Voiceprint analysis using Perceptual Linear Prediction and Support Vector Machines for detecting persons with Parkinson’s disease

ACHRAF BENBA, ABDELILAH JILBAB, AHMED HAMMOUCH
Laboratoire de Recherche en Génie Electrique, Ecole Normale Supérieure de l’Enseignement Technique, Mohammed V University, Rabat
MOROCCO
achraf.benba@um5s.net.ma

Abstract: - In the aim of developing the assessment of speech disorders for detecting patients with Parkinson’s disease (PD), we have collected 34 sustained vowel / a /, from 34 subjects including 17 PD patients. We subsequently extracted from 1 to 20 coefficients of the Perceptual Linear Prediction (PLP) from each individual. To extract the voiceprint from each individual, we compressed the frames by calculating their average value. For classification, we used the Leave-One-Subject-Out (LOSO) validation scheme along with the Support Vector Machines (SVMs) with its different types of kernels, (i.e.; RBF, Linear and polynomial). The best classification accuracy achieved was 82.35% using the first 13 and 14 coefficients of the PLP by Linear kernels SVMs.

Key-Words: - Voice analysis, Parkinson’s disease, Voiceprint, Perceptual linear prediction, Support Vector Machines, Leave One Subject Out.

1 Introduction
Parkinson's disease (PD) is currently the second most common neurological syndrome after Alzheimer’s disease. During its evolution, PD causes diverse symptoms and it influences the system which controls the execution of learned motor plans such as walking, talking or completing other simple tasks [1] [2] [3]. For this purpose, the assessment of the quality of speech, and the identification of the causes of its degradation in the context of PD based on phonological and acoustic features have become main anxieties of clinicians and speech pathologists. They have become more and more attentive to techniques or methods external to their domain, which might offer them extra information for the diagnosis and the assessment of this disease. As is known, PD generally causes voice weakening in 90% of patients [4] and touches people whose age is over 50 years, making the physical visits for diagnosis, monitoring and treatment extremely difficult [5] [6]. Clinicians and the speech pathologists have adopted subjective methods based on acoustic features to distinguish different disease states in PD patients. Recent studies use measurements of voice quality in time, spectral and cepstral domains [7] in order to develop more objective assessments to detect voice disorders in the context of PD. These measurements contain fundamental frequency of vocal oscillation (F0), absolute sound pressure level, jitter, shimmer, and harmonicity [1] [8] [9].

In this study we focused on the measurements and the assessments of speech disorders in cepstral domain by extracting the Perceptual Linear Prediction (PLP) coefficients. This method has been conventionally used in speaker recognition and identification applications. The PLP technique was first proposed by Hynek Hermansky [10]. In the last few years, the usage of the PLP has been extended to the assessment of speech quality for clinical applications. In this study, we have extracted PLP coefficients from the speech signals provided in a database and calculated the average value of the frames to get the voiceprint of each individual. We then used a Leave One Subject Out validation scheme with Support Vector Machines (SVMs) for feature classification in order to discriminate patients with PD from healthy subjects.

This paper is organized as follows: the subject database is described in section II. The PLP processes are presented in section III. The methodology of this study is presented in section IV. The obtained results are presented in Section V and conclusion in Section VI.

2 Data Acquisition
Dysarthria is the set of speech disorders associated with disturbances of muscular control of the speech organs. Dysarthria includes all malfunctions related to breathing, phonation, articulation, nasalization and prosody. These indications can be measured and detected by analyzing various features of voice. The data collected in the context of this study belong to 17 Parkinsonian patients (6 female, 11 male) and 17 healthy individuals (8 female, 9 male). Voice signals were recorded through a standard microphone at a sampling frequency of 44,100 Hz using a 16-bit sound card in a desktop computer. The microphone was placed at a 15 cm distant from subjects and they were asked to say sustained vowel /a/ at a comfortable level. All the recordings were made in mono-channel mode and saved in WAVE format; acoustic analyses were done on these recordings. All the voice samples were collected by Mr. M. Erdem Isenkul of Department of Computer Engineering at Istanbul University, Istanbul, Turkey.

3 The PLP Processes

Our first purpose in this section was to transform the speech waveform to some type of parametric representation for advanced analysis [11]. The speech signal is a real signal and it is finite in time; therefore, a processing is only possible on finite number of samples [12]. The first step of PLP process is to weight the speech segment by Hamming window [10] in order to reduce signal discontinuities, and make the ends smooth enough to connect with the beginnings [12]. This was done by applying Hamming window to taper the signal to zero in the beginning and in the end of each frame, by applying the following formula to the samples [10]:

$$W(n) = \left\{ 0.54 - 0.46 \cdot \cos \left( \frac{2\pi n}{N-1} \right) \right\}$$  \hspace{1cm} (1)

where $N$ is the length of the Hamming window, with a length about 20 ms.

