Annales UMCS Informatica AI 6 (2007) 117-126

Annales UMCS
Informatica
Lublin-Polonia
Sectio AI

http://www.annales.umcs.lublin.pl/

Cepstral analysis of speech signals in the process of automatic pathological voice assessment

Anna Samborska-Owczarek*

Faculty of Computer Science, Szczecin University of Technology,
Żołnierska 49, 71-210 Szczecin, Poland

Abstract

The paper describes the problem of cepstral speech analysis in the process of automated voice disorder probability estimation. The author proposes to derive two of the most diagnostically significant voice features: quality of harmonic structure and degree of subharmonic from cepstrum of speech signal. Traditionally, these attributes are estimated auricularly or by spectrum (or spectrogram) observation, hence this analysis often lacks accuracy and objectivity. The introduced parameters were calculated for the recordings from Disordered Voice Database (Kay, model 4337 version 2.7.0) which consists of 710 voice samples (657 pathological, 53 healthy) recorded in the laboratory environment and described with diagnosis and a number of additional attributes (such as age, sex, nationality).

The proposed cepstral voice features were compared to similar voice parameters derived from Multidimensional Voice Program (Kay, model 5105 version 2.7.0) in respect to their diagnostic significance and presented graphically. The results show that cepstral features are more correlated with decision and better discriminate clusters of healthy and disordered voices. Additionally, both parameters are obtained by single cepstral transform and do not require to perform F0 tracking earlier as it is derived simultaneously.

1. Introduction

Nowadays, computer-aided medical diagnostic techniques and decision support systems for medical specialists are becoming very popular. Designers of such systems compile their knowledge of medicine with information science: pattern recognition, machine learning and data mining etc.

Medical data is particularly difficult to process and analyze because of its specific features:

– often insufficient number of cases in classes; healthy class tends to be more occupied; some disorders are less common than others; the symptoms may be identical for different disorders,

* *E-mail address*: asamborska@wi.ps.pl

– asymmetrical error cost – false acceptance (disordered person classified as healthy one) is more risky than false rejection (healthy recognized as disordered).

An example of such difficult data is a speech sample for pathological voice symptoms assessment. Traditionally voice quality is evaluated by voice specialist (phoniatrist, laryngologist etc.) auricularly and classified to one of three categories [1]:

1. *euphonia* – no symptoms of vocal fatigue, vibrant and clear voice,
2. *dysphonia* – symptoms such as: hoarseness, voice breaks, vocal fatigue,
3. *aphonia* – total inability of vocal folds to vibrate, breathing, whispering voice.

When an equipment and software for recording and analyzing speech signal became easy to access, voice specialists were able to observe voice in the form of oscillogram or spectrogram. Hence both spectrum and spectrogram are still frequently used to visually estimate features that are diagnostically significant but difficult to measure by hearing, such as [1]:

a) fundamental frequency F0,
b) formant frequencies,
c) quality of harmonic structure,
d) degree of F0 period doubling (subharmonic).

The main problem is that spectrum operates in the frequency domain and thus its resolution is usually insufficient to express the features above with proper accuracy. Additionally, the phenomenon known as ‘spectrum leakage’ makes the evaluation less robust.

Since automated acoustic analysis software appeared, the medicine specialists have been able to attain more accurate and objective voice quality estimation. Unfortunately, there are particular difficulties of the process of both gathering and describing voice samples. First of all, they ought to be recorded with high quality equipment in constant and undisturbed environment and the spoken phrase (most often sustained vowel /ah/) must be the same for all the recordings. The next problem is the large amount of data, which should be processed, before and during automatic classification.

Nowadays, a standard for acoustic analysis is Multidimensional Voice Program (Kay, model 5105 version 2.7.0). MDVP extracts 33 voice features (related to voice disorder risk) from voice sample, estimates voice quality by using predefined thresholds and presents the results both numerically and graphically. MDVP extracts, among others, parameters related to quality of harmonic structure (Noise-to-Harmonic Ratio) and degree of F0 period doubling (Degree of Subharmonic).

