Proceedings of the 28th IEEE
EMBS Annual International Conference
New York City, USA, Aug 30-Sept 3, 2006

Automatic Assessment of Voice Quality According to the *GRBAS* Scale

Nicolás Sáenz-Lechón, Juan I. Godino-Llorente, Víctor Osma-Ruiz, Manuel Blanco-Velasco, Fernando Cruz-Roldán

Abstract— **Nowadays, the most extended techniques to measure the voice quality are based on perceptual evaluation by well trained professionals. The *GRBAS* scale is a widely used method for perceptual evaluation of voice quality. The *GRBAS* scale is widely used in Japan and there is increasing interest in both Europe and the United States. However, this technique needs well-trained experts, and is based on the evaluator's expertise, depending a lot on his own psycho-physical state. Furthermore, a great variability in the assessments performed from one evaluator to another is observed. Therefore, an objective method to provide such measurement of voice quality would be very valuable. In this paper, the automatic assessment of voice quality is addressed by means of short-term Mel Cepstral parameters (*MFCC*), and Learning Vector Quantization (*LVQ*) in a pattern recognition stage. Results show that this approach provides acceptable results for this purpose, with accuracy around 65% at the best.**

I. Introduction

There are two dominant approaches in the literature and clinic to evaluate the voice quality: acoustic and perceptual analysis. Acoustic voice analysis is an effective and non-invasive tool for screening and early detection of vocal and voice diseases and an objective support of the diagnostics, proven by many experimental researches. However, these acoustic and objective measurements are usually supplemented with perceptual judgments carried out by otolaryngologists (*ENT*) or speech therapists (*SALT*).

Acoustic analysis is a non-invasive technique based on the digital processing of the speech signal. By means of this processing, a series of temporal or spectral features can be extracted from the voice register, which are supposed to be related with its quality. Classically, a large amount of long-term parameters have been introduced to measure the quality and "degree of normality" of voice records [1]. But up to now, rigorous studies about their application to large populations are lacking in order to fix the values related with normality for every control group regarding sex and age.

On the other hand, *ENT* clinicians can provide some perceptual evaluations regarding the quality of the voice, based on its perceptual and psycho-acoustic features. Perceptual analysis is the most practiced method for evaluation and clinical management of voice disorders. Some common protocols are [2]: (a) the Buffalo Voice Profile Analysis (*BVP*); (b) the Hammarberg scheme; (c) the Vocal Profile Analysis scheme (*VPA*); and (d) the *GRBAS* scale. Combinations of the measures established in the above methods are also used towards the same objective. However, the *GRBAS* scheme is recommended as a standard for practicing voice clinicians [2]. It has been demonstrated that, on the basis of low intra-rater and inter-rater variances, the *GRBAS* scale parameters seem to be the most reliable and relevant perceptual voice quality ratings [3].

There is a need for convergence between both techniques. This paper aims to combine acoustic and perceptual assessment as a whole, in order to provide an objective assessment of voice quality according to a perceptual scale such as *GRBAS*. The underlying idea is to model the *ENT* expertise to train a system that could achieve an automatic evaluation.

The study is focused on organic pathologies affecting the vocal folds, appearing as a modification of the morphology of the excitation (i.e. vocal folds -increasing the distribution of masses-) and producing a more irregular vibration pattern. This group may include pathologies such as polyps, nodules, cysts, sulcus, edemas, carcinoma, etc.

II. The GRBAS scale

Proposed by Hirano [4] 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.

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 *fo*,

This work was supported under the grant: TIC-2003-08956-C02-00 from the Ministry of Science and Technology of Spain.

N. Sáenz-Lechón is with the Dpto. de Ingeniería de Circuitos y Sistemas, Universidad Politécnica de Madrid, Ctra. de Valencia Km. 7, 28031, Madrid, Spain (phone: +34-91-336-5527; fax: +34-91-336-5527; e-mail: nicolas.saenz@upm.es).

