Automatic classification possibilities of the voices of children with dysphonia

Miklós Gábriel Tulics and Klára Vicsi

***Abstract*—Dysphonia is a common complaint, almost every fourth child produces a pathological voice. A mobile based filtering system, that can be used by pre-school workers in order to recognize dysphonic voiced children in order to get professional help as soon as possible, would be desired. The goal of this research is to identify acoustic parameters that are able to distinguish healthy voices of children from those with dysphonia voices of children. In addition, the possibility of automatic classification is examined. Two sample T-tests were used for statistical significance testing for the mean values of the acoustic parameters between healthy voices and those with dysphonia. A two-class classification was performed between the two groups using leave-one-out cross validation, with support vector machine (SVM) classifier. Formant frequencies, mel-frequency cepstral coefficients (MFCCs), Harmonics-to-Noise Ratio (HNR), Soft Phonation Index (SPI) and frequency band energy ratios, based on intrinsic mode functions ($IMF_{entropy}$) measured on different variations of phonemes showed statistical difference between the groups. A high classification accuracy of 93% was achieved by SVM with linear and rbf kernel using only 8 acoustic parameters. Additional data is needed to build a more general model, but this research can be a reference point in the classification of voices using continuous speech between healthy children and children with dysphonia.**

***Index Terms* — voice disorder, statistical analysis, acoustic parameters, dysphonia, classification**

I. INTRODUCTION

Dysphonia is a common complaint, reported in nearly one-third of the population at some point in their life. It affects the formation of clear and distinct sounds in speech as a complex function, a pathological condition showing various symptoms due to several etiologic factors and pathogenesis diversity [1]. The term dysphonia is often incorrectly used when referring to hoarseness, however hoarseness is a symptom of altered voice quality reported in patients, while dysphonia can be defined as altered pitch, loudness, or vocal quality or effort that impairs communication as assessed by a clinician and affects the patients' life [2]. The development of cheap, easy-to-use and lightweight methods that alert subjects of possible health problems is desired.

Mobile technology is attractive, since it is easy to use and it is a nearly constant feature of daily life. The number of mobile applications is growing in healthcare. A smartphone or tablet could be an ideal mobile tool to use with complex methods that can offer clues for general physicians in identifying the early stages of dysphonia.

Researchers target such applications for the early diagnosis of pathological voices in case of adults. In the work of [3] a mobile health (m-Health) application is presented for voice screening of adults by using a mobile device. The system is able to distinguish healthy voices from pathological ones using a noise-aware method that provides a robust estimation of the fundamental frequency during a sustained production of the vowel /a/.

Some systems record more than voice disorders, also recording other details regarding the general health of a patient. In [4] a healthcare framework based on the Internet of Things (IoT) and cloud computing, the system is able to capture voice, body temperature, electrocardiogram, and ambient humidity.

Most of the research on the subject currently focuses on the accurate estimation of dysphonia, rather than the development of practical applications.

Dysphonia affects patients of all ages, however research suggests that risks are higher in pediatric and elderly (>65 years of age) populations. 23.4% of pediatric patients have dysphonia at some point during their childhood [5], [6], [7], [8]. The data therefore suggests that almost every fourth child produces a pathological voice. Studies agree that dysphonia is more often reported among boys than girls, the ratio being 70-30%.

In the last 10-20 years many studies focused on dysphonia in adults, not only on sustained vowels, but on running speech as well [9], [10]. However, in the literature we can find some studies focusing on the dysphonic voices of children.

Previous studies regarding the analysis of pathological children's voice focused mainly on sustained vowels. Researchers mostly work with small sample sizes because it is difficult to collect recordings from children. Janete Coelho and his colleagues [11] analyzed the perceptual and acoustic vocal parameters of school age children with vocal nodules and to compared them with a group of children without vocal nodules. Five children were examined from both genders, aged from 7 to 12 years. The Mann–Whitney U test, with $p < 0.05$ significance level was used in their work. Statistically significant differences were registered between the group of vocal nodules vs. the group without vocal nodules, on the following parameters: fundamental frequency, shimmer, HNR , maximum phonation time for /a/ e /z/, s/z coefficient and GRBASI (Grade, Roughness, Breathiness, Asthenia, Strain, Instability). On jitter and maximum phonation time for /s/ there were no statistically significant differences.

