SPEECH DISORDER ANALYSIS USING MATCHING PURSUIT AND KOHONEN SELF-ORGANIZING MAPS

*Marek Bártů**

Abstract: The method described in the following text was developed to analyze disordered children speech. The diagnosis of the children is developmental dysphasia. Since developmental dysphasia has impact on children's speech ability, the classification of utterances helps to determine whether treatment and medication are appropriate. The paper describes the method developed to provide classification based on utterances but without any additional demands on speech preprocessing (e.g. labeling). The method uses Matching Pursuit algorithm for speech parameterization and Kohonen Self-Organizing Maps for extraction of features from utterances. Features extracted from the utterances of healthy children are then compared to features obtained from the speech of children suffering from the illness.

Key words: *Developmental dysphasia, neurology, children's speech, Matching Pursuit, Kohonen Self-Organizing Maps, Batch Map*

Received: August 1, 2012
Revised and accepted: November 12, 2012

1. Introduction

The ability to communicate by speech is one of the most important attributes of human beings. Although there are several other means of communication, the speech is hard to substitute in everyday life. In our work we deal with a method that should evaluate the progress of the disease that complicates and finally could prevent children from learning to speak.

Roughly 5 percent of the paediatric population suffers from developmental dysphasia (DD). Such an occurrence puts this disease into the group of the most frequently occurring neurodevelopmental disorders that affect children [1]. DD is often described as an inability to acquire and learn normal communication skills in proportion to age. This happens despite the fact that a child has adequate peripheral hearing, is proportionately intelligent and deficits of broad sensomotoric

*Marek Bártů
PhD student, Department of Circuit Theory, FEE CTU Prague, E-mail: marek.bartu@gmail.com

or congenital malformation of the speech or vocal system are not noticed. Often the disease negatively affects aspects of child's personality and its development.

The relation between DD and the degree of perception and impairment of the speech was observed [2]. Utterances pronounced by dysphatic children are different from utterances pronounced by healthy children at the same age. This difference could be observed by a trained therapist. The therapist is also capable of determining whether the disease recedes or deteriorates. Furthermore, there are several other methods that help in diagnosing and checking the progress of treatment like MR tractography or EEG analysis [3]. Although these methods are very precise, feasibility of repeating examination is limited due to its demands, stress and discomfort to patients.

Our aim is to develop a software that could assist and support physicians in the process of treating DD. Since the disease has direct impact on the ability to pronounce correctly, the software based on analysis of these aspects should be able to determine a degree of the disorder.

2. Method

Description of the method is divided into three parts. Each part corresponds to the one of main steps in analyzing utterances. At the first step, parameterization of utterances is carried out. We make use of the Matching Pursuit [4, 5] algorithm. For a given signal, the algorithm finds the set of waveforms that approximate the signal. These waveforms (called atoms) are picked out from a redundant dictionary. The signal is then replaced by a set of waveforms. The replacement preserves all the important information included in the authentic signal. The parameterization is adjusted to enable easy extraction of the information in the next step.

Analysis continues with feature extraction. The sets of atoms representing particular utterances are employed as training data for Kohonen Self-Organizing Maps (KSOM) [6]. During training, characteristic features for each set are found. Finally, characteristic features obtained from the sets are compared and distortions are observed and measured.

2.1 Parameterization of utterances

In order to perform various analyses on the signal, the amount of data has to be reduced while maintaining important characteristics. A parameterization based on the Matching Pursuit algorithm is suitable for analysis and comparison of utterances (two and more syllabic words) without any need for preceding segmentation of the signal. This reduces overhead experienced when performing analysis in a clinical practice.

2.1.1 Matching Pursuit

The Matching Pursuit (MP) algorithm transforms any signal into a linear expansion of waveforms $g(t)$ [4]. The waveforms are selected from redundant dictionary of given functions to best match the signal structure. A signal is then represented with a finite set of waveforms $g_n(t)$ (1).

$$f(t) \cong \tilde{f}_N(t) = \sum_{n=0}^{N-1} \alpha_n g_n(t). \tag{1}$$

The redundant and over-complete set of time-limited functions g_n is called dictionary $\mathcal{D}$. Functions $g_n(t)$ themselves are called atoms. The choice of content of a dictionary (functions) is arbitrary. A dictionary might be adjusted to the particular application. Approximation of a signal by the functions from a suitable dictionary often gives better representation compared to transformations based on unitary basis.

Although atoms might be of arbitrary choice [5], often Gabor functions (2) are utilized [4].

$$g(t) = \frac{1}{\sqrt{s}}\, g\left(\frac{t-u}{s}\right) e^{i\xi t}. \tag{2}$$

Function g is Gaussian window, equation for discrete variant with the length of T samples is in (3). Parameter σ influences shape of the window. The range of σ is given as $\sigma \leq 0.5$.

$$g[t] = e^{-\frac{1}{2}\left(\frac{t-(T-1)/2}{\sigma(T-1)/2}\right)^2}. \tag{3}$$

Fig. 1 shows two atoms $g_1[t]$ (top pane, right) and $g_2[t]$ (top pane, left) concatenated in a simple signal. Atom $g_1[t]$ starts at $t_0 = 0.1$ s, has length = 0.3 s, amplitude = 500, $\xi = 50$ Hz and $\sigma = 0.22$, atom $g_2[t]$ starts at $t_0 = 0.4$ s, has length = 0.4 s, amplitude = 1000, $\xi = 250$ Hz and $\sigma = 0.22$, sampling frequency $f_s = 2000$ Hz. Spectrum of the signal $g_1[t] + g_2[t]$ is in the bottom of Fig. 1, the plot is limited to maximal frequency $\xi = 400$ Hz. Artefacts in the spectrogram (upper left corner) are produced by conversion from RGB format to gray scale.

