Classification of Healthy Subjects and Patients with Essential Vocal Tremor using Empirical Mode Decomposition of High Resolution Pitch Contour

Mekhala H S†, Yamini B.K†, Ketan J‡, Pal P‡, Shivashankar N† and Prasanta Kumar Ghosh §
†Department of Electronics and Communications, RVCE, Bangalore-560060
‡Department of Speech Pathology and Audiology, National Institute of Mental Health and Neurosciences (NIMHANS), Bangalore-560029
§Department of Neurology, National Institute of Mental Health and Neurosciences (NIMHANS), Bangalore-560029
hsmekhala.04@gmail.com, yaminikh@gmail.com, ketanjhunjhunjhunwala3@yahoo.co.in
palpramod@hotmail.com, n_shivashankar@yahoo.com, prasantg@ee.iisc.ernet.in

Abstract—We consider the task of automatic classification of healthy subjects and patients with essential vocal tremor (EVT) from a recording of sustained phonation. For the classification task, we propose a new set of acoustic features called pitch oscillation characteristics (POC) using empirical mode decomposition of high resolution pitch contour and its derivative. Classification experiments are performed on 25 healthy controls (HC) and 20 EVT patients using a support vector machine classifier and the proposed POC features. Experiments are also performed using a set of baseline features computed from the multi-dimensional voice program (MDVP). Classification accuracy obtained from the human experts are used for comparison too. The classification accuracy from human expert is found to be better than those from the automatic classification. However, it is found that, the average classification accuracy using a combination of the POC and baseline features is 63.66 % closer to the classification accuracy obtained from the experts compared to that using baseline features alone.

I. INTRODUCTION

Essential voice tremor (EVT) is a voice disorder that results from dyscoordination within the laryngeal musculature, which negatively impacts the symmetrical motion of the vocal folds [1]. This may result in low-frequency fluctuations of voice frequency, or amplitude or both [2]. EVT, which is known to be a chronic disease, currently has no cure. Research has shown that 18-30% of individuals diagnosed with Essential Tremor (ET) also develop EVT. Currently it has a global prevalence of 0.4% across all ages, and increases with age, rising to a prevalence of 6.4% after the age of 65 years [3]. It has been known to affect the quality of life [4] and timely intervention can assist in appropriate medication. In this study, automatic early detection of EVT symptoms in a non invasive manner from a person’s voice is attempted, which can be used by individuals as a preliminary test prior to consulting a medical expert.

Acoustic analysis of speech with EVT can be grouped into two categories: temporal and spectral analysis. There are a number of techniques for time series analysis of frequency and amplitude modulations for various studies to analyse vocal tremor [5], to determine the source of tremor [6], to understand perception of vocal tremor [7] and the physiology of change in vocal tract in general [8]. Similarly, spectral analysis of airflow and acoustic signals was used to quantify the frequency and amplitude of voice tremor and potentially to distinguish pathological tremor from normal ones [9]. Spectral measures of the amplitude envelope were also used to discriminate various sources of EVT [10]. Morlet wavelet transform was used to differentiate between parkinsonian and normophonic speakers based on tremor frequency and amplitude [11].

Typically, vocal cords vibrate at varying frequency (pitch) during speech. The pitch modulation in the presence of EVT has a perceptual characteristic different from that in the absence of EVT. The relation between pitch and EVT has been studied before in the literature [8][12]. Some popular measures that include features from pitch trajectories are jitter (frequency perturbation), shimmer (intensity perturbation), and harmonic-to-noise ratio (H/N), phonation range, maximum phonation time, signal to noise ratio (SNR) [13][14]. They are usually computed using a software called Multi-Dimensional Voice Program (MDVP, Kay Elemetrics Corp.) [15][16].

