Acoustic Analysis of Speech for Detection of Laryngeal Pathologies

Martínez César E., Rufiner Hugo L.

Laboratorio de Cibernética - Facultad de Ingeniería, Bioingeniería
Universidad Nacional de Entre Ríos

Abstract - It is well known that most laryngeal diseases and vocal fold pathologies cause significant changes in speech. Different procedures of clinical application for laryngeal examination exist, being all of them of invasive nature.

In the evaluation of quality of speech, acoustic analysis of normal and pathological voices have become increasingly interesting to researchers in laryngology and speech pathologies because of its nonintrusive nature and its potential for providing quantitative data with reasonable analysis time.

In this article, the implementation of a system for automatic detection of laryngeal pathologies using acoustic analysis of speech in the frequency domain is described. Different processing techniques of speech signal are applied: cepstrum, mel-cepstrum, delta cepstrum and delta mel-cepstrum, and FFT. The obtained data feed to neural networks, which classify the voice patterns. Two types of neural network were examined: a system trained to distinguish between normal and pathological voices (no matter the pathology); and a more complex system, trained to classify normal, bicyclic and rough voice.

High percentages of recognition are obtained, being the cepstral analysis the processing technique that achieves the highest actings. This indicates that this analysis type provides a characterization of the voice in pathological condition in a direct and noninvasive way. The obtained results make promissory the application of this alternative as a support tool for the diagnosis of pathologies of the vocal system.

Keywords - acoustic analysis, laryngeal pathologies, neural networks, speech

I. Introduction

It is well known that the presence of pathologies in the vocal folds causes significant changes in the normal vibratory patterns of they, that which impacts in the resulting quality of the voice production.

The problems in the production of the voice can arise from [1,2]: 1) functional disorder (due to the abuse or wrong use of the anatomical and physiologically intact vocal system), which are corrected by means of voice therapy; or 2) laryngeal pathologies (nodules of vocal folds, polyps, ulcers, carcinomas and paralysis of the laryngeal nerve), which can be corrected by means of voice therapy, surgery and, in some cases, radiotherapy.

Diverse routine procedures exist for examination of the larynx with clinical or investigation purposes, which include flexible and rigid fiberscopic laryngoscopy (examination with a fiber-optic instrument), video stroboscopy (strobe illumination of the larynx, useful for the visualization of movements), electromyography (indirect observation of the functional state of the larynx) and videofluoroscopy (radiographic technique in which the patient ingests mouthfuls of a radio-opaque substance to assess the swallowing function).

In the last years the interest for the acoustic analysis of normal and pathological voices as alternative method of diagnosis has grown. This type of analysis demonstrates advantages on the methods of current examination due to its noninvasive nature and to its potential to provide quantitative data about the clinical state of the functions of the larynx and the vocal tract, with appropriate times of analysis.

In the field of the Automatic Recognition of Pathologies of the Vocal System diverse architectures of artificial neural networks (ANNs) have been used, as likewise mathematical models of the vocal tract and the larynx [3,5].

The ANN is excellent classification system and it specialize in working with noisy, incomplete, overlapped data, etc. The recognition of patterns of pathological voices is a classification task of data that has all these characteristics, making the ANN an attractive alternative to the described approach [6].

II. Materials and Methods

The patterns for training the ANNs were obtained from recordings of people's voices with normal fonation and patients with pathologies of the vocal system. Each signal is a recording of the sustained phonation of a vowel or a vocalic phoneme. The use of a vocal type stimulus has certain advantages. First, the isolated vowels are used in the routine of clinical practice for evaluation of the quality of pathological voices. Second, the objective measures are relatively direct, compared with the continuous speech. Also, they allow an easy and effective separation among normal and pathological voices [3]. The study of the continuous speech is a superior objective and an evident next step. However, first valid results are required based on stimuli of smaller complexity.

0-7803-6465-1/00/$10.00 ©2000 IEEE

The speech signals of normal voice were obtained from the TIMIT continuous speech corpus [7]. From the sentence SA1.WAV, of the original group of sentences of training, the signal portions were extracted that contain the phoneme /aa/. When creating the different sets of patterns, the speakers were selected at random among the dialectical regions DR1 at DR8, in such a way of having represented a wide variety of dialects and not to repeat the patterns for the training of different nets.

Signals of pathological voice were obtained from a library of recordings of voices taken in VA Hospital (West L.A.) by investigators of the Speech Processing and Auditory Perception Laboratory (SPAPL), UCLA. The signals were recorded with a miniature microphone mounted on the head AKG C410, placed to 4 cm of the patient's lips. The signals were gone by a lowpass filter of 8 Khz, digitized directly to 20 Khz and sampled to 10 Khz. A segment of 1 second was extracted of the half portion of each recording [8].