To best fit the function being approximated, atoms are translated by the factor u and scaled by s so that term $1/\sqrt{s}$ normalizes $g(t)$ to the norm of 1. ξ represents frequency modulation (range (0; $f_S/2$), where f_S is the sampling frequency of the signal). All the factors (u, s and ξ) are determined by the algorithm.

The expansion maintains energy, which guarantees convergence of the algorithm [7]. Matching criterion is based on inner product of the signal $f(t)$ and functions (atoms) in dictionary $g(t)$. Approximation for N-th step (or as well by N atoms) is written as (4).

$$f(t) = \tilde{f}_N(t) + R^N f = \sum_{n=0}^{N-1} \langle R^n f, g_n \rangle g_n + R^N f \tag{4}$$

During N-th iteration of algorithm, the approximation $\tilde{f}_{N-1}(t)$ of the signal is improved by adding an atom g_N for which an inner product with residual signal $R^N f$ minimal square error (5) exists.

$$max\left[\langle R^N f, g_n \rangle\right]_{n=0}^{N} \rightarrow g_N \tag{5}$$

The signal is equal to a combination of N scaled and translated atoms g_n and residual signal $R^N f$. To simplify notation, each atom is written as vector γ (6).

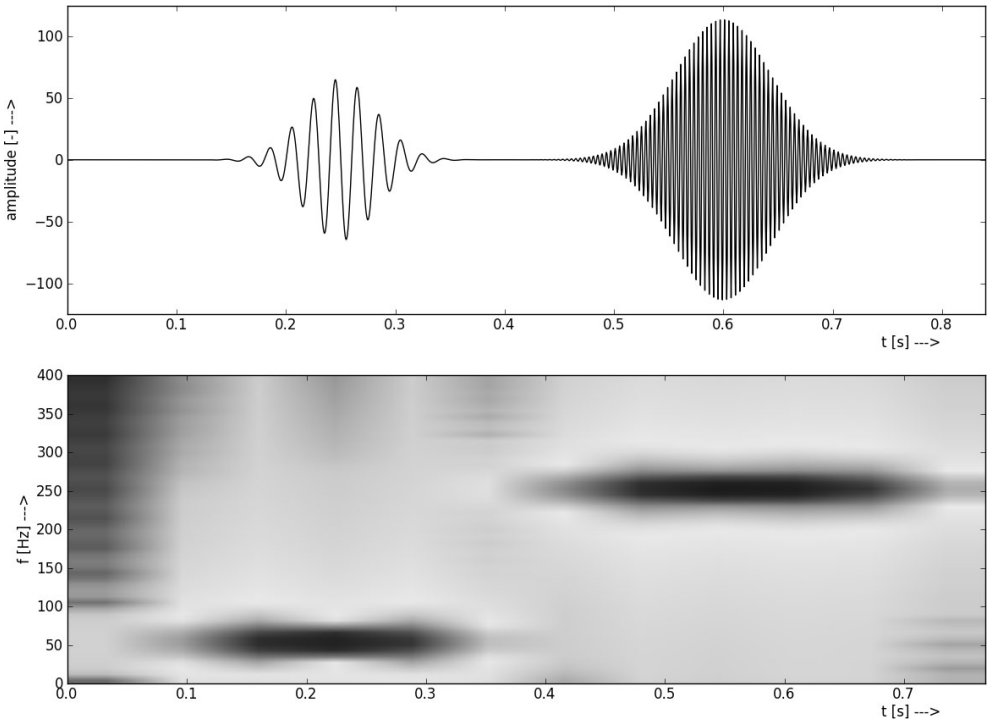

Fig. 1 *Two atoms and corresponding spectrogram - x axes (time) are in equal scale.*

$$g_n(t) = \alpha_n \frac{1}{\sqrt{s_n}}\, g\left(\frac{t-u_n}{s_n}\right) e^{i\xi_n t} \mapsto \gamma_n = (\alpha_n, u_n, \xi_n) \tag{6}$$

The number of atoms N is chosen prior to decomposition, often an estimate is done on empirical basis. The iterative process of decomposition also might be stopped due to some criterion, usually based on energy of the residuum $|R^N f|$.

For better illustration of the algorithm Figs. 2 and 3 present spectrograms obtained for two different approximations ($\tilde{f}_{500}(t)$ for $N = 500$ atoms and $\tilde{f}_{1000}(t)$ for $N = 1000$ atoms) of the same utterance ("televize"– television). Each Figure consists of map of atoms g_n in time-frequency plane (top) and of spectrogram of approximation $\tilde{f}_N(t)$ (bottom). In the time-frequency plane in the top of the Figures, each atom g_n is drawn symbolically as a line at frequency ξ_n (y-axis) located appropriately in time (x-axis). This gives a simple overview of the density of atoms approximating the utterance. No considerations about bandwidth are taken into account.

Approximations $\tilde{f}_{500}(t)$ (Fig. 2) and $\tilde{f}_{1000}(t)$ (Fig. 3) were obtained by the Matching Pursuit algorithm without any further modifications (described above). Atoms that approximate the utterance were determined by iterating process described by equation (4). The process was set to decompose the signal to 500, resp. 1000, atoms.