M. G. Tulics and K. Vicsi are with the Department of Telecommunications and Media Informatics, Budapest University of Technology and Economics, Magyar Tudósok krt. 2., Budapest 1117, Hungary
(email: tulics,vicsi@tmit.bme.hu., url: http://www.tmit.bme.hu)

DOI: 10.36244/ICJ.2018.3.5

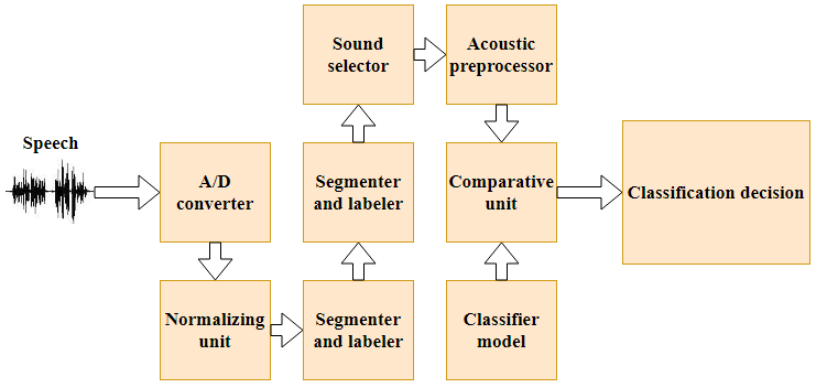

Fig. 1. Proposed framework for the recognition of dysphonic voiced children

The study of Gopi Kishore Pebbili and his colleagues [12] aimed to document the Dysphonia Severity Index (DSI) scores of 42 Indian children aged 8–12 years. DSI values were found to be significantly higher (p=0.027) in girls than in boys. DSI attempts to measure the severity of dysphonia based on the sustained production of a vowel, using a weighted combination of maximum phonation time, highest frequency, lowest intensity, and jitter (%) of an individual.

In [13] correlation between perceptual and acoustic data was examined to identify measures that are useful in determining the severity of voice deviation in children. Recordings from 71 children (aged 3–9 years) were used, containing the sustained sound /ɛ/ and the counting of numbers from 1 to 10. Results showed that F0 measures correlate with strain to phonate; shimmer and GNE parameters correlate with general degree of voice deviation.

In our earlier research [14] continuous speech was examined, where we investigated the relationship between the voices of healthy children and those with functional dysphonia (FD). The statistical analyses drew the conclusion that variations of jitter and shimmer values with HNR (Harmonics-to-Noise Ratio) and the first component (c1) of the mel-frequency cepstral coefficients (referred to as 'MFCC01') are good indicators to separate healthy voices from voices with FD in the case of children. Samples from healthy children and adult voices were also compared giving a clear conclusion that differences exist in the examined acoustical parameters even between the two groups. It is necessary to carry out the investigations separately on children's voices as well; we cannot use adult voices to draw any conclusions regarding children's voices.

The goal of this research is to identify further acoustic parameters that are able to distinguish healthy voices of children from ones with dysphonia. For this reason statistical analyses was prepared, followed by a detailed classification experiment. Thus, setting a basis of a future mobile health application for the early recognition of dysphonia in the case of children.

Section 2 briefly describes the speech material used in the experiments, followed by the description of the measured acoustic parameters, the statistical evaluation, parameter reduction and model building. Our results are shown in Section 3, followed by the discussion and the future direction in Section 4.

II. MATERIALS AND METHODS

A diagnostic support system for the early recognition of dysphonia in the voices of children would follow the logic described in Fig 1. The A/D converter digitizes the analogue speech signal of the child, after which the signal is normalized. In continuous speech, the measuring locations must be determined. Since in this work the acoustic parameters are measured on phonemes, phoneme level segmentation is required. Acoustic parameters (described in paragraph B) are extracted from the selected phonemes and arranged into a feature vector. The feature vector is given to a classifier to perform binary classification (healthy or unhealthy). Prior knowledge is gained by the processing of a carefully built speech database (described in paragraph A) and an optimal classification model using the acoustic parameters with great distinguishing power (classifier model). The system produces an output; this decision is shown on the user interface of the application. This study focuses on the automatic classification of voices of children with dysphonia.

