Complementary approaches for voice disorder assessment

JF Bonastre [1], *C. Fredouille*[1], *A. Ghio*[2], *A. Giovanni*[3], *G. Pouchoulin*[1], *J. Révis*[3], *B. Teston*[2], *P. Yu* [3]

[1] LIA, Avignon University, France ; [2] LPL, CNRS, Aix-Marseille University (France)
[3] LAPEC, Aix-Marseille University (France)

corinne.fredouille@lia.univ-avignon.fr, alain.ghio@lpl.univ-aix.fr, agiovann@ap-hm.fr

Abstract

This paper describes two comparative studies of voice quality assessment based on complementary approaches. The first study was undertaken on 449 speakers (including 391 dysphonic patients) whose voice quality was evaluated in parallel by a perceptual judgment and objective measurements on acoustic and aerodynamic data. Results showed that a non-linear combination of 7 parameters allowed the classification of 82% voice samples in the same grade as the jury. The second study relates to the adaptation of Automatic Speaker Recognition (ASR) techniques to pathological voice assessment. The system designed for this particular task relies on a GMM based approach, which is the state-of-the-art for ASR. Experiments conducted on 80 female voices provide promising results, underlining the interest of such an approach. We benefit from the multiplicity of theses techniques to evaluate the methodological situation which points fundamental differences between these complementary approaches (bottom-up vs. top-down, global vs. analytic). We also discuss some theoretical aspects about relationship between acoustic measurement and perceptual mechanisms which are often forgotten in the performance race.

1. Introduction

Within the framework of voice disorder assessment, the stage of the evaluation aims at allowing comparisons between several pathologies, several patients or several therapeutic approaches. The ways to assess such a disorder are very various, from the perceptual scores or the auto-evaluation questionnaires to the objective instrumental methods. The purpose of this work is to present a methodology of voice analysis combining:

1. perceptual judgment,
2. analytic measurements with multiparametric data (acoustic and aerodynamic)
3. a method based on Automatic Speaker Recognition adapted to dysphonia.

We ended to this association because these methods are complementary in their approaches, their principles, the results that they provide and in their capabilities to inform the failures from/to each other. We are aware that this process can appear unsuited to the clinical routine. However, it belongs to a long-term research project inside which we have been seeking the best performance but also a better understanding of the laryngeal mechanisms and of the relationship between perception and acoustic.

2. Context

Despite the progress in objective voice assessment, perceptual dimension of voice remains the most important factor of the voice quality. It is true that most patients consult because of changes in the sound of the voice, (e.g. hoarseness), and not because they estimate that their jitter is too high. A second point lies in the fact that therapeutic results are judged based on improvement in sound: perception is the first and the most available way to assess voice quality by clinicians. Last but not the least: humans remain the best to decode speech even if machines improve more and more their performance. For all these reasons, perceptual analysis of connected speech is widely used for voice quality assessment [1, 2]. However, this method is largely controversial and demonstrates various drawbacks. First of all, the perceptual judgment has to be performed by an expert jury to increase the reliability of the analysis. Nevertheless, due to the lack of universal assessment scales and other factors like professional background and experience of the experts, the perceptual judgment may involve large intra and inter-variability in the judgments. Besides, a reliable perceptual analysis (with many listeners and several sessions) is very costly in time and human resources and cannot be planned regularly. To cope with these issues, an objective approach, relying on measurement-based analysis, has been proposed.

