Quantification of Linear and Non-linear Acoustic Analysis Applied to Voice Pathology Detection

Daria Panek, Andrzej Skalski, and Janusz Gajda

AGH University of Science and Technology,
Department of Measurement and Electronics,
Al. Mickiewicza 30, 30-059 Krakow
{dpanek,skalski,jgajda}@agh.edu.pl

Abstract. Present development of digital registration and methods of recorded voice processing are useful in detection of most pathologies and diseases of a human vocal tract. The recognition of the voice condition requires the creation of a model which is comprised of different acoustic parameters of speech signal. In this study a vector consisting of 31 parameters for analysing the speech signal was created. The speech parameters were extracted from time, frequency and cepstral domains. Using Principal Components Analysis the number of the parameters was reduced to 17. In order to validate the detection of the pathological voice signal, a tenfold cross-validation and confusion matrix were used. The goal and novelty of this work was the analysis of applicability of the parameters selectively used to assess the pathology.

Keywords: acoustics analysis, cepstral analysis, pathology detection, dysphonia, principal component analysis, cross validation.

1 Introduction

Currently, European standards emphasize the need for a comprehensive assessment of voice disorders with regard to objective methods. The value of acoustic analysis is increasingly appreciated as a diagnostic test for non-invasive and objective examination. In the broad sense of laryngeal and phoniatric diagnosis of voice and speech disorders acoustic analysis provides supporting and complementary studies. The value of these studies increased significantly throughout the last years with the introduction of high-speed digital voice analyzers. The detection accuracy of voice and speech disorders located in the larynx and in voice channel increases.

Acoustical analysis allows us to make physical description of the waveforms generated and emitted by the organ of the human voice and correlates well with the phoniatric state of proper and pathological voice. There are two methods to conduct voice analysis: classic, based on subjective assessment of voice examination and modern, based on objective acoustic analysis, spectrographic images, sonographic or time recording speech signal. The physical characteristics of the voice are determined to use the latest digital technology in the acoustic

E. Pietka, J. Kawa, and W. Wieclawek (eds.), *Information Technologies in Biomedicine, Volume 4,* Advances in Intelligent Systems and Computing 284,
DOI: 10.1007/978-3-319-06596-0_33,

analysis and a detailed statistical analysis of the results. Fidelity recording and processing the digital audio signal, which is a stochastic process, promotes the development of a growing number of measurable characteristics (characteristic for the different subjective characteristics) of human voices and allows good accuracy, objective assessment of discrete, unobtrusive overflow method, voice and speech disorders [1].

Currently, the best measurement methodology models, algorithms and different approaches for classification that could discriminate between normal and pathological voices are still being sought [2]. The performance of these systems is not perfect, but they are useful as an additional source of information for other laryngoscopial examinations [3,5].

In this work, the articulation in speech and its pathological deformation was examined. This includes tools and techniques used to detect deformations in the voice signal vocalised by an ill person and a healthy one. The aim was to select and describe those signal parameters that contain the most valuable information and show the highest sensitivity to speech deformations.

2 Material

The Saarbruecken Voice Database has been published online by Institute of Phonetics of the University of the Saarland [4]. It is a collection of voice recordings collected from more than 2000 people. Each of the recording sessions contains recordings of the vowels /a/, /i/, /u/ produced at normal, high, low, low-high-low pitch.

The length of the recordings with sustained vowels amounts to 1 – 4 seconds. All the recordings are sampled at 50 000 Hz with a resolution of 16-bit. The database contains 71 different well defined pathologies. In our work we used the recordings of the vowel /a/ of 850 women, of which 425 were healthy and 425 suffered from different voice disorders (167 suffered from hyperfunctional dysphonia,139 had vocal cord paresis, 119 suffered from other pathologies listed in the database) and 510 men, of which 255 were healthy and 255 were diagnosed with different pathologies, of which 46 men suffered from hyperfunctional dysphonia, 74 suffered from vocal cord paresis, 83 experienced laryngitis of which some also had leukoplakia. The rest of the men who underwent this examination suffered from other pathologies listed in the database. Recordings that were missing or damaged were excluded from the dataset. Only for women the analysis of the vowel /a/ was extended to include all intonations – low and high pitch. Because of the intrinsic differences in voice behaviour between men and women (and because the number of male and female speakers was not equal in all groups), parameters were statistically analysed for males and females separately.

