Automatic Separation of Various Disease Types by Correlation Structure of Time Shifted Speech Features

Dávid Sztahó, Gábor Kiss, Miklós Gábriel Tulics, and Klára Vicsi
Department of Telecommunications and Media Informatics
Budapest University of Technology and Economics
Budapest, Hungary
sztaho@tmit.bme.hu, kiss.gabor@tmit.bme.hu, tulics@tmit.bme.hu, vicsi@tmit.bme.hu

***Abstract*—Special disease types may affect the complex mechanisms of speech production in different ways, causing various speech disorders. This is the reason why extraction of biomarkers from speech could be reliable indicators of those diseases. The present paper aims to separate healthy speech samples and different groups of disordered speech of patients with various disease types, namely depression, Parkinson, morphological alteration of vocal organs, functional dysphonia and recurrent paresis. The correlation matrices of the time shifted values of formant frequencies (F1, F2, F3), mel-filter band energy values, mel-frequency cepstral coefficients (MFCCs), fundamental frequency (F0) and intensity were used as input for the classification of the diseases. Support vector machines and k-nearest neighbor methods were utilized to compare performances. In six-class classification experiment, the best overall accuracy was 54.75%, and the accuracy was 77.59% using re-categorization of disorders into four classes. Based on the achieved results, a speech-based diagnostic tool can be created that helps clinical staff by giving them a novel marker for diagnosis.**

***Keywords*—depression; diagnostics; Parkinson's disease; pathological speech; speech biomarkers; support vector machines**

I. INTRODUCTION

The use of biomarkers is becoming increasingly popular since these markers can be a measurable indicator of the severity or the presence of some disease state. Speech is one of the biomarkers that may indicate a variety of the related illnesses. It can provide cheap, noninvasive and effective methods to help the work of professionals.

Dysphonia is the disturbance of the produced sound: the clear sound of the voice is accompanied by noise. Dysphonic voice is characterized to be hoarse, breathy, harsh or rough [1][2]. There are two different reasons for developing this disease: the voice problem in the absence of a physical condition, referred to as functional dysphonia (FD); and the physiological distortion in one of the subsystems of speech production, referred to as morphological alteration of vocal organs (MA). Vocal cord nodules, polyps, gastroesophageal reflux disease (GERD), cyst and vocal cord paralysis (recurrent paresis, (RP)) can all be categorized into structural organic disorders, while diseases such as stroke, Parkinson's disease (PD) or multiple sclerosis are included in neurological voice disorders. Acoustic features like jitter, shimmer, HNR (Harmonics-to-Noise Ratio) are useful in the automatic classification of healthy and dysphonic voices, using continuous speech [3][4][5].

Depression is a psychiatric illness. Clinical depression can be caused by experience of failure, which has emotional, cognitive, physical and motivational symptoms. Depression is generally under-recognized, and it is often facing the problem of inadequate treatment and mistreatment. It is predicted to become the second most significant cause of disability by 2020 [6][7]. Speech may be a good objective marker for depression detection as well, as numerous researchers have shown [8][9][10][11][21].

Parkinson's disease (PD) is one of the most frequent neurological disorders with associated progressive decline in motor precision and sensorimotor integration. Voice characteristics of PD include imprecise and uncoordinated articulation, decreased loudness, improved vocal tremor, variable speech rate and rushes of breath and pause segments, breathy and harsh voice quality [12][13][14][15][16][19].

Until now, works in the literature have mostly reported two-class classification for the distinctions of healthy and pathologic speech. In our earlier works, we also developed two-class classification systems to separate healthy speech from dysphonic one [3], healthy speech from depressed one [8] and healthy speech from speech of patients in Parkinson disease [13]. But in practice all of these diseases may occur among the patients. Therefore, in the present paper, we focus on separating more (4 or 6) different diseases simultaneously, using multi-class classification method. Examined disease types are: speech samples of depression, Parkinson's disease, structural morphological alteration of vocal organs, functional dysphonia and recurrent paresis.

