ByoVoz Automatic Voice Condition Analysis System for the 2018 FEMH Challenge

Julián David Arias-Londoño*, Jorge Andrés Gómez-García†, Laureano Moro-Velázquez‡, Juan Ignacio Godino-Llorente†

* *Dept. of Systems Engineering and Computer Science*
Universidad de Antioquia, Medellín, Colombia
julian.ariasl@udea.edu.co
† *Teoría de la señal y comunicaciones*
Universidad Politécnica de Madrid, Madrid, Spain
jorge.gomez.garcia@upm.es – igodino@ics.upm.es
‡ *Center for Language and Speech Processing*
Johns Hopkins University, Baltimore, USA
laureano@jhu.edu

***Abstract*—This paper presents the methods and results used by the ByoVoz team for the design of an automatic voice condition analysis system, which was submitted to the 2018 Far East Memorial Hospital voice data challenge. The proposed methodology is based on a cascading scheme that firstly discriminates between pathological and normophonic voices, and then identifies the type of disorder. By using diverse feature selection techniques, a subset of complexity, spectral/cepstral and perturbation characteristics were identified for the proposed tasks. Then, several generative classification methodologies based on Gaussian Mixture Models and Gradient Boosting were employed to provide decisions about the input voices in the binary classification, and using one-vs-one classification systems based on Random Forests for the categorization according to the type of disorder. By using a 4-folds cross-validation approach on the training partition a sensitivity=0.93 and specificity=0.74 were obtained. Similarly, an unweighted average recall of 0.63 and an accuracy of 66% was obtained for the identification task. Using the scoring metric proposed in the challenge the final resulting score considering both detection and identification is of 0.77.**

***Index Terms*—Voice pathology detection, voice pathology identification, Gradient boosting, Gaussian mixture models; Random forest**

I. Introduction

Voice impairments arise due to misuse, infections, physiological or psychogenic causes, or due to the presence of other systematic disorders (including neurological), vocal abuse, surgery, trauma, congenital anomalies, irradiation, chemicals affecting vocal folds, etc. [1]. The classical approach to detect voice impairments consists on an instrumental (objective) and perceptual (subjective) evaluation, which are complemented by other types of examinations to determine the existence of a voice disorder and its grade of impairment. The increasing need of improving the diagnosis of voice pathologies has given rise to an emerging field called *Automatic Voice Condition Analysis* (AVCA), that aims at analysing, classifying and quantifying the degree to which a patient is affected by a voice disorder, providing advantages to traditional detection procedures such as objectiveness or non-invasiveness due to the use of speech signals [2]. This analysis is performed using automatic systems that provide objective measurements of the patient's vocal condition, exploiting the close relationship that exists between acoustic features extracted from the speech and voice pathology [3]. This reduces the evaluation time and the cost of diagnosis and treatment, providing extra advantages such as the avoidance of invasive procedures thanks to the employment of speech signals which are easily recorded by inexpensive means [2].

With this in mind, the goal of this paper is to design an AVCA system based on a wide range of features, which is employed for the differentiation of pathological and normophonic states (binary detection task) and for the actual identification of the disorder (categorization of the pathologies). The methodology followed has been delineated to be as much simple as possible, and easily understandable from the physical point of view, condition considered necessary to transfer the system to the clinical setting.

Section II introduces the dataset and the methodology that was followed in this paper. Section III presents the outcomes of experiments performed on the training partition. Finally, section IV introduces the discussions and conclusions.