A. Dysphonic and Healthy Child Speech Database

Sound samples from children were collected at several kindergartens. All the recordings were made with parental consent, mostly in the presence of the children's parents. The children recited a poem entitled "The Squirrel", written by a logopedic specialist. This poem was chosen for therapeutic reasons, speech therapists using the poem during treatment, and because children in the 5-10 year old age group are very fond of the poem and it is easy for them to learn. The most frequent vowel in the poem is the vowel [o], with 16 pieces followed by 14 pieces of the vowel [O] and 9 pieces of vowel [E].

The recordings were made using a near field microphone (Monacor ECM-100), Creative Soundblaster Audigy 2 NX outer USB sound card, with 44.100 Hz sampling rate and 16-bit linear coding. The duration of the recordings is about 20 seconds each.

All recordings were annotated and segmented on phoneme level, using the SAMPA phonetic alphabet [15]. In the rest of this article, vowels and other sounds will be referred with SAMPA characters in brackets. The segmentation was made with the help of an automatic phoneme segmentator, which was developed in our laboratory, followed by manual corrections. A total of 59 recordings were used in this work: 25 voices from children with dysphonia (mean age: 6.52(±1.94)) (3 children had vocal nodes, the rest had functional dysphonia) and 34 recordings from healthy children (mean age: 5.35(±0.54)). Table I summarizes the recordings from the database used in the experiments.

B. Acoustic parameters

In our earlier study [14], statistical analyses draw the conclusion that acoustic parameters like jitter, shimmer, HNR and the first component (c1) of the mel-frequency cepstral coefficients are good indicators to separate healthy and dysphonic voices in case of children. These acoustic parameters showed significant difference on vowels [E], [o], [O], [A:]. Since the most frequent vowel in the poem is the [o], it is sufficient to extract these acoustic parameters on it.

In this work, we are attempting to expand the set of used acoustic parameters that could be helpful in the automatic classification of children with healthy voices from those with dysphonic ones.

In our earlier work [16] we demonstrated Soft Phonation Index (SPI) and Empirical mode decomposition (EMD) based frequency band ratios ($\mathrm{IMF_{entropy}}$) acoustic parameters measured on different phonetic classes (for example nasals, vowels, fricatives etc.) correlate with the severity of dysphonia in adult speech. Further parameters also needs to be investigated in continuous children speech.

The following acoustic parameters were used in this study: **Fundamental frequency** (F0) means, standard deviations and ranges were calculated on vowels [E] and [o]. The fundamental frequency calculation was done by an autocorrelation method described in [17].

Formant frequency (F1, F2, F3) means, standard deviations and ranges were calculated on vowels [E] and [o].

Formant frequency bandwidth (F1BW, F2BW, F3BW) means, standard deviations and ranges were calculated on vowels [E] and [o]. Formant frequency tracking was realized by applying Gaussian window for a 150 ms long signal at a 10 ms rate. For each frame LPC coefficients were measured. The algorithm can be found in [16]. In case of fundamental frequency and formant frequencies we wanted to examine if there is a difference between the vowel [E] and [o]. Vowel [E] was used by us in adult speech.

Jitter(ddp), shimmer(ddp), HNR (Harmonics-to-Noise Ratio) means, standard deviations and ranges were calculated on vowel [o]. Jitter is the average absolute difference between consecutive time periods (T) in speech, divided by the average time period. Calculation of jitter goes as follows:

$$jitter(ddp) = \frac{\sum_{i=2}^{N-1} |2*T_i - T_{i-1} - T_{i+1}|}{\sum_{i=2}^{N-1} T_i} * 100\ [\%] \quad (1)$$

where N is the number of periods, and T is the length of the periods. Shimmer is the average absolute difference between consecutive differences between the amplitudes of consecutive periods. Its calculation goes in a similar way:

$$shimmer(ddp) = \frac{\sum_{i=2}^{N-1} |2*A_i - A_{i-1} - A_{i+1}|}{\sum_{i=2}^{N-1} A_i} * 100\ [\%] \quad (2)$$

HNR represents the degree of acoustic periodicity. It is calculated wit the following formula:

$$HNR = 10 * log \frac{E_H}{E_N} [dB] \quad (3)$$

where E_H is the energy of the harmonic component, while E_N is the energy of the noise component.

