XIV Reunión de Trabajo Procesamiento de la Información y Control -- RPIC 2011, pp. 114-119, 2011.

Estimation of RASATI scores using objective acoustical parameters and CART

Pablo Daniel Agüero[†] Juan Carlos Tulli[†] Graciela Moscardi[‡]
Esteban Lucio Gonzalez[†] Alejandro Jose Uriz[†] Simon Bourguigne[†]

†*Facultad de Ingeniera*
Universidad Nacional de Mar del Plata, Argentina
pdaguero@fi.mdp.edu.ar
‡*Universidad FASTA, Argentina*

***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 and SAV 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. This approach uses classification and regression trees, a commonly used machine learning tool. The experiments show that such approach gives correct evaluations with ±1 error in $80-90\%$ of the cases.**

***Keywords*— speech disorders, voice analysis, machine learning, RASATI scale.**

1 INTRODUCTION

Since the last decades of the past century 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 perceptual limitations of 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 locations. This 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 [7, 8]. Seemingly, 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 (F0) data on which these calculations are based. Despite the basic nature of this parameter, there is no standardized algorithm to calculate F0, which has been adopted and implemented by all softwares.

While different methods for calculating F0 may yield relatively small differences in the F0 mean, such as peak picking or waveform matching [3], 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 determined by specific software (e.g., the data used for the radial graph in Multi-Dimensional Voice

Program, or MDVP) are comparable with values obtained with other softwares. This possible discrepancy between the results obtained by different programs was previously noticed and addressed by various researchers [7, 10, 13, 17].

In 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 of several qualitative concepts that have a complex relationship with objective parameters. One example of a subjective criterion is the RASATI scale, which indicates the severity of a pathology using a zero to three discrete scale, considering aspects such as roughness(R), harshness(A), breathiness(S), asthenia(A), strain(T) and instability (I) [14].

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. Subsections 2. 1 and 2.2 describe the objective and subjective approach to assess voice quality in speech disorders, and Subsection 2. 4 depicts the proposed methodology to find the relationships between objective and subjective scores. Section 3 shows experimental results with the acoustic parameter calculated using the software of voice analysis developed at the Engineering Faculty of University of Mar del Plata named SAV. 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 [12] 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 [16].

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 [16].

These two parameters have shown sufficient reliability (inter and intra observer reproducibility) when used in a current clinical setting [5]. 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) [16].

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 [15]. 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 [16].

The RASATI scale is the acronym proposed by Pinho et al. [14] 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.

In South America, the RASATI scale is widely used . In the RASATI scale, R (rouquidao) means roughness, A (aspereza) means harshness, S (soprosidade) means breathiness, A (astenia) means asthenia, T (tensao) means strain, and I (instabilidade) means instability. R is defined as irregular vocal fold vibrations, A (aspereza) is defined as mucosa rigidity, S is represented by an audible noise created at the glottis, A (asthenia) is related to hypofunctional voice, T is associated with an excessive vocal effort, and I is related to vocal tremor. This scale also uses a 4-point rating, similar to GRBASI [19].

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 that study speech disorders are based on the results of this software.

The Software for Voice Analysis (SAV) estimates several parameters (some of them available in MDVP), such as jitter (jittr, jitta, jittrap,jittppq5), shimmer (shimr, shima, shimrap, shimppq5), HNR, SPI, VTI, spectral flux, spectral tilt, and relationships of amplitude and slope between harmonics. 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?" [3].

2. 3 New acoustic features added to SAV

The initial setup of SAV only included some of the acoustic features mentioned in the previous sections: jitter (jittr, jitta, jittrap,jittppq5), shimmer (shimr, shima, shimrap, shimppq5), HNR and SPI. This fact can be observed in the manual of SAV available in the home web page [1].

One of the important acoustic features added to SAV is the relationship between the amplitudes of formants and harmonics. Information obtained from spectral structure is potentially more reliable than F0 or intensity for the purpose of voice quality identification. N Chasaide and Gobl [4] characterize creaky phonation as having slow and irregular glottal pulses in addition to low F0. Specifically, they state that significant spectral cues to creaky phonation are A1 (i.e., amplitude of the strongest harmonic of the first formant) much higher than H1 (i.e., amplitude of the first harmonic), and H2 (i.e., amplitude of the second harmonic) higher than H1 [20]. These amplitudes are calculated from the long-term average spectrum (LTAS). It is computed by time-averaging the short-term Fourier magnitude spectra, resulting in one feature vector for the whole speech sample. The advantage of LTAS is that it has more or less direct physical interpretation, relating to the location of the vocal tract resonances.