II. Experimental setup

A. Corpus

A subset of the corpus previously described in [4] has been supplied by the organizers of the challenge. The provided dataset has been divided into a training and testing partition for the purposes of performance evaluation. Voice samples were recorded at the Far Eastern Memorial Hospital (FEMH) in Taiwan. Each register contains samples of the sustained phonation of the vowel /a:/ recorded at a comfortable level of loudness, with a microphone-to-mouth distance of approximately 15-20 cm, using a high-quality microphone (Model: SM58, SHURE, IL), with a digital amplifier (Model: X2u, SHURE) under a

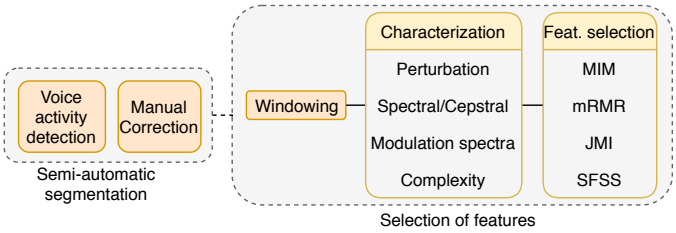

Fig. 1. Procedure for the selection of relevant features for both identification and detection purposes.

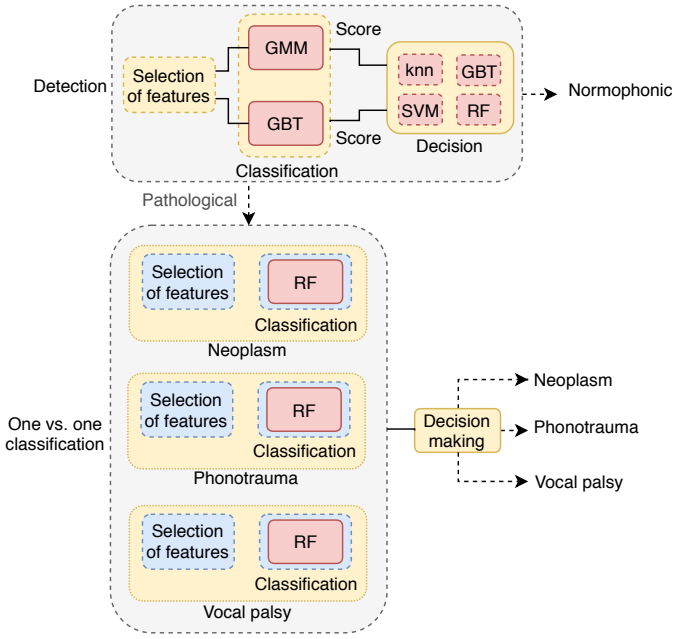

Fig. 2. Methodology for the construction of the detection and identification systems.

background noise level between 40 and 45 dBA. Recordings were sampled at 44100 Hz with 16-bit resolution, and data was stored in an uncompressed .wav format. The training partition of the dataset includes 50 normophonic and 150 dysphonic voices. The pathological voices were grouped into three clusters: *Phonotrauma* (including disorders such as vocal nodules, polyps, and cysts); *glottis neoplasm* and *unilateral vocal paralysis*. Likewise, the testing partition is comprised by 400 unlabelled voice samples belonging to the same categories as in the training partition.

B. Methodology

The system relies on a short-term analysis of the voice recordings and on an estimation of a final score for each speaker. Fig. 1 presents graphically the framework used to estimate the short-term relevant features. The initial feature set considers a wide range of complexity, noise, and spectral/cepstral based parameters, which are pruned according to different feature selection schemes to obtain the most significant feature set. Likewise, a graphic summarizing the proposed detection and identification system is presented in Fig. 2.

As observed in Fig. 1, a preprocessing stage is firstly applied. In this regard, the training and testing partitions have been semi-automatically segmented using the voice activity detection (VAD) algorithm described in [5]. This method relies on the computation of source and filter-based features and the employment of an artificial neural network to distinguish voices from low-amplitude background noise and silences. The resulting samples after having applied the VAD procedure were revised and corrected, as the algorithm is unable to eliminate high-amplitude noise samples or voices masked by background noise. Next, all registers are down-sampled to 20 kHz and max-normalized.