For the purpose of this work, the signals were resampled to 16 Khz, to obtain the same temporal reference that the signals of TIMIT.

The classification of the signals was carried out by the mentioned investigation team, being contained in the following categories: rough and rough-breathy: 11 files, bicyclic (also well-known as diplophony): 8, rough-bicyclic: 1, strained-breathy: 2, and strained-rough: 2.

For the extraction of patterns, a mobile window of 256 samples was used, with overlap of 128 samples. A window of Hamming was applied in each segment, and then the patterns were obtained extracting the first 16 cepstral coefficients [4]. Each pattern was completed with the information of 1 and 0 as it was the activation in the desired outputs of the ANN.

Other used processings were: Cepstra in Mel scale (16 coefficients), Delta Cesptra [4] and Delta Mel Cepstra (both of 32 coefficients) and FFT (128 coefficients extracted).

In figure 1 a segment of signal of normal voice and pathological voice is shown, where the temporal differences of both waves can be appreciated, while in figure 2 the spectra of the same ones are observed. A difference that is appreciated at first sight is the appearance of components of high frequency in the pathological case.

Although the patterns to classify are dynamic, they have stationary nature because the samples were taken of vocalic phonemes pronounced in sustained form, like it was mentioned previously. This makes unnecessary the use of neuronal networks with temporal delays (TDNN) [9,10,11], for what in this work a Multilayer Perceptron (MLP) of one hidden layer is used. To train the MLP the backpropagation algorithm was used [9]. The inputs were normalized for the pattern's dimension in an independent form. The training stopped in the generalization pick measured with regard to the test file.

For each classifier with a given signal processing, the optimal quantity of neurons is looked for in the hidden layer; this is, the architecture of the network that obtains the best results in classification and be not excessively big. This quantity is deduced based on two parameters: 1) patterns percentage not well classified in test, in function of training epoch; and 2) learning index for trained nets: average of successes for each class, among speaking masculine and feminine.

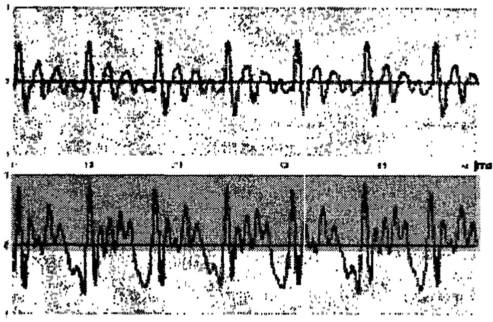

Figure 1: temporal signals corresponding to the vocalic phoneme /aa/. Up: normal voice. Below: pathological voice.

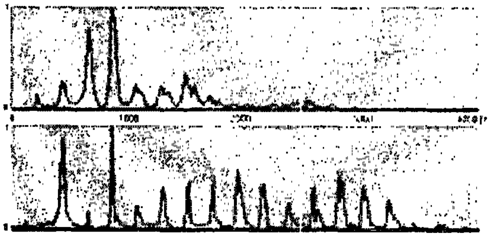

Figure 2:magnitude spectra corresponding to the temporal segments of Figure 1.

For the experiments with Cepstra and Mel Cepstra, the number of neurons in the hidden layer was fixed in 50 after carrying out a serie of experiments which results are shown in the figure 3. In the figure 4 it is shown, by an example of the learning index, the results of the calculation of this index for nets of two classes trained with patterns of Cepstra. For the ANNs trained with patterns that have the delta attached, the number of neurons in the hidden layer was fixed in 80. Because the patterns obtained by means of FFT possess dimension 128, the nets used for their classification possessed different quantity of neurons in the hidden layer that was fixed in 100 for a similar method.

Two types of different ANN were worked: one trained to distinguish among normal and pathological voice (without caring the pathology); and another to distinguish among normal, bicyclic and rough voice.

Once chosen the ideal architecture, each ANN was trained three times with the same patterns, changing the seed of aleatory initialization, being reported the best in the results.

III. Results

In the tables 1 and 2, the results of the experiments carried out for the nets of two and three classes respectively are presented. The percentage of well classified frames for the file of training (TRN) and test (TST) is shown.

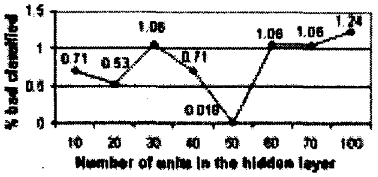

Figure 3: minimum percentage of bad classified patterns during training.

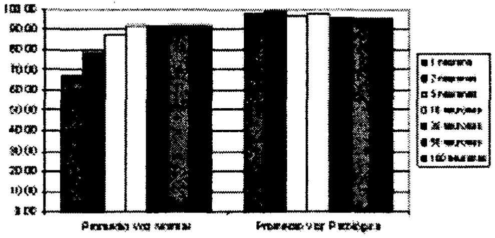

Figure 4: learning index for ANNs of two classes trained with cepstral patterns.

