Czech Technical University in Prague
Faculty of Electrical Engineering

Doctoral Thesis

February 2013

Ing. Marek Bártů

Czech Technical University in Prague
Faculty of Electrical Engineering
Department of Circuit Theory

SPEECH PARAMETERIZATION SUITABLE FOR CHILDREN WITH NEURODEVELOPMENTAL DISEASES

Doctoral thesis

Ing. Marek Bártů

Prague, February 2013

Ph.D. Programme: Electrical Engineering and Information Technology
Branch of Study: Electrical Engineering Theory

Supervisor: prof. Ing. Jana Tučková, CSc.

Acknowledgement

This work was supported by following grants:

- grant NR8287-3/2005 "**Computer Analysis of Speech and Overnight EEG in Children**"granted by IGA MH ČR agency (Science Foundation of Ministry of Health of the Czech Republic).
- grant NT11443-5/2010 "**Computer Analysis of Speech Expression, EEG Records and MR Tractography in Children with Developmental Dysphasia**"granted by IGA MH ČR agency (Science Foundation of Ministry of Health of the Czech Republic).

Contents

Chapter 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. Inabilities to appropriate communicate using speech also causing single out of the society. Problem of isolation caused by speech impairment is significant for children affected by developmental dysphasia (DD).

This work is only a part of on-going research project that brings together results from the fields of neurology, psychology, logopedics and speech processing. The aim of the research is to further advance in diagnosis of the children and help to efficiently treatment the disease.

Our team at Laboratory of Artificial Neural Network Application (LANNA) use knowledge acquired in the field of computer signal processing and utilize artificial neural networks (ANN) for speech analysis [**?**] In cooperation with the department of Paediatric Neurology in 2nd Faculty of Medicine of Charles University in Prague we are developing methods for utterance analysis that are suitable for patients with DD.

This thesis deals with speech parameterization suitable for analysis of utterances pronounced by children suffering developmental dysphasia. The aim is to develop signal representation that could be utilized in classification of speech that contains various impairments. The parameterization and subsequent classification described in the thesis are intended to be a part of a software tool evaluating progress of treatment and assist to a physician in clinical praxis.

Existing parameterization were examined and evaluated for this specific task. Since they proved not to perform sufficiently when utilized in classification of speech with impairments, new parameterization has been developed.

The parameterization was developed to avoid problems with labelling of utterances of children in age of 4 to 10 years. Labelling of the utterances is difficult because of mispronunciation, various artefact caused by the movements of fidget children.

Developed parameterization introduced in this thesis is based on matching pursuit algorithm (MP), various improvements are introduced for better performance on the speech of dysphatic children. Also simple classification method based on artificial neural networks - Kohonen Self-Organizing Maps (KSOM), is presented in following text. Both were developed with intent to reduce additional demands on speech pre-processing. This should help to diagnose advance of treatment right in consulting room and without any additional effort.

Descriptions of classification experiments are integral part of the thesis (poly-syllables words and doubled words are concerned). The experiments involve construction of evaluation methods that takes into account specific features of KSOMs. To further test performance of parameterization when utilized for classification based on KSOM. The KSOM is an effective platform for visualization of high-dimensional data, to fully understand contents of a data set it is a vital to fully understand contents exploit properties of data set [**?**].

Chapter 2

State of the Art

Developmental dysphasia is frequent and serious neurodevelopmental impairment of speech analysis and production. DD affecting five percent of paediatric population, the risk increases in premature new-borns. The impairment is often described as an inability to acquire and learn normal communication skills in proportion to age. This happens despite to the fact that the child has adequate peripheral hearing, is proportionately intelligent and deficit of broad sensomotoric 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 [**?**]. In our work we deal with method that should evaluate the progress of the disease that complicates and finally could prevent children from learning to speak.

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 getting worse. Our aim is to develop software that could assist and support physician in process of treating the disease. Since developmental dysphasia has impact on the children speech ability, the classification of utterances helps to determine whether treatment and medication are appropriate. The software based on analysis of these aspects should be able to determine a degree of the disorder and also help to validate medication. Relation between developmental dysphasia and the degree of perception and impairment of the speech was observed [**?**, **?**, **?**]. This observation allows a method based on classification of utterances to be developed for diagnosing of the disease. Roughly 5 percent of the paediatric population suffers from developmental dysphasia. Such an occurrence puts this disease into the group of the most frequently occurring neurodevelopmental disorders that affect children [**?**]. Since the linguistic message is coded into movements of the vocal tract [**?**], it is possible to classify the disease by the tractography [**?**], however this approach requires advanced equipment and still do not fit the condition

of being simple and cheap enough to use in clinical practice. Research team at workplace for special pedagogy and speech therapy pursue pathologic speech analysis in Czech republic [?].

There are several methods helping in diagnosing and determining the progress of treatment of developmental dysphasia already available like MR tractography or EEG analysis [?, ?]. These methods are accurate, however the feasibility of repeating examination is limited due to discomfort to the small patients, time requirements and limited financial resources. The aim is to develop relatively simple analysing method based on speech that do not introduce any further demands (in terms of labour and expenses) and thus might be easily used in clinical practice. The method is planned to contribute in obtaining the overall picture of patient's status and help in determination of appropriate therapy.

The work is a part of interdisciplinary research that brings together results from the fields of neurology, psychology, logopedics and speech processing. In that part, we focused on method based on speech signal processing. The main advantage is that the analysis has only a little demand on the patient compare to the complex examination (e.g. overnight EEG recording [?]). During recording the is the child located in known environment, usually in examination room of psychologist or logopedist whom he/she regularly attends, and in the form of a game repeats presented words. This approach is more convenient, the analysis is not further complicated with influence of fear from unknown environment.

Recording of utterances could be easily done with a little demand for complicated and expensive equipment. Simple recording device, in our case wireless lavaliere microphone connected to a computer, is sufficient. A patient is often fidget, walk around the room or turns. Lower quality of the microphone is compensated by utilization of artificial neural networks in processing algorithm. ANNs are not so sensitively to the noise and artefact contained in the signal.

Since utterances are recoded in a physician office, the recordings contains a lot of artefact (closing the doors) and noise. For that reason, the analysis should be based on a method robust enough to neglect all these distorting effects. For medical applications are suitable KSOM as a tool for visualising, exploring and mining large datasets [?] they perform well in prediction of seizures in epilepsy [?] or even analysing children speech with speech disorders [?, ?].

Similar reported classification methods are based on distance-based approaches and utilizes Hidden Markov Models [?]. Several distance measures and features have been published [?, ?, ?], however no single distance or feature has been reported to perform optimally for this application [?].

Our laboratory is also investigating other approaches how to tackle with the analysis [?] and [?]. The aim is to bring results obtained by several methods from different fields together and compute more precise results than these that could be given by only a single analysis method.

Utterances recorded are defined with cooperation with clinical logopedics and psychologists. The list includes types from single syllabic to complex sentences. The task of classification on the whole set is huge, therefore only isolated vowels were studied first [?] and in [?].

The initial development of classification method was based on common parameterization used in the field of speech signal processing (e.g. linear predictive coding (LPC), perceptual linear predictive analysis (PLP) or mel-frequency cepstral coefficients (MFCC) - more in [?, ?]). All parameterization mentioned above takes into account specifics of human auditory system. Unfortunately they were developed for automatic speech recognition (ASR) and optimized accordingly. Probably this is the reason why the classification based on these parameters is not right (as will be shown later, in chapter 11)

The results of analysis that utilizes KSOM-based classification and parameterizations mentioned were summarized in paper [?]. The results lead to decision to continue and extend the method, and to develop more robust parameterization that performs well with classification based on KSOMs. Also, this method should prevent manual labelling of utterances (it is required for classification based on KSOMs that dealing with the signals parameterized by the means of MFCC, PLP or LPC).

A new parameterization developed is based on matching pursuit algorithm [?, ?]. The algorithm involves finding projections of signal onto an over-complete dictionary of functions. These functions are specified in advance and could adapt the algorithm to various types of signal.

The algorithm allows to process multi-channel data and therefore is often utilized on the field of EEG signal processing [?, ?, ?, ?, ?, ?]. But application on music signals [?] and speech [?] are also reported.

Existing publication that focuses on application of matching pursuit to speech signal processing [?] deals with suitable dictionary of functions. It was reported that dictionary based on Gabor atoms [?] (described later in chapter 10.1) performs well but only when representing signal energy. Unfortunately the representation based on signal energy is not sufficient when for reliable analysis.

The method describes signal in more generic way but also allow incorporating relevant psychological phenomena as is common in contemporary parameterizations based on auditory models [?, ?, ?]. However the algorithm has been updated for speech with impairment and further classification by means

of KSOM. Updates are described in the thesis. The method will perform disordered speech assessment automatically based on features gathered from the utterances, the method is not based just only on quantitative measures [**?**].

All the updates were implemented and experimentally proven. As a starting point for implementation, recommendations in [**?**] were utilized. However the adaptation required major changes. For this reason, a supporting framework in Python programming language [**?**] has been prepared. Resulting parameterization is suitable for analysis and comparison of utterances (two and more syllabic words) without any need for preceding segmentation of the signal.

Chapter 3

Goals of the Thesis

The doctoral thesis has the following interrelated goals:

1. To verify applicability of standard speech parameterizations (LPC, PLP and MFCC) to speech of children suffering developmental dysphasia. Eventual adjustment of parameterization to best fit the utterances is allowed. Resultant parameterization should be reliable enough without any preceding manual operation on recordings (labelling). The parameterization should be applicable to isolated phonemes, syllables and words (monosyllabic and more-syllabic). The parameterization should work together with the classification to be developed (see below). In case that previous task fails, find and prepare another parameterization that comply with the conditions given above.

2. To propose a method for classification of utterances of children suffering developmental dysphasia. The method should work with utterances recorded in a consulting room, thus must be robust enough to disregard artefacts and noise present in the signal. The method should be based on artificial neural network.

3. Compare results obtained from the method with the findings of physicians and discuss eventual discrepancies. The data obtained might serve to adjust the parameters of underlying ANN. The method should be able to distinguish between healthy and ill children. Also, the method is assumed to have potential for further extension to perform fine classification and distinguish stage of the disease.

Chapter 4

Parameterization of Speech

The selection of proper parameterization of speech signal is import task when designing any system dealing with processing of speech. The usual objectives are the high level of compression of information contained in speech and eliminate all information not pertinent to analysis.

The aim of this chapter is to describe the basic parameterization of speech signal being currently used. The motivation was to obtain speech parameterization that performs well in combination with Kohonen Self-Organizing Maps. Initially, the experiments started with common methods like LPC, MFCC. Later we realized that there is a need for method that do not neglect various deviations caused by developmental dysphasia.

In order to perform various analyses on the signal, the amount of data has to be reduced while maintaining important characteristics. It is always a matter of definition which characteristics will be suppressed and which will be emphasised. For the classification of children speech we struggle for parameterization that do not significantly neglect features included by developmental dysphasia. At the same time, the parameterization should impose some degree of generalization ensuring that features characteristics for DD will be extracted rather then features for particular speakers.

Several methods described in the chapter requires the signal is divided to intervals on which is stacionarity guaranteed. Each such a frame is then processed independently and for each the frame a set of parameters is obtained.

4.1 Hamming Window

Obviously the signal is split into frames using a window function. The window function has nonzero value on some interval and zero value outside of that interval. Frequently chosen window function is Hamming window (4.1). Ham-

ming window is optimized to minimize the maximum (nearest) side lobe (see figure 4.2).

$$w[n] = 0.54 + 0.46 \cos\left(\frac{2\pi n}{N}\right) \tag{4.1}$$

The interval is chosen to maintain stacionarity, for speech it is approximately 10 to 30 ms [**?**]. The windows are overlaid by 5 to 10 ms. It is advantageous to set the exact window length as equal to the power of 2. This allows to fully utilizing computation power when dealing with subsequent FFT.

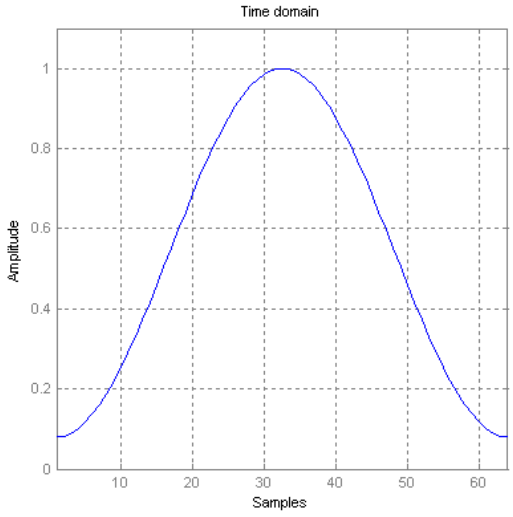

Figure 4.1: Hamming window in time plot

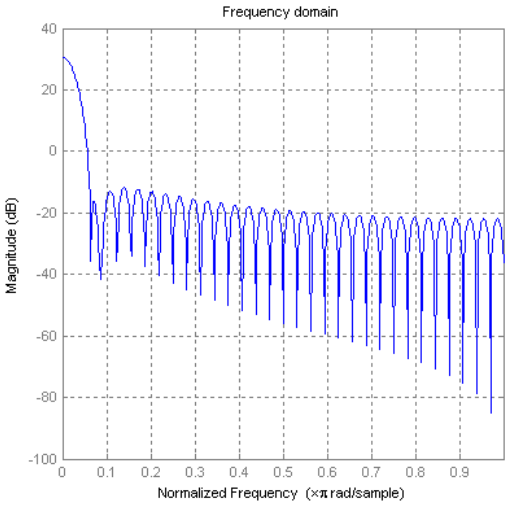

Figure 4.2: Side lobe plot for Hamming window

Chapter 5

Mel-Frequency Cepstral Coefficients (MFCC)

Mel-frequency cepstral coefficients are, as well as perceptual linear predictive coefficients [**?**] (see chapter 7), designed with regards to the human auditory specific features [**?**]. MFCC tends to respect nonlinear perception of frequencies and thus improve robustness for tasks dealing with the speech signal processing.

MFCC define the triangular-shaped filter bank arranged nonlinearly on frequency axis. Each the filter has different width. The width increases with the central frequency of filter.

The nonlinearity is introduced by mel-frequency axis (5.1), where f is frequency on linear scale in Hz and f_m is resulting frequency on non-linear mel-frequency scale in mel.

$$f_m = 2595 \ \log_{10}\left(1 + \frac{f}{700}\right) \tag{5.1}$$

The band of filters is defined linearly on mel-frequency scale (their width is constant on that scale). According to the (5.1), resulting filters are non-linearly arranged on common frequency axis.

Before filtering, the signal is pre-emphasized and split into the segments. The segments are then weighted by a window function. Common segment length is in range between 10 to 30 ms. Segments are usually weighted by the Hamming window (4.1).

In next step, amplitude spectrum $|S(f)|$ is calculated for each the segment [**?**]. Alternatively it is possible to utilize power spectrum $|S(f)|^2$ - see (5.2) (resp. (5.3)) and (5.6). Both steps are combined in (5.2) where $w[n]$ stands for windowing function (see (4.1)).

$$S(\omega, n) = \sum_{k=-\infty}^{\infty} s[k]w[n-k]e^{-j\omega k} \tag{5.2}$$

(5.2) might be rewritten as following: [**?**]

$$S(\omega, n) = e^{-j\omega n}\left(s(n) * \left(h(n)e^{j\omega n}\right)\right) \tag{5.3}$$

Comprising pre-emphasis, the initial operations done on signal writes as (5.4) - (5.4).

$$\begin{aligned} x[n] &= p\ s[n] && (5.4)\\ y[n] &= x[n]w[n] && (5.5)\\ |S(f)| &= \left|\sum_{n=1}^{N} y[n]e^{\frac{-j2\pi\omega n}{N}}\right| && (5.6) \end{aligned}$$

where $s[n]$ in (5.4) is input signal, p is amplification coefficient (might be in form of $p[n]$) and $x[n]$ is signal after pre-amplification. In (5.5) $w[n]$ stands for window function (e.g. (4.1)). For each segment $y[n]$ amplitude spectrum $|S(f)|$ is computed by (5.6). N stands for length of segments (in samples) - all samples have the same length.

A key part of the parameterization is mel filtering. The filtering is performed by a triangular-shaped filter bank that has regular spread of filters over mel-frequency scale (5.1).

Number of filters M should be chosen a priory and according to the characteristics of filtered signal. In table 5.1 are numbers recommended in [**?**] including bandwidth in Hz and mel. The recommendation considers regular spread over axis without gaps. Where it is advantageous for the task, filters in intervals where no useful signal energy is present might be left.

f_s **[kHz]**	**signal bandwidth** B **[kHz]**	**signal bandwidth** B_m **[mel]**	**number of filters** M
8	4	2146	15
11	5.5	2458	17
16	8	2840	20
22	11	3174	22
44	22	3921	27

Table 5.1: Recommendation for number of filters for MFCC parameterization (according to [**?**])

Central frequencies of filters are uniformly distributed over mel scale, their central frequencies $b_{m,i}$ are determined according to (5.7) [**?**].

$$b_{m,i} = b_{m,i} + \Delta_m \tag{5.7}$$

where $b_{m,i} = 0$ mel, $i = 1, 2, \ldots, M$ and

$$\Delta_m = \frac{B_m}{M+1} \tag{5.8}$$

where M is the number of filters (e.g. according to table 5.1).

To compute responses of the filters, filters are recomputed to the scale in Hz. All central frequencies have to be recomputed using inversion to the (5.1).

$$f = 700 \left(\exp\left(0.887 \cdot 10^{-3}\ f_m\right) - 1 \right) \tag{5.9}$$

Response of the filters then writes as (5.10).

$$Y_m = \sum_{f=b_{i-1}}^{b_{i+1}} |S(f)|\ u(f,i) \qquad i = 1, 2, \ldots, M \tag{5.10}$$

where frequencies f correspond to the frequencies used for computing Fast Fourier transform (FFT) (5.6) and $u(f,i)$ describing triangular filter (5.11) (according to [**?**]).

$$u(f,i) = \begin{cases} \frac{f - b_{i-1}}{b_i - b_{i-1}} & \text{for } b_{i-1} \leq f < b_i \\ \frac{f - b_{i+1}}{b_i - b_{i+1}} & \text{for } b_i \leq f < b_{i+1} \\ 0 & \text{else} \end{cases} \tag{5.11}$$

In figure 5.1 is the outline of distribution on the mel scale. The same filters but distributed over common frequency scale are in figure 5.2.

In the next step, the dynamics of filter outputs Y_m is reduced using logarithm. This dynamic reduction is inspired by similar feature of human hearing. The results are then transformed using Inverse Discrete Fourier Transformation (IDFT). Since the fact that power spectrum $|S(f)|^2$ is symmetric and real, IDFT is substituted by Discrete Cosine Transformation (DCT). Both operations (including the substitution) are combined in (5.12).

$$c_m(i) = \sum_{i=1}^{M} \log y_m(i) \cos\left(\frac{\pi j}{M}(i - 0.5)\right) \qquad j = 0, 1, \ldots, M \tag{5.12}$$

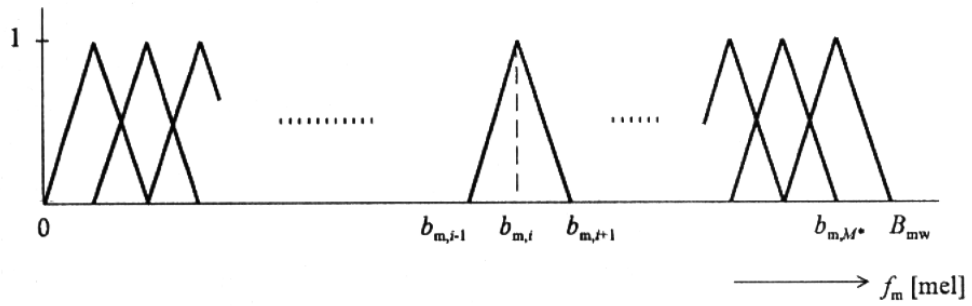

Figure 5.1: Distribution of triangular MFCC filters over mel frequency scale (reprinted from [?])

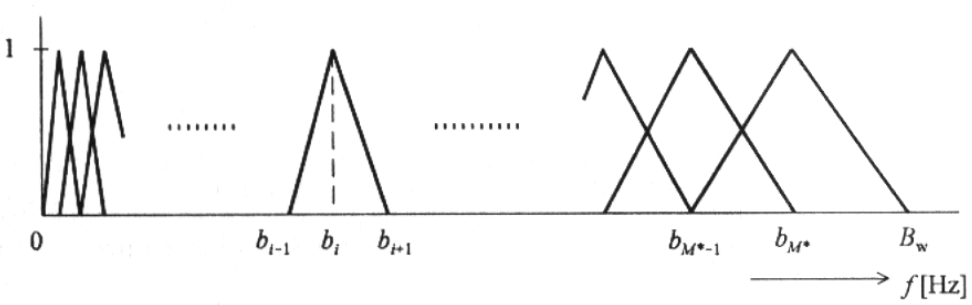

Figure 5.2: Distribution of triangular MFCC filters over common frequency scale (reprinted from [?])

Coefficient $c_m(0)$ is proportional to logarithm of energy and is often substituted by calculation directly of samples of the signal [?]:

$$c_m(0) = \log \sum_{k=0}^{N-1} \left(s(k)w(N-1-k)\right)^2 \tag{5.13}$$

MFCC values are not very robust to additive noise. It is common to normalise their values to decrease the influence of noise [?]. Proposed modifications [?] to the basic MFCC algorithm to improve robustness including raising the log-mel-amplitudes to a suitable power (around 2 or 3) before taking the DCT. This counter measurement reduces the influence of low-energy components.

MFCC coefficients are widely using in the field of speaker recognition (e.g. [?]). Despite their tendency to generalize the speech (exclude features that are specific to the speaker) their utilization in prosodic feature extraction was reported [?].

Chapter 6

Linear Predictive Coding (LPC)

Another signal parameterization is linear predictive coding [**?**]. The method determines a model of forming of speech based on short-time prediction. The parameterization is based on predicting of the n^{th} sample $s(n)$ based on linear combination the known value of preceding $n-1$ samples and excitation $u(k)$ (6.1).

$$s(n) = \sum_{m=1}^{M} a_m s(n-m) + Gu(k) \tag{6.1}$$

where a_m are predictive coefficients and M stand for order of the predictor and G is amplification. The excitation differs for voiced and unvoiced speech. For voiced speech it has a form of pulses with frequency that equals to f_0. For unvoiced speech, the excitation is assumed to be a noise with flat frequency characteristic, ideally white noise [**?**].

Equation (6.1) represents an all-pole model. If the coefficients a_m are correctly determined, the value of the n-th sample is given by equation (6.1). The short signal frame could be then described by limited number of all-pole filter coefficients a_m.

Transfer function of the model writes as (6.2):

$$H(z) = G\left(1 + \sum_{m=1}^{M} a_m z^{-m}\right)^{-1} \tag{6.2}$$

If the signal is stationary on the time frame given, then the Least square method might be utilized [**?**, **?**]. Common and widely used method for determining the coefficients is the autocorrelation method. The method is based on minimization of error e between real signal value $s(n)$ and predicted value $\hat{s(n)}$.

Function E that characterizes that error writes as (6.3).

$$E = \sum_k \left(s(k) - \hat{s}(k)\right)^2 = \sum_k \left(s(k) + \sum_{m=1}^{M} a_m s(k-m)\right)^2 \tag{6.3}$$

The error function E has minimum at

$$\frac{\partial E}{\partial a_i}, \quad 1 \leq i \leq M \tag{6.4}$$

partial derivations $\partial E/\partial a_m = 0$ leads to linear system of M equations and M unknowns [**?**, **?**, **?**] that writes as (6.5) [**?**].

$$\frac{\partial}{\partial a_\mu}\left[\sum_{n=1}^{N}\left(s(n) - \sum_{m=1}^{M} a_m s(n-m)\right)^2\right] = 0 \tag{6.5}$$

Where in (6.5) $\mu = 1, \ldots, M$. Equation (6.5) can be modified to the form of (6.6)

$$\sum_m a_m R(|m-\mu|) = R(\mu) \tag{6.6}$$

where R in (6.5) stands for autocorrelation function, $R(\mu)$ and $R(|m-\mu|)$ are defined as below

$$R(\mu) = \sum_n s(n)s(n-\mu) \tag{6.7}$$

$$R(|m-\mu|) = \sum_n s(n-m)s(n-m+m-\mu) \tag{6.8}$$

Equation (6.6) can be rewritten in the form of matrix as

$$\begin{pmatrix} R(0) & R(1) & \ldots & R(M-1) \\ R(1) & R(0) & \ldots & R(M-2) \\ \vdots & \vdots & \ddots & \vdots \\ R(M-1) & R(M-2) & \ldots & R(0) \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_M \end{pmatrix} = \begin{pmatrix} R(1) \\ R(2) \\ \vdots \\ R(M) \end{pmatrix} \tag{6.9}$$

Matrix of autocorrelation coefficients in 6.15 is a Toeplitz matrix (symmetrical and positive semi-definite). This format of the matrix allows to utilize Lewinson-Durbin algorithm [**?**, **?**] to obtain coefficients a_m.

The Lewinson-Durbin algorithm solves the system of equations (6.15) recursively for $i = 1, 2, \ldots, Q$ [?].

$$
\begin{aligned}
E^{(0)} &= R(0) && (6.10)\\
k_i &= -\frac{R(i) + \sum_{j=1}^{i-1} a_j^{(i-1)} R(i-j)}{E^{(i-1)}} && (6.11)\\
a_i^{(i)} &= k_i && (6.12)\\
a_j^{(i)} &= a_j^{(i-1)} + k_i a_{i-j}^{(i-1)} \quad where\ 1 \leq j \leq i-1 && (6.13)\\
E^{(i)} &= (1 - k_i^2) E^{(i-1)} && (6.14)
\end{aligned}
$$

Amplification G can be obtained from (6.15).

$$G^2 = R(0) + \sum_{i=1}^{Q} a_i R(i) = E \tag{6.15}$$

Coefficients a_i allow to obtain spectral envelope of native (non-sampled) signal $H(j\omega)$.

$$H(j\omega) = \frac{G}{1 + \sum_{i=1}^{Q} a_i e^{-j\omega i}} \tag{6.16}$$

Another possibility how to obtain the coefficients is the covariation method [?]. The covariation method is suitable for tasks where only a small set of samples is available. Derivation of the method is similar as for the autocorrelation method.

