Automatic detection of pathological voice on the basis of continuous speech

Klára Vicsi

Laboratory of Speech Acoustics, Budapest University of Technology and Economics
Department of Telecommunications and Media Informatics
vicsi@tmit.bme.hu

Abstract

A number of experiments were made in the field of speech diagnostic analysis in which researchers wanted to examine whether it was the acoustic characteristics of sustained voice or continuous speech that were more appropriate for distinguishing healthy from pathological voice. While in phoniatric practice, doctors mainly use continuous speech, we also wanted to concentrate on the examination of continuous speech. In this paper we present a series of classification experiments showing how it is possible to separate healthy from pathological speech automatically, on the basis of continuous speech. It is demonstrated that the results of the automatic classification of healthy vs. pathological voice improved to a large extent by a multi-step processing methodology, in which most examples in which uncertainties occurred in the measurement of the acoustic parameters can be accounted for separately.

Index Terms: pathological voice, voice disorder detection, automatic speech recognition, support vector machine

1. Introduction

Generally in voice production, there is a close connection between variation in the voice generation organs (differences in size, in tissue flexibility, etc.) and the measurable acoustic parameters (fundamental frequency, sound pressure, spectrum, etc.) of the speech product generated.

In vocal diagnostic analyses, several examinations were made in connection with the question whether it is sustained voice or continuous speech that is more effective in distinguishing healthy voice from pathological voice ([1], [2], [3], [4], [5], [6]).

In phoniatric practice, mainly continuous speech is applied by phoniatry specialists for the classification of voice quality based on hearing. This is no accident: for generating speech the cooperation of other important articulatory functions is necessary besides vibration of the vocal cords, and thus in the case of any disorder clearness of the voice can be easily disrupted. On the other hand, there is a possibility in continuous speech for the observation of suprasegmental characteristics: emphasis, intonation, and the duration of sonorants.

Let us examine the possibilities of analysis of continuous speech and sustained voice from the viewpoint of acoustic measurements. According to Rabinov et al. [7], the most reliable “tool” for the evaluation of voice quality is the human ear, after all. This can be explained by the fact that in the measurement of oscillation of amplitude and frequency of vibration of vocal cords, these parameters do not take the shape of the generated voice waves into consideration, and that the vibration of the vocal cords is accompanied by a frictional noise. These issues may contain relevant information, mainly in the case of pathological voice.

Titze and colleagues [8] suggest that acoustic measurements (jitter, shimmer) can only provide reliable information in the examination of sustained voice, because the characterization of periodicity can be determined easily due to the quasi-periodicity of the signal, and sustained voice contains enough periods for an authentic calculation of oscillations. On the other hand, in the analysis of continuous speech, where the length of examined sections is very short because of the quick voice transitions, jitter and shimmer results are less reliable.

Zhang & Jiang [1] examined the acoustic characteristics of sustained voice and continuous speech for distinguishing healthy from pathological voice. Acoustic parameters: jitter, shimmer and HNR values were taken into consideration. The authors demonstrated that continuous speech is less suitable for making a distinction between healthy and pathological voice.

While in phoniatric practice doctors mainly apply continuous speech, we also wanted to examine which kind of sound material, continuous speech or sustained vowels, are the best material for the acoustic parameters and for the automatic separation of normal and pathological speech. First, a detailed statistical analysis of acoustic parameters of vowels in continuous speech and sustained voice databases were examined, and the results were compared in healthy vs. pathological speech ([9], [10]). In this talk we present our classification experiments on how it is possible to separate healthy from pathological speech automatically on the basis of continuous speech.

2. Classification experiments

The construction of a well-designed pair of pathological and healthy speech databases was necessary for the examination and for the automatic separation of healthy and pathological voice.

2.1. Patological and healthy speech databases

The sound recordings were made in a consulting room at the Out-patients' Department of Head and Neck Surgery of the National Institute of Oncology. The following diseases occurred in the recorded database: functional dysphonia, recurrent paresis, tumors at various places of the vocal tract, gastro-oesophageal reflux disease, chronic inflammation of larynx, bulbar paresis, its symptoms (paralysis of lips, tongue, soft palate, pharynx and the muscles of larynx), amyotrophic lateral sclerosis, leukoplakia, spasmodic dysphonia and glossectomia. Recordings, for comparison, were also prepared with absolutely healthy patients who had gone to the consulting hours only for control examinations.

Speech samples were recorded by nearfield microphone (Monacor ECM-100), with Creative Soundblaster Audigy 2 NX: an outer USB sound card with 44100Hz, at a 16-bit sampling rate.