12 MFCC (mel-frequency cepstral coefficients) means, standard deviations and ranges were calculated on vowel [o]. MFCCs are widely used in automatic speech and speaker recognition, where frequency bands are equally spaced on the mel scale, that approximates the human auditory system's response. To calculate the MFCCs on needs to do the following steps: first we need to frame the signal into short frames, for each frame we need to calculate the periodogram estimate of the power spectrum. Then apply the mel filterbank to the power spectra, sum the energy in each filter and take the logarithm of all filterbank energies. MFCCs are the output of a Discrete Cosine Transform (DCT) on spectral values P_j. DCT is given by the following equation:

$$c_{k-1} = \sum_{j=1}^{N} Pj\ cos(\frac{\pi(k-1)}{N}(j - 0{,}5)) \quad (4)$$

where N represents the number of spectral values and P_j the power in dB of the jth spectral value (k runs from 1 to N).

SPI (Soft Phonation Index) and **$\mathrm{IMF_{entropy}}$** means, standard deviations and ranges were calculated on vowel [o], nasals ([m], [n] and [J]), low vowels, high vowels, voiced spirants ([v], [z] and [Z]), voiced plosives and affricates ([b], [d], [g], [dz], [dZ] and [d']). Moreover, SPI was calculated on the whole sample as well. SPI is the average ratio of energy of the speech signal in the low frequency band (70-1600 Hz) to the high frequency band (1600-4500 Hz). If the ratio is large that means the energy is concentrated in the low frequencies, indicating a softer voice [17].

$\mathrm{IMF_{entropy}}$ is an empirical mode decomposition (EMD) based frequency band ratio acoustic parameter. EMD decomposes a multicomponent signal into elementary signal components called intrinsic mode functions (IMFs) [20]. Each of these IMFs contributes both in amplitude and frequency towards generating the speech signal. The IMFs are arranged in a matrix in sorted order according to frequency. The first few IMFs are the high frequency components of the signal, the latter IMFs represent the lower frequency components. We calculate the entropy (E) for each IMF. The frequency band ratios of entropy were calculated the following way:

$$IMF_{entropy} = \frac{\sum_{d=1}^{2} E_d}{\sum_{d=2}^{D} E_d} \quad (5)$$

H_d is the value of Shannon entropy for each d = 1, 2,…D of the log-transformed IMFs. D is the total number of extracted IMFs, while the Shannon entropy for a discrete signal is defined as

$$E(p_i) = -K \sum_{i=1}^{n} p_i log\, p_i \quad (6)$$

where K is a positive constant.

Thus, in this research 124 acoustic parameters were calculated (acoustic parameter set using all 124 parameters further referred to as "starting parameters"). For the extraction of the acoustic parameters a software was used that was developed in the Laboratory of Speech Acoustics.

C. Statistical analyses of acoustic parameters and decision methods

T-test compares two averages (means) and concludes if they are different from each other. It also shows us how significant the differences are. In other words, it lets us know if those differences could have happened by chance. Two sample T-tests were used for statistical significance testing for the mean values of the acoustic parameters between healthy voices and those with dysphonia (all parameters obtained were disposed by using SPSS20.0 software). Where F tests showed significant variances of an acoustic parameter within the groups (with significance level 95% (α = 0.05), Welch's T-test was used. Welch's T-test is insensitive to equality of the variances regardless of whether the sample sizes are similar. Our assumption is that the distributions are normal, but T tests are relatively robust to moderate violations of the normality assumption.

D. Feature selection and classification

A two-class classification was performed between healthy children and children suffering from dysphonia using leave-one-out cross validation, with SVM (support vector machine) classifier. SVM is a supervised machine-learning algorithm that is used mainly for binary classification tasks (for machine-learning tests, RapidMiner Studio 7.5 was used). SVM was used in this research because the classifier has achieved good results in the classification of healthy and pathological speech in the case of adults.

While with the help of the T-test we can identify all the acoustic parameters that are significantly different in the examined two groups one-by-one, it does not give us all the acoustic parameters that are useful in the automatic classification. Significant parameters may have high correlations with each other and this examination does not say anything about possible useful parameter combinations. Subsets that are more effective for classification may exist, instead of selecting only significant parameters. In the hope of finding the best acoustic parameter subset as input vector Forward feature selection algorithm was used. This is an iterative algorithm, which chooses the best feature that improves the accuracy in regards of a cost or objective function (maximum accuracy), in each step by adding an acoustic parameter to the set of parameters already chosen. It starts with an empty set and stops after a number (here with set this number to 3) of generations without improval. In this way, the FFS algorithm also reduces dimensionality.