Earlier studies have shown that, in comparison to spoken words, sustained phonation provides better perceptual and acoustic characterisation of vocal tremor [17][18]. By using sustained phonation instead of spoken words, one can eliminate the effects of irrelevant factors such as speaking rate and temporal variation in acoustic characteristics due to the presence of different sounds which could mask the actual acoustic features indicative of EVT. Hence, in this study, we consider the pitch obtained from a sustained phonation. We hypothesise that the nature of variation of pitch at different time scales is a robust characteristic of EVT that could be different for EVT patients compared to that for healthy subjects and, thus, would aid in the classification of EVT patients and healthy subjects.
In order to obtain these robust signatures we examine pitch contour at different resolutions instead of quantifying pitch variation using parameters like jitter and shimmer. Cycle to cycle estimation of pitch based on glottal closure instants (GCIs) [19] is used unlike block based method [20] since the former is known to capture subtle pitch variations resulting in a high resolution pitch contour. This is illustrated in Figure 1, which shows an exemplary speech signal of two seconds duration for a sustained phonation from an EVT patient along with the pitch trajectories computed using both block based as well as high resolution cycle-to-cycle pitch estimation. It is clear that the fine variations observed in high resolution pitch trajectory is absent in the block based estimate although there is a gross similarity between the two.

To capture variations in different temporal scales in the cycle to cycle high resolution pitch, we use empirical mode decomposition (EMD). EMD is chosen over other popular multi-resolution techniques because of its adaptive nature that makes it an ideal tool to analyse non-stationary, non-linear real signals [21]. In our study we have chosen a bivariate implementation of EMD [22]. There have been works reported using EMD for automatic classification of normal and pathological voice [23][24] as well as Parkinsons and ET [25], however, EVT vs healthy control (HC) classification, to the best of our knowledge, has not been addressed in the past.

In this study, we have introduced a new set of attributes for acoustic analysis of EVT called pitch oscillation characteristics (POC). These features depend on the intrinsic mode functions (IMFs) obtained from the EMD of high resolution pitch contour and its first derivative. An algorithm has been developed to automatically classify EVT patients and HC based on POC features, and its performance has been compared with that based on popular acoustic (baseline) features computed from a pitch trajectory using MDVP. We have also compared the performance of automatic classification scheme with classification performance from human experts, i.e., speech pathologists. The classification accuracy from human experts was found to be higher than that obtained from the POC and baseline features. However, it was found that POC features together with baseline features resulted in a F-score, which was 63.66% closer to the F-score from human experts compared to that using baseline features alone.

II. PITCH OSCILLATION CHARACTERISTICS USING EMPIRICAL MODE DECOMPOSITION

Suppose \( y[n] \), \( 0 \leq n \leq N - 1 \) be the recorded signal (sampling rate \( F_s \)) of sustained phonation uttered by a subject where \( N \) is the total duration (in samples) of the recordings. At first, we compute a high resolution pitch contour denoted by \( p(n) \), \( 0 \leq n \leq N - 1 \). We also compute the derivative of the pitch contour denoted by \( p'(n) \equiv p(n) - p(n-1) \). We consider an analysis window of \( L \) samples and a shift of \( L_{sb} \) samples. In each analysis window, EMD of \( p(n) \) and \( p'(n) \) is computed to obtain six IMFs denoted by \( c_i(n) \) and \( c'_i(n) \), \( i = 1, \ldots, 6 \) respectively. Nine POC features (POC_P) are computed using different statistics of \( p(n) \) and \( c_i(n) \). Similarly another eight features (POC_DP) are computed from \( p'(n) \) and \( c'_i(n) \) resulting in a total of 17 POC features (POC_PDP). The steps for computing POC features are summarized in Figure 2. The high resolution pitch estimation and EMD based POC feature extraction are described in the following subsections.