In the paper the author proposes voice features alternative to NHR and DSH, derived from cepstrum of speech, that are explicitly competitive – more

correlated with decision attribute (diagnosis: healthy or disordered) and less computationally complex.

2. Cepstrum of speech signal

Cepstrum (an anagram of the word *spectrum*) is a frequency analysis of log-magnitude spectrum and is obtained via the formula [2]:

$$C(t) = \mathrm{IFFT}\left[\log\left|\mathrm{FFT}\left[x(t)\right]\right|\right], \tag{1}$$

where $x(t)$ – analyzed windowed frame.

To obtain cepstrum of speech, first one has to transfer the windowed signal frame to the frequency domain by the FFT transform, and then transfer it back (as the log-magnitude spectrum) to *quefrency* domain (an anagram of the word 'frequency'). The resulted signal is known as a *real cepstrum* (commonly called cepstrum) in opposite to its complex version[1]. *Quefrency* is measured in seconds which does not indicate time but periods of frequencies – the peaks that appear within cepstrum reveal periods of frequencies that have harmonics in the spectrum. Analogically to *quefrency*, magnitude turns into *gamnitude* and harmonic into *rahmonic*. Similarly to spectrum, cepstrum is symmetric and there is an information redundancy.

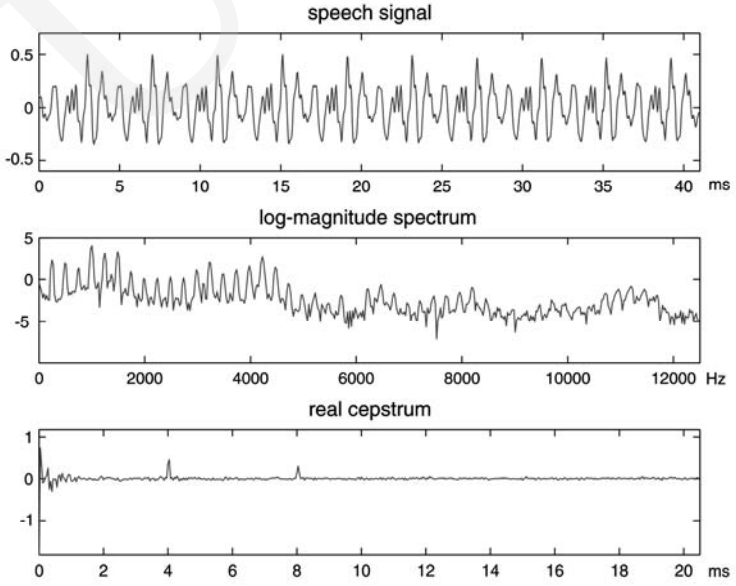

Fig. 1. A segment of speech signal, log-magnitude spectrum and cepstrum

[1]*Complex cepstrum* is derived from the same formula but takes complex logarithm.

The most common application of cepstral analysis is pitch tracking (long term F0 estimation) that is usually very effective (except strongly disordered voices). The procedure consists in cepstrum evaluating for a sequence of windowed signal frames and searching for a dominant peak in the range:

$$\left\langle \frac{f_s}{F0_{MAX}}, \frac{f_s}{F0_{MIN}} \right\rangle, \tag{2}$$

where $F0_{min}$ and $F0_{max}$ are the minimal and the maximal acceptable fundamental frequencies of voiced speech. The dominant peak theoretically indicates the period of fundamental frequency, but practically in the case of pathological voices it can be located at the double F0 period (if it is higher *gamnitude* than the F0 period peak). Cepstrum is especially useful in the analysis of speech signals, because the fundamental frequency (laryngeal tone, pitch) is present with several spectral harmonics that are an effect of passing the voice through resonance cavities (such as nasal and throat cavities).
Another common application of cepstrum is formant frequencies estimation (b) using *homomorphic speech analysis* [3]. Formant envelope is obtained with a low pass spectrum filtering by clearing low *quefrencies* in cepstrum.