J. I. Godino-Llorente is with the Dpto. de Ingeniería de Circuitos y Sistemas, Universidad Politécnica de Madrid

V. Osma-Ruíz is with the Dpto. de Ingeniería de Circuitos y Sistemas, Universidad Politécnica de Madrid.

M. Blanco-Velasco is with the Dpto. de Teoría de la Señal y Comunicaciones, Universidad de Alcalá, Ctra. Madrid-Barcelona, km 33.6, 28871, Alcalá de Henares, Madrid, Spain.

F. Cruz-Roldán is with the Dpto. de Teoría de la Señal y Comunicaciones, Universidad de Alcalá.

1-4244-0033-3/06/$20.00 ©2006 IEEE.

separately perceived acoustic impulses (as in vocal fry), and includes diplophonia and register breaks.

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 used by *SALT* and *ENT* clinicians.

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*).

The GRBAS evaluation is usually carried out based on continuous or conversational speech. However, sometimes is approached by means of sustained vowels, although there are studies demonstrating that the results might differ depending on the material used [6]. 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.

III. The voice Assessment System

The digital speech signal is framed and windowed (40 ms Hamming windows were used throughout the different experiments). Frames have been extracted with a 50% time shift. Framing is followed by an endpoint detector, allowing the separation of voiced and unvoiced segments or silences. The following step is the feature extraction module, needed to reduce the dimensionality and complexity of the patterns. The last module is a Learning Vector Quantization (*LVQ*) classifier. A similar scheme has been used before for the detection of voice impairments with good results [7].

Most of the approaches found in the literature address the automatic assessment of voice by means of long-time signal analysis. In the last recent years, new approaches using short-time analysis of the speech [8] or short-time analysis of electroglotographic (*EGG*) [9] signal can be found. In this research, the automatic assessment of voice quality is carried out by means of the well-known short-term Mel Frequency Cepstral Coefficients (*MFCC*), as an alternative to the mentioned methods. The main advantage of this parameterization scheme is that it does not show dependency on pitch estimations.

In short-term analysis, the automatic assessment is carried out on a frame basis. Every frame is represented by a vector formed by a set of features. For every speaker, the final decision is taken from the most voted class for every feature vector.

A. Computation of recognition features (MFCC)

MFCC parameters [10] are obtained calculating the discrete cosine transform (*DCT*) over the logarithm of the energy in several frequency bands as shown in (1):

$$c_m = \sum_{k=1}^{M} \log(S_k)\cdot\cos\left[n\cdot(k-0.5)\cdot\frac{\pi}{M}\right] \qquad (1)$$

where $1 \le m \le L$; *L* being the order of the *MFCC* coefficients, and S_K given by (2).

$$S_k = \sum_{j=0}^{\frac{k}{2}-1} W_k(j)\cdot X(j) \qquad (2)$$

where $1 \le k \le M$; *M* being the number of the *mel* bands in the *mel* scale, which ranges from 15 to 24. $W_k(j)$ is the triangular weighting function associated with the k_{th} *mel* band in the *mel* scale.

Each band in the frequency domain is bandwidth dependant of the central frequency of the filter. The higher the frequency, the wider is the bandwidth. Such method is based in the human perception system, establishing a logarithmic relationship between the real frequency scale (*Hz*) and the perceptual frequency scale (*mels*) [10].

A better representation showing the dynamic behavior of speech can be obtained by extending the analysis to include information about the temporal derivatives of the parameters among neighboring frames. Both first (*Δ*) and second derivatives (*ΔΔ*) have been used in the present study, computed by means of regression. To introduce temporal order into the parameter representation, let's denote the m_{th} coefficient at time *t* by $c_m(t)$:

$$\frac{\partial c_m(t)}{\partial t} = \Delta c_m(t) \approx \mu\cdot \sum_{k=-K}^{K} k\cdot c_m(t+k) \qquad (3)$$

where μ is an appropriate normalization constant and $(2K+1)$ is the number of frames over which the computation is performed.

The first and second derivatives provide information about the dynamics of the time-variation in *MFCC* parameters. *A priori*, these features have been considered significant because, due to the presence of disorders, there is a lower stability in the speech signal, therefore larger time variations of the parameters may be expected in pathological speech compared with normal speech.