The proposed methodology is based on a short-time analysis of the voice for which a framing and windowing methodology is then followed. Since the optimum length of the window depends on the characteristic that is extracted, the different sets of features -all descriptors of vocal quality- that are considered in this work are presented firstly [2]:

- *Perturbation features* that measure the presence of additive noise resulting from an incomplete glottal closure of the vocal folds, and the presence of modulation noise which is the result of irregularities in the movements of the vocal folds: *Normalised Noise Entropy* (NNE), *Cepstral Harmonics-to-Noise Ratio* (CHNR) and *Glottal-to-Noise Excitation Ratio* (GNE).
- *Spectral and cepstral features* that measure the harmonic components of the voice: *Mel-Frequency Cepstral Coefficients* (MFCC), *Smoothed Cepstral Peak Prominence* (CPPS) and *Low-to-High Frequency Spectral Energy Ratio* (LHr).
- *Features based on modulation spectrum* relying on the computation of the modulation spectrum, to characterize the modulation and acoustic frequencies of input voices: *Modulation Spectrum Homogeneity* (MSH), *Cumulative Intersection Point* (CIL), *Rate of Points above Linear Average* (RALA) and *Modulation Spectrum Percentiles* (MSP) [6], [7].
- *Complexity features* that characterize the dynamics of the system and its structure. These include dynamic invariants such as the *Correlation Dimension* (D2), the *Largest Lyapunov Exponent* (LLE), and the *Recurrence Period Density Entropy* (RPDE); features which measure long-range correlations, such as *Hurst Exponent* (He) and *Detrended Fluctuation Analysis* (DFA); regularity estimators such as *Approximate Entropy* (ApEn), *Sample Entropy* (SampEn), *Modified Sample Entropy* (mSampEn), *Gaussian Kernel Sample Entropy* (GSampEn) and *Fuzzy Entropy* (FuzzyEn); and other entropy/complexity estimators such as the *Permutation Entropy* (PE), the *Lempel-Ziv Complexity* (LZC) and the Shannon (s) and Rényi (r) estimators of the *Markov Chain Entropy* (H_{MC}), *Conditional Hidden Markov Process Entropy* (H_{HMP}) and *Recurrence State Entropy* (H_{RSE}) [8], [9]. Similarly

to the ApEn and mSampEn estimators, which use the correlation sum for two different embedding dimensions, some modifications of the measures H_{MC}, H_{HMP} and H_{RSE}, which consist of averaging the entropy estimations over two different embedding dimensions are also considered. These measures are called *Averaged Markov Chain Entropy* (A_{MC}), *Averaged Conditional Hidden Markov Process Entropy* (A_{HMP}) and *Averaged Recurrence State Entropy* (A_{RSE}).

Hamming windows of 40 ms are employed for the perturbation and spectral/cepstral features to ensure that each frame contains at least three pitch periods. Likewise, windows of 55 ms length are used with the complexity features as suggested in [8]. Finally, for the experiments in the modulation spectrum set, frames of 180 ms are utilized as suggested in [6], [7].

To determine the most significant features three filter selection methodologies were employed: *Maximal Information Maximisation* (MIM), *Minimal Redundancy Maximal Relevance* (mRMR) and *Joint Mutual Information* (JMI) [10]. Additionally, a wrapper feature selection was also considered: *Sequential Floating Feature selection* (SFFS), using a quadratic linear discriminant analysis function for performance evaluation purposes.

As observed in Fig. 2, the proposed system relies on a cascading scheme on which a discrimination between normophonic and dysphonic states is firstly carried out (binary detection task). The actual identification of the pathologies is followed in a further classification stage (categorization of pathologies task).

In this manner, and during the binary detection task, 28 features were chosen according to a consensus of the filter and wrapper feature selection algorithms, having selected those characteristics that were included by any of the 4 considered feature selection criteria. Then, *Gaussian Mixture Model* (GMM) and a *Gradient Boosting Tree* (GBT) (80 trees) binary classifiers are employed for decision making purposes, producing scores obtained in a per-file basis. For the final detection of pathology, the scores obtained from the binary classifiers were used to feed a second classification stage based on a soft voting procedure with 4 classifiers: a GBT using 10 trees, a radial-basis function *Support Vector Machine* (SVM), a *k-nearest neighbor classifier* (knn) considering 5 neighbors and a *Random Forest* (RF) with 20 trees. The final detection decision is then obtained according to the class that provides the largest joint probability among the 4 classifiers. The aim of this combination is to reduce the variance of the detector. Moreover, by merging the outputs of every single detector, a reduction in the bias can also be achieved.

