SLOPE AT ZERO CROSSINGS (ZC) OF SPEECH SIGNAL FOR MULTI-SPEAKER ACTIVITY DETECTION

V. Subba Ramaiah[1] and R. Rajeswara Rao[2]

[1]Department of CSE, MGIT, JNTUH, Hyderabad, A.P
[2]Department of CSE, JNTUK-UCEV, T.S

ABSTRACT

Multi-Speaker activity (MSA) detection helps in detecting the presence of whether the speech signals has a single speaker or multiple speaker speeches in the speech signal. It is easy to calculate the slope at ZCs (zero crossings) of the speech signal and makes a comparison with a suitable threshold (Th). Multi-speaker is declared as and when the zero crossing value exceeds the threshold. The impact of the proposed technique is compared to the existing technique by calculating the sample-by-sample ZCR (Zero crossing rate) value is demonstrated. Experimental results prove that the proposed ZCR technique achieves accurate results than the traditional techniques for MSA detection that uses the cepstrum resynthesis residual magnitude (CRRM) in the literature.

KEYWORDS

Feature Extraction, Multi-Speaker Activity Detection, Zero Crossing Rate.

1. INTRODUCTION

The task of Multi-Speaker activity (MSA) detection from speech refers to whether the input speech has single speaker speech or a multiple number of speakers speech. Many studies have tried to address this problem [1]. There are a number of applications for MSA detection ranging from speaker identification or recognition to speech recognition in multi-speaker speech scenario [2]-[4].

MSA detection task is useful in speaker recognition [5], [6], in the sense that if the input speech has either single or multiple number of speakers. First there is a need for identifying the input speech whether it has a single speaker or multiple speakers speech and then after identifying the multi-speaker. Separation of the speech of the individual speakers from that of the multi-speaker speech signals are essential [7].

Once the region of multi-speaker speech is identified, there is a need for detecting the number of speakers available in the multi-speaker speech information [8]-[9]. This problem is referred in the literature as a detecting number of speakers from the speech signal. For this problem, a number of speaker detection, the task of MSA detection is like preprocessing stage before detecting the number of speakers.

Once the number of speakers are identified, we can separate each speaker speech information for speaker recognition or speech recognition. Generally, from the studies [10], we can observe that

DOI: 10.5121/ijci.2016.5415

there is a need for enhancing each speaker speech before further proceeding for speaker recognition or speech recognition. The study of literature shows voice activity detection in multi-speaker speech scenarios such as data available at meetings, cocktail party problem, etc [11].
In [12], the authors have presented Double talk (DT) detection method in Acoustic sound cancellers (AEC). Handling DT detection is the major issue that has been renamed at the forefront is AEC development. If both the speakers speak simultaneously, then DT happens. In order to cancel far-field echo, an adaptive filter is developed in an AEC. The convergence is strongly influenced by the very presence of near-field signal. In the commercial AECs, DT detectors are commonly available. DT detector methods can be reviewed in [13]-[17].

Recently in [18], the authors have approached a method based on the measurement of the ZCR. A comparison has been done for DT detection. They employed normalized least-mean square (NLMS) formed AEC for DT detection method and is therefore, zero crossing rate is estimated over a samples of window and then it is updated. This method requires cost effective and its convergence rate is slow.

In this study, we have chosen the solution using the slope at ZCR of the speech signal. It is observed that, lower number of ZCR can be seen for a single speaker, whereas the higher number of ZCR is observed for multiple speakers.

The outline of the paper is as follows: The database used for this study is presented in next Section. In Section III, the feature extraction methods which include ZCR and cepstrum resynthesis residual magnitude (CRRM) are discussed. The proposed methodology for ZCR is presented in Section IV. The detailed experimental results are presented in Section V. Finally, the conclusion of the present work is mentioned Section VI.