For each time frame *t*, the result of the analysis is a vector of *Q* cepstral coefficients, *Q* delta cepstral coefficients, *Q* delta-delta coefficients, the energy, one delta energy and one delta-delta energy. Thus, the dimensionality of the feature space is $D=3\cdot Q+3$, as the following:

$$\begin{aligned} o(t) = (&E(t), c_1(t), c_2(t), \ldots, c_Q(t), \\ &\Delta E(t), \Delta c_1(t), \Delta c_2(t), \ldots, \Delta c_Q(t), \\ &\Delta\Delta E(t), \Delta\Delta c_1(t), \Delta\Delta c_2(t), \ldots, \Delta\Delta c_Q(t)) \end{aligned} \qquad (4)$$

where $o(t)$ is a feature vector with *D* elements.

B. The LVQ detector

The architecture of the *LVQ* net is composed by an input layer and a Kohonen layer, fully connected between them [11]. The Kohonen layer is partitioned into groups of neurons, each one associated to a class. The number of neurons per class is assumed to be *N*. Once the net has been adequately trained, each node represents one of the *N* prototypes generated by the net. The classification is then performed by choosing the label of the nearest codebook vector selected.

LVQ is an iterative probabilistic gradient method that guarantees asymptotic minimization of the average classification error.

1) *Net size*

The number of neurons in the input layer is adjusted to the number of input parameters. The number of hidden units is a parameter to be tuned during the training phase.

2) *Training and simulation*

When the input vector *o(t)* is presented to the net, each neuron computes the distance between its weight vector and *o(t)*. The winner neuron (i.e. that showing the minimum distance) is positively reinforced (i.e its weight is increased). The others are negatively reinforced (i.e weights are decreased) in a quantity proportional to the learning rate, the distance itself, and other neighborhood criteria.

Training was carried out during 8000 epochs. The codebook vectors of *LVQ* were initialized by finding a group of vectors that satisfied the *K*-nearest neighbor (*KNN*) criterion as suggested in [11]. The *KNN* criterion states that from the *K*-nearest neighbors in the training data the majority must belong to the same class as the tested vector.

Adjustments were made according to the supervised learning law *OLVQ1* (Optimized *LVQ*) [11]. The learning rate is monotonically reduced during learning. *OLVQ1* allows an optimized learning rate defined individually for each node. The feature vectors were used in random order to update the originally random valued weights. The training data were used 5 times during which the learning rate was reduced monotonically from 0.2 to 0.01 and neighborhood radius from 4 to 1. At that point the approximation reached the smallest *SMS* error. Weights were randomly initialized. Input data were normalized and unbiased before they were presented to the network, subtracting the mean and dividing by the standard deviation.

IV. Database

The employed database was recorded to *CDROM* by the Hospital Príncipe de Asturias, from Alcalá de Henares (Madrid). The database contains 648 speakers (433 normal and 215 pathological). The acoustic samples are the sustained phonation of vowel /a/ (3-4 seconds long), and a short sentence ("*es hábil un sólo día*") from patients with normal voices and a wide variety of organic, neurological, traumatic, and psychogenic voice disorders in different stages (from early to mature) (Table 4.1). The speech samples were collected in a controlled environment and sampled with a 50 *kHz* sampling rate and 16 bits of resolution. Every sample has been labeled according to the *GRBAS* scale by three different *ENT* clinicians, in order to minimize the inter-evaluator variability. The final labels were established according to the number of votes.

For every control group (*G-R-B-A-S*), data files were split randomly into two subsets: the first one for training (with 70% of the samples), and the second one (30%) to simulate and validate results, keeping the same proportion for each class (0-1-2-3). The division into training and evaluation datasets was carried out in a file basis (not in a frame basis) in order to check and prevent the system to learn speaker related features. Due to the lack of data, this is a mistake that can be found in some of the systems in the literature. As *GRBAS* evaluation is gender independent, both male and female voices have been mixed together in the training and validation sets.

