Continuous Speech Classification Systems for Voice Pathologies Identification

Hugo Cordeiro[1,2(✉)], Carlos Meneses[2], and José Fonseca[1]

[1] Department of Electrical Engineering, Faculty of Sciences and Technology of the New University of Lisbon, 2829-516 Caparica, Portugal
jmf@uninova.pt
[2] Department of Electronics and Telecommunications and Computers, High Institute of Engineering of Lisbon, 1959-007 Lisbon, Portugal
{hcordeiro,cmeneses}@deetc.isel.ipl.pt

Abstract. Voice pathologies identification using speech processing methods can be used as a preliminary diagnostic. The aim of this study is to compare the performance of sustained vowel /a/ and continuous speech task in identification systems to diagnose voice pathologies. The system recognizes between three classes consisting of two different pathologies sets and healthy subjects. The signals are evaluated using MFCC (Mel Frequency Cepstral Coefficients) as speech signal features, applied to SVM (Support Vector Machines) and GMM (Gaussian Mixture Models) classifiers. For continuous speech, the GMM system reaches 74% accuracy rate while the SVM system obtains 72% accuracy rate. For the sustained vowel /a/, the accuracy achieved by the GMM and the SVM is 66% and 69% respectively, a lower result than with continuous speech.

Keywords: Voice pathologies identification · Continuous speech · Gaussian mixture models · Support vector machines

1 Introduction

Voice pathologies are caused by different factors such as the extensive or incorrect use of voice, stress, tobacco smoke inhalation, gastric reflux or hormonal problems. These pathologies typically affect the vocal folds and are detectable by direct laryngoscopy, which is the visualization of the vocal folds using a camera. However, this method is invasive, uncomfortable for the patient and may, depending on the equipment used, require a local anesthetic. The diagnostic equipment is also expensive and has high maintenance costs, requiring sterilization between diagnoses.

The use of an efficient, non-invasive, easy and fast method in pathologies recognition may be useful as an initial evaluation or as a complementary method for the diagnosis of voice pathologies. A method of voice pathologies diagnosis based in speech signals is especially useful in screening situations because it doesn't involve specialized equipment.

Several works in pathological voice identification discriminate between healthy and unhealthy subjects, without the identification of the pathology. Some works in pathological voice identification [1-4] use features like pitch jitter, shimmer,

L.M. Camarinha-Matos et al. (Eds.): DoCEIS 2015, IFIP AICT 450, pp. 217–224, 2015.
DOI: 10.1007/978-3-319-16766-4_23

harmonic-to-noise ratio (HNR) and energy spectrum to discriminate unhealthy subjects. Other works use features like MFFC [5] and wavelet analysis [6]. In [7] a comparison between different systems and features in pathological voice identification can be found.

There are few works in voice pathologies identification and all use the sustained vowel /a/ as speech signal. In this task, features like jitter and shimmer alone are not sufficient to discriminate pathologies [8]. In [9], spectral modulation and SVM are used to detect polyps pathology among three pathologies and healthy voices, using the vowel /a/ from the MEEI corpus [10]. The system achieves an average recognition rate of 90%. The features are obtained by spectral modulation where the discrete spectrum of the signal is modeled in sub-bands. The signal obtained from patients with polyps presents a higher energy in the pitch band. After feature extraction, an algorithm based on principal component analysis is used for dimensionality reduction. Finally, feature selection based on mutual information is used to sort only the most relevant characteristics.

In [11], the authors compare the results obtained in previous work [9] with those obtained with the MFCC as parameters. The results obtained are approximately 25% lower than the results obtained by spectral modulation.

