Medical Engineering and Physics 37 (2015) 210–218

Contents lists available at ScienceDirect

Medical Engineering and Physics

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

Data dependent random forest applied to screening for laryngeal disorders through analysis of sustained phonation: Acoustic versus contact microphone

A. Verikas[a,b,*], A. Gelzinis[a], E. Vaiciukynas[a], M. Bacauskiene[a], J. Minelga[a], M. Hållander[b], V. Uloza[c], E. Padervinskis[c]

[a] *Department of Electric Power Systems, Kaunas University of Technology Studentu 50, LT-51368, Kaunas, Lithuania*
[b] *IS-Lab, Halmstad University, Box 823, S-30118 Halmstad, Sweden*
[c] *Department of Otolaryngology, Lithuanian University of Health Sciences Eiveniu 2, LT-50009, Kaunas, Lithuania*

ARTICLE INFO

Article history:
Received 30 May 2014
Revised 31 October 2014
Accepted 31 December 2014

Keywords:
Laryngeal disorder
Sustained phonation
Voice
Random forest
Committee
Decision confidence

ABSTRACT

Comprehensive evaluation of results obtained using acoustic and contact microphones in screening for laryngeal disorders through analysis of sustained phonation is the main objective of this study. Aiming to obtain a versatile characterization of voice samples recorded using microphones of both types, 14 different sets of features are extracted and used to build an accurate classifier to distinguish between *normal* and *pathological* cases. We propose a new, data dependent random forests-based, way to combine information available from the different feature sets. An approach to exploring data and decisions made by a random forest is also presented. Experimental investigations using a mixed gender database of 273 subjects have shown that the perceptual linear predictive cepstral coefficients (PLPCC) was the best feature set for both microphones. However, the linear predictive coefficients (LPC) and linear predictive cosine transform coefficients (LPCTC) exhibited good performance in the acoustic microphone case only. Models designed using the acoustic microphone data significantly outperformed the ones built using data recorded by the contact microphone. The contact microphone did not bring any additional information useful for the classification. The proposed data dependent random forest significantly outperformed the traditional random forest.

1. Introduction

Automated analysis of voice signals is used increasingly in screening for laryngeal disorders [1–6]. Several parameters computed from voice signals are a convenient way of documentation and quantification of dysphonia changes and outcomes of therapeutic and surgical treatment of laryngeal disorders [4,7–10]. Although voice recordings have been carried out for many years in clinical practice, the debate on microphone selection is still going on [11–13].

Vibrations from the vocal folds, generated during voice production, are transmitted through the vocal tract to the skin surface and can be sensed by contact microphones [14,15]. Thus, both acoustic and contact microphones can be used to record vibrations produced by vocal folds. Contact microphones are considered being useful for extraction of voice fundamental frequency [10], detecting glottal vibrations [16], recording subglottal pressure waves [17], estimating sound pressure levels of voiced speech [16], and mapping neck surface vibrations during vocalized speech [18].

Validity and reliability of acoustic measurements are highly affected by a background noise [12,19]. Due to its vicinity to the voice source, a contact microphone is less sensitive to background noises and provides enhanced voice signal clarity in noisy environments [15,16,20–22]. It is suggested that an acoustic environment should have a signal-to-noise ratio of at least 30 dB to produce valid results in audio analysis [12]. This recommendation can be fulfilled easily when voice recordings are performed in a special sound-proof booth. However, this requirement can become not feasible when voice recordings are obtained in an ordinary environment for voice disorders screening task.

However, several studies with contact microphones revealed decreased speech signal intelligibility compared to headset microphones [15,21,22]. Moreover, contact microphones are not very

* Corresponding author. Tel.: +46 35 167140.
E-mail addresses: antanas.verikas@hh.se (A. Verikas), adas.gelzinis@ktu.lt (A. Gelzinis), evaldas.vaiciukynas@ktu.lt (E. Vaiciukynas), marija.bacauskiene@ktu.lt (M. Bacauskiene), jonasmin@gmail.com (J. Minelga), magnus.hallander@hh.se (M. Hållander), virgilijus.ulozas@kmuk.lt (V. Uloza), evaldas.padervinskis@kmuk.lt (E. Padervinskis).

http://dx.doi.org/10.1016/j.medengphy.2014.12.005

effective in transmitting consonant sounds and high frequencies [23]. The elasticity properties of underlying human body tissues acting as a low-pass filter with a 3 kHz cut-off frequency [22], limit the frequency range of the resulting signal.

It was demonstrated that in case of non-stationary background noise, use of contact microphones can significantly improve accuracy of separation between voice recordings obtained from healthy subjects and subjects experiencing voice-related problems [24–26]. By using recordings from both types of microphones, Dupont et al. [22] achieved 80% recognition accuracy when discriminating between pathological and normal cases. Mubeen et al. [27] achieved some increase in performance when combining features of one type (weighted linear predictive cepstral coefficients) extracted from both types of recordings. Erzin [28] proposed a new framework, which learns joint sub-phone patterns of contact and acoustic microphone recordings using a parallel branch HMM structure. Application of this technique resulted in significant improvement of throat-only speech recognition.

