Anna SAMBORSKA-OWCZAREK
WEST POMERANIAN UNIVERSITY OF TECHNOLOGY,
ul. Żołnierska 49, Szczecin

Diagnostic significance of phase spectrum in acoustic analysis of pathological voice

Ph.D. eng. Anna SAMBORSKA-OWCZAREK

The author works at Department of Computer Science and Information Technology at the West Pomeranian University of Technology in Szczecin, Poland. Her area of research is voice and speech signal processing and analysis for medical purposes.

e-mail: asamborska@wi.zut.edu.pl

Abstract

The paper regards the possibility of using new numerical features extracted from the phase spectrum of a speech signal for voice quality estimation in acoustic analysis for medical purposes. This novel approach does not require detection or estimation of the fundamental frequency and works on all types of speech signal: euphonic, dysphonic and aphonic as well. The experiment results presented in the paper are very promising: the developed F0-independant voice features are strongly correlated with two voice quality indicators: grade of hoarseness G (r>0.8) and roughness R (r>0.75) from GIRBAS scale, and exceed the standard voice parameters: jitter and shimmer.

Keywords: acoustic analysis, voice signal, speech processing, fundamental frequency, F0, phase spectrum, features extraction, GIRBAS.

Diagnostyczne znaczenie widma fazowego w analizie akustycznej głosu patologicznego

Streszczenie

Artykuł dotyczy możliwości ekstrakcji cech numerycznych z widma fazowego sygnału mowy w celu wykorzystania w analizie akustycznej na potrzeby medyczne. Podejście to umożliwia uzależnienie analizy akustycznej od zawodnych metod wykrywania/wyznaczania częstotliwości podstawowej (tonu krtaniowego) i dzięki temu przeznaczone jest do badania wszystkich typów sygnału mowy (również afonicznych). Wyniki eksperymentu są bardzo obiecujące – proponowane cechy Ph1 i Ph2 są silnie skorelowane z dwoma kategoriami percepcyjnymi: stopniem chrypki (r>0.8) oraz szorstkością głosu (r>0.75) ze skali GIRBAS, wykazując silniejsze znaczenie diagnostyczne niż znane i stosowane od dawna wskaźniki *jitter* i *shimmer*. Proponowane podejście oprócz skuteczności charakteryzuje się szeregiem dodatkowych korzyści: algorytm metody z powodu niskiej złożoności jest szybki i niekosztowny, interpretacja matematyczna jest prosta i jednoznaczna oraz spójna z obserwowanym obrazem widma fazowego głosu. Ponadto uniezależnienie od detekcji częstotliwości podstawowej sprawia, że algorytm jest deterministyczny oraz efektywny dla każdego typu sygnału mowy.

Słowa kluczowe: analiza akustyczna, sygnał mowy, przetwarzanie mowy, częstotliwość podstawowa, widmo fazowe, ekstrakcja cech, GIRBAS.

1. Introduction

There are three known approaches to voice disorders diagnostics in laryngology/ voice pathology. The first is an instrumental approach that includes usage of highly specialized equipment: laryngoscope, endoscope, etc. This is an expensive and in most cases fatiguing method but gives immediate and accurate inspection of the voice tract organ efficiency. The second method is acoustic analysis – the cheapest and the quickest way to estimate the voice quality. It is also very comfortable for a patient. Unfortunately, it lacks objectivity and depends completely on voice specialist/therapist experience, memory and hearing acuity [1].

The third approach to voice diagnostics is automated acoustic analysis of a recorded voice sample by a computer system. The algorithm processes the speech signal and extracts a number of numerical parameters that correspond to the level of pathological symptoms in speech. The biggest advantage of this method it its objectivity and independency of human factors. Moreover, it is cheaper than instrumental diagnostics, completely non-invasive and supports long term control of the treatment process. Hence, it seems to be perfect for screening test or for self-control performed frequently by voice professionals.

Computer programs for voice analysis have been widely used since early 90's but their origins reach as early as the first decade of the 20th century when first observation of acoustic wave of speech was documented and the relationship between the wave shape and voice quality was noticed and explained [2]. Since then there have been several dozen numerical voice features and extraction algorithms developed (the most significant are *jitter* and *shimmer*). Almost all of them base on the fundamental frequency of a voice sample. *The fundamental frequency* (also called F0, laryngeal tone or pitch) corresponds to the frequency of vocal folds vibrations and its estimation is now an indispensable step in automated acoustic analysis [3].

