Workshop on Speech, Music and Mind 2018
01 September 2018, Hyderabad, India

Time-frequency spectral error for analysis of high arousal speech

P. Gangamohan, Suryakanth V Gangashetty, and B. Yegnanarayana

Speech Processing Laboratory
Language Technologies Research Center (LTRC)
International Institute of Information Technology, Hyderabad-500032, India
gangamohan.p@research.iiit.ac.in, svg@iiit.ac.in, yegna@iiit.ac.in

Abstract

High arousal speech is produced by speakers when they raise their loudness levels. There are deviations from neutral speech, especially in the excitation component of the speech production mechanism in the high arousal mode. In this study, a parameter, called the time-frequency spectral error (T_{Fe}) is derived using the single frequency filtering (SFF) spectrogram. It is used to characterize the high arousal regions in speech signals. The proposed parameter captures the fine temporal and spectral variations due to changes in the excitation source.

Index Terms: Excitation source, glottal closure instant, high arousal speech, single frequency filtering, time-frequency spectral error

1. Introduction

The single frequency filtering (SFF) method proposed in [1] has been shown to be effective in discriminating speech regions versus non-speech regions. In this method, the frequency shifted version of the speech signal is passed through a single pole filter, located at $\frac{f_s}{2}$ (i.e., half of the sampling frequency f_s) close to the unit circle in the z-plane. The amplitude envelope of the output signal is obtained at each frequency. This results in high frequency resolution in the 2-dimensional (2-D) time-frequency spectrogram. It helps in exploring the high signal-to-noise ratio (SNR) regions in the speech signal [1]. The 2-D time-frequency spectrogram computed from the SFF method has also been used for epoch extraction from the speech signals [2, 3].

During the production of voiced speech, the significant excitation occurs due to rapid closure of the vocal folds, which results in an impulse-like excitation [4]. The SFF based studies in [2, 3] exploit the nature of these impulse-like excitations.

In this paper, we propose a parameter, called the time-frequency spectral error (T_{Fe}) to capture the discontinuities in the 2-D time-frequency representation of the speech signal. Within a glottal cycle, the impulse-like excitation at the glottal closure instant (GCI) gives the highest time-frequency spectral error. The error is lower in the other region of the glottal cycle. This pattern is observed in the case of neutral speech signal, where the glottal excitation source follows a standard pattern of rapid glottal closure and gradual opening.

The T_{Fe} parameter is also examined for the case of high arousal speech signals. High arousal voice quality refers to increase in the loudness level of speech [5, 6]. High arousal speech is generally produced by a speaker in emotionally charged states, such as anger, happiness and shout. In the literature, it is reported that there are three major factors in the production of high arousal speech: Increase in the subglottal air pressure, increase in the abruptness of the closure and increase in the closed phase of glottis within a glottal cycle [7, 8, 9, 10]. These factors in the production of high arousal speech contribute to complex excitation pattern.

The paper is organized as follows: Section 2 presents the computation of T_{Fe} parameter. Analysis of neutral speech and high arousal speech using the T_{Fe} parameter is presented in Sections 3 and 4, respectively. Section 5 presents a summary of the paper.

2. Time-frequency spectral error parameter using single frequency filtering

In the SFF method, the speech signal $s[n]$ is multiplied with a complex sinusoid signal $e^{-j\frac{2\pi \hat{f}_k}{f_s}n}$. The resultant signal $s_k[n]$ is a frequency shifted (by $\hat{f}_k$) version of the signal $s[n]$. The signal $s_k[n]$ is passed through a single-pole filter $H(z)$. The pole of the filter is located on the negative real axis near to the unit circle in the z-plane, corresponding to $\frac{f_s}{2}$. The following are the steps involved in SFF:

$$H(z) = \frac{1}{1 + rz^{-1}}, \tag{1}$$

$$Y_k(z) = H(z)S_k(z), \tag{2}$$

$$y_k[n] = -ry_k[n-1] + s_k[n], \tag{3}$$

$$o[n,k] = \sqrt{y_{kr}^2[n] + y_{ki}^2[n]}, \tag{4}$$

where $y_k[n]$ is the output signal from the single-pole filter $H(z)$, $y_{kr}[n]$ and $y_{ki}[n]$ are the real and imaginary components of $y_k[n]$, respectively. Since filtering of $s_k[n]$ is carried out at $\frac{f_s}{2}$, the envelope corresponds to the frequency (f_k) component of the signal $s[n]$,

$$f_k = \frac{f_s}{2} - \hat{f}_k. \tag{5}$$

The 2-D representation of the SFF spectrogram is denoted as $o[n,k]$. The high frequency resolution of the SFF spectrum is due to narrow bandwidth of the filter, and this helps to obtain high SNR regions in time and frequency [1].

In recent studies [2, 3], the 2-D SFF spectrogram is shown to be effective in capturing the impulse-like excitation property embedded in the speech signal. For this purpose, a parameter called the spectral gain [2] also termed as time marginal [3] is used. At each sampling instant, the spectral gain is given by

$$\mu[n] = \frac{1}{K}\sum_{k=1}^{K} o[n,k], \tag{6}$$

where $n = 1, 2, 3, \ldots, N$, $k = 1, 2, 3, \ldots, K$, N is the total number of samples in the signal and K is the total number

10.21437/SMM.2018-4

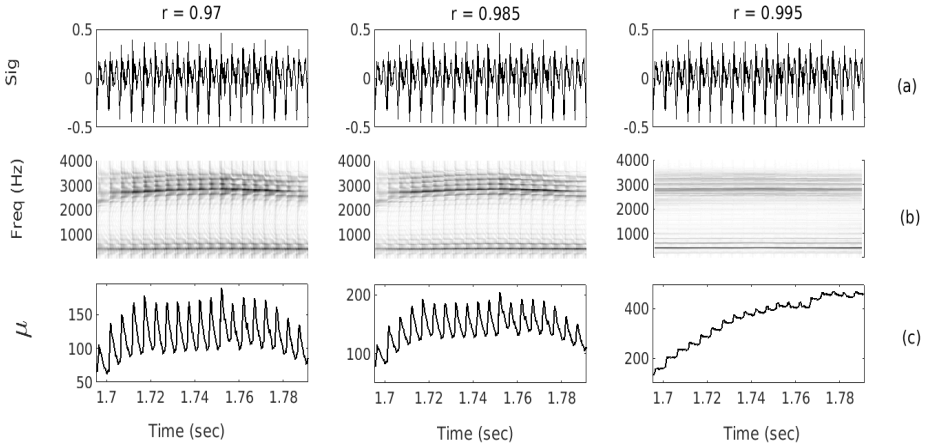

Figure 1: *Illustration of the effect of the pole radius (r) on the spectral gain (μ) parameter. First, second and third columns give the plots corresponding to r = 0.97, 0.985 and 0.995, respectively. Row (a) gives the speech signal segments (Sig). Row (b) gives the corresponding 2-D SFF spectrogram. Row (c) gives the corresponding μ contours.*

of desired frequencies. Let us briefly review the effect of r on the spectral gain parameter. In Fig.1, the speech signal with an average pitch period of 4.8 ms is passed through a single frequency filter using different values of r (r = 0.97, r = 0.985 and r = 0.995). Each column in Fig. 1 corresponds to a particular value of r. The speech signal is shown in Fig. 1(a), the SFF spectrogram in Fig. 1(b), and the spectral gain (μ) in Fig. 1(c). It can be seen form Fig. 1(c) that the amplitude values of $\mu[n]$ are smoothened out as the r value increases.