The important difference of the method is in characterization of overdetermined set of equation. Whereas results given by autocorrelation method are remains stable with increasing order, coefficients given by covariation method may lead to unstable system.

Critical is then to correctly determine order of the filter (predictor). Optimal is $M = 12$, for order higher then $M = 16$ is no improvement [?]. Mutual relation between the order of all-pole filter and error of prediction was reported in [?] (see figure 6.1).

[?] recommend the order Q of the model to be determined with the respect to sampling frequency f_s in kHz as

$$Q = f_s + 4 \tag{6.17}$$

[?] recommends order in the range of $Q = 7$ to 20. The order should be chosen according to task solved with respect to the sampling frequency of the signal, bandwidth and accuracy of approximation.

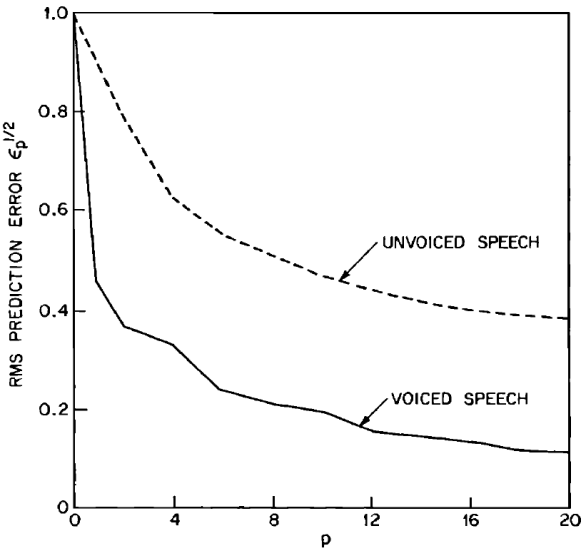

Figure 6.1: Variation of the minimum value of the RMS prediction error with p, the number of predictor coefficients (reprinted from [**?**])

Chapter 7

Perceptual Linear Predictive Analysis (PLP)

When the order of the LP model is well chosen, the parameters approximate the areas of high-energy concentration with sufficient precision. The important part of the spectrum is then well described whereas less relevant details are neglect. Satisfying these conditions, LP is an efficient tool for spectral analysis. However, in practical situation is often not easy to choose the degree in such a manner. To obtain more robust parameterization for speech recognition, the specific perception qualities of human auditory system should be taken into account, and the parameterization procedure correspondingly updated. One of such an extended method, perceptual linear predictive analysis (PLP, seldom referred as PLP-LP), was introduced by Hermansky [**?**].

Perceptual linear predictive analysis is based on linear predictive analysis. The idea of PLP is to approximate the auditory spectrum of speech by an all-pole model. Before approximation by the model, several modifications to the spectrum regards to theories of the psychophysics of hearing are made. Concepts utilized are the equal loudness curve, the intensity-loudness power law and the critical-band spectral resolution. Utilization of these concepts better adapts the LP model to properties of human auditory perceptions and improves performance in tasks of speech recognition. The implementation is described in details in [**?**], following text is intended only as an overview of PLP and the implementation of the concepts mentioned.

The analysis doesn't work on the whole signal at once, but on the segments. The speech signal is segmented and weighted by Hamming window (4.1), and then the analysis is performed for each the segment independently. The typical length of the segment is 20 ms, usually windows are overlapping by 5 to 10 ms.

f_s [kHz]	number of filters	step [bark]
8	15 + 2	973
11	17 + 2	971
16	19 + 2	985
22	21 + 2	983
44	25 + 2	991

Table 7.1: Recommendation for number of filters approximating critical-band masking curves (according to [**?**])

For each segment is then estimated power spectrum. Utilizing Discrete Fourier Transformation (DFT), the real and imaginary components are then summed up and short-term power spectrum (7.1) is obtained for each segment.

$$P(\omega) = |S(\omega)|^2 = Re[S(\omega)]^2 + Im[S(\omega)]^2 \tag{7.1}$$

To approximate nonlinear perception of acoustic signal by humans, the nonlinear transformation of frequency axis is performed. The power spectrum $P(\omega)$ is then warped into the Bark frequency axis $\Omega(\omega)$ by (7.2).

$$\Omega(\omega) = 6\ \ln\left(\frac{\omega}{1200\pi} + \sqrt{\left(\frac{\omega}{1200\pi}\right)^2 + 1}\right) \tag{7.2}$$

Warped power spectrum $P(\Omega)$ is masked by a set of filters (band passes) simulating critical band masking curve Ψ given by (7.3). Application of these filters simulates the critical-bands of spectral resolution.

$$\Psi(\Omega) = \begin{cases} 0 & \text{for } \Omega < -1.3 \\ 10^{2.5(\Omega+0.5)} & \text{for } -1.3 \leq \Omega \leq -0.5 \\ 1 & \text{for } -0.5 < \Omega < 0.5 \\ 10^{-1.0(\Omega-0.5)} & \text{for } 0.5 \leq \Omega \leq 2.5 \\ 0 & \text{for } \Omega > 2.5 \end{cases} \tag{7.3}$$

Piece-wise shape of Ψ is approximation of the asymmetric masking curve. Convolution of signal spectrum $P(\omega)$ with approximation of masking curve $\Psi(\Omega)$ yields critical-band power spectrum $\Phi(\Omega)$ (7.4).

$$\Phi(\Omega_i) = \sum_{\Omega=-1.3}^{2.5} P(\Omega - \Omega_i)\Psi(\Omega) \tag{7.4}$$

Convolution of $P(\omega)$ with relative broad critical-band masking curves results in reduced spectral resolution of $\Omega(\omega)$. The step between the filters is chosen to cover the whole analysis band. Filters are distributed linearly in the spectrum, with the step of approximately 1 bark. This span is consequence of reduction of spectral resolution critical-band power spectrum $P(\Omega)$.

In [?] are presented recommendation for appropriate number of filters covering the spectrum and the step. The key to determine these values is sampling frequency f_s. The recommendation is reproduced in 7.1. These recommendations came from practical experience on the speech recognition tasks.

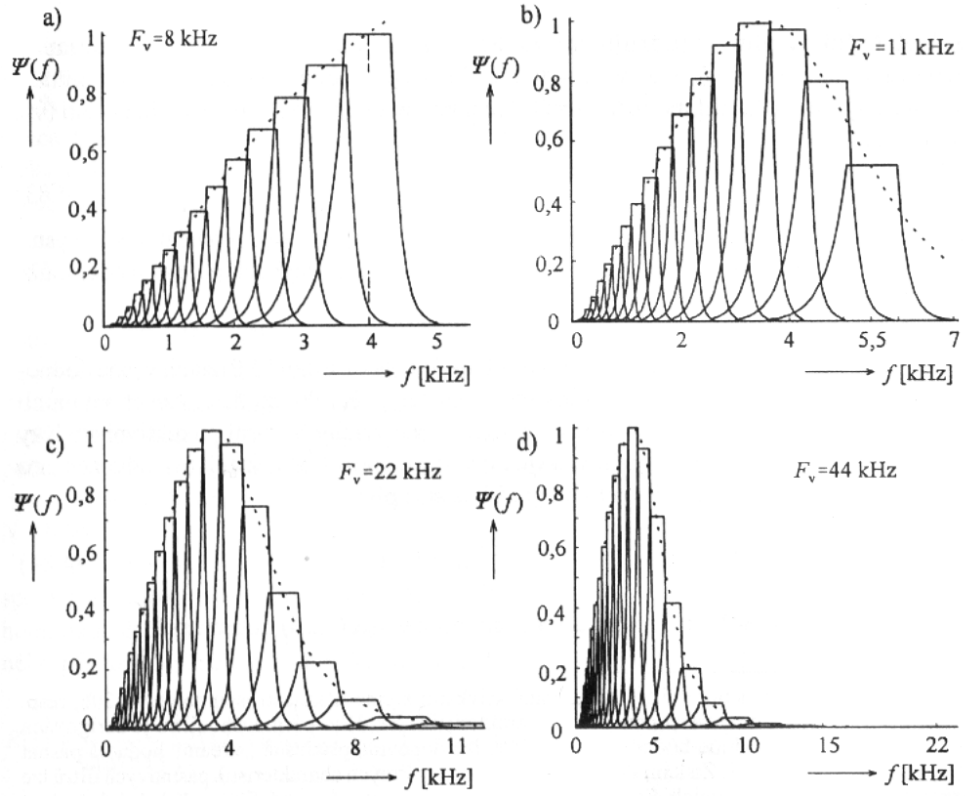

Figure 7.1: Illustration of the bank of filters for $f_s = 4$kHz (a), $f_s = 11$ kHz (b), $f_s = 22$ kHz (c) and $f_s = 44$ kHz (reprinted from [?])

The resulting spectrum Ω is pre-emphasized by equal-loudness curve, resp. its approximation $E(\omega)$. $E(\omega)$ takes into account difference in sensitivity of human hearing that depends on the frequency of the sound. Approximation adopted for PLP writes as (7.5). Figures 7.2, 7.3 and 7.4 shows the example of bank of filters for $f_s = 16$ kHz.

$$E(\omega) = \frac{\omega^4 \left(\omega^2 + 56.8 \times 10^6\right)}{\left(\omega^2 + 6.3 \times 10^6\right)^2 \left(\omega^2 + 0.38 \times 10^9\right) \left(\omega^6 + 9.58 \times 10^{26}\right)} \tag{7.5}$$

Following operation on the signal is made to approximate relation between the intensity of sound and perceived loudness (Intensity-loudness power law). This relation is nonlinear, and it is approximated by amplitude compression (7.6).

$$\Phi(\Omega) = \Xi(\Omega)^{0.33} \tag{7.6}$$

Finally, $\Phi(\Omega)$ is approximated by the spectrum of an all-pole model. Autocorrelation function of $\Phi(\Omega)$ is yielded by application of inverse DFT (IDFT). Values of autocorrelation function are then used to solve Yule-Walker equations. IDFT is preferred to FFT since only a few values are needed.

A side impact of the compression of amplitude that approximates Intensity-loudness power law (7.6) is the reduction of spectral amplitude variation in the critical-band spectrum. The reduction enables the autoregressive modelling $\Phi(\Omega)$ being done with relatively low model order. Papers [**?**] shows that degree Q of 5 is sufficient.

Different experience is reported in [**?**] for speech recognition tasks. It is recommended to use much higher degree Q of predictor.

The computational requirements are comparable to the requirements of LP. The most demanding operation is spectral calculation (FFT). Other expensive operations are the critical-band spectral integration and the cubic-root compression. The cost of autoregressive modelling for original approach described in [**?**] is negligible due to the low number of spectral samples.

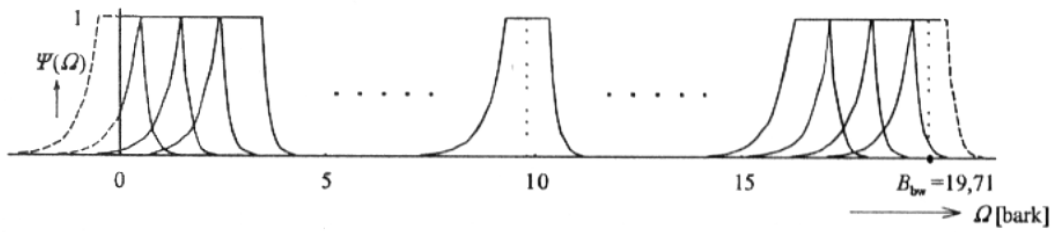

Figure 7.2: Illustration of the bank of filters for f_s= 16 kHz (reprinted from [**?**])

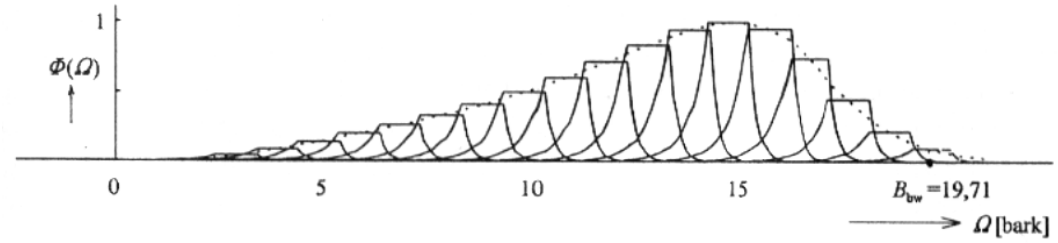

Figure 7.3: Bank of the filters taking into account the equal-loudness curve (in Barks) (reprinted from [**?**])

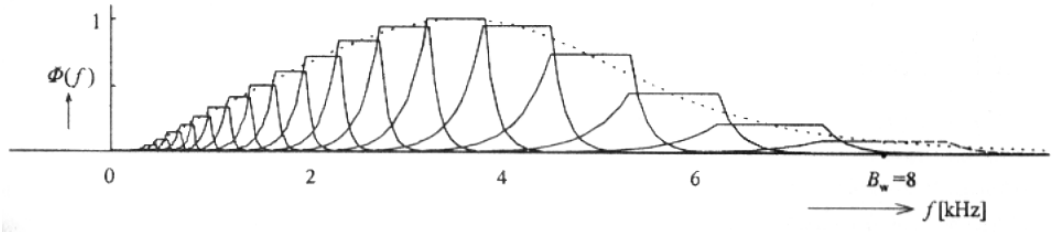

Figure 7.4: Bank of the filters taking into account the equal-loudness curve (in Hz) (reprinted from [**?**])

Chapter 8

Relative Spectral Representation (RASTA)

PLP parameterization was later expanded to RASTA parameterization (also RASTA-PLP) [**?**]. The RASTA method was developed to improve overall performance of speech recognition systems based where parameterization of speech signal is done by PLP. The aim was to develop robust parameterization insensitive to the effects caused by communication environment. RASTA abbreviation comes from RelAtive SpecTrAl representation.

It further exploits the fact that human perception tends to react more to the relative value of the change that to absolute value of an input stimulus. The linguistic message is coded into movements of the vocal tract [**?**]. The rate of change of non-linguistic component are assumed to lie outside the typical rate of change of the vocal tract shape. It is possible then to separate signal into part containing speech and the other one that contains only the non-information component of the signal. Since the human perception is invariant to slow changes, these components are then suppressed. RASTA suppresses the spectral components that change more slowly or quickly that the typical range of change of speech.

The parameterization focuses of suppression of additive and convolutional noise. RASTA is based upon PLP parameterization. PLP is supplemented with mean that suppress the non-information components before $\Phi(\Omega)$ (7.6) is approximated by the spectrum of an all-pole model.

Additional steps introduced by RASTA are similar to blind deconvolution. A critical-band short-term spectrum $\Theta(\Omega)$ from PLP (7.4) is replaced by a spectrum estimate in which each channel is band-pass filtered by a filter with a sharp spectral zero at the zero frequency. This spectral estimate is less sensitive to slow variations in the short-term spectrum.

The procedure of RASTA-PLP is described in the steps below [**?**]. On each frame of signal to be analysed following operations are performed:

1. Compute the critical-band power spectrum $\Theta(\Omega_i)$ (as in PLP, see (7.4)).
2. Transform spectral amplitude through a compression static nonlinear transformation. This step is introduced in RASTA. The nonlinear transformation mentioned is supposed to be a logarithmic one.
3. Filter the time trajectory of each transformed spectral component.
4. Transform the filtered signal through expanding static nonlinear transformation (exponential).
5. Simulate the power law of hearing (see (7.6)).
6. Approximate resulting spectrum by an all-pole model (6.1).

The procedure of the computing is in figure 8.1.

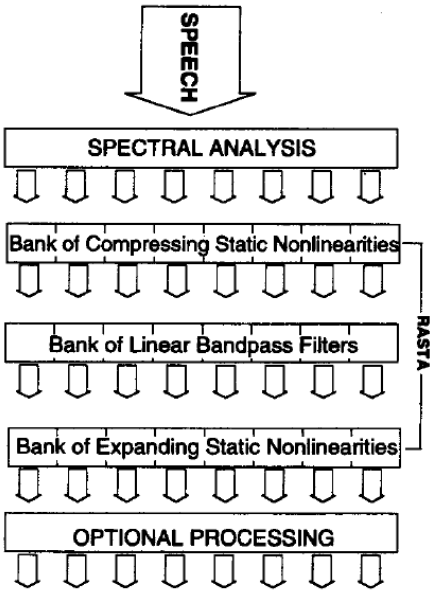

Figure 8.1: Block diagram of RASTA speech processing technique (reprinted from [**?**])

To suppress constant factors in each spectral component prior to estimation of the all-pole model, the signal is filtered by filter with transfer function specified $H(z)$ given by (8.1) (resp. by differential equation (8.2)).

$$H(z) = 0.1z^4 \, \frac{2 + z^{-1} - z^{-3} - 2z^{-4}}{1 - 0.98z^{-1}} \tag{8.1}$$

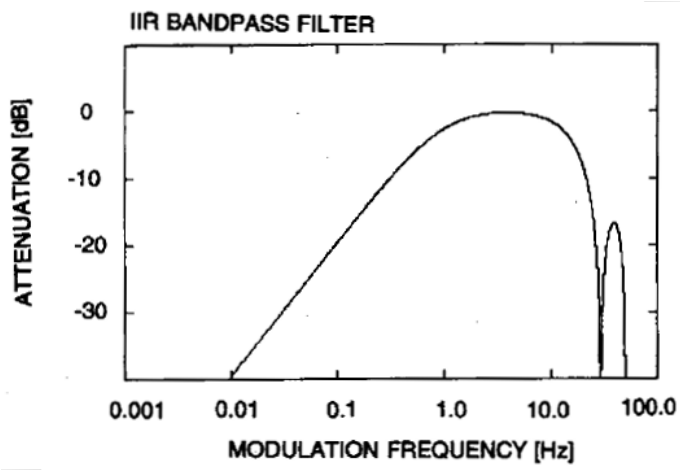

Figure 8.2: Frequency response of RASTA band filter (reprinted from [**?**])

Causal variant of the filter (8.1) writes as differential equation (8.2) (according to [**?**]).

$$\begin{aligned} y[k] =& 0.98\ y[k-1] + 0.2\ x[k] + 0.1\ x[k-1] \\ & - 0.1\ x[k-3] - 0.2\ x[k-4] \end{aligned} \tag{8.2}$$

The high-pass portion of the filter is expected to alleviate the effect of convolutional noise in the channel. The integration constant is roughly equal to the 500 ms. This means that the analysis result depends on its history that exceeds the range of a single frame.

Further experiments [**?**] showed that the integration constant of 160 ms is sufficient. Than the transfer function of the filter from (8.1), resp. (8.2) writes as (8.3), resp. as (8.4).

$$H'(z) = 0.1z^4\ \frac{2 + z^{-1} - z^{-3} - 2z^{-4}}{1 - 0.94z^{-1}} \tag{8.3}$$

The differential equation for filter given by $H'(z)$ writes as:

$$\begin{aligned} y'[k] =& 0.94\ y[k-1] + 0.2\ x[k] + 0.1\ x[k-1] \\ & - 0.1\ x[k-3] - 0.2\ x[k-4] \end{aligned} \tag{8.4}$$

Utilization of RASTA reduces requirements to the subsequent stochastic analysis. The size of a set required to train/analysis with given precision is in general lower compared to the size of the set when only PLP parameterization is utilized. RASTA parameterization was reported to works well in tasks with whole word models or phoneme-based recognizers that use triphones or broad temporal input context [**?**].

The processing involved increases the dependence of the data on its previous context, therefore the performance of simple context-independent subword-unit recognizers can be degraded. According to [**?**], RASTA is not suitable for the tasks where the speech signal without additional disturbances (clear) is to be classified.

8.1 J-RASTA

RASTA processing using logarithmical function to perform compression by nonlinear static transformation. Unfortunately this technique cannot divest signal of additive noise. To deal with additive noise, the alternative called J-RASTA was suggested in [**?**].

Instead of logarithmic compression function, J-RASTA utilizes transformation as (8.5).

$$y = \ln(1 + Jx) \tag{8.5}$$

where J is a signal-dependent positive constant [**?**]. This modification helps to suppress additive noise as well as transfer channel distortion. Since transformation (8.5) is almost linear for $J \ll 1$, the signal with substantial additive noise is transformed linearly and it is then possible to remove the noise in further spectral processing. The parts with small contribution of additive noise will be transformed logarithmically ($J \gg 1$) and then it is possible to remove impact of channel distortion during spectral processing.

Proper setting of J as indirectly proportional to the energy of additional noise [**?**].

Inversion transformation to (8.5) writes as:

$$x = \frac{e^y - 1}{J} \tag{8.6}$$

When performing spectral subtraction, the exact inverse (8.6) of transformation (8.5) is not guaranteed to be positive. To avoid such a situation, usage if approximate inverse transform (8.7) was suggested in [**?**].

$$x = \frac{e^y}{J} \tag{8.7}$$

The RASTA and J-RASTA was intended as an extension to the original PLP method. To better perform when processing signal with noise and distortions, it leaves the idea of frame-by-frame processing and incorporates concepts that depends on overall signal characteristics. To fully utilize advantages of this parameterization, the signal processing chain should consist of additional block. For J-Rasta, the value J (8.5) should be obtained by this block. RASTA might be updated as well, by alternating trajectory filter $H(z)$ (8.1)) as was published in [**?**] for $H(z)$ in (8.3).

The approach represented by RASTA or J-RASTA might be applied on another parameterizations, e.g. MFCC.

Chapter 9

Signal Approximation

To construct a different parameterization we have used a slightly different view to the speech signal. The idea is to utilize approximative algorithm that is modified and takes into account various speech-specific features. This is idea that stands behind mentioned parameterizations. MFCC could be described as a band of filters designed according to the features of human hearings. PLP is improves tradition LPC with the equal loudness curve, the intensity-loudness power law and the critical-band spectral resolution.

The speech is understood only as a signal. In signal processing orthogonal (ad orthonormal) bases are common because they lead to efficient approximation of certain types of signal with just a few vectors [**?**].

Better approximations are obtained by choosing the M basis vectors with the respect to the signal. A signal can be presented with M parameter in an orthonormal basis by keeping M inner products with vectors chosen a priori.

In Hilbert space $\mathbf{H}$ any $f \in \mathbf{H}$ can be decomposed regarding to orthonormal base $\mathcal{B} = \{g_m\}_{m \in N}$ [**?**].

$$f = \sum_{m=0}^{+\infty} \langle f, g_m \rangle g_m \tag{9.1}$$

To get just an approximation of the function f we use only first M inner products

$$f_M = \sum_{m=0}^{M-1} \langle f, g_m \rangle g_m \tag{9.2}$$

The approximation error is

$$\epsilon_M = \|f - f_M\|^2 = \sum_{m=M}^{+\infty} |\langle f, g_m \rangle|^2 \tag{9.3}$$

The fact that

$$\|f\|^2 = \sum_{m=0}^{+\infty} |\langle f, g_m \rangle|^2 < +\infty \tag{9.4}$$

implies that error decays to zero

$$\lim_{m \to +\infty} \epsilon_M = 0 \tag{9.5}$$

However, the decay rate of ϵ_M as M increases depends on the decay of $|\langle f, g_m \rangle|$ as m increases [**?**].

The Fourier basis can approximate uniformly regular signals with few low-frequency sinusoidal waves.

$$f(t) = \sum_{m=-\infty}^{+\infty} \langle f(u), e^{i2\pi mu} \rangle e^{i2\pi mt} \tag{9.6}$$

where $\{e^{i2\pi mt}\}_{m \in \mathbf{Z}}$ is an orthonormal basis and

$$\langle f(u), e^{i2\pi mu} \rangle = \int_0^1 f(u) e^{-i2\pi mu} du \tag{9.7}$$

the approximation error is then (for differentiable functions in the sense of Sobolev) [**?**]:

$$\epsilon_M = \|f - f_M\|^2 = \int_0^1 |f(t) - f_M(t)|^2 dt = \sum_{|m| > M/2} |\langle f(u), e^{i2\pi mu} \rangle|^2 \tag{9.8}$$

To localize Fourier series approximations over intervals, we multiply f by smooth window of compact support (see (4.1) or (10.2)). Regarding this situation case, we can construct orthonormal base by replacing complex exponential by cosine functions:

$$\left\{ g_{p,k}(t) = g_p(t) \sqrt{\frac{2}{l_p}} \cos \left[\pi \left(k + \frac{1}{2} \right) \frac{t - a_p}{l_p} \right] \right\} \tag{9.9}$$

which is equivalent to segmenting $f(t)$ into several windowed components $f_p(t) = f(t) g_p(t)$. Then the approximation

$$f_{p,M} = \sum_{k=0}^{M-1} \langle f, g_{p,k} \rangle g_{p,k} \tag{9.10}$$

yields an error that depends on local regularity of f over each window support [**?**].

9.1 Linear Approximation with Multiresolution Analysis

Linear approximations of f are equivalent to finite element approximations over uniform grids. The approximation of f over the first M wavelets and scaling functions writes as

$$f_M = \sum_{j=l+1}^{J} \sum_{n=0}^{2^{-j}-1} \langle f, \psi_{j,n} \rangle \psi_{j,n} + \sum_{n=0}^{2^{-j}-1} \langle f, \phi_{J,n} \rangle \phi_{J,n} \tag{9.11}$$

this can be rewritten as

$$f_M = \sum_{n=0}^{2^{-l}-1} \langle f, \phi_{l,n} \rangle \phi_{l,n} \tag{9.12}$$

The approximation error is the energy of wavelet coefficients at scales finer than 2^l:

$$\epsilon_M = \|f - f_M\|^2 = \sum_{j=-\infty}^{l} \sum_{n=0}^{2^{-j}-1} |\langle f, \psi_{j,n} \rangle|^2 \tag{9.13}$$

The relative approximation error $\|f - f_M\|/\|f\|$ is usually the same as in case of Fourier basis.

9.2 Speech Signal Parameterization Based on Wavelets

From the point of approximation error, there is almost none difference between multiresolution analysis and Fourier series. However, signals with isolated singularities are well approximated in a wavelet basis. Wavelets take advantage of time-frequency localization property [**?**, **?**].

Transients may play important role for identifying and discriminating speech sounds, identification of transients is described in [**?**] (extraction plosives, fricatives and segment speech into 4 classes) and [**?**].

