Analysis of Glottal Activity of High Arousal and Falsetto Voice

P. Gangamohan¹, Botsa Kishore Kumar² and Suryakanth V Gangashetty³

Signal Processing and Machine Learning Laboratory¹, Speech Processing Laboratory², Speech Signal Processing Laboratory³, Koneru Lakshamaiah Education Foundation, Hyderabad, India International Institute of Information Technology, Hyderabad², India Koneru Lakshamaiah Education Foundation, Vaddeswaram³, Vijayawada
ganga39@klh.edu.in¹, kishore.botsa@research.iiit.ac.in², svg@kluniversity.in³

Abstract

Speech produced in emotionally charged states like anger, happiness, shout, etc., belongs to high arousal. There are changes in the voice quality of production, mainly in the excitation source characteristics. The main features observed from the acoustic signal are increase in fundamental frequency ($F_0$), and changes in spectral characteristics. Increase in $F_0$ can also be intentional as in falsetto register. The voice in falsetto register is produced with different quality. A systematic analysis on the glottal activity of high arousal speech (angry, happy and shout) and falsetto is performed in this study. Changes in the excitation source characteristics are examined by analyzing the speech signals using the zero time windowing (ZTW) method. This method gives fine temporal variations of the spectral features. The ratio of high frequency spectral energy to low frequency spectral energy is used for this analysis. The results of the study show that the derived spectral features discriminate falsetto and high arousal speech effectively.

Index Terms: Modal register, falsetto register, high arousal speech, zero time windowing, glottal activity.

1. Introduction

In a naturally uttered sentence, pitch fluctuates throughout its production. These pitch fluctuations form the overall intonation pattern of the sentence. A speaker has a normal pitch range in which he/she changes the pitch as dictated by the intonation pattern. This could be considered as the modal ‘pitch key’ of the speaker. However, the same speaker may adopt a high or very high key wherein the whole pitch range is shifted up. All these pitch variations operate in ‘modal register’ [1].

From the perspective of psychology of speech perception, the ‘communication of affect’ is described mainly by three dimensions, arousal or activation, pleasure or valence, power or dominance [2], [3], [4]. The arousal dimension refers to the degree of voice/vocal intensity. High arousal/loud speech is produced by a speaker in situations like emotionally charged, communication over a long distance, speech produced in the presence of background noise, etc. The extreme degree of arousal level of speech is termed as shout. Speech produced in anger and happiness states also belong to high arousal state.

Pitch and loudness are interdependent attributes of human speech, although either of these quantities can be altered independent of each other [5], [6], [7]. It is a general tendency of speakers to raise their pitch while increasing voice intensity in natural communication [5]. This phenomenon has been verified by the studies [8], [9], [10], [11], [12], where the fundamental frequency ($F_0$, an acoustical correlate of pitch) is used for analysis of high arousal voices produced in shout, anger, happiness states. However, speakers can even produce a loud voice with controlled pitch, or a normal speech with higher pitch key [13]. Such voices are difficult to produce generally, but can be effectively done by trained speakers (specially singers) [14]. On the other hand, speakers trained/untrained might shift to ‘falsetto register’, where normal voice intensity speech is produced with higher pitch key. Falsetto register is generally studied in the ‘higher-than-speaking’ pitch range [15], [14]. Speakers can however vary their pitch key in falsetto, which are referred as low, mid and high-pitched falsetto cases [16], [17]. The $F_0$ of low and mid-pitched falsetto register might sometimes match with some segments of normal (modal) voice and its average $F_0$ might as well match with that of high arousal modal voices [16], [18].

The present study focuses on discriminating high arousal and falsetto voices using acoustically derived features. In either of these cases, an increase in average $F_0$ with respect to normal (modal) voice is observed. Such a discriminating algorithm can be utilized as a complementary information for emotion recognition systems, which attempt to discriminate emotions among low and high arousal levels. The current study may also be helpful in synthesizing voices with different voice qualities in...
modal and falsetto registers.

