A METHODOLOGY TO EVALUATE PATHOLOGICAL VOICE DETECTION SYSTEMS

Nicolás Sáenz-Lechón[1], Juan I. Godino-Llorente[2], Víctor Osma-Ruiz[2], Pedro Gómez-Vilda[3], Santiago Aguilera-Navarro[1]

[1] Dept. Tecnología Fotónica, E.T.S.I. Telecomunicación, Universidad Politécnica de Madrid,
Ciudad Universitaria, 28040 Madrid (Spain)
[2] Dept. Ingeniería de Circuitos y Sistemas, Universidad Politécnica de Madrid, Spain.
[3] Dept. Arquitectura y Tecnología de Sistemas Informáticos, Universidad Politécnica de Madrid, Spain.

***Abstract:* This paper describes some methodological issues to be considered when designing systems for automatic detection of voice pathology, in order to allow comparisons with previous or future experiments.**

The proposed methodology is built around Kay Elemetrics voice disorders database, which is the only one commercially available. Discussion about key points on this database is included.

Any experiment should have a cross-validation strategy, and results should supply, along with the final confusion matrix, confidence intervals for all measures. Detector performance curves such as DET plots are also considered.

An example of the methodology is provided, with an experiment based on short-term parameters and Multi-layer Perceptrons.

***Keywords:* Voice pathology detection, pathological voice databases, cross-validation, Multi-Layer Perceptrons.**

I. INTRODUCTION

In the last decade there has been a lot of work done on automatic detection and classification of voice pathologies, by means of acoustic analysis, parametric and non parametric feature extraction, automatic pattern recognition or statistical methods. A lot of research groups in speech technology have addressed in some moment these problems. However, there is a lack of uniformity in these approaches that makes very difficult to estate valid conclusions throughout the proposed methods.

As it is impossible to compare results when the experiments are performed with a private database, we have decided to concentrate on works with Kay Elemetrics database [1], which is rather extended. But even when this database was employed in the state of the art, there were many differences in the way the files were chosen and handled. Also, the experiments were carried out with such different criteria, that comparisons were fruitless. We aim to develop a method that allows comparing results from different classifiers and features.

Detection of voice pathology is much related to a speaker verification task, where a candidate sample is compared against two different models (target and impostors vs. normal and pathological). The system must provide a hard decision and a confidence score about to which model belongs the sample. So we have adopted some methodological issues that are usual in speaker verification [2].

The paper is organized as follows: Section II covers the Kay Elemetrics database and discusses some of its particularities. Section III contains an overview of previous work on pathological voice detection using this database. Sections IV and V present the proposed methodology and describe a simple experiment of detection based upon it. Finally, Section VI presents discussion and conclusions.

II. KAY'S DATABASE OVERVIEW

Kay Elemetrics database [1] was delivered in 1994. It was recorded by the MEEI Voice and Speech Lab. and in Kay Elemetrics. It contains recordings of sustained phonation of vowel /a/ (53 normal and 657 pathological) and continuous speech, (53 normal and 661 pathological). For this description we will focus on the former ones.

The database also includes clinical and personal details of the subjects and acoustic analysis data for the recordings, extracted with the *Multi-Dimensional Voice Program* (MDVP). The recordings were performed in matching acoustic conditions, using *Kay Computerized Speech Lab* (CSL). Every subject was asked to produce a sustained phonation of vowel /a/ at a comfortable pitch at loudness for at least 3 seconds. The process was repeated three times for each subject, and a speech pathologist chose the best sample for the database.

Although the database is the most widespread and available of all the voice quality databases, it has some key points that should be carefully taken into account when used for research purposes:

- Not all the pathological patients have a corresponding recording nor diagnose, and there are some patients with more than one recording, from different visits to the clinic. Fig. 1 shows detailed information about the pathological subset of recordings of vowel /a/.
- The files have different sampling frequencies. Normal and a small percentage of pathological files have 50 kHz, whereas most of the pathological ones have 25 kHz. All files should be down-sampled to 25 kHz before further processing.