In [12] five different pathologies are identified, assuming that the voice is pathologic. Two Arabic vowels equivalent to /a/ and /i/, where tested from a corpus consisting of numbers from 1 to 10, with 72 patients, thus generating 720 speech signals, where 80% are used for training and 20% for testing. From each vowel the value of the first and second formant are extracted taking 4 or 5 frames in the middle of the speech signal (where it is more stable). Using these four features a neuronal network achieves a recognition rate of 67.8% for male patients and 52.5% for female patients. Tests were also conducted with a classifier based on vector quantization but the results were inferior to those obtained by the neural network. It is reported in that work that, in some diseases, a deviation in the formants mean values allows the discrimination of the pathologies. In [13] wavelets are used in a MLP (Multi Layer Perception) neural network for identification between nodules and Reinke's edema. The accuracy reaches 87% for Reinke's edema and 86% for nodules using 80% of the data for training and 20% for independent testing. In [14] the authors use jitter in the wavelets components with SVM (Support Vector Machines) to perform distinction between nodules and Reinke's edema. Using a similar dataset as in [13], 82% accuracy has been achieved. Also for this database SVM classifiers were used to evaluate the impact of the first peak bandwidth presented in [15], pitch jitter and pitch in Reinke's edema vs. nodules identification. With the vowel /a/ the results obtained using the first LPC spectral peak achieves an overall accuracy 84.6% and do not improve by using additional features like pitch jitter. The results obtained are in line and sometimes overlap the results reported in previous publications with the same corpus, but in this work the test uses 50% of the data.

In the previous referenced works the vowel /a/ speech signal was used in pathological voice classification. In [5], continuous speech signals were also used and the results are compared with the obtained with the vowel /a/. The classifiers use MFCC and HMM and show 99.5% accuracy for the vowel /a/ and 98.6% for continuous speech. In this work, it was found that pathological voice recognition can be done based on continuous speech. However, continuous speech signals analysis has never been tested for voice pathologies identification.

The main objective of this work is to evaluate which voice signal has higher potential for voice pathologies identification: the vowel /a/ or continuous speech. The latter assumes that voice pathology have speech characteristic that are phoneme independent. To inspect that, the two signals are evaluated using MFCC as speech signal features, applied to SVM and GMM classifiers.

2 Relationship to Cloud-Based Solutions

As mentioned in the previous section, the area of pathological voice recognition is growing but there are few works on pathologies identification. One of the reasons is the lack of a reference database for the area development. Pathological voice recognition has numerous advances and in many cases recognition rates over 90% are achieved. However, for pathologies identification there is a considerable lack of data, since it must be divided by pathologies. The data shared in the cloud make possible to have more data and more important provide the same data to define strategies for studies comparison. For that, it is possible to establish a criterion for the acquisition of speech signals from subjects with pathologies. This can be done in innumerous institutions that can disseminate and share this information in the cloud. The Research from this topic always involves multidisciplinary teams and sharing information in the cloud will certainly produce important advances in the automatic diagnostic of voice pathologies.

3 Database

For training and testing the systems the MEEI database [10] was used. It contains 53 healthy subjects and 724 subjects with voice disorders. For these two set of subjects, the sustained vowel /a/ and a continuous speech excerpt, "rainbow passage" are available. Of all subjects with pathologies, only 477 have the information about the pathologie. This work uses the nodules, edema, unilateral vocal folds paralysis pathologies diagnostics and healthy subjects. The signals were recorded with a sampling rate of 25 kHz or 50 kHz. For this work all the files were down sampled to 25 kHz. There are 17 files from healthy subjects from the "rainbow passage" (continuous speech) recorded at 10 kHz that have been removed, resulting in a total of 36 healthy subjects.

For this work, two sets of pathologies were defined. A first set with subjects diagnosed with the physiological disorders edema and nodules. A second set was created with subjects diagnosed with unilateral vocal fold paralysis, a neuromuscular disorder. Each of these sets contains 59 files. A third set have 36 healthy subjects. Table 1 shows the database distribution by gender.

Table 1. Database gender distribution

Pathology	Males	Females	Total
Healthy	14	22	36
Nodules	1	18	19
Vocal Fold Edema	9	31	40
Vocal Fold Unilateral Paralysis	29	30	59

3.1 Nodules

Vocal fold nodules are benign lesions that can occur in both vocal folds, in places where friction is higher. Its location is usually at the junction of the anterior third and the posterior two thirds. When the vocal folds vibrate, impact zones exist and it is precisely in these areas where the lesions that cause nodules appear. This disease is common in people who intensively use their voice, as politicians, teachers, singers and children that often yell. The nodules are small, about the size of a pinhead and can appear over a given area on groups. The main symptoms are dysphonia, with the change in the voice tone, via a variable hoarseness. In extreme cases it may be possible to achieve a complete aphonia. Nodules do not allow complete closure of the vocal folds thus adding noise to the speech signal. To compensate for this effect the subjects tend to increase the tension in the muscle, further increasing collision that forces the vocal folds.