TABLE 1: RESULTS FOR THE TWO CLASSES ANNS

	% WELL CLASSIFIED FRAMES	
	TRN	TST
CEPSTRA	96.30	91.30
MEL CEPSTRA	87.05	83.84
FOURIER	61.00	63.50

TABLE 2: RESULTS FOR THE THREE CLASSES ANNS

	% WELL CLASSIFIED FRAMES	
	TRN	TST
CEPSTRA	91.43	81.91
MEL CEPSTRA	81.74	78.45
DELTA CEPSTRA	92.46	87.73
DELTA MEL CEPSTRA	80.87	77.41
FOURIER	48.37	46.86

In [12] they worked with patterns constituted by parameters obtained from the cepstra (jitter, shimmer and a new parameter designed by the investigators, HNRR -harmonics to noise ratio from residuals–), and two neuronal networks of two classes as the classification system. The first network carry out the classification among normal and pathological voice, and then the pathological patterns are passed through another network that classifies the voice according to the pathology. The results obtained in that work indicate a percentage of correct classification of 90%. They also say it was not possible to distinguish the three classes at such high correctness using a single neural network.

In [3], the investigators use a set of acoustic parameters to screening laryngeal diseases by means of modified SOM classifiers [13]. The overall classification accuracy was 93.5%.

To understand the classification task that was carried out by the neuronal networks, it was realized an analysis of the first 3 formants of the speech signals used in the experiments with ANN of 3 classes. This analysis want to explain the learning facility for nets with minimum architecture (3 to 5 neurons in the hidden layer), and the good results obtained in classification. These values were obtained of the spectral picks of the 20 poles LPC (Linear Predictive Coding), preprocessed with a Hamming window of 1024 samples. The obtained results will be discussed in the next section.

The averages of speakers of both sexes were calculated to emulate the task that it is executed by neural nets, which classify input patterns only for the quality of the voice, without making distinction of the speaker's sex. The following figures show these values in the usual form of representation: the formants planes.

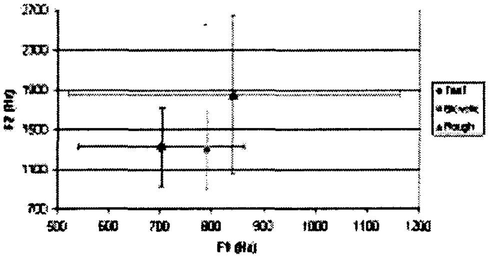

Figure 6: F1-F2 plane for average formants of masculine and feminine voice of each output class.

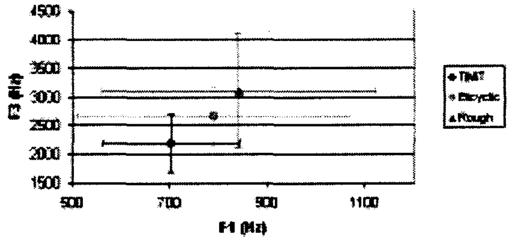

Figure 7: F2-F3 plane for average formants of masculine and feminine voice of each output class.

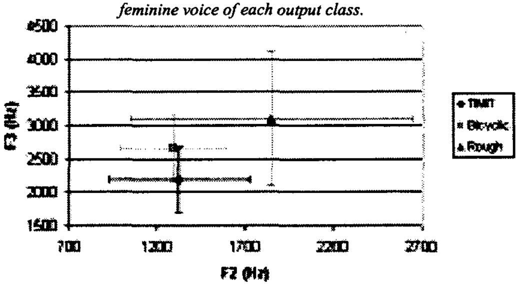

Figure 8: F1-F3 plane for average formants of masculine and feminine voice of each output class.

4. DISCUSSION AND CONCLUSIONS

In this paper an alternative for the automatic diagnosis of laryngeal pathologies is presented, based on the extraction of acoustic characteristics of speech signals and the classification of these patterns by means of static neural networks.

As it can be appreciated of the tables 1 and 2, the cepstral analysis is the processing that achieves the highest performance. This is due to that the information that allows to carry out the distinction among pathologies is in the envelope of the magnitude spectrum of the signals, which is contained in the first cepstral coefficients.

In the case of Mel Cepstra, the integration per bands can affect this information, while in the case of Fourier the increase in the quantity of dimensions of the patterns makes more difficult the task of training the nets and possibly prone to fall in local minima. The delta of each dimension of the patterns calculated by means of cepstra and mel cepstra, did not report improvements as for percentages of successes in classification for networks with the same number of neurons in the hidden layer. This is owed partly, to the fact that the input patterns have double longitude than the initial ones; while, on the other hand, it should be considered that being the input patterns belonging to sustained vocalic sounds, practically static patterns are obtained in the course of time, what subtracts influence to the information from the delta to the classification task. However, this parameter should be taken in consideration when the patterns are calculated for continuous speech.