Models and analysis of vocal emissions for biomedical applications. 4th international workshop. October 29-31, 2005 – Firenze, Italy.
Edited by C. Manfredi. ISBN 88-8453-320-1 (online) © 2005 Firenze University Press

- Normal and pathological recordings were made at different locations, assumedly under the same acoustic conditions, but there's no guarantee that this fact has no influence in an automatic detection system.
- Normal subjects were not clinically evaluated, although according to [3], none of them had "complaints or history of voice disorders".
- The files are already edited to include only the stable part of them. Several studies [4] consider that onset and offset parts of the phonation contain more acoustic information than stable parts.
- Normal and pathological files have different lengths, maybe due to the fact that is difficult for some pathological subjects to phonate for a long time. When training automatic models, one has to assure that the length is not used as a parameter for discriminating between classes.
- There is only one phonation per patient. Sometimes is useful to dispose of several samples of the same vowel to model intra-speaker variability or samples of different vowels [5].
- There are a heterogeneous number of pathologies in the database, probably because they were included as they were captured in the clinical practice.
- There are a lot of files labelled with several diagnoses, pertaining sometimes to different categories (e.g. physical and neuromuscular). According to [6], the only mutually exclusive possible categorization is at the highest level (i.e. "normal" and "pathological").
- There are a scarce number of normal recordings, compared to the number of pathological ones. This is a problem for training supervised pattern recognition systems, which work best with large amounts of data and well balanced between the different classes.
- There is no perceptual evaluation of the recordings, which would be very useful for research purposes. For this matter, there should be a similar number of recordings of each perceptual rank.
- There are no video recordings (stroboscopy, endoscopy). The importance of this kind of material is highlighted in [7].
- There are no electroglottographic data with the voice registers. EGG signals have demonstrated to be an important complement for acoustic analysis and detection of pathology [8;9].

	# Visits	# Patients
Pathological data	720	617
With audio recording	657	566
Without diagnosis	306	253
Diagnosis "normal"	6	6
Remainder files	345	307

Fig. 1: Pathological recordings of vowel /a/ in Kay database.

III. PATHOLOGICAL VOICE DETECTION

This section presents an overview of previous works in the literature using Kay Elemetrics database. The objective here is to concentrate on the way they handle the database and how they design and evaluate the results of the experiments.

In [10], Qi and Hillman employed 48 voices from Kay to test an algorithm to compute a harmonics to noise ratio (HNR) in the spectral domain. They employed some of the original files, not publicly available, before being edited.

In 1998, Cheol-Woo *et al.* [11] proposed two novelty measures, based on the wavelet transform, and compared their discriminative power against some MDVP features.

In her paper of 1998, Wester [12] compared linear regression techniques and hidden Markov models to detect voice pathologies. She employed 36 normal and 607 pathological voices from the running speech files. Some HNR-based features were extracted by acoustic analysis every 10 ms. 80% of the data were used to train the system and the rest were for testing. The word "sunlight" was segmented from each file, and perceptually evaluated by two expert listeners. Results were favourable to HMMs yielding best results of nearly 65% of correct classification rate.

Parsa and Jamieson, in 2000 [3] broached the detection task based on 6 different noise measurements. They employed 53 normal and 173 pathological voices, enumerated in an appendix. All files were down-sampled at 25 kHz, were chosen to have a diagnosis and the age distributions of both groups were similar. They only used the first second in each file. Discrimination results were obtained comparing the histograms of the two classes and ROC curves were employed to compare them. They yield a best accuracy of 98.7%.

Hadjitodorov and Mitev in 2002 [13] describe a system or acoustic analysis of voice, which also allows the automatic detection of pathology, using jitter, shimmer and noise measures. Classification is achieved by means of Linear Discriminant Analysis (LDA) and Nearest Neighbours clustering. They employed 106 normal ("two phonations by each non pathological speaker") and 638 pathological files. The total accuracy of the system was 92.7%.

Dibazar and Narayanan [14] presented some of the best results in pathology detection with this database. They used all the files in the database, along with MDVP parameters, and short-term MFCCs and F0. They classified the voices with HMMs, to achieve a best accuracy of 98.3%, though they don't give many methodological details due to the great amount of experiments broached.

Maguire *et al.*, 2003 [15] propose a pathology detector, based on sustained phonation, combining long-term acoustic, spectral and *cepstral* parameters. They used 58 normal and 573 pathological voices. The classifier was LDA with a 10 folds cross-validation strategy. They achieved 87.16% accuracy with a subset of the MDVP parameters.

Godino *et al.* have several papers using Kay's database. In [16] they employed 53 normal files and 82 pathological files, the latter chosen randomly among the whole database. All files were down-sampled to 25 kHz. The files were short-term parameterised using MFCCs and their derivatives, and the detector system was based on neural networks (MLP and LVQ). The training test was composed with 70% of the files from each class. Results were presented with confusion matrices, providing confidence intervals for the measurements.

Moran *et al.* [6] presented a telephone system for detecting voice pathologies, with the same data and classifying scheme as

[15]. They used 36 short-term parameters based on jitter, shimmer and noise measures. The system yielded 89.1% accuracy for the original data and 74.15% for simulated telephone data.

