Contents lists available at Mapta Publishing Group

Mapta Journal of Electrical and Computer Engineering (MJECE)

journal homepage: http://maptapublishing.com/index.php/mjece

Vol. 1, No. 2, August, 2018

ISSN: 2631-5718

Speech Detection Methods Combination in Noisy and Clean Environments

Sina Mojtahedi*

Biomedical Science and Engineering Department, Faculty of Engineering, KOÇ University, Istanbul, Turkey

Article Info

Article history:

Received Jul 1st, 2018
Revised Aug 15th, 2018
Accepted Sep 13th, 2018

Keyword:

Signal Processing
Voice Activity Detector
Speech Detector
Noise
Noisy Environment

Abstract

Voice Activity Detector is a branch of signal processing science and is one of the most important areas of audio signal processors that are used in many telecommunication and audio systems such as speech recognition, speech compression, noise estimation and removal in speech enhancement system, wireless communication systems, and many other communication systems. We combine a number of relatively new methods for speech recognition in this paper. The experiments performed on clean and noisy data (with moderate noise) show that the proposed method reduces the EER error rate to the best of the other methods examined in this paper by about 85.3%.

1. Introduction

In speech processing to recognize speech or the speaker, in the real world, we need to process the speech impacted by various environmental noises. Not using the Voice Activity Detection system makes word recognition more complicated and increases the speech recognition errors. In many speech signal processing applications, Voice Activity Detection (VAD) plays an essential role in separating a Voice stream with intervals of speech and non-speech activity, and is one of the most useful tools for improving and coding speech in speech coding fields [1], Automatic speech recognition [2], speech enhancement, or speaker and language recognition [3]. Speech recognition from silence also plays a very important role in speech enhancement systems, because in these types of systems it is necessary to obtain an example of pure noise found in silent regions. In speech communication, speech recognition from silence also plays an important role, since a significant part of the conversation is silence, and if it is possible to detect silent areas, there is no need to send signals related to silent areas, thus channel dedicated bit rate will be saved.

In the second section of the paper, we will briefly review the papers presented in this field. In the third section, we will first explain the proposed method and then our simulations, and in the fourth section we will examine and compare the proposed method with other available results and, in accordance with the actual data, we will review its qualitative details. As last section, we will discuss our conclusions and suggestions.

2. A review of previous studies

The beginning of the VAD studies was the first attempt at word recognition systems in 1970. At that time, simple features such as energy, the period of the signal [4], Autocorrelation [5], and Zero pass rate [6, 7] were investigated for VAD. Over the next decades, the use of complex features to achieve proper detection increased, and the results were further challenged in noisy conditions, and speech features such as spectral form [8] and harmonic structure of speech were investigated. One of the constraints of many of the early VAD algorithms was that they only compute the internal of the current frame. Ramirez et al showed the benefits of recognition based on long-term information about the speech signal. By extending the time span of the data used in the decision making, long-term features of speech, such as stability degree, were obtained [9]. Then modulation was identified as an important aspect in human perception of speech

* Corresponding author: sna.mojtahedi@gmail.com

recognition. Other features, such as time spectrograph modulation or amplitude modulation spectrum, reflect the presence of speech in a manner similar to human perception. The sub-band information [10] is another method used in VADs and examines the energy of the frame under various sub-bands that discuss multi-band performance of the features. Other features that are used in this area are Teager Energy, which is based on a nonlinear operation on the signal [11], modified group delay, which has been obtained by using the information phase against the net range of information given by the power spectrum [12] and the spectral flux examines the local spectrum differences between two adjacent frames [13]. On the other hand, the research carried out in this regard suggests that, due to the diversity of speech features, a combination of complementary features may also be desirable in practice [14].

3. Suggested method

In addition to existing basic methods, new methods utilize the fact that speech and noise signals must have different properties. Some studies have used a combination of several features. In combination methods, two or more features are defined at the window level for decision making. Many studies have shown how different ideas can increase the efficiency of speech processing methods.

In this paper, we also found better results by combining the filtering and source based features in the Krugman method [15] with the characteristics of long-term PTSD [16], LTSV [17] and MTV [18] according to the block diagram below, than the existing methods.