We hypothesize that these diseases influence the correlation matrices of the time shifted values of formant frequencies (F1, F2, F3), mel-filter band energy values, mel-frequency cepstral coefficients (MFCCs), fundamental frequency (F0) and intensity differently. (Correlation structure features have been used even before, but only for two-class classification [17][18][20][22].)

II. DATABASES

Three databases (one for each disease type) and a healthy control dataset were used. All databases contained a read text,

The research was partly funded by the Postdoctoral Fellowship Programme of the Hungarian Academy of Sciences (POSTDOC-77).

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

titled 'The North Wind and the Sun', which were selected for the experiments. This short text is used in clinical practice for running speech analysis. The average length of a recording was 41 seconds. All subjects gave signed consent for their voices to be recorded. The number of recordings and descriptive statistics of the databases are summarized in Table I.

A. Phonation disorder Speech Database (PSDb)

Phonation disorder speech samples were collected during patient consultations in a consulting room at the Department of Head and Neck Surgery of the National Institute of Oncology, Budapest. The recordings were categorized according to the following disease types: morphological alteration (MA), diseases such as: tumors at different points of vocal tract, gastroesophageal reflux disease (GERD), chronical inflammation of the larynx, cysts, tractional stenosis, vocal node, laryngitis, laryngeal paralysis, closure insufficiency; functional dysphonia (FD); recurrent paresis (RP). Speech defect severity was determined by the clinician who set up the diagnosis during the consultations. The severity of dysphonia is given by the RBH scale [23] by an expert, where R stands for roughness, B for breathiness, H for overall hoarseness.

B. Depressed Speech Database (DSDb)

The DSDb is a collection of recordings from Hungarians who are suffering from depression. Speech samples were collected jointly with the Psychiatric and Psychotherapeutic Clinic of Semmelweis University, Budapest. The recordings are from patients ranging from mild depression to severe depression. The patients diagnosed with depression by a neurologist had been not diagnosed with any other neurological diseases. In order to measure depression and to classify the recordings, Beck Depression Inventory II (BDI) scale was used [24].

C. Parkinson's Speech Database (PSDb)

PSDb is a collection of recordings from Hungarian patients who suffers from PD. Speech samples were collected in two health institutes in Budapest: Virányos Clinic and Semmelweis University. The severity of PD is given by the Hoehn & Yahr scale (H-Y) [25].

D. Healthy Control(HC)

The subjects of the healthy control population had no known diseases and were not under any medical treatment. These recordings were made using the same text material as in case of the disordered speech samples. 190 healthy subjects were recorded: 85 male 105 female speakers.

TABLE I. DATABASE AGE AND DISEASE SEVERITY STATISTICS

Dataset	*Measurement*	*Gender*	*Number of speech samples*	*Severity score*	*Age*
PSDb - MA	mean RBH (0-3)	male	52	2.17(±0.88)	55.4(±12.8)
		female	70	1.83(±0.82)	48.8(±15.3)
PSDb - FD	mean RBH (0-3)	male	20	1.45(±0.69)	56.2(±14.5)
		female	48	1.31(±0.59)	53.1(±17.3)
PSDb - RP	mean RBH (0-3)	male	22	2.50(±0.80)	50.2(±15.4)
		female	51	1.86(±0.83)	58.2(±10.6)
DSDb	BDI (0-61)	male	20	26.6(±8.9)	44.1(±14.3)
		female	35	28.2(±10.2)	43.4(±13.5)
PSDb	H-Y (0-5)	male	40	2.74(±1.05)	64(±9.5)
		female	36	2.74(±1.10)	65.4(±9.4)
HC	-	male	85	-	44.7(±18.7)
		female	105	-	47.7(±13.8)

III. METHODS

A. Low-level descriptors

Several acoustic features have been selected, which represent underlying changes in vocal tract shape and dynamics due to the voice pathologies. These features are named as low-level descriptors (LLDs), using the following sets: a set of formant frequencies (F1, F2, F3), a set of mel-filtered band energy values (27 bands from 60 Hz to 8 kHz), a set of mel-frequency cepstral coefficients (MFCCs with 12 coefficients), and a set involving fundamental frequency (F0) and intensity. All features are computed at 10-ms data frames using Praat [26].