A. High resolution pitch extraction using glottal closure instants

We compute a high resolution pitch contour using the glottal closure instants (GCIs). GCIs are computed by speech event detection using the residual excitation and a mean-based signal (SEDREAMS) [26], a technique that has been shown to be effective for GCI detection [27]. Suppose there are \( M \) GCIs in the recorded signal \( y(n) \). These are denoted by \( \eta_i, \ i = 1, \ldots, M, \) where \( 1 \leq \eta_1 < \eta_2 < \cdots < \eta_M \leq N \). Thus, the pitch at the \( i \)-th pitch cycle is computed as

\[
\pi_i = \frac{F_s}{\eta_{i+1} - \eta_i}, \quad 1 \leq i \leq M - 1
\]

Note that, \( \eta_i, \ 1 \leq i \leq M \) are non-uniformly placed. Thus, in order to obtain an estimate of the pitch contour at every sample index \( n \), we first perform linear interpolation between two successive GCI locations as follows:

\[
z(n) = \pi_i + \frac{\pi_{i+1} - \pi_i}{\eta_{i+1} - \eta_i} (n - \eta_i), \quad \eta_i \leq n \leq \eta_{i+1} \quad (2)
\]

\( z(n) \) represents the cycle-to-cycle variations of the pitch values. Since we intend to capture of the subtle variations in the pitch trajectory across cycles, we low-pass filter \( z(n) \) with

\[1\text{A Matlab implementation of SEDREAMS [28] is used in this work.}\]
a cut-off frequency of 60Hz. The cut-off frequency of 60Hz is chosen to preserve the fine pitch variations while at the same time discard noisy oscillations. A cut-off frequency lower than 60Hz is found to cause loss in the fine pitch variations. This filtered high resolution pitch contour \( p(n) \) and its first derivative \( p' (n) \) are used for computing POC features in every analysis window.

**B. Pitch oscillation characteristics (POC) features extraction**

A 6-level bivariate EMD [22,3] was applied on the high resolution pitch \( p(n) \), and its derivative \( p' (n) \) in an analysis window. EMD decomposes the input signal into IMFs, where each IMF represents oscillatory modes embedded in the signal. More than six levels of IMFs do not show oscillatory characteristics and, hence, are not included in this work. For illustration, the IMFs from a high resolution pitch contour in an analysis window are shown in Figure 4. Each of these IMFs is frequency and amplitude modulated and is non-stationary [21].

The \( p(n) \) and its derivative \( p'(n) \) are represented as a linear combination of 6 level IMFs \( c_i(n) \) and \( c'_i(n) \), \( i = 1,\ldots,6 \) respectively and the respective monotonic residue functions \( r(n) \) and \( r'(n) \) [21][30] as follows:

\[
p(n) = \sum_{i=1}^{6} c_i(n) + r(n),
\]

\[
p'(n) = \sum_{i=1}^{6} c'_i(n) + r'(n), \quad n = 1,\ldots,L
\]

The IMFs are extracted from the signal with the aid of the sifting algorithm [21]. To make sure that each \( c_i(n) \) holds enough amplitude and frequency modulation, stopping criteria from [31] is used.

Given an analysis window of duration \( L \) samples, nine POC_P features are computed using mean, log variance and log range of \( p(n) \) and the log variance of \( c_i(n) \), \( i = 1,\ldots,6 \) as follows:

\[
f_1 = \frac{1}{L} \sum_{n=1}^{L} p(n),
\]

\[
f_2 = \log \left( \frac{1}{L} \sum_{n=1}^{L} \left( p(n) - \frac{1}{L} \sum_{n=1}^{L} p(n) \right)^2 \right),
\]

\[
f_3 = \log \left( \max_{1 \leq n \leq L} p(n) - \min_{1 \leq n \leq L} p(n) \right),
\]

\[
f_i = \log \left( \frac{1}{L} \sum_{n=1}^{L} c_i(n) \right), \quad i = 4,\ldots,9
\]

Similarly, eight POC_DP features are computed using log variance and log range of \( p'(n) \) and the log variance of \( c'_i(n) \).

These are computed in a manner similar to \( f_i, \ 2 \leq i \leq 9 \) defined in equation (4). POC_P and POC_DP together yield the 17-dimensional POC_PDF features. Thus, given a recording of \( N \) samples, we obtain 17-dimensional features from each of the \( \frac{N-L}{F_{sch}} \) analysis windows.