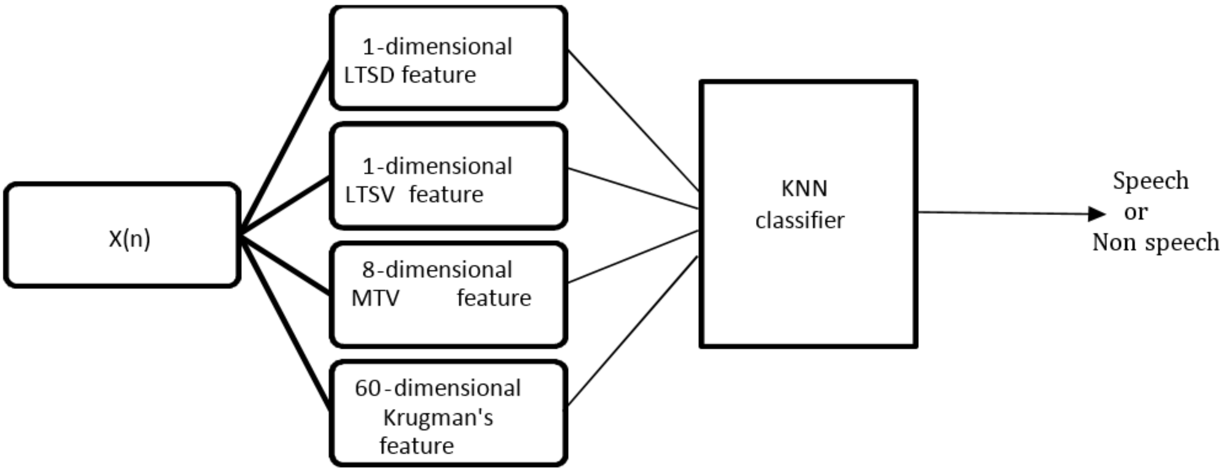
1-dimensional LTSD feature
1-dimensional LTSV feature
8-dimensional MTV feature
60-dimensional Krugman's feature
X(n)
KNN classifier
Speech or Non speech

Figure 1. Block diagram of the suggested combination method for speech detection

LTSD and LTSV each have one feature, MTV includes eight features in eight bands and Drugman includes 13 MFCC, 4 Sajjadi and 3 Drugman suggested features and their first and second derivatives. Then we have a total of a 70-character attribute vector and used to classify the attribute vector derived from the KNN category.

4. Performed Tests

4.1. Database

Speech data files contain 30 files from the TIMIT data files that are in a clean, noiseless mode and in the noisy state with seven selected noises: Volvo, pink, factory2, f16, buccaneer2, babble and white from Noisex data noises is accumulated at 10 dB and 20 dB of SNRs. Non-speech data includes various types of environmental sounds such as market, kitchen, street and station sounds, provided by Guoning Hu from Ohio university.

4.2. Evaluation Criteria

Usually, the performance of the VAD algorithm is evaluated by the Receiver Operating Characteristic (ROC) system. For this, the possibility of True Speech Detection versus the possibility of False Speech Detection for the different threshold values is plotted. To express the curve to a single value, the Area Under Curve ROC (AUC) is calculated. Based on this, the efficiency of the algorithm is comparable to other methods. Another criterion is the so-called Equal Error Rate (EER), which is related to the area where the False Alarm rate and the False reject rate are equal. The lesser the EER and the higher AUC, make the proposed method a better method and, given these quantities, the proposed method has achieved better results compared to other existing methods.

4.3. Test Results

The experiments are conducted in clean and noisy conditions and it is observed that the EER error criterion is less than other methods and the AUC is more accurate than other methods.

Table 1. The output of the average efficiency of algorithms in terms of EER error rate and AUC accuracy rate in a noisy environment with a signal-to-noise ratio of 20 dB with different noises

Method	AUC (%)	EER (%)
Proposed Method	97.48	2.23
Drugman	97.41	2.31
LTSD	62.74	33.80
LTSV	69.01	30.07
MBLTSV	97.12	2.79
Sohn	55.80	44.19

Table 2.The Average efficiency of algorithms in terms of EER error rate and AUC accuracy rate in a noisy environment with a signal-to-noise ratio of 10 dB with different noises

Method	AUC (%)	EER (%)
Proposed Method	95.35	4.13
Drugman	95.24	4.25
LTSD	55.45	40.40
LTSV	66.17	32.83
MBLTSV	95.34	4.54
Sohn	54.18	45.81

Table 3. The Average efficiency of algorithms in terms of EER error rate and AUC accuracy rate in a clean environment.

Method	AUC (%)	EER (%)
Proposed Method	96.50	3.36
Drugman	96.29	3.57
LTSD	78.79	20.18
LTSV	72.79	26.78
MBLTSV	96.27	3.70
Sohn	65.21	34.79

Table 4.The Average efficiency of algorithms in terms of EER error rate and AUC accuracy rate for different noises and different SNRs

Method	AUC (%)	EER (%)
Proposed Method	96.44	3.24
Drugman	96.31	3.37
LTSD	65.66	31.46
LTSV	69.32	29.89
MBLTSV	96.24	3.67
Sohn	58.39	41.59

As can be seen, the proposed method increases the accuracy of speech detection in clean and medium noisy environments.