Several previous studies in the literature have focused on exploring the physiological changes in falsetto voice with respect to modal voice. It has been observed that the vocal folds configuration is significantly distinct for modal and falsetto registers [15], [19]. The closed quotient \((C_q)\) of vocal fold vibrations is observed to be relatively lesser for segments of falsetto voice compared to modal voice. There have also been a few studies exploring the relation between the vocal intensity and the open quotient \((O_q = 1 - C_q)\), which have reported a decrease in \(O_q\) values with increase in vocal intensity levels [20], [21], [12]. In these studies, the \(C_q\) and \(O_q\) values are measured with the help of electroglossograph (EGG) signal.

There have also been a few studies towards analysis of falsetto voice using features derived from the acoustic signal. The spectral features such as H1-H2 (difference in the magnitude of harmonics \(1^{st}\) and \(2^{nd}\)), H2-H4, spectral tilt and subharmonic to harmonic ratio, are analyzed for regions with similar \(F_0\) for modal and falsetto registers in [18]. The variations in most of these spectral features have largely been attributed to the interdependence with \(F_0\). The subharmonic-to-harmonic measure appears to vary between registers, independent of \(F_0\). There are similar studies on analyzing variations in spectral features of angry, happy and shout speech with respect to neutral/normal speech [22], [23], [12]. Similar trends have been reported for the features like formant frequencies, their respective bandwidths, spectral tilt, etc., for angry and happy speech [22], [23]. However, there are have not been many studies highlighting the differences among high arousal voice and falsetto voice.

This paper presents a systematic analysis on glottal activity for normal, falsetto and high arousal speech. We use the \(O_q\) for parameterization of the glottal activity, which is computed from the EGG signals. An other acoustically derived feature called the ratio of high to low frequency spectral energy \((\beta)\) computed over a spectra obtained from the zero time windowing (ZTW) [12], is used for the analysis. The ZTW method helps in computing the spectral characteristics at good temporal resolution [24]. In study [12], the \(\beta\) feature has been proposed for the analysis of loud and shout speech.

2. Database and the glottal activity analysis

The speech data used for this study is collected from a total of 5 speakers (3 males and 2 females), each is a research student in the Speech and Vision Laboratory at IIIT, Hyderabad. Recordings were made using a close speaking microphone and EGG equipment. The data is recorded using a sampling rate of 48000 samples per second at a resolution of 16-bits per sample. Each speaker is asked to speak on their own framed sentences in anger, happy, shout, neutral and falsetto (low, mid and high) modes. Part of this data has been used in earlier studies for the analysis of emotion speech [25].

The glottal parameter \(O_q\) extracted from the corresponding EGG signals is used for analysis of anger, happy, shout and falsetto voices. A glottal pulse with open and closed phases is shown in Fig. 1. Within a glottal cycle, the magnitude of a differenced EGG signal at glottal closing instant (GCI) observes a sharp gradient. This is due to the abruptness in closing mechanism of vocal folds. The parameter \(O_q (\gamma)\) used for analysis is given by:

\[
\gamma = \frac{t_0}{t}
\]  

where \(t_0\) and \(t\) are the open duration and the total duration within a glottal cycle. The average values of \(\gamma\) \((= A_\gamma)\) along with the average values of \(F_0\) \((= A_{F_0})\) of anger, happy, shout, neutral and falsetto utterances for five speakers are given in Table 1. The \(F_0\) contour of an utterance is extracted using the zero-frequency filtering (ZFF) method [26].

It can be observed from Table 1 that the values of \(A_\gamma\) decreases for shout and angry speech when compared to neutral speech. Although the \(A_\gamma\) values in the case of happy speech are in decreasing trend, they are closer to neutral. Yet it is a general observation that speech in shout, anger and happiness states has increasing trend in the average values of \(F_0\).