**III. DATASET**

In this study, a group of 45 volunteers, consisting of 18 male (of age 48.41±9.81 years) and 27 female (of age 48.63±13.14 years) was used as subjects. There were 20 subjects with EVT (11 female, 9 male) and 25 HC (16 female, 9 male). Each subject’s voice was recorded in a quiet room. Care was taken to ensure that there was no electrical interference. The participants were seated comfortably, with their head held upright. Since phonation of sustained sounds has been shown to provide the best perceptual and acoustic characterisation of vocal tremor [17][18], the subjects were instructed to utter a sustained phonation of /a/ (the vowel in the word ‘car’), for as long as they could. The distance between the subject’s mouth and the microphone was kept approximately six inches throughout the recording. This was done to reduce the variation in the intensity of the sustained phonation due to varying mouth-microphone distance. The recordings were done at a sampling frequency of 16kHz, and later down sampled to 8kHz for processing. 50 milliseconds of audio at the beginning and end of each recording was discarded as these parts typically correspond to rise and fall of voice intensity in uttering sustained phonation. A total of 103 recordings were obtained, with an average recording duration of 4s. Number of recordings per subject, and duration of each recording is graphically shown in Figure 3. The number of recordings for a subject varied from 1 to 6. The maximum duration of a recording was 12.04 seconds for a female patient with EVT. Similarly, the minimum duration of a recording was 1.94 seconds for the first recording of a male patient with EVT.

**IV. EXPERIMENTATION AND RESULTS**

**A. Experimental setup**

For the experiments, the analysis window duration was chosen to be 2 seconds, i.e., \( L=16000 \) (\( F_s=8kHz \)). Similarly, the shift between consecutive windows was chosen to be 400 milliseconds, i.e., \( L_{sh}=3200 \). POC features in every analysis window were computed and used for classifying the respective window into patient with EVT or HC. Finally the decisions from multiple windows in a recording were combined using a majority rule. If the number of windows \( \frac{N-L}{F_{sch}} \) was even, the last window in a recording was excluded from the experiment to avoid a potential tie in the majority rule based classification. If the duration of the recording was less than 2 seconds, then the entire signal was considered as a single window.

All 103 recordings were divided into four groups each containing approximately equal number of male and female HC as well as EVT patients. A summary of the number of control and patient recordings in each group is given in Table I. A 4 fold cross validation was used to examine accuracy.

---

2 A 4th order lowpass digital Chebyshev filter with stopband ripple 40dB lower than passband ripple, and stopband-edge frequency of 60Hz was used.

3 A Matlab implementation of bivariate EMD [29] is used in this work.
of the classification. For each fold of cross validation, one group was used as the test set, and the remaining groups were used for training the support vector machine (SVM) classifier. A non-linear SVM using Gaussian radial basis function (RBF) was used. RBF kernel parameter ($\gamma$) and the cost of misclassification (c) for training the SVM were separately optimized for each fold of cross-validation. For this purpose, we used two different sets of linearly spaced values for $\gamma$ and c in the logarithmic scale, namely, $\{\log_2\gamma = -1 : 10\}$ and $\{\log_2c = -1 : 20\}$ respectively. The best choices of $\gamma$ and c were jointly determined by maximizing the accuracy on the training set. The trained SVM was then used to classify the recordings from the test set. It should be noted that each feature used in the classification was normalized to have zero mean and unit standard deviation.

As the baseline features in this study, we used six features from MDVP (Kay Elemetrics Corp), namely percent jitter, percent shimmer, Frequency Tremor Intensity Index (FTRI), $F_0$-Tremor Frequency (Fftr), Amplitude Tremor Intensity Index (ATRI), and Amplitude-Tremor Frequency (Fatr) [16]. These six features were extracted using Praat software [32][33]. Praat was chosen for computing these features because it has been shown in the past that the features obtained from Praat are either equally or more accurate compared to the MDVP [34][35][36][37][38][39].