The following tasks were recorded from each patient: 3 [e] vowels sustained for a long time, with a deep breath taken before the utterance of each of them, and reading out a folk tale, frequently used in the phoniatric practice, “The North Wind and the Sun”.

The recorded sound samples were classified by a leading phoniatry specialist by the sound perception evaluation scale RBH*, a popular scale in the practice of phoniatry. The scale classifies the voice samples into four classes on the basis of subjectively felt parameters provided by the RBH code. This scale was used to differentiate the degree of voice generation disorders in the database. Speech samples of the patients were labeled on the basis of this numerical scale.

Since we intended to process predefined voiced sequences of the continuous speech material, phoneme-level segmentation of voice files was necessary. It was made in a semiautomatic way, using our own automatic speech recognizer.

The continuous speech (folk tale) samples of 59 speakers were used for the classification experiment (33 pathological and 26 healthy speakers).

**RBH ((Rauhigkeit)* (roughness) *(Behauchtkeit)* (breathiness) *(Heiserkeit)* (hoarseness)*): 0 = normal voice quality, 3 = heavy huskiness.*

2.2. Measured acoustic parameters

Earlier we examined [10] which acoustic analyzing methods best reflected the degree of voice generation disorders (or which fitted the RBH scale of sound perception evaluation the most closely). Statistical distributions of the acoustic parameters were examined by measuring these parameters in sustained vowels and at the middle of vowels in continuous speech. In this experiment it was found that the selected acoustic parameters in the quasi-stationary part of the vowels in continuous speech could replicate the perceptual classification of experts much better than those in the traditionally used steady state sounds. The measured and analyzed parameters were used for the classification experiments, too:

jitter: This is the average absolute difference between consecutive time periods (T) in speech, divided by the average time period. Generally two forms of jitter are in use:

$$jitter_{local} = \frac{\sum_{i=1}^{N-1} |T_i - T_{i+1}|}{\sum_{i=1}^{N-1} T_i} \cdot 100\ [\%] \quad (1)$$

$$jitter_{ddp} = \frac{\sum_{i=2}^{N-1} |2T_i - T_{i-1} - T_{i+1}|}{\sum_{i=2}^{N-1} T_i} \cdot 100[\%]] \quad (2)$$

where N is the number of periods, and T is the length of the periods.

shimmer: This is the average absolute difference between consecutive differences between the amplitudes of consecutive periods.

$$shimmer_{local} = \frac{\sum_{i=1}^{N-1} |A_i - A_{i+1}|}{\sum_{i=1}^{N-1} A_i} \cdot 100\ [\%] \quad (3)$$

$$shimmer_{dda} = \frac{\sum_{i=2}^{N-1} |2A_i - A_{i-1} - A_{i+1}|}{\sum_{i=2}^{N-1} A_i} \cdot 100[\%] \quad (4)$$

where A is the amplitude of the period.

HNR (Harmonics-to-Noise Ratio) represents the degree of acoustic periodicity.

$$HNR = 10 * log\frac{E_H}{E_Z}\ [dB] \quad (5)$$

where E_H and E_Z are the energy of the harmonic and noise component, respectively.

2.3. Classification of healthy and pathological voices

All of the sounds [e] in the reading test were used for the classification. At the middle of the [e] sound, the following acoustic parameters were measured: local jitter, ddp jitter, local shimmer, dda shimmer and HNR values. The average, the spread, min, max, and median of the measured values were calculated. These vectors were the input of the classifier. A special neural net, the Super Vector Machine (LibSVM (http://www.csie.ntu.edu.tw/~cjlin/libsvm/)) was used for the classification, and the Leave-One-Out Cross-Validation (LOOCV) technics was used for training and testing. For the selection of the most important parameters, a series of pilot experiments were conducted. Different groups of the acoustic parameters were used and the recognition (classification) results were examined.

Table 1. *Testing the classifier with various incoming acoustic parameters (jitter = jitter(local) and jitter(ddp) together; shimmer = shimmer(local) and shimmer(dda) together)*

acoustic parameters	statistics for the sound [e]	
	average	average,spread, min, max, median
jitter, shimmer, hnr, mfcc	73%	63%
jitter, shimmer, mfcc	73%	63%
jitter, shimmer	79%	79%
jitter(local), shimmer(local)	79%	79%
jitter(ddp), shimmer(dda)	**84%**	79%
jitter, shimmer, hnr	73%	73%
hnr, mfcc	68%	63%
mfcc	73%	63%

The best classification results of healthy and pathological speech were obtained when jitter(ddp) and shimmer(dda) averages were the incoming acoustic parameters. See Table 1.

