# Estimating RASATI scores using acoustical parameters

**Pablo Daniel Agüero, Juan Carlos Tulli, Graciela Moscardi, Esteban Gonzalez and Alejandro Uriz**

Facultad de Ingenieria - Universidad Nacional de Mar del Plata

E-mail: pdaguero@fi.mdp.edu.ar

**Abstract.** Acoustical analysis of speech using computers has reached an important development in the latest years. The subjective evaluation of a clinician is complemented with an objective measure of relevant parameters of voice. Praat, MDVP (Multi Dimensional Voice Program) and SAV (Software for Voice Analysis) are some examples of software for speech analysis. This paper describes an approach to estimate the subjective characteristics of RASATI scale given objective acoustical parameters. Two approaches were used: linear regression with non-negativity constraints, and neural networks. The experiments show that such approach gives correct evaluations with $\pm 1$ error in 80% of the cases.

## 1. Introduction

In the latest years the acoustical analysis of speech has reached an important development thanks to the progress of computers. The main advantage of computer analysis of speech is the non-invasive and objective assessment of the voice.

The human auditory system is one of the main obstacles in the perceptual diagnostic of voice by the clinician ear. Humans are fundamentally prepared to perceive the voice as a whole, which is particularly advantageous from the point of view of linguistic communication. However, this ability is limited when it is necessary to individualize relevant aspects from a clinical perspective.

It is often difficult to determine the origin of certain anomalies of the voice using a perceptual procedure. For example, Baken et al. [1] show that some aspects of the pitch are more related to resonant frequencies of the vocal tract rather than to the frequency of vibration of vocal chords. The hypernasality of voice can be a consequence of the desynchronization in the timing of velar occlusion instead of an incomplete occlusion. Hence, the same attribute or alteration of the vocal quality may have its origin in different subsystems which can not be easily isolated with the audition of an expert.

In other cases, an adequate perception can not be quantized with the degree of precision of a numerical measure. For example, it is possible to measure the degree of breathiness of a breathy voice through the corresponding speech parameter, the index of turbulence of voice (or VTI). In this way, the subjective evaluation of a clinician is complemented with an objective measure of relevant parameters of voice. As a consequence, the objectivity of the report is enhanced, and it is possible to measure the degree of progress more accurately.

Validity and reliability of acoustic analysis performed with different tools is affected by many factors. These include microphone type, noise levels, data acquisition system, sampling rate

and software used for analysis [4, 5]. Ostensibly, the values of the commonly used frequency and amplitude perturbation measures should not be dependent on the software used to obtain them. Jitter and shimmer, for example, are defined by relatively simple and standardized formulas [2]. The differences observed between numerical values obtained for these measures using different softwares apparently stem from the raw fundamental frequency ($f_0$) data on which these calculations are based. Despite the basic nature of this parameter, there is no standardized algorithm to calculate $f_0$, which has been adopted and implemented by all programs.

While different methods for calculating $f_0$ may yield relatively small differences in the mean value of $f_0$, they may influence the perturbation measures to a far greater extent. This introduces a difficulty for the clinical voice specialist, because different programs which are available for conducting voice analysis could report different values when analyzing identical voice samples. Moreover, it is not clear whether normative data which are presented by specific software (e.g., the data used for the radial graph in Multi-Dimensional Voice Program, or MDVP) are comparable with values obtained in other programs. This possible discrepancy between the results obtained by different programs was previously noticed and addressed by various researchers [8, 4, 12, 6].

In the voice analysis of speech disorders exists a gap between the subjective and objective measures. On the one hand, objective measures are based on time and frequency calculations. On the other hand, subjective scores are based on scales with more complex concepts that have a correlate with objective parameters. One example of a subjective criterion is the RASATI scale, which considers aspects such as hoarseness(R), rough (A), breath (S), asthenic (A), strain (T) and instability (I). In that scale the severity of a pathology is measured using a zero to three discrete scale.

This paper explores the relationships between the perturbation measures and the subjective scale RASATI. The main goal is to estimate the values of the subjective evaluation given objective measures, such as jitter, shimmer, HNR, etc. In this way, it would be possible to trace the progress of a patient using more human perceivable factors, such as hoarseness or strain.

This paper is organized as follows. Sections 2.1 and 2.2 describe the objective and subjective approach to assess voice quality in speech disorders, and Section 2.3 depicts the proposed methodology to find the relationships between objective and subjective scores. Section 3 shows the experimental results with the acoustic parameter calculated by SAV and PRAAT. Finally, conclusions and future work are drawn in Section 4.