Various feature sets were used for the experiments in this work, namely, the 6 baseline features (MDVP), the 9 features of high resolution pitch (POC_P) and 8 features of pitch derivative (POC_DP) (as explained in Section II-B), and a 22-dimensional feature set (MDVP_POC_PDP) combining 17 features of POC_PDP alongwith the 6 baseline features. The optimized input parameters $\gamma$ and c for each fold of MDVP, POC_P and POC_DP feature sets is shown in Table II. In case of MDVP_POC_PDP based classification, a decision level fusion was performed where the decisions from MDVP, POC_P, POC_DP were combined using a majority rule.

Apart from the SVM classification, expert evaluation was conducted with the aid of three speech pathologists from National Institute of Mental Health and Neurosciences (NIMHANS), Bangalore. An interactive online form was provided to the experts which allowed them to listen to an individual clip as many times as required. After listening to each clip, they were asked to submit their decisions on whether the voice was normal or had to be diagnosed as EVT. In order to examine the consistency in expert evaluation, a randomly chosen set of 20 recordings was given for evaluation in addition to 103 recordings. These 20 repeated recordings were used to find out the consistency of an expert by finding the number of repeated recordings where an expert provided a decision identical to that given to the same clip in the set of 103 recordings. For three experts, the consistency (in percentage) was found to be 70%, 85% and 100%.

For evaluation of the classification performance, F-score defined by precision (P) and recall (R) was used. Precision measures the ability of the classifier not to label an EVT sample as healthy, and recall measures the ability of the classifier to find all the EVT samples in the test set. F-score was calculated as the harmonic mean of precision and recall, with equal weightage provided to both as below:

\[
F = \frac{2 \times P \times R}{P + R}
\]
TABLE II

<table>
<thead>
<tr>
<th>Feature</th>
<th>Fold No</th>
<th>Avg (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MDVP</td>
<td>0.9000</td>
<td>0.8867</td>
</tr>
<tr>
<td>POC_P</td>
<td>0.9032</td>
<td>0.9333</td>
</tr>
<tr>
<td>POC_DP</td>
<td>0.9231</td>
<td>0.9300</td>
</tr>
<tr>
<td>MDVP</td>
<td>0.9500</td>
<td>0.9572</td>
</tr>
</tbody>
</table>

Thus, different IMFs capture different characteristics of pitch variations from two classes and, thus, provide different amount of discrimination. Due to discrimination captured at different time scales, the POC features when combined could provide more discrimination compared to those from the baseline MDVP features.

From Table III, it is clear that the average F-score from human expert is higher than that from F-scores obtained using different feature sets in this work. The F-score from human expert is 1.175 (absolute) higher than that using MDVP features while it is only 0.427 (absolute) higher than that using the combined features resulting in a 63.66% improvement in F-score with reference to that from the human expert when POC features are used compared to when they are not. This highlight the benefit of the POC features.

V. CONCLUSIONS & FUTURE WORKS

We propose POC features from high resolution pitch contour for classification of healthy subjects and patients with EVT. Through classification experiments using sustained phonation data from 25 HC and 20 patients with EVT, we show that the proposed features are complementary to the baseline features computed from MDVP. However, the contribution of each POC and MDVP feature has not been studied. Understanding of relative contribution of each feature could help in selecting features or performing weighted combination of features for improving classification performance further.

Fractional sample interpolation of GCIs obtained from the SEDREAMS could be used to improve resolution of fine variation in pitch. More varied data can be collected and used to study the influence of several relevant thresholds and parameters used in this study, to further automate the system by selecting a suitable range without human intervention. The role of other variants of EMD such as complementary ensemble EMD in place of bivariate EMD could also be studied.

REFERENCES

Fig. 5. Normalized histogram of each of the proposed POC and baseline MDVP features for EVT patients and HC. Title in each subplot indicate the feature for the respective.