B. Correlation structure features

The calculation of the correlation and covariance structure and their derived features were performed according to the work of Williamson et al. [17][18]. The before mentioned time series of LLDs were used as channels (according to the notations of [17] and [18]) and were grouped into sets: 'formants' (F1, F2, F3), 'melfilters' (27 bands of mel-filtered band energy values), 'mfccs' (12 coefficients of mfcc coefficients), 'enf0' (intensity and f0).

The channel-delay correlation matrices were computed for each speech sample from the multiple time delays of series of LLDs. Each channel-delay matrix contains a $k \times k$ matrix, which contains further matrices with dimensions $n \times n$, where n is the number of time-delays of one LLD (for example of F1 formant frequency) and k is the number of features in a channel (for example k=3 in the 'formants' set) for a given correlation matrix. The matrices are computed at four separate time shifts scales, in which time series of LLDs in a channel are shifted with different frame spacing. A detailed description of the correlation structure approach can be found in [22] and its application to speech signals in [17][18].

10 frame delays (df) were used for 'melfilters', 'mfccs' and 'enf0' and 30 frame delays were used for 'formants' channel. Four different timescales (1, 2, 4, 8) were applied for the calculation of the correlation and covariance structure of each given channel. Fig. 1 shows the correlation matrices computed for the 'formants' channel using scale 1. All together 16 matrices were computed per speech sample.

At the end the following correlation structure features were calculated and derived from each correlation and covariance matrices for each scale: the eigenvalues of the correlation matrices, the entropy of the eigenvalues, and the power of the eigenvalues of the covariance matrices. These features were used as inputs of the classifiers.

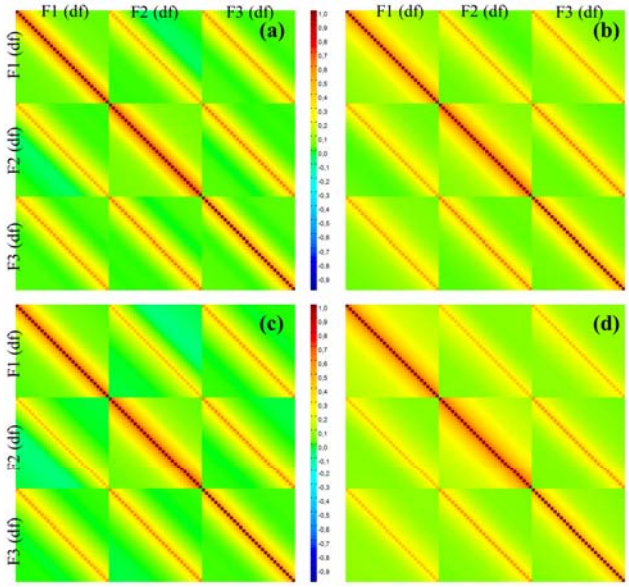

Fig. 1. Correlation matrices of 'formants' channel (feature set) using scale 1. (a)-healthy, (b)-depression, (c)-morphological alteration, (d)-Parkinson

C. Classification

RapidMiner Studio 7.5 was used for machine-learning tests. The following classification methods were applied: k-nearest neighbors (k-NN, with k set to 9) and support vector machines, c-SVC with linear (with parameter C set to 1) and radial basis kernel function (with parameters C and set to $\#features$ and $\frac{1}{\#features}$ correspondingly). The tests were performed by 10-fold cross validation.

At the first set of classification target classes, we aimed to differentiate HC, DE, PD speech and (MA), (FD) and (RP) speech from the Phonation Disorder Speech Database, a total of 6 classes. Three different groups from Phonation Disorder Speech Database were selected as separate classes, because it was found in our earlier work [5] that RP, MA and RP samples could be possible to separate.