## 2. Evaluation of speech disorders

The evaluation of speech disorders can be performed using both subjective and objective scores. The former use some ratings to measure different aspects of the voice quality. On the other hand, objective measures use acoustic parameters obtained with different computer algorithms to grade the voice pathology.

### *2.1. Subjective evaluation of speech disorders*

Proposed by Hirano [7] and accepted as standard by the Japanese Society of Logopedics and Phoniatrics and the European Group on the Larynx, the GRBAS scale comprises five qualitative characteristics: Grade of dysphony (G), Roughness (R), Breathiness (B), Asthenicity (A), and Strainess (S). For each one, a value in the range 0-3 is considered, where 0 corresponds to healthy voice, 1 to light disease, 2 to moderate and 3 to severe. Despite some limitations, GRBAS is simple and fast, and has a good correlation with some acoustic parameters [11].

The severity of hoarseness is quantified under the parameter G (Grade) integrating all deviant components. Two main components of hoarseness can be identified: Breathiness (B), which is the audible impression of turbulent air leakage through an insufficient glottal closure, and it may

include short aphonic moments (unvoiced segments); and Roughness (R), which is an audible impression of irregular glottic pulses, abnormal fluctuations in F0, separately perceived acoustic impulses (as in vocal fry), and includes diplophonia and register breaks [11].

These two parameters have shown sufficient reliability (inter and intra observer reproducibility) when used in a current clinical setting [3]. The behavioral parameters A (Asthenicity) and S (Strain) are commonly less reliable and sometimes are omitted from the basic protocol. R and B features are associated to organic lesions in which there is a lowering of vibration (R) and default of closure (B), whereas features A and S are associated to functional disorders, related with vocal tiredness (A) and hyperphonic emission (S) [11].

The GRBAS evaluation is usually carried out based on continuous or conversational speech. However, sometimes it is approached by means of sustained vowels, although there are studies demonstrating that the results might differ depending on the material used [10]. They conclude that the evaluation from sustained vowels is less severe (i.e. dysphony is underestimated) than that carried out from continuous speech, especially in those patients with severe dysphony. The same study calls the attention over the variability of each of the five GRBAS parameters. The most consistent parameter is G, whereas scales A and S demonstrated a strong variability, due to the fact that these concepts are more complex to evaluate, even by a human expert [11].

The RASATI scale is the acronym proposed by Pinho et al. [9] to replace the English acronym GRBAS, and incorporates another factor named instability. Some authors consider that Instability corresponds to the tremor of the structure of the vocal tract, and must not be included in the analysis of alterations in the glottal source. The RASAT scale (not including Instability) is the standard adopted by the "Sociedad Argentina de la Voz" to measure voice quality.

### *2.2. Objective evaluation of speech disorders*

The literature on voice analysis reveals that one or two voicing parameters alone, such as jitter and shimmer, are not sufficient to accurately describe an aberration in a patient's voice. Jitter values may be within normal limits in a patient who demonstrates a breathy voice quality, and periodic modulation over many glottal periods (tremor) should be differentiated from cycle-to-cycle modulation. Similarly, turbulence caused by incomplete glottal closure can contribute a different type of "noise" compared to noise from aperiodic vibration; and, longterm periodic modulation of amplitude (amplitude tremor) may have physiological causes that differ from those of long-term periodic modulation of frequency. The analysis of voice requires a multi-dimensional approach which is followed by many software applications. Examples of applications that agree with this direction are the Multi-Dimensional Voice Program (MDVP) of KayPentax, and the Software for Voice Analysis (SAV).

Some of the parameters estimated by MDVP are: Jita (Absolute Jitter), Jitt (Jitter percent), RAP (Relative Average Perturbation), PPQ (Pitch Perturbation Quotient), sPPQ (Smoothed Pitch Perturbation Quotient), vFo (Fundamental frequency variation), ShdB (Shimmer in decibels), Shim (Shimmer percent), APQ (Amplitude Perturbation Quotient), sAPQ (Smoothed Amplitude Perturbation Quotient), vAm (Peak-to-Peak Amplitude Variation), NHR (Noise Harmonic Ratio), VTI (Voice Turbulence Index), SPI (Soft Phonation Index), FTRI (Fo-Tremor Intensity Index), ATRI (Amplitude Tremor Intensity Index), DVB (Degree of Voice Breaks), DSH (Degree of Sub-harmonics), and DUV (Degree of Voiceless). Many of these parameters have become standards in the analysis of voice, and several papers about the study of speech disorders are based on the results of this software.

