Hierarchical Classification and System Combination for Automatically Identifying Physiological and Neuromuscular Laryngeal Pathologies

***,†Hugo Cordeiro, *José Fonseca, ‡Isabel Guimarães, and †Carlos Meneses,** **Caparica, †Lisbon, and ‡Alcabideche, Portugal*

Summary: Objectives. Speech signal processing techniques have provided several contributions to pathologic voice identification, in which healthy and unhealthy voice samples are evaluated. A less common approach is to identify laryngeal pathologies, for which the use of a noninvasive method for pathologic voice identification is an important step forward for preliminary diagnosis. In this study, a hierarchical classifier and a combination of systems are used to improve the accuracy of a three-class identification system (healthy, physiological larynx pathologies, and neuromuscular larynx pathologies).

Method. Three main subject classes were considered: subjects with physiological larynx pathologies (vocal fold nodules and edemas: 59 samples), subjects with neuromuscular larynx pathologies (unilateral vocal fold paralysis: 59 samples), and healthy subjects (36 samples). The variables used in this study were a speech task (sustained vowel /a/ or continuous reading speech), features with or without perceptual information, and features with or without direct information about formants evaluated using single classifiers. A hierarchical classification system was designed based on this information.

Results. The resulting system combines an analysis of continuous speech by way of the commonly used sustained vowel /a/ to obtain spectral and perceptual speech features. It achieved an accuracy of 84.4%, which represents an improvement of approximately 9% compared with the stand-alone approach. For pathologic voice identification, the accuracy obtained was 98.7%, and the identification accuracy for the two pathology classes was 81.3%.

Conclusions. Hierarchical classification and system combination create significant benefits and introduce a modular approach to the classification of larynx pathologies.

Key Words: Hierarchical classification–Pathologic voice identification–Larynx pathology identification–Continuous speech–Sustained vowel.

INTRODUCTION

The use of a noninvasive method for pathologic voice identification is an important step forward for a preliminary or complementary diagnosis of larynx pathologies. The primary assessment of voice in health-care practice conventionally involves invasive instruments (eg, laryngoscopy, viewing the vocal folds with a camera) to assess the biomechanical behavior of the vocal folds and analyze the color, texture, composition, and extent of a lesion (voice pathology). However, laryngoscopies involve expensive, sensitive equipment, and are a specific controlled health protocol allowed only to ear, nose, and throat medical doctors and with the examined subject's cooperation. Under such critical circumstances, an alternative noninvasive, inexpensive method based on speech signals, using portable equipment, may be very helpful for mass screening purposes, for example, in schools and companies.

Accepted for publication September 8, 2016.

Conflicts of interest: The authors do not declare any potential conflicts of interest in this study.

From the *Department of Electrical Engineering, Faculty of Sciences and Technology of the New University of Lisbon, 2829-516 Caparica, Portugal; †Department of Electronics, Telecommunications and Computers, Higher Institute of Engineering of Lisbon, 1959-007 Lisbon, Portugal; and the ‡Department of Speech and Language Therapy, School of Health Sciences at Alcoitão, 2649-506 Alcabideche, Portugal.

Address correspondence and reprint requests to Hugo Cordeiro, Department of Electrical Engineering, Faculty of Sciences and Technology of the New University of Lisbon (FTC–UNL), Caparica, Portugal; Department of Electronics and Telecommunications and Computers at the Higher Institute of Engineering of Lisbon (ISEL), Lisbon, Portugal. E-mail: hcordeiro@deetc.isel.ipl.pt

Journal of Voice, Vol. 31, No. 3, pp. 384.e9–384.e14

0892-1997

http://dx.doi.org/10.1016/j.jvoice.2016.09.003

Acoustic analysis methods provide noninvasive and perturbation measures such as jitter and shimmer.[1–3] However, in Titze,[4] it is suggested that perturbation analysis may not be applicable to aperiodic signals. In recent years, different techniques based on nonlinear dynamic models have been applied to investigate the dynamic behavior of biomedical systems, including voice signals.[5,6] Research shows that acoustic measures in the frequency and time domain, such as fundamental frequency, jitter, shimmer, and harmonics-to-noise ratio, are associated with perceived deviation from normal voice quality.

For *pathologic voice identification*, in which healthy and unhealthy voice samples are identified, spectral features such as energy spectrum,[7] Mel-frequency cepstral coefficients (MFCC),[8–12] and formant analysis[13,14] have been demonstrated to achieve an accuracy rate of more than 90%. One of the advantages of this approach is that pitch determination, a difficult task when dealing with voice disorders, is not required, as shown in Arias-Londoño et al.[10] Typically, the sample used for feature extraction is the sustained vowel /a/, because it is produced with the vocal tract completely open and is highly correlated with the electroglottograph.[15] In the study by Lee et al,[14] formants are used as features for detecting voice pathologies. The authors concluded that patients change the vocal tract in different ways to compensate for the vocal fold handicap.