3 Method

This study was carried out to assess the suitability of several methods for mapping speech signals in diagnostics of pathological speech. Each method was used

with reference to both correct and pathological speech samples. Firstly, the focus of examination was put on preliminary transformation of speech waveforms into a set of parameters whose values represented a basis for a diagnosis of the patient's disease.

The attention was drawn to the fact that the acoustic signal processing for a set of features whose values are so called parameters, is the basis for a description of the object's state in terms of diagnosis. Registration itself and its preliminary signal processing does not make it fully useful to the process of identifying and assessing changes in deformation and pathology. Therefore, it becomes necessary to develop and describe recorded phonetic tests using a set of parameters, which then, sorted out in the corresponding structure - a feature vector will be used to develop models of speech deformation. Such models can be the foundation of the recognition process, assessment of pathological changes or rehabilitation process. Analysis of the speech signals (Fig. 1) was performed with 31 parameters: root

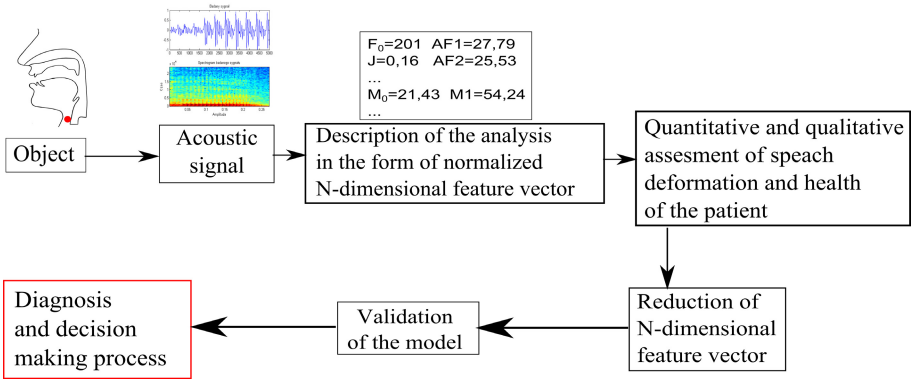

Fig. 1. Block diagram of conducted acoustic analysis

mean square value (RMS), fundamental frequency, jitter (J) and shimmer (S) coefficients, mean value of the signal, energy, average power, zero-order, first-order, second-order and third order moments, kurtosis, power factor, 1st, 2nd, 3rd formant's amplitude, 1st, 2nd, 3rd formant's frequency, maximum and minimum value of the signal and 10 mel-frequency cepstrum coefficients (MFCC). MFCC coefficients are widely used in speech recognition because they reflect well the auditory sensations by enhancing the audible frequency and are less sensitive to noise [6,7,8,9]. MFCC are designed to reflect the natural response of the auditory system to stimulation of the auditory speech sound.

A root-mean-square (RMS) value of the signal is used to estimate its loudness. In the first step of the analysis we proceeded with the normalisation where the aim was to bring the average amplitude to a target level. The normalisation was done for all signals so that their RMS value of each signal was the same.

Cepstral analysis was performed using the Fourier transformation of the sound to calculate the fundamental frequency. In order to facilitate the extraction of

spectral components before converting from time domain to the frequency domain, the signal was subjected to a windowing operation using a Hamming window. In order to improve the performance of this calculation and further purification of the spectrum the non-maximum elements were removed. The location of the maximum of the spectrum corresponds to the fundamental frequency [10]. Therefore, knowing the position of the maximum, we could determine the fundamental frequency of the analysed sample using the formula:

$$f_0 = \frac{l_s z}{wm} \tag{1}$$

where l_s – is the number of elements of the spectrum, z – is the audio sample rate,w – is the width of analysed window (number of samples), m – is the position of maximum cepstrum. The jitter coefficients (J) was calculated as an average deviation of the fundamental frequency from frequency f_0 in consecutive cycles. The jitter coefficient is presented in equation 2:

$$J = \frac{\sqrt{\frac{1}{2N-1}\sum_{i=1}^{2N-1}(f_i - f_{i-1})^2}}{\frac{1}{2N}\sum_{i=1}^{2N-1} f_i} 100\% \tag{2}$$

where f_i - the frequency in the i-th cycle. The shimmer coefficient (S) defines the variations of the fundamental tone amplitude from the average amplitude A_i in consecutive cycles, and is described by:

$$S = \frac{\sqrt{\frac{1}{2N-1}\sum_{i=1}^{2N-1}(A_i - A_{i-1})^2}}{\frac{1}{2N}\sum_{i=1}^{2N-1} A_i} 100\% \tag{3}$$

where the average amplitude is defined as:

$$\overline{A_0} = \frac{1}{N}\sum_{i=1}^{N} A_i \tag{4}$$

By integrating the square of the signal the variable E_x is formed, which is a measure of the energy carried by the signal x:

$$E_x = \int_{t_1}^{t_2} x^2(t)dt \tag{5}$$

where t_1 and t_2 are the beginning and the end of the sustained signal. The average power of the signal is defined as the average time derivative of the energy.

$$P_x = \frac{\int_{t_1}^{t_2}(x^2(t))^2 dt}{t_2 - t_1} \tag{6}$$

A large number of details, which are carried by a spectral analysis make it difficult to interpret and recognise the relevant information contained in the signal. Therefore, we determined from the frequency spectrum those features

that are useful in the analysis. Having defined signal time-frequency domain $G(t, f)$ parameters describing its shape can be determined. The first parameter describing the shape of the spectrum is the zero-order spectral moment which takes the form of:

$$M_0(t) = \sum_{i=0}^{\infty} G(t, f_i) \tag{7}$$

The zero-order moment is used to normalise the higher-order moments. The first-order moment can be interpreted as the centre of gravity of the spectrum (frequency-weighted average). This moment is used in the formulas for the calculation of higher-order central moments.

$$M_1(t) = \frac{\sum_{i=0}^{\infty} G(t, f_i) f_i}{M_0(t)} \tag{8}$$

Second-order moment is interpreted as the square of the spectrum width.

$$M_2(t) = \frac{\sum_{i=0}^{\infty} G(t, f_i)[f_i - M_1(t)]^2}{M_0(t)} \tag{9}$$

The third-order moment can be interpreted as the asymmetry of the spectrum – skewness. Standardised higher-order spectral moments are less suitable because they are correlated with each other.

$$M_3(t) = \frac{\sum_{i=0}^{\infty} G(t, f_i)[f_i - M_1(t)]^3}{M_0(t)} \tag{10}$$

Another parameter in the analysis was kurtosis, defined as flattening the spectrum measurement:

$$kurtosis = \frac{M_4(t)}{M_2(t)^2}. \tag{11}$$

Further calculation was done to get power factor characterised as the ratio of the relative power of the signal in the desired frequency f_0 wide-band $<f_d, f_g>$ to signal power across the bandwidth $<f_0, f_\infty>$, and is shown in eq. 12.

$$W_m(t) = \frac{\sum_{t=t_b}^{t_g} \sum_{f=f_{d2}}^{f_{g2}} G(t, f_i)}{\sum_{t=t_b}^{t_g} \sum_{f=f_{d1}}^{f_{g1}} G(t, f_i)} \tag{12}$$

where f_{g1}, f_{d1}- are the lower and upper frequency of the power wide-band, f_{g2}, f_{d2} – are the upper and lower frequency range of the selected frequency wide-band, t_b, t_g – are the beginning and the end of the recorded voice sample. The band selection followed the formants structure of the vowels calculated using the set of vowels recorded from the patients involved in this examination. The power coefficient was calculated only for the 1st power signal. The next computed parameters were the formants. Formant level can be defined from the envelope which can be drawn to enclose smoothly the harmonics within the