Numerous sets of features, emphasizing different properties of voice signals, have been proposed for characterizing voice recordings [29]. Some feature sets may be more suitable for acoustic while others for contact microphones. This study, based on a variety of different features sets, investigates this issue. We also investigate if significant gains in classification performance can be achieved from various combinations of information obtained from microphones of the two types. A way to explore decisions made by an automated system, usually called a "black box", is also suggested.

We have chosen sustained phonation of vowel /*a*/ for the analysis, since steady-state phonation is simple, reduces variance in sustained vowels and enables reliable computation of acoustic features [9,30]. Moreover, sustained vowels are not influenced by speech rate and stress, they typically do not contain voiceless phonemes, fast voice onsets and terminations, and prosodic fluctuations in pitch and amplitude [8]. Sustained vowel phonation is rather insulated from aspects related to different languages.

2. Voice database

Voice samples were recorded in a sound-proof booth simultaneously using an acoustic and contact microphones. An acoustic cardioid microphone AKG Perception 220 (AKG Acoustics, Vienna, Austria) with frequency range from 20 Hz to 20 kHz was used in this study. The acoustic microphone was placed at a 10 cm distance from the mouth (the subjects were seated with a head rest), keeping at about 90° microphone-to-mouth angle. An omni-directional Triumph PC (Clearer Communications Inc., Burnaby, Canada), placed on the projection of lamina of thyroid cartilage and fixed with elastic bail, was used as a contact microphone. The frequency range of the contact microphone is from 100 Hz to 16 kHz. The audio format was wav (dual-channel PCM, 16 bit samples at 44 kHz rate), providing the Nyquist frequency $F_{max} = 22$ kHz.

A mixed gender database of 273 subjects (163 normal voices and 110 pathological voices), ranging from 19 to 85 years in age, was used. The *normal voice subgroup* was composed of healthy volunteer individuals who considered their voice as normal. They had no complaints concerning their voice and no history of chronic laryngeal diseases or other long-lasting voice disorders. The voices of this group of individuals were also evaluated as healthy voices by clinical voice specialists. Furthermore, no pathological alterations in the larynx of the subjects of the normal voice subgroup group were found during video laryngostroboscopy. The *pathological voice subgroup* consisted of patients who represented a rather common, clinically discriminative collection of laryngeal diseases, that is, mass lesions of vocal folds and paralysis.

Table 1
Features sets used in the study, 927 features in total.

#	Type of extracted features	Size
1.	Pitch and amplitude perturbation measures	24
2.	Frequency (0–5000 Hz)[a]	100
3.	Mel-frequency bands[b]	35
4.	Cepstral energy[c]	100
5.	Mel-frequency cepstral coefficients	35
6.	Autocorrelation[d]	80
7.	Harmonics to noise ratio in spectral domain	11
8.	Harmonics to noise ratio in cepstral domain	11
9.	Linear predictive coefficients	77
10.	Linear predictive cosine transform coefficients	77
11.	Shape of signal envelope[e]	128
12.	Levinson–Durbin reflection coefficients	24
13.	Vocal tract area irregularity	71
14.	Perceptual linear predictive cepstral coefficients	154

[a] Spectral energy in non-overlapping frequency bands of equal width.
[b] *i*th feature is given by the weighted spectral energy in the *i*th mel-window.
[c] Cepstral energy in non-overlapping frequency bands of equal width.
[d] Autocorrelation sequence over the lag range from zero to the half period of the main frequency.
[e] Several periods of a voice signal are averaged and represented by the amplitude at 128 equally spaced points.

3. Methodology

The main goal of this study is a comprehensive comparison of usefulness of a large number of feature sets extracted from voice recordings acquired with acoustic and contact microphones in a laryngeal pathology detection task. We also investigate if significant improvement in pathology detection accuracy can be achieved by combining information obtained from microphones of the two types. We use a random forest (RF) [31] as a basic model to detect laryngeal pathology and demonstrate how information available from this "black box" type model can be used to explore data and decisions made by the model.

3.1. Feature set

Aiming to obtain a comprehensive description, each audio recording is represented by 14 feature subsets resulting in a feature vector of 927 elements, see Table 1. Technical details of feature subsets 1–11 can be found in [29]. A short description of the 12th, 13th and 14th feature subsets is given below. The last three feature subsets were added to the previously used [29] aiming to increase diversity of features.

3.1.1. Reflection coefficients and vocal tract area irregularity features

Vocal tract is modelled by M tubes and feature computation is based on Mth order linear prediction filter. For each frame of voice recording, mth order prediction error E^m and area of mth tube A_m are computed using the Levinson–Durbin recursion algorithm:

$$A_m = A_{m+1}\frac{1+k_m}{1-k_m}, \quad m = M, \dots, 2, 1 \tag{1}$$

where $A_{M+1} = 1$ and k_m is the so-called Levinson–Durbin reflection coefficient. The 12th feature set is given by the Levinson–Durbin reflection coefficients. To obtain the 13th feature set, for each tube, the mean area $\bar{A}_m$, the variance of tube area S_m and the variance of area ratio S_{mr} are calculated using tube area values A_{mk} computed for different frames k of a voice recording [32]:

$$\bar{A}_m = \frac{1}{K}\sum_{k=1}^{K} A_{mk}, \quad m = 1, \dots, M \tag{2}$$