For extraction of excitation-related information, a parameter called the time-frequency spectral error is proposed. The following are the steps involved in the computation of the proposed time-frequency spectral error:

1. The speech signal is passed through a filter with two zeros, one zero is located on the unit circle at 0 Hz and the other zero is located on the unit circle at $\frac{fs}{2}$ Hz in the z-plane. This filter de-emphasizes the effects of low frequency channel disturbances and of aliasing.
2. The 2-D time-frequency SFF spectrogram is obtained.
3. The error values at the instant n_1 is computed as

$$e[n_1, k] = o[n_1, k] - \frac{1}{M} \sum_{i=0}^{M-1} o[n_1 - i, k], \; \forall \, k = 1, 2, ..., K, \tag{7}$$

   where M is chosen to be 0.3 times the length of the local pitch period.
4. The total error is called the time-frequency spectral error (T_{Fe}) at the instant n_1, and is given by

$$T_{Fe}[n_1] = \frac{1}{K} \sum_{k=1}^{K} |e[n_1, k]|. \tag{8}$$

The effect of the r value on the T_{Fe} contour is illustrated in Fig. 2. For the same speech signal as in Fig. 1, the SFF spectrograms and T_{Fe} contours for different values of r are shown in Fig. 2. Changes in the T_{Fe} contour are not significant for different values of r, even though r = 0.995 gives a smooth SFF spectrogram. As expected, the absolute values of T_{Fe} decrease with increase in the r value.

The speech signals (of neutral mode) with different average pitch periods, and their T_{Fe} contours obtained using r = 0.9999 are shown in Fig. 3. This helps in understanding the effect of the average pitch period on the shape of the T_{Fe} contour.

3. Analysis of neutral speech using T_{Fe} parameter

In this study, the speech signals corresponding to neutral, happiness, anger and shout modes are considered. The data covering these cases is collected from three databases, namely, the IIITT-H Telugu emotion database [11], the Berlin emotion database (EMO-DB) [12], and the IIIT-H shout data [10]. Across these databases, the number of utterances corresponding to neutral, happiness, anger and shout are 196, 110, 179 and 97, respectively.

The nature of the T_{Fe} contour is that it captures discontinuities in the speech signal. The value of T_{Fe} is higher at the glottal closure instants (GCIs) as shown in Fig. 3. It is also observed that the values of T_{Fe} at the GCIs vary for different voiced sounds. In any spoken utterance, there are intrinsic variations among the sound units [13, 14, 15, 16, 17]. In [13], it is shown that the narrow stricture and complete closure in the vocal tract during the production of voiced fricatives and voiced plosives, respectively, result in increase in intra oral air pressure. This creates abducting forces on the upper surface of the vocal folds. As a result, the glottal closure is less abrupt in the voiced fricatives and voiced plosives when compared to the vowel regions [14]. The effect of abruptness of the glottal closure in the vowel and non-vowel regions is studied using the amplitude of T_{Fe} at the GCIs. Five utterances (of neutral mode) corresponding to a female speaker are taken. The vowel and non-vowel voiced regions are segmented manually. The distribution of T_{Fe} at the GCIs in the vowel and non-vowel voiced regions is shown in Fig. 4. From this figure, it is clear that the values of T_{Fe} are higher for the vowel regions. This indicates that the T_{Fe} values capture the discontinuities in the speech signal caused due to excitation.

