Vocal fold activity detection from speech related biomedical signals: a preliminary study

Ariel E. Stassi[†], Gabriel A. Alzamendi[†, ‡], Gastón Schlotthauer[†, ‡] and María E. Torres[†, ‡]

[†] Laboratorio de Señales y Dinámicas no Lineales, Facultad de Ingeniería, Universidad Nacional de Entre Ríos, Argentina
[‡] National Scientific and Technical Research Council (CONICET), Argentina

Abstract — *In vocal load estimation, detection of voiced speech regions is required in order to quantify the total time of vocal fold activity. This is a more difficult problem than voice activity detection, due to it involves the detection not only of the presence of speech but also of a periodic behavior at glottal level. In this work, we propose to use linear discriminant analysis in order to detect voiced speech periods. Here, three different signals, related to vocal fold activity, are considered: voice, electroglottogram and skin vibrations of the neck. For each signal, different sets of features are tested in order to find the corresponding optimal one. In this introductory study, the cross-validation procedures suggest that the proposed method is a suitable approach for voiced speech activity detection, independently of the considered signal, showing accuracies greater than 95 % and robustness to intersubject variability.*

Introduction

In this work, we present an approach based on linear classification to automatically detect periods of vocal fold activity in three different signals, simultaneously acquired, related to vocal folds: the voice wave (VW), electroglottogram (EGG), and skin vibrations of the neck (SVN). The method here proposed performs voiced speech detection, which is the first stage required to quantify the total phonation time in vocal dosimetry [5].

Materials and Methods

Database

Following an *ad hoc* protocol, we carried out a database acquisition. It is composed by simultaneous records of VW, EGG and SVN, from 43 subjects (14 females and 29 males) with non-pathological voices [1]. All the signals were obtained in an anechoic chamber and digitalized at a $50\,\mathrm{kHz}$ sampling frequency and $16\,bits$ quantization resolution.

Manual labeling of EGG records

Here, we considered EGG as the most suitable signal to detect periods of vocal fold activity. The software Audacity 2.0.3 was employed for manually labeling, by visual inspection, each EGG record and thus obtaining the corresponding reference class sequence (RCS), here considered as the "gold standard" method to assess classification performance. By definition, the RCS was constructed as follows: segments of vocal fold activity were labeled as 1, while the segments of silence or unvoiced speech were labeled as 0.

Feature extraction and linear discriminant analysis

Over each signal, it was performed a high-pass zero-phase digital filtering. For this, we considered an order 2 Butterworth approximation with a $75\,\mathrm{Hz}$ cut-off frequency. Hereafter, $x_m[n]$ refers to $20\,ms$ long signal frame (Hann window, $50\,\%$ overlap). From each frame, the following features were extracted and used by a Fisher's linear discriminant:

Full-band energy: $E_f = 10 \cdot \log_{10}\left(r[l]|_{l=0}/L\right)$, where $r[l]$ is the autocorrelation series of $x_m[n]$ –at lag l– and L is the frame length [2].

Low-band energy: $E_{lf}^{(i)} = 10 \cdot \log_{10}\left(\mathbf{h}_i^{\mathrm{T}}\,\mathbf{R}\,\mathbf{h}_i\,/L\right)$, where $\mathbf{R}$ is the Toeplitz autocorrelation matrix of $x_m[n]$ and $\mathbf{h}_i$ is the impulse response of a FIR filter with cut-off frequency at F_i Hz, where $F_i = 2^i \cdot 150$, with $i = 0, 1, \ldots, 4$ [2].

Normalized low-band energy ratio: $NER_{lf}^{(i)} = \dfrac{\mathbf{h}_i^{\mathrm{T}}\,\mathbf{R}\,\mathbf{h}_i}{r[l]|_{l=0}}$, where $\mathbf{h}_i$, $\mathbf{R}$ and $r[l]$ have already been described.

Zero-crossing rate: $ZCR = \dfrac{\sum_{n=1}^{L-1} |sf\{x_m[n]\} - sf\{x_m[n-1]\}|}{2L}$, where $sf\{\cdot\}$ is the signum function [2].