In contradiction to common parameterizations used in the field of speech pro-

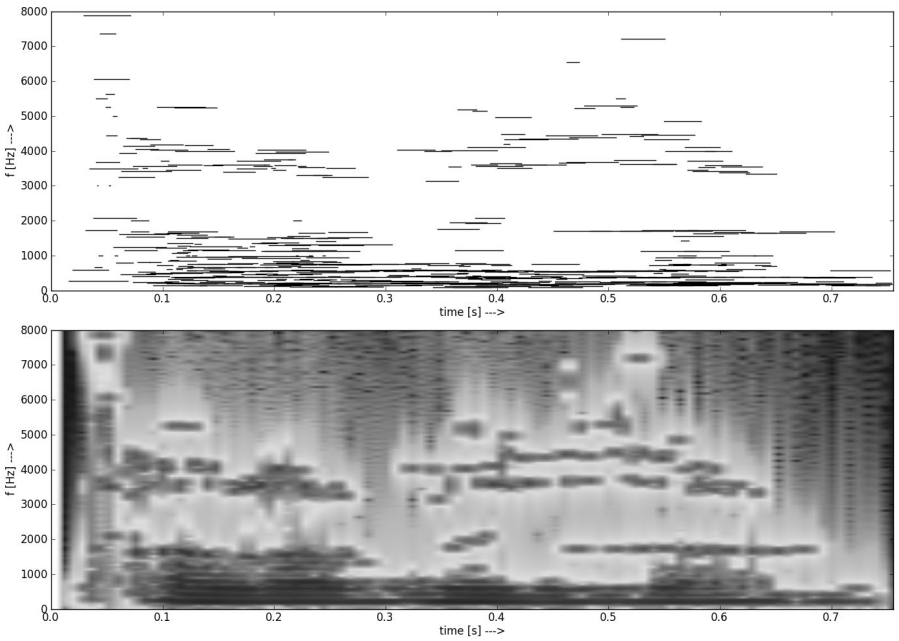

Fig. 2 *Approximation $\tilde{f}_{500}(t)$ of utterance "televize"– television ($N = 500$ atoms).*

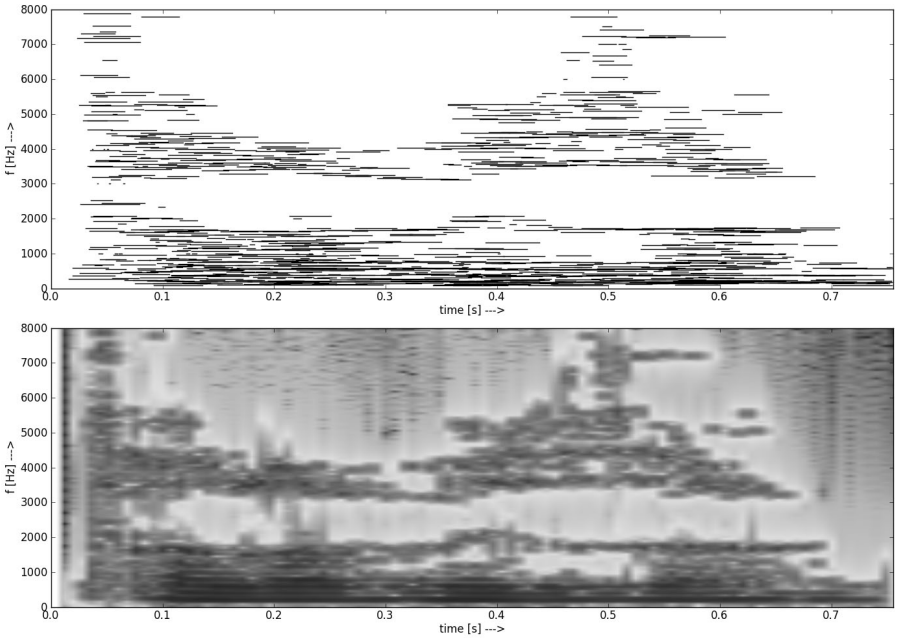

Fig. 3 *Approximation $\tilde{f}_{1000}(t)$ of utterance "televize"– television ($N = 1000$ atoms).*

cessing (e.g. MFCC, PLP) atoms g_n obtained can be used to synthesize back the approximated signal $\tilde{f}_N(t)$. Spectrograms constructed for $\tilde{f}_{500}(t)$ and $\tilde{f}_{1000}(t)$ are at the bottom of the respective Figures. Spectrograms of approximation may be compared to the spectrogram of original signal in Fig. 4.