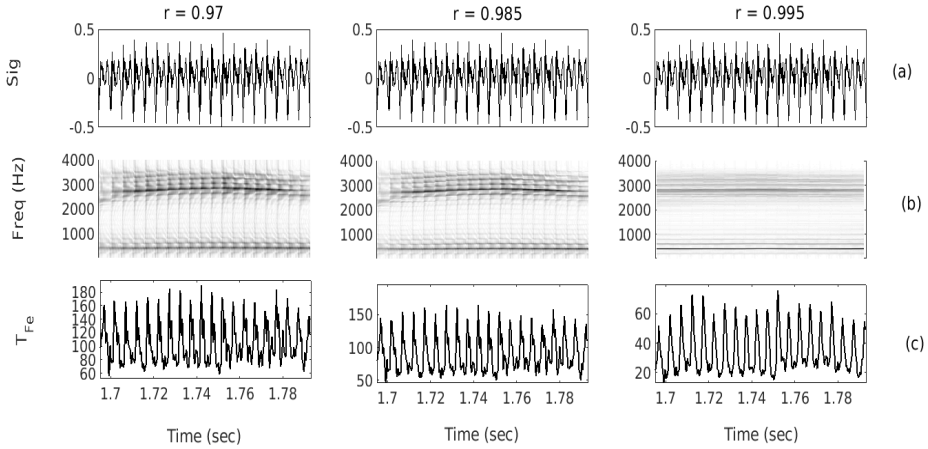

Figure 2: *Illustration of the effect of the r value on the time-frequency spectral error (T_{Fe}) parameter. First, second and third columns give plots corresponding to r = 0.97, 0.985 and 0.995, respectively. Row (a) gives the speech signal segments (Sig). Row (b) gives the corresponding 2-D SFF spectrogram. Row (c) gives the corresponding T_{Fe} contours.*

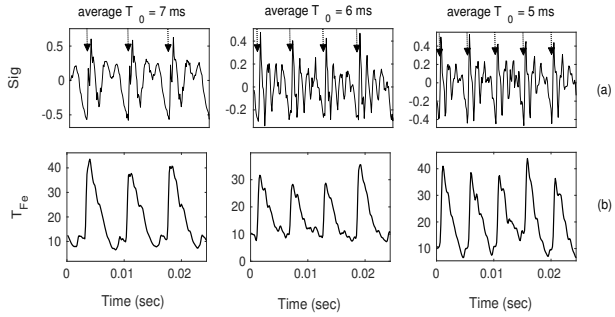

Figure 3: *Illustration of the effect of the pitch period (T_0) on the T_{Fe} parameter. First, second and third columns give plots corresponding to average T_0 = 7 ms , 6 ms and 5 ms, respectively. Row (a) gives the speech signal segments (Sig). Row (b) gives the corresponding T_{Fe} contours. The GCIs in the speech signals are marked. Note that the T_{Fe} contours are given for r = 0.9999.*

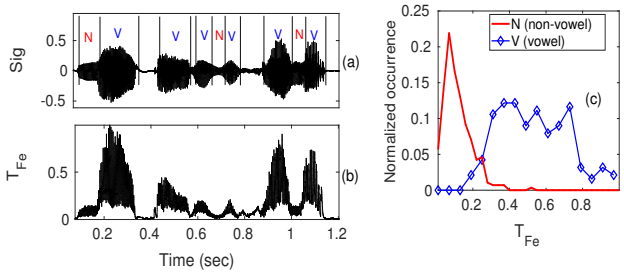

Figure 4: *(a) A sample speech signal with segmented vowel (V) and non-vowel (N) regions. (b) T_{Fe} contour of the signal shown in (a). (c) Distribution of T_{Fe} (obtained at the GCIs) for the vowel and non-vowel segments. Note: The T_{Fe} contour of an utterance is normalized with its maximum value.*

4. Analysis of high arousal speech using T_{Fe} parameter

The high arousal quality in the speech signal is expected to be reflected more in the vowel-like regions. In this section, the signal regions with higher values of T_{Fe} are chosen for analysis.

Changes in the T_{Fe} contour in a glottal cycle are illustrated in Fig. 5. The T_{Fe} contour of each glottal cycle (starting at the GCI) is amplitude normalized. Figs. 5(a), 5(b), 5(c) and 5(d) show superposition of normalized T_{Fe} contours of glottal cycles corresponding to neutral, happiness, anger and shout cases, respectively. Note that each glottal cycle is resampled to 5 ms using a discrete Fourier transform (DFT) based resampling method [18]. The data considered for this plot covers 5 utterances each of neutral, happiness, anger and shout cases of a female speaker.

A general observation is that the T_{Fe} values are relatively higher across the glottal cycle in the case of high arousal speech when compared to neutral speech. To capture the deviation in the normalized T_{Fe} contour of a glottal cycle, a reference contour is used. Fig. 6 shows the reference contour R_{Fe}, which is derived using the neutral speech signals of 12 (6 male and 6 female) speakers across all the databases. The deviation at each instant in the glottal cycle is obtained by subtracting the R_{Fe} contour from the normalized T_{Fe} contours. The positive values in the resultant contour d_e implies higher time-frequency error in the glottal cycle.