Marinaki *et al*. [17] implemented a system to distinguish between 21 normal voices and 42 voices with two different pathologies (vocal fold paralysis and edema). Patients had also others pathologies. They use short-term LPC parameters, Principal Components and LDA to classify the voices. Results yielded nearly 85% of accuracy and were presented through ROC curves.

Although all these works represent novel contributions to pathological voice detection or voice quality assessment, using the same database also, their achievements and conclusions are not easily comparable, due to a lack of uniformity when computing and presenting the results.

IV. METHODOLOGY

Having in mind all of the considerations presented in the previous sections, we aimed to develop a fixed methodology for designing experiments to detect pathological voices from normal ones. This method should allow comparisons between different experiments, in order to outline the benefits of each approach.

The first thing to fix is the database. We decided to use Kay Elemetrics', due to its availability. We have considered only a subset of all the possible files, 53 normal and 173 pathological voices, according to [3]. Features sex and age are uniformly distributed between the two classes.

Files are arranged in two sets, one for training and one for testing and validating the results. We have chosen a 70%-30% split for these sets. Feature extraction from the files is accomplished after these sets are built.

Once the system is trained, the test set is employed to estimate the performance of the detector. The final results are presented through confusion matrices (Fig. 2), where we define the next measures: *True positive* (TP) is the ratio between pathological files correctly classified and the total number of pathological voices. *False negative* (FN) is the ratio between pathological files wrongly classified and the total number of pathological files. *True negative* (TN) is the ratio between normal files correctly classified and the total number of normal files. *False positive* (FP) is the ratio between normal files wrongly classified and the total number of normal files. The final accuracy of the system is the sum of TP and TN.

		Actual diagnosis	
		Pathological	Normal
Detector's decision	Pathological	TP	FP
	Normal	FN	TN

Fig. 2: Typical aspect of a confusion matrix. TP, FP, FN and TN stand for True Positive, False Positive, False Negative and True Negative respectively. See text for definitions.

We have adopted a cross-validation scheme, namely the *bootstrap* method [18; chapter 9] to assess the generalization of the model. Each experiment is repeated N times, with a different test set, randomly chosen from the whole set of files. The final results are averaged across these repetitions, and confidence intervals are computed using the standard deviation of the measures.

When we use short-term parameters, such as MFCCs, accuracies for both frames and files are presented.

During the system testing, a score representing the likelihood of the input vector for belonging to the desired class (i.e. pathological voice) is produced. These scores are compared to a threshold value in order to compute the confusion matrix. If we move this threshold we obtain a set of possible operating points for the system, which can be represented through a *Detector Error Tradeoff* (DET) plot [19], widely used in speaker verification. In this plot, the false positives are plotted against the false negatives, for different threshold values (Fig. 3). Another choice is to represent the false positives in terms of the true positives in a *Receiver Operating Characteristi*c (ROC) [20].

V. AN EXAMPLE DETECTOR

The goal of the following experiment is not to improve the results of previous works in the state of the art, but to illustrate the proposed methodology with a brief example. We have designed an automatic system based on 18 short-term MFCCs parameters, following [16], using 20 ms windows with 50% overlapping. The detector is a basic MLP with a hidden layer of 12 neurons. Learning is carried out by backpropagation algorithm with momentum [21, chapter 6]. The input layer has as many inputs as MFCC parameters and the output layer has two neurons.

We repeat the experiment 10 times, combining the files detailed in [3] in the training and test sets randomly. Fig. 3 shows the mean and standard deviation values of the confusion matrix.

		Actual diagnosis	
		Pathological	Normal
Detector's decision	Pathological	91.36±5.34	16.72±5.02
	Normal	8.64±5.34	83.28±5.02

Fig. 3: Results of the classification (in %) given in a frame basis (mean ± std dev).

The total accuracy of the system is 87.49%±2.80. The accuracy on file basis (percentage of recordings correctly classified) is 88.97%±4.12. The DET plot on Fig. 4 shows the overall performance of the detector, the chosen point of operation (marked with a star) and the point of minimum error rate (small circle). The DET is drawn from the scores obtained with the 10 test sets.

VI. CONCLUSIONS

The only way to improve and to profit from others works is to have objective means to measure the efficiency of different approaches. We have described a set of requirements that a detector of voice pathologies should meet to allow comparisons between systems.

As far as we know, there were no previous works in the literature addressing these issues. We intend to continue the research in pathological voice detection and classification using the presented methodology.

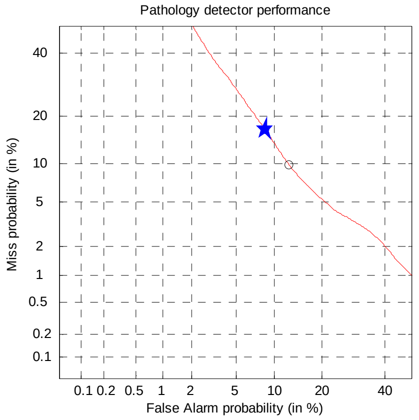

Fig. 4: DET plot for the designed detector.