Often are wavelets uses in systems for speech recognition [**?**] where outperforms MFCC [**?**, **?**, **?**, **?**]. Mel filter bank is a mature technology. Sub-band in the mel-frequency filter bank are similar to those in wavelet decomposition, they increase logarithmically in size as the frequency increases. They are noise robust [**?**] and might improve MFCC based front-end system performance by more than 44 percent [**?**].

They often provides starting point for various parameterization, e.g. mel frequency discrete wavelet coefficients MFDWC [?]. Bark wavelet transform [?] performs better as MFCC [?], further improvement might be reached by zero-crossing and peak detector [?]. Application of wavelets described in [?] improves speech recognition improvement by 15 percent by utilizing compounding wavelets.

Various Daubechies' wavelets for speech recognition were studied [?], unfortunately classic decomposition schemes (dyadic DWT, packet wavelet WP) do not provide sufficient number of frequency bands for effective speech analysis [?]. Best wavelet for speech signal were reported Meyer [?, ?], good results might be reached also with Daubechies, Meyer, Biorthogonal, Coiflets or Symlets. Often, algorithms based on orthonormal set of the wavelet packet decomposition of the signal (local cosine packet) for several reasons. Physical model of cochlea suggests that it acts as a continuous wavelet transform in that different portions of the membrane respond to different frequency excitations logarithmically [?].

Systems based on wavelets are robust [?] regarding to the noise. successful for denoising [?, ?]. They might be successfully combined with ANN, e.g. in [?] was studied Gamma Tone Filter Bank and Wavelet Packet as front-end system for Back Propagation Neural Networks. In [?] was studied speech signal enhancement using ANN (adaline) and wavelet transform.

Also systems for feature extraction and phoneme recognition utilizing best-basis method were studied. The organize wavelets bases into a binary tree [?] or perform adaptation of wavelet packet base [?, ?]. In [?] was reported feature extraction approach based on wavelet packet entropy that is robust against noise.

9.3 Nonlinear Approximations

In Fourier and wavelet bases such a linear approximation is efficient only if the signal is uniformly regular [?] which is not the case of speech signal. Following section is an introduction to nonlinear approximations. The text is based on theory introduced in [?].

Approximating f by the first M vectors of $\mathcal{B}$ according to (9.2) is not always precise, the vectors are not necessarily the best ones to approximate f. Non-linear approximations calculate with vectors that are chosen adaptively. A further degree of freedom is introduced when the basis is chosen adaptively, according to the signal properties. Best basis outlines important signal structure and characterize their time-frequency properties.

A signal $f \in \mathbf{H}$ approximated with M vectors selected to best approximate f (adaptively, a priory) writes as

$$f_M = \sum_{m \in A_M} \langle f, g_m \rangle g_m \tag{9.14}$$

where vectors g_m belongs to adaptive basis $\mathcal{B} = \{g_m\}_{m \in N}$. Approximation f_M in (9.14) is projection of f over M vectors from $\mathcal{B}$. The difference to linear approximation (9.2) is that g_m are not taken one by one by index, but belongs to a priory determined subset A_M. Set A_M contains M vectors, so the number of approximating function is the same as for linear approximation (9.2).

Approximation error regards to $\mathcal{B}$ is the sum of the remaining coefficients.

$$\epsilon(M) = \|f - f_M\|^2 = \sum_{m \notin A_M} |\langle f, g_m \rangle|^2 \tag{9.15}$$

To minimize error $\epsilon(M)$, A_M must contain vectors that best correlate with f and have largest inner product amplitude $|\langle f, g_m \rangle|$. Resulting error is smaller than the error of linear approximation 9.3 [**?**].

It is possible to define non-linear approximation in a wavelet orthonormal basis [**?**]. Equation (9.14) rewrites as

$$f_M = \sum_{(j,n) \in A_M} \langle f, \psi_{j,n} \rangle \psi_{j,n} \tag{9.16}$$

and the approximation error is

$$\epsilon_M = \|f - f_M\|^2 = \sum_{(j,n) \notin A_M} |\langle f, \psi_{j,n} \rangle|^2 \tag{9.17}$$

The error is always smaller than the error of linear approximation [**?**]. If f is piecewise regular then it could be shown that ϵ_M has a fast decay as M increases [**?**]. The more regular is f between its discontinuities, the larger the improvement is.

9.4 Adaptive Basis Selection

Obvious method how to get A_M is to sort $\{|\langle f, g_m \rangle|\}_{m \in N}$ and apply threshold function [**?**]. We denote sorted g_m by one more index k, where

$$|\langle f, g_{m,k} \rangle| \geq |\langle f, g_{m,k+1} \rangle| \tag{9.18}$$

When applying threshold in form of

$$\theta_T(x) = \begin{cases} x & \text{if } |x| \geq T \\ 0 & \text{if } |x| < T \end{cases} \tag{9.19}$$

we obtain for non-linear approximation (with threshold T):

$$f_M = \sum_{m=0}^{+\infty} \theta_T \left(\langle f, g_m \rangle \right) g_m \tag{9.20}$$

The minimum non-linear approximation error is then

$$\epsilon_M = \|f - f_M\|^2 = \sum_{k=m+1}^{+\infty} |\langle f, g_{m,k} \rangle|^2 \tag{9.21}$$

Another possibility is to get usage of dynamical programming and minimizes a concave cost function.

Consider dictionary as union of several orthonormal bases in signal space of finite dimension N.

$$\mathcal{D} = \bigcup_{\lambda \in \Lambda} \mathcal{B}^\lambda \tag{9.22}$$

where each orthonormal basis is a family of N vectors

$$\mathcal{B}^\lambda = \{g_m^\lambda\}_{1 \leq m \leq N} \tag{9.23}$$

If we want to optimize non-linear approximation of f we have to choose M vectors from $\mathcal{D}$ (resp. $\mathcal{B}^\lambda$) that maximize $|\langle f, g_m^\lambda \rangle|$. The best non-linear approximation then writes as

$$f_M^\lambda = \sum_{m \in A_M^\lambda} \langle f, g_m^\lambda \rangle g_m^\lambda \tag{9.24}$$

The approximation error is

$$\epsilon_M^\lambda = \sum_{m \notin A_M^\lambda} |\langle f, g_m^\lambda \rangle|^2 = \|f\|^2 - \sum_{m \in A_M^\lambda} |\langle f, g_m^\lambda \rangle|^2 \tag{9.25}$$

The definition of approximation error can be directly used to compare two bases. We can say that base $\mathcal{B}^\alpha = \{g_m^\alpha\}_{1 \leq m \leq N}$ is a better basis than $\mathcal{B}^\beta = \{g_m^\beta\}_{1 \leq m \leq N}$ when

$$\epsilon_M^\alpha \leq \epsilon_M^\beta \tag{9.26}$$

this might be rewritten as

$$\sum_{m \in A_M^\alpha} |\langle f, g_m^\alpha \rangle|^2 \geq \sum_{m \in A_M^\beta} |\langle f, g_m^\beta \rangle|^2 \tag{9.27}$$

In practice, two bases are compared using a single concave function [**?**]. The cost of approximating f in basis $\mathcal{B}^\lambda$ is defined by Schur concave sum.

$$C(f, \mathcal{B}^\lambda) = \sum_{m=1}^{N} \Phi \left(\frac{|\langle f, g_m^\lambda \rangle|^2}{\|f\|^2} \right) \tag{9.28}$$

It is possible to prove [?] that $\mathcal{B}^\alpha$ is better basis than $\mathcal{B}^\beta$ for approximating f when

$$C(f, \mathcal{B}^\alpha) \leq C(f, \mathcal{B}^\beta) \tag{9.29}$$

this condition is necessary but not sufficient to guarantee the statement because (9.29) tests only a single concave function.

Coifman and Wickerhauser [?] find a best basis $\mathcal{B}^\alpha$ in $\mathcal{D}$ by minimizing the cost of f:

$$C(f, \mathcal{B}^\alpha) = \min_{\lambda \in \Lambda} C(f, \mathcal{B}^\lambda) \tag{9.30}$$

then there exist no better basis in $\mathcal{D}$ to minimize f [?]. However, often there are basis that are equivalent. The choice of the particular one then depends on function Φ.

For wavelets it is possible to utilize local cosine basis or wavelet packet. These orthonormal bases include different types of time-frequency atoms and resulting decomposition of the signal is efficient. Dictionaries of wavelet packet or local cosine bases include more than $2^{N/2}$ bases (N stands for size of the signal). The best basis minimizes the cost function

$$C(f, \mathcal{B}^\lambda) = \sum_{m=0}^{N-1} \Phi \left(\frac{|f, g_m^\lambda|^2}{\|f\|^2} \right) \tag{9.31}$$

Optimal basis according to (9.31) using brute force approach would require more than $N2^{N/2}$ operations. Fast dynamic programming algorithm presented by Coifman and Wickerhauser [?] takes advantage of the tree structure parsing and finds the best basis with $O(N \log_2 N)$ operations. The performance of best approximations depends on the time-frequency properties of f.

A wavelet packet basis divides the frequency axis into interval of varying size. Frequency tiling is made by a wavelet packet function that is translated uniformly in time. Best wavelet packet could be shortly described as a "best" frequency segmentation [?].

Best wavelet packet approximates best signals with similar energy structure - time-frequency spread. The time translation of the wavelet packet is well adapted to approximation of signal structure in this frequency range that appear at different times.

Application of wavelet packets to pattern recognition remains difficult because they bases are not translation invariant. When applied to translated signal, the minimization of cost function may yield a different basis. This remark applies to local cosine bases as well.

An example published in [?] shows signal for which is not possible to get well adapted wavelet packet basis. Signal contain different type of high energy structures located at different times u_0 and u_1 but in the same frequency interval. It is sum of four transients:

$$\begin{aligned} f(t) =& \frac{K_0}{\sqrt{s_0}}\, g\left(\frac{t-u_0}{s_0}\right) \exp(i\xi_0 t) + \frac{K_1}{\sqrt{s_1}}\, g\left(\frac{t-u_1}{s_1}\right) \exp(i\xi_0 t) \\ &+ \frac{K_2}{\sqrt{s_1}}\, g\left(\frac{t-u_0}{s_1}\right) \exp(i\xi_1 t) + \frac{K_3}{\sqrt{s_0}}\, g\left(\frac{t-u_1}{s_0}\right) \exp(i\xi_1 t) \end{aligned} \tag{9.32}$$

Function g is smooth window whose energy is concentrated at low frequencies. Fourier transform of f shows that the energy of function is concentrated in frequency band centred at ξ_0 and ξ_1

$$\begin{aligned} \hat{f}(\omega) =& K_0\sqrt{s_0}\, \hat{g}\left(s_0\, (\omega-\xi_0)\right)\, \exp(-iu_0(\omega-\xi_0)) \\ &+ K_1\sqrt{s_1}\, \hat{g}\left(s_1\, (\omega-\xi_0)\right)\, \exp(-iu_1(\omega-\xi_0)) \\ &+ K_2\sqrt{s_1}\, \hat{g}\left(s_1\, (\omega-\xi_1)\right)\, \exp(-iu_0(\omega-\xi_1)) \\ &+ K_3\sqrt{s_0}\, \hat{g}\left(s_0\, (\omega-\xi_1)\right)\, \exp(-iu_1(\omega-\xi_1)) \end{aligned} \tag{9.33}$$

The time and frequency spread of the transients depends on values of s_0 and s_1. This is illustrated in figure 9.1. The best wavelet packet is adapted to the transient of highest energy, the energy of the smallest transient is then spread across many wavelet packets.

The example of signals whose energy distribution schemes change rapidly in time and thus are not suitable for decomposition based on wavelet packets are speech signals. Two different distants in the same frequency neighbourhood might have very different energy distributions. A best wavelet is not adopted to such a variation and thus gives poor non-linear approximations [?]. Similar holds for natural scene images, although for specific class of images such as fingerprints, it is possible to find wavelet packet that outperforms the wavelet basis [?].

A local cosine bases are based on the idea similar to wavelet packet. But instead of frequency division, they divide time axis into intervals of varying sizes. To obtain best basis, one must adopt that time segmentation to the variation of the signal time-frequency structures. The price paid is loss of frequency flexibility - time and frequency bounds are reversed when comparing to wavelet packet. A best local cosine basis is then adapted to signal which includes structures of very different time and frequency spread at any given time.

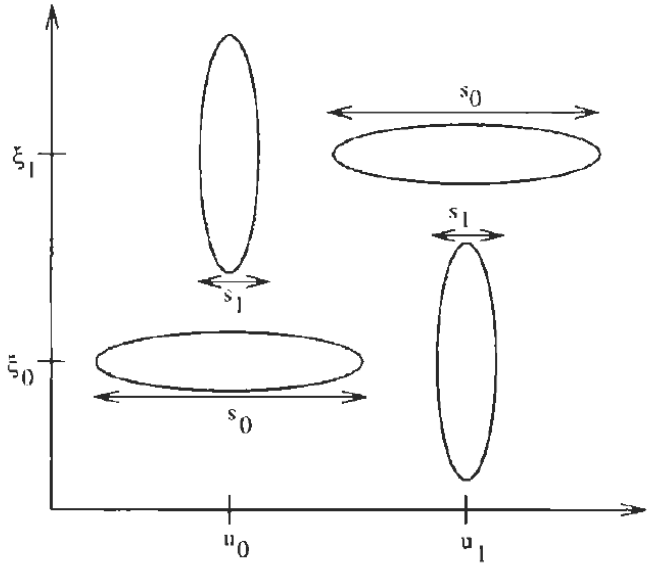

Figure 9.1: The time and frequency spread of the transients from (9.32) (reprinted from [**?**])

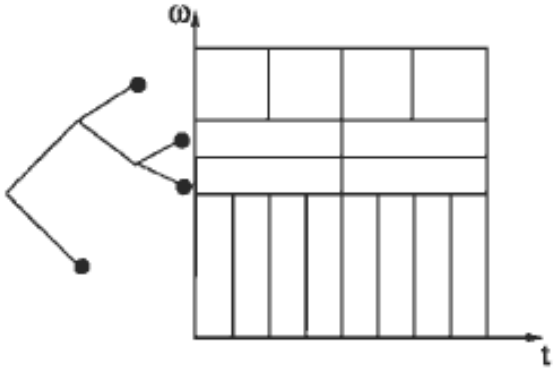

Figure 9.2: The wavelet packet (on left) determines frequency spread of Heisenberg boxes (right) (reprinted from [**?**])

The sum of four transients (9.32) is not efficiently represents in a wavelet packet but neither in a best local cosine basis. Since the scales s_0 and s_1 are very different and signal contains two transients at frequencies ξ_0 and ξ_1 that have different frequency spread and are located in different time u_0 and u_1. The size of the window is adapted to the transient of highest energy and the energy of the second transient is spread across best basis vectors.

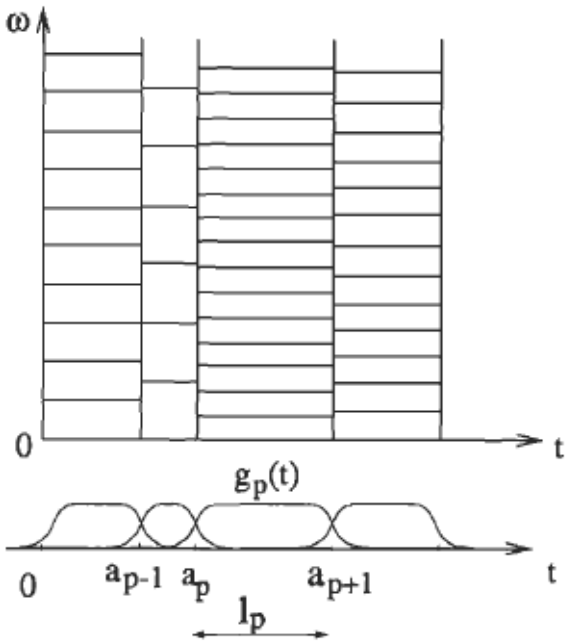

Figure 9.3: Time-frequency spread (Heisenberg boxes) of local cosine base vectors (reprinted from [**?**])

9.5 Approximation with Pursuits

The set of non-orthonormal bases is much larger than the set of orthonormal bases. If we would take an advantage of approximation in best basis sense, we must introduce approach that deals with the complexity.

Let $\mathcal{D} = \{g_p\}_{0 \leq p < P}$ is a redundant dictionary of P vectors. Consider P is higher ($P > N$) than the size of space of signals N. For any $M \geq 1$, an approximation f_M of f might be calculated as a linear combination of any M vectors from dictionary $\mathcal{D}$:

$$f_M = \sum_{m=0}^{M-1} a[p_m] g_{p_m} \tag{9.34}$$

The vector from $\mathcal{D}$ might be chosen freely. For dictionaries of $P > N$ vectors, computing the approximation f_M is an NP-hard problem [**?**]. To solve approximation that minimizes $\|f - f_M\|$, there is no known polynomial algorithm.

Pursuit algorithms are fast, greedy algorithms that generalize these adaptive approximations and reduce computational complexity. Approximative vectors are selected from redundant dictionaries of time-frequency boxes with no orthogonally constraints. The price paid is that the approximation is not optimal, but only sub-optimal. The relative good approximation is provided with $O(N^{3.5} \log_2^{3.5} N)$ operations.

Basic Pursuit performs decomposition of f to best basis $\mathcal{B}$ by (9.34). If restriction to orthonormal bases is applied, then the basis would be optimized by minimizing (9.35).

$$C(f, \mathcal{B}) = \sum_{m=0}^{N-1} \Phi \left(\frac{|a[p_m]|^2}{\|f\|^2} \right) \tag{9.35}$$

where Φ is concave. This result does not hold for general (non-orthogonal) bases. Basic Pursuit searches for a best basis that minimizes (9.35) for $\Phi(x) = x^{1/2}$:

$$C(f, \mathcal{B}) = \frac{1}{\|f\|} \sum_{m=0}^{N-1} |a[p_m]| \tag{9.36}$$

Minimizing the I^1 norm of the decomposition coefficients avoids diffusing the energy of f among many vectors [**?**]. Minimization procedure reduces cancelation between the vectors $a[p_m]g_{p_m}$. The cancelation increase the cost (9.36) by increasing $|a[p_m]|$. The minimization of an I^1 norm is related to linear programming that leads to fast computation algorithms [**?**]. Theory of interior points had led to a large collection of algorithms. The approach is summarized in [**?**].

Basic Pursuit algorithm is relatively computationally extensible. The reason is that the algorithm minimizes a global cost function over all dictionary vectors.

Chapter 10

Matching Pursuit

Matching pursuit provides parameterization that expands the signal into waveforms whose time-frequency structures are adapted to the local signal structure. Compared to Basic Pursuit algorithm, matching pursuit algorithm utilized a greedy strategy to reduce computation complexity. The algorithm was proposed and fully described in [**?**, **?**], this section only gives a short introduction and discuss application to speech analysis.

Previously described parameterizations like where the time-frequency characteristics of speech were taken into account prior to definition of parameterization, the structure of the signal parameterized by matching pursuit is derived by algorithm. Time-frequency characteristic of particular utterance is derived after parameterization and it is specific for each utterance. Adoption to the local signal structure allows tuning method for tracking particular features in the speech.

Important advantage of parameterization based on matching pursuit algorithm is that it is suitable for analysis and comparison of utterances without any need for preceding segmentation of the signal. The utterance is analysed at once and corresponding atoms (see section 10.1) are chosen to represent important features in signal. It is possible to fine-tune the parameterization to emphasize particular features. This might be also reached by proper selection of atoms within dictionary.

The main problem with utilization of matching pursuit is the computational complexity. To provide spatial representation of the signal is necessary to pass through all the functions in the dictionaries.

The advantages of matching pursuit are utilized in signal, image and video coding: [**?**, **?**], shape representation and recognition [**?**], 3D objects coding [**?**] and biomedical signal analysis (EEG) [**?**, **?**, **?**, **?**, **?**] and (ECG) [**?**, **?**, **?**, **?**, **?**, **?**, **?**].

10.1 Matching Pursuit Algorithm

Matching pursuit algorithm [**?**] transforms any signal f from Hilbert space into a linear expansion of waveforms $g(t)$. 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 $g_n(t)$ (10.1). Approximation f_N of f is given by (10.1), where α_n are scalar coefficients (see (9.34)).

$$f(t) \cong \tilde{f}_N(t) = \sum_{n=0}^{N-1} \alpha_n g_n(t) \tag{10.1}$$

Approximation $\tilde{f}_N$ of a signal by the functions from a suitable dictionary often gives better representation compare to transformations based on unitary basis [**?**].

The redundant and over-complete set of time-limited functions g_n is called dictionary $\mathcal{D} = \{g_n\}_{n\in\Gamma}$. Functions $g_n(t)$ are called atoms. The choice of content of a dictionary (functions) is arbitrary. A dictionary might be adjusted to the particular application by choice of atoms (as was discussed in chapter 9.3).

Although atoms might be of arbitrary choose [**?**], often Gabor dictionaries (10.2) are utilised [**?**]. The dictionary is constructed by modulating, translating and scaling a Gauss window $\hat{g}(t) = 2^{1/4}\exp(-\pi t^2)$:

$$g(t) = \frac{1}{\sqrt{s}}\,\hat{g}\left(\frac{t-u}{s}\right) e^{i\xi t} \tag{10.2}$$

Resulting Gabor dictionary is time and frequency translation invariant modulo period of a discrete Gauss window N.

Matching pursuit begins by projecting f on vector (atom) $g_0 \in \mathcal{D}$:

$$f = \langle f, g_0\rangle\, g_0 + Rf \tag{10.3}$$

Because g_0 is orthogonal to Rf

$$\|f\|^2 = |\langle f, g_0\rangle|^2 + \|Rf\|^2 \tag{10.4}$$

Atom $g_0 \in \mathcal{D}$ has to be chosen so that $|\langle f, g_0 \rangle|$. This selection minimizes $\|Rf\|$. To further decrease computation demands, this operation is usually replaced by finding vector that is almost optimal so

$$|\langle f, g_0 \rangle| \geq \alpha \sup_{\gamma \in \Gamma} |\langle f, g_\gamma \rangle| \tag{10.5}$$

where $\alpha \in (0, 1]$ is an optimality vector.

Next iteration of the algorithm repeats this procedure to decompose residue. Assume that $f = R^0 f$ and that the m-th order residue is already computed. The next iteration chooses $g_{\gamma_m} \in \mathcal{D}$ so that

$$|\langle R^m f, g_{\gamma_m} \rangle| \geq \alpha \sup_{\gamma \in \Gamma} |\langle R^m f, g_\gamma \rangle| \tag{10.6}$$

and projects $R^m f$ on g_{γ_m}:

$$R^m f = \langle R^m f, g_{\gamma_m} \rangle g_{\gamma_m} + R^{m+1} f \tag{10.7}$$

The orthogonality of g_{γ_m} and R^{m+1} implies that

$$\|R^m f\|^2 = |\langle R^m f, g_{\gamma_m} \rangle|^2 + \|R^{m+1} f\|^2 \tag{10.8}$$

(10.7) implies, when summing m between 0 and $M - 1$:

$$f = \sum_{m=0}^{M-1} \langle R^m f, g_{\gamma_m} \rangle g_{\gamma_m} + R^M f \tag{10.9}$$

similarly (10.8) gives:

$$\|f\|^2 = \sum_{m=0}^{M-1} |\langle R^m f, g_{\gamma_m} \rangle|^2 + \|R^M f\|^2 \tag{10.10}$$

It can be proven that $\|R^m f\|$ converges exponentially to 0 when m tends to infinity [**?**]. With increasing size of the signal space N, the convergence rate decreases. An infinite number of iterations is necessary to completely reduce the residue, even in finite dimensions [**?**]. In most signal processing application, only sufficiently precise approximation of the signal is needed, thus only N iterations are performed.

Matching pursuit with a relative precision ϵ is implemented with the following steps [**?**].

1. Initialization. Set $m = 0$ and compute $\{\langle f, g_\gamma \rangle\}_{\gamma \in \Gamma}$
2. Best match. Find $g_{\gamma_m} \in \mathcal{D}$ such that

$$|\langle R^m f, g_{\gamma_m} \rangle| \geq \alpha \sup_{\gamma \in \Gamma} |\langle R^m f, g_\gamma \rangle| \tag{10.11}$$

3. Update. For all $g_\gamma \in \mathcal{D}$ with $\langle g_{\gamma_m}, g_\gamma \rangle \neq 0$

$$\langle R^{m+1} f, g_\gamma \rangle = \langle R^m f, g_\gamma \rangle - \langle R^m f, g_{\gamma_m} \rangle \langle g_{\gamma_m}, g_\gamma \rangle \tag{10.12}$$

4. Stopping rule. The run might be stop after certain number of iterations or as well some condition is fulfilled, e.g.

$$\|R^{m+1} f\|^2 = \|R^m f\|^2 - |\langle R^m f, g_{\gamma_m} \rangle|^2 \leq \epsilon^2 \|f\|^2 \tag{10.13}$$

It is possible to reduce directory and include only atoms that matches with signal. This helps to reduce computation in steps 2 and 3. The sub-directory $\mathcal{D}_\int$ is constructed in the way that all function in a sub-directory $\mathcal{D}_\int$ where $\mathcal{D}_\int = \{g_\gamma\}_{\gamma \in \Gamma}$ maximizes $|\langle f, g_\gamma \rangle|$. Selection of sub-directory - choice of the particular directory strongly depends on the application for which is intended.

It is also possible to perform orthonormalization of dictionary [**?**]. The advantage is then that matching pursuit converges quickly - when the number of iterations gets close to N, the residues of orthogonal matching pursuit (resp. its norms) decrease faster than for non-orthogonal matching pursuit. In practice this is seldom used, since the orthonormalization is an expensive operation.

To best fit the function being approximated, atoms (10.2) 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 (see (10.11) - (10.13)).