The next processing step consists on converting each frame of N samples from time domain into frequency domain by using the Fast Fourier Transform (FFT) [11]. We applied the FFT for the reason that; it is a fast algorithm to implement the Discrete Fourier Transform (DFT) [11]. As known, the DFT is defined on the set of $N$ samples $(S_n)$ as follow [11]:

$$S_n = \sum_{k=0}^{N-1} s_k e^{-2\pi i k n / N} , n = 0,1,2,...,N-1$$  \hspace{1cm} (2)

The short-term power spectrum is calculated by adding the square of the real and imaginary components of short-term speech spectrum, as
3.2 Critical Band analysis
The short-term power spectrum \( P(\omega) \) is warped along its frequency axis \( \omega \) where \( (\omega=2\pi f) \), into Bark frequency \( \Omega \) by applying the following equation [10]:

\[
\Omega(\omega) = 6\ln\left(\frac{\omega}{1200\pi} + \sqrt{\left(\frac{\omega}{1200\pi}\right)^2 + 1}\right)
\]

(4)

\[
\Omega(f) = 6\ln\left(\frac{f}{600} + \sqrt{\left(\frac{f}{600}\right)^2 + 1}\right)
\]

(5)

\[
\Omega(f) = 6\sinh^{-1}\left(\frac{f}{600}\right)
\]

(6)

where \( \omega \) is the angular frequency in [rad/s], and \( f \) is the frequency in [Hz]. The next step consist on convolving the resulting warped power with the power spectrum of the simulated critical-band masking curve \( \Psi(\Omega) \) approximated by Hynek Hermansky [10] as follow:

\[
\Psi(\Omega) = \begin{cases} 
0 & \text{for } \Omega < -1.3 \\
10^{2.5(\Omega-0.5)} & \text{for } -1.3 \leq \Omega \leq -0.5 \\
1 & \text{for } -0.5 < \Omega < 0.5 \\
10^{-1.0(\Omega-0.5)} & \text{for } 0.5 \leq \Omega \leq 2.5 \\
0 & \text{for } \Omega > 2.5 
\end{cases}
\]

(7)

It is a rather curd approximation of the shape of auditory filters.

The samples of the critical-band power spectrum are produced by doing the discrete convolution of \( \Psi(\Omega) \) with \( P(\omega) \) using the following equation [10]:

\[
\Theta(\Omega) = \sum_{\omega=-1.3}^{2.5} P(\omega - \Omega) \Psi(\Omega)
\]

(8)

The convolution between the relatively broad critical-band masking curve \( \Psi(\Omega) \) and the short-term power spectrum \( P(\omega) \), decreases the spectral resolution of \( \Theta(\Omega) \) in comparison with the original \( P(\omega) \) [10].

3.3 Equal-loudness Preemphasis
The next step in this process is to preemphasis the samples \( \Theta(\Omega(\omega)) \) using the simulated equal-loudness curve as follow [10]:

\[
\Xi[\Omega(\omega)] = E(\omega) \times \Theta[\Omega(\omega)]
\]

(9)

where, \( E(\omega) \) is an approximation to the non-equal sensitivity of human ear perception at different frequencies. The practical approximation used in this research was adopted by Hynek Hermansky [10] and was first proposed by Makhol and Cosell [13] as shown in the following equation:

\[
E(\omega) = \frac{(\omega^2 + 56.8 \times 10^4)\omega^4}{(\omega^2 + 6.3 \times 10^4)^2 \times (\omega^2 + 0.38 \times 10^5)}
\]

(10)

\[
E(f) = \left[\frac{f^2}{f^2 + 1.6 \times 10^5}\right]^2 \times \left[\frac{f^2 + 1.44 \times 10^5}{f^2 + 9.6 \times 10^5}\right]
\]

(11)

3.4 Intensity-loudness Power Law
The next step is the cubic-root amplitude compression. The following formula approximates the power law of human hearing and simulates the non-linear relation between the intensity of sound and its perceived loudness [10]:

\[
\Phi(\Omega) = \Xi(\Omega)^{0.33}
\]

(12)

3.5 Autoregressive Modeling
In the last step of the Perceptual Linear Prediction process, \( \Phi(\Omega) \) is approximated by the spectrum of an all-pole model using the autocorrelation method of all-pole spectral modelling, this technique is called Linear Prediction [10] [14], in which the signal spectrum is modelled by an all-pole spectrum. In this study we used the Linear Predictive Coefficient analysis to compute the autoregressive model from spectral magnitude samples. The autoregressive coefficients are converted to cepstral coefficients of the all-pole model; this was realized by converting the LPC of 'n' coefficients into frames of cepstra [10].

3.6 Liftering
The important advantage of cepstral coefficients is that they are uncorrelated [12] [15]. However, the problem with them is that the higher order cepstra are fairly small [12] [15], as shown in Figure 3. For this purpose, it is important to rescale the cepstral coefficients to have quite similar scales (Figure 4) [12] [15]. This is realized by liftering the cepstra according to the following formula [12] [15]:

\[ c'_n = \left(1 + \frac{L}{2} \sin \left(\frac{\pi n}{L}\right) \right) c_n \]  

(13)

Where \( L \) is the Cepstral sine lifter parameter. In this work, we used \( L=0.6 \).