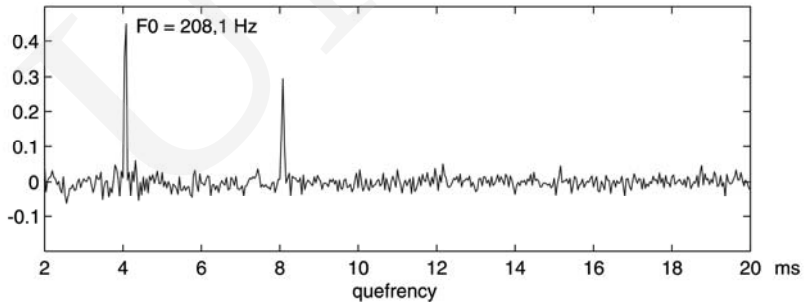

Fig. 2. F0 frequency estimation by cepstrum of speech signal

3. Evaluation of NHR and DSH parameters (MDVP)

The considered voice features quality of harmonic structure and degree of F0 period doubling are traditionally estimated visually from the spectrum or the spectrogram chart. The structure of harmonic content in spectrum indicates F0 quality and its contribution to the whole signal. High degree of F0 period doubling is definitely a pathological symptom that manifests itself by the presence of subharmonic frequencies in the spectrum. It is related to the ability of vocal folds to work steady and is typical of diplophonic voices and those with vocal fry [4].

Multidimensional Voice Program (Kay, model 5105 version 2.7.0) extracts 33 voice parameters with 2 of them corresponding to the voice features considered before [5]:

1. NHR – Noise-to-Harmonic Ratio (category: 'noise and tremor evaluation measurements') is an average ratio of the inharmonic spectral energy to the harmonic spectral energy in the frequency range 70-4200 Hz. It is computed using the pitch-synchronous frequency-domain method. The spectrum is separated into the harmonic and inharmonic components synchronously with the average fundamental frequency of the current block, and the ratio computed. When all blocks are processed, the average value is reported as the NHR.
2. DSH – Degree of Subharmonic Component (category: 'voice breaks and sub-harmonic measurements') is an estimated relative evaluation of the subharmonic to F0 component in the voice sample. It is computed as a ratio of the number of autocorrelation segments where the pitch was found to be subharmonic of the real pitch to the total number of autocorrelation segments.

The parameters were calculated for 684 voice recordings and published in Disordered Voice Database[2] (Kay, model 4337 version 2.7.0). The database consists of 710 speech samples (657 diagnosed as pathological, 53 recognized as healthy) with sustained vowel /ah/ and the same number of 12-second reading of the 'Rainbow Passage' [6]. Each recording is described by a number of attributes (patient's age and sex, etc.), diagnosed voice disorder[3] and 33 numerical parameters derived from MDVP (in several cases fewer). Pathological voices were recorded with the sampling frequency 25 kHz and the healthy ones with 50k Hz.

The analysis of features NHR and DSH extraction process revealed a number of problematic questions. Firstly, NHR requires robust F0 tracking to be performed beforehand and it is very sensitive to temporary F0 errors. Secondly, the pitch extraction range is defined to search for either 70-625 Hz or 200-1000 Hz frequencies in the algorithm of NHR calculation. The research revealed that there are 31 samples with F0 below the threshold of 70 Hz in the disordered group. Similarly, DSH uses the pitch range that is insufficient for F0 estimation within the pathological group. Additionally, DSH does not really mean the degree of subharmonic content in the spectrum but a percentage of

[2]The data was collected at the Massachusetts Eye and Ear Infirmary Voice and Speech Lab., Boston, MA.

[3]There is a significant difference between laryngeal and voice disorders: laryngeal disorders develop within vocal tract but not always affect voice quality whereas voice disorders affect the voice with origins not always located within vocal tract (for example spasmodic dysphonia or gastric reflux).

autocorrelation segments with F0 period estimated as doubly longer than the real F0 period. All the healthy patients and more than 77% of disordered voices have DSH equal to zero. The accurate information about the subharmonic to harmonic ratio is erased while even a slight increase of this value may result in audible hoarseness [1]!

4. Cepstral features extraction

The author proposes alternative parameters derived directly from cepstrum that are independent of F0 tracking results performed earlier and the fundamental frequency is estimated simultaneously like the 'side effect'.