For the identification of pathologies (categorization of pathologies task) a one-vs-one classification scheme was followed, using a RF with 100 trees. This approach was chosen because it allows each binary classifier to be trained with a different subset of features. In this case, different subsets were chosen according to the feature selection procedure described previously, but considering pairs of impairments.

Finally, and for the experiments involving the testing partition, the whole system was retrained 10 times, obtaining each time a decision about the membership of the testing samples. The mode of the repetitions is then used as a final label.

The evaluation metrics defined in the challenge for the detection of pathologies are the sensitivity and specificity, while the *unweighted average recall* (UAR) is employed for the classification task. Additionally, a score assessing the global performance of the system is defined as a weighted combination of sensitivity (40%), specificity (20%), and UAR (40%).

III. Results

Table I lists the features that were chosen for the different binary detection tasks: the normophonic vs. dysphonic detection; and for each of the one-vs-one systems. Likewise, the results of the detection and identification systems are presented in Table III using the metrics specified in the challenge (sensitivity, specificity, UAR, and score). The variability due to the cross-validation is also represented. The resulting confusion matrix of the identification system is shown in Fig. 3.

TABLE I
FEATURES CHOSEN BY THE FEATURE SELECTION ALGORITHMS AND EMPLOYED IN THE DETECTION AND IDENTIFICATION SYSTEMS.

Normophonic vs. Pathological	Neoplasm vs. Phonotrauma	Phonotrauma vs. Vocal Palsy	Neoplasm vs. Vocal Palsy
MFCC(2)	MFCC(1)	MFCC(1)	MFCC(1)
MFCC(3)	MFCC(4)	MFCC(2)	MFCC(3)
MFCC(4)	MFCC(5)	MFCC(5)	MFCC(4)
MFCC(5)	MFCC(6)	MFCC(6)	MFCC(6)
MFCC(7)	MFCC(8)	MFCC(7)	MFCC(7)
MFCC(8)	MFCC(9)	MFCC(8)	MFCC(9)
MFCC(9)	MFCC(11)	MFCC(9)	MFCC(10)
MFCC(10)	MFCC(12)	MFCC(12)	MFCC(11)
MFCC(11)	MFCC(14)	MFCC(13)	MFCC(12)
MFCC(12)	MSP_{75}	CPPS	MFCC(14)
MFCC(13)	CPPS	LHr	LHr
CPPS	LZC	CIL	A_{RSEs}
LHr	RPDE	MSH	A_{HMPr}
GNE	CIL	RALA	He
D_2	CHNR	DFA	CIL
DFA			
PE			
H_{MCs}			
H_{MCr}			
A_{MCs}			
A_{MCr}			
A_{MMPs}			
A_{HMPr}			
A_{RSEs}			
LZC			
MSH			
CIL			
RALA			
MSP_{95}			

IV. Discussions and conclusions

This paper has presented the methodology that has been followed by the ByoVoz team in the 2018 FEMH challenge. The proposed system relies in a cascading scheme that firstly differentiates between normophonic and dysphonic voices, and

TABLE II
RESULTS OF THE DETECTION

Detection	Sensitivity	0.93 ± 0.05
	Specificity	0.74 ± 0.07
Identification	UAR	0.63 ± 0.05
	Accuracy	0.66 ± 0.03
	Score	0.77 ± 0.03