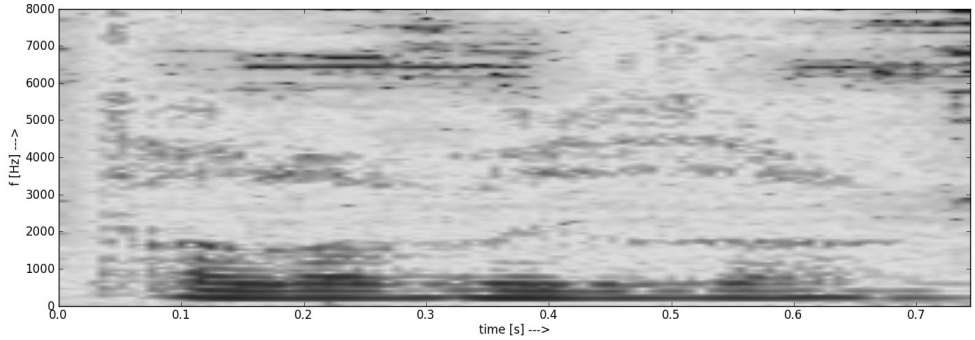

Fig. 4 *Spectrogram of utterance "televize"– television (source signal $f(t)$).*

Comparison of Figs. 2 and 3 gives an overview of the decomposition advance during algorithm progress. As could be easily noticed, $\tilde{f}_{1000}(t)$ in Fig. 3 approximates the signal $f(t)$ in more detail especially in the term of higher frequencies. This is related to the synergy of algorithm's feature to preserve energy and properties of human hearing. Since MP tends to decompose signal starting from parts containing the most energy, the approximation of a speech signal suffers from one unpleasant consequence where the approximations results in coverage that does not approximate equally all parts of the spectra, but preferring the lower frequencies that carry more energy. This could be observed in the Figs. 2 and 3. From the beginning, the algorithm tends to approximate lower parts of spectra. A reasonable approximation of the higher frequency bands is obtained only by increasing the number of atoms.

To better illustrate spectral composition of approximative signal $\tilde{f}_N(t)$ during algorithm iterations, Fig. 5 shows histograms of four approximations that differ in the different number of atoms N. As could be observed from the histograms, approximation of higher-frequency parts is being more precise with increasing number of approximative atoms. This effect is caused by a combination of previously mentioned feature of MP algorithm and attributes of human speech and is unpleasant when dealing with speech signals. Neglecting middle and high-frequency parts noticeably reduces information remaining in approximation $\tilde{f}_N(t)$ and adversely affects the task of classification.

The algorithm has been extended to avoid the consequences of the effect. The extension relies on definition of M non-overlaping frequency bands. The decomposition of the signal is performed for each of M bands separately. An arbitrary number of atoms might be found in each band but for simplicity equal number of N/M atoms is set. Bands were defined according to the recommendation in [8]: there are 24 bands covering range of 0 to 15500 Hz. A width of the band is dependent on the frequency, higher frequency bands have wider bandwidth. This

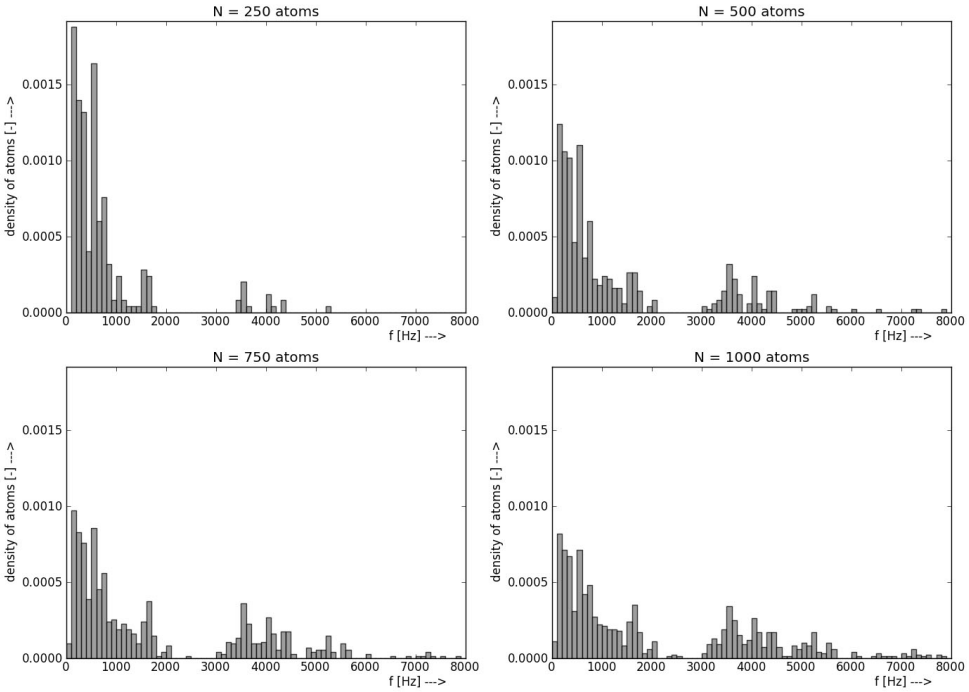

Fig. 5 *Frequency distribution of atoms for approximations consisting of different number of atoms N (standardized histograms, the scale of axes is equal).*

corresponds with the characteristics and properties of human hearing. The ceiling of 15000 Hz is sufficient despite the fact that critical frequency $f_s/2$ is higher. Atoms width $\xi \geq 16000$ Hz are rarely found in the approximation of a speech signal. This extension to the original algorithm helps to balance the content of the spectra in favor of higher-frequency components neglected by the original algorithm.

The overview of the results of decomposition based on frequency bands is in Fig. 6. The layout of the Fig. 6 is the same as for Figs. 2 and 3. Upper part shows distribution of atoms g_n in time-frequency plane (top), lower part contains spectrogram of approximation $\tilde{f}_N(t)$ (bottom). Fig. 7 shows spectral composition of approximative signal $\tilde{f}_N(t)$ for both the original (left) and modified algorithm (right). On the left is approximation according to the original algorithm given by (4), right Figure shows the distribution of atoms when modified algorithm was utilized. Both approximations were calculated for 500 atoms ($N = 500$). The extension helps to obtain approximation with balanced spectral components.

Approximation of a signal in the terms of bands allows to perform comparison based on these bands. Employment of the bands within the classification of speech of dysphatic children is being studied and the results will be published separately.