TABLE I
THE CHILD VOICE DATABASE USED

	Diagnosis		
Sex	Dysphonia	Healthy	Sum
Girl	5	15	20
Boy	20	19	39
Sum	25	34	**59**

III. RESULTS

A. Statistical analysis of healthy voices and voices of children suffering from dysphonia

In the statistical analysis, the null hypothesis states that the means of the two groups are equal. The calculated p-value is a probability that measures the evidence against the null hypothesis. A smaller p-value provides stronger evidence against the null hypothesis. To determine whether the difference between the two group means is statistically significant we compare the p-value to a significance level. In practice, 0.1, 0.05 and 0.01 significance levels are used. The significance level of 0.05 represents a 5% risk and concludes that there is a difference when there is no real difference.

F0, jitter(ddp) and shimmer(ddp) means, standard deviations and ranges did not show significant difference between the two groups.

Formant frequencies, MFCCs, HNR, SPI and $IMF_{entropy}$ showed significant difference in more cases, presented in Table II. The table shows summary statistics for acoustic parameters significant at 0.1 significance level. Formant frequencies were significant in case of vowel [o], but not in case of [E]. This can be explained with the difference in the number of occurrences of the vowels.

B. Classification results

In the binary classification experiment, several cases were examined, trying out different input vectors. For classification an SVM classifier was used with linear and radial basis function (rbf) kernel. Each parameter is scaled to [0, 1].

First, all the parameters calculated were used as input. Acoustic parameters which showed significant difference with at least $p<0.1$ significance level were selected separately and used as input vector as well. Note that if an acoustic parameter does not show significant difference between the voices of healthy children and the voices of ones with dysphonia it can still have great distinguishing power. For this reason, the Forward feature selection (FFS) algorithm was used.

TABLE I I
ACOUSTIC PARAMETERS SIGNIFICANT AT 0.1 LEVEL.

