Long-Term Spectral Pseudo-Entropy (LTSPE): A New Robust Feature for Speech Activity Detection

Mohammad Rasoul Kahrizi *
Department of Computer Engineering and Information Technology, Razi University, Kermanshah, Iran
mr.kahrizi@chmail.ir
Seyed Jahanshah Kabudian
Department of Computer Engineering and Information Technology, Razi University, Kermanshah, Iran
kabudian@razi.ac.ir

Received: 20/Dec/2018 Revised: 19/Apr/2019 Accepted: 29/May/2019

Abstract

Speech detection systems are known as a type of audio classifier systems which are used to recognize, detect, or mark parts of an audio signal including human speech. Applications of these types of systems include speech enhancement, noise cancellation, identification, reducing the size of audio signals in communication and storage, and many other applications. Here, a novel robust feature named Long-Term Spectral Pseudo-Entropy (LTSPE) is proposed to detect speech and its purpose is to improve performance in combination with other features, increase accuracy and to have acceptable performance. To this end, the proposed method is compared to other new and well-known methods of this context in two different conditions, with uses a speech enhancement algorithm to improve the quality of audio signals and without using speech enhancement. In this research, the MUSAN dataset is used, which includes a large number of audio signals in the form of music, speech, and noise. Also, various known methods of machine learning are used. As well as criteria for measuring accuracy and error in this paper are the criteria for F-Score and Equal-Error Rate (EER), respectively. Experimental results on MUSAN dataset show that if the proposed feature LTSPE is combined with other features, the performance of the detector is improved. Moreover, the proposed feature has higher accuracy and lower error compared to similar ones.

Keywords: Audio Signal Processing; Speech Processing; Speech Activity Detection (SAD); Speech Recognition; Voice Activity Detection (VAD); Robust Feature; LTSPE.

1. Introduction

One of the most critical issues in audio signal processing is processing audio signals in which there is a combination of human speech with other sounds like various types of noises, animals' sound and various sounds of different environments. For example, audio signals recorded from speeches, radio, TV, and satellite or different conversations can be mentioned. These audio signals include various types of speech sound signals.

In some applications like the file size reduction, quality enhancement [2-4], compression, bandwidth usage optimization, detection & identification [5-11], and other applications [12-16], it is needed to detect human speech or remove silence and environmental noises from human speech. Speech detection systems are known as a type of audio signal classifier systems which are used to separate, detect, or mark parts of an audio signal which includes human speech.

2. Literature Review

In this section, some of the methods and features for speech detection are mentioned, which are well-known and applicable in speech processing context, and are used here to compare their performances with the proposed method.

One of the most popular and oldest features is Mel-Frequency Cepstral Coefficients (MFCC) [17].

Long-Term Signal Variability (LTSV) and LTSVG (LTSV Gammatone) features [18] are other features which are used for comparison. The LTSV for each frame of audio signals is equal to the entropy variance of each of the frequency bins in that frame. Also, the initial idea of LTSPE feature is inspired by this algorithm.

Another method is Multi-Band Long-Term Signal Variability (MBLTSV) [19], which is a type of LTSV in which frequency scale is warped [20]. The spectrum is divided into predetermined parts which are known as bands, and LTSV is applied to each band. This process improves MBLTSV significantly. Another feature which is used in this research is Long-Term Spectrum Divergence (LTSD) [21].

Other new features and methods that have been employed in audio signal processing context are also used. One of these methods is the method has been proposed by Sadjadi [22] which includes four features called Harmonicity (a.k.a. harmonics-to-noise ratio), Clarity, linear prediction (LP) error and harmonic product spectrum (HPS). And the other is a method that has been proposed by Drugman [23] which includes three features called Cepstral Peak Prominence (CPP), Summation of the Residual Harmonics (SRH) and SRH*. Readers are encouraged to refer to the references for more details.

* Corresponding Author
ORCiD: 0000-0001-6954-1452
The source code of the LTSPE has been registered as ”Long-Term Spectral Pseudo-Entropy (LTSPE) Feature” and DOI ”10.21227/H2G05K” on IEEE Dataport [1].