Fig. 1 shows the distribution of the simulation set according to its labels. The same proportion is shown in the training set. The most frequent label among the five control groups is 0 (healthy voice), followed by 1, 2 and 3. This is a great handicap, because the same amount of patterns of each class is needed to appropriately train the system. So, in order to adjust the number of features to be equal for every control group, the training dataset has been limited. The same amount of frames for every class has been extracted randomly (1000 frames). With this approach, most frequent classes are characterized with vectors from a larger amount of speakers, so a better generalization capability for classes 1 and 2 will be expected.

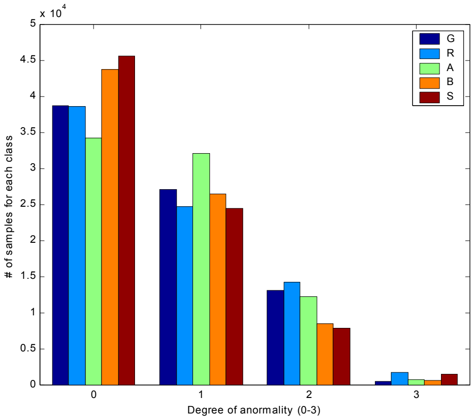

Fig. 1: Distribution of the simulation set (number of frames) according to the five GRBAS classes. Most frequent class is 0, followed by 1, 2, and 3.

Table 4.1.
Table of Pathologies

Diagnosis	N. of cases	Diagnosis	N. of cases
Neurological	15	Carcinoma	6
Chronic laryngitis	20	Paralysis	21
Cyst	20	Reinke's edema	31
Hypofunction	3	Vocal fold trauma	2
Keratosis / leukoplakia	14	Vocal fold polyp	29
Sulcus vocalis	25	Vocal nodules	29
		Normal voice	433
		TOTAL	648

V. Results

Best results were obtained using 15 MFCC parameters and 180 nodes in the LVQ net. The accuracy obtained in the training step rose over 85%, but the results in the simulation step fell dramatically. The results for the simulation step are represented in Table 5.1. There is a big difference between the results for the training and simulation steps, which could be interpreted in terms of the ability of the system to model speaker features non related with the perceived degree of abnormality.

TABLE 5.1
CONFUSION MATRIX REPRESENTING THE BEST OBTAINED RESULTS.
FOR EVERY CLASS AND CONTROL GROUP THE TOTAL NUMBER OF FILES (#) AND THE PERCENTAGE (%) ARE REPORTED

			Predicted results																			
			Feature "G"				***Feature "R"***				***Feature "A"***				***Feature "B"***				***Feature "S"***			
			0	**1**	**2**	**3**	**0**	**1**	**2**	**3**	**0**	**1**	**2**	**3**	**0**	**1**	**2**	**3**	**0**	**1**	**2**	**3**
Perceptual Label	**0**	**#**	49	11	1	0	47	10	4	0	40	19	0	0	45	11	4	0	47	7	3	0
		%	75.4	21.2	12.5	0	72.3	24.4	22.2	0	67.8	34.5	0	0	69.2	26.2	22.2	0	72.3	17.1	16.7	0
	1	**#**	15	32	2	1	16	25	6	0	18	24	2	1	19	22	8	1	17	25	7	1
		%	23.1	61.5	25	50	24.6	61	33.3	0	30.5	43.6	25	20	29.2	52.4	44.4	50	26.2	61	38.9	33.3
	2	**#**	1	9	4	0	2	5	7	3	1	11	6	2	1	8	5	1	1	8	7	2
		%	1.5	17.3	50	0	3.1	12.2	38.9	100	1.69	20	75	40	1.5	19	27.8	50	1.5	19.5	38.9	66.7
	3	**#**	0	0	1	1	0	1	1	0	0	1	0	2	0	1	1	0	0	1	1	0
		%	0	0	12.5	50	0	2.4	5.6	0	0	1.8	0	40	0	2.4	5.6	0	0	2.4	5.6	0
			Efficiency: 68%				*Efficiency: 63%*				*Efficiency: 55%*				*Efficiency: 57%*				*Efficiency: 63%*			