Acoustic parameter	Group	Mean	Std. Deviation	p-value	Acoustic parameter	Group	Mean	Std. Deviation	p-value
F2.mean_[o]	Healthy	1162.925	184.223	0.085	MFCC.std_[o]_9	Healthy	8.488	1.006	0.000
	Dysphonia	1086.651	133.979			Dysphonia	10.096	1.274	
F2.std_[o]	Healthy	265.705	73.338	0.082	MFCC.range_[o]_9	Healthy	48.096	6.023	0.012
	Dysphonia	233.993	60.123			Dysphonia	53.218	9.103	
F2.range_[o]	Healthy	956.021	263.749	0.061	HNR.mean_[o]	Healthy	11.010	2.113	0.081
	Dysphonia	833.932	208.948			Dysphonia	10.039	2.024	
F2BW.mean_[o]	Healthy	299.412	94.392	0.078	HNR.std_[o]	Healthy	2.522	1.081	0.026
	Dysphonia	262.145	50.294			Dysphonia	3.259	1.394	
F3BW.mean_[o]	Healthy	763.094	244.427	0.018	HNR.range_[o]	Healthy	8.926	3.496	0.049
	Dysphonia	606.681	243.545			Dysphonia	10.799	3.578	
F3BW.range_[o]	Healthy	1264.323	290.436	0.085	SPI.raw	Healthy	0.942	0.108	0.060
	Dysphonia	1099.791	429.988			Dysphonia	1.003	0.134	
MFCC.mean_[o]_1	Healthy	277.215	16.839	0.025	SPI.mean_[o]	Healthy	1.515	0.190	0.048
	Dysphonia	289.069	22.780			Dysphonia	1.611	0.163	
MFCC.std_[o]_1	Healthy	19.432	2.988	0.056	SPI.std_[o]	Healthy	0.295	0.044	0.006
	Dysphonia	22.158	6.380			Dysphonia	0.340	0.067	
MFCC.range_[o]_1	Healthy	98.992	16.438	0.005	SPI.range_[o]	Healthy	1.391	0.232	0.002
	Dysphonia	114.361	24.042			Dysphonia	1.604	0.258	
MFCC.range_[o]_2	Healthy	93.164	17.010	0.040	SPI.range_[m-n-J]	Healthy	1.359	0.248	0.035
	Dysphonia	103.124	19.280			Dysphonia	1.495	0.227	
MFCC.range_[o]_3	Healthy	92.916	13.919	0.058	SPI.mean_[O-A:-o-u]	Healthy	1.321	0.164	0.069
	Dysphonia	100.211	14.896			Dysphonia	1.396	0.140	
MFCC.std_[o]_4	Healthy	15.072	2.618	0.001	SPI.std_[O-A:-o-u]	Healthy	0.380	0.062	0.002
	Dysphonia	17.304	2.185			Dysphonia	0.432	0.059	
MFCC.range_[o]_4	Healthy	76.543	14.040	0.001	SPI.range_[O-A:-o-u]	Healthy	1.868	0.386	0.017
	Dysphonia	88.539	10.948			Dysphonia	2.063	0.215	
MFCC.std_[o]_5	Healthy	12.461	1.801	0.031	SPI.mean_[v-z-Z]	Healthy	0.908	0.240	0.031
	Dysphonia	13.712	2.550			Dysphonia	1.059	0.285	
MFCC.range_[o]_5	Healthy	66.465	9.360	0.024	SPI.mean_[b-d-g-dz-dZ-d']	Healthy	1.148	0.273	0.071
	Dysphonia	74.682	15.462			Dysphonia	1.271	0.227	
MFCC.mean_[o]_6	Healthy	-13.625	6.726	0.003	IMF_ENTROPY.std_[o]	Healthy	0.234	0.056	0.091
	Dysphonia	-8.662	5.205			Dysphonia	0.208	0.059	
MFCC.range_[o]_6	Healthy	59.224	7.114	0.035	IMF_ENTROPY.range_[o]	Healthy	0.886	0.252	0.017
	Dysphonia	65.273	12.366			Dysphonia	0.736	0.203	
MFCC.std_[o]_7	Healthy	10.604	1.856	0.053	IMF_ENTROPY.mean_[m-n-J]	Healthy	1.290	0.294	0.082
	Dysphonia	11.657	2.233			Dysphonia	1.148	0.318	
MFCC.mean_[o]_8	Healthy	-7.773	4.003	0.079					
	Dysphonia	-9.634	3.877						

TABLE I II
TWO-CLASS CLASSIFICATION RESULTS.

Case number	Acoustic parameters	Number of parameters	Kernel	Hyper-parameters	Accuracy (%)
1	Starting parameters	124	linear	C=1	88.13%
2	Starting parameters	124	rbf	C= 124; gamma= 0.008	86.44%
3	Significant parameters (p<0.1)	37	linear	C=1	72.88%
4	Significant parameters (p<0.1)	37	rbf	C= 37; gamma= 0.027	69.49%
5	FFS with linear kernel	8	linear	C=1	93.22%
6	FFS with rbf kernel	8	rbf	C=10; gamma=0.1	93.22%

In this experiment, in the case of rbf kernel the hyperparameter C is set to the number of parameters, while gamma is set to 1/number of parameters. When using FSS we cannot know how many parameters will be chosen by the algorithm, the hyper-parameters are chosen by intuition. Leave-one-out cross validation was used in all cases. Classification results are summarized in Table III.

As Table III shows that the highest accuracy of 93% was reached using linear and rbf kernel. The features selection algorithm reduced the input dimensionality to 8 acoustic parameters, while achieving higher accuracy than the case when the starting parameters were used. The acoustic parameters selected by the FFS algorithm in case of linear kernel are the following: F3.range_[o], MFCC.range_[o]_3, MFCC.range_[o]_4, MFCC.mean_[o]_5, MFCC.range_[o]_8, MFCC.std_[o]_9, MFCC.mean_[o]_11, HNR.mean_[o].

When using rbf kernel the parameter selection algorithm also selected 8 parameters, namely: F2.range_[o], MFCC.mean_[o]_4, MFCC.std_[o]_9, MFCC.std_[o]_10, MFCC.mean_[o]_12, SHIMMER.range_[o], SPI.std_[O-A:-o-u], SPI.range_[v-z-Z]. This combination also achieved 93% accuracy.