As it was mentioned before, the cepstrum plot of voiced speech should contain one dominant peak located at F0 period *quefrency* and (dependently on voice quality and window length) the next dominant peaks situated on F0 period *rahmonics*. The first peak amplitude (*gamnitude*) is proportional to a number and magnitude of F0 harmonics in the spectrum. As it is known, normal voiced speech consists of excitation signal (laryngeal tone) that is filtered by vocal tract and its resonance cavities. That is why, theoretically, the spectrum of normal speech should not contain harmonics different from F0 multiples. Additional noise (considered as harmonics of pathological origin) certainly affects the resulted voice, so g*amnitude* of F0 period *quefrency* peak is expected to be a reliable measure of F0 contribution to the whole signal. This voice attribute was named ***Fundamental Harmonic Index*** (**FHI**) and we assume that it is proportional to the relation:

$$\frac{\text{F0 harmonic}}{\text{signal}} = \frac{\text{F0 harmonic}}{\text{F0 harmonic} + \text{noise}} \tag{3}$$

and thus it represents the alterative voice parameter to Harmonic-to-Noise Ratio derived from MDVP.

While the first dominant peak relates to F0 harmonics in the spectrum, its second *rahmonic* relates to a number and magnitude of subharmonic frequencies in the spectrum. When transferring back to time domain – every second F0 period seems to be more intense that indicates unstable vocal folds activity – undoubtedly a pathological symptom.

Despite the second F0 *rahmonic* is supposed to be higher in the case of disordered voices than the normal ones, the research reveals the opposite tendency – moderate linear correlation (0.571) between the *gamnitude* and the decision. The reason is that pathological voices usually contain a strong noise component that interferes with the subharmonic frequency and its harmonics, as well as F0 frequency. It appears that the relation between the second and the first F0 *rahmonic* is more diagnostically significant than the second *rahmonics* itself. The author proposes to divide the *gamnitude* of the second by the first F0

rahmonic to evaluate the ***Subharmonic-to-Harmonic Ratio*** (**SHR**) – an alternative to Degree of Subharmonic proposed by MDVP.

The parameters of extraction process are very important elements in the features' evaluation, especially a proper window length and sampling frequency. It is essential to take into consideration the accessibly longest F0 period that occurs and the shortest possible frame length to keep the signal stationary (theoretically, voice frame is supposed to be stationary in the range of 10-100 ms [2]). In the case of Disordered Voice Database, the speech samples are recorded with 25 kHz sampling frequency for pathological and 50 kHz for normal voices, so they have to be unified (healthy recordings resampled).

The lowest fundamental frequency during NHR evaluation – 70 Hz appeared to be too high for some of pathological voices, so for further consideration we assume it equals to 50 Hz. The window length must be long enough for cepstrum to contain the second F0 *rahmonic*, so at least 80 ms (since cepstrum is symmetrical and redundant). In the case of 25 Hz sampling frequency the shortest window length that satisfies the conditions and is the power of 2 is 2048 samples long (81.92 ms). The lower border of F0 range is now 48.83 Hz that is sufficient for 100% of healthy and 99.58% of pathological voice samples from the database.

5. Results

Before calculating the new features, the Disordered Voice Database was analyzed in respect to all the voice samples that were either not suitable for searching for F0 (aphonic voices) or with F0 below 70 Hz (there was a high risk of miscalculating NHR and DSH values). For further evaluation 677 of 710 recordings left. The attributes FHI and SHR were evaluated for frames 2048 samples long, windowed with the Hamming function with 64 ms overlapping and the average values were reported as global features.

As mentioned before, both peaks reveal positive correlation with decision (fig. 3a) but their quotient is clearly correlated linearly with the fundamental frequency (0.570) despite the fact the clusters are distributed promisingly (fig. 3b). Finally, the author proposed the second version of Subharmonic-to-Harmonic Ratio:

$$\mathrm{S}HR = \frac{2^{\mathrm{nd}}\ \text{F0 rahmonic gamnitude}}{(1^{\mathrm{nd}}\ \text{F0 rahmonic gamnitude})^2} \tag{4}$$

which minimalized the correlation with F0 to 0.02.