Function $\hat{g}$ is Gaussian window (10.2), equation for discrete variant with the length of T samples is in (10.14). Parameter σ influences a shape of the window. The range of σ is given as $\sigma \leq 0.5$. Time plot of discrete window is in figure 10.1.

$$\hat{g}[t] = e^{-\frac{1}{2}\left(\frac{t-(T-1)/2}{\sigma(T-1)/2}\right)^2} \tag{10.14}$$

Figure 10.3 illustrate signal made of two different atoms. Atom g_1 (top pane, right) and g_2 (top pane, left) are concatenated in a simple signal. Atom g_1 starts at $t_0 = 0.1$ s, has length 0.3 s, amplitude = 500, $\xi = 50$ Hz and $\sigma = 0.22$, atom g_2 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 $g_1 + g_2$ is in the bottom of figure 10.3, the plot is limited to maximal frequency $\xi = 400$ Hz.

The expansion maintains energy, which guaranties convergence of the algorithm [**?**]. 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) writes as (10.15).

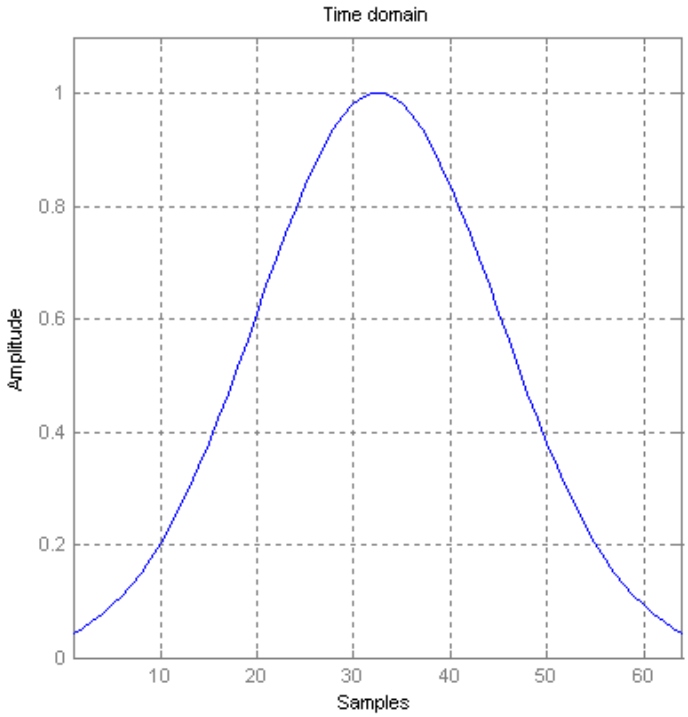

Figure 10.1: Gauss window in time plot (N=60)

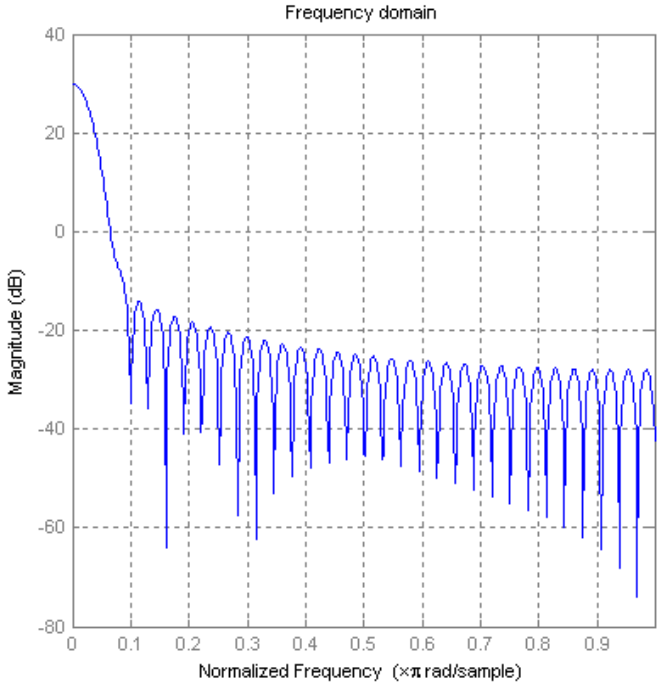

Figure 10.2: Side lobe plot for Gauss window

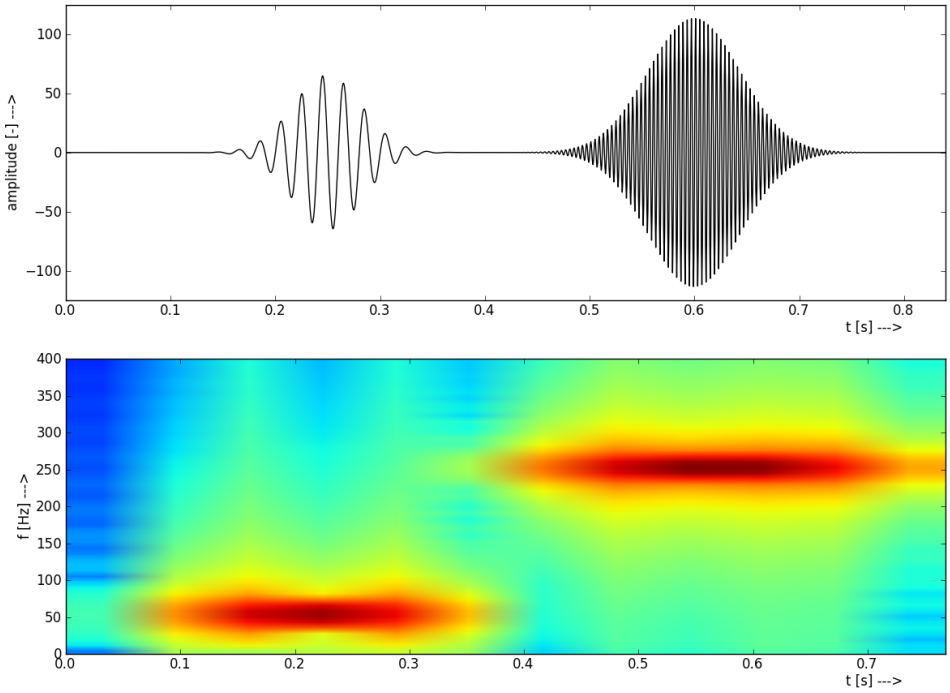

Figure 10.3: Two atoms and corresponding spectrogram - x axes (time) are in equal scale

$$f(t) = \tilde{f}_N(t) + R^N f = \sum_{n=0}^{N-1} \langle R^n f, g_n \rangle g_n + R^N f \tag{10.15}$$

During N-th iteration of algorithm (10.11) - (10.13), the approximation $\tilde{f}_{N-1}(t)$ of the signal is improved by adding an atom g_N for which has inner product with residual signal $R^N f$ minimal square error (10.16).

$$\max \left[\langle R^N f, g_n \rangle \right]_{n=0}^{N} \to g_N \tag{10.16}$$

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 writes as vector γ (10.17).

$$g_n(t) = \alpha_n \frac{1}{\sqrt{s_n}} g\left(\frac{t - u_n}{s_n}\right) e^{i\xi_n t} \mapsto \gamma_n = (\alpha_n, u_n, \xi_n) \tag{10.17}$$

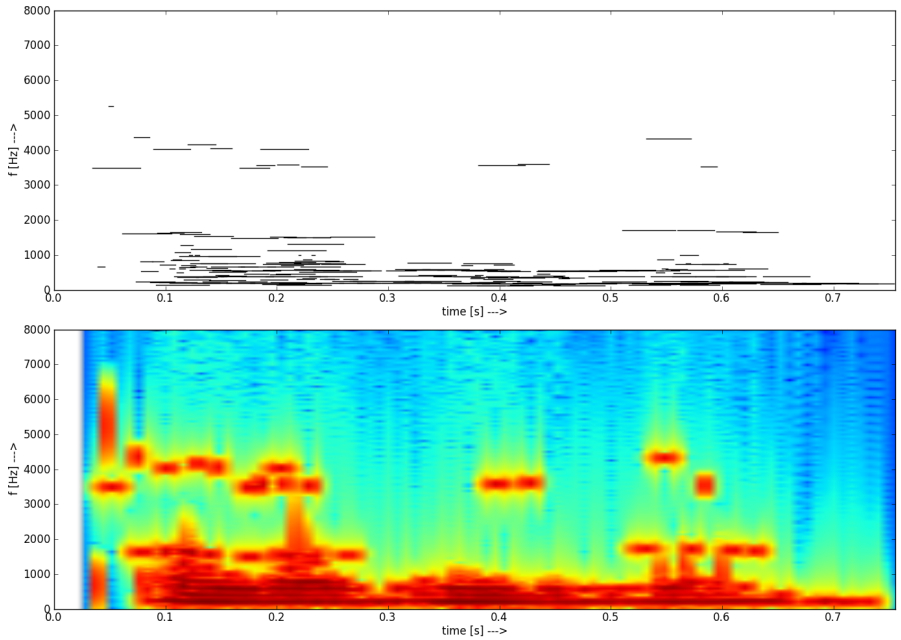

Figure 10.4: Approximation $\tilde{f}_{250}(t)$ of utterance "televize"– television ($N = 250$ atoms)

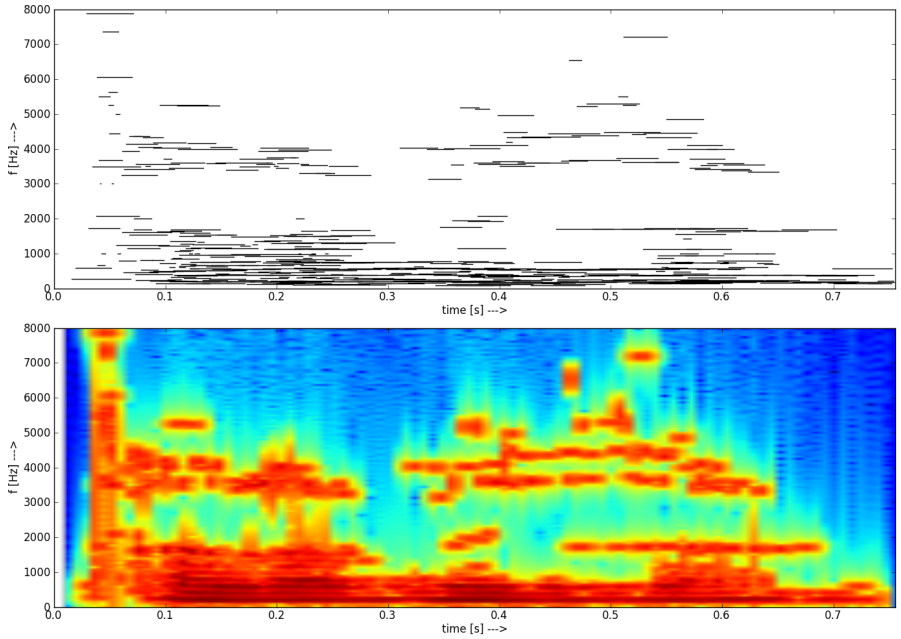

Figure 10.5: Approximation $\tilde{f}_{500}(t)$ of utterance "televize"– television ($N = 500$ atoms)

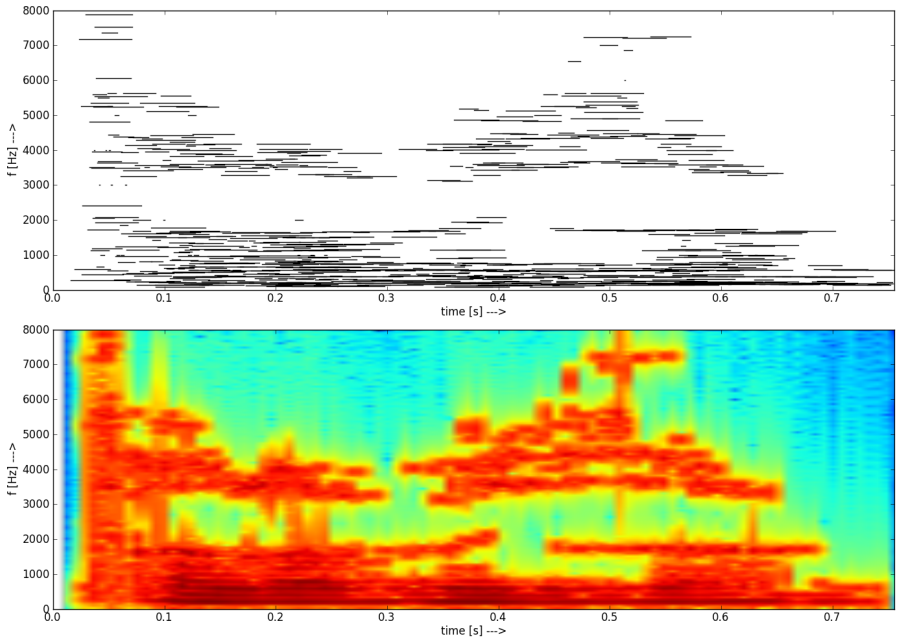

Figure 10.6: Approximation $\tilde{f}_{750}(t)$ of utterance "televize"– television ($N = 750$ atoms)

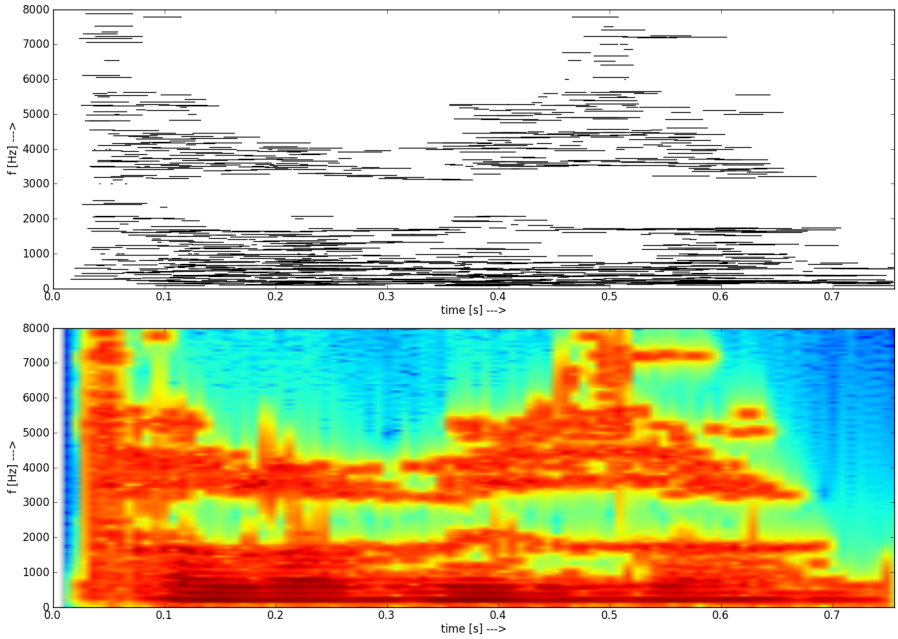

Figure 10.7: Approximation $\tilde{f}_{1000}(t)$ of utterance "televize"– television ($N = 1000$ atoms)

Matching pursuit decomposes real signals by grouping atoms $g_{\gamma+}$ and $g_{\gamma-}$ from Gabor dictionary. Indices $\gamma_{\pm} = (\alpha_n, \pm u_n, \xi_n)$. In each step, instead of projecting R^n f over and atom g_γ, algorithm computes its projection on the plane generated by $g_{\gamma+}$ and $g_{\gamma-}$.

For better illustration of the algorithm are in figures 10.4, 10.5, 10.6 and 10.7. spectrograms obtained for different approximations ($\tilde{f}_{250}(t)$, $\tilde{f}_{500}(t)$, $\tilde{f}_{750}(t)$ and $\tilde{f}_{1000}(t)$) 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 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 simple overview of the density of atoms approximating the utterance. In this figure no considerations about bandwidth are taken into account.

The approximations $\tilde{f}(t)$ (figure 10.4 to figure 10.7) were obtained by the matching pursuit algorithm without any further modifications. Atoms that approximate the utterance were determined by iterating process described by equation (10.15). The process was set to decompose the signal to 250, resp. 500, 750 and 1000 atoms.

In contradiction to common parameterizations used in the field of speech processing (e.g. MFCC, PLP) atoms g_n obtained can be used to synthesize back the approximated signal $\tilde{f}_N(t)$. Spectrograms constructed for $\tilde{f}(t)$ are at the bottom of respective figures. Spectrograms of approximation may be compared to the spectrogram of original signal $f(t)$ in figure 10.8.

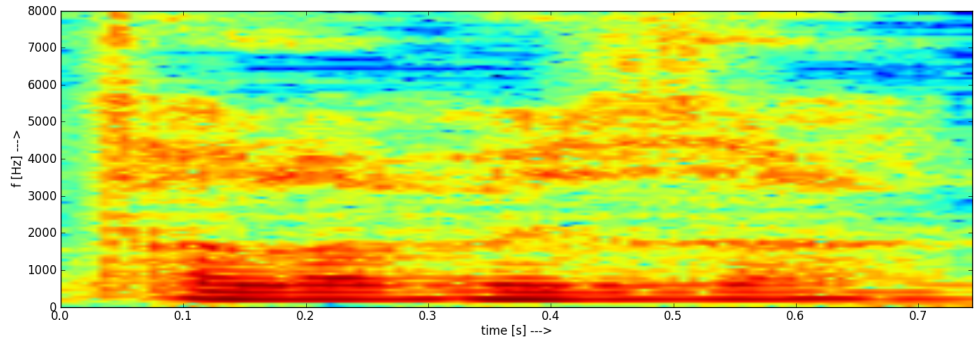

Figure 10.8: Spectrogram of utterance "televize"– television (source signal $f(t)$)

Comparison of figures 10.4 to 10.7 gives an overview of the decomposition advance during algorithm progress. As could be easily noticed, $\tilde{f}_{1000}(t)$ in figure 10.7 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 of energy, the approximation

of a speech signal suffers with one unpleasant consequence where the approximations results in coverage that does not approximate equally all parts of spectra, but preferring the lower frequencies that carry more energy as could be observed in the figures 10.4 to 10.7. 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 and iterations.

To better illustrate spectral composition of approximated signal $\tilde{f}_N(t)$ during algorithm iterations, in figure 10.9 are shown histograms of four approximations that differ in number of atoms N. As could be observed from the histograms, approximation of higher-frequency parts is being more precise with increasing number of approximating atoms. This effect is caused by combination of previously mentioned feature of matching pursuit 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 negatively adverse classification.

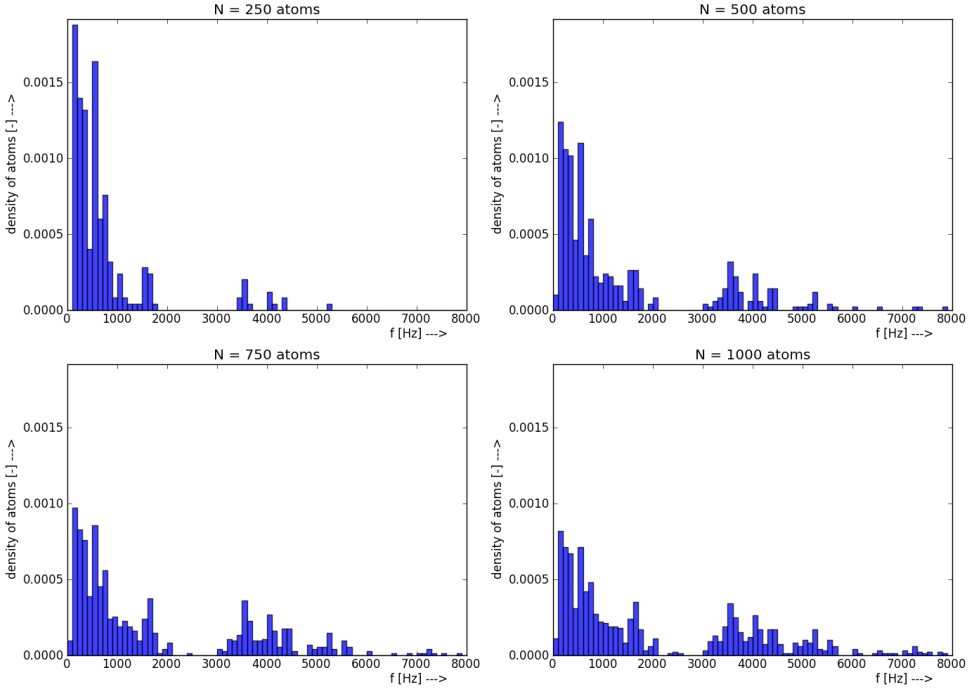

Figure 10.9: Frequency distribution of atoms for approximations consisting of different number of atoms N (normed histograms, the scale of axes is equal)

Chapter 11

Matching Pursuit for Speech Parameterization

In order to perform various analyses on the signal, the amount of data has to be reduced while maintaining important characteristics. To optimally describe and measure differences in speech, the parameterization has to respect subtle differences between patterns occurring in the signal that usually represent word with close meanings [**?**].

Common parameterizations like LPC, PLP and MFCC are adapted to specific features present in human speech. Unfortunately these parameterizations consider only speech without any significant disorder. As was experimentally proven (see chapter 13), when disorders have to be taken into account, utilization of parameterization that is less optimized is inevitable.

Low level signal representation must provide explicit information on very different properties while giving simple cues to differentiate close patterns. This concept is important for natural speech processing and parameterizations follows it. When dealing with speech of children with disorders, the importance of the concepts is even higher. Flexible decomposition is important for characterizing patterns that vary in time and frequency.

The variance of pattern for children suffering developmental dysphasia is higher compare to the patterns that are found in the speech of healthy children. To correctly determine the progress of treatment, it is important to correctly track these patterns and characterize differences. Ability to track subtle differences becomes even more important when the progress of one speaker is to be determined.

The MP algorithm might be as well adapted to better deal with a speech signal. Matching pursuit algorithm has been extended to avoid the consequences of the effect of prioritizing low parts of spectra as was described in chapter 10.1. Another technique that helps to deliver better results consider-

ing speech signal is alternating of matching criterion (10.16). In general, the alternation might be written in form of frequency-dependent weighting function h (11.1). 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.

$$\max \left[h \left(\langle R^n f, g_n \rangle \right) \right]_{n=0}^{N} \rightarrow g_n \tag{11.1}$$

Function h in (11.1) has to be chosen with respect to specific features of human auditory system. As an example, techniques implemented in PLP might be used. Function h then could include concepts of equal loudness curve, intensity-loudness power law and critical-band spectral resolution. This will improve results obtained when parameterising speech in general. However presented form is not suitable for parameterization of speech with disparities, especially speech od children with severe form of DD. As is discussed in chapter 13, PLP doesn't outperform LPC when dealing with the children with DD.

Discussion about adaptation to energy distribution scheme of signal discussed in chapter 9.3 leads to solution that allows setting up the areas within the signal. An area in signal is determined by vector $(t_{start}, t_{stop}, f_{start}, f_{stop})$. Definition of several areas makes a tiling scheme on signal. Tiling scheme does not to cover whole time-frequency plane, partial coverage is allowed. Matching pursuit is then run separately on each of the area. Results - atoms found in areas are then gathered and represent the signal.

Size of an area is limited to be at least equal to Heisenberg box. Maximum size is limited by signal itself (we consider implementation with isolated words) or by computation demands. In case of the infinite signal (streaming of voice) or for signal with significant length, it is recommended to split the signal into time frames. Splitting might be based on voice activity detector (VAD).

Since the application considered in this thesis is analysis of the utterances where these utterances are stored as separated words in database, splitting in time is no longer considered. Only frequency splitting is assumed without detriment to generality.

The idea of splitting a signal into frequency bands and then analyse these band independently comes from MFCC algorithm (see chapter 5). The extension relies on definition of M non-overlapping frequency bands. The decomposition of the signal is performed for each of M bands separately. 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 recommendation in [**?**]: 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 corresponds with characteristics and properties of human hearing. Definition of bands is in table 11.1.

Lower frequency f_L [Hz]	**Upper frequency f_H [Hz]**	**Central frequency f [Hz]**	**Bandwidth B [Hz]**
0	100	50	100
100	200	150	100
200	300	250	100
300	400	350	100
400	510	450	110
510	630	570	120
630	770	700	140
770	920	840	150
920	1080	1000	160
1080	1270	1170	190
1270	1480	1370	210
1480	1720	1600	240
1720	2000	1850	280
2000	2320	2150	320
2320	2700	2500	380
2700	3150	2900	450
3150	3700	3400	550
3700	4400	4000	700
4400	5300	4800	900
5300	6400	5800	1100
6400	7700	7000	1300
7700	9500	8500	1800
9500	12000	10500	2500
12000	15500	13500	3500

Table 11.1: Frequency band definition (according to [**?**])

The ceiling of 15500 Hz is sufficient despite the fact that critical frequency $f_s/2$ is higher. Atoms with $\xi \geq 15500$ Hz are rarely found it the approximation of a speech signal. This extension to the original algorithm helps to balance the content of the spectra in favour of higher-frequency components neglected by the original algorithm.

The overview of the results of decomposition based on frequency bands is in figure 11.1. The layout of the figure 11.1 is the same as for figures 10.4 to 10.7. Upper part shows distribution of atoms g_n in time-frequency plane (top), lower part contains spectrogram of approximation $\tilde{f}'_N(t)$ (bottom).