We wanted to analyse this result further, in terms of how the healthy samples were separated from the pathological cases on the basis of these two parameters. Thus we plotted the spread values as a function of average values in the case of
jitter(ddp) and shimmer(dda). See Fig.1. and Fig. 2.

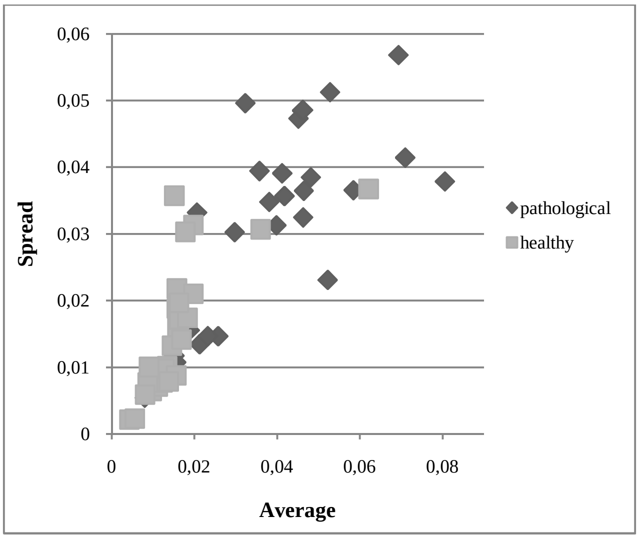

Fig. 1. Spread of jitter (ddp) as a function of the average
in the case of the sound [e]

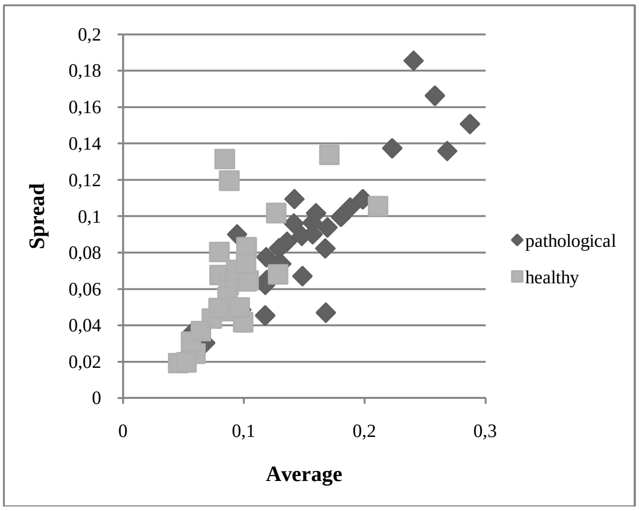

Fig. 2. Spread of shimmer (dda) as a function of the average,
in the case of the sound [e]

There are a few salient examples in case of the spread of both jitter and shimmer. Analyzing these examples one by one, it turned out that all of those salient examples originated in measuring problems. Either the fundamental frequency was too low or too high, or the voiced period was very short. When the fundamental frequency was too high, the fluctuation in % was smaller than expected. In cases of too low frequencies, the measurement of the fundamental frequencies became ambiguous, yielding mistakes. Both jitter and shimmer measurements are based on the measurement of fundamental frequency, thus in the cases mentioned we got nonsensical results, yielding salient data. In the case when the voiced period was very short, much shorter than the voiceless one, inadequate examples caused mistakes. Although it seems that Zhang and Jiang [1] were right when they concluded that continuous pathological speech contained too short examples for the authentic calculation of the jitter and shimmer parameters, the calculation difficulty occurred only with some of the examples, and those examples could be selected and evaluated in a different way.

2.4. Separation of the uncertain examples

As a first step, it was necessary to decide the threshold of the voiced/voiceless frame rate under which the examples are selected. Thus voiced/voiceless frame rates were calculated in the continuous part of the speech of the patients. A 75ms window was used with 18.75ms frame steps. The voiced/unvoiced frame rate was calculated as follows:

$$Fr_{v/u} = \frac{\sum_{i=1}^{N} Fr(v)_i}{\sum_{i=1}^{N} Fr(u)_i} \qquad (6)$$

where $Fr(v)_i$ is the number of voiced frames, and $Fr(u)_i$ is the number of voiceless frames of the i-th sound. The distribution of these voiced/unvoiced frame rates in case of healthy and pathological voices is presented in Figure 3.