2. DATABASE USED FOR STUDY

The database consists of 280 speech samples recorded by 28 speakers. Each speaker utters an isolated digit ranging from 0 to 9 (10 digits), in English. Out of 28 speakers, 14 are male and 14 are female. This database includes five conversations for each of the single male, single female, male-male, female-female, and male-female (mixed) speaker conversations. In this study, we considered multiple speakers data as male-male (14 combinations), female-female (14 combinations), and male-female (14 combinations) speech data. Hence, total 42 multiple speaker speech data are considered. The speech samples are recorded directly over an android cellular phone in a sound proof room. All the audio signals are stored in the Wav format with an 8 kHz sampling rate, bit rate of 16 bits and in mono (single channel) format. The average duration of the samples is about 3 seconds per speaker.

3. FEATURE EXTRACTION

3.1. Zero Crossing Rate (ZCR)

In discrete-time signals, when 2 successive samples show against signs, zero crossing happens. The total number of zero crossings per sample is measured by ZCR. In an analysis, M samples of window size are used. The window size is calculated using ZCR divided by M. The ‘i’ is an instance of time, the ZCR of a discrete-time signal x(i) and slope (μ) are described as

$$ZCR(y(i)) = \frac{1}{2M} \sum_{j=i-M+1}^{i} \left| \sin(y(j)) - \sin(y(j-1)) \right| . W(i-j) \quad (1)$$

Where signum (sin) function is given as

$$\sin(y(i)) = \begin{cases} 1, & y(i) \geq 0 \\ 0, & y(i) < 0 \end{cases} \quad (2)$$

$$Slope(\mu) = \frac{1}{ZC} \sum_{i=1}^{ZC} \left| y(i+1) - y(i-1) \right| \quad (3)$$

Here the ZC is number of zero crossings and window parameter is $W(i)$. When 2 signal samples have different sign and absolute function is equivalent to 2, then zero crossing occurs. The right-hand side summation of equation (1) is equal half of the ZCR. After deciding zero crossing rates from one end to other of M window samples, get the next estimate of short-time window by incrementing K samples.

The short-time autocorrelation and short-time energy are not only used to show time-domain signals, but also for ZCR. In past, zero crossings have been employed to show the discrimination between speech and speaker. Frequency approximation of sinusoidal signals is one of the prime importances of ZCR. If there is a higher frequency, it results in higher short-time ZCR, and if there is a low-frequency signal, it results low short-time ZCR.

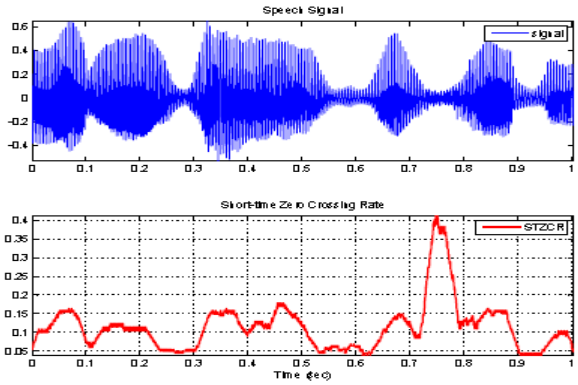

Figure 1. Speech signal and its short-time zero crossing rate for a single female speaker.

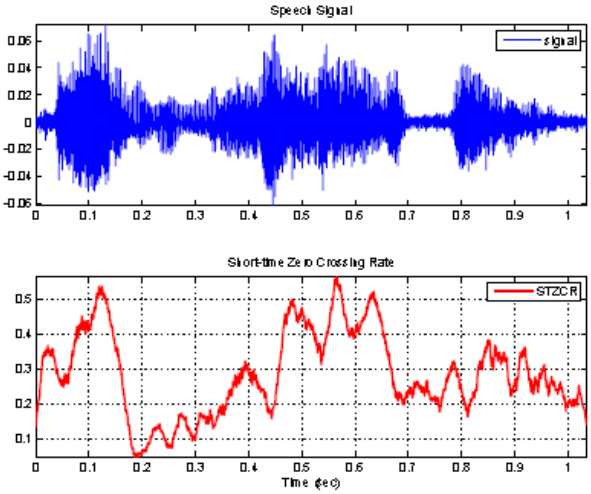

Figure 2. Speech signal and its short-time zero crossing rate for a single male speaker.