3.2 Vocal Folds Edema

The vocal folds edema increases the volume of the vocal folds. Causal factors are the use and abuse of the voice in which in some cases the subject may suffers from nodules. This pathology can be also due to the use of tobacco or drugs, excessive coughing, menstruation, menopause or pregnancy. The increase in the volume of the vocal folds causes changes in its elasticity with the consequent change in the voice timbre.

3.3 Vocal Folds Paralysis

The vocal folds paralysis is a peripheral neurology disorder. The tension and position of the vocal cords are controlled by the laryngeal muscles, which in turn are controlled by the nervous system. The vocal cord paralysis arises when these muscles cannot perform its function. The paralysis is in the muscle although the vocal folds can continue to vibrate in a not controlled way. The paralysis can be caused by various phenomena, such as compression of muscles due to infections, tumors or intoxications. The paralysis typically affects only one vocal fold. This implies that each fold vibrates at a different frequency. In these cases, the voice has a bimodal sound and the patient cannot speak loudly losing the power of vocal amplification. In this work, only unilateral paralysis is evaluated.

4 Implemented Systems

In this section the systems implementation and the data organization used to perform the tests between the sustained vowel and continuous speech signals are described. In the end results are presented.

As previously mentioned, speech signals used on this study were acquired with a sampling frequencies of 25 kHz or 50 kHz. To normalize the training set, all samples were down sampled to 25 kHz.

The features extracted from the speech signal were MFCC, energy and delta-energy. MFCC coefficients were estimated using filter banks scale 25-mel in 20 ms frames with 10 ms overlap. Several tests were conducted from order 8^{th} to 24^{th}. All the data was normalized with zero mean and unit variance. In the vowel /a/, with a duration between 3 and 5 seconds, all frames are analyzed. In the continuous speech, the silence zones were removed using the endpoint algorithm proposed in [16].

4.1 Dataset Organization

Two classes of pathologies were created: physiological lesions composed of edema in the vocal folds in a total of 59 subjects; and neuromuscular disorder with 59 subjects with unilateral paralysis of vocal folds; a third class with 36 healthy subjects is used.

The systems were trained with 75% of the data and tested with the remaining 25%, resulting in approximately 15 test subjects in classes with pathologies and 8 speakers in the healthy subjects. The *k*-fold cross-validation method [17] was used. In total 4 systems were created in order to rotate the train and test set and evaluated all data set.

4.2 SVM System

Support Vector Machines (SVM) classifiers are typically used has a two class classifier. In order to have a multiclass classifier, three SVM models were trained. This approach his known as one-against-one [18] and in this technique for N classes $N(N\text{-}1)/2$ SVM classifiers are trained. In this case, classifier #1 compares physiological pathologies vs. healthy, classifier #2 computes neuromuscular pathologies vs. healthy and classifier #3 computes physiological pathologies vs. healthy.

The test is done by single frame classification with both classifiers #1 and #2 classifying all frames as healthy/unhealthy. If the percentage of frames classified as healthy by classifier #1 averaged with classifier #2 is higher than 50% the subject is classified as healthy. Otherwise, the subject is classified as unhealthy and the classifier #3 identifies the pathology.

4.3 Gaussians Mixtures Models

Gaussians Mixtures Models (GMM) allows the train of N independent models. These systems had great impact on solutions for speaker recognition [19]. The system implemented for this work consists in one model for each class. Each model is represented by M Gaussian mixtures with covariance diagonal matrix. In this case, it was

found that the best results are obtained with 16 Gaussian mixtures for each class. For each test segment the likelihood is computed in each class and classified in the class that maximizes this value.