The MP algorithm might be as well adopted to better deal with a speech signal. The adaptation could be made by alternating of the matching criterion (5). In general, the alternation might be written in the form of frequency-dependent weighting

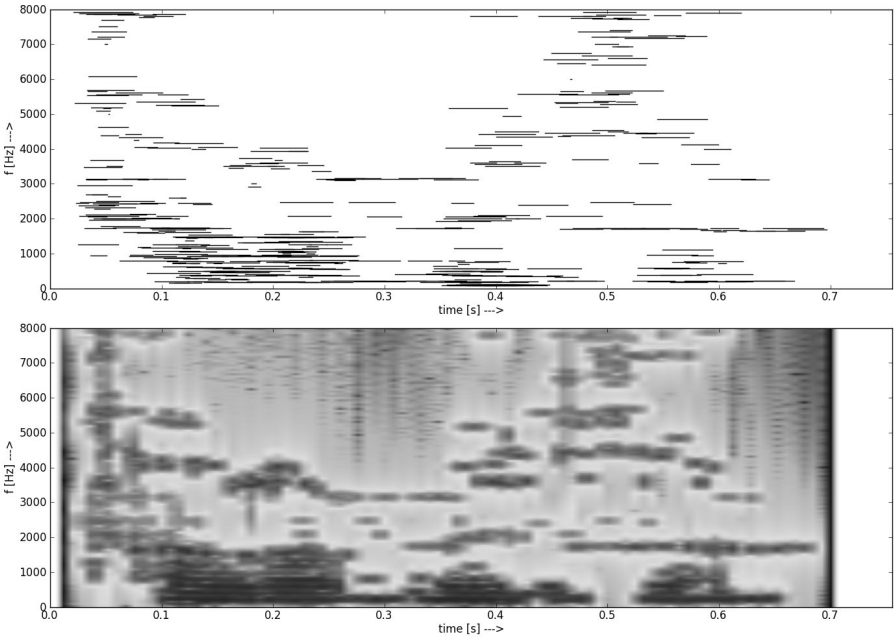

Fig. 6 *Approximation $\tilde{f}_{500}(t)$ of utterance "televize"– television, with atoms ($N = 500$) equally distributed over frequency bands.*

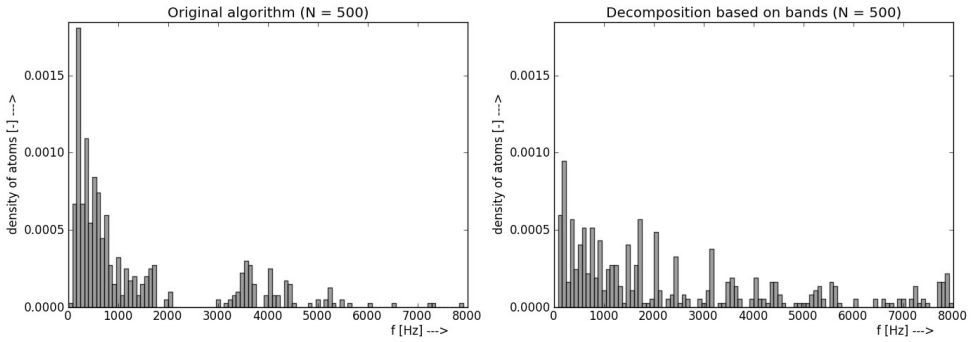

Fig. 7 *Comparison of histograms for $\tilde{f}_{500}(t)$ approximations utterance "televize"– television obtained by original algorithm (left) and algorithm based on frequency bands (right) – standardized histograms, the scale of axes is equal.*

function h (7). Function h within the equation would respect specific properties of human hearing. The function should adjust overall results of $\langle R^n f, g_n \rangle$ so that these properties are taken into account. A potential of modifications to better deal with speech signal is subject of further interest in an ongoing grant and will be published separately.

$$max\left[h\left(\langle R^n f, g_n\rangle\right)\right]_{n=0}^{N} \rightarrow g_n \tag{7}$$

2.2 Feature extraction

After parameterization, an utterance is represented by a finite set of vectors γ (6). As discussed previously, Matching Pursuit reduces greatly the amount of data for each utterance, however the set is still too large to perform direct comparison. Therefore, it is necessary to introduce another processing step that reduces the size of data and preserves all relevant information (features).

The method makes use of Kohonen Self-Organizing Maps. Vectors γ obtained from a given set of utterances serve as input data set for training maps. To train maps we use data sets that consist of the same utterances pronounced by several speakers. After training phase, map approximates the distribution of γ vector in the training set. The internal weights of the neurons are then extracted and serve as features vectors $\mathcal{F}$ for further processing and then for comparison. Dimension of the feature vectors $\mathcal{F}$ is the same for arbitrary data set and is given by the dimension of γ vectors.

The features are not extracted for each single utterance, but for a set of utterances (data set). A set consists entirely of the utterances of healthy children or only of the utterances of children suffering from developmental dysphasia. For the most experiments, a set consists only of the same kind of utterances. An experiment could require several sets to be utilized.

Particular speaker is selected according to the demands of actual experiment (e.g. age, gender). Using utterances obtained from different speakers guarantee that features represent significant characteristics of the utterance(s) and do not adapt to a particular speaker. Utterances are stored in database and the software is capable of choosing particular subset of all available records, based on gender, age and heath status. To keep the generalization, as large as possible a training data set is desirable.

Several parameters have impact on the training and subsequently on information carried by the features. The maps are trained by the Batch Map algorithm [6], so the order of vector γ does not influence the results. Appropriate size of the map has to be chosen, the shape is always rectangular. A selection of proper map size influences comparison of maps.