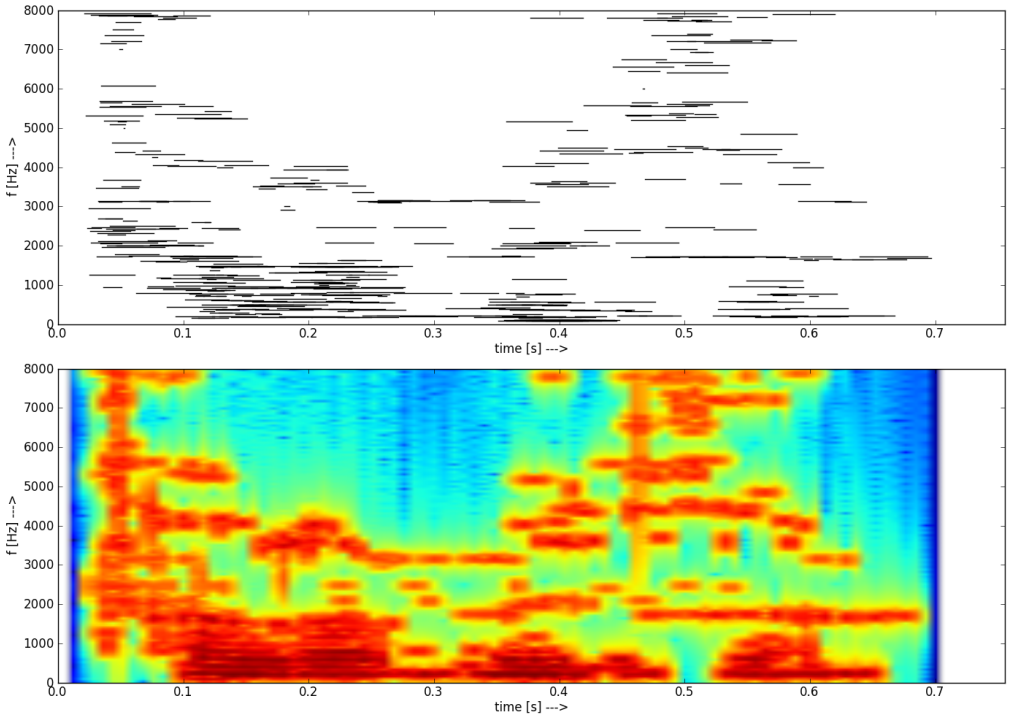

Figure 11.1: Approximation $\tilde{f}'_{500}(t)$ of utterance "televize"– television, with atoms ($N = 500$) distributed over whole frequency bands

Figure 11.2 shows spectral composition of approximated signal $\tilde{f}_N(t)$ for the original (left) and $\tilde{f}'_N(t)$ modified algorithm (right). On the left is approximation according to the original algorithm given by (10.15), right figure shows the distribution of atoms when modified algorithm was utilised. Both approximations were calculated for 500 atoms (N=500).

The extension helps to obtain approximation with balanced spectral components. This helps in classification task dealing with utterances of healthy children and children with developmental dysphasia.

The number of atoms N is chosen prior to decomposition, often an estimate is done on empirical basis. MP allows performing analysis with excess number of atoms, performing analysis of power and distribution of atoms and

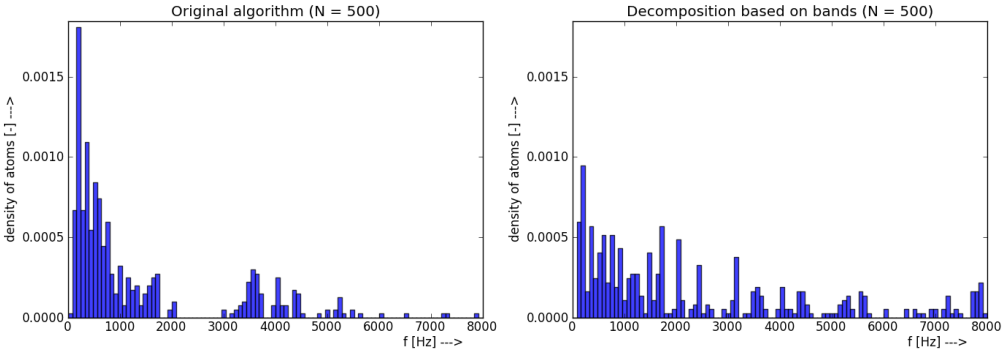

Figure 11.2: Comparison of histograms for approximation of utterance "televize"– television obtained by original algorithm (left – $\tilde{f}_{500}(t)$) and algorithm based on frequency bands ($\tilde{f'}_{500}(t)$ – right) – normed histograms, the scale of axes is equal

then shrink the number of atoms accordingly. The iterative process of decomposition also might be stopped due to some criterion, e.g. based on energy of the residuum $|R^N f|$ (see (10.13)).

Approximation of a signal in the terms of bands allows performing comparison based on these bands. Comparison based only on bands is not sufficient for the task. Classification based on KSOM delivers better results since it involves the overall utterance (see experiment described in chapter 14).

Variability of KSOM allows to create maps that are adapted to particular speakers and perform the classification regarding specific feature of these speakers. This is helpful when dealing with task of classification of children with DD according to degree of disease. Overall classification is then based on several classificators where each is trained on speech with different degree of impairment or different utterances.

Updates to original matching pursuit algorithm described in [**?**, **?**] allows to utilize this algorithm in task of classification of speech. Analysis is performed on variable time-frequency boxes instead of the whole signal. In general, this helps to adjust the analysis to the nature of the signal. For classification of isolated utterances (words) the scheme based only on frequency bands proved to be sufficient (see experiment described in chapter 14 and 15).

Chapter 12

Speech Classification Based on Artificial Neural Network

Natural data are not always describable by low-order (first and second order) statistical parameters; their distributions are non-Gaussian; and their statistics are non-stationary [**?**]. The functional relations between natural data elements are often nonlinear. Under these conditions neural-network computing methods are more effective and economic than traditional ones. Neural networks are often suitable for nonlinear estimation and control tasks in which the classical probabilities methods fail [**?**].

Only ANN-based models rely on redundancy of representations in space and time. Such models are able to describe features present in the input vectors as well as catch the relationship between features. ANN has ability to ignore individual appearing features that are deviated from the standard set. In other words ANN concentrates on collective properties and neglects the role of individual signals and patterns which is advantageous for speech classification.

Specifically this feature helps in classification of speech signals that is imperfect (i.e. contains noise and various artefacts). The perception of speech-like sounds is dependent on the preceding sound, namely, that it depends on the spectral difference between the current sound and the preceding sound [**?**]. Models will describe the sounds and the difference between models is proportional to the spectral difference which is in distance between models.

12.1 Kohonen Self-Organizing Maps

Kohonen Self-Organizing Maps are artificial neural networks trained using unsupervised learning algorithm. Training algorithm forms representation of the input space called that represents distribution function of input vectors - the

topological properties of the input space. There are two opposing tendencies in the self-organizing process (during performing training algorithm). First, the set of weight vectors tends to describe the density function of the input vectors. Second, local interactions between processing units tend to preserve continuity in sequences of weight vectors. The reference vector distribution tending to approximate a smooth hyper surface and introduce a kind of features that describes the overall set when conserving and generalising all important qualities - features [**?**].

The main applications of KSOM are in the visualization of complex data in a two-dimensional display and creation of abstraction like in many other clustering techniques. The mapping is ordered and descriptive of the distribution of input vectors. The collection of models is ordered by definition, if each model is equal to the average of input data mapped to its neighbourhood. KSOM is a kind of nonlinear projection of the probability density function of the high-dimensional input data vectors onto the two-dimensional display. Another dimension of output space is possible, however more often is two-dimensional projection.

We make use of these properties and prepare classificator based on KSOM. It allows to process complex data gathered by analysis of speech signal and classify the speaker (see experiments in chapter 14). The ability to generalize allows constructing parameterization that is redundant and let the KSOM to choose collective (and more important) features. This specifically ability allows features presented in the speech to be expressed implicitly (as numbers) and as well explicitly (as a distribution of vectors in input set)

KSOM is an effective platform for visualization of high-dimensional data. This helps to fully understand contents of a data set and it is a vital to fully understand contents exploit properties of data set. KSOM allows transforming whole input set containing overwhelming number of data to a small set of features vector. For example, a set of word pronounced by healthy children containing several tens of utterances might be transformed to a set of several (tens) of vectors - features [**?**]. This compact representation is still sufficient to converse all important features within speech and comparable with similar set of features extracted in a similar manner without losing important information (e.g. the speaker with unknown status might be then classified as healthy or unhealthy).

The visualization feature of KSOM helps to verify results obtained from numerical analysis. During our cooperation with specialist physicians we found out that several of them prefer form of figures (two-dimensional U-map) to the table of numbers. The characteristics of one speaker do not have to be expressed as a set of numbers (relatively small, to allow human compare various set in between), but might be expressed as well as picture. We gained experience that is possible to utilize figures instead of numbers just only after a short

introduction that shows several cases and comment them. Two-dimensional grid allows visual representation and interpretation of the clusters. Clustering is a way of extraction the most important features from trained KSOM [**?**]. For example see experiment described in chapter 13. Here the clustering based on k-means is utilized to obtain high level of abstraction over speech signal [**?**]. Every cluster represent different main feature, these features are isolated and then compared.

Utilization of KSOM brings another advantage to the processing of natural language. Since we have to record all the recording not in studio, the recordings contains a lot of noise and other artefacts. KSOM ability to concentrate on collective features allows using these recordings without any preceding modification (e.g. denoising). The most common features presented are emphasized whereas seldom occurring non-speech signals are ignored and has no influence on the overall classification.

Ability to concentrate on general features prioritizes KSOM. Several studies dealing with competing Hidden Markov Models (HMM) on KSOM were published (e.g. [**?**]). For several applications there are exist modification to the KSOM algorithm, like Deep Neural networks (DNNs) that have many hidden layers and are trained using new method have been shown to outperform Gaussian mixture models (GMMs) on a variety of speech recognition benchmarks, sometimes by a large margin [**?**].

On the pure form, the SOM defines an elastic net of points (parameter, reference or codebook vector) that are fitted to the input signal space to approximate its density function in an ordered fashion [**?**].

12.2 KSOM Training Algorithms

Kohonen describes two algorithms for training KSOM [**?**]. Firstly was introduced iterative algorithm that performs training by utilizing vectors from training set on one-by-another basis. Second, improved, algorithm is based on performing mean operation on the subset of the training vectors (batch training).

The iterative algorithm performs following steps for each of the vector from training set:

1. Initialization of map M - random initialization is suggested as the best and also the fastest policy. It is strongly recommended to use it in practice.

2. Take a vector from training set (denoted as x) and find best-matching model (neuron) m_c (also referred as winner) from all the neurons m_i

in the map (M is number of the neurons in the net). Each time a new vectors is taken to perform following steps with, the epoch (discrete-time coordinate) denoted as t increments. Best-matching node is defined to have the smallest Euclidean distance d from the vector.

$$d = \|x - m_i\| \tag{12.1}$$

so we are looking for neuron m_c that satisfies (12.2)

$$c = \operatorname{argmin}(\|x - m_i\|) \tag{12.2}$$

this might be rewritten as

$$\|x - m_c\| = \min\{\|x - m_i\|\} \tag{12.3}$$

3. update value of the winner (m_c) for next epoch

$$m_i(t+1) = m_i(t) + h_{ci}(t)\,[x(t) - m_i(t)] \tag{12.4}$$

where h_{ci} acts as so-called neighbourhood function, a smoothing kernel defined over the lattice points.

For convergence it is necessary that $h_{ci} \to 0$ when $t \to \infty$. One of the neighbourhood kernels is in (12.5).

$$h_{ci}(t) = \alpha(t) \exp\left(-\frac{\|r_c - r_i\|^2}{2\sigma^2(t)}\right) \tag{12.5}$$

where $\alpha(t)$ is scalar-values learning-rate factor. Parameter $\sigma(t)$ defines the width of the kernel - it corresponds to the radius of the neighbourhood. Both $\alpha(t)$ and $\sigma(t)$ are monotonically decreasing functions of time. Their settings have major impact on the results of the training, improper settings (too quick decreasing) might result in malformation of the map.

The number of epochs is multiple to the number of vectors in the training set. To guarantee generalization it is desired that each of the training vectors inputs several times into training process with random order, so $t \gg M$. This is because learning is a stochastic process and the final statistical accuracy of the mapping depends on the number of steps in the final convergence phase. The phase must be reasonably long, so the selection of optimal $\alpha(t)$ is required to both conserve statistical accuracy and minimize learning time.

The iterative algorithm was studied first (see experiment described in chapter 13), and later was abandoned for the batch algorithm. The batch algorithm provides comparable results, it is faster and doesn't require so precise setting of the parameter $\alpha(t)$ for training.

12.3 KSOM Batch Training Algorithm

The algorithm resembles Linde-Buzo-Gray algorithm [**?**] where all the trainings samples are assumed to be available when learning begins.

Notation defined in preceding section is valid here as well: KSOM is defined a set of M models (neurons) m_i. Each model m_i is represented by a vector with the dimension that is equal to the dimension of input data vectors. A set of m_i is ordered, i.e. exist relation that order the set in output space. All the models are ordered into the space of dimension N_{out}, usually $N_{out} = 2$. The ordering then defines occupations of the neurons of the point in N_{out}-dimensional lattice.

Near points on the lattice forms a neighbourhood when the distance d (see (12.1)) from central model is lower that certain value N_c.

Each part of the vector represents different variable. These variables might have different dynamic range. Normalization is not necessary in principle, but it may improve numerical accuracy because the resulting reference vectors then tend to have the same dynamic range. Learning algorithm performs the steps in the following order:

1. Random initialization - it is suggested as the best and also the fastest policy (it is the same as for iterative algorithm). It is strongly recommended to use it in practice. It is also possible to take first M vectors from training set.

2. For each model m_i, collect a list L of copies of all those training samples x whose nearest reference vector belongs to m_i, so $d(m_i, x)$ is minimal.

3. Take for each new reference vectors the mean over the union of the list made in step 2.

$$m_i(t+1) = \text{mean}(L) \tag{12.6}$$

4. Repeat from 2 a few times.

The batch training algorithms is more efficient especially for application with large input data sets. Utilization of this algorithm avoids problems connected with correct setting of $\alpha(t)$ for training.

Implementation of the batch algorithm suppose that all training vectors are known before the start of training, whereas when utilizing the iterative algorithm, the data might be supplied during the training (but must be stored in order to perform multiple passes and guarantee correct generalization).

To measure how well is a map trained, a quantization error is introduced. Quantization error is norm of difference of signal vector from the closest model (neuron) in signal space [**?**]. To verify whether the training was appropriate, a topological error might be utilized.

Topological error is the average number of times when the second best vector was not classified to belong in the neighbourhood of the same neuron as the best vector.

12.4 Visualization

As was mentioned in the beginning of the section, KSOM is an effective platform for visualization of high-dimensional data. This unique ability helps to fully understand contents of a data set and exploit contents in more convenient way.

Two-dimensional KSOMs (vast majority) are usually visualised in the form of U-matrix (Unified Distance Matrix). U-matrix is a representation of a self-organizing map (SOM) where the Euclidean distance between the codebook vectors of neighbouring neurons and the internal vector of the neuron is depicted in a colour-scale image. It is used to visualize the data in a high-dimensional space on a 2-dimensional space (more examples in [**?**]). Example of an U-matrix is in figure 12.1.

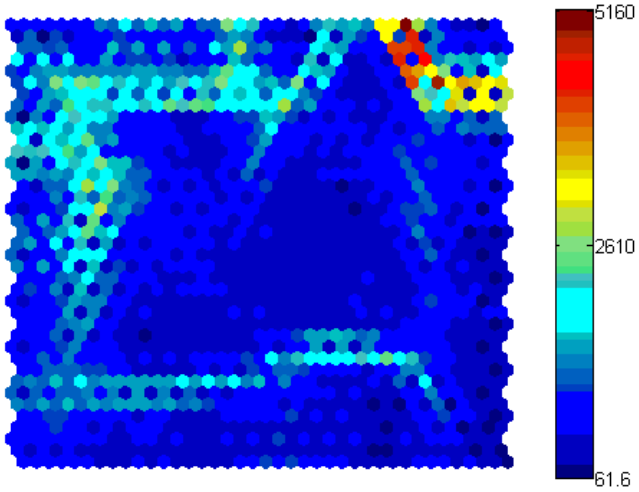

Figure 12.1: U-matrix

Each hexagon on odd position (the first, third, ...) represents by its colour the internal vector of a neuron. The colour of the hexagon that lies between two neighbours represents distance d between them. Different colours (and their shades) tends to represent distances.

12.5 Comparison of Maps

Comparison of the maps is performed using distance-based approach [**?**]. Unlike visualization using U-matrix, distance between neurons in one net is not taken into account. The distance for comparison of the maps is distance between two neurons, each being of the different map.

To compare two different KSOMs, two diverse criteria to compare maps were suggested [**?**, **?**]. Both the 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 references 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 pared 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 (12.7).

$$d(b,c) = d(\mathcal{F}_b, \mathcal{F}_c) = \sqrt{\sum_{n \in |B|,|C|} (\mathcal{F}_b[n] - \mathcal{F}_c[n])^2} \tag{12.7}$$

In (12.7), 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 (12.8), where P is the number of neuron pairs. P equals to the number of neurons in smaller map.

$$D(B,C) = \frac{1}{P} \sum_{b \in B, c \in C} d(b,c) \tag{12.8}$$

The second criterion is more restrictive (further references 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 have 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 $\mathcal{G}$ (12.8). To distinguish between the criterion used, the overall distance will be references as $D_\mathcal{G}$ for general criterion, resp. $D_\mathcal{R}$ for the restrictive one.

12.6 Implementation of KSOM

This section deals with modifications to the original algorithm regards the implementation [**?**]. Not only implementation in toolbox [**?**] or in software sense [**?**, **?**, **?**], conclusions presented are also valid for hardware implementation of KSOM in a form of standalone accelerator (e.g. in Field-Programmable Gate Arrays (FPGA)) [**?**, **?**, **?**, **?**].

Considering software implementation, both iterative and batch training algorithms might be implemented. The disadvantage when implementing batch algorithms is that it might require large memory to store the data and subsequent getting very slow operations on such a large datasets.

This is almost discriminative for batch algorithm when dealing with implementation in hardware when the sources are much limited. However the iterative algorithm might be implemented very efficiently.

KSOM algorithm itself is very robust. It is possible to introduce several changes to the algorithm that slightly influence obtained results whereas lead to the much efficient operation.

The first problem when implementing algorithm is common for every programmer that have ever tried to write a code that implements some of the common digital signal processing operations. The problem is final resolution of the numbers in computer - variable types. To get an efficient implementation, it is an advantage to utilize fixed point data types instead floating. Some means of implementation also allows to specify the width of variable, computers allow to encode the variable 8 or 16 bit wide. The question is even complicated on platforms based on FPGA. The length of the word might be set arbitrarily, with indirect impact to the performance (more bit - less performance).

KSOM algorithm is sensitive to the length of data word. Improper selection might cause training to diverge - worse the selection of proper width, less frequently divergence of the algorithm occurs. The effect of quantization can be derived in advance, the features of the implementation also depend on the distribution function of data vectors [**?**]. Only possible solution how to derive sufficient quantization is to simulate it on the representative set of data. Proper quantization is then obtained empirically from these results. Disturbing quantization effects can't be described generally because they tightly coupled with distribution function of the data.

Further simplifications leads to replacement of the neighbour function. Neighbourhood function h_{ci} (e.g. (12.5)) also known as "Mexican hat"can be replaced by rectangular or triangular window. This avoids computation of hyperbolomic functions. The influence to the results is minimal [**?**]. Euclidian distance (12.1) in could be replaced by Manhattan distance (12.9).

$$d(x, m) = \sum_i |m_i - x_i| \tag{12.9}$$

Utilization of fixed point arithmetic leads to noticeably higher amplitude of quantization error during training (see section). The example of the difference is shown in figure 12.2. This effect masks the edges observable in the quantization error curve for badly trained maps. Therefore advanced methods of poorly trained net have to be used. More information (including code examples in Matlab) are provided in [**?**].

Described techniques helps when dealing with implementation of KSOMs. Unfortunately, the nature of the disturbing effect is complicated enough that each implementation should be tested on the representative data set.

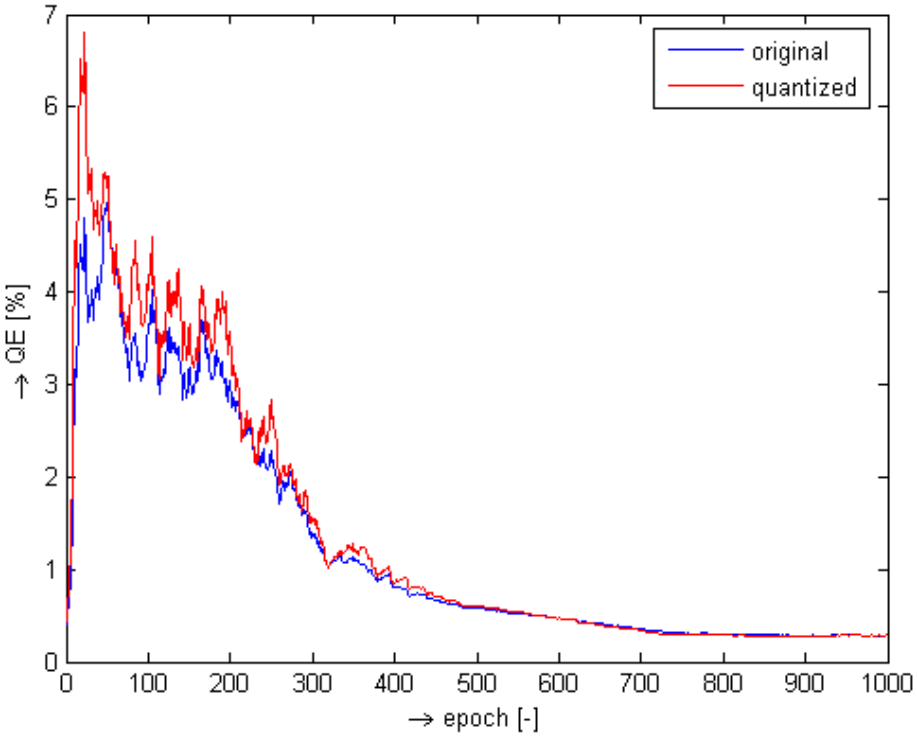

Figure 12.2: Quantization error comparison during training with iterative algorithm

12.7 Other Types of ANN in Speech Signal Processing

Several hundreds or thousands of experiments that utilize ANNs for speech signal processing were published. Intention was always to improve processing of speech signal. Current section contains only the selected overview. The aim is to illustrate possibilities different than KSOM.

Gemello, Albesano, Mana and Moisa presented in [**?**] multi-source neural network aimed to find the optimal combination of features for classification.

Most of the papers deals with representation of signal based on wavelet transform and its combination with ANN. For example, Daqrouq, Abu-Isbeih and Alfauri presented speech signals enhancement system that using neural network and wavelet transform [**?**]. They compare system based on ADALINE, feed-forward neural network and hybrid system based on combination of wavelet transform with ADALINE. The most enhancements were obtained when utilizing ADALINE (10dB) then by hybrid system consisting of wavelet parameterization and ADALINE (8dB). Less improvement by provided by the system based on feed-forward neural network (3dB).

Similar paper from Gandhiraj and Sathidevi [**?**] presents Auditory-Based Wavelet Packet Filter bank for Speech Recognition Using Neural Network. The idea is to evaluate Gamma Tone Filter Bank and Wavelet Packet as front-end system for Back Propagation Neural Network.

Since ANN represents alternative to the systems based on Hidden Markov Models and gaussian mixture models (GMM). Several studies were published touching this topic (e.g. [**?**]). Several variations of ANN were made to process data speech as for example Deep Neural networks (DNNs) [**?**]. DNNs have many hidden layers and are trained using new training method. It was shown that DNNs outperform Gaussian mixture models (GMMs) on a variety of speech recognition benchmarks, sometimes by a large margin.

12.8 KSOM Variants

Kohonen describes in his book [**?**] modification to the KSOM that is intended to to generate filters for time-domain speech waveforms. The derivation is known as the adaptive-subspace self-organizing maps (ASSOM). In this map each model represents a wavelet filter. The filters (models) are formed automatically in the ASSOM process using following algorithm - for each learning episode $S(t)$ consisting of successive time instants $t_p \in S(t)$ do the following:

1. Find the winner (indexed by c):

$$c = \text{argmax} \left\{ \sum_{t_p \in S(t)} \|\hat{x}^{(i)}(t_p)\|^2 \right\} \tag{12.10}$$

2. For each sample $x(t_p)$, where $t_p \in S(t)$ rotate the basis vector of the module

$$b_h^{(i)}(t+1) = \left[I + \lambda(t) h_c^{(i)}(t) \frac{x(t_p)x(t_p)^T}{\|\hat{x}^{(i)}(t_p)\| \|x(t_p)\|} \right] b_h^{(i)}(t) \tag{12.11}$$

3. For each sample $x(t_p)$, where $t_p \in S(t)$ dissipate the components $b_{hj}^{(i)}$ of the basis vectors $b_h^{(i)}$:

$$b_{hj}^{(i)`} = \text{sgn}\left(b_{hj}^{(i)}\right) \max\left(0, |b_{hj}^{(i)}| - \epsilon\right) \tag{12.12}$$

where

$$\epsilon = \epsilon_h^{(i)}(t) = \alpha \left| b_h^{(i)}(t) - b_h^{(i)}(t-1) \right| \tag{12.13}$$

4. Orthonormalize the basis vector for each module.

The algorithm results in producing smooth, asymptotically stable, single peaked band pass filter with continuous distribution of their mid-frequencies over the range of speech frequencies.

Described experiment was motivation for several subsequent applications that leads to connection of the speech signal and KSOM. They were published sequentially firstly experiments with operator maps that each of the model represents the filter [**?**, **?**, **?**] and subsequent improvements [**?**]. This research lead to the utilization of matching pursuit algorithm [**?**] that is described in this work.

There were several modification to the KSOM published. For example Expanding Self-Organizing Map (ESOM) published in [**?**]. This modification deals with better topology correspondence between the input data and output grid. It reaches lower topological and quantization error (see chapter 12.6). This modification addresses mainly large data set and visualization issues. Another variant that helps with clustering of the map. Growing Hierarchical Self-Organizing Maps [**?**] constructs the tree based hierarchy of clusters right during the training process.

The other noticeable variation is the FastSOM [**?**]. Main issue addressed is the speed of iterative training algorithm. The disadvantage of the algorithm is higher instability of the quantization and network error during the training. According to the experiments that were done in our laboratory, the stability of the algorithm is the issue and it is more probable that the training phase will lead to unaligned network that is suboptimal.

12.9 Resulting Method - Utilization of KSOM in Classification

A children ability to pronounce vowels and isolated words will be investigated. The utterances will be parameterized. the output of parameterization method is supposed to be in a set of vector, where for each of the utterances will be produced set consisting of several vectors.

Set for particular utterances might be grouped. This grouping takes places before the training of the map and will be done on the criteria supplied externally (e.g. same utterances for several speakers).

For each of the group, the trained map is provided. Parameters of the map (size, etc.) are to be specified externally. It is supposed that correct settings will be found experimentally during the work with the data.

Since there will be several groups, it is vital to be able to compare resulting KSOMs (trained on the data) with another network. Comparison is to be done based on criteria in section 12.5.

By comparison of the maps previously trained on specified (sub)set by the map trained on the speaker, the overall classification of the speaker will be derived. This area is out-of-scope of this work, there still on-going debate how exactly perform the overall classification.