Another spectral feature added to SAV is the spectral flux. The spectral flux is a descriptor aiming at quantifying the variation of the spectrum along time. This temporal variation is computed from the normalized cross-correlation between two successive amplitude spectra, and it is aggregated using the average over all speech signal to obtain a single value for the entire phonation [9].

Breathiness has been associated with an increase in the intensity of aspiration noise as well as changes in spectral slope [11]. The spectral slope is another representation of the amount of decreasing of the amplitude spectrum. It is computed by linear regression of the spectrum.

Voice Turbulence Index (VTI) is an indicator that mostly correlates with the turbulence components caused by incomplete or loose adduction of the vocal folds. VTI calculated as the ratio of the spectral in-harmonic high-frequency energy (2800-5800Hz) to the spectral harmonic energy (70-4500Hz) [6].

2. 4 Relationships between objective and subjective evaluations

The study of the relationships between objective and subjective evaluations can be carried out with several machine learning techniques, such as classification and regression trees or support vector machines. This paper proposes to estimate the values of the subjective evaluation given the objective parameters using classification and regression trees.

Classification and regression trees (CART), or decision trees, classify data according to a series of hierarchical questions, asked at each branch in the tree about features that describe the data points (input features). The training algorithm chooses which feature provides the highest predictive capability at each

[1] http://elaf1.fi.mdp.edu.ar/pegasus/SAV.html

node and grows the tree accordingly until some stopping criterion is met.

Figure 1 shows a scheme of a classification and regression tree, with a root node and intermediate nodes with questions about input features. The terminal nodes or leaves contain the decision. After all previous questions were answered and lead to a particular terminal node, the tree can indicate the certainty of the decision with a probability. The gray scale in the right indicates that the certainty in the decision increases from top (the root) to bottom (terminal nodes).

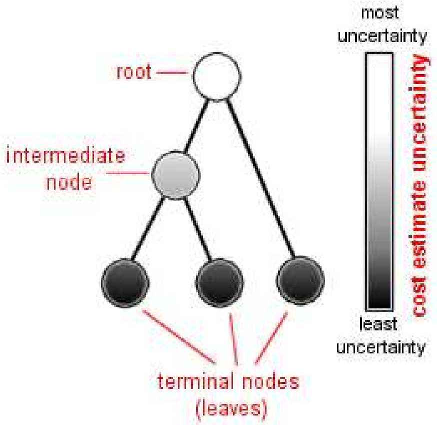

Figure 1: Classification and regression trees

The advantages of decisions trees are that they can handle categorical and numerical values, and that they choose the ordering of features (questions) for prediction automatically. They create models that provide information about the data since the resulting tree can be read as a hierarchy of questions. This is different from the models created using neural networks which are very difficult to interpret.

For the experiments was used the *wagon* decision tree implementation that is part of the Festival speech synthesizer [18]. The input features of the classification and regression tree are the parameters calculated in the objective evaluation, and the output in the leaves is the most probable RASATI scale value.

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 the Software for Voice Analysis (SAV).

3. 1 Experimental setup

The experiments to estimate the subjective values of the RASATI scale given the objective parameters were performed using a database of 143 recordings. 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, from both sexes with different pathologies, such as dysfunctional dysphonia, nodules, laryngitis, polypus or papilloma. Some recordings correspond to several sessions of the same patient.

The experiments were performed using classification and regression trees, with individual trees for each RASATI parameter. 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 objective parameters were estimated using SAV. The acoustical parameters estimated with the Software for Voice Analysis in this experiment are jittr, jitta, jittrap,jittppq5, shimr, shima, shimrap and shimapq5, HNR, SPI, VTI, spectral flux, spectral tilt, and relationships of amplitude and slope between harmonics.