The obligation of F0 detection before voice features are calculated is a serious limitation of the automated voice analysis. First, only the voiced signals can be analyzed and the aphonic ones are rejected after recognized as unvoiced. Second, despite of many available F0 estimation methods this is still a complex and not robust process. Common errors include F0 halving or doubling, significant inaccuracy or even invalid voiced/unvoiced decision. Hence, there is still a considerable risk that the feature extraction algorithms working on incorrect F0 data will result in inaccurate voice parameters. The earlier research performed on benchmark database *Disordered Voice Database* [4] proved that the leading voice analysis systems (Multidimensional Voice Program [5, 6], Voice Analysis and Screening System [7, 8], Voxmetria [9]) lacked their reliability because of erroneous or inaccurate F0 detection/estimation with the following error rates [10]:

MDVP: 9,0 %,
VASS: 6,0 %,
Voxmetria: 16,4%.

The only approach that intentionally considered the problem of F0 estimation effect on final results is *Hoarseness Diagram* (germ. *Heiserkeits-Diagramm*) developed at Göttingen University [11, 12]. The authors postulate that all voice signals can be robustly analyzed with their method because it does not require F0 estimation. In fact, the basic F0 estimation algorithm is performed in Hoarseness Diagram to calculate 3 of 4 parameters used in the analysis but with disregard of formal F0 detection. In the result, all voice signals are analyzed but the process reliability is quite poor and the obtained voice features are not interpretable in terms of phonation conditions without contradictions [10]. The only actually F0-independent element of this approach is the voice parameter *GNE Glottal-to-Noise Excitement Ratio* that is very competitive to standard HNR measures [13] [14]. Unfortunately, the GNE algorithm is also quite complex (including FFT, inverse filtering, band filtering and Hilbert transform) and, as a result, computationally expensive.

The state of art in acoustic analysis indicates the necessity of developing new, possibly effortless voice features extraction algorithms that are independent of F0 estimation and work on voiced and unvoiced signals as well.

2. Phase spectrum in acoustic analysis

In the voice diagnostics systems the feature extraction is processed after F0 estimation and usually concerns F0/ amplitude/ energy perturbation and signal-to-noise ratio measures. The parameters are calculated in time or frequency domain for separate or overlapping short windowed frames. A vector of these partial results is then used for calculation of long term features such as mean, maximum or differential of the vector. For example, maximal signal amplitude within a single pitch period is a short term feature and its statistic *Relative Amplitude Perturbation* (RAP [15]) is a long term parameter, used for voice quality evaluation.

The proposed in the paper original F0-independent voice features result from observation of normal and disordered voices **phase contour**. The extensive literature survey conducted by the author [10] proved that the phase spectrum had never been used for voice parameters extraction for medical purposes[1] and there was found only one statement concerning this problem [16]:

Because phase information is of little importance or interest in speech analysis it is rarely calculated and for everyday clinical purposes may be ignored.

Rejection of the phase spectrum as useless and complete lack of information in literature on its potential relation to the voice quality was a motivation for the author to perform the experiment presented in the next section that eventually proved usefulness of the phase contour in voice analysis.

In Figs. 1- 3. there are shown the sample phase spectra of normal (Fig. 1), moderately disordered (Fig. 2) and severely pathological, aphonic (Fig. 3.) speech from *Disordered Voice Database* (stable segment of sustained vowel /a/) and the corresponding radial charts depicting the unwrapped phase (after 2π correction). On the first two charts (Figs. 1 and 2) frequent phase jumps are present, with the first one close to the fundamental frequency. On the third chart (Fig. 3) the phase changes rapidly and seems to be random. According to [19], the negative phase jumps indicate formant frequencies and that is correct for the first normal sample (Fig. 1) only, because the other two are partially (Fig. 2) or completely (Fig. 3) chaotic.

Hence, the unwrapped phase perturbation level is probably an indicator of voice quality degradation and one can assume that the sooner phase spectrum appears random, the more disordered the voice sample is. What is really significant, the relationship between the phase spectrum contour and the voice quality occurs only for rectangular windowing that is rarely used in speech processing. In case of using Hamming, Blackmann or other common function, the effect is not present because of signal and window convolution in frequency domain and that is probably why the phenomenon has not been noticed and used in voice diagnostics so far.

