Applied Soft Computing 18 (2014) 91–103

Contents lists available at ScienceDirect

Applied Soft Computing

journal homepage: www.elsevier.com/locate/asoc

Fusion of voice signal information for detection of mild laryngeal pathology

Evaldas Vaiciukynas[a,*], Antanas Verikas[a,b], Adas Gelzinis[a], Marija Bacauskiene[a], Zvi Kons[c], Aharon Satt[c], Ron Hoory[c]

[a] Department of Electrical & Control Equipment, Kaunas University of Technology, Studentu 50, LT-51368 Kaunas, Lithuania
[b] Intelligent Systems Laboratory, Halmstad University, Box 823, S 301 18 Halmstad, Sweden
[c] IBM Haifa Research Laboratory, Haifa University Campus, Mount Carmel, 31905 Haifa, Israel

ARTICLE INFO

Article history:
Received 7 August 2012
Accepted 4 January 2014
Available online 24 January 2014

Keywords:
Random forest
SVM
Angle modulated differential evolution
Feature selection
Ensemble of classifiers
Pathological voice

ABSTRACT

Detection of mild laryngeal disorders using acoustic parameters of human voice is the main objective in this study. Observations of sustained phonation (audio recordings of vocalized /a/) are labeled by clinical diagnosis and rated by severity (from 0 to 3). Research is exclusively constrained to healthy (severity 0) and mildly pathological (severity 1) cases – two the most difficult classes to distinguish between.

Comprehensive voice signal characterization and information fusion constitute the approach adopted here. Characterization is obtained through diverse feature set, containing 26 feature subsets of varying size, extracted from the voice signal. Usefulness of feature-level and decision-level fusion is explored using support vector machine (SVM) and random forest (RF) as basic classifiers. For both types of fusion we also investigate the influence of feature selection on model accuracy. To improve the decision-level fusion we introduce a simple unsupervised technique for ensemble design, which is based on partitioning the feature set by k-means clustering, where the parameter k controls the size and diversity of the prospective ensemble.

All types of the fusion resulted in an evident improvement over the best individual feature subset. However, none of the types, including fusion setups comprising feature selection, proved to be significantly superior over the rest. The proposed ensemble design by feature set decomposition discernibly enhanced decision-level and significantly outperformed feature-level fusion. Ensemble of RF classifiers, induced from a cluster-based partitioning of the feature set, achieved equal error rate of $13.1 \pm 1.8\%$ in the detection of mildly pathological larynx. This is a very encouraging result, considering that detection of mild laryngeal disorder is a more challenging task than a common discrimination between healthy and a wide spectrum of pathological cases.

1. Introduction

Identification of laryngeal disorders in clinical practice is a rather complex diagnostic procedure, involving evaluation of patient's complaints, case-record, and data of instrumental as well as histological examination. Complaints are generally summarized by answers to the questionnaire, while the instrumental examination results into a sequence of laryngeal images and/or voice recordings. Non-invasive measurements, such as questionnaire data or voice recordings, can facilitate early detection of potential larynx-related disorders and therefore be of great value in the preventive care as well as for voice quality assessment.

The validity of acoustic analysis approach rests on the complex relationship between physiological function of human larynx and the concomitant properties of voice signal. Voice production is an elaborate process that involves muscle movements, respiration, and the brain control as well as hearing sensory system feedback. Properties of pathological voice are generally induced by mass increase, a lack of closure, or elasticity change of the vocal folds. The result is that the movement of the vocal folds is not balanced and an incomplete closure of the vocal folds may appear in glottal cycles. The subglottal airstream is modulated by this unbalanced vocal fold movement. Irregular air pulses emerge from larynx, propagate through pharynx and oral and nasal cavities, and radiate from the mouth and nose. The resultant acoustic signal is thus affected by a physiological disturbance in the larynx, and this information may be used to measure the disturbance [1].

* Corresponding author. Tel.: +370 67642585; fax: +370 37351409.
E-mail addresses: evaldas.vaiciukynas@ktu.lt, auksaplaukis@gmail.com (E. Vaiciukynas), antanas.verikas@hh.se (A. Verikas), adas.gelzinis@ktu.lt (A. Gelzinis), marija.bacauskiene@ktu.lt (M. Bacauskiene), zvi@il.ibm.com (Z. Kons), aharonsa@il.ibm.com (A. Satt), hoory@il.ibm.com (R. Hoory).

1568-4946/$ – see front matter © 2014 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.asoc.2014.01.012

Laryngeal pathologies are relatively common (affecting ~5% of the population [2]) and are found in varying degrees of progression and severity. Severity of laryngeal pathology could range from a very low to an extremely high, where the mild pathology is hardest to distinguish, and the more severe cases are easier to diagnose. Exclusively resorting to discrimination between healthy and mildly pathological voices is interesting from the preventive care perspective as well as more challenging in pattern recognition sense. Detection of pathological larynx from acoustic analysis of voice can be summarized in the following 2 steps:

1. extraction and pre-processing of meaningful features from the signal;
2. using features to distinguish between healthy and pathological cases.

Various machine learning approaches proved to be useful for such a task. Researchers apply either generative methods, like Gaussian mixture model (GMM), hidden Markov model (HMM), linear and quadratic discriminant analysis, or discriminative methods, like decision tree, *k*-nearest neighbors (*k*-NN), multi-layer perceptron (MLP), and support vector machine (SVM). Ensemble methods, which combine separate classifiers into a multiple classifier system (MCS), are also sometimes used.