The experiments were performed under two conditions:

- **Original setup of SAV**: jitter (jittr, jitta, jittrap,jittppq5), shimmer (shimr, shima, shimrap, shimppq5), HNR and SPI.
- **New setup of SAV**: jitter (jittr, jitta, jittrap,jittppq5), shimmer (shimr, shima, shimrap, shimppq5), HNR, SPI, VTI, spectral flux, spectral tilt, and relationships of amplitude and slope between harmonics.

3. 2 Experimental results

The experimental results with the classification and regression tree approach using the original setup are shown in Fig. 2. The columns are named R, A, S, A, T and I according to each RASATI qualitative characteristic. Each column shows the proportions of cases with different colors where 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 results shows that the approach achieves an error around ±1 in the $70 - 80\%$ of the cases for R, S, T and I scores.

Figure 3 shows the results using the new setup with additional spectral features. The results show that the new features produce a better performance in the final classifier, with an error around ±1 in the $80-90\%$ of the cases for R, S, T and I scores. 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 ±1. In these experiments, fluctuations in the order of ±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.

Another relevant aspect of the results is the reduction of severe errors, such as +2, -2, +3 and -3. Fig. 3 shows a smaller proportion of such errors compared to the performance depicted in Fig. 2. This improvement is more significant for Strain (T), where dark blue (-3) and brown errors (+3) were highly reduced.

Roughness(R), Breathiness(S) and Instability(I) show the best results. However, although Strain(T) shows promising improvements, it is still observable the difficulties mentioned in the introduction: S (Strain) is commonly one of the less reliable qualitative characteristics in the RASATI scale.

Both figures do not show any relevant results for Harshness (A) and Asthenia (A), because the speech disorders database only has few recordings with harshness or asthenia values superior to zero.

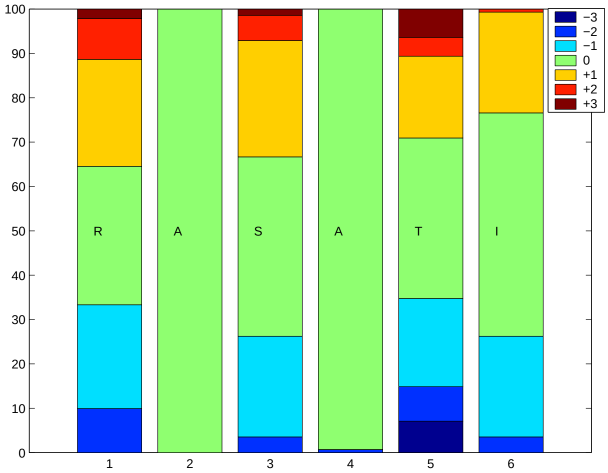

Figure 2: Distribution of the errors for the classification and regression tree approach (original setup)

The analysis of the relevance of the acoustic features to estimate RASATI scores by means of SAV parameters are (ordered by relevance):

- R : HNR, shimr, SPI and srap.
- A : too few cases available.
- S : spectral tilt, srap, spectral flux and jppq5.

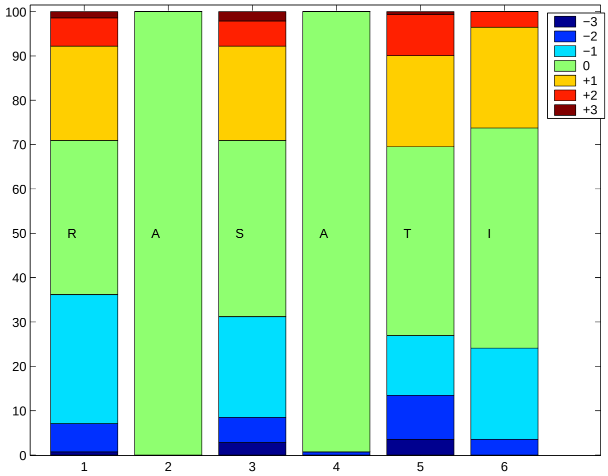

Figure 3: Distribution of the errors for the classification and regression tree approach (new setup)

- A : too few cases available.
- T : shimr, srap, jitta, H2H3 slope, H3H4 slope, HNR and H3H4.
- I : srap, vti, H1H3, H1H4 and spectral tilt.