The objective analysis using a multiparametric approach consists in qualifying and quantifying the vocal dysfunction by analyzing acoustical, aerodynamic and physiological measurements. These measurements may be directly extracted from patient's speech utterance using special devices designed for the recording and the study of many parameters of the speech and voice production. All the investigations made on the objective measurement-based analysis demonstrate the requirement of combining different measurements in order to cope with the multidimensional character of the voice and to increase the reliability of the analysis [3]. Since 1990, we work on and improve a specific equipment and methodology which could take up the challenge to put at the service of the clinician an instrument of measurement and expertise on voice quality which could "replace" easily a jury of experts, impossible to install during a consultation [4, 5]. But like the perceptual judgment, the usual objective analysis has some limitations. First of all, most of the objective analysis is based on sustained vowels, which are not representative of the continuous speech [6]. Besides, the objective analysis often relies on statistical approaches (like linear discriminant analysis, correlation estimation...) applied on the collection of measurements, which may be strongly dependent on the observed patient population in terms of quality and quantity. It means that by changing the clinical cohort, results can appear as not perfectly reliable. Finally, the use of special devices for measurement gathering may be expensive and costly in time. Therefore, these systems are in limited use in routine examination.

This is the reason why, recently, we investigated the adaptation, for dysphonic voice assessment, of automatic techniques largely used in Automatic Speaker Recognition [7]. We have based our work on the assumption that dysphonia may be considered *mutatis mutandis* similarly to a regional accent for instance; ie dysphonia has to be considered not as an homogeneous degradation along the signal but as sporadic phenomena, superposed to the phonetic and linguistic characteristics of utterances. Under this assumption, we propose to characterize these phenomena through an acoustic analysis by deriving automatic tools drawing upon the speech processing domain. Compared with traditional instrumental methods, the originalities of this approach based on a statistical modelling, are:

- its capacity to analyze continuous speech (and not sustained vowels only) near to natural elocution;
- its capacity to process large corpora, permitting to under-

take study on greater scales and to obtain significant statistical data;
- an acoustic, simple and automatic analysis, leading to a simplicity of instrumentation and a low human cost.

Preliminary studies, for instance with a simple cepstral-based analysis coupled with an automatic statistical classification system (derived from the speaker recognition technologies), show very encouraging results for dysphonic voice assessment.
In the following parts, we will present results obtained by the three approaches that we are exploring simultaneously two by two. At the end, we will focus on the general methodology and we will introduce some trails we plan to explore.

3. Voice assessment with perceptual and multiparametric objective analysis

This part describes a comparative study involving an objective voice evaluation using a multiparametric protocol including aerodynamic, linear and nonlinear acoustic parameters and a perceptual voice analysis performed by a jury [8].

3.1. Patients

Voice recordings were retrospectively selected in the data bank of the ENT Department of the Timone University Hospital Center in Marseille. A total of 449 samples were selected including 391 patients with pathological voices (308 women and 141 men) and 58 controls with normal voices (38 women and 20 men). Patients presented a variety of voice disorders typically encountered in clinical practice (96 nodules, 91 polyps, 65 paralytic dysphonia, 55 Reinke's oedema, 27 cysts, 24 functional dysphonia, 19 Dysplasia, 14 Sulcus Glottidis)

3.2. Perceptual assessment

Subjects were instructed to read a standardized text at a comfortable pitch and loudness as naturally as possible. Recorded utterances were evaluated by a jury composed of 4 experienced listeners. Three listening sessions were carried out per week. So, each voice sample was evaluated a total of 12 times. Listeners were instructed to score the G, R, and B components of the GRBAS scale [9] but only the G component was used in this study. They used a visual analogue scale converted then in a numerical scale as detailed in [8]. The conversion scale is weighted to allow more subtle differences for grades 2 and 3 as opposed to grades 1 and 4. Our preliminary findings showed that this method enhanced listener performance and reduced variability between listeners while improving agreement between instrumental and perceptual analysis.

3.3. Multiparametric objective analysis

Objective voice analysis was carried out on a sustained vowel /a/ using the EVA® (SQ-Lab, Aix en Provence, France, [5]) workstation. This system enables the simultaneous measurement of acoustic and aerodynamical parameters making use of a specific mouth-piece including a microphone and a pneumotachograph. Intra oral pressure was measured from a built-in pressure sensor.
The subject was instructed to pronounce three consecutive sustained vowels (/a/), which are analyzed afterwards through Fo (in Hz), intensity (in SPL dB), jitter factor (in %), shimmer (in %), signal ratio (in %), oral airflow (OAF in cm^3/s) and Lyapunov coefficient (Lya). Three measurement sessions were performed for each parameter and reported data per parameter correspond to the mean of the three observation values. Subglottic pressure (ESGP) was estimated with the airway interrupted method using a PVC probe located in the subject's mouth and connected to the pressure sensor device of the EVA® workstation while the subject was instructed to pronounce eight consecutive /pa/ at normal pitch and loudness. Finally, Vocal Range (lowest and highest possible pitch) and Maximum Phonation Time (MPT) were measured.