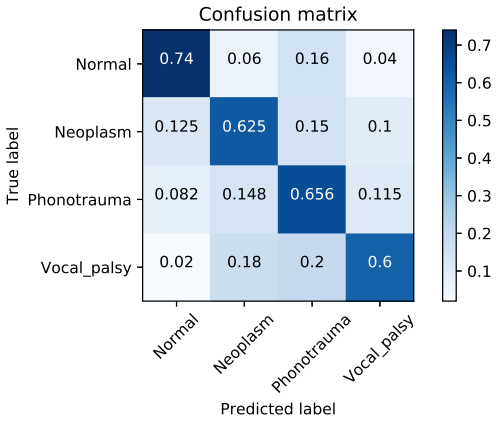

Fig. 3. Confusion matrix

then identifies the actual pathologies (neoplasm, phonotrauma or vocal palsy) in a second classification stage.

As the supplied corpus was contaminated by background noise, a VAD procedure was employed to differentiate voices from silences and low-amplitude noise. This procedure permitted a rapid preprocessing of the input dataset, although it did not serve to differentiate highly contaminated voice samples or signals suffering from saturation. It was then necessary to manually correct the recordings in order to provide reasonable speech samples that then served as input to the classification machines.

One interesting observation extracted from Table I is the importance of MFCC features in both detection and classification tasks as assessed by the feature selection algorithms. In all cases there were regarded as highly relevant for differentiation purposes. Another relevant pair of features in the spectral/cepstral set are the $CPPS$ and the LHr, which have been found relevant in all the classification tasks With regards to the modulation spectrum features, CIL was deemed as highly relevant for both detection and classification. Finally, and in reference to the complexity features it can be observed that the entropy measures based on Markov-Chains are highly relevant in detection tasks where 7 of them are considered informative according to the feature selection algorithms. By contrasts, its importance for identification is diminished.

To provide information about the performance of the identification task, the resulting scores after having divided the training partition into a training and a validation set are presented in Fig. 4 and Fig. 5 respectively. As observed, there is a large separability between normophonic and pathological scores in the training partition, but a more irregular behaviour can be observed in the validation set. In general terms the separability is good, although is difficult to establish a frontier that serves to separate normophonic from dysphonic behaviour, specially due to the large likelihood scores that are present in the normophonic class (see Fig. 5). Having chosen a margin (results in Table III) that maximizes the sensitivity served to provide a good sensitivity (0.93) at the expense of a decreased specificity (0.74).

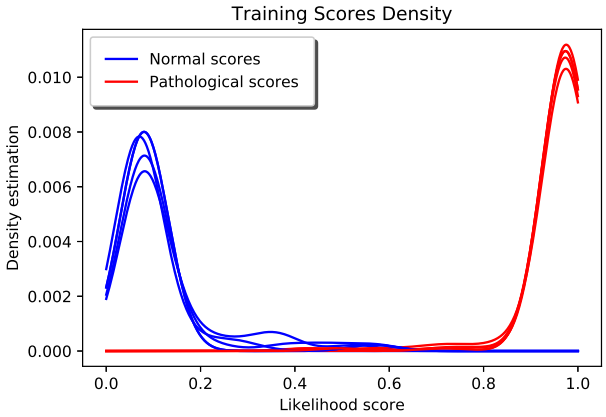

Fig. 4. Scores distribution of the training sample of the training partition.

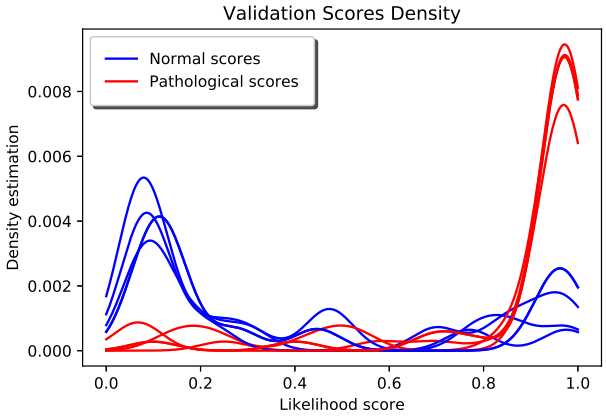

Fig. 5. Scores distribution of the validation sample of the training partition.