The results for FHI and SHR (the final version) are presented in tab 1 and fig 4. In comparison to NHR and DSH they reveal much better correlation with the diagnosis and the variation.

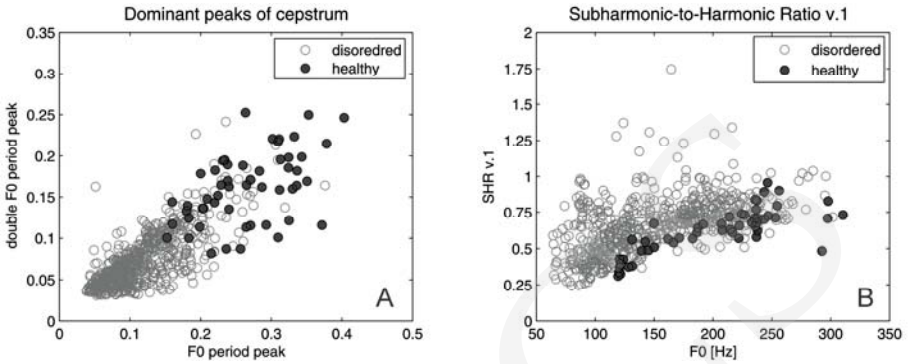

Fig. 3. Distribution of a) cepstral dominant peaks b) Subharmonic-to-Harmonic Ratio for Disordered Voice Database

Table 1. Correlation with the diagnosis for considered voice features

	correlation coefficient		*multiple correlation*
MDVP	**DSH**	**NHR**	**DSH + NHR**
	0.11	0.16	0.18
cepstrum	**SHR**	**FOI**	**SHR + FOI**
	0.24	0.61	0.64

6. Conclusions

In the article the author proposed two voice features that describe voice quality – Fundamental Harmonic Index and Subharmonic-to-Harmonic Ratio, both derived from single cepstral transformation. The features were calculated for 677 recordings from Disordered Voice Database and compared to the similar voice parameters extracted by Multidimensional Voice Program – Noise-to-Harmonic Ratio and Degree of Subharmonic.

The research proved the significant advantage of using the new voice features in comparison to the NHR and DSH not only for their better correlation with the diagnosis but also for better distribution in the 2-dimensional space (of the current feature and the fundamental frequency). In the case of MDVP features, we observe that the central points of the clusters for both pathological and normal voices are almost equal, that makes discrimination very inefficient. The next advantage of the introduced features is the extraction algorithm that evaluates both features using single cepstral transformation and simple dominant values search. It is not required to perform F0 tracking earlier (but if it was prepared, the search could speed up) because F0 frequency is calculated simultaneously, like a ‘side effect’.