3.4. Results

Pertinence of measured values and discriminant analysis have been detailed in [8]. Each variable was selected to determine its effect on the overall discrimination ability. Using a "stepwise backward" technique in which all variables are introduced and then withdrawn one by one according to their relative importance in the model, we were able to identify seven (in the female patient population) and six (in the male population) pertinent predictors of the dysphonia severity (grade G). They were Vocal Range, Lya, ESGP, MPT, OAF, SRf>1kHz and Fo for female and Vocal Range, Lya, MPT, ESGP, Fo, SNR for male. Using this method we could determine the "objective grade" for each patient and compare this value with the jury's perceptual staging. Discriminant analysis is reported on table 1.

Table 1. Comparison between objective and perceptual grading

	Obj Group 0	Obj Group 1	Ob Group 2	Obj Group 3	Total	% correct
Grade 0	**67**	5	0	0	72	93%
Grade 1	7	**94**	8	0	109	86%
Grade 2	2	29	**146**	21	198	74%
Grade 3	0	0	7	**61**	68	90%
Total	76	128	161	82	447	82%
% correct	88%	73%	91%	74%		

It has to be noted that 70% of the speakers involved in this study belong to Grade 1 & 2 which makes the task more complex (Grades 1 & 2 are the most difficult grades to classify). Nevertheless, this study conducted on a large corpus of data (449 speakers) draws similar conclusions to some previous experiments: about 80% of matching between perceptual judgment and instrumental measurement seems to be the limit of such an approach. We will discuss in a following part possible explanation of this phenomenon.

4. Voice assessment with perceptual analysis and Automatic Recognition System

4.1. Patients and perceptual assessment

The corpus used in this study is composed of 80 voices of females: 20 voices are normal (G_0), 20 have been perceptually graded 1 (G_1), 20 G_2 and 20 G_3 where G is the global judgment of dysphonia on the GRBAS scale [9]. These perceptual grades were determined by a jury composed of 3 expert listeners. The speech material is obtained by reading the same short text which the duration varies from 13.5 to 77.7 seconds (mean: 18.7s).

4.2. Classification system

The principle retained in this study consists in adapting a classical speaker recognition system to the dysphonic voice classification [7]. A speaker recognition system is a supervised classification system able to differentiate speech signals into classes. In our case, a class corresponds to either a grade of dysphonic patients or normal subjects. The speaker recognition technique used in this study is based on a GMM-based approach, which is the state-of-the-art for speaker recognition. This approach needs three phases: parameterization, model training and classification.
Parameterization consists in extracting information from speech signal. Each signal frame is characterized by 16 MEL frequency cepstral coefficients (MFCC) obtained from 24 filterbank coefficients applied on 20ms Hamming windowed frames at a 10ms frame rate. The first derivatives of the MFCC coefficients are added to the parameter vectors to take into account temporal dynamic of speech.
The class model is learnt using data from a set of speakers who

belong to the same grade. This training phase is based on the EM/ML algorithm, able to extract statistical information for each class. Obviously, the voices used for the class training could not be included in the test set in order to differentiate pathology detection from speaker recognition. During the classification phase, an input signal is presented to the system, compared with the model of each class and assigned to the closest class in terms of similarity measure (likelihood).