At the second set of classification target classes, a reduced set of the previous classes was used. The MA, FD and RP classes were merged into a single class (referred as phonation disorder, PhoD), thus in case of second set of classification 4 classes were differentiated as HC, DE, PD and PhoD.

Feature selection (where it was applied, see in Results) was done using Forward Selection. Overall accuracy was chosen as cost function and the maximal number of selected features was 20.

IV. Results

The overall accuracy ($\frac{\#\ of\ correctly\ recognized\ samples}{\#\ of\ total\ samples}$) values obtained using the 6-class and 4-class classifications are shown in Table II. The table shows the results for each feature set and for all features altogether. The different features sets have different separation performances. On average, 'enf0' had the lowest performance in all cases. It seems that auto- and cross-correlation of intensity and pitch do not have significant separation ability. The other three sets achieved higher accuracy values. Among them, 'melfilters' were the best performing set.

TABLE II. Classification results using 6-class / 4-class classification

		accuracy		
Feature set	**Scale**	**k-nn**	**svm-linear**	**svm-rbf**
enf0	1	37.85 / 51.94	41.20 / 54.05	41.73 / 56.51
	2	38.73 / 54.23	42.43 / 54.05	41.20 / 54.93
	4	35.92 / 49.65	40.32 / 53.00	42.08 / 57.75
	8	32.92 / 46.30	36.17 / 48.06	35.21 / 47.71
	all	38.73 / 53.87	34.51 / 55.89	35.21 / 56.34
formants	1	38.03 / 55.11	46.13 / 64.26	44.72 / 65.49
	2	36.27 / 53.87	43.31 / 61.80	43.31 / 62.15
	4	37.50 / 57.75	45.95 / 62.68	43.49 / 63.56
	8	38.38 / 57.39	47.71 / 63.56	45.25 / 65.85
	all	38.38 / 58.10	42.78 / 64.96	42.08 / 64.61
melfilters	1	35.21 / 51.58	44.54 / 60.56	45.95 / 63.91
	2	38.56 / 51.76	48.06 / 63.73	49.12 / 69.54
	4	39.44 / 55.28	49.82 / 65.32	47.54 / 66.73
	8	42.43 / 52.28	50.53 / 67.08	49.47 / 70.25
	all	41.55 / 56.34	51.06 / 72.36	50.35 / 74.12
mfccs	0	36.97 / 50.53	41.55 / 57.22	40.32 / 57.57
	1	36.97 / 53.87	42.78 / 60.21	42.25 / 63.03
	2	39.44 / 54.93	41.78 / 59.15	39.61 / 57.75
	3	40.14 / 59.68	41.20 / 64.79	41.55 / 65.49
	all	42.08 / 60.74	43.84 / 68.66	44.89 / 69.37
all	0	39.26 / 57.22	45.42 / 72.01	45.42 / 71.48
	1	44.54 / 59.68	50.00 / 74.30	46.30 / 74.47
	2	45.42 / 62.68	46.65 / 67.25	47.01 / 68.84
	3	46.65 / 63.03	48.94 / 75.00	47.01 / 73.42
	all	47.54 / 63.20	48.77 / **76.23**	48.42 / **77.64**
all with feature selection	0	43.13 / 65.49	46.48 / 72.76	53.87 / 72.18
	1	44.89 / 61.27	54.93 / 72.40	52.64 / 72.36
	2	48.06 / 66.55	**53.52** / 69.24	51.94 / 69.72
	3	48.77 / 68.31	52.46 / 72.15	52.64 / 71.83
	all	**51.41 / 71.30**	53.32 / 76.17	**54.75** / 77.59

TABLE III. Confusion matrix using all features (Feature set: all, Scale: all in Table II) with svm-rbf. Values in cells are percentages.