Zero crossing rate (ZCR) means the number of times the signal level crosses 0 during a constant period of time (i.e 1sec.) and is used not only for speech but also used for different detection applications. Similarly to amplitude level, a ratio of the input frame to noise is used for this feature. For this application rate at which zero crossing happens was calculated by taking a window of 20 msec. For an illustration, we show a single female speaker speech signal and its corresponding ZCR in Figure 1. Here the first Figure is a speech signal and the next Figure is a short-time zero crossing rate (STZCR). Similarly, we show a single male speaker and mixing of two male speaker speech signals and its corresponding STZCR in Figure 2 and Figure 3.

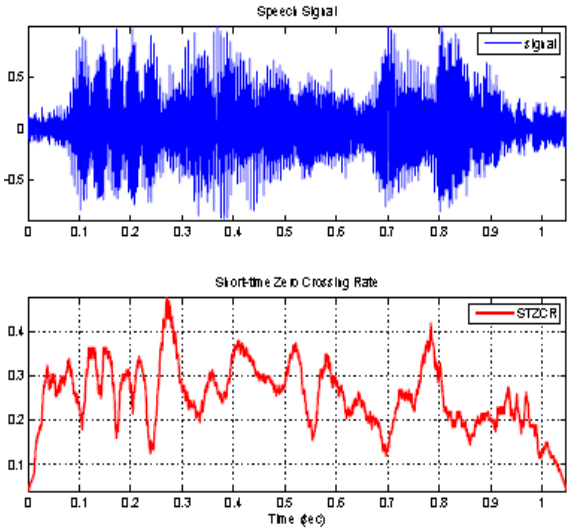

Figure 3. Speech signal and its short-time zero crossing rate for mixing of two male speakers

2.2. Cepstrum Resynthesis Residual Magnitude (CRRM)

CRRM is outlined as the L2-norm of the distinction between the absolute of the smoothed spectrum (M) and the Short Time Fourier Transform spectrum (S) evaluated at a speech sound frame. M is calculated using the real cepstrum (C):

$$C = real(FFT^{-1}(\log(|S|))) \tag{4}$$

$$C^1 = C.W \tag{5}$$

$$M = \exp(FFT(C^1)) \tag{6}$$

Here W is a window function with value 0 or 1. Only the first 'n' coefficients and the latest 'n' coefficients of C are considered. A 1048-point Fast Fourier Transform is performed on C and an 8 kHz sampling rate with n = 50 is used. M is basically a low pass filtering of the spectrum (S), so that is a better fit for noise signals than for harmonic signals is obtained.

4. METHODOLOGY

The Sequence of steps in the proposed method for detection of a single speaker or multi-speaker is shown in Table 1. The speech signal was first split up into non-overlapping short-term windows (frames) of 20 msec. length. The updated and calculated zero crossing rate for every entering sample, using M = 256 rectangular window samples corresponds to 20 msec. with 8 kHz sampling rate. For instance, the authors in [19] calculate 2 ZCRs, first one is for the far-field speech signal and the other is for the near-field speech signals that are then compared to the happening of multiple speakers. For a given input speech signal, the ZCR was calculated using equation (1). From Table 1, it was observed that if the input speech signal has a single speaker, the number of zero crossings is less. Similarly from Figure 2 and Figure 3, it was clear that if the input speech signal has multiple speakers, the number of zero crossings is high. In this study, we considered two speakers data as multiple speakers data for the analysis and evaluation of the proposed method. Based on this logic, we implemented Multi-Speaker activity (MSA) detection method shown in Table 1.