Spectral Flatness Measure: $SFM = 10 \cdot \log_{10}\left(\dfrac{GM\{|X_m[k]|\}}{AM\{|X_m[k]|\}}\right)$, where $X_m[k]$ is the discrete Fourier transform of $x_m[n]$, and $GM\{\cdot\}$ and $AM\{\cdot\}$ denote the calculation of geometric and the arithmetic means, respectively [4].

Results

At First, it was considered all the cases for $E_{lf}^{(i)}$ and $NER_{lf}^{(i)}$. Based on the best individual features criteria (implemented by Friedman and multiple comparison tests), only $E_{lf}^{(i)}$ (for $i = 0, 1, 2$) were preserved as the most representative features of the low-band frequency phenomena associated to vocal fold activity. Secondly, it was performed an exhaustive search with the 6 remaining features: E_f, $E_{lf}^{(i)}$ (for $i = 0, 1, 2$), ZCR and SFM. The tested cases were obtained by combining these 6 features in groups from 1 to 4 elements, subject to employ only one $E_{lf}^{(i)}$ each time. This resulted in 31 combinations to compare in order to decide on the best one. Two experiments of cross-validation were carried out in order to find the best feature set for each signal and characterize its performance.

In the first experiment, the subsets were obtained by random splitting of the whole used dataset. The results are shown in Table 1.

Signal	Feature Set	Accuracy	CI $(\alpha = 0{,}01)$
VW	$\{E_f; E_{lf}^{(1)}; ZCR; SFM\}$	$0{,}9535$	$(0{,}9521; 0{,}9548)$
EGG	$\{E_f; E_{lf}^{(0)}; ZCR; SFM\}$	$0{,}9604$	$(0{,}9592; 0{,}9616)$
SVN	$\{E_f; E_{lf}^{(1)}; ZCR\}$	$0{,}9547$	$(0{,}9534; 0{,}9560)$

Table 1 : Performance of the best linear classifier for each signal, in case of randomly mixed and split data. CI: confidence interval, calculated according to [3].

In the second experiment, the subsets were obtained dividing the whole used dataset by subjects. The results are shown in Table 2.

Signal	Feature Set	Accuracy	CI $(\alpha = 0{,}01)$
VW	$\{E_f; E_{lf}^{(1)}; ZCR; SFM\}$	$0{,}9532$	$(0{,}9519; 0{,}9545)$
EGG	$\{E_f; E_{lf}^{(0)}; ZCR; SFM\}$	$0{,}9601$	$(0{,}9588; 0{,}9613)$
SVN	$\{E_f; E_{lf}^{(1)}; ZCR\}$	$0{,}9543$	$(0{,}9530; 0{,}9556)$

Table 2 : Performance of the best linear classifier for each signal, in case of data divided by registered subjects. CI: confidence interval, calculated according to [3].

In the Fig. 1, it can be observed the performance of the selected classifiers over each signal, when a male volunteer reads a short phonetically balanced sentence.