spectral maximum, i.e. the sound pressure level in dB of the envelope peak [5,9]. The frequency of the formant is measured as the frequency position of the envelope maximum (sequentially designated as $F1$, $F2$, $F3$... etc). Currently, a classical spectral analysis of voice signals methods is often supplemented with methods such as linear predictive analysis, wavelet analysis or homomorphic in the field of so-called cepstrum [6,9,11,12]. Cepstrum is determined as the inverse Fourier Transform of the logarithm of the signal spectrum giving a better picture of the structure of harmonic signal and allowing for the separation of existing noise in the transformed signal of the harmonic components in the same signal [7,13]. Many authors emphasize the importance of the cepstral factors in the diagnostic evaluation of pathological changes in the glottis [14,15]. In the present work the process of determination of the cepstral coefficients was extended to a so-called Melow filtration, which consists of an additional non-linear frequency scale signal spectrum transformation, yielding MFCC coefficients (Mel Frequency Cepstral Coefficients). Obtaining the MFCC coefficients required taking a few steps. Firstly the signal was divided into frames, the amplitude was obtained from each frame and the log was taken of these spectrums. Afterwards the results were converted into the Mel scale and the Discrete Cosine Transform (DCT) was applied. Mel filtering was applied using triangular band pass filters corresponding to the Mel scale. The number of filters was set to 10.

4 Feature Selection

All the parameters discussed above had different results. Statistically, much of this data is redundant and it is therefore useful to identify the method that can extract most significant information from the collected data. While sorting features according to their discriminant capacity it is necessary to get a stable and consistent result, which is reflected in the overall performance of the system [18]. For feature selection we used a method called Principal Component Analysis (PCA). Essentially, PCA transforms data orthonormally so that the variance of the data remains constant, but is concentrated in the lower dimensions [8]. The matrix of data being transformed consisted of all calculated parameters for every voice sample in this examination. Thus, there was a single matrix of data with all parameters. The covariance matrix of the data was created. PCA for the data set was calculated to determine the eigenvectors, which are necessary for PCA [16]. As a result, a set of principal components was obtained, with the variance order from the highest to the lowest, which means that the most important data was extracted with minimum disruption to the original data collection. Once the data had been reduced and the principal components values extracted, we took 17 out of 31 parameters that covered up to 90% variance of the initial parameters.

5 Validation

The next step in the implementation of this study was the evaluation of the operator classification quality using cross-validation [19,20]. The operator

classification represents the result of the analysis of individual principal components taken into further analysis. For validation, we used a tenfold cross validation in which we randomly selected roughly 90% of the data, which represented the set of learners and then we used the remaining 10% of cases to test for classification. We iterated this procedure 10 times. To classify the data a K-means function was used. Cross validation was done separately using recordings of vowel /a/ at normal pitch for female and male recordings with healthy and pathology state and for female /a/ vowel at high and low pitch. As a result of the cross-validation a confusion matrix was constructed. A confusion matrix is a tool used to analyse the operation of the classifier, here parameters (Table 1).

Table 1. An example of confusion matrix used in the analysis

		Results from classification		
		healthy	**pathology**	
Diagnosed	**healthy**	True Positive (TP)	False Positive (FP)	**Precision Positive predictive value** = TP/(TP+FP)
	pathology	False Negative (FN)	True Negative (TN)	**Precision Negative predictive value** = TN/(FN+TN)
		Sensitivity =TP/(TP+FN)	**Specificity** =TN/(TN+FP)	

In order to make a qualitative assessment, the accuracy, precision, sensitivity and specificity were calculated (Table 2). The accuracy of the classifier determines what percentage of parameters from the test set were correctly assigned to their respective classes and obtained at the stage of testing. The precision meant the ratio of the number of cases actually correct, which have been classified by the system as correct to all classified by the system as correct. The sensitivity relates to the test's ability to indentify positive results, whereas the specificity relates to identifying negative results.

Due to unsatisfactory results for female data set, the test was repeated for the vowel /a/ at a high and low pitch. The results are shown in Table 3.

