Jacek GRYGIEL[1], Paweł STRUMIŁŁO[1], Ewa NIEBUDEK-BOGUSZ[2]

Technical University of Lodz, Institute of Electronics, Łódź, Poland (1)
Department of Audiology and Phoniatrics, Nofer Institute of Occupational Medicine, Łódź, Poland (2)

Application of Mel Cepstral Representation of Voice Recordings for Diagnosing Vocal Disorders

Abstract. *The aim of this study was to assess the applicability of Mel Frequency Cepstral Coefficients (MFCC) of voice samples in diagnosing vocal nodules and polyps. Patients' voice samples were analysed acoustically with the measurement of MFCC and values of the first three formants. Classification of mel coefficients was performed by applying the Sammon Mapping and Support Vector Machines. For the tests conducted on 95 patients, voice disorders were detected with accuracy reaching approx. 80%.*

Abstract*. Celem niniejszej pracy była ocena możliwości zastosowania analizy tzw. współczynników cepstralnych (ang. Mel Cepstral Coefficients (MFCC)) dla próbek rejestrowanego głosu pacjentów we wspomaganiu diagnozy guzów i polipów. Rejestracje mowy pacjentów poddane zostały analizie akustycznej, w której zastosowano parametry MFCC oraz wartości trzech pierwszych formantów. Do klasyfikacji współczynników cepstralnych zastosowano odwzorowanie Sammona oraz tzw. Maszynę Wektorów Nośnych. W testach wykonanych dla 95 rejestracji mowy pacjentów, zaburzenia głosu zostały wykryte z ok. 80% dokładnością. (****Zastosowanie reprezentacji Mel Cepstralnej sygnału mowy do badania zaburzeń głosu).***

Keywords: MFCC, SVN, voice disorders, Sammon mapping
Słowa kluczowe: MFCC, SVN, zaburzenia głosu, odwzorowanie Sammona

Introduction

In recent years the effectiveness of diagnosing vocal cord disorders has improved considerably due to novel diagnostic approaches. Thus, there is also and increasing need for development of novel methods for objective assessment of patients' voice quality. In a number of studies it was shown that digital signal processing methods can offer an additional aid for improving the quality of diagnosing vocal cord disorders [7].

In particular, parameterization of patient's voice by means of Mel Frequency Cepstral Coefficients MFCC analysis proved to be a successful tool in speech modelling. Application of this method of speech parameterization for voice disorders was reported by Godino-Llorente [1]. He described more than 90% of accuracy in recognizing pathologies on the basis of voice recordings of sustained vowels. In order to reduce dimensionality features the GMM (Gaussian Mixture Models) was used in relation to Fisher discriminant ratio [2].

Approaches that combine classical parameterization techniques with data classifiers were also reported in literature. Application of Support Vector Machines (SVM) and GMM (Gaussian Mixture Models) gave 98.23% diagnostic accuracy [3]. Also application of MFCC and measurement of speech pitch dynamics were modelled by an HMM (Hidden Markov Model) classifier with 99.44% correct classification rates [4].

Most of the worldwide research on recognition of voice disorders is based on the speech recordings from the Massachusetts Eye and Ear Infirmary (MEEI) database. The study reported in this paper, on the other hand, was carried out with the use of the database delivered from the Nofer Institute of Occupational Medicine in Lodz, Poland.

In our approach the MFCC analysis for parameterization of voice disorders was adopted. Support Vector Machines were used for classifying speech feature vectors obtained from MFCC speech modelling. Signal acquisition methods used for recording of the speech samples and methods for feature extraction from voice signals are also outlined.

Voice recordings

Voice samples for the reported study were recorded at the Audiology and Phoniatrics Clinic, Nofer Insitute of Occupational Medicine in Lodz, Poland. The database included 40 samples from healthy patients, 36 from patients with vocal nodules and 19 samples from patients with vocal polyps. The recordings consist of the following sentences pronounced in Polish (in brackets English translation are given): „Ten dzielny żołnierz był z nim razem"(eng. „*This brave soldier was together with him*"), „Czy jestem zdrowy?"(eng. „*Am I healthy*?"), „Tak jestem zdrowy!"(eng. „*Yes I am healthy*"), „już jestem zdrowy!"(eng. „*I am already healthy*") and of a sustained vowel „a". The Polish sentences contain specific phonemes that have proved to expose characteristic voice disorders.