4 Methodology

The first phase in this study was to build a dataset containing voice samples recordings of normal individuals and patients with Parkinson’s disease. Ultimately, we collected 17 voices from both groups which gave us 34 records. All individuals (Normal and PD) were invited to pronounce the sustained vowel / a / at a comfortable level. We then extracted from each voice sample, multi cepstral coefficients of the PLP. The extracted number of coefficients ranged from 1 to 20. We proceeded in this way to get the optimal number of coefficients needed for the best classification accuracy. The PLP coefficients extracted from each voice sample contains a large number of frames which demand an extensive processing time for classification and prevents making the correct diagnostic decision [15]. To overcome this problem, we calculated the average value of these frames to get the voiceprint of each individual. To train and validate our classifier, we used a method of classification called Leave One Subject Out (LOSO), i.e., we left out all the compressed frames of the PLP coefficients of one individual to be used for validation as if it were an unobserved individual, and trained a classifier on the rest of the compressed frames of other individuals [6] [15] [16]. We used the LOSO classification scheme iteratively for each coefficient per subject until all 20 coefficients per subject. In this work, we used the SVM classifier with its different types of kernels, i.e.; RBF, Linear and polynomial. To measure the success of our classifier and select the best coefficients needed for the best diagnosis accuracy, we used an evaluation metrics which contain accuracy, sensitivity and specificity. Accuracy is the ratio of correctly classified instances divided to the whole instances [6] [16]:
where TP is the number of true positives (healthy), TN true negatives (pathological), FP false positives (pathological but it shown as healthy), and FN false negatives (healthy but it shown as pathological). Sensitivity is a statistical measure of correctly classified positive and Specificity is a statistical measure of negative instances [6] [16]:

\[
\text{Sensitivity} = \frac{TP}{TP + FN} \tag{15}
\]

\[
\text{Specificity} = \frac{TN}{TN + FP} \tag{16}
\]

### 5 Obtained results

Based on the obtained results, it is clear from figure 6 that when we use a larger coefficient number, the accuracy of diagnosis decreases using a classification with RBF kernels SVMs, this is also valid for sensitivity (Figure 7) and specificity results (Figure 8). The maximum accuracy achieved with RBF classification was 70.59% using the first coefficient.

The classification results using Polynomial Kernels SVMs are represented in Figure 7. A maximum classification accuracy of 70.59% was achieved using the first and the 13\textsuperscript{th} coefficients of the PLP. With this type of kernel we got the same maximum classification accuracy obtained using RBF kernel.

From the Figure 6, it is clearly observable that a maximum classification accuracy of 82.35% has been achieved using the first 13 and 14 coefficients of the PLP by Linear Kernels SVMs.

### 6 Conclusion

Dysarthria symptoms associated with Parkinson’s are a slow process whose first stages may go unnoticed. To enhance the assessment of Parkinson’s disease we collected a variety of voice recordings from different individuals during the pronunciation of sustained vowel /a/. The extracted PLP coefficients from different participants contain many frames which take maximum processing time in the classification process, and prevent making correct diagnosis.

The compression of the PLP frames using their average value to extract the voiceprints from individuals, has shown to be a good parameter for the detection of voice disorder in the context of Parkinson’s disease, showing a maximum classification accuracy of 82.35% using the first 13
and 14 coefficients of the PLP by Linear Kernels SVMs.

Acknowledgment

The authors would like to thank Mr. Erdem Isenkul from Department of Computer Engineering at Istanbul University. Thomas R. Przybeck and Daniel Wood, United States Peace Corps Volunteers (Morocco 2013-2015), and all of the participants involved in the dataset collection process.

References


BENBA Achraf, received his Master's degree in Electrical Engineering from “Ecole Normale Supérieure de l’Enseignement Technique” ENSET, Rabat Mohammed V University, Morocco, in 2013 he is a research student of Sciences and Technology of the Engineer in Ecole Nationale Supérieure d’Informatique et d’Analyse des Systèmes ENSIAS, Research Laboratory in Electrical Engineering LRGE, Research Team in Computer and Telecommunication ERIT at ENSET, Mohammed V University, Rabat, Morocco. His interests are in Signal processing for detection neurological disorders.
Abdelilah JILBAB is a teacher at the Ecole Normale Supérieure de l’Enseignement Technique de Rabat, Morocco; He acquired his PhD in Computer and Telecommunication from Mohammed V Agdal University, Rabat, Morocco in February 2009. His thesis is concerned with the Filtering illegal sites on the Internet: Contribution to the type of image recognition based on the Principle of Maximum Entropy. Since 2003 he is a member of the laboratory LRIT (Unit associated with the CNRST, FSR, Mohammed V University, Rabat, Morocco).

Ahmed HAMMOUCH received the master degree and the PhD in Automatic, Electrical, Electronic by the Haute Alsace University of Mulhouse (France) in 1993 and the PhD in Signal and Image Processing by the Mohammed V Agdal University of Rabat (Morocco) in 2004. From 1993 to 2013 he was professor in the Mohammed V-Souissi University in Morocco. Since 2009 he manages the Research Laboratory in Electronic Engineering. He is an author of several papers in international journals and conferences. His domains of interest include multimedia data processing and telecommunications. He is currently head of Department for Scientific and Technical Affairs in National Center for Scientific and Technical Research in Rabat (Morocco).