3. Long-Term Spectral Pseudo-Entropy (LTSPE)

The purpose here is to introduce a new feature called long-term spectral pseudo-entropy to recognize and detect speech. Due to long-term characteristics and noise resistance, this feature can be recognized as a robust feature. LTSPE method is inspired by LTSV [18]. The main difference of the proposed method with LTSV is that the proposed method calculates entropy along the frequency axis, but in LTSV, entropy is calculated along the time axis. Other differences between the proposed method and the similar methods are the differences in the initialization of parameters and measurement intervals. Also, the reasons for naming the proposed method to pseudo-entropy are the same differences with the usual entropy method. In principle, the LTSPE calculates a modified long-term entropy from the frequency spectrum of audio signals.

After dividing the audio signal into frames with a predetermined size, Fourier transform of each frame is calculated, and the power spectrum of each frame is obtained, then these spectra are normalized. This process can be seen in Equations 1 to 4.

$$X(k,n)=\sum_{l=0}^{N_w-1} w(l)\; x(l+(n-1).N_{sh})\; e^{-j\frac{2\pi kl}{N_w}} \tag{1}$$

Where *w(l)* is window function and *X(k, n)* is a short time Fourier transform (STFT) for the n^{th} frame and k^{th} frequency bin. Moreover, N_w is the number of samples per window, and N_{sh} is the number of samples that are considered for frameshift.

$$S_x(k,n)=\left|X(k,n)\right|^2 \tag{2}$$

Where S_x shows the power spectrum.

$$S_M(f,n)=\frac{1}{M+1}\sum_{k=f-\frac{M}{2}}^{f+\frac{M}{2}} S_x(k,n) \tag{3}$$

$$S_R(f,n)=\frac{S_M(f,n)}{\sum_{k=f-\frac{R}{2}}^{f+\frac{R}{2}} S_M(k,n)} \tag{4}$$

Where S_M is the smoothed power spectrum, and S_R is normalized smoothed power spectrum. *k* shows frequency bin and *n* is the frame index. *M* and *R* are even and positive numbers which specify smoothing and normalization intervals for the power spectrum of each frame.

After all these steps, it is time to calculate entropy for each frame. Eq. 5 shows how entropy is calculated for each frame.

$$\xi(n)=-\sum_{k=1}^{K}\left(S_R(k,n)\log S_R(k,n)\right) \tag{5}$$

Finally, the variance of obtained entropies in a predetermined interval is calculated as in Eq. 6.

$$LTSPE(n)=\frac{1}{R_2+1}\sum_{m=n-R_2/2}^{n+R_2/2}\left(\xi(m)-\overline{\xi(n)}\right)^2 \tag{6}$$

Where R_2 specifies the variance interval, also mean entropy $\overline{\xi(n)}$ is calculated as in Eq. 7.

$$\overline{\xi(n)}=\frac{1}{R_2+1}\sum_{m=n-R_2/2}^{n+R_2/2}\xi(m) \tag{7}$$

As already mentioned, the main idea of LTSPE is inspired by the LTSV method. To better compare the LTSV with LTSPE, the calculation method of the LTSV is shown in equations 8 & 9.

$$LTSV_x(m)=\frac{1}{K}\sum_{k=1}^{K}\left(\xi_k^x(m)-\frac{1}{K}\sum_{k=1}^{K}\xi_k^x(m)\right)^2 \tag{8}$$

$$\xi_k^x(m)=-\sum_{n=m-R+1}^{m}\left(\frac{\left|X(n,\omega_k)\right|^2}{\sum_{l=m-R+1}^{m}\left|X(l,\omega_k)\right|^2} *\log\left(\frac{\left|X(n,\omega_k)\right|^2}{\sum_{l=m-R+1}^{m}\left|X(l,\omega_k)\right|^2}\right)\right) \tag{9}$$

For further details of these equations, and more familiarity with LTSV, refer to [18,19].

4. Evaluation and Results

In order to evaluate performance and to obtain higher accuracy and less error in the speech detection process, the proposed method is compared with some of the robust and well-known features in speech detection context which were introduced in Section 2.