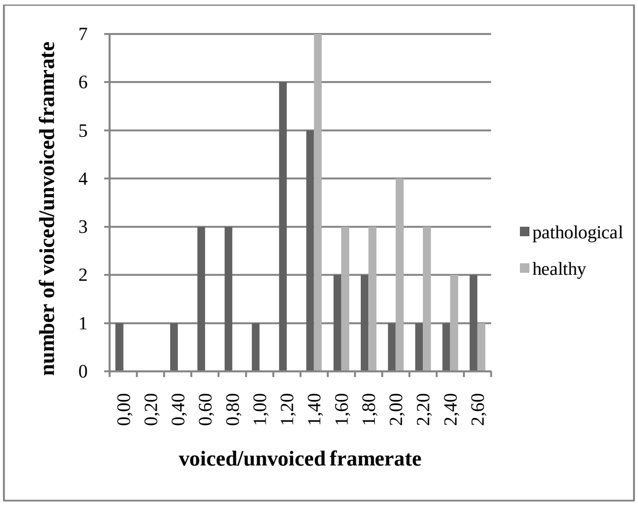

Fig. 3. Distribution of the voiced/voiceless frame rates of the healthy and pathological voices

It is quite clear from the measurement that healthy speech does not exist under a 1.4 frame rate. The quality of these sounds is the worst. The error of the measurement of the fundamental frequency is high. Examples in which the voiced/voiceless frame rate was less than 0.5 were filtered out, as they are surely pathological voices. Then the classification was repeated, and now the classification exactness increased to 86%.

As a second step, on the basis of the distribution of fundamental frequencies, those examples were separated off where the fundamental frequency was higher than 160Hz in case of men, and 270Hz in the case of women. These examples are also pathological voices, because in these cases the patients wanted to make their hoarse voices better by increasing the fundamental frequency. The classification was repeated again, and now its exactness increased to 88%.

3. Conclusions

It was impressive for us how easily we could increase the exactness of the classification by avoiding some measuring problems. In spite of the opinion of Rabinov et al. [7] that the most reliable “tool” for the evaluation of voice quality is the human ear, and the results of Zhang and Jiang [1], it is clear for us that it is worth going further in our way and use continuous speech for the detection and classification of pathological voices. Of course, far more data collection has to be undertaken for obtaining better results, and further investigation is necessary to decide which acoustic parameters are the best for the classification.

4. References

Yu Zhang, and Jack J. Jiang, “Acoustic Analyses of Sustained and Running Voices From Patients With Laryngeal Pathologies”, accepted for publication: 2006, Journal of Voice, Vol. 22, No. 1: 1-9, 0892-1997.

Ce Peng, Wenxi Chen, Xin Zhu, Baikun Wan and Daming Wei, “Pathological voice classification based on a single vowel’s acoustic features”, 0-7695-2983-6/07, IEEE. 2007.

Parsa, V. and Jamieson, D.G., “Acoustic discrimination of pathological voice: Sustained vowels versus continuous speech”, Journal of Speech, Language, and Hearing Research 44: 327–339, 2001.

Anders G. Askenfelt and Britta Hammarberg, “Speech Waveform Perturbation Analysis, A Perceptual-Acoustical Comparison of Seven Measures”, Journal of Speech and Hearing Research Vol. 29: 50-64, 1986.

Ce Peng, Wenxi Chen, Xin Zhu, Baikun Wan and Daming Wei, “Pathological Voice Classification Based on a Single Vowels’s Acoustic Features”, 7th. International Conf. on Computer and Information Techn, IEEE: 1106-1110, 2007.

R.T. Ritchings, M. McGillion and C.J. Moore, “Pathological voice quality assessment using artificial neural networks”, 2002, Medical Engineering & Physics 24: 561-564, 2002.

C. Rose Rabinov, Jody Kreiman, Bruce R. Gerratt and Steven Bielamowicz, ”Comparing Reliability of Perceptual Ratings of Roughness and Acoustic Measures of Jitter”, Journal of Speech and Hearing Research, Volume 38: 26-32, 1995.

Titze, I., Wong, D., Milder, M., Hensley, S., and Ramig, L., “Comparison between clinician-assisted and fully automated procedures for obtaining a voice range profile”, J. Speech Hear. Res., 35: 526-535. 1995.

Imre, Viktor, “Acoustical examination of pathological voices”, MSc thesis. Budapest University of Technology and Economics 2009.

Vicsi, Klára and Imre, Viktor, “Voice disorder detection on the base of continuous speech”, 4th Advanced Voice Function Assessment Workshop, COST Action 2103. York 2010: 42, 2010.