MFCC Analysis

The algorithm implementing MFCC for signal modelling is shown in Fig. 1.

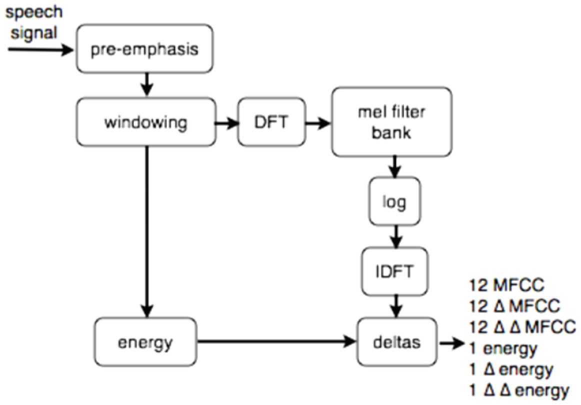

Fig.1. MFCC modelling scheme of speech

Pre-emphasis refers to a signal processing method designed to increase the magnitude of higher frequencies with respect to the magnitude of lower frequencies. The spectrum energy of pronounced vowels is mainly concentrated in the lower frequencies. This is due to the anatomical features of the larynx. By increasing the energy of high frequencies, signal components containing the higher formants can be emphasized. It was shown that application of pre-emphasis improves the accuracy of speech acoustic models [5]. Pre-emphasis was performed by using the following difference equation (1).

$$y[n] = x[n] - \alpha x[n-1] \quad (1)$$

where: $x[n]$ – are the input signal samples, $y[n]$ – are the output samples, $0.9 \leq \alpha \leq 1.0$

Extraction of features from the speech signal is difficult due to large variations of the signal spectrum in time. Hence, windowing is the procedure that is used for dividing the speech signal into segments with intervals between 20 ms and 30 ms. In such short intervals the signal can be assumed stationary. The process of windowing is characterized by the following parameters:
1) Window width
2) Window shape, e.g. the Hamming window

$$w(n) = 0.54 - 0.46\cos(2\pi n/(N-1)) \tag{2}$$

where: N – represents the width, in samples, of a discrete time window function, $0 \leq \alpha \leq N$-1

3) Time shift between successive time windows

The next step of the MFCC algorithm is the computation of the Fast Fourier Transform. Pre-scaling the signal by the Hamming window improves accuracy of the obtained spectrum. The Discrete Fourier Transform for signal $x(n)$, n=0,1,…,N-1 (3) is given by:

$$X[k] = \sum_{n=0}^{N-1} x[n] e^{\frac{-2\pi i}{N}nk} \tag{3}$$

where: k=0,1,...,N-1 and $X[k]$ – are the Fourier coefficients

Mel scale refers to pitch – that is, the auditory impression of identifying the location of the tone in the frequency scale. It was computed referring to research, which examined human hearing [5]. The mel frequency scale is defined in equation (4). The relationship between the mel scale and the physical frequency given in Hz is non-linear. Using the mel scale it is possible to compute a filter bank which makes the non-linear frequency analysis similar to the manner that takes place in the human ear i.e.:

$$f_{Mel} = 2595\log_{10}(1 + f_{Hz}/700) \tag{4}$$

where: f_{Hz} – stands for the signal's physical frequency

Equations 5÷9 define the scheme for building a bank of mel filters:

$$\Delta\varphi = (\varphi_{\max} - \varphi_{\min})/(M+1) \tag{5}$$

where: Φ_{max}, Φ_{min} – stand for cut-off frequencies of the filter, *M* is a numer of filters in the mel scale