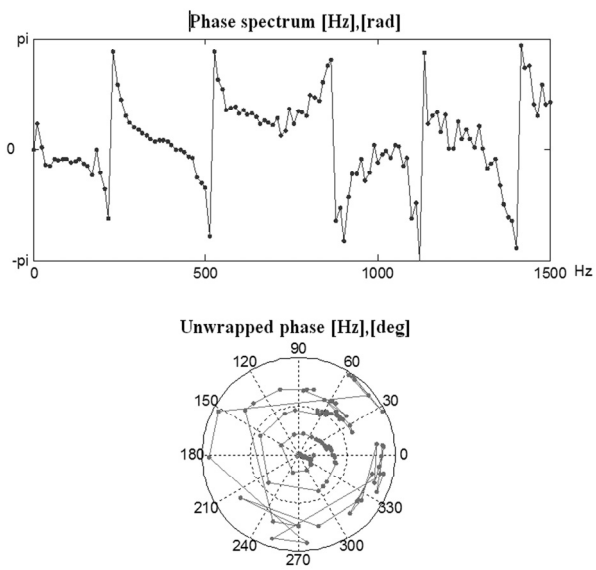

Fig. 1. Example of phase spectrum (rectangular windowing) and unwrapped phase radial chart for a normal voice
Rys. 1. Przykład widma fazowego (okienkowanie prostokątne) oraz wykres radialny skorygowanej fazy dla głosu normalnego

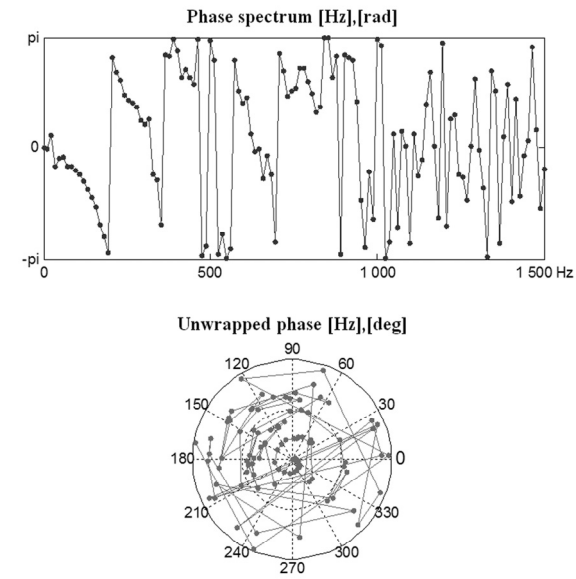

Fig. 2. Example of phase spectrum (rectangular windowing) and unwrapped phase radial chart for a disordered voice
Rys. 2. Przykład widma fazowego (okienkowanie prostokątne) oraz wykres radialny skorygowanej fazy dla głosu zaburzonego

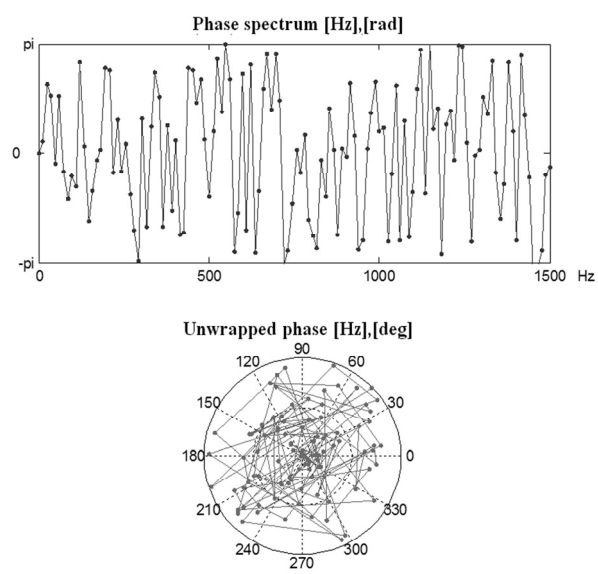

Fig. 3. Example of phase spectrum (rectangular windowing) and unwrapped phase radial chart for a pathological voice (aphonic)
Rys. 3. Przykład widma fazowego (okienkowanie prostokątne) oraz wykres radialny skorygowanej fazy dla głosu patologicznego (afonicznego)

[1] In other applications of voice analysis, phase information has been used in formant detection [17] and automatic speech recognition [18]