In the d_e contour of a glottal cycle, the maximum amplitude (M_d) value in the region between 1 ms to 4 ms is considered. Distribution of M_d for two speakers (male and female), for all the categories is shown in Fig. 7. From Figs. 7(a) and 7(b), it can be observed that there is an increasing trend in the M_d values in the case of happiness, anger and shout. A threshold value $\epsilon_d = 0.25$ is used in discriminating the high arousal segments, especially in the case of anger and shout cases. The following are the steps implemented in the proposed algorithm:

- In a speech signal, obtain the voiced regions and the GCIs using the zero frequency filtering (ZFF) method [19].
- Compute the T_{Fe} contour, and obtain the normalized T_{Fe} contour for each glottal cycle.

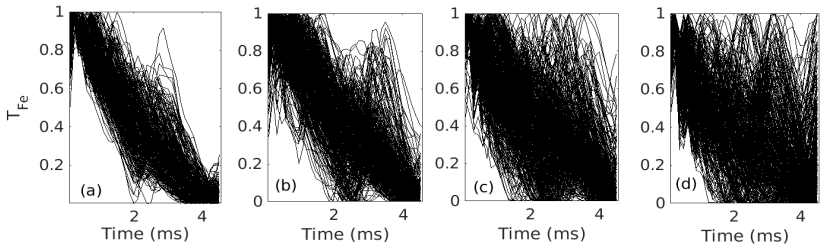

Figure 5: *Superposition of T_{Fe} contours of glottal cycles corresponding to (a) neutral, (b) happy, (c) angry, (d) shout utterances.*

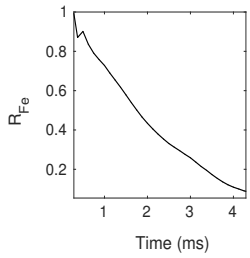

Figure 6: *A reference function R_{Fe}.*

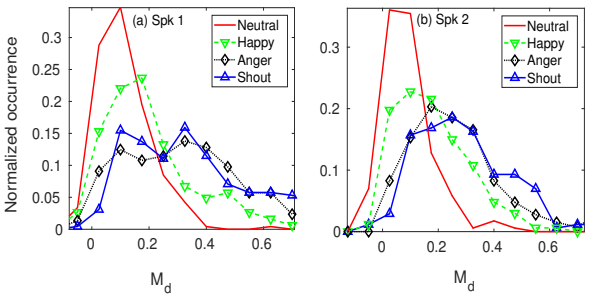

Figure 7: *Distribution of M_d for neutral, happy, angry, and shout utterances of male speaker (a) and female speaker (b).*

- Obtain the M_d value for each glottal cycle.
- In a segmented voiced region, if 30% of the glottal cycles have $M_d > 0.25$ then it is termed as high arousal.

Performance of the algorithm in terms of number of voiced regions detected as high arousal regions is given in Table 1. The percentage of regions detected as high arousal in the cases of neutral, happiness, anger and shout are 2%, 18%, 31% and 44%, respectively. It is important to note that all the voiced regions in the high arousal utterance might not carry high arousal characteristics. The results indicate that the time-frequency spectral error parameter gives information related to the excitation source.

Table 1: *Number of voiced regions detected as high arousal regions across all the speech signals of corresponding categories.*

	Total segmented voiced regions	Identified high arousal regions
Neutral	822	18 (2%)
Happy	411	72 (18%)
Angry	603	188 (31%)
Shout	384	169 (44%)

5. Summary

In this paper, high arousal speech signals are analyzed using the time-frequency spectral error (T_{Fe}) parameter. This parameter is obtained from the 2-D single frequency filtering (SFF) spectrogram. For different values of pole radius (r) of the resonator and pitch period (of the speech signal), changes in the T_{Fe} contour are not significant. In the case of neutral speech, the highest time-frequency spectral error (T_{Fe}) is observed at the glottal closure instants (GCIs). The error is observed lower in the remaining region of the glottal cycle. In the case of high arousal speech signals, the T_{Fe} values are observed relatively higher in the region other than around the GCI in a glottal cycle. The results from this study indicate that the proposed parameter is useful in identifying high arousal regions in the speech signals.