MUSAN corpus [24] is used for evaluation. K-Nearest Neighbors (KNN) and Gaussian Mixture Models (GMM) are used for classification.

In GMM method, 16 components are used to model each of the classes, and the number of repetitions of the EM algorithm is set to 100, and also the variance floor is set to 0.01.

Mahalanobis distance is used in the KNN method, and also one neighbor (k = 1) is considered for the number of neighbors in this classifier. The reason for this choice is the higher accuracy in the experimental evaluation results. Moreover, in the reference [25], it has been stated that in the case where the number of training data goes to infinity, it is guaranteed in the classification with 1-NN (k = 1) that the probability of classification error is less than twice of the probability of Bayes error. And to the other word, the probability of classification error is less than twice the probability of optimal error.

Also, to improve quality and decrease the noise of audio signals, Optimally Modified Log-Spectral Amplitude (OMLSA) speech enhancement algorithm [26] is used to improve the quality of speech signals. Of course, in principle, OM-LSA is a speech estimator for non-stationary noise environments that it is employed in this paper.

For experiments, speech signals are selected from MUSAN corpus. The selected speech signals are used in experiments in two conditions: with OMLSA speech enhancement and without speech enhancement. Also, to train non-speech (silence & noise) class, which is here the

second class in classification algorithms, all noise and non-speech signals of the MUSAN corpus are used. Different feature extraction algorithms are employed. The obtained features are used to train speech and non-speech classes by GMM & KNN. Frame size and frameshift are 25 & 12.5 ms, respectively. Parameters *M*, *R*, and R_2 are set to 24, 16, and 60. 13-dimensional Mel-Frequency Cepstral Coefficients are extracted from audio signals (c_0-c_{12} are extracted).

Finally, after creating the training models which are obtained by using KNN and GMM, the error of each method and classification accuracy are evaluated using EER (Equal Error Rate) and F-Score criteria in a 10-fold mode. Equations 10 and 11 show how these criteria are calculated. It is important to note that the values of all parameters and intervals are determined experimentally and in the optimum mode. Also, all the results shown in the tables and figures are obtained by the authors with implementing the methods in the MATLAB application.

$$EER = 1 - \frac{|True\ Accepted\ Frames|}{|Total\ Frames|} \tag{10}$$

Where the *EER* value is between 0 and 1. when the value of *EER* is zero, the best performance and the highest accuracy are achieved, of course, it should be noted, when the false acceptance rate (FAR) is as high as possible equal to the false rejection rate (FRR), this equation shows the EER.

$$F-Score = \frac{2*Precision*Recall}{Precision+Recall} \tag{11}$$

Where *F-Score* has a value between 0 and 1, and it is a kind of harmonic average of *Precision* and *Recall* criteria, and when the value of this criterion is 1, best performance and the highest accuracy are earned. Moreover, according to [27], the calculation method of precision and recall is shown in Fig. 1.

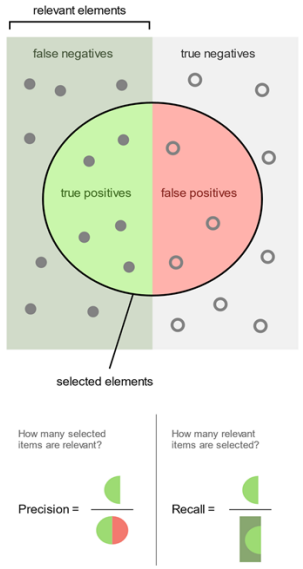

Fig. 1. Method of calculating the precision and recall criteria[1].

1. Available: https://en.wikipedia.org/w/index.php?title=Precision_and_recall&oldid=896900450. Accessed: May. 25, 2019.

Obtained results are presented in Tables 1 and 2 and Figures 2 and 3. As can be seen in results, by considering ERR and F-Score, when the methods in non-combinational mode are used, the best performance is obtained with MBLTSV. However, if other methods and features are used in combinational mode, better options would be obtained like combining LTSPE and MFCC or fusion LTSPE and MBLSTV which give better results compared to when the methods are used non-combinational.