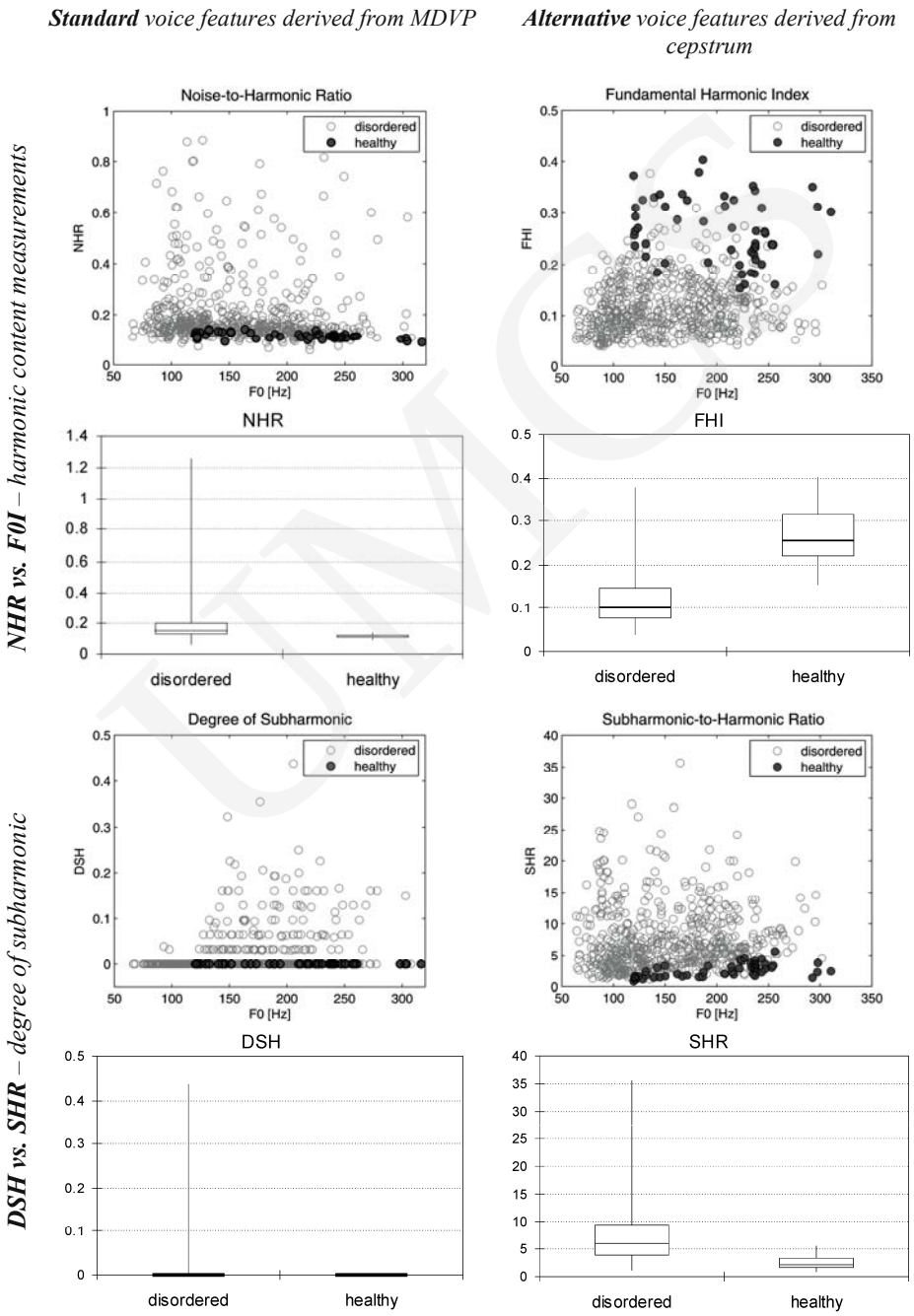

Fig. 4. Distribution and statistics for: NHR (MDVP), DSH (MDVP), FHI (cepstral), SHR (cepstral)

Assuming the maximal possible F0 period-to-period fluctuation we can control the accuracy of the algorithm.

Similarly to NHR and DSH, the results cannot be reliable for aphonic voices and will differ if the parameters of the algorithm or voice sample are changed, so it is very important to satisfy both stationary condition and appropriate accepted F0 range.

Summing up, the introduced new voice features appear to be more diagnostically significant than those derived from Multidimensional Voice Program and certainly could take substantial part in automatic system of pathological voice assessment.

The project is co-financed from European Social Fund and State budget of the Republic of Poland in the frame of the Integrated Regional Operational Programme.

References

[1] Mathieson L., *Greene and Mathieson's the Voice and its Disorders, 6th Ed*. Whurr Publishers Ltd., London, (2006).
[2] Owens F.J., *Signal processing of speech*. The Maximillian Press LTD, Londyn, (1993).
[3] Zieliński T., *Od teorii do cyfrowego przetwarzania sygnałów*, Wydział EAIiE AGH, Kraków, (2002), in Polish.
[4] Verdolini K., Rosen C.A., Branski R.C., *Classification manual for voice disorders – I.* Lawrance Erlbaum Associates, London, (2006).
[5] *Multi-Dimensional Voice Program (MDVP) Model 5105* – software instruction manual, KayPENTAX, Lincoln Park, NJ, (2006).
[6] *Disordered Voice Database Model 4337* – operations manual, KayPENTAX, Lincoln Park, NJ, (2002).