4.3. Results

The experiment consists in classifying a given voice following the four classes relating to the G dimension of the GRBAS scale. The confusion matrix (table 2) shows that the confusions involve, in most of the cases, the adjacent grades. These results were published two years ago [7] and recently 80% of correct matching between perceptual judgment and automatic classification has been reached through a deeper acoustical analysis (submitted). These studies demonstrate that dysphonic information may be caught by a GMM-based system, even if very few speech materials are available for the training phase.
Furthermore, it is important to notice that the current approach allows to investigate in smaller units than an entire voice utterance for the decision making. Therefore, future work will focus on the behaviour of the classification system at a segmental level (phoneme or shorter events) in order to evaluate if the dysphonic phenomena are uniformly spread over the speech production or more located at some specific zones in the speech signal.

Table 2: Confusion matrix between automatic classification and perceptual judgment

	Automatic classification					
	0	1	2	3	Total	% correct
Grade 0	**19**	1	0	0	20	95%
Grade 1	3	**14**	2	1	20	70%
Grade 2	2	7	**9**	2	20	45%
Grade 3	0	0	7	**13**	20	65%
Total	24	23	20	19	80	69%

5. Discussion

We started voice quality assessment in a clinical context seventeen years ago and we can discuss now our methodology in retrospect.

5.1. Top-down vs. bottom-up, global vs. analytic approaches

Analytical instrumental evaluations such as EVA[5] have been designed originally in order to provide solution in the form of one or more measures to a well-defined physio-pathological question. Let us take the case of laryngeal paralysis. The immobility of the vocal cords induces a large glottal leakage and may be treated by medialization (eg: Goretex). Question: has surgery reduced the leakage suitably and how much? Subsidiary question: in a functional point of view, should this particular reparative surgery be preferred to another (eg: collagen injection, thyroplasty). To measure the air leakage, the best device remains a flow meter. Of course, it is possible to measure the correlation between the amount of expired air and the acoustic noise of the air leakage during phonation. However, the method is still indirect and tortuous in its principle. On the other hand, a suitable measure of the air flow before and after surgery provides directly an estimate of the glottis closure during phonation and a measurement of the impact of the surgical act. As another example, a patient is presumed to speak with vocal abuse, using excessive pulmonary energy, and therefore requesting vocal cords to work in an immoderate manner. The estimated sub-glottal pressure (ESGP) is measured and compared: is the value around the normality level (7hPa) or does it reach immoderate levels (15hPa). If the vocal abuse is effective and the patient is under speech therapy, it will be interesting to check the ESGP periodically in order to observe a decrease of its value. Other examples could be provided with acoustic measures. In such an approach, the methodology is clearly ***analytical*** and ***top-down***: a clear question, one or more measures related to the question, a precise answer.
The variability issues related to the perceptual evaluation lead clinicians to turn towards objective approaches as a substitute of perceptual judgment. So, top-down analytical instrumental methods have been investigated by clinicians as a methodology, which can be qualified as bottom-up, global and blind: ***blind*** or opaque methodology since clinicians have requested that the measurement device provides measures permitting to classify patients according to grades of dysphonia severity (0, 1, 2 or 3) without clearly defining to what a voice assessed as grade 1 or 2 corresponds from a physical or physiological point of view. ***Global*** because such is the process used in the routine clinical examination: anamnesis, visual observation of the cords, global perceptual hearing (only the grade G(lobal) of Hirano's GRBAS scale) can be considered as global analysis. ***Bottom-up*** because most of the studies related to dysphonia assessment are based on a set of various measures (data-driven) in order to make rising possible clusters. Nevertheless, a top-down analytical approach applied for global, blind and bottom-up purposes can only be limited, which may explain reservations expressed by clinicians regarding the classical instrumental approaches.
This methodological point can be compared with changes which occurred during the ‘90s in the automatic speech recognition domain. At that time, researchers worked on analytical top-down systems like expert systems, which had been rapidly overcome by statistical modelling-based approaches. The latter can be qualified as global, blind, bottom-up and data-driven systems. Indeed, if phoneticians are able to measure, foresee and explain co-articulation mechanisms on /s/ or /z/ in the word "suzie", they are not able to process and to analyse, from an exhaustive manner, continuous speech, which exhibits highly variable forms. From an identical manner, even if classical instrumental devices perform quite well, remaining very useful in an analytical approach to provide a precise measure to a clear and restricted issue related to a speech dysfunction, they do not bring a sufficient, robust, reliable and reproducible solution to the general issue of global dysphonia assessment and its multiple variants.
For these reasons, we turned towards stochastic approaches which have been proven successful for recognition tasks (identification, verification, classification …) on continuous speech. We can consider that these techniques are similar, in a methodological point of view, to dysphonia perceptual assessment because they are bottom-up, global and blind. They have the capability to integrate in a statistical model a huge amount of "undefined" information. On the other hand, a main drawback may be pointed out, especially for the clinicians: these methods work classically like a black-box, for which it may be difficult to understand explicitly or to explain precisely the recognition mechanisms. This is the reason why we try now to interpret the results of the automatic recognition process (decoding, classification or identification) in a clinical and phonetic perspective. Here, the objective is to provide physiological and/or linguistic justifications to any "output" (decision, reliability of parameters used as inputs of the automatic system ...), yielded by the automatic system in order to better understand the dysphonia phenomena in speech production and finally to facilitate the adoption of the automatic system by the clinical community.