A previous study on the older version of the same database we use here introduced several diverse subsets of acoustic features, applied wrapper-based feature selection and explored various combination strategies for committees of *k*-NN and SVM classifiers [3]. In the current study context is narrowed by constraining exclusively to healthy and mildly pathological observations, information abundance is considerably increased by expanding the feature set, influence of feature selection on information fusion with RF and SVM classifiers is investigated and, to improve the accuracy of detector, feature set decomposition technique for ensemble design is proposed.

2. Related work

Objective measurement of the severity of dysphonia typically requires signal processing algorithms applied to acoustic recordings [4]. Various features were researched to provide effective distinction between normophonic/dysphonic voices. For example, in [5], the high resolution of modulation spectral representation yielded quite distinctive patterns depending on the type and severity of voice pathology, allowing thus a finer than normal/abnormal distinction.

Most scientific studies on voice pathology detection do not take into account the level of severity, with the exception of [6–11]. Furthermore, even if subjects are split up according to the level of severity, this is performed no more than subjectively, usually by the specific perceptual rating scale, like VHI (voice handicap index) or GRBAS (grade, roughness, breathiness, asthenicity, and strain). According to [8], results of experiments attempting to define and highlight the pathological voices contained within a mixed set of normal and abnormal voice data, showed that it is indeed possible to make those splits, highlighting which of the voices had been previously diagnosed as presenting a mild – first stage – vocal fold pathology. However, success is very much dependent upon the choice of parameters, namely features, describing the voice.

Findings in [12] suggest that voice type, namely breathy, hoarse, strained, is associated with the interaction of spectral noise, fundamental frequency, and signal irregularity, and that dysphonic severity is associated with similar parameters as well. A number of studies, for example [13], have shown that the amplitude of the first rahmonic peak in the cepstrum is a global descriptor of voice quality. In [14] it was reported, that the cepstral peak prominence (CPP) correlates well with perceptual ratings of breathiness, while [15] concluded that CPP exhibits a better correlation with overall voice quality compared to other cues, such as jitter, shimmer, and several spectral tilt and noise measures.

Appealing to the non-linear behaviour, involved in voice production process, study in [16] applied non-linear dynamic analysis by exploring a complex characterization of voice signals. Non-uniform embedding technique, using two features, namely, correlation dimension and approximate entropy, was introduced and classification carried out by means of *k*-NN and SVM classifiers. This interesting dynamic feature space transformation was applied to classic short-term noise parameters and Mel-frequency cepstral coefficients (MFCCs) in [17] and resulted in significant improvement of the performance with no addition of new features to the original input space.

While some studies on the detection of pathological voice investigate and compare variously derived features, others fuse diverse representations of voice signal and report successful application of feature selection. For example, in [18], subject's voice samples of continuous speech and sustained vowels were concatenated and analysed using 13 acoustic measures based on fundamental frequency perturbation, amplitude perturbation, spectral and cepstral analyses. Stepwise multiple regression yielded a six-variable acoustic model for the multi-parametric assessment of overall voice quality of the concatenated samples (with a cepstral measure as

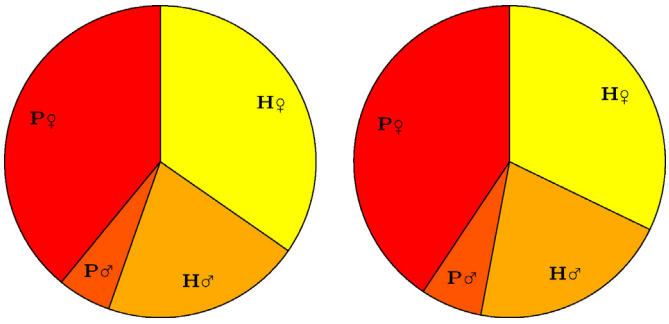

Fig. 1. Distribution of 251 subjects (left) and 625 recordings (right) in the database: healthy female (H♀), healthy male (H♂), pathological male (P♂), pathological female (P♀).

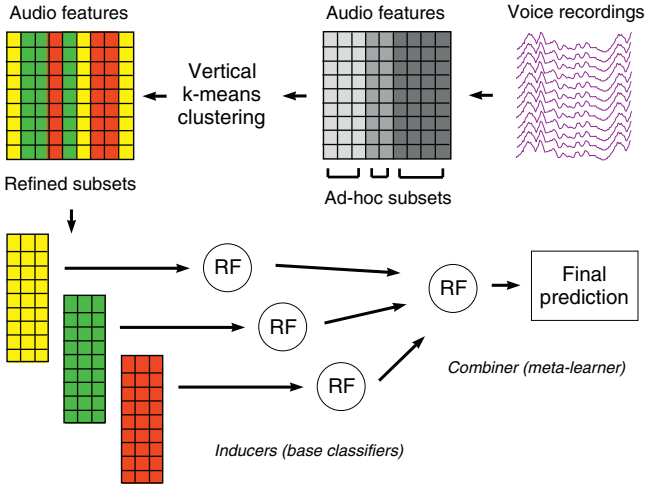

Fig. 2. Ensemble design by feature oriented decomposition: cluster-based partitioning.