It can be understood from results, that enhancing quality and reducing the noise of speech signals (speech enhancement) before speech detection, has opposite effect unexpectedly and decreases the accuracy of the speech detection system and increases classification error. This result must be accurate and acceptable because some details of the audio signals in the speech enhancement process are eliminated, which reduce the accuracy of the speech activity detection. Also, it should be considered that in this research, the speech enhancement process is entirely separate from speech activity detection.

As it turns out from figures and tables, all in all, the LTSPE has the best results in comparison with similar features such as LTSV, LTSD, and LTSVG. Of course, it should be noted that the similarity of these features is comparable in that the output of these features is only one number per frame.

Table 1. Comparing methods without applying OMLSA (%)

Features	KNN				GMM			
	EER	F-Score			EER	F-Score		
		Overall	Speech	Non-speech		Overall	Speech	Non-speech
Drugman [23]	15.91	63.45	35.97	90.92	21.53	61.90	36.78	87.02
LTSD [21]	15.94	63.49	36.10	90.89	23.55	66.57	48.40	84.74
LTSV [18]	17.79	59.62	29.42	89.82	16.44	66.47	42.53	90.41
LTSVG	15.83	64.09	37.25	90.94	12.56	73.74	54.77	92.71
Proposed LTSPE	14.95	66.04	40.63	91.45	17.16	72.56	55.76	89.36
Sadjadi [22]	18.06	59.32	28.98	89.66	32.72	57.41	36.92	77.91
MBLTSV [19]	3.86	90.89	83.99	97.80	11.72	70.37	47.33	93.40
MBLTSV+Proposed LTSPE	**2.40**	**94.50**	**90.36**	**98.63**	13.29	76.66	61.34	91.97
MFCC [17]	5.41	86.99	77.05	96.93	8.65	83.39	71.90	94.88
MFCC+Proposed LTSPE	3.90	90.78	83.77	97.79	**7.29**	**86.04**	**76.40**	**95.69**
MFCC+LTSV	5.06	87.90	78.67	97.13	9.33	82.72	71.00	94.44

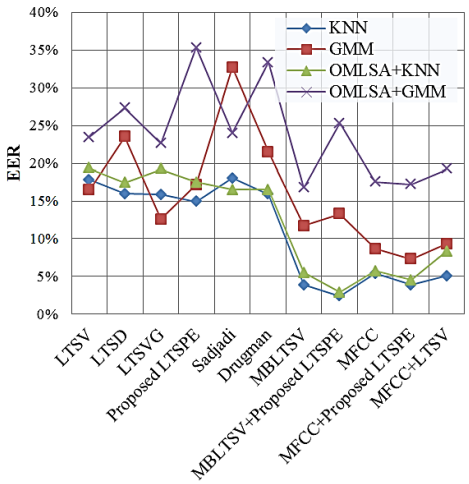

Fig. 2. Comparing classification error in terms of EER in different conditions using different methods (%)

Table 2. Comparing methods by applying OMLSA (%)

Features	KNN				GMM			
	EER	F-Score			EER	F-Score		
		Overall	Speech	Non-speech		Overall	Speech	Non-speech
Drugman [23]	16.51	62.27	33.97	90.57	33.38	55.43	33.04	77.81
LTSD [21]	17.43	60.56	31.09	90.02	27.32	63.03	44.16	81.91
LTSV [18]	19.40	55.82	22.74	88.91	23.50	58.99	32.19	85.79
LTSVG	19.27	56.43	23.89	88.97	22.68	63.30	40.61	85.99
Proposed LTSPE	17.48	60.48	30.97	89.99	35.36	65.87	56.36	75.37
Sadjadi [22]	16.48	62.52	34.47	90.58	24.03	64.10	43.44	84.76
MBLTSV [19]	5.51	87.11	77.35	96.86	**16.79**	63.38	36.44	90.32
MBLTSV+Proposed LTSPE	**2.91**	**93.42**	**88.49**	**98.34**	25.32	63.74	43.50	83.97
MFCC [17]	5.71	87.26	77.79	96.72	17.53	81.86	69.78	93.93
MFCC+Proposed LTSPE	4.51	89.87	82.19	97.55	17.20	**82.76**	**71.39**	**94.13**
MFCC+LTSV	8.32	80.50	65.39	95.61	19.28	79.78	66.89	92.68