The Software for Voice Analysis developed estimates several parameters (some of them available in MDVP), such as jitter (jittr, jitta, jittrap,jittppq5), shimmer (shimr, shima, shimrap, shimppq5), HNR and SPI. The algorithm to estimate the values of pitch period is similar to the one used by PRAAT, which is based on the autocorrelation method. MDVP pitch estimation

model is based on peak picking, and such approach is under some controversy in the paper of one of the authors of PRAAT: "Should jitter be measured by peak picking or by waveform matching?".

*2.3. Relationships between objective and subjective evaluations*

The study of the relationships between objective and subjective evaluations can be carried out with several statistical techniques, such as logistic regression and correlation studies. This paper proposes to estimate the values of the subjective evaluation given the objective parameters using two approaches: linear regression with non-negativity constraints and feed-forward neural networks.

Linear regression is a modelling approach to estimate a linear relationship between a scalar variable $y$ and one or more variables denoted $X$. In linear regression, data are modeled using linear functions ($y = \beta_0 + \sum_i^N \beta_i x_i$), and unknown model parameters are estimated from the data ($\beta_0, \beta_1, ..., \beta_N$). $y$ is called the dependent variable, and $x_i$ are called predictor or independent variables. In this case the dependent variable $y$ is one of the RASATI values that ranges from zero to three, and the independent variables $x_i$ are the parameters of the objective evaluation, such as relative jitter or HNR.

The main drawback of linear models is the lack of a non-negativity constraint in the weights ($\beta_i$). Some $\beta$ values may be negative, and such linear function will have a distorted behaviour if it is considered each weight as a measure of the importance of each parameter. For example, a negative weight would mean that the higher the jitter the lower the severity of the pathology, which is completely unlogical. Because of that the linear regression uses a non-negativity constraint in the least squares optimization of the weights.

$$y = \beta_0 + \sum_i^N \beta_i x_i, \beta_k \geq 0 \tag{1}$$

The neural network approach is also explored in this paper. An artificial neural network (ANN), usually called neural network (NN), is a mathematical model that is inspired by the structure and functional aspects of biological neural networks. A neural network consists of an interconnected group of artificial neurons, and it processes information using a connectionist approach to computation.

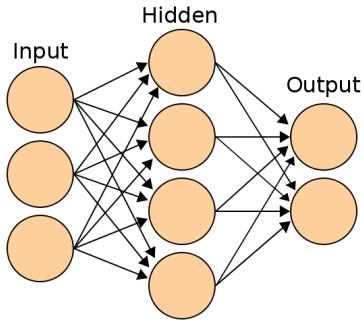


**Figure 1.** Example of a neural network with three layers: input, hidden and output

Mathematically, a neuron's network function $f(x)$ is defined as a composition of other functions $g_i(x)$, which can further be defined as a composition of other functions. This can be conveniently represented as a network structure, with arrows depicting the dependencies between variables, as shown in Figure 1. A widely used type of composition is the nonlinear

weighted sum, where $f(x) = K\left(\sum_i w_i g_i(x)\right)$ , where $K$ (commonly referred to as the activation function) is some predefined function, such as the hyperbolic tangent. This work in this paper uses feed-forward networks, because their graph is a directed acyclic graph, as shown in Figure 1.

The inputs of the neural network are the parameters calculated in the objective evaluation, and the outputs is the value of the objective evaluation in the RASATI scale. The first neuron corresponds to the value zero, and the last one to the value three.

## 3. Experiments

This section shows the experimental results of the estimation of the subjective values of the RASATI scale with the parameters calculated from the audio signal with two different acoustical analysis applications: SAV and PRAAT.

### *3.1. Experimental setup*

The experiments of the two approaches to estimate the subjective values of the RASATI scale given the objective parameters were performed using a database of 105 samples. Each recording has a sampling frequency of 44100 samples per second, and 16 bits of amplitude resolution. Each vocalization was evaluated by the speech therapist to score each qualitative characteristic of the RASATI scale from zero to three, with a rounding to plus infinity when some score fall between two possible values. For example, a value of one to two was scored as two.

Each patient uttered a sustained vowel (/a/ or /e/) with an approximate duration of eight seconds. The patients range from the age of 16 to 88, and there are patients from both sexes with different pathologies, such as disfunctional dysphonia, nodules, laryngitis, polypus or papilloma. Some recordings correspond to several sessions of the same patient.

The experiments were performed using the two approaches: linear regression with non-negativity constraints in the weights, and feed-forward neural networks. In both cases the leave-one-out cross-validation technique was used because of the limited available data in the experiments. Leave-one-out cross-validation involves using a single observation from the original sample as the validation data, and the remaining observations as the training data. This is repeated such that each observation in the sample is used once as the validation data. Leave-one-out cross-validation is usually very expensive from a computational point of view because of the large number of times the training process is repeated. The small amount of training data in the experiments is suited to be used with this cross-validation approach.