5. Conclusion

In recent years, the combination of basic methods approach because of its comprehensiveness has begun to improve VAD and it seems to be possible to achieve better results in comparison with other methods and approaches, and in this area, many ideas can be found with desirable goals in it. Based on this, we chose this approach for this article and after the implementation, we came to the conclusion that the proposed method, despite the computational complexity, has high accuracy and for the clean environment and high and medium SNRs that most systems work in this SNR, has achieved better results. And we were able to combine a number of new methods in this area, using clean and noisy tested data, to relatively improve the EER and AUC by 85.3% and 13.0%, respectively compared to the best method from other methods.

References

[1] A. Benyassine, E. Shlomot, H.-Y. Su, D. Massaloux, C. Lamblin, J.-P. Petit. ITU-T Recommendation G. 729 Annex B: a silence compression scheme for use with G. 729 optimized for V. 70 digital simultaneous voice and data applications. IEEE Communications Magazine. 35 (1997) 64-73.

[2] D. Vlaj, B. Kotnik, B. Horvat, Z. Kačič. A computationally efficient mel-filter bank VAD algorithm for distributed speech recognition systems. EURASIP Journal on Advances in Signal Processing. 2005 (2005) 561951.

[3] I. McCowan, D. Dean, M. McLaren, R. Vogt, S. Sridharan. The delta-phase spectrum with application to voice activity detection and speaker recognition. IEEE Transactions on Audio, Speech, and Language Processing. 19 (2011) 2026-38.

[4] S.O. Sadjadi, J.H. Hansen. Unsupervised speech activity detection using voicing measures and perceptual spectral flux. IEEE Signal Processing Letters. 20 (2013) 197-200.

[5] T. Kristjansson, S. Deligne, P. Olsen. Voicing features for robust speech detection. Ninth European Conference on Speech Communication and Technology2005.

[6] B. Kotnik, Z. Kacic, B. Horvat. A multiconditional robust front-end feature extraction with a noise reduction procedure based on improved spectral subtraction algorithm. Seventh European Conference on Speech Communication and Technology2001.

[7] L. Lamel, L. Rabiner, A. Rosenberg, J. Wilpon. An improved endpoint detector for isolated word recognition. IEEE Transactions on Acoustics, Speech, and Signal Processing. 29 (1981) 777-85.

[8] L. Rabiner, M. Sambur. Application of an LPC distance measure to the voiced-unvoiced-silence detection problem. IEEE Transactions on Acoustics, Speech, and Signal Processing. 25 (1977) 338-43.

[9] S. Graf, T. Herbig, M. Buck, G. Schmidt. Features for voice activity detection: a comparative analysis. EURASIP Journal on Advances in Signal Processing. 2015 (2015) 91.

[10] M. Van Segbroeck, A. Tsiartas, S. Narayanan. A robust frontend for VAD: exploiting contextual, discriminative and spectral cues of human voice. INTERSPEECH2013. pp. 704-8.

[11] Y. Ephraim, D. Malah. Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. IEEE Transactions on acoustics, speech, and signal processing. 32 (1984) 1109-21.

[12] G. Ying, C. Mitchell, L. Jamieson. Endpoint detection of isolated utterances based on a modified Teager energy measurement. 1993 IEEE International Conference on Acoustics, Speech, and Signal Processing. IEEE1993. pp. 732-5.

[13] R.M. Hegde, H.A. Murthy, V. Gadde. The modified group delay feature: a new spectral representation of speech. Proceedings of 8th International Conference on Spoken Language Processing (INTERSPEECH'04)2004. pp. 913-6.

[14] N. Cho, E.-K. Kim. Enhanced voice activity detection using acoustic event detection and classification. IEEE Transactions on Consumer Electronics. 57 (2011) 196-202.

[15] T. Drugman, Y. Stylianou, Y. Kida, M. Akamine. Voice activity detection: Merging source and filter-based information. IEEE Signal Processing Letters. 23 (2016) 252-6.

[16] J. Ramırez, J.C. Segura, C. Benıtez, A. De La Torre, A. Rubio. Efficient voice activity detection algorithms using long-term speech information. Speech communication. 42 (2004) 271-87.

[17] A.C.D.-o. Soil, Rock. Standard test methods for one-dimensional consolidation properties of soils using incremental loading. ASTM International2004.

[18] A. Tsiartas, T. Chaspari, N. Katsamanis, P.K. Ghosh, M. Li, M. Van Segbroeck, et al. Multi-band long-term signal variability features for robust voice activity detection. Interspeech2013. pp. 718-22.