2.3 Classification

The method described in the paper is being developed to determine whether a speaker suffers from developmental dysphasia or not. In the affirmative, it should be possible to particularize the stage of the disease. Utterances of the speaker examined are acquired and then parameterized. Comparison takes place after parameterization, feature extraction and classification. Usually comparison is based on two sets: one made of utterances of healthy children and the other consists of utterances of children with developmental dysphasia. However, this scheme is not obligatory – more than two sets may be used. It is possible to compare an utterance of the one speaker to a number of different sets. The number and extent of the training sets are specified separately for each particular experiment. Generally, comparison is performed separately for each different utterance type.

As a basis for comparison, the database of utterances of healthy children (of different age, both genders) and children suffering from developmental dysphasia

is maintained. A set of different utterances is kept for each child as well as health status. For most experiments, all the utterances meeting the requirements of that particular experiment are split into two sets: one set is made of utterances of the healthy children only and the other set of utterances of dysphatic children only. Also, splitting up the set of utterances of dysphatic children to several smaller sets according to the degree of the disease is possible.

Classification is performed using distance-based approach. Distance of feature vectors obtained by training Kohonen maps (resp. internal weights of neurons within a map) for each set are compared one with one another.

Two diverse criteria to compare maps were suggested. Both criteria distinguish between features obtained from a base map (further referenced as B) and futures gathered from a map to compare (C). Generally, a swap of the maps leads to a different result. Both criteria are based on pairing neurons (models) within maps according to their internal weights.

The first criterion is more general (and, therefore, further referenced as $\mathcal{G}$). For each feature neuron b in a base map B neuron c in map to compare C is found. The c is chosen regardless whether it was previously paired with another vector from B or not, c itself could be paired with one or more than one b or with no neuron as well. No restrictions are applied for the pairing. The pair is made with respect to minimal Euclidean distance between the vectors of internal weight of neuron $\mathcal{F}_b$ from B and vector of internal weights of neuron $\mathcal{F}_c$ from C (8).

$$d(b,c) = d(\mathcal{F}_b, \mathcal{F}_c) = \sqrt{\sum_{n \in |B|,|C|} (\mathcal{F}_b[n] - \mathcal{F}_c[n])^2} \tag{8}$$

In (8), vectors $\mathcal{F}_b$, resp. $\mathcal{F}_c$, represent internal weights of a neuron from B, resp. C. The overall distance $D(B,C)$ between maps B and C is defined as average distance between paired neurons (9), where P is the number of neuron pairs. P equals the number of neurons in a smaller map.

$$D(B,C) = \frac{1}{P} \sum_{b \in B, c \in C} d(b,c) \tag{9}$$

The second criterion is more restrictive (further referenced as $\mathcal{R}$). It allows each neuron c from map C to be paired only with no more than one neuron b from the base map B. The criterion has to be evaluated twice for each two nets X and Y, separately for X being a base B and then for net Y being a base. When the number of neurons in the base map $|B|$ is equal to the number of neurons in map to compare $|C|$ ($|B| = |C|$), the distances $D(B,C)$ and $D(C,B)$ are equal ($D(B,C) = D(C,B)$) regardless of the map taken as a base map. The overall distance between maps D is then determined in the same manner as for criterion $\mathcal{G}$ (9).

To distinguish between the criterion used, the overall distance will be referenced as $D_\mathcal{G}$ for general criterion, resp. $D_\mathcal{R}$ for the restrictive one.

3. Experiment

As an initial experiment, a simple task that proves the convenience is presented. The aim of the experiment is to determine whether the method can distinguish between utterances pronounced by healthy and ill children. The experiment represents only a simple application of the method described above.

The experiment deals only with two-syllabic words. To simplify the analysis, utterances involved are limited to following: "papír"(paper), "pivo"(beer) and "sokol"(falcon).

Source data for the experiment are utterances obtained from 65 healthy children (43 female, 22 male) and 44 children suffering from developmental dysphasia (14 female, 30 male). Only the healthy children without any additional speech impediment are included. Since the number of dysphatic children in our database is relatively low, all the speakers suffering from developmental dysphasia who were able to pronounce the utterances requested were involved in the experiment. Degree of the disease varies a lot through that set, children with all three degrees we internally differentiate (light handicap, medium handicap and serious handicap) are involved. All the children are between the ages of 4 and 10. For each speaker, all three utterances were obtained.

Two sets are constructed for each of the utterance: the first consists only of the utterances pronounced by healthy children (further referenced as H). The other set is made of the utterances pronounced by dysphatic children (further referenced as I). Each utterance is parameterized using matching pursuit on frequency bands with equal number of atoms in each band (35 atoms in each of the 24 bands). As a representation of each utterance, 840 atoms are obtained. There are 54600 γ vectors for each of the utterances "papír"(paper), "pivo"(beer) and "sokol"(falcon) in the training set H. For set I, exactly the same method as for set H is utilized and 36960 atoms are obtained.

To train Kohonen Maps only the γ vectors representing particular atom approximating an utterance are used. Two maps were trained: one for set H and another one for set I. Feature vectors $\mathcal{F}_H$ obtained from the maps trained on set H were then compared to the features vectors $\mathcal{F}_I$ given by the maps trained on set I.

The size of the maps is determined with respect to the previous experience gained when solving similar tasks [9]. To explore influence of the size on the results of comparison, three maps with dimensions of 30×30, 40×40 and 50×50 are trained for both sets. Results of comparisons are described in following sections, separately for criterion G and R.