However, the vital role in the classification plays the parameterization. As is described in following section, classification based on MFCC, PLP or LPC is not sufficient for the task. The idea is that the parameterization should provide a broad range of parameters that describes utterance(s). Thanks to generalization property of KSOM, the shrinkage of the range then might be done to further focus important features contained in signal.

Chapter 13

Classification Based on LPC, PLP and MFCC Parameterizations

This chapter focuses on method developed for classifications of the speech with disorders. Described classification method is based on children's speech signal analysis and allows observing the trend of the speech disorder during therapy. The parameterization included in this experiment was chosen as initial. The intent is to compare and select parameterization that work best when dealing with the speech of children with developmental dysphasia.

The classification is based on the fact that the disorder has a direct impact on speech production (i.e. movement of vocal tract). Thus, we can measure the trend of the disorders comparing patterns obtained from speech of healthy children to the patterns obtained from children with disorder. Classification utilized in the experiment is based on cluster analysis of Kohonen Self-Organizing Maps [**?**] trained on parameterized speech signals. The main advantage of using artificial neural network is adaptability to specific attributes of the signal and tolerance for the noise contained in recordings.

The aim of described method is to distinguish between healthy and ill children and describe advance of disorder during therapy. The method is based on a comparison of the differences in the parameterized speech. Purpose of the parameterization is not to give perfect representation for recognition, but describe the differences.

The process of analysis is divided into the two phases. In the first phase, the samples from selected subset of healthy children are taken and the patterns are worked out (see figure 13.1). Input set is divided in two disjunctive parts. The first part is used for training KSOMs. The second part serve as an input for computing the patterns using trained KSOMs.

There is one map for each class of samples from database. Since the complexity involved, only results for vowels are presented. The choice of these classes is not arbitrary, the effect of developmental dysphasia to the vocal triangle was already described [?, ?, ?] and in [?].

In the second phase, patterns for selected ill children are calculated and compared with the patterns for training set (see figure 13.2). Whereas in the first part utilized only samples from healthy children, the computation in the second part is done using the samples from only one (ill) child acquired during one session. Comparing these two results we can observe the trend of the disorder. Important conditions are sufficient size of the training set and using different samples for computing the patterns. Satisfy these conditions ensures generalization of children's speech signal and avoid adaptation on individual speakers. Results of the experiment were partially published in [?].

13.1 Description of Method

The patterns are estimated from the subset consisting only samples (speech) from healthy children. The speech is processed common way, firstly, the signal is divided into segments and weighted by the Hamming window (4.1). Then, every segment is represented in the terms of standard parameters: MFCC, PLP [?] and LPC coefficients [?] and [?].

MFCC and PLP coefficients are utilized generally in the various tasks in the field of speech signal processing. Whereas MFCC and PLP coefficients were created for recognition and thus, they tend to generalize, LPC coefficients clearly describe parameters of vocal tract with respect for these differences. LPC shows the best option when dealing with speech of 4 to 6 years old children. Generated vectors are used to train KSOMs.

For each segment, three different vectors are created, one for each type of parameters. Whole speech is then represented by the series of such vectors. These series are then processed using artificial neural network, namely by Kohonen Self-Organizing Maps [?].

There are three independent networks, one for each type of representation (i.e. one KSOM for MFCC, one for PLP and another for LPC). The greater part of the input data (training set) is used as the input for the training. The rest is for calculation of the patterns. We tend to preserve ration roughly $3 : 1$ between the size of training set and verification set. Overview of the phase is in figure 13.1.

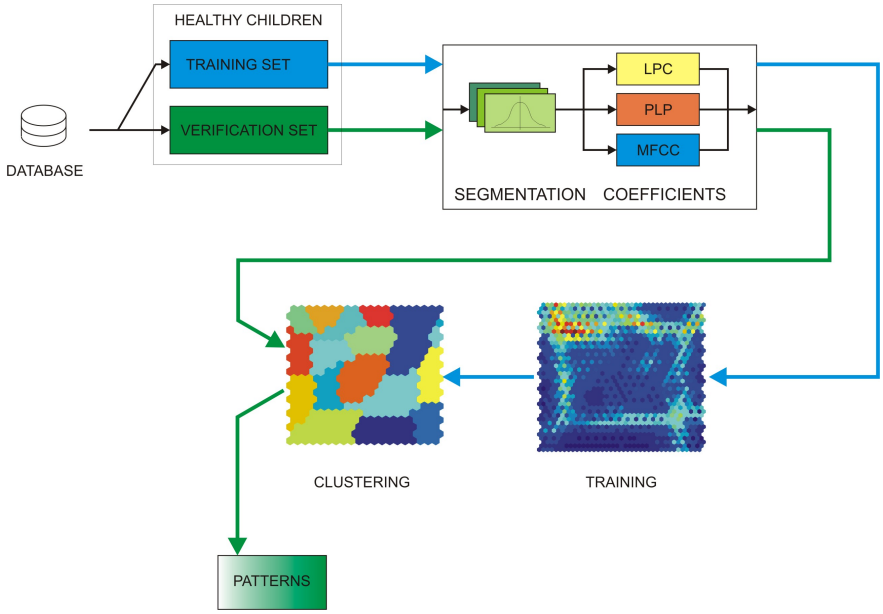

Figure 13.1: Overview of the first phase - preparing patterns

After training, the cluster analysis of each KSOM is performed. For this purpose we utilize k-means algorithm. The analysis is utilized to obtain higher level of abstraction of the speech - every cluster represents specific features in the speech.

Using the clusters, patterns are generated. Patterns are vectors, which dimension is equal to the number of clusters in particular KSOM. Each component of the pattern represents the percent occurrence of the input vectors in the corresponding cluster. The patterns are calculated using the rest of the samples from the input set (verification set). These vectors are important for following comparison. The patterns are derived from map trained only on healthy children's speech.

The patterns are generated solely from speech of healthy children. In this, initial, experiment the influence of various artefact contained in recordings (doors bumping, traffic on corridor, parents, etc.) and as well in the speech itself (various defect independent from developmental dysphasia) must have been supressed as much as possible.

sets	**boys**	**girls**	**speakers**	**utterances**
training	7	14	21	**90**
verification	2	5	7	**35**
children with disorder	2	1	3	**43**
total	11	20	31	**168**

Table 13.1: Overview of the number of utterances in experiment

Because of that very strict rules for selecting the speakers for experiments were applied which reduced the size of a set of proper speakers (see table 13.1).

Then we compute patterns for the child with disorder (see figure 13.2). After comparison with patterns estimated from healthy children's speech, we get observable differences. The measure of those differences (Euclidean distance of these two vectors) qualifies differences between the ill child's speech and the speech of healthy children. The Euclidean distance between representative vectors and the particular vector serves as a main criterion. Observing distances on the various classes of speech, we can approximate the trend of disease. Reciprocal usage of method (training on ill children and evaluate for healthy children) is possible, however, the set of available utterances was very small (as mentioned above).

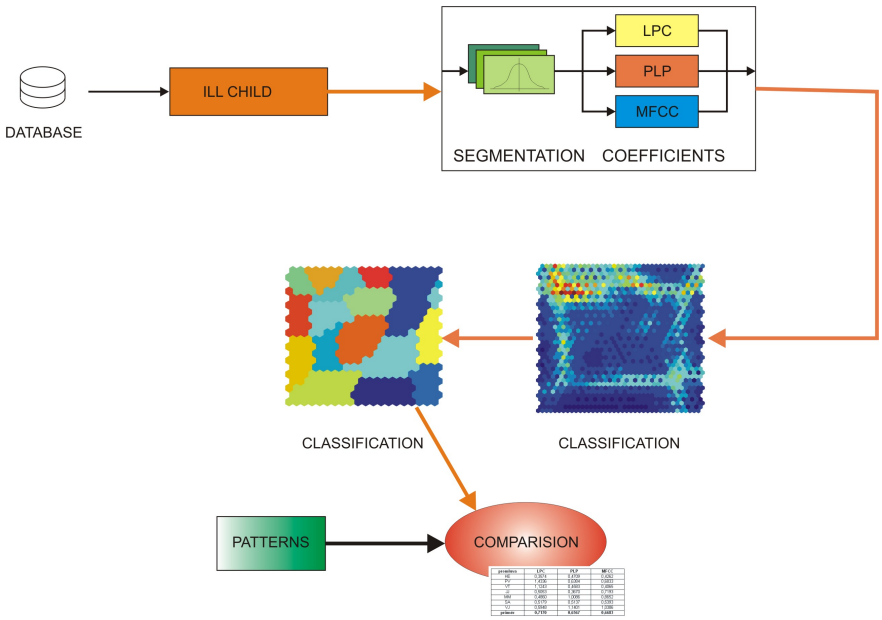

Figure 13.2: Overview of the second phase - classification

13.2 Parameterization of Utterances

Settings and results for the experiment with classification of vowels "a", "e", "i", "o" and "u" is described in the following text. For the training set consisted of recording (samples) taken from twenty-one children (see table 13.1). As was described above, whole training set was divided into two parts.

There was also verification set consisting of seven healthy children (two boys and five girls) and three children with speech disorder (two boys and one girl). The samples in this set were intended to confirm the results obtained from

the described method. For every ill children, three recordings from various sessions (i.e. recordings made on different days) were used. The overview of the division of the children taken to the experiment is in table 13.1.

Speech signal was pre-processed in the following way: whole recordings of the vowels were segmented by the 30ms Hamming window with 10ms overlay. The MFCC, PLP and LPC coefficients were calculated from these segments. For MFCC and PLP coefficients, the basic band (22kHz) was divided into twenty sub bands. The LPC coefficients were of the 8^{th} order, the MFCC and PLP coefficients were counted for 20 bands.

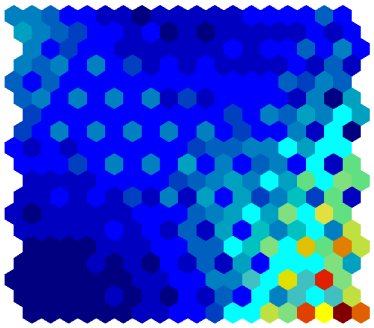

Figure 13.3: U-matrix for KSOM 10×10 trained on MFCC coefficients

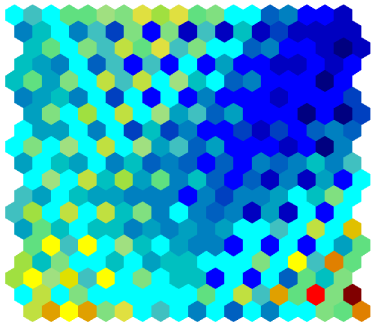

Figure 13.4: U-matrix for KSOM 10×10 trained on PLP coefficients

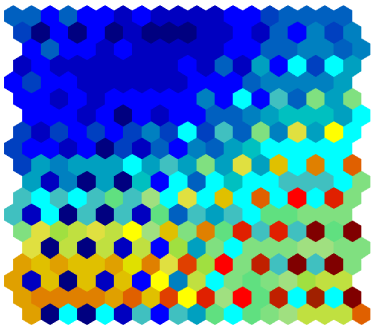

Figure 13.5: U-matrix for KSOM 10×10 trained on LPC coefficients

13.3 KSOM Training

We used three different sizes of the maps in the experiment 10×10 neurons, 20×20 neurons and 30×30 neurons. All the computation was done using Matlab and SOM Toolbox [**?**].

The maps of 10×10 neurons were too small (see figures 13.3, 13.4 and 13.5). The exceedingly generalisation was performed and therefore the distribution function has not been captured in details.

Better results were obtained using maps with 20×20 neurons (see figures 13.9, 13.10 and 13.11) , where the best results were observed for map trained on LPC coefficients. The approximation in the map trained for MFCC was worse compared to the maps trained on PLP or LPC.

Here is completely different situation with the maps containing 30×30 neurons (see figures 13.6, 13.7 and 13.8). Maps trained on MFCC and PLP coefficients generalized excessively. This is suitable for recognition, but not for our purposes. In the analyses of the differences in the speech maps trained for LPC gave better results.

Nevertheless, the maps with dimension 30 neurons are not suitable for described training group because of the effect of limited generalization. For classification for described purposes, the 20×20 maps are appropriate.

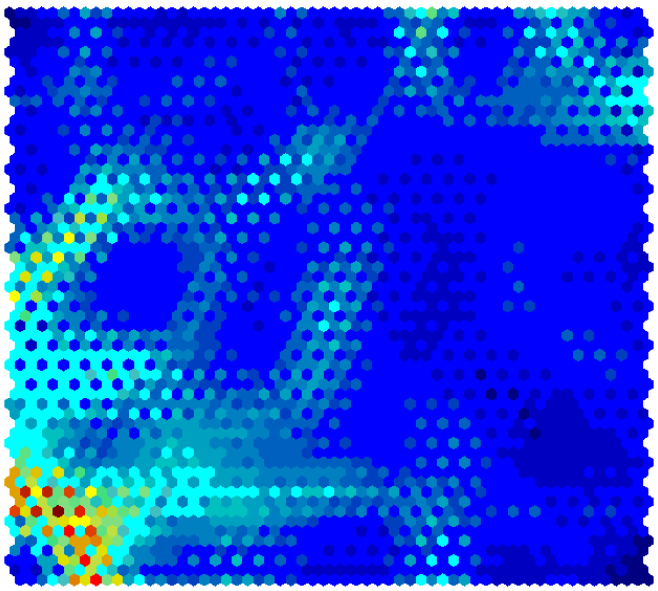

Figure 13.6: U-matrix for KSOM 30×30 trained on MFCC coefficients

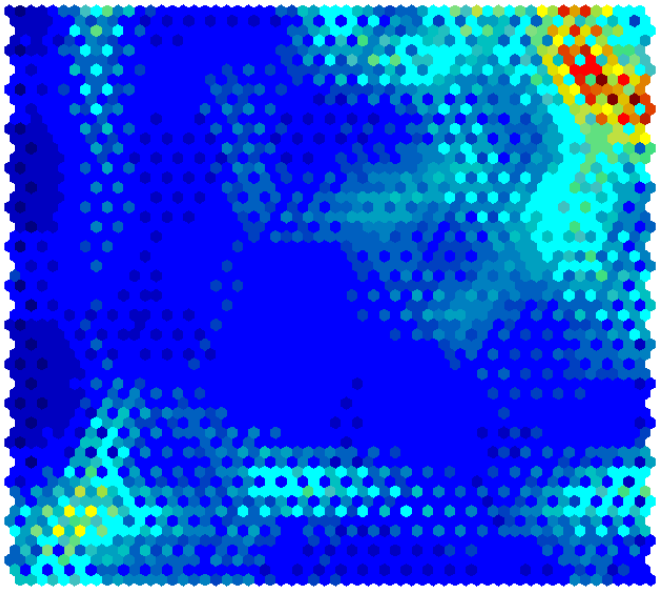

Figure 13.7: U-matrix for KSOM 30 × 30 trained on PLP coefficients

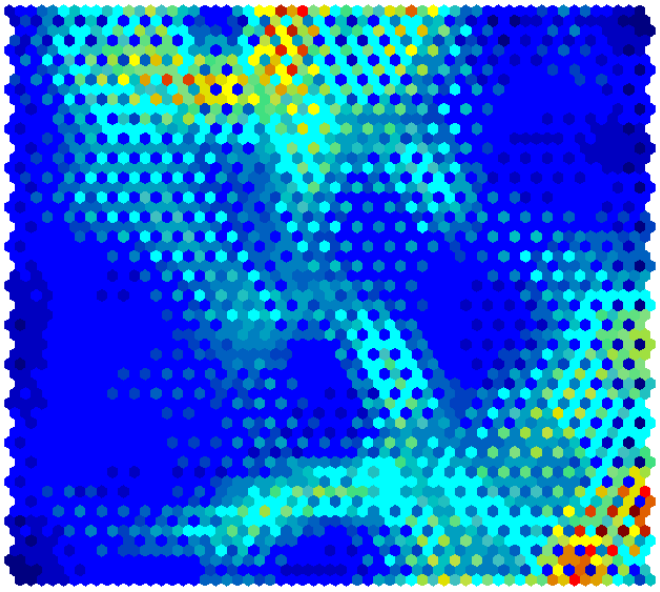

Figure 13.8: U-matrix for KSOM 30 × 30 trained on LPC coefficients

13.4 Clustering of Maps

The k-means algorithm was then utilized for getting clusters in a map. More clusters contribute to better classification of a signal. The calculations were randomly initialized. The lowest number of clusters was extracted from maps trained on MFCC coefficients - from 10 to 14 clusters. For the maps trained on LPC and PLP the k-means sensitivity was better. For LPC coefficients there

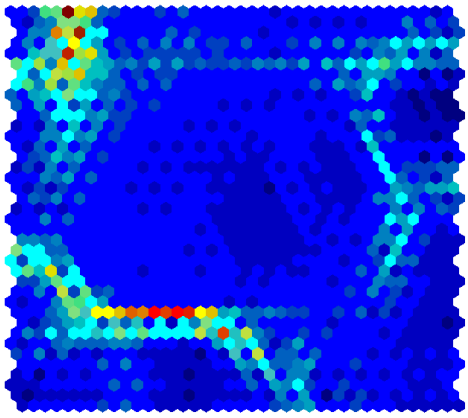

Figure 13.9: U-matrix for KSOM 20 × 20 trained on MFCC coefficients

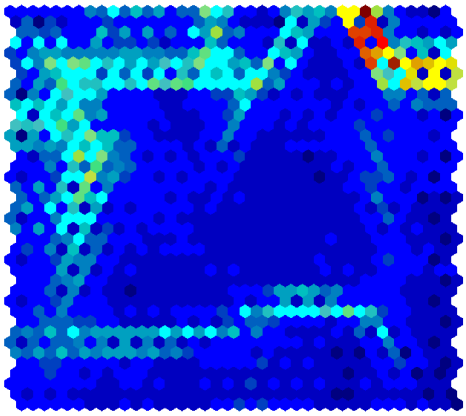

Figure 13.10: U-matrix for KSOM 20 × 20 trained on PLP coefficients

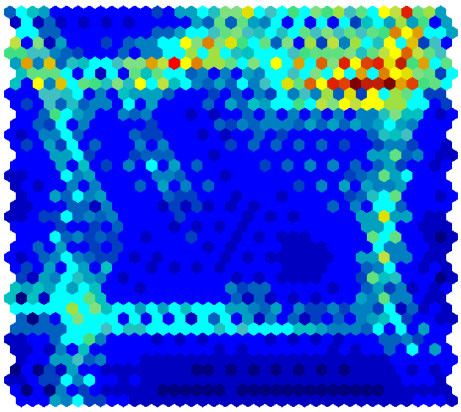

Figure 13.11: U-matrix for KSOM 20 × 20 trained on LPC coefficients

were extracted from 19 to 24 clusters, for PLP coefficients algorithm found between 20 and 24 clusters.

Examples of clustered maps are in figures 13.12, 13.13 and 13.14. Each area filled in different colour represents one cluster. The cluster is a group of the

Figure 13.12: Clusters in MFCC-trained map

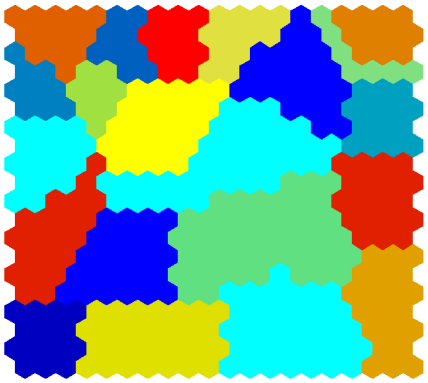

Figure 13.13: Clusters in PLP-trained map

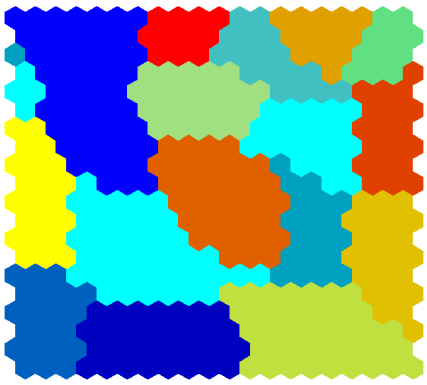

Figure 13.14: Clusters in LPC-trained map

one or more neurons, which consist of neurons that represents similar features in the speech. Dividing the trained network into cluster bring a higher level of generalization into analysis – instead 400 (almost - considering 20×20 map) similar prototypes represented by the neurons, there will be only about 20 distinctive features represented by the clusters. As could be seen in figure 13.12, MFCC parameterization gives fewer clusters. It means that this parameterization does not describing so much details as LPC based parameterization (see figure 13.14). The reasons of such differences are mentioned in section 13.3.

13.5 Classification

The patterns were calculated using the second part of the training set. Then the samples from the verification set and also the samples obtained from children with disease was compared against these vectors. In table 13.2 and table 13.3, there are results of comparison for the vowel A. In table 13.2, each row represents results from different speaker from verification set. In table 13.3, each row represents one recording session form one of the three children with disorder. The number after R is the number of recording - sessions were recorded in 3 months interval.

speaker	LPC	PLP	MFCC
H1	1.4878	1.1767	1.1939
H2	1.1741	0.4844	0.6514
H3	0.8275	1.3865	1.1705
H4	0.9739	0.5795	0.7437
H5	0.7824	0.4418	0.4904
H6	1.3211	1.3218	1.1320
H7	0.6218	0.6435	0.7373
average	**1.0269**	**0.8620**	**0.8742**

Table 13.2: Results for healthy children (vowel A)

As could be seen, there is difference between results obtained from various parameterizations. For MFCC coefficients, there is not satisfactory variability to distinguish healthy and ill children and therefore these coefficients are not suitable for classification of children with disorders using KSOMs. For this task, the LPC and PLP parameterization giving better results. It is in accordance with the reasons mentioned above.

The average results obtained for both groups of children and all parameterisations are in table 13.4. The aim of the method is to distinguish between healthy and ill children - greater difference means more suitable parameterization. The results are distinguished by the method used to obtain coefficients

speaker	LPC	PLP	MFCC
D1R1	1.1273	1.1983	1.7149
D1R2	1.3461	1.2415	1.7149
D1R3	1.4625	1.1907	1.7149
D2R1	1.5789	1.2703	1.7151
D2R2	1.4989	1.2553	1.7149
D2R3	1.5127	1.2208	1.7149
D3R1	1.5613	1.3012	1.7149
D3R2	1.5663	1.3012	1.7149
average	**1.4568**	**1.2474**	**1.7149**

Table 13.3: Results for the children with disorder (vowel A)

from segments. As one could be seen, the LPC parameterization gives the best result in distinguish between healthy children and children with disorder. The PLP is also possibly useful, but MFCC parameterization giving unsatisfactory results that are not suitable for described purposes.

vowel	group	LPC	PLP	MFCC
“a”	healthy	1.0269	0.8620	0.8742
	disorder	1.4568	1.2474	1.7149
“e”	healthy	1.3924	1.2464	1.2868
	disorder	1.7652	1.2795	2.0644
“i”	healthy	1.9056	0.4446	0.7204
	disorder	2.0020	0.9438	1.4936
“o”	healthy	1.1604	0.9426	1.0637
	disorder	1.4143	1.2510	1.4122
“u”	healthy	0.7170	0.6567	0.6683
	disorder	0.5177	1.3335	1.6071

Table 13.4: Results for both groups of children – all parameterizations

13.6 Discussion

The results are strongly depended on the type of speech units. We suppose that vowels are the simplest units to analyse and in the more complicated cases, the result could be only worse. The method well detect ordinary problems in speech of the children with developmental dysphasia – interchange of the high vowels in the vocalic triangular. The performance of described method depends on using proper KSOMs size according to the size of a training set.

Utilization of the KSOM allows modifications of the process that distinguish between speakers, to the process where common attributes are extracted and allows to distinguish between healthy children and children with disorder. The method discussed in this chapter has to be extended in order to get more information about children's speech and to reliable describe trend of the disorder.

Purpose of the parameterization is not to give perfect representation for recognition, but describe the differences. Therefore, the MFCC parameterization is not convenient, as was shown above. Classification based on MFCC could be used in the task of distinguishing whether a child is healthy or suffers developmental dysphasia, however, the parameterization is not convenient for the determining the progress of the disease.

For the task of deterring the progress of the disease, PLP and LPC parameterizations seem to be more suitable. Classification based on PLP coefficients has similar problems as MFCC, but the results are not so much influenced as in the case of MFCC. There also observed when dealing with vowels "o", "u"and sometimes "i". As could be seen in table 13.4, the difference in in averages for healthy and ill children is minimal.

We intentionally utilize LPC-based parameterization, which is not suitable for recognition. However, LPC coefficients are suitable for our analysis because it describing the differences in the speech better than MFCC coefficients. To perform final classification, it is suggested to perform both with LPC and PLP parameterizations.

13.7 Proper Size of KSOMs

The maps were trained by batch training algorithm which reduces convergence issues. However, as could be seen in the results presented, proper size of the network remains a vital factor.

After realization of several other experiments with the method described, formula (13.1) was derived to help establish the correct size settings. In (13.1) N represents the number of neurons that constitute an edge of a square Kohonen Map with hexagonal topology. Parameter n_t represent overall number of vectors obtained by parameterization of all utterances in training set, n_u represents the number of utterances in the training set. Parameter n_s stand for the number of speakers whose utterances are in the training set, n_{ut} represent number of types of utterances that are included in training set. The assumption is training set contains almost are utterances (one for each type) from almost all speakers.

$$N \cong \sqrt{k\frac{n_t}{n_u}} \cong \sqrt{k\frac{n_t}{n_s\ n_{ut}}} \tag{13.1}$$

Coefficients k represents the average number of vectors from training set that should be represented by a neuron. According to the experience gained, k is dependent on the parameterization used. The value of k is summarized in table 13.5.

parameterization	***k***
MFCC	16 – 18
PLP	8 – 18
LPC	10 – 24

Table 13.5: Recommended values of k (see (13.1))

Values of coefficient k were determined empirically and agreed with observation that the size of a map trained on vectors obtained by MFCC parameterization are proper maps larger that for vectors obtained by LPC and PLP parameterizations. On the other hand, for PLP parameterization better suits slightly smaller maps.

Another conclusion obtained considers sensitivity to size of a map. As could be seen in table 13.5 LPC parameterization shows the least sensitivity, whereas MFCC has the highest.