The most accurate representation was obtained for control groups *G* (68%) and *R* (63%). Also more accurate results were obtained for class 0 and 1. The better accuracy for class 0 and 1 can be explained in terms of the size of the dataset for every class in terms of the number of speakers stored in the database. The higher the number of speakers, the better is the expected accuracy of the system because a more precise model is achieved. Unless we had used the same amount of frames for every class, the most frequent classes contains higher inter-speaker variability. Classes 2 and 3 contain a lower amount of speakers, so they are supposed to yield a good ability to model the inter-speaker variability, but not the intra-speaker variability. Such idea should be the explanation to the decrease of the accuracy.

VI. CONCLUSIONS

The proposed scheme can be used for the assessment of voice quality. As it was expected, concerning the classification error, the most consistent parameters revealed to be *G,* followed by *R,* due to the fact that these are the easiest to be evaluated by a human expert, and are supposed to be labeled more accurately.

The lack of data in class 3 makes the efficiency for that class to fall under the efficiency of the other classes. The generalization ability of the system is better for classes 1 and 2 because the vectors used to train classes 1 and 2 have larger inter-speaker variability.

The modest scores could be due whether to the ability to discriminate of the *MFCC* features, or to the *LVQ* algorithm that is not able to separate the prototypes correctly. However, it can be seen that most of the times, the classifier misses with the nearest class. When interpreting these scores it has to be kept in mind that the labeling was made by perceptual evaluation, and sometimes the experts do not agree on the evaluation of a voice sample. It is well known that there is intra and inter-evaluator variability, due to the fact that the judgment depends a lot on their own expertise and subjective criteria about how a normal voice should be.

Despite of the modest scores, this system is able to provide an objective approach to the assessment of voice quality. For the future work, it should be tested with a larger database, especially for classes 2 and 3, to improve the accuracy of the system, and it has to be tested using running speech. Only a small tuning should be required in the endpoint detector to avoid not only silences, but also unvoiced frames.

VII. REFERENCES

[1] Baken, R. J. and Orlikoff, R., *Clinical measurement of speech and voice*, 2 ed., Singular Publishing Group, 2000.

[2] Carding, P., Carlson, E., Epstein, R., Mathieson, L., and Shewell, C., "Formal perceptual evaluation of voice quality in the United Kingdom," *Logopedics & Phoniatrics Vocology*, vol. 25, no. 3, pp. 133-138, 2000.

[3] Dejonckere, P. H., Obbens, C., de Moor, G. M., and Wienke, G. H., "Perceptual evaluation of disphonia: reliability and relevance," *Folia Phoniatrica*, vol. 45, pp. 76-83, 1993.

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

[5] Dejonckere, P. H., "Effect of slightly louder voicing on acoustical measurements in different etiological categories of disphonia," in *Proceedings de Voicedata'98*, pp. 86-91, 1998.

[6] Revis, J., Giovanni, A., and Wuyts, F., "Comparison of different types of vowel fragments for the evaluation of voice quality," in *Proceedings of Voicedata'98*, pp. 80-85, 1998.

[7] Godino-Llorente, J. I. and Gómez-Vilda, P., "Automatic detection of voice impairments by means of short-term cepstral parameters and neural network based detectors," *IEEE Transactions on Biomedical Engineering*, vol. in press, 2003.

[8] Cairns, D., Hansen, J. H., and Riski, J., "A Noninvasive Technique for Detecting Hypernasal Speech Using a Nonlinear Operator," *IEEE Trans. on Biomedical Engineering*, vol. 43, no. 1, pp. 33-45, 1996.

[9] Childers, D. G. and Sung-Bae, K., "Detection of laryngeal function using speech and electroglottographic data," *IEEE Transactions on Biomedical Engineering*, vol. 39, no. 1, pp. 19-25, Jan.1992.

[10] Deller, J. R., Proakis, J. G., and Hansen, J., *Discrete-Time Processing of Speech Signals*, New York: McMillan, 1993.

[11] Kohonen, T., *Self-Organising Maps*, Berlin: Springer-Verlag, 1997.