With regards to the identification task, the confusion matrix in Fig. 3 indicates the performance of the system in relation to the true and the predicted labels. Results indicate that the neoplasm class was most likely confused with the normophonic and the phonotrauma class, whereas for phonotrauma, the largest errors arose in relation to the neoplams and the vocal palsy classes. Similarly for the vocal palsy class, most of the errors appeared with the phonotrauma and neoplasm classes, whereas the normophonic class was most likely confused with the phonotrauma class. Results in Table III are in line with the diagonals of the confusion matrix, presenting an accuracy of

0.66 and UAR of 0.63. At the end, the score that resulted from the proposed system is 0.77.

The present paper has presented an AVCA system based on the extraction of descriptors of vocal quality from voice samples. Through different signal processing and machine learning methodologies it was possible to design a scheme that differentiates between normophonic states and pathology with an sensitivity of 0.94 and specificity of 0.74. In identification tasks an UAR of 0.63 was obtained.

ACKNOWLEDGMENT

This work was supported by the Ministry of Economy and Competitiveness of Spain under grant DPI2017-83405-R1.

REFERENCES

[1] J. B. Snow and J. J. Ballenger, *Ballenger's Otorhinolaryngology Head and Neck Surgery*, B. Decker, Ed., 2003.

[2] J. Gómez-García, L. Moro-Velázquez, and J. Godino-Llorente, "On the design of automatic voice condition analysis systems. part ii: Review of speaker recognition techniques and study on the effects of different variability factors," *Biomedical Signal Processing and Control*, vol. 48, pp. 128–143, 2019.

[3] S. N. Awan, N. Roy, and C. Dromey, "Estimating dysphonia severity in continuous speech: application of a multi-parameter spectral/cepstral model." *Clinical linguistics & phonetics*, vol. 23, no. 11, pp. 825–841, 2009.

[4] S.-H. Fang, Y. Tsao, M.-J. Hsiao, J.-Y. Chen, Y.-H. Lai, F.-C. Lin, and C.-T. Wang, "Detection of pathological voice using cepstrum vectors: A deep learning approach," *Journal of Voice*, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S089219971730509X

[5] T. Drugman, Y. Stylianou, Y. Kida, and M. Akamine, "Voice activity detection: Merging source and filter-based information," *IEEE Signal Processing Letters*, vol. 23, no. 2, pp. 252–256, Feb 2016.

[6] L. Moro-Velázquez, J. A. Gómez-García, J. I. Godino-Llorente, and G. Andrade-Miranda, "Modulation Spectra Morphological Parameters: A New Method to Assess Voice Pathologies according to the GRBAS Scale," *BioMed Research International*, vol. 2015, 2015.

[7] L. Moro-Velázquez, J. A. Gómez-García, and J. I. Godino-Llorente, "Voice Pathology Detection Using Modulation Spectrum-Optimized Metrics," *Frontiers in Bioengineering and Biotechnology*, vol. 4, no. 1, 2016.

[8] J. D. Arias-Londoño, J. I. Godino-Llorente, N. Sáenz-Lechón, V. Osma-Ruiz, and G. Castellanos-Domínguez, "Automatic detection of pathological voices using complexity measures, noise parameters, and mel-cepstral coefficients," *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 2, pp. 370–379, 2011.

[9] J. D. Arias-Londoño and J. I. Godino-Llorente, "Entropies from Markov Models as Complexity Measures of Embedded Attractors," *Entropy*, vol. 17, no. 6, pp. 3595–3620, 2015.

[10] G. Brown, A. Pocock, M.-J. Zhao, and M. Luján, "Conditional Likelihood Maximisation: A Unifying Framework for Information Theoretic Feature Selection," *Journal Machine Learning Research*, vol. 13, pp. 27–66, 2012.