3. Experiment

The objective of the presented experiment was to obtain an accurate numerical measure of the phase contour perturbation level and estimate its correlation with voice quality degradation. To make the results comparable, the research was performed on *Disordered Voice Database* from Massachusetts Eye and Ear Infirmary. After duplicates removal, the database contains 708 unique digitalized voice samples of a sustained vowel /a/ collected from about 650 patients of the clinic. The first step was to unify all the recordings: trim them to 1 sec and downsample to 25 kHz. After that all samples were evaluated by three independent voice pathologists from The International Center of Hearing and Speech in Kajetany (www.ifps.org.pl) using perceptual scale GIRBAS [20]. This scale consists of 6 voice quality categories that were individually rated 0-3 points: from 0 (normal) to 3 (severely pathological):

G – Grade of hoarseness,
I – Instability,
R – Roughness,
B – Breathiness,
A – Asthenia,
S – Strain.

The inter-rater agreement was performed using *Intraclass Correlation Coefficient* ICC(2,k) [21, 22] and then pairs of ratings of the highest reliability (G, R, B, S) were selected and summed up to produce four vectors of voice quality data (Tab. 1).

Tab. 1. The results of raters' agreement analysis ICC(2,k) for each pair of raters. The marked cells indicate the acceptable reliability level (4 of 6 categories)
Tab. 1. Wyniki analizy zgodności ICC(2,k) dla każdej pary oceniających. Zaznaczono wyniki akceptowalnej zgodności (4 z 6 kategorii)

ICC(2,k) / Category	*Rater 1 Rater 2*	*Rater 1 Rater 3*	*Rater 2 Rater 3*
G – *grade*	0,889	0,089	0,093
I – *instability*	0,007	0,024	0,066
R – *roughness*	0,142	0,679	0,109
B – *breathiness*	0,791	0,128	0,117
A – *asthenia*	0,365	0,001	0,002
S – *strain*	0,058	0,064	0,811

The parameters were expected to be F0-independent, computationally low-cost and easily interpretable. In order to select the optimal frame length and the phase frequency range, 32 versions were investigated and then compared to voice quality ratings.

The proposed short term voice features measured a sum of the absolute velocity (1) or acceleration (2) of the unwrapped phase contour changes for a signal frame in a specified frequency range f_r:

$$Ph1\left(l_{seg}, f_r\right) = \sum_k \left|\varphi'(k)\right| \tag{1}$$

$$Ph2\left(l_{seg}, f_r\right) = \sum_k \left|\varphi''(k)\right| \tag{2}$$

where $\varphi(k)$ is the unwrapped phase contour vector in the frequency range f_r calculated for l_{seg} long rectangularly windowed signal frame, where l_{seg}= {1024, 2048, 4096} and f_r takes the value of the following:

A: 60 – 1500 Hz;
B: 60 – 4500 Hz;
C: 60 – 6000 Hz;
D: 1500 – 4500 Hz;
E: 1500 – 6000 Hz;
F: 3000 – 6000 Hz.

*Ph*1 and *Ph*2 were calculated for a series of successive l_{seg} long signal frames with 50% overlapping separately for each frequency range f_r. The long term versions of *Ph*1 and *Ph*2 were defined as an **average value for all frames**.

4. Results

The diagnostic significance level of long term *Ph*1 and *Ph*2 was calculated using the Spearman rank correlation coefficient, because of interval scale of the voice ratings data. The results are presented in Tab. 2. Without any doubt, the relationship between the phase contour perturbation and the perceptually perceived voice quality is significant – for selected options as high as 0.8. The proposed features are correlated strongly with the grade of hoarseness and voice roughness, moderately with the breathiness and weakly with the voice strain. There was hardly any noticeable difference in *Ph*1 and *Ph*2 performance or between the frame lengths but there were significant differences for frequency ranges in favour to the lowest range 60 – 1500 Hz (the range with the fundamental frequency and its harmonics).