Table 1. Algorithm: Steps for incorporating multi-speaker activity detection

[1]. Record the Speech signal
[2]. Read the Normalization of the speech signal
[3]. Perform the segmentation (window) for the normalization of the speech signal
[4]. Compute the ZCR and slope
[5]. Assign the average Zero crossing (ZC) value is equal to the Zero crossing threshold(ZCTH) and average slope value as STH
[6]. If (ZC value < ZCTH) and (slope < STH) \ * ZCTH=500 and STH= 6 */
Mode= Single Speaker
else
Mode= Multi-Speaker

5. EXPERIMENTAL RESULTS

The performance of zero crossing rate methodology is decided by the parameters threshold (Th), K, and M. A window length M which is smaller contributes to less smoothening and there is every possibility to take incorrect decision. A larger window contributes to a smoothened ZCR and assists in overcoming any error as ZCR is nearer to the threshold, particularly in identifying the actual begin and end of regions. Experimental results proved that 256 samples sized window is more suitable. When ZCR gets updated the K parameter defines the number of samples. Once zero crossing rate is changed by every entering sample, usually K equal to 1 is expected. M bits are required, if K is equal to 1, to cache M zero crossing choices. Computational savings and memory are created when K exceeds 1 is selected. Lastly, the Th value is rigorously selected to prevent any warning or misdetection. Throughout this study, signals that are tested proved that, when window size M = 256 is more suitable for the threshold = 500.

Table 2. Number of zero crossings on number of speakers with gender

Number of speakers	ZCR
single male	398
single female	464
Two male (mixing of two signals)	763
Two female (mixing of two signals)	915
single male + single female (mixing of two signals)	837

The result of the study has been presented in Table 2. Here, the recognition rate is outlined as the ratio of the number of relevant speakers identified to the total number of speakers tested. The percentage of a correctly classified speech signal is given in Tables 3, 4, 5 and 6 respectively. Table 3 shows performance obtained when using the features for male speakers. Table 4 presents performance obtained when using the features for female speakers. Table 5 shows performance obtained when using features for mixed speakers. Finally, Table 6 shows that the performance comparison of the proposed ZCR method is compared with the cepstrum resynthesis residual magnitude (CRRM) method. Considering the three Tables, the differences in terms of recognition rate between the cross-testing are small, indicating that the classification scheme in use as a good generalization behavior. The results are also compared with CRRM of the speech signal. It is observed that slope with zero crossing rate (ZCR) outperforms the CRRM.

Table 3. Recognition rate for male speakers

No. of speakers	Recognition rate (%)
Single	85.51%

Multiple	87.72%

Table 4. Recognition rate for female speakers

No. of speakers	Recognition rate (%)
Single	86.38%
Multiple	89.95%

Table 5. Recognition performance for male-female (mixed) speakers

No. of speakers	Recognition rate (%)
Single	83.71%
Multiple	89.89%

Table 6. Comparing the Performance of proposed method with an existing method

Measure	Male	Female	Mixed
ZCR+Slope	86.61%	88.16%	86.6%
CRRM [1]	82.18%	83.75%	79.08%

6. CONCLUSIONS

In this work, we have presented a feature extraction methodology for slope at ZCR for multi-speaker activity detection and discussed with other existing feature extraction techniques. The zero crossing rates are very easy to calculate and proved that as an efficient discriminant of multi-speaker activity detection. The 3 parameters M, K, and Th are simple to modify and need a nominal tuning. It has also shown improved performance over the single speaker or multiple speakers activity detector. In this paper, finally, it is observed that multi-speaker activity detection using ZCR outperformed with the cepstrum resynthesis residual magnitude (CRRM) method. The results have been presented on the multi-speaker activity detection evaluation.

REFERENCES

[1] Rossignol S. and Pietquin O., (2010) "Single-speaker/Multi-speaker co-channel speech classification" In Proceedings of the International Conference on Speech Communication and Technologies, Makuhari (Japan), pp. 2322-2325.