VII. ACKNOWLEDGEMENTS

This work was supported under the grants: TIC-2003-08956-C02-00 and TIC-2002-0273 from the *Ministry of Science and Technology* and AP2001-1278 from the *Ministry of Education* of Spain.

REFERENCES

[1] Massachusetts Eye and Ear Infirmary, *Voice Disorders Database*, *Version.1.03* [CD-ROM], Lincoln Park, NJ: Kay Elemetrics Corp, 1994.

[2] Campbell, J. P., "Speaker recognition: a tutorial," *IEEE Proceedings*, vol. 85, no. 9, pp. 1437-1462, Sept.1997.

[3] Parsa, V. and Jamieson, D. G., "Identification of pathological voices using glottal noise measures," *Journal of Speech, Language and Hearing Research*, vol. 43, no. 2, pp. 469-485, Apr.2000.

[4] de Krom, G., "Consistency and reliability of voice quality ratings for different types of speech fragments," *Journal of Speech and Hearing Research*, vol. 37, no. 5, pp. 985-1000, Oct.1994.

[5] Horii, Y., "Jitter and shimmer in sustained vocal fry phonation," *Folia Phoniatrica*, vol. 37, pp. 81-86, 1985.

[6] Reilly, R. B., Moran, R., and Lacy, P. D., "Voice pathology assessment based on a dialogue system and speech analysis," in *Proceedings of the American Association of Artificial Intelligence Fall Symposium on Dialogue Systems for Health Communication*, Washington DC, USA, Nov.2004.

[7] Fröhlich, M., Michaelis, D., and Kruse, E., "Image sequences as necessary supplement to a pathological voice database," in *Proceedings of Voicedata '98*, Utretch, Netherlands, pp. 64-69, Jan.1998.

[8] Ritchings, R. T., McGillion, M. A., and Moore, C. J., "Pathological voice quality assessment using artificial neural networks," *Medical Engineering & Physics*, vol. 24, no. 8, pp. 561-564, 2002.

[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] Qi, Y. and Hillman, R. E., "Temporal and spectral estimations of harmonics-to-noise ratio in human voice signals," *Journal of the Acoustical Society of America*, vol. 102, no. 1, pp. 537-543, 1997.

[11] Cheol-Woo, J. and Dae-Hyun, K., "Analisys of disordered speech signal using wavelet transform," in *Proceedings of ICSLP '98*, Sidney, Australia, 1998.

[12] Wester, M., "Automatic classification of voice quality: comparing regression models and hidden Markov models," in *Proceedings of Voicedata '98*, Utretch, Netherlands, pp. 92-97, Jan.1998.

[13] Hadjitodorov, S. and Mitev, P., "A computer system for acoustic analysis of pathological voices and laryngeal disease screening," *Medical Engineering & Physics*, vol. 24, no. 6, pp. 419-429, July2002.

[14] Dibazar, A. A., Narayanan, S., and Berger, T. W., "Feature analysis for automatic detection of pathological speech," in *Proceedings of the Second Joint EMBS/BMES Conference*, vol. 1, Houston, TX, USA, pp. 182-183, Nov.2002.

[15] Maguire, C., deChazal, P., Reilly, R. B., and Lacy, P. D., "Identification of voice pathology using automated speech analysis," in *Proceedings of MAVEBA 2003*, Florence, Italy, pp. 259-262, Dec.2003.

[16] 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. 51, no. 2, pp. 380-384, Feb.2004.

[17] Marinaki, M., Kotropoulos, C., Pitas, I., and Maglaveras, N., "Automatic detection of vocal fold paralysis and edema," in *Proceedings of ICSLP '04*, Jeju Island, South Korea, Nov.2004.

[18] Duda, R. O., Hart, P. E., and Stork, D. G., *Pattern classification*, 2 ed., Wiley Interscience, 2000.

[19] Martin, A. F., Doddington, G. R., Kamm, T., Ordowski, M., and Przybocki, M. A., "The DET curve in assessment of detection task performance," in *Proceedings of Eurospeech '97*, vol. IV, Rhodes, Crete, pp. 1895-1898, 1997.

[20] Hanley, J. A. and McNeil, B. J., "The meaning and use of the area under a receiver operating characteristic (ROC) curve," *Radiology*, vol. 143, no. 1, pp. 29-36, Apr.1982.

[21] Haykin, S., *Neural networks*, New York: Macmillan, 1994.