Tab. 2. The results of Spearman rank correlation between Ph1/Ph2 (window length 1024-4096, frequency range A-F) and voice quality categories G/R/B/S from GIRBAS scale. The cells marked with * indicate the highest correlation coefficient: * (r>0.75), **(r>0.8)
Tab. 2. Wyniki korelacji rangowej Spearmana miedzy Ph1/Ph2 (długość okna 1024-4096, zakres częstotliwości A-F) oraz kategoriami percepcyjnymi G/R/B/S ze skali GIRBAS. Znaczniki * wskazują na najwyższe wskaźniki korelacji: * (r>0.75), **(r>0.8)

		mean Ph1				*mean Ph2*			
		G	**R**	**B**	**S**	**G**	**R**	**B**	**S**
1024	*A*	*0,798	*0,760	0,694	0,352	**0,819	*0,763	0,724	0,362
	B	*0,778	0,720	0,692	0,355	*0,797	0,733	0,717	0,348
	C	*0,754	0,696	0,675	0,344	*0,775	0,711	0,699	0,339
	D	0,696	0,624	0,635	0,311	0,701	0,629	0,645	0,286
	E	0,669	0,604	0,612	0,298	0,676	0,610	0,624	0,277
	F	0,564	0,500	0,519	0,245	0,570	0,503	0,526	0,226
2048	*A*	**0,819	*0,762	0,718	0,371	**0,822	*0,762	0,731	0,370
	B	*0,791	0,722	0,698	0,378	*0,798	0,733	0,715	0,363
	C	*0,766	0,697	0,682	0,361	*0,775	0,711	0,700	0,346
	D	0,675	0,597	0,605	0,315	0,667	0,603	0,611	0,277
	E	0,653	0,579	0,590	0,297	0,651	0,585	0,600	0,263
	F	0,523	0,447	0,472	0,221	0,531	0,465	0,487	0,192
4096	*A*	**0,807	*0,771	0,699	0,374	*0,796	*0,767	0,699	0,356
	B	*0,771	0,725	0,672	0,379	*0,763	0,725	0,677	0,341
	C	0,747	0,702	0,663	0,369	0,740	0,704	0,667	0,326
	D	0,601	0,555	0,538	0,281	0,594	0,547	0,548	0,214
	E	0,596	0,544	0,545	0,281	0,587	0,543	0,552	0,212
	F	0,487	0,427	0,448	0,239	0,489	0,447	0,459	0,167

In order to compare the diagnostic significance of the new voice features and the standard parameters, there were calculated *jitter*, *shimmer* and GNE for all the recordings from *Disordered Voice Database*. GNE was calculated with *Hoarseness Diagram* software [11], while *jitter* and *shimmer* were based on the short term perturbation factor formula [23]:

$$jitter = \frac{100\%}{N-1}\sum_{n=1}^{N-1}\left|\frac{u(n)-u(n-1)}{u(n)}\right| \quad (3)$$

where $u(n)$ is a successive pitch period vector, and n=0…N-1;

$$shimmer = \frac{100\%}{N-1}\sum_{n=1}^{N-1}\left|\frac{v(n)-v(n-1)}{v(n)}\right| \quad (4)$$

where $v(n)$ is a successive amplitude maximum vector, and n=0… N-1.

The comparison results are given in Tab. 3. Clearly, the proposed phase based features have a big advantage over the standard parameters for all categories but the voice strain.

Tab. 3. The results of Spearman rank correlation between jitter/ shimmer/ GNE/ Ph2(2048, A) and voice quality categories G/R/B/S
Tab. 3. Wyniki korelacji rangowej Spearmana pomiędzy jitter/ shimmer/ GNE/ Ph2(2048, A) oraz kategoriami percepcyjnymi G/R/B/S

	G	R	B	S
jitter	0,632	0,576	0,473	0,407
shimmer	0,731	0,680	0,611	0,345
GNE	0,601	0,452	0,641	0,182
mean Ph2(2048,A)	0,822	0,762	0,731	0,370

5. Conclusions

In the paper probably the first attempt of voice quality estimation by the phase spectrum is presented. The experiment results are very promising and encourage further research for medical voice analysis purposes. There is one potential weakness of the introduced parameter: it has not been interpreted in terms of phonation conditions yet but, nevertheless, it can be considered as a simple and universal voice quality parameter because:

1. The extraction algorithm is effortless (low complexity, low cost).
2. It is easily mathematically interpretable: it is a measure of phase contour perturbation based on its differential.
3. There is no need for prior F0 detection or estimation, hence:
   – it is deterministic and fully reliable (unlike *jitter* or *shimmer*), independent of any pitch detection algorithms;
   – all intentionally voiced signals (euphonic, dysphonic or aphonic) can be analyzed and treatment progress observation is possible, even when starting with aphonic voice.
4. The parameter is significantly correlated with perceptually perceived voice quality degradation (generally higher than best standard features).
5. Ph1/Ph2 is derived directly from the phase contour and its visual observation could support voice quality estimation as well.