3.1 Results for general criterion $\mathcal{G}$

Results for general criterion G are in Tabs. I (utterance "papír"- paper), II ("pivo"- beer) and III ("sokol"- falcon). Each Table contains results of comparison between the maps trained for the utterances given. Both data sets are taken into account: healthy (denoted as H) and ill children (denoted as I).

It could be seen that the distance between maps $D_{\mathcal{G}}$ tends to be lower when comparing maps of the same size. For utterance "papír"(paper), this is valid with

the exception when comparing maps from the group of dysphatic children (I) of sizes 40×40 and 50×50. The same results are obtained for utterance "pivo"(beer) and, as well, for utterance "sokol"(falcon).

This leads to the conclusion that for a given utterance and given parameterization (840 γ-s) the set of feature vectors for dimensions sizes 40×40 and 50×50 is inevitably large and the features contained are not generalized enough. The result is influenced by the number of utterances in each of the group, so it is not possible to conclude that for these utterances maps of 40×40 and 50×50 neurons are too large. Also, it is not possible to distinguish whether the issue is in the maps trained on utterances of healthy children (H), or in the maps for dysphatic children (I) or both.

Assumption is that the convenient size of a net is proportional to the number of γ vectors obtained for each utterance. However, the experiment presented here does not consist of enough data to prove the assumption.

Diagonal of a table (i.e. when comparing the set to itself) must be equal to 0, or might be very small numbers – errors caused by a rounding during computation.

↓ C; B →	**H 30×30**	**H 40×40**	**H 50×50**	**I 30×30**	**I 40×40**	**I 50×50**
H 30×30	0.0000	0.0389	0.0321	0.0484	0.0401	0.0352
H 40×40	0.0433	0.0000	0.0328	0.0486	0.0386	0.0365
H 50×50	0.0447	0.0419	0.0000	0.0495	0.0437	0.0371
I 30×30	0.0534	0.0498	0.0391	0.0000	0.0421	0.0357
I 40×40	0.0509	0.0436	0.0390	0.0473	0.0000	0.0361
I 50×50	0.0550	0.0534	0.0420	0.0483	0.0447	0.0000

Tab. I *Distances $D_{\mathcal{G}}$ for utterance "papír"(paper).*

↓ C; B →	**H 30×30**	**H 40×40**	**H 50×50**	**I 30×30**	**I 40×40**	**I 50×50**
H 30×30	0.0000	0.0303	0.0267	0.0400	0.0359	0.0335
H 40×40	0.0369	0.0000	0.0297	0.0416	0.0387	0.0328
H 50×50	0.0375	0.0343	0.0000	0.0462	0.0411	0.0362
I 30×30	0.0598	0.0527	0.0494	0.0000	0.0370	0.0360
I 40×40	0.0664	0.0590	0.0532	0.0458	0.0000	0.0376
I 50×50	0.0635	0.0532	0.0500	0.0464	0.0410	0.0000

Tab. II *Distances $D_{\mathcal{G}}$ for utterance "pivo"(beer).*

↓ C; B →	**H 30×30**	**H 40×40**	**H 50×50**	**I 30×30**	**I 40×40**	**I 50×50**
H 30×30	0.0000	0.0253	0.0260	0.0316	0.0307	0.0272
H 40×40	0.0299	0.0000	0.0208	0.0349	0.0321	0.0266
H 50×50	0.0358	0.0251	0.0000	0.0384	0.0337	0.0294
I 30×30	0.0380	0.0361	0.0333	0.0000	0.0314	0.0284
I 40×40	0.0437	0.0394	0.0343	0.0392	0.0000	0.0299
I 50×50	0.0458	0.0406	0.0369	0.0408	0.0357	0.0000

Tab. III *Distances $D_{\mathcal{G}}$ for utterance "sokol"(falcon).*

3.2 Results for restrictive criterion $\mathcal{R}$

Results obtained for the restrictive criterion $\mathcal{R}$ are in Tabs. IV ("papír"- paper), V ("pivo"- beer) and VI ("sokol"- falcon). Structure of the Table is the same as for Tables described in Section 3.1.

The second criterion shows a different phenomenon. The sensitivity varies proportionally to the difference in the size of maps. The resolution is best when comparing maps with the same dimensions. In that case, the results are better than using criterion G.

↓ C; B →	**H 30×30**	**H 40×40**	**H 50×50**	**I 30×30**	**I 40×40**	**I 50×50**
H 30×30	0.0000	0.0431	0.0329	0.0957	0.0442	0.0366
H 40×40	0.0431	0.0000	0.0388	0.0622	0.0972	0.0469
H 50×50	0.0329	0.0388	0.0000	0.0415	0.0470	0.0822
I 30×30	0.0957	0.0622	0.0415	0.0000	0.0466	0.0367
I 40×40	0.0442	0.0972	0.0470	0.0466	0.0000	0.0437
I 50×50	0.0366	0.0469	0.0822	0.0367	0.0437	0.0000

Tab. IV *Distances $D_{\mathcal{R}}$ for utterance "papír"(paper).*

↓ C; B →	**H 30×30**	**H 40×40**	**H 50×50**	**I 30×30**	**I 40×40**	**I 50×50**
H 30×30	0.0000	0.0321	0.0275	0.1044	0.0398	0.0352
H 40×40	0.0321	0.0000	0.0329	0.0604	0.1272	0.0374
H 50×50	0.0275	0.0329	0.0000	0.0537	0.0717	0.0888
I 30×30	0.1044	0.0604	0.0537	0.0000	0.0395	0.0372
I 40×40	0.0398	0.1272	0.0717	0.0395	0.0000	0.0446
I 50×50	0.0352	0.0374	0.0888	0.0372	0.0446	0.0000