[2] S. Maraboina, D. Kolossa, P. Bora and R. Orglmeister "Multi-speaker voice activity detection using ICA and beam pattern analysis" In Proceedings of the European signal processing conference, Florence, Italy, 2006.

[3] T. Pfau, D.P.W. Ellis, and A. Stolcke, (2001) "Multi-speaker speech activity detection for the ICSI meeting recorder" In Proceedings of IEEE ASRU Workshop, pp. 107–110.

[4] K. Laskowski, Q. Jin, and T. Schultz "Cross-correlation based multi-speaker speech activity detection" In Proceedings of INTERSPEECH-2004 (ICSLP), Jeju Island, Korea, 2004.
[5] AF Martin and MA. Przybocki, (2001) "Speaker Recognition in a Multi-Speaker Environment" In Proceedings of INTERSPEECH-2001, pp. 787-790.
[6] Rosenberg A, Magrin-Chagnolleau I., Parthasarathy S., and Huang Q. "Speaker detection in broadcast speech databases," In Proceedings of the International Conference on Spoken Language Processing, 1998.
[7] B. Yegnanarayana, R. Kumara Swamy and S.R.M Prasanna, (2005) "Separation of multi-speaker speech using excitation information" In Proceedings of ISCA workshop on NOLISP-2005, Spain, pp. 11-18.
[8] R. Kumara Swamy, K. Sri Rama Murty and B. Yegnanarayana, (2007) "Determining the number of speakers from multi-speaker speech signals using excitation source information" IEEE Signal Processing Letters, vol. 14, No. 7, pp. 481-484.
[9] B. Yegnanarayana, R. Kumara Swamy, and K. Sri Rama Murty "Determining mixing parameters from Multi-speaker data using speech-specific information" IEEE Transactions on Audio Speech and Language Processing, 2009.
[10] B. Yegnanarayana, S.R. Prasanna, K. Sreenivasa Rao "Speech enhancement using excitation source information" In Proceedings of the ICASSP, 2002.
[11] Bronkhorst and Adelbert W., (2000) "The Cocktail Party Phenomenon: A Review on Speech Intelligibility in Multiple-Talker Conditions" Acta Acustica united with Acustica, vol. 86, No. 1, pp. 117–128.
[12] T. Gansler, S.L. Gay, M.M. Sondhi, and J. Benesty, (2000) "Double-talk robust fast converging algorithms for network echo cancellation," IEEE Trans. Speech and Audio Processing, vol. 8, No. 6, pp. 656–663.
[13] J. Benesty, D. R. Morgan, and J. H. Cho, (2000) "A new class of double-talk detectors based on cross-Correlation " IEEE Trans. Speech and Audio Processing, vol. 8, No. 2, pp. 168–172.
[14] H. Buchner, J. Benesty, T. Gansler, and W. Kellermann, (2006) "Robust extended multi-delay filter and double-talk detector for acoustic echo cancellation " IEEE Trans. Audio, Speech, and Language Processing, vol. 14, No. 5, pp. 1633–1643.
[15] S. Y. Low, S. Venkatesh and S. Nordholm, (2012) "A spectral slit approach to doubletalk detection" IEEE Trans. Speech and Audio Processing, vol. 20, No. 3, pp. 1074–1080.
[16] C. Schuldt, F. Lindstrom and I. Claesson, (2012) "A delay-based double talk detector" IEEE Trans. Speech and Audio Processing, vol. 20, no. 6, pp. 1725–1733.
[17] R. Cheng, C. Zhang and C.Wei "Method and apparatus for double-talk detection" U.S. Patent 8 160 238 B2, Apr.17, 2012.
[18] Ikram, Muhammad Z. "Double-talk detection in acoustic echo cancellers using zero-crossings rate" ICASSP 2015.
[19] C. Paleologu, S. Ciochina and J. Benesty, (2008) "Double-talk robust VSS-NLMS algorithm for under-modeling acoustic echo cancellation " In Proceedings of IEEE Int. Conf. Acoust., Speech, Signal Processing, pp. 245–248.