Voice pathology identification, in which laryngeal pathologies are identified, is more challenging. Few studies[16–21] can be found in this field in comparison with pathologic voice identification. Most of these studies use the sustained vowel /a/ as a speech signal. In Muhammad et al,[16] five different larynx

pathologies are identified: vocal fold cyst, gastroesophageal reflux disease, paralysis, polyp, and sulcus. Frames of each vowel in the middle of the speech signal (stable area) were extracted to determine the first and second formant values. Using these features, a neural network achieves an accuracy rate of 67.8% for male patients and 52.5% for female patients. In Fonseca and Pereira,[17] the authors used jitter in wavelet components with a support vector machine (SVM) to distinguish nodules and Reinke edema, achieving an accuracy rate of 82%. In Markaki and Stylianou,[18] spectral modulation with the SVM classifier was used to detect polyp pathologies among five pathologies (keratosis, nodules, paralysis, adductor, and keratosis) and healthy voices. The features were obtained by spectral modulation, in which the discrete spectrum of the signal is modeled in sub-bands. This system achieved an average accuracy rate of 90%.

Continuous speech (CS) analysis for pathologic voice identification was introduced in Parsa and Jamieson.[22] The authors compared glottal features extracted from the vowel /a/ and voiced segments of continuous speech. They concluded that the vowel /a/ produces better results because detection in voiced segments was unreliable owing to the hoarseness of pathologic voices. To overcome this problem, continuous speech was also used in Dibazar et al[8] for pathologic voice detection using voiced and unvoiced segments, with results similar to those obtained using the vowel /a/.

In Cordeiro et al,[19] a continuous speech signal was introduced into the identification of laryngeal pathologies. A three-class system was implemented to distinguish between signals from healthy subjects and those with paralysis or nodules and edemas. Two systems were implemented: one based on SVM and another based on Gaussian mixture model (GMM) classifiers, both using MFCC features extracted from the sustained vowel /a/ and continuous reading speech. For continuous speech, the GMM system achieved a 74% accuracy rate, whereas the SVM system obtained a 72% accuracy rate. For the sustained vowel /a/, the accuracy rates achieved by the GMM and SVM systems were 66% and 69% respectively, which are lower results than when continuous speech is used. Continuous reading speech produced better results for the two systems with the GMM classifier, which outperformed the SVM classifier. The SVM classifier using continuous speech in particular achieved better results in the identification of healthy subjects. With the sustained vowel /a/, the best results were achieved for the identification of unilateral vocal fold paralysis.

New formant features were also evaluated for the classifiers described in Cordeiro et al.[20] The first of these features are line spectral frequencies (LSFs), which are linear predictive coding representations that contain direct information about the formant frequencies and bandwidths that represent the vocal tract. The other features are Mel line spectral frequencies (MLSFs),[23] an LSF feature that contains perceptual information, with a Mel filter bank applied to the signal spectrum. The LSF, MLSF, and MFCC features were tested for several classifiers using the sustained vowel /a/ and continuous speech to identify healthy subjects and those with unilateral vocal fold paralysis or nodules and edemas. The best accuracy rate (77.9%) was obtained by a GMM classifier using the MLSF feature extracted from continuous speech.

In Ali et al,[21] cepstral coefficients from continuous speech features applied to a GMM were tested in a similar subset of the same database used in Markaki and Stylianou.[18] The accuracy rate for identifying pathologies is 86%. In total, Ali et al,[21] study uses 173 unhealthy subjects distributed over five pathologies (keratosis, nodules, paralysis, adductor, and keratosis).

In Cordeiro et al,[20] it is clear that the same speech task (sustained vowel /a/ or continuous reading speech), feature (with or without perceptual information and with or without direct formants information), and classifier (SVM, GMM, linear discriminant analysis) have different influences on the detection of healthy voices and each of the pathologies. One of the objectives of this study is to find the best combination for distinguishing between healthy voices, physiological larynx pathologies (vocal fold edemas or nodules), and neuromuscular larynx pathologies (unilateral vocal fold paralysis). The variables used in this study are (1) the speech task (sustained vowel /a/ or continuous reading speech), (2) features with or without perceptual information, and (3) features with or without direct information about formants. With this information, the second objective is to test the hypothesis that a hierarchical classification system, in which each node is implemented in combination, performs better than the best single system.

MATERIALS AND METHODS

Our study uses the Massachusetts Eye and Ear Infirmary (MEEI) database,[24] developed by the MEEI Voice and Speech Laboratory and commercialized by the Kay Elemetrics Corp. All ethical issues, data protection, and pathology labeling are responsibility of these companies. No other speech signals were used and no clinical trials were performed.

The MEEI database is commonly used in the field of automatic pathologic voice identification[8–13,18,21,22,25] and is the only commercial database that includes continuous speech signals of a significant duration (about 12 seconds each). This database is composed of signals from 53 healthy subjects and 724 subjects with voice disorders. The files were acquired at sample rates of 10 kHz, 25 kHz, and 50 kHz.

Material

Three main classes were considered: physiological larynx pathologies (PLP), with 59 samples (vocal fold nodules or edemas); neuromuscular larynx pathologies (NLP), with 59 samples (unilateral vocal fold paralysis); and healthy voices, with 36 samples. Of the 53 samples of healthy subjects reading speech files, 17 samples recorded at 10 kHz were discarded. All the files sampled at 50 kHz were downsampled to 25 kHz.

The overall gender distribution is 34.4% men and 65.6% women (Table 1). This database subset was chosen to maximize the number of unhealthy subjects. In total there are 118 unhealthy subjects divided equally into two classes. Unilateral vocal fold paralysis is the most common larynx pathology in the MEEI database. Edemas are the second most common pathology in the MEEI database and can be a preliminary disease for nodules,[26] which is why they were merged into one class. At least three subjects in the database are diagnosed as having these two pathologies simultaneously.