Tab. V *Distances $D_{\mathcal{R}}$ for utterance "pivo"(beer).*

↓ C; B →	**H 30×30**	**H 40×40**	**H 50×50**	**I 30×30**	**I 40×40**	**I 50×50**
H 30×30	0.0000	0.0264	0.0268	0.0791	0.0331	0.0281
H 40×40	0.0264	0.0000	0.0224	0.0397	0.0920	0.0304
H 50×50	0.0268	0.0224	0.0000	0.0351	0.0414	0.0783
I 30×30	0.0791	0.0397	0.0351	0.0000	0.0348	0.0294
I 40×40	0.0331	0.0920	0.0414	0.0348	0.0000	0.0340
I 50×50	0.0281	0.0304	0.0783	0.0294	0.0340	0.0000

Tab. VI *Distances $D_{\mathcal{R}}$ for utterance "sokol"(falcon).*

But for this case the nets must be of the same size. It would further complicate analysis, but only if optimal net size is also significantly dependent on the amount of vector in a training set.

If there is only relatively small difference between the number of γ vectors for each set of utterances, the criterion $\mathcal{R}$ should be preferred over criterion $\mathcal{G}$.

4. Conclusion

This paper describes first steps on the way to get reliable and robust classification for utterances pronounced by children with developmental dysphasia. The overall processing starting from parameterization of utterances to classification of speaker based on several utterances has been described. The parameterization is based on the Matching Pursuit algorithm [4] and feature extraction using Kohonen Self-Organizing Maps [6]. The ability of KSOM to neglect disturbing effects like noise and speech artefacts [10] is utilised.

Matching Pursuit performs parameterization that is adjusted right to the signal. The only prerequisite is a proper dictionary of functions. The dictionary should be large enough to represent a signal, but a large dictionary slows computation. The difference and potential disadvantage is that the signal is not parameterized in vectors that represent it in an equidistant manner. Representation in terms of atoms is closer to an analytic description. Because of that successive processing (classification, etc.) is being adopted.

Since internal weights of neurons in maps have a similar meaning as vectors γ, the features extracted might be resynthesized to the form of a signal. The signal might be then assessed by a speech therapist and the results obtained (based on empirical experience of a trained specialist) compared to the results of the method described.

The presented experiment shows that the method has an ability to distinguish between utterances pronounced by healthy children and children suffering from developmental dysphasia. The simple criteria were chosen only to show the potential of the method. There is a still a lot of degrees of freedom (e.g. size of γ vectors, number of features (size of F), etc.) that must be carefully examined and their influence described. Assumption that the convenient size of the net is proportional to the number of γ vectors obtained for each one utterance should be a starting point. For the presented experiment, these parameters were set according to the previous experience when solving similar problems [10, 9].

The parameterization using the Matching Pursuit algorithm and further feature extraction by Kohonen Maps has the potential to be further extended for software intended to clinical practice. The method is being developed to provide more precise results that allow to classify children with developmental dysphasia into several groups based on the degree of disease. Matching Pursuit parameterization is being used besides the common parameterizations (LPC, MFCC) and all the results obtained will be included in a speaker-overall classification.

The overall classification would then be compared to the results of psychological and logopedical examination as well as to the result of methods based on Electroencephalography (EEG) analysis and Magnetic Resonance Analysis (MRI) and further adjusted. The aim is to provide a software that allows fast classification and performs processing of utterances recorded during examination. This will provide a doctor feedback during therapy and also offer cheaper, instant and children friendly way how to verify treatment during or immediately after examination.

Acknowledgement

This work was supported by the Grant IGA MH CR agency, No. NT11443-5/2010.

References

[1] Dlouhá O.: Opožděný vývoj řeči a vývojové poruchy řeči – delayed development of speech and developmental disorders of speech, http://www.dysfazie.info, (in Czech).

[2] Pospíšilová L.: Diagnostické otázky k vývojové dysfázii - diagnostics questions of developmental dysphasia, Vox Pediatrie, Journal of General Practitioner for Children and Young, **5**, 1, 2005, pp. 25–27, (in Czech).

[3] Hrncir Z., Komarek V.: Analyses of EEG recordings, Neural Network World, Int. Journal on Non-Standard Computing and Artificial Intelligence, **14**, 1, 2004, pp. 21–25.

[4] Zhang Z., Mallat S., Matching pursuit with time-frequency dictionaries, IEEE Transactions on Signal Processing, **41**, 12, Dec. 1993, pp. 3397–3415.

[5] Gibson J. D., Sturm B. L.: Matching pursuit decompositions of non-noisy speech signals using several dictionaries. In: Proc. Int. Conf. Acoust. Speech Signal Process. (ICASSP'06), Toulouse, France, **3**, May 2006.

[6] Kohonen T.: Self-Organizing Maps, Springer-Verlag, 3rd edition, 2001.

[7] Mallat S.: A Wavelet Tour of Signal Processing, Academic Press, 2nd edition, 1999.

[8] Psutka J.: Komunikace s počítačem mluvenou řečí (Communication with the Computer in Spoken Language), Academia, 1995, (in Czech).

[9] Zetocha P., Tuckova J.: Speech analysis of children with developmental dysphasia by supervised som, Neural Network World, **16**, 6, 2006, pp. 533–545.

[10] Tuckova J., Komarek V.: Effectiveness of speech analysis by self-organizing maps in children with developmental language disorders, Neuroendocrinology Letters, **29**, 6, 2009, pp. 939–948.