The role of experiment described in chapter 13 is, apart from the comparison of parameterizations, to illustrate the influence of size of a map. Therefore it was not intended only the size obtained by the formula (13.1). The values of coefficient k for the maps in the experiment are summarized in table 13.6.

parameterization	***k***
MFCC	3.61
PLP	14.45
LPC	32.53

Table 13.6: Values of k for maps in the experiment

Formula (13.1) as well as accompany values of k in table 13.5 express empirical experience gained when working with the parameterizations mentioned.

Chapter 14

Classification Based on Matching Pursuit on Spectral Bands

The method described in the following text was developed to analyse disordered children speech. Since developmental dysphasia has impact on the children speech ability, the classification of utterances helps to determine whether treatment and medication is appropriate. The paper describes method developed to provide classification based on utterances but without any additional demands on speech pre-processing (e.g. labelling). 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 the illness.

As an initial experiment, simple task that prove the convenience is presented. The aim of the experiment is to determine whether the method can distinguish between utterances pronounced by healthy and ill childrenThe experiment represents just only a simple application of the method described above. Results were published in [**?**].

14.1 Description of the Method

Description of the method is divided into three parts. Each part corresponds to the one of main steps in analysing utterances. At the first step, parameterization of utterances is carried out. Matching pursuit [**?**, **?**] algorithm with usage of spectral band is utilized 11. 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 atoms. The replacement preserves all important information included in the authentic signal. The parameterization is adjusted so that information can be easily extracted in the next steps.

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

14.1.1 Feature Extraction

After parameterization, an utterance is represented by a finite set of vectors γ (10.17). As was 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 KSOM. 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 suffer developmental dysphasia. For the most experiments, a set consists only of the same kind of utterances. An experiment could require several sets to be utilised.

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 health status. To keep the generalization, the large as possible 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 [**?**], 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.

14.1.2 Classification

The method described in the paper is being developed to determine whether a speaker suffers developmental dysphasia or not. In case of positive answer it is should be possible to particularize the stage of the disease. Utterances of the speaker examined are acquired and then parameterized. After parameterization, feature extraction and classification a comparison takes place. Usually the 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 extend of the training sets are specified separately for each particular experiment. Generally, the comparison is performed separately for each different utterance type.

As a basis for comparison, the database of utterances of healthy children (in different age, both genders) and children suffers developmental dysphasia is maintained. A set of different utterances is being kept for each child as well as health status. For most experiments all the utterances meeting the requirement of an experiment are split into two sets: one set is made of utterances of the healthy children only and the 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 KSOMs (resp. internal weights of neurons within a map) for each the set is compared one to each other. Criteria to compare maps were already described in section 12.5, but since the definitions are straightforward, they are duplicated here.

Both the 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 references 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 pared 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 (12.7).

$$d(b,c) = d(\mathcal{F}_b, \mathcal{F}_c) = \sqrt{\sum_{n \in |B|,|C|} (\mathcal{F}_b[n] - \mathcal{F}_c[n])^2} \tag{14.1}$$

In (14.1), 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 (14.2), where P is the number of neuron pairs. P equals to the number of neurons in smaller map.

$$D(B,C) = \frac{1}{P} \sum_{b \in B, c \in C} d(b,c) \tag{14.2}$$

The second criterion is more restrictive (further references 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 have 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 $\mathcal{G}$ (12.8) resp. (14.2).

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

14.2 Healthy and Ill Children Distinction

The experiment deals only with two-syllabic words. To simplify the analysis, utterances involved are limited to only 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 developmental dysphasia (14 female, 30 male). These experiments are supported by extended database of children utterances. Previous experiments were done on database with lower number of utterances. These (previous) experiments were considered to be an introductory study.

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 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. For each the 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). Another set is made of the utterances pronounced by dysphatic children (further references 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 training set H. For set I, exactly the same method as for set H is utilised and 36960 atoms are obtained.

To train KSOMs 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 the respect to the previous experience gained when solving similar tasks [**?**]. To explore influence of the size to 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.

14.2.1 Results for General Criterion $\mathcal{G}$

Results for general criterion G are in tables 14.1 (utterance "papír"- paper), 14.2 ("pivo"- beer) and 14.3 ("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 exception when comparing maps from 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 conclusion, that for 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 are maps of 40×40 and 50×50 neurons too large. Also, it is not possible to distinguish whether the issue is in 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 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 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

Table 14.1: 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

Table 14.2: 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

Table 14.3: Distances $D_{\mathcal{G}}$ for utterance "sokol"(falcon)

14.2.2 Results for Restrictive Criterion $\mathcal{R}$

Result obtained for the restrictive criterion $\mathcal{R}$ are in tables 14.4 ("papír"- paper), 14.5 ("pivo"- beer) and 14.6 ("sokol"- falcon). Structure of the table is the same as for tables described in section 14.2.1.

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

Table 14.4: 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

Table 14.5: 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

Table 14.6: 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 in if optimal net size is also significantly dependent on the amount of vector in 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 prefer over criterion $\mathcal{G}$.

14.3 Discussion

This section 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 matching pursuit algorithm [**?**] improved by matching on spectral bands.

Feature extraction is based on Kohonen Self-Organizing Maps [**?**]. The ability of KSOM to neglect disturbing effect like noise and speech artefacts [**?**] 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 large dictionary slows the computation. The difference and potential disadvantage is that the signal is not parameterized in vectors that represents it in equidistant manner. Representation in the 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 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 in empirical experience of a trained specialist) compared to the results of the method described.

Presented experiment shows that the method has ability to distinguish between utterances pronounced by healthy children and children suffering developmental dysphasia. The simple criteria were chosen to only 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 the starting point. For the presented experiment, these parameters were set according to the previous experience when solving similar problems [**?**, **?**].

The parameterization using matching pursuit algorithm and further feature extraction by Kohonen Self-Organizing Maps has potential to be further extended for software intended to clinical praxis. The method is being developed to provide more precise results that allows 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, PLP and MFCC) and all the results obtained will be included in a speaker-overall classification.

The overall classification then would be compared to 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 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 right during or right after examination.

Chapter 15

Fine Tuning of Classification

Previous chapter described experiment where modified matching pursuit algorithm is utilized to distinguish between health and ill children. Matching Pursuit on spectral bands described in chapter 11 proved to have ability to detect features specific to speech of children with DD.

This chapter deals with extent of algorithm towards improvement od algorithm that allows fine classification. The aim is to possess whether the algorithm is sufficient for the task.

This experiment utilizes parameterization matching pursuit on spectral bands. Since utterances are taken from database where are stored as separated - one file, one utterance with aligned start, there is no need to introduce time slicing. So, for the experiment only frequency are bands are considered (as was discussed in chapter 11).

The scheme of the experiment is following: records of two utterances (television - "televize"and multi-coloured - "různobarevný"). Selection of utterances is not arbitrary, children with heavy forms of dysphasia have troubles to pronounce word with 4 or more syllables.

This utterance was reported to cause pronunciation problems to children with impairment. Utterance "televize"(television) is include because of having regular structure (three same vowels, two of them in a row) and should also cause problems to children with DD.

Utterance are taken from database and it is known in advance whether they are pronounced by healthy or ill children. The task does not deal with classification of healthy/ill as was described in chapter 14, however the method should be also used for this. The difference to the experiment described in chapter 14 is that we deals not with comparing two networks (each trained on particular set) but with single utterance. The results of the experiment should show whether the method has potential to perform fine classification and reveal what are the factors that have to be carefully adjusted to get reliable results.

15.1 Description of the Method

Experiment involves utterances of both healthy and ill children in age of 4 to 10 years. Number of utterances and speakers is summarized in table 15.1.

	children		**utterances**	
utterance	healthy	ill	healthy	ill
"různobarevný" (multi-coloured)	70	67	140	290
"televize"(television)	70	67	140	264

Table 15.1: Number of speakers and utterance involved in experiment

Each utterance is parameterized using 35 atoms for each spectral band. The definition of spectral bands is according to table 11.1. The number of atoms within one band is overwhelming, so only 10 first most significant atoms are included in training set.

In figures 15.1 and 15.2 are boxplot diagrams that shows distribution of atoms within one spectral band. Each of the box in figure represent distribution of atoms that are determined on particular place - order of the atoms is on x-axis. Edges of the box represents 25th percentile, red line in the centre of a box represents medians. The whiskers shows most extreme data points that are not being outliers. Plots help to compare distribution and range of atoms. Since the plots serves only to adjust the method, only atoms gathered from healthy speakers are considered.

By detailed comparison of the diagrams for each of the band, we might obtain bands and the number of significant atoms that shows the most differences. This might later help to adjust the method, for example only several bands would be considered for classification. However this experiment involves all bands.

After parameterization of all utterances, KSOM are trained separately for utterances of healthy and utterances of ill children. These networks then serve as pattern and single utterances of particular children are compared to them. The difference to the experiment described in chapter 14 is that there is no comparison between maps trained to healthy children and children with impairment. Rather, there is comparison of atoms gathered from utterances of single children which are compared to both networks.

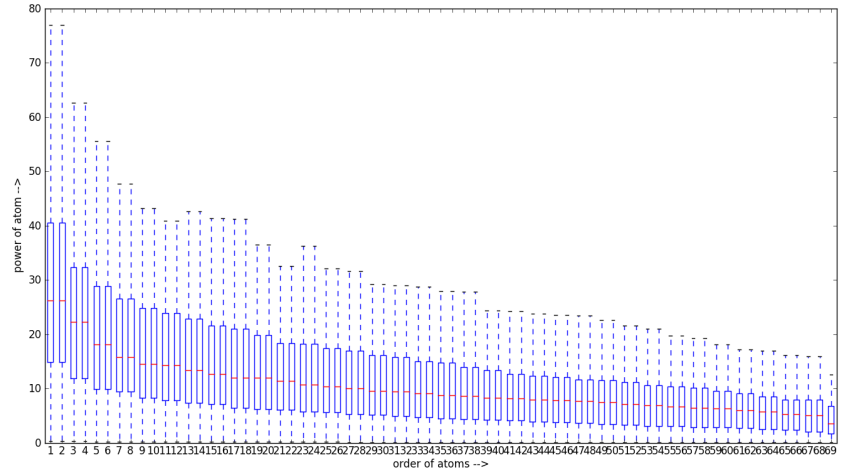

Figure 15.1: Distribution of atoms obtained by MP from utterances of healthy children (band with frequency range 400 to 510 Hz)

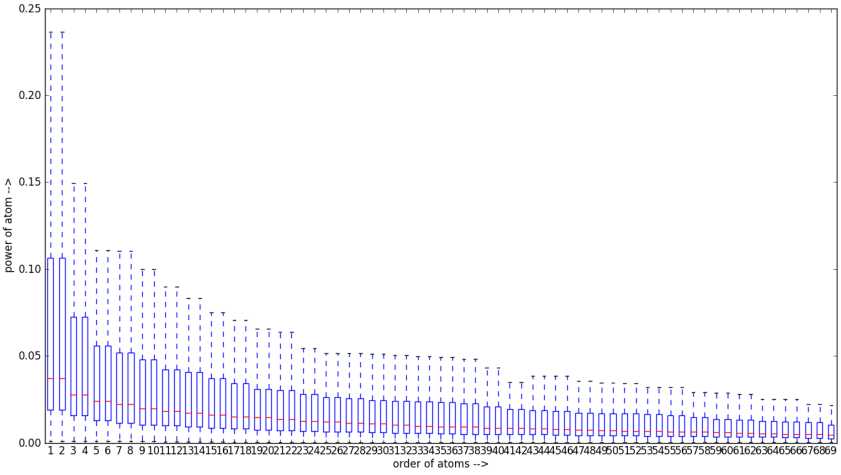

Figure 15.2: Distribution of atoms obtained by MP from utterances of healthy children (band with frequency range 3150 to 3700 Hz)

15.2 Classification of an Utterance

Single utterances are compared to the nets based on distance (15.1).

$$d_u = \min \left\{ \sqrt{\sum_{n=0}^{N} (m_n - g_{u,n})^2} \right\}_M \tag{15.1}$$

Where M is the number of atoms and where N is dimension of atoms g_u (10.17). Neural network is represented by its neurons m which have the same dimension as atoms g_u. Square root is not necessary because results are only for comparison. Equation (15.1) then becomes

$$d_u = \min \left\{ \sum_{n=0}^{N} (m_n - g_{u,n})^2 \right\}_M \tag{15.2}$$

Distance d_u is computed for each single atom that comes from utterance being compared. Individual distance are then summed up from and forms overall distance of utterance D.

$$D(U) = \sum_{u \in U} d_u \tag{15.3}$$

Where U is set of atoms that comes from parameterization of utterance.

15.3 Results

Distance D for several arbitrarily chosen utterances that were excluded from training set and then utilized as verification vectors are in figures 15.3 and 15.3. Figure 15.3 shows result for utterance "televize"(television). In figure 15.4 are results for utterance "různobarevný"(multi-coloured). Particular utterances from healthy children start with H, whereas utterances from ill children start with L. Number after the letter denotes speaker.

15.4 Discussion

The experiment was carried out to find out whether the parameterization based on matching pursuit algorithm is suitable for classification of utterances pronounced by children with developmental dysphasia. The parameterization is based on matching pursuit algorithm [**?**] improved by matching on spectral bands. Parameterized speech is further used to train Kohonen Self-Organizing

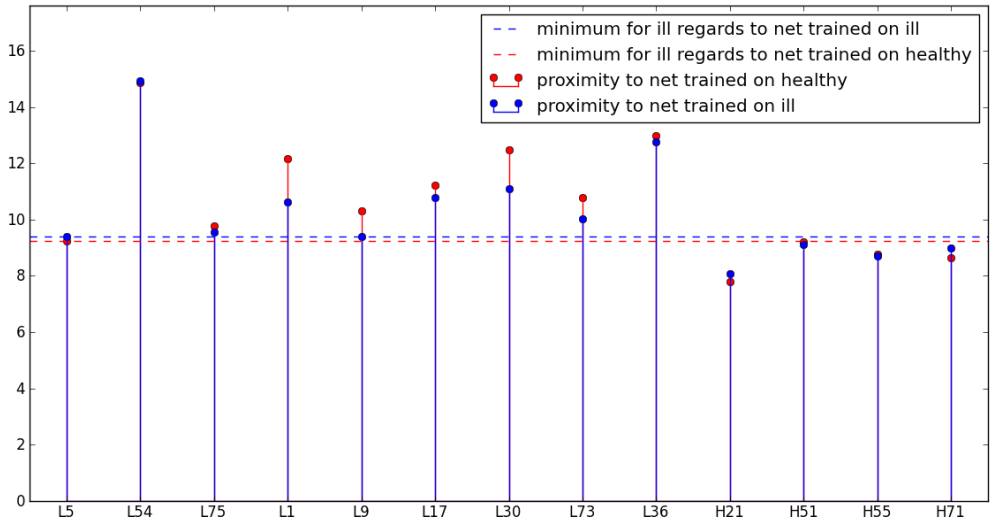

Figure 15.3: "televize" (television)

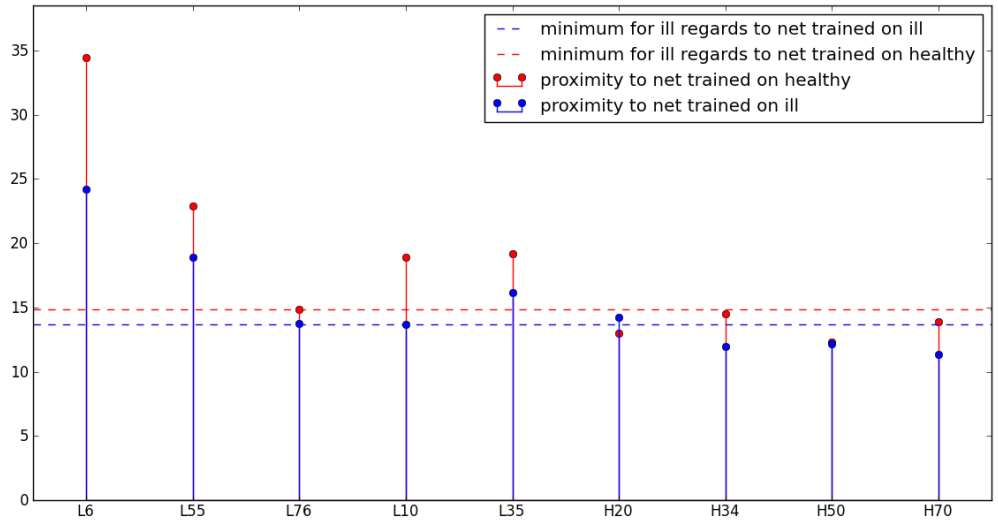

Figure 15.4: "různobarevný" (multi-coloured)

Maps [**?**]. KSOM delivers two generalized set of features contained in speech of healthy children and children with developmental dysphasia. Single utterances are then compared to these patterns.

The comparison is influenced by proper selection of number of atoms in each frequency band. Results presented were obtained for constant number of atoms (10) in each the band. Possible improvement might be obtained by determining threshold based on power distribution in atoms according to the order in which are they determined by modified MP algorithm (see figures 15.1 and 15.1).

Presented experiment shows that the method has ability to distinguish between utterances pronounced by healthy children and children suffering developmental dysphasia. This ability was also demonstrated in experiment described in chapter 14. However the difference in the procedure allows classifying just one speaker.

Experiment was worked out only for two types of utterances. Results of the experiment shows that the method has have potential for further extension to perform fine classification and distinguish stage of the disease. To gather more robust classification for one speaker it is substantial to judge several utterances and deliver final classification of the speaker based on results obtained for all these utterances.

To obtain more stable results and also to sort out speakers to the groups according to the degree of impairment, it is expedient to use more than two generalized sets (trained KSOMs).

Limiting factor for performing more experimental work and perform all adjustment is implementation. The algorithms were implemented in Python language [**?**]. The advantage is that all changes and adjustment might be done relatively easy, however extensive computation demands with conjunction of interpreted language produce relatively long computation times. It takes approximately 16 hours to compute results presented in this chapter .

Chapter 16

Conclusion

The method described in the thesis was developed to analyse disordered children speech. We focused on children with developmental dysphasia and believe that utilization of the method in clinical practise will bring more insight into progression and treatment of the disease and help to think out efficiently treatment of the disease.

This work is only a part of on-going research project focused on treatment of developmental dysphasia. In cooperation with the department of Paediatric Neurology in 2nd Faculty of Medicine of Charles University in Prague we are developing methods for utterance analysis that further advance diagnosis of the children with DD and help to find the most efficient therapy.

Laboratory of Artificial Neural Network Applications is focused on development of classification method based on artificial neural network. ANN, namely Kohonen Self-Organizing Maps, were chosen because of their robustness to artefacts and noise present in the signal. This specific feature helps to develop methods that might be later used directly in clinical practise without any special and expensive equipment. Also, children examined is not forced to change the known environment and might stay in examination room of specialist that is well known for him. This advantage is vital for not insignificant part of patients that refuses to communicate if they are in unknown environment.

Another advantage is that KSOM is an effective platform for visualization of high-dimensional data. This unique ability helps to fully understand contents of a data set and exploit contents in a way that is more convenient for physicians.

Thesis started with testing of commonly used speech parameterization (MFCC, LPC and PLP). We observed unreliable classification results, especially in the case where MFCC and PLP were utilized. Results reported shown that these parameterization are not suitable for speech with major impairment.

Classification based on MFCC could be used in the task of distinguishing whether a child is healthy or suffers developmental dysphasia. However, MFCC is not convenient for the determining the progress of the disease.

For the task of deterring the progress of the disease, PLP and LPC parameterizations seem to be more suitable. Classification based on PLP coefficients shows similar inaccuracy as MFCC, but the results are not so much influenced as in the case of MFCC.

Explanation to the observation is that parameterization used were developed and optimized for speech recognition. Developmental dysphasia causes impairment that is over limits that were taken into account when the parameterizations were designed and thus the main features present in the speech of children suffering DD are suppressed. For the task given, purpose of the parameterization is not to give perfect representation for recognition, but rather describe the differences. Because of that we decided to develop parameterization that is suitable for speech of dysphatic children and possibly be adaptable to another impairment.

We focused on parameterization based on wavelets and, after several introductory experiments, decided to further explore class of pursuit algorithms. Idea was to adjust some of the algorithms to perform well on speech with impairment and avoid time-consuming time of labelling utterances. Labelling of the utterances is difficult because of mispronunciation, various artefact caused by the movements of fidget children. Intention was to develop algorithm that does that autonomously or doesn't rely on labelling.

Finally, matching pursuit algorithm was chosen. The algorithm is reported to be utilized in fields of signal, image and video coding, shape representation and recognition and as well in the field of biomedical signal analysis (EEG and ECG). The algorithm must have been improved to perform better on speech signal.

We decided not implement improvements that take into consideration features of vocal or auditory tract - i.e. how the speech is produced and perceived and its advanced implementation. Reason for declination was that these techniques are utilized in parameterization mentioned and these parameterization were observed not to work perfectly on speech with impairment.

Rather, we adopted simple time and frequency dividing scheme and ensure that decomposition obtained from MP are not focused only on the lower part of the spectral where the energy of the signal is concentrated.

The parameterization using modified MP algorithm and feature extraction by Kohonen Self-Organizing Maps has potential to be further extended for software intended to clinical praxis. The method has been developed to provide more precise results that allows to classify children with developmental

dysphasia into several groups based on the degree of disease. Presented experiment shows that the method has ability to distinguish between utterances pronounced by healthy children and children suffering developmental dysphasia.

16.1 Further Development

To gather more robust classification for one speaker it is substantial to judge several utterances and deliver final classification of the speaker based on results obtained for all these utterances. To find and adjust overall classification scheme, it is necessary to compare results obtained from the method with the findings of physicians and discuss eventual discrepancies. Final aim is to develop method that will classify children to three groups: healthy, with DD (light form, medium form and hard form). Final method should incorporate comparison of one utterance to several groups to improve robustness.

Since no features specific for DD were taken into account during development and the method was developed in a way that allows adaptation to features in speech in general, it is probable that the classification based on matching pursuit and Kohonen Self-Organizing Maps will perform well for another kind of impairment.

Also, in classification method presented is a still a lot of degrees of freedom either for matching pursuit (size of γ vectors, number of features, extent of dictionary, etc.) as well as for KSOM (size of the network, adjusting of gain during training). Settings of these parameters was adjusted for utterances in experiments. Although settings would probably work for the rest of utterances that are on list (and recorder during each session with therapist), it is desirable that influence and impacts of variation of this parameterization has to be evaluated and described.

Before extensive testing must be rewritten to further decrease computation demands. The current situation is that the algorithm is implemented in python [**?**]. This is advantageous when dealing with development, since it allows software to be easily debugged and modified. But to fine tune algorithm to perform fine classification task on DD (determine advance of disease) an extended set of computation is to be performed. Nowadays the one series of computation for one set of parameters takes approximately 5 to 13 hours on powerful server.

The algorithm is now being rewritten in C++ language to be computed in multithreaded way on many cores in one time. Also, implementation on Epiphany parallel architecture accelerator [**?**] is being considered. Since both main algorithmic parts, matching pursuit and KSOM, are easily parallelized, this allows setting of all parameters to finally perform well enough to perform such extensive experiments.

References

[1] Tučková J., Komárek V., "Effectiveness of speech analysis by self-organizing maps in children with developmental language disorders," *Neuroendocrinology Letters*, vol. 29, no. 6, pp. 939 – 948, 2008, ISSN 0172-780X.

[2] Vesanto J., Alhoniemi E., "Clustering of the self-organizing map," *IEEE Transactions on Neural Networks*, vol. 11, no. 3, pp. 586 – 600, may 2000, ISSN 1045-9227.

[3] Tallal P., Piercy M., "Defects of auditory perception in children with developmental dysphasia," in *Developmental Dysphasia*, Wyke M.A., Ed. Academic Press, 1978.

[4] Archer L. A., "Manual motor functions in developmental dysphasia," M.S. thesis, McMaster University, Aug. 1980, Open Access Disertation and Theses, Paper 2744.

[5] Hoeffner J. H., McClelland J. L., "Can a perceptual processing deficit explain the impairment of inflectional morphology in developmenal dysphasia? A computational investigation.," in *Proceedings of the 25th Annual Child Language Research Forum*. Center for the study of Language and Information, Stanford University, 1993.

[6] Norbury C. F., Tomblin J. B., Bishop D. V. M., *Understanding developmental language disorders: from theory to practice*, chapter G. Baird: Assessment and investigation of children with developmental language disorder, Psychology, Hove, 2008, ISBN 9781841696669.

[7] Pospíšilová L., "Diagnostické otázky k vývojové dysfázii - diagnostics questions of developmental dysphasia," *Vox Pediatrie, Journal of General Practitioner for Children and Young*, vol. 5, no. 1, pp. 25 – 27, 2005, ISSN 1213-2241 (in Czech).

[8] Singh S., Bookless T., "Analysing spontaneous speech in dysphasic adults," *International Journal of Applied Linguistics*, vol. 7, no. 2, pp. 165 – 182, 1997, ISSN 1473-4192.

[9] Nasr J. T., Gabis L., Savatic M., Andriola M. R., "The electroencephalogram in children with developmental dysphasia," *Epilepsy & Behavior*, vol. 2, pp. 115 – 118, 2001, ISSN 1525-5050.

[10] Hermansky H., Morgan N., "RASTA processing of speech," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 578 – 589, Oct. 1994, ISSN 1558-7916.

[11] Durka P. J., Matysiak A., Montes E. M., Sosa P. V., Blinowska K. J., "Multichannel matching pursuit and EEG inverse solutions," *Journal of Neuroscience Methods*, vol. 148, no. 1, pp. 49 – 59, Oct. 2005, ISSN 0165-0270.

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

[13] Neuschlová L., Štěrbová K., Žáčková J., Komárek V., "Epileptiform activity in children with developmental dysphasia: quantification of discharges in overnight sleep video-EEG," *Epileptic Disord*, vol. 9, pp. S28 – 35, Dec. 2007, Supplement 1, ISSN 1950-6945.

[14] Komárek V., Hrnčír Z., "Analyses of EEG recordings," *Neural Network World*, vol. 14, no. 1, pp. 21 – 25, 2004, ISSN 1210-0552.