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 classification and regression trees are a good machine learning technique to estimate subjective evaluations using objective parameters, when using both time and spectral acoustic features. The error in the estimation of the RASATI value is around ±1 in the $80-90\%$ 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 ±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] R. Baken and R. Orlikoff. Clinical measurement of speech and voice. In *Second Edition. San Diego, CA: Singular Publishing Group*, 2000.

[2] R.J. Baken. Clinical measurement of speech and voice. In *Allyn and Bacon, Needham Heights, MA*, 1987.

[3] P. Boersma. Should jitter be measured by peak picking or by waveform matching? In *Folia Phoniatrica et Logopaedica*, pages 305–308, 2009.

[4] N. Chasaide and A. Gobl. Voice source variation. In *The Handbook of Phonetic Sciences. Blackwell Publishers, Oxford*, pages 1–11, 1997.

[5] P.H. Dejonckere. Effect of slightly louder voicing on acoustical measurements in different etiological categories of disphonia. In *Proceedings de Voicedata*, pages 86–91, 1998.

[6] D.D. Deliyski. Acoustic model and evaluation of pathological voice production. In *Proceedings of Eurospeech'93*, pages 1969–1972, 1993.

[7] D.D. Deliyski, H.S. Shaw, and M.K. Evans. Influence of sampling rate on accuracy and reliability of acoustic voice analysis. In *Logopedics, Phoniatrics, Vocology*, volume 30, pages 55–62, 2005.

[8] D.D. Deliyski, H.S. Shaw, and M.K. Evans. Regression tree approach to studying factors influencing acoustic voice analysis. In *Folia Phoniatrica et Logopaedica*, volume 58, pages 274–288, 2006.

[9] T. Dubuisson, T. Dutoit, B. Gosselin, and M. Remacle. On the use of the correlation between acoustic descriptors for the normal/pathological voices discrimination. In *EURASIP Journal on Advances in Signal Processing*, pages 1–19, 2009.

[10] J.I. Godino-Llorente, V. Osma-Ruiz, and N. Saenz-Lechon. Acoustic analysis of voice using wpcvox: a comparative study with multi dimensional voice program. In *European archives of otorhinolaryngology*, volume 265, pages 465–476, 2008.

[11] B. Hammarberg, B. Fritzell, J. Gauffin, J. Sundberg, and L. Wedin. Perceptual and acoustic correlates of abnormal voice qualities. In *Acta Oto-Laryngol.*, volume 90, pages 441–451, 1980.

[12] M. Hirano. Psycho-acoustic evaluation of voice. In *New York: Springer-Verlag*, 1981.

[13] M.P. Karnell, K.D. Hall, and K.L. Landahl. Comparison of fundamental frequency and perturbation measures among three analysis systems. In *Journal of Voice*, volume 9, pages 383–393, 1995.

[14] S.M.R. Pinho and P. Pontes. Musculos intrinsecos da laringe e dinamica vocal serie desvendando os segredos da voz. In *Revinter*, 2008.

[15] J. Revis, A. Giovanni, and F. Wuyts. Comparison of different types of vowel fragments for the evaluation of voice quality. In *Proceedings de Voicedata*, pages 80–85, 1998.

[16] N. Saenz-Lechon, J.I. Godino-Llorente, V. Osma-Ruiz, M. Blanco-Velasco, and F. Cruz-Roldan. Automatic assessment of voice quality according to the GRBAS scale. In *International Conference of the IEEE*, pages 2478–2481, 2006.

[17] I. Smith, P. Ceuppens, and M.S. De Bodt. A comparative study of acoustic voice measurements by means of Dr. Speech and Computarized Speech Lab. In *Journal of Voice*, volume 19, pages 187–196, 2005.

[18] P. Taylor, R. Caley, A.W. Black, and S. King. Edinburgh speech tools library, system documentation edition 1.2. 1999.

[19] E. J. Yamauchi, S. Imaizumi, H. Maruyama, and T. Haji. Perceptual evaluation of pathological voice quality: A comparative analysis between the rasati and grbasi scales. In *Logopedics Phoniatrics Vocology*, volume 35, pages 121–128, 2010.

[20] T. Yoon, X. Zhuang, J. Cole, and M. Hasegawa-Johnson. Voice quality dependent speech recognition. In *Linguistic Patterns in Spontaneous Speech (Language and Linguistics Monograph Series)*, 2008.