5.2. Relationship between perception and physics

A second point that we can discuss is the general relationship between perception and physics. It is well known that the perception mechanisms are not linear functions of the real world.

The projection of the physical parameters on the perceptual space is complex. We can list the logarithmic link between loudness and amplitude, the isosonic curves which can be considered as a model taking into account the complex relation between perception of loudness and signal amplitude+ frequencies [10]. We can also mention Bark or Mel scales which are complex mathematical transformations, including spectral integration with critical bands in order to obtain a spectral analysis similar to the one performed by the human perceptual system.
In [3], objectives relating to a direct relationship between acoustic and perception are highlighted by the authors. If finally authors obtain concordant classification between acoustic and perceptual analysis only in 56% of cases, it can be explained by several facts. First of all, relationship in [3] is a very simple linear regression equation. We have seen previously that very often, the projection of acoustic on perceptual space is highly non linear. It means that if we hope for improving our knowledge and performance, it will be necessary to model data with non-linear functions. Automatic Recognition Systems are well suitable for this task. Another method could be the conversion of physical data in order to obtain a linear metric in the perceptual space. SPL dB conversion is one famous example where perception = Log(excitation).
To conclude this paragraph, we should keep in mind that measurement quantifies differences in signal but not necessary in perception. Some information can be picked up by instrument without perceptual relevance. On the other hand, no significant difference on measures can have a large perceptual impact.

5.3. Is GRBAS a good metric space?

GRBAS [9] is the most used scale in perceptual assessment. It is globally accepted that **B**reathiness is linked to the impression of air leakage extent through glottis and **R**oughness can be defined by the impression of vocal fold vibration irregularity. These two dimensions are globally well defined, are linked to physiopathology (B ⇔non closure, R ⇔ instability) and can be considered as globally independent. **A**sthenia is linked to weakness or lack of power whereas **S**train is the impression of a hyperfunctional state of phonation. First of all, these "definitions" are vague, which explains the unreliability of the observed results and the reserve of clinicians. Secondly, these two dimensions can be considered as members of the same axis with opposite valences: A ⇔Hypo, S ⇔Hyper. This H&H functional state of phonation has some similarities with well-known theories of Hypo&Hyper speech[11]. Finally, according to us, the grade G can be considered not as another independent axis but as the consequence of the other dimensions. From a geometrical point of view, B, R and A+S can be assimilated as the three axes of a metric space (Fig. 1). Patients are located though 3 coordinates in this "perceptual" space and finally, G can be referred to as a scalar distance. But with the same distance, which means the same Grade, the localisation in the perceptual space but also the functional state can be very different (see G' and G" in Fig.1). We think that all these badly defined considerations take part in the variability observed.