$$\varphi_c(m) = m \cdot \Delta\varphi \tag{6}$$

where: Φ_c – defines frequency value for the m-filter

$$f_c(m) = 700(10^{\varphi_c(m)/2595} - 1) \tag{7}$$

where: f_c – is the filter frequency

$$f_k(k) = kf_s/N \tag{8}$$

where: f_k – is a middle frequency of the filter

$$H(k,m) = \left\{ \begin{array}{lll} 0 & for & f(k) < f_c(m-1) \\ \dfrac{f(k) - f_c(m-1)}{f_c(m) - f_c(m-1)} & for & f_c(m-1) \leq f(k) < f_c(m) \\ \dfrac{f(k) - f_c(m+1)}{f_c(m) - f_c(m+1)} & for & f_c(m) \leq f(k) < f_c(m+1) \\ 0 & for & f(k) \geq f_c(m+1) \end{array} \right\} \tag{9}$$

The filter bank given by matrix $H(k,m)$ is shown in Fig. 2.

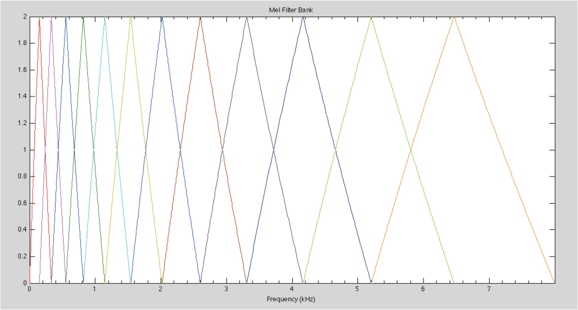

Fig. 2. The mel filter bank *H(k,m)*

The cepstrum coefficients are obtained by calculating the inverse Fourier Transform. Then the maximum amplitude of the signal is chosen and used as the cepstrum coefficients.

MFC Coefficients are real numbers hence in order to reduce the cost of calculations the Discrete Cosine Transform (DCT) was implemented:

$$C(l) = \sum_{m=1}^{M} X(m)\cos[l\frac{\pi}{M}(m - \tfrac{1}{2})] \tag{10}$$

where: $C(l)$ – are the cepstral coefficients

Properties of the MFC Coefficients of the recorded speech are the following:

- coefficients of small order are related to the transmittance of voice along the vocal tract – these coefficients are the most important in identifying the speakers, because they are not very sensitive to noise and interferences generated in the vocal tract,
- coefficients of high order, represent signal noise interferences.

DCT performs partial de-correlation of the obtained coefficients, which is important in the modelling of their probability distributions.

Additional features commonly used in the analysis of the speech signals are then computed:

1) Energy:

$$E = \sum_{t=t_1}^{t_2} x^2[t] \tag{11}$$

2) Delta:

$$\Delta(t) = \frac{C(t+1) - C(t-1)}{2} \tag{12}$$

3) Delta Delta:

$$\Delta\Delta(t) = \frac{\Delta(t+1) - \Delta(t-1)}{2} \tag{13}$$

Speech signal parameters were computed for each frame lasting 20ms. Then the parameters corresponding to each frame are averaged to form a 39-element vector representing data assigned to one patient (as shown in Fig. 1)

The Sammon mapping

In this study the Sammon mapping was additionally applied to reduce vector dimensionality of the MFCC characteristics, which facilitates its graphical interpretation as well as calculations of data classification task. The Sammon mapping performs reduction of data dimension

while retaining its statistical structure and making data clusters belonging to the same class better concentrated.
This approach was proposed because of difficulties in separating individual parameters which not clearly describe a particular class of patients. In the conducted statistical analysis of mel cepstral coefficients, similar scattering parameters in the classes of healthy patients and patients with disorders were observed (Fig. 3.).

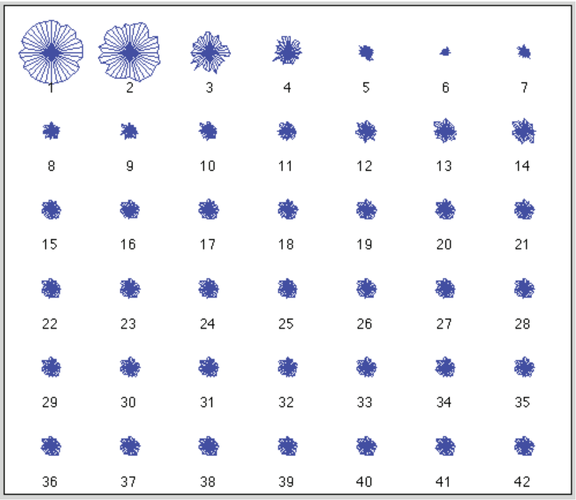

Fig.3. Illustration of the dispersion parameters for all computed MFCC coefficients for the set of patients with nodules, (averaged parameters are diagrammed after correction using the standard deviation).