[15] Geach J. E., "Unsupervised self-organised mapping: a versatile empirical tool for object selection, classification and redshift estimation in large surveys," *Monthly Notices of the Royal Astronomical Society*, vol. 419, no. 3, pp. 2633 – 2645, Jan. 2012, ISSN 1365-2966.

[16] Tetzlaff R., Senger V., "The seizure prediction problem in epilepsy: Cellular nonlinear networks," *IEEE Circuits and Systems Magazine*, vol. 12, no. 4, fourth quarter 2012, ISSN 1531-636X.

[17] Singh S., "Quantitative classification of conversational language using artificial neural networks," *Aphasiology*, vol. 11, no. 9, pp. 829 – 844, 1997, ISSN 1464-5041.

[18] Chaloupka Z., Uhlíř J., "Speech perception deficits and the underlying nature of developmental dysphasia," *Radioengineering*, vol. 16, no. 1, 2007, ISSN 1210-2512.

[19] Oller D. K., Niyogi P., Gray S., Richards J. A., Gilkerson J., Xu D., Yapanel U., Warren S. F., "Automated vocal analysis of naturalistic recordings from children with autism, language delay, and typical development," in *Proceedings of the National Academy of Sciences of the United States of Americe*, July 2010, pp. 13354 – 13359.

[20] Nejepsová M., Janda J., Čmejla R., Vokřál J., "The severity rating of developmental dysphasia by utterances from 5–7 years old patients," in *Proceedings of the International Conference on Applied Electronics (AE)*, Sept. 2012, pp. 191 – 194.

[21] Ananth I., Ofoegbu U., Yantorno R., Smolenski B., "Speaker distinguishing distances: A comparative study," *International Journal of Speech Technology*, vol. 10, 2007, ISSN 1381-2416.

[22] Vavřina J., Zetocha P., Tučková J., "Detection of degree of developmental dysphasia based on methods of vowel analysis," in *Proceedings of the 35th international conference on Telecommunications and Signal Processing (TSP)*, July 2012, pp. 503 – 507.

[23] Tučková J., Zetocha P., "Speech analysis of children with developmental dysphasia by supervised SOM," *Neural Network World*, vol. 16, no. 6, pp. 533 – 545, 2006, ISSN 1210-0552.

[24] Psutka J., Müller L., Matoušek J., Radová V., *Mluvíme s počítačem česky (Talking with Computer in Czech)*, Academia, 2006, ISBN 80-200-1309-1 (in Czech).

[25] Rabiner L., Juang B. H., *Fundamentals of Speech Recognition*, Prentice-Hall, 1993, ISBN 978-0130151575.

[26] Mallat S., Zhang Z., "Matching pursuit with time-frequency dictionaries," *IEEE Transactions on Signal Processing*, vol. 41, no. 12, pp. 3397 – 3415, Dec. 1993, ISSN 1053-587X.

[27] Mallat S., *A Wavelet Tour of Signal Processing*, Academic Press, 2nd edition, 1999, ISBN 978-0124666061.

[28] Schneider R., Lau S., Kuhlmann L., Vogrin S., Gratkowski M., Cook M., Haueisen J., "Matching pursuit based removal of cardiac pulse-related artifacts in EEG/fMRI," in *World Academy of Science, Engineering and Technology*, 2011.

[29] Vařeka L., "Matching pursuit for p300-based brain-computer interfaces," in *Proceedings of the 35th Internation Conference on Telecommunications and Signal Processing (TSP)*, July 2012, pp. 513 – 516.

[30] Bénar C. G., Papadopoulo T., Torrésani B., Clerc M., "Consensus matching pursuit for multi-trial EEG signals," *Journal of Neuroscience Methods*, vol. 180, no. 1, pp. 161 – 170, May 2009, ISSN 0165-0270.

[31] Durka P. J., Blinowska K. J., "Analysis of EEG transients by means of matching pursuit," *Annals of Biomedical Engineering*, vol. 23, pp. 608 – 611, 1995, ISSN 1573-9686.

[32] Durka P., *Matching Pursuit and Unification in EEG Analysis*, Artech House, 2007, ISBN 978-1580533041.

[33] Grais E. M., Erdogan H., "Single channel speech-music separation using matching pursuit and spectral masks," in *IEEE 19th Conference on Signal Processing and Communications Applications (SIU)*, Apr. 2011, pp. 323 – 326.

[34] Sturm B. L., Gibson J. D., "Matching pursuit decompositions of non-noisy speech signals using several dictionaries," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP'06)*, May 2006, vol. 3.

[35] Gabor D., "Theory of communication," *Journal of I.E.E.*, vol. 93, no. 26, pp. 429 – 441, 1946.

[36] Stern R. M., Morgan N., "Hearing is believing: Biologically inspired methods for robust automatic speech recognition," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 34 – 43, nov 2012, ISSN 1053-5888.

[37] Chi T., Ru P., Shamma S. A., "Multiresolution spectrotemporal analysis of complex sounds," *Journal of the Acoustical Society of America*, vol. 118, no. 2, pp. 887 – 906, aug 2005, ISSN 0001-4966.

[38] Kim C., Stern R. M., "Power-normalized cepstral coefficients (PNCC) for robust speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, march 2012, pp. 4101 – 4104, ISSN 1520-6149.

[39] Gu L., Harris J. G., Shrivastav R., Sapienza Ch., "Disordered speech assessment using automatic methods based on quantitative measures," *EURASIP Journal on Applied Signal Processing*, vol. 2005, no. 9, pp. 1400 – 1409, Jan. 2005, ISSN 1687-6180.

[40] Krstulovic S., Gribonval R., "MPTK: Matching Pursuit made tractable," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP'06)*, May 2006, vol. 3, pp. 496 – 499.

[41] Lutz M., *Programming Python*, O'Reilly, 4th edition, 2011, ISBN 978-0596158101.

[42] Hermansky H., "Perceptual linear predictive (PLP) analysis of speech," *Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738 – 1752, Apr. 1990, ISSN 0001-4966.

[43] Davis S. B., Mermelstein P., "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,"

IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 28, no. 4, pp. 357 – 366, Aug. 1980, ISSN 0096-3518.

[44] Milner B., Shao X., "Clean speech reconstruction from MFCC vectors and fundamental frequency using an integrated front-end," *Speech Communication*, vol. 48, no. 6, pp. 697 – 715, June 2006, ISSN 0167-6393.

[45] Tyagi V., Wellekens C., "On desensitizing the mel-cepstrum to spurious spectral components for robust speech recognition," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'05)*, 2005, vol. 1, pp. 529 – 532.

[46] Zhou X., Garcia-Romero D., Duraiswami R., Espy-Wilson C., Shamma S., "Linear versus mel frequency cepstral coefficients for speaker recognition," *IEEE Automatic Speech Recognition and Understanding Workshop*, 2011.

[47] Singh N., Khan R. A., Shree R., "MFCC and prosodic feature extraction techniques: A comparative study," *International Journal of Computer Applications*, vol. 54, no. 1, pp. 9–13, September 2012, Published by Foundation of Computer Science, New York, USA, ISSN 0975 - 8887.

[48] Markel J. D., Gray A. H., *Linear Prediction of Speech*, Springer-Verlag, 1976, ISBN 0-13-007444-6.

[49] Sovka P., Pollák P., *Vybrané metody číslicového zpracování signálů - Selected Method of Digital SIgnal Processing*, CTU publishing, 2003, ISBN 80-01-02821-6 (in Czech).

[50] Sigmund M., *Analýza řečových signálů: přednášky - Analysis of Speech Signals: lectures*, Vysoké Učení Technické, Fakulta elektrotechniky a informatiky, Ústav radioelektroniky, 2000, ISBN 80-214-1783- 8 (in Czech).

[51] Makhoul J., "Linear prediction: A tutorial review," in *Proceedings of the IEEE*, 1975, pp. 561 – 580.

[52] Atal B. S., Hanauer S. L., "Speech analysis and synthesis by linear prediction of the speech wave," *Journal of the Acoustical Society of America*, vol. 50, pp. 637 – 655, Aug. 1971, ISSN 0001-4966.

[53] Hermansky H., Morgan N., "RASTA extensions: robustness to additive and convolutional noise," in *Speech Processing in Adverse Conditions (SPAC-1992)*, Cannes-Mandelieu, France, Nov. 1992, pp. 115 – 118.

[54] Gowdy J. N., Tufekci Z., "Mel-scaled discrete wavelet coefficients for speech recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing ICASSP*, 2000, vol. 3, pp. 1351 – 1354, ISSN 1520-6149.

[55] Mporas I., Ganchev T., Siafarikas M., Fakotakis N., "Comparison of speech features on the speech recognition task," *Journal of Computer Science*, vol. 3, no. 8, pp. 608–616, 2007, ISSN 1860-4749.

[56] Tan B. T., Lang R., Schroder H., Spray A., Dermody P., "Applying wavelet analysis to speech segmentation and classification," in *Proceedings of Wavelet Applications*, 1994, pp. 750 – 761.

[57] Rasetshwane D. M., Boston J. R., Li C.-C., "Identification of speech transients using variable frame rate analysis and wavelet packets," in *Proceedings of the 28th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS)*, Sept. 2006, pp. 1727 – 1730, ISSN 1557-170X.

[58] Favero R. F., King R. W., "Wavelet parameterization for speech recognition: variations in translation and scale parameters," in *Proceedings of the International Symposium on Speech, Image Processing and Neural Networks ISSIPNN*, Apr. 1994, vol. 2, pp. 694 – 697.

[59] Ricotti L. P., "Multitapering and a wavelet variant of MFCC in speech recognition," *IEEE Proceedings - Vision, Image and Signal Processing*, vol. 152, no. 1, pp. 29 – 35, Feb. 2005, ISSN 1350-245X.

[60] Korba M. Ch. A., Messadeg D., Djemili R., Bourouba H., "Robust speech recognition using perceptual wavelet denoising and mel-frequency product spectrum cepstral coefficient features," *Informatica (Slovenia)*, pp. 283–288, 2008, ISSN 1854-3871.

[61] Kadambe S., Srinivasan P., "Application of adaptive wavelets for speech coding," in *Proceedings of the IEEE-SP International Symposium on Time-Frequency and Time-Scale Analysis*, oct 1994, pp. 632 – 635.

[62] Siafarikas M., Ganchev T., Fakotakis N., "Wavelet packet bases for speaker recognition," in *Proceedings of the 19th IEEE International Conference on Tools with Artificial Intelligence ICTAI*, oct. 2007, vol. 2, pp. 514 – 517.

[63] Kotnik B., Kacic Z., Horvat B., "The usage of wavelet packet transformation in automatic noisy speech recognition systems," in *Proceedings of the IEEE Region 8 EUROCON. Computer as a Tool.*, Sept. 2003, vol. 2, pp. 131 – 134.

[64] Kotnik B., Vlad D., Kacic Z., Horvat B., "Robust MFCC feature extraction algorithm using efficient additive and convolutional noise reduction procedures," in *Proceedings of the 7th International Conference on Spoken Language Processing ICSLP - EUROSPEECH*, 2002, pp. 445 – 448.

[65] Tufekci Z., Gowdy J. N., "Feature extraction using discrete wavelet transform for speech recognition," in *Proceedings of the IEEE Southeastcon*, 2000, pp. 116 – 123.

[66] Fu Q., Yi K. C., "Bark wavelet transform of speech and it's application in speech recognition," *Journal of Electronics*, vol. 10, no. 28, Oct. 2000, ISSN 1993-0615.

[67] Zhang X.-Y., Bai J., Liang W.-Z., "The speech recognition system based on bark wavelet MFCC," in *Proceedings of the 8th International Conference on Signal Processing*, 2006, vol. 1.

[68] Zhang X.-Y., Bai J., "The speech recognition based on the bark wavelet and CZCPA features," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, Oct. 2006, pp. 318 – 321.

[69] Favero R. F., "Compound wavelets: wavelets for speech recognition," in *Proceedings of the IEEE-SP International Symposium on Time-Frequency and Time-Scale Analysis*, Oct. 1994, pp. 600 – 603.

[70] Kidae K., Youn D. H., Lee Ch., "Evaluation of wavelet filters for speech recognition," in *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics*, 2000, vol. 4, pp. 2891 – 2894.

[71] Galka J., Ziolko M., "Wavelet parameterization for speech recognition," in *Proceedings of AN ISCA Tutorial and Research Workshop on Nonlinear Speech Processing (NOLISP)*, 2009.

[72] Long Y., Gang L., Jun G., "Selection of the best wavelet base for speech signal," in *Proceedings of 2004 International Symposium on Intelligent Multimedia, Video and Speech Processing*, oct. 2004, pp. 218 – 221.

[73] Agbinya J. I., "Discrete wavelet transform techniques in speech processing," in *Proceedings of the IEEE Digital Signal Processing Applications TENCON*, nov 1996, vol. 2, pp. 514 – 519.

[74] Daubechies I., Maes S., "A nonlinear squeezing of the continuous wavelet transform based on auditory nerve models," in *Wavelets in Medicine and Biology*, A. Aldroubi, M. Unser, Ed., pp. 527 – 546. CRC Press, 1996, ISBN 978-0849394836.

[75] Gupta M., Gilbert A., "Robust speech recognition using wavelet coefficient features," in *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding ASRU*, 2001, pp. 445 – 448.

[76] Pham T. V., Kubin G., Rank E., "Robust speech recognition using adaptive noise threshold estimation and wavelet shrinkage," in *Proceedings of*

the Second International Conference on Communications and Electronics ICCE, June 2008, pp. 206 – 211.

[77] Gandhiraj R., Sathidevi P. S., "Auditory-based wavelet packet filter-bank for speech recognition using neural network," in *Proceedings of the International Conference on Advanced Computing and Communications (ADCOM)*, Dec. 2007, pp. 666 – 673.

[78] Daqrouq K., Abu-Isbeih I. N., Alfauri M., "Speech signal enhancement using neural network and wavelet transform," in *Proceedings of the 6th International Multi-Conference on Systems, Signals and Devices (SSD)*, Mar. 2009, pp. 1 – 6.

[79] Long C. J., Datta S., "Wavelet based feature extraction for phoneme recognition," in *Proceedings of the Fourth International Conference on Spoken Language ICSLP*, oct 1996, vol. 1, pp. 264 – 267.

[80] Wickerhauser M. V., "Best-adapted wavelet packet bases," in *Different Perspectives on Wavelets*, I. Daubechies, Ed., number 47 in Proceedings of Symposia in Applied Mathematics, pp. 155–171. American Mathematical Society, San Antonio, Texas, Jan. 1993, Minicourse lecture notes, ISBN 0-8218-5503-4.

[81] Coifman R. R., Wickerhauser M. V., "Entropy-based algorithms for best basis selection," *IEEE Transactions on Information Theory*, vol. 38, no. 2, pp. 713 – 718, Mar. 1992, ISSN 0018-9448.

[82] Youssef S. M., "A robust automated speech classification using hybrid wavelet-based architecture," in *Proceedings of the National Radio Science Conference NRSC*, mar 2008, pp. 1 – 8.

[83] Coifman R. R., Wickerhauser M. V., "Entropy-based algorithms for best basis selection," *IEEE Transactions on Information Theory*, vol. 38, no. 2, pp. 713 – 718, mar 1992, ISSN 0018-9448.

[84] Brislaw C. M., "Fingerprints go digital," *Notices of the AMS*, vol. 42, no. 11, pp. 1272 – 1282, 1995, ISSN 1088-9477.

[85] Davis G., Mallat S., Avellaneda M., "Adaptive greedy approximations," *Constructive Approximation*, vol. 13, pp. 57 – 98, 1997, ISSN 0176-4276.

[86] Lustig I. J., Marsten R., Shanno D.F., "Interior point methods for linear programming: Computational state of the art," *ORSA Journal on Computing*, vol. 6, no. 1, pp. 1 – 14, 1994, ISSN 0899-1499.

[87] Bergeaud F., Mallat S., "Matching pursuit of images," in *Proceedings of International Conference on Image Processing*, 1995, vol. 1, pp. 53 – 56.

[88] Neff R., Zakhor A., "Very low bit-rate video coding based on matching pursuits," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 7, no. 1, pp. 158 – 171, 1997, ISSN 1051-8215.

[89] Mendels F., Vandergheynst P., Thiran J. P., "Matching pursuit-based shape representation and recognition using scale-space," *International Journal of Imaging Systems and Technology*, vol. 16, no. 5, pp. 162 – 180, 2006, ISSN 1098-1098.

[90] Tosic I., Frossard P., Vandergheynst P., "Progressive low bit rate coding of simple 3D objects with matching pursuit," in *Proceedingd of the Data Compression Conference DCC*, Mar. 2005.

[91] Jalaleddine S. M., Hutchens C. G., Strattan R. D., Coberly W. A., "ECG data compression techniques - a unified approach," *IEEE Transactions on Biomedical Engineering*, vol. 37, no. 4, pp. 329 – 343, May 1990, ISSN 0018-9294.

[92] Nakashizuka M., Niwa K., Kikuchi H., "ECG data compression by matching pursuit with multiscale atoms," *IEICE Transaction Fundamentals*, vol. E84-A, no. 8, pp. 1919 – 2024, Aug. 2001, ISSN 1745-1337.

[93] Pantelopoulos A., Bourbakis N., "Efficient single-lead ECG beat classification using matching pursuit based features and an artificial neural network," in *Proceeding of the 10th IEEE International Conference on Information Technology and Applications in Biomedicine (ITAB)*, Nov. 2010, pp. 1 – 4.

[94] Tan Q., Fang B., Wang P., "Improved simultaneous matching pursuit for multi-lead ECG data compression," in *Proceedings of the International Conference on Measuring Technology and Mechatronics Automation (ICMTMA)*, Mar. 2010, vol. 2, pp. 438 – 441.

[95] Skretting K., Engan K., Husoy J. H., "ECG compression using signal dependent frames and matching pursuit," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'05)*, Mar. 2005, vol. 4, pp. 585 – 588.

[96] Bardoňová J., Provazník I., Nováková M., "Matching pursuit decomposition for detection of frequency changes in experimental data - application to heart signal recording analysis," *Scripta Medica*, vol. 79, no. 5 – 6, pp. 279 – 288, Dec. 2006, ISSN 1211-3395.

[97] Zhao Z., Yang L., "ECG identification based on matching pursuit," in *Proceedings of the 4th International Conference on Biomedical Engineering and Informatics (BMEI)*, Oct. 2011, vol. 2, pp. 721 – 724.

[98] Kohonen T., *Self-Organizing Maps*, Springer-Verlag, 3rd edition, 2001, ISBN 3540679219.

[99] Schulerud H., Albregtsen F., "Many are called, but few are chosen. feature selection and error estimation in high dimensional spaces," *Computer Methods and Programs in Biomedicine*, vol. 73, no. 2, pp. 91 – 99, 2004, ISSN 0169-2607.

[100] Somervuo P., "Competing hidden markov models on the self-organizing map," in *Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks IJCNN*, 2000, vol. 3, pp. 169 – 174, ISSN 1098-7576.

[101] Hinton G., Deng L., Yu D., Dahl G. E., Mohamed A., Jaitly N., Senior A., Vanhoucke V., Nguyen P., Sainath T. N., Kingsbury B., "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82 – 97, nov 2012, ISSN 1053-5888.

[102] Linde Y., Buzo A., Gray R., "An algorithm for vector quantizer design," *IEEE Transactions on Communications*, vol. 28, no. 1, pp. 84 – 95, jan 1980, ISSN 0090-6778.

[103] Vesanto J., Himberg J., Alhoniemi E., Parhankangas J., "SOM toolbox for matlab 5, report a57," Tech. Rep., Helsinky University of Technology, 2000.

[104] Thiran P., Peiris V., Heim P., Hochet B., "Quantization effects in digitally behaving circuit implementations of kohonen networks," *IEEE Transactions on Neural Networks*, vol. 5, no. 3, pp. 450 – 458, May 1994, ISSN 1045-9227.

[105] Gemello R., Albesano D., Mana F., Moisa L., "Multi-source neural networks for speech recognition: a review of recent results," in *Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks (IJCNN)*, 2000, vol. 5, pp. 265 – 270, ISSN 1098-7576.

[106] Shum W.-H., Jin H.-D., Leung K.-S., Wong M.-L., "A self-organizing map with expanding force for data clustering and visualization," in *Proceedings of the IEEE International Conference on Data Mining ICDM*, 2002, pp. 434 – 441.

[107] Rauber A., Merkl D., Dittenbach M., "The growing hierarchical self-organizing map: exploratory analysis of high-dimensional data," *IEEE Transactions on Neural Networks*, vol. 13, no. 6, pp. 1331 – 1341, Nov. 2002, ISSN 1045-9227.

[108] Su M.-Ch., Chang H.-T., "Fast self-organizing feature map algorithm," *IEEE Transactions on Neural Networks*, vol. 11, no. 3, pp. 721 – 733, may 2000, ISSN 1045-9227.

[109] Tvrdík V., Žák V., Tučková J., "Parametrizations of a children speech and their impact on classification of vowels," in *Proceedings of the Digital Technologies 2007*. 2007, University of Žilina, Fakulty of electrical engineering, (Proceedings on CD).

[110] Rabiner L., Shafer R. W., *Digital Processing of Speech Signals*, Prentice-Hall, 1978, ISBN 978-0132136037.

[111] Gwennap L., "Adapteva: More flops, less watts," *Microprocessor Report*, jun 2011, www.mpronline.com.

List of author's publications related to the thesis

[112] Tučková J., Bártů M., Zetocha P., Grill P., "Self-organizing maps as data classifier in medical applications," in *Proceedings of the International Conference on Neural Computation Theory and Applications*, Madeira, 2011, pp. 422 – 429, SciTePress, ISBN 978-989-8425-84-3.

[113] Bártů M., "Zpracování signálů neuronovými sítěmi - signal processing using neural networks," in *Analýza a zpracován signálů VI*, Praha, 2005, vol. 1, pp. 1–11, ČVUT FEL, Katedra teorie obvodů, ISBN 80-01-03217-5, (in Czech).

[114] Bártů M., Tučková J., "A classification method of children with developmental dysphasia based on disorder speech analysis," in *Proceedings of the 18th international conference on Artificial Neural Networks (ICANN'08)*, Heidelberg, 2008, vol. 1, pp. 822 – 828, Springer, Part II, ISBN 978-3-540-87558-1, ISSN 0302-9743.

[115] Bártů M., "Implementation of Kohonen Self-Organizing Map," in *ECMS 2007 & Doctoral School (EDSYS, GEET)*, Liberec, 2007, pp. 40–43, Technická univerzita v Liberci, ISBN 978-80-7372-202-9.

[116] Bártů M., "Speech disorder analysis using matching pursuit and kohonen self-organizing maps," *Neural Network World*, vol. 22, no. 6, pp. 519 – 533, 2012, ISSN 1210-0552.

[117] Bártů M., "Zjednodušení algoritmu pro výpočet Kohonenovy mapy - simplification of algorithm for kohonen map computation," in *Analýza a zpracování řečových a biologických signálů*, Praha, 2007, pp. 1–6, ČVUT v Praze, ISBN 978-80-01-03940-3, (in Czech).

[118] Tučková J., Bártů M., Zetocha P., *Aplikace umělých neuronových sítí při zpracování signálů - Application of Artifical Neural Network in Signal Processing*, chapter 8, pp. 94 – 126, CTU publishing, 2009, (in Czech).

[119] Bártů M., "Framework for Kohonen Self-Organizing Maps," in *Digital Technologies 2007-Book of Abstracts*, Žilina, 2007, p. 8, Slovenská elektrotechnická společnost, ISBN 80-8070-777-4.

[120] Bártů M., "Implementation of Kohonen Self-Organizing Map," in *The 22nd International Symposium on Computer and Information Sciences (ICTS - Program & Abstracts)*, Ankara, 2007, p. 54, Middle East Technical University.

[121] Bártů M., Tučková J., Šťastný J., "An Accelerator for Kohonen Self-Organizing Maps," in *Digital Technologies 2006 - 3rd International Workshop* , Žilina, 2006, vol. 1, pp. 1–6, University of Žilina, Faculty of Electrical Engineering, ISBN 80-8070-637-9.

[122] Bártů M., "Akcelerátor pro KSOM - an accelerator for KSOM," in *Analýza a zpracování řečových a biologických signálů - Sborník prací 2006*, Praha, 2006, vol. 1, pp. 1–7, ČVUT, ISBN 80-01-03621-9, (in Czech).

[123] Bártů M., Tučková J., Šťastný J., "A concept of accelerator for Kohonen Self-Organizing Maps," in *Proceedings EDS'06 IMAPS CS International Conference*, Brno, 2006, vol. 1, pp. 1–5, VUT v Brně, FEKT, ISBN 80-214-3246-2.

[124] Bártů M., "Speech Analysis Using Operator Maps," in *Digital Technologies 2008*, Žilina, 2008, vol. 1, University of Žilina, Faculty of Electrical Engineering, ISBN 978-80-8070-953-2.

[125] Bártů M., "Využití Kohonenových map pro analýzu řečových signálů dětí s vývojovou dysfázií - utilization of kohonen maps for analysis of speech signal of children with developmental dysphasia," in *Analýza a zpracování řečových a biologických signálů - sborník prací 2008*, Praha, 2008, pp. 1–6, České vysoké učení technické v Praze, ISBN 978-80-01-04243-4, (in Czech).

[126] Bártů M., "Speech Parametrization for Operator Maps," in *Digital Technologies 2009*, Žilina, 2009, vol. 1, TU v Žilině, ISBN 978-80-554-0150-8.

[127] Bártů M., "Možnosti využití algoritmů nelineární aproximace pro parametrizaci řečových signálů - possibilities of utilization of nonlinear approximation algorithms for parameterisation of speech signals," in *Analýza a zpracování řečových a biologických signálů - sborník prací 2009*, Praha, 2009, České vysoké učení technické v Praze, ISBN 978-80-01-04474-2, (in Czech).

[128] Zetocha P., Bártů M., Žůrek M., Tučková J., "Analýza řeči dětí s vývojovou dysfázií - analysis of speech of children with developmental dysphasia," in *Trendy v biomedicínském inženýrství, Sborník 7. česko-slovenské konference*, Praha, 2007, pp. 186–189, ČVUT v Praze, Fakulta biomedicínského inženýrství, ISBN 978-80-01-03777-5, (in Czech).