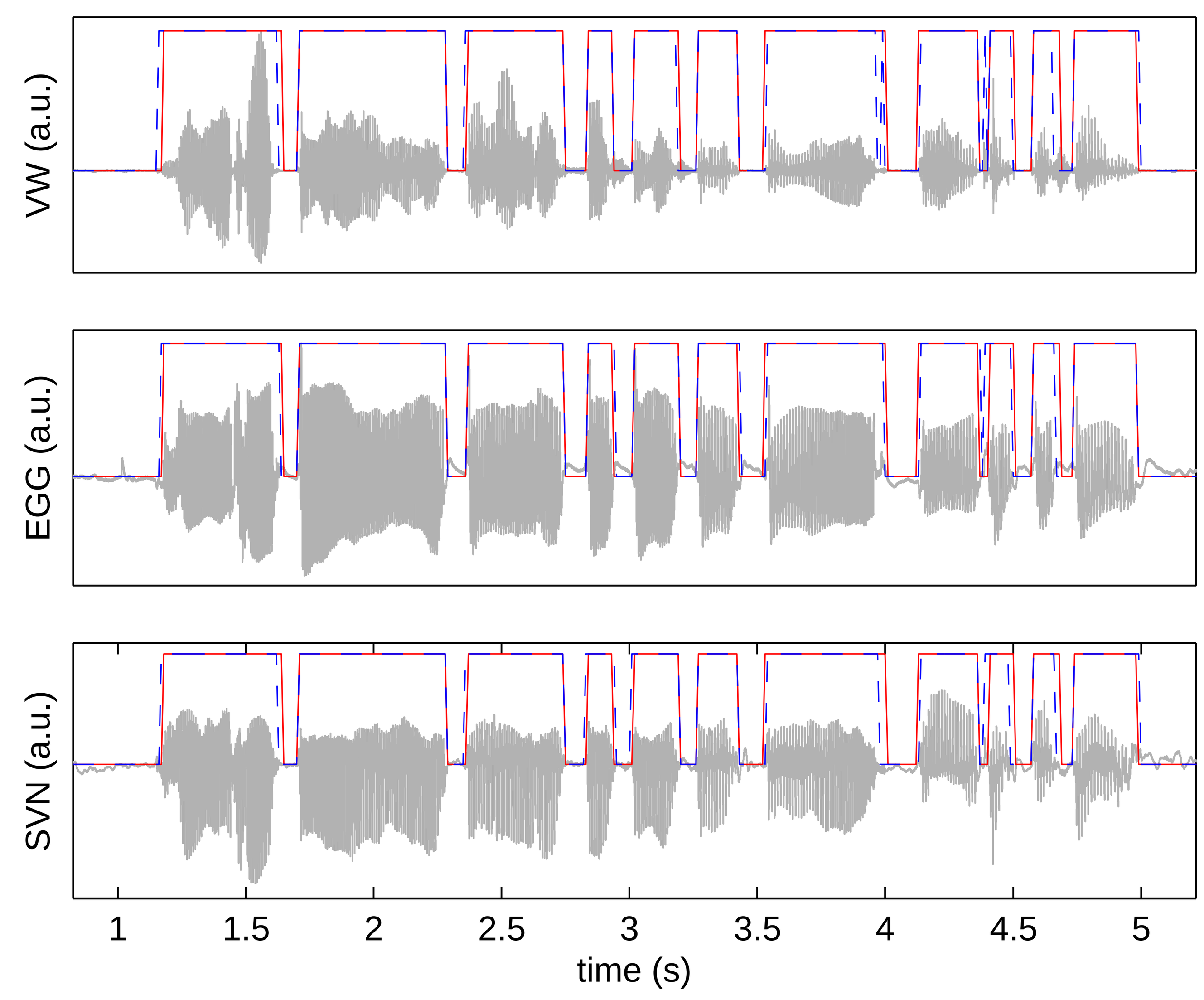

Figure 1 : Performance of our VSD method over each signal (blue dashed line) along with the corresponding RCS (red solid line).

From the first experiment, we can appreciate that the generalization capabilities of the obtained classifiers are very clear, showing the appropriateness of this method for this application. From the second experiment, we can conclude that our classification method does not depend on the subject considered.

Conclusions

In this work, we presented a method for the detection of voiced speech activity periods, based on a linear classification technique. From the cross-validation procedures, we can conclude that the method here proposed has a very good performance, with 0.9604 as the best accuracy value in EGG signal. Moreover, we showed that, for voiced speech detection, the SVN provides as much information as the VW. Nevertheless, it is known that SVN has demonstrated to be more practical than VW for ambulatory monitoring applications.

References

[1] Cheyne H. *Estimating glottal voicing source characteristics by measuring and modeling the acceleration of the skin on the neck.* Tufts University 2002.

[2] ITU-T G.729: Coding of speech at 8 kbit/s using conjugate-structure algebraic-code-excited linear prediction (CS-ACELP) *Annex B: A silence compression scheme for ITU-T G.729 optimized for terminals conforming to ITU-T V.70.* 2012.

[3] Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection in *14th IJCAI;*(Montreal, Quebec, Canada):1137–1143 1995.

[4] Moattar M, Homayounpour M. A simple but efficient real-time voice activity detection algorithm in *17th EUSIPCO;*(Glasgow, Scotland):2549–2553 2009.

[5] Svec J, Popolo P, Titze I. Measurement of vocal doses in speech: experimental procedure and signal processing *Logoped Phoniatr Vocol.* 2003;28(4):181–192.