6. Acknowledgements

The authors would like to thank Tata Consultancy Services (TCS) for funding the first author for his PhD programme.

7. References

[1] G. Aneeja and B. Yegnanarayana, "Single frequency filtering approach for discriminating speech and nonspeech," *IEEE Trans. Audio, Speech and Language Processing*, vol. 23, no. 4, pp. 705–717, 2015.

[2] S. R. Kadiri and B. Yegnanarayana, "Epoch extraction from emotional speech using single frequency filtering approach," *Speech Communication*, vol. 86, pp. 52–63, 2017.

[3] C. M. Vikram and S. R. M. Prasanna, "Epoch extraction from telephone quality speech using single pole filter," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 3, pp. 624–636, March 2017.

[4] K. S. R. Murty and B. Yegnanarayana, "Epoch extraction from speech signals," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 16, no. 8, pp. 1602–1613, 2008.

[5] K. R. Scherer, "Vocal correlates of emotional arousal and affective disturbance," *Handbook of social psychophysiology*, pp. 165–197, 1989.

[6] M. Schröder, "Speech and emotion research: An overview of research frameworks and a dimensional approach to emotional speech synthesis," Ph.D. dissertation, Saarland University, Germany, 2004.

[7] P. Ladefoged and N. P. McKinney, "Loudness, sound pressure, and subglottal pressure in speech," *J. Acoust. Soc. Am.*, vol. 35, no. 4, pp. 454–460, 1963.

[8] I. R. Titze, "On the relation between subglottal pressure and fundamental frequency in phonation," *J. Acoust. Soc. Am.*, vol. 85, no. 2, pp. 901–906, 1989.

[9] C. Dromey, E. T. Stathopoulos, and C. M. Sapienza, "Glottal airflow and electroglottographic measures of vocal function at multiple intensities," *J. Voice*, vol. 6, no. 1, pp. 44–54, 1992.

[10] V. K. Mittal and B. Yegnanarayana, "Effect of glottal dynamics in the production of shouted speech," *J. Acoust. Soc. Am.*, vol. 133, no. 5, pp. 3050–3061, 2013.

[11] S. R. Kadiri, P. Gangamohan, S. V. Gangashetty, and B. Yegnanarayana, "Analysis of excitation source features of speech for emotion recognition," in *Proc. INTERSPEECH*, Dresden, Germany, 2015, pp. 1032–1036.

[12] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *Proc. INTERSPEECH*, Lisbon, Portugal, 2005, pp. 1517–1520.

[13] K. N. Stevens, *Acoustic Phonetics*. MIT Press, 1998, vol. 30.

[14] V. K. Mittal, B. Yegnanarayana, and P. Bhaskararao, "Study of the effects of vocal tract constriction on glottal vibration," *J. Acoust. Soc. Am.*, vol. 136, no. 4, pp. 1932–1941, 2014.

[15] D. H. Whalen and A. G. Levitt, "The universality of intrinsic f 0 of vowels," *Journal of Phonetics*, vol. 23, no. 3, pp. 349–366, 1995.

[16] H. M. Hanson, "Effects of obstruent consonants on fundamental frequency at vowel onset in englisha)," *J. Acoust. Soc. Am.*, vol. 125, no. 1, pp. 425–441, 2009.

[17] D. B. Fry, *The physics of speech*. Cambridge University Press, 1979.

[18] K. S. Rao and B. Yegnanarayana, "Prosody modification using instants of significant excitation," *IEEE Trans. Audio, Speech and Language Processing*, vol. 14, no. 3, pp. 972–980, 2006.

[19] N. Dhananjaya and B. Yegnanarayana, "Voiced/nonvoiced detection based on robustness of voiced epochs," *IEEE Signal Processing Letters*, vol. 17, no. 3, pp. 273–276, 2010.