Comparing the calculated characteristics in the confusion matrix for the vowel /a/ using PCA with different intonations, we came to the conclusion that the analysis of the vowel /a/ at a high pitch gave the most accurate indications of the patient's healthy condition and the pathological one – the accuracy of 77.5%,

Table 2. Results obtained from the confusion matrix using 31 primary parameters and PCA for healthy(H) and pathological(P) voice samples

vowel /a/ *normal pitch*	**31 param**		**PCA**		**31 param**		**PCA**	
	female				***male***			
	H	***P***	***H***	***P***	***H***	***P***	***H***	***P***
accuracy [%]	81.4		74.8		100.0		100.0	
precision [%]	74.6	88.0	80.4	69.1	100.0	100.0	100.0	100.0
sensitivity [%]	86.0	–	72.2	–	100.0	–	100.0	–
specificity [%]	–	78.0	–	78.0	–	100.0	–	100.0

Table 3. Results obtained from the confusion matrix for healthy and pathological female voices with different intonations

female ***vowel /a/***	*high pitch*		*low pitch*	
	healthy	***pathology***	***healthy***	***pathology***
accuracy [%]	77.5		73.3	
precision [%]	83.7	71.3	79.0	67.5
sensitivity [%]	73.4	–	72.1	–
specificity [%]	–	79.9	–	78.6

precision of almost 84% for the healthy females and 71 pointing the pathology ones. The same dependence showed the sensitivity and specificity that were around 1% higher for a high pitch than normal one. Results from the vowel /a/ with a low pitch showed slightly lower results than a normal pitch, expect for the specificity that was slightly higher, but at the same time it was still lower than the one associated with a high pitch. There was no need to repeat the test for male recordings due to 100% correct classification Results based on 31 original parameters for female recordings show higher accuracy and sensitivity than PCA. While analysing PCA the precision for healthy women was higher than the one for all original parameters, pathology precision at the same time was still lower. The specificity for both methods for normal pitch was the same.

6 Conclusion

The open and free database available online has been used in the context of pathology detection. The substantial amount and various types of recordings included in this database made it possible to conduct different and interesting tests. Based on the conducted studies it was shown that the use of the calculated parameters and their subsequent reduction significantly differentiated between acoustic characteristics of the pathological speech and those of the healthy one. An integrated acoustical analysis of deformed pathological speech was discussed in this paper. The analysis employed among groups of patients showed that

speech pathology caused by various laryngeal diseases can be computed using acoustical methods. In order to expedite and simplify the calculations the Principal Components Analysis was conducted and led to obtaining 17 out of 31 parameters. The final analysis included just these factors. Analysis of the 17 parameters analysed for women showed lower accuracy and sensitivity than 31 original parameters at normal pitch when determining whether the patient was healthy or ill. The precision of pathology detection was higher when PCA was used. The specificity did not change.

According to the statistical accuracy of the pathological voice diagnosis obtained satisfactory results showing 100% compatibility classification obtained for the male voices analysing original vector of parameters and PCA, whereas for the female ones it proved more complicated.

The best results were achieved with high intonation for the female recordings giving an approximately 3% higher result of accuracy than normal intonation and almost 4% higher than low intonation. Other calculations like precision, specificity and sensitivity showed an upward trend. The reason might be that for the same vowel spoken at different pitches the relationship between the second harmonic and the first formant can change, causing the amplitude of this harmonic to be artificially amplified or attenuated [21]. The results show that the techniques discussed here could be used for detecting pathological voices.

Acknowledgement. This work was funded by the Ministry of Science and Higher Education in Poland under the Diamond Grant program, decision number 0136/DIA/2013/42 (AGH 68.68.120.364).