We can conclude that these parameter combinations have great power to distinguish healthy from dysphonic voices of children.

IV. DISCUSSION AND CONCLUSIONS

The present study investigated the relationship between healthy and dysphonic voices of children using continuous speech. The classification results are good; in the long term it is worth developing a tool for the automatic detection of dysphonic voices among children. Mobile devices are suitable for implementing this method and using it in practice.

Mobile health applications are usually designed for smartphones or tablets, on some occasions smartwatches. They allow users to access information when and where they need it; reducing time wasted searching for specific data. These devices are cheap, easy-to-use and lightweight. Voice samples, metadata, acoustic parameter values and the classifier output can be collected and uploaded to a cloud server. In this way, we can monitor the quality of the children's voice over the long term.

A. Database

It was essential to create a well-structured speech database containing children's speech samples, both from healthy children and children suffering from dysphonia. The database used in this research contains 59 recordings: 25 voices from children with dysphonia and 34 healthy children. Three children from the dysphonic group had vocal nodes, the rest had functional dysphonia.

Earlier research confirmed that is necessary to carry out the investigations separately on children's voices as well, we cannot use adult voices to make any conclusions to children's voices.

B. Statistical analysis

Through statistical analyses we drew the conclusion that formant frequencies, MFCCs, HNR, SPI and $IMF_{entropy}$ measured on different variations of phonemes are good indicators to separate healthy and dysphonic voices in the case of children. F0, jitter and shimmer means, standard deviations and ranges did not show significant difference between the two groups.

C. Two-class classification and parameter selection

During classification experiments, a high classification accuracy of 93% was reached using SVM with linear and rbf kernel and reducing dimensionality to 8 acoustic parameters. It is worth mentioning that selecting only the acoustic parameters that showed significant difference did not improve the classification accuracy.

Due to small sample size overfitting might be problem. This occurs when a model begins to "memorize" the detail and noise in the training data, rather than "learning", to generalize from a trend. Although overfitting happens more often with nonparametric nonlinear models, our highest accuracy was reached with a linear model as well. We cannot conclude that the problem is solved; much more data is needed to obtain better and more general results.

The trend however is clear and promising; the automatic separation of healthy from pathological voices in the case of children is possible. This research can be a reference point in the classification of the voices of healthy children and voices of children with dysphonia using continuous speech.

D. Future work

The goal is to build a filtering system that can be used by pre-school workers. If a child with dysphonic voice can be filtered in time, they have a better chance of getting a professional help from an ear, nose and throat (ENT) specialist or a speech therapist.

Future work includes collecting further speech records to generalize the classification model on a larger dataset. There are possibilities to optimize the model as well, for example by tuning the hyper-parameters of an estimator with grid-search. The automatic annotation and segmentation of the speech recordings implemented in a smartphone-based system, and the automatic assessment of the severity of dysphonia is also desirable. We also believe that the results are generalizable to other languages.

V. ACKNOWLEDGEMENT

We would like to thank Mária Ágostházy from the Speech Therapy and Vocational Education Service of Újbuda and Beke-Nádas Éva from the Cseresznyevirág Art Kindergarten for helping us construct the Dysphonic and Healthy Child Speech Database.

REFERENCES

[1] Hirschberg J., Hacki T. és Mészáros K. (szerk.), "Foniátria és társtudományok: A hangképzés, a beszéd és a nyelv, a hallás és a nyelés élettana, kórtana, diagnosztikája és terápiája (I. kötet)". Budapest: *Eötvös Kiadó*. 2013.

[2] Stachler, Robert J., et al. "Clinical Practice Guideline: Hoarseness (Dysphonia)(Update) Executive Summary." *Otolaryngology–Head and Neck Surgery* 158.3 (2018): 409-426.

[3] Verde, Laura, et al. "An m-health system for the estimation of voice disorders." *Multimedia & Expo Workshops (ICMEW)*, 2015 IEEE International Conference on. IEEE, 2015.

[4] Muhammad, Ghulam, et al. "Smart health solution integrating IoT and cloud: a case study of voice pathology monitoring." *IEEE Communications Magazine* 55.1, 2017: 69-73.

[5] Bhattacharya N. "The prevalence of pediatric voice and swallowing problems in the United States", *Laryngoscope*, 2015;125:746-750.