The task of non-linear transformation that is applied in the Sammon mapping is to choose such vectors that minimize the error function [6]:

$$E = \frac{1}{c}\sum_{i<j}^{n}[d_{ij}^{*} - d_{ij}]^{2} / d_{ij}^{*} \tag{14}$$

where: d_{ij}^{*} – is a distance between *i* and *j* object in the input dimension, d_{ij} is a distance between i and j object in the reduced dimension.

Data classification

The Support Vector Machine as the main classifier and the Minimum Distance Classifier (MDC) for visual data separation was employed in the study. SVM is a classifier that belongs to the group of methods, which transform the data nonlinearly (a quadratic kernel function was used) in order to make the classification problem linearly separable [9]. As indicated earlier, the MDC was also implemented with a separation surface between centroids of 2 clusters representing two different data classes (e.g. data representing voice samples taken from healthy patients and data from patients with disorders). These centroids were determined as the so called prototype vectors belonging to the given class. The decision function for separating the class models is given by equation (15).

$$d_{j}(\mathbf{x}) = \mathbf{x}^{T}\mathbf{m}_{j} - \frac{1}{2}\mathbf{m}_{j}^{T}\mathbf{m}_{j} \tag{15}$$

where: $\boldsymbol{m}$ is a centroid vector for appropriate data class. Vector $\boldsymbol{x}$ is assigned to class $\boldsymbol{\omega_j}$ if its decision function $\boldsymbol{d_j(x)}$ is the largest.

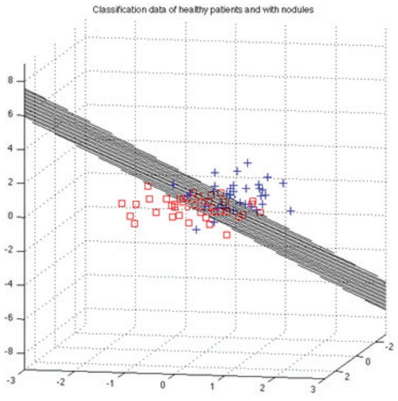

Fig.4. Visualisation of 3D data obtained by applying the Sammon mapping to a 39 dimensional feature vector set; squares – healthy patients, crosses – patients with nodules.

Results

By applying the MFCC algorithm to each of the recorded speech signal the vector of 39 parameters was obtained. The vector was augmented by the first three formants. These additional parameters were calculated by using standard Linear Prediction Coding (LPC) algorithm [8]. All database samples were randomized. The training set consisted of 2/3 of all available voice samples and the remaining samples were used for testing. First the Sammon mapping was used for reducing the data vector to three dimensions. Fig. 4 illustrates voice recordings that were classified to two separate classes corresponding to healthy and ill patients. The error function for the Sammon mapping is approx. $1.26 \cdot 10^{-14}$. Parameters of the receiver operating characteristic were used for evaluating the quality of the classifiers and are shown in Table 1. The mean accuracy is estimated at approx. 80% correct classification rates.

Table 1. Parameters of the receiver operating characteristic

Classifier	Sensitivity	Specificity
MDC	0.7528	0.8197
SVM	0.9091	0.8333

Table 2. Confusion Matrix for healthy patients and patients with vocal nodules

True\Predicted	Healthy	Ill
Healthy	11	2
Ill	2	10

Table 3. Confusion Matrix for healthy patients and patients with vocal polyps

True\Predicted	Healthy	Ill
Healthy	7	4
Ill	2	16

The SVM classifier applied in this research yields superior data separation results in the examined groups of patients (nodules, polyps, healthy). Appropriate confusion matrices are given in Table 2 and Table 3.