References

1. Deliyski, D.D.: Multi-dimensional acoustic analysis of spasmodic dysphonia. In: Proc. in the ASHA Convention, Atlanta (1991)
2. Gogh, C.D.L., Festen, J.M., Verdonck-de Leeuw, I.M., Parker, A.J.: Acoustical analysis of tracheoesophageal voice. Speech Communication 47, 160–168 (2005) ISSN 0167-6393
3. Martinez, D., Lleida, E., Ortega, A., Miguel, A., Villalba, J.: Voice Pathology Detection on the Saarbruecken Voice Database with Calibration and Fusion of Scores Using MultiFocal Toolkit. Advances in Speech and Language Technologies for Iberian Languages Commu-nications in Computer and Information Science 328, 99–109 (2012)
4. Barry, W.J., Putzer, M.: Saarbruecken Voice Database. Institute of Phonetics, University of Saarland, `http://www.stimmdatenbank.coli.uni-saarland.de/`
5. Fant, G.: Acoustic Theory of Speech Production With Calculations based on X-Ray Studies of Russian Articulations, Mouton, The Hague (1970) ISBN: 9027916004
6. Maciel, C.D., Pereira, J.: Identifying healthy and pathologically affected voice signals. IEEE Signal Processing Magazine 27(1), 120–123 (2010)
7. Arroyave, J.R.O.A., Bonilla, J.F.V., Trejos, E.D.: Acoustic Analysis and Non Linear Dynamics Applied to Voice Pathology Detection: A Review. Recent Patents on Signal Processing (2012)

8. Loughran, R., Walker, J., O'Neill, M., O'Farrell, M.: The Use of Mel-frequency Cepstral Coefficients in Musical Instrument Identification. Routes/Roots, Michigan (2008)
9. Engel, Z.W., Klaczynski, M., Wszolek, W.: A Vibroacoustic Model of Selected Human Larynx Diseases. International Journal of Occupational Safety and Ergonomics (JOSE) 13(4), 367–379 (2007)
10. Wuyts, F.L., De Bodt, M.S., Molenberghs, G., Remacle, M., Heylen, L., Millet, B., Van Lierde, K., Jan, R., Van de Heyning, P.H.: The Dysphonia Severity Index: An Objective Measure of Vocal Quality Based on a Multiparameter Approach. Journal of Speech, Language and Hearing Research 43, 796–809 (2000) ISSN 1092-4388
11. Osowski, S.: Sieci neuronowe w ujęciu algorytmicznym, WNT, Warszawa (1996)
12. Tadeusiewicz, R., Izworski, A., Wszolek, W.: Pathological speech evaluation using the artificial intelligence methods. Med. Biol. Eng. Comput. (2007)
13. Noll, A.: Short-term spectrum and 'cepstrum' techniques for vocal pitch detection. J. Acoust. Soc. Am. 41, 293–300 (1964)
14. Awan, S.N., Giovinco, A., Owens, J.: Effects of vocal intensity and vowel type on cepstral analysis of voice. In: Presented at the 39th Annual Symposium: Care of the Professional Voice, Philadelphia (2010)
15. Mehta, D.D., Deliyski, D.D., Zeitels, S.M., Quatieri, T.F., Hillman, E.R.: Voice Production Mechanisms Following Phonosurgical Treatment of Early Glottic Cancer. Ann. Otol. Rhinol. Laryngol. 119(1), 1–9 (2010)
16. Bishop, C.M.: Pattern Recognition and Machine Learning, pp. 559–599. Springer Science, Singapore (2006)
17. Methods based on Principal Components Analysis and the concept of Eigenface, Metody oparte na Analizie Glownych Skladowych i koncepcji Eigenface, http://icsolutions.pl/
18. Orozco-Arroyave, J.R., Murillo-Rendon, S., Alvares-Meza, A.M., Arias-Londono, J.D., Delgado-Trejos, E., Vargas-Bonilla, J.F., Castellanos-Domingues, C.G.: Automatic Selection of Acoustic and Non-linear Dynamic Features in Voice Signals for Hypernasality Detection. In: Interspeech, pp. 529–532 (2011)
19. Refaeilzadeh, P., Tang, L., Liu, H.: Cross Validation, Encyclopedia of Database Systems (EDBS), p. 6. Arizona State University, Springer (2009)
20. Delgado-Trejos, E., Sepulveda-Sepulveda, F.A., Castellanos-Domnguez, G.: Robustness Improvement of Hypernasal Speech Detection by Acoustic Analysis and the Rademacher Complexity Model. In: Advances in Biomed. Research, pp. 159–162
21. Epstein, M.A., Payri, B.G.: The effects of vowel quality and pitch on spectral and glottal flow measurements of the voice source, Lecture, University of California, Los Angeles.