The linear regression with non-negativity constraints was calculated using the *LSQNONNEG* command of MATLAB to estimate the weights using the training data. LSQNONNEG returns the vector X that minimizes NORM(C*X-d) subject to $X \geq 0$. The non-negativity constraints are necessary to estimate the weights in order to avoid negative values that would indicate an inverse behaviour. For example, if the weight of the jitta parameter is negative, it would mean that the greater this value, the smaller the pathology, which is clearly wrong.

Feed-forward neural networks were trained using the backpropagation optimization algorithm. The neural networks have multiple inputs and four outputs, each one corresponding to a possible value of the RASATI evaluation. The hidden layer consisted of 10 neurons with a tansig activation function.

Two applications to estimate the objective parameters were used: SAV and PRAAT. PRAAT was chosen because is one state-of-the-art free software for the analysis of speech in phonetics written by Paul Boersma and David Weenink of Phonetic Sciences of University of Amsterdam (The Netherlands).

The acoustical parameters estimated with the Software for Voice Analysis in this experiment are jittr, jitta, jittrap,jittppq5, shimr, shima, shimrap and shimapq5. The parameters estimated with PRAAT were jittr, jitta, rap, ppq5, ddp, shimr, shimdB, apq3, apq5, apq11 and dda.

*3.2. Experimental results*

The experimental results with the linear regression with non-negativity constraints approach is shown in Figure 2. The columns that end with the letter S correspond to the results obtained with SAV, and the columns with an ending P correspond to the results obtained using PRAAT. Each column shows in different colours the proportions of cases were the estimation of the RASATI values using the objective parameters was lower than the value decided by the speech therapist (-1, -2 or -3), higher (+1, +2 or +3), or equal. The similar results obtained with the parameters estimated with PRAAT and SAV shows that the linear regression with non-negativity constraints achieves an error around $\pm 1$ in the 80% of the cases. Such errors are not so severe because RASATI scale is a subjective assessment procedure with only four possible values, and any slight difference of perception produces a $\pm 1$. In these experiments, fluctuations in the order of $\pm 1$ are possible due to the sensitivity of objective measures. Nevertheless, future work should be devoted to better match the subjective opinion or uncover their origin in each case.

Figure 3 shows the experimental results using the feed-forward neural networks. The distributions of the errors are similar to the results obtained with the linear regression with non-negativity constraints approach. The first approach shows a smaller error than the neural networks for all conditions, but such difference is not significant due to the small amount of data used in the experiments.

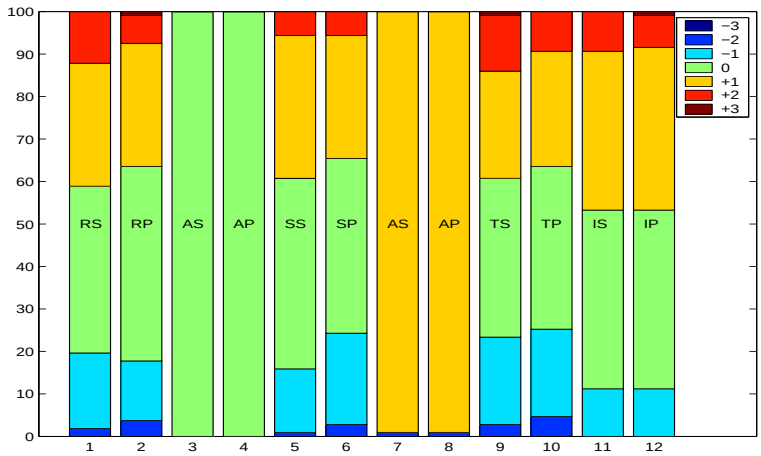


**Figure 2.** Distribution of the absolute errors for the linear regression with non-negativity constraints approach

The analysis of the weights shows that the important parameters to estimate RASATI scores by means of SAV parameters are (ordered by relevance):

- R : jittr, jitta, shimr and shimapq5.
- A : too few cases available.
- S : jitta, jittr, shimr and shimapq5.
- A : too few cases available.
- T : jittr, shimapq5 and shimr.
- I : jittr, shimapq5 and shimrap.