[6] Duff MC, Proctor A, Yairi E., "Prevalence of voice disorders in African American and European American preschoolers", *J Voice*, 2004;18:348-353.

[7] Carding PN, Roulstone S, Northstone K, et al., "The prevalence of childhood dysphonia: a cross-sectional study", *J Voice*, 2006;20:623-630.

[8] Silverman EM, "Incidence of chronic hoarseness among schoolage children", *J Speech Hear Disord*. 1975;40:211-215.

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

[10] Grygiel, J., Strumołło, P., Niebudek-Bogusz, E., "Application of mel cepstral representation of voice recordings for diagnosing vocal disorders", *Delta 12*, 2, 2012.

[11] Coelho, Janete, et al. "Vocal nodules in school age children." *Revista de Logopedia, Foniatría y Audiología* 36.3 (2016): 103-108.

[12] Pebbili, Gopi Kishore, Juhi Kidwai, and Srushti Shabnam. "Dysphonia Severity Index in typically developing Indian children." *Journal of Voice* 31.1, 2017: 125-e1.

[13] Lopes, Leonardo Wanderley, et al. "Severity of voice disorders in children: correlations between perceptual and acoustic data." *Journal of Voice* 26.6, 2012: 819-e7.

[14] Tulics, Miklós Gábriel, Ferenc Kazinczi, and Klára Vicsi. "Statistical analysis of acoustical parameters in the voice of children with juvenile dysphonia." *International Conference on Speech and Computer*. Springer, Cham, 2016.

[15] Klára, Vicsi: „SAMPA computer readable phonetic alphabet, Hungarian," 2008.

[16] Tulics, Miklós Gábriel, and Klára, Vicsi. "Phonetic-class based correlation analysis for severity of dysphonia." *Cognitive Infocommunications (CogInfoCom)*, 2017 8th IEEE International Conference on. IEEE, 2017.

[17] Paul Boersma. "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ration of a sampled sound", *Proceedings of the Institute of Phonetic Sciences* 17. 1193, pp. 97- 110.

[18] Press, Wiliam H., et al. "Numerical Recipes in C: The Art of Scientific Computing (10.5) Cambridge University Press." *Cambridge, t992*, 1992.

[19] Roussel, N.C., Lobdell, M., "The clinical utility of the soft phonation index", *Clinical linguistics & phonetics 20*, 181–186, 2006.

[20] Huang, N.E., Shen, Z., Long, S.R., Wu, M.C., Shih, H.H., Zheng, Q., Yen, N.C., Tung, C.C., Liu, H.H., "The empirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis", in: *Proceedings of the Royal Society of London* A: mathematical, physical and engineering sciences, The Royal Society. pp. 903–995, 1998.

Miklós Gábriel Tulics was born in Baia Mare, Romania, on Aug. 15, 1988. He is currently a technical assistant and a Ph.D. candidate at the Laboratory of Speech Acoustics, Budapest University of Technology and Economics, focusing on automatic speech recognition, machine learning and pathological voice disorder recognition. He earned his Master's degree in Electrical Engineering in 2015. Tulics has been working part time for the Department of Telecommunications and Media Informatics, Budapest University of Technology and Economics since 2013.

Klára Vicsi DSc Ph.D. is currently a lead research professor and was the Chief of the Laboratory of Speech Acoustics of the Department of Telecommunications and Media Informatics at Budapest University of Technology and Economics from 2002 until 2018. She earned her Master's degree at the Loránd Eötvös University of Sciences, Budapest in 1966-1971. She earned her DSc degree at the Hungarian Academy of Sciences in 2005, her PhD degree at the Technical University of Budapest in 1992, and her Doctor's degree at Loránd Eötvös University of Sciences in 1982. She has worked as a researcher and lecturer, participated in conferences and congresses in several countries such as Germany, California USA, Finland and Poland. She is responsible for the organization of many international conferences, workshops and summer schools. She holds project manager and participant member status in several international research projects such as: Contact person, of CLARIN, FLaReNet, ELSNET. She is the leader of the Acoustical Complex Committee of the Hungarian Academy of Sciences, a Member of ISCA and a member of the scientific board of journals. She has more than 122 publications in peer-reviewed journals, 67 refereed conference proceedings and is the owner of three patents.