Predicted\True	**HC**	**DE**	**PD**	**FD**	**MA**	**RP**
HC	**73.16**	25.45	16.88	26.87	11.61	1.49
DE	8.42	**56.36**	9.09	1.49	1.79	0.00
PD	5.79	10.91	**72.73**	0.00	1.79	1.49
FD	5.26	0.00	0.00	**32.84**	17.86	8.96
MA	6.84	5.45	1.30	28.36	**19.64**	80.60
RP	0.53	1.82	0.00	10.45	47.32	**7.46**

In the case of 6-class classification task, the MA, FD and RP classes were usually confused with each other. The confusion matrix in case of using all features and svm-rbf is presented as an example in Table III.. This phenomenon was the reason why MA, FD and RP classes were merged into a single class. Eventually, in case of 4-class classification task, using svm-rbf and all scales as inputs, 74.12% was achieved.

In general, using all timescales as inputs to the classifiers, increased overall accuracies values were obtained. Using all sets

as inputs to the classification, resulted the highest accuracies were reached. The highest values were obtained using all scales and all feature set altogether, applying feature selection: 54.75% in the case of 6-class classification and 77.59% in the case of 4-class classification. Feature selection improved the accuracy in case of k-nn method. It is worth noticing, that with feature selection, a simple method, such as k-nn, did approach the more complex classification method, such as support vector machines.

V. Conclusion

The paper investigates an automatic approach for the classification of several types of pathological voice disorders with correlation matrices of the time shifted values of formant frequencies (F1, F2, F3), mel-filter band energy values, mel-frequency cepstral coefficients (MFCCs), fundamental frequency (F0) and intensity. Using multiple delay scales and different classification methods, the best overall accuracies was 77.59% in 4-class classification tasks. This is a remarkable achievement, especially because there are not much speech samples per class of illness was available for the training.This suggests that there are indeed correlation differences in the time domain signals of the measured features due to the articulation abnormalities of the examined 4 disease types. Based on the results, the correlation structure features can be integrated into an automatic complex diagnostic system.

The examination of the error matrices in the 6-class classification experiment showed that there are many confusions among the various types of phonation disorders as (MA), (FD) and (RP). So if these types have to be distinguished from each other, additional parameters must be selected.

References

[1] Tulics, M.G., Kazinczi, F., Vicsi, K., "Statistical analysis of acoustical parameters in the voice of children with juvenile dysphonia," in: International Conference on Speech and Computer, Springer. 2016, pp. 667–674.

[2] Ruotsalainen, J., Sellman, J., Lehto, L., Verbeek, J., "Systematic review of the treatment of functional dysphonia and prevention of voice disorders," Otolaryngology-Head and Neck Surgery 138, 2008, pp. 557–565.

[3] Kazinczi, F., Mészáros, K., Vicsi, K., "Automatic detection of voice disorders," in: International Conference on Statistical Language and Speech Processing, Springer, 2015, pp. 143–152.

[4] Grygiel J. and Strumillo P., "Application of Mel Cepstral Representation of Voice Recordings for Diagnosing Vocal Disorders," Przeglad Elektrotechzny (Electrical Review), 2012.

[5] Tulics, M.G., and Vicsi, K., "Phonetic-class based correlation analysis for severity of dysphonia," in: Cognitive Infocommunications (CogInfoCom), 2017 8th IEEE Conference on, IEEE, 2017, pp. 21-26.

[6] Kessler, R.C., Bromet, E.J., "The epidemiology of depression across cultures," Annual review of public health 34, 2013, pp. 119–138.

[7] Lépine, J.P., Briley, M., "The increasing burden of depression," Neuropsychiatric disease and treatment 7, 2011, pp 3.

[8] Kiss, G., Vicsi, K., "Mono-and multi-lingual depression prediction based on speech processing," International Journal of Speech Technology, 2017, pp. 1–17.

[9] 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, 2015, pp. 10–49.

[10] Asgari, M., Shafran, I., "Improvements to harmonic model for extracting better speech features in clinical applications," Computer Speech & Language 47, 2018, pp. 298–313.

[11] Valstar, M., Schuller, B., Smith, K., Eyben, F., Jiang, B., Bilakhia, S., Schnieder, S., Cowie, R., Pantic, M., "Avec 2013: the continuous audio/visual emotion and depression recognition challenge," in: Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge, ACM. 2013., pp. 3–10.

