Language independent detection possibilities of depression by speech," Recent Advances in Nonlinear Speech Processing, pp. 103-114. Springer International Publishing., 2016

[4] Alghowinem, S., Goecke, R., Epps, J., Wagner, M. and Cohn, J., "Cross-Cultural Depression Recognition from Vocal Biomarkers," Interspeech 2016, pp. 1943-1947, 2016

[5] Liu, Z., Hu, B., Yan, L., Wang, T., Liu, F., Li, X., & Kang, H., "Detection of depression in speech," Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on, IEEE, pp. 743-747, 2015

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

[7] Cummins, N., Sethu, V., Epps, J., Schnieder, S., & Krajewski, J., "Analysis of acoustic space variability in speech affected by depression," Speech Communication, 75:27-49, 2015

[8] Alghowinem, S., Goecke, R., Wagner, M., Epps, J., Breakspear, M., Parker, G.: "Detecting Depression – A Comparison Between Spontaneous and Read Speech", 38th International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2013

[9] Gupta, R., Malandrakis, N., Xiao, B., Guha, T., Van Segbroeck, M., Black, M., Potamianos, A., Narayanan, S.: "Multimodal prediction of affective dimensions and depression in human-computer interactions", Proceedings of the 4th ACM International Workshop on Audio/Visual Emotion Challenge (AVEC '14). ACM, Orlando, Florida, USA, 2014, pp. 33-40

[10] Pérez, H., Escalente, H.J. Villasenor-Pineda, L., Montesy-y-Gómez, M., Pinto-Avedano, D., Reyes-Meza, V.: "Fusing affective dimensions and audio-visual features from segmented video for depression recognition interactions", Proceedings of the 4th ACM International Workshop on Audio/Visual Emotion Challenge (AVEC '14). ACM, Orlando, Florida, USA, 2014, pp. 49-55

[11] Sidorov, M., Minker, W.: "Emotion recognition and depression diagnosis by acoustic and visual features: a multimodal approach interactions", Proceedings of the 4th ACM International Workshop on Audio/Visual Emotion Challenge (AVEC '14). ACM, Orlando, Florida, USA, 2014, pp. 81-86

[12] Kiss, G., Simin, L., Vicsi, K., "Estimation of the severity of depression based on speech processing on Hungarian langugae (original title: Depresszió súlyosságának becslése beszédjel alapján magyar nyelven)," XIII. Magyar Számítógépes Nyelvészeti Konferencia, (MSZNY2017). Conference, pp. 125-135, 2017

[13] Kiss, G., Sztahó, D., & Vicsi, K., "Language independent automatic speech segmentation into phoneme-like units on the base of acoustic distinctive features," Cognitive Infocommunications (CogInfoCom), 2013 IEEE 4th International Conference on, IEEE, pp. 579-582, 2013

[14] B. Tarján, G. Sárosi, T. Fegyo and P. Mihajlik. "Improved recognition of Hungarian call center conversations." In The 7th International Conference on Speech Technology and Human-Computer Dialogue (SpeD 2013), pp. 1-6.

[15] Kiss, G., & Vicsi, K., "Physiological and cognitive status monitoring on the base of acoustic-phonetic speech parameters," International Conference on Statistical Language and Speech Processing, pp. 120-131. Springer International Publishing, 2014

[16] Cummins, N., Scherer, S., Krajewski, J., Schnieder, S., Epps, J., & Quatieri, T. F., "A review of depression and suicide risk assessment using speech analysis," Speech Communication, 71:10-49, 2015

[17] Cortes, C., & Vapnik, V., "Support-vector networks," Machine learning, 20(3): 273-297, 1995

[18] Chang, C. C., & Lin, C. J. (2011). "LIBSVM: a library for support vector machines", ACM Transactions on Intelligent Systems and Technology (TIST), 2(3):27, 2011

[19] Gósy M.: "Phonetics, the speech science" (original title: Fonetika, a beszéd tudománya). Osiris; 2004.

[20] Markó, A., Gráczi, T. and Bóna, J.,: "The realisation of voicing assimilation rules in Hungarian spontaneous and read speech: Case studies.", Acta Linguistica Hungarica, 57(2-3), pp.210-238, 2010