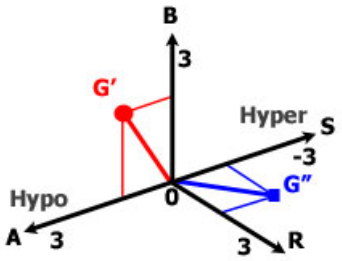

Figure 1: *GRBAS metric space.*

Moreover, a patient evaluated as R0;B2;A1 (Fig.1, G') will have globally the same G=1 than another one quoted R2;B0; S1 (Fig.1, G"). But from a point of view of physiopathological and also physical measures, these two patients are very different. We can easily understand why physical measures cannot match perfectly the Global Grade.

5.4. A fuzzy perceptual space ?

The subjectivity of the perceptual analysis is not really a problem: the practices on intelligibility show that human remains powerful in discrimination or identification tasks. The only necessities are well-defined instructions and shared references. However, the problem of the perceptual dysphonia assessment is that these conditions are not really met. Nobody is able to describe in a clear way what is a G1 or G2 voice. There is no consensual referent as phonemes or lexical units have to be in the case of intelligibility measurements. As long as this formalization will remain vague, the results based on perceptual analysis will only be fuzzy.

6. Conclusions

By these experiments, our final goal is of course to provide reliable methodology and tools for voice quality assessment. But the identification of acoustic-perceptual relationships and the generation of a comprehensive model of voice quality is probably a good way to achieve this aim.

7. References

[1] Dejonckere P, Bradley P, Clemente P, Cornut G, Crevier-Buchman L, Friedrich G, Van de Heyning P, Remacle M, Woisard V. "A basic protocol for functional assessment of voice pathology, especially for investigating the efficacy of (phonosurgical) treatments and evaluating new assessment techniques.", Eur Arch Otorhinolar, 258:77-82, 2001

[2] Revis J, Giovanni A, Wuyts FL, Triglia JM., "Comparison of different voice samples for perceptual analysis", Folia Phoniatr Logop., 51:108-116, 1999

[3] Wuyts FL, De Bodt MS, Molenberghs G, Remarcle M, Heylen L, Millet B, Lierde KV, Raes J, Van de Heyning PH., "The dysphonia severity index: An objective measure of vocal quality based on a multiparameter approch", J Speech Hear Res.,43:796-809, 2000.

[4] Giovanni A, Robert D, Estublier N, Teston B, Zanaret M, Cannoni M., "Objective evaluation of dysphonia: Preliminary results of a device allowing simultaneous acoustic and aerodynamic measures.", Folia Phoniatr Logop., 48:175-185, 1996.

[5] Teston B., Galindo, B. "A diagnosis of rehabilitation aid workstation for speech and voice pathologies", Proc. European Conference on Speech Communication and Technology (Eurospeech), pp. 1883– 1886, 1995

[6] Parsa, V., Jamieson, D. G. "Acoustic discrimination of pathological voice: sustained vowels versus continuous speech", J Speech Hear Res , 44 : 327–339, 2001

[7] Fredouille, C., Pouchoulin, G., Bonastre, J.-F., Azzarello, M., Govanni, A., Ghio, A., "Application of Automatic Speaker Recognition techniques to pathological voice assessment (dysphonia)", Proc. Eurospeech, Lisboa, ISCA, p. 149-152, 2005

[8] Yu P., Garrel R., Nicollas R., Ouaknine M., Giovanni A., "Objective voice analysis in dysphonic patients. New data including non linear measurements", Folia Phoniatrica et Logopaedica, 59:20-30, 2007.

[9] Hirano M. Clinical Examination of Voice. Wien, Springer Verlag, , 1981

[10] Fletcher H.F, Munson W.A "Loudness, its definition, measurement and calculation", J. Acoust. Soc. Am 5 : 82-108, 1933

[11] Lindblom B., "Explaining Phonetic Variation: A Sketch Of The H&h Theory". In Hardcastle, Marchal, Speech Production And Speech Modelling, 1990, pp. 403-439