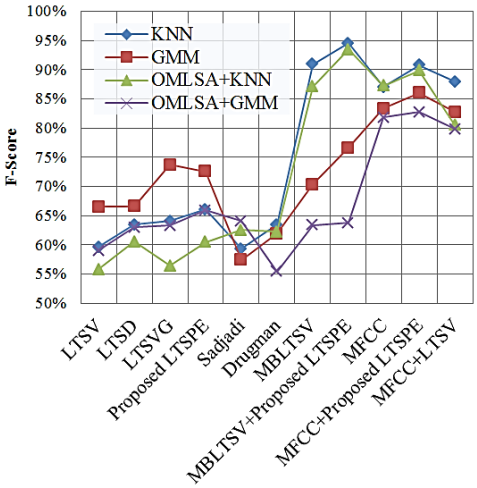

Fig. 3. Comparing classification accuracy in terms of F-Score using different methods (%)

5. Conclusion

The goal of this paper was to introduce a new feature for speech activity detection with the name, long-term spectral pseudo entropy (LTSPE). The LTSPE has been inspired by the LTSV method, and by making some changes to it has been proposed. One of the strengths of the proposed method is that, when combined with other methods, it improves the performance of those methods, even the methods with the highest accuracy, like MBLTSV and MFCC in evaluations. It should be mentioned that comparison in evaluations has been performed only in terms of accuracy and if the size of output data, processing time and calculations of the methods were considered for comparison, results would be different. For example, only four methods have been used in this research where their outputs, feature matrixes, have one column for each audio signal, i.e., four of the mentioned methods give a scalar for each input frame of audio signals, which indicates the smaller size of output data, fewer calculations and faster processing time. These four methods consisted of LTSV, LTSVG, LTSD, and LTSPE, that among them, and on average in all investigated moods, the proposed method LTSPE has had higher accuracy.

References

[1] M. R. Kahrizi, "Long-Term Spectral Pseudo-Entropy (LTSPE) Feature," IEEE Dataport, 2017. [Online]. Available: http://dx.doi.org/10.21227/H2G05K. Accessed: May. 27, 2019.

[2] W. Wang, H. Liu, J. Yang, G. Cao, and C. Hua, "Speech enhancement based on noise classification and deep neural network," Modern Physics Letters B, p. 1950188, 2019.

[3] H. Wang, Z. Ye, and J. Chen, "A Front-End Speech Enhancement System for Robust Automotive Speech Recognition," in 2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP), 2018, pp. 1-5: IEEE.

[4] K. Dinesh, R. Prakash, and M. P. Madhan, "Real-time Multi-Source Speech Enhancement for Voice Personal Assistant by using Linear Array Microphone based on Spatial Signal Processing," in 2019 International Conference on Communication and Signal Processing (ICCSP), 2019, pp. 0965-0967: IEEE.

[5] I. Ariav, D. Dov, and I. Cohen, "A deep architecture for audio-visual voice activity detection in the presence of transients," Signal Processing, vol. 142, pp. 69-74, 2018.

[6] M. Pal, D. Paul, and G. Saha, "Synthetic speech detection using fundamental frequency variation and spectral features language," Computer Speech, vol. 48, pp. 31-50, 2018.

[7] B. Mouaz, B. H. Abderrahim, and E. Abdelmajid, "Speech Recognition of Moroccan Dialect Using Hidden Markov Models," Procedia Computer Science, vol. 151, pp. 985-991, 2019.

[8] H. Chen, "Speaker Identification: Time-Frequency Analysis With Deep Learning," ETD Collection for Tennessee State University, 2018.

[9] P. Vecchiotti, G. Pepe, E. Principi, and S. Squartini, "Detection of activity and position of speakers by using Deep Neural Networks and Acoustic Data Augmentation," Expert Systems with Applications, 2019.