In the case of PRAAT acoustical parameters the order of relevance are:

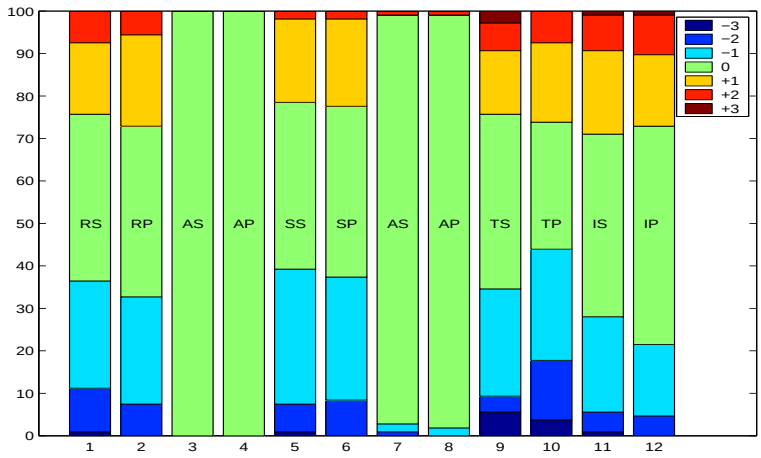


**Figure 3.** Distribution of the absolute errors for the feed-forward neural network approach

- R : jittr, apq11, shimdB and jitta.
- A : too few cases available.
- S : apq11, ppq5 and shimdB
- A : too few cases available.
- T : apq11, jitta and shimdB.
- I : jittr, apq11 and ppq5.

In both cases is observed changes in the acoustical features and also in the order of relevance. This fact is important to model each RASATI characteristic individually.

## 4. Conclusions

In this paper it was made a set of comparative experiments to study the relationships between objective and subjective evaluations of speech disorders. The subjective measure is the RASATI scale, the standard chosen by the "Sociedad Argentina de la Voz" to measure voice quality.

Experimental results shown that linear regression with non-negativity constraints and neural networks achieve similar estimation performances. The error in the estimation of the RASATI value is around $\pm 1$ in the 80% of the cases. Such errors are not so severe because RASATI scale has only four possible values, and any slight difference of perception produces a $\pm 1$. These fluctuations are possible due to the sensitivity of objective measures.

Future work will focus in the evaluation of additional objective acoustical features to improve the estimation of the six qualitative characteristics of RASATI scale, and to uncover the origin of the differences between the estimation and the opinion of the speech therapist in each case.

## References

[1] Baken, R., Orlikoff, R.: Clinical measurement of speech and voice. In: Second Edition. San Diego, CA: Singular Publishing Group (2000)
[2] Baken, R.: Clinical measurement of speech and voice. In: Allyn and Bacon, Needham Heights, MA (1987)
[3] Dejonckere, P.: Effect of slightly louder voicing on acoustical measurements in different etiological categories of disphonia. In: Proceedings de Voicedata. pp. 86–91 (1998)
[4] Deliyski, D., Shaw, H., Evans, M.: Influence of sampling rate on accuracy and reliability of acoustic voice analysis. In: Logopedics, Phoniatrics, Vocology. vol. 30, pp. 55–62 (2005)
[5] Deliyski, D., Shaw, H., Evans, M.: Regression tree approach to studying factors influencing acoustic voice analysis. In: Folia Phoniatrica et Logopaedica. vol. 58, pp. 274–288 (2006)

[6] Godino-Llorente, J., Osma-Ruiz, V., Saenz-Lechon, N.: Acoustic analysis of voice using wpcvox: a comparative study with multi dimensional voice program. In: European archives of otorhinolaryngology. vol. 265, pp. 465–476 (2008)
[7] Hirano, M.: Psycho-acoustic evaluation of voice. In: New York: Springer-Verlag (1981)
[8] Karnell, M., Hall, K., Landahl, K.: Comparison of fundamental frequency and perturbation measures among three analysis systems. In: Journal of Voice. vol. 9, pp. 383–393 (1995)
[9] Pinho, S., Pontes, P.: Musculos intrinsecos da laringe e dinamica vocal serie desvendando os segredos da voz. In: Revinter (2008)
[10] Revis, J., Giovanni, A., Wuyts, F.: Comparison of different types of vowel fragments for the evaluation of voice quality. In: Proceedings de Voicedata. pp. 80–85 (1998)
[11] Saenz-Lechon, N., Godino-Llorente, J., Osma-Ruiz, V., Blanco-Velasco, M., Cruz-Roldan, F.: Automatic assessment of voice quality according to the grbas scale. In: International Conference of the IEEE. pp. 2478–2481 (2006)
[12] Smith, I., Ceuppens, P., Bodt, M.D.: A comparative study of acoustic voice measurements by means of dr. speech and computarized speech lab. In: Journal of Voice. vol. 19, pp. 187–196 (2005)