4.4 Metrics and Results

As described above, tests with the sustained vowel /a/ and with continuous speech used SVM and GMM classifiers. The systems evaluation was performed by the overall system ACC (accuracy) (1), the class sensitivity TPR (True Positive Rate) (2) and the class precision PPV (Positive Predictive Rate) (3). These measures are computed using the TP (True Positive), FP (False Positive), FN (False Negative) and TN (True Negative) approach. Table 2 and 3 show the obtained results.

$$ACC = (TP + TN) / (TP + TN + FP + FN) \quad (1)$$

$$TPR = TP / (TP + FN) \quad (2)$$

$$PPV = TP / (TP + FP) \quad (3)$$

In general, the system implemented with continuous speech using GMM achieved better results (74% accuracy), as well as best average precision and sensitivity rates (76.2% and 77.9% respectively).

The systems that used the sustained vowel /a/ achieved about 20% in absolute value less precision rate for healthy subjects. These results are particularly important since the unhealthy subjects are not diagnosed with any of the pathologies what can be considered an important disadvantage of these systems.

Continuous speech based systems have typically better results than the vowel /a/ based systems. Only in two instances the results obtained with the sustained vowel are better. In the first case, in table 2, the sensitivity rate in unilateral paralysis using the SVM classifier is 69.4% against 56% for continuous speech. With the GMM classifier the precision rate is 61%. In the second case, the precision for nodules and edema is 66.6% when used the vowel /a/ in the SVM classifier. However, when using continuous speech in the GMM classifier this value is 66.2%, decreasing only 0.4%.

Table 2. SVM system best results for continuous speech and Vowel /a/ in parentheses. ACC: 72% (69%), both systems use MFCC 20th order.

Classification / Original	Healthy	Nodules and Vocal Fold Edema	Unilateral Paralysis	Sensitivity (TPR)
Healthy	34 (33)	0 (1)	2 (2)	**94.4%** (91.6%)
Nodules and Vocal Fold Edema	1 (9)	44(32)	14 (18)	74.5% (54.2%)
Unilateral Paralysis	2 (3)	24 (15)	33 (41)	56% **(69.4%)**
Precision (PPV)	91.8% (73%)	64.7% **(66.6%)**	67.3% (67.2%)	

Table 3. GMM system best results for continuous speech and Vowel /a/ in parentheses. ACC: 74% (66%), systems use MFCC 12th order for continuous speech and 8th order for the vowel /a/.

Classification / Original	Healthy	Nodules and Vocal Fold Edema	Unilateral Paralysis	Sensitivity (TPR)
Healthy	33 (30)	3 (2)	0 (4)	91.6% (83.3%)
Nodules and Vocal Fold Edema	1 (10)	45 (32)	13 (17)	**76.2%** (54.2%)
Unilateral Paralysis	1 (4)	20 (15)	36 (40)	61% (67.7%)
Precision (PPV)	**94.2%** (68%)	66.2% (65.3%)	**73.4%** (65.5%)	

5 Conclusions and Future Work

This work presents two voice pathologies identification systems using continuous speech and the sustained vowel /a/. Three classes were created: healthy subjects, subjects diagnosed with vocal fold edema and nodules (physiological disorders) and subjects diagnosed with unilateral vocal folds paralysis (neuromuscular disorder).

The main objective was to compare the performance of voice pathologies identification using continuous speech with the sustained vowel /a/ typically used in these applications. Two classifiers systems were created, using SVM and GMM, both using MFFC parameters as speech signal features. The results showed that continuous speech allows better results for the two systems and that GMM models classifier overcame the SVM classifier. In particular GMM and continuous speech have better precision values in the identification of healthy subjects. However, the sustained vowel /a/ has the best results in the sensitivity for unilateral vocal fold paralysis. Considering this fact, some research in system fusion will be done in the future to verify if, this way, it is possible to improve the overall performance. Continuous speech shows high potential for voice pathologies identification. In the future this performance may be improved using other speech signal features and different classifiers.

There are few studies in pathological voice identification. Different databases and pathologies studied make the comparison with other woks inconsistent. For that a cloud sharing database approach certain will bring relevant improvements in this task.