Discussion and conclusions

It was shown that the proposed implementation of the MFCC algorithm has proved effective in parameterizing speech signals recorded from both healthy subjects and patients with voice disorders. MFFC are well known to be used in speaker recognition systems but performed equally well in the presented application. However, there are two specific problems that were noticed in the conducted study. The main issue is the acoustic environment in which the voice samples were recorded. During analysis it was found that patients' speech was interfered by noise coming from the recording environment. These signal contaminations impede the final quality of classification.

The best results obtained in this research were approx. 90% of true positives (better rates were obtained for the vocal cords affected by the nodules than for the polyps). These results were obtained for groups of 40 healthy subjects, 36 patients with nodules and 19 patients with polyps. The number of subjects that were diagnosed as positive is still too small to draw general conclusions about the method's true diagnostic value. Currently, a new series of measurements is underway in order to collect a larger number of samples for further analyses.

Another observation made on the basis of the conducted study is that separating speech features taken from patients with nodules from the same speech features originating from patients with polyps is very unreliable.

REFERENCES

[1] J.I. Godino-Llorente, Ruben Fraile, N. Sãenz-Lechõn, Osma-Ruiz, P. Gomez-Vilda, Automatic detection of voice impairments from text-dependent running speech, Biomedical Signal Processing and Control, pp. 176-182, March 2011

[2] J.I. Godino-Llorente, P. Gomez-Vilda, Manuel Blanco-Velasco, Dimensionality Reduction of a Pathological Voice Quality Assessment System Based on Gaussian Mixture Models and Short-Term Cepstral Parameters, IEEE Transactions on Biomedical Engineering, vol. 53, no. 10, pp. 1943–1953, October 2006.

[3] J.D. Arias-Londoño, J.I. Godino-Llorente, N. Sãenz-Lechõn, V. Osma-Ruiz, G. Castellanos-Domínguez, Automatic Detection of Pathological Voices Using Complexity Measures, Noise Parameters, and Mel-Cepstral Coefficients, IEEE Transactions on Biomedical Engineering, vol. 58, no. 2, pp. 370–370, February 2011.

[4] A.A. Dibazar, S. Narayanan, T.W. Berger, Feature Analysis for Automatic Detection of Pathological Speech, Proceedings of the Second Joint EMBS/BMES Conference Houston, TX, USA, October 23–26, 2002.

[5] J. H. Martin, D. Jurafsky, Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition. Prentice Hall, 2 edition, June 2008.

[6] S. Osowski, Neural Networks: an algorithmic approach (in Polish), WNT Warszawa, 1996.

[7] C. Maciel, J. Pereira, and D. Stewart, “Identifying healthy and pathologically affected voice signals,” IEEE Signal Process. Mag., vol. 27, no. 1, pp. 120–123, Jan. 2010.

[8] Ce Peng, Wenxi Chen, Xin Zhu, Baikun Wan, Darning Wei, "Pathological Voice Classification Based on a Single Vowel's Acoustic Features," 7th IEEE International Conference on Computer and Information Technology, CIT 2007, pp.1106–1110, 16-19 Oct. 2007.

[9] I.R. Titze, “Workshop on acoustic voice analysis summary statement”, in Proc Workshop on Acoustic Voice Analysis, Denver, Colorado, February, 1994.

***Authors**: mgr inż. Jacek Grygiel, Instytut Elektroniki Politechniki Łódzkiej, ul. Wólczańska 211/215, 90-924 Łódź, e-mail: jacek.grygiel@gmail.com; dr hab. inż. Paweł Strumiłło, prof. PŁ, Instytut Elektroniki Politechniki Łódzkiej, ul. Wólczańska 211/215, 90-924 Łódź, e-mail: pawel.strumillo@p.lodz.pl; dr hab. n. med. Ewa Niebudek-Bogusz, Klinika Audiologii i Foniatrii Instytutu Medycyny Pracy w Łodzi, ul. św. Teresy 8, 91-348 Łódź, e-mail: ebogusz@imp.lodz.pl.*