[12] Sztahó D, Vicsi, K., "Estimating the severity of Parkinson's disease using voiced ratio and nonlinear parameters," in: Pavel Král, Carlos Martín-Vide, Statistical Language and Speech Processing: 4th International Conference, SLSP 2016, Proceedings. Springer International Publishing, 2016. pp. 96-107.

[13] An, G., Brizan, D. G., Ma, M., Morales, M., Syed, A. R., & Rosenberg, A., "Automatic Recognition of Unified Parkinson's Disease Rating from Speech with Acoustic, i-Vector and Phonotactic Features, " Sixteenth Annual Conference of the International Speech Communication Association. 2015.

[14] Naranjo, L., Pérez, C.J., Campos-Roca, Y., Martín, J., "Addressing voice recording replications for parkinson's disease detection," Expert Systems with Applications 46, 2016, pp. 286–292.

[15] Mekyska, J., Smekal, Z., Galaz, Z., Mzourek, Z., Rektorova, I., Faundez-Zanuy, M., López-de Ipiña, K., "Perceptual features as markers of parkinson's disease: the issue of clinical interpretability," in: Recent Advances in Nonlinear Speech Processing. Springer, 2016, pp. 83–91.

[16] Pompili, A., Abad, A., Romano, P., Martins, I.P., Cardoso, R., Santos, H., Carvalho, J., Guimarães, I., Ferreira, J.J., "Automatic detection of parkinson's disease: An experimental analysis of common speech production tasks used for diagnosis," in: International Conference on Text, Speech, and Dialogue, Springer, 2017, pp. 411–419.

[17] J. R. Williamson, T. F. Quatieri, B. S. Helfer, G. Ciccarelli, and D. D. Mehta, "Vocal and facial biomarkers of depression based on motor incoordination and timing," in Proceedings of the 4th ACM International Workshop on Audio/Visual Emotion Challenge (AVEC), 2014, pp. 65–72.

[18] J. R. Williamson, T. F. Quatieri, B. S. Helfer, R. Horwitz, B. Yu, and D. D. Mehta, "Vocal biomarkers of depression based on motor incoordination," in Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge, 2013, pp. 41–48.

[19] Williamson, James R., et al. "Segment-dependent dynamics in predicting Parkinson's disease." Sixteenth Annual Conference of the International Speech Communication Association. 2015.

[20] B. Yu, T. F. Quatieri, J. W. Williamson, and J. Mundt, "Prediction of cognitive performance in an animal fluency task based on rate and articulatory markers," in 15th Annual Conference of the International Speech Communication Association, September 9–13, Portland, Oregon, Proceedings, 2014.

[21] B. S. Helfer, T. F. Quatieri, J. R. Williamson, L. Keyes, B. Evans, W. N. Greene, J. Palmer, and K. Heaton, "Articulatory dynamics and coordination in classifying cognitive change with preclinical mTBI," in 15th Annual Conference of the International Speech Communication Association, September 9–13, Portland, Oregon, Proceedings, 2014.

[22] J. R. Williamson, D. Bliss, D. W. Browne, and J. T. Narayanan, "Seizure prediction using EEG spatiotemporal correlation structure," Epilepsy and Behavior, vol. 25, no. 2, 2012, pp. 230–238.

[23] Wendler, J., Rauhut, A., Kruger, H., "Classification of voice qualities," Journal of Phonetics 14, 1986, pp. 483–488.

[24] Beck, A.T., Steer, R.A., Ball, R., Ranieri, W.F., "Comparison of beck depression inventories-ia and-ii in psychiatric outpatients," Journal of personality assessment 67, 1996, pp. 588–597.

[25] Hoehn, M.M., Yahr, M.D., "Parkinsonism onset, progression, and mortality," Neurology 17, 1967, pp. 427–427.

[26] Boersma, Paul & Weenink, David (2018). Praat: doing phonetics by computer [Computer program]. Version 6.0.39, retrieved 3 April 2018 from http://www.praat.org/