Acknowledgments. The authors would like to thank Prof. Ana Mendes from Polytechnic Institute of Setúbal for sharing the database used in this study and the High Institute of Engineering of Lisbon by the scholarship that allowed the work progress.

References

1. Lieberman, P.: Some acoustic measures of the fundamental periodicity of normal and pathologic larynges. J. Acoust. Soc. Amer. **35**, 344–353 (1963)
2. Iwata, S.: Periodicities of pitch perturbations in normal and pathological larynges. J. Acoust. Soc. Amer. **45**, 344–353 (1972)

3. Shama, K., Krishna, A., Niranjan Cholayya, N.U.: Study of harmonics-to-noise ratio and critical-band energy spectrum of speech as acoustic indicators of laryngeal and voice pathology. EURASIP Journal on Advances in Signal Processing 1 (2007)
4. Cordeiro, H., Fonseca, J., Meneses C.: Spectral Envelope and Periodic Component in Classification Trees for Pathological Voice Diagnostic. In: 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 4607–4610 (2014)
5. Dibazar, A., Narayanan S.: A system for automatic detection of pathological speech. In: 36th Asilomar Conf., Signal, Systems & Computers (2002)
6. Fonseca, E.S., Guido, R.C., Scalassara, P.R., Maciel, C.D., Pereira, J.C.: Wavelet time–frequency analysis and least squares support vector machines for the identification of voice disorders. Comput. Biol. Med. **37**, 571–578 (2006)
7. Sáenz-Lechón, N., Godino-Llorente, J.I., Osma-Ruiz, V., Gómez-Vilda, P.: Methodological issues in the development of automatic systems for voice pathology detection. Biomedical Signal Processing and Control 1, 120–128 (2006)
8. Scalassara, P.R., Dajer, M.E., Maciel, C.D., Guido, R.C., Pereira, J.C.: Relative entropy measures applied to healthy and pathological voice characterization. Applied Mathematics and Computation **207**, 95–108 (2009)
9. Markaki M., Stylianou Y.: Using modulation spectra for voice pathology detection and classification. In: Proc. IEEE EMBC 2009, Minneapolis, pp. 2514–2517 (2009)
10. Key Elemetrics, Elemetrics Disordered Voice Database (1994)
11. Markaki, M., Stylianou, Y.: Voice Pathology Detection and Discrimination Based on Modulation Spectral Features. IEEE Transactions on Audio, Speech, and Language Processing **19**, 1938–1948 (2011)
12. Muhammad, G., Alsulaiman, M., Mahmood, A., Ali, Z.: Automatic voice disorder classification using vowel formants. In: IEEE Int. Conf. Multimedia and Expo (ICME) (2011)
13. Fonseca, E.S., Pereira, J.C.: Normal versus pathological voice signals. IEEE Engineering in Medicine and Biology Magazine **28**, 44–48 (2009)
14. Carvalho, R.T.S., Cavalcante, C.C., Cortez, P.C.: Wavelet transform and artificial neural networks applied to voice disorders identification. In: Third World Congress on Nature and Biologically Inspired Computing (NaBIC), pp. 371–376 (2011)
15. Cordeiro, H., Fonseca, J., Meneses, C.: Edema and Nodules Identification in vowels using spectral features and jitter. In: CETC 2013, Conference on Electronics, Telecommunications and Computers, Procedia Technology, vol. 17, pp. 202–208 (2014)
16. Lamel, L., Rabiner, L., Rosenberg, A., Wilpon, J.: An Improved Endpoint Detector for Isolated Word Recognition. IEEE Transactions on Acoustics, Speech, and Signal Processing **29**, 777–785 (1981)
17. Kohavi, R.: A study of cross-validation and bootstrap for accuracy estimation and model selection. IJCAI 14 (1995)
18. Chih-Wei, H., Chih-Jen, L.: A comparison of methods for multiclass support vector machines. IEEE Transactions on Neural Networks **13**, 415–425 (2002)
19. Reynolds, D.: Speaker identification and verification using Gaussian mixture speaker models. Speech Communications **17**, 91–108 (1995)