The separation task of the input patterns into three classes gave as a result a smaller performance due to a bigger difficulty of this task. However, the possibility of training different nets for each pathology exists, what would allow to increase the total number of classes without losing too much precision, and to add pathologies without re-training the system as well.

The diagrams of formant spaces reveal that, for the analyzed parameters, a great separation of the formant averages exists, which is translated into a bigger facility of the neural networks to divide the space of solutions. These results confirm the exposed argument about the easiness of division of the input patterns space with relatively few neurons in the hidden layer, which are those in charge of adding the limits of decision.

It seemed also possible to use only three parameters as input to the classifiers (F1, F2 and F3). Nevertheless, it should be noticed the complexity of the diagrams as for the maximum deviations for speakers of the same class. This complexity increases when adding pathologies to the system, reason why it is justified the use of cepstral patterns that represent the whole spectrum of the speech signal in a small number of coefficients.

The results obtained in classification makes promissory the application of this alternative as a support tool for the diagnosis of pathologies of the vocal system. In addition, it is possible for each medical professional to make his own database with pathologies that are of his interest or whose incidence is bigger in its influence area.

Among the applications to be explored with this tool is the possibility to not only carry out a diagnosis or qualitative analysis, but also a quantitative analysis based on the percentage of well classified frames for a set of patterns of unique class. This would allow, for example, to follow the evolution of some rehabilitation therapy or medication.

A natural step to advance in the study and application of this technique is the use of continuous speech as input to the classification system, instead of sustained vocalic phonemes. This would allow the patient to have more naturalness to record the speech signal, since the continuous speech allows to vary the pitch of the emitted sounds, making more flexible the study. In the classification systems for continuous speech, the architectures of dynamic neuronal networks are involved.

References

[1] Paul W. Flint, Charles W. Cummings, "The John Hopkins Center for Laryngeal and Voice Disorders". [http://www.med.jhu.edu/voice/index.html]. Department of Otolaryngology-Head & Neck Surgery, John Hopkins University. Baltimore, Maryland, August 1997.

[2] James A. Koufman, Gregory N. Postma, "Center for Voice Disorders". [http://www.bgsm.edu/voice]. Wake Forest University, November 13, 1998.

[3] B. Boyanov, S. Hadjitodorov, "Acoustic analysis of pathological voices", IEEE Engineering in Medicine and Biology, pp. 74-82, july/august 1997..

[4] J. R. Deller, J. G. Proakis, J. H. L. Hansen, Discrete-Time Processing of Speech Signals, Macmillan Series. Englewood Cliffs, NJ: Prentice-Hall, 1993.

[5] John H. L. Hansen, "A nonlinear operator-based speech feature analysis method with application to vocal fold pathology assessment", IEEE Transactions on Biomedical Engineering, vol. 45, no. 3, pp. 300-312, 1998.

[6] Judith A. Markowitz, Using speech recognition, Prentice-Hall, NJ, 1996.

[7] Garofolo, Lamel, Fisher, Fiscus, Pallett, Dahlgren, DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus Documentation, National Institute of Standards and Technology, February 1993.

[8] A. Alwan, P. Bangayan, J. Kreiman, and C. Long, "Time and Frequency Synthesis Parameters for Severe Pathological Voice Qualities", Proc. of ICPhS, Stockholm, Sweden, Vol. 2, 250-253, August 1995

[9] Mohamad H. Hassoun, Fundamentals of Artificial Neural Networks, The MIT Press, 1995.

[10] J.L. Elman, "Finding structure in time", Cognitive Science 14 *(1990) 179-211.*

[11] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, K. Lang; "Phoneme Recognition Using Time-Delay Neural Networks". IEEE Trans. ASSP Vol. 37, No 3 (1989).

[12] Cheol-Woo Jo, Dae-Hyun Kim, "Classification of pathological voice into normal/benign/malign state", Proceedings of EuroSpeech 99.

[13] S. Hadjitodorov, B. Boyanov, T. Ivanov and N. Dalakchieva, "Text-independent speaker identification using neural nets and AR-models", *Electronic Letters*, vol. 30, no 11, pp. 838-840, 1994.

E-mail:
César E. Martínez: cesarmart@arnet.com.ar
Hugo L. Rufiner: lrufiner@arcride.edu.ar.
Address for correspondence: Laboratorio de Cibernética-Facultad de Ingeniería (UNER). Casilla de Correo 47 - Sucursal 3 (CP 3100). Paraná (Entre Ríos). ARGENTINA

0-7803-6465-1/00/$10.00 ©2000 IEEE