The potential of the phase spectrum in voice quality analysis has not been fully discovered yet. There are probably more significant information in the phase contour that can be represented objectively (for example specific shape features) to be investigated.

6. References

[1] Kent R.D., Ball, M.J: Voice quality measurement. San Diego, Singular, 2000.

[2] Buder E.H.: Acoustic Analysis of Voice Quality: A Tabulation of Algorithms 1902-1990. [in Ball M.J., Kent R.D, Voice Quality Measurement]. San Diego, Singular, 2000.

[3] Titze I.: Workshop on Acoustic Voice Analysis: Summary Statement. Denver, 1995.

[4] KayPENTAX. Disordered Voice Database Model 4337 - operations manual. Lincoln Park, NJ, KayPENTAX, 2002.

[5] KayPENTAX. Multidimensional Voice Program, model 5105 - software instruction manual. Lincoln Park NY, KayPENTAX, 2002.

[6] Deliyski D. D.: Acoustic Model and Evaluation of Pathological Voice Production. Proceeding of EUROSPEECH'93. Berlin, 1993.

[7] Mitev P.: System for acoustic analysis of the pathological voices. PhD thesis. Sofia, Center on Biomedical Engineering, Bulgarian Academy of Sciences, 2000.

[8] Hadjitodorov S., Mitev P. A: computer system for acoustic analysis of pathological voices and laryngeal diseases screening. Medical Engineering and Physics. 2002, Vol 24, 6.

[9] CTS Informática. VoxMetria - Voice Analysis and Vocal Quality. Voice, Speech and Language Software. [online] http://www.ctsinformatica.com.br/english/#voxMetria.html.

[10] Samborska-Owczarek A.: Heuristic classification methods for vocal tract efficiency diagnostics, Phd Thesis (polish title: Metody heurystycznej klasyfikacji we wspomaganiu diagnozowania wydolności traktu głosowego). Szczecin, West Pomeranien University of Technology, 2009.

[11] Fröhlich M., et al.: Acoustic voice quality description: Case studies for different regions of the hoarseness diagram. Advances in Quantitative Laryngoscopy, 2nd 'Round Table'. Erlangen, 1997.

[12] Michaelis D.: Das Göttinger Heiserkeits-Diagramm - Entwicklung und Prüfung eines akustischen Verfahrens zur objektiven Stimmgütebeurteilung pathologischer Stimmen. PhD thesis. Göttingen, Georg-August-Universitat zu Göttingen, 1999.

[13] Michaelis D., Gramss T., Strube H. W:. Glottal to noise excitation ratio - a new measure for describing pathological voices. Acta acustica. 1997, Vol 83.

[14] Fröhlich M., Michaelis D. Strube H. W.: Acoustic "breathiness measures" in the description of pathologic voices. In Proceedings ICASSP'98. 1998.

[15] Kiritani S.: High-speed digital image recording for observing vocal fold vibration. [in R. D. Kent, M. J. Ball, Voice Quality Measurement]. San Diego, Singular, 2000.

[16] Baken R.J., Orlikoff R.F: Clinical Measurement of Speech and Voice. Cengage Learning, 2000.

[17] Bozkurt B., at al.: Improved Differential Phase Spectrum Processing For Formant Tracking. Proc. Icslp. Jeju Island, 2004.

[18] Paliwal K.K.: Usefulness of Phase in Speech Processing. Proc. IPSJ Spoken Language Processing Workshop. Gifu, Japan, Feb. 2003.

[19] O'Shaughnessy D.: Speech Communications: Human and Machine. IEEE Press, 2000.

[20] Webb A.L., at al.: The reliability of three perceptual evaluation scales for dysphonia. European Archives of Oto-Rhino-Laryngology. 2004.

[21] von Eye A.: Mun Young, E., Analyzing Rater Agreement: Manifest Variable Methods. Routledge, 2005.

[22] Shrout P.E., Fleiss J.L.: Intraclass correlations: uses in assessing rater reliability. Psychological Bulletin. 1979, Vol 86, 2.

[23] Kasuya H., Endo Y., Saliu S.: Novel acoustic measurements of jitter and shimmer characteristics from pathological voice. In Eurospeech'93. 1993, Vol 3.

otrzymano / received: 21.09.2010
przyjęto do druku / accepted: 01.11.2010 *artykuł recenzowany*