[10] F. Tao, "Advances in Audiovisual Speech Processing for Robust Voice Activity Detection and Automatic Speech Recognition," 2018.

[11] A. Ivry, B. Berdugo, and I. Cohen, "Voice Activity Detection for Transient Noisy Environment Based on Diffusion Nets," IEEE Journal of Selected Topics in Signal Processing, 2019.

[12] V. Andrei, H. Cucu, and C. Burileanu, "Overlapped speech detection and competing speaker counting-humans vs. deep learning," IEEE Journal of Selected Topics in Signal Processing, 2019.

[13] C. Vikram, N. Adiga, and S. M. Prasanna, "Detection of Nasalized Voiced Stops in Cleft Palate Speech Using Epoch-Synchronous Features," IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2019.

[14] N. Mansour, M. Marschall, T. May, A. Westermann, and T. Dau, "A method for conversational signal-to-noise ratio estimation in real-world sound scenarios," The Journal of

the Acoustical Society of America, vol. 145, no. 3, pp. 1873-1873, 2019.
[15] M. Pandharipande, R. Chakraborty, A. Panda, and S. K. Kopparapu, "Robust Front-End Processing For Emotion Recognition In Noisy Speech," in 2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP), 2018, pp. 324-328: IEEE.
[16] Y. Malviya, S. Kaul, and K. Goyal, "Music Speech Discrimination," CS229 Final Project, 2016.
[17] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," Speech, and Signal Processing, vol. 28, no. 4, pp. 357-366, 1980.
[18] P. K. Ghosh, A. Tsiartas, and S. Narayanan, "Robust voice activity detection using long-term signal variability," IEEE Transactions on Audio, Speech, and Language Processing, vol. 19, no. 3, pp. 600-613, 2011.
[19] A. Tsiartas et al., "Multi-band long-term signal variability features for robust voice activity detection," in Interspeech, 2013, pp. 718-722.
[20] A. Makur and S. K. Mitra, "Warped discrete-Fourier transform: Theory and applications," IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications, vol. 48, no. 9, pp. 1086-1093, 2001.
[21] J. Ramırez, J. C. Segura, C. Benıtez, A. De La Torre, and A. Rubio, "Efficient voice activity detection algorithms using long-term speech information," Speech Communication, vol. 42, no. 3-4, pp. 271-287, 2004.
[22] S. O. Sadjadi and J. H. Hansen, "Unsupervised speech activity detection using voicing measures and perceptual spectral flux," IEEE Signal Processing Letters, vol. 20, no. 3, pp. 197-200, 2013.
[23] T. Drugman, Y. Stylianou, Y. Kida, and M. Akamine, "Voice activity detection: Merging source and filter-based information," IEEE Signal Processing Letters, vol. 23, no. 2, pp. 252-256, 2016.
[24] D. Snyder, G. Chen, and D. Povey, "MUSAN: A Music, Speech, and Noise Corpus," CoRR, vol. abs/1510.08484, 2015.
[25] T. Cover and P. Hart, "Nearest neighbor pattern classification," IEEE Transactions on Information Theory, vol. 13, no. 1, pp. 21-27, 1967.
[26] I. Cohen and B. Berdugo, "Speech enhancement for non-stationary noise environments," Signal Processing, vol. 81, no. 11, pp. 2403-2418, 2001.
[27] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, 2016.

Mohammad Rasoul Kahrizi received the M.Sc. degree in software engineering from the Department of Computer Engineering and Information Technology, Faculty of Engineering, Razi University, Kermanshah, Iran, in 2017. His research interests are in programming, optimization algorithms, data science, signal processing, and machine learning.

Seyed Jahanshah Kabudian is currently an assistant professor at the Department of Computer Engineering and Information Technology in Razi University, Kermanshah, Iran. He received the Ph.D. degree in artificial intelligence from the Amir Kabir University of Technology, Tehran, Iran, in 2010. His research interests lie in signal processing, speech processing, natural language processing, pattern recognition, and machine learning.