From the values of \(A_{F_0}\), it is evident that low, mid and high falsetto voices can be produced with distinct pitch variations. The values of \(A_{F_0}\) in the mid and high falsetto cases are clearly higher than values within the ‘speaking pitch range’. For all the cases of low falsetto, a significant increment in the values of \(A_\gamma\) can be observed. The \(A_\gamma\) values are also expected to follow an incremental trend with the ascending degree of falsetto. It can be seen that the production of high voice intensity speech is coupled with a decrease in the \(O_q\) values, and the production
of falsetto speech observes an increase in $O_d$. It is due to the cumulative effect of high intensity and production of falsetto, that the $A_o$ values do not increase much for the cases of mid falsetto, and have somewhat decreasing trend for the cases of high falsetto compared to low falsetto cases.

3. Acoustic feature analysis

We extend our analysis on falsetto and high intensity voices by investigating the excitation parameters. A general approach to derive the source characteristics from speech signals is using the inverse filtering (IF) technique [27]. However there are several limitations in inverse filtering, deriving accurate transfer function to cancel the effect of vocal tract system and obtaining accurate duration for open phase of a glottal cycle being a few [28] [29].

It is important to note that the glottal vibration characteristics are reflected in the vocal tract system response. The effective vocal tract length changes with the opening and closing of the vocal folds. This can be attributed to the periodic coupling and de-coupling of the subglottal cavity with the supra-glottal cavity during glottal vibrations. Such a behavior of the excitation characteristics can therefore be computed from a speech analysis technique with high spectral and temporal resolution.

We use the ZTW method which has the ability to capture spectral changes at a finer temporal resolution. In the ZTW method, a heavily decaying window in the temporal domain given by $h[n]$ in (2), is used to segment the speech signal.

$$h[n] = \frac{1}{8\sin\left(\frac{2\pi n}{N}\right)^2}$$  \hspace{1cm} (2)

The Hilbert envelope of the numerator of group delay function (HNGD) is computed at every sampling instant. The resultant spectrum exhibits a good spectral and temporal resolution.

The $\beta$ feature, which is the ratio of energies in the high (800–5000 Hz) and low (0–400 Hz) frequency bands of the HNGD spectra, averaged over a duration of 10 ms has been used for the study of loud and shout speech [12]. For these two cases, the effect of increased $C_q$ is observed in lowering the energy in the frequency band ranging from 0 to 400 Hz.

The $\beta$ contour for a few glottal cycles is shown in Fig. 2. The sharp gradients in the $\beta$ contour as shown in the figure is due to relatively larger magnitudes in higher frequency components. These large magnitudes in the high frequency band occurs due to the impulse-like behavior around the GCI locations. From Fig. 2, we observe that there are sharp peaks in the $\beta$ contour around the GCI which is otherwise smooth.

![Figure 2: (a) Segment of a speech signal. (b) Differenced EGG signal. (c) $\beta$ contour.](image)

The highest and the lowest values within in a glottal cycle are denoted by $\beta_c$ and $\beta_o$, respectively, depicting the respective strengths of glottal closing and opening. The values of $\beta_c$ and $\beta_o$ are obtained from segments with a duration twice of the average $F_0$ for voiced segments of speech signal. The average values of $\beta_c$ ($=A_{c\beta}$) and $\beta_o$ ($=A_{o\beta}$) from high SNR regions of voiced speech segments for normal, shout, angry, happy and falsetto modes are given in Table 2.
Table 2: The average values of $\beta_c$ ($A_{\beta_c}$) and $\beta_o$ ($A_{\beta_o}$) for normal, shout, angry, happy and falsetto modes.

<table>
<thead>
<tr>
<th></th>
<th>Neutral</th>
<th>Shout</th>
<th>Anger</th>
<th>Happy</th>
<th>Low falsetto</th>
<th>Mid falsetto</th>
<th>High falsetto</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speaker1</td>
<td>0.21</td>
<td>0.03</td>
<td>0.43</td>
<td>0.17</td>
<td>0.41</td>
<td>0.11</td>
<td>0.37</td>
</tr>
<tr>
<td>Speaker2</td>
<td>0.49</td>
<td>0.09</td>
<td>1.40</td>
<td>0.36</td>
<td>1.21</td>
<td>0.31</td>
<td>0.63</td>
</tr>
<tr>
<td>Speaker3</td>
<td>0.27</td>
<td>0.04</td>
<td>0.45</td>
<td>0.19</td>
<td>0.39</td>
<td>0.09</td>
<td>0.33</td>
</tr>
<tr>
<td>Speaker4</td>
<td>0.51</td>
<td>0.10</td>
<td>1.11</td>
<td>0.29</td>
<td>0.82</td>
<td>0.24</td>
<td>0.55</td>
</tr>
<tr>
<td>Speaker5</td>
<td>0.23</td>
<td>0.03</td>
<td>0.39</td>
<td>0.14</td>
<td>0.29</td>
<td>0.05</td>
<td>0.23</td>
</tr>
</tbody>
</table>

From Table 2, it is evident that there is significant increment in the $A_{\beta_c}$ values in shout and angry speech with respect to neutral speech. A similar, but not significant trend, can be observed in the case of happy speech. For these cases, we can infer that a decrease in $O_q$ always corresponds to an increase in $A_{\beta_c}$. This can be explained by the fact that a smaller value of $O_q$ (or $A_c$) results in higher glottal closure strength, and therefore a higher value of $A_{\beta_c}$.

For the case of low falsetto voice, small changes are observed in $A_{\beta_c}$ with respect to its values for neutral speech. This can be attributed to the fact that higher $O_q$ (or $A_c$) correspond to weaker glottal closure strength. Similar observations are expected from the cases of mid and high falsetto voices, but there appears to be an anomaly. The $A_{\beta_c}$ values for mid and high falsetto voices increase, and also they lie in the same range as in the anger, shout cases. This behavior may occur because of the implicit increment in intensity factor while attempting high pitch in falsetto voices which may affect the $\beta$ parameter. Also it might be because of relative decrease in $O_q$ in cases of mid and high falsetto with respect to low falsetto.

The $A_{\beta_o}$ values might appear in the same range for cases of anger, shout, mid falsetto and high falsetto cases. There is, however, a significant difference in the values of $A_{\beta_o}$ for these cases. The $A_{\beta_o}$ values for the cases of mid and high falsetto are much higher when compared with that for the anger and shout cases. From the above discussions, it is evident that the spectral features of speech obtained using the ZTW method indeed can highlight the source characteristics which can be utilized to discriminate among high arousal and falsetto speech.

4. Summary

In this paper, a systematic analysis of the glottal activity of high arousal speech and falsetto speech is carried out using electoglottograph (EGG) signals. In both the cases, the speech is produced with higher pitch, while latter being produced in controlled way. It is shown that there is significant differences in the glottal excitation characteristics like open quotient and close quotient among these cases. There is decrease in the open quotient in the case of high arousal speech, and increase in the open quotient in the case of falsetto speech. It is also interesting to note there is relative decrease in the open quotient in the high falsetto case with respect to the low falsetto.

The ratio of high frequency spectral energy to low frequency spectral energy ($\beta$) derived using the zero time windowing (ZTW) method is used for analysis. The ZTW method helps in computing the spectral characteristics with good temporal resolution. It is observed that when there is decrease in the open quotient (or increase in close quotient) in the case of high arousal speech, there is increase in values of average $\beta$ near the glottal closure instants. There is some decrease in $\beta$ values near the glottal closure instants in the low falsetto with respect to neutral speech, where there is increase in open quotient. Although the $\beta$ values near the glottal closure instants are in the same range among anger, shout, mid falsetto and high falsetto cases, there is significant difference in the values of $\beta$ values in the glottal open regions in these cases. This research study may provide complementary information for emotion recognition systems which try to discriminate among high arousal and neutral voices.

5. Acknowledgement

The authors would like to thank the Department of Science and Technology (DST), India for supporting Paidi Gangamohan through the project SRG/2020/001363.

6. References


